cemmap

centre for microdata methods and practice

Inference under covariate-adaptive randomization

Federico Bugni
Ivan Canay
Azeem Shaikh

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP45/15

E·S·R·C
ECONOMIC
& SOCIAL
RESEARCH
COUNCIL

An ESRC Research Centre

Inference under Covariate-Adaptive Randomization*

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

August 6, 2015

Abstract

This paper studies inference for the average treatment effect in randomized controlled trials with covariate-adaptive randomization. Here, by covariate-adaptive randomization, we mean randomization schemes that first stratify according to baseline covariates and then assign treatment status so as to achieve "balance" within each stratum. Such schemes include, for example, Efron's biased-coin design and stratified block randomization. When testing the null hypothesis that the average treatment effect equals a pre-specified value in such settings, we first show that the usual two-sample t-test is conservative in the sense that it has limiting rejection probability under the null hypothesis no greater than and typically strictly less than the nominal level. In a simulation study, we find that the rejection probability may in fact be dramatically less than the nominal level. We show further that these same conclusions remain true for a naïve permutation test, but that a modified version of the permutation test yields a test that is non-conservative in the sense that its limiting rejection probability under the null hypothesis equals the nominal level. The modified version of the permutation test has the additional advantage that it has rejection probability exactly equal to the nominal level for some distributions satisfying the null hypothesis. Finally, we show that the usual t-test (on the coefficient on treatment assignment) in a linear regression of outcomes on treatment assignment and indicators for each of the strata yields a non-conservative test as well. In a simulation study, we find that the non-conservative tests have substantially greater power than the usual two-sample t-test.

KEYWORDS: Covariate-adaptive randomization, stratified block randomization, Efron's biased-coin design, treatment assignment, randomized controlled trial, permutation test, two-sample t-test, strata fixed effects

JEL classification codes: C12, C14

[*] We would like to thank Lori Beaman for helpful comments on this paper. The research of the second author was supported by National Science Foundation Grants SES-1530534. The research of the third author was supported by National Science Foundation Grants DMS-1308260, SES-1227091, and SES-1530661. This research was supported in part through the computational resources and staff contributions provided for the Social Sciences Computing cluster (SSCC) at Northwestern University.

1 Introduction

This paper studies inference for the average treatment effect in randomized controlled trials with covariate-adaptive randomization. Here, by covariate-adaptive randomization, we mean 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 in randomized controlled trials in economics and the social sciences more generally. Duflo et al. (2007) and Bruhn and McKenzie (2008) provide a review focused on methods used in randomized controlled trials in development economics. In this paper, we take as given the use of such a treatment assignment mechanism satisfying weak assumptions and study its consequences for testing the null hypothesis that the average treatment effect equals a pre-specified value in such settings.

Our first result establishes that the usual two-sample t-test is conservative in the sense that it has limiting rejection probability under the null hypothesis no greater than and typically strictly less than the nominal level. We additionally provide a characterization of when the limiting rejection probability under the null hypothesis is in fact strictly less than the nominal level. As explained further in Remark 4.4 below, our result substantially generalizes a related result obtained by Shao et al. (2010), who established this phenomenon under much stronger assumptions and for only one specific treatment assignment mechanism. We show further that these conclusions remain true for a naïve permutation test. In a simulation study, we find that the rejection probability of these tests may in fact be dramatically less than the nominal level, and, as a result, they may have very poor power when compared to other tests. Intuitively, the conservative feature of these tests is a consequence of the dependence in treatment status across units and between treatment status and baseline covariates resulting from covariate-adaptive randomization.

Motivated by these results, we go on to show that a modified version of the permutation test which only permutes treatment status for units within the same stratum yields a test that is non-conservative in the sense that its limiting rejection probability under the null hypothesis equals the nominal level. We refer to this test as the covariate-adaptive permutation test. As explained further in Remark 4.10 below, this test or closely related tests have been previously proposed and justified in finite samples for a much narrower version of the null hypothesis when treatment status is determined using very specific randomization schemes. See, for example, Rosenberger and Lachin (2004, Section 7.4), Rosenbaum (2007), and Heckman et al. (2011). Exploiting recent results on the large-sample behavior of permutation tests by Chung and Romano (2013), our results, in contrast, asymptotically justify the use of these tests for testing the null hypothesis that the average treatment effect equals a pre-specified value for a much wider variety of randomization schemes, while retaining the finite-sample validity for the narrower version of the null hypothesis.

We additionally consider the usual t-test (on the coefficient on treatment assignment) in a linear regression of outcomes on treatment assignment and indicators for each of the strata. We refer to this test as the t-test with strata fixed effects. Remarkably, this simple modification of the usual two-sample t-test yields a test that is non-conservative as well. On the other hand, this test does not enjoy the finite-sample validity of the covariate-adaptive permutation test for the narrower version of the null hypothesis, though it remains valid asymptotically for an even wider variety of randomization schemes.

While all of our results apply much more generally, it is important to emphasize that they apply in particular to stratified block randomization. In stratified block randomization, units are first stratified according to baseline covariates and then a subset of the units within each strata are chosen at random to be assigned to treatment. In a sense made more precise in Example 3.4 below, when approximately one half of the units within each strata are chosen to be assigned to treatment, this treatment assignment mechanism exhibits the best finite-sample "balancing" properties. It has therefore become increasingly popular, especially in development economics. Indeed, many very recent papers in development economics use this particular randomization scheme, including, for example, Dizon-Ross (2014, footnote 13), Duflo et al. (2014, footnote 6), Callen et al. (2015, page 24), and Berry et al. (2015, page 6).

The remainder of the paper is organized as follows. In Section 2, we describe our setup and notation. In particular, there we describe the weak assumptions we impose on the treatment assignment mechanism. In Section 3, we discuss several examples of treatment assignment mechanisms satisfying these assumptions, importantly including stratified block randomization. Our main results about the four tests mentioned above are contained in Section 4. In Section 5, we examine the finite-sample behavior of these tests as well as some other tests via a small simulation study. 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 an indicator for whether the ith unit is treated or not, and Z_i denote observed, baseline covariates for the ith unit. Further denote by $Y_i(1)$ the potential outcome of the ith unit if treated and by $Y_i(0)$ the potential outcome of the ith unit if not treated. As usual, the (observed) outcome and potential outcomes are related to treatment assignment by the relationship

$$Y_i = Y_i(1)A_i + Y_i(0)(1 - A_i) \ . \tag{1}$$

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

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

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.

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 $(Y_i(1), Y_i(0), Z_i)$. We further restrict Q to satisfy the following, mild requirement:

Assumption 2.1. For some $\delta > 0$, Q satisfies

$$E[|Y_i(1)|^{2+\delta}] < \infty \text{ and } E[|Y_i(0)|^{2+\delta}] < \infty \ .$$

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. To this end, let $S : \text{supp}(Z_i) \to \mathcal{S}$, where $\mathcal{S}$ is a finite set, be the function used to construct strata and, for $1 \leq i \leq n$, let $S_i = S(Z_i)$. Denote by $S^{(n)}$ the vector of strata $(S_1, \ldots, S_n)$ and denote by $A^{(n)}$ the vector of treatment assignments $(A_1, \ldots, A_n)$. For $s \in \mathcal{S}$, let $p(s) = P\{S_i = s\}$ and

$$D_n(s) = \sum_{1 \leq i \leq n} A_i^* I\{S_i = s\} \ , \tag{2}$$

where

$$A_i^* = 2A_i - 1 \ .$$

Note that $D_n(s)$ as defined in (2) is simply a measure of the imbalance in stratum s. In order to rule out trivial strata, we, of course, assume that $p(s) > 0$ for all $s \in \mathcal{S}$. Our other requirements on the treatment assignment mechanism are summarized in the following assumption:

Assumption 2.2. The treatment assignment mechanism is such that

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

(b) $E[A_i | S^{(n)}] = \frac{1}{2} + O_{a.s}(\frac{1}{n})$ for all $1 \leq i \leq n$,

(c) $\left\{ \left\{ \frac{D_n(s)}{\sqrt{n}} \right\}_{s \in \mathcal{S}} \middle| S^{(n)} \right\} \xrightarrow{d} N(0, \Sigma_D)$ a.s., where $\Sigma_D = \text{diag}\{\sigma_D^2(s) : s \in \mathcal{S}\}$ and

$$\sigma_D^2(s) = p(s)\tau(s) \text{ with } 0 \leq \tau(s) \leq 1 \text{ for all } s \in \mathcal{S} \ ,$$

(d) $\text{Var}[D_n(s)] \leq np(s)$ for all $s \in \mathcal{S}$.

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)–(d) are additional requirements that are satisfied by a wide variety of randomization schemes. In the following section, we provide several important examples of treatment assignment mechanisms satisfying this assumption, including many that are used routinely in economics and other social sciences.

Our object of interest is the average effect of the treatment on the outcome of interest, defined to be

$$\theta(Q) = E[Y_i(1) - Y_i(0)] \ . \tag{3}$$

For a pre-specified choice of θ_0, the testing problem of interest is

$$H_0 : \theta(Q) = \theta_0 \text{ versus } H_1 : \theta(Q) \neq \theta_0 \tag{4}$$

at level $\alpha \in (0, 1)$.

3 Examples

In this section, we briefly describe several different randomization schemes that satisfy our Assumption 2.2. A more detailed review of these methods and their properties can be found in Rosenberger and Lachin (2004). In our descriptions, we make use of the notation $A^{(k-1)} = (A_1, \ldots, A_{k-1})$ and $S^{(k)} = (S_1, \ldots, S_k)$ for $1 \leq k \leq n$, where $A^{(0)}$ is understood to be a constant.

Example 3.1. *(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

$$P\{A_k = 1 | S^{(n)}\} = P\{A_k = 1\} = \frac{1}{2} \tag{5}$$

for $1 \leq k \leq n$. In this case, Assumption 2.2.(a) follows immediately from (5), Assumption 2.2.(b) follows from $E[A_i] = \frac{1}{2}$, Assumption 2.2.(c) follows from the central limit theorem with $\tau(s) = 1$ for all $s \in \mathcal{S}$, and Assumption 2.2.(d) holds with equality for all $s \in \mathcal{S}$. Note that $E[D_n(s)] = 0$ for all $s \in \mathcal{S}$, so SRS ensures "balance" on average, yet in finite samples $D_n(s)$ may be far from zero. ■

Example 3.2. *(Biased-Coin Design)* A biased-coin design is a generalization of simple random sampling originally proposed by Efron (1971) with the aim of improving "balance" in finite samples. In this randomization scheme, treatment assignment is determined recursively for $1 \leq k \leq n$ as follows:

$$P\{A_k = 1 | S^{(k)}, A^{(k-1)}\} = \begin{cases} \frac{1}{2} & \text{if } D_{k-1}(S_k) = 0 \\ \pi & \text{if } D_{k-1}(S_k) < 0 \\ 1 - \pi & \text{if } D_{k-1}(S_k) > 0 \end{cases}, \tag{6}$$

where $D_{k-1}(S_k) = \sum_{1 \leq i \leq k-1} A_i^* I\{S_i = S_k\}$, and $\frac{1}{2} \leq \pi \leq 1$. Here, $D_0(S_1)$ is understood to be zero. When $\pi = \frac{1}{2}$, the scheme is just SRS; otherwise, it adjusts the probability with which the kth unit is assigned to treatment in an effort to improve "balance" in the corresponding stratum in finite samples. In this case, Assumption 2.2.(a) follows immediately from (6), Assumption 2.2.(b) follows from Rosenberger and Lachin (2004, Section 3.6), and Assumption 2.2.(c)-(d) follow from Markaryan and Rosenberger (2010, Proposition 4.1), which implies in particular that $D_n(s) = O_P(1)$ for all $s \in \mathcal{S}$, so that Assumption 2.2.(c) holds with $\tau(s) = 0$ for all $s \in \mathcal{S}$. In this sense, we see that biased-coin design provides improved "balance" relative to simple random sampling. ■

Example 3.3. *(Adaptive Biased-Coin Design)* An adaptive biased-coin design, also known as Wei's urn design, is an alternative generalization of SRS originally proposed by Wei (1978). This randomization scheme is similar to a biased-coin design, except that the probability π in (6) depends on $D_{k-1}(S_k)$, the magnitude of imbalance in the corresponding stratum. More precisely, in this randomization scheme, treatment assignment is determined recursively for $1 \leq k \leq n$ as follows:

$$P\{A_k = 1 | S^{(k)}, A^{(k-1)}\} = \varphi\left(\frac{D_{k-1}(S_k)}{k-1}\right) , \tag{7}$$

where $\varphi(x) : [-1,1] \to [0,1]$ is a pre-specified non-increasing function satisfying $\varphi(-x) = 1 - \varphi(x)$. Here, $\frac{D_0(S_1)}{0}$ is understood to be zero. In this case, Assumption 2.2.(a) follows immediately from (7), Assumption 2.2.(b) follows from Rosenberger and Lachin (2004, Section 3.7), Assumption 2.2.(c) follows from Wei (1978, Theorem 3), and Assumption 2.2.(d) follows from the proof of Theorem 4 in Baldi Antognini (2008). In particular, Assumption 2.2.(c) holds with $\tau(s) = (1 - 4\varphi'(0))^{-1} \in (0,1)$. In this sense, adaptive biased-coin designs provide improved "balance" relative to simple random sampling (i.e., $\tau(s) < 1$), but to a lesser extent than biased-coin designs (i.e., $\tau(s) > 0$). ■

Example 3.4. *(Stratified Block Randomization)* An early discussion of stratified block randomization is provided by Zelen (1974). 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 $n_1(s) \leq n(s)$ be given. In this randomization scheme, $n_1(s)$ units in stratum s are assigned to treatment and the remainder are assigned to control, where all

$$\binom{n(s)}{n_1(s)}$$

possible assignments are equally likely and treatment assignment across strata are independent. By setting

$$n_1(s) = \left\lfloor \frac{n(s)}{2} \right\rfloor , \tag{8}$$

this scheme ensures $|D_n(s)| \leq 1$ for all $s \in \mathcal{S}$ and therefore exhibits the best "balance" in finite samples among the methods discussed here. In this case, Assumption 2.2.(a) holds immediately and Assumption 2.2.(b)-(d) follow from the analysis in Hallstrom and Davis (1988). In particular, as in Example 3.2, Assumption 2.2.(c) holds with $\tau(s) = 0$ for all $s \in \mathcal{S}$. ■

Example 3.5. *(Minimization Methods)* Minimization methods were originally proposed by Pocock and Simon (1975) and more recently extended and further studied by Hu and Hu (2012). In Hu and Hu (2012), treatment assignment is determined recursively for $1 \leq k \leq n$ as follows:

$$P\{A_k = 1 | S^{(k)}, A^{(k-1)}\} = \begin{cases} \frac{1}{2} & \text{if } \mathrm{Imb}_k = 0 \\ \pi & \text{if } \mathrm{Imb}_k < 0 \\ 1 - \pi & \text{if } \mathrm{Imb}_k > 0 \end{cases} , \tag{9}$$

where $\frac{1}{2} \leq \pi \leq 1$ and $\mathrm{Imb}_k = \mathrm{Imb}_k(S^{(k)}, A^{(k-1)})$ is a weighted average of different measures of imbalance. See Hu and Hu (2012) for expressions of these quantities. In this case, Assumption 2.2.(a) holds immediately and Assumption 2.2.(b)–(d) can be established using arguments in Hu and Hu (2012). In particular, Hu and

Hu (2012, Theorem 3.2 and Remark 3.1) show that $D_n(s) = O_P(1)$ for all $s \in \mathcal{S}$, so, as in as in Examples 3.2 and 3.3, Assumption 2.2.(c) holds with $\tau(s) = 0$ for all $s \in \mathcal{S}$. ■

Remark 3.1. Another treatment assignment mechanism for randomized controlled trials that has received considerable attention is re-randomization. See, for example, Bruhn and McKenzie (2008) and Lock Morgan and Rubin (2012). In this case, as explained by Lock Morgan and Rubin (2012), the properties of $D_n(s)$ depend on the rule used to decide whether to re-randomize and how to re-randomize. As a result, the analysis of such randomization schemes is necessarily case-by-case, and we do not consider them further in this paper. ■

Remark 3.2. Our framework does not accommodate response-adaptive randomization schemes. In such randomization schemes, units are assigned to treatment sequentially and treatment assignment for the ith unit, A_i, depends on $Y_1, \ldots, Y_{i-1}$. This feature leads to a violation of part (a) of our Assumption 2.2. It is worth emphasizing that response-adaptive randomization schemes are only feasible when at least some of the outcomes are observed at some point of the treatment assignment process, which is unusual in experiments in economics and other social sciences. ■

4 Main Results

4.1 Two-Sample t-Test

In this section, we consider using the two-sample t-test to test (4) at level $\alpha \in (0,1)$. In order to define this test, for $a \in \{0,1\}$, let

$$\begin{aligned} \bar{Y}_{n,a} &= \frac{1}{n_a} \sum_{1 \le i \le n} Y_i I\{A_i = a\} \\ \hat{\sigma}^2_{n,a} &= \frac{1}{n_a} \sum_{1 \le i \le n} (Y_i - \bar{Y}_{n,a})^2 I\{A_i = a\} \ , \end{aligned}$$

where $n_a = \sum_{1 \le i \le n} I\{A_i = a\}$. The two-sample t-test is given by

$$\phi_n^{t\text{-test}}(X^{(n)}) = I\{|T_n^{t\text{-test}}(X^{(n)})| > z_{1-\frac{\alpha}{2}}\} \ , \tag{10}$$

where

$$T_n^{t\text{-test}}(X^{(n)}) = \frac{\bar{Y}_{n,1} - \bar{Y}_{n,0} - \theta_0}{\sqrt{\frac{\hat{\sigma}^2_{n,1}}{n_1} + \frac{\hat{\sigma}^2_{n,0}}{n_0}}} \tag{11}$$

and $z_{1-\frac{\alpha}{2}}$ is the $1-\frac{\alpha}{2}$ quantile of a standard normal random variable. This test may equivalently be described as the usual t-test (on the coefficient on treatment assignment) in a linear regression of outcomes on treatment assignment with heteroskedasticity-robust standard errors. It is used routinely throughout economics and the social sciences, including settings with covariate-adaptive randomization. Note that further results on linear regression are developed in Section 4.4 below.

The following theorem describes the asymptotic behavior of the two-sample t-statistic defined in (11) and, as a consequence, the two-sample t-test defined in (10) under covariate-adaptive randomization. In particular, the theorem shows that the limiting rejection probability of the two-sample t-test under the null hypothesis is generally strictly less than the nominal level.

Theorem 4.1. *Suppose Q satisfies Assumption 2.1 and the treatment assignment mechanism satisfies Assumption 2.2. Then,*

$$\frac{\bar{Y}_{n,1} - \bar{Y}_{n,0} - \theta(Q)}{\sqrt{\frac{\hat{\sigma}^2_{n,1}}{n_1} + \frac{\hat{\sigma}^2_{n,0}}{n_0}}} \xrightarrow{d} N(0, \sigma^2_{t\text{-test}}) \ ,$$

where $\sigma^2_{t\text{-test}} \leq 1$. Furthermore, $\sigma^2_{t\text{-test}} < 1$ unless

$$(1 - \tau(s))(E[m_1(Z_i)|S_i = s] + E[m_0(Z_i)|S_i = s])^2 = 0 \text{ for all } s \in \mathcal{S} \ , \tag{12}$$

where

$$m_a(Z_i) = E[Y_i(a)|Z_i] - E[Y_i(a)] \tag{13}$$

for $a \in \{0, 1\}$. Thus, for the problem of testing (4) at level $\alpha \in (0, 1)$, $\phi^{t\text{-test}}_n(X^{(n)})$ defined in (10) satisfies

$$\limsup_{n \to \infty} E[\phi^{t\text{-test}}_n(X^{(n)})] \leq \alpha \tag{14}$$

whenever Q additionally satisfies the null hypothesis, i.e., $\theta(Q) = \theta_0$. Furthermore, the inequality in (14) is strict unless (12) holds.

Remark 4.1. Note that the two-sample t-test defined in (10) uses the $1 - \frac{\alpha}{2}$ quantile of a standard normal random variable instead of the corresponding quantile of a t-distribution. Theorem 4.1 remains true with such a choice of critical value provided that the rule for choosing the degrees of freedom for the t-distribution diverges (in probability) with the sample size. See Imbens and Kolesar (2012) for a recent review of some such degrees of freedom adjustments. ■

Remark 4.2. While we generally expect that (12) will fail to hold, there are some important cases in which it does hold. First, as explained in Example 3.1, for simple random sampling Assumption 2.2 holds with $\tau(s) = 1$ for all $s \in \mathcal{S}$. Hence, (12) holds, and Theorem 4.1 implies, as one would expect, that the two-sample t-test is not conservative under simple random sampling. Second, if stratification is irrelevant for potential outcomes in the sense that $E[Y_i(a)|S_i] = E[Y_i(a)]$ for all $a \in \{0, 1\}$, then $E[m_a(Z_i)|S_i] = 0$ for $a \in \{0, 1\}$. Hence, (12) again holds, and Theorem 4.1 implies that the two-sample t-test is not conservative when stratification is irrelevant for potential outcomes. Note that a special case of irrelevant stratification is simply no stratification, i.e., S_i is constant. ■

Remark 4.3. In the proof of Theorem 4.1 in the Appendix, it is shown that

$$\sigma^2_{t\text{-test}} = \frac{\sigma^2_{\tilde{Y}} + \sigma^2_H + \sigma^2_A}{\sigma^2_Y} \ , \tag{15}$$

where

$$\begin{aligned}
\sigma^2_{Y} &= 2(\mathrm{Var}[Y_i(1)] + \mathrm{Var}[Y_i(0)]) && (16)\\
\sigma^2_{\tilde{Y}} &= 2(\mathrm{Var}[\tilde{Y}_i(1)] + \mathrm{Var}[\tilde{Y}_i(0)]) && (17)\\
\sigma^2_{H} &= E[(E[m_1(Z_i)|S_i] - E[m_0(Z_i)|S_i])^2] && (18)\\
\sigma^2_{A} &= E[\tau(S_i)(E[m_1(Z_i)|S_i] + E[m_0(Z_i)|S_i])^2] && (19)
\end{aligned}$$

with $\tilde{Y}_i(a) = Y_i(a) - E[Y_i(a)|S_i]$. From (15), we see that three different sources of variation contribute to the variance. The first quantity, $\sigma^2_{\tilde{Y}}$, reflects variation in the potential outcomes; the second quantity, σ^2_H, reflects variation due to heterogeneity in the responses to treatment, i.e., $m_1 \neq m_0$; and the third quantity, σ^2_A, reflects variation due to "imperfectly balanced" treatment assignment, i.e., $\sigma^2_D(s) > 0$ in Assumption 2.2. ■

Remark 4.4. Under substantially stronger assumptions than those in Theorem 4.1, Shao et al. (2010) also establish conservativeness of the two-sample t-test for a specific covariate-adaptive randomization scheme. Shao et al. (2010) require, in particular, that $m_a(Z_i) = \gamma' Z_i$, that $\mathrm{Var}[Y_i(a)|Z_i]$ does not depend on Z_i, and that the treatment assignment rule is a biased-coin design, as described in Example 3.2. Theorem 4.1 relaxes all of these requirements. ■

Remark 4.5. While Theorem 4.1 characterizes when the limiting rejection probability of the two-sample t-test under the null hypothesis is strictly less than the nominal level, it does not reveal how significant this difference might be. In our simulation study in Section 5, we find that the rejection probability may in fact be dramatically less than the nominal level and that this difference translates into substantial power losses when compared with non-conservative tests studied in Sections 4.3 and 4.4. ■

4.2 Naïve Permutation Test

In this section, we consider using a naïve permutation test to test (4) at level $\alpha \in (0,1)$. In order to define this test, let $\mathbf{G}_n$ to be the group of permutations of n elements. Define the action of $g \in \mathbf{G}_n$ on $X^{(n)}$ as follows:

$$gX^{(n)} = \{(Y_i, A_{g(i)}, Z_i) : 1 \leq i \leq n\} ,$$

i.e., $g \in \mathbf{G}_n$ acts on $X^{(n)}$ by permuting treatment assignment. For a given choice of test statistic $T_n(X^{(n)})$, the naïve permutation test is given by

$$\phi_n^{\text{naïve}}(X^{(n)}) = I\{T_n(X^{(n)}) > \hat{c}_n^{\text{naïve}}(1-\alpha)\} , \tag{20}$$

where

$$\hat{c}_n^{\text{naïve}}(1-\alpha) = \inf\left\{x \in \mathbf{R} : \frac{1}{|\mathbf{G}_n|}\sum_{g \in \mathbf{G}_n} I\{T_n(gX^{(n)}) \leq x\} \geq 1-\alpha\right\} . \tag{21}$$

The following theorem describes the asymptotic behavior of naïve permutation test defined in (20) with

$T_n(X^{(n)})$ given by (11) under covariate-adaptive randomization. In particular, it shows that the naïve permutation test, like the two-sample t-test, also has limiting rejection probability under the null hypothesis generally strictly less than the nominal level.

Theorem 4.2. *Suppose Q satisfies Assumption 2.1 and the treatment assignment mechanism satisfies Assumption 2.2. For the problem of testing (4) at level $\alpha \in (0,1)$, $\phi_n^{naïve}(X^{(n)})$ defined in (20) with $T_n(X^{(n)})$ given by (11) satisfies*

$$\limsup_{n \to \infty} E[\phi_n^{naïve}(X^{(n)})] \leq \alpha \tag{22}$$

whenever Q additionally satisfies the null hypothesis, i.e., $\theta(Q) = \theta_0$. Furthermore, the inequality in (22) is strict unless (12) holds.

Remark 4.6. This result essentially follows from Theorem 4.1, which establishes the asymptotic behavior of the two-sample t-statistic, and results in Janssen (1997) and Chung and Romano (2013), which establish the asymptotic behavior of $\hat{c}_n^{\text{naïve}}(1-\alpha)$ defined in (21). ■

Remark 4.7. It may often be the case that $\mathbf{G}_n$ is too large to permit computation of $\hat{c}_n^{\text{naïve}}(1-\alpha)$ defined in (21). In such situations, a stochastic approximation to the test may be used by replacing $\mathbf{G}_n$ with $\hat{\mathbf{G}}_n = \{g_1, \ldots, g_B\}$, where g_1 equals the identity permutation and $g_2, \ldots, g_B$ are i.i.d. Uniform$(\mathbf{G}_n)$. Theorem 4.2 remains true with such an approximation provided that $B \to \infty$ as $n \to \infty$. ■

Remark 4.8. While Theorem 4.2 characterizes when the limiting rejection probability of the naïve permutation test under the null hypothesis is strictly less than the nominal level, it does not reveal how significant this difference might be. In our simulation study in Section 5, we find that, like the two-sample t-test studied in the previous section, the rejection probability may in fact be dramatically less than the nominal level and that this difference translates into substantial power losses when compared with non-conservative tests studied in Sections 4.3 and 4.4. ■

4.3 Covariate-Adaptive Permutation Test

It follows from Theorems 4.1-4.2 and Remark 4.2 that the two-sample t-test and naïve permutation test are conservative in the sense that their limiting rejection probability under the null hypothesis is generally strictly less than the nominal level. As explained in Remarks 4.5 and 4.8, the finite-sample rejection probability may in fact be dramatically less than the nominal level. In this section, we propose a modified version of the permutation test, which we term the covariate-adaptive permutation test, that is not conservative in this way.

In order to define the test, we require some further notation. Define

$$\mathbf{G}_n(S^{(n)}) = \{g \in \mathbf{G}_n : S_{g(i)} = S_i \text{ for all } 1 \leq i \leq n\} , \tag{23}$$

i.e., $\mathbf{G}_n(S^{(n)})$ is the subgroup of permutations of n elements that only permutes indices within strata. Define the action of $g \in \mathbf{G}_n(S^{(n)})$ on $X^{(n)}$ as before. For a given choice of test statistic $T_n(X^{(n)})$, the covariate-adaptive permutation test is given by

$$\phi_n^{\text{cap}}(X^{(n)}) = I\{T_n(X^{(n)}) > \hat{c}_n^{\text{cap}}(1-\alpha)\} \ , \tag{24}$$

where

$$\hat{c}_n^{\text{cap}}(1-\alpha) = \inf\left\{x \in \mathbf{R} : \frac{1}{|\mathbf{G}_n(S^{(n)})|} \sum_{g \in \mathbf{G}_n(S^{(n)})} I\{T_n(gX^{(n)}) \le x\} \ge 1-\alpha\right\} \ . \tag{25}$$

The following theorem describes the asymptotic behavior of the covariate-adaptive permutation test defined in (24) with $T_n(X^{(n)})$ given by (11) under covariate-adaptive randomization. In particular, it shows that the limiting rejection probability of the proposed test under the null hypothesis equals the nominal level. As a result of this, we show in our simulations that the test has dramatically greater power than either the two-sample t-test or the naïve permutation test. In comparison with our preceding results, the theorem further requires that $\tau(s) = 0$ for all $s \in \mathcal{S}$, but, as explained in Section 3, this property holds for a wide variety of treatment assignment mechanisms, including biased-coin designs, stratified block randomization, and the minimization method proposed by Hu and Hu (2012).

Theorem 4.3. *Suppose Q satisfies Assumption 2.1 and the treatment assignment mechanism satisfies Assumption 2.2 with $\tau(s) = 0$ for all $s \in \mathcal{S}$. For the problem of testing (4) at level $\alpha \in (0,1)$, $\phi_n^{cap}(X^{(n)})$ defined in (24) with $T_n(X^{(n)})$ given by (11) satisfies*

$$\lim_{n\to\infty} E[\phi_n^{cap}(X^{(n)})] = \alpha$$

whenever Q additionally satisfies the null hypothesis, i.e., $\theta(Q) = \theta_0$.

Remark 4.9. An additional advantage of the covariate-adaptive permutation test is that it satisfies

$$E[\phi_n^{\text{cap}}(X^{(n)})] \le \alpha \tag{26}$$

for any Q such that

$$Y_i(0)|S_i \overset{d}{=} Y_i(1)|S_i \tag{27}$$

and treatment assignment mechanism such that

$$gA^{(n)}|S^{(n)} \overset{d}{=} A^{(n)}|S^{(n)} \text{ for all } g \in \mathbf{G}_n(S^{(n)}) \ . \tag{28}$$

This property clearly holds, for example, for simple random sampling and stratified block randomization. Moreover, if one uses a randomized version of the test, as described in Chapter 15 of Lehmann and Romano (2005), then the inequality in (26) holds with equality. ■

Remark 4.10. For testing the much narrower null hypothesis that (27) holds and for very specific randomization schemes, the use of the test in (24) has been proposed previously. See, for example, Rosenberger and Lachin (2004, Section 7.4), Rosenbaum (2007), and Heckman et al. (2011). Theorem 4.3 asymptotically justifies the use of (24) for testing (4) for a wide variety of treatment assignment mechanisms while retaining this finite-sample validity. The proof of Theorem 4.3 exploits recent developments in the literature on the asymptotic behavior of permutation tests. In particular, we employ a novel coupling construction following the approach put forward by Chung and Romano (2013) in order to show that the test statistic $T_n(X^{(n)})$ in (11) and the group of permutations $\mathbf{G}_n(S^{(n)})$ in (23) satisfy the conditions in Hoeffding (1952). ■

Remark 4.11. As with the naïve permutation test, it may often be the case that $\mathbf{G}_n(S^{(n)})$ is too large to permit computation of $\hat{c}_n^{\text{cap}}(1-\alpha)$ defined in (25). In such situations, a stochastic approximation to the test may be used by replacing $\mathbf{G}_n(S^{(n)})$ with $\hat{\mathbf{G}}_n = \{g_1, \ldots, g_B\}$, where g_1 equals the identity permutation and $g_2, \ldots, g_B$ are i.i.d. Uniform$(\mathbf{G}_n(S^{(n)}))$. Theorem 4.3 remains true with such an approximation provided that $B \to \infty$ as $n \to \infty$. ■

4.4 Linear Regression with Strata Indicators

In this section, we consider using the usual t-test (on the coefficient on treatment assignment) in a linear regression of outcomes on treatment assignment and indicators for each of the strata. As mentioned previously, we refer to this test as the t-test with strata fixed effects. We consider tests with both homoskedasticity-only and heteroskedasticity-robust standard errors. Note that the two-sample t-test studied in Section 4.1 can be viewed as the usual t-test (on the coefficient on treatment assignment) in a linear regression of outcomes on treatment assignment only with heteroksedasticity-robust standard errors. It follows from Theorem 4.1 and Remark 4.2 that such a test is conservative in the sense that the limiting rejection probability under the null hypothesis is generally strictly less than the nominal level. Remarkably, in this section, we show that the addition of strata fixed effects results in a test is not conservative in this way, regardless of whether homoskedasticity-only or heteroskedasticity-robust standard errors are used.

In order to define the test, consider estimation of the equation

$$Y_i = \beta A_i + \sum_{s \in \mathcal{S}} \delta_s I\{S_i = s\} + \epsilon_i \tag{29}$$

by ordinary least squares. Denote by $\hat{\beta}_n$ the resulting estimator of β in (29). Let

$$T_n^{\text{sfe}}(X^{(n)}) = \frac{\sqrt{n}(\hat{\beta}_n - \theta_0)}{\sqrt{\hat{V}_{n,\beta}}} ,$$

where $\hat{V}_{n,\beta}$ equals either the usual homoskedasticity-only or heteroskedasticity-robust standard error for $\hat{\beta}_n$. See (A-59) and (A-61) in the Appendix for exact expressions. Using this notation, the test of interest is given by

$$\phi_n^{\text{sfe}}(X^{(n)}) = I\{|T_n^{\text{sfe}}(X^{(n)})| > z_{1-\frac{\alpha}{2}}\} . \tag{30}$$

The following theorem describes the asymptotic behavior of the proposed test. In particular, it shows that its limiting rejection probability under the null hypothesis equals the nominal level. In the simulation results below, we show that, like the covariate-adaptive permutation test, the test has dramatically greater power than either the two-sample t-test or the naïve permutation test. Note that, in contrast to our preceding result on the covariate-adaptive permutation test, the theorem does not require $\tau(s) = 0$ for all $s \in \mathcal{S}$. On the other hand, the t-test with strata fixed effects studied here does not share with the covariate-adaptive permutation test the finite-sample validity explained in Remark 4.9.

Theorem 4.4. *Suppose Q satisfies Assumption 2.1 and the treatment assignment mechanism satisfies Assumption 2.2. Then,*

$$\sqrt{n}(\hat{\beta}_n - \theta(Q)) \xrightarrow{d} N(0, \sigma^2_{sfe}) \ .$$

Furthermore,

$$\hat{V}_{n,\beta} \xrightarrow{P} \sigma^2_{sfe} \ ,$$

where $\hat{V}_{n,\beta}$ equals either the usual homoskedasticity-only or heteroskedasticity-robust standard error for $\hat{\beta}_n$. Thus, for the problem of testing (4) at level $\alpha \in (0,1)$, $\phi_n^{sfe}(X^{(n)})$ defined in (30) with either choice of $\hat{V}_{n,\beta}$ satisfies

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

for Q additionally satisfying the null hypothesis, i.e., $\theta(Q) = \theta_0$.

Remark 4.12. In the proof of Theorem 4.4 in the Appendix, it is shown that

$$\sigma^2_{\text{sfe}} = \sigma^2_{\tilde{Y}} + \sigma^2_H \ , \tag{31}$$

where $\sigma^2_{\tilde{Y}}$ and σ^2_H are defined as in (17) and (18), respectively. Remarkably, from (31), we see that variation due to "imperfectly balanced" treatment assignment, i.e., $\sigma^2_D(s) > 0$ in Assumption 2.2, does not contribute to the variance σ^2_{sfe}. It is for this reason that Theorem 4.4, as opposed to Theorem 4.3, does not require $\tau(s) = 0$ for all $s \in \mathcal{S}$ and so $\phi_n^{\text{sfe}}(X^{(n)})$ remains valid for randomization schemes like adaptive biased-coin designs; see Example 3.3. ■

Remark 4.13. Imbens and Rubin (2015, Ch. 9.6) examine the limit in probability of $\hat{\beta}_n$ under a specific randomization assignment, namely, stratified block randomization; see Example 3.4. In contrast to our results, they do not impose the requirement that $n_1(s)$ is chosen to achieve "balance" as, for example, in (8). As a result, Assumption 2.2.(b)-(c) do not necessarily hold, and they conclude that $\hat{\beta}_n$ is generally not consistent for the average treatment effect, $\theta(Q)$. By exploiting Assumption 2.2.(b)-(c), we not only conclude that $\hat{\beta}_n$ is consistent for $\theta(Q)$, but the test $\phi_n^{\text{sfe}}(X^{(n)})$ has limiting rejection probability under the null hypothesis equal to the nominal level. Importantly, Imbens and Rubin (2015) do not include results on $\phi_n^{\text{sfe}}(X^{(n)})$. Note that the required arguments are involved due to $A^{(n)}$ not being i.i.d., relying in particular on non-standard convergence results, such as Lemma B.2 in the Appendix. ■

Remark 4.14. As in the literature on linear panel data models with fixed effects, $\hat{\beta}_n$ may be equivalently computed using ordinary least squares and the deviations of Y_i and A_i from their respective means within strata. However, it is important to note that the resulting standard errors are not equivalent to the standard errors associated with ordinary least squares estimation of (29). We therefore do not recommend computing $\hat{\beta}_n$ using the deviations of Y_i and A_i from their respective means within strata. ■

Remark 4.15. It is important to point out that the asymptotic validity of neither the covariate-adaptive permutation test nor the t-test with strata fixed effects discussed in this section rely on a particular model of (potential) outcomes. In the simulations below, we see that when such additional information is available, it may be possible to exploit it to devise even more powerful methods (e.g., linear regression of outcomes on treatment assignment and covariates). However, these methods may perform quite poorly when this information is incorrect. ■

5 Simulation Study

In this section, we examine the finite-sample performance of several different tests of (4), including those introduced in Section 4, with a simulation study. For $a \in \{0,1\}$ and $1 \leq i \leq n$, potential outcomes are generated in the simulation study according to the equation:

$$Y_i(a) = \mu_a + m_a(Z_i) + \sigma_a(Z_i)\epsilon_{a,i} \ . \tag{32}$$

where μ_a, $m_a(Z_i)$, $\sigma_a(Z_i)$, and $\epsilon_{a,i}$ are specified as follows. In each of the following specifications, $n = 100$ and $\{(Z_i, \epsilon_{0,i}, \epsilon_{1,i}) : 1 \leq i \leq n\}$ are i.i.d. and $(Z_i, \epsilon_{0,i}, \epsilon_{1,i})$ are mutually independent.

Model 1: $Z_i \sim \text{Beta}(2,2)$ (re-centered and re-scaled 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 $\epsilon_{1,i} \sim N(0,1)$; $m_0(Z_i) = m_1(Z_i) = \gamma Z_i$. Note that 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 $m_0(Z_i) = m_1(Z_i) = \sin(\gamma Z_i)$.

Model 3: As in Model 2, but $m_1(Z_i) = \sin(\gamma Z_i) + \sqrt{Z_i + 2.25}$.

Model 4: As in Model 3, but $\sigma_0(Z_i) = Z_i^2$ and $\sigma_1(Z_i) = Z_i^2\sigma_1$.

Model 5: As in Model 4, but $\epsilon_{0,i} \sim \text{Pareto}(1,2)$ and $\epsilon_{1,i} \sim \text{Pareto}(1,2)$ (both re-centered to have mean zero); $Z_i \sim \text{Uniform}(-2,2)$; and

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

For each of the above specifications of $m_a(Z_i)$, $\sigma_a(Z_i)$, and $\epsilon_{i,a}$, we consider both $(\gamma, \sigma_1) = (2,1)$ and $(\gamma, \sigma_1) = (4, \sqrt{2})$. For each resulting specifications, we additionally consider both $(\mu_0, \mu_1) = (0,0)$ (i.e., under the null hypothesis) and $(\mu_0, \mu_1) = (0, \frac{1}{2})$ (i.e., under the alternative hypothesis).

Treatment assignment is generated according to one of the four different covariate-adaptive randomization schemes. In each of the schemes, strata are determined by dividing the support of Z_i (which is a closed interval in all specifications) into ten intervals of equal length and having $S(Z_i)$ be the function that returns the interval in which Z_i lies. In particular, $|\mathcal{S}| = 10$ in all specifications.

SRS: Treatment assignment is generated as in Example 3.1.

BCD: Treatment assignment is generated as in Example 3.2 with $\pi = \frac{2}{3}$.

WEI: Treatment assignment is generated as in Example 3.3 with $\phi(x) = 1 - x^2$.

SBR: Treatment assignment is generated as in Example 3.4 with blocks of size $\lfloor \frac{n(s)}{2} \rfloor$.

In all cases, observed outcomes Y_i are generated according to (1).

In the simulation results below, we consider the following five different tests:

***t*-test**: The usual two-sample t-test studied in Section 4.1.

Naïve: The naïve permutation test studied in Section 4.2.

Reg: The usual t-test (on the coefficient on treatment assignment) in a linear regression of outcomes Y_i on treatment assignment A_i and covariates Z_i using heteroskedasticity-robust standard errors.

SYZ: The bootstrap-based test proposed by Shao et al. (2010).

CAP: The covariate-adaptive permutation test studied in Section 4.3.

SFE: The t-test with strata fixed effects studied in Section 4.4. In this case, we consider both homoskedasticity-only and heteroskedasticity-robust standard errors.

In all cases, rejection probabilities are computed using 10^4 replications.

Table 1 displays the results of the simulation study for $(\gamma, \sigma_1) = (2, 1)$ and Table 2 displays the results of the simulation study for $(\gamma, \sigma_1) = (4, \sqrt{2})$. In the 'SFE' column in both tables, the first number corresponds to homoskedasticity-only standard errors and the second number corresponds to the heteroskedasticity-robust standard errors. We organize our discussion of the results by test:

***t*-test**: As expected in light of Theorem 4.1 and Remark 4.2, we see the two-sample t-test has rejection probability under the null hypothesis very close to the nominal level under simple random sampling, but has rejection probability under the null hypothesis strictly less than the nominal level under the more complicated randomization schemes. Indeed, in some instances, the rejection probability under the null hypothesis is close to zero. Moreover, for all specifications, the two-sample t-test has nearly the lowest rejection probability under the alternative hypothesis. Remarkably, this difference in power is pronounced even under simple random sampling.

Naïve: The results for the naïve permutation test are very similar to those for the two-sample t-test.

		Rejection rate under null - $\theta = 0$						Rejection rate under alternative - $\theta = 1/2$					
Model	CAR	t-test	Naïve	Reg	SYZ	CAP	SFE	t-test	Naïve	Reg	SYZ	CAP	SFE
1	SRS	5.39	4.81	5.45	4.92	5.04	5.26/5.07	20.03	18.86	68.18	19.08	44.94	61.73/61.38
	WEI	0.69	0.57	5.01	5.39	4.98	5.13/5.32	13.25	11.94	69.71	35.7	50.61	65.46/64.02
	BCD	0.90	0.68	4.94	6.05	5.00	4.99/4.90	13.08	11.63	69.50	35.87	50.02	65.26/64.96
	PB	0.01	0.01	5.20	4.93	5.05	5.04/5.13	5.30	4.42	70.10	59.49	61.45	66.47/66.14
2	SRS	5.69	5.03	5.20	5.11	5.01	5.29/5.49	53.80	51.97	55.13	52.36	57.71	62.65/63.85
	WEI	3.05	2.65	3.23	5.91	4.51	4.88/5.15	55.12	53.00	56.64	65.17	63.32	66.63/66.12
	BCD	3.08	2.65	3.28	5.89	5.00	4.96/4.90	54.99	52.84	56.16	65.19	63.59	65.94/67.13
	PB	2.19	1.95	2.37	5.63	4.75	4.57/5.22	55.95	53.58	57.12	70.71	67.73	67.81/68.09
3	SRS	5.43	4.80	5.05	4.98	5.03	4.98/5.42	49.32	47.20	54.15	47.89	55.05	62.15/63.43
	WEI	2.74	2.41	3.51	6.10	4.85	5.02/4.69	49.92	47.70	55.06	61.72	59.93	64.22/65.79
	BCD	2.56	2.22	3.39	5.69	4.81	4.90/5.16	48.80	46.47	54.43	61.33	60.76	64.70/65.65
	PB	1.66	1.45	2.40	6.24	4.98	4.85/4.83	50.08	47.95	55.85	68.92	65.39	66.44/66.01
4	SRS	5.18	4.51	5.09	4.96	4.77	5.80/5.03	34.47	32.59	37.80	32.89	39.11	46.90/41.97
	WEI	3.54	3.03	4.08	6.67	5.18	5.83/5.25	33.92	32.21	36.93	42.79	40.99	46.43/44.45
	BCD	3.17	2.80	3.91	5.78	4.73	5.84/5.07	33.74	31.91	37.31	42.09	40.68	46.80/44.69
	PB	2.91	2.52	3.75	7.29	5.06	6.06/5.88	33.66	31.69	36.99	47.52	42.20	46.26/45.20
5	SRS	4.36	3.96	3.93	3.94	4.90	3.13/3.62	14.50	13.54	13.93	13.78	19.55	19.02/18.63
	WEI	2.41	2.20	2.11	4.65	4.87	3.51/3.04	11.54	10.86	10.88	18.51	21.09	20.12/18.48
	BCD	2.17	1.91	1.85	4.42	4.95	3.37/3.39	12.23	11.22	11.20	18.26	20.96	20.13/19.27
	PB	1.27	1.00	1.05	4.20	4.85	3.29/3.22	9.52	8.98	8.88	21.91	22.90	19.36/19.51

Table 1: Parameter values: $\gamma = 2$, $\sigma_1 = 1$.

		Rejection rate under null - $\theta = 0$						Rejection rate under alternative - $\theta = 1/2$					
Model	CAR	t-test	Naïve	Reg	SYZ	CAP	SFE	t-test	Naïve	Reg	SYZ	CAP	SFE
1	SRS	5.89	5.27	5.13	5.37	5.30	5.26/4.86	9.73	8.81	52.02	9.11	31.81	42.70/41.80
	WEI	0.37	0.31	5.15	5.46	4.93	5.04/5.20	2.20	1.80	51.55	15.03	32.97	43.78/44.08
	BCD	0.56	0.50	4.69	5.22	4.61	4.63/4.99	2.32	1.98	52.57	15.20	33.54	44.43/43.99
	PB	0.00	0.00	5.21	3.62	5.09	4.98/5.27	0.01	0.01	51.95	30.43	39.08	45.56/46.90
2	SRS	5.65	5.14	5.04	5.24	4.96	4.96/4.73	43.22	41.42	41.47	41.97	43.99	46.52/45.18
	WEI	4.24	3.78	3.83	6.43	5.26	5.22/4.43	42.83	40.78	40.74	48.89	46.38	47.76/47.32
	BCD	4.16	3.63	3.65	6.08	5.20	5.04/4.88	42.84	40.57	40.63	49.16	47.04	47.14/48.42
	PB	3.42	3.22	3.10	6.57	5.01	5.26/5.35	42.82	41.82	40.81	53.62	49.16	49.27/49.96
3	SRS	5.51	4.85	4.89	5.04	5.03	5.06/5.36	42.17	40.45	40.78	40.81	42.93	45.06/46.27
	WEI	4.23	3.73	3.86	6.20	5.20	5.02/5.13	42.26	40.21	40.86	48.50	46.02	47.78/48.51
	BCD	3.73	3.33	3.39	5.67	4.85	4.77/5.14	41.63	39.78	40.44	47.85	45.86	47.37/48.10
	PB	3.21	2.96	2.83	6.31	4.81	4.93/5.31	42.12	41.14	40.58	53.48	48.86	48.98/49.10
4	SRS	5.44	4.66	4.98	4.79	5.27	5.95/5.09	28.54	26.91	28.06	27.48	29.92	34.38/31.53
	WEI	4.07	3.51	3.77	5.74	4.74	5.36/5.43	27.68	26.02	27.27	31.76	30.26	34.50/33.57
	BCD	4.19	3.65	3.95	6.01	4.86	5.70/4.96	27.97	26.49	27.44	32.25	30.86	34.52/33.30
	PB	3.80	3.58	3.58	7.53	5.00	6.23/6.77	27.53	26.51	26.68	36.58	31.13	33.65/34.91
5	SRS	5.04	4.62	4.69	4.67	5.74	4.93/4.92	7.31	6.58	6.75	6.79	8.94	7.24/7.85
	WEI	1.93	1.72	1.81	5.72	6.05	5.02/5.21	3.34	2.93	3.04	8.13	9.20	7.62/8.01
	BCD	1.61	1.40	1.47	5.36	6.12	4.72/4.82	3.36	2.89	2.96	8.08	9.80	7.71/7.54
	PB	0.70	0.58	0.55	5.71	6.37	5.03/5.10	1.59	1.53	1.41	9.39	9.84	7.00/8.12

Table 2: Parameter values: $\gamma = 4$, $\sigma_1 = \sqrt{2}$

Reg: The usual t-test (on the coefficient on treatment assignment) in a linear regression of outcomes Y_i on treatment assignment A_i and covariates Z_i using heteroskedasticity-robust standard errors has rejection probability under the null hypothesis very close to the nominal level for Model 1, i.e., when the linear regression is correctly specified. Interestingly, even though the linear regression is incorrectly specified for all other models, the rejection probability of the test under the null hypothesis never exceeds the nominal level, though it is frequently much less than the nominal level. Not surprisingly, for Model 1, the test also has the highest rejection probability under the alternative hypothesis. For all other models, the rejection probability of the test under the alternative hypothesis is lower than that of some of the other tests considered below.

SYZ: For most specifications, the bootstrap-based test proposed by Shao et al. (2010) has rejection probability under the null hypothesis very close to the nominal level, though in some instances the rejection probability under the null hypothesis mildly exceeds the nominal level (e.g., 7.53% under Model 4 and stratified block randomization with $\gamma = 4$ and $\sigma_1 = \sqrt{2}$). Its rejection probability under the alternative hypothesis is often considerably lower than that of the other tests considered below. Recall, however, that Shao et al. (2010) only justify the use of this test for biased-coin designs.

CAP: As expected in light of Theorem 4.3, the covariate-adaptive permutation test has rejection probability under the null hypothesis very close to the nominal level in all specifications. Indeed, among all the tests considered here, it arguably has rejection probability under the null hypothesis closest to the nominal level across all specifications. As explained in Remark 4.9, its rejection probability under the null hypothesis even equals the nominal level in finite-samples for some specifications. Furthermore, its rejection probability under the alternative hypothesis typically exceeds that of all the tests considered previously and often by a considerably margin. On the other hand, its rejection probability under the alternative hypothesis is typically less than that of the following test.

SFE: As expected in light of Theorem 4.4, the t-test with strata fixed effects has rejection probability under the null hypothesis very close to the nominal level in nearly all specifications. The only specification for which this is not true is Model 5 with $\gamma = 2$ and $\sigma_1 = 1$, in which case the test has rejection probability under the null hypothesis mildly less than the nominal level. Its rejection probability under the alternative hypothesis typically exceeds that of all the tests considered previously and often by a considerable margin. Note that the results using homoskedasticity-only and heteroskedasticity-robust standard errors are nearly identical.

Appendix A Proof of the main results

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 $a \in \{0,1\}$, $E[Y_i(a)]$
$\tilde{Y}_i(a)$	For $a \in \{0,1\}$, $Y_i(a) - E[Y_i(a)\|S_i]$
$m_a(Z_i)$	For $a \in \{0,1\}$, $E[Y_i(a)\|Z_i] - \mu_a$
σ^2_Y	$2(\sigma^2_{Y(1)} + \sigma^2_{Y(0)})$
$\sigma^2_{\tilde{Y}}$	$2(\sigma^2_{\tilde{Y}(1)} + \sigma^2_{\tilde{Y}(0)})$
σ^2_H	$\sum_{s\in\mathcal{S}} p(s)(E[m_1(Z)\|S=s] - E[m_0(Z)\|S=s])^2$
σ^2_A	$\sum_{s\in\mathcal{S}} \sigma^2_D(s)(E[m_1(Z)\|S=s] + E[m_0(Z)\|S=s])^2$
$n(s)$	Number of individuals in strata $s \in \mathcal{S}$
$n_1(s)$	Number of individuals in the treatment group in strata $s \in \mathcal{S}$

Table 3: Useful notation

A.1 Proof of Theorem 4.1

We start the proof by showing that the numerator of the root in the statement of the theorem satisfies

$$\sqrt{n}\left(\bar{Y}_{n,1} - \bar{Y}_{n,0} - \theta(Q)\right) \xrightarrow{d} N(0, \sigma^2_{\tilde{Y}} + \sigma^2_H + \sigma^2_A)\ . \tag{A-33}$$

Consider the following algebraic derivation,

$$\begin{aligned}
\sqrt{n}(\bar{Y}_{n,1} - \bar{Y}_{n,0} - \theta(Q)) &= \sqrt{n}(\frac{1}{n_1}\sum_{i=1}^{n}(Y_i(1)-\mu_1)A_i - \frac{1}{n_0}\sum_{i=1}^{n}(Y_i(0)-\mu_0)(1-A_i)) \\
&= R^*_{n,1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[(1-\frac{D_n}{n})(Y_i(1)-\mu_1)A_i - (1+\frac{D_n}{n})(Y_i(0)-\mu_0)(1-A_i)\right] \\
&= R^*_{n,1}(R^*_{n,2} + R^*_{n,3})\ ,
\end{aligned}$$

where we used $\mu_a = E[Y_i(a)]$, $n = n_0 + n_1$, $n/n_1 = 2(D_n/n+1)^{-1}$, $D_n = \sum_{s\in\mathcal{S}} D_n(s)$, and the following definitions,

$$\begin{aligned}
R^*_{n,1} &\equiv \frac{2}{1-(D_n/n)^2}\ , \\
R^*_{n,2} &\equiv \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[(Y_i(1)-\mu_1)A_i - (Y_i(0)-\mu_0)(1-A_i)\right]\ , \\
R^*_{n,3} &\equiv -\frac{D_n}{\sqrt{n}}\frac{1}{n}\sum_{i=1}^{n}\left[(Y_i(1)-\mu_1)A_i + (Y_i(0)-\mu_0)(1-A_i)\right]\ .
\end{aligned}$$

By Assumption 2.2.(c), $D_n/n \xrightarrow{p} 0$ and this implies $R^*_{n,1} \xrightarrow{p} 2$. Lemma B.1 implies $R^*_{n,2} \xrightarrow{d} N(0, (\sigma^2_{\tilde{Y}} + \sigma^2_H + \sigma^2_A)/4)$. Lemma B.3 implies $R^*_{n,3} \xrightarrow{p} 0$. Elementary properties of stochastic convergence complete the proof of this first step.

We next prove that

$$\sqrt{n}\sqrt{\frac{\hat{\sigma}^2_{n,1}}{n_1} + \frac{\hat{\sigma}^2_{n,0}}{n_0}} \xrightarrow{p} \sqrt{2(\sigma^2_{Y(1)} + \sigma^2_{Y(0)})}\ . \tag{A-34}$$

This follows from showing that $n\hat{\sigma}^2_{n,a}/n_a \xrightarrow{p} 2\sigma^2_{Y(a)}$ for $a \in \{0,1\}$. We only show $a=1$, as the result for $a=0$ is analogous. Start by writing $\bar{Y}_{n,1}$ as follows,

$$\bar{Y}_{n,1} \equiv \frac{1}{n_1}\sum_{i=1}^{n} A_i Y_i = \mu_1 + \frac{n}{n_1}\frac{1}{n}\sum_{i=1}^{n} A_i(Y_i(1) - \mu_1)\ . \tag{A-35}$$

Then consider the following derivation,

$$\begin{aligned} \frac{n\hat{\sigma}^2_{n,1}}{n_1} &= \frac{n}{n_1}\frac{1}{n_1}\sum_{i=1}^{n}(Y_i - \bar{Y}_{n,1})^2 A_i = \frac{n}{n_1}\frac{1}{n_1}\sum_{i=1}^{n}(\mu_1 - \bar{Y}_{n,1} + Y_i(1) - \mu_1)^2 A_i \\ &= \frac{n}{n_1}\left\{\frac{n}{n_1}\frac{1}{n}\sum_{i=1}^{n}[Y_i(1) - \mu_1]^2 A_i - (\mu_1 - \bar{Y}_{n,1})^2\right\} = \left(\frac{n}{n_1}\right)^2 R^{\star}_{n,4} - \left(\frac{n}{n_1}\right)^3 R^{\star 2}_{n,5}\ , \end{aligned}$$

where we used (A-35) and the following definitions,

$$\begin{aligned} R^{\star}_{n,4} &\equiv \frac{1}{n}\sum_{i=1}^{n}[Y_i(1) - \mu_1]^2 A_i\ , \\ R^{\star}_{n,5} &\equiv \frac{1}{n}\sum_{i=1}^{n}(Y_i(1) - \mu_1) A_i\ . \end{aligned}$$

Since $n/n_1 = 2(D_n/n + 1)^{-1}$ and $D_n/n \xrightarrow{p} 0$ by Assumption 2.2.(c), it follows that $n/n_1 \xrightarrow{p} 2$. The result follows from showing that $R^{\star}_{n,4} \xrightarrow{p} \frac{1}{2}\sigma^2_{Y(1)}$ and $R^{\star}_{n,5} \xrightarrow{p} 0$.

Start with $R^{\star}_{n,4}$. By Assumption 2.2.(a) and $W^{(n)}$ consisting of n i.i.d. observations we have that, conditionally on $\{S^{(n)}, A^{(n)}\}$, $\{A_i(Y_i(1) - \mu_1)^2 : 1 \le i \le n\}$ is an independent sequence. It follows that

$$\begin{aligned} E[R^{\star}_{n,4}|S^{(n)}, A^{(n)}] &= \frac{1}{2n}\sum_{i=1}^{n}(A_i^* + 1)E[(Y_i(1) - \mu_1)^2|S_i] \\ &= \frac{1}{2}\sum_{s\in\mathcal{S}} E[(Y_i(1) - \mu_1)^2|S = s]\left(\frac{D_n(s)}{n} + \frac{n(s)}{n}\right) \\ &\xrightarrow{p} \frac{1}{2}\sum_{s\in\mathcal{S}} E[(Y_i(1) - \mu_1)^2|S = s]p(s) \\ &= \frac{1}{2}\sigma^2_{Y(1)}\ , \end{aligned} \tag{A-36}$$

where we used Assumption 2.2.(a), $A_i = (1 + A^*)/2$, $D_n(s)/n \xrightarrow{p} 0$ and $n(s)/n \xrightarrow{p} p(s)$ for all $s \in \mathcal{S}$. Then, for any $\varepsilon > 0$,

$$\begin{aligned} &P\left\{|R^{\star}_{n,4} - E[R^{\star}_{n,4}|S^{(n)}, A^{(n)}]| > \varepsilon|S^{(n)}, A^{(n)}\right\}\varepsilon^{1+\delta/2} \\ &\le E\left[\left|\frac{1}{n}\sum_{i=1}^{n} A_i(Y_i(1) - \mu_1)^2 - E\left[\frac{1}{n}\sum_{i=1}^{n} A_i(Y_i(1) - \mu_1)^2\right| S^{(n)}, A^{(n)}\right]\right|^{1+\delta/2}\left| S^{(n)}, A^{(n)}\right] \to 0\ , \end{aligned} \tag{A-37}$$

where the convergence is the result of $E[|A_i(Y_i(1) - \mu_1)|^{2+\delta}|S^{(n)}, A^{(n)}] < \infty$, which follows from Assumption 2.1, and Lemma B.5. By definition, (A-37) implies that,

$$R^{\star}_{n,4} - E\left[R^{\star}_{n,4}\right| S^{(n)}, A^{(n)}\left.\right] \xrightarrow{p} 0\ . \tag{A-38}$$

By combining (A-36) and (A-38) we can conclude that $R^{\star}_{n,4} \xrightarrow{p} \frac{1}{2}\sigma^2_{Y(1)}$.

The arguments for $R^{\star}_{n,5}$ are analogous. In fact, replacing $A_i(Y_i(1) - \mu_1)^2$ with $A_i(Y_i(1) - \mu_1)$ in the derivations leading to (A-36), (A-37) and (A-38), results in $R^{\star}_{n,5} \xrightarrow{p} E[(Y_i(1) - \mu_1)]/2 = 0$.

To prove that $\sigma^2_{\tilde{Y}} + \sigma^2_H + \sigma^2_A \leq \sigma^2_Y$ holds with strict inequality unless (12) holds, notice that for $a \in \{0, 1\}$,

$$\sigma^2_{\tilde{Y}(a)} = \sigma^2_{Y(a)} - \sum_{s \in \mathcal{S}} E[(Y(1) - \mu_1)|S = s]^2 p(s) = \sigma^2_{Y(a)} - \sum_{s \in \mathcal{S}} E[m_a(Z_i)|S = s]^2 p(s) \ . \tag{A-39}$$

Using (A-39) and some algebra shows that,

$$\begin{aligned} \sigma^2_Y - \sigma^2_{\tilde{Y}} - \sigma^2_H - \sigma^2_A &= 2(\sigma^2_{Y(1)} + \sigma^2_{Y(0)} - \sigma^2_{\tilde{Y}(1)} - \sigma^2_{\tilde{Y}(0)}) \\ &\quad - \sum_{s \in \mathcal{S}} p(s)[E[m_1(Z)|S = s] - E[m_0(Z)|S = s]]^2 \\ &\quad - \sum_{s \in \mathcal{S}} \sigma^2_D(s)[E[m_1(Z)|S = s] + E[m_0(Z)|S = s]]^2 \\ &= \sum_{s \in \mathcal{S}} p(s) \, [1 - \tau(s)] \, [E[m_1(Z)|S = s] + E[m_0(Z)|S = s]]^2 \ , \end{aligned}$$

where, by Assumption 2.2.(c), $\sigma^2_D(s) = p(s)\tau(s)$ with $\tau(s) \in [0, 1]$. The RHS is non-negative and it is zero if and only if (12) holds, as required.

A.2 Proof of Theorem 4.3

We divide the proof in two parts. The first part shows that our test statistic and group of permutations satisfies the so-called Hoeffding's condition, cf. Lehmann and Romano (2005, Eq. (15.10)). The second part uses the first one to prove the convergence of the critical value and the conclusion of the theorem.

Part I. Let $G^{(n)}$ and $G^{(n)\prime}$ be two independent random variables that are uniform on $\mathbf{G}_n(S^{(n)})$ and independent of $X^{(n)}$, and let

$$\sigma^2_{\text{cap}} = \frac{\sigma^2_{\tilde{Y}} + \sigma^2_H}{\sigma^2_Y} \ . \tag{A-40}$$

We will show that, for Q such that $\theta(Q) = \theta_0$,

$$(T_n(G^{(n)} X^{(n)}), T_n(G^{(n)\prime} X^{(n)})) \xrightarrow{d} (T, T') \ ,$$

where T and T' are independent with common c.d.f. $\Phi(t/\sigma_{\text{cap}})$.

Step 1: We start the proof of part I by using a coupling construction, following Chung and Romano (2013), that links the random variables in Lemma B.6 with those in $X^{(n)}$ in a way that, except for ordering, most observations in $\{V_j : 1 \leq j \leq n\}$ and $\{Y_i : 1 \leq i \leq n\}$ are identically the same, conditional on $S^{(n)}$. We do this using the following three-step algorithm, where we initiate $\mathcal{Y}_n = \{Y_i : 1 \leq i \leq n\}$:

1. Given $s \in \mathcal{S}$, draw an index $C_j \in \{0, 1\}$ such that $P\{C_j = 1\} = 1/2$.
2. Conditional on $C_j = a$, set $V_j = Y_i - \theta_0 A_i$ for $Y_i \in \mathcal{Y}_n$ such that $A_i = a$ and $S_i = s$, if one such observation is available, and remove Y_i from $\mathcal{Y}_n$. If there is no such Y_i in $\mathcal{Y}_n$, draw a new independent observation from $Q_a(s)$ and set it equal to V_j.
3. Repeat steps 1-2 $n(s)$ times for each $s \in \mathcal{S}$.

The algorithm above generates random variables $V^{(n)} = \{V_j : 1 \leq j \leq n\}$ using observation from $Y^{(n)} = \{Y_i : 1 \leq i \leq n\}$, when possible, while making sure that the random variables generated in this way have the properties in

Lemma B.6. In addition, by construction there exists $g_0 \in \mathbf{G}_n(S^{(n)})$ such that $g_0 V^{(n)}$ has the elements in common with $Y^{(n)}$ in exactly the same position. In fact, if we denote by K_n the number of observations in $g_0 V^{(n)}$ and $Y^{(n)}$ that differ (which equals the random number of independently generated variables from $Q_a(s)$, $a \in \{0,1\}$, in step 2), it follows that

$$K_n = \sum_{i=1}^{n} I\{V_{g_0(i)} \neq Y_i - \theta_0 A_i\} = O_p(n^{1/2}) \ . \tag{A-41}$$

A proof of the above result follows similar arguments to those in Chung and Romano (2013) so we omit it here. We next divide the proof in two steps.

Step 2: we now prove that for Q such that $\theta(Q) = \theta_0$ and

$$T_n^U(X^{(n)}) \equiv \sqrt{n}(\frac{1}{n_1}\sum_{i=1}^{n} Y_i A_i - \frac{1}{n_0}\sum_{i=1}^{n} Y_i(1-A_i) - \theta(Q)) \ , \tag{A-42}$$

it follows that

$$(T_n^U(G^{(n)}X^{(n)}), T_n^U(G^{(n)\prime}X^{(n)})) \xrightarrow{d} (T^U, T^{U\prime}) \ , \tag{A-43}$$

where T^U and $T^{U\prime}$ are independent with common distribution given by $N(0, \sigma^2_{\tilde{Y}} + \sigma^2_H)$. We do this using a similar approach to that in the proof of Lemma 5.1 in Chung and Romano (2013). This requires verifying two conditions, together with an application of Slutsky's theorem. The first condition requires us to show that

$$(T_n^U(G^{(n)}\tilde{X}_n), T_n^U(G^{(n)\prime}\tilde{X}_n)) \xrightarrow{d} (T^U, T^{U\prime}) \ , \tag{A-44}$$

where $\tilde{X}^{(n)} = \{(V_i, A_i, Z_i) : 1 \leq i \leq n\}$. The second condition requires us to show that, for any $g \in \mathbf{G}_n(S^{(n)})$,

$$T_n^U(gg_0\tilde{X}_n) - T_n^U(gX^{(n)}) = o_p(1) \ . \tag{A-45}$$

Given (A-44) and (A-45), (A-43) immediately follows from Slutsky's theorem.

To verify (A-44), let $A_i^* = 2A_i - 1$ and let $g(\cdot)$ and $g'(\cdot)$ be the two random permutations associated with $G^{(n)}$ and $G^{(n)\prime}$, respectively. By (B-77) and the fact that permuting $\tilde{X}_n$ is equivalent to permuting $\{A_i^* : 1 \leq i \leq n\}$, we get

$$(T_n^U(G^{(n)}\tilde{X}_n), T_n^U(G^{(n)\prime}\tilde{X}_n)) = \left(\frac{2}{\sqrt{n}}\sum_{i=1}^{n} V_i A^*_{g(i)}, \frac{2}{\sqrt{n}}\sum_{i=1}^{n} V_i A^*_{g'(i)}\right) + o_p(1) \ .$$

By Lemma B.7 and the Cramer-Wold device, it suffices to show that for any $a \in \mathbf{R}$ and $b \in \mathbf{R}$,

$$\frac{2}{\sqrt{n}}\sum_{i=1}^{n} V_i(aA^*_{g(i)} + bA^*_{g'(i)}) \xrightarrow{d} \sqrt{a^2+b^2}N(0, \sigma^2_{\tilde{Y}} + \sigma^2_H + \sigma^2_A) \ .$$

In order to do this, note that the left hand side in the above display equals

$$\frac{2}{\sqrt{n}}\sum_{i=1}^{n} V_i^*(aA^*_{g(i)} + bA^*_{g'(i)}) + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(E[m_0(Z_i)|S_i] + E[m_1(Z_i)|S_i])(aA^*_{g(i)} + bA^*_{g'(i)}) \ , \tag{A-46}$$

where $V_i^* = V_i - \frac{1}{2}(E[m_0(Z_i)|S_i]) + E[m_1(Z_i)|S_i])$ has mean zero conditional on $S^{(n)}$. Denote by $\tilde{R}_{n,1}$ and $\tilde{R}_{n,2}$ the first and second term in (A-46), respectively.

Now let $B_i = aA^*_{g(i)} + bA^*_{g'(i)}$. Conditional on $\{S^{(n)}, A^{(n)}\}$, B_i and V_i^* are independent and V_i^* is i.i.d. with conditional mean equal to zero and conditional variance as defined in Lemma B.6. It then follows from Lemma 11.3.3

in Lehmann and Romano (2005) and the proof of Lemma B.7 that

$$\{\tilde{R}_{n,1}|S^{(n)},A^{(n)}\} \xrightarrow{d} \sqrt{a^2+b^2}N(0,\sigma^2_{\tilde{Y}}+\sigma^2_H)\ a.s.\ , \tag{A-47}$$

provided that

$$\left\{ \frac{\max_{1\le i\le n} B_i^2/n}{\frac{1}{n}\sum_{j=1}^n B_j^2} \middle| S^{(n)},A^{(n)} \right\} \xrightarrow{p} 0\ a.s. \tag{A-48}$$

Since conditional $\{S^{(n)},A^{(n)}\}$, $A^*_{g(j)}$ and $A^*_{g'(j)}$ are independent, similar arguments to those in Lehmann and Romano (2005, Eq. (15.16)) show that

$$\frac{1}{n}\sum_{j=1}^n B_j^2 = a^2+b^2+2ab\frac{1}{n}\sum_{j=1}^n A^*_{g(j)}A^*_{g'(j)} \xrightarrow{p} a^2+b^2\ a.s.,$$

where the convergence occurs conditional on $\{S^{(n)},A^{(n)}\}$. Thus, (A-48) follows from $\max_{1\le i\le n} B_i^2 \le (|a|+|b|)^2$.

Next, by the definition of $\mathbf{G}_n(S^{(n)})$, $g(\cdot)$ and $g'(\cdot)$ do not permute observations across strata and therefore

$$D_n(s) = \sum_{i=1}^n A^*_{g(i)}I\{S_i=s\} = \sum_{i=1}^n A^*_{g'(i)}I\{S_i=s\}\ .$$

It follows that

$$\tilde{R}_{n,2} = \sum_{s\in\mathcal{S}}(E[m_0(Z)|S=s]+E[m_1(Z)|S=s])(a+b)\frac{D_n(s)}{\sqrt{n}}$$

and so

$$\{\tilde{R}_{n,2}|S^{(n)}\} \xrightarrow{d} (a+b)N(0,\sigma^2_A)\ a.s.\ ,$$

where the convergence follows from Assumption 2.2.(c). Finally, by similar arguments to those used in the proof of Lemma B.1, $\tilde{R}_{n,1}$ and $\tilde{R}_{n,2}$ are asymptotically independent and so it follows that

$$\frac{2}{\sqrt{n}}\sum_{i=1}^n V_i(aA^*_{g(i)}+bA^*_{g'(i)}) \xrightarrow{d} N(0,(a^2+b^2)(\sigma^2_{\tilde{Y}}+\sigma^2_H+\sigma^2_A)+2ab\,\sigma^2_A)\ .$$

We conclude that the Cramer-Wold condition holds for any $a\in\mathbf{R}$ and $b\in\mathbf{R}$ if and only if $\sigma^2_A=0$, which in turn follows from $\tau(s)=0$ for all $s\in\mathcal{S}$. Condition (A-44) then holds.

To verify (A-45), let $g(\cdot)$ be the random permutation associated with a random $G^{(n)}\in\mathbf{G}_n(S^{(n)})$ and note that

$$T_n^U(G^{(n)}g_0\tilde{X}_n)-T_n^U(G^{(n)}X^{(n)}) = \sqrt{n}\left[\frac{1}{n_1}\sum_{i=1}^n(V_{gg_0(i)}-Y_{g(i)}+\theta(Q))A_i - \frac{1}{n_0}\sum_{i=1}^n(V_{gg_0(i)}-Y_{g(i)})(1-A_i)\right]\ .$$

Due to our coupling construction, all of the terms in the above two sums are zero except for at most K_n of them, where K_n satisfies (A-41). But any nonzero term like $(V_{gg_0(i)}-Y_{g(i)})(1-A_i)$ satisfies,

$$E[(V_{gg_0(i)}-Y_{g(i)})(1-A_i)|K_n,g,A^{(n)}]=0$$
$$\mathrm{Var}[(V_{gg_0(i)}-Y_{g(i)})(1-A_i)|K_n,g,A^{(n)}] \le \max_{a\in\{0,1\}}\mathrm{Var}[Y_1(a)]<\infty\ .$$

It follows that $E[T_n^U(G^{(n)}g_0\tilde{X}_n)-T_n^U(G^{(n)}X^{(n)})|K_n,g,A^{(n)}]=0$, with conditional variance

$$\mathrm{Var}[T_n^U(G^{(n)}g_0\tilde{X}_n)-T_n^U(G^{(n)}X^{(n)})|K_n,g,A^{(n)}] \le \frac{2n}{\min\{n_1^2,n_0^2\}}\max_{a\in\{0,1\}}\mathrm{Var}[Y_1(a)]K_n\ , \tag{A-49}$$

and therefore the unconditional variance is bounded above by

$$O_p(n^{-1}) \max_{a\in\{0,1\}} \mathrm{Var}[Y_1(a)] O_p(n^{1/2}) = O_p(n^{-1/2}) \ . \tag{A-50}$$

By an application of Chebyshev's inequality, (A-45) follows. Finally, by invoking Slutsky's theorem, (A-43) follows and this completes the proof of step 2.

Step 3: note that we can write $T_n(G^{(n)}X^{(n)})$ as

$$T_n(G^{(n)}X^{(n)}) = \frac{1}{T_n^L(G^{(n)}X^{(n)})} T_n^U(G^{(n)}X^{(n)}) \ ,$$

where $T_n^L(X^{(n)})$ is as in (B-80) and $T_n^U(X^{(n)})$ is as in (B-75). By Lemma B.8 and the continuous mapping theorem, we have

$$\frac{1}{T_n^L(\tilde{X}_n)} \xrightarrow{p} \frac{1}{\sqrt{\sigma_Y^2}} \ ,$$

where $\sigma_Y^2 = 2(\sigma_{Y(1)}^2 + \sigma_{Y(0)}^2)$. It follows from Chung and Romano (2013, Lemma 5.3) that

$$\frac{1}{T_n^L(G^{(n)}X^{(n)})} \xrightarrow{p} \frac{1}{\sqrt{\sigma_Y^2}} \ , \tag{A-51}$$

for any random $G^{(n)} \in \mathbf{G}_n(S^{(n)})$. This allows us to deduce the behavior of the statistic T_n^L under the permutation distribution on $X^{(n)}$ from the behavior of T_n^L under the i.i.d. random variables $\tilde{X}_n$. Combining (A-51), (A-43), and Lemma B.9, we conclude that

$$(T_n(G^{(n)}X^{(n)}), T_n(G^{(n)\prime}X^{(n)})) \xrightarrow{d} (T, T') \ , \tag{A-52}$$

where T and T' are independent with common c.d.f. $\Phi(t/\sigma_{\text{cap}})$ and σ_{cap}^2 as in (A-40). The completes the proof of Part I.

Part II. Let $c_n^{\text{cap}}(1-\alpha)$ denote the critical value of the CAP test in (24), defined as

$$\hat{c}_n^{\text{cap}}(1-\alpha) = \inf\left\{ x \in \mathbf{R} : \frac{1}{|G_n(S^{(n)})|} \sum_{g \in \mathbf{G}_n(S^{(n)})} I\{T_n(gX^{(n)}) \le x\} \ge 1-\alpha \right\} \ . \tag{A-53}$$

By the result in Part I and Lehmann and Romano (2005, Theorem 15.2.3), it follows that

$$\hat{c}_n^{\text{cap}}(1-\alpha) \xrightarrow{p} \inf\{t : \Phi(t/\sigma_{\text{cap}}) \ge t\} \ .$$

Combining this last result with Theorem 4.1, we get $\lim_{n\to\infty} E[\phi_n^{\text{cap}}(X^{(n)})] = \alpha$, under the null hypothesis $\theta(Q) = \theta_0$, whenever the treatment assignment mechanism is such that $\tau(s) = 0$ for all $s \in \mathcal{S}$. This completes the proof of the theorem.

A.3 Proof of Theorem 4.4

Part I. Let $\hat{\beta}_n$ be the least squares estimator of β in (29). We first prove that

$$\sqrt{n}(\hat{\beta}_n - \theta(Q)) \xrightarrow{d} N(0, \sigma_{\text{strata}}^2) \ . \tag{A-54}$$

To do this, write $\hat{\beta}_n$ as

$$\hat{\beta}_n = \frac{\sum_{i=1}^n \tilde{A}_i Y_i}{\sum_{i=1}^n \tilde{A}_i^2} ~,$$

where $\tilde{A}_i$ is the projection of A_i on the strata indicators and it equals $\tilde{A}_i = A_i - \frac{n_1(S_i)}{n(S_i)}$, with

$$\frac{n_1(S_i)}{n(S_i)} = \sum_{s \in \mathcal{S}} I\{S_i = s\} \frac{n_1(s)}{n(s)} ~.$$

We prove the result by showing that $\sqrt{n}(\hat{\beta}_n - \theta(Q)) = 2R_{n,1} + 2R_{n,3} + o_p(1)$ where $R_{n,1}$ and $R_{n,3}$ are defined as in (B-64) and (B-66). To this end, consider the following derivation,

$$\begin{aligned}
\sqrt{n}(\hat{\beta}_n - \theta(Q)) &= \frac{\sqrt{n}}{\frac{1}{n}\sum_{i=1}^n \tilde{A}_i^2} \left[\left(\frac{1}{n} \sum_{i=1}^n \tilde{A}_i Y_i \right) - \theta(Q) \left(\frac{1}{n} \sum_{i=1}^n \tilde{A}_i^2 \right) \right] \\
&= \frac{1}{\frac{1}{n}\sum_{i=1}^n \tilde{A}_i^2} \left[\left(\sqrt{n} \left(\frac{1}{n} \sum_{i=1}^n \tilde{A}_i Y_i - \frac{\theta(Q)}{4} \right) \right) - \theta(Q) \left(\sqrt{n} \left(\frac{1}{n} \sum_{i=1}^n \tilde{A}_i^2 - \frac{1}{4} \right) \right) \right] \\
&= \sqrt{n} \left(\frac{1}{n} \sum_{i=1}^n 4\tilde{A}_i Y_i - \theta(Q) \right) + o_p(1) ~,
\end{aligned}$$

where the last step uses $\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n \tilde{A}_i^2 - 1/4\right) = o_p(1)$, which follows from step 1 below.

Step 1: Show that $\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n \tilde{A}_i^2 - 1/4\right) = o_p(1)$. This follows from

$$\begin{aligned}
\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n \tilde{A}_i^2 - 1/4\right) &= \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n \left(A_i - \sum_{s\in\mathcal{S}} I\{S_i = s\}\frac{n_1(s)}{n(s)}\right)^2 - 1/4\right) \\
&= \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n \left(A_i - 2A_i \sum_{s\in\mathcal{S}} I\{S_i = s\}\frac{n_1(s)}{n(s)} + \sum_{s\in\mathcal{S}} I\{S_i = s\}\left(\frac{n_1(s)}{n(s)}\right)^2\right) - 1/4\right) \\
&= \sqrt{n}\frac{1}{n}\sum_{i=1}^n (A_i - 1/2) - \sqrt{n}\sum_{s\in\mathcal{S}} \frac{(n_1(s)/n)^2}{n(s)/n} + \sqrt{n}\frac{1}{4} \\
&= \frac{1}{2}\sqrt{n}\sum_{s\in\mathcal{S}} D_n(s)/n - \sum_{s\in\mathcal{S}} \sqrt{n}\left(n_1(s)/n - p(s)/2\right) + o_p(1) \\
&= \frac{1}{2}\sqrt{n}\left[\sum_{s\in\mathcal{S}} D_n(s)/n - \sum_{s\in\mathcal{S}} \left[\sqrt{n}\left(n(s)/n - p(s)\right) + D_n(s)/n\right]\right] + o_p(1) ~, \qquad \text{(A-55)}
\end{aligned}$$

where the fourth equality follows from

$$\sqrt{n}\left(\frac{(n_1(s)/n)^2}{n(s)/n} - p(s)/4\right) = \sqrt{n}\left(n_1(s)/n - p(s)/2\right) - \sqrt{n}\left((n(s)/n) - p(s)\right)/4 + o_p(1) ~,$$

$\sqrt{n}\left(n_1(s)/n - p(s)/2\right) = O_p(1)$ and $\sqrt{n}\left(n(s)/n - p(s)\right) = O_p(1)$. The result follows from the fact that the term in brackets in (A-55) is numerically equal to zero, after using $\sum_{s\in\mathcal{S}} p(s) = 1$ and $\sum_{s\in\mathcal{S}} n(s) = n$.

Step 2: Show that $\sqrt{n}(\frac{1}{n}\sum_{i=1}^n 4\tilde{A}_i Y_i - \theta(Q)) = 2R_{n,1} + 2R_{n,3} + o_p(1)$. Note that

$$\sqrt{n}(\frac{1}{n}\sum_{i=1}^n 4\tilde{A}_i Y_i - \theta(Q)) = \frac{1}{\sqrt{n}}\sum_{i=1}^n 4A_i Y_i - \sqrt{n}\theta(Q) - \frac{2}{\sqrt{n}}\sum_{s\in\mathcal{S}}\sum_{i=1}^n \frac{2n_1(s)}{n(s)} I\{S_i = s\} Y_i ~. \qquad \text{(A-56)}$$

Consider the first two terms. Use that $2A_i = A_i^* + 1$ and the definition of Y_i to get

$$\begin{aligned}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} 4A_iY_i - \sqrt{n}\theta(Q) &= \frac{2}{\sqrt{n}}\sum_{i=1}^{n} A_i^*Y_i + \frac{2}{\sqrt{n}}\sum_{i=1} Y_i - \sqrt{n}\theta(Q)\\ &= \frac{2}{\sqrt{n}}\sum_{i=1}^{n}[(Y_i(1)-\mu_1)A_i - (Y_i(0)-\mu_0)(1-A_i)] + \frac{2}{\sqrt{n}}\sum_{i=1} Y_i\\ &\quad + \frac{2}{\sqrt{n}}\sum_{i=1}^{n}[A_i\mu_1 - (1-A_i)\mu_0] - \sqrt{n}(\mu_1-\mu_0)\\ &= 2R_{n,1} + 2R_{n,3} + 2R_{n,2} + \frac{2}{\sqrt{n}}\sum_{i=1} Y_i + \frac{2}{\sqrt{n}}\sum_{i=1}^{n}[A_i^*(\mu_1+\mu_0)]\ , \end{aligned} \tag{A-57}$$

where the last equality follows from the proof of Lemma B.1. Now consider the third term in (A-56) and use that $2n_1(s)/n(2) = D_n(s)/n(s) + 1$,

$$\begin{aligned}\frac{2}{\sqrt{n}}\sum_{s\in\mathcal{S}}\sum_{i=1}^{n}\frac{2n_1(s)}{n(s)}I\{S_i=s\}Y_i &= \frac{2}{\sqrt{n}}\sum_{s\in\mathcal{S}}\sum_{i=1}^{n}\frac{D_n(s)}{n(s)}I\{S_i=s\}Y_i + \frac{2}{\sqrt{n}}\sum_{s\in\mathcal{S}}\sum_{i=1}^{n}I\{S_i=s\}Y_i\\ &= \sum_{s\in\mathcal{S}}\frac{D_n(s)}{\sqrt{n}}\frac{2n}{n(s)}\frac{1}{n}\sum_{i=1}^{n}I\{S_i=s\}Y_i + \frac{2}{\sqrt{n}}\sum_{i=1}^{n}Y_i\\ &= \sum_{s\in\mathcal{S}}\frac{D_n(s)}{\sqrt{n}}\frac{2n}{n(s)}\left\{\frac{p(s)}{2}\left[\mu_1+\mu_0+E[m_1(Z)+m_0(Z)|S=s]\right]+o_p(1)\right\}\\ &\quad + \frac{2}{\sqrt{n}}\sum_{i=1}^{n}Y_i\\ &= 2R_{n,2} + \frac{2}{\sqrt{n}}\sum_{i=1}^{n}Y_i + \frac{2}{\sqrt{n}}\sum_{i=1}^{n}[A_i^*(\mu_1+\mu_0)] + o_p(1)\ , \end{aligned} \tag{A-58}$$

where the third equality follows from $\frac{1}{n}\sum_{i=1}^{n} I\{S_i = s\}Y_i = \frac{p(s)}{2}\left[\mu_1 + \mu_0 + E[m_1(Z) + m_0(Z)|S = s]\right] + o_p(1)$, and the last equality follows from $n/n(s) = 1/p(s) + o_p(1)$, $D_n(s)/\sqrt{n} = O_p(1)$, and the definition of $R_{n,2}$ in (B-65). Invoking (A-57), (A-58), and (A-56), we conclude that $\sqrt{n}(\hat{\beta}_n - \theta(Q)) = 2R_{n,1} + 2R_{n,3} + o_p(1)$. The proof of part I is completed by invoking (B-67) and (B-68) in the proof of Lemma B.1 which give $2R_{n,1} + 2R_{n,3} \xrightarrow{d} N(0, \sigma_{\tilde{Y}}^2 + \sigma_H^2)$.

Part II. We first prove the result for homoskedasticity-only standard errors, i.e.,

$$\hat{V}_{\text{homo}} = \left(\frac{1}{n}\sum_{i=1}^{n}\hat{u}_i^2\right)\left(\frac{\mathbb{C}_n'\mathbb{C}_n}{n}\right)^{-1}_{[1,1]} \xrightarrow{p} \sigma_{\tilde{Y}}^2 + \sigma_H^2\ , \tag{A-59}$$

where $\mathbb{C}_n$ is a $n \times |\mathcal{S}| + 1$ matrix with the treatment assignment vector $\mathbb{A}_n$ in the first row and the strata indicators vector in the rest of the rows, and $\hat{u}_i$ is the least squares residual of the regression in (29).

Next note that $\frac{1}{n}\mathbb{C}_n'\mathbb{C}_n \xrightarrow{p} \Sigma_{\mathbb{C}}$ where

$$\Sigma_{\mathbb{C}} \equiv \begin{bmatrix} 1/2 & \frac{1}{2}p(1) & \frac{1}{2}p(2) & \cdots & \frac{1}{2}p(|\mathcal{S}|) \\ \frac{1}{2}p(1) & p(1) & 0 & \cdots & 0 \\ \frac{1}{2}p(2) & \vdots & \ddots & & \vdots \\ \vdots & \vdots & & \ddots & \vdots \\ \frac{1}{2}p(|\mathcal{S}|) & 0 & \cdots & \cdots & p(|\mathcal{S}|) \end{bmatrix} \text{ and } \Sigma_{\mathbb{C}}^{-1} = \begin{bmatrix} 4 & -2 & -2 & \cdots & -2 \\ -2 & 1+\frac{1}{p(1)} & 1 & \cdots & 1 \\ -2 & \vdots & \ddots & & \vdots \\ \vdots & \vdots & & \ddots & \vdots \\ -2 & 1 & \cdots & \cdots & 1+\frac{1}{p(|\mathcal{S}|)} \end{bmatrix}. \tag{A-60}$$

The convergence in probability follows from $\frac{1}{n}\sum_{i=1}^{n} A_i \xrightarrow{p} 1/2$, $n_1(s)/n \xrightarrow{p} p(s)/2$, and $n(s)/n \xrightarrow{p} p(s)$ for all $s \in \mathcal{S}$.

The second result follows from analytically computing the inverse matrix, which we omit here. It follows that the $[1,1]$ component of $\Sigma_{\mathbb{C}}^{-1}$ equals 4. By Lemma B.11, we know $n^{-1}\sum_{i=1}^{n}\hat{u}_i^2 \xrightarrow{p} \left(\sigma_{\tilde{Y}}^2+\sigma_H^2\right)/4$, and the result in (A-59) immediately follows.

We now prove the result for heteroskedasticity-robust standard errors, i.e.,

$$\hat{V}_{\text{hc}} = \left[\left(\frac{\mathbb{C}_n'\mathbb{C}_n}{n}\right)^{-1}\left(\frac{\mathbb{C}_n'\operatorname{diag}(\{\hat{u}_i^2\}_{i=1}^n)\mathbb{C}_n}{n}\right)\left(\frac{\mathbb{C}_n'\mathbb{C}_n}{n}\right)^{-1}\right]_{[1,1]} \xrightarrow{p} \sigma_{\tilde{Y}}^2+\sigma_H^2 \ . \tag{A-61}$$

First note that

$$\frac{\mathbb{C}_n'\operatorname{diag}(\{\hat{u}_i^2\}_{i=1}^n)\mathbb{C}_n}{n} = \frac{1}{n}\begin{bmatrix} \sum_{i=1}^n \hat{u}_i^2 A_i & \sum_{i=1}^n \hat{u}_i^2 A_i I\{S_i=1\} & \cdots & \sum_{i=1}^n \hat{u}_i^2 A_i I\{S_i=|\mathcal{S}|\} \\ \sum_{i=1}^n \hat{u}_i^2 A_i I\{S_i=1\} & \sum_{i=1}^n \hat{u}_i^2 I\{S_i=1\} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^n \hat{u}_i^2 A_i I\{S_i=|\mathcal{S}|\} & 0 & \cdots & \sum_{i=1}^n \hat{u}_i^2 I\{S_i=|\mathcal{S}|\} \end{bmatrix} .$$

It follows from Lemma B.11 that

$$\frac{\mathbb{C}_n'\operatorname{diag}(\{\hat{u}_i^2\}_{i=1}^n)\mathbb{C}_n}{n} \xrightarrow{p} \Omega \tag{A-62}$$

where each component of the matrix Ω corresponds to the respective limits in Lemma B.11. It follows that

$$\left(\frac{\mathbb{C}_n'\mathbb{C}_n}{n}\right)^{-1}\left(\frac{\mathbb{C}_n'\operatorname{diag}(\{\hat{u}_i^2\}_{i=1}^n)\mathbb{C}_n}{n}\right)\left(\frac{\mathbb{C}_n'\mathbb{C}_n}{n}\right)^{-1} \xrightarrow{p} \Sigma_{\mathbb{C}}^{-1}\Omega\Sigma_{\mathbb{C}}^{-1} = \begin{bmatrix} \sigma_{\tilde{Y}}^2+\sigma_H^2 & \mathbb{V}_{12} \\ \mathbb{V}_{12}' & \mathbb{V}_{22} \end{bmatrix} , \tag{A-63}$$

where $\mathbb{V}_{12}$ is an $|\mathcal{S}|\times 1$ matrix with sth element equal to

$$-(\sigma_{\tilde{Y}}^2+\sigma_H^2)/2+\sigma_{\tilde{Y}(1)}^2(s)-\sigma_{\tilde{Y}(0)}^2(s) \ ,$$

and $\mathbb{V}_{22}$ is an $|\mathcal{S}|\times|\mathcal{S}|$ matrix with $(s,\tilde{s})$th element equal to

$$\begin{aligned}&\frac{1}{4}\left(\sigma_{\tilde{Y}}^2+\sigma_H^2\right)+\frac{1}{2}\left[\sigma_{\tilde{Y}(0)}^2(s)-\sigma_{\tilde{Y}(1)}^2(s)\right]+\frac{1}{2}\left[\sigma_{\tilde{Y}(0)}^2(\tilde{s})-\sigma_{\tilde{Y}(1)}^2(\tilde{s})\right]+2\left(\sigma_{\tilde{Y}(1)}^2(\tilde{s})+\sigma_{\tilde{Y}(0)}^2(\tilde{s})\right)\\ &\quad +I\{\tilde{s}=s\}\frac{1}{4}p(\tilde{s})\left[(E\left[m_1(Z)-m_0(Z)|S=\tilde{s}\right])^2\right] \ .\end{aligned}$$

From here, (A-61) follows immediately. Putting together the results of parts I and II completes the proof of the theorem.

Appendix B Auxiliary Results

Lemma B.1. *Let Assumptions 2.1 and 2.2 hold. Then,*

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}[(Y_i(1)-\mu_1)A_i-(Y_i(0)-\mu_0)(1-A_i)] \xrightarrow{d} N(0,(\sigma_{\tilde{Y}}^2+\sigma_H^2+\sigma_A^2)/4) \ ,$$

where $\sigma_{\tilde{Y}}^2$, σ_H^2, and σ_A^2 are defined in Table 3.

Proof. Let $\tilde{Y}_i(a) \equiv Y_i(a) - E[Y_i(a)|S_i]$, $m_a(Z_i) \equiv E[Y_i(a)|Z_i] - \mu_a$, and consider the following derivation,

$$
\begin{aligned}
&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}[(Y_i(1)-\mu_1)A_i - (Y_i(0)-\mu_0)(1-A_i)] \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\tilde{Y}_i(1)A_i - \tilde{Y}_i(0)(1-A_i)\right] + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}[E[m_1(Z_i)|S_i]A_i - E[m_0(Z_i)|S_i](1-A_i)] \\
&= R_{n,1} + \frac{1}{2\sqrt{n}}\sum_{i=1}^{n}A_i^*[\sum_{s\in\mathcal{S}}E[m_1(Z_i)|S_i=s]I\{S_i=s\} + \sum_{s\in\mathcal{S}}E[m_0(Z_i)|S_i=s]I\{S_i=s\}] \\
&\quad + \frac{1}{2\sqrt{n}}\sum_{i=1}^{n}[\sum_{s\in\mathcal{S}}E[m_1(Z_i)|S_i=s]I\{S_i=s\} - \sum_{s\in\mathcal{S}}E[m_0(Z_i)|S_i=s]I\{S_i=s\}] \\
&= R_{n,1} + R_{n,2} + R_{n,3}\ ,
\end{aligned}
$$

where we used $A_i = (1+A^*)/2$ and the following definitions,

$$R_{n,1} \equiv \frac{1}{\sqrt{n}}\sum_{i=1}^{n}[\tilde{Y}_i(1)A_i - \tilde{Y}_i(0)(1-A_i)]\ , \qquad \text{(B-64)}$$

$$R_{n,2} \equiv \frac{1}{2}\sum_{s\in\mathcal{S}}\frac{D_n(s)}{\sqrt{n}}[E[m_1(Z)|S=s] + E[m_0(Z)|S=s]]\ , \qquad \text{(B-65)}$$

$$R_{n,3} \equiv \frac{1}{2}\sum_{s\in\mathcal{S}}\sqrt{n}[\frac{n(s)}{n} - p(s)][E[m_1(Z)|S=s] - E[m_0(Z)|S=s]]\ . \qquad \text{(B-66)}$$

The result follows from the continuous mapping theorem once we show that $R_n \equiv (R_{n,1}, R_{n,2}, R_{n,3}) \xrightarrow{d} (\zeta_{R_1}, \zeta_{R_2}, \zeta_{R_3})$ where $\{\zeta_{R_1}, \zeta_{R_2}, \zeta_{R_3}\}$ are independent and satisfy $\zeta_{R_1} \sim N(0, \sigma^2_{\tilde{Y}}/4)$, $\zeta_{R_2} \sim N(0, \sigma^2_A/4)$, and $\zeta_{R_3} \sim N(0, \sigma^2_H/4)$.

Start with $R_{n,1}$. Conditionally on $\{S^{(n)}, A^{(n)}\}$, $\{\tilde{Y}_i(1)A_i - \tilde{Y}_i(0)(1-A_i) : 1 \le i \le n\}$ is an independent sequence with mean zero that, by Lemma B.4, satisfies the Lyapunov condition a.s. Then, the Lyapunov CLT implies that

$$\left\{ \left. \frac{\sum_{i=1}^{n}\left[\tilde{Y}_i(1)A_i - \tilde{Y}_i(0)(1-A_i)\right]}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\text{Var}[\tilde{Y}_i(1)A_i + \tilde{Y}_i(0)(1-A_i)|S^{(n)}, A^{(n)}]}} \right| S^{(n)}, A^{(n)} \right\} \xrightarrow{d} N(0,1) \text{ a.s.}$$

By using Slutsky's theorem and Lemma B.2,

$$\{R_{n,1}|S^{(n)}, A^{(n)}\} \xrightarrow{d} \zeta_{R_1} \sim N(0, \sigma^2_{\tilde{Y}}/4) \text{ a.s.} \qquad \text{(B-67)}$$

where $\sigma^2_{\tilde{Y}} = 2(\text{Var}[\tilde{Y}(1)] + \text{Var}[\tilde{Y}(0)])$. For $R_{n,2}$, Assumption 2.2.(c) implies that

$$\{R_{n,2}|S^{(n)}\} \xrightarrow{d} \zeta_{R_2} \sim N(0, \frac{1}{4}\sum_{s\in\mathcal{S}}\sigma^2_D(s)\,[E[m_1(Z)|S=s] + E[m_0(Z)|S=s]]^2]) \text{ a.s.}$$

For $R_{n,3}$, note that $n(s)/n = n^{-1}\sum_{i=1}^{n}I\{S_i=s\}$, so that by $W^{(n)}$ being i.i.d. and the CLT, it follows that

$$\left\{\sqrt{n}\left(\frac{n(s)}{n} - p(s)\right)\right\}_{s\in\mathcal{S}} \xrightarrow{d} N(\mathbf{0}_{|\mathcal{S}|}, \Sigma)\ ,$$

where $\Sigma_{[s,\tilde{s}]} = p(s)I\{s=\tilde{s}\} - p(s)p(\tilde{s})$. In turn, this implies that

$$R_{n,3} \xrightarrow{d} \zeta_{R_3} \sim N(0, \frac{1}{4}\sum_{s\in\mathcal{S}}p(s)\,(E[m_1(Z)|S=s] - E[m_0(Z)|S=s])^2)\ . \qquad \text{(B-68)}$$

To conclude the proof, we show that $R_n \equiv (R_{n,1}, R_{n,2}, R_{n,3}) \xrightarrow{d} (\zeta_{R_1}, \zeta_{R_2}, \zeta_{R_3})$ with $\{\zeta_{R_1}, \zeta_{R_2}, \zeta_{R_3}\}$ independent random variables. In what follows fix $h = (h_1, h_2, h_3) \in \mathbf{R}^3$ s.t. $P\{\zeta_{R_1} \leq h_1\}P\{\zeta_{R_2} \leq h_2\}P\{\zeta_{R_3} \leq h_3\}$ is continuous. Notice that $P\{\zeta_{R_1} \leq h_1\}$ is continuous, so discontinuities could be caused by discontinuities of $P\{\zeta_{R_2} \leq h_2\}$ or $P\{\zeta_{R_3} \leq h_3\}$. According to these possibilities, we divide the proof into cases.

First, suppose that both $P\{\zeta_{R_2} \leq \cdot\}$ and $P\{\zeta_{R_3} \leq \cdot\}$ are continuous at (h_2, h_3). Conditional on $\{S^{(n)}, A^{(n)}\}$, $R_{n,2}$ and $R_{n,3}$ are non-stochastic and, hence, conditionally, $R_{n,1}$, $R_{n,2}$, and $R_{n,3}$ are independent. Then, $P\{R_n \leq h|S^{(n)}, A^{(n)}\} = P\{R_{n,1} \leq h_1|S^{(n)}, A^{(n)}\}P\{R_{n,2} \leq h_2|S^{(n)}, A^{(n)}\}P\{R_{n,3} \leq h_3|S^{(n)}, A^{(n)}\}$. Our previous derivations show that $\{R_{n,1}|S^{(n)}, A^{(n)}\} \xrightarrow{d} \zeta_{R_1}$ a.s. and $\{R_{n,2}|S^{(n)}\} \xrightarrow{d} \zeta_{R_2}$ a.s. Then, repeated applications of the dominated convergence theorem and some algebra imply that

$$P\{R_n \leq h\} = P\{\zeta_{R_1} \leq h_1\}P\{\zeta_{R_2} \leq h_2\}P\{\zeta_{R_3} \leq h_3\} + o(1) \ .$$

Second, suppose $P\{\zeta_{R_3} \leq \cdot\}$ is discontinuous at h_3. Since ζ_{R_3} is normally distributed, the discontinuity at h_3 implies that $\text{Var}[\zeta_{R_3}] = 0$ (this occurs if and only if $E[m_1(Z)|S = s] = E[m_0(Z)|S = s]$ for all $s \in \mathcal{S}$). This implies that $R_{n,3} = \zeta_{R_3} = 0$ and $h_3 = 0$ and so,

$$P\{R_n \leq h\} = P\{\zeta_{R_1} \leq h_1\}P\{\zeta_{R_2} \leq h_2\} + o(1) = P\{\zeta_{R_1} \leq h_1\}P\{\zeta_{R_2} \leq h_2\}P\{\zeta_{R_3} \leq h_3\} + o(1) \ ,$$

where we have used that $P\{\zeta_{R_3} \leq h_3\} = P\{R_{n,3} \leq h_3\} = 1$ and the convergence follows from the same argument as in the first case.

Third, suppose $P\{\zeta_{R_2} \leq \cdot\}$ is discontinuous at h_2 and $P\{\zeta_{R_3} \leq \cdot\}$ is continuous at h_3. The fact that $P\{\zeta_{R_1} \leq h_1\}P\{\zeta_{R_2} \leq h_2\}P\{\zeta_{R_3} \leq h_3\}$ is continuous at h implies that $P\{\zeta_{R_2} \leq h_2\} = 0$. Since ζ_{R_2} is normally distributed, the discontinuity at h_2 implies that $\text{Var}[\zeta_{R_2}] = 0$ (this occurs if and only if $\sigma_D^2(s)\,[E[m_1(Z)|S = s] + E[m_0(Z)|S = s]]^2 = 0$ for all $s \in \mathcal{S}$). This implies that $\zeta_{R_2} = 0$ and $h_2 = 0$, but then $P\{\zeta_{R_2} \leq h_2\} = 1$, which is a contradiction. This completes the proof for the last case. ■

Lemma B.2. *Let Assumptions 2.1 and 2.2 hold. Then,*

$$\frac{1}{n}\sum_{i=1}^{n} \text{Var}[\tilde{Y}_i(1)A_i + \tilde{Y}_i(0)(1 - A_i)|S^{(n)}, A^{(n)}] \overset{a.s.}{\to} \frac{1}{2}(\text{Var}[\tilde{Y}(1)] + \text{Var}[\tilde{Y}(0)]) \ .$$

Proof. For any random variable X, let $\sigma_X^2(S_i) = \text{Var}[X|S_i]$. Now note that

$$\begin{aligned}\text{Var}[\tilde{Y}_i(1)A_i + \tilde{Y}_i(0)(1 - A_i)|S^{(n)}, A^{(n)}] &= E[\tilde{Y}_i(1)^2A_i + \tilde{Y}_i(0)^2(1 - A_i) + 2A_i(1 - A_i)\tilde{Y}_i(1)\tilde{Y}_i(0)|S^{(n)}, A^{(n)}] \\ &= A_i\sigma^2_{\tilde{Y}(1)}(S_i) + (1 - A_i)\sigma^2_{\tilde{Y}(0)}(S_i) \ ,\end{aligned}$$

where have used Assumption 2.2.(a) and $E[\tilde{Y}_i(a)|S_i] = 0$. The result then follows by proving that

$$\frac{1}{n}\sum_{i=1}^{n} A_i\sigma^2_{\tilde{Y}(1)}(S_i) \overset{a.s.}{\to} \frac{1}{2}\,\text{Var}[\tilde{Y}(1)] \ ,$$

since the second terms involves similar arguments. Note that $\{A_i\sigma^2_{\tilde{Y}(1)}(S_i) : 1 \leq i \leq n\}$ is a sequence of nonnegative random variables and with finite variance by Assumption 2.1 and the fact that $\mathcal{S}$ is a finite set. For any $n \in \mathbf{N}$, let

$\Psi_n \equiv \sum_{i=1}^n A_i \sigma^2_{\tilde{Y}(1)}(S_i)$ and so, for any $n, \tilde{n} \in \mathbf{N} \cup \{0\}$ with $n > \tilde{n}$,

$$E[\Psi_n - \Psi_{\tilde{n}}] = \sum_{i=\tilde{n}}^n \sum_{s \in \mathcal{S}} E[A_i \sigma^2_{\tilde{Y}(1)}(s) I\{S_i = s\}] = \sum_{i=\tilde{n}}^n \sum_{s \in \mathcal{S}} E[A_i | S_i = s] \sigma^2_{\tilde{Y}(1)}(s) p(s) \leq (n - \tilde{n}) \sigma^2_{\tilde{Y}(1)}(1/2 + C) \ ,$$

for some constant C, where we used $\sigma^2_{\tilde{Y}(1)} \equiv \sum_{s \in \mathcal{S}} \sigma^2_{\tilde{Y}(1)}(s) p(s)$ and $E[A_i | S_i = s] = 1/2 + O(n^{-1})$ from Assumption 2.2.(b). For $\tilde{n} = 0$, this derivation shows that $E[n^{-1} \sum_{i=1}^n A_i \sigma^2_{\tilde{Y}(1)}(S_i)] = \sigma^2_{\tilde{Y}(1)}/2 + O_{a.s}(n^{-1}) \xrightarrow{a.s.} \sigma^2_{\tilde{Y}(1)}/2$ as $n \to \infty$.

Next, note that for $s, \tilde{s} \in \mathcal{S}$ with $s \neq \tilde{s}$, the Cauchy-Schwarz inequality gives

$$\begin{aligned} \mathrm{Cov}[\sum_{i=1}^n A_i I\{S_i = s\}, \sum_{j=1}^n A_j I\{S_j = \tilde{s}\}] &\leq (\mathrm{Var}(\sum_{i=1}^n A_i I\{S_i = s\}) \,\mathrm{Var}(\sum_{i=1}^n A_i I\{S_i = \tilde{s}\}))^{1/2} \\ &\leq \frac{n}{4} \sqrt{p(s)(1 - p(s))} \sqrt{p(\tilde{s})(1 - p(\tilde{s}))} \ , \end{aligned} \tag{B-69}$$

where the last step follows from Assumption 2.2.(d). Therefore,

$$\begin{aligned} \mathrm{Var}\left(\sum_{i=1}^n A_i \sigma^2_{\tilde{Y}(1)}(S_i)\right) &= \mathrm{Var}\left(\sum_{s \in \mathcal{S}} \sigma^2_{\tilde{Y}(1)}(s) \sum_{i=1}^n A_i I\{S_i = s\}\right) \\ &= \sum_{s \in \mathcal{S}} (\sigma^2_{\tilde{Y}(1)}(s))^2 \,\mathrm{Var}\left(\sum_{i=1}^n A_i I\{S_i = s\}\right) + \sum_{s \neq \tilde{s}} \sigma^2_{\tilde{Y}(1)}(s) \sigma^2_{\tilde{Y}(1)}(\tilde{s}) \,\mathrm{Cov}[\sum_{i=1}^n A_i I\{S_i = s\}, \sum_{j=1}^n A_j I\{S_j = \tilde{s}\}] \\ &\leq \frac{n}{4} \sum_{s \in \mathcal{S}} (\sigma^2_{\tilde{Y}(1)}(s))^2 p(s)(1 - p(s)) + \frac{n}{4} \sum_{s \neq \tilde{s}} \sigma^2_{\tilde{Y}(1)}(s) \sigma^2_{\tilde{Y}(1)}(\tilde{s}) \sqrt{p(s)(1 - p(s))} \sqrt{p(\tilde{s})(1 - p(\tilde{s}))} \ , \end{aligned}$$

where we used (B-69). From here, we conclude that,

$$\sum_{n=1}^\infty \frac{1}{n^3} \mathrm{Var}\left(\sum_{i=1}^n A_i \sigma^2_{\tilde{Y}(1)}(S_i)\right) \leq |\mathcal{S}|^2 \max_{s \in S} \{(\sigma^2_{\tilde{Y}(1)}(s))^2 p(s)(1 - p(s))\} \sum_{n=1}^\infty \frac{1}{4n^2} < \infty \ .$$

By combining these findings and using the theorem in Petrov (2011), the desired result follows. ■

Lemma B.3. *Let Assumptions 2.1 and 2.2 hold. Then,*

$$\frac{1}{n} \sum_{i=1}^n [(Y_i(1) - \mu_1) A_i + (Y_i(0) - \mu_0)(1 - A_i)] \xrightarrow{p} 0 \ .$$

Proof. By similar steps to those in the proof of Lemma B.1, it follows that

$$\frac{1}{n} \sum_{i=1}^n [(Y_i(1) - \mu_1) A_i + (Y_i(0) - \mu_0)(1 - A_i)] = \bar{R}_{n,1} + \bar{R}_{n,2} + \bar{R}_{n,3} \ ,$$

where $\bar{R}_{n,1} = R_{n,1}/\sqrt{n}$, $\bar{R}_{n,2} = R_{n,2}/\sqrt{n}$, $\bar{R}_{n,3} = R_{n,3}/\sqrt{n}$, and $(R_{n,1}, R_{n,2}, R_{n,3})$ are defined in (B-64)-(B-66). Note that $\{Z_i : 1 \leq i \leq n\}$ is i.i.d. and, hence, so is $\{I\{S_i = s\} : 1 \leq i \leq n\}_{s \in \mathcal{S}}$ with mean $\{p(s)\}_{s \in \mathcal{S}}$. The SLLN implies that $\bar{R}_{n,3} \xrightarrow{a.s.} 0$. Second, $\{\{D_n(s)/\sqrt{n}\}_{s \in \mathcal{S}} | S^{(n)}\} \xrightarrow{d} \zeta_D$ a.s. implies that $\{\bar{R}_{n,2} | S^{(n)}\} \xrightarrow{p} 0$ a.s. which implies $\bar{R}_{n,2} \xrightarrow{p} 0$. Finally, by Lemma B.1 we know $\{R_{n,1} | S^{(n)}, A^{(n)}\} = O_p(1)$ a.s., and this immediately implies $\bar{R}_{n,1} \xrightarrow{p} 0$. ■

Lemma B.4. *Let Assumptions 2.1 and 2.2 hold. Then, conditional on $\{S^{(n)}, A^{(n)}\}$, $\{A_i \tilde{Y}_i(1) + (1 - A_i) \tilde{Y}_i(0) : 1 \leq i \leq n\}$ satisfies the Lyapunov condition, i.e.,*

$$\frac{\left(\sum_{i=1}^n E[|A_i \tilde{Y}_i(1) + (1 - A_i) \tilde{Y}_i(0)|^{2+\delta} | S^{(n)}, A^{(n)}]\right)^{1/(2+\delta)}}{\left(\sum_{i=1}^n \mathrm{Var}[A_i \tilde{Y}_i(1) + (1 - A_i) \tilde{Y}_i(0) | S^{(n)}, A^{(n)}]\right)^{1/2}} \xrightarrow{a.s.} 0 \ .$$

Proof. By Assumption 2.2.(a) and $W^{(n)}$ consisting of n i.i.d. observations, conditionally on S_i, $(\tilde{Y}_i(1), \tilde{Y}_i(0))$ is independent of $\{S^{(n)}, A^{(n)}\}$. It follows from Minkowski's inequality and Assumptions 2.1 and 2.2.(a) that

$$\begin{aligned} E[|\tilde{Y}(a)|^{2+\delta}|S^{(n)}, A^{(n)}] &\leq 2^{1+\delta}E[|Y_i(a)|^{2+\delta}|S^{(n)}, A^{(n)}] + 2^{1+\delta}E[|E[Y_i(a)|S_i]|^{2+\delta}|S^{(n)}, A^{(n)}] \\ &= 2^{1+\delta}(E[|Y_i(a)|^{2+\delta}|S_i] + E[|E[Y_i(a)|S_i]|^{2+\delta}|S_i]) < \infty \ , \end{aligned} \tag{B-70}$$

where we have used that $\mathcal{S}$ is a finite set. We thus conclude that

$$E[|A_i\tilde{Y}_i(1) + (1 - A_i)\tilde{Y}_i(0)|^{2+\delta}|S^{(n)}, A^{(n)}] \leq \max_{a\in\{0,1\}} E[|\tilde{Y}_i(a)|^{2+\delta}|S^{(n)}, A^{(n)}] < \infty \tag{B-71}$$

and

$$\frac{1}{n}\sum_{i=1}^{n} E[|A_i\tilde{Y}_i(1) + (1 - A_i)\tilde{Y}_i(0)|^{2+\delta}|S^{(n)}, A^{(n)}] < \infty \ .$$

By combining the previous equation with the result in Lemma B.2, it follows that the expression in the statement of the lemma is $O_{a.s.}(n^{-\delta/(4+2\delta)}) = o_{a.s.}(1)$. ■

Lemma B.5. *Let $\{U_i\}_{i\geq 1}$ be an arbitrary sequence of random variables s.t. $E|U_i|^{2+\delta} \leq B < \infty$ for some $\delta, B > 0$. Then,*

$$\lim_{n\to\infty} E\left[\left|\frac{1}{n}\sum_{i=1}^{n} U_i^2 - E\left(\frac{1}{n}\sum_{i=1}^{n} U_i^2\right)\right|^{1+\delta/2}\right] = 0 \ .$$

Proof. First we show that

$$E\left[\left|\frac{1}{n}\sum_{i=1}^{n} U_i^2 - E\left(\frac{1}{n}\sum_{i=1}^{n} U_i^2\right)\right|^{1+\delta/2}\right] \leq 2E\left[\left|\frac{1}{n}\sum_{i=1}^{n} U_i^2\right|^{1+\delta/2}\right] \ .$$

Define $\Psi_n \equiv \frac{1}{n}\sum_{i=1}^{n} U_i^2$. By the triangle and Hölder's inequality,

$$|\Psi_n - E(\Psi_n)|^{1+\delta/2} \leq (|\Psi_n| + |E[\Psi_n]|)^{1+\delta/2} \leq C(|\Psi_n|^{1+\delta/2} + |E[\Psi_n]|^{1+\delta/2}) \ ,$$

where $C = 2^{(1+\delta/2)(1-1/(1+\delta/2))}$. Taking expectations on both side implies

$$E\left[|\Psi_n - E[\Psi_n]|^{1+\delta/2}\right] \leq CE\left[|\Psi_n|^{1+\delta/2}\right] + C\left(E[|\Psi_n|]\right)^{1+\delta/2} \leq 2CE\left[|\Psi_n|^{1+\delta/2}\right] \ , \tag{B-72}$$

where we used $E[\Psi_n] \leq E|\Psi_n|$ and Hölder's inequality.

Second, note that

$$E\left[\left|\frac{1}{n}\sum_{i=1}^{n} U_i^2\right|^{1+\delta/2}\right] \leq E\left[\left(\sum_{i=1}^{n}\left|\frac{1}{\sqrt{n}}U_i\right|^{2+\delta}\right)\right] = \left(\frac{1}{n}\right)^{\delta/2} E\left[\frac{1}{n}\sum_{i=1}^{n}|U_i|^{2+\delta}\right] \ , \tag{B-73}$$

where the inequality follows Hölder's inequality (for the counting measure). Using (B-72), (B-73), and $E|U_i|^{2+\delta} \leq B$, we conclude that,

$$E\left[\left|\frac{1}{n}\sum_{i=1}^{n} U_i^2 - E\left(\frac{1}{n}\sum_{i=1}^{n} U_i^2\right)\right|^{1+\delta/2}\right] \leq 2C\left(\frac{1}{n}\right)^{\delta/2} E\left[\frac{1}{n}\sum_{i=1}^{n}|U_i|^{2+\delta}\right] \leq 2C\left(\frac{1}{n}\right)^{\delta/2} B \ .$$

By taking limits as $n \to \infty$, the result follows. ■

Lemma B.6. *For any $s \in \mathcal{S}$, let $\bar{Q}(s) = \frac{1}{2}Q_1(s) + \frac{1}{2}Q_0(s)$ where $Q_1(s)$ is the conditional on $S = s$ distribution of $Y_i(1) - \theta(Q)$ and $Q_0(s)$ is the conditional on $S = s$ distribution of $Y_i(0)$. Let $V^{(n)} = \{V_i : 1 \leq i \leq n\}$ be an i.i.d. sequence of random variables such that*

$$\{V_i | S^{(n)}\} \overset{d}{=} \{V_i | S_i\} \sim \bar{Q}(S_i) = \frac{1}{2}Q_1(S_i) + \frac{1}{2}Q_0(S_i) \ . \tag{B-74}$$

Then, the conditional mean and variance of V_i are given by,

$$\begin{aligned} E[V_i|S^{(n)}] &= \frac{1}{2}(\mu_0 + E[m_0(Z_i)|S_i]) + \frac{1}{2}(\mu_1 - \theta(Q) + E[m_1(Z_i)|S_i]) \\ \mathrm{Var}[V_i|S^{(n)}] &= \frac{1}{2}(\sigma^2_{\tilde{Y}(1)}(S_i) + \sigma^2_{\tilde{Y}(0)}(S_i)) + \frac{1}{4}(E[m_1(Z_i)|S_i] - E[m_0(Z_i)|S_i])^2 \ . \end{aligned}$$

Proof. Let $\mu(P)$ and $\sigma^2(P)$ denote the mean and variance of a distribution P, and note that for $a \in \{0, 1\}$,

$$\mu(Q_a(s)) = \mu_a + E[m_a(Z)|S = s] - \theta(Q)I\{a = 1\} \text{ and } \sigma^2(Q_a(s)) = \sigma^2_{Y(a)}(s) \ .$$

The expression for $E[V_i|S^{(n)}]$ follows immediately after noticing $E[V_i|S^{(n)}] = E[V_i|S_i]$.

For the conditional variance of V_i, let C_i be a random variable that selects whether $V_i \sim Q_0(S_i)$ or $V_i \sim Q_1(S_i)$. By the law of total variance,

$$\mathrm{Var}[V_i|S^{(n)}] = E[\mathrm{Var}[V_i|C_i, S^{(n)}]|S^{(n)}] + \mathrm{Var}[E[V_i|C_i, S^{(n)}]|S^{(n)}] \ .$$

Let's first consider the expectation of the variance,

$$\begin{aligned} E[\mathrm{Var}[V_i|C_i, S^{(n)}]|S^{(n)}] &= \frac{1}{2}\,\mathrm{Var}[V_i|C_i = 1, S^{(n)}] + \frac{1}{2}\,\mathrm{Var}[V_i|C_i = 0, S^{(n)}] \\ &= \frac{1}{2}(\sigma^2_{Y(1)}(S_i) + \sigma^2_{Y(0)}(S_i)) \\ &= \frac{1}{2}(\sigma^2_{\tilde{Y}(1)}(S_i) + \sigma^2_{\tilde{Y}(0)}(S_i)) \ , \end{aligned}$$

where in the last step we used $\sigma^2_{Y(a)}(S_i) = \sigma^2_{\tilde{Y}(a)}(S_i)$ for $\tilde{Y}_i(a) = Y_i(a) - E[Y_i(a)|S_i]$ and $a \in \{0, 1\}$.

The expression for the variance of the expectation below completes the proof,

$$\begin{aligned} \mathrm{Var}[E[V_i|C_i, S^{(n)}]|S^{(n)}] &= \frac{1}{2}(\mu_1 - \theta(Q) + E[m_1(Z_i)|S_i])^2 + \frac{1}{2}(\mu_0 + E[m_0(Z_i)|S_i])^2 \\ &\quad - \frac{1}{2}(\mu_1 - \theta(Q) + E[m_1(Z_i)|S_i] + \mu_0 + E[m_0(Z_i)|S_i])^2 \\ &= \frac{1}{4}(E[m_1(Z_i)|S_i] - E[m_0(Z_i)|S_i])^2 \ . \end{aligned}$$

■

Lemma B.7. *Let Assumptions 2.1 and 2.2 hold and define*

$$T^U_n(X^{(n)}) \equiv \sqrt{n}(\frac{1}{n_1}\sum_{i=1}^{n} Y_i A_i - \frac{1}{n_0}\sum_{i=1}^{n} Y_i(1 - A_i) - \theta(Q)) = \sqrt{n}(\frac{1}{n_1}\sum_{i=1}^{n}(Y_i - \theta(Q))A_i - \frac{1}{n_0}\sum_{i=1}^{n} Y_i(1 - A_i)) \ . \tag{B-75}$$

It follows that

$$T^U_n(\tilde{X}^{(n)}) \xrightarrow{d} N(0, \sigma^2_{\tilde{Y}} + \sigma^2_H + \sigma^2_A) \ , \tag{B-76}$$

where $\tilde{X}^{(n)} = \{(V_i, A_i, Z_i) : 1 \leq i \leq n\}$, $V^{(n)} = \{V_i : 1 \leq i \leq n\}$ are the random variables defined in Lemma B.6, and $\sigma^2_{\tilde{Y}}$, σ^2_H, and σ^2_A are defined in Table 3.

Proof. Since $T^U_n(X^{(n)})$ in (B-75) is invariant under the same shift for all the observations in $Y^{(n)} = \{Y_i : 1 \leq i \leq n\}$, we can assume without loss of generality that $E[V_i] = \frac{1}{2}(\mu_0 + \mu_1 - \theta(Q)) = 0$. Commputing $T^U_n(\cdot)$ with the sample $\tilde{X}^{(n)}$ we get,

$$T^U_n(\tilde{X}^{(n)}) = \sqrt{n}(\frac{1}{n_1}\sum_{i=1}^{n} V_i A_i - \frac{1}{n_0}\sum_{i=1}^{n} V_i(1 - A_i)) = \frac{2}{1 - (D_n/n)^2}\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n} V_i A^*_i + \frac{D_n}{n}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} V_i\right\}$$
$$= \frac{2}{\sqrt{n}}\sum_{i=1}^{n} V_i A^*_i + o_p(1) \ , \qquad \text{(B-77)}$$

where we used $A^*_i = 2A_i - 1$ and $D_n/n = o_p(1)$. Adding and subtracting the conditional mean derived in Lemma B.6, we get

$$T^U_n(\tilde{X}^{(n)}) = \frac{2}{\sqrt{n}}\sum_{i=1}^{n} V^*_i A^*_i + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(E[m_0(Z_i)|S_i]) + E[m_1(Z_i)|S_i])A^*_i + o_p(1) \ , \qquad \text{(B-78)}$$

where $V^*_i = V_i - \frac{1}{2}(E[m_0(Z_i)|S_i]) + E[m_1(Z_i)|S_i])$ has mean zero conditional on $S^{(n)}$. Denote by $R^U_{n,1}$ and $R^U_{n,2}$ the first and second term in (B-78), respectively.

We first find the limit distribution of $R^U_{n,1}$. Lemma B.6 and similar arguments to those in Lemma B.2 show that

$$\frac{1}{n}\sum_{i=1}^{n} \text{Var}[2V^*_i|S^{(n)}, A^{(n)}] \to^{a.s.} \sigma^2_{\tilde{Y}} + \sigma^2_H \ , \qquad \text{(B-79)}$$

using $\text{Var}[V^*_i] = E[\text{Var}[V^*_i|S_i]]$. Conditional on $\{S^{(n)}, A^{(n)}\}$, $\{V^*_i : 1 \leq i \leq n\}$ is an independent sequence with mean zero that, by arguments similar to those in Lemma B.4, satisfies the Lyapunov condition a.s. It follows that

$$\{R^U_{n,1}|S^{(n)}, A^{(n)}\} \xrightarrow{d} N(0, \sigma^2_{\tilde{Y}} + \sigma^2_H) \text{ a.s.}$$

We now find the limit distribution of $R^U_{n,2}$. First note that,

$$R^U_{n,2} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} A^*_i\left[\sum_{s\in\mathcal{S}}(E[m_0(Z)|S = s]) + E[m_1(Z)|S = s])I\{S_i = s\}\right]$$
$$= \frac{1}{\sqrt{n}}\sum_{s\in\mathcal{S}}\sum_{i=1}^{n} A^*_i I\{S_i = s\}\left[(E[m_0(Z)|S = s]) + E[m_1(Z)|S = s])\right]$$
$$= \sum_{s\in\mathcal{S}}\frac{D_n(s)}{\sqrt{n}}\left[E[m_0(Z)|S = s] + E[m_1(Z)|S = s]\right] \ .$$

Assumption 2.2.(c) then implies that

$$\{R^U_{n,2}|S^{(n)}\} \xrightarrow{d} N(0, \sigma^2_A) \text{ a.s.}$$

By similar arguments to those is the proof of Lemma B.1, we conclude that (B-76) follows. ■

Lemma B.8. *Let Assumption 2.1 and 2.2 hold and define*

$$T^L_n(X^{(n)}) \equiv \sqrt{\frac{n}{n_1}\frac{1}{n_1}\sum_{i=1}^{n}(Y_i - \bar{Y}_{n_1})^2 A_i + \frac{n}{n_0}\frac{1}{n_0}\sum_{i=1}^{n}(Y_i - \bar{Y}_{n_0})^2(1 - A_i)} \ . \qquad \text{(B-80)}$$

It follows that

$$T_n^L(\tilde{X}^{(n)}) \xrightarrow{p} \sqrt{2(\sigma^2_{Y(1)} + \sigma^2_{Y(0)})} \ ,$$

where $\tilde{X}^{(n)} = \{(V_i, A_i, Z_i) : 1 \leq i \leq n\}$ *and* $V^{(n)} = \{V_i : 1 \leq i \leq n\}$ *are the random variables defined in Lemma B.6.*

Proof. Let $\bar{V}_{n_0} = \frac{1}{n_0}\sum_{i=1}^n V_i(1-A_i)$ and $\bar{V}_{n_1} = \frac{1}{n_1}\sum_{i=1}^n V_i A_i$, so that

$$T_n^L(\tilde{X}^{(n)}) = \sqrt{(\frac{n}{n_1})^2 \frac{1}{n}\sum_{i=1}^n (V_i - \bar{V}_{n_1})^2 A_i + (\frac{n}{n_0})^2 \frac{1}{n}\sum_{i=1}^n (V_i - \bar{V}_{n_0})^2(1-A_i)} \ .$$

We focus on the first term in the above expression, since the second follows analogous arguments. Since $T_n^L(X^n)$ is invariant under the same shift for all the observations in $Y^{(n)} = \{Y_i : 1 \leq i \leq n\}$, we can assume without loss of generality that $E[V_i] = \frac{1}{2}(\mu_0 + \mu_1 - \theta(Q)) = 0$, which implies $\mu_0 = 0$ and $\mu_1 = \theta(Q)$ by virtue of $\theta(Q) = \mu_1 - \mu_0$. Re-write this term as

$$\frac{1}{n}\sum_{i=1}^n (V_i - \bar{V}_{n_1})^2 A_i = \frac{1}{n}\sum_{i=1}^n V_i^2 A_i - \bar{V}_{n_1}^2 \frac{1}{n}\sum_{i=1}^n A_i \ . \tag{B-81}$$

Note that V_i and A_i are independent conditional on $S^{(n)}$ by Assumption 2.2.(a), so that

$$\begin{aligned}
E\left[\frac{1}{n}\sum_{i=1}^n V_i^2 A_i \middle| S^{(n)}\right] &= \frac{1}{n}\sum_{i=1}^n E[V_i^2|S^{(n)}]E[A_i|S^{(n)}] = (\frac{1}{2} + O(n^{-1}))\frac{1}{n}\sum_{i=1}^n E[V_i^2|S^{(n)}] \\
&= (\frac{1}{2} + O_{a.s.}(n^{-1}))\frac{1}{n}\sum_{i=1}^n \left\{\frac{1}{2}E[(Y_i(1) - \theta(Q))^2|S^{(n)}] + \frac{1}{2}E[Y_i(0)^2|S^{(n)}]\right\} \\
&= (\frac{1}{4} + O_{a.s.}(n^{-1}))\frac{1}{n}\sum_{i=1}^n \left\{\sigma^2_{Y(1)}(S_i) + E[Y_i(1) - \theta(Q)|S_i]^2 + \sigma^2_{Y(0)}(S_i) + E[Y_i(0)|S_i]^2\right\} \\
&\xrightarrow{a.s.} \frac{1}{4}(\sigma^2_{Y(1)} + \sigma^2_{Y(0)}) \ ,
\end{aligned}$$

where the second equality follows from Assumption 2.2.(b), the fourth equality follows from $\mu_0 = 0$, $\mu_1 = \theta(Q)$, and the law of total variance, and the last step from arguments similar to those in the proof of Lemma B.2.

It remains to be shown that the second term in (B-81) is $o_p(1)$. In order to see this, note that

$$\begin{aligned}
E\left[\frac{1}{n}\sum_{i=1}^n V_i A_i \middle| S^{(n)}, A^{(n)}\right] &= \frac{1}{n}\sum_{i=1}^n A_i E[V_i|S^{(n)}, A^{(n)}] \\
&= \frac{1}{n}\sum_{i=1}^n A_i \left\{\frac{1}{2}(\mu_1 - \theta(Q) + E[m_1(Z_i)|S_i]) + \frac{1}{2}(\mu_0 + E[m_0(Z_i)|S_i])\right\} \\
&\xrightarrow{p} 0 \ a.s. \ ,
\end{aligned}$$

and

$$\begin{aligned}
\text{Var}\left[\frac{1}{n}\sum_{i=1}^n V_i A_i \middle| S^{(n)}, A^{(n)}\right] &= \frac{1}{n^2}\sum_{i=1}^n A_i \,\text{Var}[V_i|S^{(n)}, A^{(n)}] \\
&= \frac{1}{n^2}\sum_{i=1}^n A_i \left\{\frac{1}{2}(\sigma^2_{\tilde{Y}(1)}(S_i) + \sigma^2_{\tilde{Y}(0)}(S_i)) + \frac{1}{4}(E[m_1(Z_i)|S_i] - E[m_0(Z_i)|S_i])^2\right\} \\
&\xrightarrow{p} 0 \ a.s. \ ,
\end{aligned}$$

where the convergence holds by arguments similar to those in the proof of Lemma B.2. Then,

$$\bar{V}_{n_1} = \frac{n}{n_1}\frac{1}{n}\sum_{i=1}^{n} V_i A_i \xrightarrow{p} 0 \ ,$$

where we used $n/n_1 = 2 + o_p(1)$. Combining the previous steps,

$$(\frac{n}{n_1})^2 \frac{1}{n}\sum_{i=1}^{n}(V_i - \bar{V}_{n_1})^2 A_i = (\frac{n}{n_1})^2 (\frac{1}{n}\sum_{i=1}^{n} V_i^2 A_i - \bar{V}_{n_1}^2 \frac{1}{n}\sum_{i=1}^{n} A_i) \xrightarrow{p} \sigma^2_{Y(1)} + \sigma^2_{Y(0)} \ ,$$

and the result follows. ■

Lemma B.9. *Let $G^{(n)}$ and $G^{(n)\prime}$ be two independent random transformations from $\mathbf{G}_n(S^{(n)})$, also independent from $X^{(n)}$. Let L_n and T_n be sequences of random variables satisfying*

$$L_n(G^{(n)}X^{(n)}) \xrightarrow{p} L \ , \tag{B-82}$$

and

$$(T_n(G^{(n)}X^{(n)}), T_n(G^{(n)\prime}X^{(n)})) \xrightarrow{d} (T, T') \ , \tag{B-83}$$

where T and T' are independent, each with common cdf $F^T(\cdot)$. Then, the randomization distribution of $L_n T_n$ converges to LT in probability, i.e.,

$$\hat{F}_n^{LT}(t) \equiv \frac{1}{|\mathbf{G}_n(S^{(n)})|}\sum_{g\in\mathbf{G}_n(S^{(n)})} I\{L_n(gX^{(n)})T_n(gX^{(n)}) \le t\} \xrightarrow{p} F^{LT}(t) \ , \tag{B-84}$$

if F^{LT} is continuous at t, where F^{LT} denotes the corresponding c.d.f. of LT.

Proof. The proof of this lemma is a simple adaptation of the one in Chung and Romano (2013, Lemma A.3) that allows for $L = 0$ and so we omit it here. ■

Lemma B.10. *Let Assumptions 2.1 and 2.2 hold and $\gamma = (\beta, \delta_1, \ldots, \delta_{|\mathcal{S}|})'$ be the parameters in the regression (29). Let $\hat{\gamma}_n$ be the least squares estimator of γ. Then,*

$$\hat{\gamma}_n \xrightarrow{p} \gamma \equiv \begin{bmatrix} \theta(Q) \\ \mu_0 + E[m_1(Z) + m_0(Z)|S=1]/2 \\ \vdots \\ \mu_0 + E[m_1(Z) + m_0(Z)|S=|\mathcal{S}|]/2 \end{bmatrix} .$$

Proof. First note that $\hat{\gamma}_n = (\mathbb{C}_n'\mathbb{C}_n)^{-1}\mathbb{C}_n'\mathbb{Y}_n$, where $\mathbb{C}_n$ is an $n \times |\mathcal{S}|+1$ matrix with the treatment assignment vector $\mathbb{A}_n$ in the first row and the strata indicators vector in the rest of the rows, and $\mathbb{Y}_n$ is an $n \times 1$ vector of outcomes. The $(s+1)$th element of $\frac{1}{n}\mathbb{C}_n'\mathbb{Y}_n$ equals $\frac{1}{n}\sum_{i=1}^{n} A_i Y_i$ if $s = 0$ and $\frac{1}{n}\sum_{i=1}^{n} I\{S_i = s\}Y_i$ for $s \in \mathcal{S}$. In turn, this last term satisfies

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} I\{S_i = s\}Y_i &= \frac{n_1(s)}{n}(\mu_1 + E[m_1(Z)|S=s]) + (\frac{n(s)}{n} - \frac{n_1(s)}{n})(\mu_0 + E[m_0(Z)|S=s]) \\ &\quad + \frac{1}{n}\sum_{i=1}^{n} A_i I\{S_i = s\}\tilde{Y}_i(1) + \frac{1}{n}\sum_{i=1}^{n}(1-A_i)I\{S_i = s\}\tilde{Y}_i(0) \\ &= \frac{1}{2}p(s)(\mu_1 + \mu_0 + E[m_1(Z) + m_0(Z)|S=s]) + o_p(1) \ ,\end{aligned}$$

where in the last step we used $n_1(s)/n \xrightarrow{p} p(s)/2$, $n(s)/n \xrightarrow{p} p(s)$, and that $n^{-1}\sum_{i=1}^n A_i I\{S_i = s\}\tilde{Y}_i(a) \xrightarrow{p} 0$ for $a \in \{0,1\}$ by similar arguments to those in the proof of Lemma B.3. Analogous arguments show that,

$$\frac{1}{n}\sum_{i=1}^n A_i Y_i = \frac{1}{2}\mu_1 + o_p(1) \ ,$$

so that we conclude that

$$\frac{1}{n}\mathbb{C}_n'\mathbb{Y}_n \xrightarrow{p} \frac{1}{2}\begin{bmatrix} \mu_1 \\ p(1)(\mu_1 + \mu_0 + E[m_1(Z) + m_0(Z)|S = 1]) \\ \vdots \\ p(|\mathcal{S}|)(\mu_1 + \mu_0 + E[m_1(Z) + m_0(Z)|S = |\mathcal{S}|]) \end{bmatrix} .$$

The result then follows from the above display, (A-60), and some additional algebra. ■

Lemma B.11. *Let Assumptions 2.1 and 2.2 hold, $C_i \equiv [A_i, I\{S_i = 1\}, \ldots, I\{S_i = |\mathcal{S}|\}]'$ be the ith row of the matrix $\mathbb{C}_n$ formed by stacking the treatment assignment vector $\mathbb{A}_n$ in the first row and the strata indicators vector in the rest of the rows, $\hat{u}_i$ be the least squares residuals of the regression in (29), and $\hat{\gamma}_n$ be the least squares estimator of the regression coefficients $\gamma = (\beta, \delta_1, \ldots, \delta_{|\mathcal{S}|})'$. Then,*

$$\hat{u}_i = \frac{1}{2}\sum_{s\in\mathcal{S}} I\{S_i = s\}A_i^* E\left[m_1(Z) - m_0(Z)|S = s\right] + \tilde{Y}_i(1)A_i + \tilde{Y}_i(0)(1 - A_i) + C_i(\gamma - \hat{\gamma}_n) \ .$$

Furthermore,

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^n \hat{u}_i^2 &\xrightarrow{p} \frac{1}{4}(\sigma_{\tilde{Y}}^2 + \sigma_H^2) \\
\frac{1}{n}\sum_{i=1}^n \hat{u}_i^2 A_i &\xrightarrow{p} \frac{1}{8}\sigma_H^2 + \frac{1}{2}\sigma_{\tilde{Y}(1)}^2 \\
\frac{1}{n}\sum_{i=1}^n \hat{u}_i^2 I\left\{S_i = s\right\} &\xrightarrow{p} \frac{1}{4}p(s)\left[\left(E\left[m_1(Z) - m_0(Z)|S = s\right]\right)^2 + 2(\sigma_{\tilde{Y}(1)}^2(s) + \sigma_{\tilde{Y}(0)}^2(s))\right] \\
\frac{1}{n}\sum_{i=1}^n \hat{u}_i^2 I\left\{S_i = s\right\}A_i &\xrightarrow{p} \frac{1}{4}p(s)\left[\frac{1}{2}\left(E\left[m_1(Z) - m_0(Z)|S = s\right]\right)^2 + 2\sigma_{\tilde{Y}(1)}^2(s)\right] \ .
\end{aligned}$$

Proof. Consider the following derivation,

$$\begin{aligned}
Y_i = Y_i(1)A_i + Y_i(0)(1 - A_i) &= \theta(Q)A_i + \tilde{Y}_i(1)A_i + \tilde{Y}_i(0)(1 - A_i) \\
&+ \sum_{s\in\mathcal{S}} I\{S_i = s\}\left[\mu_0 + E[m_1(Z) + m_0(Z)|S = s]/2\right] + \frac{A_i^*}{2}\sum_{s\in\mathcal{S}} I\{S_i = s\}E\left[m_1(Z) - m_0(Z)|S = s\right] \ .
\end{aligned}$$

Using Lemma B.10 and some algebra shows that

$$u_i = Y_i - C_i\gamma = \frac{1}{2}\sum_{s\in\mathcal{S}} I\{S_i = s\}A_i^* E\left[m_1(Z) - m_0(Z)|S = s\right] + \tilde{Y}_i(1)A_i + \tilde{Y}_i(0)\left(1 - A_i\right) \ , \quad \text{(B-85)}$$

and since $\hat{u}_i = u_i + C_i(\gamma - \hat{\gamma}_n)$, this proves the first part of the lemma.

To prove the second part we note that for any univariate random variable X_i such that

$$\frac{1}{n}\sum_{i=1}^n \left[C_i'C_i \otimes X_i\right] = O_p(1) \text{ and } \frac{1}{n}\sum_{i=1}^n C_i u_i X_i = O_p(1), \quad \text{(B-86)}$$

it follows that

$$\frac{1}{n}\sum_{i=1}^{n}\hat{u}_i^2X_i=\frac{1}{n}\sum_{i=1}^{n}u_i^2X_i+(\gamma-\hat{\gamma}_n)'\frac{1}{n}\sum_{i=1}^{n}\left[C_i'C_i\otimes X_i\right](\gamma-\hat{\gamma}_n)+2(\gamma-\hat{\gamma}_n)\frac{1}{n}\sum_{i=1}^{n}C_iu_iX_i=\frac{1}{n}\sum_{i=1}^{n}u_i^2X_i+o_p(1)\ ,$$

where we used $(\gamma-\hat{\gamma}_n)\xrightarrow{p}0$ from Lemma B.10. Since the condition in (B-86) certainly holds for $X_i=1$, $I\{S_i=s\}$, and A_i, we can focus on averages involving u_i^2 as opposed to $\hat{u}_i^2$ in what follows. We therefore need an expression for u_i^2. Using (B-85) and some algebra shows that,

$$\begin{aligned}u_i^2=&\frac{1}{4}\sum_{s\in\mathcal{S}}I\{S_i=s\}\left(E\left[m_1(Z)-m_0(Z)|S=s\right]\right)^2+\tilde{Y}_i(1)^2A_i+\tilde{Y}_i(0)^2\left(1-A_i\right)\\&+\sum_{s\in\mathcal{S}}I\{S_i=s\}A_i^*E\left[m_1(Z)-m_0(Z)|S=s\right]\tilde{Y}_i(1)A_i\\&+\sum_{s\in\mathcal{S}}I\{S_i=s\}A_i^*E\left[m_1(Z)-m_0(Z)|S=s\right]\tilde{Y}_i(0)(1-A_i)\ .\end{aligned}$$

We now use this expression to prove the four last statements of the lemma. In all cases, the convergence in probability holds from similar arguments to those used in the proof of Lemma B.1 so we omit them here.

The first average satisfies,

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}u_i^2&=\frac{1}{4}\sum_{s\in\mathcal{S}}p(s)(E[m_1(Z)-m_0(Z)|S=s])^2+\frac{1}{2}(\sigma_{\tilde{Y}(1)}^2+\sigma_{\tilde{Y}(0)}^2)+o_p(1)\\&=\frac{1}{4}(\sigma_{\tilde{Y}}^2+\sigma_H^2)+o_p(1)\ .\end{aligned}$$

The second average satisfies,

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}u_i^2A_i=&\frac{1}{4}\sum_{s\in\mathcal{S}}\frac{1}{n}\sum_{i=1}^{n}I\{S_i=s\}A_i\left(E\left[m_1(Z)-m_0(Z)|S=s\right]\right)^2+\frac{1}{n}\sum_{i=1}^{n}\tilde{Y}_i(1)^2A_i\\&+\sum_{s\in\mathcal{S}}\frac{1}{n}\sum_{i=1}^{n}\tilde{Y}_i(1)A_iI\{S_i=s\}A_i^*E\left[m_1(Z)-m_0(Z)|S=s\right]\\=&\frac{1}{8}\sigma_H^2+\frac{1}{2}\sigma_{\tilde{Y}(1)}^2+o_p(1)\ .\end{aligned}$$

The third average satisfies,

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}u_i^2I\left\{S_i=s\right\}=&\frac{1}{n}\sum_{i=1}^{n}I\{S_i=s\}\left(E\left[m_1(Z)-m_0(Z)|S=s\right]\right)^2/4\\&+\frac{1}{n}\sum_{a\in\{0,1\}}\sum_{i=1}^{n}\tilde{Y}_i(a)^2I\{S_i=s\}I\{A_i=a\}\\&+\frac{1}{n}\sum_{a\in\{0,1\}}\sum_{i=1}^{n}I\{S_i=s\}A_i^*E\left[m_1(Z)-m_0(Z)|S=s\right]\tilde{Y}_i(a)I\{A_i=a\}\\=&\frac{p(s)}{4}\left[\left(E\left[m_1(Z)-m_0(Z)|S=s\right]\right)^2+2\left(\sigma_{\tilde{Y}(1)}^2(s)+\sigma_{\tilde{Y}(0)}^2(s)\right)\right]+o_p(1)\ .\end{aligned}$$

The last average is similar to the third one and so we omit it here. This completes the proof. ■

References

Baldi Antognini, A. (2008). A theoretical analysis of the power of biased coin designs. *Journal of Statistical Planning and Inference*, **138** 1792–1798.

Berry, J., Karlan, D. S. and Pradhan, M. (2015). The impact of financial education for youth in Ghana. *Working paper*.

Bruhn, M. and McKenzie, D. (2008). In pursuit of balance: Randomization in practice in development field experiments. *World Bank Policy Research Working Paper*, **4752**.

Callen, M., Gulzar, S., Hasanain, A., Khan, Y. and Rezaee, A. (2015). Personalities and public sector performance: Evidence from a health experiment in Pakistan. Tech. rep., Working paper.

Chung, E. and Romano, J. P. (2013). Exact and asymptotically robust permutation tests. *The Annals of Statistics*, **41** 484–507.

Dizon-Ross, R. (2014). Parents' perceptions and children's education: Experimental evidence from Malawi. Manuscript, M.I.T.

Duflo, E., Dupas, P. and Kremer, M. (2014). Education, HIV, and early fertility: Experimental evidence from kenya. Tech. rep., National Bureau of Economic Research.

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.

Hallstrom, A. and Davis, K. (1988). Imbalance in treatment assignments in stratified blocked randomization. *Controlled Clinical Trials*, **9** 375–382.

Heckman, J. J., Pinto, R., Shaikh, A. M. and Yavitz, A. (2011). Inference with imperfect randomization: The case of the Perry Preschool. Manuscript.

Hoeffding, W. (1952). The large-sample power of tests based on permutations of observations. *The Annals of Mathematical Statistics*, **23** pp. 169–192. URL http://www.jstor.org/stable/2236445.

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

Imbens, G. W. and Kolesar, M. (2012). Robust standard errors in small samples: some practical advice. Tech. rep., National Bureau of Economic Research.

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

Janssen, A. (1997). Studentized permutation tests for non-iid hypotheses and the generalized Behrens-Fisher problem. *Statistics & probability letters*, **36** 9–21.

Lehmann, E. and Romano, J. P. (2005). *Testing Statistical Hypotheses*. 3rd ed. Springer, New York.

Lock Morgan, K. and Rubin, D. (2012). Rerandomization to improve covariate balance in experiments. *The Annals of Statistics*, **40** 1263–1282.

Markaryan, T. and Rosenberger, W. (2010). Exact properties of efron's biased coin randomization procedure. *The Annals of Statistics*, **38** 1546–1567.

Petrov, V. (2011). On the strong law of large numbers for a sequence of nonnegative random variables. *Journal of Mathematical Sciences*, **176** 207–208.

Pocock, S. and Simon, R. (1975). Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial. *Biometrics* 103–115.

Rosenbaum, P. R. (2007). Interference between units in randomized experiments. *Journal of the American Statistical Association*, **102**.

Rosenberger, W. F. and Lachin, J. M. (2004). *Randomization in clinical trials: theory and practice*. John Wiley & Sons.

Shao, J., Yu, X. and Zhong, B. (2010). A theory for testing hypotheses under covariate-adaptive randomization. *Biometrika*, **97** 347–360.

Wei, L. (1978). The adaptive biased coin design for sequential experiments. *The Annals of Statistics*, **6** 92–100.

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