Inference under Covariate-Adaptive Randomization with Multiple Treatments *

Federico A. Bugni
Department of Economics
Duke University
federico.bugni@duke.edu

Ivan A. Canay
Department of Economics
Northwestern University
iacanay@northwestern.edu

Azeem M. Shaikh
Department of Economics
University of Chicago
amshaikh@uchicago.edu

January 17, 2019

*We would like to thank Lori Beaman, Joseph Romano, Andres Santos, and seminar participants at various institutions for helpful comments on this paper. Yuehao Bai, Jackson Bunting, Mengsi Gao, Max Tabord-Meehan, Vishal Kamat, and Winnie van Dijk provided excellent research assistance. The research of the first author was supported by National Institutes of Health Grant 40-4153-00-0-85-399 and the National Science Foundation Grant SES-1729280. The research of the second author was supported by National Science Foundation Grant SES-1530534. The research of the third author was supported by National Science Foundation Grants SES-1308260, SES-1227091, and SES-1530661.

Abstract

This paper studies inference in randomized controlled trials with covariate-adaptive randomization when there are multiple treatments. More specifically, we study in this setting inference about the average effect of one or more treatments relative to other treatments or a control. As in Bugni et al. (2018), covariate-adaptive randomization refers to randomization schemes that first stratify according to baseline covariates and then assign treatment status so as to achieve "balance" within each stratum. Importantly, in contrast to Bugni et al. (2018), we not only allow for multiple treatments, but further allow for the proportion of units being assigned to each of the treatments to vary across strata. We first study the properties of estimators derived from a "fully saturated" linear regression, i.e., a linear regression of the outcome on all interactions between indicators for each of the treatments and indicators for each of the strata. We show that tests based on these estimators using the usual heteroskedasticity-consistent estimator of the asymptotic variance are invalid in the sense that they may have limiting rejection probability under the null hypothesis strictly greater than the nominal level; on the other hand, tests based on these estimators and suitable estimators of the asymptotic variance that we provide are exact in the sense that they have limiting rejection probability under the null hypothesis equal to the nominal level. For the special case in which the target proportion of units being assigned to each of the treatments does not vary across strata, we additionally consider tests based on estimators derived from a linear regression with "strata fixed effects," i.e., a linear regression of the outcome on indicators for each of the treatments and indicators for each of the strata. We show that tests based on these estimators using the usual heteroskedasticity-consistent estimator of the asymptotic variance are conservative in the sense that they have limiting rejection probability under the null hypothesis no greater than and typically strictly less than the nominal level, but tests based on these estimators and suitable estimators of the asymptotic variance that we provide are exact, thereby generalizing results in Bugni et al. (2018) for the case of a single treatment to multiple treatments. A simulation study and an empirical application illustrate the practical relevance of our theoretical results.

KEYWORDS: Covariate-adaptive randomization, multiple treatments, stratified block randomization, Efron's biased-coin design, treatment assignment, randomized controlled trial, strata fixed effects, saturated regression

JEL classification codes: C12, C14

1 Introduction

This paper studies inference in randomized controlled trials with covariate-adaptive randomization when there are multiple treatments. As in Bugni et al. (2018), covariate-adaptive randomization refers to randomization schemes that first stratify according to baseline covariates and then assign treatment status so as to achieve "balance" within each stratum. Many such methods are used routinely when assigning treatment status in randomized controlled trials in all parts of the sciences. See, for example, Rosenberger and Lachin (2016) for a textbook treatment focused on clinical trials and Duflo et al. (2007) and Bruhn and McKenzie (2009) for reviews focused on development economics. Importantly, in contrast to Bugni et al. (2018), we not only allow for multiple treatments, but further allow the target proportion of units being assigned to each of the treatments to vary across strata. In this paper, we take as given the use of such a treatment assignment mechanism and study its consequences for inference about the average effect of one or more treatments relative to other treatments or a control. Our main requirement is that the randomization scheme is such that the fraction of units being assigned to each treatment within each stratum is suitably well behaved in a sense made precise by our assumptions below as the sample size n tends to infinity. See, in particular, Assumptions 2.2.(b) and 4.1.(c). Importantly, these requirements are satisfied by most commonly used treatment assignment mechanisms, including simple random sampling and stratified block randomization. The latter treatment assignment scheme is especially noteworthy because of its widespread use recently in development economics. See, for example, Dizon-Ross (2018, footnote 13), Duflo et al. (2015, footnote 6), Callen et al. (2019, page 24), and Berry et al. (2018, page 6).

We first study the properties of ordinary least squares estimation of a "fully saturated" linear regression, i.e., a linear regression of the outcome on all interactions between indicators for each of the treatments and indicators for each of the strata. We emphasize that tests based on these estimators were not considered previously in Bugni et al. (2018). We show that tests based on these estimators using the usual heteroskedasticity-consistent estimator of the asymptotic variance are invalid in the sense that they may have limiting rejection probability under the null hypothesis strictly greater than the nominal level. As explained further below, this phenomenon contrasts sharply with the analysis in Bugni et al. (2018) of other tests that were found to be conservative in the sense that their limiting rejection probabilities were no greater than the nominal level. We then exploit our characterization of the behavior of the ordinary least squares estimator of the coefficients in such a regression under covariate-adaptive randomization to develop a consistent estimator of the asymptotic variance. Our main result about the "fully saturated" linear regression shows that tests based on these estimators and our new estimator of the asymptotic variance are exact in the sense that they have limiting rejection probability under the null hypothesis equal to the nominal level. In a simulation study, we find that tests using the usual heteroskedasticity-consistent estimator of the asymptotic variance may have rejection probability under the null hypothesis dramatically larger than the nominal level. On the other hand, tests using the new estimator of the asymptotic variance have rejection probability under the null hypothesis very close to the nominal level.

We additionally consider tests based on ordinary least squares estimation of a linear regression with "strata fixed effects," i.e., a linear regression of the outcome on indicators for each of the treatments and

indicators for each of the strata. As emphasized by Imbens and Rubin (2015, Ch. 9) in the case of a single treatment, such estimators need not even be consistent for the average treatment effect when the target proportion of units being assigned to treatment varies across strata, so in our analysis of tests based on these estimators we restrict attention to the special case in which the target proportion of units being assigned to each of the treatments does not vary across strata. Based on simulation evidence and earlier assertions by Kernan et al. (1999), the use of this test has been recommended by Bruhn and McKenzie (2009). More recently, Bugni et al. (2018) provided a formal analysis of the properties of tests based on these estimators in the case of a single treatment. In this paper, we extend the analysis in Bugni et al. (2018) about these tests to multiple treatments. We show that tests based on these estimators using the usual heteroskedasticity-consistent estimator of the asymptotic variance are conservative in the sense that they have limiting rejection probability under the null hypothesis no greater than, and typically strictly less than, the nominal level. Once again, we exploit our characterization of the behavior of the ordinary least squares estimator of the coefficients in such a regression under covariate-adaptive randomization to develop a consistent estimator of the asymptotic variance. Our main result about the linear regression with "strata fixed effects" shows that tests based on these estimators and our new estimator of the asymptotic variance are exact in the sense that they have limiting rejection probability under the null hypothesis equal to the nominal level. In a simulation study, we find that tests using the usual heteroskedasticity-consistent estimator of the asymptotic variance may have rejection probability under the null hypothesis dramatically less than the nominal level and, as a result, may have very poor power when compared to other tests. On the other hand, tests using the new estimator of the asymptotic variance have rejection probability under the null hypothesis very close to the nominal level.

The remainder of the paper is organized as follows. In Section 2, we describe our setup and notation. In particular, there we describe the assumptions we impose on the treatment assignment mechanism. Our main results concerning the "fully saturated" linear regression are contained in Section 3. Our main results concerning the linear regression with "strata fixed effects" are contained in Section 4. In Section 5, we discuss our results in the special case where there is only a single treatment, which facilitates a comparison of our results with those in Imbens and Rubin (2015, Chapter 9). In Section 6, we examine the finite-sample behavior of all the tests we consider in this paper via a small simulation study. In Section 7, we provide recommendations for empirical practice. Finally, in Section 8, we provide an empirical illustration of our results. Proofs of all results are provided in the Appendix.

2 Setup and Notation

Let Y_i denote the (observed) outcome of interest for the ith unit, A_i denote the treatment received by the ith unit, and Z_i denote observed, baseline covariates for the ith unit. The list of possible treatments is given by $\mathcal{A} = \{1, \ldots, |\mathcal{A}|\}$, and we say there are multiple treatments when $|\mathcal{A}| > 1$. Without loss of generality we assume there is a control group, which we denote as treatment zero, and use $\mathcal{A}_0 = \{0\} \cup \mathcal{A}$ to denote the list of treatments that includes the control group. Denote by $Y_i(a)$ the potential outcome of the ith unit under treatment $a \in \mathcal{A}_0$. As usual, the (observed) outcome and potential outcomes are related to treatment

assignment by the relationship

$$Y_i = \sum_{a \in \mathcal{A}_0} Y_i(a) I\{A_i = a\} = Y_i(A_i) \ . \tag{1}$$

Denote by P_n the distribution of the observed data

$$X^{(n)} = \{(Y_i, A_i, Z_i) : 1 \leq i \leq n\}$$

and denote by Q_n the distribution of

$$W^{(n)} = \{(Y_i(0), Y_i(1), \ldots, Y_i(|\mathcal{A}|), Z_i) : 1 \leq i \leq n\} \ .$$

Note that P_n is jointly determined by (1), Q_n, and the mechanism for determining treatment assignment. We therefore state our assumptions below in terms of assumptions on Q_n and assumptions on the mechanism for determining treatment status. Indeed, we will not make reference to P_n in the sequel and all operations are understood to be under Q_n and the mechanism for determining treatment status.

Strata are constructed from the observed, baseline covariates Z_i using a function $S : \text{supp}(Z_i) \to \mathcal{S}$, where $\mathcal{S}$ is a finite set. For $1 \leq i \leq n$, let $S_i = S(Z_i)$ and denote by $S^{(n)}$ the vector of strata $(S_1, \ldots, S_n)$.

We begin by describing our assumptions on Q_n. We assume that $W^{(n)}$ consists of n i.i.d. observations, i.e., $Q_n = Q^n$, where Q is the marginal distribution of $(Y_i(0), Y_i(1), \ldots, Y_i(|\mathcal{A}|), Z_i)$. In order to rule out trivial strata, we henceforth assume that $p(s) = P\{S_i = s\} > 0$ for all $s \in \mathcal{S}$. We further restrict Q to satisfy the following mild requirement.

Assumption 2.1. Q satisfies

$$\max_{a \in \mathcal{A}_0} E[|Y_i(a)|^2] < \infty$$

and for all $a \in \mathcal{A}_0$

$$\max_{s \in \mathcal{S}} \text{Var}[Y_i(a) | S_i = s] > 0 \ .$$

We note that the second requirement in Assumption 2.1 is made only to rule out degenerate situations and is stronger than required for our results.

Next, we describe our assumptions on the mechanism determining treatment assignment. As mentioned previously, in this paper we focus on covariate-adaptive randomization, i.e., randomization schemes that first stratify according baseline covariates and then assign treatment status so as to achieve "balance" within each stratum. In order to describe our assumptions on the treatment assignment mechanism more formally, we require some further notation. Let $A^{(n)}$ be vector of treatment assignments $(A_1, \ldots, A_n)$. For any $(a, s) \in \mathcal{A}_0 \times \mathcal{S}$, let $\pi_a(s) \in (0, 1)$ be the target proportion of units to assign to treatment a in stratum s, let

$$n_a(s) = \sum_{1 \leq i \leq n} I\{A_i = a, S_i = s\}$$

be the number of units assigned to treatment a in stratum s, and let

$$n(s) = \sum_{1 \leq i \leq n} I\{S_i = s\}$$

be the total number of units in stratum s. Note that $\sum_{a \in \mathcal{A}_0} \pi_a(s) = 1$ for all $s \in \mathcal{S}$. The following assumption summarizes our main requirement on the treatment assignment mechanism for the analysis of the "fully saturated" linear regression.

Assumption 2.2. The treatment assignment mechanism is such that

(a) $W^{(n)} \perp\!\!\!\perp A^{(n)} | S^{(n)}$.

(b) $\frac{n_a(s)}{n(s)} \xrightarrow{P} \pi_a(s)$ as $n \to \infty$ for all $(a, s) \in \mathcal{A} \times \mathcal{S}$.

Assumption 2.2.(a) simply requires that the treatment assignment mechanism is a function only of the vector of strata and an exogenous randomization device. Assumption 2.2.(b) is an additional requirement that imposes that the (possibly random) fraction of units assigned to treatment a and stratum s approaches the target proportion $\pi_a(s)$ as the sample size tends to infinity. This requirement is satisfied by a wide variety of randomization schemes; see Bugni et al. (2018), Rosenberger and Lachin (2016, Sections 3.10 and 3.11), and Wei et al. (1986, Proposition 2.5). Before proceeding, we briefly discuss two popular randomization schemes that are easily seen to satisfy Assumption 2.2.

Example 2.1. *(Simple Random Sampling)* Simple random sampling (SRS), also known as Bernoulli trials, refers to the case where $A^{(n)}$ consists of n i.i.d. random variables with

$$P\{A_k = a | S^{(n)}, A^{(k-1)}\} = P\{A_k = a\} = \pi_a \tag{2}$$

for $1 \leq k \leq n$ and $\pi_a \in (0, 1)$ satisfying $\sum_{a \in \mathcal{A}_0} \pi_a = 1$. In this case, Assumption 2.2.(a) follows immediately from (2), while Assumption 2.2.(b) follows from the weak law of large numbers. If (2) is such that the target probabilities π_a vary by strata, then

$$P\{A_k = a | S^{(n)}, A^{(k-1)}\} = P\{A_k = a | S_k = s\} = \pi_a(s) \ ,$$

which is equivalent to simple random sampling within each stratum. ■

Example 2.2. *(Stratified Block Randomization)* An early discussion of stratified block randomization (SBR) is provided by Zelen (1974) for the case of a single treatment. This randomization scheme is sometimes also referred to as block randomization or permuted blocks within strata. In order to describe this treatment assignment mechanism, for $s \in \mathcal{S}$, denote by $n(s)$ the number of units in stratum s and let

$$n_a(s) = \lfloor n(s)\pi_a(s) \rfloor$$

for $a \in \mathcal{A}$ with $n_0(s) = n(s) - \sum_{a \in \mathcal{A}} n_a(s)$. In this randomization scheme, independently for each each

stratum s, $n_a(s)$ units are assigned to each treatment a, where all

$$\binom{n(s)}{n_0(s), n_1(s), \ldots, n_{|\mathcal{A}|}(s)}$$

possible assignments are equally likely. Assumptions 2.2.(a) and 2.2.(b) follow by construction in this case. ■

We note that our analysis of the linear regression with "strata fixed effects" requires an assumption that is mildly stronger than Assumption 2.2 above. It is worth emphasizing that this stronger assumption parallels the assumption made in Bugni et al. (2018) for the analysis of linear regression with "strata fixed effects" in the case of a single treatment and is also satisfied by a wide variety of treatment assignment mechanisms, including Examples 2.1 and 2.2 above. See Assumption 4.1 and the subsequent discussion there for further details.

Our object of interest is the vector of average treatment effects (ATEs) on the outcome of interest. For each $a \in \mathcal{A}$, we use

$$\theta_a(Q) \equiv E[Y_i(a) - Y_i(0)] \tag{3}$$

to denote the ATE of treatment a relative to the control and

$$\theta(Q) \equiv (\theta_a(Q) : a \in \mathcal{A}) = (\theta_1(Q), \ldots, \theta_{|\mathcal{A}|}(Q))'$$

to denote the $|\mathcal{A}|$-dimensional vector of such ATEs. Our results permit testing a variety of hypotheses on smooth functions of the vector $\theta(Q)$ at level $\alpha \in (0, 1)$. In particular, hypotheses on linear functionals can be written as

$$H_0 : \Psi\theta(Q) = c \text{ versus } H_1 : \Psi\theta(Q) \neq c \ , \tag{4}$$

where Ψ is a full-rank $(r \times |\mathcal{A}|)$-dimensional matrix and c is a r-dimensional column vector. This framework accommodates, for example, hypotheses on a particular ATE,

$$H_0 : \theta_a(Q) = c \text{ versus } H_1 : \theta_a(Q) \neq c \ , \tag{5}$$

as well as hypotheses comparing treatment effects,

$$H_0 : \theta_a(Q) = \theta_{a'}(Q) \text{ versus } H_1 : \theta_a(Q) \neq \theta_{a'}(Q) \text{ for any } a, a' \in \mathcal{A} \ . \tag{6}$$

Note that $\theta_a(Q) = \theta_{a'}(Q)$ if and only if $E[Y_i(a)] = E[Y_i(a')]$. We note further that it is also possible to use our results to test smooth non-linear hypotheses on $\theta(Q)$ via the Delta method, but, for ease of exposition, we restrict our attention to linear restrictions as described above in what follows.

Finally, we often transform objects that are indexed by $(a, s) \in \mathcal{A} \times \mathcal{S}$ into vectors or matrices, using the following conventions. For $X(a)$ being a scalar object indexed over $a \in \mathcal{A}$, we use $(X(a) : a \in \mathcal{A})$ to denote the $|\mathcal{A}|$-dimensional column vector $(X(1), \ldots, X(|\mathcal{A}|))'$. For $X_a(s)$ being a scalar object indexed by $(a, s) \in \mathcal{A} \times \mathcal{S}$ we use $(X_a(s) : (a, s) \in \mathcal{A} \times \mathcal{S})$ to denote the $(|\mathcal{A}| \times |\mathcal{S}|)$-dimensional column vector where

the order of the indices matter: first we iterate over a and then over s, i.e.,

$$(X_a(s) : (a,s) \in \mathcal{A} \times \mathcal{S}) \equiv (X_1(1), \ldots, X_{|\mathcal{A}|}(1), X_1(2), \ldots, X_{|\mathcal{A}|}(2), \ldots)' \ .$$

Remark 2.1. The term "balance" is often used in a different way to describe whether the distributions of baseline covariates Z_i in the treatment and control groups are similar. For example, this might be measured according to the difference in the means of Z_i in the treatment and control groups. Our usage follows the usage in Efron (1971) or Hu and Hu (2012), where "balance" refers to the extent to which the of fraction of treated units within a strata differs from the target proportion $\pi_a(s)$. ■

3 "Fully Saturated" Linear Regression

In this section, we study the properties of ordinary least squares estimation of a linear regression of the outcome on all interactions between indicators for each of the treatments and indicators for each of the strata under covariate-adaptive randomization. We then study the properties of different tests of (4) based on these estimators. As already noted, these tests have not been previously considered in Bugni et al. (2018). We consider tests using both the usual homoskedasticity-only and heteroskedasticity-robust estimators of the asymptotic variance. Our results show that neither of these estimators are consistent for the asymptotic variance, and, as a result, both lead to tests that are asymptotically invalid in the sense that they may have limiting rejection probability under the null hypothesis strictly greater than the nominal level. In light of these results, we exploit our characterization of the behavior of the ordinary least squares estimator of the coefficients in such a regression under covariate-adaptive randomization to develop a consistent estimator of the asymptotic variance. Furthermore, tests using our new estimator of the asymptotic variance are exact in the sense that they have limiting rejection probability under the null hypotheses equal to the nominal level.

In order to define the tests we study, consider estimation of the equation

$$Y_i = \sum_{s \in \mathcal{S}} \delta(s) I\{S_i = s\} + \sum_{(a,s) \in \mathcal{A} \times \mathcal{S}} \beta_a(s) I\{A_i = a, S_i = s\} + u_i \tag{7}$$

by ordinary least squares. For all $s \in \mathcal{S}$, denote by $\hat{\delta}_n(s)$ and $\hat{\beta}_{n,a}(s)$ the resulting estimators of $\delta(s)$ and $\beta_a(s)$, respectively. The corresponding estimator of the ATE of treatment a is given by

$$\hat{\theta}_{n,a} = \sum_{s \in \mathcal{S}} \frac{n(s)}{n} \hat{\beta}_{n,a}(s) \ , \tag{8}$$

and the resulting estimator of $\theta(Q)$ is thus given by

$$\hat{\theta}_n = (\hat{\theta}_{n,a} : a \in \mathcal{A}) \equiv (\hat{\theta}_{n,1}, \ldots, \hat{\theta}_{n,|\mathcal{A}|})' \ . \tag{9}$$

Let $\hat{\mathbb{V}}_n$ be an estimator of the asymptotic covariance matrix of $\hat{\theta}_n$. For testing the hypotheses in (4), we

consider tests of the form

$$\phi_n^{\text{sat}}(X^{(n)}) = I\{T_n^{\text{sat}}(X^{(n)}) > \chi^2_{r,1-\alpha}\} \ , \tag{10}$$

where

$$T_n^{\text{sat}}(X^{(n)}) = n(\Psi\hat{\theta}_n - c)'(\Psi\hat{\mathbb{V}}_n\Psi')^{-1}(\Psi\hat{\theta}_n - c)$$

and $\chi^2_{r,1-\alpha}$ is the $1-\alpha$ quantile of a χ^2 random variable with r degrees of freedom. In order to study the properties of this test, we first derive in the following theorem the asymptotic behavior of $\hat{\theta}_n$.

Theorem 3.1. *Suppose Q satisfies Assumption 2.1 and the treatment assignment mechanism satisfies Assumption 2.2. Then,*

$$\sqrt{n}(\hat{\theta}_n - \theta(Q)) \xrightarrow{d} N(0, \mathbb{V}_{\text{sat}}) \ ,$$

where $\mathbb{V}_{\text{sat}} = \mathbb{V}_H + \mathbb{V}_{\tilde{Y}}$,

$$\mathbb{V}_H \equiv \sum_{s\in\mathcal{S}} p(s)\left(E[m_a(Z_i) - m_0(Z_i)|S_i = s] : a \in \mathcal{A}\right)\left(E[m_a(Z_i) - m_0(Z_i)|S_i = s] : a \in \mathcal{A}\right)' \tag{11}$$

$$\mathbb{V}_{\tilde{Y}} \equiv \sum_{s\in\mathcal{S}} \frac{p(s)\sigma^2_{\tilde{Y}(0)}(s)}{\pi_0(s)}\iota_{|\mathcal{A}|}\iota'_{|\mathcal{A}|} + \text{diag}\left(\sum_{s\in\mathcal{S}} \frac{p(s)\sigma^2_{\tilde{Y}(a)}(s)}{\pi_a(s)} : a \in \mathcal{A}\right) \ , \tag{12}$$

$\iota_{|\mathcal{A}|}$ is a $|\mathcal{A}|$-dimensional vector of ones, and

$$\begin{aligned} m_a(Z_i) &\equiv E[Y_i(a)|Z_i] - E[Y_i(a)] \\ \sigma^2_{\tilde{Y}(a)}(s) &\equiv \text{Var}[\tilde{Y}_i(a)|S_i = s] \\ \tilde{Y}_i(a) &\equiv Y_i(a) - E[Y_i(a)|S_i = s] \ . \end{aligned}$$

Remark 3.1. For each $a \in \mathcal{A}$, note that

$$\begin{aligned} \sqrt{n}(\hat{\theta}_{n,a} - \theta_a(Q)) &= \sum_{s\in\mathcal{S}}\left(\sqrt{n}\left(\frac{n(s)}{n} - p(s)\right)\hat{\beta}_{n,a}(s) + \sqrt{n}(\hat{\beta}_{n,a}(s) - \beta_a(s))p(s)\right) \\ &= \sum_{s\in\mathcal{S}}\left(\sqrt{n}\left(\frac{n(s)}{n} - p(s)\right)\beta_a(s) + \sqrt{n}(\hat{\beta}_{n,a}(s) - \beta_a(s))p(s)\right) + o_P(1) \ , \end{aligned}$$

where the second equality exploits a novel law of large numbers that accounts for covariate-adaptive randomization (see Lemma C.4) and the central limit theorem. It is therefore straightforward to derive the conclusion of Theorem 3.1 from the limit in distribution of

$$\left(\sqrt{n}\left(\frac{n(s)}{n} - p(s)\right), \sqrt{n}(\hat{\beta}_{n,a}(s) - \beta_a(s)) : (a,s) \in \mathcal{A}\times\mathcal{S}\right) \ . \tag{13}$$

The derivation of the limit in distribution of (13) does not follow from conventional central limit theorems due to covariate-adaptive randomization. These difficulties are overcome in Lemma C.1 in the Appendix using a novel coupling-like argument in combination with results about partial sums. ■

The following theorem characterizes the limits in probability for the usual homoskedasticity-only and heteroskedasticity-robust estimators of the asymptotic variance. It shows, in particular, that neither $\hat{\mathbb{V}}_{\text{ho}}$ nor $\hat{\mathbb{V}}_{\text{hc}}$ are consistent for the asymptotic variance of $\hat{\theta}_n$, $\mathbb{V}_{\text{sat}}$.

Theorem 3.2. *Suppose Q satisfies Assumption 2.1 and the treatment assignment mechanism satisfies Assumption 2.2. Let $\hat{\mathbb{V}}_{\text{ho}}$ be the homoskedasticity-only estimator of the asymptotic variance defined in (B-35) and $\hat{\mathbb{V}}_{\text{hc}}$ be the heteroskedasticity-consistent estimator of the asymptotic variance defined in (B-36). Then,*

$$\hat{\mathbb{V}}_{\text{ho}} \xrightarrow{P} \sum_{(a,s)\in\mathcal{A}_0\times\mathcal{S}} p(s)\pi_a(s)\sigma^2_{\tilde{Y}(a)}(s) \left[\sum_{s\in\mathcal{S}} \frac{p(s)}{\pi_0(s)} \iota_{|\mathcal{A}|}\iota'_{|\mathcal{A}|} + \text{diag}\left(\sum_{s\in\mathcal{S}} \frac{p(s)}{\pi_a(s)} : a \in \mathcal{A}\right)\right]$$

and

$$\hat{\mathbb{V}}_{\text{hc}} \xrightarrow{P} \sum_{s\in\mathcal{S}} \frac{p(s)\sigma^2_{\tilde{Y}(0)}(s)}{\pi_0(s)} \iota_{|\mathcal{A}|}\iota'_{|\mathcal{A}|} + \text{diag}\left(\sum_{s\in\mathcal{S}} \frac{p(s)\sigma^2_{\tilde{Y}(a)}(s)}{\pi_a(s)} : a \in \mathcal{A}\right) \ .$$

Remark 3.2. In the special case with a single treatment, i.e. $|\mathcal{A}| = 1$, we show in Section 5 that the limit in probability of $\hat{\mathbb{V}}_{\text{hc}}$ could be strictly smaller than $\mathbb{V}_{\text{sat}}$. Therefore, testing (4) using (10) with $\hat{\mathbb{V}}_n = \hat{\mathbb{V}}_{\text{hc}}$ could lead to over-rejection. In our simulation study in Section 6, we find that the rejection probability may in fact be substantially larger than the nominal level. ■

Remark 3.3. It is important to note that in the special case where $|\mathcal{A}| = 1$ and $\pi_1(s) = \frac{1}{2}$ for all $s \in \mathcal{S}$, both $\hat{\mathbb{V}}_{\text{ho}}$ and $\hat{\mathbb{V}}_{\text{hc}}$ are consistent for $\mathbb{V}_{\text{sat}}$. The particular properties of this special case have been already highlighted by Bugni et al. (2018) in the cases of the two-sample t-test, t-test with strata fixed effects, and covariate-adaptive permutation tests. ■

Even though $\hat{\mathbb{V}}_{\text{hc}}$ is generally inconsistent for $\mathbb{V}_{\text{sat}}$, the proof of Theorem 3.2 reveals that

$$\hat{\mathbb{V}}_{\text{hc}} \xrightarrow{P} \mathbb{V}_{\tilde{Y}} \ , \tag{14}$$

under the same assumptions. We exploit this observation in the following theorem to construct a consistent estimator of the asymptotic variance. The theorem further establishes that tests using this new estimator of the asymptotic variance are exact in the sense that they have limiting rejection probability under the null hypotheses equal to the nominal level.

Theorem 3.3. *Suppose Q satisfies Assumption 2.1 and the treatment assignment mechanism satisfies Assumption 2.2. Let $\hat{\mathbb{V}}_{\text{hc}}$ be the heteroskedasticity-consistent estimator of the asymptotic variance defined in (B-36) and let*

$$\hat{\mathbb{V}}_H = \sum_{s\in\mathcal{S}} \frac{n(s)}{n} \left(\hat{\beta}_{n,a}(s) - \hat{\theta}_{n,a} : a \in \mathcal{A}\right)\left(\hat{\beta}_{n,a}(s) - \hat{\theta}_{n,a} : a \in \mathcal{A}\right)' \ , \tag{15}$$

where $\hat{\theta}_{n,a}$ is as in (8) and $\hat{\beta}_{n,a}(s)$ is the ordinary least squares estimator of $\beta_a(s)$ in (7). Then,

$$\hat{\mathbb{V}}_{\text{sat}} = \hat{\mathbb{V}}_H + \hat{\mathbb{V}}_{\text{hc}} \xrightarrow{P} \mathbb{V}_{\text{sat}} = \mathbb{V}_H + \mathbb{V}_{\tilde{Y}} \ . \tag{16}$$

In addition, for the problem of testing (4) at level $\alpha \in (0,1)$, $\phi_n^{\text{sat}}(X^{(n)})$ defined in (10) with $\hat{\mathbb{V}}_n = \hat{\mathbb{V}}_{\text{sat}}$

satisfies

$$\lim_{n\to\infty} E[\phi_n^{\text{sat}}(X^{(n)})] = \alpha \tag{17}$$

for Q additionally satisfying the null hypothesis, i.e., $\Psi\theta(Q) = c$.

4 Linear Regression with "Strata Fixed Effects"

In this section, we study the properties of ordinary least squares estimation of a linear regression of the outcome on indicators for each of the treatments and indicators for each of the strata under covariate-adaptive randomization. We then study the properties of different tests of (4) based on these estimators. As before, we consider tests using both the usual homoskedasticity-only and heteroskedasticity-robust estimators of the asymptotic variance, and our results show that neither of these estimators are consistent for the asymptotic variance. We therefore exploit, as in the previous section, our characterization of the behavior of the ordinary least squares estimator of the coefficients in such a regression under covariate-adaptive randomization to develop a consistent estimator of the asymptotic variance, which leads to tests that are exact in the sense that they have limiting rejection probability under the null hypotheses equal to the nominal level.

In order to define the tests we study, consider estimation of the equation

$$Y_i = \sum_{s\in\mathcal{S}} \delta_s^* I\{S_i = s\} + \sum_{a\in\mathcal{A}} \beta_a^* I\{A_i = a\} + u_i \tag{18}$$

by ordinary least squares. Denote by $\hat{\beta}_{n,a}^*$ the resulting estimator of β_a^* in (18). The corresponding estimator of the ATE of treatment a is simply given by $\hat{\beta}_{n,a}^*$, and the resulting estimator of $\theta(Q)$ is thus given by

$$\hat{\theta}_n^* = (\hat{\beta}_{n,a}^* : a \in \mathcal{A}) \equiv (\hat{\beta}_{n,1}^*, \ldots, \hat{\beta}_{n,|\mathcal{A}|}^*)' \ . \tag{19}$$

Let $\hat{\mathbb{V}}_n^*$ be an estimator of the asymptotic variance of $\hat{\theta}_n^*$. For testing the hypotheses in (4), we consider tests of the form

$$\phi_n^{\text{sfe}}(X^{(n)}) = I\{T_n^{\text{sfe}}(X^{(n)}) > \chi_{r,1-\alpha}^2\} \ , \tag{20}$$

where

$$T_n^{\text{sfe}}(X^{(n)}) = n(\Psi\hat{\theta}_n^* - c)'(\Psi\hat{\mathbb{V}}_n^*\Psi')^{-1}(\Psi\hat{\theta}_n^* - c)$$

and $\chi_{r,1-\alpha}^2$ is the $1-\alpha$ quantile of a χ^2 random variable with r degrees of freedom. In order to study the properties of this test, we first derive the asymptotic behavior of $\hat{\theta}_n^*$. As mentioned earlier, in order to do so, we impose instead of Assumption 2.2 the following assumption, which mildly strengthens it. We emphasize again that this stronger assumption parallels the assumption made in Bugni et al. (2018) for the analysis of linear regression with "strata fixed effects" in the case of a single treatment and is also satisfied by a wide variety of treatment assignment mechanisms, including Examples 2.1 and 2.2.

Assumption 4.1. The treatment assignment mechanism is such that

(a) $W^{(n)} \perp\!\!\!\perp A^{(n)} | S^{(n)}$.

(b) $\pi_a(s) = \pi_a \in (0,1)$ for all $(a,s) \in \mathcal{A} \times \mathcal{S}$.

(c) $\left\{ \left(\sqrt{n} \left(\frac{n_a(s)}{n(s)} - \pi_a \right) : (a,s) \in \mathcal{A} \times \mathcal{S} \right) \middle| S^{(n)} \right\} \xrightarrow{d} N(0, \text{diag}(\Sigma_D(s)/p(s) : s \in \mathcal{S}))$ a.s. where for each $s \in \mathcal{S}$ and some $\tau(s) \in [0,1]$,

$$\Sigma_D(s) = \tau(s) \left[\text{diag}(\pi_a : a \in \mathcal{A}) - (\pi_a : a \in \mathcal{A})(\pi_a : a \in \mathcal{A})' \right] . \tag{21}$$

Assumption 4.1.(a) is the same as Assumption 2.2.(a) and requires that the treatment assignment mechanism is a function only of the vector of strata and an exogenous randomization device. Assumption 4.1.(b) requires the target proportion $\pi_a(s)$ to be constant across strata. This restriction is required for consistency of $\hat{\theta}_n^*$ for $\theta(Q)$. Finally, Assumption 4.1.(c) is stronger than Assumption 2.2.(b) and requires that the (possibly random) fraction of units assigned to treatment a and stratum s is asymptotically normal as the sample size tends to infinity. In the case of simple random sampling, where each unit is randomly assigned to each treatment with probability π_a, Assumption 4.1.(c) holds with $\tau(s) = 1$ for all $s \in \mathcal{S}$. In this sense, the assumption requires that the treatment assignment mechanism improves "balance" relative to simple random sampling. At the other extreme, we say that the treatment assignment mechanism achieves "strong balance" when $\tau(s) = 0$ for all $s \in \mathcal{S}$, which leads to $\Sigma_D(s)$ being a null matrix. It is straightforward to show that stratified block randomization satisfies Assumption 4.1.(c) with $\tau(s) = 0$, i.e., that it achieves "strong balance."

The following theorem derives the asymptotic behavior of $\hat{\theta}_n^*$:

Theorem 4.1. *Suppose Q satisfies Assumption 2.1 and the treatment assignment mechanism satisfies Assumption 4.1. Then,*

$$\sqrt{n}(\hat{\theta}_n^* - \theta(Q)) \xrightarrow{d} N(0, \mathbb{V}_{\text{sfe}}) ,$$

where $\mathbb{V}_{\text{sfe}} = \mathbb{V}_H + \mathbb{V}_{\tilde{Y}} + \mathbb{V}_A$, $\mathbb{V}_H$ is as in (11), $\mathbb{V}_{\tilde{Y}}$ is as in (12) with $\pi_a(s) = \pi_a$ for all $(a,s) \in \mathcal{A} \times \mathcal{S}$, and

$$\begin{aligned} \mathbb{V}_A \equiv \Bigg(\sum_{s \in \mathcal{S}} p(s) \Bigg(& \xi_a(s)\xi_{a'}(s) \frac{\Sigma_D(s)_{[a,a']}}{\pi_a \pi_{a'}} - \xi_a(s)\xi_0(s) \frac{\Sigma_D(s)_{[a,0]}}{\pi_a \pi_0} \\ & - \xi_{a'}(s)\xi_0(s) \frac{\Sigma_D(s)_{[a',0]}}{\pi_{a'} \pi_0} + \xi_0(s)\xi_0(s) \frac{\Sigma_D(s)_{[0,0]}}{\pi_0 \pi_0} \Bigg) : (a,a') \in \mathcal{A} \times \mathcal{A} \Bigg) \end{aligned} \tag{22}$$

and

$$\xi_a(s) \equiv E[m_a(Z_i)|S_i = s] - \sum_{a' \in \mathcal{A}_0} \pi_{a'} E[m_{a'}(Z_i)|S_i = s] . \tag{23}$$

Lemmas C.6 and C.7 in the Appendix derive the limit in probability of the usual homoskedasticity-only and heteroskedasticity-consistent estimators of the asymptotic variance of $\hat{\theta}_n^*$. As in the preceding section, these results show that neither of these estimators are consistent for the asymptotic variance of $\hat{\theta}_n^*$. In the special case with only one treatment (i.e., $|\mathcal{A}| = 1$), however, the heteroskedasticity-consistent estimator of the asymptotic variance leads to tests that are asymptotically conservative in the sense that they have limiting rejection probability under the null hypothesis no greater than the nominal level. See Bugni et al. (2018, Theorem 4.3) and Section 5 below for further discussion. In light of these results, the following theorem

constructs a consistent estimator of the asymptotic variance of $\hat{\theta}_n^*$. The theorem further establishes that tests using this new estimator of the asymptotic variance are exact in the sense that they have limiting rejection probability under the null hypotheses equal to the nominal level. Before proceeding, we note, however, that the theorem imposes the additional requirement that the randomization scheme achieves "strong balance," i.e,. that $\tau(s) = 0$ for all $s \in \mathcal{S}$. While it is possible to derive consistent estimators of the asymptotic variance of $\hat{\theta}_n^*$ even when this is not the case, it follows from Theorem D.1 in the Appendix that when each test is used with a consistent estimator for the appropriate asymptotic variance, $\phi_n^{\text{sfe}}(X^{(n)})$ is in general less powerful along a sequence of local alternatives than $\phi_n^{\text{sat}}(X^{(n)})$ except in the case of "strong balance." Indeed, it follows immediately from Theorems 3.1 and 4.1 that the asymptotic variance of $\hat{\theta}_n^*$ coincides with the asymptotic variance of $\hat{\theta}_n$ for randomization schemes that achieve "strong balance." For this reason, we view the case of randomization schemes that achieve "strong balance" as being the most relevant.

Theorem 4.2. *Suppose Q satisfies Assumption 2.1 and the treatment assignment mechanism satisfies Assumption 4.1 with $\tau(s) = 0$ for all $s \in \mathcal{S}$. Let $\hat{\mathbb{V}}_{\text{hc}}$ be the heteroskedasticity-consistent estimator of the asymptotic variance defined in (B-36) and let $\hat{\mathbb{V}}_H$ be defined as in (15). Then,*

$$\hat{\mathbb{V}}_{\text{sfe}} = \hat{\mathbb{V}}_H + \hat{\mathbb{V}}_{\text{hc}} \xrightarrow{P} \mathbb{V}_{\text{sfe}} = \mathbb{V}_H + \mathbb{V}_{\tilde{Y}} \ . \tag{24}$$

In addition, for the problem of testing (4) at level $\alpha \in (0,1)$, $\phi_n^{\text{sfe}}(X^{(n)})$ defined in (20) with $\hat{\mathbb{V}}_n = \hat{\mathbb{V}}_{\text{sfe}}$ satisfies

$$\lim_{n\to\infty} E[\phi_n^{\text{sfe}}(X^{(n)})] = \alpha \tag{25}$$

for Q additionally satisfying the null hypothesis, i.e., $\Psi\theta(Q) = c$.

5 The Case of a Single Treatment

In this section we consider the special case where $|\mathcal{A}| = 1$ to better illustrate the results we derived for the general case and to compare them to those in Imbens and Rubin (2015). When $|\mathcal{A}| = 1$, $\theta(Q)$ is a scalar parameter and the asymptotic variances in Theorems 3.1 and 4.1 become considerably simpler.

Consider first the the "fully saturated" linear regression. Applying Theorem 3.1 to the case $|\mathcal{A}| = 1$ shows that $\sqrt{n}(\hat{\theta}_n - \theta(Q))$ tends in distribution to a normal random variable with mean zero and variance equal to

$$\mathbb{V}_{\text{sat}} = \varsigma_H^2 + \varsigma_{\tilde{Y}}^2 \ ,$$

where

$$\varsigma_H^2 \equiv \sum_{s\in\mathcal{S}} p(s) \left(E[m_1(Z_i) - m_0(Z_i)|S_i = s]\right)^2 \tag{26}$$

$$\varsigma_{\tilde{Y}}^2 \equiv \sum_{s\in\mathcal{S}} p(s) \left(\frac{\sigma^2_{\tilde{Y}(0)}(s)}{\pi_0(s)} + \frac{\sigma^2_{\tilde{Y}(1)}(s)}{\pi_1(s)} \right) \ . \tag{27}$$

In addition, it follows from Theorem 3.2 and (14) that the usual heteroskedasticity-consistent estimator of

the asymptotic variance of $\hat{\theta}_n$ converges in probability to $\varsigma^2_{\tilde{Y}}$. As a result, tests based on $\hat{\theta}_n$ and this estimator for the asymptotic variance lead to over-rejection under the null hypothesis whenever $\varsigma^2_H > 0$.

Imbens and Rubin (2015, Ch. 9) study the properties of $\hat{\theta}_n$ when $|\mathcal{A}| = 1$ and the treatment assignment mechanism is stratified block randomization, which satisfies the hypotheses of Theorem 3.1. In contrast to our results, Imbens and Rubin (2015, Theorem 9.2, page 207) conclude that $\sqrt{n}(\hat{\theta}_n - \theta(Q))$ tends in distribution to a normal random variable with mean zero and variance equal to $\varsigma^2_{\tilde{Y}}$. In other words, the results in Imbens and Rubin (2015) coincide with our results when the model is sufficiently homogeneous in the sense that $\varsigma^2_H = 0$. This condition can be alternatively written as

$$E[Y_i(1) - Y_i(0)|S_i = s] = E[Y_i(1) - Y_i(0)] \quad \text{for all } s \in \mathcal{S} \ . \tag{28}$$

When this condition does not hold, however, our results differ from those in Imbens and Rubin (2015) and lead to tests that are asymptotically exact under arbitrary heterogeneity. In Section 6, we show further that tests based on $\hat{\theta}_n$ and a consistent estimator of $\varsigma^2_{\tilde{Y}}$ only may over-reject dramatically when ς^2_H is indeed positive.

Now consider the linear regression with "strata fixed effects." Applying Theorem 4.1 to the case $|\mathcal{A}| = 1$ shows that $\sqrt{n}(\hat{\theta}^*_n - \theta(Q))$ tends in distribution to a normal random variable with mean zero and variance equal to

$$\mathbb{V}_{\text{sfe}} = \varsigma^2_H + \varsigma^2_{\tilde{Y}} + \varsigma^2_A \ ,$$

where ς^2_H is as in (26), $\varsigma^2_{\tilde{Y}}$ is as in (27), and

$$\varsigma^2_A = \frac{(1-2\pi_1)^2}{\pi_1(1-\pi_1)} \sum_{s \in \mathcal{S}} \tau(s)p(s) \left(E[m_1(Z)|S = s] - E[m_0(Z)|S = s]\right)^2 \ . \tag{29}$$

For treatment assignment mechanisms that achieve "strong balance," we have in particular that $\mathbb{V}_{\text{sfe}} = \varsigma^2_H + \varsigma^2_{\tilde{Y}}$. Furthermore, applying Lemmas C.6 and C.7 in the Appendix to the case $|\mathcal{A}| = 1$ and $\tau(s) = 0$ shows that the usual homoskedasticity-only estimator of the asymptotic variance is generally inconsistent for $\mathbb{V}_{\text{sfe}}$, while the heteroskedasticity-consistent estimator of the variance, $\hat{\mathbb{V}}^*_{\text{hc}}$, satisfies

$$\hat{\mathbb{V}}^*_{\text{hc}} \xrightarrow{P} \left[\frac{1}{\pi_1(1-\pi_1)} - 3\right] \varsigma^2_H + \varsigma^2_{\tilde{Y}} \ , \tag{30}$$

which is strictly greater than $\mathbb{V}_{\text{sfe}}$, unless $\varsigma^2_H = 0$ or $\pi_1 = \frac{1}{2}$. In other words, when $|\mathcal{A}| = 1$ and $\tau(s) = 0$ for all $s \in \mathcal{S}$, tests of (4) based on $\hat{\theta}^*_n$ and the usual the heteroskedasticity-consistent estimator of the asymptotic variance $\hat{\mathbb{V}}^*_{\text{hc}}$ are asymptotically conservative unless $\varsigma^2_H = 0$ or $\pi_1 = \frac{1}{2}$. See Bugni et al. (2018, Theorem 4.3) for a formal statement of this result.

Imbens and Rubin (2015, Ch. 9) also study the properties of $\hat{\theta}^*_n$ when $|\mathcal{A}| = 1$ and the treatment assignment mechanism is stratified block randomization, which satisfies the hypotheses of Theorem 4.1. In particular, stratified block randomization satisfies Assumption 4.1 with $\tau(s) = 0$ for all $s \in \mathcal{S}$, so $\varsigma^2_A = 0$. In contrast to our results, Imbens and Rubin (2015, Theorem 9.1, page 206) conclude that $\sqrt{n}(\hat{\theta}^*_n - \theta(Q))$ tends in distribution to a normal random variable with mean zero and variance that can be expressed in our

notation as

$$\left[\frac{1}{\pi_1(1-\pi_1)} - 3\right]\varsigma_H^2 + \varsigma_Y^2 \ .$$

This asymptotic variance is strictly greater than $\mathbb{V}_{\text{sfe}}$ unless $\varsigma_H^2 = 0$ or $\pi_1 = \frac{1}{2}$, and it coincides with the limit in probability of the heteroskedasticity-consistent estimator of the asymptotic variance in (30). As in the case of the "fully saturated" linear regression, the results in Imbens and Rubin (2015) coincide with our results when the model is sufficiently homogeneous in the sense that condition (28) holds. When this condition does not hold, however, our results differ from those in Imbens and Rubin (2015) and lead to tests that are asymptotically exact under arbitrary heterogeneity. In Section 6, we again show that tests based on $\hat{\theta}_n^*$ and the usual heteroskedasticity-consistent estimator of the asymptotic variance may over-reject dramatically under the null hypothesis.

Remark 5.1. An inspection of the proofs of Theorems 3.1 and 4.1 reveals that the ς_H^2 term in the expressions for the variances of our limiting distributions of $\sqrt{n}(\hat{\theta}_n - \theta(Q))$ and $\sqrt{n}(\hat{\theta}_n^* - \theta(Q))$ stems from the contribution of a term involving $\left(\sqrt{n}\left(\frac{n(s)}{n} - p(s)\right) : s \in \mathcal{S}\right)$. It follows from this observation that it may be possible to reconcile the differences between our analysis and that in Imbens and Rubin (2015, Ch. 9) by considering an alternative sampling framework where $\frac{n(s)}{n}$ is constant with n. ■

6 Monte Carlo Simulations

In this section, we examine the finite-sample performance of several tests for the hypotheses in (4), including those introduced in Sections 3 and 4, with a simulation study. For $a \in \mathcal{A}$ and $1 \leq i \leq n$, potential outcomes are generated in the simulation study according to the equation:

$$Y_i(a) = \mu_a + (m_a(Z_i) - M_a) + \sigma_a(Z_i)\epsilon_{a,i} \ , \tag{31}$$

where μ_a, $m_a(Z_i)$, $\sigma_a(Z_i)$, M_a, and $\epsilon_{a,i}$ are defined below. In each specification, $n = 500$, $\{(Z_i, \epsilon_{0,i}, \epsilon_{1,i}) : 1 \leq i \leq n\}$ are i.i.d. with Z_i, $\epsilon_{0,i}$, and $\epsilon_{1,i}$ all being independent of each other, and $M_a = E[m_a(Z_i)]$. We focus on the case $|\mathcal{A}| = 1$ with $\pi_1(s) = \pi$ for all $s \in \mathcal{S}$ in order to be able to compare the tests studied in Sections 3 and 4; but also consider the case where $\pi_1(s) \neq \pi_1(s')$ for $s \neq s'$.

Model 1: $Z_i \sim \text{Beta}(2,2)$ (re-centered and re-scaled by the population mean and variance to have mean zero and variance one); $\sigma_0(Z_i) = \sigma_0 = 1$ and $\sigma_1(Z_i) = \sigma_1$; $\epsilon_{0,i} \sim N(0,1)$ and $\epsilon_{1,i} \sim N(0,1)$; $m_0(Z_i) = m_1(Z_i) = \gamma Z_i$. In this case,

$$Y_i = \mu_0 + (\mu_1 - \mu_0)A_i + \gamma Z_i + \eta_i \ ,$$

where

$$\eta_i = \sigma_1 A_i \epsilon_{1,i} + \sigma_0(1 - A_i)\epsilon_{0,i}$$

and $E[\eta_i | A_i, Z_i] = 0$.

Model 2: As in Model 1, but $m_0(Z_i) = -\gamma \log(Z_i + 3) I\{Z_i \leq \frac{1}{2}\}$.

Model 3: As in Model 2, but $\sigma_a(Z_i) = \sigma_a|Z_i|$.

Model 4: $Z_i \sim \text{Unif}(-2, 2)$; $\epsilon_{0,i} \sim \frac{1}{3}t_3$ and $\epsilon_{1,i} \sim \frac{1}{3}t_3$; $\sigma_a(Z_i) = \sigma_a|Z_i|$; and

$$m_0(Z_i) = \begin{cases} \gamma Z_i^2 & \text{if } Z_i \in [-1, 1] \\ \gamma Z_i & \text{otherwise} \end{cases} \quad \text{and} \quad m_1(Z_i) = \begin{cases} \gamma Z_i & \text{if } Z_i \in [-1, 1] \\ \gamma Z_i^2 & \text{otherwise} \end{cases} .$$

Treatment status is determined according to one of the following four different covariate-adaptive randomization schemes:

SRS: Treatment assignment is generated as in Example 2.1.

SBR: Treatment assignment is generated as in Example 2.2.

In each case, strata are determined by dividing the support of Z_i into $|\mathcal{S}|$ intervals of equal length and letting $S(Z_i)$ be the function that returns the interval in which Z_i lies. In all cases, observed outcomes Y_i are generated according to (1). Finally, for each of the above specifications, we consider different values of $(|\mathcal{S}|, \pi, \gamma, \sigma_1)$ and consider both $(\mu_0, \mu_1) = (0, 0)$ (i.e., under the null hypothesis that $\theta = \mu_1 - \mu_0 = 0$) and $(\mu_0, \mu_1) = (0, 0.2)$ (i.e., under the alternative hypothesis with $\theta = 0.2$).

The results of our simulations are presented in Tables 1–4 below. Rejection probabilities are computed using 10^4 replications. Columns are labeled in the following way:

SAT: The t-test from the "fully saturated" linear regression studied in Section 3. We report results for this test using the homoskedasticity-only ('HO'), heteroskedasticity-robust ('HC'), and the new ('NEW') consistent (as in Theorem 3.3), estimators of the asymptotic variance.

SFE: The t-test from the linear reression with "strata fixed effects" studied in Section 4. We report results for this test using the homoskedasticity-only ('HO'), heteroskedasticity-robust ('HC'), and the new ('NEW') consistent (as in Theorem 3.3), estimators of the asymptotic variance.

Table 1 displays the results of our baseline specification, where $(|\mathcal{S}|, \pi, \gamma, \sigma_1) = (10, 0.3, 1, 1)$. Table 2 displays the results for $(|\mathcal{S}|, \pi, \gamma, \sigma_1) = (10, 0.3, 2, 1)$, to explore sensitivity to changes in γ. Tables 3 and 4 replace $\pi = 0.3$ with $\pi = 0.7$, so $(|\mathcal{S}|, \pi, \gamma, \sigma_1) = (10, 0.7, 1, 1)$ and $(|\mathcal{S}|, \pi, \gamma, \sigma_1) = (10, 0.7, 2, 1)$. Finally, Table 5 considers the baseline specification but with $\pi_1(s) \neq \pi_1(s')$ for $s \neq s'$, i.e.,

$$(\pi_1(1), \ldots, \pi_1(|\mathcal{S}|)) = (0.20, 0.25, 0.30, 0.35, 0.40, 0.60, 0.65, 0.70, 0.75, 0.80) . \tag{32}$$

We organize our discussion of the results by test:

SAT: As expected in light of Theorems 3.1 and 3.2, the test $\phi_n^{\text{sat}}(X^{(n)})$ in (10) when $\hat{\mathbb{V}}_n$ is either the homoskedasticity-only or heteroskedasticity-consistent estimator of the asymptotic variance may over-reject under the null hypothesis. Indeed, in some cases (Model 4 in Table 2) the rejection probability

		Rejection rate under H_0: $\theta = 0$						Rejection rate under H_1: $\theta = 0.2$					
		SAT			SFE			SAT			SFE		
M	CAR	HO	HC	NEW	HO	HC	NEW	HO	HC	NEW	HO	HC	NEW
1	SRS	5.13	5.30	5.27	5.08	5.14	5.17	81.96	82.11	82.08	82.01	82.06	82.15
	SBR	4.74	4.98	4.92	4.71	4.88	4.93	82.25	82.44	82.32	82.21	82.17	82.31
2	SRS	6.65	6.84	4.93	6.31	5.05	5.08	80.18	80.77	75.71	75.91	72.58	72.66
	SBR	6.75	4.63	4.60	4.74	3.58	4.63	79.63	79.94	75.14	75.75	71.91	75.77
3	SRS	7.69	7.79	5.17	6.25	4.86	4.89	84.84	84.93	80.87	80.10	76.98	77.06
	SBR	7.19	4.59	4.52	4.53	3.34	4.59	85.11	85.16	80.58	81.14	77.75	81.08
4	SRS	20.04	19.22	5.06	10.80	5.12	5.13	92.44	91.93	79.17	76.45	65.00	65.11
	SBR	19.92	19.16	5.19	5.92	2.21	5.35	92.91	92.37	79.10	80.19	67.16	78.98

Table 1: Treatment assignment implemented via simple random sampling (SRS) and stratified block randomization (SBR). SAT and SFE tests implemented with homoskedastic-only (HO), heteroskedasticity-consistent (HC), and newly developed (NEW) standard errors. Parameter values: $(|\mathcal{S}|, \pi, \gamma, \sigma_1) = (10, 0.3, 1, 1)$.

		Rejection rate under H_0: $\theta = 0$						Rejection rate under H_1: $\theta = 0.2$					
		SAT			SFE			SAT			SFE		
M	CAR	HO	HC	NEW	HO	HC	NEW	HO	HC	NEW	HO	HC	NEW
1	SRS	8.57	5.06	5.07	8.41	4.85	4.87	66.73	58.45	58.55	67.22	58.37	58.47
	SBR	8.51	5.10	5.05	8.42	5.00	5.06	67.57	59.03	58.79	67.43	58.64	58.80
2	SRS	14.35	10.16	5.31	10.85	5.39	5.44	65.42	58.17	45.91	53.33	39.88	39.93
	SBR	14.58	9.80	5.06	7.50	3.15	5.10	65.87	58.93	46.96	54.53	39.72	47.68
3	SRS	14.73	10.45	5.25	10.23	5.09	5.10	69.79	63.22	49.71	56.39	43.53	43.64
	SBR	15.02	10.55	4.88	6.96	2.89	4.97	71.28	64.39	49.93	57.48	41.88	51.10
4	SRS	31.22	26.06	5.28	12.35	5.39	5.41	73.57	69.41	36.25	42.20	26.50	26.56
	SBR	32.00	26.69	5.00	6.56	1.82	5.09	74.30	69.97	36.60	40.38	21.48	36.56

Table 2: Treatment assignment implemented via simple random sampling (SRS) and stratified block randomization (SBR). SAT and SFE tests implemented with homoskedastic-only (HO), heteroskedasticity-consistent (HC), and newly developed (NEW) standard errors. Parameter values: $(|\mathcal{S}|, \pi, \gamma, \sigma_1) = (10, 0.3, 2, \sqrt{2})$.

under the null hypothesis could be as high as 32% for the homoskedasticity-only case and 30% for the heteroskedasticity-consistent case. This over-rejection happens both, under simple random sampling and stratified block randomization. Finally, and consistent with the results in Section 5, whenever Q is such that $\mathbb{V}_H = 0$, as it is the case in Model 1, the test with the heteroskedasticity-consistent estimator of the asymptotic variance is asymptotically exact.

According to Theorem 3.3, the test $\phi_n^{\text{sat}}(X^{(n)})$ in (10) when $\hat{\mathbb{V}}_n$ is given by the new consistent estimator of the asymptotic variance in (16) is asymptotically exact across all the specifications we consider. Indeed, the rejection probability under the null hypothesis is very close to the nominal level in all models and all tables. The rejection probability under the alternative hypothesis is the highest under simple random sampling among the tests that are asymptotically exact and do not over-reject under the null hypothesis. Under stratified block randomization, and given that in this case $\tau(s) = 0$ for all $s \in \mathcal{S}$, the rejection probability under the alternative hypothesis is effectively the same as that of $\phi_n^{\text{sfe}}(X^{(n)})$ with the new consistent estimator of the asymptotic variance in (24). These results are in line with the theoretical results described in Section 4. Finally, Table 5 illustrates that the results for $\phi_n^{\text{sat}}(X^{(n)})$ with the new consistent estimator of the asymptotic variance are not affected by whether $\pi_1(s)$ is the same across strata $s \in \mathcal{S}$ or not.

SFE: As expected from Theorem 4.1 and the subsequent discussion, the test $\phi_n^{\text{sfe}}(X^{(n)})$ in (20) when

		Rejection rate under H_0: $\theta = 0$						Rejection rate under H_1: $\theta = 0.2$					
		SAT			SFE			SAT			SFE		
M	CAR	HO	HC	NEW	HO	HC	NEW	HO	HC	NEW	HO	HC	NEW
1	SRS	5.08	5.29	5.23	4.96	5.01	5.02	81.75	82.12	82.00	81.99	81.97	82.01
	SBR	5.02	5.10	5.06	4.95	4.95	5.00	82.76	82.93	82.79	82.65	82.73	82.82
2	SRS	6.72	6.94	4.83	6.26	5.01	5.03	79.85	80.08	75.32	74.87	71.56	71.63
	SBR	7.05	7.11	5.08	4.99	3.93	5.05	80.46	80.54	76.61	75.77	72.26	76.04
3	SRS	7.23	7.58	5.03	6.44	5.03	5.05	85.81	85.82	81.28	80.35	77.09	77.12
	SBR	7.56	7.70	5.14	5.07	3.92	5.16	87.56	87.62	83.07	82.40	78.71	82.75
4	SRS	18.46	19.91	5.43	10.02	5.20	5.21	92.45	93.12	80.79	76.88	66.84	66.95
	SBR	18.25	19.63	5.93	5.21	2.09	5.83	92.98	93.33	82.57	81.27	71.75	82.77

Table 3: Treatment assignment implemented via simple random sampling (SRS) and stratified block randomization (SBR). SAT and SFE tests implemented with homoskedastic-only (HO), heteroskedasticity-consistent (HC), and newly developed (NEW) standard errors. Parameter values: $(|\mathcal{S}|, \pi, \gamma, \sigma_1) = (10, 0.7, 1, 1)$.

		Rejection rate under H_0: $\theta = 0$						Rejection rate under H_1: $\theta = 0.2$					
		SAT			SFE			SAT			SFE		
M	CAR	HO	HC	NEW	HO	HC	NEW	HO	HC	NEW	HO	HC	NEW
1	SRS	2.72	5.55	5.45	2.79	5.35	5.38	58.45	68.64	68.35	59.02	68.51	68.62
	SBR	2.66	5.23	5.17	2.64	5.13	5.14	58.79	68.91	68.79	58.79	68.74	68.80
2	SRS	7.18	11.48	5.28	6.22	5.44	5.47	58.35	66.71	51.98	47.35	45.08	45.21
	SBR	7.18	11.19	4.99	3.19	2.80	5.02	58.95	66.52	53.69	45.17	43.14	52.74
3	SRS	8.00	12.36	5.13	6.43	5.24	5.29	64.51	71.87	56.25	51.30	47.55	47.61
	SBR	7.63	11.88	4.99	3.35	2.83	5.00	65.91	73.20	58.83	50.41	47.03	57.71
4	SRS	24.98	30.67	5.12	10.82	5.61	5.62	69.65	74.39	39.07	39.87	27.80	27.86
	SBR	24.81	30.72	6.01	4.49	1.50	5.81	70.74	75.42	41.60	37.57	24.20	41.41

Table 4: Treatment assignment implemented via simple random sampling (SRS) and stratified block randomization (SBR). SAT and SFE tests implemented with homoskedastic-only (HO), heteroskedasticity-consistent (HC), and newly developed (NEW) standard errors. Parameter values: $(|\mathcal{S}|, \pi, \gamma, \sigma_1) = (10, 0.7, 2, \sqrt{2})$.

$\hat{\mathbb{V}}_n$ is the homoskedasticity-only estimator of the asymptotic variance could lead to over-rejection or under-rejection, depending on the specification. For example, the rejection probability under the null hypothesis in Table 2 could be as high as 12.25%, while in Table 4 could be as low as 2.64%. On the other hand, when $\hat{\mathbb{V}}_n$ is the heteroskedasticity-consistent estimator of the asymptotic variance, this test is asymptotically conservative; in line with the results in Bugni et al. (2018) and Section 5. Indeed, the rejection probability under the null hypothesis is close to 2% in Model 4 under stratified block randomization for all the specifications we consider. Finally, and consistent with the results in Section 5, whenever Q is such that $\mathbb{V}_H = 0$, as it is the case in Model 1, the test with the heteroskedasticity-consistent estimator of the asymptotic variance is asymptotically exact.

According with Theorem 4.2, the test $\phi_n^{\text{sfe}}(X^{(n)})$ in (20) when $\hat{\mathbb{V}}_n$ is given by the new consistent estimator of the asymptotic variance in (24) is asymptotically exact across all the specifications we consider. The rejection probability under the null hypothesis is very close to the nominal level in all models and all tables. The rejection probability under the alternative hypothesis is similar to that of $\phi_n^{\text{sat}}(X^{(n)})$ with $\hat{\mathbb{V}}_n = \hat{\mathbb{V}}_{\text{sat}}$ under stratified block randomization, but often below the rejection probability of that same test under simple random sampling. These results are again in line with the theoretical results discuss in Section 4. Finally, Table 5 illustrates that $\phi_n^{\text{sfe}}(X^{(n)})$ is only a valid test for the null in (4) when $\pi_1(s) = \pi$ for all $s \in \mathcal{S}$ and may otherwise over-reject under the null hypothesis.

		Rejection rate under H_0: $\theta = 0$						Rejection rate under H_1: $\theta = 0.2$					
		SAT			SFE			SAT			SFE		
M	CAR	HO	HC	NEW	HO	HC	NEW	HO	HC	NEW	HO	HC	NEW
1	SRS	5.20	5.47	5.47	5.08	5.12	5.15	81.63	82.48	82.48	82.80	82.71	82.75
	SBR	5.27	5.39	5.39	5.32	5.42	5.44	83.15	83.48	83.48	83.49	83.43	83.58
2	SRS	6.74	7.18	5.70	9.05	7.13	9.51	79.53	80.14	76.98	87.24	84.66	87.61
	SBR	7.18	7.33	5.63	8.92	7.05	9.08	80.57	80.91	77.23	90.72	88.61	90.91
3	SRS	8.89	8.14	6.34	9.49	8.18	8.99	85.19	84.10	81.04	92.03	90.57	91.54
	SBR	8.24	7.56	5.53	9.03	7.53	8.37	86.51	85.38	81.77	94.92	93.76	94.42
4	SRS	19.74	18.16	6.41	60.82	45.51	59.43	91.77	90.90	80.14	12.92	5.62	12.42
	SBR	19.71	18.14	6.69	67.13	48.22	66.08	91.61	90.77	80.78	4.42	1.12	4.00

Table 5: Treatment assignment implemented via simple random sampling (SRS) and stratified block randomization (SBR). SAT and SFE tests implemented with homoskedastic-only (HO), heteroskedasticity-consistent (HC), and newly developed (NEW) standard errors. Parameter values: $(|\mathcal{S}|, \pi, \gamma, \sigma_1) = (10, \pi_1(s), 1, 1)$ with $\pi_1(s)$ as in (32).

7 Implications for Empirical Practice

When the target proportion of units being assigned to each treatment varies across strata, we recommend using the test ϕ_n^{sat} based on ordinary least squares estimation of the "fully saturated" linear regression and the consistent estimator of the asymptotic variance that we derive in Theorem 3.3. Importantly, tests based on these estimators with the usual heteroskedasticity-consistent estimator of the asymptotic variance may be invalid in the sense that they may have limiting rejection probability under the null hypothesis strictly greater than the nominal level. When the target proportion of units being assigned to each treatment does not vary across strata, one may additionally consider use of the test ϕ_n^{sfe} based on ordinary least squares estimation of the linear regression with "strata fixed effects" and the consistent estimator of the asymptotic variance that we derive in Theorem 4.2. Our theoretical results results reveal that for a given function mapping Z_i into strata fixed, the power of ϕ_n^{sfe} is highest when using a randomization schemes that satisfies Assumption 4.1.(c) with $\tau(s) = 0$ for all $s \in \mathcal{S}$, such as stratified block randomization. On the other hand, ϕ_n^{sat} is in general weakly preferred to ϕ_n^{sfe} and may be strictly preferred for randomization schemes that satisfy Assumption 4.1.(c) with $\tau(s) > 0$ for some $s \in \mathcal{S}$. For simplicity, it may therefore be preferable to use ϕ_n^{sat}.

In this paper, we do not consider further questions about "optimal" treatment assignment, but, in conclusion, we mention two recent papers on this topic. Building upon our results, Tabord-Meehan (2018) considers optimization of the power of ϕ_n^{sat} over different functions mapping Z_i into strata using stratification trees. Bai (2018), on the other hand, considers minimization of the mean squared error of the difference-in-means estimator of the average treatment effect over a general class of randomization mechanisms that, importantly, includes mechanisms with a "large" number of strata.

8 Empirical Illustration

We conclude our paper with an empirical illustration using data from Chong et al. (2016), who study the effect of iron deficiency anemia (i.e., anemia caused by a lack of iron) on school-age children's educational

attainment and cognitive ability in Peru. The data used in this experiment are publicly available in the AEA website at https://www.aeaweb.org/articles?id=10.1257/app.20140494.

8.1 Empirical Setting

We now briefly summarize the empirical setting; see Chong et al. (2016) for a more detailed description. According to the medical literature, iron deficiency anemia may impair cognitive function, memory, and attention span. In this way, iron deficiency anemia may significantly increase the cost of human capital accumulation for school-age children and lead to nutrition-based poverty traps. Chong et al. (2016) investigate whether showing students promotional videos can incentivize them to increase their iron intake and thus improve their academic performance.

The units in this experiment are 219 students in a rural secondary school in the impoverished Cajamarca district of Peru between October and December in 2009. During this period, these students were exposed to short instructional videos when logging into their personal computers at school. Each student was randomly assigned to one of three types of videos: two treatments and a control. The first treatment video featured a popular soccer player encouraging the students to consume iron supplements to maximize their energy. The second treatment video featured a doctor encouraging them to consume iron supplements for their overall health. Finally, the control video featured a dentist who encouraged oral hygiene without mentioning iron in any way. Throughout this experiment, researchers additionally stocked the local clinic with iron supplements, which were provided for free to any student who requested them.

Students were assigned to one of the three types of videos using stratified block randomization, where stratification occurred by grade, taking values $s \in \mathcal{S} = \{1, 2, 3, 4, 5\}$. As explained in footnote 17 of Chong et al. (2016), within each grade, the researchers assigned one third of the students to each video type, i.e., $\pi_a(s) = 1/3$ for all $a \in \mathcal{A}_0 = \{0, 1, 2\}$ and $s \in \mathcal{S}$. Table 6 describes the sample sizes for each combination of stratum and treatment. Note that the sample consists of 215 students rather than 219 students because four students were excluded from the study for various reasons; see, for example, footnote 24 in Chong et al. (2016), which explains that two students failed to turn in a required consent form. We conjecture that these exclusions explain the discrepancies between the observed treatment proportions and $\pi_a(s)$ observed in Table 6. Note further that since in this case $\pi_a(s)$ does not depend on s, our results imply that we could analyze the experiment using either the "fully saturated" linear regression described in Section 3 or the linear regression with "strata fixed effects" described in Section 4. Below we focus on the former, but note that the latter provides similar results.

	$s = 1$	$s = 2$	$s = 3$	$s = 4$	$s = 5$	total
$a = 0$ (placebo video)	15	19	16	12	10	72
$a = 1$ (soccer video)	16	19	15	10	10	70
$a = 2$ (doctor video)	17	20	15	11	10	73
total	48	58	46	33	30	215

Table 6: Sample sizes for each combination of stratum and treatment.

8.2 Results

Chong et al. (2016) examine the effect of the treatment videos relative to the control video on a variety of cognitive ability and educational attainment outcomes. We focus on academic achievement, as measured by a student's average grade during the last two quarters of the 2009 academic year in five subjects: math, foreign language, social science, science, and communications. As explained by the authors, this constitutes one of the primary outcomes of interest in Chong et al. (2016).

We present our results in Table 7, which was computed using our `car_sat` Stata package available at `https://bitbucket.org/iacanay/car-stata`. In both the top and bottom half of Table 7, the first column reports point estimates of $\theta_a(Q)$ for the two treatment videos $a \in \mathcal{A} = \{1, 2\}$ that we obtained from the "fully saturated" linear regression, i.e.,

$$\hat{\theta}_{n,a} = \sum_{s=1}^{5} \frac{n(s)}{n} \hat{\beta}_{n,a}(s) \ ,$$

where $\hat{\beta}_{n,a}(s)$ is the ordinary least squares estimator of $\beta_a(s)$ in the equation

$$Y_i = \sum_{s=1}^{5} \delta(s) I\{S_i = s\} + \sum_{a=1}^{2} \sum_{s=1}^{5} \beta_a(s) I\{A_i = a, S_i = s\} + u_i \ .$$

The remaining columns report standard errors, the resulting t-statistic, a p-value for a two-sided test of the null hypothesis that $\theta_a(Q) = 0$; and a 95% confidence interval for $\theta_a(Q)$. The top and bottom half of Table 7 differ only through the standard errors. The top half reports results with the "new" standard errors computed using our estimator of the asymptotic variance, $\hat{\mathbb{V}}_{\text{sat}}$ defined in (16). For our subsequent discussion, it is useful to recall that

$$\hat{\mathbb{V}}_{\text{sat}} = \hat{\mathbb{V}}_H + \hat{\mathbb{V}}_{\text{hc}} \ ,$$

where, in the context of our applicaton,

$$\hat{\mathbb{V}}_H = \sum_{s=1}^{5} \frac{n(s)}{n} \begin{pmatrix} \hat{\beta}_{n,1}(s) - \hat{\theta}_{n,1} \\ \hat{\beta}_{n,2}(s) - \hat{\theta}_{n,2} \end{pmatrix} \begin{pmatrix} \hat{\beta}_{n,1}(s) - \hat{\theta}_{n,1} \\ \hat{\beta}_{n,2}(s) - \hat{\theta}_{n,2} \end{pmatrix}' \tag{33}$$

and $\hat{\mathbb{V}}_{\text{hc}}$ is the usual heteroskedasticity-consistent estimator of the asymptotic variance defined in (B-36). The bottom half of Table 7 reports results with the standard errors computed using $\hat{\mathbb{V}}_{\text{hc}}$.

Since $\hat{\mathbb{V}}_{\text{sat}} = \hat{\mathbb{V}}_H + \hat{\mathbb{V}}_{\text{hc}}$ and $\hat{\mathbb{V}}_H$ is positive semidefinite, the "new" standard errors are larger than the usual heteroskedasticity-consistent standard errors. The differences, however, in this instance are small and do not lead to any meaningful differences in terms of the conclusions we draw from the experiment when testing either the null hypothesis that $\theta_1(Q) = 0$ or the null hypothesis that $\theta_2(Q) = 0$ at the conventional 5% significance level. To gain further insight into the magnitude of these differences, it is instructive to

	SAT regression: "new" standard errors					
	Coef.	s.e.	t-stat	p-value	[95% Conf.	Int.]
$\hat{\theta}_{n,1}$ (soccer video)	-0.051	0.206	-0.248	0.805	-0.458	0.356
$\hat{\theta}_{n,2}$ (doctor video)	0.409	0.206	1.981	0.049	-0.002	0.816
	SAT regression: hc standard errors					
	Coef.	s.e.	t-stat	p-value	[95% Conf.	Int.]
$\hat{\theta}_{n,1}$ (soccer video)	-0.051	0.206	-0.248	0.804	-0.457	0.354
$\hat{\theta}_{n,2}$ (doctor video)	0.409	0.203	2.013	0.046	-0.008	0.810

Table 7: Inference about the average effect of treatments $a \in \mathcal{A} = \{1, 2\}$ (relative to the control) on academic achievement. "New" standard errors correspond to the ones we derive in this paper, while hc standard errors are the default "robust" standard errors in Stata.

examine $\hat{\mathbb{V}}_H$ and $\hat{\mathbb{V}}_{\text{hc}}$ in more detail, which are displayed below:

$$\hat{\mathbb{V}}_H = \begin{pmatrix} 0.0630 & 0.0385 \\ 0.0385 & 0.291 \end{pmatrix}, \quad \hat{\mathbb{V}}_{\text{hc}} = \begin{pmatrix} 9.101 & 4.503 \\ 4.503 & 8.879 \end{pmatrix}.$$

We see that $\hat{\mathbb{V}}_H$ is close to zero and at least an order of magnitude smaller than $\hat{\mathbb{V}}_{\text{hc}}$. By inspecting (33), we see that $\hat{\mathbb{V}}_H$ being close to zero implies that $\hat{\beta}_{n,1}(s)$ and $\hat{\beta}_{n,2}(s)$ are nearly constant across the five strata, which in turn suggests that stratification is nearly irrelevant in this particular application in the sense that $E[Y_i(a) - Y_i(0)|S_i]$ nearly equals $E[Y_i(a) - Y_i(0)]$ for each $a \in \{1, 2\}$.

Appendix A Additional Notation

Throughout the Appendix we employ the following notation, not necessarily introduced in the text.

$\sigma^2_X(s)$	For a random variable X, $\sigma^2_X(s) = \text{Var}[X\|S = s]$
σ^2_X	For a random variable X, $\sigma^2_X = \text{Var}[X]$
μ_a	For $a \in \mathcal{A}_0$, $E[Y_i(a)]$
$\tilde{Y}_i(a)$	For $a \in \mathcal{A}_0$, $Y_i(a) - E[Y_i(a)\|S_i]$
$m_a(Z_i)$	For $a \in \mathcal{A}_0$, $E[Y_i(a)\|Z_i] - \mu_a$
$n(s)$	Number of individuals in strata $s \in \mathcal{S}$
$n_a(s)$	Number of individuals in treatment $a \in \mathcal{A}_0$ in strata $s \in \mathcal{S}$
$\iota_{\|\mathcal{A}\|}$	$\|\mathcal{A}\|$-dimensional column vector of ones
$\mathbb{O}$	$(\|\mathcal{A}\| \times \|S\|)$-dimensional matrix of zeros
$\mathbb{I}_{\|\mathcal{A}\|}$	$\|\mathcal{A}\|$-dimensional identity matrix
$\mathbb{J}_s$	$(\|\mathcal{S}\| \times \|\mathcal{S}\|)$-dimensional matrix with a 1 on the (s, s)th coordinate and zeros otherwise

Table 8: Useful notation

In addition, we often transform objects that are indexed by $(a, s) \in \mathcal{A} \times \mathcal{S}$ into vectors or matrices, using the

following conventions. For $X(a)$ being a scalar object indexed over $a \in \mathcal{A}$, we use $(X(a) : a \in \mathcal{A})$ to denote the $|\mathcal{A}|$-dimensional vector $(X(1), \ldots, X(|\mathcal{A}|))'$. For $X_a(s)$ being a scalar object indexed by $(a, s) \in \mathcal{A} \times \mathcal{S}$ we use $(X_a(s) : (a, s) \in \mathcal{A} \times \mathcal{S})$ to denote the $(|\mathcal{A}| \times |\mathcal{S}|)$-dimensional column vector where the order of the indices is as follows,

$$(X_a(s) : (a, s) \in \mathcal{A} \times \mathcal{S}) = (X_1(1), \ldots, X_{|\mathcal{A}|}(1), X_1(2), \ldots, X_{|\mathcal{A}|}(2), \ldots)' .$$

Finally throughout the appendix we use $L_{n,a}^{(j)}(s)$ and $\mathbb{L}_n^{(j)}$ for $j = 1, 2, \ldots$, to denote scalar objects and matrices/vectors that may be redefined from theorem to theorem.

Appendix B Proof of Main Theorems

B.1 Proof of Theorem 3.1

Let $\mathbb{C}_n$ be the matrix of covariate associated with the regression in (7), i.e., the matrix with ith row given by

$$C_i = [(I\{S_i = s\} : s \in \mathcal{S})', (I\{A_i = a, S_i = s\} : (a, s) \in \mathcal{A} \times \mathcal{S})'] .$$

Let $\mathbb{R}_n$ be a matrix with $|\mathcal{A}|$ rows and $(|\mathcal{S}| + |\mathcal{A}| \times |\mathcal{S}|)$ columns defined as

$$\mathbb{R}_n = \left[\mathbb{O}, \frac{n(1)}{n}\mathbb{I}_{|\mathcal{A}|}, \ldots, \frac{n(|\mathcal{S}|)}{n}\mathbb{I}_{|\mathcal{A}|}\right] , \tag{B-34}$$

where $\mathbb{O}$ and $\mathbb{I}_{|\mathcal{A}|}$ are defined in Table 8. Using this notation, we can write

$$\hat{\theta}_n = \mathbb{R}_n \begin{bmatrix} (\hat{\delta}_n(s) : s \in \mathcal{S}) \\ (\hat{\beta}_{n,a}(s) : (a, s) \in \mathcal{A} \times \mathcal{S}) \end{bmatrix}$$

where $\hat{\delta}_n(s)$ and $\hat{\beta}_{n,a}(s)$ are the resulting estimators of $\delta(s)$ and $\beta_a(s)$ in (7), respectively. Now consider the following derivation,

$$\begin{aligned}
\sqrt{n}(\hat{\theta}_n - \theta(Q)) &= \sqrt{n}\left(\mathbb{R}_n \left(\frac{1}{n}\mathbb{C}_n'\mathbb{C}_n\right)^{-1} \frac{1}{n}\mathbb{C}_n'\mathbb{Y}_n - \theta(Q)\right) \\
&= \left(\sum_{s\in\mathcal{S}} \frac{n(s)}{n_a(s)}\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n} I\{A_i = a, S_i = s\}\tilde{Y}_i(a)\right] - \sum_{s\in\mathcal{S}} \frac{n(s)}{n_0(s)}\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n} I\{A_i = 0, S_i = s\}\tilde{Y}_i(0)\right]\right. \\
&\quad \left. + \sum_{s\in\mathcal{S}} \sqrt{n}\left(\frac{n(s)}{n} - p(s)\right) E\left[m_a(Z) - m_0(Z)|S = s\right] : a \in \mathcal{A}\right) \\
&= \left(\sum_{s\in\mathcal{S}} \left(L_{n,a}^{(1)}(s) - L_{n,0}^{(1)}(s)\right) : a \in \mathcal{A}\right) + \left(\sum_{s\in\mathcal{S}} L_{n,a}^{(2)}(s) : a \in \mathcal{A}\right) + o_P(1)
\end{aligned}$$

where for $(a, s) \in \mathcal{A} \times \mathcal{S}$,

$$\begin{aligned}
L_{n,a}^{(1)}(s) &\equiv \frac{1}{\pi_a(s)}\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n} I\{A_i = a, S_i = s\}\tilde{Y}_i(a)\right] \\
L_{n,a}^{(2)}(s) &\equiv \sqrt{n}\left(\frac{n(s)}{n} - p(s)\right) E\left[m_a(Z) - m_0(Z)|S = s\right] .
\end{aligned}$$

By Lemma C.1 and some additional calculations, it follows that

$$\begin{pmatrix} \left(\sum_{s\in\mathcal{S}}\left(L^{(1)}_{n,a}(s)-L^{(1)}_{n,0}(s)\right) : a\in\mathcal{A}\right) \\ \left(\sum_{s\in\mathcal{S}} L^{(2)}_{n,a}(s) : a\in\mathcal{A}\right) \end{pmatrix} \xrightarrow{d} N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \mathbb{V}_{\tilde{Y}} & 0 \\ 0 & \mathbb{V}_H \end{pmatrix}\right) ,$$

where $\mathbb{V}_{\tilde{Y}}$ is as in (12) and $\mathbb{V}_H$ is as in (11). Importantly, to get $\mathbb{V}_H$ for the second term we used that $\sum_{s\in\mathcal{S}} p(s)E\left[m_a(Z)-m_0(Z)|S=s\right]=0$ for all $a\in\mathcal{A}$.

B.2 Proof of Theorem 3.2

The homoskedasticity-only estimator of the asymptotic variance for the regression in (7) is

$$\hat{\mathbb{V}}_{\text{ho}} = \left(\frac{1}{n}\sum_{i=1}^{n}\hat{u}_i^2\right)\mathbb{R}_n\left(\frac{1}{n}\mathbb{C}_n'\mathbb{C}_n\right)^{-1}\mathbb{R}_n' , \tag{B-35}$$

where $\{\hat{u}_i : 1\leq i\leq n\}$ are the least squares residuals. The result then follows immediately from

$$\frac{1}{n}\sum_{i=1}^{n}\hat{u}_i^2 \xrightarrow{P} \sum_{(a,s)\in\mathcal{A}_0\times\mathcal{S}} p(s)\pi_a(s)\sigma^2_{\tilde{Y}(a)}(s) ,$$

which follows from Lemma C.5, and

$$\mathbb{R}_n\left(\frac{1}{n}\mathbb{C}_n'\mathbb{C}_n\right)^{-1}\mathbb{R}_n' \xrightarrow{P} \left[\sum_{s\in\mathcal{S}}\frac{p(s)}{\pi_0(s)}\iota_{|\mathcal{A}|}\iota_{|\mathcal{A}|}' + \text{diag}\left(\sum_{s\in\mathcal{S}}\frac{p(s)}{\pi_a(s)} : a\in\mathcal{A}\right)\right]$$

which follows from Lemma C.3, (B-34), and some additional calculations.

The heteroskedasticity-consistent estimator of the asymptotic variance for the regression in (7) is

$$\hat{\mathbb{V}}_{\text{hc}} = \mathbb{R}_n\left[\left(\frac{1}{n}\mathbb{C}_n'\mathbb{C}_n\right)^{-1}\left(\frac{1}{n}\mathbb{C}_n'\,\text{diag}\left(\hat{u}_i^2 : 1\leq i\leq n\right)\mathbb{C}_n\right)\left(\frac{1}{n}\mathbb{C}_n'\mathbb{C}_n\right)^{-1}\right]\mathbb{R}_n' . \tag{B-36}$$

First note that $\frac{1}{n}\mathbb{C}_n'\,\text{diag}\left(\hat{u}_i^2 : 1\leq i\leq n\right)\mathbb{C}_n$ equals

$$\begin{bmatrix} \text{diag}(\frac{1}{n}\sum_{i=1}^n \hat{u}_i^2 I\{S_i=s\} : s\in\mathcal{S}) & \sum_{s\in\mathcal{S}}\mathbb{J}_s\otimes(\frac{1}{n}\sum_{i=1}^n \hat{u}_i^2 I\{A_i=a,S_i=s\} : a\in\mathcal{A})' \\ \sum_{s\in\mathcal{S}}\mathbb{J}_s\otimes(\frac{1}{n}\sum_{i=1}^n \hat{u}_i^2 I\{A_i=a,S_i=s\} : a\in\mathcal{A}) & \text{diag}(\frac{1}{n}\sum_{i=1}^n \hat{u}_i^2 I\{A_i=a,S_i=s\} : (a,s)\in\mathcal{A}\times\mathcal{S}) \end{bmatrix} ,$$

which follows from Lemma C.3. By Lemma C.4, this matrix converges in probability to

$$\begin{bmatrix} \text{diag}(\sum_{a\in\mathcal{A}_0} p(s)\pi_a(s)\sigma^2_{\tilde{Y}(a)}(s) : s\in\mathcal{S}) & \sum_{s\in\mathcal{S}}\mathbb{J}_s\otimes(p(s)\pi_a(s)\sigma^2_{\tilde{Y}(a)}(s) : a\in\mathcal{A})' \\ \sum_{s\in\mathcal{S}}\mathbb{J}_s\otimes(p(s)\pi_a(s)\sigma^2_{\tilde{Y}(a)}(s) : a\in\mathcal{A}) & \text{diag}(p(s)\pi_a(s)\sigma^2_{\tilde{Y}(a)}(s) : (a,s)\in\mathcal{A}\times\mathcal{S}) \end{bmatrix} .$$

The result follows by combining this with Lemma C.3 and doing some additional calculations.

B.3 Proof of Theorem 3.3

By Theorem 3.2, it follows that

$$\hat{\mathbb{V}}_{\text{hc}} \xrightarrow{P} \sum_{s\in\mathcal{S}}\frac{p(s)\sigma^2_{\tilde{Y}(0)}(s)}{\pi_0(s)}\iota_{|\mathcal{A}|}\iota_{|\mathcal{A}|}' + \text{diag}\left(\sum_{s\in\mathcal{S}}\frac{p(s)\sigma^2_{\tilde{Y}(a)}(s)}{\pi_a(s)} : a\in\mathcal{A}\right) .$$

By Lemma C.3 and for any $a \in \mathcal{A}$,

$$\left(\hat{\beta}_{n,a}(s) - \hat{\theta}_{n,a}\right) \xrightarrow{P} E\left[m_a(Z) - m_0(Z)|S = s\right],$$

which in turn implies that

$$\begin{aligned}\hat{\mathbb{V}}_{\mathrm{H}} &= \sum_{s\in\mathcal{S}} \frac{n(s)}{n} \left(\hat{\beta}_{n,a}(s) - \hat{\theta}_{n,a} : a \in \mathcal{A}\right)\left(\hat{\beta}_{n,a}(s) - \hat{\theta}_{n,a} : a \in \mathcal{A}\right)' \\ &\xrightarrow{P} \sum_{s\in\mathcal{S}} p(s) \left(E[m_a(Z) - m_0(Z)|S = s] : a \in \mathcal{A}\right)\left(E[m_a(Z) - m_0(Z)|S = s] : a \in \mathcal{A}\right)' ,\end{aligned}$$

where we used $\frac{n(s)}{n} \xrightarrow{P} p(s)$. By the continuous mapping theorem, we conclude that $\hat{\mathbb{V}}_{\mathrm{sat}} \xrightarrow{P} \mathbb{V}_{\mathrm{sat}}$. By Theorem 3.1, $\lim_{n\to\infty} E[\phi_n^{\mathrm{sat}}(X^{(n)})] = \alpha$ follows immediately whenever Q is such that $\Psi\theta(Q) = c$.

B.4 Proof of Theorem 4.1

Let $\mathbb{M}_n \equiv \mathbb{I}_n - \mathbb{S}_n(\mathbb{S}_n'\mathbb{S}_n)^{-1}\mathbb{S}_n'$ denote the projection on the orthogonal complement of the column space of $\mathbb{S}_n$, where $\mathbb{S}_n$ is the matrix with ith row given by $(I\{S_i = s\} : s \in \mathcal{S})'$. By the Frisch-Waugh-Lovell Theorem,

$$\hat{\theta}_n^* = (\mathbb{A}_n'\mathbb{M}_n'\mathbb{M}_n\mathbb{A}_n)^{-1}(\mathbb{A}_n'\mathbb{M}_n'\mathbb{Y}_n) ,$$

where $\mathbb{Y}_n = (Y_i : 1 \le i \le n)$ and $\mathbb{A}_n$ is the matrix with ith row given by $(I\{A_i = a\} : a \in \mathcal{A})'$. Next, notice that

$$\mathbb{M}_n\mathbb{A}_n = \left(\left(I\{A_i = a\} - \sum_{s\in\mathcal{S}} I\{S_i = s\}\frac{n_a(s)}{n(s)} : a \in \mathcal{A}\right)' : 1 \le i \le n\right)$$

is an $(n\times|A|)$-dimensional matrix, where we have used that $\mathbb{S}_n'\mathbb{S}_n = \mathrm{diag}\,(n(s) : s \in \mathcal{S})$ and that $\mathbb{S}_n'\mathbb{A}_n$ is an $(|\mathcal{S}|\times|\mathcal{A}|)$-dimensional matrix with (s,a)th element given by $n_a(s)$. It follows from the above derivation and Assumption 4.1 that the $(a,\tilde{a})$ element of $\frac{1}{n}\mathbb{A}_n'\mathbb{M}_n'\mathbb{M}_n\mathbb{A}_n$ satisfies

$$I\{a = \tilde{a}\}\sum_{s\in\mathcal{S}} \frac{n_a(s)}{n} - \sum_{\tilde{s}\in\mathcal{S}} \frac{n_a(\tilde{s})n_{\tilde{a}}(\tilde{s})}{n(\tilde{s})n} \xrightarrow{P} I\{a = \tilde{a}\}\pi_a - \pi_a\pi_{\tilde{a}} ,$$

and so by the continuous mapping theorem we get

$$\left(\frac{1}{n}\mathbb{A}_n'\mathbb{M}_n'\mathbb{M}_n\mathbb{A}_n\right)^{-1} \xrightarrow{P} \mathrm{diag}\left(\frac{1}{\pi_a} : a \in \mathcal{A}\right) + \frac{1}{\pi_0}\iota_{|\mathcal{A}|}\iota_{|\mathcal{A}|}' .$$

Now consider the matrix $\frac{1}{n}\mathbb{A}_n'\mathbb{M}_n'\mathbb{Y}_n$. Simple manipulations shows that

$$\begin{aligned}\frac{1}{n}\mathbb{A}_n'\mathbb{M}_n'\mathbb{Y}_n = \Bigg(&\sum_{s\in\mathcal{S}}\frac{1}{n}\sum_{i=1}^n I\{A_i = a, S_i = s\}\tilde{Y}_i(a) - \sum_{s\in\mathcal{S}}\sum_{\tilde{a}\in\mathcal{A}_0}\frac{n_a(s)}{n(s)}\frac{1}{n}\sum_{i=1}^n I\{A_i = \tilde{a}, S_i = s\}\tilde{Y}_i(\tilde{a}) \\ &+ \sum_{s\in\mathcal{S}}\frac{n_a(s)}{n(s)}\frac{n(s)}{n}E[m_a(Z)|S = s] - \sum_{\tilde{a}\in\mathcal{A}_0}\sum_{s\in\mathcal{S}}\frac{n_a(s)}{n(s)}\frac{n_{\tilde{a}}(s)}{n(s)}\frac{n(s)}{n}E[m_{\tilde{a}}(Z)|S = s] : a \in \mathcal{A}\Bigg)\end{aligned}$$

We conclude that

$$\sqrt{n}(\hat{\theta}_n^* - \theta(Q)) = \left(\mathrm{diag}\left(\frac{1}{\pi_a} : a \in \mathcal{A}\right) + \frac{1}{\pi_0}\iota_{|\mathcal{A}|}\iota_{|\mathcal{A}|}' + o_P(1)\right)\frac{1}{\sqrt{n}}\mathbb{A}_n'\mathbb{M}_n'\mathbb{Y}_n .$$

Next, we derive the limiting distribution of $\frac{1}{\sqrt{n}}\mathbb{A}_n'\mathbb{M}_n'\mathbb{Y}_n$. In order to do this, write

$$\frac{1}{\sqrt{n}}\mathbb{A}_n'\mathbb{M}_n'\mathbb{Y}_n = \overline{\mathbb{L}}_n + o_P(1) \ ,$$

where

$$\begin{aligned}
\overline{\mathbb{L}}_n = \Bigg(& \sum_{s\in\mathcal{S}} \frac{1}{\sqrt{n}} \sum_{i=1}^n I\{A_i = a, S_i = s\}\tilde{Y}_i(a) - \pi_a \sum_{s\in\mathcal{S}}\sum_{\tilde{a}\in\mathcal{A}_0} \frac{1}{\sqrt{n}} \sum_{i=1}^n I\{A_i = \tilde{a}, S_i = s\}\tilde{Y}_i(\tilde{a}) \\
& + \pi_a \sum_{s\in\mathcal{S}} \sqrt{n}\left(\frac{n(s)}{n} - p(s)\right)\left[E[m_a(Z)|S=s] - \sum_{\tilde{a}\in\mathcal{A}_0}\pi_{\tilde{a}} E[m_{\tilde{a}}(Z)|S=s]\right] \\
& + \sum_{s\in\mathcal{S}} \sqrt{n}\left(\frac{n_a(s)}{n(s)} - \pi_a\right) p(s)\left[E[m_a(Z)|S=s] - \sum_{\tilde{a}\in\mathcal{A}_0}\pi_{\tilde{a}} E[m_{\tilde{a}}(Z)|S=s]\right] \\
& - \pi_a \sum_{\tilde{a}\in\mathcal{A}_0}\sum_{s\in\mathcal{S}} \sqrt{n}\left(\frac{n_{\tilde{a}}(s)}{n(s)} - \pi_{\tilde{a}}\right) p(s) E[m_{\tilde{a}}(Z)|S=s] : a \in \mathcal{A}\Bigg) \ .
\end{aligned}$$

Since the right-hand side is $O_P(1)$, then Slutzky's theorem and some simple manipulations shows that

$$\begin{aligned}
\sqrt{n}(\hat{\theta}_n^* - \theta(Q)) &= \left(\text{diag}\left(\frac{1}{\pi_a} : a \in \mathcal{A}\right) + \frac{1}{\pi_0}\iota_{|\mathcal{A}|}\iota_{|\mathcal{A}|}'\right)\overline{\mathbb{L}}_n + o_P(1) \\
&= \left(\sum_{s\in\mathcal{S}}\left(\bar{L}_{n,a}^{(1)}(s) - \bar{L}_{n,0}^{(1)}(s)\right) : a \in \mathcal{A}\right) + \left(\sum_{s\in\mathcal{S}} \bar{L}_{n,a}^{(2)}(s) : a \in \mathcal{A}\right) \\
&\quad + \left(\sum_{s\in\mathcal{S}} (\bar{L}_{n,a}^{(3)}(s) - \bar{L}_{n,0}^{(3)}(s)) : a \in \mathcal{A}\right) + o_P(1) \ ,
\end{aligned}$$

where for $(a,s) \in \mathcal{A}\times\mathcal{S}$,

$$\begin{aligned}
\bar{L}_{n,a}^{(1)}(s) &\equiv \frac{1}{\pi_a}\left[\frac{1}{\sqrt{n}}\sum_{i=1}^n I\{A_i = a, S_i = s\}\tilde{Y}_i(a)\right] \\
\bar{L}_{n,a}^{(2)}(s) &\equiv \sqrt{n}\left(\frac{n(s)}{n} - p(s)\right) E\left[m_a(Z) - m_0(Z)|S=s\right] \\
\bar{L}_{n,a}^{(3)}(s) &\equiv \sqrt{n}\left(\frac{n_a(s)}{n(s)} - \pi_a\right)\frac{p(s)}{\pi_a}\left[E[m_a(Z)|S=s] - \sum_{\tilde{a}\in\mathcal{A}}\pi_{\tilde{a}} E[m_{\tilde{a}}(Z)|S=s]\right] \ .
\end{aligned}$$

By Lemma C.2 and some additional calculations, it follows that

$$\begin{pmatrix} \left(\sum_{s\in\mathcal{S}}\left(\bar{L}_{n,a}^{(1)}(s) - \bar{L}_{n,0}^{(1)}(s)\right) : a \in \mathcal{A}\right) \\ \left(\sum_{s\in\mathcal{S}} \bar{L}_{n,a}^{(2)}(s) : a \in \mathcal{A}\right) \\ \left(\sum_{s\in\mathcal{S}} (\bar{L}_{n,a}^{(3)}(s) - \bar{L}_{n,0}^{(3)}(s)) : a \in \mathcal{A}\right) \end{pmatrix} \xrightarrow{d} N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \mathbb{V}_{\tilde{Y}} & 0 & 0 \\ 0 & \mathbb{V}_H & 0 \\ 0 & 0 & \mathbb{V}_A \end{pmatrix}\right) \ ,$$

where $\mathbb{V}_{\tilde{Y}}$ is as in (12) with $\pi_a(s) = \pi_a$ for all $(a,s) \in \mathcal{A}_0 \times \mathcal{S}$, $\mathbb{V}_H$ is as in (11), and

$$\begin{aligned}
\mathbb{V}_A = \Bigg(& \sum_{s\in\mathcal{S}} p(s)\Bigg(\xi_a(s)\xi_{a'}(s)\frac{\Sigma_D(s)_{[a,a']}}{\pi_a\pi_{a'}} - \xi_a(s)\xi_0(s)\frac{\Sigma_D(s)_{[a,0]}}{\pi_a\pi_0} \\
& - \xi_{a'}(s)\xi_0(s)\frac{\Sigma_D(s)_{[a',0]}}{\pi_{a'}\pi_0} + \xi_0(s)\xi_0(s)\frac{\Sigma_D(s)_{[0,0]}}{\pi_0\pi_0}\Bigg) : (a,a') \in \mathcal{A}\times\mathcal{A}\Bigg)
\end{aligned}$$

with

$$\xi_a(s) \equiv E[m_a(Z_i)|S_i = s] - \sum_{a' \in \mathcal{A}_0} \pi_{a'} E[m_{a'}(Z_i|S_i = s)] \ .$$

Importantly, to get $\mathbb{V}_H$ for the second term we used that $\sum_{s \in \mathcal{S}} p(s) E\left[m_a(Z) - m_0(Z)|S = s\right] = 0$ for all $a \in \mathcal{A}$.

Appendix C Auxiliary Results

Lemma C.1. *Suppose Q satisfies Assumption 2.1 and the treatment assignment mechanism satisfies Assumption 2.2. Define*

$$\mathbb{L}_n^{(1)} \equiv \left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} I\{A_i = a, S_i = s\} \tilde{Y}_i(a) : (a,s) \in \mathcal{A}_0 \times \mathcal{S} \right) \tag{C-37}$$

$$\mathbb{L}_n^{(2)} \equiv \left(\sqrt{n} \left(\frac{n(s)}{n} - p(s) \right) : s \in \mathcal{S} \right) \ , \tag{C-38}$$

and $\mathbb{L}_n = (\mathbb{L}_n^{(1)\prime}, \mathbb{L}_n^{(2)\prime})'$. It follows that

$$\mathbb{L}_n \xrightarrow{d} N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{pmatrix} \right) \ ,$$

where

$$\Sigma_1 = \text{diag}\left(\pi_a(s) p(s) \sigma^2_{\tilde{Y}(a)}(s) : (a,s) \in \mathcal{A}_0 \times \mathcal{S} \right)$$
$$\Sigma_2 = \text{diag}\left(p(s) : s \in \mathcal{S} \right) - \left(p(s) : s \in \mathcal{S} \right) \left(p(s) : s \in \mathcal{S} \right)' \ .$$

Proof. To prove our result, we first show that

$$\left\{ \mathbb{L}_n^{(1)}, \mathbb{L}_n^{(2)} \right\} \overset{d}{=} \left\{ \mathbb{L}_n^{*(1)}, \mathbb{L}_n^{(2)} \right\} + o_P(1) \ ,$$

for a random vector $\mathbb{L}_n^{*(1)}$ satisfying $\mathbb{L}_n^{*(1)} \perp\!\!\!\perp \mathbb{L}_n^{(2)}$ and $\mathbb{L}_n^{*(1)} \xrightarrow{d} N\left(0, \Sigma_1\right)$. We then combine this result with the fact that $\mathbb{L}_n^{(2)} \xrightarrow{d} N\left(0, \Sigma_2\right)$, which follows from $W^{(n)}$ consisting of n i.i.d. observations and the CLT.

Under the assumption that $W^{(n)}$ is i.i.d. and Assumption 2.2.(a), the distribution of $\mathbb{L}_n^{(1)}$ is the same as the distribution of the same quantity where units are ordered first by strata $s \in \mathcal{S}$ and then ordered by treatment assignment $a \in \mathcal{A}$ within strata. In order to exploit this observation, it is useful to introduce some further notation. Define $N(s) \equiv \sum_{i=1}^n I\{S_i < s\}$, $N_a(s) \equiv \sum_{i=1}^n I\{A_i < a, S_i = s\}$, $F(s) \equiv P\{S_i < s\}$, and $F_a(s) \equiv P\{A_i < a, S_i = s\}$ for all $(a,s) \in \mathcal{A} \times \mathcal{S}$. Furthermore, independently for each $(a,s) \in \mathcal{A} \times \mathcal{S}$ and independently of $(A^{(n)}, S^{(n)})$, let $\{\tilde{Y}_i^s(a) : 1 \leq i \leq n\}$ be i.i.d. with marginal distribution equal to the distribution of $\tilde{Y}_i(a)|S_i = s$. With this notation, define

$$\tilde{\mathbb{L}}_n^{(1)} \equiv \left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} I\{A_i = a, S_i = s\} \tilde{Y}_i^s(a) : (a,s) \in \mathcal{A}_0 \times \mathcal{S} \right) = \left(\frac{1}{\sqrt{n}} \sum_{i=n \frac{N(s)+N_a(s)}{n} + 1}^{n \frac{N(s)+N_{a+1}(s)}{n}} \tilde{Y}_i^s(a) : (a,s) \in \mathcal{A}_0 \times \mathcal{S} \right) \ .$$

By construction, $\{\tilde{\mathbb{L}}_n^{(1)} | S^{(n)}, A^{(n)}\} \overset{d}{=} \{\mathbb{L}_n^{(1)} | S^{(n)}, A^{(n)}\}$ and so $\tilde{\mathbb{L}}_n^{(1)} \overset{d}{=} \mathbb{L}_n^{(1)}$. Since $\mathbb{L}_n^{(2)}$ is only a function of $S^{(n)}$, we

further have that $\left\{\mathbb{L}_n^{(1)}, \mathbb{L}_n^{(2)}\right\} \stackrel{d}{=} \left\{\tilde{\mathbb{L}}_n^{(1)}, \mathbb{L}_n^{(2)}\right\}$. Next, define

$$\mathbb{L}_n^{*(1)} \equiv \left(\frac{1}{\sqrt{n}} \sum_{i=\lfloor n(F(s)+F_a(s))\rfloor+1}^{\lfloor n\left(F(s)+F_{a+1}(s)\right)\rfloor} \tilde{Y}_i^s(a) : (a,s) \in \mathcal{A}_0 \times \mathcal{S} \right) .$$

Note that $\mathbb{L}_n^{*(1)} \perp\!\!\!\perp \mathbb{L}_n^{(2)}$. Using similar partial sum arguments as those in Bugni et al. (2018, Lemma B.1), it follows that

$$L_{n,a}^{*(1)}(s) = \frac{1}{\sqrt{n}} \sum_{i=\lfloor n(F(s)+F_a(s))\rfloor+1}^{\lfloor n\left(F(s)+F_{a+1}(s)\right)\rfloor} \tilde{Y}_i^s(a) \stackrel{d}{\to} N\left(0, \pi_a(s)p(s)\sigma^2_{\tilde{Y}(a)}(s)\right) ,$$

for all $(a,s) \in \mathcal{A}_0 \times \mathcal{S}$, where we used that $F_{a+1}(s) - F_a(s) = \pi_a(s)p(s)$. By the independence of the components, it follows that $\mathbb{L}_n^{*(1)} \stackrel{d}{\to} N(0, \Sigma_1)$. We conclude the proof by arguing that

$$\tilde{L}_{n,a}^{(1)}(s) - L_{n,a}^{*(1)}(s) \stackrel{P}{\to} 0 ,$$

for all $(a,s) \in \mathcal{A}_0 \times \mathcal{S}$, where

$$\tilde{L}_{n,a}^{(1)}(s) = \frac{1}{\sqrt{n}} \sum_{i=n\frac{N(s)+N_a(s)}{n}+1}^{n\frac{N(s)+N_{a+1}(s)}{n}} \tilde{Y}_i^s(a) .$$

This in turn follows from

$$\left(\frac{N(s)}{n}, \frac{N_a(s)}{n}\right) \stackrel{P}{\to} (F(s), F_a(s))$$

for all $(a,s) \in \mathcal{A}_0 \times \mathcal{S}$ and again invoking similar arguments to those in Bugni et al. (2018, Lemma B.1). ■

Lemma C.2. *Suppose Q satisfies Assumption 2.1 and the treatment assignment mechanism satisfies Assumption 4.1. Define*

$$\mathbb{L}_n^{(1)} \equiv \left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} I\{A_i = a, S_i = s\} \tilde{Y}_i(a) : (a,s) \in \mathcal{A}_0 \times \mathcal{S} \right) \tag{C-39}$$

$$\mathbb{L}_n^{(2)} \equiv \left(\sqrt{n}\left(\frac{n(s)}{n} - p(s)\right) : s \in \mathcal{S} \right) , \tag{C-40}$$

$$\mathbb{L}_n^{(3)} \equiv \left(\sqrt{n}\left(\frac{n_a(s)}{n(s)} - \pi_a\right) : (a,s) \in \mathcal{A}_0 \times \mathcal{S} \right) , \tag{C-41}$$

and $\mathbb{L}_n = (\mathbb{L}_n^{(1)\prime}, \mathbb{L}_n^{(2)\prime}, \mathbb{L}_n^{(3)\prime})'$. It follows that

$$\mathbb{L}_n \stackrel{d}{\to} N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \Sigma_1 & 0 & 0 \\ 0 & \Sigma_2 & 0 \\ 0 & 0 & \Sigma_3 \end{pmatrix} \right) ,$$

where

$$\Sigma_1 = \text{diag}\left(\pi_a(s)p(s)\sigma^2_{\tilde{Y}(a)}(s) : (a,s) \in \mathcal{A}_0 \times \mathcal{S}\right)$$
$$\Sigma_2 = \text{diag}\left(p(s) : s \in \mathcal{S}\right) - \left(p(s) : s \in \mathcal{S}\right)\left(p(s) : s \in \mathcal{S}\right)'$$
$$\Sigma_3 = \text{diag}\left(\Sigma_D(s)/p(s) : s \in \mathcal{S}\right) .$$

Proof. To prove our result, we first show that

$$\left\{\mathbb{L}_n^{(1)}, \mathbb{L}_n^{(2)}, \mathbb{L}_n^{(3)}\right\} \stackrel{d}{=} \left\{\mathbb{L}_n^{*(1)}, \mathbb{L}_n^{(2)}, \mathbb{L}_n^{(3)}\right\} + o_P(1) \ ,$$

for a random vector $\mathbb{L}_n^{*(1)}$ satisfying $\mathbb{L}_n^{*(1)} \perp\!\!\!\perp (\mathbb{L}_n^{(2)}, \mathbb{L}_n^{(3)})$ and $\mathbb{L}_n^{*(1)} \stackrel{d}{\to} N(0, \Sigma_1)$. We then combine this result with the fact that $\mathbb{L}_n^{(2)} \stackrel{d}{\to} N(0, \Sigma_2)$, which follows from $W^{(n)}$ consisting of n i.i.d. observations and the CLT, and the fact that conditional on $S^{(n)}$, $\mathbb{L}_n^{(3)} \stackrel{d}{\to} N(0, \Sigma_3)$, which follows from Assumption 4.1. The proof of (C) follows from similar arguments to those used in the proof of Lemma C.1 and so we omit them here. ■

Lemma C.3. *Suppose Q satisfies Assumption 2.1 and the treatment assignment mechanism satisfies Assumption 2.2. Let*

$$\mathbb{C}_n'\mathbb{C}_n = \begin{bmatrix} \text{diag}(n(s) : s \in \mathcal{S}) & \sum_{s\in\mathcal{S}} \mathbb{J}_s \otimes (n_a(s) : a \in \mathcal{A})' \\ \sum_{s\in\mathcal{S}} \mathbb{J}_s \otimes (n_a(s) : a \in \mathcal{A}) & \text{diag}(n_a(s) : (a,s) \in \mathcal{A}\times\mathcal{S}) \end{bmatrix} , \tag{C-42}$$

and

$$\mathbb{C}_n'\mathbb{Y}_n = \begin{bmatrix} \left(\sum_{a\in\mathcal{A}_0}\sum_{i=1}^n I\{A_i = a, S_i = s\}\tilde{Y}_i(a) + \sum_{a\in\mathcal{A}_0} n_a(s)\left(E[m_a(Z)|S=s] + \mu_a\right) : s \in \mathcal{S}\right) \\ \left(\sum_{i=1}^n I\{A_i = a, S_i = s\}\tilde{Y}_i(a) + n_a(s)(E[m_a(Z)|S=s] + \mu_a) : (a,s) \in \mathcal{A}\times\mathcal{S}\right) \end{bmatrix} , \tag{C-43}$$

where $\mathbb{Y}_n \equiv (Y_i : 1 \le i \le n)$. It follows that

$$\frac{1}{n}\mathbb{C}_n'\mathbb{C}_n \stackrel{P}{\to} \Sigma_C \equiv \begin{bmatrix} \text{diag}(p(s) : s \in \mathcal{S}) & \sum_{s\in\mathcal{S}} \mathbb{J}_s \otimes (\pi_a(s)p(s) : a \in \mathcal{A})' \\ \sum_{s\in\mathcal{S}} \mathbb{J}_s \otimes (\pi_a(s)p(s) : a \in \mathcal{A}) & \text{diag}(\pi_a(s)p(s) : (a,s) \in \mathcal{A}\times\mathcal{S}) \end{bmatrix} ,$$

and

$$\frac{1}{n}\mathbb{C}_n'\mathbb{Y}_n \stackrel{P}{\to} \begin{bmatrix} \left(p(s)\sum_{a\in\mathcal{A}_0}\pi_a(s)\left(E[m_a(Z)|S=s] + \mu_a\right) : s \in \mathcal{S}\right) \\ \left(p(s)\pi_a(s)(E[m_a(Z)|S=s] + \mu_a) : (a,s) \in \mathcal{A}\times\mathcal{S}\right) \end{bmatrix} .$$

In addition,

$$\Sigma_C^{-1} = \begin{bmatrix} \text{diag}\left(\frac{1}{\pi_0(s)p(s)} : s \in \mathcal{S}\right) & \sum_{s\in\mathcal{S}} \mathbb{J}_s \otimes \left(\frac{-1}{\pi_0(s)p(s)} : a \in \mathcal{A}\right)' \\ \sum_{s\in\mathcal{S}} \mathbb{J}_s \otimes \left(\frac{-1}{\pi_0(s)p(s)} : a \in \mathcal{A}\right) & \sum_{s\in\mathcal{S}} \mathbb{J}_s \otimes \left(\text{diag}\left(\frac{1}{\pi_a(s)p(s)} : a \in \mathcal{A}\right) + \frac{1}{\pi_0(s)p(s)}\iota_{|\mathcal{A}|}\iota_{|\mathcal{A}|}'\right) \end{bmatrix} .$$

Proof. The first result follows immediately from Assumption 2.2.(b) and the fact that $\frac{n(s)}{n} \stackrel{P}{\to} p(s)$ and $\frac{n_a(s)}{n} = \frac{n_a(s)}{n(s)}\frac{n(s)}{n} \stackrel{P}{\to} \pi_a(s)p(s)$ for all $(a,s) \in \mathcal{A}\times S$. For the second result, consider the following argument,

$$\begin{aligned} \frac{1}{n}\mathbb{C}_n'\mathbb{Y}_n &= \frac{1}{n}\sum_{i=1}^n \begin{bmatrix} (I\{S_i = s\}Y_i : s \in \mathcal{S}) \\ (I\{A_i = a, S_i = s\}Y_i : (a,s) \in \mathcal{A}\times\mathcal{S}) \end{bmatrix} \\ &= \frac{1}{n}\sum_{i=1}^n \begin{bmatrix} \left(\sum_{a\in\mathcal{A}_0} I\{A_i = a, S_i = s\}\left[\tilde{Y}_i(a) + E\left[m_a(Z)|S_i = s\right] + \mu_a\right] : s \in \mathcal{S}\right) \\ \left(I\{A_i = a, S_i = s\}\left[\tilde{Y}_i(a) + E\left[m_a(Z)|S_i = s\right] + \mu_a\right] : (a,s) \in \mathcal{A}\times\mathcal{S}\right) \end{bmatrix} \\ &= \begin{bmatrix} \left(p(s)\sum_{a\in\mathcal{A}_0}\pi_a(s)(E\left[m_a(Z)|S=s\right] + \mu_a) : s \in \mathcal{S}\right) \\ (p(s)\pi_a(s)(E\left[m_a(Z)|S=s\right] + \mu_a) : (a,s) \in \mathcal{A}\times\mathcal{S}) \end{bmatrix} + o_P(1) \end{aligned}$$

where we used $\frac{1}{n}\sum_{i=1}^n I\{A_i = a, S_i = s\} = \frac{n_a(s)}{n} \stackrel{P}{\to} \pi_a(s)p(s)$, and $\frac{1}{n}\sum_{i=1}^n I\{A_i = a, S_i = s\}\tilde{Y}_i(a) \stackrel{P}{\to} 0$ for all $(a,s) \in \mathcal{A}_0\times\mathcal{S}$. Finally, the last result follows from simple manipulations that we omit. ■

Lemma C.4. *Suppose Q satisfies Assumption 2.1 and the treatment assignment mechanism satisfies Assumption*

2.2. Let $W_i = f((Y_i(a) : a \in \mathcal{A}), S_i)$ for some function $f(\cdot)$ satisfy $E[|W_i|] < \infty$. Then, for all $a \in \mathcal{A}_0$,

$$\frac{1}{n}\sum_{i=1}^{n} W_i I\{A_i = a\} \xrightarrow{P} \sum_{s\in\mathcal{S}} p(s)\pi_a(s)E[W_i] \ . \tag{C-44}$$

Proof. Fix $a \in \mathcal{A}_0$. By arguing as in the proof of Lemma C.1, note that

$$\frac{1}{n}\sum_{i=1}^{n} W_i I\{A_i = a\} \stackrel{d}{=} \sum_{s\in\mathcal{S}} \frac{1}{n}\sum_{i=1}^{n_a(s)} W_i^s \ ,$$

where, independently for each $s \in S$ and independently of $(A^{(n)}, S^{(n)})$, $\{W_i^s : 1 \le i \le n\}$ are i.i.d. with marginal distribution equal to the distribution of $W_i|S_i = s$. In order to establish the desired result, it suffices to show that

$$\frac{1}{n}\sum_{i=1}^{n_a(s)} W_i^s \xrightarrow{P} p(s)\pi_a(s)E[W_i^s] \ . \tag{C-45}$$

From Assumption 2.2.(b), $\frac{n_a(s)}{n} \xrightarrow{P} p(s)\pi_a(s)$, so (C-45) follows from

$$\frac{1}{n_a(s)}\sum_{i=1}^{n_a(s)} W_i^s \xrightarrow{P} E[W_i^s] \ . \tag{C-46}$$

To establish (C-46), use the almost sure representation theorem to construct $\frac{\tilde{n}_a(s)}{n}$ such that $\frac{\tilde{n}_a(s)}{n} \stackrel{d}{=} \frac{n_a(s)}{n}$ and $\frac{\tilde{n}_a(s)}{n} \to p(s)\pi_a(s)$ a.s. Using the independence of $(A^{(n)}, S^{(n)})$ and $\{W_i^s : 1 \le i \le n\}$, we see that for any $\epsilon > 0$,

$$\begin{aligned}
P\left\{\left|\frac{1}{n_a(s)}\sum_{i=1}^{n_a(s)} W_i^s - E[W_i^s]\right| > \epsilon\right\} &= P\left\{\left|\frac{1}{n\frac{n_a(s)}{n}}\sum_{i=1}^{n\frac{n_a(s)}{n}} W_i^s - E[W_i^s]\right| > \epsilon\right\} \\
&= P\left\{\left|\frac{1}{n\frac{\tilde{n}_a(s)}{n}}\sum_{i=1}^{n\frac{\tilde{n}_a(s)}{n}} W_i^s - E[W_i^s]\right| > \epsilon\right\} \\
&= E\left[P\left\{\left|\frac{1}{n\frac{\tilde{n}_a(s)}{n}}\sum_{i=1}^{n\frac{\tilde{n}_a(s)}{n}} W_i^s - E[W_i^s]\right| > \epsilon \middle| \frac{\tilde{n}_a(s)}{n}\right\}\right] \\
&\to 0 \ ,
\end{aligned}$$

where the convergence follows from the dominated convergence theorem and

$$P\left\{\left|\frac{1}{n\frac{\tilde{n}_a(s)}{n}}\sum_{i=1}^{n\frac{\tilde{n}_a(s)}{n}} W_i^s - E[W_i^s]\right| > \epsilon \middle| \frac{\tilde{n}_a(s)}{n}\right\} \to 0 \text{ a.s. .} \tag{C-47}$$

To see that the convergence (C-47) holds, note that the weak law of large numbers implies that

$$\frac{1}{n_k}\sum_{i=1}^{n_k} W_i^s \xrightarrow{P} E[W_i^s] \tag{C-48}$$

for any subsequence $n_k \to \infty$ as $k \to \infty$. Since $n\frac{\tilde{n}_a(s)}{n} \to \infty$ a.s., (C-47) follows from the independence of $\frac{\tilde{n}_a(s)}{n}$ and $\{W_i^s : 1 \le i \le n\}$ and (C-48). ■

Lemma C.5. *Suppose Q satisfies Assumption 2.1 and the treatment assignment mechanism satisfies Assumption*

2.2. Let $\hat{u}_i = Y_i - C_i\hat{\gamma}_n$ *and* $\hat{\gamma}_n = \left((\hat{\delta}_n(s) : s \in \mathcal{S})', (\hat{\beta}_{n,a}(s) : (a,s) \in \mathcal{A} \times \mathcal{S})'\right)'$*, where* C_i *is as in* (B.1)*, be the least squares residuals associated with the regression in* (7)*. Then,*

$$\frac{1}{n}\sum_{i=1}^{n} \hat{u}_i^2 \xrightarrow{P} \sum_{(a,s)\in\mathcal{A}_0\times\mathcal{S}} p(s)\pi_a(s)\sigma^2_{\tilde{Y}(a)}(s)$$

$$\frac{1}{n}\sum_{i=1}^{n} \hat{u}_i^2 I\{A_i = a, S_i = s\} \xrightarrow{P} p(s)\pi_a(s)\sigma^2_{\tilde{Y}(a)}(s)$$

$$\frac{1}{n}\sum_{i=1}^{n} \hat{u}_i^2 I\{S_i = s\} \xrightarrow{P} \sum_{a\in\mathcal{A}_0} p(s)\pi_a(s)\sigma^2_{\tilde{Y}(a)}(s)$$

$$\frac{1}{n}\sum_{i=1}^{n} \hat{u}_i^2 I\{A_i = a\} \xrightarrow{P} \sum_{s\in\mathcal{S}} p(s)\pi_a(s)\sigma^2_{\tilde{Y}(a)}(s) \ .$$

Proof. First note that, by definition of $\tilde{Y}_i(a)$, we can write.

$$Y_i = \sum_{(a,s)\in\mathcal{A}_0\times S} I\{A_i = a, S_i = s\}[\tilde{Y}_i(a) + E[m_a(Z)|S = s] + \mu_a] \ .$$

In addition, for $\gamma = ((\delta(s) : s \in \mathcal{S})', (\beta_a(s) : (a,s) \in \mathcal{A} \times \mathcal{S})')'$

$$\begin{aligned} C_i\gamma = &\sum_{s\in\mathcal{S}} I\{S_i = s\}\left(E\left[m_0(Z)|S = s\right] + \mu_0\right) \\ &+ \sum_{(a,s)\in\mathcal{A}\times\mathcal{S}} I\{A_i = a, S_i = s\}\left[E[m_a(Z) - m_0(Z)|S = s] + \theta_a\right] \ . \end{aligned}$$

We can therefore write the error term u_i as

$$u_i = Y_i - C_i\gamma = \sum_{(a,s)\in\mathcal{A}_0\times\mathcal{S}} I\{A_i = a, S_i = s\}\tilde{Y}_i(a) \ ,$$

and its square as

$$u_i^2 = \sum_{(a,s)\in\mathcal{A}_0\times\mathcal{S}} I\{A_i = a, S_i = s\}\tilde{Y}_i^2(a) \ .$$

By arguments similar to those in Bugni et al. (2018, Lemma B.8), it is enough to show the results with u_i^2 in place of $\hat{u}_i^2$. Since $E[u_i^2] = p(s)\pi_a(s)\sigma^2_{\tilde{Y}(a)}(s)$, the results follow immediately by invoking Lemma C.4 repeatedly. We therefore omit the arguments here. ■

Lemma C.6. *Suppose* Q *satisfies Assumption 2.1 and the treatment assignment mechanism satisfies Assumption 4.1. Let* $\hat{\mathbb{V}}^*_{\text{ho}}$ *be the homoskedasticity-only estimator of the asymptotic variance for the regression in* (18)*, defined as*

$$\hat{\mathbb{V}}^*_{\text{ho}} = \left(\frac{1}{n}\sum_{i=1}^{n}\hat{u}_i^2\right)\mathbb{R}^*\left(\frac{1}{n}\mathbb{C}_n^{*\prime}\mathbb{C}_n^*\right)^{-1}\mathbb{R}^{*\prime} \ , \tag{C-49}$$

where $\{\hat{u}_i : 1 \le i \le n\}$ *are the least squares residuals,* $\mathbb{C}_n^*$ *is the matrix with* i*th row given by*

$$C_i^* = [(I\{S_i = s\} : s \in \mathcal{S})', (I\{A_i = a\} : a \in \mathcal{A})'] \ ,$$

and $\mathbb{R}^*$ *is a matrix with* $|\mathcal{A}|$ *rows and* $|\mathcal{S}| + |\mathcal{A}|$ *columns defined as* $\mathbb{R}^* = \left[\mathbb{O}, \mathbb{I}_{|\mathcal{A}|}\right]$*, where* $\mathbb{O}$ *and* $\mathbb{I}_{|\mathcal{A}|}$ *are defined in*

Table 8. Then.

$$\hat{\mathbb{V}}^*_{\text{ho}} \xrightarrow{P} \left(\sum_{(a,s)\in\mathcal{A}_0\times\mathcal{S}} p(s)\pi_a\sigma^2_{\tilde{Y}(a)}(s) + \sum_{s\in\mathcal{S}} p(s)\varsigma^2_H(s) \right) \left[\frac{1}{\pi_0}\iota_{|\mathcal{A}|}\iota'_{|\mathcal{A}|} + \text{diag}\left(\frac{1}{\pi_a} : a\in\mathcal{A} \right) \right]$$

where

$$\varsigma^2_H(s) = \sum_{a\in\mathcal{A}_0} \pi_a \left(E[m_a(Z_i)|S=s] \right)^2 - \left(\sum_{a\in\mathcal{A}_0} \pi_a E[m_a(Z_i)|S=s] \right)^2 .$$

Proof. The proof is similar to that of Theorem 3.2 and therefore omitted. ■

Lemma C.7. *Suppose Q satisfies Assumption 2.1 and the treatment assignment mechanism satisfies Assumption 4.1. Let $\hat{\mathbb{V}}^*_{\text{he}}$ be the heteroskedasticity-consistent estimator of the asymptotic variance for the regression in (18), defined as*

$$\hat{\mathbb{V}}^*_{\text{he}} = \mathbb{R}^* \left[\left(\frac{\mathbb{C}^{*\prime}_n \mathbb{C}^*_n}{n} \right)^{-1} \left(\frac{\mathbb{C}^{*\prime}_n \,\text{diag}(\{\hat{u}^2_i\}^n_{i=1})\mathbb{C}^*_n}{n} \right) \left(\frac{\mathbb{C}^{*\prime}_n \mathbb{C}^*_n}{n} \right)^{-1} \right] \mathbb{R}^{*\prime} , \tag{C-50}$$

where $\{\hat{u}_i : 1\le i\le n\}$ are the ordinary least squares residuals, and $\mathbb{C}^*_n$ and $\mathbb{R}^*$ are defined as in Lemma C.6. Then.

$$\hat{\mathbb{V}}^*_{\text{he}} \xrightarrow{P} \mathbb{V}^*_1 + \mathbb{V}^*_2 ,$$

where

$$\mathbb{V}^*_1 = \text{diag}\left(\sum_{s\in\mathcal{S}} \frac{p(s)}{\pi_a} \left[\sigma^2_{\tilde{Y}(a)}(s) + \left(E[m_a(Z)|S=s] - \sum_{\tilde{a}\in\mathcal{A}_0} \pi_{\tilde{a}} E[m_{\tilde{a}}(Z)|S=s] \right)^2 \right] : a\in\mathcal{A} \right)$$

$$\mathbb{V}^*_2 = \iota_{|\mathcal{A}|}\iota'_{|\mathcal{A}|} \sum_{s\in\mathcal{S}} \frac{p(s)}{\pi_0} \left[\sigma^2_{\tilde{Y}(0)}(s) + \left(E[m_0(Z)|S=s] - \sum_{\tilde{a}\in\mathcal{A}_0} \pi_{\tilde{a}} E[m_{\tilde{a}}(Z)|S=s] \right)^2 \right] .$$

Proof. The proof is similar to that of Theorem 3.2 and therefore omitted. ■

Appendix D Results on Local Power

Let $\{Q^*_n : n\ge 1\}$ be a sequence of local alternatives to the null hypothesis in (4) that satisfies

$$\sqrt{n}(\Psi\theta(Q^*_n) - c) \to \lambda \tag{D-51}$$

as $n\to\infty$, for λ and c being r-dimensional column vectors and Ψ being a $(r\times|\mathcal{A}|)$-dimensional matrix such that $\text{rank}(\Psi) = r$. Consider a test of the form

$$\phi_n(X^{(n)}) = I\{T_n(X^{(n)}) > \chi^2_{r,1-\alpha}\} ,$$

where

$$T_n(X^{(n)}) = n(\Psi\hat{\theta}_n - c)'(\Psi\hat{\mathbb{V}}_n\Psi')^{-1}(\Psi\hat{\theta}_n - c) ,$$

$\hat{\theta}_n$ is an estimator of $\theta(Q)$ satisfying

$$\sqrt{n}(\hat{\theta}_n - \theta(Q^*_n)) \xrightarrow{d} N(0,\mathbb{V}) \text{ under } Q^*_n \tag{D-52}$$

for some asymptotic variance $\mathbb{V}$, $\hat{\mathbb{V}}_n$ is a matrix intended to Studentize the test statistic that satisfies

$$\hat{\mathbb{V}}_n \xrightarrow{P} \mathbb{V}_{\text{stud}} \text{ under } Q_n^* \tag{D-53}$$

for some $\mathbb{V}_{\text{stud}}$, and $\chi^2_{r,1-\alpha}$ is the $1-\alpha$ quantile of a χ^2 random variable with r degrees of freedom. The next theorem summarizes our main result.

Theorem D.1. *Let $\{Q_n^* : n \geq 1\}$ be the sequence of local alternatives satisfying (D-51), $\hat{\theta}_n$ be an estimator satisfying (D-52), and $\hat{\mathbb{V}}_n$ be a random matrix satisfying (D-53). Assume that $\mathbb{V}$ and $\mathbb{V}_{\text{stud}}$ are positive definite, that $\mathbb{V}_{\text{stud}} - \mathbb{V}$ is positive semi-definite, and that rank$(\Psi) = r$. Then,*

$$\lim_{n\to\infty} E[\phi_n(X^{(n)})] = P\left\{(\xi+\tilde{\lambda})'(\Psi\mathbb{V}\Psi')^{1/2}(\Psi\mathbb{V}_{\text{stud}}\Psi')^{-1}(\Psi\mathbb{V}\Psi')^{1/2}(\xi+\tilde{\lambda}) > \chi^2_{r,1-\alpha}\right\}, \tag{D-54}$$

under Q_n^*, where $\xi \sim N(0, \mathbb{I}_r)$ and $\tilde{\lambda} = (\Psi\mathbb{V}\Psi')^{-1/2}\lambda$. In addition, the following three statements follow under Q_n^*.

(a) Under the assumptions above,

$$\limsup_{n\to\infty} E[\phi_n(X^{(n)})] \leq \; P\left\{(\xi+\tilde{\lambda})'(\xi+\tilde{\lambda}) > \chi^2_{r,1-\alpha}\right\} \; .$$

(b) If $\mathbb{V} = \mathbb{V}_{\text{stud}}$, then

$$\lim_{n\to\infty} E[\phi_n(X^{(n)})] = P\left\{(\xi+\tilde{\lambda})'(\xi+\tilde{\lambda}) > \chi^2_{r,1-\alpha}\right\} \geq \alpha \; ,$$

where the inequality is strict if and only if $\lambda \neq 0$.

(c) If $\phi_n^1(X^{(n)})$ and $\phi_n^2(X^{(n)})$ are two tests such that $\phi_n^1(X^{(n)})$ is based on an estimator with $\mathbb{V}^1 = \mathbb{V}^1_{\text{stud}}$ and $\phi_n^2(X^{(n)})$ is based on an estimator with $\mathbb{V}^2 = \mathbb{V}^2_{\text{stud}}$, then

$$\lim_{n\to\infty} E[\phi_n^1(X^{(n)})] \geq \lim_{n\to\infty} E[\phi_n^2(X^{(n)})] \; ,$$

provided $\mathbb{V}^2 - \mathbb{V}^1$ is positive semi-definite. In addition, the inequality becomes strict if and only if $\lambda \neq 0$ and $\mathbb{V}^2 - \mathbb{V}^1$ is positive definite.

Proof. Notice that

$$\sqrt{n}(\Psi\hat{\theta}_n - c) = \sqrt{n}(\Psi\hat{\theta}_n - \Psi\theta(Q_n^*)) + \sqrt{n}(\Psi\theta(Q_n^*) - c) \xrightarrow{d} N(\lambda, \Psi\mathbb{V}\Psi') \text{ under } Q_n^* \; .$$

By Slutsky's theorem,

$$\begin{aligned}(\Psi\hat{\mathbb{V}}_n\Psi')^{-1/2}\sqrt{n}(\Psi\hat{\theta}_n - c) &\xrightarrow{d} N\left((\Psi\mathbb{V}_{\text{stud}}\Psi')^{-1/2}\lambda, (\Psi\mathbb{V}_{\text{stud}}\Psi')^{-1/2}(\Psi\mathbb{V}\Psi')(\Psi\mathbb{V}_{\text{stud}}\Psi')^{-1/2}\right) \\ &\sim (\Psi\mathbb{V}_{\text{stud}}\Psi')^{-1/2}(\Psi\mathbb{V}\Psi')^{1/2}(\xi+\tilde{\lambda}) \; ,\end{aligned}$$

under Q_n^*, with $\xi \sim N(0, \mathbb{I}_r)$ and $\tilde{\lambda} = (\Psi\mathbb{V}\Psi')^{-1/2}\lambda$. From here we conclude that

$$T_n(X^{(n)}) \xrightarrow{d} (\xi+\tilde{\lambda})'(\Psi\mathbb{V}\Psi')^{1/2}(\Psi\mathbb{V}_{\text{stud}}\Psi')^{-1}(\Psi\mathbb{V}\Psi')^{1/2}(\xi+\tilde{\lambda}) \; ,$$

and (D-54) follows.

Part (a). This follows immediately from Lemma D.1.

Part (b). Note that

$$P\{(\xi+\tilde{\lambda})'(\xi+\tilde{\lambda}) > \chi^2_{r,1-\alpha}\} = \Lambda_{\frac{r}{2}}\left(\sqrt{\mu}, \sqrt{\chi^2_{r,1-\alpha}}\right) , \tag{D-55}$$

where $\Lambda_m(a,b)$ is the Marcum-Q-function and $\mu \equiv \tilde{\lambda}'\tilde{\lambda} = \lambda'(\Psi\mathbb{V}\Psi')^{-1}\lambda \geq 0$. By the fact that $\Lambda_m(a,b)$ is increasing in a (see Temme (2014, p. 575) and (Sun and Baricz, 2008, Theorem 3.1)), $\Lambda_{\frac{r}{2}}(\sqrt{\mu}, \sqrt{\chi^2_{r,1-\alpha}}) \geq \Lambda_{\frac{r}{2}}(0, \sqrt{\chi^2_{r,1-\alpha}}) = \alpha$, with strict inequality if and only if $\mu > 0$. Since $\mathbb{V}$ is positive definite and Ψ is full rank, $\Psi\mathbb{V}\Psi'$ is positive definite and, thus, non-singular. Then, $\mu > 0$ if and only if $\lambda \neq 0$.

Part (c). We only show the strict inequality, as the weak inequality follows from weakening all the inequalities. For $d = 1, 2$, since $\mathbb{V}^d$ is positive definite and Ψ is full rank, $\Psi\mathbb{V}^d\Psi'$ is positive definite and, thus, non-singular. Since $\mathbb{V}^2 - \mathbb{V}^1$ is positive definite and Ψ is full rank, $\Psi\mathbb{V}^2\Psi' - \Psi\mathbb{V}^1\Psi'$ is positive definite and so $(\Psi\mathbb{V}^2\Psi')^{-1} - (\Psi\mathbb{V}^1\Psi')^{-1}$ is negative definite. By this and the fact that $\lambda \neq 0$, we conclude that

$$\mu^2 - \mu^1 = \lambda'(\Psi\mathbb{V}^2\Psi')^{-1}\lambda - \lambda'(\Psi\mathbb{V}^1\Psi')^{-1}\lambda = \lambda'((\Psi\mathbb{V}^2\Psi')^{-1} - (\Psi\mathbb{V}^1\Psi')^{-1})\lambda < 0 .$$

By (D-55) and the fact that $\Lambda_m(a,b)$ is increasing in a, the result follows. ■

Lemma D.1. *Suppose that $\mathbb{V} - \mathbb{V}_{\text{stud}} \in \mathbf{R}^{|\mathcal{A}|\times|\mathcal{A}|}$ is negative semi-definite, $\mathbb{V}_{\text{stud}}$ is non-singular, and rank$(\Psi) = r$. Then, $(\Psi\mathbb{V}_{\text{stud}}\Psi')^{-1/2}(\Psi\mathbb{V}\Psi')(\Psi\mathbb{V}_{\text{stud}}\Psi')^{-1/2} - \mathbb{I}_r$ is negative semi-definite.*

Proof. Since Ψ is full rank and $\mathbb{V}_{\text{stud}}$ is non-singular, $(\Psi\mathbb{V}_{\text{stud}}\Psi')^{1/2}$ is well defined and non-singular. Let a be an arbitrary r-dimensional column vector. We wish to show that

$$a'((\Psi\mathbb{V}_{\text{stud}}\Psi')^{-1/2}(\Psi\mathbb{V}\Psi')(\Psi\mathbb{V}_{\text{stud}}\Psi')^{-1/2} - \mathbb{I}_r)a \leq 0 . \tag{D-56}$$

Let $b = (\Psi\mathbb{V}_{\text{stud}}\Psi')^{-1/2}a$ and note that (D-56) is equivalent to

$$b'(\Psi\mathbb{V}_{\text{stud}}\Psi')^{1/2}((\Psi\mathbb{V}_{\text{stud}}\Psi')^{-1/2}(\Psi\mathbb{V}\Psi')(\Psi\mathbb{V}_{\text{stud}}\Psi')^{-1/2} - \mathbb{I}_r)(\Psi\mathbb{V}_{\text{stud}}\Psi')^{1/2}b \leq 0$$

which, in turn, is equivalent to $(\Psi'b)'(\mathbb{V} - \mathbb{V}_{\text{stud}})(\Psi'b) \leq 0$. This last inequality holds because $\mathbb{V} - \mathbb{V}_{\text{stud}}$ is negative semi-definite. ■

References

Bai, Y. (2018). On optimal stratification in randomized controlled trials. Manuscript. The University of Chicago.

Berry, J., Karlan, D. S. and Pradhan, M. (2018). The impact of financial education for youth in Ghana. *World Development*, **102** 71 – 89.

Bruhn, M. and McKenzie, D. (2009). In pursuit of balance: Randomization in practice in development field experiments. *American Economic Journal: Applied Economics*, **1** 200–232.

Bugni, F. A., Canay, I. A. and Shaikh, A. M. (2018). Inference under covariate-adaptive randomization. *Journal of the American Statistical Association,* forthcoming.

Callen, M., Gulzar, S., Hasanain, A., Khan, Y. and Rezaee, A. (2019). Personalities and public sector performance: Evidence from a health experiment in Pakistan. NBER Working Paper No. 21180.

Chong, A., Cohen, I., Field, E., Nakasone, E. and Torero, M. (2016). Iron deficiency and schooling attainment in peru. *American Economic Journal: Applied Economics*, **8** 222–255.

Dizon-Ross, R. (2018). Parents' beliefs about their children's academic ability: implications for educational investments. Manuscript, University of Chicago Booth School of Business.

Duflo, E., Dupas, P. and Kremer, M. (2015). Education, HIV, and early fertility: Experimental evidence from Kenya. *American Economics Review*, **105** 2757–2797.

Duflo, E., Glennerster, R. and Kremer, M. (2007). Using randomization in development economics research: A toolkit. *Handbook of development economics*, **4** 3895–3962.

Efron, B. (1971). Forcing a sequential experiment to be balanced. *Biometrika*, **58** 403–417.

Hu, Y. and Hu, F. (2012). Asymptotic properties of covariate-adaptive randomization. *Annals of Statistics, forthcoming.*

Imbens, G. W. and Rubin, D. B. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction.* Cambridge University Press.

Kernan, W. N., Viscoli, C. M., Makuch, R. W., Brass, L. M. and Horwitz, R. I. (1999). Stratified randomization for clinical trials. *Journal of clinical epidemiology*, **52** 19–26.

Rosenberger, W. F. and Lachin, J. M. (2016). *Randomization in clinical trials: theory and practice.* 2nd ed. John Wiley & Sons.

Sun, Y. and Baricz, Á. (2008). Inequalities for the generalized marcum q-function. *Applied Mathematics and Computation*, **203** 134–141.

Tabord-Meehan, M. (2018). Stratification trees for adaptive randomization in randomized controlled trials. Manuscript. Northwestern University.

Temme, N. M. (2014). *Asymptotic methods for integrals*, vol. 11. World Scientific.

Wei, L., Smythe, R. and Smith, R. (1986). K-treatment comparisons with restricted randomization rules in clinical trials. *The Annals of Statistics* 265–274.

Zelen, M. (1974). The randomization and stratification of patients to clinical trials. *Journal of chronic diseases*, **27** 365–375.