Jahrbücher für Nationalökonomie und Statistik manuscript No.
(will be inserted by the editor)

OLS and 2SLS in Randomized and Conditionally Randomized Experiments

Jason Ansel, Han Hong, and Jessie Li

C1, C8, C9

Big Data, data science

Abstract. We investigate estimation and inference of the (local) average treatment effect parameter when a binary instrumental variable is generated by a randomized or conditionally randomized experiment. Under i.i.d. sampling, we show that adding covariates and their interactions with the instrument will weakly improve estimation precision of the (local) average treatment effect, but the robust OLS (2SLS) standard errors will no longer be valid. We provide an analytic correction that is easy to implement and demonstrate through monte carlo simulations and an empirical application the interacted estimator's efficiency gains over the unadjusted estimator and the uninteracted covariate adjusted estimator. We also generalize our results to covariate adaptive randomization where the treatment assignment is not i.i.d., thus extending the recent contributions of Bugni et al. (2017a) and Bugni et al. (2017b) to allow for the case of non-compliance. [1]

1 Introduction

With the advent of the internet and large online datasets, randomized and conditionally randomized experiments are becoming increasingly common. Despite the vast literature on treatment effect analysis under conditional independence and monotonicity assumptions, the surge of recent interest in these experiments (for example, Lin (2013); Freedman (2008);Bugni et al. (2017a);Bugni et al. (2017b)) suggests a need to clarify the relationship to the previous literature and for understanding new results that are relevant in these settings.

This paper contributes to the vast literature on treatment effect analysis (Rosenbaum and Rubin (1983), Imbens and Angrist (1994), among others) by examining the role

[1] This paper was inspired by a discussion with Patrick Kline, who also provided key references. We thank Joe Romano for very helpful discussions, and particular the editors and referee for insightful comments and constructive suggestions. This research was supported by a Faculty Research Grant awarded by the Committee on Research from the University of California, Santa Cruz, the National Science Foundation (SES 1658950), and SIEPR. Correspondence can be sent to jeqli@ucsc.edu.

that covariates play in improving the efficiency of treatment effect estimates. Under the conventional assumption of independent and identically-distributed (i.i.d.) sampling from an infinite population, we provide an extensive analysis of the efficiency gains from including covariates interacted with the binary instrument Z for the treatment. The efficiency analysis applies under both full compliance and partial compliance. The full compliance case, where the treatment indicator D coincides with the binary instrument of treatment eligibility Z, is governed by a model of conditional independence (unconfoundedness) in Rosenbaum and Rubin (1983). The partial compliance case, where D may not be equal to Z, is governed by the local average treatment effect (LATE) model of Imbens and Angrist (1994). Because the LATE model is more general, most of the results in this paper are presented under the partial compliance LATE model. Results for the full compliance model under unconfoundedness are special cases of the LATE model when $D = Z$.

Under the i.i.d. sampling framework, we find that including additional covariates X and their interactions with Z in a 2SLS regression of the outcome Y on the treatment indicator D will weakly improve efficiency of the estimator for the LATE parameter. Including the covariates X only, without interacting with the instrument Z, might improve or reduce efficiency. The intuition for the efficiency gains from including the interaction term comes from examining its relationship with the sieve estimator for ATE as discussed in Chen, Hong and Tarozzi (CHT 2008) and for LATE as discussed in Frolich (2006). If X were replaced with a sieve expansion then the interacted estimator coincides with the sieve estimator.

We also extend our efficiency comparisons to covariate adaptive randomization schemes. Following Bugni et al. (2017a) (BCS 2017a) and Bugni et al. (2017b) (BCS 2017b), we admit strata-level differences in the randomization scheme, perform efficiency comparisons, and extend their results to allow for noncompliance. In general, the interacted estimator is more efficient than the unadjusted and uninteracted estimators, except when the sampling scheme exhibits strong balance, such as in stratified block randomization, in which case the interacted estimator is equally efficient as the uninteracted covariate adjusted estimator.

In addition to providing monte carlo simulations comparing the standard errors of the various estimators, we also include an empirical application. The data come from GoDaddy, a domain name registrar responsible for managing sales of internet domain names through a variety of formats such as auctions and direct negotiation between buyers and sellers. We observe a sample of auctions which underwent a simple randomized experiment in which some auctions were assigned a valuation determined by a machine learning algorithm for the domain name that is being sold. The question is whether the act of seeing the valuation for the domain name would induce bidders to submit higher bids and thereby raise the sale price. We find that there is indeed a positive effect of seeing the valuation on the sale price, and we show how the standard error of the average treatment effect can decrease by including covariates interacted with the treatment.

In section 2, we present the model, the estimators, efficiency comparisons, and consistent inference under the assumption of i.i.d. sampling. In section 3, we discuss semiparametric efficiency and semiparametric estimation methods. Section 4 generalizes

our results to covariate adaptive randomization. We provide monte carlo simulations in section 5 and the empirical application in section 6. Section 7 concludes.

2 Theoretical Model and Parametric Efficiency

Consider the causal LATE model of Imbens and Angrist (1994), and its special case the conditional independence (CI) model of Rosenbaum and Rubin (1983). The counterfactual outcomes are denoted Y_1, Y_0, and let $Z \in \{0, 1\}$ be the dummy instrumental variable indicating the eligibility for treatment. D_1 and D_0 are the counterfactual treatment statuses corresponding to $Z = 1$ and $Z = 0$ respectively. The sample contains Y, D, Z, and possibly additional covariates X, such that

$$\begin{aligned} D =& D_1 Z + D_0 (1 - Z) \\ Y =& Y_1 D + Y_0 (1 - D) = Y_1^* Z + Y_0^* (1 - Z) \end{aligned}$$

where $Y_1^* = Y_1 D_1 + Y_0 (1 - D_1)$ and $Y_0^* = Y_1 D_0 + Y_0 (1 - D_0)$. We begin with the usual i.i.d sampling with replacement framework:

Assumption 1 (i.i.d sampling) *$Y_{1i}, Y_{0i}, D_{1i}, D_{0i}, X_i, Z_i$ are drawn i.i.d from an underlying population.*

Both the CI and LATE models are assumed to satisfy two assumptions. First, the instrumental variable is independent of the potential outcomes, counterfactual treatment statuses, and covariates. Second, the counterfactual treatment status corresponding to $Z = 1$ is weakly greater than the counterfactual treatment status corresponding to $Z = 0$ with probability 1 and strictly greater with positive probability.

Assumption 2 (CI.1, LATE.1, Independence) *$Y_1, Y_0, D_1, D_0, X \perp Z$.*

Assumption 3 (CI.2, LATE.2, Monotonicity) *$P(D_1 \geq D_0) = 1$, and $P(D_1 > D_0) > 0$.*

The CI model additionally satisfies

Assumption 4 (CI.3, Full compliance) *$D_1 = 1$ and $D_0 = 0$, or $D = Z$.*

The CI Model often is stated without reference to Z since $D = Z$.

Assumption 5 (CI.1-3) *$Y_1, Y_0, X \perp D$.*

The population LATE parameter of interest is given by Imbens and Angrist (1994):

$$\beta_0 = E(Y_1 - Y_0 | D_1 > D_0) = Cov(Y, Z)/Cov(D, Z) = \frac{(E(Y|Z = 1) - E(Y|Z = 0))}{(E(D|Z = 1) - E(D|Z = 0))} \tag{1}$$

which becomes the average treatment effect (ATE) parameter under CI:

$$\beta_0 = E(Y_1 - Y_0) = (E(Y|Z = 1) - E(Y|Z = 0)).$$

Let $\hat{\beta}_1$ be the coefficient on D when running 2SLS of Y on D instrumented by Z:

$$\hat{\beta}_1 = \sum_{i=1}^{n} Y_i \left(Z_i - \bar{Z}\right) / \sum_{i=1}^{n} D_i \left(Z_i - \bar{Z}\right)$$

This corresponds to the following Stata command:

```
ivreg Y (D=Z)
```

When $Z = D$, $\hat{\beta}_1$ becomes the coefficient on D in an OLS regression of Y on D:

```
reg Y D
```

In both the CI and the LATE models, it is well known that $\hat{\beta}_1$ is consistent for β_0 and the nominal (robust) OLS and 2SLS standard errors consistently estimate the asymptotic variance for $\sqrt{n}\left(\hat{\beta}_1 - \beta_0\right)$. Next, consider applying OLS or 2SLS to regress Y_i on D_i and X_i, where D is instrumented by Z. Let $\hat{\beta}_2$ be the coefficient on D_i in

```
ivreg Y (D=Z) X      or     reg Y D X
```

It can be shown that $\hat{\beta}_2$ is consistent for β_0 under both CI and LATE, but $\hat{\beta}_2$ can be either more or less efficient than $\hat{\beta}_1$. The nominal (robust) OLS and 2SLS standard errors remain valid for $\hat{\beta}_2$.

Finally, consider using OLS or 2SLS to regress Y_i on D_i and X_i, and the interaction between $X_i - \bar{X}$ and Z_i. Let $\hat{\beta}_3$ be the coefficient on D in

```
ivreg Y (D=Z) X Z*(X-X̄) or  reg Y D X D*(X-X̄)
```

It can be shown that $\hat{\beta}_3$ is consistent for β_0 under both CI and LATE, and $\hat{\beta}_3$ is always no less efficient than $\hat{\beta}_1$ and $\hat{\beta}_2$.

Under i.i.d. sampling, the efficiency comparison holds under both the correlation model and the causal model.

Assumption 6 (correlation model) *Y and X have finite $\delta > 4$ moments, and $X \perp Z$.*

Theorem 1 *Under Assumptions 1 and 6, for $j = 1, 2, 3$,*

$$\sqrt{n}\left(\hat{\beta}_j - \beta_0\right) \xrightarrow{d} N\left(0, \sigma_j^2\right), \qquad \text{where } \sigma_3 \leq \sigma_k \text{ for } k = 1, 2.$$

Assumption 7 (causal model) *Y_{1i}, Y_{0i} and X_i have finite $4 + \delta$ moments.*

Corollary 1 *Theorem 1 holds under Assumptions 1, 2, 3, and 7.*

These results include the model in Lin (2013) as a special case when $Z_i = D_i$, but differ in that we are concerned about superpopulation asymptotics while Lin (2013) is not concerned about the variation in X_i. Adding covariates or functions of covariates into the interaction estimator $\hat{\beta}_3$ will also further improve efficiency. Let $\hat{\beta}_4$ be defined as $\hat{\beta}_3$ except that we replace X with a subset of it denoted X_s:

`ivreg Y (D=Z)` X_s `Z*(`X_s`-`$\bar{X}_s$`) or reg Y D` X_s `D*(`X_s`-`$\bar{X}_s$`)`

Corollary 2 *Under the conditions of either Theorem 1 or Corollary 1,*

$$\sqrt{n}\left(\hat{\beta}_4 - \beta_0\right) \xrightarrow{d} N\left(0, \sigma_4^2\right), \qquad \text{where } \sigma_4 \geq \sigma_3.$$

For $\hat{\beta}_1$ and $\hat{\beta}_2$, nominal robust 2SLS standard errors (when $D \neq Z$, and robust OLS standard errors when $Z = D$) reported by Stata are asymptotically valid. In contrast, robust 2SLS standard errors for $\hat{\beta}_3$ underestimate its asymptotic variance. In particular, nominal robust standard errors for $\hat{\beta}_k, k = 1, 2, 3$, denoted $\hat{\sigma}_k^2$, are given by the $(2,2)$ element of $\hat{A}^{-1}\hat{B}\hat{A}^{-1}$, where

$$\hat{A} = \frac{1}{n}\sum_{i=1}^{n} \bar{W}_{ik}\bar{V}_{ik} \qquad \hat{B} = \frac{1}{n}\sum_{i=1}^{n} \hat{\epsilon}_{ik}^2 \bar{W}_{ik}\bar{W}_{ik}'. \tag{2}$$

In the above, $\bar{W}_{i1} = (1\ Z_i)'$, $\bar{V}_{i1} = (1\ D_i)$, $\bar{W}_{i2} = (1\ Z_i\ X_i)$, $\bar{V}_{i2} = (1\ D_i\ X_i)$, $\bar{W}_{i3} = \left(1\ Z_i\ X_i\ Z_i\left(X_i - \bar{X}\right)\right)$, $\bar{V}_{i3} = \left(1\ D_i\ X_i\ Z_i\left(X_i - \bar{X}\right)\right)$, and $\hat{\epsilon}_{ik}$ is the regression residual corresponding to $\hat{\beta}_k$. Furthermore, define

$$\bar{\epsilon}_{i3} = (Z_i - \hat{p}_z)\,\hat{\epsilon}_{i3} + \hat{p}_z\,(1 - \hat{p}_z)\,\hat{\phi}\left(X_i - \bar{X}\right). \tag{3}$$

where $\hat{\phi}$ are the coefficients on $Z_i\left(X_i - \bar{X}\right)$ and $\hat{p}_z = \bar{Z}$. For $\widehat{Cov}_{Z,D} = \frac{1}{n}\sum_{i=1}^{n} Z_i D_i - \bar{Z}\bar{D}$,

$$\bar{\sigma}_3^2 = \widehat{Cov}_{Z,D}^{-2}\frac{1}{n}\sum_{i=1}^{n}\bar{\epsilon}_{i3}^2 \tag{4}$$

Corollary 3 *Under the conditions of either Theorem 1 or Corollary 1,* $\hat{\sigma}_k^2 \xrightarrow{p} \sigma_k^2$ *for* $k = 1, 2$*, and plim* $\hat{\sigma}_3^2 \leq \sigma_3^2$*. But* $\bar{\sigma}_3^2 \xrightarrow{p} \sigma_3^2$*.*

It is possible to give a GMM interpretation to the interactive estimator $\hat{\beta}_3$. By independence between Z and X, the moment conditions $E\phi_i(\alpha_0, \beta_0, \mu_{0x}) = 0$ hold, where

$$\phi_i(\alpha, \beta, \mu_x) = \begin{pmatrix} Z_i \\ 1 - Z_i \end{pmatrix} \otimes \begin{pmatrix} y_i - \alpha - \beta D_i \\ X_i - \mu_x \end{pmatrix}$$

Let $\widehat{\mathrm{Var}}(\phi_i(\cdot)) = \mathrm{Var}(\phi_i(\alpha_0, \beta_0, \mu_{0x})) + o_P(1)$ and $\hat{\phi}(\alpha, \beta, \mu_x) = \frac{1}{n}\sum_{i=1}^{n}\phi_i(\alpha, \beta, \mu_x)$. It can then be shown that the GMM estimator, defined through

$$\left(\hat{\alpha}, \hat{\beta}_{GMM}, \hat{\mu}_x\right) = \arg\min_{\alpha,\beta,\mu_x} \hat{\phi}(\alpha, \beta, \mu_x)'\,\widehat{\mathrm{Var}}(\phi_i(\alpha, \beta, \mu_x))^{-1}\,\hat{\phi}(\alpha, \beta, \mu_x) \tag{5}$$

coincides asymptotically with the interactive IV estimator $\hat{\beta}_3$.

Proposition 1 $\hat{\beta}_{GMM} = \hat{\beta}_3 + o_P\left(\frac{1}{\sqrt{n}}\right)$.

3 Semiparametric Estimation and Efficiency

The asymptotic variance of the interactive estimator $\hat{\beta}_3$ decreases when more regressors are added. If we replace X_i by its sieve expansion, denoted $V_i = V(X_i)$, where $\dim(V_i) \to \infty$ as $n \to \infty$ at a suitable rate, then our interactive estimator is exactly the sieve ATE estimate in Chen, Hong and Tarozzi (CHT 2008) when $D = Z$, and is a sieve version of the average LATE estimator of Frolich (2006) when $D \neq Z$. We denote by $\hat{\beta}_\infty$ our interactive estimator using $V_i = V(X_i)$ in place of X_i.

We show this equivalence first for $D = Z$. The CHT 2008 ATE estimator uses two linear regressions:

$$\begin{aligned}(1 - D_i) Y_i =& (1 - D_i) \left[\hat{\gamma}_0 + \hat{\vartheta}_0 V_i + e_{0i}\right] \\ D_i Y_i =& D_i \left[\hat{\gamma}_1 + \hat{\vartheta}_1 V_i + e_{1i}\right]\end{aligned}$$

and is based on the following relations:

$$\begin{aligned}\hat{E}(Y_0|X = x) =& \hat{\gamma}_0 + \hat{\vartheta}_0 V(x) \quad \hat{E}(Y_1|X = x) = \hat{\gamma}_1 + \hat{\vartheta}_1 V(x) \\ \widehat{ATE} =& \frac{1}{n} \sum_{i=1}^{n} \left(\hat{E}(Y_1|X = X_i) - \hat{E}(Y_0|X = X_i)\right) \\ =& \frac{1}{n} \sum_{i=1}^{n} \left(\hat{\gamma}_1 + \hat{\vartheta}_1 V_i - \hat{\gamma}_0 - \hat{\vartheta}_0 V_i\right) = \hat{\gamma}_1 + \hat{\vartheta}_1 \bar{V} - \hat{\gamma}_0 - \hat{\vartheta}_0 \bar{V}.\end{aligned}$$

Our interactive regression can be rewritten as

$$Y_i = \hat{\alpha} + \hat{\beta}_\infty D_i + \hat{\eta} V_i + \hat{\phi} D_i \left(V_i - \bar{V}\right) = \hat{\alpha} + \hat{\eta} V_i + D_i \left(\hat{\beta}_\infty - \hat{\phi} \bar{V} + \hat{\phi} V_i\right).$$

The following equalities hold between the two different parameterizations:

$$\hat{\gamma}_0 = \hat{\alpha}, \quad \hat{\vartheta}_0 = \hat{\eta}, \quad \hat{\gamma}_1 = \hat{\alpha} + \hat{\beta}_\infty - \hat{\phi} \bar{V}, \quad \hat{\vartheta}_1 = \hat{\eta} + \hat{\phi}.$$

Therefore the following equivalence relation holds:

$$\widehat{ATE} = \hat{\gamma}_1 + \hat{\vartheta}_1 \bar{V} - \hat{\gamma}_0 - \hat{\vartheta}_0 \bar{V} = \hat{\alpha} + \hat{\beta}_\infty - \hat{\phi} \bar{V} + \left(\hat{\eta} + \hat{\phi}\right) \bar{V} - \hat{\alpha} - \hat{\eta} \bar{V} = \hat{\beta}_\infty. \tag{6}$$

In the special case when V_i are cluster dummy variables, equation (6) is identical to a fully saturated regression of the outcome on the treatment and cluster dummies with full interactions and computes the cluster-weighted average of the cluster-level estimates. More precisely, let $V_i(s) = 1(X_i \in s)$ for all clusters $s = 1, \ldots, S$ and let $\bar{V}(s) = \frac{1}{n} \sum_{i=1}^{n} 1(X_i \in s)$. Then equation (6) becomes

$$\hat{\gamma}_1 + \sum_{s=2}^{S} \hat{\vartheta}_{1s} \bar{V}(s) - \left(\hat{\gamma}_0 + \sum_{s=2}^{S} \hat{\vartheta}_{0s} \bar{V}(s)\right) \tag{7}$$

Let $\hat{\xi}_{11} = \hat{\gamma}_1$, $\hat{\xi}_{01} = \hat{\gamma}_0$, $\hat{\xi}_{1s} = \hat{\gamma}_1 + \hat{\vartheta}_{1s}$, and $\hat{\xi}_{0s} = \hat{\gamma}_0 + \hat{\vartheta}_{0s}$ for $s = 2, \ldots, S$. Then

$$(7) = \sum_{s=1}^{S} \left(\hat{\xi}_{1s} - \hat{\xi}_{0s}\right) \frac{\sum_{i=1}^{n} 1(X_i \in s)}{n}. \tag{8}$$

But $\hat{\xi}_{1s} - \hat{\xi}_{0s}$ is exactly the difference in the cluster s levels in Y_i for $D_i = 1$ and $D_i = 0$. This estimator achieves the semiparametric efficiency bound when only cluster indicators are observable, but is not fully efficient when the covariates X_i are also observable.

We now consider the LATE model when $D \neq Z$, and show that the interactive IV estimator $\hat{\beta}_\infty$ is a sieve implementation of the semiparametrically efficient average LATE estimator in Frolich (2006), which takes the form of

$$\widehat{AvgLATE} = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{E}\left(Y|Z=1,X=X_i\right) - \hat{E}\left(Y|Z=0,X=X_i\right)\right)}{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{E}\left(D|Z=1,X=X_i\right) - \hat{E}\left(D|Z=0,X=X_i\right)\right)}.$$

A sieve implementation of this estimator is based on the following relations:

$$\begin{aligned}
\hat{E}\left(Y|Z=0,X=x\right) =& \hat{\gamma}_0 + \hat{\vartheta}_0 V\left(x\right), \quad \hat{E}\left(Y|Z=1,X=x\right) = \hat{\gamma}_1 + \hat{\vartheta}_1 V\left(x\right) \\
\hat{E}\left(D|Z=0,X=x\right) =& \hat{\tau}_0 + \hat{\zeta}_0 V\left(x\right), \quad \hat{E}\left(D|Z=1,X=x\right) = \hat{\tau}_1 + \hat{\zeta}_1 V\left(x\right),
\end{aligned}$$

and uses the following four linear regressions:

$$\begin{aligned}
\left(1-Z_i\right)Y_i =& \left(1-Z_i\right)\left[\hat{\gamma}_0 + \hat{\vartheta}_0 V_i + e_{0i}\right], \quad Z_i Y_i = Z_i\left[\hat{\gamma}_1 + \hat{\vartheta}_1 V_i + e_{1i}\right] \\
\left(1-Z_i\right)D_i =& \left(1-Z_i\right)\left[\hat{\tau}_0 + \hat{\zeta}_0 V_i + e_{0i}\right], \quad Z_i D_i = Z_i\left[\hat{\tau}_1 + \hat{\zeta}_1 V_i + e_{1i}\right].
\end{aligned} \tag{9}$$

We can then write

$$\widehat{AvgLATE} = \frac{\hat{\gamma}_1 - \hat{\gamma}_0 + \left(\hat{\vartheta}_1 - \hat{\vartheta}_0\right)'\bar{V}}{\hat{\tau}_1 - \hat{\tau}_0 + \left(\hat{\zeta}_1 - \hat{\zeta}_0\right)'\bar{V}}. \tag{10}$$

Proposition 2 *$\widehat{AvgLATE} = \hat{\beta}_\infty$ for the interactive instrumental variable estimator $\hat{\beta}_\infty$.*

When V_i are cluster indicators, the LATE analog of (8) becomes

$$\hat{\beta}_\infty = \frac{\sum_{s=1}^{S}\left(\hat{\xi}_{1s} - \hat{\xi}_{0s}\right)\frac{\sum_{i=1}^{n}1(X_i \in s)}{n}}{\sum_{s=1}^{S}\left(\hat{\zeta}_{1s} - \hat{\zeta}_{0s}\right)\frac{\sum_{i=1}^{n}1(X_i \in s)}{n}}, \tag{11}$$

where $\hat{\xi}_{1s} - \hat{\xi}_{0s}$ is exactly the difference in the cluster s levels in Y_i between $Z_i = 1$ and $Z_i = 0$, and $\hat{\zeta}_{1s} - \hat{\zeta}_{0s}$ is the difference in the cluster s levels in D_i between $Z_i = 1$ and $Z_i = 0$. This estimator achieves the semiparametric efficiency bound when only cluster indicators are observable, but is not fully efficient when the covariates X_i are also observable. Under CI where $D = Z$, semiparametric efficiency bounds are calculated in, among others, Hahn (1998) and Chen et al. (2008). In the more general LATE case when $D \neq Z$, the LATE efficiency bound is calculated in Frolich (2006) as well as Hong and Nekipelov (2010) (Lemma 1 and Theorem 4). These results from the previous literature confirm the following efficiency comparison:

Proposition 3 $\sqrt{n}\left(\hat{\beta}_\infty - \beta_0\right) \xrightarrow{d} N\left(0, \sigma^2_\infty\right)$, *where* $\sigma^2_\infty \leq \sigma^2_3$.

Under suitable regularity conditions, a consistent estimate of σ^2_∞ can be obtained by an analog of (3) and (4) when X_i and $\bar{X}$ are replaced by sieve expansions V_i and $\bar{V}$:

$$\bar{\epsilon}_{i\infty} = (Z_i - \hat{p}_z)\,\hat{\epsilon}_{i\infty} + \hat{p}_z\,(1 - \hat{p}_z)\,\hat{\phi}'\left(V_i - \bar{V}\right).$$

where $\hat{\phi}$ are the coefficients on $Z_i\left(V_i - \bar{V}\right)$ and $\hat{p}_z = \bar{Z}$. For $\widehat{Cov}_{Z,D} = \frac{1}{n}\sum Z_i D_i - \bar{Z}\bar{D}$, a consistent estimate of σ^2_∞ is given by

$$\bar{\sigma}^2_\infty = \widehat{Cov}^{-2}_{Z,D}\frac{1}{n}\sum_{i=1}^{n}\bar{\epsilon}^2_{i\infty}$$

4 Covariate Adaptive Randomization

We now move beyond the i.i.d. sampling framework and consider the covariate adaptive randomization scheme discussed in Bugni et al. (2017a) and Bugni et al. (2017a) (BCS 2017a, BCS 2017b) where units are first assigned to a finite number of strata using baseline covariates and then are assigned treatment status using the instrument. Unlike BCS 2017b, we do not allow for multiple treatments, so we will describe how our notation differs from BCS 2017a.

1. Our treatment variables are D_1 and D_0 corresponding to $Z = 1$ and $Z = 0$ respectively while BCS 2017a use A to denote the treatment in the case where $D = Z$.
2. Our baseline covariates which determine stratum membership are denoted by X while BCS 2017a use Z.
3. Our target proportion of units assigned to treatment in each stratum is denoted by p_z while BCS 2017a use π.

We also note that our notation will differ from chapter 9 of Imbens and Rubin (2015) in the following ways:

1. Our treatment variables are D_1 and D_0 while Imbens and Rubin (2015) assume $D = Z$ and use W to denote the treatment.
2. We follow BCS 2017a and use $S_i \in \{1, 2, ..., S\}$ to denote the stratum of unit i while Imbens and Rubin (2015) define a variable $B_i(j)$ which is an indicator for unit i belonging in stratum j for $j \in \{1, ..., J\}$.
3. We use $p_z(s)$ to denote the proportion of treated units in stratum s while Imbens and Rubin (2015) use $e(j)$.

With more than one stratum, BCS 2017b have already shown that interacting the instrument with strata indicators improves efficiency, and they allow for different conditional targeted randomization probabilities across strata. Our contribution is twofold. First we extend BCS 2017a and BCS 2017b to allow for non-compliance by operating under the LATE framework and IV regression. Second we show how

additional covariates beyond strata indicators further enhance efficiency and derive an efficient semiparametric sieve based estimator. Results in the previous sections correspond to the single stratum case and coincide in the special case of $D = Z$ with simple OLS and adjusted but non-interacted OLS.

Similar to BCS 2017a, consider a sampling scheme where $(Y_{1i}, Y_{0i}, D_{1i}, D_{0i}, X_i)$ are drawn i.i.d from a superpopulation and are first assigned to a finite set of clusters using a function $S : \text{supp}(X_i) \to \mathcal{S}$ based on the value of X_i before treatment status is assigned using Z_i. As in BCS 2017a, let $S_i = S(X_i)$, $S^{(n)} = (S_1, \ldots, S_n)$, $Z^{(n)} = (Z_1, \ldots, Z_n)$, $p(s) = P(S_i = s)$, and define a measure of imbalance in stratum s relative to the target proportion p_z as

$$Z_n(s) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} 1(X_i \in s)(Z_i - p_z).$$

Assumption 8 *1.* $Z^{(n)} \perp (Y_{1i}, Y_{0i}, D_{1i}, D_{0i}, X_i, i = 1, \ldots, n) | S^{(n)}$.
2. $P\left(Z_i = 1 | S^{(n)}\right) = p_z + O_{a.s.}\left(\frac{1}{n}\right)$ *for all* $1 \leq i \leq n$.
3. $\{Z_n(s)_{s \in \mathcal{S}} | S^{(n)}\} \xrightarrow{d} N(0, \Sigma_Z)$, *where* $\Sigma_Z = diag\{\sigma_Z^2(s) : s \in \mathcal{S}\}$ *and*

$$\sigma_Z^2(s) = p(s)\tau(s) \quad \text{and} \quad 0 \leq \tau(s) \leq p_z(1 - p_z), \forall s \in \mathcal{S}.$$

$\tau(s)$ is a strata-specific scalar that equals 0 for all $s \in \mathcal{S}$ in the case of strong balance. An example of a sampling scheme that achieves strong balance is stratified block randomization (see example 3.4 of BCS 2017a). Assumption 8 is modeled after Assumption 2.2 in BCS 2017a but is different. It allows for non-compliance in the sense that $D \neq Z$.

Consider first the simple IV estimator $\hat{\beta}_1$ with instrument Z and regressor D. Define for $t_{1i} = Y_{1i}^* - \beta_0 D_{1i}$ and $t_{0i} = Y_{0i}^* - \beta_0 D_{0i}$,

$$\omega_i = \left[\frac{t_{1i} - Et_{1i}}{p_z} + \frac{t_{0i} - Et_{0i}}{1 - p_z}\right] \quad \text{and} \quad \omega(s) = E[\omega_i | X_i \in s].$$

Proposition 4 *Under Assumption 8,* $\sqrt{n}\left(\hat{\beta}_1 - \beta_0\right) \xrightarrow{d} N\left(0, \sigma_{1fs1}^2 + \sigma_{1fs2}^2 + \sigma_\infty^2\right)$, *where*

$$\sigma_{1fs1}^2 = \frac{p_z(1 - p_z)}{P(D_1 > D_0)^2} \sum_{s=1}^{S} p(s) Var(\omega_i | s), \sigma_{1fs2}^2 = \frac{\sum_{s=1}^{S} \omega(s)^2 p(s) \tau(s)}{P(D_1 > D_0)^2}, \sigma_\infty^2 = \frac{Var[t_{1i} - t_{0i}]}{P(D_1 > D_0)^2}.$$

Furthermore, $plim\, \hat{\sigma}_1^2 \geq \sigma_{1fs}^2 + \sigma_\infty^2$ *where the inequality is strict when* $\tau(s) < p_z(1 - p_z)$.

Therefore, 2SLS nominal standard errors are in general conservatively valid and only asymptotically accurate when $\tau(s) = p_z(1 - p_z)$. We note that Proposition 4 has already been shown in Theorem 4.1 of BCS 2017a when $D = Z$. Our contribution is to extend their results to the case of $D \neq Z$.

Next consider the adjusted regression $\hat{\beta}_2$, where X_i is replaced by cluster dummies

$$V_i = \{1(X_i \in s), s \in \mathcal{S}\}.$$

Proposition 5 *Under Assumption 8,* $\sqrt{n}\left(\hat{\beta}_2 - \beta_0\right) \xrightarrow{d} N\left(0, \sigma^2_{2fs1} + \sigma^2_{2fs2} + \sigma^2_{\infty}\right)$, *where* $\sigma^2_{2fs1} = \sigma^2_{1fs1}$, *and*

$$\sigma^2_{2fs2} = P\left(D_1 > D_0\right)^{-2} \sum_{s=1}^{S} p\left(s\right) \tau\left(s\right) \left(\frac{1-2p_z}{p_z\left(1-p_z\right)} \left(t_1\left(s\right) - t_0\left(s\right)\right)\right)^2.$$

where $t_1\left(s\right) = E\left(t_{1i} | X_i \in s\right)$ *and* $t_0\left(s\right) = E\left(t_{0i} | X_i \in s\right)$. *Furthermore,* $plim\, \hat{\sigma}^2_2 = \sigma^2_{2fs1} + \bar{\sigma}^2_{2fs2} + \sigma^2_{\infty}$, *where* $\bar{\sigma}^2_{2fs2} \geq \sigma^2_{2fs2}$,

$$\bar{\sigma}^2_{2fs2} = P\left(D_1 > D_0\right)^{-2} \sum_{s=1}^{S} p\left(s\right) \frac{1}{p_z\left(1-p_z\right)} \left(\left(1-2p_z\right)\left(t_1\left(s\right) - t_0\left(s\right)\right)\right)^2.$$

Consequently, nominal 2SLS standard errors are generally conservative and only asymptotically valid when either $\tau\left(s\right) \equiv p_z\left(1-p_z\right)$ or $p_z = \frac{1}{2}$. We emphasize that the case of $D = Z$ has already been shown in BCS 2017a. Our contribution is only to allow for noncompliance with $D \neq Z$.

Proposition 6 *Under Assumption 8,* $\sqrt{n}\left(\hat{\beta}_3 - \beta_0\right) \xrightarrow{d} N\left(0, \sigma^2_{3fs} + \sigma^2_{\infty}\right)$, *where* $\sigma^2_{3fs} = \sigma^2_{2fs1}$.*Furthermore,* $plim\, \hat{\sigma}^2_3 \in \left[\sigma^2_{3fs}, \sigma^2_{3fs} + \sigma^2_{\infty}\right]$.

The asymptotic variance of $\hat{\beta}_3$ is smaller than the variances of $\hat{\beta}_1$ and $\hat{\beta}_2$ by the amount σ^2_{1fs2} and σ^2_{2fs2}, respectively, except in the cases of $p_z = 1/2$ or $\tau\left(s\right) = 0$, in which case the asymptotic variances of $\hat{\beta}_3$ and $\hat{\beta}_2$ are the same. As before the case of $D = Z$ in Proposition 6 follows from results that have already been shown in BCS 2017b. Our contribution is to allow for noncompliance.

In addition, even if the targeted randomization probability is specific to each cluster, namely $p_z\left(s\right)$ can differ across clusters, if we replace Assumption 8.2 by

$$P\left(Z_i = 1 | S^{(n)}, X_i \in s\right) = p_z\left(s\right) + O_{a.s.}\left(\frac{1}{n}\right) \quad \text{and} \quad \tau\left(s\right) \leq p_z\left(s\right)\left(1 - p_z\left(s\right)\right),$$

$\hat{\beta}_3$ continues to be consistent and Proposition 6 continues to hold with

$$\sigma^2_{3fs} = P\left(D_1 > D_0\right)^{-2} \sum_{s \in \mathcal{S}} p\left(s\right) p_z\left(s\right)\left(1 - p_z\left(s\right)\right) Var\left(\omega_i | s\right).$$

When $p_z\left(s\right) \equiv p_z$, nominal 2SLS robust standard errors overestimate σ^2_{3fs} but underestimate $\sigma^2_{3fs} + \sigma^2_{\infty}$. A consistent estimate for $\sigma^2_{3fs} + \sigma^2_{\infty}$ can be obtained by

$$\widehat{Cov}^{-2}_{Z,D} \frac{1}{n} \sum_{i=1}^{n} \left(\left(Z_i - \hat{p}_z\right)\hat{\epsilon}_{i\infty} + \hat{p}_z\left(1 - \hat{p}_z\right)\hat{\phi}'\left(V_i - \bar{V}\right)\right)^2$$

where $\hat{\phi}$ are the coefficients on $Z_i\left(V_i - \bar{V}\right)$, $\hat{p}_z = \bar{Z}$, and $\widehat{Cov}_{Z,D} = \frac{1}{n}\sum Z_i D_i - \bar{Z}\bar{D}$. Additional covariates X_i can be utilized to improve efficiency beyond the cluster indicators. Asymptotic efficiency is obviously maximized by the semiparametric estimators in section 3, e.g. Chen et al. (2008) and Frolich (2006), where both the cluster

dummies and sieve transformations of increasing dimensions of X_i and their interactions are included in V_i when defining the sieve 2SLS estimator $\hat{\beta}_\infty$. Here we investigate the efficiency improvement from interacting the finite dimensional functions of X_i (which we denote for convenience as X_i) with Z_i and the cluster dummies. We can equivalently rewrite this estimator as

$$\hat{\beta}_S = \frac{\sum_{s\in\mathcal{S}} \hat{p}(s)\left(\hat{\gamma}_{1s} - \hat{\gamma}_{0s} + \left(\hat{\vartheta}_{1s} - \hat{\vartheta}_{0s}\right)' \bar{X}_s\right)}{\sum_{s\in\mathcal{S}} \hat{p}(s)\left(\hat{\tau}_{1s} - \hat{\tau}_{0s} + \left(\hat{\zeta}_{1s} - \hat{\zeta}_{0s}\right)' \bar{X}_s\right)}$$

where $\hat{p}(s) = \frac{1}{n}\sum_{i=1}^n 1(X_i \in s)$, $\bar{X}_s = \frac{1}{n}\sum_{i=1}^n 1(X_i \in s) X_i/\hat{p}(s)$, and the coefficients are obtained using four sets of cluster-specific regressions using only the s cluster:

$$\begin{aligned}(1-Z_i)Y_i &= (1-Z_i)\left[\hat{\gamma}_{0s} + \hat{\vartheta}_{0s}X_i + e_{0i}\right], \quad Z_iY_i = Z_i\left[\hat{\gamma}_{1s} + \hat{\vartheta}_{1s}X_i + e_{1i}\right] \\ (1-Z_i)D_i &= (1-Z_i)\left[\hat{\tau}_{0s} + \hat{\zeta}_{0s}X_i + e_{0i}\right], \quad Z_iD_i = Z_i\left[\hat{\tau}_{1s} + \hat{\zeta}_{1s}X_i + e_{1i}\right].\end{aligned} \tag{12}$$

Proposition 7 *For* $\sigma^2_{Sfs} < \sigma^2_{3fs}$, $\sqrt{n}\left(\hat{\beta}_S - \beta_0\right) \xrightarrow{d} N\left(0, \sigma^2_{Sfs} + \sigma^2_\infty\right)$,

In the special case of $D = Z$, $\hat{\xi}_{1s} = 1$, $\hat{\xi}_{0s} = 0$, $\hat{\zeta}_{1s} = \hat{\zeta}_{0s} = 0$. Therefore,

$$\hat{\beta}_S = \sum_{s\in\mathcal{S}} \hat{p}(s)\left(\hat{\gamma}_{1s} - \hat{\gamma}_{0s} + \left(\hat{\vartheta}_{1s} - \hat{\vartheta}_{0s}\right)' \bar{X}_s\right)$$

A consistent estimate of $\sigma^2_{Sfs} + \sigma^2_\infty$ can again be obtained using an analytical expression.

5 Monte Carlo Simulations

The purpose of these simulations is to perform efficiency comparisons for the three different estimators in both the CI and LATE models, with and without covariate adaptive randomization.

The data generating process for the CI model is as follows:

$$Y_i = \beta_0 + \beta_1 D_i + \beta_2 D_i X_i^2 + \beta_3 X_i + \epsilon_i, \epsilon_i \sim N(0,1), \epsilon_i \perp D_i$$

where $\beta_0 = 1$, $\beta_1 = 0.5$, and $\beta_2 = \beta_3 = -1$. The single covariate is generated as $X_i \sim N(\mu_x = 10, \sigma_x = 5), X_i \perp D_i, \epsilon_i$.

Without covariate adaptive randomization, the treatment D_i is generated as $D_i \sim Bern(0.5)$.

With covariate adaptive randomization, there are two strata $S_i = 1(X_i > \mu_x)$. Under block randomization, $p_z \# \{S_i = 0\}$ elements are assigned to treatment in strata

0 and $p_z\#\{S_i = 1\}$ elements are assigned to treatment in strata 1. Under simple randomization, $D_i|S_i = s \sim Binomial\left(\#\{S_i = s\}, p_z\right)$ for $s = 0, 1$.

For the LATE model, the data generating process is

$$D_i = \gamma_0 + \gamma_1 Z_i + \nu_i > 0$$
$$Y_i = \beta_0 + \beta_1 D_i + \beta_2 D_i X_i^2 + \beta_3 X_i + \epsilon_i$$
$$\begin{pmatrix} \nu_i \\ \epsilon_i \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 2 & 0.5 \\ 0.5 & 2 \end{pmatrix}\right)$$

where $\gamma_0 = 1$, $\gamma = 10$, $\beta_0 = 1$, $\beta_1 = 0.5$, and $\beta_2 = \beta_3 = -1$.

Without covariate adaptive randomization, the instrument Z_i is generated as $Z_i \sim Bern\,(0.5)$. With covariate adaptive randomization, there are two strata $S_i = 1\,(X_i > \mu_x)$. Under block randomization, $p_z\#\{S_i = 0\}$ elements are assigned to treatment in strata 0 and $p_z\#\{S_i = 1\}$ elements are assigned to treatment in strata 1. Under simple randomization, $Z_i|S_i = s \sim Binomial\left(\#\{S_i = s\}, p_z\right)$ for $s = 0, 1$.

We consider the simple (L)ATE estimator without covariates, the (L)ATE estimator with the covariate, and the (L)ATE estimator with the covariate and the interaction between the instrument and the covariate. In the case of covariate adaptive randomization, the covariate is the stratum indicator V.

The first row of Table 1 shows the average monte carlo standard errors for the ATE estimates which are the standard deviations of the estimates across 20000 monte carlo simulations. The next two rows show the ATE average standard errors and average confidence interval length using nominal robust OLS standard errors for $\hat{\beta}_1$ and $\hat{\beta}_2$ and the analytic correction for $\hat{\beta}_3$ in Corollary 3. The number of observations is $n = 2000$.

We can see that relative to the baseline model with just the treatment on the right hand side, the monte carlo standard error decreases when we include the uninteracted covariate and when we also include the interaction between the covariate and the treatment. The same pattern holds for the average standard errors. Additionally, the average confidence interval length decreases when we add in the uninteracted covariate and when we also include the interaction between the covariate and the treatment.

Tabelle 1 ATE without Covariate Adaptive Randomization

	`reg` Y D	`reg` Y D X	`reg` Y D X $D * (X - \bar{X})$
Monte Carlo Standard Err	3.49	2.49	2.49
Avg Standard Err	3.50	2.50	2.50
Avg Length of CI	13.74	9.80	9.79

The first rows of Tables 2 and 3 show the average monte carlo standard errors which are the standard deviations of the estimates across 20000 monte carlo simulations. The next two rows show the ATE average standard errors and average confidence interval length under covariate adaptive block and simple randomization with $p_z = 0.3$ using nominal robust OLS standard errors for $\hat{\beta}_1$ and $\hat{\beta}_2$ and the analytic correction for $\hat{\beta}_3$. The number of observations is $n = 2000$.

For covariate adaptive block randomization, the monte carlo standard error is the same for all three estimators. Nominal OLS standard errors for the unadjusted and uninteracted estimators are conservative, while standard errors obtained using the analytic correction are not.

For covariate adaptive simple randomization, the monte carlo standard error is highest for the unadjusted model and is lowest for the interacted model. Nominal OLS and analytically corrected standard errors are not conservative and exhibit the same pattern as the monte carlo standard errors.

Tabelle 2 ATE with Covariate Adaptive Block Randomization, $p_z = 0.3$

	`reg` $Y\ D$	`reg` $Y\ D\ V$	`reg` $Y\ D\ V\ D*(V-\bar{V})$
Monte Carlo Standard Err	3.46	3.46	3.46
Avg Standard Err	4.52	3.78	3.45
Avg Length of CI	17.70	14.82	13.51

Tabelle 3 ATE with Covariate Adaptive Simple Randomization, $p_z = 0.3$

	`reg` $Y\ D$	`reg` $Y\ D\ V$	`reg` $Y\ D\ V\ D*(V-\bar{V})$
Monte Carlo Standard Err	4.52	3.79	3.46
Avg Standard Err	4.52	3.79	3.45
Avg Length of CI	17.72	14.85	13.53

The first rows of Tables 4 and 5 show the average monte carlo standard errors which are the standard deviations of the estimates across 20000 monte carlo simulations. The next two rows show the ATE average standard errors and average confidence interval length under covariate adaptive block and simple randomization with $p_z = 0.5$ using nominal robust OLS standard errors for $\hat{\beta}_1$ and $\hat{\beta}_2$ and the analytic correction for $\hat{\beta}_3$. The number of observations is $n = 2000$.

For covariate adaptive block randomization, the monte carlo standard error is the same for all three estimators. Nominal OLS standard errors are conservative for the unadjusted estimator but not for the uninteracted estimator. Standard errors obtained using the analytic correction are also not conservative.

For covariate adaptive simple randomization, the monte carlo standard error is highest for the unadjusted model and decreases when we add covariates. Unlike the case of $p_z = 0.3$, now the uninteracted estimator and the interacted estimator have the same standard errors. Nominal OLS and analytically corrected standard errors are not conservative and exhibit the same pattern as the monte carlo standard errors.

Tabelle 4 ATE with Covariate Adaptive Block Randomization, $p_z = 0.5$

	reg Y D	reg Y D V	reg Y D V $D * (V - \bar{V})$
Monte Carlo Standard Err	2.90	2.90	2.90
Avg Standard Err	3.50	2.90	2.90
Avg Length of CI	13.74	11.38	11.38

Tabelle 5 ATE with Covariate Adaptive Simple Randomization, $p_z = 0.5$

	reg Y D	reg Y D V	reg Y D V $D * (V - \bar{V})$
Monte Carlo Standard Err	3.51	2.90	2.90
Avg Standard Err	3.51	2.91	2.90
Avg Length of CI	13.74	11.39	11.38

The first rows of Tables 6 and 7 show the average monte carlo standard errors which are the standard deviations of the estimates across 20000 monte carlo simulations. The next two rows show the ATE average standard errors and average confidence interval length under covariate adaptive block and simple randomization with $p_z = 0.7$ using nominal robust OLS standard errors for $\hat{\beta}_1$ and $\hat{\beta}_2$ and the analytic correction for $\hat{\beta}_3$. The number of observations is $n = 2000$.

For covariate adaptive block randomization, the monte carlo standard error is the same for all three estimators. However, just like for $p_z = 0.3$, nominal OLS standard errors for the unadjusted and uninteracted estimators are conservative, while standard errors obtained using the analytic correction are not.

For covariate adaptive simple randomization, the monte carlo standard error is highest for the unadjusted model and is lowest for the interacted model. Nominal OLS and analytically corrected standard errors are not conservative and exhibit the same pattern as the monte carlo standard errors.

Tabelle 6 ATE with Covariate Adaptive Block Randomization, $p_z = 0.7$

	reg Y D	reg Y D V	reg Y D V $D * (V - \bar{V})$
Monte Carlo Standard Err	2.63	2.63	2.63
Avg Standard Err	2.97	3.06	2.63
Avg Length of CI	11.63	12.01	10.33

Tabelle 7 ATE with Covariate Adaptive Simple Randomization, $p_z = 0.7$

	reg Y D	reg Y D V	reg Y D V $D * (V - \bar{V})$
Monte Carlo Standard Err	2.98	3.05	2.63
Avg Standard Err	2.97	3.06	2.64
Avg Length of CI	11.63	12.01	10.33

The first row of Table 8 shows the average monte carlo standard errors for the LATE estimates which are the standard deviations of the estimates across 20000 monte carlo simulations. The next two rows show the LATE average standard errors and average confidence interval length using nominal robust 2SLS standard errors for $\hat{\beta}_1$ and $\hat{\beta}_2$ and the analytic correction for $\hat{\beta}_3$ in Corollary 3. The number of observations is $n = 2000$. We can see that relative to the baseline model with just the treatment on

the right hand side, the monte carlo standard error decreases when we include either the uninteracted covariate by itself or with the interaction between the covariate and the instrument. The same pattern holds for the average standard errors. Additionally, the average confidence interval length decreases when we add in just the uninteracted covariate and when we also include the interaction between the covariate and the instrument.

Tabelle 8 LATE without Covariate Adaptive Randomization

	ivreg Y $(D = Z)$	ivreg Y $(D = Z)$ X	ivreg Y $(D = Z)$ X $Z * \left(X - \bar{X}\right)$
Monte Carlo Standard Err	19.31	8.66	8.67
Avg Standard Err	19.49	8.67	8.67
Avg Length of CI	76.40	33.99	34.00

The first rows of Tables 9 and 10 show the average monte carlo standard errors which are the standard deviations of the estimates across 20000 monte carlo simulations. The next two rows show the LATE average standard errors and average confidence interval length under covariate adaptive block and simple randomization with $p_z = 0.3$ using nominal robust 2SLS standard errors for $\hat{\beta}_1$ and $\hat{\beta}_2$ and the analytic correction for $\hat{\beta}_3$. The number of observations is $n = 2000$.

For covariate adaptive block randomization, the monte carlo standard error is the same for all three estimators. However, the nominal 2SLS standard errors for the unadjusted estimator is very conservative, while the nominal 2SLS standard errors for the uninteracted estimator and the analytic correction standard errors for the interacted estimator are not particularly conservative.

For covariate adaptive simple randomization, the monte carlo standard error is highest for the unadjusted model and is lowest for the interacted model. Nominal 2SLS and analytically corrected standard errors are not particularly conservative and exhibit the same pattern as the monte carlo standard errors.

Tabelle 9 LATE with Covariate Adaptive Block Randomization, $p_z = 0.3$

	ivreg Y $(D = Z)$	ivreg Y $(D = Z)$ V	ivreg Y $(D = Z)$ V $Z * \left(V - \bar{V}\right)$
Monte Carlo Standard Err	14.78	14.78	14.78
Avg Standard Err	21.76	14.92	14.84
Avg Length of CI	85.30	58.50	58.18

Tabelle 10 LATE with Covariate Adaptive Simple Randomization, $p_z = 0.3$

	ivreg Y $(D = Z)$	ivreg Y $(D = Z)$ V	ivreg Y $(D = Z)$ V $Z * \left(V - \bar{V}\right)$
Monte Carlo Standard Err	21.71	14.85	14.79
Avg Standard Err	21.79	14.94	14.86
Avg Length of CI	85.42	58.57	58.23

The first rows of Tables 11 and 12 show the average monte carlo standard errors which are the standard deviations of the estimates across 20000 monte carlo simulations.

The next two rows show the LATE average standard errors and average confidence interval length under covariate adaptive block and simple randomization with $p_z = 0.5$ using nominal robust 2SLS standard errors for $\hat{\beta}_1$ and $\hat{\beta}_2$ and the analytic correction for $\hat{\beta}_3$. The number of observations is $n = 2000$.

For covariate adaptive block randomization, the monte carlo standard error is the same for all three estimators. Nominal 2SLS standard errors for the unadjusted estimator are conservative, while nominal 2SLS standard errors for the uninteracted estimator are not. Standard errors obtained using the analytic correction are also not conservative.

For covariate adaptive simple randomization, the monte carlo standard error is highest for the unadjusted model and decreases when we add covariates. Unlike the case of $p_z = 0.3$, now the uninteracted estimator and the interacted estimator have the same standard errors. Nominal 2SLS and analytically corrected standard errors are not conservative and exhibit the same pattern as the monte carlo standard errors.

Tabelle 11 LATE with Covariate Adaptive Block Randomization, $p_z = 0.5$

	`ivreg` Y $(D = Z)$	`ivreg` Y $(D = Z)$ V	`ivreg` Y $(D = Z)$ V $Z * (V - \bar{V})$
Monte Carlo Standard Err	13.60	13.60	13.60
Avg Standard Err	19.46	13.63	13.63
Avg Length of CI	76.30	53.43	53.43

Tabelle 12 LATE with Covariate Adaptive Simple Randomization, $p_z = 0.5$

	`ivreg` Y $(D = Z)$	`ivreg` Y $(D = Z)$ V	`ivreg` Y $(D = Z)$ V $Z * (V - \bar{V})$
Monte Carlo Standard Err	19.49	13.58	13.58
Avg Standard Err	19.49	13.64	13.63
Avg Length of CI	76.40	53.46	53.45

The first rows of Tables 13 and 14 show the average monte carlo standard errors which are the standard deviations of the estimates across 20000 monte carlo simulations. The next two rows show the LATE average standard errors and average confidence interval length under covariate adaptive block and simple randomization with $p_z = 0.7$ using nominal robust 2SLS standard errors for $\hat{\beta}_1$ and $\hat{\beta}_2$ and the analytic correction for $\hat{\beta}_3$. The number of observations is $n = 2000$.

For covariate adaptive block randomization, the monte carlo standard error is the same for all three estimators. However, just like in Table 9, nominal 2SLS standard errors for the unadjusted estimator is very conservative, while standard errors for the uninteracted and interacted estimators are not.

For covariate adaptive simple randomization, the monte carlo standard error is highest for the unadjusted model, decreases when we add the uninteracted covariate, and is the lowest for the interacted model. Nominal 2SLS and analytically corrected standard errors are not conservative and exhibit the same pattern as the monte carlo standard errors.

Tabelle 13 LATE with Covariate Adaptive Block Randomization, $p_z = 0.7$

	`ivreg` $Y\ (D = Z)$	`ivreg` $Y\ (D = Z)\ V$	`ivreg` $Y\ (D = Z)\ V\ Z * (V - \bar{V})$
Monte Carlo Standard Err	14.94	14.93	14.93
Avg Standard Err	20.72	14.95	14.87
Avg Length of CI	81.21	58.60	58.27

Tabelle 14 LATE with Covariate Adaptive Simple Randomization, $p_z = 0.7$

	`ivreg` $Y\ (D = Z)$	`ivreg` $Y\ (D = Z)\ V$	`ivreg` $Y\ (D = Z)\ V\ Z * (V - \bar{V})$
Monte Carlo Standard Err	20.91	15.03	14.95
Avg Standard Err	20.75	14.95	14.87
Avg Length of CI	81.35	58.62	58.27

6 Empirical Application

We investigate the efficiency improvement in ATE estimates after including the interaction between the exogenous treatment and the demeaned covariate in a dataset with over 2 million observations. Each observation is an expiry auction for a particular domain name listed on GoDaddy, an online platform where domain names which are no longer maintained by an individual are auctioned off in an open-bid English auction with a minimum bid of $12 and a duration of approximately 10 days. One interesting fact about these auctions is that the majority of participants are speculators who have no intrinsic use of the domain name except turning a profit when they resell the name in an aftermarket. Another interesting fact is that very few of the English auctions result in sale, partly due to the sheer volume of domain names that are listed for sale. For example, only 1.3% of auctions with a start time on or after May 12th, 2017 and before July 11th, 2017 had bids at or above the minimum bid.

Starting on May 12th, 2017, GoDaddy implemented a simple randomized experiment where some domain names would receive a valuation metric provided by a machine learning algorithm using deep learning. The idea was to provide auction participants with a better sense of the value of a domain name using features such as the length of the domain name, how many words in the domain name are part of the English dictionary, and whether the domain name is a .com, .net, or .org. The algorithm performed better than many existing approaches for predicting whether the domain name would sell and if so, at what price. At the start of the experiment, the treatment probability was 50%, but starting June 1st, 2017, it became 75%. Then on July 11th, 2017, the treatment probability became 100%.

In table 15, we look at the average treatment effect of including the valuation on the sale price conditional on at least one bidder meeting the minimum bid requirement for auctions with start times between May 12th, 2017 and June 1st, 2017. Of the 812026 auctions which occurred during this time frame, only 9283 auctions met the minimum bid requirement. In the first column, we estimate the ATE without any covariates to be 1.6754 with a standard error of 6.1526. In the second column, we add different combinations of the following covariates: the length of the domain name, whether the top level domain is .com, .net, or .org, and whether the domain name contains any words in the English dictionary. After including all of these covariates, the ATE

becomes 0.7859 and the standard error increases to 6.1966. In the third column, we include the covariates and the interactions between the treatment and the demeaned covariates. The ATE becomes 0.7894, and the standard error decreases to 6.1301.

In table 16, we look at the average treatment effect of including the valuation on the sale price conditional on at least one bidder meeting the minimum bid requirement for auctions with start times between June 1st, 2017 and July 11th, 2017. Of the 1366161 auctions in this time frame, only 19165 auctions met the minimum bid requirement. In the first column, we estimate the ATE without any covariates to be 12.6569 with a standard error of 4.4770. In the second column, we add different combinations of the following covariates: the length of the domain name, whether the top level domain is .com, .net, or .org, and whether the domain name contains any words in the English dictionary. After including all of these covariates, the ATE becomes 14.3432 and the standard error increases to 4.6244. In the third column, we include the covariates and the interactions between the treatment and the demeaned covariates. The ATE becomes 13.1778, and the standard error decreases to 4.4617.

In table 17, we look at the average treatment effect of including the valuation on the sale price conditional on at least one bidder meeting the minimum bid requirement for auctions with start times between May 12th, 2017 and July 11th, 2017. Of the 2178187 auctions in this time frame, only 28448 auctions met the minimum bid requirement. In the first column, we estimate the ATE without any covariates to be 6.6893 with a standard error of 3.8301. In the second column, we add different combinations of the following covariates: the length of the domain name, whether the top level domain is .com, .net, or .org, and whether the domain name contains any words in the English dictionary. After including all of these covariates, the ATE becomes 6.7003 and the standard error increases to 3.8848. In the third column, we include the covariates and the interactions between the treatment and the demeaned covariates. The ATE becomes 6.4064, and the standard error decreases to 3.8183.

Tabelle 15 ATE of valuation on sale price conditional on entry for 9283 auctions from 5.12.17 to 6.01.17

	reg Y D	reg Y D X	reg Y D X D*$(X - \bar{X})$
X=length of domain name			
$\hat{\beta}$	1.6754	1.0673	1.0682
$se\left(\hat{\beta}\right)$	6.1526	6.1325	6.1402
X=length,is.com,is.net,is.org			
$\hat{\beta}$	1.6754	0.8519	0.8584
$se\left(\hat{\beta}\right)$	6.1526	6.1165	6.1301
X=length,is.com,is.net,is.org,contains English word			
$\hat{\beta}$	1.6754	0.7859	0.7894
$se\left(\hat{\beta}\right)$	6.1526	6.1966	6.1301

Tabelle 16 ATE of valuation on sale price conditional on entry for 19165 auctions from 6.01.17 to 7.11.17

	reg Y D	reg Y D X	reg Y D X D*$(X-\bar{X})$
X=length of domain name			
$\hat{\beta}$	12.6569	14.1856	13.8739
$se\left(\hat{\beta}\right)$	4.4770	4.5152	4.4655
X=length,is.com,is.net,is.org			
$\hat{\beta}$	12.6569	14.8170	13.8605
$se\left(\hat{\beta}\right)$	4.4770	4.5488	4.4627
X=length,is.com,is.net,is.org,contains English word			
$\hat{\beta}$	12.6569	14.3432	13.1778
$se\left(\hat{\beta}\right)$	4.4770	4.6244	4.4617

Tabelle 17 ATE of valuation on sale price conditional on entry for 28448 auctions from 5.12.17 to 7.11.17

	reg Y D	reg Y D X	reg Y D X D*$(X-\bar{X})$
X=length of domain name			
$\hat{\beta}$	6.6893	7.2388	7.1763
$se\left(\hat{\beta}\right)$	3.8301	3.8316	3.8226
X=length,is.com,is.net,is.org			
$\hat{\beta}$	6.6893	6.9263	6.7882
$se\left(\hat{\beta}\right)$	3.8301	3.8377	3.8186
X=length,is.com,is.net,is.org,contains English word			
$\hat{\beta}$	6.6893	6.7003	6.4064
$se\left(\hat{\beta}\right)$	3.8301	3.8848	3.8183

7 Conclusion

This paper has compared the relative efficiencies of different types of OLS and 2SLS estimators in randomized or conditionally randomized experiments. Although the results are presented in the context of (local) average treatment effects, they can be generalized to nonlinear parameters including quantile treatment effects. Further extensions include propensity score regression and regression discontinuity models.

Literatur

Bugni, F., I. A. Canay, and A. M. Shaikh (2017a): "Inference under covariate-adaptive randomization," Tech. rep.

——— (2017b): "Inference under covariate-adaptive randomization with Multiple Treatments," Tech. rep.

Chen, X., H. Hong, and A. Tarozzi (2008): "Semiparametric efficiency in GMM models with auxiliary data," *Annals of Statistics*, 36, 808–843.
Freedman, D. A. (1981): "Bootstrapping regression models," *The Annals of Statistics*, 9, 1218–1228.
——— (2008): "On regression adjustments to experimental data," *Advances in Applied Mathematics*, 40, 180–193.
Frolich, M. (2006): "Nonparametric IV estimation of local average treatment effects with covariates," *Journal of Econometrics*, 139, 35–75.
Hahn, J. (1998): "On the Role of Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects," *Econometrica*, 66, 315–332.
Hong, H. and D. Nekipelov (2010): "Semiparametric efficiency in nonlinear LATE models," *Quantitative Economics*, 1, 279–304.
Imbens, G. and J. Angrist (1994): "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62, 467–475.
Imbens, G. W. and D. B. Rubin (2015): *Causal inference in statistics, social, and biomedical sciences*, Cambridge University Press.
Imbens, G. W. and J. M. Wooldridge (2009): "Recent Developments in the Econometrics of Program Evaluation," *Journal of Economic Literature*, 47, 5–86.
Lin, W. (2013): "Agnostic notes on regression adjustments to experimental data: Reexamining Freedman critique," *The Annals of Applied Statistics*, 7, 295–318.
Newey, W. (1994): "The Asymptotic Variance of Semiparametric Estimators," *Econometrica*, 62, 1349–82.
——— (1997): "Convergence Rates and Asymptotic Normality for Series Estimators," *Journal of Econometrics*, 79, 147–168.
Rosenbaum, P. and D. Rubin (1983): "The central role of the propensity score in observational studies for causal effects," *Biometrika*, 70, 41.
Shao, J., X. Yu, and B. Zhong (2010): "A theory for testing hypotheses under covariate-adaptive randomization," *Biometrika*, 97, 347–360.

A Proof of Theorem 1

For $k = 1, 2, 3$, let W_{ik} denote the instruments, let V_{ik} and U_{ik} denote the regressors, and let θ_k denote the parameters of the instrumental variable moment condition $EW_{ik}\left(Y_i - V_{ik}'\theta_{0k}\right) = 0$. The estimators $\hat{\theta}_k$ are defined by the sample estimating equations:

$$\frac{1}{n}\sum_{i=1}^{n} W_{ik}\left(Y_i - V_{ik}'\hat{\theta}_k\right) + \frac{1}{n}\sum_{i=1}^{n} W_{ik}U_{ik}\hat{\phi}'\left(\bar{X} - \mu_x\right) = 0. \tag{13}$$

For $\hat{\beta}_1$, let $U_{i1} = 0$, $p_z = P\left(Z_i = 1\right)$, $p_d = P\left(D_i = 1\right)$, $W_{i1} = \left(1 \;\; Z_i - p_z\right), V_{i1} = \left(1 \;\; D_i - p_d\right)$, and $\theta_1 = (\alpha, \beta)$. For $\hat{\beta}_2$, let $U_{i2} = 0$, $W_{i2} = \left(1 \;\; Z_i - p_z \;\; X_i - \mu_x\right)$ $V_{i2} = \left(1 \;\; D_i - p_d \;\; X_i - \mu_x\right)$, $\theta_2 = (\alpha, \beta, \eta)$. For $\hat{\beta}_3$, let $U_{i3} = Z_i - p_z, \theta_3 = (\alpha, \beta, \eta, \phi), W_{i3} = \left(1 \;\; Z_i - p_z \;\; X_i - \mu_x \;\; \left(Z_i - p_z\right)\left(X_i - \mu_x\right)\right)$, and $V_{i3} = \left(1 \;\; D_i - p_d \;\; X_i - \mu_x \;\; \left(Z_i - p_z\right)\left(X_i - \mu_x\right)\right)$. (13) leads to the following influence function representation of :

$$\sqrt{n}\left(\hat{\theta}_k - \theta_{0k}\right) = \left(EW_{ik}V_{ik}\right)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(W_{ik}\left(Y_i - V_{ik}'\theta_0\right) + EW_{ik}U_{ik}\phi_0\left(X_i - \mu_x\right)\right) + o_P\left(1\right).$$

It can be calculated that the second row of $E\left(W_{ik}V_{ik}\right)^{-1}, k = 1, 2, 3$ takes the forms of

$$\begin{pmatrix} 0 & Cov\left(D, Z\right)^{-1} \end{pmatrix} \quad \begin{pmatrix} 0 & Cov\left(D, Z\right)^{-1} & 0 \end{pmatrix} \quad \begin{pmatrix} 0 & Cov\left(D, Z\right)^{-1} & 0 & 0 \end{pmatrix}.$$

Therefore for $k = 1, 2, 3$,

$$\begin{aligned}\sqrt{n}\left(\hat{\beta}_k - \beta_0\right) =& Cov\left(Z, D\right)^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_k\left(Y_i, Z_i, X_i, W_{ik}, V_{ik}, U_{ik}\right) + o_P\left(1\right), \\ \psi_k\left(Y_i, Z_i, X_i, W_{ik}, U_{ik}, V_{ik}\right) =& \underbrace{\left(Z_i - p_z\right)\left(Y_i - V_{ik}'\theta_0\right)}_{\psi_{ik}^1} + \underbrace{E\left[\left(Z_i - p_z\right)U_{ik}\right]\phi_0\left(X_i - \mu_x\right)}_{\psi_{ik}^2} \equiv \psi_{ik}\end{aligned} \tag{14}$$

Consequently, $\sqrt{n}\left(\hat{\beta}_k - \beta_0\right) \xrightarrow{d} N\left(0, Cov(D, Z)^{-2} Var\left(\psi_{ik}\right)\right)$. It remains to show that for $j = 1, 2$, $Var\left(\psi_{i3}\right) \leq Var\left(\psi_{ij}\right)$. This can be done by showing that $Cov\left(\psi_{ij} - \psi_{i3}, \psi_{i3}\right) = 0$. For this purpose, consider first $j = 1$. Note that

$$\begin{aligned}\psi_{i1} - \psi_{i3} &= \left(Z_i - p_z\right)\left[\eta_0'\left(X_i - \mu_x\right) + \phi_0'\left(X_i - \mu_x\right)\left(Z_i - p_z\right)\right] - \left(p_z - p_z^2\right)\left(X_i - \mu_x\right)'\phi \\ &= \underbrace{\left(Z_i - p_z\right)\eta_0'\left(X_i - \mu_x\right)}_{\Delta\psi_{i13}^1} + \underbrace{\left(Z_i - p_z\right)^2\phi_0'\left(X_i - \mu_x\right)}_{\Delta\psi_{i13}^2} - \underbrace{\left(p_z - p_z^2\right)\left(X_i - \mu_x\right)'\phi}_{\Delta\psi_{i13}^3}.\end{aligned} \tag{15}$$

It follows from $Z_i^2 = Z_i$ and $EW_{i3}\left(Y_i - V_{i3}'\theta_0\right) = 0$ that

$$Cov\left(\psi_{i3}^1, \Delta\psi_{i13}^k\right) = 0, \quad k = 1, 2, 3$$

By independence of Z_i from X_i, $Cov\left(\psi_{i3}^2, \Delta\psi_{i13}^1\right) = 0$. Finally, we check that $Cov\left(\psi_{i3}^2, \Delta\psi_{i13}^2\right) = \left(p_z - p_z^2\right)^2\phi_0' Var\left(X\right)\phi_0$, and $Cov\left(\psi_{i3}^2, \Delta\psi_{i13}^3\right) = \left(p_z - p_z^2\right)^2\phi_0' Var\left(X\right)\phi_0$, so that

$$Cov\left(\psi_{i3}^2, \Delta\psi_{i13}^2 - \Delta\psi_{i13}^3\right) = 0.$$

We have verified that $Cov\left(\Delta\psi_{i13}, \psi_{i3}\right) = 0$, and $\hat{\beta}_3$ is more efficient than $\hat{\beta}_1$ asymptotically.

Next turn to $\hat{\beta}_2$ and $\psi_{i2} = \left(Z_i - p_z\right)\left(Y_i - \alpha_0 - \beta_0\left(D_i - p_d\right) - \eta_0'\left(X_i - \mu_x\right)\right)$. We want to show $Var\left(\psi_{i3}\right) \leq Var\left(\psi_{i2}\right)$ by verifying that $Cov\left(\Delta\psi_{i23}, \psi_{i3}\right) = 0$, where

$$\begin{aligned}\Delta\psi_{i23} &= \left(Z_i - p_z\right)^2\phi_0'\left(X_i - \mu_x\right) - \left(p_z - p_z^2\right)\phi_0'\left(X_i - \mu_x\right) \\ &= \underbrace{\left(\left(1 - 2p_z\right)Z_i + p_z^2\right)\phi_0'\left(X_i - \mu_x\right)}_{\Delta\psi_{i23}^1} - \underbrace{\left(p_z - p_z^2\right)\phi_0'\left(X_i - \mu_x\right)}_{\Delta\psi_{i23}^2}.\end{aligned} \tag{16}$$

By the moment conditions $EW_{i3}\left(Y_i - V_{i3}'\theta_0\right) = 0$,

$$Cov\left(\psi_{i3}^1, \Delta\psi_{i23}^k\right) = 0, \quad , k = 1, 2.$$

By independence between Z and X

$$Cov\left(\psi_{i3}^2, \Delta\psi_{i23}^1\right) = \left(p_z - p_z^2\right)^2\phi_0' Var\left(X\right)\phi_0.$$

Therefore since also $Cov\left(\psi_{i3}^2, \Delta\psi_{i23}^2\right) = \left(p_z - p_z^2\right)^2\phi_0' Var\left(X\right)\phi_0$, it follows that

$$Cov\left(\psi_{i3}^2, \Delta\psi_{i23}^1 - \Delta\psi_{i23}^2\right) = 0.$$

So that $Cov\left(\Delta\psi_{i23}, \psi_{i3}\right) = 0$, and $Var\left(\psi_{i3}\right) \leq Var\left(\psi_{i1}\right)$; $\hat{\beta}_3$ is also more efficient than $\hat{\beta}_2$.

However, there is no efficiency ranking between $\hat{\beta}_1$ and $\hat{\beta}_2$. Note that

$$\Delta\psi_{i12} \equiv \psi_{i1} - \psi_{i2} = \left(Z_i - p_z\right)\eta_0'\left(X_i - \mu_x\right).$$

There is no guarantee of either $Cov\left(\Delta\psi_{i12},\psi_{i2}\right)=0$ or $Cov\left(\Delta\psi_{i12}\left(W\right),\psi_{i1}\left(W\right)\right)=0$. This is because the moment conditions for $\hat{\beta}_2$ do not impose that

$$EZX\left(Y-\alpha_0-\beta_0 D-\eta_0' X\right)=0,$$

and the moment conditions for $\hat{\beta}_1$ do not impose

$$EZX\left(Y-\alpha_0-\beta_0 D\right)=0 \quad \text{or} \quad EX\left(Y-\alpha_0-\beta_0 D\right)=0.$$

B Proof of Corollary 1

Under the causal model, the parameter β_0 and the influence functions for $\hat{\beta}$ can be written using the counterfactuals. Recall that $\beta_0=E\left(Y_1-Y_0|D_1>D_0\right)=E\left(Y_1^*-Y_0^*\right)/E\left(D_1-D_0\right)$. Define $t_1=Y_1^*-\beta_0 D_1$, $t_0=Y_0^*-\beta_0 D_0$. Then

$$\begin{aligned}\alpha_0-\beta_0 p_d=&EY-\beta_0 ED=p_z Et_1+\left(1-p_z\right)Et_0\\ \psi_1=&\left(Z-p_z\right)\left(Y-\alpha_0-\beta_0\left(D-p_d\right)\right)\\ =&\left(Z-p_z\right)\left(\left(1-p_z\right)\left(t_1-Et_1\right)+p_z\left(t_0-Et_0\right)\right)+p_z\left(1-p_z\right)\left(t_1-t_0\right),\end{aligned}$$

where by definition $Et_1-Et_0=0$. Next consider $\hat{\beta}_2$. It follows from the 3rd moment equation $E\left(X-\mu_x\right)\left(Y-\alpha_0-\beta_0\left(D-p_d\right)-\eta_0\left(X-\mu_X\right)\right)=0$ that

$$\eta_0=Var\left(X\right)^{-1}Cov\left(X,Y-\beta_0 D\right)=Var\left(X\right)^{-1}\left[p_z Cov\left(X,t_1\right)+\left(1-p_z\right)Cov\left(X,t_0\right)\right],$$

and that

$$\begin{aligned}\psi_2=&\left(Z-p_z\right)\left(Y-\alpha_0-\beta_0\left(D-p_d\right)-\eta_0\left(X-\mu_x\right)\right)\\ =&\left(Z-p_z\right)\left(\left(1-p_z\right)\left(t_1-Et_1\right)+p_z\left(t_0-Et_0\right)-\eta_0\left(X-\mu_x\right)\right)+p_z\left(1-p_z\right)\left(t_1-t_0\right),\end{aligned}$$

Next consider $\hat{\beta}_3$. It follows from the 4th moment condition

$$E\left(Z-p_z\right)\left(X-\mu_x\right)\left(Y-\alpha_0-\beta_0\left(D-p_d\right)-\eta_0\left(X-\mu_X\right)-\phi_0\left(Z-p_z\right)\left(X-\mu_x\right)\right)=0$$

that $\phi_0=Var\left(X\right)^{-1}Cov\left(X,t_1-t_0\right)$. Therefore,

$$\begin{aligned}\psi_3^1=&\left(Z-p_z\right)\left(Y-\alpha_0-\beta_0\left(D-p_d\right)-\eta_0'\left(X-\mu_x\right)-\phi_0'\left(Z-p_z\right)\left(X-\mu_x\right)\right)\\ =&\left(Z-p_z\right)\Bigg(\left(1-p_z\right)\left(t_1-Et_1-Cov\left(t_1,X\right)Var\left(X\right)^{-1}\left(X-\mu_x\right)\right)\\ &+p_z\left(t_0-Et_0-Cov\left(t_0,X\right)Var\left(X\right)^{-1}\left(X-\mu_x\right)\right)\Bigg)\\ &+p_z\left(1-p_z\right)\left(\left(t_1-t_0\right)-Cov\left(t_1-t_0,X\right)Var\left(X\right)^{-1}\left(X-\mu_x\right)\right)\end{aligned}$$

and $\psi_3^2=p_z\left(1-p_z\right)Cov\left(t_1-t_0,X\right)Var\left(X\right)^{-1}\left(X-\mu_x\right)$. Therefore

$$\begin{aligned}\psi_3=&\left(Z-p_z\right)\left(Y-\alpha_0-\beta_0\left(D-p_d\right)-\eta_0'\left(X-\mu_x\right)-\phi_0'\left(Z-p_z\right)\left(X-\mu_x\right)\right)\\ =&\left(Z-p_z\right)\Bigg(\left(1-p_z\right)\left(t_1-Et_1-Cov\left(t_1,X\right)Var\left(X\right)^{-1}\left(X-\mu_x\right)\right)\\ &+p_z\left(t_0-Et_0-Cov\left(t_0,X\right)Var\left(X\right)^{-1}\left(X-\mu_x\right)\right)\Bigg)+p_z\left(1-p_z\right)\left(t_1-t_0\right)\end{aligned}$$

Using $Z\perp\left(t_1,t_0,X\right)$, it can then be verified that

$$Cov\left(\psi_1-\psi_3,\psi_3\right)=0 \quad \text{and} \quad Cov\left(\psi_2-\psi_3,\psi_3\right)=0.$$

In the special case when $D = Z$, $t_1 = Y_1 - \beta_0$, $t_0 = Y_0$, $\beta_0 = E(Y_1 - Y_0)$, then

$$\begin{aligned}\psi_3 =& (Z - p_z)\left((1-p_z)\left(Y_1 - EY_1 - Cov(Y_1, X)\, Var(X)^{-1}(X - \mu_x)\right)\right.\\ &\left. + p_z\left(Y_0 - EY_0 - Cov(Y_0, X)\, Var(X)^{-1}(X - \mu_x)\right)\right) + p_z(1-p_z)(Y_1 - Y_0 - \beta_0)\\ \psi_2 =& (Z - p_z)\left((1-p_z)(Y_1 - EY_1)\right.\\ &\left. + p_z(Y_0 - EY_0) - \eta_0(X - \mu_x)\right) + p_z(1-p_z)(Y_1 - Y_0 - \beta_0).\end{aligned} \tag{17}$$

for $\eta_0 = Var(X)^{-1}[p_z Cov(X, Y_1) + (1-p_z)\, Cov(X, Y_0)]$.

C Proof of Corollary 2

Replace μ_x by $\mu_{xs} = EX_s$. Then it can be shown that $\hat{\beta}_3$ is more efficient than $\hat{\beta}_4$. Similar calculations as those for $\hat{\beta}_3$ show that

$$\begin{aligned}\sqrt{n}\left(\hat{\beta}_4 - \beta_0\right) =& Cov(D, Z)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_{i4} + o_P(1), \quad \text{where}\\ \psi_{i4} =& (Z_i - p_z)\left(Y_i - \rho_0 - \beta_0(D_i - p_d) - \eta_{0s}'(X_{si} - \mu_{xs}) - \phi_{0s}'(X_{si} - \mu_{xs})(Z_i - p_z)\right)\\ &+ \left(p_z - p_z^2\right)\phi_{0s}'(X_{si} - \mu_{xs})\end{aligned}$$

Then we can write, for $\bar{\eta}_0, \bar{\phi}_0$ possibly different from both η_0, ϕ_0 and η_{0s}, ϕ_{0s},

$$\begin{aligned}\Delta\psi_{i43} &= \psi_{i4} - \psi_{i3}\\ &= (Z_i - p_z)\left[\eta_0'(X_i - \mu_x) - \eta_{0s}'(X_{si} - \mu_{xs}) + \left(\phi_0'(X_i - \mu_x) - \phi_{0s}'(X_{si} - \mu_{xs})\right)(Z_i - p_z)\right]\\ &\quad + \left(p_z - p_z^2\right)(X_{si} - \mu_{xs})'\phi_{0s} - \left(p_z - p_z^2\right)(X_i - \mu_x)'\phi_0\\ &= (Z_i - p_z)\left[\bar{\eta}_0'(X_i - \mu_x) + \bar{\phi}_0'(X_i - \mu_x)(Z_i - p_z)\right] - \left(p_z - p_z^2\right)(X_i - \mu_x)'\bar{\phi}_0\\ &= \underbrace{(Z_i - p_z)\bar{\eta}_0'(X_i - \mu_x)}_{\Delta\psi_{i43}^1} + \underbrace{(Z_i - p_z)^2\bar{\phi}_0'(X_i - \mu_x)}_{\Delta\psi_{i43}^2} - \underbrace{\left(p_z - p_z^2\right)(X_i - \mu_x)'\bar{\phi}_0}_{\Delta\psi_{i43}^3}.\end{aligned}$$

It follows from $Z_i^2 = Z_i$, and the instrumental variable moment equations that

$$Cov\left(\psi_{i3}^1, \Delta\psi_{i43}^k\right) = 0, \quad , k = 1, 2, 3$$

By independence of Z_i and X_i, $Cov\left(\psi_{i3}^2, \Delta\psi_{i43}^1\right) = 0$. Finally, we check that

$$Cov\left(\psi_{i3}^2, \Delta\psi_{i43}^2\right) = \left(p_z - p_z^2\right)^2\phi_0' Var(X)\,\bar{\phi}_0,$$

and $Cov\left(\psi_{i3}^2, \Delta\psi_{i43}^3\right) = \left(p_z - p_z^2\right)^2\psi_0' Var(X)\,\bar{\psi}_0$. so that

$$Cov\left(\psi_{i3}^2, \Delta\psi_{i43}^2 - \Delta\psi_{i43}^3\right) = 0.$$

We have verified that $Cov(\Delta\psi_{i43}, \psi_{i3}) = 0$, and $\hat{\beta}_3$ is more efficient than $\hat{\beta}_4$ asymptotically. The same result can also be verified using the counter-factual model as in Corollary 1.

D Proof of Corollary 3

Note $\hat{A}_k^{-1}\hat{B}_k\hat{A}_k^{-1} \xrightarrow{p} \text{Var}\left(\Lambda_k \left(EW_{ik}V_{ik}'\right)^{-1} W_{ik}\left(Y_i - V_{ik}'\theta_{0k}\right)\right)$, where

$$\Lambda_1 = \begin{bmatrix} 1 & -p_d \\ 0 & 1 \end{bmatrix} \quad \Lambda_2 = \begin{bmatrix} 1 & -p_d & -\mu_x \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad \Lambda_3 = \begin{bmatrix} 1 & -p_d & -\mu_x & p_z\mu_x \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & -p_z \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

Using the spare structure of $EW_{ik}V_{ik}$, the $(2,2)$ elements of $A_k^{-1}B_kA_k^{-1}$ are then given by

$$\text{Var}\left(Cov\left(Z_i, D_i\right)^{-1}\left(Z_i - p_z\right)\left(Y_i - V_{ik}'\theta_{0k}\right)\right)$$

For $k = 1, 2$, this coincides with the asymptotic variance σ_k^2 in Theorem 1. Theorem 1 also shows the asymptotic variance of $\hat{\beta}_3$ as

$$\sigma_3^2 = \text{Var}\left(Cov\left(Z_i, D_i\right)^{-1}\left[\left(Z_i - p_z\right)\left(Y_i - V_{ik}'\theta_{0k}\right) + p_z\left(1 - p_z\right)\phi_0\left(X_i - \mu_x\right)\right]\right)$$

By the moment condition $E\left(Z_i - p_z\right)\left(X_i - \mu_x\right)\left(Y_i - V_{ik}'\theta_{0k}\right) = 0$, σ_3^2 is at least as large as

$$\text{plim}\hat{\sigma}_3^2 = \text{Var}\left(Cov\left(Z_i, D_i\right)^{-1}\left(Z_i - p_z\right)\left(Y_i - V_{ik}'\theta_{0k}\right)\right) \tag{18}$$

A similar calculation shows that $\bar{\sigma}_3^2 \xrightarrow{p} \sigma_3^2$. Of course one can also bootstrap.

E Proof of Proposition 1

Consider first the case of $D = Z$. For $\alpha + \beta = E\left(Y|Z = 1\right)$, $\alpha = E\left(Y|Z = 0\right)$, and $\mu_x = EX$, the moment conditions are $E\phi_i\left(\alpha_0, \beta_0, \mu_{0x}\right) = 0$, where

$$\phi_i\left(\alpha, \beta, \mu_x\right)' = \left(Z_i\left(Y_i - \alpha - \beta\right),\ Z_i\left(X_i - \mu_x\right),\ \left(1 - Z_i\right)\left(Y_i - \alpha\right),\ \left(1 - Z_i\right)\left(X_i - \mu_x\right)\right),$$

such that for $A_{11} = Var\left(Y_i|Z_i = 1\right)$, $A_{12} = Cov\left(Y_i, X_i|Z_i = 1\right) = A_{21}'$, $B_{11} = Var(Y_i|Z_i = 0)$, $B_{12} = Cov\left(Y_i, X_i|Z_i = 0\right) = B_{21}'$, $A_{22} = B_{22} = Var\left(X_i\right)$,

$$Var\left(\phi_i\left(\cdot\right)\right) = \begin{pmatrix} p_zA & 0 \\ 0 & \left(1 - p_z\right)B \end{pmatrix} \quad A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} \quad B = \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix} \tag{19}$$

Then $\widehat{\text{Var}}\left(\phi_i\left(\cdot\right)\right)$ is similar to $Var\left(\phi\left(\cdot\right)\right)$ with p_z, A, B replaced by $\hat{p}_z, \hat{A}, \hat{B}$.

An application of the partitioned matrix inversion formula shows that the solution to (5) is given by, for $F_2 = \left(A_{22} - A_{21}A_{11}^{-1}A_{12}\right)^{-1}$ and $G_2 = \left(B_{22} - B_{21}B_{11}^{-1}B_{12}\right)^{-1}$,

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} Z_i\left(Y_i - \alpha - \beta\right) - \hat{A}_{12}\hat{A}_{22}^{-1}\frac{1}{n}\sum_{i=1}^{n} Z_i\left(X_i - \mu_x\right) = 0 \\
&\frac{1}{n}\sum_{i=1}^{n}\left(1 - Z_i\right)\left(Y_i - \alpha\right) - \hat{B}_{12}\hat{B}_{22}^{-1}\frac{1}{n}\sum_{i=1}^{n}\left(1 - Z_i\right)\left(X_i - \mu_x\right) = 0 \\
&- \hat{F}_2A_{21}A_{11}^{-1}\frac{1}{n}\sum_{i=1}^{n} Z_i\left(Y_i - \alpha - \beta\right) + F_2\frac{1}{n}\sum_{i=1}^{n} Z_i\left(X_i - \mu_x\right) \\
&- \hat{G}_2\hat{B}_{21}\hat{B}_{11}^{-1}\frac{1}{n}\sum_{i=1}^{n}\left(1 - Z_i\right)\left(Y_i - \alpha\right) + \hat{G}_2\frac{1}{n}\sum_{i=1}^{n}\left(1 - Z_i\right)\left(X_i - \mu_x\right) = 0.
\end{aligned} \tag{20}$$

Substitute the first two equations into the third and simplify to

$$\hat{A}_{22}^{-1}\frac{1}{n}\sum_{i=1}^{n}Z_i\left(X_i-\mu_x\right)+\hat{B}_{22}^{-1}\frac{1}{n}\sum_{i=1}^{n}\left(1-Z_i\right)\left(X_i-\mu_x\right)=0 \tag{21}$$

Since $\hat{A}_{22}=Var\left(X_i\right)+O_P\left(\frac{1}{\sqrt{n}}\right)=\hat{B}_{22}+O_P\left(\frac{1}{\sqrt{n}}\right)$, this can be used to show that $\hat{\mu}_x=\bar{X}+o_P\left(\frac{1}{\sqrt{n}}\right)$. And then

$$\hat{\alpha}+\hat{\beta}=\left(\sum_{i=1}^{n}Z_iY_i-\hat{A}_{12}\hat{A}_{22}^{-1}\sum_{i=1}^{n}Z_i\left(X_i-\mu_x\right)\right)/\sum_{i=1}^{n}Z_i+o_P\left(\frac{1}{\sqrt{n}}\right)$$

$$\hat{\alpha}=\left(\sum_{i=1}^{n}\left(1-Z_i\right)Y_i-\hat{B}_{12}\hat{B}_{22}^{-1}\sum_{i=1}^{n}\left(1-Z_i\right)\left(X_i-\mu_x\right)\right)/\sum_{i=1}^{n}\left(1-Z_i\right)+o_P\left(\frac{1}{\sqrt{n}}\right)$$

Up to $o_P\left(\frac{1}{\sqrt{n}}\right)$ terms, these are the intercept terms in separate regressions of Y_i on $X_i-\bar{X}$ among the control and treatment groups.

These calculations can be extended to the LATE GMM model in (5), where we now define $A_{11}=Var\left(Y-\alpha_0-\beta_0D|Z=1\right)$, $A_{12}=Cov\left(Y-\alpha_0-\beta_0D,X|Z=1\right)=A_{21}'$, $B_{11}=Var(Y-\alpha_0-\beta_0D|Z=0)$, $B_{12}=Cov\left(Y-\alpha_0-\beta_0D,X|Z=0\right)=B_{21}'$, $A_{22}=B_{22}=Var\left(X\right)$, and let $\hat{A}_{jk},\hat{B}_{jk}$ denote their $\sqrt{n}$ consistent estimates. Then (19) and (21) both continue to hold, leading to $\hat{\mu}_x=\bar{X}+o_P\left(\frac{1}{\sqrt{n}}\right)$. The first two equations in (20) now become

$$\frac{1}{n}\sum_{i=1}^{n}Z_i\left(Y_i-\alpha-\beta D_i\right)-\hat{A}_{12}\hat{A}_{22}^{-1}\frac{1}{n}\sum_{i=1}^{n}Z_i\left(X_i-\mu_x\right)=0$$

$$\frac{1}{n}\sum_{i=1}^{n}\left(1-Z_i\right)\left(Y_i-\alpha-\beta D_i\right)-\hat{B}_{12}\hat{B}_{22}^{-1}\frac{1}{n}\sum_{i=1}^{n}\left(1-Z_i\right)\left(X_i-\mu_x\right)=0,$$

Note that given α and β, $\hat{A}_{12}\hat{A}_{22}^{-1}$ and $\hat{B}_{12}\hat{B}_{22}^{-1}$ are precisely the profiled $\hat{\phi}$ and $\hat{\eta}$ implied by the estimating equations (13) for β_3. In other words, the above two equations are the concentrated estimating equations for α and β implied by (13).

F Proof of Proposition 2

Let $W_{1i}=\left(Z_i,\ Z_iV_i\right)^T$ and $W_{0i}=\left(\left(1-Z_i\right),\ \left(1-Z_i\right)V_i\right)^T$. Then the normal equations corresponding to (9) are

$$\frac{1}{n}\sum_{i=1}^{n}W_{1i}\left(Y_i-\hat{\gamma}_1-\hat{\vartheta}_1V_i\right)=0,\quad\frac{1}{n}\sum_{i=1}^{n}W_{0i}\left(Y_i-\hat{\gamma}_0-\hat{\vartheta}_0V_i\right)=0.$$

$$\frac{1}{n}\sum_{i=1}^{n}W_{1i}\left(D_i-\hat{\tau}_1-\hat{\zeta}_1V_i\right)=0,\quad\frac{1}{n}\sum_{i=1}^{n}W_{0i}\left(D_i-\hat{\tau}_0-\hat{\zeta}_0V_i\right)=0.$$

Taking a linear combination using $\hat{\beta}_{AL}=\widehat{AvgLATE}$,

$$\frac{1}{n}\sum_{i=1}^{n}W_{1i}\left(Y_i-\hat{\gamma}_1-\hat{\vartheta}_1V_i-\hat{\beta}_{AL}\left(D_i-\hat{\tau}_1-\hat{\zeta}_1V_i\right)\right)=0$$

$$\frac{1}{n}\sum_{i=1}^{n}W_{0i}\left(Y_i-\hat{\gamma}_0-\hat{\vartheta}_0V_i-\hat{\beta}_{AL}\left(D_i-\hat{\tau}_0-\hat{\zeta}_0V_i\right)\right)=0$$

We rearrange this into

$$\frac{1}{n}\sum_{i=1}^{n} W_{1i}\left(Y_i - \hat{\gamma}_1 + \hat{\beta}_{AL}\hat{\tau}_1 + \left(\hat{\beta}_{AL}\hat{\zeta}_1 - \hat{\vartheta}_1\right)\bar{V} - \hat{\beta}_{AL}D_i - \left(\hat{\beta}_{AL}\hat{\zeta}_1 - \hat{\vartheta}_1\right)\left(V_i - \bar{V}\right)\right) = 0$$
$$\frac{1}{n}\sum_{i=1}^{n} W_{0i}\left(Y_i - \hat{\gamma}_0 + \hat{\beta}_{AL}\hat{\tau}_0 + \left(\hat{\beta}_{AL}\hat{\zeta}_0 - \hat{\vartheta}_0\right)\bar{V} - \hat{\beta}_{AL}D_i - \left(\hat{\beta}_{AL}\hat{\zeta}_0 - \hat{\vartheta}_0\right)\left(V_i - \bar{V}\right)\right) = 0$$

By the definition in (10),

$$\hat{\nu} = \hat{\gamma}_1 - \hat{\beta}_{AL}\hat{\tau}_1 - \left(\hat{\beta}_{AL}\hat{\zeta}_1 - \hat{\vartheta}_1\right)\bar{V} = \hat{\gamma}_0 - \hat{\beta}_{AL}\hat{\tau}_0 - \left(\hat{\beta}_{AL}\hat{\zeta}_0 - \hat{\vartheta}_0\right)\bar{V}.$$

The normal equations therefore take the form of

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} W_{1i}\left(Y_i - \hat{\nu} - \hat{\beta}_{AL}D_i - \left(\hat{\beta}_{AL}\hat{\zeta}_1 - \hat{\vartheta}_1\right)\left(V_i - \bar{V}\right)\right) = 0 \\
&\frac{1}{n}\sum_{i=1}^{n} W_{0i}\left(Y_i - \hat{\nu} - \hat{\beta}_{AL}D_i - \left(\hat{\beta}_{AL}\hat{\zeta}_0 - \hat{\vartheta}_0\right)\left(V_i - \bar{V}\right)\right) = 0
\end{aligned} \tag{22}$$

Next, consider the normal equations determining the interactive $\hat{\beta}_\infty$. For $W_i = (W_{1i}, W_{0i})^T$,

$$\frac{1}{n}\sum_{i=1}^{n} W_i\left(Y_i - \hat{\alpha} - \hat{\beta}_\infty D_i - \hat{\eta}\left(V_i - \bar{V}\right) - \hat{\phi}Z_i\left(V_i - \bar{V}\right)\right) = 0.$$

This can be rewritten as

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} W_{1i}\left(Y_i - \hat{\alpha} - \hat{\beta}_\infty D_i - \left(\hat{\eta} + \hat{\phi}\right)\left(V_i - \bar{V}\right)\right) = 0 \\
&\frac{1}{n}\sum_{i=1}^{n} W_{0i}\left(Y_i - \hat{\alpha} - \hat{\beta}_\infty D_i - \hat{\eta}\left(V_i - \bar{V}\right)\right) = 0
\end{aligned} \tag{23}$$

Then (23) can be satisfied through (22) by setting

$$\hat{\alpha} = \hat{\nu}, \quad \hat{\beta}_\infty = \hat{\beta}_{AL}, \quad \hat{\eta} = \hat{\beta}_{AL}\hat{\zeta}_0 - \hat{\vartheta}_0, \quad \hat{\phi} = \hat{\beta}_{AL}\hat{\zeta}_1 - \hat{\vartheta}_1 - \hat{\eta}.$$

G Proof of Proposition 3

When $D = Z$, Hahn (1998) shows that $\sigma_\infty^2 = Var(\psi_\infty)$, where

$$\begin{aligned}
\psi_\infty =& \frac{D}{p}\left(Y_1 - E\left(Y_1|X\right)\right) - \frac{1-D}{1-p}\left(Y_0 - E\left(Y_0|X\right)\right) + \left(E\left(Y_1 - Y_0|X\right) - E\left(Y_1 - Y_0\right)\right) \\
=& (D-p)\left[\frac{Y_1 - E\left(Y_1|X\right)}{p} + \frac{Y_0 - E\left(Y_0|X\right)}{1-p}\right] + Y_1 - Y_0 - E\left(Y_1 - Y_0\right)
\end{aligned}$$

We can then use ψ_3 in the proof of Corollary 1 to show that

$$Cov\left(\psi_3/\left(p_z\left(1-p_z\right)\right) - \psi_\infty, \psi_\infty\right) = 0.$$

More generally when $Z \neq D$, the LATE efficiency bound was calculated in Frolich (2006) and Hong and Nekipelov (2010) (Lemma 1 and Thm 4), with $\sigma_\infty^2 = Var(\psi_\infty)$, and

$$\begin{aligned}\psi_\infty = \frac{1}{P(D_1 > D_0)}\Big\{ & \frac{Z}{p_z}(Y - E(Y|Z=1,X)) + E(Y|Z=1,X) \\ & - \frac{1-Z}{1-p_z}(Y - E(Y|Z=0,X)) - E(Y|Z=0,X) \\ & - \Big(\frac{Z}{p_z}(D - E(D|Z=1,X)) + E(D|Z=1,X) \\ & \qquad - \frac{1-Z}{1-p_z}(D - E(D|Z=0,X)) - E(D|Z=0,X)\Big)\beta\Big\},\end{aligned}$$

where $P(D_1 > D_0) = P(D=1|Z=1) - P(D=1|Z=0)$. We can rewrite this as

$$\begin{aligned}P(D_1 > D_0)\,\psi_\infty =& \frac{Z}{p_z}(t_1 - E(t_1|X)) - \frac{1-Z}{1-p_z}(t_0 - E(t_0|X)) + E(t_1 - t_0|X) \\ =& (Z - p_z)\left\{\frac{t_1 - E(t_1|X)}{p_z} + \frac{t_0 - E(t_0|X)}{1-p_z}\right\} + t_1 - t_0.\end{aligned} \tag{24}$$

Again comparing this to ψ_3 in the proof of Corollary 1 shows that

$$Cov\left(\psi_3 / \left(P(D_1 > D_0)\, p_z (1-p_z)\right) - \psi_\infty, \psi_\infty\right) = 0.$$

The comparison between ψ_2 and ψ_3 in (17) can also be understood in the context of doubly robust estimators, which use influence functions of the form similar to ψ_∞ but without requiring p_z to be constant. Define $Q(X) \equiv P(Z=1|X)$. In the case of $D = Z$,

$$\begin{aligned}\phi_\infty =& \frac{D}{Q(X)}(Y - E(Y_1|X)) - \frac{1-D}{1-Q(X)}(Y - E(Y_0|X)) + (E(\Delta Y|X) - E\Delta Y) \\ =& (D - Q(X))\left[\frac{Y_1 - E(Y_1|X)}{Q(X)} + \frac{Y_0 - E(Y_0|X)}{1-Q(X)}\right] + (Y_1 - Y_0 - \beta)\end{aligned}$$

The estimators with influence function ϕ_∞ are consistent as long as either $Q(X)$ or the pair of $E(Y_1|X), E(Y_0|X)$ are correctly specified. Under complete randomization and with $Q(X)$ specified as p_z, the P-score model is obviously correctly specified. Therefore $E(Y_1|X)$ and $E(Y_0|X)$, being linear projections on $(1 \quad V(X))$, have no effect on consistency. However, between two misspecified conditional mean models, the first pair in ψ_3 is a more efficient projection that induces a smaller variance than the linear projection in ψ_2. Similarly, in the general LATE case when $D \neq Z$, doubly robust estimators use influence functions of the form

$$\phi_\infty = \left(D - \tilde{Q}(X)\right)\left[\frac{t_1 - E(t_1|X)}{Q(X)} + \frac{t_0 - E(t_0|X)}{1-Q(X)}\right] + (t_1 - t_0 - \beta).$$

where $E(t_1|X) = E(Y_1^*|X) - \beta_0 E(D_1|X)$ and $E(t_0|X) = E(Y_0^*|X) - \beta_0 E(D_0|X)$. These estimators are consistent as long as either $Q(X)$ or the set of

$$E(Y_1^*|X), E(Y_0^*|X), E(D_1|X), E(D_0|X).$$

are correctly specified. Among different misspecified linear approximations to $E(t_1|X)$ and $E(t_0|X)$, the least square projection is more efficient.

Similar to (3) and (4), σ_∞^2 can be consistently estimated under suitable regularity conditions (such as those in Newey (1997)) by

$$\bar{\sigma}_\infty^2 = \widehat{Cov}_{Z,D}^{-2}\frac{1}{n}\sum_{i=1}^n \bar{\epsilon}_{i\infty}^2 \text{ where } \bar{\epsilon}_{i\infty} = (Z_i - \hat{p}_z)\,\hat{\epsilon}_{i\infty} + \hat{p}_z(1-\hat{p}_z)\,\hat{\phi}_\infty\left(V_i - \bar{V}\right)$$

and $\hat{\epsilon}_{i\infty} = Y_i - \hat{\alpha} - \hat{\beta}_\infty D_i - \hat{\eta}\left(V_i - \bar{V}\right) - \hat{\phi} Z_i \left(V_i - \bar{V}\right)$. If we write

$$Y_i - \hat{\beta}_\infty D_i = (1 - Z_i)\left(\hat{\alpha} + \hat{\eta}\left(V_i - \bar{V}\right)\right) + Z_i\left(\hat{\alpha} + \left(\hat{\eta} + \hat{\phi}\right)\left(V_i - \bar{V}\right)\right) + \hat{\epsilon}_{i\infty},$$

then we expect that uniformly in X_i,

$$\begin{aligned}\hat{\alpha} + \hat{\eta}\left(V_i - \bar{V}\right) =& E\left(Y - \beta D | Z = 0, X_i\right) + o_P(1) = E\left(t_{0i}|X_i\right) + o_P(1)\\ \hat{\alpha} + \left(\hat{\eta} + \hat{\phi}\right)\left(V_i - \bar{V}\right) =& E\left(Y - \beta D | Z = 1, X_i\right) + o_P(1) = E\left(t_{1i}|X_i\right) + o_P(1).\end{aligned}$$

Therefore $\hat{\phi}\left(V_i - \bar{V}\right) = E\left(t_{1i} - t_{0i}|X_i\right) + o_P(1)$, $\hat{\epsilon}_{i\infty} = Z_i\left(t_{1i} - E\left(t_{1i}|X_i\right)\right) + (1 - Z_i)\left(t_{0i} - E\left(t_{0i}|X_i\right)\right)$,

$$\bar{\epsilon}_{i\infty} = (Z_i - p_z)\left((1 - p_z)\left(t_{1i} - E\left(t_{1i}|X_i\right)\right) + p_z\left(t_{0i} - E\left(t_{0i}|X_i\right)\right)\right) + p_z(1 - p_z)\left(t_{1i} - t_{0i}\right) + o_P(1),$$

which coincides with the semiparametric asymptotic influence function, and includes the CI model as a special case when $D = Z$.

H Proof of Proposition 4

Recall that $\sqrt{n}\left(\hat{\beta}_1 - \beta_0\right) = Cov_n(Z, D)^{-1}\sqrt{n} Cov_n(Z, Y - D\beta_0)$. It can be shown that

$$\begin{aligned}Cov_n(Z, D) =& \frac{1}{n}\sum_{i=1}^n Z_i D_i - \frac{1}{n}\sum_{i=1}^n Z_i \frac{1}{n}\sum_{i=1}^n D_i\\ =& \hat{p}_z(1 - \hat{p}_z)\left[\frac{\frac{1}{n}\sum_{i=1}^n Z_i D_{1i}}{\hat{p}_z} - \frac{\frac{1}{n}\sum_{i=1}^n (1 - Z_i) D_{0i}}{1 - \hat{p}_z}\right]\\ =& p_z(1 - p_z) P(D_1 > D_0) + o_P(1).\end{aligned}$$

where the last line follows from Assumption 8.2. Furthermore,

$$Cov_n(Y, Z) = \frac{1}{n}\sum_{i=1}^n Z_i Y_i - \frac{1}{n}\sum_{i=1}^n Z_i \frac{1}{n}\sum_{i=1}^n Y_i = \hat{p}_z(1 - \hat{p}_z)\left[\frac{\frac{1}{n}\sum_{i=1}^n Z_i Y_{1i}^*}{\hat{p}_z} - \frac{\frac{1}{n}\sum_{i=1}^n (1 - Z_i) Y_{0i}^*}{1 - \hat{p}_z}\right]$$

Next we consider

$$\begin{aligned}&\sqrt{n}\left(Cov_n(Y, Z) - Cov_n(D, Z)\beta_0\right)\\ &= \hat{p}_z(1 - \hat{p}_z)\frac{1}{\sqrt{n}}\sum_{i=1}^n\left[\frac{Z_i t_{1i}}{\hat{p}_z} - \frac{(1 - Z_i) t_{0i}}{1 - \hat{p}_z}\right]\\ &= \hat{p}_z(1 - \hat{p}_z)\frac{1}{\sqrt{n}}\sum_{i=1}^n\left(\frac{Z_i t_{1i}}{\hat{p}_z} + \frac{Z_i t_{0i}}{1 - \hat{p}_z} - (t_{1i} - t_{0i}) - \frac{t_{0i}}{1 - \hat{p}_z}\right) + \hat{p}_z(1 - \hat{p}_z)\frac{1}{\sqrt{n}}\sum_{i=1}^n (t_{1i} - t_{0i})\\ &= \hat{p}_z(1 - \hat{p}_z)\frac{1}{\sqrt{n}}\sum_{i=1}^n (Z_i - \hat{p}_z)\left(\frac{t_{1i}}{\hat{p}_z} + \frac{t_{0i}}{1 - \hat{p}_z}\right) + \hat{p}_z(1 - \hat{p}_z)\frac{1}{\sqrt{n}}\sum_{i=1}^n (t_{1i} - t_{0i})\\ &= p_z(1 - p_z)\frac{1}{\sqrt{n}}\sum_{i=1}^n (Z_i - p_z)\left(\frac{t_{1i} - \mathbb{E}[t_{1i}]}{p_z} + \frac{t_{0i} - \mathbb{E}[t_{0i}]}{1 - p_z}\right) + p_z(1 - p_z)\frac{1}{\sqrt{n}}\sum_{i=1}^n (t_{1i} - t_{0i}) + R_n\end{aligned}$$

where

$$
\begin{aligned}
R_n = & \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ \left(Z_i - \hat{p}_z\right) \left(\left(p_z - \hat{p}_z\right) \left(t_{1i} - t_{0i}\right) + \left(\left(1 - p_z\right) \mathbb{E}\left[t_{1i}\right] + p_z \mathbb{E}\left[t_{0i}\right] \right) \right) \right\} \\
& + \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(p_z - \hat{p}_z\right) \left(\left(1 - p_z\right) t_{1i} + p_z t_{0i} - \left(\left(1 - p_z\right) \mathbb{E}\left[t_{1i}\right] + p_z \mathbb{E}\left[t_{0i}\right] \right) \right) \\
& + \left[\hat{p}_z \left(1 - \hat{p}_z\right) - p_z \left(1 - p_z\right)\right] \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(t_{1i} - t_{0i}\right) \\
= & \left(p_z - \hat{p}_z\right) \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(Z_i - p_z\right) \left(t_{1i} - t_{0i}\right) + \left(p_z - \hat{p}_z\right)^2 \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(t_{1i} - t_{0i}\right) \\
& + \left(p_z - \hat{p}_z\right) \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(\left(1 - p_z\right) \left(t_{1i} - \mathbb{E}\left[t_{1i}\right]\right) + p_z \left(t_{0i} - \mathbb{E}\left[t_{0i}\right]\right) \right) \\
& + \left[\hat{p}_z \left(1 - \hat{p}_z\right) - p_z \left(1 - p_z\right)\right] \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(t_{1i} - t_{0i}\right)
\end{aligned}
$$

Using Assumption 8, each term can be shown to be $o_P(1)$, so that $R_n = o_P(1)$.
From this point on the variance becomes different depending on whether $S = 1$ or $S > 1$. Recall that $\omega_i = \left[\frac{t_{1i} - Et_{1i}}{p_z} + \frac{t_{0i} - Et_{0i}}{1 - p_z}\right]$ and $\omega(s) = E\left[\omega_i | X_i \in s\right]$. First note that $\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(t_{1i} - t_{0i}\right)$ is asymptotically orthogonal to $\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(Z_i - p_z\right) \omega_i$ under assumptions 8.1 and 8.2.

$$
\begin{aligned}
& Cov\left[\sum_{s \in \mathcal{S}} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} 1\left(X_i \in s\right) \left(Z_i - p_z\right) \omega_i, \sum_{s \in \mathcal{S}} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} 1\left(X_i \in s\right) \left(t_{1i} - t_{0i}\right) \right] \\
= & \sum_{s \in \mathcal{S}} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathbb{E}\left[1\left(X_i \in s\right) \left(Z_i - p_z\right) \omega_i \left(t_{1i} - t_{0i}\right)\right] \\
= & \sum_{s \in \mathcal{S}} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathbb{E}\left[1\left(X_i \in s\right) \left(\mathbb{E}\left[\, Z_i \middle| \, X_i \in s, Y_{1i}, Y_{0i}, D_{1i}, D_{0i}\right] - p_z\right) \omega_i \left(t_{1i} - t_{0i}\right)\right] \\
= & \sum_{s \in \mathcal{S}} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathbb{E}\left[1\left(X_i \in s\right) \left(\mathbb{E}\left[\, Z_i \middle| \, X_i \in s\right] - p_z\right) \omega_i \left(t_{1i} - t_{0i}\right)\right] \\
= & O_{a.s.}\left(\frac{1}{\sqrt{n}}\right)
\end{aligned}
$$

Then write the first part of the influence function as

$$
\begin{aligned}
& \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(Z_i - p_z\right) \omega_i \\
= & \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(Z_i - p_z\right) \left(\omega_i - \omega\left(s\right)\right) + \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(Z_i - p_z\right) \omega\left(s\right) \qquad (25) \\
= & \sum_{s \in S} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} 1\left(X_i \in s\right) \left(Z_i - p_z\right) \left(\omega_i - \omega\left(s\right)\right) + \sum_{s \in S} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} 1\left(X_i \in s\right) \left(Z_i - p_z\right) \omega\left(s\right).
\end{aligned}
$$

First note that the two sums are orthogonal:

$$
\begin{aligned}
&Cov\left[\sum_{s\in\mathcal{S}}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}1\left(X_i\in s\right)\left(Z_i-p_z\right)\left(\omega_i-\omega(s)\right),\sum_{s\in\mathcal{S}}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}1\left(X_i\in s\right)\left(Z_i-p_z\right)\omega(s)\right]\\
&=\sum_{s\in\mathcal{S}}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}Cov\left[1\left(X_i\in s\right)\left(Z_i-p_z\right)\left(\omega_i-\omega(s)\right),1\left(X_i\in s\right)\left(Z_i-p_z\right)\omega(s)\right]\\
&=\sum_{s\in\mathcal{S}}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}E\left[1\left(X_i\in s\right)\left(Z_i-p_z\right)^2\left(\omega_i-\omega(s)\right)\omega(s)\right]\\
&=\sum_{s\in\mathcal{S}}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}E\left[1\left(X_i\in s\right)\left(Z_i-p_z\right)^2\left(\omega_i-E\left[\omega_i|X_i\in s,Z_i\right]\right)E\left[\omega_i|X_i\in s,Z_i\right]\right]\\
&=\sum_{s\in\mathcal{S}}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}E\left[1\left(X_i\in s\right)\left(Z_i-p_z\right)^2\left(\omega_i-E\left[\omega_i|X_i\in s\right]\right)E\left[\omega_i|X_i\in s\right]\right]\\
&=0
\end{aligned}
$$

We now use arguments similar to those in Lemma B.2 of BCS 2017a to derive the limiting distribution of (25). The distribution of $U=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(Z_i-p_z\right)\left(\omega_i-\omega\left(s\right)\right)$ is the same as the distribution of the same quantity where the observations are first ordered by strata and then by $Z_i=1$ and $Z_i=0$ within strata. Let $n_z(s)$ be the number of observations in strata s which have $Z_i=z\in\{0,1\}$, and let $p\left(s\right)=P\left(X_i\in s\right)$, $N(s)=\sum_{i=1}^{n}I\left\{S_i<s\right\}$, and $F(s)=P\left\{S_i<s\right\}$. Independently for each s and independently of $\left(Z^{(n)},S^{(n)}\right)$, let $\{\omega_i^s:1\le i\le n\}$ be i.i.d. with marginal distribution equal to the distribution of $\omega_i|X_i\in s$. Define

$$
\tilde{U}=\frac{1}{\sqrt{n}}\sum_{s\in\mathcal{S}}\left(\sum_{i=n\frac{N(s)}{n}+1}^{n\left(\frac{N(s)}{n}+\frac{n_1(s)}{n}\right)}\left(\omega_i^s-\omega\left(s\right)\right)\left(1-p_z\right)+\sum_{i=n\left(\frac{N(s)}{n}+\frac{n_1(s)}{n}\right)+1}^{n\left(\frac{N(s)}{n}+\frac{n(s)}{n}\right)}\left(\omega_i^s-\omega\left(s\right)\right)\left(-p_z\right)\right)
$$

By construction, $\left\{U|S^{(n)},Z^{(n)}\right\}\overset{d}{=}\left\{\tilde{U}|S^{(n)},Z^{(n)}\right\}$ which implies $U\overset{d}{=}\tilde{U}$. Next define

$$
U^*=\frac{1}{\sqrt{n}}\sum_{s\in\mathcal{S}}\left(\sum_{i=\lfloor nF(s)\rfloor+1}^{\lfloor n(F(s)+p(s)p_z)\rfloor}\left(\omega_i^s-\omega\left(s\right)\right)\left(1-p_z\right)+\sum_{i=\lfloor n(F(s)+p(s)p_z)\rfloor+1}^{\lfloor n(F(s)+p(s))\rfloor}\left(\omega_i^s-\omega\left(s\right)\right)\left(-p_z\right)\right)
$$

Using properties of Brownian motion,

$$
\frac{1}{\sqrt{n}}\sum_{i=\lfloor nF(s)\rfloor+1}^{\lfloor n(F(s)+p(s)p_z)\rfloor}\left(\omega_i^s-\omega\left(s\right)\right)\left(1-p_z\right)\overset{d}{\to}N\left(0,p(s)p_z\left(1-p_z\right)^2\mathbb{E}\left[\left(\omega_i^s-\omega\left(s\right)\right)^2\right]\right)
$$

$$
\frac{1}{\sqrt{n}}\sum_{i=\lfloor n(F(s)+p(s)p_z)\rfloor+1}^{\lfloor n(F(s)+p(s))\rfloor}\left(\omega_i^s-\omega\left(s\right)\right)\left(-p_z\right)\overset{d}{\to}N\left(0,p(s)\left(1-p_z\right)\left(p_z\right)^2\mathbb{E}\left[\left(\omega_i^s-\omega\left(s\right)\right)^2\right]\right)
$$

Since the two sums are independent, $\omega_i^s-\omega\left(s\right)$ are independent across i and s, and $\mathbb{E}\left[\left(\omega_i^s-\omega\left(s\right)\right)^2\right]=\mathbb{E}\left[\left(\omega_i-\omega\left(s\right)\right)^2\middle|X_i\in s\right]$,

$$
U^*\overset{d}{\to}N\left(0,p_z\left(1-p_z\right)\sum_{s\in\mathcal{S}}p(s)\mathbb{E}\left[\left(\omega_i-\omega\left(s\right)\right)^2\middle|X_i\in s\right]\right)
$$

Furthermore, since $\left(\frac{N(s)}{n}, \frac{n_1(s)}{n}\right) \xrightarrow{p} (F(s), p_z p(s))$, by the continuous mapping theorem,

$\tilde{U} - U^* \xrightarrow{p} 0$

Therefore,

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (Z_i - p_z)(\omega_i - \omega(s)) \xrightarrow{d} N\left(0, \underbrace{p_z(1-p_z) \sum_{s \in \mathcal{S}} p(s) \mathbb{E}\left[(\omega_i - \omega(s))^2 \middle| X_i \in s\right]}_{\Omega_1}\right)$$

For the second term, it suffices to use Assumption 8.2 to show that

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (Z_i - p_z)\omega(s) = \sum_{s \in \mathcal{S}} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} 1(X_i \in s)(Z_i - p_z)\omega(s) \xrightarrow{d} N\left(0, \underbrace{\sum_{s \in \mathcal{S}} \tau(s) p(s) \omega(s)^2}_{\Omega_2}\right)$$

Lastly, note that $\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (t_{1i} - t_{0i}) \xrightarrow{d} N\left(0, \underbrace{Var[t_{1i} - t_{0i}]}_{\Omega_3}\right)$. Then $\sqrt{n}\left(\hat{\beta}_1 - \beta_0\right) \xrightarrow{d}$ $N\left(0, P(D_1 > D_0)^{-2}(\Omega_1 + \Omega_2 + \Omega_3)\right)$.

As in section 2, it is straightforward to show that the 2SLS robust variance is consistent for $P(D_1 > D_0)^{-2}$ times

$$\begin{aligned}
&\operatorname{plim} \frac{1}{n} \sum_{i=1}^{n} [(Z_i - p_z)[\omega_i] + (t_{1i} - t_{0i})]^2 \\
&= \operatorname{plim} \frac{1}{n} \sum_{i=1}^{n} ((Z_i - p_z)\omega_i)^2 + \operatorname{plim} \frac{1}{n} \sum_{i=1}^{n} (t_{1i} - t_{0i})^2 \\
&= \operatorname{plim} \frac{1}{n} \sum_{i=1}^{n} (Z_i - p_z)^2 (\omega_i - \omega(s))^2 + \operatorname{plim} \frac{1}{n} \sum_{i=1}^{n} (Z_i - p_z)^2 \omega(s)^2 + \operatorname{plim} \frac{1}{n} \sum_{i=1}^{n} (t_{1i} - t_{0i})^2
\end{aligned}$$

Independently for each s and independently of $\left(Z^{(n)}, S^{(n)}\right)$, let $\{\omega_i^s : 1 \le i \le n\}$ be i.i.d with marginal distribution equal to the distribution of $\omega_i | X_i \in s$. Using similar arguments as those in Lemma B.3 of BCS 2017a,

$$\begin{aligned}
&\frac{1}{n} \sum_{i=1}^{n} (Z_i - p_z)^2 (\omega_i - \omega(s))^2 \\
&= \sum_{s \in \mathcal{S}} \left(\frac{1}{n} \sum_{i=1}^{n_1(s)} (1 - p_z)^2 (\omega_i^s - \omega(s))^2 + \frac{1}{n} \sum_{i=1}^{n_0(s)} (-p_z)^2 (\omega_i^s - \omega(s))^2 \right) \\
&= \sum_{s \in \mathcal{S}} \left(\frac{n_1(s)}{n} \frac{1}{n_1(s)} \sum_{i=1}^{n_1(s)} (1 - p_z)^2 (\omega_i^s - \omega(s))^2 + \frac{n_0(s)}{n} \frac{1}{n_0(s)} \sum_{i=1}^{n_0(s)} (-p_z)^2 (\omega_i^s - \omega(s))^2 \right) \\
&\xrightarrow{p} \sum_{s \in \mathcal{S}} \left\{ p_z p(s) (1 - p_z)^2 \mathbb{E}\left[(\omega_i^s - \omega(s))^2\right] + (1 - p_z) p(s) (-p_z)^2 \mathbb{E}\left[(\omega_i^s - \omega(s))^2\right] \right\} \\
&= p_z (1 - p_z) \sum_{s \in \mathcal{S}} p(s) \mathbb{E}\left[(\omega_i - \omega(s))^2 \middle| X_i \in s\right]
\end{aligned}$$

The key steps are to use the Almost Sure Representation theorem to construct $\frac{\tilde{n}_1(s)}{n} \stackrel{d}{=} \frac{n_1(s)}{n}$ such that $\frac{\tilde{n}_1(s)}{n} \stackrel{a.s.}{\to} p_z p(s)$ and then to note that by independence of $\left(Z^{(n)}, S^{(n)}\right)$ and $\{\omega_i^s : 1 \leq i \leq n\}$, for any $\epsilon > 0$,

$$\begin{aligned}
&\mathbb{P}\left\{\left|\frac{1}{n_1(s)}\sum_{i=1}^{n_1(s)}\left(\omega_i^s - \omega(s)\right)^2 - \mathbb{E}\left[\left(\omega_i^s - \omega(s)\right)^2\right]\right| > \epsilon\right\} \\
&= \mathbb{E}\left[\mathbb{P}\left\{\left|\frac{1}{n\frac{\tilde{n}_1(s)}{n}}\sum_{i=1}^{n\frac{\tilde{n}_1(s)}{n}}\left(\omega_i^s - \omega(s)\right)^2 - \mathbb{E}\left[\left(\omega_i^s - \omega(s)\right)^2\right]\right| > \epsilon \middle| \frac{\tilde{n}_1(s)}{n}\right\}\right]
\end{aligned}$$

Also, note that by the weak law of large numbers, for any sequence $n_k \to \infty$ as $k \to \infty$,

$$\frac{1}{n_k}\sum_{i=1}^{n_k}\left(\omega_i^s - \omega(s)\right)^2 \stackrel{p}{\to} \mathbb{E}\left[\left(\omega_i^s - \omega(s)\right)^2\right]$$

Since $n\frac{\tilde{n}_1(s)}{n} \to \infty$ almost surely, by independence of $\frac{\tilde{n}_1(s)}{n}$ and $\{\omega_i^s : 1 \leq i \leq n\}$,

$$\mathbb{P}\left\{\left|\frac{1}{n\frac{\tilde{n}_1(s)}{n}}\sum_{i=1}^{n\frac{\tilde{n}_1(s)}{n}}\left(\omega_i^s - \omega(s)\right)^2 - \mathbb{E}\left[\left(\omega_i^s - \omega(s)\right)^2\right]\right| > \epsilon \middle| \frac{\tilde{n}_1(s)}{n}\right\} \stackrel{a.s.}{\to} 0$$

Therefore, the first and third terms coincide with Ω_1 and Ω_3. The second term converges to

$$\operatorname{plim}\frac{1}{n}\sum_{i=1}^{n}\left(Z_i - p_z\right)^2\omega(s)^2 = \sum_{s=1}^{S}\omega(s)^2 p(s) p_z\left(1 - p_z\right)$$

This is larger than Ω_2 as long as $\tau(s) \leq p_z(1 - p_z)$ for all $s \in \mathcal{S}$, and strictly so for some $s \in \mathcal{S}$.

I Proof of Proposition 5

The sample normal equations for this regression are given by

$$\tau_n\left(\hat{\beta}_2, \hat{\eta}\right) = \frac{1}{n}\sum_{i=1}^{n}\begin{bmatrix} 1\left(X_i \in s\right)_{s\in\mathcal{S}} \\ \left(Z_i - p_z\right)\end{bmatrix}\left(Y_i - \hat{\beta}_2 D_i - \sum_{s=1}^{S}\hat{\eta}_s 1\left(X_i \in s\right)\right) = 0.$$

We can write $\left(\hat{\beta}_2 - \beta_0, \hat{\eta} - \eta_0\right) = \hat{A}^{-1}\tau_n(\beta_0, \eta_0)$ if we let $\eta_0 = (\eta_{0s}, s \in \mathcal{S})$, $t_1(s) = E(t_{1i}|X_i \in s)$, $t_0(s) = E(t_{0i}|X_i \in s)$,

$$\begin{aligned}
\eta_{0s} =& E(Y|s) - E(D|s)\beta_0 = p_z t_1(s) + (1 - p_z)t_0(s) \\
&= (1 - p_z)t_1(s) + p_z t_0(s) - (1 - 2p_z)\left[t_1(s) - t_0(s)\right].
\end{aligned}$$

and

$$\hat{A} = \frac{1}{n}\sum_{i=1}^{n}\begin{bmatrix} 1\left(X_i \in s\right)_{s\in\mathcal{S}} D_i & \operatorname{diag}\left(1\left(X_i \in s\right)_{s\in\mathcal{S}}\right) \\ \left(Z_i - p_z\right)D_i & \left(Z_i - p_z\right)1\left(X_i \in s\right)_{s\in\mathcal{S}}.\end{bmatrix}$$

Using Assumption 8.1 and 8.2 we can show that $\hat{A} = A + o_P(1)$, where

$$A = \begin{bmatrix} p(s) E(D|s) & \text{diag}(p(s), s \in \mathcal{S}) \\ p_z(1-p_z) P(D_1 > D_0) & 0 \end{bmatrix}$$

In the following we will show that $\tau_n(\beta_0, \eta_0) = O_p\left(\frac{1}{\sqrt{n}}\right)$, which by non-singularity of A implies that $\left(\hat{\beta}_2 - \beta_0, \hat{\eta} - \eta_0\right) = O_P\left(\frac{1}{\sqrt{n}}\right)$. Then the second row of the relation

$$(A + o_P(1)) \left(\hat{\beta}_2 - \beta_0, \hat{\eta} - \eta_0\right)' = \tau_n(\beta_0, \eta_0)$$

implies that, using the above η_{0s},

$$\begin{aligned}
&P(D_1 > D_0) \sqrt{n} \left(\hat{\beta}_2 - \beta_0\right) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{(Z_i - p_z)}{p_z(1-p_z)} \left(Y_i - \beta_0 D_i - \sum_{s=1}^{S} \eta_{0s} 1(X_i \in s)\right) + o_P(1) \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Bigg[(Z_i - p_z) \left(\frac{t_{1i} - Et_{1i}}{p_z} + \frac{t_{0i} - Et_{0i}}{1-p_z} \right. \\
&- \left. \sum_{s \in \mathcal{S}} \left(\frac{E(t_{1i} - Et_{1i} | X_i \in s)}{p_z} + \frac{E(t_{0i} - Et_{0i} | X_i \in s)}{1-p_z} - \frac{1-2p_z}{p_z(1-p_z)} [t_1(s) - t_0(s)] \right) 1(X_i \in s) \right) + (t_{1i} - t_{0i}) \Bigg] + o_P(1) \\
&= \sum_{s \in \mathcal{S}} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} 1(X_i \in s)(Z_i - p_z)(\omega_i - \omega(s)) + \sum_{s \in \mathcal{S}} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} 1(X_i \in s)(Z_i - p_z) \frac{1-2p_z}{p_z(1-p_z)} (t_1(s) - t_0(s)) \\
&\qquad + \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (t_{1i} - t_{0i}) + o_P(1).
\end{aligned}$$

where we recall that $\omega_i = \frac{t_{1i} - Et_{1i}}{p_z} + \frac{t_{0i} - Et_{0i}}{1-p_z}$ and $\omega(s) = E[\omega_i | X_i \in s]$. Using similar arguments to those in proposition 4,

$$\sum_{s \in \mathcal{S}} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} 1(X_i \in s)(Z_i - p_z)(\omega_i - \omega(s)) \xrightarrow{d} N\left(0, \underbrace{p_z(1-p_z) \sum_{s \in \mathcal{S}} p(s) \mathbb{E}\left[(\omega_i - \omega(s))^2 \middle| X_i \in s\right]}_{\Omega_1}\right)$$

$$\sum_{s \in \mathcal{S}} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} 1(X_i \in s)(Z_i - p_z) \frac{1-2p_z}{p_z(1-p_z)} (t_1(s) - t_0(s)) \xrightarrow{d} N\left(0, \underbrace{\sum_{s \in \mathcal{S}} p(s) \tau(s) \left(\frac{1-2p_z}{p_z(1-p_z)} (t_1(s) - t_0(s))\right)^2}_{\Omega_2}\right)$$

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (t_{1i} - t_{0i}) \xrightarrow{d} N\left(0, \underbrace{Var[t_{1i} - t_{0i}]}_{\Omega_3}\right)$$

Note that the first two sums in the influence function are orthogonal:

$$
\begin{aligned}
&Cov\left[\sum_{s\in\mathcal{S}}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}1\left(X_i\in s\right)\left(Z_i-p_z\right)\left(\omega_i-\omega\left(s\right)\right),\sum_{s\in\mathcal{S}}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}1\left(X_i\in s\right)\left(Z_i-p_z\right)\frac{1-2p_z}{p_z\left(1-p_z\right)}\left(t_1\left(s\right)-t_0\left(s\right)\right)\right]\\
&=\sum_{s\in\mathcal{S}}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbb{E}\left[1\left(X_i\in s\right)\left(Z_i-p_z\right)^2\left(\omega_i-\omega\left(s\right)\right)\frac{1-2p_z}{p_z\left(1-p_z\right)}\left(t_1\left(s\right)-t_0\left(s\right)\right)\right]\\
&=\sum_{s\in\mathcal{S}}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbb{E}\left[1\left(X_i\in s\right)\left(Z_i-p_z\right)^2\frac{1-2p_z}{p_z\left(1-p_z\right)}\left(\mathbb{E}\left[\left(\omega_i-\omega\left(s\right)\right)\left(t_1\left(s\right)-t_0\left(s\right)\right)\middle|X_i\in s,Z_i\right]\right)\right]\\
&=\sum_{s\in\mathcal{S}}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbb{E}\left[1\left(X_i\in s\right)\left(Z_i-p_z\right)^2\left(\omega_i-\mathbb{E}\left[\omega_i\middle|X_i\in s\right]\right)\frac{1-2p_z}{p_z\left(1-p_z\right)}\left(t_1\left(s\right)-t_0\left(s\right)\right)\right]\\
&=0
\end{aligned}
$$

And the third sum is orthogonal to the first two sums by the same arguments in proposition 4. Therefore, $P\left(D_1>D_0\right)\sqrt{n}\left(\hat{\beta}_2-\beta_0\right)\xrightarrow{d}N\left(0,\Omega_1+\bar{\Omega}_2+\Omega_3\right)$. It is also easy to show using similar arguments to those in proposition 4 that the 2SLS nominal variance consistently estimates $P\left(D_1>D_0\right)^{-2}$ times

$$
\begin{aligned}
&\operatorname{plim}\frac{1}{n}\sum_{i=1}^{n}\left[\frac{\left(Z_i-p_z\right)}{p_z\left(1-p_z\right)}\left(Y_i-\beta_0D_i-\sum_{s=1}^{S}\eta_{0s}1\left(X_i\in s\right)\right)\right]^2\\
&=\operatorname{plim}\sum_{s\in\mathcal{S}}\frac{1}{n}\sum_{i=1}^{n}1\left(X_i\in s\right)\left(Z_i-p_z\right)\left(\omega_i-\omega\left(s\right)\right)^2\\
&+\operatorname{plim}\sum_{s\in\mathcal{S}}\frac{1}{n}\sum_{i=1}^{n}1\left(X_i\in s\right)\left(Z_i-p_z\right)\left(\frac{1-2p_z}{p_z\left(1-p_z\right)}\right)^2\left(t_1\left(s\right)-t_0\left(s\right)\right)^2\\
&+\operatorname{plim}\frac{1}{n}\sum_{i=1}^{n}\left(t_{1i}-t_{0i}\right)^2\\
&=\Omega_1+\tilde{\Omega}_2+\Omega_3
\end{aligned}
$$

where

$$
\tilde{\Omega}_2=\sum_{s\in\mathcal{S}}p\left(s\right)p_z\left(1-p_z\right)\left(\frac{1-2p_z}{p_z\left(1-p_z\right)}\left(t_1\left(s\right)-t_0\left(s\right)\right)\right)^2
$$

which is larger than $\bar{\Omega}_2$ if $p_z\left(1-p_z\right)>\tau\left(s\right)$ for some s, unless $S=1$ or $p_z=\frac{1}{2}$.

J Proof of Proposition 6

We choose to work with the representation in (11), using which we write

$$
\sqrt{n}\left(\hat{\beta}_3-\beta_0\right)=\sqrt{n}\frac{\sum_{s=1}^{S}\left(\hat{\xi}_{1s}-\hat{\xi}_{0s}-\beta_0\left(\hat{\zeta}_{1s}-\hat{\zeta}_{0s}\right)\right)\frac{\sum_{i=1}^{n}1\left(X_i\in s\right)}{n}}{\sum_{s=1}^{S}\left(\hat{\zeta}_{1s}-\hat{\zeta}_{0s}\right)\frac{\sum_{i=1}^{n}1\left(X_i\in s\right)}{n}} \tag{26}
$$

For the denominator, under Assumption 8, Lemma B.3 of BCS 2017a implies that

$$\hat{\zeta}_{1s} = \frac{\frac{1}{n}\sum_{i=1}^{n} 1(X_i \in s) Z_i D_i}{\frac{1}{n}\sum_{i=1}^{n} 1(X_i \in s) Z_i} \xrightarrow{p} P(D_1 = 1|s),$$
$$\hat{\zeta}_{0s} = \frac{\frac{1}{n}\sum_{i=1}^{n} 1(X_i \in s)(1 - Z_i) D_i}{\frac{1}{n}\sum_{i=1}^{n} 1(X_i \in s)(1 - Z_i)} \xrightarrow{p} P(D_0 = 1|s).$$

Together with $\frac{1}{n}\sum_{i=1}^{n} 1(x_i \in s) \xrightarrow{p} p(s) \equiv p(x_i \in s)$,

$$\sum_{s=1}^{S} \left(\hat{\zeta}_{1s} - \hat{\zeta}_{0s}\right) \frac{\sum_{i=1}^{n} 1(X_i \in s)}{n} \xrightarrow{p} P(D_1 = 1) - P(D_0 = 1) = P(D_1 > D_0).$$

Using $\hat{p}(s) = \frac{1}{n}\sum_{i=1}^{n} 1(X_i \in s)$, $\hat{p}(s)\hat{p}_z(s) = \frac{1}{n}\sum_{=1}^{n} 1(X_i \in s) Z_i$, $t_1(s) = E[t_{1i}|X_i \in s]$, and $t_0(s) = E[t_{0i}|X_i \in s]$

$$\begin{aligned}
&\sum_{s=1}^{S} \left(\hat{\xi}_{1s} - \hat{\xi}_{0s} - \beta_0 \left(\hat{\zeta}_{1s} - \hat{\zeta}_{0s}\right)\right) \frac{\sum_{i=1}^{n} 1(X_i \in s)}{n} \\
=&\sum_{s=1}^{S} \hat{p}(s) \left[\frac{\frac{1}{n}\sum_{i=1}^{n} t_{1i} 1(X_i \in s) Z_i}{\hat{p}(s)\hat{p}_z} - \frac{\frac{1}{n}\sum_{i=1}^{n} t_{0i} 1(X_i \in s)(1 - Z_i)}{\hat{p}(s)(1 - \hat{p}_z)}\right] \\
=&\sum_{s=1}^{S} \frac{1}{n}\sum_{i=1}^{n} 1(X_i \in s) \left[\frac{(t_{1i} - t_1(s)) Z_i}{\hat{p}_z} - \frac{(t_{0i} - t_0(s))(1 - Z_i)}{(1 - \hat{p}_z)}\right] \\
&+ \sum_{s=1}^{S} \frac{1}{n}\sum_{i=1}^{n} 1(X_i \in s)(t_1(s) - t_0(s)) \\
=&\sum_{s=1}^{S} \frac{1}{n}\sum_{i=1}^{n} 1(X_i \in s) \left[\frac{(t_{1i} - t_1(s)) Z_i}{p_z} - \frac{(t_{0i} - t_0(s))(1 - Z_i)}{(1 - p_z)}\right] + \sum_{s \in \mathcal{S}} (R_{1ns} + R_{2ns}) \\
&+ \sum_{s=1}^{S} \frac{1}{n}\sum_{i=1}^{n} 1(X_i \in s)(t_1(s) - t_0(s)) \\
=&\sum_{s=1}^{S} \frac{1}{n}\sum_{i=1}^{n} 1(X_i \in s)(Z_i - p_z) \left[\frac{t_{1i} - t_1(s)}{p_z} + \frac{t_{0i} - t_0(s)}{1 - p_z}\right] + \sum_{s \in \mathcal{S}} (R_{1ns} + R_{2ns}) \\
&+ \frac{1}{n}\sum_{i=1}^{n} (t_{1i} - t_{0i})
\end{aligned} \tag{27}$$

In the above

$$R_{1ns} = \frac{p_z - \hat{p}_z}{\hat{p}_z(1 - \hat{p}_z)} \frac{1}{n}\sum_{i=1}^{n} 1(X_i \in s) \left[(t_{1i} - t_1(s)) Z_i + (t_{0i} - t_0(s))(1 - Z_i)\right]$$
$$R_{2ns} = \left(\frac{1}{\hat{p}_z(1 - \hat{p}_z)} - \frac{1}{p_z(1 - p_z)}\right) \times \frac{1}{n}\sum_{i=1}^{n} 1(X_i \in s) \left[(1 - p_z)(t_{1i} - t_1(s)) Z_i - p_z(t_{0i} - t_0(s))(1 - Z_i)\right].$$

Rewriting,

$$R_{1ns} = \frac{p_z - \hat{p}_z}{\hat{p}_z\left(1-\hat{p}_z\right)}\left\{\frac{1}{n}\sum_{i=1}^{n} 1\left(X_i \in s\right)\left(Z_i - p_z\right)\left[\left(t_{1i} - t_1\left(s\right)\right) - \left(t_{0i} - t_0\left(s\right)\right)\right] \right.$$
$$\left. + p_z \frac{1}{n}\sum_{i=1}^{n} 1\left(X_i \in s\right)\left(t_{1i} - t_1\left(s\right)\right) + \left(1-p_z\right)\frac{1}{n}\sum_{i=1}^{n} 1\left(X_i \in s\right)\left(t_{0i} - t_0\left(s\right)\right)\right\}$$

$$R_{2ns} = \left(\frac{1}{\hat{p}_z\left(1-\hat{p}_z\right)} - \frac{1}{p_z\left(1-p_z\right)}\right)\left\{\frac{1}{n}\sum_{i=1}^{n} 1\left(X_i \in s\right)\left(Z_i - p_z\right)\left[\left(t_{1i} - t_1\left(s\right)\right) + \left(t_{0i} - t_0\left(s\right)\right)\right]\right.$$
$$\left. + p_z \frac{1}{n}\sum_{i=1}^{n}\left(1 - Z_i\right) 1\left(X_i \in s\right)\left(t_{1i} - t_1\left(s\right)\right) - \left(1-p_z\right)\frac{1}{n}\sum_{i=1}^{n} Z_i 1\left(X_i \in s\right)\left(t_{0i} - t_0\left(s\right)\right)\right\}$$
$$= \left(\frac{1}{\hat{p}_z\left(1-\hat{p}_z\right)} - \frac{1}{p_z\left(1-p_z\right)}\right)\left\{\frac{1}{n}\sum_{i=1}^{n} 1\left(X_i \in s\right)\left(Z_i - p_z\right)\left[\left(t_{1i} - t_1\left(s\right)\right) + \left(t_{0i} - t_0\left(s\right)\right)\right]\right.$$
$$+ p_z \frac{1}{n}\sum_{i=1}^{n}\left(1 - Z_i - \left(1-p_z\right)\right) 1\left(X_i \in s\right)\left(t_{1i} - t_1\left(s\right)\right) - \left(1-p_z\right)\frac{1}{n}\sum_{i=1}^{n}\left(Z_i - p_z\right) 1\left(X_i \in s\right)\left(t_{0i} - t_0\left(s\right)\right)$$
$$\left. + p_z\left(1-p_z\right)\frac{1}{n}\sum_{i=1}^{n} 1\left(X_i \in s\right)\left(t_{1i} - t_1\left(s\right)\right) - \left(1-p_z\right)p_z\frac{1}{n}\sum_{i=1}^{n} 1\left(X_i \in s\right)\left(t_{0i} - t_0\left(s\right)\right)\right\}$$

Using Assumption 8, Lemmas B.2 and B.3 of BCS 2017a, and arguments similar to those in propositions 4 and 5, we can show that $\sum_{s\in\mathcal{S}} R_{1ns} = o_P\left(1\right) O_P\left(\frac{1}{\sqrt{n}}\right) = o_P\left(\frac{1}{\sqrt{n}}\right)$ and $\sum_{s\in\mathcal{S}} R_{2ns} = o_P\left(1\right) O_P\left(\frac{1}{\sqrt{n}}\right) = o_P\left(\frac{1}{\sqrt{n}}\right)$.

Since $\frac{t_{1i}-t_1(s)}{p_z} + \frac{t_{0i}-t_0(s)}{1-p_z} = \frac{t_{1i}-Et_{1i}-(t_1(s)-Et_{1i})}{p_z} + \frac{t_{0i}-Et_{0i}-(t_0(s)-Et_{0i})}{1-p_z} = \omega_i - \omega\left(s\right)$, (27) can be written as

$$\sum_{s=1}^{S}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} 1\left(X_i \in s\right)\left(Z_i - p_z\right)\left[\omega_i - \omega\left(s\right)\right] + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(t_{1i} - t_{0i}\right) + o_P\left(1\right), \tag{28}$$

The first part of this influence function corresponds exactly to the first term in (25). Therefore regardless of p_z there is no need to worry about the variation induced by the sampling scheme for Z_i within the cluster.

In the special case of unconfoundedness, (28) becomes

$$\sum_{s=1}^{S}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} 1\left(X_i \in s\right)\left(Z_i - p_z\right)\left[\frac{Y_{1i}}{p_z} + \frac{Y_{0i}}{1-p_z} - E\left[\frac{Y_{1i}}{p_z} + \frac{Y_{0i}}{1-p_z}\middle| X_i \in s\right]\right]$$
$$+ \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(Y_{1i} - Y_{0i}\right) - \left(\mu_1 - \mu_0\right) + o_P\left(1\right), \tag{29}$$

Using Assumption 8, $\hat{\sigma}_3^2$ is consistent for the plim of $P\left(D_1 > D_0\right)^{-2}$ times $\frac{1}{n}\sum_{i=1}^{n}\psi_i^2$ where

$$\psi_i = \left(Z_i - p_z\right)\left(\left(\frac{t_{1i}-\bar{t}_1}{p_z} + \frac{t_{0i}-\bar{t}_0}{1-p_z} - \Sigma_{n,X,\frac{t_1}{p_z}+\frac{t_0}{1-p_z}}\Sigma_{n,X}^{-1}\left(X_i - \bar{X}\right)\right)\right)$$
$$+ p_z\left(1-p_z\right)\left(t_{1i} - t_{0i} - \bar{t}_1 + \bar{t}_0 - \Sigma_{n,X,t_1-t_0}\Sigma_{n,X}^{-1}\left(X_i - \bar{X}\right)\right)$$
$$= \left(Z_i - p_z\right)\left(\left(\omega_i - \bar{\omega} - \Sigma_{n,X,\omega}\Sigma_{n,X}^{-1}\left(X_i - \bar{X}\right)\right)\right)$$
$$+ p_z\left(1-p_z\right)\left(t_{1i} - t_{0i} - \bar{t}_1 + \bar{t}_0 - \Sigma_{n,X,t_1-t_0}\Sigma_{n,X}^{-1}\left(X_i - \bar{X}\right)\right)$$

for $\Sigma_{n,X} = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(X_i - \bar{X}\right)'$ and $\Sigma_{n,X,t} = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)(t_i - \bar{t})$. With X_i being the cluster dummies, $\omega_i - \bar{\omega} - \Sigma_{n,X,\omega}\Sigma_{n,X}^{-1}\left(X_i - \bar{X}\right)$ is the residual from a saturated regression of ω_i on the cluster dummies, and converges to $\sum_{s\in\mathcal{S}} 1(x_i \in s)(\omega_i - \omega(s))$. For the same reason, $t_{1i} - t_{0i} - \bar{t}_1 + \bar{t}_0 - \Sigma_{n,X,t_1-t_0}\Sigma_{n,X}^{-1}\left(X_i - \bar{X}\right)$ is the residual from a saturated regression of $t_{1i} - t_{0i}$ on the cluster dummies, and converges to

$$\sum_{s\in\mathcal{S}} 1(x_i \in s)(t_{1i} - t_{0i} - E(t_{1i} - t_{0i}|s)).$$

Therefore $\frac{1}{n}\sum_{i=1}^{n}\psi_i^2$ is in turn consistent for the variance of

$$\sum_{s=1}^{S}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} 1(X_i \in s)(Z_i - p_z)[\omega_i - \omega(s)] + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\sum_{s\in\mathcal{S}} 1(x_i \in s)(t_{1i} - t_{0i} - E(t_{1i} - t_{0i}|s)),$$

which is asymptotically smaller than the variance of (28) but larger than the variance of its first component. Next we will need to add a consistent estimate of

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{s\in\mathcal{S}} 1(x_i \in s) E(t_{1i} - t_{0i}|s)^2.$$

This is obtained by $\hat{\phi}'\frac{1}{n}\sum_{i=1}^{n}\left(V_i - \bar{V}\right)\left(V_i - \bar{V}\right)'\hat{\phi}$, which is the variance of the fitted value of the saturated cluster dummy regression. We can then use $\bar{\sigma}_3^2$ in Corollary 3 to obtain a consistent estimate of the variance of (28).

We can also directly estimate the variance of $\hat{\beta}_3$ by estimating the first representation of the influence function in (27). Let $\hat{t}_{1i}Z_i = \left(Y_i - D_i\hat{\beta}_3\right)Z_i$, $\hat{t}_{0i}(1 - Z_i) = \left(Y_i - D_i\hat{\beta}_3\right)(1 - Z_i)$,

$$\hat{t}_1(s) = \sum_{i=1}^{n}\hat{t}_{1i}Z_i 1(X_i \in s) / \sum_{i=1}^{n} Z_i 1(X_i \in s)$$
$$\hat{t}_0(s) = \sum_{i=1}^{n}\hat{t}_{0i}(1 - Z_i) 1(X_i \in s) / \sum_{i=1}^{n}(1 - Z_i) 1(X_i \in s)$$

and construct

$$\hat{\Omega} = \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{s\in\mathcal{S}} 1(X_i \in s)\left[\frac{\left(\hat{t}_{1i}Z_i - \hat{t}_1(s)Z_i\right)}{\hat{p}_z} - \frac{\left(\hat{t}_{0i}(1 - Z_i) - \hat{t}_0(s)(1 - Z_i)\right)}{(1 - \hat{p}_z)}\right]\right)^2$$
$$+ \frac{1}{n}\sum_{i=1}^{n}\left[\sum_{s=1}^{S} 1(X_i \in s)\left(\hat{t}_1(s) - \hat{t}_0(s)\right)\right]^2$$

Lemma B.3 of BCS 2017a and the continuous mapping theorem imply that $\hat{t}_1(s) \xrightarrow{p} t_1(s)$ and $\hat{t}_0(s) \xrightarrow{p} t_0(s)$. Slutsky's theorem then implies that $\hat{\Omega}$ consistently estimates the variance of (27).

K Proof of Proposition 7

This estimator can be implemented using OLS and 2SLS by fully interacting Z_i, the cluster dummies, and the additional regressors X_i. To simplify notation we denote $W_i = (1\ X_i)$ and

the regression functions in (12) as $\hat{\gamma}_{0s}'W_i$, $\hat{\gamma}_{1s}'W_i$, $\hat{\tau}_{0s}'W_i$ and $\hat{\tau}_{1s}'W_i$. Consider first the OLS case under Assumption 5.

$$\hat{\beta}_S = \sum_{s\in\mathcal{S}} \hat{p}(s)\,\bar{W}_s\,(\hat{\gamma}_{1s} - \hat{\gamma}_{0s})$$

where $\bar{W}_s = \frac{1}{n}\sum_{i=1}^n 1(X_i \in s)\,W_i/\hat{p}(s)$, $\hat{\gamma}_{0s} \xrightarrow{p} \gamma_{0s} = (E(WW'|s))^{-1}(E(WY_0|s))$, and $\hat{\gamma}_{1s} \xrightarrow{p} \gamma_{1s} = (E(WW'|s))^{-1}(E(WY_1|s))$, for

$$\hat{\gamma}_{1s} = H_{1n}^{-1}\left(\frac{1}{n}\sum_{i=1}^n 1(X_i \in s)\,Z_iW_iY_i\right) \quad \text{and} \quad H_{1n} = \left(\frac{1}{n}\sum_{i=1}^n 1(X_i \in s)\,Z_iW_iW_i'\right). \tag{30}$$

$$\hat{\gamma}_{0s} = H_{0n}^{-1}\left(\frac{1}{n}\sum_{i=1}^n 1(X_i \in s)\,(1-Z_i)\,W_iY_i\right) \quad \text{and} \quad H_{0n} = \left(\frac{1}{n}\sum_{i=1}^n 1(X_i \in s)\,(1-Z_i)\,W_iW_i'\right). \tag{31}$$

In the normal equations $EW(Y_j - W'\gamma_{js}|s) = 0$ for $j = 0, 1$, and W includes the constant term. Therefore $E(Y_j - W'\gamma_{js}|s) = 0$ for $j = 0, 1$, so that $\hat{\beta}_S \xrightarrow{p} \beta_0 = \Delta = E(Y_1 - Y_0)$. In the following, we will not require $p_z(s) \equiv p_z$. Note that

$$\hat{\beta}_S - \beta_0 = \underbrace{\sum_{s\in\mathcal{S}} \hat{p}(s)\,\bar{W}_s\,[\hat{\gamma}_{1s} - \gamma_{1s} - \hat{\gamma}_{0s} + \gamma_{0s}]}_{(1)} + \underbrace{\sum_{s\in\mathcal{S}} \hat{p}(s)\,\bar{W}_s\,(\gamma_{1s} - \gamma_{0s}) - \Delta}_{(2)},$$

where we can write (1) as

$$\sum_{s\in\mathcal{S}} \hat{p}(s)\,\bar{W}_s\left[H_{1n}^{-1}\frac{1}{n}\sum_{i=1}^n 1(X_i \in s)\,W_iZ_i\left(Y_{1i} - W_i'\gamma_{1s}\right) - H_{0n}^{-1}\frac{1}{n}\sum_{i=1}^n 1(X_i \in s)\,W_i(1-Z_i)\left(Y_{0i} - W_i'\gamma_{0s}\right)\right].$$

Using $\hat{p}(s) \xrightarrow{p} p(s)$, $\bar{W}_s \xrightarrow{p} E(W|s)$, $E(W'|s)\,E(WW'|s)^{-1} = (1, 0, ...)$,

$$H_{1n} \xrightarrow{p} p(s)\,p_z(s)\,E\left(WW'|s\right), \quad H_{0n} \xrightarrow{p} p(s)\,(1-p_z(s))\,E\left(WW'|s\right),$$

$$\frac{1}{n}\sum_{i=1}^n 1(X_i \in s)\,W_iZ_i\left(Y_{1i} - W_i'\gamma_{1s}\right) = O_P\left(\frac{1}{\sqrt{n}}\right), \quad \frac{1}{n}\sum_{i=1}^n 1(X_i \in s)\,W_i(1-Z_i)\left(Y_{0i} - W_i'\gamma_{0s}\right) = O_P\left(\frac{1}{\sqrt{n}}\right),$$

we can write (1) as

$$\begin{aligned}
&\sum_{s\in\mathcal{S}} E(W|s)\,E\left(WW'|s\right)^{-1}\frac{1}{n}\sum_{i=1}^n 1(X_i \in s)\,W_i\left(Z_i\frac{Y_{1i} - W_i'\gamma_{1s}}{p_z(s)} - (1-Z_i)\,\frac{Y_{0i} - W_i'\gamma_{0s}}{1-p_z(s)}\right) + o_P\left(\frac{1}{\sqrt{n}}\right)\\
&= \sum_{s\in\mathcal{S}}\frac{1}{n}\sum_{i=1}^n 1(X_i \in s)\left[(Z_i - p_z(s))\left(\frac{Y_{1i} - W_i'\gamma_{1s}}{p_z(s)} + \frac{Y_{0i} - W_i'\gamma_{0s}}{1-p_z(s)}\right)\right.\\
&\qquad\qquad \left. + \left[Y_{1i} - Y_{0i} - W_i'(\gamma_{1s} - \gamma_{0s})\right]\right] + o_P\left(\frac{1}{\sqrt{n}}\right)
\end{aligned}$$

Therefore,

$$\begin{aligned}
(1) + (2) = &\sum_{s\in\mathcal{S}}\frac{1}{n}\sum_{i=1}^n 1(X_i \in s)\,(Z_i - p_z(s))\left(\frac{Y_{1i} - W_i'\gamma_{1s}}{p_z(s)} + \frac{Y_{0i} - W_i'\gamma_{0s}}{1-p_z(s)}\right)\\
&+\frac{1}{n}\sum_{i=1}^n (Y_{1i} - Y_{0i} - \Delta) + o_P\left(\frac{1}{\sqrt{n}}\right)
\end{aligned}$$

This obviously is more efficient than (29) since $W_i'\gamma_{js}, j = 0, 1$ is the linear projection of $Y_{ij} - E(Y_{ij}|s)$ within cluster s, and results in a smaller variance.

Next we generalize the above to LATE. Consider

$$\hat{\beta}_S = \frac{\sum_{s\in\mathcal{S}} \hat{p}(s)\bar{W}_s(\hat{\gamma}_{1s} - \hat{\gamma}_{0s})}{\sum_{s\in\mathcal{S}} \hat{p}(s)\bar{W}_s(\hat{\tau}_{1s} - \hat{\tau}_{0s})}$$

so that for $\beta_0 = E(Y_1 - Y_0|D_1 > D_0)$,

$$\hat{\beta}_S - \beta_0 = \frac{\sum_{s\in\mathcal{S}} \hat{p}(s)\bar{W}_s\left(\hat{\gamma}_{1s} - \hat{\gamma}_{0s} - (\hat{\tau}_{1s} - \hat{\tau}_{0s})'\beta_0\right)}{\sum_{s\in\mathcal{S}} \hat{p}(s)\bar{W}_s(\hat{\tau}_{1s} - \hat{\tau}_{0s})}$$

Since the denominator is $E(D_1 - D_0) + o_P(1) = P(D_1 > D_0) + o_P(1)$, we focus on the numerator, and write

$$(P(D_1 > D_0) + o_P(1))\left(\hat{\beta}_S - \beta_0\right) = \sum_{s\in\mathcal{S}} \hat{p}(s)\bar{W}_s\left(\hat{\gamma}_{1s} - \hat{\gamma}_{0s} - (\hat{\tau}_{1s} - \hat{\tau}_{0s})'\beta_0\right).$$

γ_{1s} and γ_{0s} are defined by

$$\hat{\gamma}_{1s} = H_{1n}^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} 1(X_i \in s) Z_i W_i Y_i\right) \xrightarrow{p} \gamma_{1s} = \left(E\left(WW'|s\right)\right)^{-1}(E(WY_1^*|s))$$

$$\hat{\gamma}_{0s} = H_{0n}^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} 1(X_i \in s)(1 - Z_i) W_i Y_i\right) \xrightarrow{p} \gamma_{0s} = \left(E\left(WW'|s\right)\right)^{-1}(E(WY_0^*|s)),$$

and τ_{1s} and τ_{0s} are analogously defined by

$$\hat{\tau}_{1s} = H_{1n}^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} 1(X_i \in s) Z_i W_i D_i\right) \xrightarrow{p} \tau_{1s} = \left(E\left(WW'|s\right)\right)^{-1}(E(WD_1|s))$$

$$\hat{\tau}_{0s} = H_{0n}^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} 1(X_i \in s)(1 - Z_i) W_i D_i\right) \xrightarrow{p} \tau_{0s} = \left(E\left(WW'|s\right)\right)^{-1}(E(WD_0|s)).$$

Define $\hat{\eta}_{js} = \hat{\gamma}_{js} - \hat{\tau}_{js}'\beta_0$ for $j = 0, 1$, so that $\hat{\eta}_{js} \xrightarrow{p} \eta_{js} = E(WW'|s)^{-1} E(Wt_j|s)$, where

$$\hat{\eta}_{1s} = \left(\frac{1}{n}\sum_{i=1}^{n} 1(X_i \in s) Z_i W_i W_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} 1(X_i \in s) Z_i W_i \underbrace{\left(Y_{1i}^* - D_{1i}'\beta_0\right)}_{t_{1i}}\right)$$

$$\hat{\eta}_{0s} = \left(\frac{1}{n}\sum_{i=1}^{n} 1(X_i \in s)(1 - Z_i) W_i W_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} 1(X_i \in s)(1 - Z_i) W_i \underbrace{\left(Y_{0i}^* - D_{0i}'\beta_0\right)}_{t_{0i}}\right)$$

Then we proceed similar as the ATE case to write the numerator as

$$\underbrace{\sum_{s\in\mathcal{S}} \hat{p}(s)\bar{W}_s[\hat{\eta}_{1s} - \eta_{1s} - \hat{\eta}_{0s} + \eta_{0s}]}_{(1)} + \underbrace{\sum_{s\in\mathcal{S}} \hat{p}(s)\bar{W}_s(\eta_{1s} - \eta_{0s})}_{(2)},$$

where by noting that

$$\frac{1}{n}\sum_{i=1}^{n} 1\left(X_i \in s\right) W_i Z_i \left(t_{1i} - W_i'\eta_{1s}\right) = O_P\left(\frac{1}{\sqrt{n}}\right), \quad \frac{1}{n}\sum_{i=1}^{n} 1\left(X_i \in s\right) W_i \left(1 - Z_i\right)\left(t_{0i} - W_i'\eta_{0s}\right) = O_P\left(\frac{1}{\sqrt{n}}\right),$$

we can write (1) as

$$\begin{aligned}
&\sum_{s\in\mathcal{S}} E\left(W|s\right) E\left(WW'|s\right)^{-1} \frac{1}{n}\sum_{i=1}^{n} 1\left(X_i \in s\right) W_i \left(Z_i \frac{t_{1i} - W_i'\eta_{1s}}{p_z\left(s\right)} - \left(1 - Z_i\right)\frac{t_{0i} - W_i'\eta_{0s}}{1 - p_z\left(s\right)}\right) + o_P\left(\frac{1}{\sqrt{n}}\right) \\
&= \sum_{s\in\mathcal{S}} \frac{1}{n}\sum_{i=1}^{n} 1\left(X_i \in s\right)\left[\left(Z_i - p_z\left(s\right)\right)\left(\frac{t_{1i} - W_i'\eta_{1s}}{p_z\left(s\right)} + \frac{t_{0i} - W_i'\eta_{0s}}{1 - p_z\left(s\right)}\right) + \left[t_{1i} - t_{0i} - W_i'\left(\eta_{1s} - \eta_{0s}\right)\right]\right] + o_P\left(\frac{1}{\sqrt{n}}\right)
\end{aligned}$$

Therefore,

$$\begin{aligned}
(1) + (2) = &\sum_{s\in\mathcal{S}} \frac{1}{n}\sum_{i=1}^{n} 1\left(X_i \in s\right)\left(Z_i - p_z\left(s\right)\right)\left(\frac{t_{1i} - W_i'\eta_{1s}}{p_z\left(s\right)} + \frac{t_{0i} - W_i'\eta_{0s}}{1 - p_z\left(s\right)}\right) \\
&+ \frac{1}{n}\sum_{i=1}^{n}\left(t_{1i} - t_{0i}\right) + o_P\left(\frac{1}{\sqrt{n}}\right)
\end{aligned} \tag{32}$$

Again this ought to be more efficient than (27) since $W_i'\eta_{js}$ is the within cluster linear projection of $t_{ji} - t_j\left(s\right)$. The more variables the projection is on, the smaller the variance. As $\dim\left(W\right) \to \infty$ at an appropriate rate, $W_i'\eta_{js} \to E\left(t_{ji}|W_i\right)$ for $j = 0, 1$, so that the above equation becomes the efficient influence function in (24) conditional on both the cluster indicators and the extra regressors.