Essays on Causal Inference in Randomized Experiments

by

Winston Lin

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy

in

Statistics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Jasjeet S. Sekhon, Co-chair
Professor Terence P. Speed, Co-chair
Professor Dylan S. Small
Professor Deborah Nolan
Professor Justin McCrary

Spring 2013

Essays on Causal Inference in Randomized Experiments

Abstract

Essays on Causal Inference in Randomized Experiments

by

Winston Lin

Doctor of Philosophy in Statistics

University of California, Berkeley

Professor Jasjeet S. Sekhon, Co-chair

Professor Terence P. Speed, Co-chair

This dissertation explores methodological topics in the analysis of randomized experiments, with a focus on weakening the assumptions of conventional models.

Chapter 1 gives an overview of the dissertation, emphasizing connections with other areas of statistics (such as survey sampling) and other fields (such as econometrics and psychometrics).

Chapter 2 reexamines Freedman's critique of ordinary least squares regression adjustment in randomized experiments. Using Neyman's model for randomization inference, Freedman argued that adjustment can lead to worsened asymptotic precision, invalid measures of precision, and small-sample bias. This chapter shows that in sufficiently large samples, those problems are minor or easily fixed. OLS adjustment cannot hurt asymptotic precision when a full set of treatment–covariate interactions is included. Asymptotically valid confidence intervals can be constructed with the Huber–White sandwich standard error estimator. Checks on the asymptotic approximations are illustrated with data from a randomized evaluation of strategies to improve college students' achievement. The strongest reasons to support Freedman's preference for unadjusted estimates are transparency and the dangers of specification search.

Chapter 3 extends the discussion and analysis of the small-sample bias of OLS adjustment. The leading term in the bias of adjustment for multiple covariates is derived and can be estimated empirically, as was done in Chapter 2 for the single-covariate case. Possible implications for choosing a regression specification are discussed.

Chapter 4 explores and modifies an approach suggested by Rosenbaum for analysis of treatment effects when the outcome is censored by death. The chapter is motivated by a randomized trial that studied the effects of an intensive care unit staffing intervention on length of stay in the ICU. The proposed approach estimates effects on the distribution of a composite outcome measure based on ICU mortality and survivors' length of stay, addressing concerns about selection bias by

comparing the entire treatment group with the entire control group. Strengths and weaknesses of possible primary significance tests (including the Wilcoxon–Mann–Whitney rank sum test and a heteroskedasticity-robust variant due to Brunner and Munzel) are discussed and illustrated.

For my mother and in memory of my father

Acknowledgments

I owe many thanks to Jas Sekhon, Terry Speed, and Dylan Small for their kind advice and support. All three of them helped me a great deal with large and small matters during my years in graduate school. Jas made my studies at Berkeley more interesting and enjoyable, encouraged me to write Chapter 2 when I would otherwise have given up, and helped me think about the right questions to ask and how to write about them. Terry introduced our class to sampling, ratio estimators, and Cochran's wonderful book and gave me valuable feedback on all the dissertation essays, but at least as valuable have been his kindness, wisdom, and humor. Dylan very generously gave me thoughtful comments on outlines and drafts of Chapter 2, suggested the topic of Chapter 4 and guided my work on it, and introduced me to many interesting papers.

I only met David Freedman once, but he was very generous to me with unsolicited help and advice after I sent him comments on three papers. He encouraged me to study at Berkeley even though (or perhaps because) he knew my thoughts on adjustment were not the same as his. (As always, he was also a realist: "You have to understand that the Ph.D. program is a genteel version of Marine boot camp. Some useful training, some things very interesting, but a lot of drill and hazing.") I remain a big fan of his oeuvre, and I hope it's clear from Chapter 2's "Further remarks" and final footnote that the chapter is meant as not only a dissent, but also a tribute. His books and papers have been a great pleasure to read and extremely valuable in my education, thanks to the care he took as a scholar, writer, and teacher.

Erin Hartman and Danny Hidalgo were wonderful Graduate Student Instructors and gave me valuable comments and advice in the early stages of my work on Chapter 2. I am also grateful to Deb Nolan and Justin McCrary for helpful conversations and for serving on my exam and dissertation committees. Deb organized my qualifying exam, asked thoughtful questions, and helped me fit my unwieldy talk into the time available. Justin kindly allowed me to audit his very enjoyable course in empirical methods, which helped keep me sane (I hope) during a heavy semester. Pat Kline's applied econometrics course was also very interesting and useful for my research. I have greatly enjoyed discussions with my classmates Alex Mayer and Luke Miratrix, and I appreciate all their help and friendship.

Chapter 2 is reprinted from *Annals of Applied Statistics*, vol. 7, no. 1 (March 2013), pp. 295–318, with permission from the Institute of Mathematical Statistics. I had valuable discussions with many people who are acknowledged in the published article, and with Richard Berk and seminar participants at Abt Associates and MDRC after it went to press.

Chapters 3 and 4 are motivated by ongoing collaborations with Peter Aronow, Don Green, and Jas Sekhon on regression adjustment and with Scott Halpern, Meeta Prasad Kerlin, and Dylan Small on ICU length of stay. I am grateful to all of them for their insights and interest, but any errors are my own.

Beth Cooney, Nicole Gabler, and Michael Harhay provided data for Chapter 4, and Tamara Broderick kindly helped with a query about notation. Paul Rosenbaum was helpful in an early discussion of the topic.

Ani Adhikari, Roman Yangarber, Monica Yin, Howard Bloom, Johann Gagnon-Bartsch, and Ralph Grishman were generous with encouragement, advice, and help when I was considering graduate school.

I am deeply grateful to all my family and friends for their support, especially Jee Leong Koh and Mingyew Leung. Most of all, I would like to thank my parents for all their love, care, and sacrifices and for everything they have taught me.

Contents

List of Tables

Chapter 1

Overview

The essays in this dissertation are about the statistics of causal inference in randomized experiments, but they draw on ideas from other branches of statistics and other fields. In presentations to public policy researchers, I've mentioned an excellent essay by the economist Joshua Angrist (2004) on the rise of randomized experiments in education research. Commenting on the roles of outsiders from economics, psychology, and other fields in this quiet revolution, Angrist writes that "education research is too important to be left entirely to professional education researchers." Those may be fighting words, but I like to draw a conciliatory lesson: Almost any community can benefit from an outside perspective. Statistics is too important to be left entirely to statisticians, and causal inference is too important to be left entirely to causal inference researchers.

1.1 Regression adjustment

David Freedman was a great statistician and probabilist, but he argued for more humility about what statistics can accomplish. One of his many insightful essays is a critique of the use of regression for causal inference in observational studies [Freedman (1991)]. Four of his final publications extend his critique to ordinary least squares regression adjustment in randomized experiments [Freedman (2008ab)], logistic and probit regression in experiments [Freedman (2008d)], and proportional hazards regression in experiments and observational studies [Freedman (2010)]. Chapter 2 of this dissertation responds to Freedman (2008ab) on OLS adjustment.[1]

Random assignment is intended to create comparable treatment and control groups, reducing the need for dubious statistical models. Nevertheless, researchers often use linear regression models to "adjust" for random treatment–control differ-

[1] I largely agree with the other papers in Freedman's quartet. Some of the issues with logits and probits are also discussed in Firth and Bennett (1998), Lin (1999), Gelman and Pardoe (2007), and pp. 323–324 of my Appendix D to Bloom et al. (1993).

ences in baseline characteristics. The classic rationale, which assumes the regression model is true, is that adjustment tends to improve precision if the covariates are correlated with the outcome and the sample size is much larger than the number of covariates [e.g., Cox and McCullagh (1982)]. In contrast, Freedman (2008ab) uses Neyman's (1923) potential outcomes framework for randomization inference, avoiding dubious assumptions about functional forms, error terms, and homogeneous treatment effects. He shows that (i) adjustment can actually hurt asymptotic precision; (ii) the conventional OLS standard error estimator is inconsistent; and (iii) the adjusted treatment effect estimator has a small-sample bias. He writes, "The reason for the breakdown is not hard to find: randomization does not justify the assumptions behind the OLS model."

Chapter 2 argues that in sufficiently large samples, the problems Freedman raised are minor or easily fixed. Under the Neyman model with Freedman's regularity conditions, I show that (i) OLS adjustment cannot hurt asymptotic precision when a full set of treatment $\times$ covariate interactions is included, and (ii) the Huber–White sandwich standard error estimator is consistent or asymptotically conservative. I briefly discuss the small-sample bias issue, and I give an empirical example to illustrate methods for estimating the bias and checking the validity of confidence intervals.

The theorems in Chapter 2 are not its only goal.[2] The chapter also offers intuition and perspective on Freedman's and my results, borrowing insights from econometrics and survey sampling. In econometrics, regression is sometimes studied and taught from an "agnostic" view that assumes random sampling from an infinite population but does not assume a regression model. As Goldberger (1991, p. xvi) writes, "Whether a regression specification is 'right' or 'wrong' ... one can consider whether or not the population feature that [least squares] does consistently estimate is an interesting one." Moreover, the sandwich standard error estimator remains consistent [Chamberlain (1982, pp. 17–19)]. This view of regression is not often taught in statistics, although Buja et al. (2012) and Berk et al. (2013) are notable recent exceptions.

In survey sampling, the design-based, model-assisted approach studies regression-adjusted estimators of population means in a similar spirit [Cochran (1977); Särndal, Swennson, and Wretman (1992); Fuller (2002)]. Adjustment may achieve greater precision improvement when the regression model fits well, but as Särndal et al. write (p. 227): "The basic properties (approximate unbiasedness, validity of the variance formulas, etc.) ... are not dependent on whether the model ξ holds or not. Our procedures are thus *model assisted*, but they are not model dependent."

[2]The mathematician William Thurston argued against overemphasis on "theorem-credits," writing that "we would be much better off recognizing and valuing a far broader range of activity" [Thurston (1994)]. Rereading math textbooks after the field had "come alive" for him, he "was stunned by how well their formalism and indirection hid the motivation, the intuition and the multiple ways to think about their subjects: they were unwelcoming to the full human mind" [Thurston (2006)].

I argue that the parallels between regression adjustment in experiments and regression estimators in sampling are underexplored and that the sampling analogy naturally suggests adjustment with treatment $\times$ covariate interactions.[3]

Chapter 2 is not designed to serve as a guide to practice, although I hope it gives some helpful input for future guides to practice. It focuses on putting Freedman's critique in perspective and responding to the specific theoretical issues he raised. I give a bit more discussion of practical implications in a companion blog essay [Lin (2012ab)].

Chapter 3 gives additional results on the small-sample bias of OLS adjustment, which received less attention in Chapter 2 than Freedman's other two issues. In Chapter 2, I showed how to estimate the leading term in the bias of OLS adjustment for a single covariate (with and without the treatment $\times$ covariate interaction), using the sample analogs of asymptotic formulas from Cochran (1977) and Freedman (2008b). Chapter 3 derives and discusses the leading term in the bias of adjustment for multiple covariates, which turns out to involve the diagonal elements of the hat matrix [Hoaglin and Welsch (1978)] and can be estimated from the data. The theoretical expression for the leading term may also be relevant to choosing a regression specification when the sample is small.

As Efron and Tibshirani (1993, p. 138) write in the bootstrap literature, "Bias estimation is usually interesting and worthwhile, but the exact use of a bias estimate is often problematic." Using a bias estimate to "correct" the original estimator can do more harm than good: the reduction in bias is often outweighed by an increase in variance. Thus, I am only suggesting bias estimation for a ballpark idea of whether small-sample bias is a serious problem.

1.2 Censoring by death and the nonparametric Behrens–Fisher problem

Chapter 4 is motivated by a specific application, but focuses on methodological issues that may be of broader interest. The SUNSET-ICU trial [Kerlin et al. (2013)] studied the effectiveness of 24-hour staffing by intensivist physicians in an intensive care unit, compared to having intensivists available in person during the day and by phone at night. The primary outcome was length of stay in the ICU. (Longer ICU stays are associated with increased stress and discomfort for patients and their families, as well as increased costs for patients, hospitals, and society.) A significant proportion of patients die in the ICU, and there are no reliable ways to disentangle an intervention's effects on length of stay from its effects on mortality. Conventional approaches (e.g., analyzing only survivors, pooling survivors

[3] Fienberg and Tanur (1987, 1996) discuss many parallels between experiments and sampling and argue that the two fields drifted apart because of the rift between R. A. Fisher and Jerzy Neyman.

and non-survivors, or proportional hazards modeling) depend on assumptions that are often unstated and difficult to interpret or check.

Chapter 4 explores an approach adapted from Rosenbaum (2006) that avoids selection bias and makes its assumptions explicit. In our context, the approach requires a "placement of death" relative to survivors' possible lengths of stay, such as "Death in the ICU is the worst possible outcome" or "Death in the ICU and a survivor's 100-day ICU stay are considered equally undesirable." Given a placement of death, we can compare the entire treatment group with the entire control group to estimate the intervention's effects on the median outcome and other quantiles. As researchers, we cannot decide the appropriate placement of death, but we can show how the results vary over a range of placements.

Rosenbaum's original proposal appeared in a comment on Rubin (2006a) and has not been used in empirical studies (to my knowledge). Rosenbaum derives exact, randomization-based confidence intervals for a nonstandard estimand; as Rubin (2006b) notes, the proposal is "deep and creative" but may be "difficult to convey to consumers." Chapter 4 discusses ways to construct approximate confidence intervals for more familiar estimands (treatment effects on quantiles of the outcome distribution or on proportions of patients with outcomes better than various cutoff values). Simulation evidence on the validity of bootstrap confidence intervals for quantile treatment effects is presented.

Recommended practice for analysis of clinical trials includes pre-specification of a primary outcome measure. As stated in the CONSORT explanation and elaboration document, "Having several primary outcomes ... incurs the problems of interpretation associated with multiplicity of analyses ... and is not recommended" [Moher et al. (2010, p. 7)]. In the approach of Chapter 4, the same principle may suggest designating one quantile as primary. The median may seem a natural choice, but some interventions may be intended to shorten long ICU stays without necessarily reducing the median. It may be difficult to predict which points in the outcome distribution are likely to be affected.

An alternative strategy is to pre-specify that the primary significance test is a rank test with some sensitivity to effects throughout the outcome distribution.[4] Rubin (2006b) comments that the Wilcoxon–Mann–Whitney rank sum test could be combined with Rosenbaum's approach. More broadly, the econometricians Guido Imbens and Jeffrey Wooldridge (2009, pp. 21–23) suggest the Wilcoxon test as an omnibus test for "establishing whether the treatment has any effect" in randomized experiments. Imbens has explained his views in presentations and in blog comments that merit quoting at length:

- "Why then do I think it is useful to do the randomization test using average ranks as the statistic instead of doing a t-test? I think rather than being in-

[4] In general I agree with the notion that confidence intervals should be preferred to tests. In Chapter 4's empirical example, I report Brunner and Munzel's (2000) test together with a confidence interval for the associated estimand.

terested very specifically in the question whether the average effect differs from zero, one is typically interested in the question whether there is evidence that there is a positive (or negative) effect of the treatment. That is a little vague, but more general than simply a non-zero average effect. If we can't tell whether the average effect differs from zero, but we can be confident that the lower tail of the distribution moves up, that would be informative. I think this vaguer null is well captured by looking at the difference in average ranks: do the treated have higher ranks on average than the controls. I would interpret that as implying that the treated have typically higher outcomes than the controls (not necessarily on average, but typically)." [Imbens (2011a)]

- "Back to the randomization tests. Why do I like them? I think they are a good place to start an analysis. If you have a randomized experiment, and you find that using a randomization test based on ranks that there is little evidence of any effect of the treatment, I would be unlikely to be impressed by any model-based analysis that claimed to find precise non-zero effects of the treatment. It is possible, and the treatment could affect the dispersion and not the location, but in most cases if you don't find any evidence of any effects based on that single randomization based test, I think you can stop right there. I see the test not so much as answering whether in the population the effects are all zero (not so interesting), rather as answering the question whether the data are rich enough to make precise inferences about the effects." [Imbens (2011b)]

I think Imbens's advice is very well thought out, but I would prefer a different test. Chapter 4 discusses the properties of the Wilcoxon test and a heteroskedasticity-robust variant due to Brunner and Munzel (2000). The Wilcoxon test is valid for the strong null hypothesis that treatment has no effect on any patient, but whether researchers should be satisfied with a test of the strong null is debatable. The Mann–Whitney form of the test statistic naturally suggests the weaker null hypothesis that if we sample the treated and untreated potential outcome distributions independently, a random outcome under treatment is equally likely to be better or worse than a random outcome in the absence of treatment. There is an interesting, somewhat neglected literature on the "nonparametric Behrens–Fisher problem" of testing the weak null, extending from Pratt (1964) to recent work by the econometrician EunYi Chung and the statistician Joseph Romano [Romano (2009); Chung and Romano (2011)].[5]

The chapter gives simulations that illustrate and support Pratt's (1964) asymptotic analysis. The Wilcoxon test is not a valid test of the weak null, even when the

[5]This literature is not explicitly causal. An example of a descriptive application is the null hypothesis that a random Australian is equally likely to be taller or shorter than a random Canadian. The psychometrician Andrew Ho (2009) gives a helpful discussion of a related literature on nonparametric methods for comparing test score distributions, trends, and gaps.

design is balanced. It is valid for the strong null, but it is sensitive to certain kinds of departures from the strong null and not others. These properties complicate the test's interpretation and are probably not well-known to most of its users. In contrast, the Brunner–Munzel test is an approximately valid test of the weak null in sufficiently large samples.[6] In simulations based on the SUNSET-ICU trial data, the two tests have approximately equal power.

An illustrative example reanalyzes the SUNSET-ICU data. I find no evidence that the intervention affected the distribution of patients' outcomes, regardless of whether death is considered the worst possible outcome or placed as comparable to a length of stay as short as 30 days. Since there was little difference in ICU mortality between the treatment and control groups, it is not surprising that this conclusion is similar to the original findings of Kerlin et al. (2013).

It should be noted that Chapter 4's placement-of-death approach does not estimate treatment effects on ICU length of stay per se. Instead, it estimates effects on the distribution of a composite outcome measure based on ICU mortality and survivors' lengths of stay. Researchers may understandably want to disentangle effects on length of stay from effects on mortality, but opinions may differ on whether this can be done persuasively, since stronger assumptions would be needed. Thus, the placement-of-death approach does not answer all relevant questions, but it may be a useful starting point. It addresses concerns about selection bias by comparing the entire treatment group with the entire control group, and it can provide evidence of an overall beneficial or harmful effect.

[6] Neubert and Brunner (2007) propose a permutation test based on the Brunner–Munzel statistic. Their test is exact for the strong null and asymptotically valid for the weak null.

Chapter 2

Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique

2.1 Introduction

One of the attractions of randomized experiments is that, ideally, the strength of the design reduces the need for statistical modeling. Simple comparisons of means can be used to estimate the average effects of assigning subjects to treatment. Nevertheless, many researchers use linear regression models to adjust for random differences between the baseline characteristics of the treatment groups. The usual rationale is that adjustment tends to improve precision if the sample is large enough and the covariates are correlated with the outcome; this argument, which assumes that the regression model is correct, stems from Fisher (1932) and is taught to applied researchers in many fields. At research firms that conduct randomized experiments to evaluate social programs, adjustment is standard practice.[1]

In an important and influential critique, Freedman (2008ab) analyzes the behavior of ordinary least squares regression-adjusted estimates without assuming a regression model. He uses Neyman's (1923) model for randomization inference: treatment effects can vary across subjects, linearity is not assumed, and random assignment is the source of variability in estimated average treatment effects. Freedman shows that (i) adjustment can actually worsen asymptotic precision, (ii) the conventional OLS standard error estimator is inconsistent, and (iii) the adjusted treatment effect estimator has a small-sample bias. He writes [Freedman (2008a)], "The reason for the breakdown is not hard to find: randomization does not justify

[1] Cochran (1957), Cox and McCullagh (1982), Raudenbush (1997), and Klar and Darlington (2004) discuss precision improvement. Greenberg and Shroder (2004) document the use of regression adjustment in many randomized social experiments.

the assumptions behind the OLS model."

This chapter offers an alternative perspective. Although I agree with Freedman's (2008b) general advice ("Regression estimates ... should be deferred until rates and averages have been presented"), I argue that in sufficiently large samples, the statistical problems he raised are either minor or easily fixed. Under the Neyman model with Freedman's regularity conditions, I show that (i) OLS adjustment cannot hurt asymptotic precision when a full set of treatment $\times$ covariate interactions is included, and (ii) the Huber–White sandwich standard error estimator is consistent or asymptotically conservative (regardless of whether the interactions are included). I also briefly discuss the small-sample bias issue and the distinction between unconditional and conditional unbiasedness.

Even the traditional OLS adjustment has benign large-sample properties when subjects are randomly assigned to two groups of equal size. Freedman (2008a) shows that in this case, adjustment (without interactions) improves or does not hurt asymptotic precision, and the conventional standard error estimator is consistent or asymptotically conservative. However, Freedman and many excellent applied statisticians in the social sciences have summarized his papers in terms that omit these results and emphasize the dangers of adjustment. For example, Berk et al. (2010) write: "Random assignment does not justify any form of regression with covariates. If regression adjustments are introduced nevertheless, there is likely to be bias in any estimates of treatment effects and badly biased standard errors."

One aim of this chapter is to show that such a negative view is not always warranted. A second aim is to help provide a more intuitive understanding of the properties of OLS adjustment when the regression model is incorrect. An "agnostic" view of regression [Angrist and Imbens (2002); Angrist and Pischke (2009, ch. 3)] is adopted here: without taking the regression model literally, we can still make use of properties of OLS that do not depend on the model assumptions.

Precedents

Similar results on the asymptotic precision of OLS adjustment with interactions are proved in interesting and useful papers by Yang and Tsiatis (2001), Tsiatis et al. (2008), and Schochet (2010), under the assumption that the subjects are a random sample from an infinite superpopulation.[2] These results are not widely known, and Freedman was apparently unaware of them. He did not analyze adjustment with interactions, but conjectured, "Treatment by covariate interactions can probably be accommodated too" [Freedman (2008b, p. 186)].

Like Freedman, I use the Neyman model, in which random assignment of a finite population is the sole source of randomness; for a thoughtful philosophical

[2] Although Tsiatis et al. write that OLS adjustment *without* interactions "is generally more precise than ... the difference in sample means" (p. 4661), Yang and Tsiatis's asymptotic variance formula correctly implies that this adjustment may help or hurt precision.

discussion of finite- vs. infinite-population inference, see Reichardt and Gollob (1999, pp. 125–127). My purpose is not to advocate finite-population inference, but to show just how little needs to be changed to address Freedman's major concerns. The results may help researchers understand why and when OLS adjustment can backfire. In large samples, the essential problem is omission of treatment $\times$ covariate interactions, not the linear model. With a balanced two-group design, even that problem disappears asymptotically, because two wrongs make a right (underadjustment of one group mean cancels out overadjustment of the other).

Neglected parallels between regression adjustment in experiments and regression estimators in survey sampling turn out to be very helpful for intuition.

2.2 Basic framework

For simplicity, the main results in this chapter assume a completely randomized experiment with two treatment groups (or a treatment group and a control group), as in Freedman (2008a). Results for designs with more than two groups are discussed informally.

The Neyman model with covariates

The notation is adapted from Freedman (2008b). There are n subjects, indexed by $i = 1, \ldots, n$. We assign a simple random sample of fixed size n_A to treatment A and the remaining $n - n_A$ subjects to treatment B. For each subject, we observe an outcome Y_i and a row vector of covariates $\mathbf{z}_i = (z_{i1}, \ldots, z_{iK})$, where $1 \leq K < \min(n_A, n - n_A) - 1$. Treatment does not affect the covariates.

Assume that each subject has two potential outcomes [Neyman (1923); Rubin (1974, 2005); Holland (1986)], a_i and b_i, which would be observed under treatments A and B, respectively.[3] Thus, the observed outcome is $Y_i = a_i T_i + b_i(1 - T_i)$, where T_i is a dummy variable for treatment A.

Random assignment is the sole source of randomness in this model. The n subjects are the population of interest; they are not assumed to be randomly drawn from a superpopulation. For each subject, a_i, b_i, and $\mathbf{z}_i$ are fixed, but T_i and thus Y_i are random.

Let $\overline{a}$, $\overline{a}_A$, and $\overline{a}_B$ denote the means of a_i over the population, treatment group A, and treatment group B:

$$\overline{a} = \frac{1}{n} \sum_{i=1}^{n} a_i, \qquad \overline{a}_A = \frac{1}{n_A} \sum_{i \in A} a_i, \qquad \overline{a}_B = \frac{1}{n - n_A} \sum_{i \in B} a_i.$$

Use similar notation for the means of b_i, Y_i, $\mathbf{z}_i$, and other variables.

[3] Most authors use notation such as $Y_i(1)$ and $Y_i(0)$, or Y_{1i} and Y_{0i}, for potential outcomes. Freedman's (2008b) choice of a_i and b_i helps make the finite-population asymptotics more readable.

Our goal is to estimate the average treatment effect of A relative to B:

$$\text{ATE} = \overline{a} - \overline{b}.$$

Estimators of average treatment effect

The unadjusted or difference-in-means estimator of ATE is

$$\widehat{\text{ATE}}_{\text{unadj}} = \overline{Y}_A - \overline{Y}_B = \overline{a}_A - \overline{b}_B.$$

The usual OLS-adjusted estimator of ATE is the estimated coefficient on T_i in the OLS regression of Y_i on T_i and $\mathbf{z}_i$. (All regressions described in this chapter include intercepts.) Let $\widehat{\text{ATE}}_{\text{adj}}$ denote this estimator.

A third estimator, $\widehat{\text{ATE}}_{\text{interact}}$, can be computed as the estimated coefficient on T_i in the OLS regression of Y_i on T_i, $\mathbf{z}_i$, and $T_i(\mathbf{z}_i - \overline{\mathbf{z}})$. Section 2.3 motivates this estimator by analogy with regression estimators in survey sampling. In the context of observational studies, Imbens and Wooldridge (2009, pp. 28–30) give a theoretical analysis of $\widehat{\text{ATE}}_{\text{interact}}$, and a related method is known as the Peters–Belson or Oaxaca–Blinder estimator.[4] When $\mathbf{z}_i$ is a set of indicators for the values of a categorical variable, $\widehat{\text{ATE}}_{\text{interact}}$ is equivalent to subclassification or poststratification [Miratrix, Sekhon, and Yu (2013)].

2.3 Connections with sampling

Cochran (1977, ch. 7) gives a very readable discussion of regression estimators in sampling.[5] In one example [Watson (1937)], the goal was to estimate $\overline{y}$, the average surface area of the leaves on a plant. Measuring a leaf's area is time-consuming, but its weight can be found quickly. So the researcher weighed all the leaves, but measured area for only a small sample. In simple random sampling, the sample mean area $\overline{y}_S$ is an unbiased estimator of $\overline{y}$. But $\overline{y}_S$ ignores the auxiliary data on leaf weights. The sample and population mean weights ($\overline{z}_S$ and $\overline{z}$) are both known, and if $\overline{z} > \overline{z}_S$, then we expect that $\overline{y} > \overline{y}_S$. This motivates a "linear regression estimator"

$$\widehat{\overline{y}}_{\text{reg}} = \overline{y}_S + q(\overline{z} - \overline{z}_S) \tag{2.3.1}$$

where q is an adjustment factor. One way to choose q is to regress leaf area on leaf weight in the sample.

Regression adjustment in randomized experiments can be motivated analogously under the Neyman model. The potential outcome a_i is measured for only a simple random sample (treatment group A), but the covariates $\mathbf{z}_i$ are measured for the

[4] See Cochran (1969), Rubin (1984), and Kline (2011). Hansen and Bowers (2009) analyze a randomized experiment with a variant of the Peters–Belson estimator derived from logistic regression.

[5] See also Fuller (2002, 2009).

whole population (the n subjects). The sample mean $\bar{a}_A$ is an unbiased estimator of $\bar{a}$, but it ignores the auxiliary data on $\mathbf{z}_i$. If the covariates are of some help in predicting a_i, then another estimator to consider is

$$\hat{\bar{a}}_{\text{reg}} = \bar{a}_A + (\bar{\mathbf{z}} - \bar{\mathbf{z}}_A)\mathbf{q}_a \tag{2.3.2}$$

where $\mathbf{q}_a$ is a $K \times 1$ vector of adjustment factors. Similarly, we can consider using

$$\hat{\bar{b}}_{\text{reg}} = \bar{b}_B + (\bar{\mathbf{z}} - \bar{\mathbf{z}}_B)\mathbf{q}_b \tag{2.3.3}$$

to estimate $\bar{b}$ and then $\hat{\bar{a}}_{\text{reg}} - \hat{\bar{b}}_{\text{reg}}$ to estimate ATE $= \bar{a} - \bar{b}$.

The analogy suggests deriving $\mathbf{q}_a$ and $\mathbf{q}_b$ from OLS regressions of a_i on $\mathbf{z}_i$ in treatment group A and b_i on $\mathbf{z}_i$ in treatment group B—in other words, separate regressions of Y_i on $\mathbf{z}_i$ in the two treatment groups. The estimator $\hat{\bar{a}}_{\text{reg}} - \hat{\bar{b}}_{\text{reg}}$ is then just $\widehat{\text{ATE}}_{\text{interact}}$. If, instead, we use a pooled regression of Y_i on T_i and $\mathbf{z}_i$ to derive a single vector $\mathbf{q}_a = \mathbf{q}_b$, then we get $\widehat{\text{ATE}}_{\text{adj}}$.

Connections between regression adjustment in experiments and regression estimators in sampling have been noted but remain underexplored.[6] All three of the issues that Freedman raised have parallels in the sampling literature. Under simple random sampling, when the regression model is incorrect, OLS adjustment of the estimated mean still improves or does not hurt asymptotic precision [Cochran (1977)], consistent standard error estimators are available [Fuller (1975)], and the adjusted estimator of the mean has a small-sample bias [Cochran (1942)].

2.4 Asymptotic precision

Precision improvement in sampling

This subsection gives an informal argument, adapted from Cochran (1977), to show that in simple random sampling, OLS adjustment of the sample mean improves or does not hurt asymptotic precision, even when the regression model is incorrect. Regularity conditions and other technical details are omitted; the purpose is to motivate the results on completely randomized experiments in the next subsection.

First imagine using a "fixed-slope" regression estimator, where q in Eq. (2.3.1) is fixed at some value q_0 before sampling:

$$\hat{\bar{y}}_f = \bar{y}_S + q_0(\bar{z} - \bar{z}_S).$$

[6] Connections are noted by Fienberg and Tanur (1987), Hansen and Bowers (2009), and Middleton and Aronow (2012) but are not mentioned by Cochran despite his important contributions to both literatures. He takes a design-based (agnostic) approach in much of his work on sampling, but assumes a regression model in his classic overview of regression adjustment in experiments and observational studies [Cochran (1957)].

If $q_0 = 0$, $\widehat{\overline{y}}_f$ is just $\overline{y}_S$. More generally, $\widehat{\overline{y}}_f$ is the sample mean of $y_i - q_0(z_i - \overline{z})$, so its variance follows the usual formula with a finite-population correction:

$$\text{var}\left(\widehat{\overline{y}}_f\right) = \frac{N-n}{N-1}\frac{1}{n}\frac{1}{N}\sum_{i=1}^{N}\left[(y_i - \overline{y}) - q_0(z_i - \overline{z})\right]^2$$

where N is the population size and n is the sample size.

Thus, choosing q_0 to minimize the variance of $\widehat{\overline{y}}_f$ is equivalent to running an OLS regression of y_i on z_i in the population. The solution is the "population least squares" slope,

$$q_{\text{PLS}} = \frac{\sum_{i=1}^{N}(z_i - \overline{z})(y_i - \overline{y})}{\sum_{i=1}^{N}(z_i - \overline{z})^2},$$

and the minimum-variance fixed-slope regression estimator is

$$\widehat{\overline{y}}_{\text{PLS}} = \overline{y}_S + q_{\text{PLS}}(\overline{z} - \overline{z}_S).$$

Since the sample mean $\overline{y}_S$ is a fixed-slope regression estimator, it follows that $\widehat{\overline{y}}_{\text{PLS}}$ has lower variance than the sample mean, unless $q_{\text{PLS}} = 0$ (in which case $\widehat{\overline{y}}_{\text{PLS}} = \overline{y}_S$).

The actual OLS regression estimator is almost as precise as $\widehat{\overline{y}}_{\text{PLS}}$ in sufficiently large samples. The difference between the two estimators is

$$\widehat{\overline{y}}_{\text{OLS}} - \widehat{\overline{y}}_{\text{PLS}} = (\widehat{q}_{\text{OLS}} - q_{\text{PLS}})(\overline{z} - \overline{z}_S)$$

where $\widehat{q}_{\text{OLS}}$ is the estimated slope from a regression of y_i on z_i in the sample. The estimation errors $\widehat{q}_{\text{OLS}} - q_{\text{PLS}}$, $\overline{z}_S - \overline{z}$, and $\widehat{\overline{y}}_{\text{PLS}} - \overline{y}$ are of order $1/\sqrt{n}$ in probability. Thus, the difference $\widehat{\overline{y}}_{\text{OLS}} - \widehat{\overline{y}}_{\text{PLS}}$ is of order $1/n$, which is negligible compared to the estimation error in $\widehat{\overline{y}}_{\text{PLS}}$ when n is large enough.

In sum, in large enough samples,

$$\text{var}\left(\widehat{\overline{y}}_{\text{OLS}}\right) \approx \text{var}\left(\widehat{\overline{y}}_{\text{PLS}}\right) \leq \text{var}\left(\overline{y}_S\right)$$

and the inequality is strict unless y_i and z_i are uncorrelated in the population.

Precision improvement in experiments

The sampling result naturally leads to the conjecture that in a completely randomized experiment, OLS adjustment with a full set of treatment $\times$ covariate interactions improves or does not hurt asymptotic precision, even when the regression model is incorrect. The adjusted estimator $\widehat{\text{ATE}}_{\text{interact}}$ is just the difference between two OLS regression estimators from sampling theory, while $\widehat{\text{ATE}}_{\text{unadj}}$ is the difference between two sample means.

The conjecture is confirmed below. To summarize the results:

1. $\widehat{\text{ATE}}_{\text{interact}}$ is consistent and asymptotically normal (as are $\widehat{\text{ATE}}_{\text{unadj}}$ and $\widehat{\text{ATE}}_{\text{adj}}$, from Freedman's results).

2. Asymptotically, $\widehat{\text{ATE}}_{\text{interact}}$ is at least as efficient as $\widehat{\text{ATE}}_{\text{unadj}}$, and more efficient unless the covariates are uncorrelated with the weighted average

$$\frac{n-n_A}{n}a_i + \frac{n_A}{n}b_i.$$

3. Asymptotically, $\widehat{\text{ATE}}_{\text{interact}}$ is at least as efficient as $\widehat{\text{ATE}}_{\text{adj}}$, and more efficient unless (a) the two treatment groups have equal size or (b) the covariates are uncorrelated with the treatment effect $a_i - b_i$.

Assumptions for asymptotics

Finite-population asymptotic results are statements about randomized experiments on (or random samples from) an imaginary infinite sequence of finite populations, with increasing n. The regularity conditions (assumptions on the limiting behavior of the sequence) may seem vacuous, since one can always construct a sequence that contains the actual population and still satisfies the conditions. But it may be useful to ask whether a sequence that preserves any relevant "irregularities" (such as the influence of gross outliers) would violate the regularity conditions. See also Lumley (2010, pp. 217–218).

The asymptotic results in this chapter assume Freedman's (2008b) regularity conditions, generalized to allow multiple covariates; the number of covariates K is constant as n grows. One practical interpretation of these conditions is that in order for the results to be applicable, the size of each treatment group should be sufficiently large (and much larger than the number of covariates), the influence of outliers should be small, and near-collinearity in the covariates should be avoided.

As Freedman (2008a) notes, in principle, there should be an extra subscript to index the sequence of populations: for example, in the population with n subjects, the ith subject has potential outcomes $a_{i,n}$ and $b_{i,n}$, and the average treatment effect is ATE_n. Like Freedman, I drop the extra subscripts.

Condition 1. There is a bound $L < \infty$ such that for all $n = 1, 2, \ldots$ and $k = 1, \ldots, K$,

$$\frac{1}{n}\sum_{i=1}^{n} a_i^4 < L, \qquad \frac{1}{n}\sum_{i=1}^{n} b_i^4 < L, \qquad \frac{1}{n}\sum_{i=1}^{n} z_{ik}^4 < L.$$

Condition 2. Let $\mathbf{Z}$ be the $n \times (K+1)$ matrix whose ith row is $(1, \mathbf{z}_i)$. Then $n^{-1}\mathbf{Z}'\mathbf{Z}$ converges to a finite, invertible matrix. Also, the population means of a_i, b_i, a_i^2, b_i^2, $a_i b_i$, $a_i \mathbf{z}_i$, and $b_i \mathbf{z}_i$ converge to finite limits. For example, $\lim_{n\to\infty} n^{-1}\sum_{i=1}^{n} a_i \mathbf{z}_i$ exists and is a finite vector.

Condition 3. The proportion n_A/n converges to a limit p_A, with $0 < p_A < 1$.

Asymptotic results

Let $\mathbf{Q}_a$ denote the limit of the vector of slope coefficients in the population least squares regression of a_i on $\mathbf{z}_i$. That is,

$$\mathbf{Q}_a = \lim_{n\to\infty} \left[\left(\sum_{i=1}^{n} (\mathbf{z}_i - \overline{\mathbf{z}})'(\mathbf{z}_i - \overline{\mathbf{z}}) \right)^{-1} \sum_{i=1}^{n} (\mathbf{z}_i - \overline{\mathbf{z}})'(a_i - \overline{a}) \right].$$

Define $\mathbf{Q}_b$ analogously.

Now define the prediction errors

$$a_i^* = (a_i - \overline{a}) - (\mathbf{z}_i - \overline{\mathbf{z}})\mathbf{Q}_a, \qquad b_i^* = (b_i - \overline{b}) - (\mathbf{z}_i - \overline{\mathbf{z}})\mathbf{Q}_b$$

for $i = 1, \ldots, n$.

For any variables x_i and y_i, let σ_x^2 and $\sigma_{x,y}$ denote the population variance of x_i and the population covariance of x_i and y_i. For example,

$$\sigma_{a^*,b^*} = \frac{1}{n}\sum_{i=1}^{n} \left(a_i^* - \overline{a^*}\right)\left(b_i^* - \overline{b^*}\right) = \frac{1}{n}\sum_{i=1}^{n} a_i^* b_i^*.$$

Theorem 2.1 and its corollaries are proved in Appendix A.

Theorem 2.1. *Assume Conditions 1–3. Then* $\sqrt{n}(\widehat{\text{ATE}}_{\text{interact}} - \text{ATE})$ *converges in distribution to a Gaussian random variable with mean 0 and variance*

$$\frac{1-p_A}{p_A} \lim_{n\to\infty} \sigma_{a^*}^2 + \frac{p_A}{1-p_A} \lim_{n\to\infty} \sigma_{b^*}^2 + 2 \lim_{n\to\infty} \sigma_{a^*,b^*}.$$

Corollary 2.1.1. *Assume Conditions 1–3. Then* $\widehat{\text{ATE}}_{\text{unadj}}$ *has at least as much asymptotic variance as* $\widehat{\text{ATE}}_{\text{interact}}$. *The difference is*

$$\frac{1}{np_A(1-p_A)} \lim_{n\to\infty} \sigma_E^2$$

where $E_i = (\mathbf{z}_i - \overline{\mathbf{z}})\mathbf{Q}_E$ *and* $\mathbf{Q}_E = (1-p_A)\mathbf{Q}_a + p_A\mathbf{Q}_b$. *Therefore, adjustment with* $\widehat{\text{ATE}}_{\text{interact}}$ *helps asymptotic precision if* $\mathbf{Q}_E \neq \mathbf{0}$ *and is neutral if* $\mathbf{Q}_E = \mathbf{0}$.

Remarks. (i) $\mathbf{Q}_E$ can be thought of as a weighted average of $\mathbf{Q}_a$ and $\mathbf{Q}_b$, or as the limit of the vector of slope coefficients in the population least squares regression of $(1-p_A)a_i + p_A b_i$ on $\mathbf{z}_i$.

(ii) The weights may seem counterintuitive at first, but the sampling analogy and Eqs. (2.3.2–2.3.3) can help. Other things being equal, adjustment has a larger effect on the estimated mean from the smaller treatment group, because its mean covariate values are further away from the population mean. The adjustment added to $\overline{a}_A$ is

$$(\overline{\mathbf{z}} - \overline{\mathbf{z}}_A)\widehat{\mathbf{Q}}_a = \frac{n - n_A}{n}(\overline{\mathbf{z}}_B - \overline{\mathbf{z}}_A)\widehat{\mathbf{Q}}_a$$

while the adjustment added to $\overline{b}_B$ is

$$(\overline{\mathbf{z}} - \overline{\mathbf{z}}_B)\widehat{\mathbf{Q}}_b = -\frac{n_A}{n}(\overline{\mathbf{z}}_B - \overline{\mathbf{z}}_A)\widehat{\mathbf{Q}}_b,$$

where $\widehat{\mathbf{Q}}_a$ and $\widehat{\mathbf{Q}}_b$ are OLS estimates that converge to $\mathbf{Q}_a$ and $\mathbf{Q}_b$.

(iii) If the covariates' associations with a_i and b_i go in opposite directions, it is possible for adjustment with $\widehat{\text{ATE}}_{\text{interact}}$ to have no effect on asymptotic precision. Specifically, if $(1-p_A)\mathbf{Q}_a = -p_A\mathbf{Q}_b$, the adjustments to $\overline{a}_A$ and $\overline{b}_B$ tend to cancel each other out.

(iv) In designs with more than two treatment groups, estimators analogous to $\widehat{\text{ATE}}_{\text{interact}}$ can be derived from a separate regression in each treatment group, or equivalently a single regression with the appropriate treatment dummies, covariates, and interactions. The resulting estimator of (for example) $\overline{a} - \overline{b}$ is at least as efficient as $\overline{Y}_A - \overline{Y}_B$, and more efficient unless the covariates are uncorrelated with both a_i and b_i. Appendix A gives a proof.

Corollary 2.1.2. *Assume Conditions 1–3. Then* $\widehat{\text{ATE}}_{\text{adj}}$ *has at least as much asymptotic variance as* $\widehat{\text{ATE}}_{\text{interact}}$. *The difference is*

$$\frac{(2p_A-1)^2}{np_A(1-p_A)} \lim_{n\to\infty} \sigma_D^2$$

where $D_i = (\mathbf{z}_i - \overline{\mathbf{z}})(\mathbf{Q}_a - \mathbf{Q}_b)$. *Therefore, the two estimators have equal asymptotic precision if* $p_A = 1/2$ *or* $\mathbf{Q}_a = \mathbf{Q}_b$. *Otherwise,* $\widehat{\text{ATE}}_{\text{interact}}$ *is asymptotically more efficient.*

Remarks. (i) $\mathbf{Q}_a - \mathbf{Q}_b$ is the limit of the vector of slope coefficients in the population least squares regression of the treatment effect $a_i - b_i$ on $\mathbf{z}_i$.

(ii) For intuition about the behavior of $\widehat{\text{ATE}}_{\text{adj}}$, suppose there is a single covariate, z_i, and the population least squares slopes are $Q_a = 10$ and $Q_b = 2$. Let $\widehat{Q}$ denote the estimated coefficient on z_i from a pooled OLS regression of Y_i on T_i and z_i. In sufficiently large samples, $\widehat{Q}$ tends to fall close to $p_A Q_a + (1-p_A)Q_b$. Consider two cases:

- If the two treatment groups have equal size, then $\overline{z} - \overline{z}_B = -(\overline{z} - \overline{z}_A)$, so when $\overline{z} - \overline{z}_A = 1$, the ideal linear adjustment would add 10 to $\overline{a}_A$ and subtract 2 from $\overline{b}_B$. Instead, $\widehat{\text{ATE}}_{\text{adj}}$ uses the pooled slope estimate $\widehat{Q} \approx 6$, so it tends to underadjust $\overline{a}_A$ (adding about 6) and overadjust $\overline{b}_B$ (subtracting about 6). Two wrongs make a right: the adjustment adds about 12 to $\overline{a}_A - \overline{b}_B$, just as $\widehat{\text{ATE}}_{\text{interact}}$ would have done.

- If group A is 9 times larger than group B, then $\overline{z} - \overline{z}_B = -9(\overline{z} - \overline{z}_A)$, so when $\overline{z} - \overline{z}_A = 1$, the ideal linear adjustment adds 10 to $\overline{a}_A$ and subtracts $9 \cdot 2 = 18$ from $\overline{b}_B$, thus adding 28 to the estimate of ATE. In contrast, the pooled

adjustment adds $\widehat{Q} \approx 9.2$ to $\overline{a}_A$ and subtracts $9\widehat{Q} \approx 82.8$ from $\overline{b}_B$, thus adding about 92 to the estimate of ATE. The problem is that the pooled regression has more observations of a_i than of b_i, but the adjustment has a larger effect on the estimate of $\overline{b}$ than on that of $\overline{a}$, since group B's mean covariate value is further away from the population mean.

(iii) The example above suggests an alternative regression adjustment: when group A has nine-tenths of the subjects, give group B nine-tenths of the weight. More generally, let $\tilde{p}_A = n_A/n$. Run a weighted least squares regression of Y_i on T_i and $\mathbf{z}_i$, with weights of $(1-\tilde{p}_A)/\tilde{p}_A$ on each observation from group A and $\tilde{p}_A/(1-\tilde{p}_A)$ on each observation from group B. This "tyranny of the minority" estimator is asymptotically equivalent to $\widehat{\text{ATE}}_{\text{interact}}$ (Appendix A outlines a proof). It is equal to $\widehat{\text{ATE}}_{\text{adj}}$ when $\tilde{p}_A = 1/2$.

(iv) The tyranny estimator can also be seen as a one-step variant of Rubin and van der Laan's (2011) two-step "targeted ANCOVA." Their estimator is equivalent to the difference in means of the residuals from a weighted least squares regression of Y_i on $\mathbf{z}_i$, with the same weights as in remark (iii).

(v) When is the usual adjustment worse than no adjustment? Eq. (23) in Freedman (2008a) implies that with a single covariate z_i, for $\widehat{\text{ATE}}_{\text{adj}}$ to have higher asymptotic variance than $\widehat{\text{ATE}}_{\text{unadj}}$, a necessary (but not sufficient) condition is that either the design must be so imbalanced that more than three-quarters of the subjects are assigned to one group, or z_i must have a larger covariance with the treatment effect $a_i - b_i$ than with the expected outcome $p_A a_i + (1-p_A)b_i$. With multiple covariates, a similar condition can be derived from Eq. (14) in Schochet (2010).

(vi) With more than two treatment groups, the usual adjustment can be worse than no adjustment even when the design is balanced [Freedman (2008b)]. All the groups are pooled in a single regression without treatment $\times$ covariate interactions, so group B's data can affect the contrast between A and C.

Example

This simulation illustrates some of the key ideas.

1. For $n = 1{,}000$ subjects, a covariate z_i was drawn from the uniform distribution on $[-4,4]$. The potential outcomes were then generated as

$$\begin{aligned} a_i &= \frac{\exp(z_i) + \exp(z_i/2)}{4} + \nu_i, \\ b_i &= \frac{-\exp(z_i) + \exp(z_i/2)}{4} + \varepsilon_i \end{aligned}$$

with ν_i and ε_i drawn independently from the standard normal distribution.

Table 2.1: Simulation (1,000 subjects; 40,000 replications)

Estimator	Proportion assigned to treatment A				
	0.75	0.6	0.5	0.4	0.25
SD (asymptotic) $\times$ 1,000					
Unadjusted	93	49	52	78	143
Usual OLS-adjusted	171	72	46	79	180
OLS with interaction	80	49	46	58	98
Tyranny of the minority	80	49	46	58	98
SD (empirical) $\times$ 1,000					
Unadjusted	93	49	53	78	142
Usual OLS-adjusted	171	73	47	80	180
OLS with interaction	81	50	47	59	99
Tyranny of the minority	81	50	47	59	99
Bias (estimated) $\times$ 1,000					
Unadjusted	0	0	0	0	-2
Usual OLS-adjusted	-3	-3	-3	-3	-5
OLS with interaction	-5	-3	-3	-4	-6
Tyranny of the minority	-5	-3	-3	-4	-6

2. A completely randomized experiment was simulated 40,000 times, assigning $n_A = 750$ subjects to treatment A and the remainder to treatment B.

3. Step 2 was repeated for four other values of n_A (600, 500, 400, and 250).

These are adverse conditions for regression adjustment: z_i covaries much more with the treatment effect $a_i - b_i$ than with the potential outcomes, and the population least squares slopes $Q_a = 1.06$ and $Q_b = -0.73$ are of opposite signs.

Table 2.1 compares $\widehat{\text{ATE}}_{\text{unadj}}$, $\widehat{\text{ATE}}_{\text{adj}}$, $\widehat{\text{ATE}}_{\text{interact}}$, and the "tyranny of the minority" estimator from remark (iii) after Corollary 2.1.2. The first panel shows the asymptotic standard errors derived from Freedman's (2008b) Theorems 1 and 2 and this chapter's Theorem 2.1 (with limits replaced by actual population values). The second and third panels show the empirical standard deviations and bias estimates from the Monte Carlo simulation.

The empirical standard deviations are very close to the asymptotic predictions, and the estimated biases are small in comparison. The usual adjustment hurts precision except when $n_A/n = 0.5$. In contrast, $\widehat{\text{ATE}}_{\text{interact}}$ and the tyranny estimator improve precision except when $n_A/n = 0.6$. [This is approximately the value of p_A where $\widehat{\text{ATE}}_{\text{interact}}$ and $\widehat{\text{ATE}}_{\text{unadj}}$ have equal asymptotic variance; see remark (iii) after Corollary 2.1.1.]

Randomization does not "justify" the regression model of $\widehat{\text{ATE}}_{\text{interact}}$, and the linearity assumption is far from accurate in this example, but the estimator solves Freedman's asymptotic precision problem.

2.5 Variance estimation

Eicker (1967) and White (1980ab) proposed a covariance matrix estimator for OLS that is consistent under simple random sampling from an infinite population. The regression model assumptions, such as linearity and homoskedasticity, are not needed for this result.[7] The estimator is

$$(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\,\text{diag}(\hat{\varepsilon}_1^2,\ldots,\hat{\varepsilon}_n^2)\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$$

where $\mathbf{X}$ is the matrix of regressors and $\hat{\varepsilon}_i$ is the ith OLS residual. It is known as the sandwich estimator because of its form, or as the Huber–White estimator because it is the sample analog of Huber's (1967) formula for the asymptotic variance of a maximum likelihood estimator when the model is incorrect.

Theorem 2.2 shows that under the Neyman model, the sandwich variance estimators for $\widehat{\text{ATE}}_{\text{adj}}$ and $\widehat{\text{ATE}}_{\text{interact}}$ are consistent or asymptotically conservative. Together, Theorems 2.1 and 2.2 in this chapter and Theorem 2 in Freedman (2008b) imply that asymptotically valid confidence intervals for ATE can be constructed from either $\widehat{\text{ATE}}_{\text{adj}}$ or $\widehat{\text{ATE}}_{\text{interact}}$ and the sandwich standard error estimator.

The vectors $\mathbf{Q}_a$ and $\mathbf{Q}_b$ were defined in Section 2.4. Let $\mathbf{Q}$ denote the weighted average $p_A\mathbf{Q}_a + (1-p_A)\mathbf{Q}_b$. As shown in Freedman (2008b) and Appendix A, $\mathbf{Q}$ is the probability limit of the vector of estimated coefficients on $\mathbf{z}_i$ in the OLS regression of Y_i on T_i and $\mathbf{z}_i$.

Mimicking Section 2.4, define the prediction errors

$$a_i^{**} = (a_i - \overline{a}) - (\mathbf{z}_i - \overline{\mathbf{z}})\mathbf{Q}, \qquad b_i^{**} = (b_i - \overline{b}) - (\mathbf{z}_i - \overline{\mathbf{z}})\mathbf{Q}$$

for $i = 1,\ldots,n$.

Theorem 2.2 is proved in Appendix A.

Theorem 2.2. *Assume Conditions 1–3. Let $\widehat{v}_{\text{adj}}$ and $\widehat{v}_{\text{interact}}$ denote the sandwich variance estimators for $\widehat{\text{ATE}}_{\text{adj}}$ and $\widehat{\text{ATE}}_{\text{interact}}$. Then $n\widehat{v}_{\text{adj}}$ converges in probability to*

$$\frac{1}{p_A}\lim_{n\to\infty}\sigma^2_{a^{**}} + \frac{1}{1-p_A}\lim_{n\to\infty}\sigma^2_{b^{**}},$$

which is greater than or equal to the true asymptotic variance of $\sqrt{n}(\widehat{\text{ATE}}_{\text{adj}} - \text{ATE})$. The difference is

$$\lim_{n\to\infty}\sigma^2_{(a-b)} = \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}[(a_i - b_i) - \text{ATE}]^2.$$

[7] See, e.g., Chamberlain (1982, pp. 17–19) or Angrist and Pischke (2009, pp. 40–48). Fuller (1975) proves a finite-population version of the result.

Similarly, $n\widehat{v}_{\text{interact}}$ *converges in probability to*

$$\frac{1}{p_A}\lim_{n\to\infty}\sigma^2_{a^*}+\frac{1}{1-p_A}\lim_{n\to\infty}\sigma^2_{b^*},$$

which is greater than or equal to the true asymptotic variance of $\sqrt{n}(\widehat{\text{ATE}}_{\text{interact}}-\text{ATE})$. *The difference is*

$$\lim_{n\to\infty}\sigma^2_{(a^*-b^*)}=\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}[(a_i-b_i)-\text{ATE}-(\mathbf{z}_i-\overline{\mathbf{z}})(\mathbf{Q}_a-\mathbf{Q}_b)]^2.$$

Remarks. (i) Theorem 2.2 generalizes to designs with more than two treatment groups.

(ii) With two treatment groups of equal size, the conventional OLS variance estimator for $\widehat{\text{ATE}}_{\text{adj}}$ is also consistent or asymptotically conservative [Freedman (2008a)].

(iii) Freedman (2008a) shows analogous results for variance estimators for the difference in means; the issue there is whether to assume $\sigma^2_a=\sigma^2_b$. Reichardt and Gollob (1999) and Freedman, Pisani, and Purves (2007, pp. 508–511) give helpful expositions of basic results under the Neyman model. Related issues appear in discussions of the two-sample problem [Miller (1986, pp. 56–62); Stonehouse and Forrester (1998)] and randomization tests [Gail et al. (1996); Chung and Romano (2011, 2012)].

(iv) With a small sample or points of high leverage, the sandwich estimator can have substantial downward bias and high variability. MacKinnon (2013) discusses bias-corrected sandwich estimators and improved confidence intervals based on the wild bootstrap. See also Wu (1986), Tibshirani (1986), Angrist and Pischke (2009, ch. 8), and Kline and Santos (2012).

(v) When $\widehat{\text{ATE}}_{\text{unadj}}$ is computed by regressing Y_i on T_i, the HC2 bias-corrected sandwich estimator [MacKinnon and White (1985); Royall and Cumberland (1978); Wu (1986, p. 1274)] gives exactly the variance estimate preferred by Neyman (1923) and Freedman (2008a): $\widehat{\sigma}^2_a/n_A+\widehat{\sigma}^2_b/(n-n_A)$, where $\widehat{\sigma}^2_a$ and $\widehat{\sigma}^2_b$ are the sample variances of Y_i in the two groups.[8]

(vi) When the n subjects are randomly drawn from a superpopulation, $\widehat{v}_{\text{interact}}$ does not take into account the variability in $\overline{\mathbf{z}}$ [Imbens and Wooldridge (2009, pp. 28–30)]. In the Neyman model, $\overline{\mathbf{z}}$ is fixed.

(vii) Freedman's (2006) critique of the sandwich estimator does not apply here, as $\widehat{\text{ATE}}_{\text{adj}}$ and $\widehat{\text{ATE}}_{\text{interact}}$ are consistent even when their regression models are incorrect.

(viii) Freedman (2008a) associates the difference in means and regression with heteroskedasticity-robust and conventional variance estimators, respectively. His

[8]For details, see Hinkley and Wang (1991), Angrist and Pischke (2009, pp. 294–304), or Samii and Aronow (2012).

rationale for these pairings is unclear. The pooled-variance two-sample t-test and the conventional F-test for equality of means are often used in difference-in-means analyses. Conversely, the sandwich estimator has become the usual variance estimator for regression in economics [Stock (2010)]. The question of whether to adjust for covariates should be disentangled from the question of whether to assume homoskedasticity.

2.6 Bias

The bias of OLS adjustment diminishes rapidly with the number of randomly assigned units: $\widehat{\text{ATE}}_{\text{adj}}$ and $\widehat{\text{ATE}}_{\text{interact}}$ have biases of order $1/n$, while their standard errors are of order $1/\sqrt{n}$. Brief remarks follow; see also Deaton (2010, pp. 443–444), Imbens (2010, pp. 410–411), and Green and Aronow (2011).

(i) If the actual random assignment yields substantial covariate imbalance, it is hardly reassuring to be told that the difference in means is unbiased over all possible random assignments. Senn (1989) and Cox and Reid (2000, pp. 29–32) argue that inference should be conditional on a measure of covariate imbalance, and that the conditional bias of $\widehat{\text{ATE}}_{\text{unadj}}$ justifies adjustment. Tukey (1991) suggests adjustment "perhaps as a supplemental analysis" for "protection against either the consequences of inadequate randomization or the (random) occurrence of an unusual randomization."

(ii) As noted in Section 2.2, poststratification is a special case of $\widehat{\text{ATE}}_{\text{interact}}$. The poststratified estimator is a population-weighted average of subgroup-specific differences in means. Conditional on the numbers of subgroup members assigned to each treatment, the poststratified estimator is unbiased, but $\widehat{\text{ATE}}_{\text{unadj}}$ can be biased. Miratrix, Sekhon, and Yu (2013) give finite-sample and asymptotic analyses of poststratification and blocking; see also Holt and Smith (1979) in the sampling context.

(iii) Cochran (1977) analyzes the bias of $\widehat{\bar{y}}_{\text{reg}}$ in Eq. (2.3.1). If the adjustment factor q is fixed, $\widehat{\bar{y}}_{\text{reg}}$ is unbiased, but if q varies with the sample, $\widehat{\bar{y}}_{\text{reg}}$ has a bias of $-\text{cov}(q, \bar{z}_S)$. The leading term in the bias of $\widehat{\bar{y}}_{\text{OLS}}$ is

$$-\frac{1}{\sigma_z^2}\left(\frac{1}{n}-\frac{1}{N}\right)\lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N} e_i(z_i-\bar{z})^2$$

where n is the sample size, N is the population size, and e_i is the prediction error in the population least squares regression of y_i on z_i.

(iv) By analogy, the leading term in the bias of $\widehat{\text{ATE}}_{\text{interact}}$ (with a single covariate z_i) is

$$-\frac{1}{\sigma_z^2}\left[\left(\frac{1}{n_A}-\frac{1}{n}\right)\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n} a_i^*(z_i-\bar{z})^2-\left(\frac{1}{n-n_A}-\frac{1}{n}\right)\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n} b_i^*(z_i-\bar{z})^2\right].$$

Thus, the bias tends to depend largely on n, n_A/n, and the importance of omitted quadratic terms in the regressions of a_i and b_i on z_i. With multiple covariates, it would also depend on the importance of omitted first-order interactions between the covariates.

(v) Remark (iii) also implies that if the adjustment factors $\mathbf{q}_a$ and $\mathbf{q}_b$ in Eqs. (2.3.2–2.3.3) do not vary with random assignment, the resulting estimator of ATE is unbiased. Middleton and Aronow's (2012) insightful paper uses out-of-sample data to determine $\mathbf{q}_a = \mathbf{q}_b$. In-sample data can be used when multiple pretests (pre-randomization outcome measures) are available: if the only covariate z_i is the most recent pretest, a common adjustment factor $q_a = q_b$ can be determined by regressing z_i on an earlier pretest.

2.7 Empirical example

This section suggests empirical checks on the asymptotic approximations. I will focus on the validity of confidence intervals, using data from a social experiment for an illustrative example.

Background

Angrist, Lang, and Oreopoulos (2009; henceforth ALO) conducted an experiment to estimate the effects of support services and financial incentives on college students' academic achievement. At a Canadian university campus, all first-year undergraduates entering in September 2005, except those with a high-school grade point average (GPA) in the top quartile, were randomly assigned to four groups. One treatment group was offered support services (peer advising and supplemental instruction). Another group was offered financial incentives (awards of $1,000 to $5,000 for meeting a target GPA). A third group was offered both services and incentives. The control group was eligible only for standard university support services (which included supplemental instruction for some courses).

ALO report that for women, the combination of services and incentives had sizable estimated effects on both first- and second-year academic achievement, even though the programs were only offered during the first year. In contrast, there was no evidence that services alone or incentives alone had lasting effects for women or that any of the treatments improved achievement for men (who were much less likely to contact peer advisors).

To simplify the example and focus on the accuracy of large-sample approximations in samples that are not huge, I use only the data for men (43 percent of the students) in the services-and-incentives and services-only groups (9 percent and 15 percent of the men). First-year GPA data are available for 58 men in the services-and-incentives group and 99 in the services-only group.

Table 2.2 shows alternative estimates of ATE (the average treatment effect of the financial incentives, given that the support services were available). The services-and-incentives and services-only groups had average first-year GPAs of 1.82 and 1.86 (on a scale of 0 to 4), so the unadjusted estimate of ATE is close to zero. OLS adjustment for high-school GPA hardly makes a practical difference to either the point estimate of ATE or the sandwich standard error estimate, regardless of whether the treatment $\times$ covariate interaction is included.[9] The two groups had similar average high-school GPAs, and high-school GPA was not a strong predictor of first-year college GPA.

Table 2.2: Estimates of average treatment effect on men's first-year GPA

	Point estimate	Sandwich SE
Unadjusted	-0.036	0.158
Usual OLS-adjusted	-0.083	0.146
OLS with interaction	-0.081	0.146

The finding that adjustment appears to have little effect on precision is not unusual in social experiments, because the covariates are often only weakly correlated with the outcome [Meyer (1995, pp. 100, 116); Lin et al. (1998, pp. 129–133)]. Examining eight social experiments with a wide range of outcome variables, Schochet (2010) finds R^2 values above 0.3 only when the outcome is a standardized achievement test score or Medicaid costs and the covariates include a lagged outcome.

Researchers may prefer not to adjust when the expected precision improvement is meager. Either way, confidence intervals for treatment effects typically rely on either strong parametric assumptions (such as a constant treatment effect or a normally distributed outcome) or asymptotic approximations. When a sandwich standard error estimate is multiplied by 1.96 to form a margin of error for a 95 percent confidence interval, the calculation assumes the sample is large enough that (i) the estimator of ATE is approximately normally distributed, (ii) the bias and variability of the sandwich standard error estimator are small relative to the true standard error (or else the bias is conservative and the variability is small), and (iii) the bias of adjustment (if used) is small relative to the true standard error.

Below I discuss a simulation to check for confidence interval undercoverage due to violations of (i) or (ii), and a bias estimate to check for violations of (iii). These checks are not foolproof, but may provide a useful sniff test.

[9] ALO adjust for a larger set of covariates, including first language, parents' education, and self-reported procrastination tendencies. These also have little effect on the estimated standard errors.

Simulation

For technical reasons, the most revealing initial check is a simulation with a constant treatment effect. When treatment effects are heterogeneous, the sandwich standard error estimators for $\widehat{\text{ATE}}_{\text{unadj}}$ and $\widehat{\text{ATE}}_{\text{adj}}$ are asymptotically conservative,[10] so nominal 95 percent confidence intervals for ATE achieve greater than 95 percent coverage in large enough samples. A simulation that overstates treatment effect heterogeneity may overestimate coverage.

Table 2.3 reports a simulation that assumes treatment had no effect on any of the men. Keeping the GPA data at their actual values, I replicated the experiment 250,000 times, each time randomly assigning 58 men to services-and-incentives and 99 to services-only. The first panel shows the means and standard deviations of $\widehat{\text{ATE}}_{\text{unadj}}$, $\widehat{\text{ATE}}_{\text{adj}}$, and $\widehat{\text{ATE}}_{\text{interact}}$. All three estimators are approximately unbiased, but adjustment slightly improves precision. Since the simulation assumes a constant treatment effect (zero), including the treatment $\times$ covariate interaction does not improve precision relative to the usual adjustment.

The second and third panels show the estimated biases and standard deviations of the sandwich standard error estimator and the three variants discussed in Angrist and Pischke (2009, pp. 294–308). ALO's paper uses HC1 [Hinkley (1977)], which simply multiplies the sandwich variance estimator by $n/(n-k)$, where k is the number of regressors. HC2 [see remark (v) after Theorem 2.2] and the approximate jackknife HC3 [Davidson and MacKinnon (1993, pp. 553–554); Tibshirani (1986)] inflate the squared residuals in the sandwich formula by the factors $(1-h_{ii})^{-1}$ and $(1-h_{ii})^{-2}$, where h_{ii} is the ith diagonal element of the hat matrix $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. All the standard error estimators appear to be approximately unbiased with low variability.

The fourth and fifth panels evaluate thirteen ways of constructing a 95 percent confidence interval. For each of the three estimators of ATE, each of the four standard error estimators was multiplied by 1.96 to form the margin of error for a normal-approximation interval. Welch's (1949) t-interval [Miller (1986, pp. 60–62)] was also constructed. Welch's interval uses $\widehat{\text{ATE}}_{\text{unadj}}$, the HC2 standard error estimator, and the t-distribution with the Welch–Satterthwaite approximate degrees of freedom.

The fourth panel shows that all thirteen confidence intervals cover the true value of ATE (zero) with approximately 95 percent probability. The fifth panel shows the average widths of the intervals. (The mean and median widths agree up to three decimal places.) The regression-adjusted intervals are narrower on average than the unadjusted intervals, but the improvement is meager. In sum, adjustment appears to yield slightly more precise inference without sacrificing validity.

[10] By Theorem 2.2, the sandwich standard error estimator for $\widehat{\text{ATE}}_{\text{interact}}$ is also asymptotically conservative unless the treatment effect is a linear function of the covariates.

Table 2.3: Simulation with zero treatment effect (250,000 replications). The fourth panel shows the empirical coverage rates of nominal 95 percent confidence intervals. All other estimates are on the four-point GPA scale.

	ATE estimator		
	Unadjusted	Usual OLS-adjusted	OLS with interaction
Bias & SD of ATE estimator			
Mean (estimated bias)	0.000	0.000	0.000
SD	0.158	0.147	0.147
Bias of SE estimator			
Classic sandwich	−0.001	−0.002	−0.002
HC1	0.000	0.000	0.000
HC2	0.000	0.000	0.000
HC3	0.001	0.002	0.002
SD of SE estimator			
Classic sandwich	0.004	0.004	0.004
HC1	0.004	0.004	0.004
HC2	0.004	0.004	0.004
HC3	0.004	0.004	0.005
CI coverage (percent)			
Classic sandwich	94.6	94.5	94.4
HC1	94.8	94.7	94.7
HC2 (normal)	94.8	94.8	94.8
HC2 (Welch t)	95.1		
HC3	95.0	95.0	95.1
CI width (average)			
Classic sandwich	0.618	0.570	0.568
HC1	0.622	0.576	0.575
HC2 (normal)	0.622	0.576	0.577
HC2 (Welch t)	0.629		
HC3	0.627	0.583	0.586

Bias estimates

One limitation of the simulation above is that the bias of adjustment may be larger when treatment effects are heterogeneous. With a single covariate z_i, the leading term in the bias of $\widehat{\text{ATE}}_{\text{adj}}$ is[11]

$$-\frac{1}{n}\frac{1}{\sigma_z^2}\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}[(a_i-b_i)-\text{ATE}](z_i-\bar{z})^2.$$

Thus, with a constant treatment effect, the leading term is zero (and the bias is of order $n^{-3/2}$ or smaller). Freedman (2008b) shows that with a balanced design and a constant treatment effect, the bias is exactly zero.

We can estimate the leading term by rewriting it as

$$-\frac{1}{n}\frac{1}{\sigma_z^2}\left[\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}(a_i-\bar{a})(z_i-\bar{z})^2-\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}(b_i-\bar{b})(z_i-\bar{z})^2\right]$$

and substituting the sample variance of high-school GPA for σ_z^2, and the sample covariances of first-year college GPA with the square of centered high-school GPA in the services-and-incentives and services-only groups for the bracketed limits. The resulting estimate of the bias of $\widehat{\text{ATE}}_{\text{adj}}$ is -0.0002 on the four-point GPA scale. Similarly, the leading term in the bias of $\widehat{\text{ATE}}_{\text{interact}}$ [Section 2.6, remark (iv)] can be estimated, and the result is also -0.0002. The biases would need to be orders of magnitude larger to have noticeable effects on confidence interval coverage (the estimated standard errors of $\widehat{\text{ATE}}_{\text{adj}}$ and $\widehat{\text{ATE}}_{\text{interact}}$ in Table 2.2 are both 0.146).

Remarks

(i) This exercise does not prove that the bias of adjustment is negligible, since it just replaces a first-order approximation (the bias is close to zero in large enough samples) with a second-order approximation (the bias is close to the leading term in large enough samples), and the estimate of the leading term has sampling error.[12] The checks suggested here cannot validate an analysis, but they can reveal problems.

(ii) Another limitation is that the simulation assumes the potential outcome distributions have the same shape. In Stonehouse and Forrester's (1998) simulations, Welch's t-test was not robust to extreme skewness in the smaller group when that

[11] An equivalent expression appears in the version of Freedman (2008a) on his web page. It can be derived from Freedman (2008b) after correcting a minor error in Eqs. (17–18): the potential outcomes should be centered.

[12] Finite-population bootstrap methods [Davison and Hinkley (1997, pp. 92–100, 125)] may also be useful for estimating the bias of $\widehat{\text{ATE}}_{\text{interact}}$, but similar caveats would apply.

group's sample size was 30 or smaller. That does not appear to be a serious issue in this example, however. The distribution of men's first-year GPA in the services-and-incentives group is roughly symmetric (e.g., see ALO, Fig. 1A).

(iii) The simulation check may appear to resemble permutation inference [Fisher (1935); Tukey (1993); Rosenbaum (2002)], but the goals differ. Here, the constant treatment effect scenario just gives a benchmark to check the finite-sample coverage of confidence intervals that are asymptotically valid under weaker assumptions. Classical permutation methods achieve exact inference under strong assumptions about treatment effects, but may give misleading results when the assumptions fail. For example, the Fisher–Pitman permutation test is asymptotically equivalent to a t-test using the conventional OLS standard error estimator. The test can be inverted to give exact confidence intervals for a constant treatment effect, but these intervals may undercover ATE when treatment effects are heterogeneous and the design is imbalanced [Gail et al. (1996)].

(iv) Chung and Romano (2011, 2012) discuss and extend a literature on permutation tests that do remain valid asymptotically when the null hypothesis is weakened. One such test is based on the permutation distribution of a heteroskedasticity-robust t-statistic. Exploration of this approach under the Neyman model (with and without covariate adjustment) would be valuable.

2.8 Further remarks

Freedman's papers answer important questions about the properties of OLS adjustment. He and others have summarized his results with a "glass is half empty" view that highlights the dangers of adjustment. To the extent that this view encourages researchers to present unadjusted estimates first, it is probably a good influence. The difference in means is the "hands above the table" estimate: it is clearly not the product of a specification search, and its transparency may encourage discussion of the strengths and weaknesses of the data and research design.[13]

But it would be unwise to conclude that Freedman's critique should always override the arguments for adjustment, or that studies reporting only adjusted estimates should always be distrusted. Freedman's own work shows that with large enough samples and balanced two-group designs, randomization justifies the traditional adjustment. One does not need to believe in the classical linear model to tolerate or even advocate OLS adjustment, just as one does not need to believe in the Four Noble Truths of Buddhism to entertain the hypothesis that mindfulness meditation has causal effects on mental health.

From an agnostic perspective, Freedman's theorems are a major contribution. Three-quarters of a century after Fisher discovered the analysis of covariance,

[13] On transparency and critical discussion, see Ashenfelter and Plant (1990), Freedman (1991, 2008c, 2010), Moher et al. (2010), and Rosenbaum (2010, ch. 6).

Freedman deepened our understanding of its properties by deriving the regression-adjusted estimator's asymptotic distribution without assuming a regression model, a constant treatment effect, or an infinite superpopulation. His argument is constructed with unsurpassed clarity and rigor. It deserves to be studied in detail and considered carefully.

Chapter 3

Approximating the bias of OLS adjustment in randomized experiments

3.1 Motivation

Chapter 2 and a companion blog essay [Lin (2012ab)] discussed Freedman's (2008ab) three concerns about OLS adjustment—possible worsening of precision, invalid measures of precision, and small-sample bias—and a further concern about ad hoc specification search [Freedman (2008c, 2010)]. Small-sample bias is probably the least important of these concerns in many social experiments, since it diminishes rapidly as the number of randomly assigned units grows. Yet the bias issue has captured the lion's share of the attention in some published and unpublished discussions of Freedman's critique. The economist Jed Friedman (2012) writes: "I and others have indeed received informal comments and referee reports claiming that adjusting for observables leads to biased inference (without supplemental caveats on small sample bias).... The precision arguments of Freedman don't seem to have settled in the minds of practitioners as much as bias."

How can applied researchers judge whether small-sample bias is likely to be a serious concern? One approach is to use the data to estimate the bias, as Freedman (2004) notes in his discussion of ratio estimators in survey sampling.[1] In the empirical example in Chapter 2, I estimated the leading term in the bias of OLS adjustment for a single covariate (with and without the treatment $\times$ covariate interaction), using the sample analogs of asymptotic formulas from Cochran (1977, pp. 198–199) and Freedman (2008b). The current chapter derives and discusses the leading term in the bias of adjustment for multiple covariates. The results may be useful for estimating the bias and may also be relevant to choosing a regression

[1] Ratio estimators of population means are a special case of regression estimators and also have a bias of order $1/n$. See, e.g., Cochran (1977, pp. 160–162, 189–190).

model when the sample is small.

3.2 Assumptions and notation

Review from Chapter 2

We assume a completely randomized experiment with n subjects, assigning n_A to treatment A and $n-n_A$ to treatment B. For each subject i, we observe an outcome Y_i and a $1 \times K$ vector of covariates $\mathbf{z}_i$. The potential outcomes corresponding to treatments A and B are a_i and b_i. Let T_i denote a dummy variable for treatment A.

The means of a_i, b_i, and $\mathbf{z}_i$ over the population (the n subjects) are $\overline{a}$, $\overline{b}$, and $\overline{\mathbf{z}}$. The average treatment effect of A relative to B is $\text{ATE} = \overline{a} - \overline{b}$. We consider two OLS-adjusted estimators, $\widehat{\text{ATE}}_{\text{adj}}$ (the estimated coefficient on T_i in the regression of Y_i on T_i and $\mathbf{z}_i$) and $\widehat{\text{ATE}}_{\text{interact}}$ [the estimated coefficient on T_i in the regression of Y_i on T_i, $\mathbf{z}_i$, and $T_i(\mathbf{z}_i - \overline{\mathbf{z}})$].

Section 2.4 discusses the scenario and regularity conditions for asymptotics. As Freedman (2008a) writes, the scenario assumes "our inference problem is embedded in an infinite sequence of such problems, with the number of subjects n increasing to infinity." The number of covariates K is held constant as n grows. Theorem 3.1 below (on the bias of $\widehat{\text{ATE}}_{\text{adj}}$) assumes Conditions 1–3 from Section 2.4.

Additional assumptions and notation

Freedman (2008b, p. 194) and Appendix A (Section A.1) note that Conditions 1–3 do not rule out the possibility that for some n and some randomizations, $\widehat{\text{ATE}}_{\text{adj}}$ or $\widehat{\text{ATE}}_{\text{interact}}$ is ill-defined because of perfect multicollinearity. The current chapter assumes that for all n above some threshold, the distribution of the covariates is such that both $\widehat{\text{ATE}}_{\text{adj}}$ and $\widehat{\text{ATE}}_{\text{interact}}$ are well-defined for every possible randomization. (It seems likely that results similar to Theorems 3.1 and 3.2 below would hold even without this assumption, since Conditions 2 and 3 imply that perfect multicollinearity becomes extremely unlikely as n grows with K fixed. But the details have not been fleshed out.)

Theorem 3.2 (on the bias of $\widehat{\text{ATE}}_{\text{interact}}$) assumes a stronger set of regularity conditions. In brief, in addition to Conditions 1–3, we assume bounded eighth moments and converging fourth moments. Details are given in the theorem's statement.

For both theorems, let $\mathbf{M} = [n^{-1}\sum_{i=1}^{n}(\mathbf{z}_i - \overline{\mathbf{z}})'(\mathbf{z}_i - \overline{\mathbf{z}})]^{-1}$, or equivalently $\mathbf{M} = n(\widetilde{\mathbf{Z}}'\widetilde{\mathbf{Z}})^{-1}$, where $\widetilde{\mathbf{Z}}$ is the $n \times K$ matrix whose ith row is $\mathbf{z}_i - \overline{\mathbf{z}}$.

Section 2.4 defined prediction errors a_i^* and b_i^* for predictions based on $\mathbf{Q}_a$ and $\mathbf{Q}_b$, the limits of the least squares slope vectors in the population regressions of a_i and b_i on $\mathbf{z}_i$. Theorem 3.2 below involves the actual population least squares slope

vectors $\widetilde{\mathbf{Q}}_a$ and $\widetilde{\mathbf{Q}}_b$ instead of their asymptotic limits:

$$\begin{aligned}\widetilde{\mathbf{Q}}_a &= \left[\sum_{i=1}^{n}(\mathbf{z}_i-\overline{\mathbf{z}})'(\mathbf{z}_i-\overline{\mathbf{z}})\right]^{-1}\sum_{i=1}^{n}(\mathbf{z}_i-\overline{\mathbf{z}})'(a_i-\overline{a}),\\ \widetilde{\mathbf{Q}}_b &= \left[\sum_{i=1}^{n}(\mathbf{z}_i-\overline{\mathbf{z}})'(\mathbf{z}_i-\overline{\mathbf{z}})\right]^{-1}\sum_{i=1}^{n}(\mathbf{z}_i-\overline{\mathbf{z}})'(b_i-\overline{b}).\end{aligned}$$

Let $\tilde{a}_i$ and $\tilde{b}_i$ denote the population least squares prediction errors:

$$\tilde{a}_i = (a_i-\overline{a})-(\mathbf{z}_i-\overline{\mathbf{z}})\widetilde{\mathbf{Q}}_a, \qquad \tilde{b}_i = (b_i-\overline{b})-(\mathbf{z}_i-\overline{\mathbf{z}})\widetilde{\mathbf{Q}}_b$$

for $i = 1,\ldots,n$.

3.3 Results

Theorems 3.1 and 3.2 are proved in Appendix B. Theorem 3.1 gives the leading term in the bias of $\widehat{\text{ATE}}_{\text{adj}}$.

Theorem 3.1. *Assume Conditions 1–3. Then*

$$\widehat{\text{ATE}}_{\text{adj}} - \text{ATE} \;=\; \eta_n \,+\, \rho_n,$$

where

$$E(\eta_n) \;=\; -\frac{1}{n-1}\frac{1}{n}\sum_{i=1}^{n}[(a_i-b_i)-ATE](\mathbf{z}_i-\overline{\mathbf{z}})\mathbf{M}(\mathbf{z}_i-\overline{\mathbf{z}})'$$

and ρ_n is of order less than or equal to $n^{-3/2}$ in probability.

Remarks. (i) With a single covariate z_i, the leading term $E(\eta_n)$ reduces to

$$-\frac{1}{n-1}\frac{1}{n}\sum_{i=1}^{n}[(a_i-b_i)-ATE][(z_i-\bar{z})/\sigma_z]^2,$$

where σ_z is the population standard deviation of z_i. In other words, the leading term is $-1/(n-1)$ times the covariance between the treatment effect a_i-b_i and the square of the standardized covariate. This expression should be interpreted with care: the covariance can be nonzero even when a_i-b_i is a linear function of z_i.

(ii) With multiple covariates, the leading term is $-1/(n-1)$ times the covariance between the treatment effect and the quadratic form $(\mathbf{z}_i-\overline{\mathbf{z}})\mathbf{M}(\mathbf{z}_i-\overline{\mathbf{z}})'$. The quadratic form is a linear combination of the squares and first-order interactions of the mean-centered covariates, and can be rewritten as nh_{ii}, where h_{ii} is the leverage

of observation i in a no-intercept regression on the mean-centered covariates [the ith diagonal element of the hat matrix $\widetilde{\mathbf{Z}}(\widetilde{\mathbf{Z}}'\widetilde{\mathbf{Z}})^{-1}\widetilde{\mathbf{Z}}'$] [Hoaglin and Welsch (1978)].

(iii) Although $E(\eta_n)$ depends on the heterogeneity in the individual treatment effects, which are unobservable, it can be estimated as in Section 2.7 after rewriting it as

$$-\frac{1}{n-1}\left[\frac{1}{n}\sum_{i=1}^{n}(a_i-\overline{a})(\mathbf{z}_i-\overline{\mathbf{z}})\mathbf{M}(\mathbf{z}_i-\overline{\mathbf{z}})'-\frac{1}{n}\sum_{i=1}^{n}(b_i-\overline{b})(\mathbf{z}_i-\overline{\mathbf{z}})\mathbf{M}(\mathbf{z}_i-\overline{\mathbf{z}})'\right].$$

(iv) A technical point: Conditions 1 and 2 ensure that the covariance between a_i-b_i and $(\mathbf{z}_i-\overline{\mathbf{z}})\mathbf{M}(\mathbf{z}_i-\overline{\mathbf{z}})'$ is bounded as n goes to infinity, so $E(\eta_n)$ is of order $1/n$ or smaller.

Theorem 3.2 gives the leading term in the bias of $\widehat{\text{ATE}}_{\text{interact}}$.

Theorem 3.2. *Assume Conditions 2 and 3, and assume there is a bound $L<\infty$ such that for all $n=1,2,\ldots$ and $k=1,\ldots,K$,*

$$\frac{1}{n}\sum_{i=1}^{n}a_i^8<L,\qquad \frac{1}{n}\sum_{i=1}^{n}b_i^8<L,\qquad \frac{1}{n}\sum_{i=1}^{n}z_{ik}^8<L.$$

Also assume that for all $k=1,\ldots,K$ and $\ell=1,\ldots,K$, the population variances and covariances of a_iz_{ik}, b_iz_{ik}, and $z_{ik}z_{i\ell}$ converge to finite limits. Then

$$\widehat{\text{ATE}}_{\text{interact}}-\text{ATE} \;=\; \tilde{\eta}_n+\tilde{\rho}_n,$$

where

$$E(\tilde{\eta}_n) \;=\; -\left[\left(\frac{1}{n_A}-\frac{1}{n}\right)\frac{1}{n-1}\sum_{i=1}^{n}\tilde{a}_i(\mathbf{z}_i-\overline{\mathbf{z}})\mathbf{M}(\mathbf{z}_i-\overline{\mathbf{z}})' \;-\; \left(\frac{1}{n-n_A}-\frac{1}{n}\right)\frac{1}{n-1}\sum_{i=1}^{n}\tilde{b}_i(\mathbf{z}_i-\overline{\mathbf{z}})\mathbf{M}(\mathbf{z}_i-\overline{\mathbf{z}})'\right]$$

and $\tilde{\rho}_n$ is of order less than or equal to $n^{-3/2}$ in probability.

Remarks. (i) $E(\tilde{\eta}_n)$ equals the difference between the leading terms in the biases of the OLS-adjusted mean outcomes under treatments A and B.

(ii) With a single covariate, $E(\tilde{\eta}_n)$ reduces to

$$-\left(\frac{1}{n_A}-\frac{1}{n}\right)\frac{1}{n-1}\sum_{i=1}^{n}\tilde{a}_i[(z_i-\overline{z})/\sigma_z]^2+\left(\frac{1}{n-n_A}-\frac{1}{n}\right)\frac{1}{n-1}\sum_{i=1}^{n}\tilde{b}_i[(z_i-\overline{z})/\sigma_z]^2.$$

The factor $1/n_A-1/n$ reflects the sample size of treatment group A and a finite-population correction. The covariance between the prediction error $\tilde{a}_i$ and the square of the standardized covariate reflects the variation in the potential outcome

a_i that is not explained by the population least squares regression of a_i on z_i but would be explained if z_i^2 were included. If the relationship between a_i and z_i is linear (e.g., if z_i is a dummy variable), this covariance is zero.

(iii) With multiple covariates, $E(\tilde{\eta}_n)$ involves the covariances of the population least squares prediction errors with the quadratic form $(\mathbf{z}_i - \bar{\mathbf{z}})\mathbf{M}(\mathbf{z}_i - \bar{\mathbf{z}})'$, which was discussed in remark (ii) after Theorem 3.1. Thus, the leading term in the bias of $\widehat{\text{ATE}}_{\text{interact}}$ reflects variation in the potential outcomes that cannot be predicted by linear functions of the original covariates but can be predicted by quadratic terms or first-order interactions.

(iv) With a balanced design, $n_A = n/2$, so $E(\tilde{\eta}_n)$ reduces to

$$-\frac{1}{n-1}\frac{1}{n}\sum_{i=1}^{n}(\tilde{a}_i - \tilde{b}_i)(\mathbf{z}_i - \bar{\mathbf{z}})\mathbf{M}(\mathbf{z}_i - \bar{\mathbf{z}})'.$$

This expression is formally similar to the leading term in the bias of $\widehat{\text{ATE}}_{\text{adj}}$ (see Theorem 3.1) but arguably easier to interpret. The term $\tilde{a}_i - \tilde{b}_i$ is the prediction error in the population least squares regression of the treatment effect $a_i - b_i$ on $\mathbf{z}_i$. By construction, $\tilde{a}_i - \tilde{b}_i$ has mean zero and is uncorrelated with $\mathbf{z}_i$. Thus, the covariance $n^{-1}\sum_{i=1}^{n}(\tilde{a}_i - \tilde{b}_i)(\mathbf{z}_i - \bar{\mathbf{z}})\mathbf{M}(\mathbf{z}_i - \bar{\mathbf{z}})'$ reflects treatment effect heterogeneity that is not linearly correlated with $\mathbf{z}_i$ but is correlated with squares or first-order interactions of the covariates. If the relationship between the treatment effect and the covariates is linear, then $E(\tilde{\eta}_n) = 0$, in contrast to remark (i) after Theorem 3.1.

(v) The regularity conditions ensure that $E(\tilde{\eta}_n)$ is of order $1/n$ or smaller.

3.4 Discussion

The leading terms derived above can be estimated by their sample analogs, as done in Section 2.7 with a single covariate. These formulas are second-order asymptotic approximations, so they may be inaccurate in very small samples. Simulations to check their accuracy would be useful.

Bootstrap methods, including finite-population bootstraps [Davison and Hinkley (1997, pp. 92–100, 125)], may also be useful for estimating the bias of OLS adjustment. Again, these methods yield second- or higher-order asymptotic approximations.

Bias estimation can help provide a ballpark sense of the magnitude of the problem, but as Efron and Tibshirani (1993, p. 138) warn, using a bias estimate to "correct" the original estimator (i.e., to reduce its bias) "can be dangerous in practice." The reduction in bias is often outweighed by an increase in variance.

Examining the leading term in the bias of regression estimators of population means, Cochran (1977, p. 198) writes: "This term represents a contribution from the *quadratic* component of the regression.... Thus, if a sample plot ... appears approximately linear, there should be little risk of major bias." Similar comments

apply to the bias of $\widehat{\text{ATE}}_{\text{interact}}$, which is just the difference between the regression estimators of the two mean potential outcomes. When one baseline characteristic is thought to be much more predictive than all others (e.g., when a baseline measure of the outcome is available), Theorem 3.2 suggests that in small samples, one possible strategy to achieve precision improvement without serious bias is to adjust only for that characteristic, but use a specification that allows some nonlinearities (e.g., including a quadratic term) and includes treatment $\times$ covariate interactions.

Chapter 4

A "placement of death" approach for studies of treatment effects on ICU length of stay

4.1 Introduction

Length of stay (LOS) in the intensive care unit (ICU) is a common outcome measure used as an indicator of both quality of care and resource use [Marik and Hedman (2000); Rapoport et al. (2003)]. Longer ICU stays are associated with increased stress and discomfort for patients and their families, as well as increased costs for patients, hospitals, and society. Recent randomized-trial reports that estimate treatment effects on LOS include Lilly et al. (2011) and Mehta et al. (2012). LOS was the primary outcome for the SUNSET-ICU trial [Kerlin et al. (2013)], which studied the effectiveness of 24-hour staffing by intensivist physicians in the ICU, compared to having intensivists available in person during the day and by phone at night.

Because a significant proportion of patients die in the ICU, conventional analytic approaches may confound an intervention's effects on LOS with its effects on mortality. Analyzing only survivors' stays is problematic: if the intervention saves the lives of some patients, but those patients have atypically long LOS, then the intervention may spuriously appear to increase survivors' LOS. It is also potentially misleading to pool the LOS data of survivors and non-survivors: a reduction in average LOS could be achieved either by helping survivors to recover faster or by shortening non-survivors' lives. Finally, time-to-event analysis can attempt to account for death by treating non-survivors' stays as censored, but this typically involves dubious assumptions and concepts (such as the existence of a latent LOS that exceeds the observed values for non-survivors and is independent of time till death).[1]

[1] See, e.g., Freedman (2010) and Joffe (2011, section 3.2.1) for critical discussions of the as-

These issues are related to the "censoring by death" problem discussed from different perspectives by Rubin (2006a) and Joffe (2011). Rubin's exposition uses the hypothetical example of a randomized trial where the outcome is a quality-of-life (QOL) score, some patients die before QOL is measured, and treatment may affect mortality. In a comment on Rubin's paper, Rosenbaum (2006) proposes an analysis of a composite outcome that equals the QOL score if the patient was alive at the measurement time and indicates death otherwise. Death need not be valued numerically; given any preference ordering that includes death and all possible QOL scores, Rosenbaum's method gives confidence intervals for treatment effects on order statistics of the distribution of the treated patients' outcomes. He notes that although researchers cannot decide the appropriate placement of death relative to the QOL scores, we can offer analyses for several different placements, "and each patient could select the analysis that corresponds to that patient's own evaluation."

This chapter explores a modified version of Rosenbaum's approach for application to randomized trials in which ICU LOS is an outcome measure. Using a composite outcome that equals the LOS if the patient was discharged alive and indicates death otherwise, we can make inferences about treatment effects on the median and other quantiles of the outcome distribution, or about effects on the proportions of patients whose outcomes are considered better than various cutoff values of LOS. Sensitivity analyses can show how the results vary according to whether death is treated as the worst possible outcome or as preferable to extremely long ICU stays. Because the approach (like Rosenbaum's) compares the entire treatment group with the entire control group, it avoids the selection bias problem that can arise in analyses of survivors' LOS data.

A multiple-comparisons issue arises when treatment effects are estimated at multiple quantiles of the outcome distribution or on proportions below multiple cutoffs. Some researchers may choose to focus on effects on the median outcome, but the expected or intended effects of an intervention may be concentrated elsewhere in the distribution (e.g., the goal may be to reduce extremely long stays). For protection against data dredging, it may be desirable to choose a primary significance test before outcome data are available. We discuss the properties of several possible primary tests, including the Wilcoxon–Mann–Whitney rank sum test and a heteroskedasticity-robust variant due to Brunner and Munzel (2000).

Section 4.2 explains Rosenbaum's proposal and our modified approach and presents simulation evidence on the validity of bootstrap percentile confidence intervals for quantile treatment effects. Section 4.3 discusses the choice of a primary significance test and reasons to prefer the Brunner–Munzel test to the Wilcoxon–Mann–Whitney, with both a review of the literature and new simulations. Section 4.4 re-analyzes the SUNSET trial data as an illustrative example. Section 4.5 discusses benefits and limitations of the approach and directions for further research.

sumptions underlying conventional time-to-event analyses.

4.2 Estimating treatment effects

Rosenbaum's original proposal

Rosenbaum (2006) considers a completely randomized experiment: out of a finite population of N patients, we assign a simple random sample of fixed size to treatment and the remainder to control. Patients' QOL scores take values in a subset Q of the real line. For those who have died before the time of QOL measurement, the outcome is "D," indicating death, instead of a real number. The analysis requires a "placement of death" determining, for each $x \in Q$, either that x is preferred to D or vice versa. For example, two possible placements are "Death is the worst outcome" and "Death is worse than x if $x \geq 2$, but better than x if $x < 2$."[2] Any placement of death, together with the assumption that higher QOL scores are preferred to lower scores, defines a total ordering of $Q \cup \{D\}$.

Rosenbaum derives exact, randomization-based confidence intervals for order statistics of the distribution of outcomes that the treatment group patients would have experienced if they had been assigned to control. For example, his method enables statements of the form: "Ranking the 400 treatment group patients' outcomes from best to worst, the 201st value was a QOL score of 4.2. We estimate that if the same 400 patients had not received the intervention and we ranked their outcomes from best to worst, the 201st value would lie in the range $[x, y]$ (95% confidence interval)." Here x and y could be real numbers, or one or both of them could be D.

As Rubin (2006b) notes, Rosenbaum's elegant and insightful proposal deserves exploration but may be "difficult to convey to consumers." In the example above, slightly complicated language is needed to describe the quantity being estimated. A statement is being made about the treatment group patients (and since they are a random set, the estimand is a random variable). We know their actual outcome distribution, and we are constructing a confidence interval for an order statistic of the distribution that would have been observed had they been assigned to control. With 400 treatment group patients, the median is not an order statistic, so the example uses the 201st value instead. These unusual features of the approach allow the derivation of exact confidence intervals.

Alternative estimands

We borrow Rosenbaum's use of placements of death and his suggestion to offer multiple analyses corresponding to different placements, but we explore alternative estimands that may be more familiar to applied audiences. Our confidence intervals for those estimands will be approximate instead of exact.

[2]The framework could easily be modified to allow placements such as "Death is equivalent to a QOL score of 2."

In the LOS context, for each patient i, let Y_i denote a composite outcome that equals her LOS if she was discharged alive from the ICU and takes the value D otherwise. We allow D to be either a real number (meaning that death and some length of stay are considered equally undesirable) or a special nonnumeric value that is considered greater (i.e., worse) than any possible LOS. Using the potential outcomes framework [Neyman (1923); Rubin (1974, 2005)], let Y_{1i} denote the outcome that would occur if patient i were assigned to treatment. If she is actually assigned to treatment, then $Y_i = Y_{1i}$; otherwise, Y_{1i} is a counterfactual. Similarly, let Y_{0i} denote the outcome that would occur if she were assigned to control.

Assume that each pair (Y_{1i}, Y_{0i}) is an independent observation from a probability distribution with marginal distribution functions $F_1(x) = P(Y_{11} \leq x)$ and $F_0(x) = P(Y_{01} \leq x)$. An intuitive interpretation of this assumption is that the patients in the trial are a random sample from an infinite population of interest. We make this assumption for mathematical convenience and compatibility with literature cited later in this chapter, but it is probably not crucial, as standard errors, significance tests, and confidence intervals that are valid from the infinite-population perspective are typically conservative from the finite-population perspective (in which the N patients in the trial are the population of interest).[3]

Define the *treatment effect on the p quantile* as

$$\mathrm{QTE}_p = \min\{x : F_1(x) \geq p\} - \min\{x : F_0(x) \geq p\}$$

if both terms on the right-hand side are real numbers; if either term is a nonnumeric placement of death, QTE_p is undefined.[4] For example, $\mathrm{QTE}_{0.5}$ (the treatment effect on the median) is the difference between the population medians of Y_{1i} and Y_{0i}, if both are real numbers.

Define the *cutoff treatment effect at cutoff c* as

$$\mathrm{CTE}_c = P(X_{11} \geq c) - P(X_{01} \geq c).$$

For example, if LOS is measured in days, then CTE_{20} is the treatment effect on the proportion of patients with outcome at least as bad as a 20-day LOS. If death is the worst possible outcome, then CTE_D is the treatment effect on the mortality rate.

Quantile treatment effects and cutoff treatment effects are different ways of summarizing effects on the outcome distribution. QTEs may be undefined in the highest quantiles (if death is considered the worst possible outcome but is not assigned a numeric value), but CTEs are defined at all cutoffs. On the other hand, there is perhaps more danger of data dredging with CTEs, since researchers may have more leeway to choose cutoffs that yield results they like than to focus on,

[3] See, e.g., Reichardt and Gollob (1999) and Chapter 2.

[4] A nonnumeric placement means that death is the worst possible outcome, but need not imply that the difference between death and a 30-day ICU stay is considered greater than the difference between 30 and 3 days. Thus, the former difference is undefined, not infinite.

say, the treatment effect on the 0.53 quantile instead of the median. In the very different context of educational test scores, Holland (2002) argues that for measuring changes over time in the gap between two distributions, analyses of differences in proportions below a cutoff score can easily mislead. He prefers analyses of differences in quantiles and recommends supplementary graphical displays. Whether analogous issues arise in the LOS context (e.g., in comparing treatment effects for different subgroups or different interventions) is a worthwhile topic for future research.

CTEs can be estimated by differences in sample proportions, with normal-approximation confidence intervals or the finite-sample improvements recommended by Agresti and Caffo (2000) or Brown and Li (2005). QTEs can be estimated by differences in sample quantiles; we have used the version of sample quantiles recommended by Hyndman and Fan (1996) ["Definition 8," which is median-unbiased of order $o(1/\sqrt{N})$]. Below we explore the use of bootstrap percentile confidence intervals for QTEs.

Confidence intervals for QTEs

The treatment–control difference in sample quantiles is a special case of a quantile regression estimator. Hahn (1995, Theorem 3) shows that bootstrap percentile confidence intervals for quantile regression coefficients have correct asymptotic coverage probabilities, under regularity conditions that in our case imply that the distributions of Y_{1i} and Y_{0i} are continuous and their densities are bounded away from zero near the quantiles of interest. In practice, we expect some discreteness in the distributions, in part because LOS data may be rounded, but most importantly because many values will be tied at the placement of death D.

Another wrinkle is that if D is nonnumeric, then the difference in sample quantiles is undefined when one or both of the sample quantiles equal D, and thus the bootstrap percentile CI is undefined when any bootstrap replication yields a treatment or control group sample quantile equal to D. (One can still report the two sample quantiles and, in some cases, a CI for one of the population quantiles.)

To examine these issues empirically, we simulated a hypothetical trial with 1,500 patients, assigning 750 to treatment and 750 to control (slightly smaller sample sizes than the SUNSET trial's). On each of 10,000 replications of the trial:

1. We generated patients' outcomes assuming that the probability of death in the ICU was 20% for control group patients and 10% for treatment group patients. Survivors' LOS values were sampled with replacement from the data for SUNSET control group patients who survived their ICU stays. (The SUNSET data are rounded to the nearest tenth of an hour.)

Table 4.1: True quantiles of outcome distributions for simulations in Section 4.2. The table assumes the placement of death D is no less than 40.9 days. The second and third columns show quantiles of the distributions that treatment and control group patients' outcomes are randomly sampled from. Lengths of ICU stay are given in days.

Quantile	Treatment	Control
0.25	1.2	1.4
0.5 (median)	2.3	2.7
0.6	3.0	4.1
0.7	4.5	5.7
0.75	5.7	10.6
0.775	6.7	16.6
0.8	7.6	40.9
0.825	9.2	Death
0.85	11.8	Death
0.9	40.9	Death
0.95	Death	Death

Table 4.2: Coverage rates (in $10,000$ replications) of nominal 95% confidence intervals (bootstrap percentile method) for quantile treatment effects, assuming placement of death $D = 40.9$ days.

Quantile	Coverage rate (percent)
0.25	95.7
0.5 (median)	95.5
0.6	95.6
0.7	95.7
0.75	95.8
0.775	95.3
0.8	87.8
0.825	96.2
0.85	95.4
0.9	95.1
0.95	100.0

Table 4.3: Empirical properties (in $10,000$ replications) of nominal 95% confidence intervals (bootstrap percentile method) for quantile treatment effects, assuming death is the worst possible outcome.

Quantile	% of confidence intervals that:		
	Cover true value	Miss true value	Are undefined
0.25	95.7	4.3	0.0
0.5 (median)	95.5	4.5	0.0
0.6	95.6	4.4	0.0
0.7	95.7	4.2	0.2
0.75	49.9	1.7	48.4
0.775	5.2	2.2	92.6
0.8	0.0	0.2	99.8
0.825	NA	NA	100.0
0.85	NA	NA	100.0
0.9	NA	NA	100.0
0.95	NA	NA	100.0

2. Nominal 95% confidence intervals were constructed using the bootstrap percentile method with 1,000 bootstrap replications. We resampled the treatment group and control group independently with fixed sample sizes, as suggested in Efron and Tibshirani (1993, pp. 88–89) and Davison and Hinkley (1997, p. 71).

Step 1 implies the population quantiles of Y_{1i} and Y_{0i} shown in Table 4.1, if D is either nonnumeric or a number of days no less than 40.9 (the highest LOS value for survivors in the SUNSET control group). On each replication of the hypothetical trial, we observe the treatment and control sample quantiles, which are estimates of the population quantiles.

One might consider Hahn's asymptotic results least reassuring near and above the 0.8 quantile, both because the population distributions of Y_{0i} and Y_{1i} put 20% and 10% probabilities on point masses at D, and because just below the 0.8 and 0.9 population quantiles, there are nonnegligible gaps between the highest numeric LOS values in the distributions (e.g., the two highest values are 40.9 and 37.8 days). Table 4.2 assumes $D = 40.9$ days and shows a below-nominal CI coverage rate (88 percent) at the 0.8 quantile, but the effect is localized and not severe.[5]

In Table 4.3, D is nonnumeric (death is the worst possible outcome). The CIs appear to be valid at the 0.7 quantile and below. At the 0.75 quantile, the bootstrap CI is undefined in 48% of the trial replications, because the control group's 0.75 quantile equaled D on at least one bootstrap replication. At the 0.8 and higher quantiles, this situation occurs frequently, and the CI is always or almost always undefined.

These results suggest that bootstrap percentile CIs for QTEs are likely to have approximate validity near the median (as long as mortality rates are well below 50%), but caution is warranted in the upper tail of the distribution, near the placement of death. Further research with more advanced methods such as BC_a bootstrap CIs [Efron and Tibshirani (1993, ch. 14); Davison and Hinkley (1997, ch. 5)] or subsampling [Politis, Romano, and Wolf (1999)] may be worthwhile.

4.3 Choosing a primary significance test

Researchers may appropriately wish to explore QTEs or CTEs at more than one quantile or more than one cutoff value of LOS. To mitigate the resulting multiple comparisons issue, pre-specification of a primary significance test may be advisable. One possibility is to designate a specific quantile or cutoff as primary and invert the corresponding CI. The median may seem a natural choice, but some interventions may be intended to shorten long ICU stays without necessarily re-

[5] When D was raised to 200 days, undercoverage occurred not at the exact 0.8 quantile, but just above it. Otherwise the results were similar.

ducing the median. It may be difficult to predict which points in the outcome distribution are likely to be affected.

Another option is to use a rank test that has some sensitivity to effects throughout the outcome distribution, such as the Wilcoxon–Mann–Whitney (WMW) rank sum test. Rank tests do not require a numeric placement of death (unlike, e.g., the two-sample t-test). Rubin (2006b), modifying Korn's (2006) proposal, comments that the WMW test could be combined with Rosenbaum's (2006) approach. More generally, Rosenbaum has extensively explored the use of rank tests in causal inference [e.g., Rosenbaum (2007, 2010)], and Imbens and Wooldridge (2009, pp. 21–23) suggest the WMW test "as a generally applicable way of establishing whether the treatment has any effect" in randomized experiments.

The WMW test is often recommended because it is believed to have more robustness of efficiency (power) than tests based on the difference in mean outcomes; Lehmann (2009) gives a helpful overview of results that support this view. However, when the classical assumption of a constant additive treatment effect is relaxed, power comparisons vary with the nature of the anticipated treatment effect [White and Thompson (2003)], and an even more fundamental issue is the need to carefully consider what hypothesis would be useful to test [Romano (2009); Chung and Romano (2011)]. The WMW test is still valid for the strong null hypothesis that treatment has no effect on any patient (or for the hypothesis that treatment does not change the outcome distribution), but whether researchers should be satisfied with a test of the strong null is debatable.[6] The Mann–Whitney form of the test statistic naturally suggests a weaker null hypothesis, and there is an interesting, somewhat neglected literature on testing the weak null.[7]

Suppose m patients are assigned to treatment and $n = N - m$ to control. Let T and C denote the sets of indices of the treated and control patients. The Wilcoxon rank sum statistic is $\sum_{i \in T} R_i$, where R_i is the rank of Y_i among the N observations (in ascending order). Ties are often handled by the midrank method: each member of a group of tied observations is given the average of the ranks they would have if they were not tied. The rank sum statistic can be rewritten as $U + m(m+1)/2$, where U is the Mann–Whitney statistic

$$U = \sum_{i \in T} \sum_{j \in C} \left[I(Y_{1i} > Y_{0j}) + \frac{1}{2} I(Y_{1i} = Y_{0j}) \right]$$

and $I(A)$ equals 1 if A occurs and 0 otherwise.[8] We can equivalently use the

[6] R. A. Fisher and Jerzy Neyman debated the relevance of the strong null in connection with the F-test. See, e.g., Fienberg and Tanur (1996) and Gail et al. (1996) for discussion.

[7] We will discuss this literature from a causal inference perspective and continue to use the potential outcomes framework. However, the literature is not explicitly causal; it assumes two independent random samples from two infinite populations and has both causal and descriptive applications.

[8] The result is derived in, e.g., Gibbons and Chakraborti (2011, pp. 292–293).

statistic $U/mn - 1/2$, which is a consistent estimate of

$$P(Y_{1i} > Y_{0j}) + \frac{1}{2}P(Y_{1i} = Y_{0j}) - \frac{1}{2} \;=\; \frac{P(Y_{1i} > Y_{0j}) - P(Y_{1i} < Y_{0j})}{2}, \qquad i \neq j.$$

An extreme positive test statistic is evidence that $P(Y_{1i} > Y_{0j}) > P(Y_{1i} < Y_{0j})$—that is, if we sample the treated and untreated potential outcome distributions independently, it is more likely that a random treated value will exceed a random untreated value than the other way around.[9] Similarly, an extreme negative test statistic is evidence that $P(Y_{1i} > Y_{0j}) < P(Y_{1i} < Y_{0j})$.

Thus, the WMW test can be reexamined as a test of the weak null hypothesis

$$H_0^w \;:\; P(Y_{1i} > Y_{0j}) = P(Y_{1i} < Y_{0j}), \qquad i \neq j.$$

Loosely speaking, H_0^w says there is no systematic tendency for a treatment group patient's outcome to be better or worse than a control group patient's outcome. Pratt (1964) shows that in general, the WMW test is not an asymptotically valid test of H_0^w, in part because heteroskedasticity can distort the significance level. Pratt's Table 2 implies that if $m = n$, the size of a two-tailed WMW test (assuming no ties) at the nominal 5% level tends to a limit between 5% and 11%. If $m \neq n$, this range widens in both directions.

Brunner and Munzel (2000) derive an asymptotically valid test of H_0^w by studentizing $U/mn - 1/2$ (i.e., dividing by a consistent estimate of its standard error). The Brunner–Munzel (BM) test allows ties (the distributions of Y_{1i} and Y_{0i} can be of any nondegenerate form). The test statistic (which can be computed from the overall ranks R_i and the ranks within the treatment and control groups) is asymptotically $N(0,1)$ under H_0^w; to improve small-sample performance, BM suggest using the t-distribution with degrees of freedom from a Welch–Satterthwaite approximation.[10] Neubert and Brunner (2007) propose a permutation test based on the BM statistic and prove its asymptotic validity. Chung and Romano (2011) derive a general theory for constructing asymptotically valid permutation tests based on two-sample U-statistics, discuss misapplications of the WMW test, and provide a studentized permutation version (for the case without ties) whose critical values can be tabled.

Simulation evidence on test validity

Table 4.4 shows the rejection rates of the WMW and BM tests (two-tailed, at the nominal 5% level) in simulations of nine null-hypothesis scenarios with 1,500 patients and 250,000 replications. For the WMW test, we used the large-sample normal approximation [e.g., Miller (1986, p. 51)]. In each panel, we show results

[9] Estimands related to $P(Y_{1i} > Y_{0j})$ have been studied in many fields. Ho (2009) gives a helpful discussion.

[10] The BM test is implemented in the R lawstat package.

Table 4.4: Rejection rates (in 250,000 replications) of the Wilcoxon–Mann–Whitney and Brunner–Munzel tests in nine null-hypothesis scenarios. All tests are two-tailed with nominal significance level 5%.

Scenario	Rejection rate (%)	
	Wilcoxon–Mann–Whitney	Brunner–Munzel
Strong null		
Balanced design	5.0	5.0
90% treated	5.1	5.1
10% treated	5.0	5.0
Weak null (no ties)		
Balanced design	6.4	4.9
90% treated	14.7	5.0
10% treated	0.3	5.0
Weak null (with ties)		
Balanced design	5.7	5.0
90% treated	11.4	5.0
10% treated	1.2	5.0

for a balanced design (i.e., with a 1:1 treatment:control allocation ratio), and two imbalanced designs (with 9:1 and 1:9 allocation ratios).

The first panel shows rejection rates under the strong null hypothesis that treatment has no effect on any patient's outcome. For both the treatment group and the control group, the data-generating process for outcomes is identical to that used for the control group in Table 4.3: the probability of death is 20%, and survivors' LOS values are sampled with replacement from the SUNSET trial's control group data. Death is placed as the worst possible outcome. As expected, the WMW and BM tests have rejection rates close to the nominal 5% significance level.

For the second and third panels, we simulated scenarios in which H_0^w holds but the strong null does not. In each case, treatment shrinks the spread of a symmetric outcome distribution without shifting its center. The second panel assumes continuous distributions [the case analyzed by Pratt (1964)], while the third panel allows a substantial number of ties.

In the second panel, the treated and control patients' outcomes are drawn from the continuous uniform distributions on [12.5, 27.5] and [5, 35], respectively. As a test of H_0^w, the WMW test rejects somewhat too often (6.4%) with a 1:1 treatment:control allocation ratio, far too often (14.7%) with a 9:1 ratio, and rarely (0.3%) with a 1:9 ratio; these rates are very close to the asymptotic limits implied by Pratt's (1964) Table 1. In contrast, the BM test's rejection rates are always close

to the nominal 5% level.

For the third panel, outcomes are drawn from mixed discrete/continuous distributions. For treated patients, the distribution puts 20% probability on a point mass at 2.5, 60% on the uniform distribution on [12.5, 27.5], and 20% on a point mass at 37.5. The control patients' distribution is similar but the point masses are at 0 and 40 and the uniform distribution's range is [5, 35]. Again, the BM test maintains the nominal significance level but the WMW test does not (its rejection rates are 5.7%, 11.4%, and 1.2%).

In sum, the WMW test is not a valid test of H_0^w. It is valid for the strong null, but it is sensitive to certain kinds of departures from the strong null and not others. For example, it is more likely to reject the null when treatment narrows the spread of the outcome distribution and there are more treated than control patients, or when treatment widens the spread and there are more control than treated patients. It is less likely to reject when the opposite is true. These properties complicate the test's interpretation and are probably not well-known to most of its users. In contrast, the BM test is an approximately valid test of H_0^w in sufficiently large samples, and a rejection of H_0^w can be understood as evidence of a general tendency for treated patients' outcomes to be better or worse (depending on the sign of the test statistic) than those of untreated patients.

On the other hand, it is not clear whether these issues are likely to be empirically important in most clinical trials. With a balanced design, the WMW test's overrejection of H_0^w in Table 4.4 is only slight, and the simulated scenarios are perhaps extreme (e.g., in the second panel, treatment halves the standard deviation of the outcome).

Simulation evidence on power

Table 4.5 compares the abilities of three tests to detect beneficial treatment effects (i.e., reducing LOS or mortality) in various scenarios. The tests are the WMW, the BM, and the significance test for $\text{QTE}_{0.5}$ (the treatment effect on the median) constructed by inverting the bootstrap percentile confidence interval (based on 1,000 bootstrap replications). In each case we used a two-tailed test (at the 5% level) but assumed that if the null hypothesis was rejected, researchers would infer the direction of the effect from the sign of (i) the difference between the Wilcoxon rank sum statistic and its expected value, (ii) the BM statistic, or (iii) the CI limits for $\text{QTE}_{0.5}$. The upper half of the table shows the rates of correctly inferring a beneficial treatment effect (in 10,000 replications of a clinical trial); the lower half shows the rates of incorrectly inferring a harmful effect, which are very low. In each scenario, the trial includes 1,500 patients with a 1:1 treatment:control allocation, and death is placed as the worst possible outcome.

In scenario A, the probability of death in the ICU is 5% for both the treatment group and the control group, but treatment and control group survivors' LOS values are sampled from two different distributions. The control group LOS dis-

Table 4.5: Rejection rates (in 10,000 replications) of three significance tests in six alternative-hypothesis scenarios. WMW = Wilcoxon–Mann–Whitney; BM = Brunner–Munzel. The $QTE_{0.5}$ test rejects if and only if the bootstrap percentile CI for the treatment effect on the median excludes zero. All tests are two-tailed with nominal significance level 5%.

	WMW	BM	$QTE_{0.5}$
Reject, correctly infer beneficial effect (%)			
Scenario A	54.8	54.7	10.3
Scenario B	30.6	30.5	61.2
Scenario C	38.3	38.2	61.6
Scenario D	7.5	7.5	4.2
Scenario E	5.1	5.1	6.9
Scenario F	7.4	7.4	7.0
Reject, incorrectly infer harm (%)			
Scenario A	0.0	0.0	0.3
Scenario B	0.0	0.0	0.0
Scenario C	0.0	0.0	0.0
Scenario D	0.7	0.7	0.1
Scenario E	0.1	0.1	0.6
Scenario F	0.7	0.7	0.6

tribution is just the empirical distribution for the SUNSET trial's control group survivors. The treatment group LOS distribution substitutes $g(x)$ for each value x in the control group distribution, where $g(x) = x$ if x is less than or equal to 2 days and $g(x) = x/2 + 1$ days if x exceeds 2 days. The underlying idea is that the intervention is not expected to affect the shortest ICU stays, because bed space availability limits the speed at which patients can be moved from the ICU to other hospital units.

The WMW and BM tests detected a beneficial treatment effect in 55% of the replications of scenario A, while the corresponding rate for the $QTE_{0.5}$ test was only 10%. In this scenario, the true QTE is small at the median and larger in the upper half of the composite outcome distribution (except the upper 5% tail, which represents death): the intervention reduces the population median from 2.2 to 2.1 days and the 95th percentile (the highest value except for death) from 40.9 to 21.5 days.

Scenario B raises the probability of death to 20% but is otherwise identical to scenario A. Because death now occupies a larger area at the upper tail of the composite outcome distribution, the median values with and without the intervention are now higher, and the intervention reduces the population median from 2.7 to

2.4 days. Thus, the true QTE at the median is higher than in scenario A, and the $QTE_{0.5}$ test has more power, detecting a beneficial effect in 61% of the replications. The corresponding rates for the WMW and BM tests have fallen to 31%; these tests lose power when there are many ties. A possible remedy, following Follmann, Fay, and Proschan (2009) and Hallstrom (2010), is to perform a WMW test after removing an equal and maximal number of observations at the extremum (here, death) from each group. However, we do not know of a way to use this approach to construct a test that would share the BM test's property of asymptotic validity for a weak null hypothesis.

In scenario C, the intervention reduces both LOS and mortality. We assume each patient in the population belongs to one of three principal strata [Frangakis and Rubin (2002); Rubin (2006a)]: 17.5% are "never-survivors," who would die in the ICU with or without the intervention; 80% are "always-survivors," who would survive with or without the intervention; and 2.5% are "responders," who would die in the ICU without the intervention but would survive with the intervention. (For simplicity, we assume the intervention does not cause any patients to die in the ICU, although this may be unrealistic.) The control group's outcomes are generated exactly as in scenario B. The treatment group's outcome-generating process puts 17.5% probability on death, 80% on the same LOS distribution as in scenarios A and B, and 2.5% on the uniform distribution with range 14 to 28 days (thus assuming that responders have atypically long ICU stays). The WMW and BM tests have somewhat higher power than in scenario B, while the $QTE_{0.5}$ test's power is essentially unchanged. (The mortality effects are irrelevant to $QTE_{0.5}$ here, since responders' outcomes are worse than the median with or without the intervention.)

Scenarios D, E, and F are identical to A, B, and C, respectively, except that the treatment group LOS distribution only substitutes $g(x)$ for a random 25% of the values x in the control group distribution. The underlying idea is that we may expect that only a minority of patients will have their outcomes affected by whether an intensivist physician is present in the ICU at night. Table 4.5 shows that all three tests have very low power in these scenarios.

The results in Table 4.5 suggest three conclusions. First, in large samples, it seems appropriate to prefer the BM test to the WMW test, since they have approximately equal power in Table 4.5 and the BM test has much more robustness of validity in Table 4.4. Second, power comparisons between the BM test and the $QTE_{0.5}$ test vary with the nature of the treatment effect. (Arguably, in the absence of any prior information about the anticipated effect, the BM test is a more robust choice, since it has some sensitivity to effects throughout the outcome distribution, and an intervention can have practically significant effects without affecting the median.) Third, if an intervention is expected to affect LOS for only a minority of patients and only those with longer ICU stays, either a sample size greater than 1,500 or a more powerful test may be needed to detect treatment effects.

4.4 Illustrative example

SUNSET-ICU [Kerlin et al. (2013)] was conducted in the medical ICU of the Hospital of the University of Pennsylvania (a 24-bed ICU). The trial enrolled patients who were admitted between September 12, 2011, and September 11, 2012. Within each two-week block during this period (except a winter holiday block), one week was randomly assigned to the intervention staffing model and the other to the control model. In both models, daytime staff included two intensivists (attending physicians who were board-certified or board-eligible in critical care medicine), and nighttime staff included three medical residents, who were expected to review all new admissions and critical events with an intensivist or critical care fellow by phone or in person. On control nights, two intensivists were available by phone. On intervention nights, one intensivist was present in the ICU.

The staffing model on the night of admission (or the night after a daytime admission) determined whether each patient was considered a member of the treatment group or the control group. In other words, the analysis estimates the effects of being admitted during an intervention week vs. a control week. Most patients experienced only one staffing model (the median LOS was about 2 days), but patients could experience both models if they stayed in the ICU long enough. After sample exclusions detailed in Kerlin et al. (2013), there were 820 patients in the treatment group and 778 in the control group.[11]

Using a proportional hazards model with death treated as a censoring event, Kerlin et al. found no effect of intervention week admission on ICU LOS. There was also no discernible effect on ICU deaths: 17.3% of the treatment group and 16.4% of the control group died in the ICU, and the difference is not statistically significant. The concern about selection bias in analyses of survivors' LOS, which was a motivation for this chapter, is therefore lessened (although it is theoretically possible that the intervention changed the composition of the survivor group). We nevertheless present a re-analysis of the trial data here as an illustrative example.[12]

With death placed as the worst possible outcome, the Brunner–Munzel test does not reject the hypothesis of no systematic tendency for a treatment group patient's outcome to be better or worse than a control group patient's (the P-value in a two-tailed test is 0.25). The associated 95% CI for $P(Y_{1i} < Y_{0j}) + 0.5\, P(Y_{1i} = Y_{0j})$—that is, the probability that a random treatment group patient's outcome is better than a random control group patient's, plus one-half the probability that they are equally desirable (or undesirable)—is $[0.455, 0.512]$. The results are similar when death and a 30-day LOS are considered equally undesirable.

[11] The trial has a matched-pair, cluster-randomized design [Imai, King, and Nall (2009)]: the patients admitted during a week are a cluster, and each two-week block is a matched pair. For simplicity, in this example we analyze the data as if individual patients were randomized without blocking.

[12] The dataset used for this chapter excludes two control group patients who had LOS exceeding 90 days. The maximum LOS in the dataset is 40.9 days.

Table 4.6: Estimated quantile treatment effects in the SUNSET trial. The second and third columns show the sample quantiles for the treatment and control groups (lengths of ICU stay are given in days). The fourth column shows their difference, the estimated QTE. The fifth column shows a 95% confidence interval (bootstrap percentile method) for the QTE. The top panel assumes death is the worst possible outcome. The bottom panel assumes death and a 30-day ICU stay are considered equally undesirable.

Quantile	Treatment	Control	Difference	95% CI
Death is worst outcome				
0.25	1.4	1.3	0.1	[−0.1, 0.3]
0.5 (median)	2.8	2.6	0.2	[−0.2, 0.7]
0.6	4.0	3.7	0.3	[−0.3, 1.3]
0.7	7.3	5.8	1.5	[−0.5, 3.2]
0.75	10.0	8.1	1.9	[−1.1, 6.6]
0.8	17.3	12.9	4.3	Undefined
0.9	Death	Death	Undefined	Undefined
Death placed at 30 days				
0.25	1.4	1.3	0.1	[−0.1, 0.3]
0.5 (median)	2.8	2.6	0.2	[−0.2, 0.7]
0.6	4.0	3.7	0.3	[−0.3, 1.3]
0.7	7.3	5.8	1.5	[−0.5, 3.2]
0.75	10.0	8.1	1.9	[−1.1, 6.6]
0.8	17.3	12.9	4.3	[−8.2, 18.7]
0.9	30.0	30.0	0.0	[0.0, 0.0]

The top panel of Table 4.6 shows estimated quantile treatment effects and 95% confidence intervals (using the bootstrap percentile method with 1,000 replications), with death placed as the worst outcome. There is no evidence that the intervention affected the median outcome or any of the other quantiles examined. The CIs for the treatment effects on the 25th to 75th percentiles of the outcome distribution all include zero. Our method is unable to perform inference for treatment effects at the 80th percentile and above. The 80th percentile outcome for the treatment group is a 17.3-day LOS, compared to a 12.9-day LOS for the control group, but one or both values are death on some of the bootstrap replications, so the method cannot produce a confidence interval without additional assumptions about how to value the difference between death and a numeric LOS. The 90th percentile is death in both groups. For a general audience, one might present the results for the 0.25 to 0.75 quantiles together with a CI for the intervention's effect

Table 4.7: Estimated cutoff treatment effects in the SUNSET trial. The second and third columns show the treatment and control group sample proportions with outcomes at least as bad as the cutoff. The fourth column shows their difference, the estimated CTE. The fifth column shows a 95% confidence interval (normal approximation) for the CTE. The top panel assumes death is the worst possible outcome. The bottom panel assumes death and a 30-day ICU stay are considered equally undesirable.

Cutoff	% with outcome at least as bad as cutoff			
	Treatment	Control	Difference	95% CI
Death is worst outcome				
1 day in ICU	84.3	83.2	1.0	$[-2.6, 4.6]$
2 days	60.4	59.0	1.3	$[-3.5, 6.2]$
3 days	48.0	44.5	3.6	$[-1.3, 8.5]$
4 days	40.1	37.8	2.4	$[-2.4, 7.1]$
1 week	30.9	27.3	3.5	$[-0.9, 8.0]$
2 weeks	22.1	19.6	2.5	$[-1.5, 6.5]$
4 weeks	18.7	16.9	1.8	$[-2.0, 5.5]$
Death	17.3	16.4	1.0	$[-2.7, 4.6]$
Death placed at 30 days				
1 day in ICU	84.3	83.2	1.0	$[-2.6, 4.6]$
2 days	60.4	59.0	1.3	$[-3.5, 6.2]$
3 days	48.0	44.5	3.6	$[-1.3, 8.5]$
4 days	40.1	37.8	2.4	$[-2.4, 7.1]$
1 week	30.9	27.3	3.5	$[-0.9, 8.0]$
2 weeks	22.1	19.6	2.5	$[-1.5, 6.5]$
4 weeks	18.7	16.9	1.8	$[-2.0, 5.5]$
30 days	18.5	16.9	1.7	$[-2.1, 5.4]$
5 weeks	0.7	0.3	0.5	$[-0.2, 1.2]$

on mortality, which will be given below.

The bottom panel of Table 4.6 repeats the analysis with death and a 30-day LOS considered equally undesirable. The results for the 25th to 75th percentiles are unchanged. At the 80th percentile, a CI can now be constructed, and it cannot rule out a strong beneficial effect (shortening LOS by 8.2 days), a strong harmful effect (lengthening LOS by 18.7 days), or no effect. The 90th percentile outcome is 30 (representing death) in both the treatment group and the control group and the CI excludes all nonzero values.

The top panel of Table 4.7 shows estimated effects on the proportions of patients with outcomes at least as bad as various cutoff values, with 95% CIs based on the

normal approximation. Death in the ICU is placed as the worst outcome, and the last row of the panel shows the estimated effect on mortality (the point estimate is 1.0 percentage point, but the CI ranges from -2.7 to 4.6 percentage points). All the CIs include zero, so there is no evidence that the intervention affected the ICU death rate or any of the other proportions. The bottom panel repeats the analysis with death and a 30-day LOS considered equally undesirable; the results are similar.

In sum, our analysis finds no evidence that the intervention affected the distribution of patients' outcomes, regardless of whether death is considered the worst possible outcome or placed as comparable to an LOS as short as 30 days. Since there was little difference in ICU mortality between the treatment and control groups, it is not surprising that the original analysis in Kerlin et al. (2013) and the re-analysis presented here yield similar conclusions.

4.5 Discussion

The placement-of-death approach does not estimate treatment effects on LOS per se. Instead, it estimates effects on the distribution of a composite outcome measure based on ICU mortality and survivors' LOS. Researchers may hope to disentangle those effects and to estimate treatment effects on the LOS of always-survivors—those patients who would have survived their ICU stays regardless of whether they were assigned to the intervention.[13] Such questions are important, but stronger assumptions are needed to study them. Thus, our approach is not a substitute for additional modeling, but it may be a useful starting point. It addresses concerns about selection bias by comparing the entire treatment group with the entire control group, and it can provide evidence of an overall beneficial or harmful effect.

The approach allows sensitivity analysis with alternative placements of death, but it does make some restrictive assumptions about valuations of LOS and death. For example, it awards no credit for reducing long ICU stays of patients who would die in the ICU with or without the intervention, although such an effect may be in accordance with some patients' wishes. Extending the approach to accomodate more complicated valuations may be a useful direction for further work. Alternatively, social cost-benefit analysis could be considered. However, the placement-of-death approach may be more appealing to some audiences because it avoids the need to assign a numeric value to death.

Of the significance tests we studied, the Brunner–Munzel test (or a permutation test based on the BM statistic) may be a reasonable choice for an omnibus primary test. Some other rank tests may have more power when there are many ties [Follmann, Fay, and Proschan (2009); Hallstrom (2010)] or when a small fraction

[13] Rubin (2006a) explains the concept of treatment effects on always-survivors and the difficulties involved in estimating them. Joffe (2011) argues for a broader focus.

of treated patients experience large treatment effects [see Rosenbaum (2007) and references therein]. It is not clear that any of those other tests can be easily converted into robust tests of weak null hypotheses, but further investigation may be worthwhile.

Extension of the approach to cover cluster-randomized trials would also be valuable. Rosenbaum's (2006) original approach provided exact confidence intervals in experiments with complete random assignment of individuals, but more complex designs create difficulties for exact inference.

Adjustment for treatment–control imbalances in the distributions of baseline covariates may be desired. One option that could be investigated is to combine the placement-of-death approach with inverse propensity score weighting.

References

AGRESTI, A. and CAFFO, B. (2000). Simple and effective confidence intervals for proportions and differences of proportions result from adding two successes and two failures. *Amer. Statist.* **54** 280–288.

ANGRIST, J. D. (2004). American education research changes tack. *Oxf. Rev. Econ. Policy* **20** 198–212.

ANGRIST, J. D. and IMBENS, G. W. (2002). Comment on "Covariance adjustment in randomized experiments and observational studies" by P. R. Rosenbaum. *Statist. Sci.* **17** 304–307.

ANGRIST, J. D., LANG, D. and OREOPOULOS, P. (2009). Incentives and services for college achievement: Evidence from a randomized trial. *Am. Econ. J.: Appl. Econ.* **1:1** 136–163.

ANGRIST, J. D. and PISCHKE, J.-S. (2009). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton Univ. Press, Princeton, NJ.

ASHENFELTER, O. and PLANT, M. W. (1990). Nonparametric estimates of the labor-supply effects of negative income tax programs. *J. Labor Econ.* **8** S396–S415.

BERK, R., BARNES, G., AHLMAN, L. and KURTZ, E. (2010). When second best is good enough: A comparison between a true experiment and a regression discontinuity quasi-experiment. *J. Exp. Criminol.* **6** 191–208.

BERK, R., BROWN, L., BUJA, A., GEORGE, E., PITKIN, E., ZHANG, K. and ZHAO, L. (2013). Misspecified mean function regression: Making good use of regression models that are wrong. Working paper, Univ. of Pennsylvania.

BLOOM, H. S., ORR, L. L., CAVE, G., BELL, S. H. and DOOLITTLE, F. with LIN, W., BOS, J., TOUSSAINT, C. and KORNFELD, R. (1993). *The National JTPA Study: Title II-A Impacts on Earnings and Employment at 18 Months.* Abt Associates, Bethesda, MD.

BROWN, L. and LI, X. (2005). Confidence intervals for two sample binomial distribution. *J. Statist. Plann. Inference* **130** 359–375.

BRUNNER, E. and MUNZEL, U. (2000). The nonparametric Behrens-Fisher problem: Asymptotic theory and a small-sample approximation. *Biom. J.* **42** 17–25.

BUJA, A., BERK, R., BROWN, L., GEORGE, E., TRASKIN, M., ZHANG, K. and ZHAO, L. (2012). A conspiracy of random X and model violation against classical inference in linear regression. Working paper, Univ. of Pennsylvania.

CHAMBERLAIN, G. (1982). Multivariate regression models for panel data. *J. Econometrics* **18** 5–46.

CHUNG, E. Y. and ROMANO, J. P. (2011). Asymptotically valid and exact permutation tests based on two-sample U-statistics. Technical Report 2011-09, Dept. of Statistics, Stanford Univ.

CHUNG, E. Y. and ROMANO, J. P. (2012). Exact and asymptotically robust permutation tests. *Ann. Statist.* To appear.

COCHRAN, W. G. (1942). Sampling theory when the sampling-units are of unequal sizes. *J. Amer. Statist. Assoc.* **37** 199–212.

COCHRAN, W. G. (1957). Analysis of covariance: Its nature and uses. *Biometrics* **13** 261–281.

COCHRAN, W. G. (1969). The use of covariance in observational studies. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **18** 270–275.

COCHRAN, W. G. (1977). *Sampling Techniques*, 3rd ed. Wiley, New York.

COX, D. R. and MCCULLAGH, P. (1982). Some aspects of analysis of covariance (with discussion). *Biometrics* **38** 541–561.

COX, D. R. and REID, N. (2000). *The Theory of the Design of Experiments*. CRC Press, Boca Raton, FL.

DAVIDSON, R. and MACKINNON, J. G. (1993). *Estimation and Inference in Econometrics*. Oxford Univ. Press, Oxford.

DAVISON, A. C. and HINKLEY, D. V. (1997). *Bootstrap Methods and Their Application*. Cambridge Univ. Press, Cambridge.

DEATON, A. (2010). Instruments, randomization, and learning about development. *J. Econ. Lit.* **48** 424–455.

EFRON, B. and TIBSHIRANI, R. J. (1993). *An Introduction to the Bootstrap*. CRC Press, Boca Raton, FL.

EICKER, F. (1967). Limit theorems for regressions with unequal and dependent errors. *Proc. Fifth Berkeley Sympos. Math. Statist. and Probability* **1** 59–82.

FIENBERG, S. E. and TANUR, J. M. (1987). Experimental and sampling structures: Parallels diverging and meeting. *Internat. Statist. Rev.* **55** 75–96. Amendment: **56** 197.

FIENBERG, S. E. and TANUR, J. M. (1996). Reconsidering the fundamental contributions of Fisher and Neyman on experimentation and sampling. *Internat. Statist. Rev.* **64** 237–253.

FIRTH, D. and BENNETT, K. (1998). Robust models in probability sampling. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **60** 3–21.

FISHER, R. A. (1932). *Statistical Methods for Research Workers*, 4th ed. Oliver and Boyd, Edinburgh.

FISHER, R. A. (1935). *The Design of Experiments*. Oliver and Boyd, Edinburgh.

FOLLMANN, D., FAY, M. P. and PROSCHAN, M. (2009). Chop-lump tests for vaccine trials. *Biometrics* **65** 885–893.

FRANGAKIS, C. E. and RUBIN, D. B. (2002). Principal stratification in causal inference. *Biometrics* **58** 21–29.

FREEDMAN, D. A. (1991). Statistical models and shoe leather (with discussion). *Socio. Meth.* **21** 291–358.

FREEDMAN, D. A. (2004). Sampling. In *Encyclopedia of Social Science Research Methods* (M. Lewis-Beck, A. Bryman, and T. F. Liao, eds.) **3** 986–990. SAGE, Thousand Oaks, CA.

FREEDMAN, D. A. (2006). On the so-called "Huber sandwich estimator" and "robust standard errors." *Amer. Statist.* **60** 299–302.

FREEDMAN, D. A. (2008a). On regression adjustments to experimental data. *Adv. in Appl. Math.* **40** 180–193.

FREEDMAN, D. A. (2008b). On regression adjustments in experiments with several treatments. *Ann. Appl. Stat.* **2** 176–196.

FREEDMAN, D. A. (2008c). Editorial: Oasis or mirage? *Chance* **21:1** 59–61. Annotated references at www.stat.berkeley.edu/~census/chance.pdf.

FREEDMAN, D. A. (2008d). Randomization does not justify logistic regression. *Statist. Sci.* **23** 237–249.

FREEDMAN, D. A. (2010). Survival analysis: An epidemiological hazard? In *Statistical Models and Causal Inference: A Dialogue with the Social Sciences* (D. Collier, J. S. Sekhon, and P. B. Stark, eds.) 169–192. Cambridge Univ. Press, Cambridge.

FREEDMAN, D. A., PISANI, R. and PURVES, R. (2007). *Statistics*, 4th ed. Norton, New York.

FRIEDMAN, J. (2012). Comment on "Regression adjustment in randomized experiments: Is the cure really worse than the disease? (Part I)" by W. Lin. Development Impact blog comment, World Bank. Available at http://blogs.worldbank.org/impactevaluations/comment/860#comment-860.

FULLER, W. A. (1975). Regression analysis for sample survey. *Sankhyā Ser. C* **37** 117–132.

FULLER, W. A. (2002). Regression estimation for survey samples. *Surv. Meth.* **28** 5–23.

FULLER, W. A. (2009). *Sampling Statistics*. Wiley, Hoboken, NJ.

GAIL, M. H., MARK, S. D., CARROLL, R. J., GREEN, S. B. and PEE, D. (1996). On design considerations and randomization-based inference for community intervention trials. *Stat. Med.* **15** 1069–1092.

GELMAN, A. and PARDOE, I. (2007). Average predictive comparisons for models with nonlinearity, interactions, and variance components. *Socio. Meth.* **37** 23–51.

GIBBONS, J. D. and CHAKRABORTI, S. (2011). *Nonparametric Statistical Inference*, 5th ed. CRC Press, Boca Raton, FL.

GOLDBERGER, A. S. (1991). *A Course in Econometrics*. Harvard Univ. Press, Cambridge, MA.

GREEN, D. P. and ARONOW, P. M. (2011). Analyzing experimental data using regression: When is bias a practical concern? Working paper, Yale Univ.

GREENBERG, D. and SHRODER, M. (2004). *The Digest of Social Experiments*, 3rd ed. Urban Institute Press, Washington, DC.

HAHN, J. (1995). Bootstrapping quantile regression estimators. *Econometric Theory* **11** 105–121.

HALLSTROM, A. P. (2010). A modified Wilcoxon test for non-negative distributions with a clump of zeros. *Stat. Med.* **29** 391–400.

HANSEN, B. B. and BOWERS, J. (2009). Attributing effects to a cluster-randomized get-out-the-vote campaign. *J. Amer. Statist. Assoc.* **104** 873–885.

HINKLEY, D. V. (1977). Jackknifing in unbalanced situations. *Technometrics* **19** 285–292.

HINKLEY, D. V. and WANG, S. (1991). Efficiency of robust standard errors for regression coefficients. *Comm. Statist. Theory Methods* **20** 1–11.

HO, A. D. (2009). A nonparametric framework for comparing trends and gaps across tests. *J. Educ. Behav. Stat.* **34** 201–228.

HOAGLIN, D. C. and WELSCH, R. E. (1978). The hat matrix in regression and ANOVA. *Amer. Statist.* **32** 17–22.

HOLLAND, P. W. (1986). Statistics and causal inference (with discussion). *J. Amer. Statist. Assoc.* **81** 945–970.

HOLLAND, P. W. (2002). Two measures of change in the gaps between the CDFs of test-score distributions. *J. Educ. Behav. Stat.* **27** 3–17.

HOLT, D. and SMITH, T. M. F. (1979). Post stratification. *J. Roy. Statist. Soc. Ser. A* **142** 33–46.

HUBER, P. J. (1967). The behavior of maximum likelihood estimates under non-standard conditions. *Proc. Fifth Berkeley Sympos. Math. Statist. and Probability* **1** 221–233.

HYNDMAN, R. J. and FAN, Y. (1996). Sample quantiles in statistical packages. *Amer. Statist.* **50** 361–365.

IMAI, K., KING, G. and NALL, C. (2009). The essential role of pair matching in cluster-randomized experiments, with application to the Mexican universal health insurance evaluation (with discussion). *Statist. Sci.* **24** 29–72.

IMBENS, G. W. (2010). Better LATE than nothing: Some comments on Deaton (2009) and Heckman and Urzua (2009). *J. Econ. Lit.* **48** 399–423.

IMBENS, G. W. (2011a). Comments on "Advice from G. Imbens on design & analysis of RCTs" by C. Samii. Available at http://cyrussamii.com/?p=837#comment-438 and http://cyrussamii.com/?p=837#comment-440.

IMBENS, G. W. (2011b). Comment on "Why it doesn't make sense in general to form confidence intervals by inverting hypothesis tests" by A. Gelman. Available at http://andrewgelman.com/2011/08/25/why_it_doesnt_m/#comment-62163.

IMBENS, G. W. and WOOLDRIDGE, J. M. (2009). Recent developments in the econometrics of program evaluation. *J. Econ. Lit.* **47** 5–86.

JOFFE, M. (2011). Principal stratification and attribution prohibition: Good ideas taken too far. *Int. J. Biostat.* **7:1** Article 35.

KERLIN, M. P., SMALL, D. S., COONEY, E., FUCHS, B. D., BELLINI, L. M., MIKKELSEN, M. E., SCHWEICKERT, W. D., BAKHRU, R. N., GABLER, N. B., HARHAY, M. O., HANSEN-FLASCHEN, J. and HALPERN, S. D. (2013). A randomized clinical trial of nighttime intensivist staffing in a medical intensive care unit. Unpublished manuscript, Univ. of Pennsylvania.

KLAR, N. and DARLINGTON, G. (2004). Methods for modelling change in cluster randomization trials. *Stat. Med.* **23** 2341–2357.

KLINE, P. (2011). Oaxaca–Blinder as a reweighting estimator. *Am. Econ. Rev.* **101:3** 532–537.

KLINE, P. and SANTOS, A. (2012). Higher order properties of the wild bootstrap under misspecification. *J. Econometrics* **171** 54–70.

KORN, E. L. (2006). Comment: Causal inference in the medical area. *Statist. Sci.* **21** 310–312.

LEHMANN, E. L. (1999). *Elements of Large-Sample Theory*. Springer, New York.

LEHMANN, E. L. (2009). Parametric versus nonparametrics: Two alternative methodologies (with discussion). *J. Nonparametr. Stat.* **21** 397–426.

LILLY, C. M., CODY, S., ZHAO, H., LANDRY, K., BAKER, S. P., MCILWAINE, J., CHANDLER, M. W. and IRWIN, R. S. (2011). Hospital mortality, length of stay, and preventable complications among critically ill patients before and after tele-ICU reengineering of critical care processes. *J. Am. Med. Assoc.* **305** 2175–2183.

LIN, W. (1999). Estimating impacts on binary outcomes under random assignment: Unadjusted, OLS, or logit? Unpublished technical note, Social Research and Demonstration Corp., Ottawa.

LIN, W. (2012a). Regression adjustment in randomized experiments: Is the cure really worse than the disease? (Part I). Development Impact blog post, World Bank. Available at http://blogs.worldbank.org/impactevaluations/regression-adjustment-in-randomized-experiments-is-the-cure-really-worse-than-the-disease.

LIN, W. (2012b). Regression adjustment in randomized experiments: Is the cure really worse than the disease? (Part II). Development Impact blog post, World Bank. Available at http://blogs.worldbank.org/impactevaluations/guest-post-by-winston-lin-regression-adjustment-in-randomized-experiments-is-the-cure-really-worse-0.

LIN, W., ROBINS, P. K., CARD, D., HARKNETT, K. and LUI-GURR, S. with PAN, E. C., MIJANOVICH, T., QUETS, G. and VILLENEUVE, P. (1998). *When Financial Incentives Encourage Work: Complete 18-Month Findings from the Self-Sufficiency Project*. Social Research and Demonstration Corp., Ottawa.

LUMLEY, T. (2010). *Complex Surveys: A Guide to Analysis Using R*. Wiley, Hoboken, NJ.

MACKINNON, J. G. (2013). Thirty years of heteroskedasticity-robust inference. In *Recent Advances and Future Directions in Causality, Prediction, and Specification Analysis: Essays in Honor of Halbert L. White Jr.* (X. Chen and N. R. Swanson, eds.) 437–461. Springer, New York.

MACKINNON, J. G. and WHITE, H. (1985). Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *J. Econometrics* **29** 305–325.

MARIK, P. E. and HEDMAN, L. (2000). What's in a day? Determining intensive care unit length of stay. *Crit. Care Med.* **28** 314–321.

MEHTA, S., BURRY, L., COOK, D., FERGUSSON, D., STEINBERG, M., GRANTON, J., HERRIDGE, M., FERGUSON, N., DEVLIN, J., TANIOS, M., DODEK, P., FOWLER, R., BURNS, K., JACKA, M., OLAFSON, K., SKROBIK, Y., HÉBERT, P., SABRI, E. and MEADE, M. (2012). Daily sedation interruption in mechanically ventilated critically ill patients cared for with a sedation protocol: A randomized controlled trial. *J. Am. Med. Assoc.* **308** 1985–1992.

MEYER, B. D. (1995). Lessons from the U.S. unemployment insurance experiments. *J. Econ. Lit.* **33** 91–131.

MIDDLETON, J. A. and ARONOW, P. M. (2012). Unbiased estimation of the average treatment effect in cluster-randomized experiments. Working paper, Yale Univ.

MILLER, R. G. (1986). *Beyond ANOVA: Basics of Applied Statistics*. Wiley, New York.

MIRATRIX, L. W., SEKHON, J. S. and YU, B. (2013). Adjusting treatment effect estimates by post-stratification in randomized experiments. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **75** 369–396.

MOHER, D., HOPEWELL, S., SCHULZ, K. F., MONTORI, V., GØTZSCHE, P. C., DEVEREAUX, P. J., ELBOURNE, D., EGGER, M. and ALTMAN, D. G. (2010). CONSORT 2010 explanation and elaboration: Updated guidelines for reporting parallel group randomised trials. *BMJ* **340** c869. Correction: **343** d6131.

NEUBERT, K. and BRUNNER, E. (2007). A studentized permutation test for the non-parametric Behrens–Fisher problem. *Comput. Statist. Data Anal.* **51** 5192–5204.

NEYMAN, J. (1923). On the application of probability theory to agricultural experiments. Translated and edited by D. M. Dabrowska and T. P. Speed (1990). *Statist. Sci.* **5** 463–480 (with discussion).

POLITIS, D. N., ROMANO, J. P. and WOLF, M. (1999). *Subsampling*. Springer, New York.

PRATT, J. W. (1964). Robustness of some procedures for the two-sample location problem. *J. Amer. Statist. Assoc.* **59** 665–680.

RAPOPORT, J., TERES, D., ZHAO, Y. and LEMESHOW, S. (2003). Length of stay data as a guide to hospital economic performance for ICU patients. *Med. Care* **41** 386–397.

RAUDENBUSH, S. W. (1997). Statistical analysis and optimal design for cluster randomized trials. *Psychol. Meth.* **2** 173–185.

REICHARDT, C. S. and GOLLOB, H. F. (1999). Justifying the use and increasing the power of a t test for a randomized experiment with a convenience sample. *Psychol. Meth.* **4** 117–128.

ROMANO, J. P. (2009). Discussion of "Parametric versus nonparametrics: Two alternative methodologies" by E. L. Lehmann. *J. Nonparametr. Stat.* **21** 419–424.

ROSENBAUM, P. R. (2002). Covariance adjustment in randomized experiments and observational studies (with discussion). *Statist. Sci.* **17** 286–327.

ROSENBAUM, P. R. (2006). Comment: The place of death in the quality of life. *Statist. Sci.* **21** 313–316.

ROSENBAUM, P. R. (2007). Confidence intervals for uncommon but dramatic responses to treatment. *Biometrics* **63** 1164–1171.

ROSENBAUM, P. R. (2010). *Design of Observational Studies*. Springer, New York.

ROYALL, R. M. and CUMBERLAND, W. G. (1978). Variance estimation in finite population sampling. *J. Amer. Statist. Assoc.* **73** 351–358.

RUBIN, DANIEL B. and VAN DER LAAN, M. J. (2011). Targeted ANCOVA estimator in RCTs. In *Targeted Learning: Causal Inference for Observational and Experimental Data* (M. J. van der Laan and S. Rose, eds.) 201–215. Springer, New York.

RUBIN, DONALD B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.* **66** 688–701.

RUBIN, DONALD B. (1984). William G. Cochran's contributions to the design, analysis, and evaluation of observational studies. In *W. G. Cochran's Impact on Statistics* (P. S. R. S. Rao and J. Sedransk, eds.) 37–69. Wiley, New York.

RUBIN, DONALD B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *J. Amer. Statist. Assoc.* **100** 322–331.

RUBIN, DONALD B. (2006a). Causal inference through potential outcomes and principal stratification: Application to studies with "censoring" due to death (with discussion). *Statist. Sci.* **21** 299–321.

RUBIN, DONALD B. (2006b). Rejoinder. *Statist. Sci.* **21** 319–321.

SAMII, C. and ARONOW, P. M. (2012). On equivalencies between design-based and regression-based variance estimators for randomized experiments. *Statist. Probab. Lett.* **82** 365–370.

SÄRNDAL, C.-E., SWENNSON, B. and WRETMAN, J. (1992). *Model Assisted Survey Sampling*. Springer, New York.

SCHOCHET, P. Z. (2010). Is regression adjustment supported by the Neyman model for causal inference? *J. Statist. Plann. Inference* **140** 246–259.

SENN, S. J. (1989). Covariate imbalance and random allocation in clinical trials. *Stat. Med.* **8** 467–475.

STOCK, J. H. (2010). The other transformation in econometric practice: Robust tools for inference. *J. Econ. Perspect.* **24:2** 83–94.

STONEHOUSE, J. M. and FORRESTER, G. J. (1998). Robustness of the *t* and *U* tests under combined assumption violations. *J. Appl. Stat.* **25** 63–74.

THURSTON, W. P. (1994). On proof and progress in mathematics. *Bull. Amer. Math. Soc. (N.S.)* **30** 161–177. Available at http://arxiv.org/abs/math/9404236.

THURSTON, W. P. (2006). Foreword. In J. H. Hubbard, *Teichmüller Theory and Applications to Geometry, Topology, and Dynamics, Vol. 1* xi–xiv. Matrix Editions, Ithaca, NY. Available at http://matrixeditions.com/Thurstonforeword.html.

TIBSHIRANI, R. (1986). Discussion of "Jackknife, bootstrap and other resampling methods in regression analysis" by C. F. J. Wu. *Ann. Statist.* **14** 1335–1339. Correction: **16** 479.

TSIATIS, A. A., DAVIDIAN, M., ZHANG, M. and LU, X. (2008). Covariate adjustment for two-sample treatment comparisons in randomized clinical trials: A principled yet flexible approach. *Stat. Med.* **27** 4658–4677.

TUKEY, J. W. (1991). Use of many covariates in clinical trials. *Internat. Statist. Rev.* **59** 123–137.

TUKEY, J. W. (1993). Tightening the clinical trial. *Contr. Clin. Trials* **14** 266–285.

WATSON, D. J. (1937). The estimation of leaf area in field crops. *J. Agr. Sci.* **27** 474–483.

WELCH, B. L. (1949). Appendix: Further note on Mrs. Aspin's tables and on certain approximations to the tabled function. *Biometrika* **36** 293–296.

WHITE, H. (1980a). Using least squares to approximate unknown regression functions. *Int. Econ. Rev.* **21** 149–170.

WHITE, H. (1980b). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica* **48** 817–838.

WHITE, I. R. and THOMPSON, S. G. (2003). Choice of test for comparing two groups, with particular application to skewed outcomes. *Stat. Med.* **22** 1205–1215.

WU, C. F. J. (1986). Jackknife, bootstrap and other resampling methods in regression analysis (with discussion). *Ann. Statist.* **14** 1261–1350.

YANG, L. and TSIATIS, A. A. (2001). Efficiency study of estimators for a treatment effect in a pretest–posttest trial. *Amer. Statist.* **55** 314–321.

Appendix A

Proofs for Chapter 2

A.1 Additional notation and definitions

In Chapter 2:

- Section 2.2 defines the basic notation;
- Section 2.4 states Conditions 1–3, defines the vectors $\mathbf{Q}_a$ and $\mathbf{Q}_b$ and the prediction errors a_i^* and b_i^*, and introduces the σ_x^2 and $\sigma_{x,y}$ notation for population variances and covariances;
- Section 2.5 defines the vector $\mathbf{Q}$ and the prediction errors a_i^{**} and b_i^{**}.

Let $\tilde{p}_A = n_A/n$ [as in remark (iii) after Corollary 2.1.2].

Extend Section 2.2's notation for population and group means to cover any scalar, vector, or matrix expression. For example:

$$\overline{ab}_A = \frac{1}{n_A}\sum_{i\in A} a_i b_i, \qquad \overline{a\mathbf{z}}_A = \frac{1}{n_A}\sum_{i\in A} a_i \mathbf{z}_i, \qquad \overline{\mathbf{z}'\mathbf{z}}_A = \frac{1}{n_A}\sum_{i\in A} \mathbf{z}_i'\mathbf{z}_i.$$

Extend Freedman's (2008b) angle bracket notation to cover all the finite limits assumed in Condition 2. For example:

$$\langle a\mathbf{z}\rangle = \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n} a_i\mathbf{z}_i, \qquad \langle \mathbf{z}'\mathbf{z}\rangle = \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n} \mathbf{z}_i'\mathbf{z}_i.$$

(The second limit exists since it is a submatrix of $\lim_{n\to\infty} n^{-1}\mathbf{Z}'\mathbf{Z}$.)

Condition 4 (centering) will sometimes be assumed for convenience. The proofs will explain why this can be done without loss of generality.

Condition 4. The population means of the potential outcomes and the covariates are zero: $\overline{a} = \overline{b} = 0$ and $\overline{\mathbf{z}} = \mathbf{0}$.

Some transformations of the regressors will be useful in the proofs. Define the pooled-slopes regression estimator of mean potential outcomes, $\widehat{\boldsymbol{\beta}}_{\text{adj}}$, as the 2×1 vector containing the estimated coefficients on T_i and $1 - T_i$ from the no-intercept OLS regression of Y_i on T_i, $1 - T_i$, and $\mathbf{z}_i - \overline{\mathbf{z}}$. Let $\widehat{\mathbf{Q}}$ denote the vector of estimated coefficients on $\mathbf{z}_i - \overline{\mathbf{z}}$ from the same regression.

The vector $\widehat{\boldsymbol{\beta}}_{\text{adj}}$ is an estimate of $\boldsymbol{\beta} = (\overline{a}, \overline{b})'$. By well-known invariance properties of least squares, $\widehat{\text{ATE}}_{\text{adj}}$ is the difference between the two elements of $\widehat{\boldsymbol{\beta}}_{\text{adj}}$.

Similarly, define the separate-slopes regression estimator of mean potential outcomes, $\widehat{\boldsymbol{\beta}}_{\text{interact}}$, as the 2×1 vector containing the estimated coefficients on T_i and $1 - T_i$ from the no-intercept OLS regression of Y_i on T_i, $1 - T_i$, $\mathbf{z}_i - \overline{\mathbf{z}}$, and $T_i(\mathbf{z}_i - \overline{\mathbf{z}})$. Then $\widehat{\text{ATE}}_{\text{interact}}$ is the difference between the two elements of $\widehat{\boldsymbol{\beta}}_{\text{interact}}$.

Let $\widehat{\mathbf{Q}}_a$ and $\widehat{\mathbf{Q}}_b$ denote the vectors of estimated coefficients on $\mathbf{z}_i$ in the OLS regressions of Y_i on $\mathbf{z}_i$ in groups A and B, respectively.

Conditions 1–3 do not rule out the possibility that under some realizations of random assignment, the regressors are perfectly collinear. The probability of this event converges to zero by Conditions 2 and 3, so it is irrelevant to the asymptotic results. For concreteness, whenever $\widehat{\text{ATE}}_{\text{adj}}$ cannot be computed because of collinearity, let $\widehat{\text{ATE}}_{\text{adj}} = \overline{Y}_A - \overline{Y}_B$, $\widehat{\mathbf{Q}} = \mathbf{0}$, and $\widehat{\boldsymbol{\beta}}_{\text{adj}} = (\overline{Y}_A, \overline{Y}_B)'$; whenever $\widehat{\text{ATE}}_{\text{interact}}$ cannot be computed, let $\widehat{\text{ATE}}_{\text{interact}} = \overline{Y}_A - \overline{Y}_B$, $\widehat{\mathbf{Q}}_a = \mathbf{0}$, $\widehat{\mathbf{Q}}_b = \mathbf{0}$, and $\widehat{\boldsymbol{\beta}}_{\text{interact}} = (\overline{Y}_A, \overline{Y}_B)'$. Other arbitrary values could be used.

A.2 Lemmas

Lemma A.1 is a finite-population version of the Weak Law of Large Numbers.

Lemma A.1. *Assume Conditions 1–3. The means over group A or group B of a_i, b_i, $\mathbf{z}_i$, a_i^2, b_i^2, $\mathbf{z}_i'\mathbf{z}_i$, $a_i b_i$, $a_i \mathbf{z}_i$, and $b_i \mathbf{z}_i$ converge in probability to the limits of the population means. For example:*

$$\begin{aligned} \overline{a}_A &\xrightarrow{p} \langle a \rangle, \\ \overline{a^2}_A \equiv \frac{1}{n_A} \sum_{i \in A} a_i^2 &\xrightarrow{p} \langle a^2 \rangle, \\ \overline{ab}_A &\xrightarrow{p} \langle ab \rangle, \\ \overline{a\mathbf{z}}_A &\xrightarrow{p} \langle a\mathbf{z} \rangle, \\ \overline{\mathbf{z}'\mathbf{z}}_A &\xrightarrow{p} \langle \mathbf{z}'\mathbf{z} \rangle. \end{aligned}$$

Proof. From basic results on simple random sampling [e.g., Freedman's (2008b) Proposition 1], $E(\overline{a}_A) = \overline{a}$ and

$$\text{var}(\overline{a}_A) = \frac{1}{n-1} \frac{1 - \tilde{p}_A}{\tilde{p}_A} \sigma_a^2.$$

As $n \to \infty$, $\tilde{p}_A \to p_A > 0$ and $\sigma_a^2 \to \langle a^2 \rangle - \langle a \rangle^2$, so $\text{var}(\overline{a}_A) \to 0$. By Chebyshev's inequality, $\overline{a}_A - \overline{a} \xrightarrow{p} 0$. Therefore,

$$\overline{a}_A \xrightarrow{p} \lim_{n\to\infty} \overline{a} = \langle a \rangle.$$

The proofs that $\overline{a^2}_A \xrightarrow{p} \langle a^2 \rangle$ and $\overline{ab}_A \xrightarrow{p} \langle ab \rangle$ are similar but rely on Condition 1 to show that $\text{var}(\overline{a^2}_A) \to 0$ and $\text{var}(\overline{ab}_A) \to 0$. First note that

$$\text{var}(\overline{a^2}_A) = \frac{1}{n-1} \frac{1-\tilde{p}_A}{\tilde{p}_A} \sigma^2_{(a^2)}$$

and

$$\text{var}(\overline{ab}_A) = \frac{1}{n-1} \frac{1-\tilde{p}_A}{\tilde{p}_A} \sigma^2_{(ab)}.$$

By Condition 1, $\sigma^2_{(a^2)}$ is bounded:

$$\sigma^2_{(a^2)} \leq \overline{a^4} < L.$$

Therefore, $\text{var}(\overline{a^2}_A) \to 0$. Next note that $\sigma^2_{(ab)}$ is bounded, using the Cauchy–Schwarz inequality:

$$\sigma^2_{(ab)} \leq \frac{1}{n} \sum_{i=1}^n a_i^2 b_i^2 \leq \left(\frac{1}{n} \sum_{i=1}^n a_i^4 \right)^{1/2} \left(\frac{1}{n} \sum_{i=1}^n b_i^4 \right)^{1/2} < L.$$

Therefore, $\text{var}(\overline{ab}_A) \to 0$.

The same logic can be used to show the remaining results. Those involving $\mathbf{z}_i$ can be proved element by element. □

Lemma A.2. *The pooled-slopes estimator of mean potential outcomes is*

$$\widehat{\boldsymbol{\beta}}_{\text{adj}} = \left[\overline{Y}_A - (\overline{\mathbf{z}}_A - \overline{\mathbf{z}}) \widehat{\mathbf{Q}}, \overline{Y}_B - (\overline{\mathbf{z}}_B - \overline{\mathbf{z}}) \widehat{\mathbf{Q}} \right]'.$$

Proof. The residuals from the regression defining $\widehat{\boldsymbol{\beta}}_{\text{adj}}$ are uncorrelated with T_i and $1 - T_i$. Therefore, the regression line passes through the points of means within groups A and B, and the result follows. □

Lemma A.3. *The separate-slopes estimator of mean potential outcomes is*

$$\widehat{\boldsymbol{\beta}}_{\text{interact}} = \left[\overline{Y}_A - (\overline{\mathbf{z}}_A - \overline{\mathbf{z}}) \widehat{\mathbf{Q}}_a, \overline{Y}_B - (\overline{\mathbf{z}}_B - \overline{\mathbf{z}}) \widehat{\mathbf{Q}}_b \right]'.$$

Proof. In the regression defining $\widehat{\boldsymbol{\beta}}_{\text{interact}}$, the coefficient on $\mathbf{z}_i - \overline{\mathbf{z}}$ is $\widehat{\mathbf{Q}}_b$ and the coefficient on $T_i(\mathbf{z}_i - \overline{\mathbf{z}})$ is $\widehat{\mathbf{Q}}_a - \widehat{\mathbf{Q}}_b$. (This can be shown from the equivalence of the minimization problems.) The rest of the proof is similar to that of Lemma A.2. □

Lemma A.4. *Assume Conditions 1–3. Then* $\widehat{\mathbf{Q}} \xrightarrow{p} \mathbf{Q}$.

Proof. We can assume Condition 4 without loss of generality: Let $\widehat{\gamma}$ be the estimated coefficient vector from a no-intercept OLS regression of Y_i on T_i, $1 - T_i$, and $\mathbf{z}_i - \overline{\mathbf{z}}$. Let $\tilde{a}_i = a_i - \overline{a}$ and $\tilde{b}_i = b_i - \overline{b}$, so that Condition 4 holds for $\tilde{a}_i$ and $\tilde{b}_i$. Let $\tilde{Y}_i = \tilde{a}_i T_i + \tilde{b}_i(1 - T_i)$. By a well-known property of OLS [e.g., Freedman's (2008b) Lemma A.1], the estimated coefficient vector from a no-intercept OLS regression of $\tilde{Y}_i$ on T_i, $1 - T_i$, and $\mathbf{z}_i - \overline{\mathbf{z}}$ is $\widehat{\gamma} - (\overline{a}, \overline{b}, 0)'$, so $\widehat{\mathbf{Q}}$ is unchanged. Similarly, $\mathbf{Q}$ is unchanged. Finally, centering $\mathbf{z}_i$ has no effect on the slope vectors $\widehat{\mathbf{Q}}$ and $\mathbf{Q}$.

By the Frisch–Waugh–Lovell theorem, $\widehat{\mathbf{Q}}$ can be computed from auxiliary regressions: Let

$$\begin{aligned} e_i &= Y_i - \overline{Y}_A T_i - \overline{Y}_B(1 - T_i), \\ \mathbf{f}_i &= \mathbf{z}_i - \overline{\mathbf{z}}_A T_i - \overline{\mathbf{z}}_B(1 - T_i). \end{aligned}$$

Then

$$\widehat{\mathbf{Q}} = \left(\frac{1}{n} \sum_{i=1}^{n} \mathbf{f}_i' \mathbf{f}_i \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} \mathbf{f}_i' e_i \right).$$

Some algebra yields

$$\frac{1}{n} \sum_{i=1}^{n} \mathbf{f}_i' \mathbf{f}_i = \overline{\mathbf{z}'\mathbf{z}} - \tilde{p}_A \overline{\mathbf{z}}_A' \overline{\mathbf{z}}_A - (1 - \tilde{p}_A) \overline{\mathbf{z}}_B' \overline{\mathbf{z}}_B.$$

By Condition 4 and Lemma A.1, $\overline{\mathbf{z}}_A \xrightarrow{p} \mathbf{0}$ and $\overline{\mathbf{z}}_B \xrightarrow{p} \mathbf{0}$. Therefore,

$$\frac{1}{n} \sum_{i=1}^{n} \mathbf{f}_i' \mathbf{f}_i \xrightarrow{p} \langle \mathbf{z}'\mathbf{z} \rangle.$$

Now note that

$$\begin{aligned} e_i &= (a_i - \overline{a}_A) T_i + (b_i - \overline{b}_B)(1 - T_i), \\ \mathbf{f}_i &= (\mathbf{z}_i - \overline{\mathbf{z}}_A) T_i + (\mathbf{z}_i - \overline{\mathbf{z}}_B)(1 - T_i). \end{aligned}$$

Therefore,

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^{n} \mathbf{f}_i' e_i &= \frac{1}{n} \sum_{i \in A} (\mathbf{z}_i - \overline{\mathbf{z}}_A)'(a_i - \overline{a}_A) + \frac{1}{n} \sum_{i \in B} (\mathbf{z}_i - \overline{\mathbf{z}}_B)'(b_i - \overline{b}_B) \\ &= \tilde{p}_A (\overline{a\mathbf{z}}_A - \overline{a}_A \overline{\mathbf{z}}_A)' + (1 - \tilde{p}_A)(\overline{b\mathbf{z}}_B - \overline{b}_B \overline{\mathbf{z}}_B)' \\ &\xrightarrow{p} p_A \langle a\mathbf{z} \rangle' + (1 - p_A) \langle b\mathbf{z} \rangle'. \end{aligned}$$

(Convergence to the last expression follows from Lemma A.1 and Conditions 3–4.)

It follows that

$$\begin{aligned}\widehat{\mathbf{Q}} &\xrightarrow{p} \langle \mathbf{z}'\mathbf{z}\rangle^{-1}\left[p_A\langle a\mathbf{z}\rangle' + (1-p_A)\langle b\mathbf{z}\rangle'\right] \\ &= p_A \lim_{n\to\infty}\left[\left(\sum_{i=1}^n \mathbf{z}_i'\mathbf{z}_i\right)^{-1}\sum_{i=1}^n \mathbf{z}_i' a_i\right] + (1-p_A)\lim_{n\to\infty}\left[\left(\sum_{i=1}^n \mathbf{z}_i'\mathbf{z}_i\right)^{-1}\sum_{i=1}^n \mathbf{z}_i' b_i\right] \\ &= p_A\mathbf{Q}_a + (1-p_A)\mathbf{Q}_b = \mathbf{Q}.\end{aligned}$$

□

Lemma A.5. *Assume Conditions 1–3. Then* $\widehat{\mathbf{Q}}_a \xrightarrow{p} \mathbf{Q}_a$ *and* $\widehat{\mathbf{Q}}_b \xrightarrow{p} \mathbf{Q}_b$.

Proof. The proof is similar to that of Lemma A.4 but simpler. Again, we can assume Condition 4 without loss of generality. By the Frisch–Waugh–Lovell theorem,

$$\widehat{\mathbf{Q}}_a = \left[\frac{1}{n_A}\sum_{i\in A}(\mathbf{z}_i - \overline{\mathbf{z}}_A)'(\mathbf{z}_i - \overline{\mathbf{z}}_A)\right]^{-1}\left[\frac{1}{n_A}\sum_{i\in A}(\mathbf{z}_i - \overline{\mathbf{z}}_A)'(a_i - \overline{a}_A)\right].$$

Some algebra, Lemma A.1, and Condition 4 yield

$$\frac{1}{n_A}\sum_{i\in A}(\mathbf{z}_i - \overline{\mathbf{z}}_A)'(\mathbf{z}_i - \overline{\mathbf{z}}_A) = \overline{\mathbf{z}'\mathbf{z}}_A - \overline{\mathbf{z}}_A'\overline{\mathbf{z}}_A \xrightarrow{p} \langle \mathbf{z}'\mathbf{z}\rangle$$

and

$$\frac{1}{n_A}\sum_{i\in A}(\mathbf{z}_i - \overline{\mathbf{z}}_A)'(a_i - \overline{a}_A) = (\overline{a\mathbf{z}}_A - \overline{a}_A\overline{\mathbf{z}}_A)' \xrightarrow{p} \langle a\mathbf{z}\rangle'$$

so

$$\begin{aligned}\widehat{\mathbf{Q}}_a &\xrightarrow{p} \langle \mathbf{z}'\mathbf{z}\rangle^{-1}\langle a\mathbf{z}\rangle' \\ &= \lim_{n\to\infty}\left[\left(\sum_{i=1}^n \mathbf{z}_i'\mathbf{z}_i\right)^{-1}\sum_{i=1}^n \mathbf{z}_i' a_i\right] = \mathbf{Q}_a.\end{aligned}$$

The proof that $\widehat{\mathbf{Q}}_b \xrightarrow{p} \mathbf{Q}_b$ is similar. □

Lemma A.6 is similar to part of Freedman's (2008b) Theorem 2.

Lemma A.6. *Assume Conditions 1–3. Then*

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_{\text{adj}} - \boldsymbol{\beta}) \xrightarrow{d} N(0, \mathbf{V})$$

where

$$\mathbf{V} = \begin{bmatrix} \frac{1-p_A}{p_A}\lim_{n\to\infty}\sigma^2_{a^{**}} & -\lim_{n\to\infty}\sigma_{a^{**},b^{**}} \\ -\lim_{n\to\infty}\sigma_{a^{**},b^{**}} & \frac{p_A}{1-p_A}\lim_{n\to\infty}\sigma^2_{b^{**}} \end{bmatrix}.$$

Proof. We can assume Condition 4 without loss of generality: Centering a_i, b_i, and $\mathbf{z}_i$ has no effect on $\widehat{\mathbf{Q}}$ and $\mathbf{Q}$, as shown in the proof of Lemma A.4, so it subtracts $(\overline{a},\overline{b})'$ from both $\widehat{\boldsymbol{\beta}}_{\text{adj}}$ (see Lemma A.2) and $\boldsymbol{\beta}$, and it has no effect on the elements of $\mathbf{V}$.

Condition 4 and Lemma A.2 imply that

$$\begin{aligned}\sqrt{n}(\widehat{\boldsymbol{\beta}}_{\text{adj}}-\boldsymbol{\beta}) &= \sqrt{n}(\overline{Y}_A-\overline{\mathbf{z}}_A\widehat{\mathbf{Q}},\overline{Y}_B-\overline{\mathbf{z}}_B\widehat{\mathbf{Q}})' \\ &= \sqrt{n}(\overline{a}_A-\overline{\mathbf{z}}_A\mathbf{Q},\overline{b}_B-\overline{\mathbf{z}}_B\mathbf{Q})'-[\sqrt{n}\overline{\mathbf{z}}_A(\widehat{\mathbf{Q}}-\mathbf{Q}),\sqrt{n}\overline{\mathbf{z}}_B(\widehat{\mathbf{Q}}-\mathbf{Q})]'.\end{aligned}$$

By a finite-population Central Limit Theorem [Freedman's (2008b) Theorem 1], $\sqrt{n}\overline{\mathbf{z}}_A$ and $\sqrt{n}\overline{\mathbf{z}}_B$ are $O_p(1)$, and by Lemma A.4, $\widehat{\mathbf{Q}}-\mathbf{Q}$ is $o_p(1)$. Therefore,

$$[\sqrt{n}\overline{\mathbf{z}}_A(\widehat{\mathbf{Q}}-\mathbf{Q}),\sqrt{n}\overline{\mathbf{z}}_B(\widehat{\mathbf{Q}}-\mathbf{Q})]' \xrightarrow{p} \mathbf{0}.$$

The conclusion follows from Freedman's (2008b) Theorem 1 with a and b replaced by $a-\mathbf{zQ}$ and $b-\mathbf{zQ}$. □

Lemma A.7 is an application of the Weak Law of Large Numbers (Lemma A.1).

Lemma A.7. *Assume Conditions 1–3. Let $\boldsymbol{\theta}$ be any $K\times 1$ vector that is constant as $n\to\infty$. Then*

$$\begin{aligned}\frac{1}{n_A}\sum_{i\in A}(a_i+\mathbf{z}_i\boldsymbol{\theta})^2 &\xrightarrow{p} \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}(a_i+\mathbf{z}_i\boldsymbol{\theta})^2, \\ \frac{1}{n-n_A}\sum_{i\in B}(b_i+\mathbf{z}_i\boldsymbol{\theta})^2 &\xrightarrow{p} \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}(b_i+\mathbf{z}_i\boldsymbol{\theta})^2.\end{aligned}$$

Proof. Using Lemma A.1,

$$\begin{aligned}\frac{1}{n_A}\sum_{i\in A}(a_i+\mathbf{z}_i\boldsymbol{\theta})^2 &= \overline{a^2}_A+2\overline{a\mathbf{z}}_A\boldsymbol{\theta}+\boldsymbol{\theta}'\overline{\mathbf{z}'\mathbf{z}}_A\boldsymbol{\theta} \\ &\xrightarrow{p} \langle a^2\rangle+2\langle a\mathbf{z}\rangle\boldsymbol{\theta}+\boldsymbol{\theta}'\langle\mathbf{z}'\mathbf{z}\rangle\boldsymbol{\theta} \\ &= \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}(a_i+\mathbf{z}_i\boldsymbol{\theta})^2.\end{aligned}$$

The proof of the other assertion is analogous. □

Lemma A.8 shows that the sandwich variance estimator for $\widehat{\text{ATE}}_{\text{adj}}$ is invariant to the transformation of the regressors that was used to define $\widehat{\boldsymbol{\beta}}_{\text{adj}}$.

Lemma A.8. *Let*

$$\mathbf{W} = (\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\left(\sum_{i=1}^{n}\hat{e}_i^2\tilde{\mathbf{x}}_i'\tilde{\mathbf{x}}_i\right)(\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}$$

where $\tilde{\mathbf{X}}$ is the $n\times(K+2)$ matrix with row i equal to $\tilde{\mathbf{x}}_i=(T_i,1-T_i,\mathbf{z}_i-\overline{\mathbf{z}})$ and $\hat{e}_i$ is the residual from the no-intercept OLS regression of Y_i on $\tilde{\mathbf{x}}_i$. Then $\widehat{v}_{\text{adj}}=W_{11}+W_{22}-2W_{12}$, where W_{ij} is the (i,j) element of $\mathbf{W}$.

Proof. By definition, $\widehat{v}_{\text{adj}}$ is the $(2,2)$ element of

$$(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\text{diag}(\hat{\varepsilon}_1^2,\ldots,\hat{\varepsilon}_n^2)\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \quad = \quad (\mathbf{X}'\mathbf{X})^{-1}\left(\sum_{i=1}^{n}\hat{\varepsilon}_i^2\mathbf{x}_i'\mathbf{x}_i\right)(\mathbf{X}'\mathbf{X})^{-1}$$

where $\mathbf{X}$ is the $n\times(K+2)$ matrix whose ith row is $\mathbf{x}_i=(1,T_i,\mathbf{z}_i)$ and $\hat{\varepsilon}_i$ is the residual from the OLS regression of Y_i on $\mathbf{x}_i$.

The OLS residuals are invariant to the linear transformation of regressors, so $\hat{e}_i=\hat{\varepsilon}_i$ for $i=1,2,\ldots,n$. Also, $\mathbf{X}=\tilde{\mathbf{X}}\mathbf{R}\mathbf{S}$ where

$$\mathbf{R}=\begin{bmatrix}\mathbf{M} & \mathbf{0}\\ \mathbf{0} & \mathbf{I}_K\end{bmatrix}, \qquad \mathbf{S}=\begin{bmatrix}\mathbf{I}_2 & \mathbf{L}\\ \mathbf{0} & \mathbf{I}_K\end{bmatrix},$$

and

$$\mathbf{M}=\begin{bmatrix}1 & 1\\ 1 & 0\end{bmatrix}=\begin{bmatrix}0 & 1\\ 1 & -1\end{bmatrix}^{-1}, \qquad \mathbf{L}=\begin{bmatrix}\overline{\mathbf{z}}\\ \mathbf{0}\end{bmatrix}.$$

Note that $\mathbf{R}$ is symmetric but $\mathbf{S}$ is not, and

$$\mathbf{S}^{-1} \quad = \quad \begin{bmatrix}\mathbf{I}_2 & -\mathbf{L}\\ \mathbf{0} & \mathbf{I}_K\end{bmatrix}.$$

Therefore,

$$(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\text{diag}(\hat{\varepsilon}_1^2,\ldots,\hat{\varepsilon}_n^2)\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \quad = \quad \mathbf{S}^{-1}\mathbf{R}^{-1}\mathbf{W}\mathbf{R}^{-1}(\mathbf{S}^{-1})'.$$

The (2, 2) element is $W_{11}+W_{22}-2W_{12}$. □

Lemma A.9 is important for the proof of Theorem 2.2.

Lemma A.9. *Assume Conditions 1–4. Let $\hat{e}_i$ denote the residual from the no-intercept OLS regression of Y_i on T_i, $1-T_i$, and $\mathbf{z}_i$. Then*

$$\frac{1}{n_A}\sum_{i\in A}\hat{e}_i^2 \xrightarrow{p} \lim_{n\to\infty}\sigma_{a^{**}}^2, \qquad \frac{1}{n-n_A}\sum_{i\in B}\hat{e}_i^2 \xrightarrow{p} \lim_{n\to\infty}\sigma_{b^{**}}^2,$$

and $n^{-1}\sum_{i\in A}\hat{e}_i^2\mathbf{z}_i$, $n^{-1}\sum_{i\in B}\hat{e}_i^2\mathbf{z}_i$, and $n^{-1}\sum_{i=1}^{n}\hat{e}_i^2\mathbf{z}_i'\mathbf{z}_i$ are all $O_p(1)$.

Proof. Let $\widehat{\beta}_{adj(1)}$ and $\widehat{\beta}_{adj(2)}$ denote the estimated coefficients on T_i and $1-T_i$, respectively. Then

$$\begin{aligned}\hat{e}_i &= Y_i-\widehat{\beta}_{adj(1)}T_i-\widehat{\beta}_{adj(2)}(1-T_i)-\mathbf{z}_i\widehat{\mathbf{Q}}\\ &= T_i[(a_i-\mathbf{z}_i\widehat{\mathbf{Q}})-\widehat{\beta}_{adj(1)}]+(1-T_i)[(b_i-\mathbf{z}_i\widehat{\mathbf{Q}})-\widehat{\beta}_{adj(2)}]\\ &= T_i[a_i^{**}-\mathbf{z}_i(\widehat{\mathbf{Q}}-\mathbf{Q})-\widehat{\beta}_{adj(1)}]+(1-T_i)[b_i^{**}-\mathbf{z}_i(\widehat{\mathbf{Q}}-\mathbf{Q})-\widehat{\beta}_{adj(2)}].\end{aligned}$$

Therefore,

$$\begin{aligned}\frac{1}{n_A}\sum_{i\in A}\hat{e}_i^2 &= \frac{1}{n_A}\sum_{i\in A}[a_i^{**}-\mathbf{z}_i(\widehat{\mathbf{Q}}-\mathbf{Q})-\widehat{\beta}_{adj(1)}]^2\\ &= S_1+S_2+S_3-2S_4-2S_5-2S_6\end{aligned}$$

where

$$\begin{aligned}S_1 &= \frac{1}{n_A}\sum_{i\in A}(a_i^{**})^2,\\ S_2 &= (\widehat{\mathbf{Q}}-\mathbf{Q})'\overline{\mathbf{z}'\mathbf{z}}_A(\widehat{\mathbf{Q}}-\mathbf{Q}),\\ S_3 &= \widehat{\beta}^2_{adj(1)},\\ S_4 &= \left(\frac{1}{n_A}\sum_{i\in A}a_i^{**}\mathbf{z}_i\right)(\widehat{\mathbf{Q}}-\mathbf{Q}),\\ S_5 &= \widehat{\beta}_{adj(1)}\overline{a^{**}}_A,\\ S_6 &= \widehat{\beta}_{adj(1)}\overline{\mathbf{z}}_A(\widehat{\mathbf{Q}}-\mathbf{Q}).\end{aligned}$$

$S_1 \xrightarrow{p} \lim_{n\to\infty}\sigma^2_{a^{**}}$ by Lemma A.7 and Condition 4.
The other terms are all $o_p(1)$:

- $S_2 \xrightarrow{p} 0$ because $\widehat{\mathbf{Q}} \xrightarrow{p} \mathbf{Q}$ (by Lemma A.4) and $\overline{\mathbf{z}'\mathbf{z}}_A \xrightarrow{p} \langle\mathbf{z}'\mathbf{z}\rangle$ (by Lemma A.1).
- $S_3 \xrightarrow{p} 0$ because $\widehat{\beta}_{adj(1)} \xrightarrow{p} \overline{a} = 0$ (by Condition 4 and Lemma A.6).
- $S_4 \xrightarrow{p} 0$ because

$$\begin{aligned}\frac{1}{n_A}\sum_{i\in A}a_i^{**}\mathbf{z}_i &= \frac{1}{n_A}\sum_{i\in A}(a_i-\mathbf{Q}'\mathbf{z}_i')\mathbf{z}_i\\ &\xrightarrow{p} \langle a\mathbf{z}\rangle-\mathbf{Q}'\langle\mathbf{z}'\mathbf{z}\rangle\end{aligned}$$

  (by Lemma A.1) and $\widehat{\mathbf{Q}} \xrightarrow{p} \mathbf{Q}$.
- $S_5 \xrightarrow{p} 0$ because $\overline{a^{**}}_A \xrightarrow{p} \langle a\rangle - \langle\mathbf{z}\rangle\mathbf{Q} = 0$ (by Lemma A.1 and Condition 4) and $\widehat{\beta}_{adj(1)} \xrightarrow{p} 0$.
- $S_6 \xrightarrow{p} 0$ because $\overline{\mathbf{z}}_A \xrightarrow{p} \mathbf{0}$ (by Lemma A.1 and Condition 4), $\widehat{\beta}_{adj(1)} \xrightarrow{p} 0$, and $\widehat{\mathbf{Q}} \xrightarrow{p} \mathbf{Q}$.

Therefore,

$$\frac{1}{n_A}\sum_{i\in A}\hat{e}_i^2 \xrightarrow{p} \lim_{n\to\infty}\sigma^2_{a^{**}}.$$

Similarly,

$$\frac{1}{n-n_A}\sum_{i\in B}\hat{e}_i^2 \quad \xrightarrow{p} \quad \lim_{n\to\infty}\sigma_{b^{**}}^2.$$

Now note that

$$\begin{aligned} n^{-1}\sum_{i\in A}\hat{e}_i^2\mathbf{z}_i &= \frac{1}{n}\sum_{i\in A}[a_i-\mathbf{z}_i\widehat{\mathbf{Q}}-\widehat{\beta}_{adj(1)}]^2\mathbf{z}_i \\ &= \mathbf{R}_1+\mathbf{R}_2+\mathbf{R}_3-2\mathbf{R}_4-2\mathbf{R}_5-2\mathbf{R}_6 \end{aligned}$$

where

$$\begin{aligned} \mathbf{R}_1 &= \frac{1}{n}\sum_{i\in A}a_i^2\mathbf{z}_i, \\ \mathbf{R}_2 &= \frac{1}{n}\sum_{i\in A}(\mathbf{z}_i\widehat{\mathbf{Q}})^2\mathbf{z}_i, \\ \mathbf{R}_3 &= \tilde{p}_A\widehat{\beta}_{adj(1)}^2\overline{\mathbf{z}}_A, \\ \mathbf{R}_4 &= \widehat{\mathbf{Q}}'\frac{1}{n}\sum_{i\in A}a_i\mathbf{z}_i'\mathbf{z}_i, \\ \mathbf{R}_5 &= \tilde{p}_A\widehat{\beta}_{adj(1)}\overline{a\mathbf{z}}_A, \\ \mathbf{R}_6 &= \tilde{p}_A\widehat{\beta}_{adj(1)}\widehat{\mathbf{Q}}'\overline{\mathbf{z}'\mathbf{z}}_A. \end{aligned}$$

$\mathbf{R}_3$, $\mathbf{R}_5$, and $\mathbf{R}_6$ are $o_p(1)$ because $\widehat{\beta}_{adj(1)} \xrightarrow{p} 0$, $\overline{\mathbf{z}}_A \xrightarrow{p} \mathbf{0}$, and $\tilde{p}_A$, $\overline{a\mathbf{z}}_A$, $\overline{\mathbf{z}'\mathbf{z}}_A$, and $\widehat{\mathbf{Q}}$ converge to finite limits (by Condition 3, Lemma A.1, and Lemma A.4).

$\mathbf{R}_1$, $\mathbf{R}_2$, and $\mathbf{R}_4$ are $O_p(1)$, by Condition 1, Lemma A.4, and repeated application of the Cauchy–Schwarz inequality. For example, for $k=1,\ldots,K$, the kth element of $\mathbf{R}_2$ is

$$\frac{1}{n}\sum_{i\in A}\left(\sum_{j=1}^K z_{ij}\widehat{Q}_j\right)^2 z_{ik} \quad = \quad \sum_{j=1}^K\sum_{\ell=1}^K\left(\widehat{Q}_j\widehat{Q}_\ell\frac{1}{n}\sum_{i\in A}z_{ij}z_{i\ell}z_{ik}\right).$$

$\widehat{Q}_j$ and $\widehat{Q}_\ell$ are $O_p(1)$, and $n^{-1}\sum_{i\in A}z_{ij}z_{i\ell}z_{ik}$ is $O(1)$:

$$\begin{aligned} \left|\frac{1}{n}\sum_{i\in A}z_{ij}z_{i\ell}z_{ik}\right| &\leq \frac{1}{n}\sum_{i=1}^n|z_{ij}||z_{i\ell}z_{ik}| \leq \left(\frac{1}{n}\sum_{i=1}^n z_{ij}^2\right)^{1/2}\left(\frac{1}{n}\sum_{i=1}^n z_{i\ell}^2z_{ik}^2\right)^{1/2} \\ &\leq \left(\frac{1}{n}\sum_{i=1}^n z_{ij}^4\right)^{1/4}\left(\frac{1}{n}\sum_{i=1}^n 1\right)^{1/4}\left(\frac{1}{n}\sum_{i=1}^n z_{i\ell}^4\right)^{1/4}\left(\frac{1}{n}\sum_{i=1}^n z_{ik}^4\right)^{1/4} \\ &< L^{3/4}. \end{aligned}$$

Therefore, $\mathbf{R}_2$ is $O_p(1)$.

Thus, $n^{-1}\sum_{i\in A}\hat{e}_i^2\mathbf{z}_i$ is $O_p(1)$. The proofs for $n^{-1}\sum_{i\in B}\hat{e}_i^2\mathbf{z}_i$ and $n^{-1}\sum_{i=1}^n\hat{e}_i^2\mathbf{z}_i'\mathbf{z}_i$ are similar. □

A.3 Proof of Theorem 2.1

We can assume Condition 4 without loss of generality, by an argument similar to that given in the proof of Lemma A.6. Then ATE = 0, and by Lemma A.3 and Condition 4,

$$\begin{aligned}\sqrt{n}(\widehat{\text{ATE}}_{\text{interact}} - \text{ATE}) &= \sqrt{n}[(\overline{a}_A - \overline{\mathbf{z}}_A\widehat{\mathbf{Q}}_a) - (\overline{b}_B - \overline{\mathbf{z}}_B\widehat{\mathbf{Q}}_b)] \\ &= \sqrt{n}[(\overline{a}_A - \overline{\mathbf{z}}_A\mathbf{Q}_a) - (\overline{b}_B - \overline{\mathbf{z}}_B\mathbf{Q}_b)] - \\ &\quad \sqrt{n}\overline{\mathbf{z}}_A(\widehat{\mathbf{Q}}_a - \mathbf{Q}_a) + \sqrt{n}\overline{\mathbf{z}}_B(\widehat{\mathbf{Q}}_b - \mathbf{Q}_b).\end{aligned}$$

By a finite-population Central Limit Theorem [Freedman's (2008b) Theorem 1], $\sqrt{n}\overline{\mathbf{z}}_A$ and $\sqrt{n}\overline{\mathbf{z}}_B$ are $O_p(1)$, and by Lemma A.5, $\widehat{\mathbf{Q}}_a - \mathbf{Q}_a$ and $\widehat{\mathbf{Q}}_b - \mathbf{Q}_b$ are $o_p(1)$. Therefore, $\sqrt{n}\overline{\mathbf{z}}_A(\widehat{\mathbf{Q}}_a - \mathbf{Q}_a)$ and $\sqrt{n}\overline{\mathbf{z}}_B(\widehat{\mathbf{Q}}_b - \mathbf{Q}_b)$ are $o_p(1)$.

The conclusion follows from Freedman's (2008b) Theorem 1 with a and b replaced by $a - \mathbf{z}\mathbf{Q}_a$ and $b - \mathbf{z}\mathbf{Q}_b$.

A.4 Proof of Corollary 2.1.1

We can assume Condition 4 without loss of generality: Centering a_i, b_i, and $\mathbf{z}_i$ has no effect on $\widehat{\text{ATE}}_{\text{interact}} - \text{ATE}$, $\widehat{\text{ATE}}_{\text{unadj}} - \text{ATE}$, $\mathbf{Q}_a$, $\mathbf{Q}_b$, or σ^2_E.

Note that:

$$\begin{aligned}\lim_{n\to\infty}\sigma^2_{a^*} &= \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}(a_i - \mathbf{z}_i\mathbf{Q}_a)^2 \\ &= \langle a^2\rangle - \langle a\mathbf{z}\rangle\langle \mathbf{z}'\mathbf{z}\rangle^{-1}\langle a\mathbf{z}\rangle', \\ \lim_{n\to\infty}\sigma^2_{b^*} &= \langle b^2\rangle - \langle b\mathbf{z}\rangle\langle \mathbf{z}'\mathbf{z}\rangle^{-1}\langle b\mathbf{z}\rangle', \\ \lim_{n\to\infty}\sigma_{a^*,b^*} &= \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}(a_i - \mathbf{z}_i\mathbf{Q}_a)(b_i - \mathbf{z}_i\mathbf{Q}_b) \\ &= \langle ab\rangle - \langle a\mathbf{z}\rangle\mathbf{Q}_b - \langle b\mathbf{z}\rangle\mathbf{Q}_a + \mathbf{Q}'_a\langle \mathbf{z}'\mathbf{z}\rangle\mathbf{Q}_b \\ &= \langle ab\rangle - \langle a\mathbf{z}\rangle\langle \mathbf{z}'\mathbf{z}\rangle^{-1}\langle b\mathbf{z}\rangle'.\end{aligned}$$

By Freedman's (2008b) Theorem 1,

$$\begin{aligned}\text{avar}(\sqrt{n}[\widehat{\text{ATE}}_{\text{unadj}} - \text{ATE}]) &= \text{avar}(\sqrt{n}[\overline{a}_A - \overline{b}_B]) \\ &= \frac{1-p_A}{p_A}\langle a^2\rangle + \frac{p_A}{1-p_A}\langle b^2\rangle + 2\langle ab\rangle.\end{aligned}$$

Let

$$\Delta = \text{avar}(\sqrt{n}[\widehat{\text{ATE}}_{\text{unadj}} - \text{ATE}]) - \text{avar}(\sqrt{n}[\widehat{\text{ATE}}_{\text{interact}} - \text{ATE}]).$$

Then

$$\begin{aligned}\Delta &= \frac{1-p_A}{p_A}\langle a\mathbf{z}\rangle\langle \mathbf{z}'\mathbf{z}\rangle^{-1}\langle a\mathbf{z}\rangle' + \frac{p_A}{1-p_A}\langle b\mathbf{z}\rangle\langle \mathbf{z}'\mathbf{z}\rangle^{-1}\langle b\mathbf{z}\rangle' + 2\langle a\mathbf{z}\rangle\langle \mathbf{z}'\mathbf{z}\rangle^{-1}\langle b\mathbf{z}\rangle' \\ &= \frac{1}{p_A(1-p_A)}\mathbf{Q}_E'\langle \mathbf{z}'\mathbf{z}\rangle\mathbf{Q}_E = \frac{1}{p_A(1-p_A)}\lim_{n\to\infty}\sigma_E^2 \geq 0.\end{aligned}$$

The matrix $\langle \mathbf{z}'\mathbf{z}\rangle$ is positive definite, so $\Delta/n = 0$ if and only if $\mathbf{Q}_E = \mathbf{0}$.

A.5 Proof of remark (iv) after Corollary 2.1.1

Suppose there are three treatment groups, A, B, and C, with associated dummy variables U_i, V_i, and W_i and potential outcomes a_i, b_i, and c_i. Let $\text{ATE} = \overline{a} - \overline{b}$, and let $\widehat{\text{ATE}}_{\text{interact}}$ be the difference between the estimated coefficients on U_i and V_i in the no-intercept OLS regression of Y_i on U_i, V_i, W_i, $\mathbf{z}_i - \overline{\mathbf{z}}$, $U_i(\mathbf{z}_i - \overline{\mathbf{z}})$, and $W_i(\mathbf{z}_i - \overline{\mathbf{z}})$.

Assume the three groups are of fixed sizes n_A, n_B, and $n - n_A - n_B$. Assume regularity conditions analogous to Conditions 1–3: for example, $n_A/n \to p_A$ and $n_B/n \to p_B$, where $p_A > 0$, $p_B > 0$, and $p_A + p_B < 1$. Without loss of generality, assume Condition 4.

Then $\sqrt{n}(\widehat{\text{ATE}}_{\text{interact}} - \text{ATE})$ converges in distribution to a Gaussian random variable with mean 0 and variance

$$\frac{1-p_A}{p_A}\lim_{n\to\infty}\sigma_{a^*}^2 + \frac{1-p_B}{p_B}\lim_{n\to\infty}\sigma_{b^*}^2 + 2\lim_{n\to\infty}\sigma_{a^*,b^*}.$$

The proof is essentially the same as that of Theorem 2.1.

Let $\widehat{\text{ATE}}_{\text{unadj}} = \overline{Y}_A - \overline{Y}_B$. By Freedman's (2008b) Theorem 1, the asymptotic variance of $\sqrt{n}(\widehat{\text{ATE}}_{\text{unadj}} - \text{ATE})$ is

$$\frac{1-p_A}{p_A}\langle a^2\rangle + \frac{1-p_B}{p_B}\langle b^2\rangle + 2\langle ab\rangle.$$

Let

$$\Delta = \text{avar}(\sqrt{n}[\widehat{\text{ATE}}_{\text{unadj}} - \text{ATE}]) - \text{avar}(\sqrt{n}[\widehat{\text{ATE}}_{\text{interact}} - \text{ATE}]).$$

Then

$$\begin{aligned}\Delta &= \frac{1-p_A}{p_A}\langle a\mathbf{z}\rangle\langle \mathbf{z}'\mathbf{z}\rangle^{-1}\langle a\mathbf{z}\rangle' + \frac{1-p_B}{p_B}\langle b\mathbf{z}\rangle\langle \mathbf{z}'\mathbf{z}\rangle^{-1}\langle b\mathbf{z}\rangle' + 2\langle a\mathbf{z}\rangle\langle \mathbf{z}'\mathbf{z}\rangle^{-1}\langle b\mathbf{z}\rangle' \\ &= \frac{1-p_A}{p_A}\langle a\mathbf{z}\rangle\langle \mathbf{z}'\mathbf{z}\rangle^{-1}\langle a\mathbf{z}\rangle' + \frac{p_A}{1-p_A}\langle b\mathbf{z}\rangle\langle \mathbf{z}'\mathbf{z}\rangle^{-1}\langle b\mathbf{z}\rangle' + 2\langle a\mathbf{z}\rangle\langle \mathbf{z}'\mathbf{z}\rangle^{-1}\langle b\mathbf{z}\rangle' + \\ &\qquad \left(\frac{1-p_B}{p_B} - \frac{p_A}{1-p_A}\right)\langle b\mathbf{z}\rangle\langle \mathbf{z}'\mathbf{z}\rangle^{-1}\langle b\mathbf{z}\rangle' \\ &= \frac{1}{p_A(1-p_A)}\lim_{n\to\infty}\sigma_E^2 + \left(\frac{1-p_B}{p_B} - \frac{p_A}{1-p_A}\right)\mathbf{Q}_b'\langle \mathbf{z}'\mathbf{z}\rangle\mathbf{Q}_b,\end{aligned}$$

where $E_i = (\mathbf{z}_i - \bar{\mathbf{z}})\mathbf{Q}_E$ and $\mathbf{Q}_E = (1-p_A)\mathbf{Q}_a + p_A\mathbf{Q}_b$.

Similarly,

$$\Delta = \frac{1}{p_B(1-p_B)} \lim_{n\to\infty} \sigma_F^2 + \left(\frac{1-p_A}{p_A} - \frac{p_B}{1-p_B} \right) \mathbf{Q}_a' \langle \mathbf{z}'\mathbf{z} \rangle \mathbf{Q}_a,$$

where $F_i = (\mathbf{z}_i - \bar{\mathbf{z}})\mathbf{Q}_F$ and $\mathbf{Q}_F = p_B\mathbf{Q}_a + (1-p_B)\mathbf{Q}_b$.

The condition $p_A + p_B < 1$ implies

$$\frac{1-p_B}{p_B} - \frac{p_A}{1-p_A} > 0, \qquad \frac{1-p_A}{p_A} - \frac{p_B}{1-p_B} > 0.$$

Also, $\langle \mathbf{z}'\mathbf{z} \rangle$ is positive definite. Therefore, $\Delta \geq 0$, and the inequality is strict unless $\mathbf{Q}_a = \mathbf{0}$ and $\mathbf{Q}_b = \mathbf{0}$.

The proof extends to designs with more than three treatment groups.

A.6 Proof of Corollary 2.1.2

Again, we can assume Condition 4 without loss of generality. By Lemma A.6,

$$\begin{aligned}
\text{avar}(\sqrt{n}[\widehat{\text{ATE}}_{\text{adj}} - \text{ATE}]) &= \frac{1-p_A}{p_A} \lim_{n\to\infty} \sigma_{a^{**}}^2 + \frac{p_A}{1-p_A} \lim_{n\to\infty} \sigma_{b^{**}}^2 + 2 \lim_{n\to\infty} \sigma_{a^{**},b^{**}} \\
&= \frac{1-p_A}{p_A}[\langle a^2 \rangle + \mathbf{Q}'\langle \mathbf{z}'\mathbf{z} \rangle \mathbf{Q} - 2\mathbf{Q}'\langle a\mathbf{z} \rangle'] + \\
&\quad \frac{p_A}{1-p_A}[\langle b^2 \rangle + \mathbf{Q}'\langle \mathbf{z}'\mathbf{z} \rangle \mathbf{Q} - 2\mathbf{Q}'\langle b\mathbf{z} \rangle'] + \\
&\quad 2[\langle ab \rangle + \mathbf{Q}'\langle \mathbf{z}'\mathbf{z} \rangle \mathbf{Q} - \mathbf{Q}'\langle a\mathbf{z} \rangle' - \mathbf{Q}'\langle b\mathbf{z} \rangle'] \\
&= \frac{1-p_A}{p_A}\langle a^2 \rangle + \frac{p_A}{1-p_A}\langle b^2 \rangle + 2\langle ab \rangle + \\
&\quad \frac{1}{p_A(1-p_A)} \mathbf{Q}'\langle \mathbf{z}'\mathbf{z} \rangle \mathbf{Q} - \frac{2}{p_A}\mathbf{Q}'\langle a\mathbf{z} \rangle' - \frac{2}{1-p_A}\mathbf{Q}'\langle b\mathbf{z} \rangle'.
\end{aligned}$$

Let

$$\Delta = \text{avar}(\sqrt{n}[\widehat{\text{ATE}}_{\text{adj}} - \text{ATE}]) - \text{avar}(\sqrt{n}[\widehat{\text{ATE}}_{\text{interact}} - \text{ATE}]).$$

Then

$$\begin{aligned}
\Delta &= \frac{1}{p_A(1-p_A)}\mathbf{Q}'\langle \mathbf{z}'\mathbf{z}\rangle\mathbf{Q} - \frac{2}{p_A}\mathbf{Q}'\langle a\mathbf{z}\rangle' - \frac{2}{1-p_A}\mathbf{Q}'\langle b\mathbf{z}\rangle' + \\
&\qquad \frac{1-p_A}{p_A}\langle a\mathbf{z}\rangle\langle \mathbf{z}'\mathbf{z}\rangle^{-1}\langle a\mathbf{z}\rangle' + \frac{p_A}{1-p_A}\langle b\mathbf{z}\rangle\langle \mathbf{z}'\mathbf{z}\rangle^{-1}\langle b\mathbf{z}\rangle' + 2\langle a\mathbf{z}\rangle\langle \mathbf{z}'\mathbf{z}\rangle^{-1}\langle b\mathbf{z}\rangle' \\
&= \left(\frac{p_A}{1-p_A} - 2 + \frac{1-p_A}{p_A}\right)\left(\langle a\mathbf{z}\rangle\langle \mathbf{z}'\mathbf{z}\rangle^{-1}\langle a\mathbf{z}\rangle' + \langle b\mathbf{z}\rangle\langle \mathbf{z}'\mathbf{z}\rangle^{-1}\langle b\mathbf{z}\rangle' - 2\langle a\mathbf{z}\rangle\langle \mathbf{z}'\mathbf{z}\rangle^{-1}\langle b\mathbf{z}\rangle'\right) \\
&= \frac{(2p_A-1)^2}{p_A(1-p_A)}(\mathbf{Q}_a-\mathbf{Q}_b)'\langle \mathbf{z}'\mathbf{z}\rangle(\mathbf{Q}_a-\mathbf{Q}_b) \\
&= \frac{(2p_A-1)^2}{p_A(1-p_A)}\lim_{n\to\infty}\sigma_D^2 \geq 0.
\end{aligned}$$

A.7 Outline of proof of remark (iii) after Corollary 2.1.2

Without loss of generality, assume Condition 4. From the proof of Theorem 2.1,

$$\sqrt{n}\widehat{\text{ATE}}_{\text{interact}} = \sqrt{n}[(\overline{a}_A - \overline{\mathbf{z}}_A\mathbf{Q}_a) - (\overline{b}_B - \overline{\mathbf{z}}_B\mathbf{Q}_b)] + o_p(1).$$

By Condition 4, $\tilde{p}_A\overline{\mathbf{z}}_A + (1-\tilde{p}_A)\overline{\mathbf{z}}_B = \mathbf{0}$. Therefore, $\overline{\mathbf{z}}_A = (1-\tilde{p}_A)(\overline{\mathbf{z}}_A - \overline{\mathbf{z}}_B)$ and $\overline{\mathbf{z}}_B = -\tilde{p}_A(\overline{\mathbf{z}}_A - \overline{\mathbf{z}}_B)$. It follows that

$$\sqrt{n}\widehat{\text{ATE}}_{\text{interact}} = \sqrt{n}\{\overline{a}_A - \overline{b}_B - (\overline{\mathbf{z}}_A - \overline{\mathbf{z}}_B)[(1-p_A)\mathbf{Q}_a + p_A\mathbf{Q}_b]\} + o_p(1).$$

Now let $\widehat{\text{ATE}}_{tyranny}$ and $\widehat{\mathbf{Q}}_{tyranny}$ be the estimated coefficients on T_i and $\mathbf{z}_i$ from a weighted least squares regression of Y_i on T_i and $\mathbf{z}_i$, with weights

$$w_i = \frac{1-\tilde{p}_A}{\tilde{p}_A}T_i + \frac{\tilde{p}_A}{1-\tilde{p}_A}(1-T_i).$$

It can be shown that $\widehat{\mathbf{Q}}_{tyranny} \xrightarrow{p} (1-p_A)\mathbf{Q}_a + p_A\mathbf{Q}_b$. The proof is similar to that of Lemma A.4, after noting that weighted least squares is equivalent to OLS with all data values (including the constant) multiplied by $\sqrt{w_i}$.

It follows that

$$\sqrt{n}\widehat{\text{ATE}}_{tyranny} = \sqrt{n}\{\overline{a}_A - \overline{b}_B - (\overline{\mathbf{z}}_A - \overline{\mathbf{z}}_B)[(1-p_A)\mathbf{Q}_a + p_A\mathbf{Q}_b]\} + o_p(1).$$

The proof is similar to arguments in the proofs of Lemmas A.2 and A.6.

Therefore, $\sqrt{n}(\widehat{\text{ATE}}_{tyranny} - \widehat{\text{ATE}}_{\text{interact}}) \xrightarrow{p} 0$.

A.8 Proof of Theorem 2.2

We can assume Condition 4 without loss of generality, by arguments similar to those given in the proofs of Lemmas A.4, A.6, and A.8.

By Lemma A.8, $n\widehat{v}_{\text{adj}} = M_{11} + M_{22} - 2M_{12}$, where

$$\mathbf{M} \;=\; (n^{-1}\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\left(n^{-1}\sum_{i=1}^{n}\hat{e}_i^2\tilde{\mathbf{x}}_i'\tilde{\mathbf{x}}_i\right)(n^{-1}\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}.$$

Using Condition 4,

$$n^{-1}\tilde{\mathbf{X}}'\tilde{\mathbf{X}} \;=\; \left[\begin{array}{cc}\mathbf{C} & \mathbf{D}\\ \mathbf{D}' & \mathbf{z}'\mathbf{z}\end{array}\right],$$

where

$$\mathbf{C} = \left[\begin{array}{cc}\tilde{p}_A & 0\\ 0 & 1-\tilde{p}_A\end{array}\right], \qquad \mathbf{D} = \left[\begin{array}{c}\tilde{p}_A\overline{\mathbf{z}}_A\\ (1-\tilde{p}_A)\overline{\mathbf{z}}_B\end{array}\right].$$

By Conditions 2–4 and Lemma A.1, $\tilde{p}_A \to p_A$, $\overline{\mathbf{z}}_A \xrightarrow{p} \mathbf{0}$, $\overline{\mathbf{z}}_B \xrightarrow{p} \mathbf{0}$, and $\langle \mathbf{z}'\mathbf{z}\rangle$ is invertible. Therefore,

$$(n^{-1}\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1} \;\xrightarrow{p}\; \left[\begin{array}{cc}\mathbf{F} & \mathbf{0}\\ \mathbf{0} & \langle \mathbf{z}'\mathbf{z}\rangle^{-1}\end{array}\right]$$

where

$$\mathbf{F} = \left[\begin{array}{cc}1/p_A & 0\\ 0 & 1/(1-p_A)\end{array}\right].$$

Also,

$$\tilde{\mathbf{x}}_i'\tilde{\mathbf{x}}_i \;=\; \left[\begin{array}{cc}\mathbf{G} & \mathbf{H}\\ \mathbf{H}' & \mathbf{z}_i'\mathbf{z}_i\end{array}\right],$$

where

$$\mathbf{G} = \left[\begin{array}{cc}T_i & 0\\ 0 & 1-T_i\end{array}\right], \qquad \mathbf{H} = \left[\begin{array}{c}T_i\mathbf{z}_i\\ (1-T_i)\mathbf{z}_i\end{array}\right].$$

So

$$n^{-1}\sum_{i=1}^{n}\hat{e}_i^2\tilde{\mathbf{x}}_i'\tilde{\mathbf{x}}_i \;=\; \left[\begin{array}{cc}\mathbf{K} & \mathbf{L}\\ \mathbf{L}' & n^{-1}\sum_{i=1}^{n}\hat{e}_i^2\mathbf{z}_i'\mathbf{z}_i\end{array}\right],$$

where

$$\mathbf{K} \;=\; \left[\begin{array}{cc}n^{-1}\sum_{i\in A}\hat{e}_i^2 & 0\\ 0 & n^{-1}\sum_{i\in B}\hat{e}_i^2\end{array}\right] = \left[\begin{array}{cc}\tilde{p}_A n_A^{-1}\sum_{i\in A}\hat{e}_i^2 & 0\\ 0 & (1-\tilde{p}_A)(n-n_A)^{-1}\sum_{i\in B}\hat{e}_i^2\end{array}\right],$$

$$\mathbf{L} \;=\; \left[\begin{array}{c}n^{-1}\sum_{i\in A}\hat{e}_i^2\mathbf{z}_i\\ n^{-1}\sum_{i\in B}\hat{e}_i^2\mathbf{z}_i\end{array}\right].$$

By Lemma A.9 and Condition 3, $\mathbf{L}$ and $n^{-1}\sum_{i=1}^{n}\hat{e}_i^2\mathbf{z}_i'\mathbf{z}_i$ are $O_p(1)$, and

$$\mathbf{K} \xrightarrow{p} \begin{bmatrix} p_A \lim_{n\to\infty}\sigma_{a^{**}}^2 & 0 \\ 0 & (1-p_A)\lim_{n\to\infty}\sigma_{b^{**}}^2 \end{bmatrix}.$$

The above results imply that the upper-left 2×2 block of $\mathbf{M}$ converges in probability to

$$\begin{bmatrix} 1/p_A & 0 \\ 0 & 1/(1-p_A) \end{bmatrix}\begin{bmatrix} p_A \lim_{n\to\infty}\sigma_{a^{**}}^2 & 0 \\ 0 & (1-p_A)\lim_{n\to\infty}\sigma_{b^{**}}^2 \end{bmatrix}\begin{bmatrix} 1/p_A & 0 \\ 0 & 1/(1-p_A) \end{bmatrix}$$
$$= \begin{bmatrix} p_A^{-1}\lim_{n\to\infty}\sigma_{a^{**}}^2 & 0 \\ 0 & (1-p_A)^{-1}\lim_{n\to\infty}\sigma_{b^{**}}^2 \end{bmatrix}.$$

Thus,

$$n\widehat{v}_{\text{adj}} \xrightarrow{p} \frac{1}{p_A}\lim_{n\to\infty}\sigma_{a^{**}}^2 + \frac{1}{1-p_A}\lim_{n\to\infty}\sigma_{b^{**}}^2.$$

Lemma A.6 implies

$$\text{avar}(\sqrt{n}[\widehat{\text{ATE}}_{\text{adj}} - \text{ATE}]) = \frac{1-p_A}{p_A}\lim_{n\to\infty}\sigma_{a^{**}}^2 + \frac{p_A}{1-p_A}\lim_{n\to\infty}\sigma_{b^{**}}^2 + 2\lim_{n\to\infty}\sigma_{a^{**},b^{**}}.$$

Let $\Delta = \text{plim}\, n\widehat{v}_{\text{adj}} - \text{avar}(\sqrt{n}[\widehat{\text{ATE}}_{\text{adj}} - \text{ATE}])$. Then

$$\begin{aligned}\Delta &= \lim_{n\to\infty}\sigma_{a^{**}}^2 + \lim_{n\to\infty}\sigma_{b^{**}}^2 - 2\lim_{n\to\infty}\sigma_{a^{**},b^{**}} \\ &= \lim_{n\to\infty}\sigma_{(a^{**}-b^{**})}^2 = \lim_{n\to\infty}\sigma_{(a-b)}^2 \geq 0.\end{aligned}$$

The proof for $n\widehat{v}_{\text{interact}}$ is similar.

Appendix B

Proofs for Chapter 3

B.1 Additional notation

Section 3.2 explains the notation used in Chapter 3. This appendix uses additional notation from Chapter 2 and Appendix A.

Let $\overline{a}$, $\overline{a}_A$, and $\overline{a}_B$ denote the means of a_i over the population, treatment group A, and treatment group B:

$$\overline{a} = \frac{1}{n}\sum_{i=1}^{n} a_i, \qquad \overline{a}_A = \frac{1}{n_A}\sum_{i\in A} a_i, \qquad \overline{a}_B = \frac{1}{n-n_A}\sum_{i\in B} a_i.$$

Use similar notation for the means of b_i, Y_i, $\mathbf{z}_i$, and other scalar, vector, and matrix variables. For example:

$$\overline{ab}_A = \frac{1}{n_A}\sum_{i\in A} a_i b_i, \qquad \overline{a\mathbf{z}}_A = \frac{1}{n_A}\sum_{i\in A} a_i \mathbf{z}_i, \qquad \overline{\mathbf{z}'\mathbf{z}}_A = \frac{1}{n_A}\sum_{i\in A} \mathbf{z}_i'\mathbf{z}_i.$$

Let $\tilde{p}_A = n_A/n$.

Let $\widehat{\mathbf{Q}}$ denote the vector of estimated coefficients on $\mathbf{z}_i$ in the OLS regression of Y_i on T_i and $\mathbf{z}_i$.

Let $\widehat{\mathbf{Q}}_a$ and $\widehat{\mathbf{Q}}_b$ denote the vectors of estimated coefficients on $\mathbf{z}_i$ in the OLS regressions of Y_i on $\mathbf{z}_i$ in groups A and B, respectively.

Condition 4 (mean-centering) will sometimes be assumed for convenience. The proof of Theorem 3.1 explains why this can be done without loss of generality.

Condition 4. The population means of the potential outcomes and the covariates are zero: $\overline{a} = \overline{b} = 0$ and $\overline{\mathbf{z}} = \mathbf{0}$.

B.2 Lemmas

Lemma B.1 generalizes Freedman's (2008b) Proposition 1.

Lemma B.1. *Let* $\mathbf{x}_i$ *and* $\mathbf{y}_i$ *be* $1 \times K$ *vectors defined for* $i = 1, \ldots, n$ *(i.e., for each subject* i *in the population). Let* $\mathbf{M}$ *be any fixed* $K \times K$ *matrix. Then*

$$\begin{aligned}
E([\overline{\mathbf{x}}_A - \overline{\mathbf{x}}]\mathbf{M}[\overline{\mathbf{y}}_A - \overline{\mathbf{y}}]') &= \frac{1}{n-1}\frac{1-\tilde{p}_A}{\tilde{p}_A}\frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \overline{\mathbf{x}})\mathbf{M}\mathbf{y}_i', \\
E([\overline{\mathbf{x}}_B - \overline{\mathbf{x}}]\mathbf{M}[\overline{\mathbf{y}}_B - \overline{\mathbf{y}}]') &= \frac{1}{n-1}\frac{\tilde{p}_A}{1-\tilde{p}_A}\frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \overline{\mathbf{x}})\mathbf{M}\mathbf{y}_i', \\
E([\overline{\mathbf{x}}_A - \overline{\mathbf{x}}]\mathbf{M}[\overline{\mathbf{y}}_B - \overline{\mathbf{y}}]') &= -\frac{1}{n-1}\frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \overline{\mathbf{x}})\mathbf{M}\mathbf{y}_i'.
\end{aligned}$$

Proof. Without loss of generality, we can assume $\overline{\mathbf{x}} = \overline{\mathbf{y}} = \mathbf{0}$, since the general case follows easily from this special case.

Note that if $i = j$, then $E(T_i T_j) = P(i \in A) = n_A/n$, and if $i \neq j$, then

$$E(T_i T_j) = P(i \in A, j \in A) = \frac{n_A}{n}\frac{n_A - 1}{n-1}.$$

Next,

$$\begin{aligned}
E([\overline{\mathbf{x}}_A - \overline{\mathbf{x}}]\mathbf{M}[\overline{\mathbf{y}}_A - \overline{\mathbf{y}}]') &= E(\overline{\mathbf{x}}_A \mathbf{M} \overline{\mathbf{y}}_A') \\
&= E\left(\left[\frac{1}{n_A}\sum_{i=1}^{n} T_i \mathbf{x}_i\right]\mathbf{M}\left[\frac{1}{n_A}\sum_{j=1}^{n} T_j \mathbf{y}_j\right]'\right) \\
&= \frac{1}{n_A^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbf{x}_i \mathbf{M}\mathbf{y}_j' E(T_i T_j) \\
&= \frac{1}{n_A^2}\left[\frac{n_A}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{M}\mathbf{y}_i' + \frac{n_A}{n}\frac{n_A-1}{n-1}\sum_{i=1}^{n}\sum_{j\neq i}\mathbf{x}_i\mathbf{M}\mathbf{y}_j'\right].
\end{aligned}$$

We can rewrite $\sum_{i=1}^{n}\sum_{j\neq i}\mathbf{x}_i\mathbf{M}\mathbf{y}_j'$ as

$$n^2 \overline{\mathbf{x}}\mathbf{M}\overline{\mathbf{y}}' - \sum_{i=1}^{n}\mathbf{x}_i\mathbf{M}\mathbf{y}_i' = -\sum_{i=1}^{n}\mathbf{x}_i\mathbf{M}\mathbf{y}_i'.$$

Therefore,

$$\begin{aligned}
E([\overline{\mathbf{x}}_A - \overline{\mathbf{x}}]\mathbf{M}[\overline{\mathbf{y}}_A - \overline{\mathbf{y}}]') &= \frac{1}{n_A}\frac{1}{n}\frac{n-n_A}{n-1}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{M}\mathbf{y}_i' \\
&= \frac{1}{n-1}\frac{1-\tilde{p}_A}{\tilde{p}_A}\frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \overline{\mathbf{x}})\mathbf{M}\mathbf{y}_i',
\end{aligned}$$

proving the first result. The second result is analogous.

Next note that if $i = j$, then $E(T_i[1 - T_j]) = E(T_i - T_i) = 0$, and if $i \neq j$, then

$$E(T_i[1 - T_j]) = P(i \in A, j \in B) = \frac{n_A}{n}\frac{n-n_A}{n-1}.$$

Therefore,

$$\begin{aligned}
E([\bar{\mathbf{x}}_A - \bar{\mathbf{x}}]\mathbf{M}[\bar{\mathbf{y}}_B - \bar{\mathbf{y}}]') &= E(\bar{\mathbf{x}}_A \mathbf{M} \bar{\mathbf{y}}_B') \\
&= E\left(\left[\frac{1}{n_A} \sum_{i=1}^{n} T_i \mathbf{x}_i \right] \mathbf{M} \left[\frac{1}{n - n_A} \sum_{j=1}^{n} (1 - T_j) \mathbf{y}_j \right]' \right) \\
&= \frac{1}{n_A} \frac{1}{n - n_A} \sum_{i=1}^{n} \sum_{j \neq i} \mathbf{x}_i \mathbf{M} \mathbf{y}_j' E(T_i[1 - T_j]) \\
&= \frac{1}{n} \frac{1}{n-1} \sum_{i=1}^{n} \sum_{j \neq i} \mathbf{x}_i \mathbf{M} \mathbf{y}_j' \\
&= -\frac{1}{n-1} \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i \mathbf{M} \mathbf{y}_i' \\
&= -\frac{1}{n-1} \frac{1}{n} \sum_{i=1}^{n} (\mathbf{x}_i - \bar{\mathbf{x}}) \mathbf{M} \mathbf{y}_i'.
\end{aligned}$$

□

Lemma B.2. *Let $\mathbf{D}_n$ and $\mathbf{F}_n$ be sequences of random, invertible matrices such that $\mathbf{D}_n = \mathbf{F}_n + O_p(g(n))$ and $\mathbf{F}_n \xrightarrow{p} \mathbf{F}$, where $g(n)$ is $o(1)$ and $\mathbf{F}$ is invertible. Then $\mathbf{D}_n^{-1} = \mathbf{F}_n^{-1} + O_p(g(n))$.*

Proof. Pre-multiply by $\mathbf{F}_n^{-1}$ and post-multiply by $\mathbf{D}_n^{-1}$ to get

$$\mathbf{F}_n^{-1} = \mathbf{D}_n^{-1} + \mathbf{F}_n^{-1} \cdot O_p(g(n)) \cdot \mathbf{D}_n^{-1}.$$

By the continuous mapping theorem, $\mathbf{F}_n^{-1} \xrightarrow{p} \mathbf{F}^{-1}$ and $\mathbf{D}_n^{-1} \xrightarrow{p} \mathbf{F}^{-1}$. Therefore, $\mathbf{F}_n^{-1} - \mathbf{D}_n^{-1}$ is $O_p(g(n))$. □

B.3 Proof of Theorem 3.1

We can assume Condition 4 without loss of generality, since centering a_i, b_i, and $\mathbf{z}_i$ has no effect on $\widehat{\text{ATE}}_{\text{adj}} - \text{ATE}$ and the expression for $E(\eta_n)$.

By Lemma A.2 and Condition 4,

$$\widehat{\text{ATE}}_{\text{adj}} - \text{ATE} = (\bar{a}_A - \bar{\mathbf{z}}_A \widehat{\mathbf{Q}}) - (\bar{b}_B - \bar{\mathbf{z}}_B \widehat{\mathbf{Q}}),$$

and from the proof of Lemma A.4, $\widehat{\mathbf{Q}} = \mathbf{D}^{-1}\mathbf{N}$, where

$$\begin{aligned}
\mathbf{D} &= \overline{\mathbf{z}'\mathbf{z}} - \tilde{p}_A \bar{\mathbf{z}}_A' \bar{\mathbf{z}}_A - (1 - \tilde{p}_A) \bar{\mathbf{z}}_B' \bar{\mathbf{z}}_B, \\
\mathbf{N} &= \tilde{p}_A (\overline{a\mathbf{z}}_A - \bar{a}_A \bar{\mathbf{z}}_A)' + (1 - \tilde{p}_A)(\overline{b\mathbf{z}}_B - \bar{b}_B \bar{\mathbf{z}}_B)'.
\end{aligned}$$

Note that $\bar{\mathbf{z}}_A$, $\bar{\mathbf{z}}_B$, $\bar{a}_A$, and $\bar{b}_B$ are $O_p(1/\sqrt{n})$ by a finite-population Central Limit Theorem [e.g., Theorem 1 in Freedman (2008b) or Theorem 2.8.2 in Lehmann

(1999)] and Conditions 1–4, while $\tilde{p}_A$, $\overline{a\mathbf{z}}_A$, and $\overline{b\mathbf{z}}_B$ are $O_p(1)$ by Lemma A.1 and Conditions 1–3. We thus get

$$\mathbf{D} = \overline{\mathbf{z}'\mathbf{z}} + O_p(1/n),$$

and by Lemma B.2 and Conditions 2 and 4, $\mathbf{D}^{-1} = \mathbf{M} + O_p(1/n)$. Therefore,

$$\begin{aligned}
\bar{\mathbf{z}}_A\widehat{\mathbf{Q}} &= \bar{\mathbf{z}}_A\mathbf{M}[\tilde{p}_A\overline{a\mathbf{z}}_A + (1-\tilde{p}_A)\overline{b\mathbf{z}}_B]' + O_p(n^{-3/2}),\\
\bar{\mathbf{z}}_B\widehat{\mathbf{Q}} &= \bar{\mathbf{z}}_B\mathbf{M}[\tilde{p}_A\overline{a\mathbf{z}}_A + (1-\tilde{p}_A)\overline{b\mathbf{z}}_B]' + O_p(n^{-3/2}).
\end{aligned}$$

Now let $\widetilde{\mathbf{Q}}$ denote the vector of slope coefficients in the population least squares regression of $\tilde{p}_A a_i + (1-\tilde{p}_A)b_i$ on $\mathbf{z}_i$, that is,

$$\widetilde{\mathbf{Q}} = \mathbf{M}[\tilde{p}_A\overline{a\mathbf{z}} + (1-\tilde{p}_A)\overline{b\mathbf{z}}]'.$$

Also, let

$$\check{\mathbf{Q}} = \mathbf{M}[\tilde{p}_A(\overline{a\mathbf{z}}_A - \overline{a\mathbf{z}}) + (1-\tilde{p}_A)(\overline{b\mathbf{z}}_B - \overline{b\mathbf{z}})]'.$$

Note that $\widetilde{\mathbf{Q}}$ is fixed, while $\check{\mathbf{Q}}$ may vary across randomizations. We have

$$\widehat{\mathrm{ATE}}_{\mathrm{adj}} - \mathrm{ATE} = \eta_n + O_p(n^{-3/2}),$$

where

$$\eta_n = [\overline{a}_A - \bar{\mathbf{z}}_A(\widetilde{\mathbf{Q}} + \check{\mathbf{Q}})] - [\overline{b}_B - \bar{\mathbf{z}}_B(\widetilde{\mathbf{Q}} + \check{\mathbf{Q}})].$$

From Condition 4 and the unbiasedness of sample means under simple random sampling, $E(\overline{a}_A) = E(\overline{b}_B) = 0$ and $E(\bar{\mathbf{z}}_A) = E(\bar{\mathbf{z}}_B) = \mathbf{0}$. Thus,

$$E(\eta_n) = -[E(\bar{\mathbf{z}}_A\check{\mathbf{Q}}) - E(\bar{\mathbf{z}}_B\check{\mathbf{Q}})].$$

Next,

$$E(\bar{\mathbf{z}}_A\check{\mathbf{Q}}) = \tilde{p}_A E(\bar{\mathbf{z}}_A\mathbf{M}[\overline{a\mathbf{z}}_A - \overline{a\mathbf{z}}]') + (1-\tilde{p}_A)E(\bar{\mathbf{z}}_A\mathbf{M}[\overline{b\mathbf{z}}_B - \overline{b\mathbf{z}}]').$$

Using Lemma B.1 and Condition 4, we get

$$\begin{aligned}
E(\bar{\mathbf{z}}_A\check{\mathbf{Q}}) &= (1-\tilde{p}_A)\frac{1}{n-1}\frac{1}{n}\left[\sum_{i=1}^n(\mathbf{z}_i-\bar{\mathbf{z}})\mathbf{M}(a_i\mathbf{z}_i)' - \sum_{i=1}^n(\mathbf{z}_i-\bar{\mathbf{z}})\mathbf{M}(b_i\mathbf{z}_i)'\right]\\
&= (1-\tilde{p}_A)\frac{1}{n-1}\frac{1}{n}\sum_{i=1}^n[(a_i-b_i)-\mathrm{ATE}](\mathbf{z}_i-\bar{\mathbf{z}})\mathbf{M}(\mathbf{z}_i-\bar{\mathbf{z}})'.
\end{aligned}$$

Similarly,

$$E(\bar{\mathbf{z}}_B\check{\mathbf{Q}}) = -\tilde{p}_A\frac{1}{n-1}\frac{1}{n}\sum_{i=1}^n[(a_i-b_i)-\mathrm{ATE}](\mathbf{z}_i-\bar{\mathbf{z}})\mathbf{M}(\mathbf{z}_i-\bar{\mathbf{z}})'$$

and thus

$$E(\eta_n) = -\frac{1}{n-1}\frac{1}{n}\sum_{i=1}^n[(a_i-b_i)-\mathrm{ATE}](\mathbf{z}_i-\bar{\mathbf{z}})\mathbf{M}(\mathbf{z}_i-\bar{\mathbf{z}})'.$$

B.4 Proof of Theorem 3.2

We can again assume Condition 4 without loss of generality. By Lemma A.3,

$$\widehat{\text{ATE}}_{\text{interact}} - \text{ATE} \quad = \quad (\overline{a}_A - \overline{\mathbf{z}}_A \widehat{\mathbf{Q}}_a) - (\overline{b}_B - \overline{\mathbf{z}}_B \widehat{\mathbf{Q}}_b).$$

Recall that $\widehat{\mathbf{Q}}_a$ is the vector of estimated slope coefficients in the OLS regression of a_i on $\mathbf{z}_i$ in group A, while $\widetilde{\mathbf{Q}}_a$ is the corresponding population least squares slope vector. By Condition 4, the population least squares prediction error $\tilde{a}_i$ is

$$\tilde{a}_i \quad = \quad a_i - \mathbf{z}_i \widetilde{\mathbf{Q}}_a.$$

Thus,

$$\begin{aligned}
\widehat{\mathbf{Q}}_a &= \left[\sum_{i\in A} (\mathbf{z}_i - \overline{\mathbf{z}}_A)'(\mathbf{z}_i - \overline{\mathbf{z}}_A)\right]^{-1} \left[\sum_{i\in A} (\mathbf{z}_i - \overline{\mathbf{z}}_A)'(a_i - \overline{a}_A)\right] \\
&= \left[\sum_{i\in A} (\mathbf{z}_i - \overline{\mathbf{z}}_A)'(\mathbf{z}_i - \overline{\mathbf{z}}_A)\right]^{-1} \left[\sum_{i\in A} (\mathbf{z}_i - \overline{\mathbf{z}}_A)'(\mathbf{z}_i \widetilde{\mathbf{Q}}_a + \tilde{a}_i - \overline{a}_A)\right] \\
&= \widetilde{\mathbf{Q}}_a + \left[\sum_{i\in A} (\mathbf{z}_i - \overline{\mathbf{z}}_A)'(\mathbf{z}_i - \overline{\mathbf{z}}_A)\right]^{-1} \left[\sum_{i\in A} (\mathbf{z}_i - \overline{\mathbf{z}}_A)'\tilde{a}_i\right]
\end{aligned}$$

and

$$\begin{aligned}
\overline{\mathbf{z}}_A \widehat{\mathbf{Q}}_a &= \overline{\mathbf{z}}_A \widetilde{\mathbf{Q}}_a + \overline{\mathbf{z}}_A \left[\overline{\mathbf{z}'\mathbf{z}}_A - \overline{\mathbf{z}}'_A \overline{\mathbf{z}}_A\right]^{-1} \left[\overline{\tilde{a}\mathbf{z}}_A - \overline{\tilde{a}}_A \overline{\mathbf{z}}_A\right]' \\
&= \overline{\mathbf{z}}_A \widetilde{\mathbf{Q}}_a + \overline{\mathbf{z}}_A \mathbf{D}^{-1} \mathbf{N},
\end{aligned}$$

where $\mathbf{D} = \overline{\mathbf{z}'\mathbf{z}}_A - \overline{\mathbf{z}}'_A \overline{\mathbf{z}}_A$ and $\mathbf{N} = (\overline{\tilde{a}\mathbf{z}}_A - \overline{\tilde{a}}_A \overline{\mathbf{z}}_A)'$.

Note that $\overline{\mathbf{z}}_A$ and $\overline{\tilde{a}}_A = \overline{a}_A - \overline{\mathbf{z}}_A \widetilde{\mathbf{Q}}_a$ are $O_p(1/\sqrt{n})$. We get $\mathbf{D} = \overline{\mathbf{z}'\mathbf{z}}_A + O_p(1/n)$, and by Lemma B.2, Lemma A.1, and Condition 2,

$$\mathbf{D}^{-1} = \overline{\mathbf{z}'\mathbf{z}}_A^{-1} + O_p(1/n).$$

Next, by a finite-population Central Limit Theorem [Theorem 1 in Freedman (2008b) or Theorem 2.8.2 in Lehmann (1999)] and our premise (specifically Condition 3, bounded eighth moments, and finite limiting variances of $z_{ik} z_{i\ell}$),

$$\overline{\mathbf{z}'\mathbf{z}}_A = \overline{\mathbf{z}'\mathbf{z}} + O_p(1/\sqrt{n}).$$

By Lemma B.2 and Conditions 2 and 4, $\overline{\mathbf{z}'\mathbf{z}}_A^{-1} = \mathbf{M} + O_p(1/\sqrt{n})$. Thus,

$$\mathbf{D}^{-1} = \mathbf{M} + O_p(1/\sqrt{n}).$$

In the population, the prediction errors $\tilde{a}_i$ are orthogonal to the covariates:

$$\begin{aligned}
\sum_{i=1}^{n} \tilde{a}_i \mathbf{z}_i' &= \sum_{i=1}^{n} (a_i - \mathbf{z}_i \widetilde{\mathbf{Q}}_a) \mathbf{z}_i' \\
&= \sum_{i=1}^{n} a_i \mathbf{z}_i' - \sum_{i=1}^{n} \mathbf{z}_i' \mathbf{z}_i \widetilde{\mathbf{Q}}_a \\
&= \sum_{i=1}^{n} a_i \mathbf{z}_i' - \left[\sum_{i=1}^{n} \mathbf{z}_i' \mathbf{z}_i \right] \left[\sum_{i=1}^{n} \mathbf{z}_i' \mathbf{z}_i \right]^{-1} \left[\sum_{i=1}^{n} \mathbf{z}_i' a_i \right] \\
&= \mathbf{0}.
\end{aligned}$$

Thus, $\overline{\tilde{a}\mathbf{z}} = \mathbf{0}$, so by a finite-population Central Limit Theorem, $\overline{\tilde{a}\mathbf{z}}_A$ is $O_p(1/\sqrt{n})$.

Therefore,

$$\overline{\mathbf{z}}_A \widehat{\mathbf{Q}}_a = \overline{\mathbf{z}}_A \widetilde{\mathbf{Q}}_a + \overline{\mathbf{z}}_A \mathbf{M} \overline{\tilde{a}\mathbf{z}}_A' + O_p(n^{-3/2}).$$

Similarly,

$$\overline{\mathbf{z}}_B \widehat{\mathbf{Q}}_b = \overline{\mathbf{z}}_B \widetilde{\mathbf{Q}}_b + \overline{\mathbf{z}}_B \mathbf{M} \overline{\tilde{b}\mathbf{z}}_B' + O_p(n^{-3/2}).$$

We thus have $\widehat{\text{ATE}}_{\text{interact}} - \text{ATE} = \tilde{\eta}_n + O_p(n^{-3/2})$, where

$$\tilde{\eta}_n = (\overline{a}_A - \overline{\mathbf{z}}_A \widetilde{\mathbf{Q}}_a - \overline{\mathbf{z}}_A \mathbf{M} \overline{\tilde{a}\mathbf{z}}_A') - (\overline{b}_B - \overline{\mathbf{z}}_B \widetilde{\mathbf{Q}}_b - \overline{\mathbf{z}}_B \mathbf{M} \overline{\tilde{b}\mathbf{z}}_B').$$

The result for $E(\tilde{\eta}_n)$ follows from Condition 4 and Lemma B.1.