Essays in Semiparametric Econometrics

Essays in Semiparametric Econometrics

A dissertation presented

by

Jann Spiess

to

The Department of Economics

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Economics

Harvard University

Cambridge, Massachusetts

May 2018

Dissertation Advisor:
Professor Sendhil Mullainathan

Author:
Jann Spiess

Essays in Semiparametric Econometrics

Abstract

This dissertation studies how the estimation of first-stage nuisance parameters associated with control and instrumental variables affects second-stage estimation of treatment-effect parameters.

Chapter 1 approaches the analysis of experimental data as a mechanism-design problem that acknowledges that researchers choose between estimators according to their own preferences. Specifically, I focus on covariate adjustments, which can increase the precision of a treatment-effect estimate, but open the door to bias when researchers engage in specification searches. I establish that unbiasedness is a requirement on the estimation of the average treatment effect that aligns researchers' preferences with the minimization of the mean-squared error relative to the truth, and that fixing the bias can yield an optimal restriction in a minimax sense. I then provide a characterization of unbiased treatment-effect estimators as sample-splitting procedures.

Chapter 2 gives two examples in which we can improve estimation by shrinking in high-dimensional nuisance parameters while avoiding or even reducing the bias in a low-dimensional target parameter. I first consider shrinkage estimation of the nuisance parameters associated with control variables in a linear model, and show that for at least three control variables the standard least-squares estimator is dominated with respect to variance in the treatment effect even among unbiased estimators when treatment is exogenous. Second, I consider shrinkage in the estimation of first-stage instrumental variable coefficients in a two-stage linear regression model. For at least four instrumental variables, I establish that the standard two-stage least-squares estimator is dominated with respect to bias.

Chapter 3 (with Alberto Abadie) considers regression analysis of treatment effects after nearest-neighbor matching on control variables. We show that standard errors that ignore the matching step are not generally valid if the second-step regression model is misspecified. We offer two easily implementable alternatives, (i) clustering the standard errors at the level of the matches, or (ii) a nonparametric block bootstrap procedure, that produce approximations to the distribution of the post-matching estimator that are robust to misspecification, provided that matching is done without replacement.

Contents

List of Tables

List of Figures

Acknowledgments

For their guidance I am indebted to Sendhil Mullainathan, Alberto Abadie, Elie Tamer, and Gary Chamberlain. For their advice I am also thankful to Edward Glaeser, Lawrence Katz, David Laibson, and Andrei Shleifer. For their feedback on the projects in this dissertation I thank Laura Blattner, Kirill Borusyak, Avi Feller, Nathaniel Hendren, Simon Jäger, Ariella Kahn-Lang, Maximilian Kasy, Gary King, Scott Kominers, Eben Lazarus, Shengwu Li, Carl Morris, Amanda Palais, Mikkel Plagborg-Møller, Ashesh Rambachan, Jonathan Roth, Elizabeth Santorella, Heather Sarsons, Neil Shephard, and James Stock. For their support during this time I am also thankful to Valentin Bolotnyy, Katharina Bürkin, Denniz Dönmez, Talia Gillis, Frank Schilbach, Chenzi Xu, my brother Julian, and my parents.

I dedicate this dissertation to those who instilled in me my love for mathematics: Chapter 1 to Anusch Taraz, who taught me discrete mathematics; Chapter 2 to the memory of Friedrich Roesler, who taught me linear algebra; Chapter 3 to Silke Rolles, who taught me probability.

Introduction

In many empirical applications, we are interested in a (typically low-dimensional) target parameter in the presence of a (sometimes high-dimensional) nuisance parameter. In my dissertation, I consider the estimation of treatment effects after a pre-processing step involving nuisance parameters associated with control or instrumental variables. I am particularly interested in how model selection in the first stage can improve the estimation of the target parameter, how we can avoid second-stage biases when researchers engage in first-stage specification searches, and how second-stage inference is affected by the first stage.

In the first chapter, I consider the estimation of an average treatment effect in an experiment in a world where researchers have discretion over estimators and engage in specification searches. Econometric analysis typically focuses on the statistical properties of fixed estimators and ignores researcher choices. In this chapter, I approach the analysis of experimental data as a mechanism-design problem that acknowledges that researchers choose between estimators, sometimes based on the data and often according to their own preferences. Specifically, I focus on covariate adjustments, which can increase the precision of a treatment-effect estimate, but open the door to bias when researchers engage in specification searches.

Having set up the estimation of the average treatment effect as a principal–agent problem, I characterize the optimal solution of a designer who puts restrictions on the estimation, as well as the second-best estimator given the designer's restrictions. First, I establish that unbiasedness is a requirement on the estimation of the average treatment effect that aligns researchers' preferences with the minimization of the mean-squared error relative to the truth, and that fixing the bias can yield an optimal restriction in a minimax sense. Second,

I provide a constructive characterization of all treatment-effect estimators with fixed bias as sample-splitting procedures. Third, I show that a researcher restricted specifically to the class of unbiased estimators of the average treatment effect solves a prediction problem. The equivalence of unbiased estimation and prediction across sample splits characterizes all admissible unbiased procedures in finite samples, leaves space for beneficial specification searches, and offers an opportunity to leverage machine learning. As a practical implication, I describe flexible pre-analysis plans for randomized experiments that achieve efficiency without bias.

In the second chapter, I consider shrinkage in high-dimensional nuisance parameters associated with control or instrumental variables when we care about the bias properties of a (typically low-dimensional) target parameter. Shrinkage estimation usually reduces variance at the cost of bias. But when we care only about some parameters of a model, I give two examples where we can improve estimation by shrinking in high-dimensional nuisance parameters while avoiding or even reducing the bias in the low-dimensional target.

Specifically, I consider two types of two-step estimators, showing in both cases how shrinkage in the first stage can improve estimation in the second. I first consider shrinkage estimation of the nuisance parameters associated with control variables, and show that the standard least-squares estimator is dominated with respect to squared-error loss in the treatment effect even among unbiased estimators when treatment is exogenous. Second, the two-stage least-squares (TSLS) estimator is known to be biased when its first-stage fit is poor, and I show that shrinkage in the first stage of a two-stage linear regression model reduces the bias of the standard TSLS estimator. Both estimators apply James–Stein-type shrinkage in first-stage high-dimensional Normal-means problems, and provide dominance in finite samples under linearity, homoscedasticity and Normality assumptions.

In the third chapter (with Alberto Abadie), we consider regression analysis of treatment effects after matching on control variables. Nearest-neighbor matching (Cochran, 1953; Rubin, 1973) is a popular nonparametric tool to create balance between treatment and control groups in observational studies. As a preprocessing step before regression analysis, matching reduces

the dependence on parametric modeling assumptions (Ho *et al.*, 2007). Moreover, matching followed by regression allows estimation of elaborate models that are useful to describe heterogeneity in treatment effects. In current empirical practice, however, the matching step is often ignored for the estimation of standard errors and confidence intervals. That is, to do inference, researchers proceed as if matching did not take place.

In this chapter, we offer tools for valid inference after matching. Specifically, we show that ignoring the matching first step results in asymptotically valid standard errors if matching is done without replacement and the regression model is correctly specified relative to the population regression function of the outcome variable on the treatment variable and *all* the covariates used for matching. However, standard errors that ignore the matching step are not valid if matching is conducted with replacement or, more crucially, if the second step regression model is misspecified in the sense indicated above. We show that two easily implementable alternatives, (i) clustering the standard errors at the level of the matches, or (ii) a nonparametric block bootstrap procedure, produce approximations to the distribution of the post-matching estimator that are robust to misspecification, provided that matching is done without replacement. These results allow robust inference for post-matching methods that use regression in the second step. A simulation study and an empirical example demonstrate the empirical relevance of our results.

As we work with increasingly high-dimensional, "big" data, and more and more methods from machine learning proliferate empirical work, the question of how we can employ machine-learning tools designed for big-data prediction in causal estimation has become a central challenge in econometrics. In this dissertation, I aim to provide some insights into how first-stage nonparametric estimation can affect and improve second-stage estimation of low-dimensional target parameters.

But as complex machine-learning methods proliferate data-driven decision making, there is also a risk that more flexible and less transparent methods increase biases from misspecified models and specification searches. While these methods promise to "let the data speak," they may also exacerbate p-hacking and publication biases. In my dissertation, I therefore develop

a decision-theoretic principal–agent perspective on estimation that explicitly considers the preferences and degrees of freedom of the analyst. I thus hope to contribute to integrating specification searches into causal inference while avoiding biases from human and machine choices.

Chapter 1

Optimal Estimation with Misaligned Preferences

1.1 Introduction

There is a tension between flexibility and robustness in empirical work. Consider an investigator who estimates a treatment effect from experimental data. If the investigator has the freedom to choose a specification that adjusts for control variables, her choice can improve the precision of the estimate. However, the investigator's specification search may also produce an estimate that reflects a preference for publication or ideology instead of a more precise guess of the truth.[1] To solve this problem, we sometimes tie the investigator's hands and restrict her to a simple specification, like a difference in averages. In contrast, this chapter characterizes flexible estimators that leverage the data and researchers' expertise, and do not also reflect researchers' preferences.

To characterize optimal estimators when researcher and social preferences are misaligned,

[1] A literature in statistics dating back to at least Sterling (1959) and Tullock (1959), and most strongly associated with the work of Edward Leamer (e.g. Leamer, 1974, 1978), acknowledges that empirical estimates reflect not just data, but also researcher motives. Fears of biases have been fueled more recently by replication failures (Open Science Collaboration, 2015), anomalies in published p-values (Brodeur *et al.*, 2016), and empirical evidence for publication biases (Andrews and Kasy, 2017). This concern is also evident in the American Economic Association's 2012 decision to establish a registry for randomized controlled trials.

I approach the analysis of experimental data as a mechanism-design problem.[2] Concretely, I consider a designer and an investigator who are engaged in the estimation of an average treatment effect. As the designer, we aim to obtain a precise estimate of the truth (which I capture in terms of mean-squared error). I assume however that the investigator may care about the value of the estimate and not only its precision. For example, the investigator may have a preference for large estimates in order to get published. The investigator picks an estimator based on her private information about the specific experiment. The designer chooses optimal constraints on the estimation by the investigator.[3]

First, I argue that we should not leave the decision over the bias of an estimator to the investigator, and motivate a restriction to estimators with fixed bias. More precisely, I prove that setting the bias aligns the incentives of the investigator and the designer and is a minimax optimal solution to the designer's problem under suitable assumptions on preferences.[4] Allowing the investigator to choose the bias can, in principle, improve overall precision through a reduction in the variance. But an investigator could use her control over the bias to reflect her preferences rather than her private information. Among unbiased estimators, for example, even an investigator who wants to obtain an estimate close to some large, fixed value will still choose an estimator that minimizes the variance.

Second, having motivated a bias restriction, I prove that every estimator of the average treatment effect with fixed bias has a sample-splitting representation. As the starting point for this representation, consider a familiar estimator that is unbiased, namely the difference in averages between treatment and control groups. We can adjust this estimator for control variables by a procedure that splits the sample into two groups. From the first group, we

[2] Like Leamer (1974, 1978), I explicitly consider researchers' degrees of freedom. Like Glaeser (2006), I also model their preferences. Like Schorfheide and Wolpin (2012, 2016), I employ a principal–agent perspective to justify data-splitting procedures.

[3] Abstractly, the designer could represent professional norms. Concretely, it could represent a journal setting standards for the analysis of randomized controlled trials, or the U.S. Food and Drug Administration (FDA) imposing rules for the evaluation of new drugs.

[4] This result echoes Frankel's (2014) characterization of simple delegation mechanisms that align an agent's choices with a principal's preferences by fixing budgets. In Section 1.5, I explore the similarities of my solution to results in the mechanism-design literature on delegation that goes back to Holmström (1978, 1984), and I exploit these parallels in the proof of my minimax result.

calculate regression adjustments that we subtract from the outcomes in the second group. The updated difference in averages is still unbiased by construction. Though this procedure appears specific, I prove that any estimator with fixed bias can be represented by multiple such sample-splitting steps. Unbiased estimators, for example, can differ from a difference in averages only by leave-one-out or leave-two-out regression adjustments of individual outcomes.[5]

Third, I focus specifically on estimation with zero bias, and show that an investigator restricted to unbiasedness will solve a prediction problem. By the sample-splitting representation, I can write every unbiased estimator of the average treatment effect in terms of a set of regression adjustments. When choosing from this restricted set of estimators, the investigator picks regression adjustments that minimize prediction risk for a specific loss function. Each optimal adjustment predicts the outcomes of one or two units from other units in the sample.

The investigator's solution reveals a finite-sample complete-class theorem that characterizes all admissible unbiased estimators of an average treatment effect as solutions to out-of-sample prediction problems. Since my results hold exactly without taking large-sample limits or relying on other approximations, I obtain a general duality between unbiased estimation and prediction without putting any essential restrictions on the distribution of the data other than random assignment of treatment. Any admissible unbiased estimator corresponds exactly to a set of admissible prediction solutions.

As a practical implication, my results motivate and describe flexible yet robust pre-analysis plans for the analysis of experimental data.[6] Having established that unbiased estimation is equivalent to a set of prediction tasks, there are two types of flexible pre-analysis plan that achieve precise estimation of treatment effects without leaving room for bias from specification searches. In the first type, the investigator commits to an algorithm that predicts outcomes from covariates. This algorithm can engage in automated specification searches to learn a

[5]In particular, for known treatment probability, I show that all unbiased estimators of the sample-average treatment effect take the form of the "leave-one-out potential outcomes" (LOOP) estimator from Wu and Gagnon-Bartsch (2017), which is a special case of Aronow and Middleton's (2013) extension of the Horvitz and Thompson (1952) estimator.

[6]Coffman and Niederle (2015), Olken (2015), and Heckman and Singer (2017) discuss the benefits, costs, and limitations of pre-analysis plans. I resolve an implicit flexibility-robustness tradeoff for one specific setting.

good model from the data.[7] Adjusting outcomes by its fitted out-of-sample predictions will yield an unbiased estimator.

There is a second, more flexible type of pre-analysis plan that achieves unbiased and precise estimation without the investigator committing to her specification searches in advance. In this second type of pre-analysis plan, the investigator only commits to splitting the data and distributing subsamples to her research team. Each researcher then engages in specification searches on a part of the data and reports back a prediction function. As my fourth main result, I characterize all unbiased estimators of the treatment effect that delegate the estimation of some or all regression adjustments in this way. Delegation to one researcher improves over simple pre-analysis plans.[8] Delegation to at least two researchers asymptotically attains the semi-parametric efficiency bound of Hahn (1998) under assumptions that apply to most parametric and many semi- and non-parametric estimators of the regression adjustments.

The results in this chapter relate to the practice of sample splitting in econometrics, statistics, and machine learning. From Hájek (1962) to Jackknife IV (Angrist *et al.*, 1999), model selection (e.g. Hansen and Racine, 2012), and time-series forecasting (see e.g. Diebold, 2015; Hirano and Wright, 2017), sample splitting is used as a tool to avoid bias by construction. Wager and Athey (2017) highlight the role of sample splitting in the estimation of heterogeneous treatment effects. Chernozhukov *et al.* (2018) show its relevance in achieving valid and efficient inference in high-dimensional observational data. My results show that sample splitting is not just an ad-hoc tool, but a feature of optimal estimators.[9] I establish that sample splitting is a necessary restriction on the investigator's estimator to achieve fixed bias and align incentives.

Moreover, I build upon an active literature in statistics on regression adjustments to experimental data. Freedman (2008) and Lin (2013) discuss the bias of linear-least squares

[7] In a similar spirit, Balzer *et al.* (2016) propose a data-adaptive procedure that selects among specifications to minimize the variance of treatment-effect estimators in experiments.

[8] My hold-out approach is similar to Dahl *et al.* (2008), Fafchamps and Labonne (2016) and Anderson and Magruder (2017), who all propose split-sample strategies to combine exploratory data analysis with valid inference. Dwork *et al.* (2015) propose a protocol to reuse the hold-out data to improve efficiency. I show that in my setting simple hold-out procedures are dominated when data can be distributed to multiple researchers.

[9] A rationale for holding out data in policy evaluation from randomized experiments has also been formalized by Schorfheide and Wolpin (2012, 2016).

regression adjustments. Most closely related to the investigator's solution in my paper, Wu and Gagnon-Bartsch (2017) propose the "leave-one-out potential outcomes" (LOOP) estimator that yields regression adjustments without bias, which coincides with the variance-minimizing unbiased estimator chosen by the researcher in my setting for the case of known treatment probability. Wager *et al.* (2016) propose a similar sample-splitting estimator based on separate prediction problems in the treatment and control groups. Relative to this literature, I motivate a bias restriction and fully characterize estimators with given bias. Specifically, I show that in my setting and the case of known treatment probability any variance-minimal unbiased estimator can be written in the form of Wu and Gagnon-Bartsch's (2017) LOOP estimator, where the adjustment terms solve a prediction problem in finite samples.

Relatedly, this chapter contributes to a growing literature that employs machine learning in program evaluation. Supervised machine learning algorithms solve prediction problems like those that I show to be equivalent to unbiased estimation (see e.g. Mullainathan and Spiess, 2017). As in Wager *et al.* (2016) and Wu and Gagnon-Bartsch (2017), the sample-splitting construction allows researchers to leverage machine learning in estimating average treatment effects in experimental data. Bloniarz *et al.* (2016) specifically use the LASSO to select among control variables in experiments. Athey and Imbens (2016) use regression trees to estimate heterogeneous treatment effects. Chernozhukov *et al.* (2017) estimate treatment effects from high-dimensional observational data. I contribute a finite-sample principal–agent framework for integrating machine learning, which is mostly agnostic about specific algorithms or asymptotic approximations.

My analysis is limited in three ways. First, I assume randomization, and thus that identification is resolved by design. My findings extend to known propensity scores, stratified and conditional randomization, and corresponding identification from quasi-experiments.[10] Second, I focus on the analysis of a single experiment, and neither on repeated interactions

[10]When treatment is not random, endogeneity creates auxiliary prediction tasks in the propensity score that interact with fitting regression adjustments (Robins and Rotnitzky, 1995; Chernozhukov *et al.*, 2018). Finite-sample unbiased estimation may then be infeasible absent strong parametric assumptions, and inference may be invalid when these additional prediction tasks are ignored (Belloni *et al.*, 2014).

between designer and investigator, nor on the publication policies that may shape investigators' preferences. Third, I characterize optimal estimators in terms of prediction tasks, but I do not discuss in depth the solution to these prediction problems. A large and active literature that straddles econometrics, statistics, and machine learning provides guidance and tools to provide efficient prediction functions.

The remaining chapter is structured as follows. Section 1.2 introduces the main ideas behind my theoretical results in a stylized example. In Section 1.3, I formally lay out the specific estimation setting and my mechanism-design approach. I preview my main theoretical results in Section 1.4. In Section 1.5, I solve for optimal restrictions on the investigator's estimation. Section 1.6 characterizes unbiased estimators and solves for the investigator's second-best choice. For the case that full ex-ante commitment is infeasible or impractical, Section 1.7 considers unbiased estimators that permit ex-post researcher input. In the Conclusion, I discuss extensions. In the Appendix, I collect the proofs of my main results and discuss asymptotic inference.

1.2 A Simple Example

I consider the estimation of a sample-average treatment effect. But the main features of my analysis are already apparent when we focus on a single unit within that sample. As an example, I discuss the estimation of the effect of random assignment to a job-training program on the earnings of one specific worker.[11]

1.2.1 Estimating the Unit-Level Causal Effect

The causal effect on unit i is $\tau_i = y_i(1) - y_i(0)$, where $y_i(1), y_i(0)$ are the potential outcomes when assigned to treatment or control, respectively. For assignment to a job-training program, $y_i(1) = \$1,190$ could be the earnings of worker i when he is offered the training program, and $y_i(0) = \$1,080$ the earnings of the *same* worker without access to this training, so $\tau_i = \$110$.

[11] Throughout, I focus on intent-to-treat effects, so I do not consider take-up or the use of random assignment as an instrument.

We do not observe both potential outcomes for one unit simultaneously, but observe only the treatment status d_i and the realized outcome

$$y_i = \begin{cases} y_i(1), & d_i = 1, \\ y_i(0), & d_i = 0. \end{cases}$$

But since treatment is assigned randomly (with probability $p = \mathrm{P}(d_i = 1)$), we can still obtain an unbiased estimate of the unit treatment effect.[12] Indeed, I will note below that $\frac{d_i - p}{p(1-p)} y$ is an unbiased estimator for τ_i. (Throughout, by "unbiased" I mean that, for fixed potential outcomes $y_i(1)$ and $y_i(0)$, the treatment-effect estimator averages out to $y_i(1) - y_i(0)$ over random draws of treatment d_i.)

In addition to the realized outcome y_i and treatment status d_i, I assume that we also have access to some pre-treatment characteristics x_i of unit i. Estimating the treatment effect $\tau_i = y_i(1) - y_i(0)$ for, say, a treated unit ($d_i = 1$) amounts to imputing the missing, counterfactual control outcome $y_i(0)$. When we have additional information about that unit, we can hope to use it together with the outcome, treatment, and characteristic data $z_{-i} = (y_j, d_j, x_j)_{j \neq i}$ of all other units to estimate $y_i(0)$, and thus τ_i. The investigator could, for example, run a linear regression of earnings on treatment, pre-assignment earnings, and some basic demographic characteristics to impute the counterfactual outcome $y_i(0)$. She could then estimate that worker's treatment effect by the difference between realized and imputed earnings.

If we do not put any restriction on estimation and investigator and social preferences agree, then the investigator's estimator will represent her expertise as well as the data. I model the investigator's expertise as a prior distribution π over potential outcomes $y_i(1), y_i(0)$ given

[12] Here, I assume that we know that treatment has been assigned with known probability $p = \mathrm{P}(d_i = 1)$. Throughout the remaining chapter, I also consider random assignment with a fixed number of treated units rather than a known ex-ante probability of treatment. The case of known number n_1 of treated units has structurally similar features, but is *not* the same as the case with known probability $p = \frac{n_1}{n}$. The reason for the difference is that knowledge of all *other* units' treatment status is not informative about a given unit's treatment status for known p, but perfectly determines the left-out unit's treatment status for known n_1. Instead of leave-one-out regression adjustments, for fixed n_1 I therefore show in Section 1.6 that leave-two-out regression adjustments fully characterize treatment-effect estimators with given bias.

characteristics x_i. (To be more precise, this prior will be over the joint distribution of the potential outcomes of all units given all their controls.) If the investigator aims to minimize the average mean-squared error $\mathrm{E}_\pi(\hat{\tau}_i - \tau_i)^2$, then for $d_i = 1$ she will estimate τ_i by

$$\hat{\tau}_i = \mathrm{E}_\pi[\tau_i | y_i, d_i, x_i, z_{-i}] = \underbrace{y_i(1)}_{\text{observed}} - \mathrm{E}_\pi[\overbrace{y_i(0)}^{\text{unobserved}} | y_i(1), x_i, z_{-i}].$$

This estimator represents the investigator's best guess of the treatment effect given her prior and all information in the data. In the training-program example, one specific prior could imply the use of Mincer polynomials in imputing the missing counterfactual outcome by its posterior expectation $\mathrm{E}_\pi[y_i(0)|y_i(1), x_i, z_{-i}]$.

1.2.2 Specification Searches and Optimal Restrictions on Estimation

If investigator and social preferences are misaligned, then the investigator's estimator may represent her incentives more than her expertise and the data. Even if the investigator commits to an estimator ex-ante, she could still choose one that is biased towards her preference rather than her prior. As the designer, we therefore should not only require that the investigator commits to an estimator before she has seen all of the data, but also restrict the estimators the investigator can choose from.

We face a tradeoff between flexibility and robustness. Constraints that are too permissive may lead to publication bias. One extreme solution would restrict the investigator to simple specifications that do not use control covariates, or use them only in simple linear regressions. Conventional pre-analysis plans often take this form. But restricting the investigator to a few estimators may forfeit experiment-specific knowledge about the relationship of control variables to outcomes in the prior, which I assume encodes the private information of the investigator.

I show that fixing the bias is a restriction on estimation that resolves this tradeoff. The bias of the first-best optimal estimator usually varies with the prior. Indeed, the posterior expectation of the treatment effect τ_i is usually biased towards the investigator's prior expectation $\mathrm{E}_\pi \tau_i$. But when we leave the decision over bias to the investigator, then the investigator

may shrink her estimator to her preferred estimate instead of her prior.

Once we restrict the investigator to, say, unbiased estimators of τ_i, even an investigator who wants to minimize mean-squared error relative to some fixed target $\tilde{\tau}_i$ (rather than the true treatment effect) will minimize average mean-squared error relative to the true treatment effect among unbiased estimators, since the investigator's average risk (or cost in the nomenclature of mechanism design) is then

$$\mathrm{E}_\pi(\hat{\tau}_i - \tilde{\tau}_i)^2 = \underbrace{\mathrm{E}_\pi(\hat{\tau}_i - \tau_i)^2}_{\text{social preference}} + \overbrace{\mathrm{E}_\pi(\tau_i - \tilde{\tau}_i)^2}^{\text{unaffected by investigator choice}}.$$

My first main result is that fixing the bias represents an optimal restriction in a minimax sense (Theorem 1.1) over a set of investigator preferences that generalize this risk function (Assumption 1.5). That is, the designer's average mean-squared error is minimal for an investigator that minimizes mean-squared error relative to some worst-case target, given some (hyper-)prior over the investigator's private information. Specifically, if an uninformed designer has little systematic information about the location of the treatment effect, they may want to set the bias close to zero.

1.2.3 Optimal Unbiased Estimation

Now that investigator and social preferences are aligned, how can the investigator choose an estimator with given bias and low variance? Focusing on the case of zero bias, a simple unbiased estimator of the unit-level treatment effect τ_i is available. Indeed, as e.g. noted by Athey and Imbens (2016) (where $\hat{\tau}_i$ is called the "transformed outcome"), the estimator

$$\hat{\tau}_i = \frac{d_i - p}{p(1-p)} y_i = \begin{cases} +\frac{1}{p} y_i & d_i = 1, \\ -\frac{1}{1-p} y_i & d_i = 0, \end{cases}$$

is unbiased because $\mathrm{E}[\hat{\tau}_i] = p\frac{1}{p}y_i(1) - (1-p)\frac{1}{1-p}y_i(0) = \tau_i$. But this estimator can have very high variance. Assume that job training is assigned with probability $p = .5$, and that the

potential earnings are $y_i(1) = \$1,190$ and $y_i(0) = \$1,080$. Then

$$\hat{\tau}_i = \begin{cases} +\$2,380 & d_i = 1, \\ -\$2,160 & d_i = 0, \end{cases}$$

is an unbiased, but extremely variable estimator of the treatment effect $\tau_i = \$110$. Indeed, the variance of $\hat{\tau}_i$ under treatment assignment is

$$\mathrm{Var}(\hat{\tau}_i) = p(1-p)(\hat{\tau}_i(d_i = 1) - \hat{\tau}_i(d_i = 0))^2,$$

so in the example the standard error amounts to $\sqrt{\mathrm{Var}(\hat{\tau}_i)} = \$2,270$.

We can modify this estimator by regression adjustments $\hat{y}_i$ to obtain

$$\hat{\tau}_i = \frac{d_i - p}{p(1-p)}(y_i - \hat{y}_i). \tag{1.1}$$

As long as $\hat{y}_i$ only uses information from x_i and $z_{-i} = (y_j, d_j, x_j)_{j \neq i}$, and not the outcome y_i or treatment effect d_i, $\hat{\tau}_i$ will still be unbiased. Averaging over all $\hat{\tau}_i$ and for an appropriate choice of the adjustments, Wu and Gagnon-Bartsch (2017) introduce this estimator as the "leave-one out potential outcomes" (LOOP) estimator. *My second main result* shows that all estimators of the treatment effect with a given bias can be written in this way (Lemma 1.1). Concretely, any unbiased estimator of the sample-average treatment effect is the average over estimators $\hat{\tau}_i$ for all i that each include an adjustment that uses data only from all other units. All unbiased estimators are thus equivalent to a repeated sample-splitting procedure. Conversely, if $\hat{y}_i$ is fitted, for example, by a regression of y on x that violates the sample-splitting construction by also including y_i, then overfitting of $\hat{y}_i$ to y_i would bias the treatment-effect estimate towards zero.

Among unbiased estimators, which regression adjustment minimizes variance? As Wu and Gagnon-Bartsch (2017) also note, the investigator would optimally set $\hat{y}_i$ to $(1-p)y_i(1)+py_i(0)$, since this leads to $\hat{\tau}_i = \tau_i$. But without using $y_i(1)$ or $y_i(0)$, the investigator's best choice is

the posterior expectation

$$\hat{y}_i = \mathrm{E}_\pi[(1-p)y_i(1) + py_i(0)|x_i, z_{-i}].$$

In the example, if the investigator's best guess of the expected potential earnings, $\frac{y_i(1)+y_i(0)}{2}$, based on her prior and data on all other units is $\hat{y}_i = \$1,100$, then

$$\hat{\tau}_i = \begin{cases} +2(\$1,190 - \$1,100) = \$180 & d_i = 1, \\ -2(\$1,080 - \$1,100) = \$40 & d_i = 0 \end{cases}$$

is still unbiased for $\tau_i = \$110$, but has much lower variance (the standard error is now $\sqrt{\mathrm{Var}(\hat{\tau}_i)} = \$70$). *My third main result* shows that among unbiased estimators the investigator's solution for the regression adjustments in general takes this form (Theorem 1.2), and as a corollary that all admissible (non-dominated) unbiased estimators can be achieved by exactly these regression adjustments (Theorem 1.3).

1.2.4 Machine Learning

By construction, the estimator in (1.1) of the unit-level treatment effect τ_i is unbiased whatever the regression adjustment is. In particular, the sample-splitting construction ensures that prior information only affects variance. Even a misspecified or dogmatic prior does not systematically bias what we learn about τ_i. As also used in Wager *et al.* (2016) and Wu and Gagnon-Bartsch (2017), this robust construction offers an opportunity to leverage tools that produce good predictions of potential outcomes even when they come with little guarantees that would otherwise ensure unbiasedness.

The optimal regression adjustments $\hat{y}_i = \mathrm{E}_\pi[(1-p)y_i(1) + py_i(0)|x_i, z_{-i}]$ solve an out-of-sample prediction problem. Take the special case $p = .5$.[13] Then $\hat{f}_i(x_i) = \mathrm{E}_\pi[.5y_i(1) + .5y_i(0)|x_i, z_{-i}]$ minimizes average prediction risk for the loss $(\hat{f}_i(x_i) - y_i)^2$ where $\hat{f}_i$ uses outcome and treatment data from all other units only. This is a regression problem where the quality of

[13] When treatment is not balanced, $p \neq .5$, additional weights in the prediction loss express that adjustments for the smaller group effectively get weighted up in (1.1). For details, see (1.2) in Section 1.6.

fit is measured at a new sample point, and not inside the training sample. Supervised machine-learning algorithms are built to solve exactly such out-of-sample prediction problems. For example, shrinkage methods like ridge regression of the LASSO can have better out-of-sample prediction performance than a linear least-squares regression that optimizes the in-sample fit.

I also obtain an intuitive formula for calculating standard errors. The variance of $\hat{\tau}_i$ is the expected loss in predicting the weighted potential outcome sum $(1-p)y_i(1) + py_i(0)$ by the adjustment $\hat{y}_i$, which can be estimated from the realized outcome y_i that has been excluded from the construction of $\hat{y}_i$. When units are sampled randomly, I show that, under mild conditions on the construction of regression adjustments, standard errors can be calculated from estimated prediction loss.

1.2.5 Unbiased Estimation without Pre-Specification

Regression adjustments incorporate flexibly the investigator's expertise as well as the data, but to ensure that they do not add bias, the investigator must commit to their construction in advance. Indeed, once the investigator has seen the full sample data, she cannot credibly claim that some adjustment uses data only from other units. Practically, the investigator could pre-specify a machine-learning algorithm that learns regression adjustments from the data. But that may be impractical when the construction of adjustments requires input by the researcher.

However, complete pre-specification is not necessary to ensure unbiasedness (or, more generally, a given bias). Instead the investigator could commit to splitting and distributing the sample. Assume there is a researcher in the investigator's research team that has not yet seen the data. To obtain a regression adjustment for unit i, the investigator could give that researcher access to data only from all other units. That researcher then takes the subsample, solves a prediction problem to obtain a good adjustment $\hat{y}_i$, and returns that regression adjustment to the investigator, who estimates the treatment effect according to (1.1). In that case, that researcher's choice will not introduce bias even if the researcher does not commit to the construction of the regression adjustments in advance.

Of course, estimating the average treatment effect on all n sample units in this way would require a team of n researchers. But my *fourth main result* characterizes all estimators with given bias that remain feasible without detailed pre-specification and when only K researchers are available (Corollary 1.1). Even ex-post analysis by a single researcher improves over simple pre-analysis plans without the need for detailed pre-specification. I also show that delegating estimation to two researchers approximates optimal estimation in that it ensures asymptotic efficiency under mild conditions.

1.3 Setup

Having given a simple example, I now lay out formally how I approach causal inference as a mechanism-design problem. A designer delegates the estimation of an average treatment effect in a randomized experiment to an investigator. The investigator receives a private signal about the distribution of potential outcomes, but has unknown preferences that can be biased. The designer does not analyze the dataset herself, but instead sets constraints on the investigator's estimator.

In this section, I first define the data-generating process and target parameter before introducing the investigator's and designer's problems. To simplify the further analysis, I then argue that we can restrict the analysis to direct restrictions by the designer on the space of estimators the investigator commits to.

1.3.1 Target Parameter

Following Neyman (1923), I am interested in the average treatment effect

$$\tau_\theta = \frac{1}{n}\sum_{i=1}^{n} \underbrace{(y_i(1) - y_i(0))}_{=\tau_i} \qquad\qquad \theta = (y_i(1), y_i(0))_{i=1}^n$$

in a given sample of n units. In the Rubin (1974, 1975, 1978) causal model interpretation, $y_i(d_i)$ is the potential outcome of unit i had they received treatment status $d_i \in \{0, 1\}$, and τ_i the respective causal effect.

The n units may be randomly sampled from a population distribution,

$$(y_i(1), y_i(0), x_i) \overset{\text{iid}}{\sim} \mathrm{P},$$

with pre-treatment characteristics $x_i \in \mathcal{X}$. In this case, my analysis will extend to the estimation of the population-average treatment effect $\tau = \mathrm{E}[y_i(1) - y_i(0)]$ and the conditional average treatment effect (given characteristics $x \in \mathcal{X}^n$) $\frac{1}{n}\sum_{i=1}^n \mathrm{E}[y_i(1) - y_i(0)|x_i]$. My main analysis is conditional on $(y_i(1), y_i(0), x_i)_{i=1}^n$ and therefore focuses on the sample-average treatment effect τ_θ, but I will return to τ when I discuss inference.

1.3.2 Experimental Setup

I assume that treatment is assigned randomly to overcome the missing-data problem central to causal inference (Holland, 1986). For a unit with treatment status d_i, we only observe the realized outcome $y_i = y_i(d_i)$. But because I assume that the distribution of treatment assignment $d \in \{0,1\}^n$ does not vary with the potential outcome vectors $y(1), y(0) \in \mathbb{R}^n$ (Cochran, 1972), we can estimate the treatment effect without bias. The stable-unit treatment effect assumption (Rubin, 1978) of no interference between units is implicit.

Assumption 1.1 (Random Treatment)**.** *Given potential outcomes $\theta = (y_i(1), y_i(0))_{i=1}^n$, the data $z = (y_i, d_i)_{i=1}^n$ is distributed according to P_θ as follows. d is generated from a known distribution over $\{0,1\}^n$ that does not depend on $(y(1), y(0))$ and is one of:*

1. *Each unit is independently assigned to treatment with known probability $p = \mathrm{P}(d_i = 1)$ (where $0 < p < 1$).*
2. *d is drawn uniformly at random from all assignments with known number $n_1 = \sum_{i=1}^n d_i$ of treated units (where $0 < n_1 < n$).*

Given d, $y_i = y_i(d_i)$ for all $i \in \{1, \ldots, n\}$.

In this notation, I do not explicitly include the covariates $x_1, \ldots, x_n$ in the data z, since I condition on the controls and therefore treat $(x_i)_{i=1}^n$ as a constant and not as a random variable. While neither of the distributions of d depends on the controls, my results will extend

to distributions that are known functions of x_i if they ensure identification of τ_θ. These include stratified or conditional random sampling, and sampling according to known propensity scores.

1.3.3 Covariate Adjustments

How can we estimate the sample-average treatment effect τ_θ from data $(y_i, d_i, x_i)_{i=1}^n$? Since treatment is exogenous, the average difference

$$\hat{\tau}^*(z) = \frac{1}{n_1 n_0} \sum_{d_i=1, d_j=0} (y_i - y_j) = \frac{1}{n_1} \sum_{d_i=1} y_i - \frac{1}{n_0} \sum_{d_i=0} y_i$$

between treatment and control outcomes is an unbiased estimator of τ_θ conditional on the number n_1 of treated units (provided $0 < n_1 < n$).

Of course this difference in averages $\hat{\tau}^*$ leaves information in the covariates $x_1, \ldots, x_n$ on the table and is likely inefficient. In econometric practice, τ_θ is therefore often estimated from a linear regression of the outcome on treatment and controls. But the researcher's choice of control strategy can bias published results. First, implicit model assumptions may bias estimates. Even simple linear regressions can be biased (Freedman, 2008), although this bias vanishes asymptotically if interactions are included (Lin, 2013). Second, if the investigator does not document that she picked among multiple covariate adjustments, an unsuspecting observer's inference may be biased towards stronger treatment effects and unjustified confidence (Lenz and Sahn, 2017).

1.3.4 Estimation Preferences

I explicitly consider the choice of the control specification in a mechanism-design framework. A designer and an investigator face a choice of an estimator

$$\hat{\tau} : \mathcal{Z} \to \mathbb{R}$$

that maps experimental data $z = (y, d) \in (\mathcal{Y} \times \{0, 1\})^n = \mathcal{Z}$ into an estimate $\hat{\tau}(z)$ of the sample-average treatment effect τ_θ. Since my analysis is conditional on the control covariates, this estimator encodes in particular how the estimate of the treatment effect is adjusted for

the realizations $x_1, \ldots, x_n$ of the control variables.

Designer and investigator preferences are expressed by risk functions $r^D, r^I : \Theta \times \mathbb{R}^{\mathcal{Z}} \to \mathbb{R}$ that encode the expected loss $r^D_\theta(\hat{\tau}), r^I_\theta(\hat{\tau})$ of an estimator $\hat{\tau} \in \mathbb{R}^{\mathcal{Z}}$ given the full matrix $\theta = (y(1), y(0)) \in \mathcal{Y}^{2n} = \Theta$ of $2n$ potential outcomes in the sample at hand. Both designer and investigator aim to minimize their respective risk given the potential outcomes θ. Throughout this chapter, I specifically assume that the designer's risk function expresses a social desire to obtain precise estimates of the true treatment effect τ_θ.

Assumption 1.2 (Social risk function)**.** *The designer's risk for an estimator $\hat{\tau} : \mathcal{Z} \to \mathbb{R}$ is the estimator's mean-squared error*

$$r^D_\theta(\hat{\tau}) = \mathrm{E}_\theta[(\hat{\tau}(z) - \tau_\theta)^2],$$

where the expectation averages over random treatment assignment given potential outcomes $\theta \in \Theta$.

Notably, I do not assume that the designer has an inherent preference for unbiased estimators.[14] While my characterization results will depend on this specific form of the social risk function, the general mechanism-design approach extends to alternative risk (or equivalently utility) functions.

The investigator's risk function can differ from the designer's risk function. For example, I will later consider risk functions that include $r^I_\theta(\hat{\tau}) = \mathrm{E}_\theta[(\hat{\tau}(z) - \tilde{\tau})^2]$, which expresses a desire to obtain a certain estimate $\tilde{\tau}$ irrespective of the true treatment effect τ_θ. The designer knows only that $r^I \in \mathcal{R}$ for some set of risk functions.

1.3.5 Prior Information

Since generally no single estimator $\hat{\tau}$ minimizes risk for all potential outcomes $\theta \in \Theta$ and θ is not known, a good estimator has to trade off risk performance across different draws of

[14] Still, the minimization of squared-error loss is associated with unbiasedness, as e.g. in Lehmann and Romano (2006, Example 1.5.6).

potential outcomes. Following Wald (1950), I assume that a prior distribution π over potential outcomes governs this tradeoff.[15]

The investigator receives the prior distribution π over potential outcomes θ as a private signal before the data z is realized. This private information models researcher expertise. For example, the investigator may have run previous studies or a pilot and synthesize relevant results in the literature. The investigator therefore has a sense which variables are important and which regression specifications are more likely to work well.

The uninformed designer does not observe the prior π, but only has a diffuse (hyper-)prior η for π. The designer therefore designs a mechanism that elicits the investigator's prior information. Optimally, the designer would want to obtain an estimator that minimizes average mean-squared error given the investigator's private prior, but since the investigator's preferences may differ from the designer's, the latter cannot generally achieve a first-best estimator.

1.3.6 Mechanism Structure and Timeline

I assume that the designer has the authority to set rules in the form of a mechanism without transfers. The designer cannot verify the investigator's risk type or private prior information. The investigator follows whatever mapping from investigator decisions to final estimator the designer sets, and the designer follows through on the mapping she commits to. Similar to Frankel's (2014) delegation setup, the game between designer and investigator plays out in the following steps:

1. The designer chooses a mechanism that consists of a message space M and a mapping from messages m into estimators $\hat{\tau}_m : \mathcal{Z} \to \mathbb{R}$.

2. The investigator observes the prior distribution π and sends a message $m(r^I, \pi)$.

[15] One alternative approach to finding a good estimator would involve putting restrictions on the distribution of potential outcomes and discussing efficient estimators under some large-sample approximation. But since researchers may reasonably disagree about these choices, this would itself add an additional degree of freedom to estimation. I instead consider estimation in an exact finite-sample decision-theoretic framework that does not restrict the distribution of potential outcomes.

3. The potential outcomes θ are realized, the data z drawn according to the experiment, and the estimate $\hat{\tau}_{m(r^I,\pi)}(z)$ formed.

In econometric terms, I think of the investigator's message as a modelling decision. The designer then restricts the space of models the investigator can choose from.

For simplicity, I assume that the investigator's message given her risk type and private information and the mapping of her message to the final estimator are deterministic, but the setup extends to stochastic actions as in Frankel (2014). By the revelation principle, the specific form of the mechanism is not a substantial restriction, since it includes direct mechanisms in which the investigator reveals her risk type and her private information (as e.g. in Holmström, 1984).

Since the investigator controls the estimator with her choice of message, we can assume without loss of generality that the message space is a set of estimators (and the mapping from message to estimator the identity). Indeed, take any estimator that is an outcome for some message. Since neither risk type nor prior are verifiable, the investigator can always choose that message to obtain said estimator.

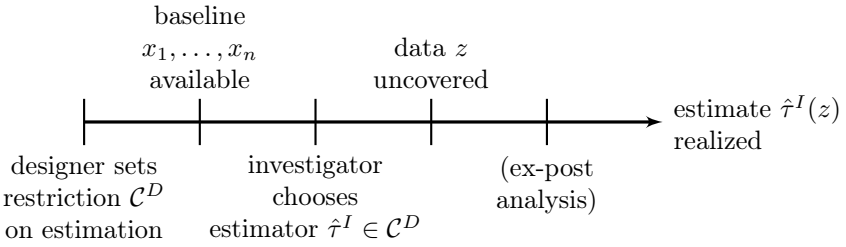

Figure 1.1: *Estimation timeline*

Hence, the designer directly restricts estimators to some set $\mathcal{C}^D$. Subject to the constraint, the investigator specifies an estimator $\hat{\tau}^I \in \mathcal{C}^D$ before data becomes available. Once the data $z \in \mathcal{Z}$ is realized, the investigator reports the estimate $\hat{\tau}^I(z)$ (Figure 1.1). Since my econometric analysis is conditional on the control variables $x_1, \ldots, x_n$, this baseline information can be available to the investigator and inform her choice of estimator.

Optimal estimation in this framework will require some degree of commitment by the

investigator before the data is available. Otherwise, any restriction on estimation would be cheap talk, since the investigator could choose an estimator ex post that justifies their preferred estimate at the realized data. But I will show that optimal commitment is less constraining than restricting the investigator to pre-analysis plans with simple specifications that are chosen ex ante. First, the investigator's estimator can still contain (automated) specification searches. Second, in Section 1.7, I show that it is not generally necessary to specify the full estimator ex ante, and that additional exploratory analysis after the data has become available can improve estimation.

1.3.7 Investigator and Designer Choices

Having set up the actions available to the investigator and designer, I now describe their preferences. The investigator chooses an estimator to minimize average risk subject to her prior.

Assumption 1.3 (Investigator's choice)**.** *Given the prior distribution π over potential outcomes $\theta \in \Theta$, the investigator minimizes average risk subject to the constraint $\mathcal{C}^D \subseteq \mathbb{R}^{\mathcal{Z}}$ set by the designer,*

$$\hat{\tau}^I = \hat{\tau}^I(\mathcal{C}^D, \pi) \in \arg\min_{\hat{\tau} \in \mathcal{C}^D} \mathrm{E}_\pi[r^I_\theta(\hat{\tau})].$$

The designer does not know the risk function of the investigator, but only assumes that it falls within some set $\mathcal{R}$ of risk functions. Adapting the maxmin criterion from the mechanism-design literature (e.g. Hurwicz and Shapiro, 1978; Frankel, 2014; Carroll, 2015), I assume that the designer chooses a constraint that minimizes average risk at a worst-case investigator type.

Definition 1.1 (Designer's minimax delegation problem)**.** *Given some set $\mathcal{R}$ of investigator risk functions, the designer picks a constraint $\mathcal{C}^D \subseteq \mathbb{R}^{\mathcal{Z}}$ to minimize average mean-squared error,*

$$\mathcal{C}^D = \mathcal{C}^D(\mathcal{R}, \eta) \in \min_{\mathcal{C} \subseteq \mathbb{R}^{\mathcal{Z}}} \sup_{r^I \in \mathcal{R}} \mathrm{E}_\eta[r^D_\theta(\hat{\tau}^I)],$$

where I assume that the investigator breaks ties in the designer's favor.

The minimax criterion can be seen as a game between designer and nature. For every choice of restriction that the designer picks, nature responds with an investigator who produces maximal average mean-squared error. In this game, the designer picks a constraint that ensures that the average risk at a worst-case outcome is minimal.

Without constraints, the investigator's estimator may be a poor fit from the designer's perspective. But if the constraints are too restrictive, for example if we reduce the allowed set of estimators to the difference in averages $\hat{\tau}^*$, we will use the investigator's expertise inefficiently. I therefore solve for constraints $\mathcal{C}^D$ that resolve this tradeoff between flexibility and robustness optimally.

1.3.8 Support Restriction

Throughout this chapter, I assume that the support of (potential) outcomes is finite, for three reasons. First, I adapt results from the mechanism-design literature that involve finite sums. Second, I use and provide complete-class theorems that fully characterize admissible (non-dominated) estimators provided their support is finite. Third, I derive intuitive combinatorial proofs for my characterization results.

Assumption 1.4 (Finite support)**.** *The support $\mathcal{Y}$ of potential outcomes $y_i(1), y_i(0)$ is finite.*

Since the number of support points is otherwise unrestricted, the finite-support assumption allows for flexible approximations to arbitrary distributions.

1.4 Overview of Main Results

In this section, I preview my main theoretical results. Under specific restrictions on investigator preferences, I show that fixing the bias is a minimax optimal constraint on estimation. I then present a representation of treatment-effect estimators with given bias, characterize the investigator's optimal choice from this restricted class for the case of unbiased estimators, and extend the analysis to estimators with limited pre-specification.

I assume that investigator risk functions express mean-squared error relative to some target which may not be the true treatment effect.

Assumption 1.5 (Investigator risk restriction)**.** *The investigator has a risk function from the set*

$$\mathcal{R}^* = \{r^I; r^I_\theta(\hat{\tau}) = \mathrm{E}_\theta[(\hat{\tau}(z) - \tilde{\tau}_\theta)^2] \text{ for some } \tilde{\tau} : \Theta \to \mathbb{R}\}.$$

The target function $\tilde{\tau}_\theta$ is unrestricted in this definition. For example, the investigator may want to achieve a constant target no matter what the true potential outcomes are ($\tilde{\tau}_\theta = \text{const.}$). Or the investigator may prefer to obtain estimates above the true treatment effect ($\tilde{\tau}_\theta = \tau_\theta + \varepsilon$).

In any of these cases, restricting investigators to unbiased estimators (or more generally, estimators with given bias, $\mathrm{E}_\theta[\hat{\tau}(z)] = \tau_\theta + \beta_\theta$) ensures that they choose among these estimators as if they had the designer's preference, i.e. they minimize average variance. Once I have established tools for asymptotically valid inference, it will also follow that unbiasedness aligns the choices of investigators who want to obtain a small standard error or a low p-value.

While fixing the bias aligns preferences, this restriction may be too strong. However, I establish that it is minimax optimal for an appropriate choice of biases.

Theorem 1.1 (Fixed bias is minimax optimal)**.** *Write $\Delta^*(\Theta)$ for all distributions over Θ with full support. For every hyperprior η with support within $\Delta^*(\Theta)$ there is a set of biases $\beta^\eta : \Theta \to \mathbb{R}$ such that the fixed-bias restriction*

$$\mathcal{C}^\eta = \{\hat{\tau} : \mathcal{Z} \to \mathbb{R}; \mathrm{E}_\theta[\hat{\tau}] = \tau_\theta + \beta^\eta_\theta\}$$

is a minimax optimal mechanism in the sense of Definition 1.1, i.e.

$$\mathcal{C}^\eta \in \arg\min_{\mathcal{C}} \sup_{r^I \in \mathcal{R}^*} \mathrm{E}_\eta \left[r^D_\theta \left(\arg\min_{\hat{\tau} \in \mathcal{C}} \mathrm{E}_\pi[r^I_\theta(\hat{\tau})] \right) \right].$$

This result implies that the designer should not leave the choice of bias to the investigator. If the designer has an informative hyperprior, she may set biases to reflect that information. But with little information on the designer's side, the designer may want to set them close to zero. I discuss setting the bias in Appendix A.6.

With a restriction to given bias (in the sense of an ex-ante fixed vector of biases), the investigator chooses the estimators to minimize variance. The next result specifically characterizes unbiased estimators, and therefore the choice set of the investigator if the designer sets the bias to zero.

Lemma 1.1 (Representation of unbiased estimators)**.** *The estimator $\hat{\tau}$ is unbiased, $\mathrm{E}_\theta[\hat{\tau}(z)] = \tau_\theta$ for all potential outcomes $\theta \in \Theta$, if and only if:*

1. *For a known treatment probability p, there exist leave-one-out regression adjustments $(\phi_i : (\mathcal{Y} \times \{0,1\})^{n-1} \to \mathbb{R})_{i=1}^n$ such that*

$$\hat{\tau}(z) = \frac{1}{n} \sum_{i=1}^{n} \frac{d_i - p}{p(1-p)} (y_i - \phi_i(z_{-i})).$$

2. *For a fixed number n_1 of treated units, there exist leave-two-out regression adjustments $(\phi_{ij} : (\mathcal{Y} \times \{0,1\})^{n-2} \to \mathbb{R})_{i<j}$ such that*

$$\hat{\tau}(z) = \frac{1}{n_1 n_0} \sum_{i<j} (d_i - d_j)(y_i - y_j - \phi_{ij}(z_{-ij})),$$

where $\phi_{ij}(z_{-ij})$ may be undefined outside $\mathbf{1}' d_{-ij} = n_1 - 1$.

For the case of known probability p, any unbiased estimator can therefore be written in the leave-one-out form that Wu and Gagnon-Bartsch (2017) obtain as a special case of the unbiased estimators introduced by Aronow and Middleton (2013).

The result directly extends to a characterization of estimators with fixed bias. Indeed, fixing the bias is equivalent to the designer choosing an estimator $\hat{\tau}^D$ with the desired biases $\mathrm{E}_\theta[\hat{\tau}^D(z)] - \tau_\theta = \beta_\theta$ for all $\theta \in \Theta$, and letting the investigator choose a zero-expectation adjustments $\hat{\delta}^I$ ($\mathrm{E}_\theta[\hat{\delta}^I(z)] = 0$ for all $\theta \in \Theta$) to form the estimator $\hat{\tau} = \hat{\tau}^D + \hat{\delta}^I$. Given $\hat{\tau}^D$, any estimator with the associated bias profile can thus be written as

$$\hat{\tau}^D(z) - \frac{1}{n} \frac{d_i - p}{p(1-p)} \sum_{i=1}^{n} \phi_i(z_{-i}), \qquad \hat{\tau}^D(z) - \frac{1}{n_1 n_0} \sum_{i<j} (d_i - d_j) \phi_{ij}(z_{-ij}),$$

respectively, with adjustments as in the lemma. The statement of the lemma corresponds to

the unbiased choices

$$\hat{\tau}^D(z) = \frac{1}{n}\sum_{i=1}^{n}\frac{d_i - p}{p(1-p)}y_i, \qquad \hat{\tau}^D(z) = \frac{1}{n_1 n_0}\sum_{i<j}(d_i - d_j)(y_i - y_j).$$

All estimators with given bias are hence sample-splitting estimators that leave one or two units out, respectively, when calculating their regression adjustments. But when is an estimator not just of this form, but also precise? As a general solution, the investigator would now pick one set of adjustments that minimize variance averaged over their prior, yielding constrained optimal solution from the perspective of the designer.

For the specific case of zero-bias estimators, the adjustments take a particularly simple form. The investigator would optimally want to set regression adjustments to the oracle solutions

$$\begin{aligned}\bar{y}_i &= (1-p)y_i(1) + py_i(0),\\ \Delta\bar{y}_{ij} &= \left(\frac{n_0}{n}y_i(1) + \frac{n_1}{n}y_i(0)\right) - \left(\frac{n_0}{n}y_j(1) + \frac{n_1}{n}y_j(0)\right),\end{aligned}$$

respectively, but since the potential outcomes are unknown, these adjustments are infeasible. Instead, I show that the investigator chooses leave-one-out or leave-two-out expectations of these adjustments.

Theorem 1.2 (Choice of the investigator from unbiased estimators)**.** *An investigator with risk $r \in \mathcal{R}^*$ and prior π over Θ chooses the following unbiased Bayes estimators:*

1. *For a known treatment probability p,*

$$\hat{\tau}(z) = \frac{1}{n}\sum_{i=1}^{n}\frac{d_i - p}{p(1-p)}(y_i - \mathrm{E}_\pi[\bar{y}_i|z_{-i}]).$$

2. *For a fixed number n_1 of treated units,*

$$\hat{\tau}(z) = \frac{1}{n_1 n_0}\sum_{i<j}(d_i - d_j)(y_i - y_j - \mathrm{E}_\pi[\Delta\bar{y}_{ij}|z_{-ij}]).$$

Hence, all optimal unbiased estimators take as regression adjustments conditional expectations of potential outcomes. These conditional expectations can be obtained as solutions to

a prediction problem.[16] Independently of the mechanism-design setup, the set of investigator solutions across different priors completely characterize the class of admissible unbiased estimators of the sample-average treatment effect.

Theorem 1.3 (Complete-class theorem for unbiased estimators)**.** *For any unbiased estimator $\hat{\tau}$ of the sample-average treatment effect that is not dominated with respect to variance, there is a converging sequence of priors $(\pi_t)_{t=1}^{\infty}$ with full support such that $\hat{\tau}$ equals the limit of the respective estimators in Theorem 1.2. Conversely, for any converging sequence of priors $(\pi_t)_{t=1}^{\infty}$ that put positive weight on every state $\theta \in \Theta$, every converging subsequence of corresponding estimators is admissible among unbiased estimators.*

Now that I have characterized the optimal solution of the designer and the investigator (with an explicit expression for the case of unbiased estimators), I return to the question of commitment. The representation of fixed-bias estimators in Lemma 1.1 requires that the construction of regression adjustments does not involve the adjusted unit. In Theorem 1.2, the investigator would therefore have to commit to their construction before she has access to the full sample. This pre-specification leaves room for automated specification searches in constructing the adjustments. But fully pre-specifying all specification searches may be impractical.

I also characterize estimators that ensure fixed bias not by the investigator fully pre-specifying adjustments, but by a commitment to a sample-splitting scheme. I consider estimation contracts that have the investigator delegate estimation tasks on subsamples to K researchers who do not share information about the data they receive.

Definition 1.2 (K-distribution contract)**.** *A K-distribution contract $\hat{\tau}^{\Phi}$ distributes data $z = (y, d) \in (\mathcal{Y} \times \{0, 1\})^n = \mathcal{Z}$ to K researchers. Researcher k receives data $g_k(z) \in A_k$ and returns the intermediate output $\hat{\phi}_k(g_k(z)) \in B_k$. The estimate is*

$$\hat{\tau}^{\Phi}((\hat{\phi}_k)_{k=1}^{K}; z) = \Phi((\hat{\phi}_k(g_k(z))_{k=1}^{K}; z).$$

[16] For known p, this exact solution mirrors Wu and Gagnon-Bartsch's (2017) LOOP estimator, for which the authors discuss estimating the adjustments using different prediction methods.

The investigator chooses the functions g_k (from data in $\mathcal{Z}$ to researcher input in A_k) and Φ (from the researcher outputs in $\bigtimes_{k=1}^K B_k$ and data in $\mathcal{Z}$ to estimates in $\mathbb{R}$) before accessing the data.

As one special case of my general representation result of fixed-bias K-distribution contracts, I characterize estimators with given bias that divide the sample into K folds and then give each researcher access to all but one of these folds. In that case, I deduce from the representation of unbiased estimators in Lemma 1.1 that the estimator always has the given bias if and only if each researcher only controls the regression adjustments for the respective left-out fold.

Corollary 1.1 (Characterization of fixed-bias K-fold distribution contracts)**.** *For K disjoint folds $\mathcal{I}_k \subseteq \{1, \dots, n\}$ with projections $g_k : (y, d) = z \mapsto z_{-\mathcal{I}_k} = (y_i, d_i)_{i \neq \mathcal{I}_k}$, a K-distribution contract $\hat{\tau}^\Phi$ has given bias if and only if:*

1. *For a known treatment probability p, there exist a fixed estimator $\hat{\tau}_0(z)$ with the given bias and regression adjustment mappings $(\Phi_k)_{k=1}^K$ such that*

$$\hat{\tau}^\Phi((\hat{\phi}_k)_{k=1}^K; z) = \hat{\tau}_0(z) - \frac{1}{n} \sum_{k=1}^K \sum_{i \in \mathcal{I}_k} \frac{d_i - p}{p(1-p)} \phi_i^k(z_{-i})$$

   where $(\phi_i^k)_{i \in \mathcal{I}_k} = \Phi_k(\hat{\phi}_k(z_{-\mathcal{I}_k}))$.

2. *For a fixed number n_1 of treated units, there exist a fixed estimator $\hat{\tau}_0(z)$ with the given bias and regression adjustment mappings $(\Phi_k)_{k=1}^K$ such that*

$$\hat{\tau}^\Phi((\hat{\phi}_k)_{k=1}^K; z) = \hat{\tau}_0(z) - \frac{1}{n_1 n_0} \sum_{k=1}^K \sum_{\{i<j\} \subseteq \mathcal{I}_k} (d_i - d_j) \phi_{ij}^k(z_{-ij}),$$

   where $(\phi_i^k)_{i \in \mathcal{I}_k} = \Phi_k(\hat{\phi}_k(z_{-\mathcal{I}_k}))$.

These sample-distribution contracts achieve the given bias without detailed commitments by the researchers.

1.5 Designer's Solution

Having set up the estimation of a sample-average treatment effect as a mechanism-design problem, I justify a restriction to estimators with fixed bias by solving the designer's delegation problem. Subject to fixed bias, the investigator pre-specifies an estimator according to the designer's preferences. I prove minimax optimality of fixed-bias restrictions, echoing a result from mechanism design on optimal delegation.

1.5.1 The Role of Bias

When there is no misalignment of preferences, then the resulting first-best estimator that minimizes average mean-squared error will generally have bias that changes with prior. To understand how being flexible on bias can improve estimation, note that both bias and variance contribute to the risk

$$r_\theta^D(\hat{\tau}) = \mathrm{E}_\theta[(\hat{\tau} - \tau)^2] = \underbrace{(\mathrm{E}_\theta[\hat{\tau}] - \tau)^2}_{\text{bias}} + \underbrace{\mathrm{Var}_\theta(\hat{\tau})}_{\text{variance}}$$

the designer aims to minimize. We can often improve an estimator with fixed bias by moving along this bias–variance tradeoff. Indeed, consider the first-best solution $\hat{\tau}_\pi = \arg\min_{\hat{\tau}} \mathrm{E}_\pi[r_\theta^D(\hat{\tau})]$ of the designer. The estimate $\hat{\tau}_\pi(z) = \mathrm{E}_\pi[\tau_\theta|z]$ comprises the posterior expectations $\mathrm{E}_\pi[y_i(1) - y_i(0)|z]$, which are usually biased towards the prior expectation of unit treatment effects when the prior is informative along this dimension.

But if the designer leaves the decision over bias to the investigator, then an investigator who has biased preferences will be inclined to bias the estimator in the direction of her preferences, not of her prior. Consider an investigator with risk

$$r_\theta^I(\hat{\tau}) = \mathrm{E}_\theta[(\hat{\tau}(z) - (\tau_\theta + \varepsilon))^2] \qquad (\varepsilon > 0)$$

who would like to show that the treatment effect is higher than it is. The investigator's unconstrained solution is now shifted upward by ε, which is added to the bias term. While reducing the variance relative an unbiased estimator, the designer's risk may also be increased through additional bias.

For choices among estimators with fixed bias, however, the investigator's and designer's preferences in this example are perfectly aligned. With bias fixed at zero, say, mean-squared error is variance, $r_\theta^D(\hat{\tau}) = \mathrm{Var}_\theta(\hat{\tau})$. The ε-biased investigator's risk is $r_\theta^I(\hat{\tau}) = \varepsilon^2 + \mathrm{Var}_\theta(\hat{\tau})$. While risks are not the same, they are shifted by a constant. There is no distortion in choices between estimators with fixed bias for this investigator loss function.

1.5.2 Fixed-Bias Estimation as Second-Best

Having motivated in an example that fixing the bias can align investigator choices, I extend alignment to a minimax result. If the investigator has constant bias, I have argued that among estimators with fixed bias she will still commit to a variance-minimizing estimator. To show that this example extends to an optimal solution, I have to establish that the bias restriction is neither too permissive nor too restrictive.

A restriction that fixes the bias, for example to zero,

$$\mathcal{C}^* = \{\hat{\tau} : \mathcal{Z} \to \mathbb{R}; \mathrm{E}_\theta[\hat{\tau}] = \tau_\theta \forall \theta \in \Theta\},$$

is not too permissive provided that investigators all choose as if they minimized mean-squared error relative to *some* target, albeit not necessarily relative to the true treatment effect.

Assumption 1.5 (Investigator risk restriction)**.** *The investigator has a risk function from the set*

$$\mathcal{R}^* = \{r^I; r_\theta^I(\hat{\tau}) = \mathrm{E}_\theta[(\hat{\tau}(z) - \tilde{\tau}_\theta)^2] \text{ for some } \tilde{\tau} : \Theta \to \mathbb{R}\}.$$

The target $\tilde{\tau}_\theta$ can vary arbitrarily with the potential outcomes. In particular, permissible risk functions include constant biases relative to the truth ($\tilde{\tau} = \tau + \varepsilon$) or fixed estimation targets ($\tilde{\tau} = \text{const.}$). $\mathcal{R}^*$ also includes the designer's risk function r^D at $\tilde{\tau} = \tau$.

Lemma 1.2 (Unbiasedness aligns estimation)**.** *If the investigator has risk from $\mathcal{R}^*$ then the investigator will choose from the unbiased estimators $\mathcal{C}^*$ according to the designer's preferences.*

Note that the result extends to restrictions to fixed bias (that can vary with θ).

Once I have established asymptotically valid inference for unbiased estimators in Section A.5, I will also show in Remark A.5 that the unbiasedness restriction aligns the choices of investigators who want to obtain small standard errors or tight confidence intervals. For a local-to-null alternative, by Remark A.6 unbiasedness also insures asymptotic alignment in large samples when the investigator wants to obtain a low p-value (that is, wants to maximize the power of a test against some null hypothesis $\tau_\theta = \tau_0$).

Note, however, that there are many risk (or equivalently utility) functions for which fixing the bias does not provide alignment. In particular, it may be a poor alignment device for non-convex loss functions. Take an investigator who wants to produce an estimate that does *not* reject some null hypothesis, for example when running a balance or robustness check. In that case, if some valid way of calculating standard errors is available, the investigator would want to obtain high variance even among unbiased estimators in order to weaken the evidence against her preferred null hypothesis.

For the class $\mathcal{R}^*$ of investigator risk functions, fixing the bias is not too restrictive because it is minimax optimal over investigator preferences. While Lemma 1.2 establishes that choices from unbiased estimators will be the same for any $r^I \in \mathcal{R}^*$, there could be a larger set of estimators that provide alignment, or full alignment of preferences could be too costly.

Theorem 1.1 (Fixed bias is minimax optimal)**.** *Write $\Delta^*(\Theta)$ for all distributions over Θ with full support. For every hyperprior η with support within $\Delta^*(\Theta)$ there is a set of biases $\beta^\eta : \Theta \to \mathbb{R}$ such that the fixed-bias restriction*

$$\mathcal{C}^\eta = \{\hat{\tau} : \mathcal{Z} \to \mathbb{R}; \mathrm{E}_\theta[\hat{\tau}] = \tau_\theta + \beta^\eta_\theta\}$$

is a minimax optimal mechanism in the sense of Definition 1.1, i.e.

$$\mathcal{C}^\eta \in \arg\min_{\mathcal{C}} \sup_{r^I \in \mathcal{R}^*} \mathrm{E}_\eta \left[r^D_\theta \left(\arg\min_{\hat{\tau} \in \mathcal{C}} \mathrm{E}_\pi[r^I_\theta(\hat{\tau})] \right) \right].$$

This minimax result shows that the gains from variance reduction of being flexible on bias are fully undone by the cost of misalignment for a worst-case risk function, for any relaxation of the fixed-bias restriction. Once we allow the bias to track the prior, it could as well reflect

the preference of a worst-case investigator. The designer therefore chooses fixed biases that reflect her hyperprior.

If the designer has a hyperprior η that is quite informative about treatment effects, she could introduce biases towards expected treatment effects under that hyperprior. Crucially, however, these biases would be fixed ex-ante and not chosen by the investigator. But when the hyperprior contains little systematic information about the treatment effect at θ, then β_θ close to zero is a natural choice. In Appendix A.6, I highlight one construction that shows how an (approximately) uninformative hyperprior delivers (approximately) unbiased estimation as the support grows. There, I also lay out how being minimax over a specific, uninformative class of hyperpriors yields zero bias.

1.5.3 Connection to Aligned Delegation

My econometric finding that fixed-biased estimation is minimax optimal (Theorem 1.1) builds upon a mechanism-design result by Frankel (2014). There, a principal delegates decisions to an agent who observes states. Frankel (2014) characterizes optimal delegation mechanisms without transfers. In a class of maxmin optimal, simple mechanisms, the agent behaves according to the principal's preferences.

In a leading example from Frankel (2014), a school principal delegates the grading of a group of students to a teacher. The teacher may prefer to give more skewed or better grades than the principal, who does not observe the students' performance. However, the principal can exploit that the teacher's biased preferences are consistent across students. If the teacher and the principal agree on the ranking of students, fixing the distribution of grades obtains a second-best grade assignment. If the teacher has a constant bias, fixing the average grade already achieves agreement between principal and teacher. In both cases, the teacher chooses from the restricted grade assignments according to the principal's preferences.

What a fixed average is to grading in Frankel (2014), constant bias is to estimation in my setting. More precisely, I identify Frankel's (2014) school principal with my designer, the teacher with the investigator, and individual students with different draws of the data. In the

school example, the performance of students is the private information of the teacher. For estimation, the prior distribution over potential outcomes is the private information of the investigator. Where the teacher chooses a grade for each student, the investigator commits to an estimator, that is, the investigator chooses an estimate for each (potential) draw of the data.

Frankel (2014) shows that fixing the average over grades is a maxmin (in utility terms) optimal mechanism for a class of biased squared-error preferences. Analogously, my fixed-bias restriction fixes weighted sums over estimates. But since fixing the bias requires setting many sums at once, and the designer's and investigator's preferences involve weights determined by the prior, additional work is required to establish the minimax optimality in Theorem 1.1. In Section A.1, I show how Frankel's (2014) result carries over to the designer's problem across all $\theta \in \Theta$, where the investigator sets all $(2|\mathcal{Y}|)^n$ values of $\hat{\tau}(y, d)$ simultaneously.

1.5.4 Design of Experiment vs. Design of Estimator

In Theorem 1.1, I have assumed that treatment is assigned randomly according to some fixed rule, but my results extend to the design of treatment assignment itself. The investigator may leverage prior knowledge about potential outcomes to adjust propensity scores (Kasy, 2016). For example, if the prior distribution of treated outcomes has larger variance than that of controls, the investigator may want to assign more units to treatment. Under the fixed-bias restriction, the investigator's preference over this additional decision remains aligned with the goal of the designer.

For $K = 1$, I show that giving one researcher (with risk function in the set $\mathcal{R}^*$) access to part of the sample for exploratory ex-post analysis can improve over simple pre-analysis plans. For $K = 2$, I show that a flexible, unbiased pre-analysis plan that specifies distribution to two researchers asymptotically achieves semi-parametric efficiency when the units are sampled iid under conditions on the population distribution.

1.6 Investigator's Solution

The designer restricts the investigator to estimators with given bias. I establish that this restriction is equivalent to splitting the sample in a particular way. In solving the investigator's constrained optimization problem in the specific case of zero bias, I show that optimal unbiased estimation is equivalent to a set of out-of-sample prediction tasks. I obtain a complete-class theorem that characterizes admissible unbiased estimators of the sample-average treatment effect.

Throughout this section, I assume that the investigator fully specifies her estimator before it is applied to outcome and treatment data $z = (y, d)$. Although the estimator is pre-specified, it can still include (automated) specification searches. The pre-specified estimator thus plays the role of a flexible pre-analysis plan. Since my results hold conditional on potential outcomes, the covariates $x_1, \ldots, x_n$ can be common knowledge before this pre-analysis plan is filed. In Section 1.7, I show how the results in this section extend when full pre-specification is impractical. There, I provide a constructive characterization of pre-analysis plans that only commit to the way the sample is split and distributed.

1.6.1 Characterization of Fixed-Bias Estimators

When does an estimator have a given bias, conditional on potential outcomes? The designer requires that the investigator provides a fixed-bias estimator. In this section, I provide an intuitive representation of estimators of a given bias that the investigator can achieve transparently by construction.

For the case of zero bias, a class of estimators that ensures unbiasedness is obtained by sample splitting. For known treatment probability p, the Horvitz and Thompson (1952) estimator $\hat{\tau}^{\mathrm{HT}} = \frac{1}{n}\sum_{i=1}^{n} \frac{d_i - p}{p(1-p)} y_i$ is unbiased for any pair of potential outcome vectors because

$$\mathrm{E}_\theta\left[\frac{d_i - p}{p(1-p)} y_i\right] = y_i(1) - y_i(0).$$

If we replace outcomes y_i by adjusted outcomes $y_i - \phi_i(z_{-i})$ with regression adjustments that do not vary with (y_i, d_i), where z_{-i} denotes the data $(y_j, d_j)_{j \neq i}$ from all units other

than i, then the resulting estimator is still unbiased. (Recall that I condition on controls $x_1, \ldots, x_n$ throughout.) Wu and Gagnon-Bartsch (2017) call the resulting estimator for known p the "leave-one-out potential outcomes" (LOOP) estimator. This estimator is a special case of Aronow and Middleton's (2013) modification of the Horvitz and Thompson (1952) estimator. Since the adjustment $\phi_i(z_{-i})$ is the same whether unit i is treated or not and $\mathrm{E}_\theta\left[\frac{d_i - p}{p(1-p)}\middle| z_{-i}\right] = 0$, their addition averages out to zero, no matter the potential outcomes or realized treatment of the other units.[17]

I show that these sample-splitting estimators are also all estimators that are unbiased conditional on potential outcomes. If an estimator cannot be written as a Horvitz and Thompson (1952) estimator with leave-one-out regression adjustments (i.e. in the form of Wu and Gagnon-Bartsch's (2017) LOOP estimator), it must have bias for some matrix of potential outcomes. If instead we considered estimators that are unbiased given some distribution of potential outcomes (for example, we may want to model noise terms in potential outcomes that we do not want to condition on), then the result would trivially extend as long as we do not restrict this distribution. If an estimator cannot be written in this leave-one-out form, it must have bias for some distribution of potential outcomes.

A leave-one-out estimator can have bias conditional on the number of treated units. If the number n_1 of treated units is known, the leave-one-out adjustment $\phi_i(z_{-i})$ implicitly depends on $d_i = n_1 - \sum_{j \neq i} d_j$. For permutation randomization, I therefore start with the difference in averages $\hat{\tau}^* = \frac{1}{n_1 n_0} \sum_{d_i=1, d_j=0} (y_i - y_j)$ and establish that all unbiased estimators differ from $\hat{\tau}^*$ only by leave-two-out regression adjustments $\phi_{ij}(z_{-ij})$. In every sample split, these unbiased estimators leave out one treated and one untreated unit.[18]

Lemma 1.1 (Representation of unbiased estimators)**.** *The estimator $\hat{\tau}$ is unbiased, $\mathrm{E}_\theta[\hat{\tau}(z)] = \tau_\theta$ for all potential outcomes $\theta \in \Theta$, if and only if:*

1. For a known treatment probability p, there exist leave-one-out regression adjustments

[17] It would not be enough to exclude the treatment status d_i from the constriction of unit i's regression adjustment, and thus use y_i, since y_i can be correlated with d_i.

[18] Wager *et al.* (2016) consider leave-one-out estimators separately in the treatment and control groups, and use a leave-two-out construction to derive asymptotic unbiasedness.

$(\phi_i : (\mathcal{Y} \times \{0,1\})^{n-1} \to \mathbb{R})_{i=1}^n$ *such that*

$$\hat{\tau}(z) = \frac{1}{n} \sum_{i=1}^{n} \frac{d_i - p}{p(1-p)} (y_i - \phi_i(z_{-i})).$$

2. *For a fixed number n_1 of treated units, there exist leave-two-out regression adjustments $(\phi_{ij} : (\mathcal{Y} \times \{0,1\})^{n-2} \to \mathbb{R})_{i<j}$ such that*

$$\hat{\tau}(z) = \frac{1}{n_1 n_0} \sum_{i<j} (d_i - d_j)(y_i - y_j - \phi_{ij}(z_{-ij})),$$

where $\phi_{ij}(z_{-ij})$ may be undefined outside $\mathbf{1}' d_{-ij} = n_1 - 1$.

While I have derived these estimators from unbiased estimators, the characterization directly carries over to estimators with fixed bias. Indeed, fixing the bias is equivalent to the designer choosing an estimator $\hat{\tau}^D$ with the desired biases $\mathrm{E}_\theta[\hat{\tau}^D(z)] - \tau_\theta = \beta_\theta$ for all $\theta \in \Theta$, and letting the investigator choose a zero-expectation adjustments $\hat{\delta}^I$ ($\mathrm{E}_\theta[\hat{\delta}^I(z)] = 0$ for all $\theta \in \Theta$) to form the estimator $\hat{\tau} = \hat{\tau}^D + \hat{\delta}^I$. Given $\hat{\tau}^D$, any estimator with the associated bias profile can thus be written as

$$\hat{\tau}^D(z) - \frac{1}{n} \sum_{i=1}^{n} \frac{d_i - p}{p(1-p)} \phi_i(z_{-i}), \qquad \hat{\tau}^D(z) - \frac{1}{n_1 n_0} \sum_{i<j} (d_i - d_j) \phi_{ij}(z_{-ij}),$$

respectively, with adjustments as in the lemma. The statement of the lemma corresponds to the unbiased choices

$$\hat{\tau}^D(z) = \frac{1}{n} \sum_{i=1}^{n} \frac{d_i - p}{p(1-p)} y_i, \qquad \hat{\tau}^D(z) = \frac{1}{n_1 n_0} \sum_{i<j} (d_i - d_j)(y_i - y_j).$$

The representations are restrictive, but not unique. In the minimal non-trivial case $n = 2$ and $|\mathcal{Y}| = 2$ for known treatment probability, the leave-one-out representation reduces the dimension of estimators $\hat{\tau} \in \mathbb{R}^{(\mathcal{Y} \times \{0,1\})^n}$ from 16 to 8. Unbiased estimators form a 7-dimensional affine linear subspace, and equivalent representations lie on lines in Euclidean space.

Notably, linear regression can not generally be represented in this way, as it is not generally unbiased in my setting (Freedman, 2008). In Section A.4, I provide a simple example of a biased OLS regression. Also, I make a connection between overfitting and bias, and show that bias can persist even under sampling from a population distribution and in large samples with

high-dimensional controls.

We usually associate sample splitting with losses in efficiency in return for robustness. Since all unbiased estimators must split the sample, this logic applies here only through the robustness of the unbiasedness assumption to any distribution of potential outcomes. As long as we do not impose additional structure, all admissible (with respect to variance or equivalently mean-squared error) unbiased estimators must be among the sample-splitting estimators.

This result implies that the set of fixed-bias estimators the investigator chooses from is characterized by prohibitions. When we represent an estimator by a sum over adjusted outcomes, then there must be one such representation for which the investigator is not allowed to use the outcome and treatment assignment of a unit to construct its adjustment. For this prohibition to apply, in practice the investigator has to commit how the adjustment is constructed before she has access to the respective outcome and treatment status. I show below that this commitment leaves room for automated specification searches, and discuss in Section 1.7 that human specification searches also remain feasible.

1.6.2 Solution to the Investigator's Problem

Given the restriction to a given bias, what is the optimal solution of the investigator? The sample-splitting representation provides an objective criterion for fixed bias. Since preferences are aligned, the investigator applies their subjective prior to minimize average variance over the regression adjustments from Lemma 1.1. The resulting estimator is a Bayes estimator in the sense of Wald (1950).

In the specific case of unbiased estimators, the adjustments take a particularly simple form as solutions to prediction problems. If the investigator knew the potential outcomes, a set of

variance-minimizing regression adjustments would be given by the infeasible oracle solutions

$$\begin{aligned}\bar{y}_i &= (1-p)y_i(1) + py_i(0),\\ \Delta\bar{y}_{ij} &= \underbrace{\Big(\frac{n_0}{n}y_i(1) + \frac{n_1}{n}y_i(0)\Big)}_{=\bar{\bar{y}}_i} - \Big(\frac{n_0}{n}y_j(1) + \frac{n_1}{n}y_j(0)\Big) = \bar{\bar{y}}_i - \bar{\bar{y}}_j.\end{aligned}$$

I establish that the respective Bayesian leave-one-out and leave-two-out posterior expectations minimize average risk.[19] The resulting estimator is a constrained Bayes estimator in the sense of Wald (1950).

Theorem 1.2 (Choice of the investigator from unbiased estimators)**.** *An investigator with risk $r \in \mathcal{R}^*$ and prior π over Θ chooses the following unbiased Bayes estimators:*

1. *For a known treatment probability p,*

$$\hat{\tau}(z) = \frac{1}{n}\sum_{i=1}^{n}\frac{d_i - p}{p(1-p)}(y_i - \mathrm{E}_\pi[\bar{y}_i|z_{-i}]).$$

2. *For a fixed number n_1 of treated units,*

$$\hat{\tau}(z) = \frac{1}{n_1 n_0}\sum_{i<j}(d_i - d_j)(y_i - y_j - \mathrm{E}_\pi[\Delta\bar{y}_{ij}|z_{-ij}]).$$

The theorem is non-trivial because one adjustment appears in the estimate for multiple draws of the data. In particular, if two sample draws only differ in one unit, then the adjustments to that unit are the same. Key to the proof (which I develop in Section A.3) is solving a system of first-order conditions jointly for all potential draws of the data.

While the objective unbiasedness restriction dictates sample splitting and guarantees preference alignment, the prior picks one suitable estimator that trades off risk optimally between different unobserved states. If the prior assigns low probability to the realized set of potential outcomes, then the estimator is still unbiased, but may have high variance. In any case, the investigator wants to reveal her best guess given prior knowledge.

[19] In the case of known p this is similar to Wu and Gagnon-Bartsch's (2017) LOOP estimator, which estimates $y_i(1)$ and $y_i(0)$ separately from all other units and then averages these estimates with weights $1-p$ and p to obtain an adjustment that estimates $\bar{y}_i$.

Sample splitting guards not just against misaligned preferences, but also against priors that are dogmatic in the treatment effect. From a Bayesian point of view, we only use the prior information orthogonal to the treatment effect. Hence, even if the investigator's prior is very informative about the treatment effect, the estimator will not reflect this ex-ante bias. The definition of investigator risk functions $\mathcal{R}^*$ as mean-squared error with respect to some pseudo-target therefore plays a second role. Alignment with respect to these preferences also implies robustness against misspecification of priors in the direction of the treatment effect. Hence, wrong preconceptions about treatment effects will not lead to systematic distortions in estimates if we restrict researchers to unbiased estimators.

1.6.3 Complete Class and Estimation-Prediction Duality

Since there is generally no single best estimator for all values of the truth, we have minimized average loss for some prior. If instead we consider admissible estimators that are not dominated by any other estimator in a purely frequentist sense, the same conclusions apply. Indeed, a duality result connects admissible unbiased estimation and admissible prediction.

For finite support any admissible estimator is the limit of a Bayes estimator that minimizes posterior loss given the data for some prior with full support (e.g. Ferguson, 1967). I extend this complete-class argument to unbiased estimators by applying it to the representation in Lemma 1.1.

Theorem 1.3 (Complete-class theorem for unbiased estimators)**.** *For any unbiased estimator $\hat{\tau}$ of the sample-average treatment effect that is not dominated with respect to variance, there is a converging sequence of priors $(\pi_t)_{t=1}^{\infty}$ with full support such that $\hat{\tau}$ equals the limit of the respective estimators in Theorem 1.2. Conversely, for any converging sequence of priors $(\pi_t)_{t=1}^{\infty}$ that put positive weight on every state $\theta \in \Theta$, every converging subsequence of corresponding estimators is admissible among unbiased estimators.*

The individual increments

$$\phi_i(z_{-i}) = \mathrm{E}_\pi[\bar{y}_i|z_{-i}],$$
$$\phi_{i;j}(z_{-ij}) = \mathrm{E}_\pi[\bar{\bar{y}}_i|z_{-ij}]$$

solve a leave-one-out and leave-two-out out-of-sample prediction problem, respectively. (The adjustment $\phi_{ij}(z_{-ij})$ is obtained as $\phi_{ij}(z_{-ij}) = \phi_{i;j}(z_{-ij}) - \phi_{j;i}(z_{-ij})$.) Indeed, ϕ_i and $\phi_{i;j}$ minimize the average of the forecast risk

$$r^i_\theta(\hat{y}_i) = \mathrm{E}_\theta[w(d_i)(\hat{y}_i - y_i)^2] \tag{1.2}$$

given the respective data and the prior π. The weights

$$w(d_i) = \left(\frac{(d_i - p)}{p(1-p)}\right)^2, \qquad\qquad w(d_i) = \left(\frac{n(d_i n - n_1)}{n_1 n_0}\right)^2$$

put higher emphasis on the smaller of the the treatment and control groups.[20]

I apply the complete-class logic to both sides of the problem to obtain a one-to-many correspondence between unbiased admissible estimation and admissible prediction.[21] The relationship is not one-to-one because different prediction solutions may correspond to the same estimator.

Corollary 1.2 (Estimation-prediction duality)**.** *Any admissible unbiased estimator can be expressed in terms of a jointly admissible solution to the prediction problems with risks r^i_θ. Conversely, any jointly admissible solution to the prediction problems defined by risks r^i_θ yields an admissible unbiased estimator of the sample-average treatment effect via the representation in Lemma 1.1. (Here, by joint admissibility I mean that the solutions to all prediction problems are the limits of average-risk minimizers with respect to the same sequence of priors.)*

[20] This mirrors Lin's (2013) "tyranny of the minority" estimator, which puts similar weights into a least-squares regression.

[21] Wager *et al.* (2016) in an asymptotic framework using a similar sample-splitting construction note that "the precision of the treatment effect estimates obtained by such regression adjustments depends only on the prediction risk of the fitted regression adjustment." Similary, Wu and Gagnon-Bartsch (2017) show that the variance of their LOOP estimator is approximately the average mean-squared error in predicting the oracle adjustments, provided that certain covariance terms are negligible. In my finite-sample Bayesian setting, the duality holds exactly.

While the estimator itself is unbiased, the implicit prediction solution of a low-variance estimator will typically have bias.

1.6.4 Constrained Cross-Fold Solutions

It may be infeasible to estimate all regression adjustments optimally. Mimicking machine-learning practice, one could instead partition the sample into K folds and estimate adjustments in one fold jointly from the units in all other folds. The resulting estimator resembles Wager *et al.*'s (2016) "cross-estimation" and Chernozhukov *et al.*'s (2017) "cross-fitting" estimator.

Remark 1.1 (Exact K-fold cross-fitting)**.** *For a partition of the sample*

$$\{1,\dots,n\} = \bigcup_{k=1}^{K} \mathcal{I}^{(k)}$$

into K folds with $n^{(k)} \geq 2$ units each of which $n_1^{(k)} > 0$ treated and $n_0^{(k)} > 0$ untreated, the estimator

$$\hat{\tau}(z) = \frac{1}{n}\sum_{k=1}^{K} n^{(k)} \sum_{i\in\mathcal{I}^{(k)}} \frac{d_i n^{(k)} - n_1^{(k)}}{n_1^{(k)} n_0^{(k)}} \left(y_i - \phi_i^{(k)}(z_{-\mathcal{I}^{(k)}})\right)$$

is unbiased for the sample-average treatment effect τ conditional on $(\mathcal{I}^{(k)})_{k=1}^{K}$ and $(n_1^{(k)})_{k=1}^{K}$ under either randomization. The investigator obtains their constrained optimal (Bayes) $\hat{\tau}$ among these estimators at

$$\phi_i^{(k)}(z_{-\mathcal{I}^{(k)}}) = \mathrm{E}_\pi[n_0^{(k)} y_i(1) + n_1^{(k)} y_i(0) | z_{-\mathcal{I}^{(k)}}]/n^{(k)}.$$

Randomization could be within folds or folds could be chosen after overall randomization. If K divides n_1 and n_0, we achieve perfect balance by stratifying folds by treatment (or the other way around), $Kn_1^{(k)} = n_1$ and $Kn_0^{(k)} = n_0$.

In particular, the optimal regression adjustments are predictions even when not all adjustments are estimated. Indeed, $\phi_i^{(k)}$ minimizes average risk r_θ^i in (1.2) with weight

$$w_i^{(k)}(d_i) = \left(\frac{n^{(k)}(d_i n^{(k)} - n_1^{(k)})}{n_1^{(k)} n_0^{(k)}}\right)^2$$

given data from other folds and the prior π. An unbiased estimator of the risk is the average loss on fold k.

1.6.5 Machine Learning Algorithms as Agents

When high-dimensional unit characteristics are available, machine learning offers a solution to the prediction problems implicit to unbiased estimation. Effectively, machine learning engages in automated specification searches to find a model that predicts well, which Wager *et al.* (2016) and Wu and Gagnon-Bartsch (2017) also leverage for variance reduction in the same setting. I take a principal–agent perspective on machine-learning algorithms to provide a formal embedding. The investigator as principal delegates to the machine-learning agent. Through sample splitting, there is no misalignment of preferences between the investigator and the machine-learning agent provided the latter minimizes prediction risk, and the investigator achieves a second-best estimation solution from first-best predictions.

For randomly sampled units, the implicit prediction solutions forecast outcomes from characteristics. If units are draw according to the population distribution $(y_i(1), y_i(0), x_i) \overset{\text{iid}}{\sim} \mathrm{P}$ that includes characteristics x_i, then

$$y_i(1), y_i(0) | x_1, \ldots, x_n \sim \mathrm{P}(x_i).$$

Increments $\phi_i(y_{T_i}, d_{T_i})$ fitted on $T_i \subseteq \{1, \ldots, n\} \setminus \{i\}$ minimize expected forecast risk

$$\mathrm{E}[r^i_\theta(\hat{y}_i) | x_1, \ldots, x_n, y_{T_i}, d_{T_i}] = \mathrm{E}[\mathrm{E}_\theta[w(d_i)(\hat{y}_i - y_i)^2 | y_i(1), y_i(0)] | x_i]$$

over $\hat{y}_i \in \mathbb{R}$. Writing $\hat{y}_i = \hat{f}_i(x_i)$ with $\hat{f}_i : \mathcal{X} \to \mathbb{R}$ a function of training data $(y_{T_i}, d_{T_i}, x_{T_i})$ evaluated on the test point x_i, $\hat{f}_i$ solves the prediction problem

$$L_i(\hat{f}) = \mathrm{E}[w(d_i)(\hat{f}(x_i) - y_i)^2 | x_i] \to \min_{\hat{f}}. \tag{1.3}$$

Here, I conflate the population distribution P with the sampling process to describe the distribution of observable data.

Supervised machine learning offers non-parametric solutions of out-of-sample prediction

problems like (1.3) that are particularly suitable for high-dimensional characteristics x_i. Since the test point (y_i, d_i, x_i) follows the same distribution as the training sample $\mathcal{T}_i$, sample-splitting techniques within the training sample allow for specification searches (in the form of model regularization and combination) to obtain good average predictions at the test point. Furthermore, the realized loss at i is an unbiased estimate of $L_i(\hat{f}_i)$.

I capture machine learning as an agent who minimizes average forecast risk for weighted loss $w(d_i)(\hat{f}(x_i) - y_i)^2$. The machine-learning agent's choice $\hat{f}_i$ may have complex structure that eludes causal interpretation and its parameters may not even be stable approximations of correlation patterns (Mullainathan and Spiess, 2017). However, the investigator as principal cares only about the forecast properties of the agent's solution.

Provided that the agent (approximately) minimizes risk, their choices are (approximately) aligned with the preferences of the investigator. There is no moral hazard from unobserved modeling decisions in the delegation of the prediction task from investigator to machine-learning agent. The machine-learning delegation task can be realized as a contract that pays the provider of the machine-learning solution according to the observed performance of prediction functions $\hat{f}_i$ on test points (y_i, d_i, x_i).

Crucially, sample splitting guards against prediction mistakes. Even when the specific prediction method does not minimize forecast risk or makes systematic mistakes, the resulting estimator is still unbiased. Worse predictions can lead to worse estimation performance, but only through variance.

1.7 Pre-Analysis Plans and Ex-Post Analysis

There are two ways in which we can guarantee that the investigator delivers an unbiased estimator (or, more generally, an estimator with fixed bias). In the previous section, I derived a representation of unbiased estimators that require that the investigator's estimator only uses one part of the sample when constructing regression adjustments for another part. Since the investigator will ultimately work with all of the data, this condition cannot be verified ex-post, but has to be guaranteed by ex-ante commitment. One way to guarantee that the

estimator fulfills this condition is to require that the investigator commits to the construction of all regression adjustments before she has seen any of the data.

In this section, I consider instead that the investigator commits to how she will split and distribute the data to one or multiple researchers who have not yet accessed the data. Detailed commitment may be infeasible for methods that require active guidance by the researcher, impractical for very complex algorithms, or inefficient when some prior uncertainty is resolved only after the initial commitment. I therefore consider sample-splitting schemes that leave some or all regression adjustments unspecified, and instead delegate their estimation. Delegating to one researcher can already improve over simple pre-specified estimators. Delegating to two researchers attains semi-parametric efficiency without any commitment beyond sample splitting.

1.7.1 Automated vs. Human Specification Searches

The results in this chapter imply a constructive characterization of robust yet flexible pre-analysis plans. The two ways of ensuring unbiasedness correspond to two different types of specification searches. The first way in which we can be flexible while also ensuring unbiasedness is that the investigator commits in her pre-analysis plan which algorithm she will use to construct regression adjustments. This algorithm then engages in automated specification searches to solve the prediction problems I have shown to be equivalent to unbiased estimation.

The second way in which specification searches remain possible applies when the investigator splits the sample and distributes it to one or multiple researchers. Then each researcher can search through specifications using his full subsample and does not have to commit to an empirical strategy ex ante. As long as the investigator commits to how she will distribute the sample and use the output from the researchers, and follows the procedures I characterize below, the resulting estimator is again guaranteed to be unbiased.

Automated and human specification searches can be combined to ensure precise and unbiased estimation under logistical constraints. An investigator who analyzes the data by

herself can split the sample into two, apply a pre-specified algorithm to the first half of the data, and search through specifications by hand only in the second half.

1.7.2 Unbiased Estimators without Full Commitment

I show that the class of unbiased estimators includes protocols that do not require full pre-commitment, but leave additional degrees of freedom open. (I formulate the results in terms of unbiased estimation, but they carry over to estimation with fixed bias in the same way as above.) The investigator commits to an estimator that includes flexible inputs by one or multiple researchers. Each researcher obtains access to a subset of the data, but does not have to pre-commit to their output.

Definition 1.2 (K-distribution contract)**.** *A K-distribution contract $\hat{\tau}^{\Phi}$ distributes data $z = (y, d) \in (\mathcal{Y} \times \{0, 1\})^n = \mathcal{Z}$ to K researchers. Researcher k receives data $g_k(z) \in A_k$ and returns the intermediate output $\hat{\phi}_k(g_k(z)) \in B_k$. The estimate is*

$$\hat{\tau}^{\Phi}((\hat{\phi}_k)_{k=1}^K; z) = \Phi((\hat{\phi}_k(g_k(z))_{k=1}^K; z).$$

The investigator chooses the functions g_k (from data in $\mathcal{Z}$ to researcher input in A_k) and Φ (from the researcher outputs in $\bigtimes_{k=1}^K B_k$ and data in $\mathcal{Z}$ to estimates in $\mathbb{R}$) before accessing the data.

While the investigator still commits which part of the data individual researchers receive and how their choices and the data form an overall estimate, the individual researchers' actions are not pre-specified. From my results in the previous section, I obtain a full characterization of K-distribution contracts that are unbiased no matter the choices of the researchers. Since the resulting estimators are always unbiased, the preferences of the researchers, the investigator, and the designer over these contracts are aligned provided that the investigator and the researchers all minimize average risk for risk functions in $\mathcal{R}^*$ and have the same prior π.

Lemma 1.3 (Characterization of unbiased K-distribution contracts)**.** *A K-distribution contract $\hat{\tau}^{\Phi}$ is unbiased for the sample-average treatment effect τ_θ for any conformable researcher input $(\hat{\phi}_k)_{k=1}^K$ if and only if:*

1. *For known treatment probability p, there exist regression adjustments $(\phi_i : (\bigtimes_{k \in C_i} B_k) \times (\mathcal{Y} \times \{0,1\})^{n-1} \to \mathbb{R})_{i=1}^n$ such that*

$$\hat{\tau}^\Phi((\hat{\phi}_k)_{k=1}^K; z) = \frac{1}{n} \sum_{i=1}^n \frac{d_i - p}{p(1-p)} (y_i - \phi_i((\hat{\phi}_k(g_k(z))_{k \in C_i}; z_{-i}))$$

for $C_i = \{k; g_k(z) = \tilde{g}(z_{-i})$ for some $\tilde{g}\}$.

2. *For fixed number n_1 of treated units, there exist regression adjustments $(\phi_{ij} : (\bigtimes_{k \in C_{ij}} B_k) \times (\mathcal{Y} \times \{0,1\})^{n-2} \to \mathbb{R})_{i<j}$ such that*

$$\hat{\tau}^\Phi((\hat{\phi}_k)_{k=1}^K; z) = \frac{1}{n_1 n_0} \sum_{i<j} (d_i - d_j)(y_i - y_j - \phi_{ij}((\hat{\phi}_k(g_k(z))_{k \in C_{ij}}; z_{-ij})),$$

for $C_{ij} = \{k; g_k(z) = \tilde{g}(z_{-ij})$ for some $\tilde{g}\}$.

In other words, the regression adjustments of a given unit are only controlled by the choices of researchers who do not have access to data from that unit. The sets C_i, C_{ij} are thus the set of researchers who have control over regression adjustments ϕ_i, ϕ_{ij}. For the special case $K = 1$, this construction resembles proposals to use hold-out sets to avoid false positives in multiple testing (Dahl *et al.*, 2008; Fafchamps and Labonne, 2016; Anderson and Magruder, 2017). For general K, the construction resembles K-fold cross-validation. Indeed, we obtain a particularly simple form if we restrict sample distribution to K-fold partitions.

Corollary 1.1 (Characterization of fixed-bias K-fold distribution contracts)**.** *For K disjoint folds $\mathcal{I}_k \subseteq \{1, \ldots, n\}$ with projections $g_k : (y, d) = z \mapsto z_{-\mathcal{I}_k} = (y_i, d_i)_{i \neq \mathcal{I}_k}$, a K-distribution contract $\hat{\tau}^\Phi$ has given bias if and only if:*

1. *For a known treatment probability p, there exist a fixed estimator $\hat{\tau}_0(z)$ with the given bias and regression adjustment mappings $(\Phi_k)_{k=1}^K$ such that*

$$\hat{\tau}^\Phi((\hat{\phi}_k)_{k=1}^K; z) = \hat{\tau}_0(z) - \frac{1}{n} \sum_{k=1}^K \sum_{i \in \mathcal{I}_k} \frac{d_i - p}{p(1-p)} \phi_i^k(z_{-i})$$

where $(\phi_i^k)_{i \in \mathcal{I}_k} = \Phi_k(\hat{\phi}_k(z_{-\mathcal{I}_k}))$.

2. *For a fixed number n_1 of treated units, there exist a fixed estimator $\hat{\tau}_0(z)$ with the given*

bias and regression adjustment mappings $(\Phi_k)_{k=1}^K$ *such that*

$$\hat{\tau}^{\Phi}((\hat{\phi}_k)_{k=1}^K; z) = \hat{\tau}_0(z) - \frac{1}{n_1 n_0} \sum_{k=1}^{K} \sum_{\{i<j\} \subseteq \mathcal{I}_k} (d_i - d_j)\phi_{ij}^k(z_{-ij}),$$

where $(\phi_i^k)_{i \in \mathcal{I}_k} = \Phi_k(\hat{\phi}_k(z_{-\mathcal{I}_k}))$.

K-fold distribution contracts are similar to K-fold cross-fitting from Remark 1.1, but different in terms of motivation and more flexible in terms of application. K-fold distribution is motivated by ensuring unbiasedness, not by computational limitations. While K-fold cross-fitting is contained as the special case where a researcher determines the regression adjustments for all units in the target fold directly from their training data (that is, no data from the target fold is used to adjust any of the units in that fold), K-fold distribution contracts also contain solutions that use additional data without bias. Indeed, for the case of known p, say, if regression adjustments take the form $\phi_i^k(\lambda_k; z_{-i})$ with a pre-determined function $\phi_i^k(\cdot;\cdot)$ and some tuning parameter λ_k, then the adjustments can be be a function of all the data in z_{-i} as long as the tuning parameter λ_k is fitted only on the other folds.[22]

1.7.3 Hybrid Pre-Analysis Plans

I apply the previous result to show that a simple pre-analysis plan is dominated by a hybrid pre-analysis plan that allows for additional discretion after part of the data is revealed. The investigator fixes some regression adjustment, but can modify others after access to a subset of the sample. Since sample splitting ensures preference alignment, the hybrid estimator will dominate if the ex-post analysis permits better implementation of prior information.

I now assume that the investigator's prior π is only realized after the data is available. Before the data is available, the investigator has a prior η^I over π. I think of η^I as a crude approximation to π. A simple ex-ante prior η^I could come from high costs of fully writing down or automating the way in which the investigator translates prior information and data

[22] This idea can be applied to the post-LASSO (Belloni and Chernozhukov, 2013) after selection on the training sample. Unlike the cross-fitted LASSO, the post-selection fitting step can include the full sample (provided all regression adjustments are fitted using a leave-one- or leave-two-out construction). Furthermore, the selection step can include researcher intervention that has not been pre-specified.

into predictions of potential outcomes. The ex-post prior π could also represent updated beliefs after the pre-analysis plan has been filed. In both cases, however, the difference does not represent the information in the collected data itself, which will be incorporated in the posterior distribution instead.

Anderson and Magruder (2017) propose a hybrid pre-analysis plan for multiple testing. The investigator pre-specifies some hypothesis they will test, and then selects additional hypotheses from a training sample. The additional hypotheses are only evaluated on the remaining hold-out sample. I adopt their proposal to my estimation setting.

Definition 1.3 (Hybrid pre-analysis plan)**.** *A hybrid pre-analysis plan is a 1-fold distribution contract, i.e. an estimator*

$$\hat{\tau}^{\Phi}(\hat{\phi}; z) = \Phi(\hat{\phi}(z_T); z)$$

that pre-specifies a mapping Φ from ex-post researcher input $\hat{\phi}(z_T)$ and realized sample data z to an estimate of the sample-average treatment effect. The researcher (which here could be the investigator herself) obtains access to training data $T \subseteq \{1, \ldots, n\}$ before the final estimator is formed.

I assume that the investigator must still pre-commit to an unbiased estimator, so Corollary 1.1 for $K = 1$ fully characterizes the plans available to the investigator. In these sample-splitting plans, the choices of the researcher after gaining access to the training sample are fully aligned with the intentions of the investigator according to their updated prior. The investigator pre-commits all adjustments in the training sample according to η^I, while the researcher chooses the remaining regression adjustments according to π and their training data.

Theorem 1.4 (Hybrid pre-analysis plan dominates rigid pre-analysis plan)**.** *Assume that investigator and researcher have risk functions in $\mathcal{R}^*$. The optimal unbiased pre-committed estimator $\hat{\tau}^{\text{pre}}$ is strictly dominated by an unbiased hybrid pre-analysis plan with respect to average variance, i.e. the hybrid plan is as least as precise on average over any ex-ante prior η^I and strictly better for many non-trivial ex-ante priors η^I.*

Since the researcher's and investigator's preference over unbiased estimators is fully aligned with the designer's goal, there is no preference misalignment and the variance captures all of their risk functions.

Remark 1.2 (Optimal hybrid pre-analysis plan)**.** *The dominating hybrid plan is:*

1. *For known treatment probability p, the researcher chooses regression adjustments $(\phi_i^{\text{post}} : (\mathcal{Y} \times \{0,1\})^{n-1} \to \mathbb{R})_{i \notin T} = \hat{\phi}(z_T)$ to obtain*

$$\hat{\tau}^{\text{hybrid}}(\hat{\phi}; z) = \hat{\tau}^{\text{pre}}(z) - \frac{1}{n} \sum_{i \notin T} \frac{d_i - p}{p(1-p)} \, \phi_i^{\text{post}}(z_{-i})$$

   where $1 \leq |T| \leq n-1$.

2. *For fixed number n_1 of treated units, the researcher chooses adjustments $(\phi_{ij}^{\text{post}} : (\mathcal{Y} \times \{0,1\})^{n-2} \to \mathbb{R})_{\{i<j\} \cap T = \emptyset} = \hat{\phi}(z_T)$ to obtain*

$$\hat{\tau}^{\text{hybrid}}(\hat{\phi}; z) = \hat{\tau}^{\text{pre}}(z) - \frac{1}{n_1 n_0} \sum_{\{i<j\} \cap T = \emptyset} (d_i - d_j) \, \phi_{ij}^{\text{post}}(z_{-ij})$$

   where $1 \leq |T| \leq n-2$.

In both cases, the investigator commits to the training sample $T \subseteq \{1, \ldots, n\}$ and the unbiased estimator $\hat{\tau}^{\text{pre}} : \mathcal{Z} \to \mathbb{R}$.

The optimal ex-post adjustments modify the implicit adjustments of the ex-ante estimator to match the solution from Theorem 1.2 on the relevant subset, i.e. they solve an out-of-sample prediction problem.

1.7.4 Many-Researcher Delegation

The hybrid pre-analysis plan is itself dominated by a plan that distributes the data to multiple researchers. If a single researcher has access to the full dataset before committing their estimator, bias can return even if the researcher represents their estimate by regression adjustments. Distribution to multiple researchers reduces inefficiency without introducing misalignment. Even when ex-ante commitment beyond a trivial estimator is infeasible or

undesirable, distribution between at least two researchers can produce an ex-post desirable estimator.

Remark 1.3 (More researchers are better)**.** *Assume that the investigator and researchers all have risk functions in $\mathcal{R}^*$, and that the researchers all share the same (ex-post) prior π. Then an optimal unbiased K-distribution contract is dominated by an unbiased $K+1$-distribution contract in the sense of Theorem 1.4.*

I now consider standard large-sample efficiency criteria for the estimation of the population-average treatment effect. There is no unique variance-minimal solution in finite samples, as the class of admissible estimators is large. In the large-sample limit, however, essentially all admissible estimators have approximately equal performance, and coordination between researchers with different (non-dogmatic) priors is resolved by a common understanding of the truth.

Under random sampling of units, the semi-parametric efficiency bound of Hahn (1998) is achieved at oracle prediction adjustments.[23] For $(y_i(1), y_i(0), x_i) \overset{\text{iid}}{\sim} \mathrm{P}$ with fixed probability p of treatment, an infeasible estimator of the population average treatment effect τ is

$$\hat{\tau}^{\mathrm{P}}(z) = \frac{1}{n}\sum_{i=1}^{n} \frac{d_i - p}{p(1-p)}(y_i - \mathrm{E}[\bar{y}_i|x_i])$$

where the oracle regression adjustments are optimal given knowledge of P. While we will not generally be able to achieve the variance of $\hat{\tau}^{\mathrm{P}}$, under assumptions we can achieve a variance that is asymptotically equivalent (i.e. $\mathrm{Var}(\hat{\tau})/\mathrm{Var}(\hat{\tau}^{\mathrm{P}}) \to 1$ as $n \to \infty$).

Remark 1.4 (Semi-parametric efficiency)**.** *If researchers use prediction algorithms $(A_n : \mathcal{Z} \to \mathbb{R}^{\mathcal{X}}, z \mapsto \hat{f}_n)_{n=1}^{\infty}$ with*

$$\mathrm{E}[(\hat{f}_n(x_i) - \mathrm{E}[\bar{y}_i|x_i])^2] \to 0$$

as $n \to \infty$, then delegation to two researchers with risk functions in $\mathcal{R}^*$ (who each obtain access to half of the data, say) without further commitment achieves both finite-sample unbiased

[23] See also Imbens (2004) for a discussion of efficient estimation of average treatment effects.

estimation of τ_θ, and large-sample semi-parametric efficient estimation of τ for the semi-parametric efficiency bound of Hahn (1998).

In other words, semi-parametric efficiency is achieved from distribution of the data to at least two independent researchers with risk-consistent predictors. Data distribution ensures that there is no misalignment.

1.8 Conclusion

By taking a mechanism-design approach to econometrics, I account for misaligned researcher incentives in causal inference. I motivate why and how we should pre-commit our empirical strategies, and demonstrate that there exist flexible pre-analysis plans that allow for exploratory data analysis and machine learning without leaving room for biases. In particular, I characterize all unbiased estimators of an average treatment effect as sample-splitting procedures that permit beneficial specification searches.

My results shed light on the role of bias and variance in treatment-effect estimation from experimental data. Allowing for bias can reduce the variance and thus improve precision. But when incentives are misaligned, giving a researcher the freedom to choose the bias may, in fact, reduce precision. However, once we restrict the researcher to fixed-bias estimators, there will again be a bias–variance tradeoff in the nuisance parameters associated with the control variables. I have shown in this chapter that unbiased estimation of a treatment effect in an experiment is equivalent to a set or prediction tasks. Inside these tasks, some bias in return for a substantial variance reduction can improve prediction quality. Better predictions in turn translate into lower variance of the unbiased estimator.

In related work, I show that under additional parametric assumptions standard treatment-effect estimators are dominated because shrinkage can reduce variance without introducing bias. In a linear model with homoscedastic, Normal noise and exogenous treatment, the usual linear least-squares estimator for the treatment effect is dominated provided that there are

at least three Normally-distributed control variables (Spiess, 2017b).[24] In that case, I reduce variance without introducing bias by James and Stein (1961) shrinkage in the underlying prediction problem.[25]

I am working on extending my mechanism-design approach to other estimation tasks in experimental or quasi-experimental data. Applications include effects on endogenously chosen subgroups, heterogeneous treatment effects, treatment effects under optimal assignment, and tests for effects on multiple outcome variables. In each case, I conjecture that my approach can motivate a design restriction by its preference alignment property, yield a representation of the resulting estimators as sample-splitting procedures, and suggest a characterization of optimal mechanisms and second-best pre-analysis plans.

One possible direction to pursue is to extend the approach of this chapter to cases where unbiased estimators are generally unavailable. In instrumental-variable estimation, unbiased estimation is possible under sign restrictions on the first stage (Andrews and Armstrong, 2017), but generally infeasible when the parameter space is unrestricted (Hirano and Porter, 2015). Still, when there are many instruments, we can improve estimation by providing better solutions to the first-stage prediction problem implicit to the two-stage linear IV model. For example, shrinkage in the first stage reduces bias relative to the standard two-stage least-squares estimator (Spiess, 2017a). This finding raises the question how the delegation of the first-stage prediction problem can be realized in a way that aligns researcher preferences.

---

[24]The usual linear least-squares estimator is, by Gauss-Markov, still variance-minimal among conditionally unbiased estimators. However, once we integrate over the distribution of control variables (and if these are orthogonal to treatment), I show that there is an unbiased estimator with lower variance.

[25]In the nonparametric setting in this paper and the Normal–linear setting in Spiess (2017b), unbiased estimation reduces to prediction problems. The results are connected because they both stem from invariances that characterize the distributions – in the case of this paper reflections and permutations, in the case of Spiess (2017b) rotations that leave the Normal distribution invariant.

Chapter 2

Shrinkage in Treatment-Effect Estimation

2.1 Introduction

Many inference tasks have the following feature: the researcher wants to obtain a high-quality estimate of one or a small set of target parameters (for example, a set of treatment effects in an RCT), but also estimates a number of nuisance parameters she does not care about separately (for example, coefficients on control variables). In these cases, can we improve estimation in the target parameter by shrinking in the estimation of possibly high-dimensional nuisance parameters? In this chapter, I give a positive answer to this question in two common program-evaluation cases, namely for adjusting for control variables in an experiment and for the first stage of a linear instrumental-variables regression.

First, in a linear regression model with homoscedastic Normal noise, I consider shrinkage estimation of the nuisance parameters associated with control variables. For at least three control variables and exogenous treatment, I establish that the standard least-squares estimator is dominated with respect to squared-error loss in the treatment effect even among unbiased estimators and even when the target parameter is low-dimensional. I construct the dominating estimator by a variant of James–Stein shrinkage in a high-dimensional Normal-means problem.

It can be interpreted as an invariant generalized Bayes estimator with an uninformative (improper) Jeffreys prior in the target parameter.

Second, in a two-stage linear regression model with Normal noise, I consider shrinkage in the estimation of the first-stage instrumental-variable coefficients. For at least four instrumental variables and a single endogenous regressor, I establish that the standard two-stage least-squares estimator is dominated with respect to bias. The dominating IV estimator applies James–Stein type shrinkage in a first-stage high-dimensional Normal-means problem followed by a control-function approach in the second stage. It preserves invariances of the structural instrumental variable equations.

My results directly build upon properties of shrinkage estimators established by James and Stein (1961). They are most closely related to previous work by Hansen (2007, 2016, 2017) on model-averaging estimators and shrinkage in instrumental variables. Relative to these existing results, I focus on bias properties of the estimators and obtain finite-sample dominance in a Normal model.

Shrinkage in control variables is discussed in Section 2.2, and shrinkage in the first stage of an instrumental-variable regression follows in Section 2.3. Section 2.4 concludes by contrasting the two dominance results.

2.2 Unbiased Shrinkage Estimation in Experimental Data

When we estimate a treatment effect in the presence of control variables, can we reduce variance in the estimation of a target parameter without inducing bias by shrinking in the estimation of possibly high-dimensional nuisance parameters? In a linear regression model with homoscedastic, Normal noise, I show that a natural application of James–Stein shrinkage to the parameters associated with at least three control variables reduces loss in the possibly low-dimensional treatment effect parameter without producing bias provided that treatment is random.

The proposed estimator effectively averages between regression models with and without control variables, similar to the Hansen (2016) model-averaging estimator and coinciding up to

a degrees-of-freedom correction with the corresponding Mallows estimator from Hansen (2007). For the specific choice of shrinkage, I contribute four finite-sample properties: First, I note that by averaging over the distribution of controls we obtain dominance of the shrinkage estimator even for low-dimensional target parameters, unlike other available results that require a loss function that is at least three-dimensional. Second, I establish that the resulting estimator remains unbiased under exogeneity of treatment. Third, I conceptualize it as a two-step estimator with a first-stage prediction component. Fourth, I show that it can be seen as a natural, invariant generalized Bayes estimator with respect to a partially improper prior corresponding to uninformativeness in the target parameter.

The linear regression model is set up in Section 2.2.1. Section 2.2.2 proposes the estimator and establishes loss improvement relative to a benchmark OLS estimator provided treatment is exogenous. Section 2.2.3 motivates the estimator as an invariant generalized Bayes estimator (with respect to an improper prior) in a suitably transformed many-means problem. Section 2.2.4 discusses the properties of the estimator in a simulation exercise.

2.2.1 Linear Regression Setup

I consider estimation of the structural parameter $\beta \in \mathbb{R}^k$ in the canonical linear regression model

$$Y_i = \alpha + X_i'\beta + W_i'\gamma + U_i \tag{2.1}$$

from n iid observations (Y_i, X_i, W_i), where $X_i \in \mathbb{R}^m$ are the regressors of interest, $W_i \in \mathbb{R}^k$ control variables, and $U_i \in \mathbb{R}$ is homoscedastic, Normal noise. α is an intercept,[1] and γ is a nuisance parameter. To obtain identification of β in Equation (2.1), I assume that U_i is orthogonal to X_i and W_i (no omitted variables).

Throughout this document, I write upper-case letters for random variables (such as Y_i) and lower-case letters for fixed values (such as when I condition on $X_i = x_i$). When I suppress

[1]We could alternatively include a constant regressor in X_i and subsume α in β. I choose to treat α separately since I will focus on the loss in estimating β, ignoring the performance in recovering the intercept α.

indices, I refer to the associated vector or matrix of observations, e.g. $Y \in \mathbb{R}^n$ is the vector of outcome variables Y_i and $X \in \mathbb{R}^{n \times m}$ is the matrix with rows X_i'.

2.2.2 Two-Step Partial Shrinkage Estimator

By assumption there are control variables W available with

$$Y|X{=}x, W{=}w \sim \mathcal{N}(\mathbf{1}\alpha + x\beta + w\gamma, \sigma^2 \mathbb{I}_n)$$

where σ^2 need not be known. We care about the (possibly high-dimensional) nuisance parameter γ only in so far as it helps us to estimate the (typically low-dimensional) target parameter β, which is our object of interest.

Given $x \in \mathbb{R}^{n \times m}$ and $w \in \mathbb{R}^{n \times k}$, where we assume that $(\mathbf{1}, x, w)$ has full rank $1{+}m{+}k \leq n$, let $q = (q_{\mathbf{1}}, q_x, q_w, q_r) \in R^{n \times n}$ orthonormal where $q_{\mathbf{1}} \in \mathbb{R}^n, q_x \in \mathbb{R}^{n \times m}, q_w \in \mathbb{R}^{n \times k}$ such that $\mathbf{1}$ is in the linear subspace of $\mathbb{R}^n$ spanned by $q_{\mathbf{1}} \in \mathbb{R}^n$ (that is, $q_{\mathbf{1}} \in \{\mathbf{1}/\sqrt{n}, -\mathbf{1}/\sqrt{n}\}$), the columns of $(\mathbf{1}, x)$ are in the space spanned by the columns of $(q_{\mathbf{1}}, q_x)$, and the columns of $(\mathbf{1}, x, w)$ are in the space spanned by the columns of $(q_{\mathbf{1}}, q_x, q_w)$. (Such a basis exists, for example, by an iterated singular value decomposition.) Then,

$$Y^* = q'Y|X{=}x, W{=}w \sim \mathcal{N}\left(\begin{pmatrix} q_{\mathbf{1}}'\mathbf{1}\alpha + q_{\mathbf{1}}'x\beta + q_{\mathbf{1}}'w\gamma \\ q_x'x\beta + q_x'w\gamma \\ q_w'w\gamma \\ \mathbf{0}_{n-1-m-k} \end{pmatrix}, \sigma^2\mathbb{I}_n\right).$$

Writing Y_x^*, Y_w^*, Y_r^* for the appropriate subvectors of Y^*, we find, in particular, that

$$\begin{pmatrix} Y_x^* \\ Y_w^* \\ Y_r^* \end{pmatrix} |X{=}x, W{=}w \sim \mathcal{N}\left(\begin{pmatrix} \mu_x + a\mu_w \\ \mu_w \\ \mathbf{0}_{n-1-m-k} \end{pmatrix}, \sigma^2\mathbb{I}_{n-1}\right)$$

where $\mu_x = q_x'x\beta \in \mathbb{R}^m$, $\mu_w = q_w'w\gamma \in \mathbb{R}^k$, and $a = q_x'w(q_w'w)^{-1} \in \mathbb{R}^{m \times k}$.[2] In transforming

[2]Alternatively, we could have denoted by μ_x the mean of Y_x^*. However, by separating out μ_x from $a\mu_w$ I feel that the role of μ_w as a relevant nuisance parameter becomes more transparent.

linear regression to this Normal-means problem, as well as in partitioning the coefficient vector into two groups, for only one of which I will propose shrinkage, I follow Sclove (1968).

Conditional on $X{=}x, W{=}w$ and given an estimator $\hat{\mu}_w = \hat{\mu}_w(Y_w^*, Y_r^*)$ of μ_w, a natural estimator of μ_x is $\hat{\mu}_x = \hat{\mu}_x(Y_x^*, Y_w^*, Y_r^*) = Y_x^* - a\hat{\mu}_w$. An estimator of β is obtained by setting $\hat{\beta} = (q_x' x)^{-1}\hat{\mu}_x$. (The linear least-squares estimator for β is obtained from $\hat{\mu}_w = Y_w^*$.) A natural loss function for $\hat{\beta}$ that represents prediction loss units is the weighted loss $(\hat{\beta} - \beta)'(x' q_x q_x' x)(\hat{\beta} - \beta) = \|\hat{\mu}_x - \mu_x\|^2$. We can therefore focus on the (conditional) expected squared-error loss in estimating μ_x, for which we find

$$\mathrm{E}[\|\hat{\mu}_x - \mu_x\|^2 | X{=}x, W{=}w] = m\sigma^2 + \mathrm{E}[\|\hat{\mu}_w - \mu_w\|^2_{a'a} | X{=}x, W{=}w]$$

with the seminorm $\|v\|_{a'a} = \sqrt{v'a'av}$ on $\mathbb{R}^k$.

For high-dimensional μ_w ($k \geq 3$), a natural estimator $\hat{\mu}_w$ with low expected squared-error loss is a shrinkage estimator of the form $\hat{\mu}_w = CY_w^*$ with scalar C, such as the James and Stein (1961) estimator for which $C = 1 - \frac{(k-2)\|Y_r^*\|^2}{(n-m-k+1)\|Y_w^*\|^2}$ (or its positive part). While improving with respect to expected squared-error loss ($a'a = \text{const.} \cdot \mathbb{I}_k$), this specific estimator may yield higher (conditional) expected loss in μ_x when the implied loss function for μ_w deviates from squared-error loss ($a'a \neq \text{const.} \cdot \mathbb{I}_k$, so the loss function is not invariant under rotations). We will show below that it is still appropriate in the case of independence of treatment and control.

For conditional inference it is known that the least-squares estimator is admissible for estimating β provided $m \leq 2$ and inadmissible provided $m \geq 3$ no matter what the dimensionality k of the nuisance parameter γ is (James and Stein, 1961), as the rank of the loss function is decisive. The above construction does not provide a counter-example to this result: the rank of $a'a = (w'q_w)^{-1}w'q_x q_x' w(q_w' w)^{-1}$ is at most m, so for $m \leq 2$, $\hat{\mu}_w = Y_w^*$ remains admissible for the loss function on the right. While we could achieve improvements for $m \geq 3$ – through shrinkage in $\hat{\mu}_w$ and/or directly in $\hat{\mu}_x$ – our interest is in the case where m is low and k is high. Conditional on $X{=}x, W{=}w$ we can thus not hope to achieve improvements that hold for any (β, γ), but we can still hope that shrinkage estimation of μ_w yields better estimates of

β on average over draws of the data.

To this end, assume that

$$\mathrm{vec}(W)|X{=}x \sim \mathcal{N}(\mathrm{vec}(\mathbf{1}\alpha_W + x\beta_W), \Sigma_W \otimes \mathbb{I}_n)$$

(that is, $W_i|X{=}x \overset{\text{iid}}{\sim} \mathcal{N}(1\alpha_W + x_i\beta_W, \Sigma_W)$). Here, $\Sigma_W \in \mathbb{R}^{k\times k}$ is symmetric positive-definite (but not necessarily known). $\alpha_W \in \mathbb{R}^{1\times k}, \beta_W \in \mathbb{R}^{m\times k}$ describe the conditional expectation of control variables given the regressors $X{=}x$. The case where x and W are orthogonal ($\beta_W = \mathbb{O}_{m\times k}$) and controls W thus not required for identification will play a special role below.

Given $X{=}x$, assume $(q_{\mathbf{1}}, q_x)$ is deterministic, and fix $q_\perp$ such that $\tilde{q} = (q_{\mathbf{1}}, q_x, q_\perp) \in \mathbb{R}^{n\times n}$ is orthonormal. Note that

$$\mathrm{vec}((q_x, q_\perp)'W)|X{=}x \sim \mathcal{N}\left(\mathrm{vec}\left(\begin{pmatrix} q_x'x\beta_W \\ \mathbb{O}_{n-1-m\times k}\end{pmatrix}\right), \Sigma_W \otimes \mathbb{I}_{n-1}\right).$$

In particular, $q_x'W \perp\!\!\!\perp q_\perp'W$. It follows with

$$(q_x, q_\perp)'Y|X{=}x, W{=}w \sim \mathcal{N}\left(\begin{pmatrix} q_x'x\beta + q_x'w\gamma \\ q_\perp'w\gamma\end{pmatrix}, \sigma^2\mathbb{I}_{n-1}\right)$$

that indeed $q_x'(Y,W) \perp\!\!\!\perp q_\perp'(Y,W)$.

Conditional on $W{=}w$ in the above derivation, $a\hat{\mu}_w - a\mu_w = q_x'w\hat{\gamma} - q_x'w\gamma$ for $\hat{\gamma} = (q_w'w)^{-1}\hat{\mu}_w$ a function of $q_\perp'w$ and $(Y_w^*, Y_r^*) = (q_w'q_\perp, q_r'q_\perp)(q_\perp'Y)$, so $\hat{\gamma} = \hat{\gamma}(q_\perp'Y, q_\perp'w)$. Assuming measurability, $\hat{\gamma}(q_\perp'Y, q_\perp'W) \perp\!\!\!\perp (q_x'Y, q_x'W)$. Now writing $\hat{\gamma} = \hat{\gamma}(q_\perp'y, q_\perp'w)$ this implies that

$$\mathrm{E}[\|\hat{\mu}_w - \mu_w\|_{a'a}^2 | X{=}x, q_\perp'(Y,W) = q_\perp'(y,w)] = \|\hat{\gamma} - \gamma\|^2_{\mathrm{E}[W'q_xq_x'W|X=x]}$$

with $\mathrm{E}[W'q_xq_x'W|X{=}x] = \beta_W'x'q_xq_x'x\beta_W + m\Sigma_W$ of full rank k. For the expectation of the implied $\hat{\beta}$, we find

$$\mathrm{E}[\hat{\beta}|X{=}x, q_\perp'(Y,W) = q_\perp'(y,w)] = \beta - \beta_W(\hat{\gamma} - \gamma).$$

We obtain the following characterization of conditional bias and squared-error loss of the implied estimator $\hat{\beta}$:

Lemma 2.1 (Properties of the two-step estimator)**.** *Let* $(\tilde{Y}, \tilde{W})$ *be jointly distributed according to*

$$\begin{aligned}\mathrm{vec}(\tilde{W}) &\sim \mathcal{N}(\mathbf{0}_{k(n-1-m)}, \Sigma_W \otimes \mathbb{I}_{n-1-m}),\\ \tilde{Y}|\tilde{W} = \tilde{w} &\sim \mathcal{N}(\tilde{w}\gamma, \sigma^2 \mathbb{I}_{n-1-m}),\end{aligned}$$

and write $\tilde{\mathrm{E}}$ *for the corresponding expectation operator. For any measurable estimator* $\hat{\gamma} : \mathbb{R}^{n-m-1} \times \mathbb{R}^{n-m-1\times k} \to \mathbb{R}^k$ *with* $\tilde{\mathrm{E}}[\|\hat{\gamma}(\tilde{Y}, \tilde{W})\|^2] < \infty$*, the estimator* $\hat{\beta}(y, w) = (q_x' x)^{-1} q_x' y - (q_x' x)^{-1} q_x' w \hat{\gamma}(q_\perp' y, q_\perp' w)$ *defined for convenience for fixed* x*, has conditional bias*

$$\mathrm{E}[\hat{\beta}(Y, W)|X{=}x] - \beta = -\beta_W(\tilde{\mathrm{E}}[\hat{\gamma}(\tilde{Y}, \tilde{W})] - \gamma)$$

and expected (prediction-norm) loss

$$\mathrm{E}[\|\hat{\beta}(Y, W) - \beta\|^2_{x' q_x q_x' x}|X{=}x] = m\sigma^2 + \tilde{\mathrm{E}}[\|\hat{\gamma}(\tilde{Y}, \tilde{W}) - \gamma\|^2_\phi]$$

for $\phi = \beta_W' x' q_x q_x' x \beta_W + m\Sigma_W$*.*

Note that this lemma does not rely on $n \geq 1 + m + k$, and indeed generalizes to the case $n > 1 + m$ for any $k \geq 1$, including $k > n$.

We consider the special case where treatment is exogenous, and thus $\beta_W = \mathbb{O}_{m\times k}$. This assumption could be justified, for example, in a randomized trial. Note that in this case in addition to the linear least-squares estimator in the "long" regression that includes controls W another natural unbiased (conditional on $X{=}x$) estimator is available, namely the coefficient $(q_x' x)^{-1} q_x' Y$ in the "short" regression without controls. The "long" and "short" regression represent special (edge) cases in the class of two-step estimators introduced above, which are all unbiased in that sense under the exogeneity assumption:

Corollary 2.1 (A class of unbiased two-step estimators)**.** *If* $\beta_W = \mathbb{O}_{m\times k}$ *then for any* $\hat{\gamma}$ *and* $\hat{\beta}$ *as in Lemma 2.1* $\mathrm{E}[\hat{\beta}(Y, W)|X{=}x] = \beta$*. Furthermore,*

$$\mathrm{E}[\|\hat{\beta}(Y, W) - \beta\|^2_{x' q_x q_x' x}|X{=}x] = m\tilde{\mathrm{E}}[(\tilde{Y}_0 - \tilde{W}_0' \hat{\gamma}((\tilde{Y}_i, \tilde{W}_i)_{i=1}^{n-1-m}))^2]$$

for $(\tilde{Y}_i, \tilde{W}_i)_{i=0}^{n-1-m}$ *iid with* $\tilde{W}_i \sim \mathcal{N}(\mathbf{0}_k, \Sigma_W)$*,* $\tilde{Y}_i|\tilde{W}_i = \tilde{w}_i \sim \mathcal{N}(\tilde{w}_i'\gamma, \sigma^2)$ *(here,* $(\tilde{Y}_i, \tilde{W}_i)_{i=1}^{n-1-m}$

is the training sample and $(\tilde{Y}_0, \tilde{W}_0)$ *an additional test point drawn from the same distribution).*

This corollary clarifies that the class of natural estimators derived above are unbiased conditional on $X{=}x$ (but not necessarily on $X{=}x, W{=}w$ jointly), with expected loss equal to the expected out-of-sample prediction loss in a prediction problem where the prediction function $\tilde{w}_0 \mapsto \tilde{w}_0'\hat{\gamma}$ is trained on $n-1-m$ iid draws, and evaluated on an additional, independent draw $(\tilde{Y}_0, \tilde{W}_0)$ from the same distribution. The "long" and "short" regressions are included as the special cases $\hat{\gamma}(\tilde{w}, \tilde{y}) = (\tilde{w}'\tilde{w})^{-1}\tilde{w}'\tilde{y}$ and $\hat{\gamma} \equiv \mathbf{0}_k$, respectively.

The covariates in training and test sample follow the same distribution, which suggests an estimator that is invariant to rotations in the corresponding k-means problem. Indeed, the dominating estimator I construct in the following results is of the form

$$\hat{\mu}_w = \left(1 - \frac{p\|Y_r^*\|^2}{\|Y_w^*\|^2}\right) Y_w^*,$$

where the standard James and Stein (1961) estimator (for unknown σ^2) is recovered at $p = \frac{k-2}{n-m-k+1}$.

Theorem 2.1 (Inadmissibility of OLS among unbiased estimators)**.** *Maintain* $\beta_W = \mathbb{O}_{m\times k}$. *Denote by* $(\hat{\alpha}^{\mathrm{OLS}}, \hat{\beta}^{\mathrm{OLS}}, \hat{\gamma}^{\mathrm{OLS}})$ *the coefficients and by* $\mathrm{SSR} = \|Y - \mathbf{1}\hat{\alpha}^{\mathrm{OLS}} - X\hat{\beta}^{\mathrm{OLS}} - W\hat{\gamma}^{\mathrm{OLS}}\|^2$ *the sum of squared residuals in a linear least-squares regression of* Y *on an* $\mathbf{1}$*,* X*, and* W*. Write* $h = \mathbb{I}_n - \mathbf{1}_n\mathbf{1}_n'/n$ *(the annihilator matrix with respect to the intercept). Assume that* $k \geq 3$ *and* $n \geq m + k + 2$*. Then, the two-step estimator* $\hat{\beta} = (X'hX)^{-1}X'h(Y - W\hat{\gamma})$ *with*

$$\hat{\gamma} = \left(1 - \frac{p\,\mathrm{SSR}}{\|\hat{\gamma}^{\mathrm{OLS}}\|^2_{W'h(\mathbb{I}-X(X'hX)^{-1}X')hW}}\right)\hat{\gamma}^{\mathrm{OLS}}$$

where $p \in \left(0, \frac{2(k-2)}{n-m-k+2}\right)$ *is unbiased for* β *given* $X{=}x$ *and dominates* $\hat{\beta}^{\mathrm{OLS}}$ *in the sense that*

$$\mathrm{E}[\|\hat{\beta} - \beta\|^2_{X'hX}|X{=}x] < \mathrm{E}[\|\hat{\beta}^{\mathrm{OLS}} - \beta\|^2_{X'hX}|X{=}x].$$

Proof. The OLS estimator in the theorem corresponds to $\hat{\gamma}^{\mathrm{OLS}}(\tilde{y}, \tilde{w}) = (\tilde{w}'\tilde{w})^{-1}\tilde{y}'\tilde{w}$ in

Lemma 2.1, which yields the maximum-likelihood estimator $\hat{\gamma}^{\text{OLS}}(\tilde{Y}, \tilde{W})$ for γ given data

$$\begin{aligned}\text{vec}(\tilde{W}) &\sim \mathcal{N}(\mathbf{0}_{k(n-1-m)}, \Sigma_W \otimes \mathbb{I}_{n-1-m}),\\ \tilde{Y}|\tilde{W} = \tilde{w} &\sim \mathcal{N}(\tilde{w}\gamma, \sigma^2 \mathbb{I}_{n-1-m}).\end{aligned}$$

By Baranchik (1973), this maximum-likelihood estimator is inadmissible with respect to the risk $\tilde{\text{E}}[\|\hat{\gamma} - \gamma\|^2_{\Sigma_W}]$ and thus for $\tilde{\text{E}}[\|\hat{\gamma} - \gamma\|^2_{\phi}$ in Lemma 2.1, as $\phi = m\Sigma_W$ for $\beta_W = \mathbb{O}_{m\times k}$. However, Baranchik (1973) also includes an intercept that is estimated, but does not enter the loss function. To formally use the result for our case without intercept in the first-step prediction exercise, I construct an augmented problem such that the dominance result in the augmented problem implies the theorem.

To this end, let

$$\begin{aligned}\text{vec}(W^a) &\sim \mathcal{N}(\mathbf{0}_{k(n-m)}, \Sigma_W \otimes \mathbb{I}_{n-m}),\\ Y^a|W^a = w^a &\sim \mathcal{N}(w^a\gamma, \sigma^2 \mathbb{I}_{n-m}).\end{aligned}$$

(which has one additional sample point, and could without loss include intercepts in W^a, Y^a). By Baranchik (1973, Theorem 1), the estimator

$$\hat{\gamma}^a = \left(1 - p\frac{(Y^a)'h^aY^a - \|\hat{\gamma}^{a,\text{OLS}}\|^2_{(W^a)'h^aW^a}}{\|\hat{\gamma}^{a,\text{OLS}}\|^2_{(W^a)'h^aW^a}}\right)\hat{\gamma}^{a,\text{OLS}}$$

strictly dominates $\hat{\gamma}^{a,\text{OLS}} = ((W^a)'h^aW^a)^{-1}(W^a)'h^aZ^a$, where $h^a = \mathbb{I}_{n-m} - \mathbf{1}_{n-m}\mathbf{1}'_{n-m}/(n-m)$, in the sense that

$$\text{E}^a[(\hat{\gamma}^a - \gamma)'\Sigma_W(\hat{\gamma}^a - \gamma)] < \text{E}^a[(\hat{\gamma}^{a,\text{OLS}} - \gamma)'\Sigma_W(\hat{\gamma}^{a,\text{OLS}} - \gamma)]$$

for any $\gamma \in \mathbb{R}^k$, provided that $p \in \left(0, \frac{2(k-2)}{n-m-k+2}\right)$ with $k \geq 3$ and $n - m \geq k + 2$.

We now show that this implies dominance of $\hat{\gamma}(\tilde{Y}, \tilde{W})$ for

$$\hat{\gamma}(\tilde{y}, \tilde{w}) = \left(1 - p\frac{\tilde{y}'\tilde{y} - \|\hat{\gamma}^{\text{OLS}}(\tilde{y}, \tilde{w})\|^2_{\tilde{w}'\tilde{w}}}{\|\hat{\gamma}^{\text{OLS}}(\tilde{y}, \tilde{w})\|^2_{\tilde{w}'\tilde{w}}}\right)\hat{\gamma}^{\text{OLS}}(\tilde{y}, \tilde{w})$$

in the original problem. Let $q^a \in \mathbb{R}^{(n-m)\times(n-m-1)}$ be such that $(q^a, \mathbf{1}_{n-m}/(n-m))$ is orthonormal (that is, the columns of q^a complete $\mathbf{1}_{n-m}/(n-m)$ to an orthonormal basis of

$\mathbb{R}^{m-n}$). This implies that $q^a(q^a)' = h^a$ and $(q^a)'q^a = \mathbb{I}_{n-m-1}$. Then, $(q^a)'(Y^a, W^a) \stackrel{d}{=} (\tilde{Y}, \tilde{W})$. In particular,

$$((Y^a)'h^a(Y^a), (Y^a)'h^a(W^a), (W^a)'h^a(W^a)) \stackrel{d}{=} (\tilde{Y}'\tilde{Y}, \tilde{Y}'\tilde{W}, \tilde{W}'\tilde{W})$$

and thus $(\hat{\gamma}^a, \hat{\gamma}^{a,\mathrm{OLS}}) \stackrel{d}{=} (\hat{\gamma}(\tilde{Y}, \tilde{W}), \hat{\gamma}^{\mathrm{OLS}}(\tilde{Y}, \tilde{W}))$. We have thus established

$$\begin{aligned}&\tilde{\mathrm{E}}[(\hat{\gamma}(\tilde{Y}, \tilde{W}) - \gamma)'\Sigma_W(\hat{\gamma}(\tilde{Y}, \tilde{W}) - \gamma)] \\ &< \tilde{\mathrm{E}}[(\hat{\gamma}^{\mathrm{OLS}}(\tilde{Y}, \tilde{W}) - \gamma)'\Sigma_W(\hat{\gamma}^{\mathrm{OLS}}(\tilde{Y}, \tilde{W}) - \gamma)].\end{aligned}$$

Note that $\hat{\gamma}^{\mathrm{OLS}}(\tilde{y}, \tilde{w}) = (\tilde{w}'\tilde{w})^{-1}\tilde{y}'\tilde{w}$ in Lemma 2.1 does indeed yield $\hat{\gamma}^{\mathrm{OLS}}$ and $\hat{\beta}^{\mathrm{OLS}}$ in the theorem, and that this extends to $\hat{\gamma}$ and $\hat{\beta}$ by

$$\hat{\gamma}(q_\perp' y, q_\perp' w) = \left(1 - p\frac{\|y\|^2_{q_\perp q_\perp'} - \|\hat{\gamma}^{\mathrm{OLS}}(\ldots)\|^2_{w'q_\perp q_\perp' w}}{\|\hat{\gamma}^{\mathrm{OLS}}(\ldots)\|^2_{w'q_\perp q_\perp' w}}\right)\hat{\gamma}^{\mathrm{OLS}}(\ldots)$$

with $q_\perp q_\perp' = h(\mathbb{I} - x(x'hx)^{-1}x')h$ and

$$\begin{aligned}\mathrm{SSR} &= \|Y - \mathbf{1}\hat{\alpha}^{\mathrm{OLS}} - X\hat{\beta}^{\mathrm{OLS}} - W\hat{\gamma}^{\mathrm{OLS}}\|^2 \\ &= \|Y - W\hat{\gamma}^{\mathrm{OLS}}\|^2_{h(\mathbb{I}-X(X'hX)^{-1}X')h} \\ &= \|Y\|^2_{h(\mathbb{I}-X(X'hX)^{-1}X')h} - \|W\hat{\gamma}^{\mathrm{OLS}}\|^2_{h(\mathbb{I}-X(X'hX)^{-1}X')h} \\ &= \|Y\|^2_{h(\mathbb{I}-X(X'hX)^{-1}X')h} - \|\hat{\gamma}^{\mathrm{OLS}}\|^2_{W'h(\mathbb{I}-X(X'hX)^{-1}X')hW}.\end{aligned}$$

Unbiasedness and dominance follow with $\beta_W = \mathbb{O}_{m\times k}$ in Lemma 2.1. □

Note that the result extends to the positive-part analog for which the shrinkage factor is set to zero whenever the expression is negative. For $m = 1$, the following dominance is immediate:

Corollary 2.2 (A non-contradiction of Gauss–Markov)**.** *For exogenous treatment, $m = 1$, $k \geq 3$, and $n \geq k + 3$, there exists an estimator $\hat{\beta}$ with $\mathrm{E}[\hat{\beta}|X{=}x] = \beta$ and $\mathrm{Var}(\hat{\beta}|X{=}x) < \mathrm{Var}(\hat{\beta}^{\mathrm{OLS}}|X{=}x)$.*

The assumption of exogenous treatment is essential for this result, as dropping conditioning on W and restricting interest to β would not suffice to break optimality of linear least-squares.

2.2.3 Invariance Properties and Bayesian Interpretation

Starting with the transformations in subsection 2.2.2, we consider the decision problem of estimating β (equivalently, μ_x). Guided by the treatment of a linear panel-data model in Chamberlain and Moreira (2009), I develop the specific estimator proposed in Theorem 2.1 as (the empirical Bayes version of) an invariant Bayes estimator with respect to a partially uninformative (improper) Jeffreys prior.

In this section, we condition on X throughout and assume that covariates W are Normally distributed given X. Writing $W_x^* = q_x'W, W_\perp^* = q_\perp'W, Y_x^* = q_x'Y, Y_\perp^* = q_\perp'Y$, the transformation developed in subsection 2.2.2 yields the joint distribution

$$\begin{aligned} \begin{pmatrix} W_x^* \\ W_\perp^* \end{pmatrix} &= \begin{pmatrix} \mu_W \\ \mathbb{O}_{s\times k} \end{pmatrix} + V_W \Sigma_W^{1/2} \qquad (V_W)_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0,1) \\ \begin{pmatrix} Y_x^* \\ Y_\perp^* \end{pmatrix} &= \begin{pmatrix} \mu_x + W_x^*\gamma \\ W_\perp^*\gamma \end{pmatrix} + V_Y \sigma^2 \qquad (V_Y)_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1) \end{aligned} \tag{2.2}$$

where $\Sigma_W^{1/2}$ is the unique symmetric positive-definite square-root of the symmetric positive-definite matrix Σ_W, and V_W and V_Y are independent. Here, in addition to $\mu_x = q_x'x\beta$, also $\mu_W = q_x'x\beta_W$, and $s = n - m - 1$. I write $\mathcal{Z} = \mathbb{R}^{m+s} \times \mathbb{R}^{(m+s)\times k}$ for the sample space from which (Y^*, W^*) is drawn according to this P_θ, where I parametrize $\theta = (\mu_x, \gamma) \in \Theta = \mathbb{R}^m \times \mathbb{R}^k$. (I take $\sigma^2, \Sigma_W, \mu_W$ to be constants.)

The action space is $\mathcal{A} = \mathbb{R}^m$, from which an estimate of μ_x is chosen. As the loss function $L : \Theta \times \mathcal{A} \to \mathbb{R}$ I take squared-error loss $L(\theta, a) = \|\mu_x - a\|^2$. An estimator $\hat{\beta} : \mathcal{Z} \to \mathcal{A}$ from the previous section is a feasible decision rule in this decision problem.

For an element $g = (g_\mu, g_x, g_W, g_\perp)$ in the (product) group $G = \mathbb{R}^m \times O(m) \times O(k) \times O(s)$, where $\mathbb{R}^m$ denotes the group of real numbers with addition (neutral element 0) and $O(k)$ the group of orthonormal matrices in $\mathbb{R}^{k\times k}$ with matrix multiplication (neutral element $\mathbb{I}_k$), consider the following set of transformations (which are actions of G on $\mathcal{Z}, \Theta, \mathcal{A}$):

- Sample space: $m_{\mathcal{Z}} : G \times \mathcal{Z} \to \mathcal{Z}$,

$$\begin{aligned}&(g, (y_x, y_\perp, w_x, w_\perp))\\ &\mapsto (g_x y_x + g_\mu, g_\perp y_\perp, g_x w_x \Sigma_W^{-1/2} g_W' \Sigma_W^{1/2}, g_\perp w_\perp \Sigma_W^{-1/2} g_W' \Sigma_W^{1/2})\end{aligned}$$

- Parameter space: $m_\Theta : G \times \Theta \to \Theta$,

$$(g, (\mu_x, \gamma)) \mapsto (g_x \mu_x + g_\mu, \Sigma_W^{-1/2} g_W \Sigma_W^{1/2} \gamma)$$

- Action space: $m_{\mathcal{A}} : G \times \mathcal{A} \to \mathcal{A}, (g, a) \mapsto g_x a + g_\mu$

For exogenous treatment, these transformations are tied together by leaving model and loss invariant. Indeed, the following is immediate from Equation 2.2:[3]

Proposition 2.1 (Invariance of model and loss)**.** *For $\mu_W = \mathbb{O}_{m\times k}$:*

1. *The model is invariant: $m_{\mathcal{Z}}(g, (Y^*, W^*)) \sim \mathrm{P}_{m_\Theta(g,\theta)}$ for all $g \in G$.*

2. *The loss is invariant: $L(m_\Theta(g, \theta), m_{\mathcal{A}}(g, a)) = L(\theta, a)$ for all $g \in G$.*

By Proposition 2.1, a natural (generalized) Bayes estimator of μ_x is derived from an improper prior on θ that is invariant under the action of G on Θ, as this will yield a decision rule $d : \mathcal{Z} \to \mathcal{A}$ that is invariant in the sense that $d(m_{\mathcal{Z}}(g, (y, w))) = m_{\mathcal{A}}(g, d((y, w)))$ for all $(g, (y, w)) \in G \times \mathcal{Z}$. This implies for μ_x as an improper prior the Haar measure with respect to the translation action (i.e. up to a multiplicative constant the σ-finite Lebesgue measure on $\mathbb{R}^m$), and for γ a prior that is uniform on ellipsoids $\gamma' \Sigma_W \gamma = \omega$. Taking $\frac{\omega}{\tau^2} \sim \chi^2_m$ with some $\tau > 0$ yields the prior $\gamma \sim \mathcal{N}(\mathbf{0}, \tau^2 \Sigma_W^{-1})$. With a product prior for θ, the resulting generalized Bayes estimator for μ_x – which minimizes posterior loss conditional on the data – is

$$\begin{aligned}\mathrm{E}[\mu_x | Y^* = y, W^* = w] &= y_x - w_x \,\mathrm{E}[\gamma | Y^* = y, W^* = w]\\ &= y_x - w_x (w_\perp' w_\perp + \sigma^2 \Sigma_W / \tau^2)^{-1} w_\perp' y_\perp.\end{aligned}$$

[3]Alternatively, we could have treated μ_W as an element of the parameter space and extend the analysis to the case of endogenous treatment. Adding $(g, \mu_W) \mapsto g_x \mu_W \Sigma_W^{-1/2} g_W' \Sigma_W^{1/2}$ to the action on the parameter space would have retained invariance.

Replacing Σ_W by the specific sample analog $W_\perp' W_\perp / s$, we obtain the estimator $Y_x^* - \frac{s\tau^2}{s\tau^2+\sigma^2} W_x^*((W_\perp^*)'W_\perp^*)^{-1}(W_\perp^*)'Y_\perp^*$. Similarly assuming that $\gamma \sim \mathcal{N}(\mathbf{0}, s\tau^2 W_\perp' W_\perp)$, an unbiased estimator of $\frac{s\tau^2}{s\tau^2+\sigma^2}$ (given W) is

$$C = 1 - \frac{(Y_\perp^*)'(\mathbb{I}_s - W_\perp^*((W_\perp^*)'W_\perp^*)^{-1}(W_\perp^*)')Y_\perp^*/(s-k)}{(Y_\perp^*)'W_\perp^*((W_\perp^*)'W_\perp^*)^{-1}(W_\perp^*)'Y_\perp^*/(k-2)}.$$

This estimator corresponds to the estimator from Theorem 2.1 at $p = \frac{k-2}{s-k} = \frac{k-2}{n-m-k-1}$. By construction, it retains the invariance of the associated generalized Bayes estimator. This is not specific to this value of p:

Proposition 2.2 (Invariance of estimator)**.** *For any p, the estimator $\hat{\beta}$ from Theorem 2.1 is invariant with respect to the above actions of G.*

2.2.4 Simulation

In this section, I study the performance of the shrinkage estimator introduced in Section 2.2.2 in a simulation exercise. I generate data according to Equation 2.1, where I normalize the variance of the error term to one and the target parameter to $\beta = 1$ (in particular, X_i is uni-dimensional). The (X_i, W_i) are drawn independently from a multivariate standard Normal distribution. I fix the sample size to $n = 80$. I vary the size $\|\gamma\|$ of the control-variable parameter, as well as the number k of controls.

On this data, I compare the performance of estimates of β from the short OLS regression (Y_i on X_i and a constant, "Short"), the long OLS regression (Y_i on X_i, W_i and a constant, "Long"), and the partial shrinkage estimator introduced in Section 2.2.2 ("JS"). For each parameter setting and estimator I obtain the root mean-squared error from 100,000 Monte-Carlo draws.

Table 2.1: *Simulation results for partial shrinkage estimator (root mean-squared error)*

$k \to$	5			15			40		
$\|\gamma\| \downarrow$	Short	Long	JS	Short	Long	JS	Short	Long	JS
0.0	0.132	0.138	0.134	0.132	0.154	0.135	0.132	0.243	0.149
0.5	0.148	0.138	0.138	0.148	0.155	0.144	0.148	0.242	0.168
1.0	0.187	0.139	0.139	0.188	0.154	0.150	0.187	0.243	0.198

Table 2.1 reports the results of the simulation exercise for $\|\gamma\| \in \{0.0, 0.5, 1.0\}$ and $k \in \{5, 15, 40\}$. The short, long, and partial shrinkage estimators are all unbiased. As predicted by the theory, the partial shrinkage estimator persistently outperforms the long OLS estimator, with higher gains when the control coefficient is small or its dimension high. The short OLS estimator performs better than the long and partial shrinkage estimators when the control variables matter very little.

2.3 Bias Reduction in Instrumental-Variable Estimation through First-Stage Shrinkage

The standard two-stage least-squares (TSLS) estimator is known to be biased towards the OLS estimator when instruments are many or weak. In a linear instrumental variables model with one endogenous regressor, at least four instruments, and Normal noise, I propose an estimator that combines James–Stein shrinkage in a first stage with a second-stage control-function approach. Unlike other IV estimators based on James–Stein shrinkage, my estimator reduces bias uniformly relative to TSLS. Unlike LIML, it is invariant with respect to the structural form and translation of the target parameter.

I consider the first stage of a two-stage least-squares estimator as a high-dimensional prediction problem, to which I apply rotation-invariant shrinkage akin to James and Stein (1961). Regressing the outcome on the resulting predicted values of the endogenous regressor directly would shrink the TSLS estimator towards zero, which could increase or decrease bias depending on the true value of the target parameter. Conversely, shrinking the TSLS estimator towards the OLS estimator can reduce risk (Hansen, 2017), but increases bias towards OLS. Instead, my proposed estimator uses the first-stage residuals as controls in the second-stage regression of the outcome on the endogenous regressor. If no shrinkage is applied, the TSLS estimator is obtained as a special case, while a variant of James and Stein (1961) shrinkage that never fully shrinks to zero uniformly reduces bias.

The proposed estimator is invariant to a group of transformations that include translation

in the target parameter. While the limited-information maximum likelihood estimator (LIML) can be motivated rigorously as an invariant Bayes solution to a decision problem (Chamberlain, 2007), these transformations rotate the (appropriately re-parametrized) target parameter and invariance applies to a loss function that has a non-standard form in the original parametrization. In particular, unlike LIML, the invariance of my estimator applies to squared-error loss.

The two-stage linear model is set up in Section 2.3.1. Section 2.3.2 proposes the estimator and establishes bias improvement relative to TSLS. Section 2.3.3 develops invariance properties of the proposed estimator. Section 2.3.4 discusses the properties of the estimator in a simulation exercise.

2.3.1 Two-Stage Linear Regression Setup

I consider estimation of the structural parameter $\beta \in \mathbb{R}$ in the standard two-stage linear regression model

$$\begin{aligned} Y_i &= \alpha + X_i'\beta + W_i'\gamma + U_i \\ X_i &= \alpha_X + Z_i'\pi + W_i'\gamma_X + V_i \end{aligned} \tag{2.3}$$

from n iid observations (Y_i, X_i, Z_i, W_i), where $X_i \in \mathbb{R}$ is the regressor of interest (assumed univariate), $W_i \in \mathbb{R}^k$ control variables, $Z_i \in \mathbb{R}^\ell$ instrumental variables, and $(U_i, V_i)' \in \mathbb{R}^2$ is homoscedastic (wrt Z_i), Normal noise. α is an intercept,[4] and γ and π are nuisance parameters. This model could be motivated by a latent variable present in both outcome and first-stage equation under appropriate exclusion restrictions as in Chamberlain (2007).[5]

Throughout this document, I write upper-case letters for random variables (such as Y_i) and lower-case letters for fixed values (such as when I condition on $X_i = x_i$). When I suppress indices, I refer to the associated vector or matrix of observations, e.g. $Y \in \mathbb{R}^n$ is the vector of outcome variables Y_i and $X \in \mathbb{R}^{n\times m}$ is the matrix with rows X_i'.

[4]We could alternatively include a constant regressor in X_i and subsume α in β. I choose to treat α separately since I will focus on the loss in estimating β, ignoring the performance in recovering the intercept α.

[5]In this section, the intercepts α, α_X could be subsumed in the control coefficients γ, γ_X without loss, but I maintain this notation to keep it consistent.

For the noise I use the notation

$$\begin{pmatrix} U_i \\ V_i \end{pmatrix} | Z_i = z_i, W_i = w_i \sim \mathcal{N}\left(\mathbf{0}_2, \begin{pmatrix} \sigma^2 & \rho\sigma\tau \\ \rho\sigma\tau & \tau^2 \end{pmatrix} \right)$$

for some $\rho \in (-1, 1)$. The reduced form is

$$\begin{pmatrix} Y \\ X \end{pmatrix} | Z = z, W{=}w \sim \mathcal{N}\left(\begin{pmatrix} \alpha + z\pi_Y + w\gamma_Y \\ \alpha_X + z\pi + w\gamma_X \end{pmatrix}, \Sigma \otimes \mathbb{I}_{2n} \right)$$

with

$$\begin{aligned} \pi_Y &= \pi\beta, \\ \gamma_Y &= \gamma + \gamma_X\beta, \end{aligned} \qquad \Sigma = \begin{pmatrix} \sigma^2 + 2\rho\beta\sigma\tau + \beta^2\tau^2 & \rho\sigma\tau + \beta\tau^2 \\ \rho\sigma\tau + \beta\tau^2 & \tau^2 \end{pmatrix}.$$

Note that there is a one-two-one mapping between reduced-form and structural-form parameters provided that the proportionality restriction $\pi_Y = \pi\beta$ holds. I develop a natural many-means form directly from the structural model, which is thus without loss, but not without consequence. Throughout, our interest will be in estimating β for many instruments (large ℓ).

We have $U_i | X_i = x_i, Z_i = z_i, W_i = w_i \sim \mathcal{N}\left(\frac{\rho\sigma}{\tau} v_i, (1-\rho^2)\sigma^2\right)$ where $v_i = x_i - \alpha_X - z_i'\pi - w_i'\gamma_X$. Given $w \in \mathbb{R}^{n\times k}$ and $z \in \mathbb{R}^{n\times\ell}$, where I assume that $(\mathbf{1}, w, z)$ has full rank $1 + k + \ell \leq n - 1$, let $q = (q_{\mathbf{1}}, q_w, q_z, q_r) \in R^{n\times n}$ orthonormal where $q_{\mathbf{1}} \in \mathbb{R}^n, q_w \in \mathbb{R}^{n\times k}, q_z \in \mathbb{R}^{n\times\ell}$ such that $\mathbf{1}$ is in the linear subspace of $\mathbb{R}^n$ spanned by $q_{\mathbf{1}} \in \mathbb{R}^n$ (that is, $q_{\mathbf{1}} \in \{\mathbf{1}/n, -\mathbf{1}/n\}$), the columns of $(\mathbf{1}, w)$ are in the space spanned by the columns of $(q_{\mathbf{1}}, q_w)$, and the columns of $(\mathbf{1}, w, z)$ are in the space spanned by the columns of $(q_{\mathbf{1}}, q_w, q_z)$. (As above, such a basis exists, for example, by an iterated singular value decomposition.) Then,

$$\begin{pmatrix} q_z'X \\ q_r'X \end{pmatrix} | Z{=}z, W{=}w \sim \mathcal{N}\left(\begin{pmatrix} q_z'z\pi \\ \mathbf{0}_{n-1-k-\ell} \end{pmatrix}, \tau^2\mathbb{I}_{n^*} \right)$$

$$\begin{pmatrix} q_z'Y \\ q_r'Y \end{pmatrix} | X{=}x, Z{=}z, W{=}w \sim \mathcal{N}\left(\begin{pmatrix} q_z'x \\ q_r'x \end{pmatrix} \beta + \begin{pmatrix} q_z'x - q_z'z\pi \\ q_r'x \end{pmatrix} \frac{\rho\sigma}{\tau}, (1{-}\rho^2)\sigma^2\mathbb{I}_{n^*} \right),$$

where $n^* = n - 1 - k$. Writing $X_z^*, X_r^*, Y_z^*, Y_r^*$ for the respective subvectors,

$$X^* = \begin{pmatrix} X_z^* \\ X_r^* \end{pmatrix}, \qquad Y^* = \begin{pmatrix} Y_z^* \\ Y_r^* \end{pmatrix},$$

$\mu = q_z' z \pi$, and $s = n - 1 - k - \ell$, we arrive at the canonical structural form

$$\begin{aligned} X^* &\sim \mathcal{N}\left(\begin{pmatrix} \mu \\ \mathbf{0}_s \end{pmatrix}, \tau^2 \mathbb{I}_{\ell+s}\right) \\ Y^* | X^* = x^* &\sim \mathcal{N}\left(x^*\beta + \left(x^* - \begin{pmatrix} \mu \\ \mathbf{0}_s \end{pmatrix}\right) \frac{\rho\sigma}{\tau}, (1-\rho^2)\sigma^2 \mathbb{I}_{\ell+s}\right), \end{aligned} \tag{2.4}$$

where I have suppressed conditioning on $Z{=}z, W{=}w$ (and omit it from here on).

2.3.2 Control-Function Shrinkage Estimator

Given an estimator $\hat{\mu} = \hat{\mu}(X^*)$ of μ, a feasible implied estimator for β in Equation 2.4 is the coefficient on X^* in a linear regression of Y^* on X^* and the control function $X^* - (\hat{\mu}', \mathbf{0}_s')'$. (The two-stage least-squares estimator $\hat{\beta}^{\text{TSLS}} = \frac{(Y_z^*)' X_z^*}{(X_z^*)' X_z^*}$ is obtained from the first-stage OLS solution $\hat{\mu}^{\text{OLS}} = X_z^*$. It is biased towards the OLS estimator $\hat{\beta}^{\text{OLS}} = \frac{(Y^*)' X^*}{(X^*)' X^*}$.)

For high-dimensional μ, a natural estimator for μ is a shrinkage estimator of the form $\hat{\mu}(X^*) = c(X^*) X_z^*$ with scalar $c(X^*)$. The conditional bias of the implied control-function estimator $\hat{\beta}$ takes a particularly simple form for this class of estimators:

Lemma 2.2 (Conditional bias of CF–shrinkage estimators)**.** *For $x^* \in \mathbb{R}^{\ell+s}$ with $c(x^*) \neq 0$,*

$$\begin{aligned} \mathrm{E}[\hat{\beta} | X^* = x^*] - \beta &= \mathrm{E}\left[\frac{\hat{\mu}'(\hat{\mu} - \mu)}{\hat{\mu}'\hat{\mu}} \middle| X^* = x^*\right] \frac{\rho\sigma}{\tau} \\ &= \left(1 - \frac{1}{c(x^*)} \frac{(x_z^*)'\mu}{(x_z^*)' x_z^*}\right) \frac{\rho\sigma}{\tau}. \end{aligned}$$

Shrinkage in the James and Stein (1961) estimator (for unknown τ^2) takes the form $c(x^*) = 1 - p\frac{\|x_r^*\|^2}{\|x_z^*\|^2}$. This shrinkage pattern (and its positive-part variant) is unappealing here, as it can cross zero, around which point the estimator diverges. A natural variant that

mitigates this problem is

$$c(x^*) = \frac{1}{1 + p\frac{\|x_r^*\|^2}{\|x_z^*\|^2}} = \frac{\|x_z^*\|^2}{\|x_z^*\|^2 + p\|x_r^*\|^2},$$

which behaves as $1 - p\frac{\|x_r^*\|^2}{\|x_z^*\|^2}$ for small $p\frac{\|x_r^*\|^2}{\|x_z^*\|^2}$, but never quite reaches zero.

Theorem 2.2 (Bias dominance through shrinkage)**.** *Assume that $\ell \geq 4$ and $p \in \left(0, 2\frac{\ell-2}{s}\right)$. Then $|\,\mathrm{E}[\hat{\beta}|Z{=}z, W{=}w] - \beta| < |\,\mathrm{E}[\hat{\beta}^{\mathrm{TSLS}}|Z{=}z, W{=}w] - \beta|$ provided $\rho \neq 0$ and $\|\mu\| \neq 0$ (otherwise equality).*

The requirement $\ell \geq 4$ is an artifact of this specific shrinkage pattern and dominance should extend to $\ell = 3$ for an appropriate modification.

Proof. For the (rescaled) bias, where $\lambda = ps$ and $M = X_z^*$, we have by Lemma 2.2 that

$$\begin{aligned} B(\lambda) = \frac{\tau}{\rho\sigma}\,\mathrm{E}[\hat{\beta} - \beta] &= \mathrm{E}\left[1 - \frac{\|X_z^*\|^2 + p\|X_r^*\|^2}{\|X_z^*\|^2}\frac{(X_z^*)'\mu}{(X_z^*)'X_z^*}\right] \\ &= \mathrm{E}\left[1 - \frac{\|X_z^*\|^2 + p\,\mathrm{E}\left[\|X_r^*\|^2\middle|X_z^*\right]}{\|X_z^*\|^2}\frac{(X_z^*)'\mu}{(X_z^*)'X_z^*}\right] \\ &= \mathrm{E}\left[1 - \frac{M'\mu}{\|M\|^2} - \lambda\tau^2\frac{M'\mu}{\|M\|^4}\right], \end{aligned}$$

provided that $\mathrm{E}\left|1 - \frac{M'\mu}{\|M\|^2} - \lambda\tau^2\frac{M'\mu}{\|M\|^4}\right| < \infty$. By the multi-dimensional version of Stein's (1981) lemma for $h(M) = \frac{1}{\|M\|^2}$,

$$-2\tau^2\,\mathrm{E}\left[\frac{M}{\|M\|^4}\right] = \tau^2\,\mathrm{E}\left[\nabla h(M)\right] = \mathrm{E}\left[(M - \mu)h(M)\right] = \mathrm{E}\left[\frac{M - \mu}{\|M\|^2}\right],$$

again provided that all moments exist.

For the existence of moments, note that by Cauchy–Schwarz and Jensen it suffices to consider $\mathrm{E}\left\|M/\|M\|^4\right\| = \mathrm{E}[\|M\|^{-3}]$. To establish that this expectation is finite, note that the distribution of $\|M\|^2/\tau^2$, a non-central χ^2 distribution with ℓ degrees of freedom and non-centrality parameter $\|\mu\|^2/\tau^2$, is first-order stochastically dominating a central χ^2 distribution with ℓ degrees of freedom, so it is sufficient to establish $\mathrm{E}[(X^2)^{-3/2}] < \infty$ where X^2 has a central χ^2 distribution with ℓ degrees of freedom. Now, the density $f(y)$ of $(X^2)^{3/2}$ is proportional to $y^{\ell/3-1}\exp(-y^{2/3}/2)$, implying $\lim_{y\searrow 0} f(y)/y^\alpha = 0$ for $\ell \geq 4$ and, say, $\alpha = 1/4 > 0$. The

existence of the inverse moment, i.e. $E[(X^2)^{-3/2}] < \infty$, follows by Piegorsch and Casella (1985).

We thus have

$$\mathrm{E}\left[\frac{M'\mu}{\|M\|^4}\right] = \frac{-1}{2\tau^2}\,\mathrm{E}\left[\frac{(M-\mu)'\mu}{\|M\|^2}\right],$$

which yields

$$\begin{aligned} B(\lambda) &= \mathrm{E}\left[1 - \frac{M'\mu}{\|M\|^2} + \frac{\lambda}{2}\frac{(M-\mu)'\mu}{\|M\|^2}\right] \\ &= \mathrm{E}\left[1 - \frac{\|M\|^2 - (M-\mu)'M}{\|M\|^2} + \frac{\lambda}{2}\frac{\|M^2\| - \|\mu\|^2 - (M-\mu)'M}{\|M\|^2}\right] \\ &= \frac{\lambda}{2} - \frac{\lambda}{2}\,\mathrm{E}\left[\frac{\|\mu\|^2}{\|M\|^2}\right] - \frac{\lambda-2}{2}\,\mathrm{E}\left[\frac{(M-\mu)'M}{\|M\|^2}\right]. \end{aligned}$$

Denote by K a Poisson random variable with mean $\kappa = \frac{\|\mu\|^2}{2\tau^2} > 0$. ($B(\lambda)$ is constant at 1 for $\|\mu\| = 0$, and there remains nothing to show.) From James and Stein (1961, (9), (16)) we have that

$$\begin{aligned} \mathrm{E}\left[\frac{\|\mu\|^2}{\|M\|^2}\right] &= \mathrm{E}\left[\frac{2\kappa}{\ell - 2 + 2K}\right] = Q(\ell), \\ \mathrm{E}\left[\frac{(M-\mu)'M}{\|M\|^2}\right] &= \mathrm{E}\left[\frac{\ell-2}{\ell-2+2K}\right] = P(\ell). \end{aligned}$$

It immediately follows from

$$B(\lambda) = P(\ell) - \frac{\lambda}{2}(P(\ell) + Q(\ell) - 1)$$

that the bias for the unshrunk reference estimator ($\lambda = 0$, TSLS) is $B(0) = P(\ell) > 0$, and that $B(\lambda)$ is decreasing in λ since $P(\ell) + Q(\ell) \geq 1$ by Jensen's inequality (with strict inequality unless $\|\mu\| = 0$). The (infeasible) bias-minimizing choice of λ is given by

$$\lambda^* = \frac{2P(\ell)}{P(\ell) + Q(\ell) - 1} = \frac{\ell - 2}{\frac{\ell-2}{2} + \kappa - 1/\,\mathrm{E}\left[(\frac{\ell-2}{2} + K)^{-1}\right]}.$$

To conclude the proof, I assert (and prove below) that, for any $a \geq 1$,

$$\mathrm{E}\left[(a+K)^{-1}\right] \leq \frac{1}{a + \nu - 1}. \tag{2.5}$$

With $a = \frac{\ell-2}{2}$ it follows that $\frac{\ell-2}{2} + \kappa - 1/\operatorname{E}\left[(\frac{\ell-2}{2} + K)^{-1}\right] \leq 1$ and thus $\lambda^* \geq \ell - 2$. We obtain $|B(\lambda)| \leq |B(0)|$ (dominance over TSLS in terms of bias) for all $\lambda \in (0, \ell - 2)$ by strict monotonicity of $B(\lambda)$, which yields the theorem.

To establish Equation 2.5, fix $a \in \mathbb{R}$ with $a \geq 1$ and note that for K Poisson with parameter ν

$$\begin{aligned}\operatorname{E}\left[\frac{\nu}{a+K}\right] &= \sum_{\iota=0}^{\infty} \frac{\nu}{a+\iota} \frac{\nu^\iota \exp(-\nu)}{\iota!} = \sum_{\iota=0}^{\infty} \frac{\iota+1}{a+\iota} \frac{\nu^{\iota+1} \exp(-\nu)}{(\iota+1)!} \\ &= \sum_{\iota=1}^{\infty} \frac{\iota}{a+\iota-1} \frac{\nu^\iota \exp(-\nu)}{\iota!}.\end{aligned}$$

For $a = 1$, thus $\operatorname{E}\left[\frac{\nu}{a+K}\right] = 1 - \exp(-\nu) \leq 1$. For $a > 1$,

$$\operatorname{E}\left[\frac{\nu}{a+K}\right] = \sum_{\iota=0}^{\infty} \frac{\iota}{a+\iota-1} \frac{\nu^\iota \exp(-\nu)}{\iota!} = \operatorname{E}\left[\frac{K}{a+K-1}\right] \leq \frac{\nu}{a+\nu-1}$$

by Jensen's inequality applied to the concave function $x \mapsto \frac{x}{a-1+x} (x \geq 0)$. In both cases, Equation 2.5 follows by dividing by ν, yielding a generalization of an inequality in Moser (2008, Theorem 6) to non-integer a. □

2.3.3 Invariance Properties

The estimator $\hat{\beta}$ developed in the previous section has invariance properties in a decision problem, where in spirit and notation I follow the treatment of LIML in Chamberlain (2007).

First I fix the sample and action spaces, as well as a class of loss functions, for the decision problem of estimating β. Starting with Equation 2.4, I write $\mathcal{Z} = (\mathbb{R}^{\ell+s})^2$ for the sample space from which (X^*, Y^*) is drawn according to P_θ, where I parametrize $\theta = (\beta, \mu, \rho, \sigma, \tau) \in \Theta = \mathbb{R} \times \mathbb{R}^\ell \times \mathbb{R}^3_{\geq 0}$. The action space is $\mathcal{A} = \mathbb{R}$, from which an estimate of β is chosen. I assume that the loss function $L : \Theta \times \mathcal{A} \to \mathbb{R}$ can be written as $L(\theta, a) = \ell(a - \beta)$ for some sufficiently well-behaved $\ell : \mathbb{R} \to \mathbb{R}$ (such as squared-error loss $L(\theta, a) = (a - \theta)^2$). The estimator $\hat{\beta} : \mathcal{Z} \to \mathcal{A}$ from the previous section is a feasible decision rule in this decision problem.

For an element $g = (g_\beta, g_z, g_r)$ in the (product) group $G = \mathbb{R} \times O(\ell) \times O(s)$, where $\mathbb{R}$

denotes the group of real numbers with addition (neutral element 0) and $O(\ell)$ the group of ortho-normal matrices in $\mathbb{R}^{\ell\times\ell}$ with matrix multiplication (neutral element $\mathbb{I}_\ell$), consider the following set of transformations (which are actions of G on $\mathcal{Z}, \Theta, \mathcal{A}$):

- Sample space: $m_{\mathcal{Z}} : G \times \mathcal{Z} \to \mathcal{Z}$,

$$(g, (x^*, y^*)) \mapsto \left(\begin{pmatrix} g_z & \mathbb{O} \\ \mathbb{O} & g_r \end{pmatrix} x^*, \begin{pmatrix} g_z & \mathbb{O} \\ \mathbb{O} & g_r \end{pmatrix} (y^* + g_\beta x^*) \right)$$

- Parameter space: $m_\Theta : G \times \Theta \to \Theta$,

$$(g, \theta) \mapsto (\beta + g_\beta, g_z\mu, \rho, \sigma, \tau)$$

- Action space: $m_{\mathcal{A}} : G \times \mathcal{A} \to \mathcal{A}, (g, a) \mapsto a + g_\beta$

These transformations are tied together by leaving model and loss invariant. Indeed, the following result is immediate from Equation 2.4:

Proposition 2.3 (Invariance of model and loss)**.**

1. *The model is invariant:* $m_{\mathcal{Z}}(g, (X^*, Y^*)) \sim \mathrm{P}_{m_\Theta(g,\theta)}$ *for all* $g \in G$.
2. *The loss is invariant:* $L(m_\Theta(g, \theta), m_{\mathcal{A}}(g, a)) = L(\theta, a)$ *for all* $g \in G$.

A decision rule $d : \mathcal{Z} \to \mathcal{A}$ is invariant if, for all $(g, (x^*, y^*)) \in G \times \mathcal{Z}$, $d(m_{\mathcal{Z}}(g, (x^*, y^*))) = m_{\mathcal{A}}(g, d((x^*, y^*)))$. The estimator $\hat{\beta}$ above is included in a class of invariant decision rules:

Proposition 2.4 (Invariance of a class of control-function estimators)**.** *Consider a control-function decision rule* $d((x^*, y^*))$ *obtained as the coefficient on* x^* *in a linear regression of* y^* *on* x^**, controlling for* $x^* - (c(\|x_z^*\|, \|x_r^*\|)(x_z^*)', \mathbf{0}')'$*, where* $c(\|x_z^*\|, \|x_r^*\|)$ *scalar (and measurable). Then* d *is an invariant decision rule with respect to the above actions of* G.

Proof. Fix $(x, y) \in \mathcal{Z}$ and consider $d((x, y))$. Note first that $c = c(\|x_z\|, \|x_r\|)$ is invariant to the action of G on $\mathcal{Z}$. The decision rule is

$$d((x, y)) = \frac{x' a(x) y}{x' a(x) x}$$

where

$$a(x) = \mathbb{I} - b(x)(b(x)'b(x))^{-1}b(x)' \text{ for } b(x) = \begin{pmatrix} (1-c)x_z \\ x_r \end{pmatrix}.$$

Now for any $g \in G$, where I write $q_g = \begin{pmatrix} g_z & \mathbb{O} \\ \mathbb{O} & g_r \end{pmatrix}$, we have $b(q_g x) = q_g b(x)$ and thus $a(q_g x) = q_g a(x) q_g'$. It is immediate that

$$\begin{aligned} d(m_{\mathcal{Z}}(g,(x,y))) = d((q_g x, q_g y + g_\beta q_g x)) &= \frac{x'a(x)y}{x'a(x)x} + g_\beta \frac{x'a(x)x}{x'a(x)x} \\ &= d((x,y)) + g_\beta = m_{\mathcal{A}}(g, d((x,y))), \end{aligned}$$

as claimed. □

2.3.4 Simulation

In this section, I study the performance of the shrinkage estimator introduced in Section 2.3.2 in a simulation exercise. I generate data according to Equation 2.3 without control variables W_i ($k = 0$), where I normalize the target parameter to $\beta = 1$, the variance of both error terms to one, and set their correlation to $\rho = .5$. The Z_i are drawn independently from a multivariate standard Normal distribution. I fix the sample size to $n = 60$. I vary the size $\|\pi\| \in \{0.5, 1.0\}$ of the first-stage parameter, as well as the number $\ell \in \{5, 10, 20\}$ of instruments.

On this data, I compare the performance of estimates of β from OLS regression (Y_i on X_i), two-stage least squares (TSLS), and the IV shrinkage estimator introduced in Section 2.3.2 (JSIV). For each parameter setting and each estimator I obtain bias, median bias, standard deviation (SD), root mean-squared error (RMSE), and inter-quartile range (IQR, the difference between the 75th and 25th percentile of the distribution of estimates) from 100,000 Monte-Carlo draws.

Table 2.2 reports the results of the simulation exercise. The two-stage least-squares estimator is biased towards the OLS estimator, which has positive bias. As predicted by the theory, the shrinkage estimator persistently reduces the bias, with higher gains when the

Table 2.2: *Simulation results for IV shrinkage estimator*

(a) $\|\pi\| = 1.0$

k	5			10			20		
	OLS	TSLS	JSIV	OLS	TSLS	JSIV	OLS	TSLS	JSIV
Bias	0.250	0.025	0.001	0.250	0.063	0.009	0.250	0.120	0.038
Median bias	0.250	0.032	0.010	0.251	0.067	0.019	0.250	0.123	0.047
SD	0.087	0.128	0.139	0.087	0.120	0.143	0.087	0.109	0.146
RMSE	0.265	0.131	0.139	0.265	0.135	0.143	0.265	0.162	0.151
IQR	0.116	0.166	0.178	0.116	0.158	0.184	0.116	0.144	0.188

(b) $\|\pi\| = 0.5$

k	5			10			20		
	OLS	TSLS	JSIV	OLS	TSLS	JSIV	OLS	TSLS	JSIV
Bias	0.400	0.096	0.013	0.400	0.188	0.074	0.399	0.281	0.181
Median bias	0.400	0.112	0.051	0.401	0.195	0.101	0.399	0.283	0.197
SD	0.105	0.240	0.342	0.105	0.200	0.309	0.105	0.161	0.259
RMSE	0.413	0.259	0.342	0.414	0.275	0.318	0.413	0.324	0.316
IQR	0.140	0.295	0.359	0.140	0.256	0.357	0.140	0.211	0.316

control coefficient is small or its dimension high. In the simulation, this pattern carries over to the median bias. At the same time, the variance of estimates increases, with an ambiguous effect on overall mean-squared error (MSE): while the MSE is consistently below OLS, MSE improves over two-stage least squares only for many instruments.

2.4 Conclusion

In this chapter, I discuss two applications of James–Stein-type shrinkage to treatment-effect estimation. First, shrinkage in control variables in a Normal linear model consistently reduces expected prediction error without introducing bias in the treatment parameter of interest provided treatment is random. In this case, the linear least-squares estimator is thus inadmissible even among unbiased estimators. Second, shrinkage in at least four instrumental variables in a canonical structural form provides consistent bias improvement over the two-stage least-squares estimator. Together, these results suggests different roles of overfitting in control and

instrumental variable coefficients, respectively: while overfitting to control variables induces variance, overfitting to instrumental variables in the first stage of a two-stage least-squares procedure induces bias.

Chapter 3

Robust Post-Matching Inference

3.1 Introduction

Matching methods are widely used to create balance between treatment and control groups in observational studies. Oftentimes, matching is followed by a simple comparison of means between treated and nontreated (Cochran, 1953; Rubin, 1973; Dehejia and Wahba, 1999). In other instances, however, matching is used in combination with regression or with other estimation methods more complex than a simple comparison of means. The combination of matching in a first step with a second-step regression estimator brings together parametric and nonparametric estimation strategies and reduces the dependence of regression estimates on modeling decisions (Ho *et al.*, 2007). Matching followed by regression allows the estimation of elaborate models, such as those that include interaction effects, that go beyond the average treatment effect.

In this chapter, we develop valid standard error estimates for linear regression after nearest-neighbor matching without replacement. The asymptotic properties of average treatment effect estimators that employ a simple comparison of mean outcomes between treated and nontreated after matching on covariates are well understood (Abadie and Imbens, 2006). However, studies that employ regression models after matching usually ignore the matching step when performing inference on post-matching regression coefficients. We show that this

practice is not generally valid if the second step regression is misspecified or if matching is done with replacement. For matching without replacement, we provide two easily implementable alternatives for post-matching linear regression coefficients that are robust to misspecification. First, we show that standard errors that are clustered at the level of the matches are valid under misspecification. Second, we show that a nonparametric block bootstrap that resamples matched pairs or matched groups, as opposed to resampling individual observations, also yields valid inference under misspecification. Furthermore, we show that standard errors that ignore the matching step can both under- or overestimate the variation of post-matching estimates. The procedures proposed in this chapter are straightforward to implement with standard statistical software.

Throughout the chapter, we will consider the following setup. Let W be a binary random variable representing exposure to the treatment or condition of interest (e.g., smoking), so $W = 1$ for the treated, and $W = 0$ for the nontreated. Y is a random variable representing the outcome of interest (e.g., forced expiratory volume) and X is a vector of covariates (e.g., gender or age). We will study the problem of estimating how the treatment affects the outcomes of the individuals in the treated population (that is, those with $W = 1$). In particular, we will analyze the properties of a two-step (first matching, then regression) estimator often used in empirical practice. This estimation strategy starts with a non-experimental sample, $\mathcal{S}$, from which treated units and their matches are extracted to create a matched sample, $\mathcal{S}^*$. Then, using data for the matched sample only, the researcher runs a regression of Y on Z, where Z is a vector of functions of W and X (e.g., individual variables plus interactions). We aim to obtain valid inferential methods for the coefficients of this regression, possibly under misspecification. To be precise, by "misspecification" we mean that there is no version of the conditional expectation of Y given W and X that follows the functional form employed in the second-step estimator.

A special case of our setup is that of the standard matching estimator for the average treatment effect on the treated, which is given by the regression coefficient on treatment W in a regression of Y on $Z = (1, W)'$. In this sense, our chapter generalizes the standard theory for

matching estimators. However, the framework allows for richer analysis, such as the analysis of linear interaction effects of the treatment with a given covariate, $Z = (1, W, WX', X')'$.

To illustrate the implications of our results, consider the simple case when $Z = (1, W)'$. As we mentioned in the previous paragraph, in this setting, the sample regression coefficient on W corresponds to the simple matching estimator often employed in applied studies, which is based on a post-matching comparison of means between treated and nontreated. Under well-known conditions this estimator is consistent for the average effect of the treatment on the treated (see, e.g., Abadie and Imbens, 2012), irrespective of the true form of the expectation of Y given W and X. Notice, however, that even in this simple scenario, our results imply that regression standard errors that ignore the matching step are not valid in general. While the expectation of Y given W always admits a linear version given that W is binary, a linear regression of Y on $Z = (1, W)'$ will be misspecified relative to the regression of Y on W and X, unless Y is mean-independent of X given W over a set of probability one.

The rest of the chapter is organized as follows. In Section 3.2 we first provide a detailed description of the setup of our investigation. We then characterize the parameters estimated by the two-step procedure described above. We show that these parameters coincide with the regression coefficients in a regression of Y on Z in a population for which the distribution of matching covariates X in the control group has been modified to coincide with that of the treated. This is similar to the generalization of the Oaxaca decomposition (Oaxaca, 1973) in DiNardo *et al.* (1996). Under selection on observables, that is, if treatment is as good as random conditional on X, these regression coefficients coincide with the population regression coefficients in an experiment where treatment is randomly assigned in a population that has the same distribution of X as the treated. We next establish consistency with respect to this vector of parameters, show asymptotic Normality, and describe the asymptotic variance of the post-matching estimator. In Section 3.3, we discuss different ways of constructing standard errors. Based on the results of Section 3.2, we show that naive standard errors that ignore the matching step are not generally valid if the regression model is misspecified, while clustered standard errors or an analogous block bootstrap procedure yield valid inference. Section 3.4

presents simulation evidence, which confirms our theoretical results. Section 3.5 applies our results to the analysis of the effect of smoking on pulmonary function. The results show how matching before the regression as well as robust standard errors can quantitatively and qualitatively alter conclusions from real data. Section 3.6 concludes. The Appendix contains proofs and extensions.

3.2 Post-Matching Inference

In this section, we discuss the asymptotic distribution of the least-squares estimator, obtained from a linear regression of Y on Z after matching on observables X.

3.2.1 Post-Matching Least Squares

Consider a standard binary treatment setting along the lines of Rubin (1974) with potential outcomes $Y(1)$ and $Y(0)$, of which we only observe $Y = Y(W)$ for treatment status $W \in \{0, 1\}$. Also assume that there are additional (pre-treatment) covariates S.

We will assume that the data consist of random samples of treated and nontreated. This assumption could be easily relaxed, and we adopt it only to simplify the discussion.

Assumption 3.1 (Random sampling)
$\mathcal{S} = \{(Y_i, W_i, S_i)\}_{i=1}^N$ *is a pooled sample obtained from* N_1 *and* N_0 *independent draws from the population distribution of* (Y, S) *for the treated* ($W = 1$) *and nontreated* ($W = 0$)*, respectively.*

We first form a new sample, $\mathcal{S}^* \subseteq \mathcal{S}$, by matching each treated unit, i, to M nontreated units, $\mathcal{J}(i)$, without replacement. Specifically, we assume that there is an $(m \times 1)$ vector of covariates $X = f(S) \in \mathcal{X} \subseteq \mathbb{R}^m$, along with some distance metric $d : \mathcal{X} \times \mathcal{X} \to [0, \infty)$ on the support $\mathcal{X}$ of the covariates, such that the sets of matches, $\mathcal{J}(i) \subseteq \{j; W_j = 0\}$, are chosen to minimize the sum of matching discrepancies

$$\sum_{i=1}^{N} W_i \sum_{j \in \mathcal{J}(i)} d(X_i, X_j),$$

where every nontreated unit appears in at most one set of matches, that is, matching is without replacement. For simplicity, we omit in our notation the dependence of $\mathcal{J}(i)$ on N and M.

The matched sample, $\mathcal{S}^*$, has size $n = (M+1)N_1$. We use a double subscript notation to refer to the observations in the matched sample. For instance, $Y_{n1}, \ldots, Y_{nn}$ refers to the values of the outcome variable for the units in $\mathcal{S}^*$, with analogous notation for other variables. Within the matched sample, observations will be rearranged so that observations $n1$ to nn are the N_1 treated observations followed by the MN_1 matches.

Let $Z = g(W, S)$ be a $(k \times 1)$ vector of functions of (W, S), and let $\widehat{\beta}$ be the vector of sample regression coefficients obtained from regressing Y on Z in the matched sample,

$$\begin{aligned}\widehat{\beta} &= \arg\min_{b \in \mathbb{R}^k} \frac{1}{n} \sum_{i=1}^{n} (Y_{ni} - Z_{ni}' b)^2 \\ &= \left(\frac{1}{n} \sum_{i=1}^{n} Z_{ni} Z_{ni}' \right)^{-1} \frac{1}{n} \sum_{i=1}^{n} Z_{ni} Y_{ni}. \end{aligned} \tag{3.1}$$

In Section 3.2.3 we will introduce a set of assumptions under which $\widehat{\beta}$ exists and is unique with probability approaching one.

As we mentioned above, when $Z = (1, W)'$ then the regression coefficient on W in the matched sample is given by

$$\begin{aligned}\widehat{\tau} &= \frac{1}{N_1} \sum_{i=1}^{n} W_{ni} Y_{ni} - \frac{1}{MN_1} \sum_{i=1}^{n} (1 - W_{ni}) Y_{ni} \\ &= \frac{1}{N_1} \sum_{i=1}^{N} W_i \left(Y_i - \frac{1}{M} \sum_{j \in \mathcal{J}(i)} Y_j \right),\end{aligned}$$

which is the usual matching estimator for the average effect of the treatment on the treated.

For reasons of concreteness and following the vast majority of applied practice, we restrict the analysis to linear regression after matching, as in Equation (3.1). We conjecture that our results extend to M-estimators under suitable regularity conditions.

3.2.2 Characterization of the Estimand

Before we study the sampling distribution of $\widehat{\beta}$, we first characterize its population counterpart, which we will denote by β. That is, our first task is to obtain a precise description of the nature of the parameters estimated by $\widehat{\beta}$.

The goal of matching is to change the distribution of the covariates in the sample of nontreated units so that it reproduces the distribution of the covariates among the treated. In order to do so it is necessary that the support of the matching variable X among the treated is a subset of the support of that variable among the nontreated.

Assumption 3.2 (Support condition)
Let $\mathcal{X}_1 = \text{supp}(X|W=1)$ *and* $\mathcal{X}_0 = \text{supp}(X|W=0)$*, then*

$$\mathcal{X}_1 \subseteq \mathcal{X}_0.$$

We now describe the population distribution targeted by the matched sample, $\mathcal{S}^*$. Let $P(\cdot|W=1)$ and $P(\cdot|W=0)$ be the *matching source* distributions of (Y,S) from where the treated and nontreated samples in $\mathcal{S}$ are respectively drawn, and let $E[\cdot|W=1]$ and $E[\cdot|W=0]$ be the corresponding expectation operators. For given $P(\cdot|W=1)$ and $P(\cdot|W=0)$ and a given number of matches, M, we define a *matching target* distribution, P^*, over the triple (Y,S,W), as follows:

$$P^*(W=1) = \frac{1}{1+M},$$

and for each measurable set, A,

$$P^*((Y,S) \in A|W=1) = P((Y,S) \in A|W=1),$$

and

$$P^*((Y,S) \in A|W=0) = E[P((Y,S) \in A|W=0,X)|W=1].$$

That is, in the matching target distribution: (i) treatment is assigned in the same proportion as in the matched sample; (ii) the distribution of outcomes (Y,S) among the treated is the

same as in the matching source; (iii) the distribution of outcomes (Y, S) among the nontreated is generated by integrating the conditional distribution of (Y, S) given X and $W = 0$ over the distribution of X given $W = 1$. As a result, under the matching target distribution, the distribution of X given $W = 0$ coincides with the distribution of X given $W = 1$.

Under regularity conditions stated below, estimation on the matched sample, $\mathcal{S}^*$, asymptotically recovers parameters of the matching target distribution, P^*, in which the treated and nontreated have the same distribution of X, but possibly different outcome and covariate distributions conditional on X. As a result, comparisons of outcomes between treated and nontreated in the matched sample, $\mathcal{S}^*$, produce the controlled contrasts of the Oaxaca-Blinder decomposition (Oaxaca, 1973; Blinder, 1973; and DiNardo *et al.*, 1996). More generally, under regularity conditions, regression coefficients of Y on Z in the matched sample, $\mathcal{S}^*$, asymptotically recover the analogous regression coefficients in the target population:

$$\begin{aligned} \beta &= \arg\min_{b \in \mathbb{R}^k} E^*[(Y - Z'b)^2] \\ &= (E^*[ZZ'])^{-1} E^*[ZY]. \end{aligned} \tag{3.2}$$

Matching methods are often motivated by a selection-on-observables assumption, that is, by the assumption that treatment assignment is as good as random conditional on the observables we match on. To formalize the assumption of selection on observables and its implications in our framework, consider source populations, expressed this time in terms of potential outcomes and covariates, $Q(\cdot|W = 1)$ and $Q(\cdot|W = 0)$, which represent the distributions of $(Y(1), Y(0), S)$ given $W = 1$ and $W = 0$, respectively. These distribution are defined in a way that $P(\cdot|W = 1)$ and $P(\cdot|W = 0)$ can be obtained by integrating out $Y(0)$ from $Q(\cdot|W = 1)$ and $Y(1)$ from $Q(\cdot|W = 0)$, respectively. For given $Q(\cdot|W = 1)$ and $Q(\cdot|W = 0)$, selection on observables means

$$(Y(1), Y(0), S)|X, W = 1 \sim (Y(1), Y(0), S)|X, W = 0$$

almost surely with respect to the distribution of $X|W = 1$. In words, the joint distribution of covariates and potential outcomes is independent of treatment assignment conditional on

the matching variables. Because in this chapter we focus on causal parameters defined for a population with distribution of the matching variables equal to $X|W=1$, for our purposes it is enough that the selection-on-observables assumption holds for the distribution of $(Y(0), S)$ only, that is,

$$(Y(0), S)|X, W=1 \sim (Y(0), S)|X, W=0. \tag{3.3}$$

Proposition 3.1 (Estimand under selection on observables)
Suppose that Assumption 3.2 holds and that β, as defined in Equation (3.2), exists and is finite. Then if selection on observables, as defined in Equation (3.3), holds, the coefficients in β are the same as the population coefficients that would be obtained from a regression of Y on Z in a setting where:

(a) $(Y(1), Y(0), S)$ has distribution $Q(\cdot|W=1)$,

(b) treatment is randomly assigned with probability $1/(M+1)$.

This result formalizes the notion that matching under selection on observables allows researchers to artificially reproduce an experimental setting under which average treatment effects can be easily evaluated through a least-squares regression of Y on Z. Notice, however, that all results in this chapter apply to the general estimand β in Equation (3.2), regardless of the validity of the selection-on-observables assumption.

3.2.3 Consistency and Asymptotic Normality

In this section, we will establish large sample properties of $\widehat{\beta}$. Throughout this chapter, we will first assume that the sum of matching discrepancies vanishes fast enough to allow asymptotic unbiasedness and root-n consistency:

Assumption 3.3 (Matching discrepancies)
As $N \to \infty$,

$$\frac{1}{\sqrt{N}} \sum_{i=1}^{N} W_i \sum_{j \in \mathcal{J}(i)} d(X_i, X_j) \xrightarrow{p} 0.$$

Abadie and Imbens (2012) derive primitive conditions on sampling the data-generating processes that guarantee Assumption 3.3 holds. Of course, in concrete empirical settings, the adequacy of matching should not rely on asymptotic results. Instead, the quality of the matches needs to be evaluated for each particular sample (e.g., using normalized differences as in Abadie and Imbens, 2011).

For any real matrix A, let $\|A\| = \sqrt{\text{tr}(A'A)}$ be the Euclidean norm of A. Next assumption collects regularity conditions on the conditional moments of (Y, Z) given (X, W).

Assumption 3.4 (Well-behavedness of conditional expectations)
For $w = 0, 1$, and some $\delta > 0$,

$$E[\|Z\|^4|W = w, X = x] \quad E[\|Z(Y - Z'\beta)\|^{2+\delta}|W = w, X = x]$$

are uniformly bounded on $\mathcal{X}_w$. Furthermore,

$$E[ZZ'|X = x, W = 0] \qquad E[ZY|X = x, W = 0] \qquad \text{Var}(Z(Y - Z'\beta)|X = x, W = 0)$$

are componentwise Lipschitz in x with respect to $d(\cdot, \cdot)$.

To ensure the existence of $\widehat{\beta}$ with probability approaching one as $n \to 0$, we assume invertibility of the Hessian, $H = E^*(ZZ')$. Notice that

$$H = \frac{E\Big[E[ZZ'|X, W = 1] + ME[ZZ'|X, W = 0]\big|W = 1\Big]}{1 + M}. \tag{3.4}$$

Assumption 3.5 (Linear independence of regressors)
H is invertible.

The next proposition establishes the asymptotic distribution of $\widehat{\beta}$.

Proposition 3.2 (Asymptotic distribution of the post-matching estimator)
Under Assumptions 3.1 to 3.5,

$$\sqrt{n}(\widehat{\beta} - \beta) \xrightarrow{d} \mathcal{N}(0, H^{-1}JH^{-1}),$$

where

$$J = \frac{\mathrm{Var}\Big(E[Z(Y - Z'\beta)|X, W = 1] + ME[Z(Y - Z'\beta)|X, W = 0]\big|W = 1\Big)}{1 + M} + \frac{E\Big[\,\mathrm{Var}(Z(Y - Z'\beta)|X, W = 1] + M\,\mathrm{Var}(Z(Y - Z'\beta)|X, W = 0)\big|W = 1\Big]}{1 + M}$$

and H is as defined in Equation (3.4).

All proofs are in Appendix B.3.

3.3 Post-Matching Standard Errors

In the previous section, we established that

$$\sqrt{n}(\widehat{\beta} - \beta) \xrightarrow{d} \mathcal{N}(0, H^{-1}JH^{-1})$$

for the post-matching OLS estimator $\widehat{\beta}$ that is obtained from a regression of Y on Z within the matched sample $\mathcal{S}^*$. In this section, our goal is to estimate the asymptotic variance,

$$H^{-1}JH^{-1},$$

in order to do inference on β.

3.3.1 OLS Standard Errors Ignoring the Matching Step

Ho *et al.* (2007) argue that matching can be seen as a preprocessing step, prior to estimation, so the matching step can be ignored in the calculation of standard errors. Here, we consider commonly applied Eicker–Huber–White (EHW or "sandwich") standard error estimates for iid data (Eicker, 1967; Huber, 1967; White, 1980a,b, 1982). EHW standard errors are robust to misspecification.

EHW standard errors can be computed as the square root of the main diagonal of the matrix $\widehat{H}^{-1}\widehat{J}_r\widehat{H}^{-1}/n$, where

$$\widehat{H} = \frac{1}{n}\sum_{i=1}^{n} Z_{ni}Z'_{ni} \tag{3.5}$$

and

$$\widehat{J}_r = \frac{1}{n}\sum_{i=1}^{n} Z_{ni}(Y_{ni} - Z'_{ni}\widehat{\beta})^2 Z'_{ni}. \tag{3.6}$$

The following proposition derives the probability limit of $\widehat{J}_r$ with data from a matched sample.

Proposition 3.3 (Convergence of J_r)

Suppose that Assumptions 3.1 to 3.5 hold. Assume also that

$$E[Z(Y - Z'\beta)^2 Z'|X = x, W = 0]$$

is Lipschitz on $\mathcal{X}_0$ and

$$E[Y^4|X = x, W = w]$$

is uniformly bounded on $\mathcal{X}_w$ for all $w \in \{0, 1\}$. Then, $\widehat{J}_r \xrightarrow{p} J_r$, where

$$J_r = \frac{E\Big[E[Z(Y - Z'\beta)^2 Z'|X, W = 1] + ME[Z(Y - Z'\beta)^2 Z'|X, W = 0]\Big|W = 1\Big]}{1 + M}.$$

Notice that $J_r = E^*[Z(Y - Z'\beta)^2 Z]$. That is, J_r is equal to the inner matrix of the EHW asymptotic variance when data are iid with distribution P^*. However, since the matched sample $\mathcal{S}^*$ is not an iid sample from P^*, $\widehat{J}_r$ is not generally consistent for J. The difference between the limit of the OLS standard errors $\widehat{H}^{-1}\widehat{J}_r\widehat{H}^{-1}$ and the actual asymptotic variance $H^{-1}JH^{-1}$ is given by $H^{-1}\Delta H^{-1}$, where

$$\Delta = \frac{-ME\big[\Gamma_0(X)\Gamma_1(X)' + \Gamma_1(X)\Gamma_0(X)'|W = 1\big] - (M - 1)ME\big[\Gamma_0(X)\Gamma_0(X)'|W = 1\big]}{M + 1},$$

and

$$\Gamma_w(x) = E\big[Z(Y - Z'\beta)|X = x, W = w\big],$$

for $w = 0, 1$.

Bias in the estimation of the variance can arise if the covariates in the regression are correlated with the error terms in the regression, conditional on the variables we have matched on, once we divide the sample between treated and control units. The following example provides a simple instance of this bias.

Example 3.1: *Inconsistency of OLS standard errors*

Assume the sample is drawn from

$$Y = \tau W + X + \varepsilon,$$

where X is scalar and also exogenous, $E[X] = E[\varepsilon] = 0$, and W and X are independent of ε. Assume that we match the values of X for N_1 treated units to N_1 untreated units ($M = 1$) without replacement. Let $j(i)$ be the index of the untreated observation that serves as a match for treated observation i. For simplicity, suppose that all matches are perfect, so $X_i = X_{j(i)}$, for every treated unit i. Within the matched sample, $\mathcal{S}^*$, we run a linear regression of Y on $Z = (1, W)'$ to obtain the regression coefficient on W,

$$\widehat{\tau} = \frac{1}{N_1} \sum_{i=1}^{N} W_i (Y_i - Y_{j(i)}).$$

$\widehat{\tau}$ is the usual matching estimator for the average effect of the treatment on the treated. Notice that $Y_i - Y_{j(i)} = \tau + \varepsilon_i - \varepsilon_{j(i)}$. Because the variation in X is taken care of through the matching, all variation in $\widehat{\tau}$ comes through the error terms ε_i, and we obtain

$$n \ \mathrm{Var}(\widehat{\tau}) = 4 \, \mathrm{Var}(\varepsilon).$$

Consider now the residuals of the OLS regression of Y_{ni} on a constant and W_{ni} in the matched sample:

$$\widehat{\varepsilon}_{ni} = Y_{ni} - \widehat{\mu} - \widehat{\tau} W_{ni} \approx X_{ni} + \varepsilon_{ni},$$

where $\widehat{\mu}$ is the intercept of the sample regression line. For this simple case, the OLS (EHW) variance estimator for $\widehat{\tau}$ is

$$n \, \widehat{\mathrm{Var}}(\widehat{\tau}) = \frac{4}{n} \sum_{i=1}^{n} \widehat{\varepsilon}_{ni}^2 \approx 4\big(\mathrm{Var}(X) + \mathrm{Var}(\varepsilon) \big).$$

As a result, EHW overestimate the variance of $\widehat{\tau}$. In this example, OLS standard errors do not take into account the correlation between regression residuals (generated by X), overestimating variance. □

Example 3.1 gives a compelling intuition for why we would generally expect that OLS standard errors are too large for post-matching estimators: If regression residuals are highly positively correlated between control units and their match (which seems plausible as long as unmodeled systematic heterogeneity in treatment is not too high), OLS standard errors for treatment or interaction-with-treatment terms will overestimate the true variation.

The following example shows, however, that OLS standard errors that ignore the matching step may also *underestimate* the variance.

Example 2: *Underestimation of the variance*
In the same setting as Example 3.1, assume that data is generated by

$$Y = \tau W + X - 2WX + \varepsilon.$$

The post-matching estimator of τ from a regression of Y on $(1, W)'$ is

$$\widehat{\tau} = \frac{1}{N_1} \sum_{i=1}^{n} W_i (Y_i - Y_{j(i)}).$$

In this case, $Y_i - Y_{j(i)} = \tau - 2X + \varepsilon_i - \varepsilon_{j(i)}$. Therefore,

$$n \operatorname{Var}(\widehat{\tau}) = 8 \operatorname{Var}(X) + 4 \operatorname{Var}(\varepsilon).$$

OLS standard errors are based on residuals,

$$\widehat{\varepsilon}_{ni} = Y_{ni} - \widehat{\mu} - \widehat{\tau} W_{ni} \approx X_i - 2W_{ni}X_{ni} + \varepsilon_{ni} = \begin{cases} X_{ni} + \varepsilon_{ni} & \text{if} \quad W_{ni} = 1, \\ -X_{ni} + \varepsilon_{ni} & \text{if} \quad W_{ni} = 0. \end{cases}$$

As a result, we obtain

$$n\widehat{\operatorname{Var}}(\widehat{\tau}) \approx 4 \operatorname{Var}(X) + 4 \operatorname{Var}(\varepsilon).$$

In this example, the OLS variance estimator does not take into account all the heterogeneity in the treatment effects, underestimating the variance of $\widehat{\tau}$. □

In both examples, OLS standard errors would be valid if we included the terms containing X in the post-matching regression. Indeed, OLS standard errors are generally valid if the regression is correctly specified in a specific sense defined in the following result.

Proposition 3.4 (Validity of OLS standard errors under correct specification)
Assume that the post-matching regression,

$$Y = Z'\beta + \varepsilon,$$

is correctly specified with respect to the conditional distribution of Y given (Z, X, W). That is, with $E[\varepsilon|Z, X, W] = 0$. Then, $J_r = J$, and the EHW variance estimator, $\widehat{H}^{-1}\widehat{J}_r\widehat{H}^{-1}$, is consistent for the asymptotic variance of $\sqrt{n}(\widehat{\beta} - \beta)$.

Note that correct specification is precisely the condition under which matching would not be required to create consistency with respect to the estimand β characterized by the idealized experiment laid out above, since a direct estimation without matching would do.

More importantly, notice that correct specification (in the sense defined above) of the post-matching regression is not required for consistent estimation of causal parameters. For example, under appropriate conditions, a simple difference in means between the treated and a matched sample of untreated units is consistent for the average effect of the treatment on the treated. Moreover, consistent estimators of the variance exist for the simple difference in means. These variance estimators are different from the OLS variance estimator, and do not rely on correct specification of the post-matching regression (see Abadie and Imbens, 2006).

3.3.2 Match-Level Clustered Standard Errors

By the previous section, we have shown that OLS standard errors are not generally valid for the post-matching least-squares estimator. In this section, we will demonstrate that when matching is done without replacement, clustered standard errors (Liang and Zeger, 1986; Arellano, 1987) can be employed to obtain valid estimates of the variance of post-matching regression coefficients. In particular, we will consider variance estimates that cluster at the level of the sets of units matched together in the first step.

Consider an estimator of the asymptotic variance of $\widehat{\beta}$ given by $\widehat{H}^{-1}\widehat{J}\widehat{H}^{-1}$, where $\widehat{H}$ is as in Equation (3.5) and $\widehat{J}$ is given by the clustered variance formula applied to the sets of units

matched together in the matching step,

$$\widehat{J} = \frac{1}{n}\sum_{i=1}^{n} W_i\Big(Z_i(Y_i - Z_i'\widehat{\beta}) + \sum_{j\in\mathcal{J}(i)} Z_j(Y_j - Z_j'\widehat{\beta})\Big) \times \Big(Z_i(Y_i - Z_i'\widehat{\beta}) + \sum_{j\in\mathcal{J}(i)} Z_j(Y_j - Z_j'\widehat{\beta})\Big)'.$$

Clustered standard errors can be readily implemented using standard statistical software. The next results shows that match-level clustered standard errors are valid in large samples for the post-matching estimator (provided matching is done without replacement).

Proposition 3.5 (Validity of clustered standard errors)
Under the assumptions of Proposition 3.3 we obtain that

$$\widehat{J} \xrightarrow{p} J.$$

In particular, the clustered estimator of the variance is consistent, i.e.,

$$\widehat{H}^{-1}\widehat{J}\widehat{H}^{-1} - n\,\mathrm{Var}(\widehat{\beta}) \xrightarrow{p} 0.$$

The intuition behind this result is that matching on covariates makes regression errors statistically dependent among units in the same match-set. Standard errors clustered at the level of the match-set take this dependency into account.

3.3.3 Matched Bootstrap

By the previous section, clustered standard errors are valid for the asymptotic variance of the post-matching estimator; however, they require calculation of $\widehat{H}, \widehat{J}$ as estimates of H, J.

In this section, we show that a clustered version of the nonparametric bootstrap (Efron, 1979) is also valid, and thereby offer a method of calculating standard errors that only requires the (repeated) calculation of regression coefficients. This bootstrap relies on a general result about the coupled resampling of martingale increments. Note that the nonparametric bootstrap on the full sample is not only computationally unattractive (as matches have to be calculated in each bootstrap iteration), but also invalid (Abadie and Imbens, 2008).

Recall that we reordered the observations in our sample, so that the first N_1 observations are the treated. Consider the nonparametric bootstrap that samples treated units together with their matching partners from $\mathcal{S}^*$ to obtain

$$\widehat{\beta}^* = \left(\frac{1}{n} \sum_{i=1}^{n} V_{ni} Z_{ni} Z_{ni}' \right)^{-1} \frac{1}{n} \sum_{i=1}^{n} V_{ni} Z_{ni} Y_{ni}$$

where $(V_{n1}, \ldots, V_{nN_1})$ has a multinomial distribution with parameters $(N_1, (1/N_1, \ldots, 1/N_1))$, and $V_{nj} = V_{ni}$ if $j > N_1$ and $j \in \mathcal{J}(i)$. In this bootstrap procedure, N_1 units are drawn at random with replacement from the N_1 treated sample units. Untreated units are drawn along with their treated match. Effectively, the matched bootstrap samples matched sets of one treated unit and M untreated units. The next proposition shows validity of the matched bootstrap.

Proposition 3.6 (Validity of the matched bootstrap)

Under the assumptions of Proposition 3.5, we have that

$$\sup_{r \in \mathbb{R}^s} \left| P\left(\sqrt{n}(\widehat{\beta}^* - \widehat{\beta}) \leq r \middle| \mathcal{S} \right) - P(\mathcal{N}(0, H^{-1} J H^{-1}) \leq r) \right| \xrightarrow{p} 0.$$

The proof of this proposition relies on a general result on the coupled resampling or martingale increments, which can be found in Appendix B.2.

Proposition 3.6 shows that the bootstrap distribution provides an asymptotically valid approximation of the limiting distribution of the post-matching estimator, but that does not necessarily imply that the associated bootstrap variance is an asymptotically valid estimate of the variance of the estimator. Indeed, the analysis of the bootstrap variance is complicated by the fact that, in forming the bootstrap estimate $\widehat{\beta}^*$, the empirical analog

$$\widehat{H}^* = \frac{1}{n} \sum_{i=1}^{n} V_{ni} Z_{ni} Z_{ni}'$$

of the Hessian H for a given bootstrap draw may be badly conditioned or even non-invertible, which happens with positive probability at any given sample size. To circumvent this issue,

we fix constants $c > 0$ and $\alpha \in (0, 1/2)$ and consider the alternative bootstrap estimator

$$\tilde{\beta}^* = \begin{cases} \widehat{\beta}^*, & \|\widehat{H}^* - \widehat{H}\| \leq \frac{c}{n^\alpha} \\ \widehat{\beta}, & \text{otherwise} \end{cases}$$

where $\|\cdot\|$ denotes the Frobenius norm. In other words, this modified bootstrap estimator coincides with the matched bootstrap estimator whenever the bootstrap Hessian $\widehat{H}^*$ is close to the empirical analog of the Hessian $\widehat{H}$ in the full matched sample, and equals the original post-matching estimator for other bootstrap draws. The threshold is chosen such that, as the sample size grows, the two bootstrap estimators coincide with probability approaching one.

We conjecture that $\tilde{\beta}^*$ allows for valid inference in large samples, including the consistent estimation of standard errors:

Conjecture 3.1 (Validity of the alternative bootstrap and bootstrap standard errors)
Under the assumptions of Proposition 3.5, the alternative bootstrap given by $\tilde{\beta}^*$ is valid in the sense of Proposition 3.6, and yields a valid estimate of the asymptotic variance of $\widehat{\beta}$, i.e.

$$n \operatorname{Var}(\tilde{\beta}^* | \mathcal{S}) \xrightarrow{p} H^{-1} J H^{-1}$$

as $n \to \infty$.

3.4 Simulations

In this section, we study the performance of the post-matching standard error estimators from Section 3.3 in a simulation exercise.

3.4.1 Setup I: Robustness to Misspecification

We generate data according to

$$Y = WX + 5X^2 + \varepsilon$$

with

$$X|W=1 \sim \mathcal{U}(0,1), \qquad X|W=0 \sim \mathcal{U}(0,1.5), \qquad \varepsilon \sim \mathcal{N}(0,\sigma^2),$$

and $\mathcal{U}(a,b)$ is the Uniform distribution on $[a,b]$. Notice that the distribution of the covariates, X, differs between treatment and control groups. We sample $N_1 = 40$ treated and $N_0 = 160$ nontreated units. We first match treated and untreated units on the covariates, X, without replacement and with $M = 1$ match per treated unit. We consider the following post-matching regression specifications.

Specification 1:

$$Y = \alpha + \tau_0 W + \tau_1 WX + \beta_1 X + \varepsilon$$

Specification 2:

$$Y = \alpha + \tau_0 W + \tau_1 WX + \beta_1 X + \beta_2 X^2 + \varepsilon$$

Specification 1 includes X as a linear control, but is misspecified, while Specification 2 is correct relative to the conditional expectation $E[Y|W,X]$. Regardless of whether the specification is correct or not, it can always be seen as an L_2 approximation to $E[Y|W,X]$ (see, e.g., White, 1980b). We focus on the regression coefficients τ_0 and τ_1.

Table 3.1 reports the results of the simulation exercise. In a regression within the full sample without matching, the estimates of τ_0 and τ_1 are biased under misspecification (Specification 1), while they are valid under correct specification (Specification 2). After matching, both specifications yield valid estimates for τ_0 and τ_1. However, OLS standard error estimates are inflated under misspecification, while average clustered and matched bootstrap standard errors are close to the standard deviation of the estimators. Under correct specification, all standard error estimates perform well.

This simulation exercise shows the role of matching before running the regression in obtaining valid estimates even under misspecification, and confirms our theoretical results. OLS standard errors are not robust to misspecification, while clustered and matched bootstrap standard error estimates are.

Table 3.1: *Simulation results for 10,000 Monte Carlo iterations for Setup I in Section 3.4.1*

(a) *Target parameter: Coefficient $\tau_0 = 0$ on W*

	full sample		post-matching		average standard error		
Specification	$E[\widehat{\tau}_0]$	std$(\widehat{\tau}_0)$	$E[\widehat{\tau}_0]$	std$(\widehat{\tau}_0)$	OLS	cluster	bootstrap
1	1.03	.237	0.00	.116	.225	.109	.112
2	0.00	.093	0.00	.115	.108	.108	.111

(b) *Target parameter: Coefficient $\tau_1 = 1$ on the interaction WX*

	full sample		post-matching		average standard error		
Specification	$E[\widehat{\tau}_1]$	std$(\widehat{\tau}_1)$	$E[\widehat{\tau}_1]$	std$(\widehat{\tau}_1)$	OLS	cluster	bootstrap
1	−1.49	.366	1.00	.199	.406	.190	.195
2	1.00	.159	1.00	.198	.187	.188	.194

3.4.2 Setup II: High Treatment-Effect Heterogeneity

In the simulation in the previous section, OLS standard errors overestimate the variation of the post-matching estimator under misspecification. In this section, we present an example in which OLS standard errors are too low. We generate data according to

$$Y = WX + 20W(X - .5)^2 - 10(X - .5)^2 + \varepsilon$$

with $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ as above. According to this data-generating process, the conditional treatment effects is non-linear with

$$E[Y|W = 1, X] - E[Y|W = 0, X] = X + 20(X - .5)^2.$$

Sample sizes, matching settings, and regression specifications are as in Setup I. Notice that both regression specifications are now misspecified, as they cannot capture non-linear conditional treatment effects. Like in Section 3.4.1, regression coefficients represent the parameters of an L_2 approximation to $E[Y|W, X]$ over the distribution of (W, X) in Proposition 3.1. Direct calculations yield $\tau_0 = 5/3$ and $\tau_1 = 1$ for both specifications.

Table 3.2: *Simulation results for 10,000 Monte Carlo iterations for Setup II in Section 3.4.2*

(a) *Target parameter: Coefficient* $\tau_0 = 1.63$ *on* W

	full sample		post-matching		average standard error		
Specification	$E[\widehat{\tau}_1]$	std$(\widehat{\tau}_1)$	$E[\widehat{\tau}_1]$	std$(\widehat{\tau}_1)$	OLS	cluster	bootstrap
1	−0.44	.449	1.63	.598	.409	.567	.598
2	1.45	.570	1.63	.597	.408	.567	.598

(b) *Target parameter: Coefficient* $\tau_1 = 1$ *on the interaction* WX

	full sample		post-matching		average standard error		
Specification	$E[\widehat{\tau}_1]$	std$(\widehat{\tau}_1)$	$E[\widehat{\tau}_1]$	std$(\widehat{\tau}_1)$	OLS	cluster	bootstrap
1	5.99	0.68	1.01	1.09	0.74	1.03	1.09
2	1.45	1.05	1.01	1.09	0.74	1.03	1.09

Table 3.2 presents the results of the simulation exercise under this alternative setup. Unlike in the previous simulation study, naive OLS standard errors that ignore the matching step now underestimate the variation of the post-matching estimator, driven by the large heterogeneity in conditional treatment effects that is not captured by either regression specification. As our theoretical results suggest, our proposed robust standard error estimates approximate the true variation adequately on average. Note that the different specifications now change inference without matching, but are essentially equivalent in post-matching regression given that the estimand β_2 is zero in the second specification.

3.5 Application

This section reports the results of an empirical application where we look at the effect of smoking on the pulmonary function of youth. The application is based on data originally collected in East Boston by Tager *et al.* (1979, 1983), and subsequently described and analyzed in Rosner (1995) and Kahn (2005). The sample contains 654 youth, $N_1 = 65$ who have ever smoked regularly ($W = 1$) and $N_0 = 589$ who never smoked regularly ($W = 0$). The outcome

of interest is the subjects' forced expiratory volume (Y), ranging from 0.791 to 5.793 liters per second (ℓ/sec). In addition, we use data on age (X_1, ranging from 3 to 19 with the youngest ever-smoker aged 9) and gender (X_2, with $X_2 = 1$ for males and $X_2 = 0$ for females).

The use of matching to study the causal effect of smoking is motivated by the likely confounding effects of age and gender. For instance, while the causal effect of smoking on respiratory volume is expected to be negative, older children smoke more and have a larger respiratory volume, which may result in a positive association between smoking and respiratory volume in this sample.

We first match every smoker in the sample to one non-smoker ($M = 1$), without replacement, based on age (X_1) and gender (X_2). Within the resulting matched sample of 65 smokers and 65 non-smokers, we run linear regressions with the following specifications:

Specification 1:

$$Y = \alpha + \tau_0 W + \varepsilon.$$

Specification 2:

$$Y = \alpha + \tau_0 W + \beta_1 X_1 + \beta_2 X_2 + \varepsilon.$$

Specification 3:

$$\begin{aligned} Y = \alpha + \tau_0 W + \tau_1 W(X_1 - E[X_1]) + \tau_2 W(X_2 - E[X_2]) \\ + \beta_1(X_1 - E[X_1]) + \beta_2(X_2 - E[X_2]) + \varepsilon. \end{aligned}$$

The first specification yields the matching estimator for the average treatment effect τ_0 as the regression coefficient on W, while the second adds linear controls ins X_1 and X_2. The third specification also includes linear interaction effects with age and gender.

The regression coefficients τ_0 on treatment W as well as the linear interactions τ_1 of treatment with age and τ_2 of treatment with gender for the three specifications (where applicable) are reported in Table 3.3. The first specification demonstrates the confounding problem: without controlling for age and gender, there appears to be a positive effect of smoking on respiratory function. After controlling for age and gender using matching the sign

Table 3.3: *OLS and post-matching estimates of τ_0 in Specifications 1, 2, 3 in Section 3.5*

	smoker			smoker×age			smoker×male		
	coeff.	st. error		coeff.	st. error		coeff.	st. error	
		OLS	clust		OLS	clust		OLS	clust
Specification 1:									
OLS	.711	.099							
post-matching	−.066	.132	.095						
Specification 2:									
OLS	−.154	.104							
post-matching	−.077	.104	.096						
Specification 3:									
OLS	.495	.187		−.182	.036		.461	.193	
post-matching	−.077	.102	.093	−.092	.054	.038	−.021	.249	.212

of the treatment coefficient is reversed. In this specification, the clustered standard error is considerably smaller than the OLS standard error.

Once we include linear controls (Specification 2), the sign of the main effect is not affected by matching any more, and clustered standard errors are similar to OLS standard errors. Both findings are consistent with our regression specification moving closer towards correct specification. Notice, however, that the magnitude of the OLS estimate with Specification 2 is double that of the magnitude with Specification 1, while the magnitude of the post-matching estimate stays roughly constant. This result illustrates the higher robustness across specifications of the post-matching estimator relative to OLS.

In Specification 3, which includes full interactions, both the use of matching and the use of robust standard errors may matter for qualitative conclusions. First, notice that the coefficient on the interaction of gender with treatment is large, significant and positive without matching, suggesting that the effect of smoking is more severe for girls than for boys. After matching, the sign flips, and the estimated effect is small and insignificant. This suggests that the interaction finding with OLS is driven by misspecification. Second, in the post-matching regression we find a negative estimate for the interaction of treatment with age. With OLS standard errors, this effect is not significant (at the 5% level); the robust clustered standard error, as our theory suggests, is smaller and lets us reject a zero interaction coefficient.

3.6 Conclusion

This chapter establishes valid inference in linear regression after nearest-neighbor matching without replacement. OLS standard errors that ignore the matching step are not generally valid if the regression specification is incorrect relative to the expectation of the outcome conditional on the treatment and the matching covariates. Notice, however, that using a correct specification relative to $E[Y|W, X]$ is not necessary to consistently estimate treatment parameters after matching. For example, a simple difference in means can identify the average treatment effect in a matched sample.

We propose two alternatives – standard errors clustered at the match level and an analogous block bootstrap – that are robust to misspecification and easily implementable with standard statistical software. A simulation study and an empirical example demonstrate the usefulness of our results.

Our analysis remains limited in three ways: First, we discuss only matching without replacement, and our results do not directly carry over to matching with replacement as in Abadie and Imbens (2006). Second, our analysis assumes that the quality of matches is good enough for matching discrepancies not to bias the asymptotic distribution of the post-matching regression estimator. Post-matching regression adjustments may, in practice, help eliminating the bias as in the bias-corrected matching estimator in Abadie and Imbens (2011). This is an angle that we do not explore in this chapter and an interesting avenue for future research. Third, the results presented in this chapter are formulated for post-matching linear least squares only, but we believe that they carry over to more general classes of regression techniques, specifically M-estimators.

References

Abadie, A. and Imbens, G. (2006). Large sample properties of matching estimators for average treatment effects. *Econometrica*, **74** (1), 235–267.

— and — (2008). On the failure of the bootstrap for matching estimators. *Econometrica*, **76** (6), 1537–1557.

— and — (2011). Bias-corrected matching estimators for average treatment effects. *Journal of Business & Economic Statistics*, **29** (1), 1–11.

— and — (2012). A martingale representation for matching estimators. *Journal of the American Statistical Association*, **107** (498), 833–843.

—, Imbens, G. W. and Zheng, F. (2014). Inference for misspecified models with fixed regressors. *Journal of the American Statistical Association*, **109** (508), 1601–1614.

Anderson, M. L. and Magruder, J. (2017). Split-sample strategies for avoiding false discoveries. *NBER working paper*.

Andrews, I. and Armstrong, T. B. (2017). Unbiased instrumental variables estimation under known first-stage sign. *Quantitative Economics*, **8** (2), 479–503.

— and Kasy, M. (2017). Identification of and correction for publication bias. *NBER working paper*.

Angrist, J. D., Imbens, G. W. and Krueger, A. B. (1999). Jackknife instrumental variables estimation. *Journal of Applied Econometrics*, **14** (1), 57–67.

Arellano, M. (1987). Computing robust standard errors for within-groups estimators. *Oxford Bulletin of Economics and Statistics*, **49** (4), 431–434.

Aronow, P. M. and Middleton, J. A. (2013). A class of unbiased estimators of the average treatment effect in randomized experiments. *Journal of Causal Inference*, **1** (1), 135–154.

Athey, S. and Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, **113** (27), 7353–7360.

Balzer, L. B., van der Laan, M. J., Petersen, M. L. and the SEARCH Collaboration (2016). Adaptive pre-specification in randomized trials with and without pair-matching. *Statistics in Medicine*, **35** (25), 4528–4545.

Baranchik, A. J. (1973). Inadmissibility of maximum likelihood estimators in some multiple regression problems with three or more independent variables. *The Annals of Statistics*, **1** (2), 312–321.

Belloni, A. and Chernozhukov, V. (2013). Least squares after model selection in high-dimensional sparse models. *Bernoulli*, **19** (2), 521–547.

—, — and Hansen, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, **81** (2), 608–650.

Blinder, A. S. (1973). Wage discrimination: Reduced form and structural estimates. *Journal of Human Resources*, **8** (4), 436–455.

Bloniarz, A., Liu, H., Zhang, C.-H., Sekhon, J. S. and Yu, B. (2016). Lasso adjustments of treatment effect estimates in randomized experiments. *Proceedings of the National Academy of Sciences*, **113** (27), 7383–7390.

Brodeur, A., Lé, M., Sangnier, M. and Zylberberg, Y. (2016). Star Wars: The empirics strike back. *American Economic Journal: Applied Economics*, **8** (1), 1–32.

Carroll, G. (2015). Robustness and linear contracts. *The American Economic Review*, **105** (2), 536–563.

Chamberlain, G. (2007). Decision theory applied to an instrumental variables model. *Econometrica*, **75** (3), 609–652.

— and Moreira, M. J. (2009). Decision theory applied to a linear panel data model. *Econometrica*, **77** (1), 107–133.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C. and Newey, W. (2017). Double/debiased/Neyman machine learning of treatment effects. *American Economic Review*, **107** (5), 261–65.

—, —, —, —, —, — and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, **21** (1), C1–C68.

Cochran, W. G. (1953). Matching in analytical studies. *American Journal of Public Health and the Nation's Health*, **43** (6 Pt 1), 684–691.

— (1972). Observational studies. *Reprinted in Observational Studies, 2015*, **1**, 126–136.

Coffman, L. C. and Niederle, M. (2015). Pre-analysis plans have limited upside, especially where replications are feasible. *Journal of Economic Perspectives*, **29** (3), 81–98.

Dahl, F. A., Grotle, M., Šaltytė Benth, J. and Natvig, B. (2008). Data splitting as a countermeasure against hypothesis fishing: with a case study of predictors for low back pain. *European Journal of Epidemiology*, **23** (4), 237–242.

Dehejia, R. H. and Wahba, S. (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association*, **94** (448), 1053–1062.

Diebold, F. X. (2015). Comparing predictive accuracy, twenty years later: A personal perspective on the use and abuse of Diebold–Mariano tests. *Journal of Business & Economic Statistics*, **33** (1), 1–9.

DiNardo, J., Fortin, N. and Lemieux, T. (1996). Labor market institutions and the distribution of wages, 1973-1992: A semiparametric approach. *Econometrica*, **64** (5), 1001–1044.

Dwork, C., Feldman, V., Hardt, M., Pitassi, T., Reingold, O. and Roth, A. (2015). The reusable holdout: Preserving validity in adaptive data analysis. *Science*, **349** (6248), 636–638.

Efron, B. (1979). Bootstrap methods: Another look at the Jackknife. *The Annals of Statistics*, **7** (1), 1–26.

Eicker, F. (1967). Limit theorems for regressions with unequal and dependent errors. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, pp. 59–82.

Fafchamps, M. and Labonne, J. (2016). Using split samples to improve inference about causal effects. *NBER working paper*.

Ferguson, T. S. (1967). *Mathematical Statistics: A Decision Theoretic Approach*. Academic Press.

Frankel, A. (2014). Aligned delegation. *American Economic Review*, **104** (1), 66–83.

Freedman, D. A. (2008). On regression adjustments to experimental data. *Advances in Applied Mathematics*, **40** (2), 180–193.

Glaeser, E. L. (2006). Researcher incentives and empirical methods. *NBER working paper*.

Hahn, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, **66** (2), 315.

Hájek, J. (1962). Asymptotically most powerful rank-order tests. *The Annals of Mathematical Statistics*, pp. 1124–1147.

Hansen, B. E. (2007). Least squares model averaging. *Econometrica*, **75** (4), 1175–1189.

— (2016). Efficient shrinkage in parametric models. *Journal of Econometrics*, **190** (1), 115–132.

— (2017). Stein-like 2SLS estimator. *Econometric Reviews*, **36** (6-9), 840–852.

— and Racine, J. S. (2012). Jackknife model averaging. *Journal of Econometrics*, **167** (1), 38–46.

Heckman, J. J. and Singer, B. (2017). Abducting economics. *American Economic Review*, **107** (5), 298–302.

Hirano, K. and Porter, J. R. (2015). Location properties of point estimators in linear instrumental variables and related models. *Econometric Reviews*, **34** (6-10), 720–733.

— and Wright, J. H. (2017). Forecasting with model uncertainty: Representations and risk reduction. *Econometrica*, **85** (2), 617–643.

Ho, D. E., Imai, K., King, G. and Stuart, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis*, **15** (3), 199–236.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, **81** (396), 945.

Holmström, B. R. (1978). *On Incentives and Control in Organizations*. Ph.D. thesis, Stanford University.

— (1984). On the theory of delegation. In *Bayesian Models in Economic Theory*.

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, **47** (260), 663–685.

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, pp. 221–233.

Hurwicz, L. and Shapiro, L. (1978). Incentive structures maximizing residual gain under incomplete information. *The Bell Journal of Economics*, pp. 180–191.

Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics*, **86** (1), 4–29.

James, W. and Stein, C. (1961). Estimation with quadratic loss. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, pp. 361–379.

Kahn, M. (2005). An exhalent problem for teaching statistics. *The Journal of Statistical Education*, **13** (2).

Kasy, M. (2016). Why experimenters might not always want to randomize, and what they could do instead. *Political Analysis*, **24** (03), 324–338.

Leamer, E. E. (1974). False models and post-data model construction. *Journal of the American Statistical Association*, **69** (345), 122–131.

— (1978). *Specification Searches: Ad Hoc Inference with Nonexperimental Data*. Wiley.

Lehmann, E. L. and Romano, J. P. (2006). *Testing Statistical Hypotheses*. Springer Science & Business Media.

Lenz, G. and Sahn, A. (2017). Achieving statistical significance with covariates. *BITSS preprint*.

Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, **73** (1), 13–22.

Lin, W. (2013). Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique. *The Annals of Applied Statistics*, **7** (1), 295–318.

Moser, S. M. (2008). Expectations of a noncentral chi-square distribution with application to iid MIMO Gaussian fading. In *Information Theory and Its Applications*, IEEE, pp. 1–6.

Mullainathan, S. and Spiess, J. (2017). Machine learning: An applied econometric approach. *Journal of Economic Perspectives*, **31** (2), 87–106.

Neyman, J. (1923). On the application of probability theory to agricultural experiments. *Translated in Statistical Science, 1990*, **5** (4), 465–472.

Oaxaca, R. (1973). Male-female wage differentials in urban labor markets. *International Economic Review*, **14** (3), 693–709.

Olken, B. A. (2015). Promises and perils of pre-analysis plans. *Journal of Economic Perspectives*, **29** (3), 61–80.

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science*, **349** (6251), aac4716–1–aac4716–8.

Pauly, M. (2011). Weighted resampling of martingale difference arrays with applications. *Electronic Journal of Statistics*, **5**, 41–52.

Piegorsch, W. W. and Casella, G. (1985). The existence of the first negative moment. *The American Statistician*, **39** (1), 60.

Robins, J. M. and Rotnitzky, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, **90** (429), 122.

Rosner, B. (1995). *Fundamentals of Biostatistics*. Duxbury Press.

Rubin, D. B. (1973). Matching to remove bias in observational studies. *Biometrics*, **29** (1), 159–183.

— (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, **66** (5), 688.

— (1975). Bayesian inference for causality: The importance of randomization. In *Proceedings of the Social Statistics Section of the American Statistical Association*, pp. 233–239.

— (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics*, **6** (1), 34–58.

Schorfheide, F. and Wolpin, K. I. (2012). On the use of holdout samples for model selection. *American Economic Review*, **102** (3), 477–81.

— and — (2016). To hold out or not to hold out. *Research in Economics*, **70** (2), 332–345.

Sclove, S. L. (1968). Improved estimators for coefficients in linear regression. *Journal of the American Statistical Association*, **63** (322), 596.

Spiess, J. (2017a). Bias reduction in instrumental variable estimation through first-stage shrinkage. *arXiv preprint arXiv:1708.06443*.

— (2017b). Unbiased shrinkage estimation. *arXiv preprint arXiv:1708.06436*.

Stein, C. M. (1981). Estimation of the mean of a multivariate Normal distribution. *The Annals of Statistics*, **9** (6), 1135–1151.

Sterling, T. D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance–or vice versa. *Journal of the American Statistical Association*, **54** (285), 30.

Tager, I. B., Weiss, S. T., Muñoz, A., Rosner, B. and Speizer, F. E. (1983). Longitudinal study of the effects of maternal smoking on pulmonary function in children. *New England Journal of Medicine*, **309** (12), 699–703.

—, —, Rosner, B. and Speizer, F. E. (1979). Effect of parental cigarette smoking on the pulmonary function of children. *American Journal of Epidemiology*, **110** (1), 15–26.

Tullock, G. (1959). Publication decisions and tests of significance—a comment. *Journal of the American Statistical Association*, **54** (287), 593–593.

Wager, S. and Athey, S. (2017). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*.

—, Du, W., Taylor, J. and Tibshirani, R. J. (2016). High-dimensional regression adjustments in randomized experiments. *Proceedings of the National Academy of Sciences*, **113** (45), 12673–12678.

Wald, A. (1950). *Statistical Decision Functions*. Wiley.

White, H. (1980a). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, **48** (4), 817–838.

— (1980b). Using least squares to approximate unknown regression functions. *International Economic Review*, **21** (1), 149–170.

— (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, **50** (1), 1–25.

Wu, E. and Gagnon-Bartsch, J. (2017). The LOOP estimator: Adjusting for covariates in randomized experiments. *arXiv preprint arXiv:1708.01229*.

Appendix A

Appendix to Chapter 1

This appendix builds up to the proofs of the main results (Theorem 1.1, Lemma 1.1, Theorem 1.2). Throughout, I restate the relevant claims from the main chapter with their original numbering.

A.1 Minimax Optimality of Fixed Bias

Lemma 1.2 (Unbiasedness aligns estimation)**.** *If the investigator has risk from $\mathcal{R}^*$ then the investigator will choose from the unbiased estimators $\mathcal{C}^*$ according to the designer's preferences.*

Proof of Lemma 1.2. Take any investigator risk function $r^I \in \mathcal{R}^*$, unbiased estimator $\hat{\tau} \in \mathcal{C}^*$, and prior $\pi \in \Delta(\Theta)$. ($\Delta(\Theta)$ denotes the unit $|\theta| - 1$-simplex in $\mathbb{R}^\Theta$.) Then, the designer's average risk is

$$\begin{aligned}
&\mathrm{E}_\pi[r_\theta^D(\hat{\tau})] \\
\overset{r^I \in \mathbb{R}^*}{=}\ &\mathrm{E}_\pi[(\hat{\tau}(z) - \tilde{\tau}_\theta)^2] \\
=\ &\mathrm{E}_\pi[((\hat{\tau}(z) - \mathrm{E}_\theta[\hat{\tau}(z)]) - (\mathrm{E}_\theta[\hat{\tau}(z)] - \tilde{\tau}_\theta))^2] \\
=\ &\mathrm{E}_\pi[(\hat{\tau}(z) - \mathrm{E}_\theta[\hat{\tau}(z)])^2] + \mathrm{E}_\pi[(\mathrm{E}_\theta[\hat{\tau}(z)] - \tilde{\tau}_\theta)^2] \\
\overset{\hat{\tau} \in \mathcal{C}^*}{=}\ &\mathrm{E}_\pi[\mathrm{Var}_\theta(\hat{\tau}(z))] + \mathrm{E}_\pi[(\tau_\theta - \tilde{\tau}_\theta)^2]
\end{aligned}$$

by a bias-variance decomposition. (I conflate P_θ into P_π.) Since $\mathrm{E}_\pi[(\tau_\theta - \tilde{\tau}_\theta)^2]$ is constant

with respect to $\hat{\tau}$ and $\mathrm{E}_\pi[\mathrm{Var}_\theta(\hat{\tau}(z))]$ does not vary with $\tilde{\tau}$, the estimation target $\tilde{\tau}$ does not affect the choice of the estimator from $\mathcal{C}^*$. Hence, choices are as if $\tilde{\tau} = \tau$. The investigator chooses from $\mathcal{C}^*$ according to the designer's risk r^D. □

Theorem 1.1 (Fixed bias is minimax optimal)**.** *Write $\Delta^*(\Theta)$ for all distributions over Θ with full support. For every hyperprior η with support within $\Delta^*(\Theta)$ there is a set of biases $\beta^\eta : \Theta \to \mathbb{R}$ such that the fixed-bias restriction*

$$\mathcal{C}^\eta = \{\hat{\tau} : \mathcal{Z} \to \mathbb{R}; \mathrm{E}_\theta[\hat{\tau}] = \tau_\theta + \beta^\eta_\theta\}$$

is a minimax optimal mechanism in the sense of Definition 1.1, i.e.

$$\mathcal{C}^\eta \in \arg\min_{\mathcal{C}} \sup_{r^I \in \mathcal{R}^*} \mathrm{E}_\eta \left[r^D_\theta \left(\arg\min_{\hat{\tau} \in \mathcal{C}} \mathrm{E}_\pi[r^I_\theta(\hat{\tau})] \right) \right].$$

Proof of Theorem 1.1. I apply the strategy from Theorem 1 in Frankel (2014) to establish that the unbiasedness restriction yields a minimax (maxmin in utility terms) optimal mechanism. Relative to the quadratic-loss constant-bias setup in Frankel (2014), average risk yields weighted sums where the prior changes weights and the bias changes across decisions (sample draws) and states (posterior expectations). Rather than using Lemma 3 on quadratic-loss constant-bias utilities in Frankel (2014) as stated there, I therefore appeal directly to the logic of his more general Theorem 1, which I extend to deal with the non-compact type and action spaces in my application.

The agent's (investigator's) actions are the estimates $\hat{\tau}(z)$ at all $N = (2|\mathcal{Y}|)^n$ sample points $z \in \mathcal{Z}$. (I assume that the covariates x are already known when the investigator commits to their estimator.) The state that only the agent observes is the investigator's prior $\pi \in \Delta(\Theta)$. π is drawn from the (hyper-)prior η.

In the parlance of Frankel (2014), I consider the Φ-moment mechanisms where the agent chooses from estimators

$$\mathcal{C}_\beta = \{\hat{\tau} : \mathcal{Z} \to \mathbb{R}; \mathrm{E}_\theta[\hat{\tau}] = \tau_\theta + \beta_\theta \forall \theta \in \Theta\}$$

for a set of fixed biases $\beta \in \mathbb{R}^\Theta$. (Each expectation – a weighted sum over actions $\hat{\tau}(z)$ – is a

map from actions to real numbers.) To show that this mechanism is maxmin optimal for some choice of β, I establish that:

1. Any feasible such Φ-moment mechanism (i.e. any bias vector β with $\mathcal{C}_\beta \neq \emptyset$) induces aligned delegation over $\mathcal{R}^*$, that is, subject to the restriction $\hat{\tau} \in \mathcal{C}_\beta$ agents of all risk types $r^I \in \mathcal{R}^*$ choose as if they were of risk type r^D.

2. $\mathcal{R}^*$ is Φ-rich, that is, for any mechanism there exists some $\overline{\beta} \in \mathbb{R}^\Theta$ and a sequence of risk types $(r^{I_k})_{k=1}^\infty \in (\mathcal{R}^*)^\mathbb{N}$ such that for all realized $\pi \in \Delta^*(\Theta)$ and all corresponding sequences $(\hat{\tau}_k)_{k=1}^\infty$ of chosen estimators, $\lim_{k\to\infty} \mathrm{E}_\theta[\hat{\tau}_k(z)] = \tau_\theta + \overline{\beta}_\theta$ for all θ in the support of π. (Unlike Frankel (2014) I do not explicitly consider mixed strategies since randomized estimators are dominated in my setting.)

Similar to Frankel's (2014) Theorem 1, the restriction $\mathcal{C}_\beta$ is then minimax optimal provided that β is chosen to minimize the designer's average risk, for some distribution (hyperprior) η over π. I will develop this deduction below for my specific case (in which type and action spaces are not compact) once I have established aligned delegation and richness.

1. Aligned delegation. For $\beta \in \mathbb{R}^\Theta$ such that $\mathcal{C}_\beta \neq \emptyset$, the average over risk $r^I \in \mathcal{R}^*$ for an estimator $\hat{\tau} \in \mathcal{C}_\beta$ over the prior $\pi \in \Delta(\Theta)$ is

$$\mathrm{E}_\pi r^I_\theta(\hat{\tau}) = \mathrm{E}_\pi[\mathrm{Var}_\theta(\hat{\tau}(z))] + \mathrm{E}_\pi[(\tau_\theta + \beta_\theta - \tilde{\tau}_\theta)^2]$$

as in the proof of Lemma 1.2. Hence, choices do not vary with the risk type of the investigator and are as if the investigator shared the designer's risk function r^D.

2. Richness. For some arbitrary, but fixed mechanism, our goal is to find a vector of biases $\bar{\beta}$ and a risk sequence $r^{I_1}, r^{I_2}, \ldots$ such that biases of mechanism outcomes along this sequence always converge to $\bar{\beta}$. I first justify assumptions on the mechanism, then pick a bias vector $\bar{\beta}$, and finally construct a suitable sequence of risk types that ensures bias convergence.

For some conformal mechanism, consider the set $\mathcal{C} \subseteq \mathbb{R}^\mathcal{Z}$ of estimators $\hat{\tau}$ that are outcomes for some investigator risk function $r^I \in \mathcal{R}^*$ and prior π in the support of η. Note that the

outcomes of the mechanism are the investigator choices

$$\hat{\tau}_\pi(r^I) \in \arg\min_{\hat{\tau}\in\mathcal{C}} \mathrm{E}_\pi r^I_\theta(\hat{\tau}) \tag{A.1}$$

where by assumption ties are broken in favor of the designer. I first show that $\mathcal{C}$ in (A.1) is wlog closed. Since the minimizers are already included in $\mathcal{C}$, taking the closure of $\mathcal{C}$ does not change investigator risk at their optimal choices. Replacing $\mathcal{C}$ by its closure thus does not affect investigator risk at choices (A.1), and can only improve outcomes for the designer, since additional ties are broken in their favor. For the analysis of minimax optimal mechanisms, we can therefore assume wlog that $\mathcal{C}$ is closed.

I first assume that $\mathcal{C}$ is also bounded. Define the set

$$\mathcal{D} = \{\theta \mapsto \mathrm{E}_\theta[\hat{\tau}(z)]; \hat{\tau} \in \mathcal{C}\} \subseteq \mathbb{R}^\Theta$$

of vectors of expectations achieved by estimators in $\mathcal{C}$. By linearity of expectation, $\mathcal{D}$ is wlog compact by the above reasoning. Fix some ordering $\theta_1, \ldots, \theta_J$ of Θ (where $J = |\Theta|$). Let δ^0 be the maximal element in $\mathcal{D}$ with respect to the corresponding lexicographic ordering (so that, in particular, $\delta^0_{\theta_1} \geq \delta_{\theta_1}$ for all $\delta \in \mathcal{D}$). For every $h \in \{2, \ldots, J\}$, there exists a function $f_h : \mathbb{R}_{>0} \to \mathbb{R}_{>0}$ such that for all $\varepsilon > 0$

$$\delta \in \mathcal{D}, \sum_{j=1}^{h-1} |\delta_{\theta_j} - \delta^0_{\theta_j}| < f_h(\varepsilon) \qquad \Rightarrow \qquad \delta_{\theta_h} < \delta^0_{\theta_h} + \varepsilon. \tag{A.2}$$

Indeed, assume not, then there must be some h and some $\varepsilon > 0$ such that for every $k \in \mathbb{N}$ there exists a $\delta^k \in \mathcal{D}$ with $\sum_{j=1}^{h-1} |\delta^\nu_{\theta_j}| < 1/k$ and $\delta^k_{\theta_h} \geq \delta^0_{\theta_j} + \varepsilon$. Since $\mathcal{D}$ is compact, δ^k must have a convergent subsequence with limit $\delta^\varepsilon \in \mathcal{D}$. But $\delta^\varepsilon_{\theta_j} = \delta^0_{\theta_j}$ for $j < h$ and $\delta^\varepsilon_{\theta_h} \geq \delta^0_{\theta_h} + \varepsilon > \delta^0_{\theta_h}$, contradicting that δ^0 is maximal in $\mathcal{D}$ with respect to the lexicographic order. Hence there exists such f_h, and we can assume wlog $\frac{f_h(\varepsilon)}{\varepsilon}$ is monotonically increasing in $\varepsilon > 0$ (otherwise we can choose an f_h that is smaller for small values of ε).

Given the target $\delta^0 \in \mathcal{D}$ and the functions $f_h, h \geq 2$, I construct a sequence of risk functions r^{I_k} such that the expectation of the corresponding investigator choices converges to δ^0 for all

$\pi \in \Delta^*(\Theta)$. Concretely, for $k \in \mathbb{N}$ define $\alpha^k \in \mathbb{R}^\Theta$ recursively by

$$\alpha^k_{\theta_J} = k \qquad\qquad \alpha^k_{\theta_j} = k / \min_{h>j} f_h(1/\alpha^k_{\theta_h}), j < J$$

and consider the sequence of investigator risk functions

$$r^{I_k}_\theta(\hat{\tau}) = \mathrm{E}_\theta[(\hat{\tau}(z) - \tilde{\tau}^k_\theta)^2], \qquad\qquad \tilde{\tau}^k_{\theta_j} = \delta^0_{\theta_j} + \alpha^k_{\theta_j}$$

which falls within $\mathcal{R}^*$.

For the case of bounded $\mathcal{C}$ and some arbitrary, but fixed $\pi \in \Delta^*(\Theta)$, it remains to show that the expectation of $\hat{\tau}_\pi(r^{I_k})$ converges to δ^0. Write $\delta^k_\theta = \mathrm{E}_\theta \hat{\tau}_\pi(r^{I_k})$. Assume for contradiction that δ^k_θ does not converge to δ^0_θ. Since also $\delta^k_\theta \in \mathcal{D}$ for all k and $\mathcal{D}$ compact, $(\delta^k_\theta)_{k=1}^\infty$ must have a converging subsequence $(\delta^{k_\ell}_\theta)_{h=1}^\infty$ with $\delta^{k_\ell} \to \delta^1 \in \mathcal{D} \setminus \{\delta^0\}$ as $h \to \infty$. The average investigator loss along the sequence is

$$\mathrm{E}_\pi r^{I_{k_\ell}}_\theta(\hat{\tau}_\pi(r^{I_{k_\ell}})) = \mathrm{E}_\pi \underbrace{\mathrm{Var}_\theta(\hat{\tau}_\pi(r^{I_{k_\ell}}))}_{\leq \text{const. } (\mathcal{C} \text{ bounded})} + \mathrm{E}_\pi(\delta^{k_\ell}_\theta - (\delta^0_\theta + \alpha^{k_\ell}_\theta))^2. \tag{A.3}$$

Note that an estimator $\hat{\tau}^0$ with expectation $\delta^0 \in \mathcal{D}$ would also have been available in $\mathcal{C}$ by definition of $\mathcal{D}$, and the difference in risk between the chosen subsequence and the alternative is

$$\begin{aligned}
\Delta_\ell &= \mathrm{E}_\pi r^{I_{k_\ell}}_\theta(\hat{\tau}_\pi(r^{I_{k_\ell}})) - \mathrm{E}_\pi r^{I_{k_\ell}}_\theta(\hat{\tau}^0) \\
&\overset{(A.3)}{=} \mathrm{E}_\pi(\delta^{k_\ell}_\theta - (\delta^0_\theta + \alpha^{k_\ell}_\theta))^2 - \mathrm{E}_\pi(\alpha^{k_\ell}_\theta)^2 + \mathcal{O}(1) \\
&= \mathrm{E}_\pi \underbrace{(\delta^{k_\ell}_\theta - \delta^0_\theta)^2}_{\to(\delta^1_\theta - \delta^0_\theta)^2} - 2\,\mathrm{E}_\pi(\delta^{k_\ell}_\theta - \delta^0_\theta)\alpha^{k_\ell}_\theta + \mathcal{O}(1) \\
&= -2\sum_{j=1}^J \pi(\theta_j)\alpha^{k_\ell}_{\theta_j}(\delta^{k_\ell}_{\theta_j} - \delta^0_{\theta_j}) + \mathcal{O}(1).
\end{aligned}$$

Denote by h the smallest index of for which $\delta^0_{\theta_h} \neq \delta^1_{\theta_h}$. Since δ^0 is maximal with respect to the lexicographic ordering of $\mathcal{D}$ and δ^1 also in $\mathcal{D}$, we must have $\delta^0_{\theta_h} - \delta^1_{\theta_h} > 0$. By revealed

preference and since $\alpha^k_{\theta_{j+1}} = o(\alpha^k_{\theta_j})$ for all j, it follows that

$$0 \geq \Delta_\ell / \alpha^{k_\ell}_{\theta_h} = -2 \sum_{j=1}^{h-1} \pi(\theta_j) \frac{\alpha^{k_\ell}_{\theta_j}}{\alpha^{k_\ell}_{\theta_h}} (\delta^{k_\ell}_{\theta_j} - \delta^0_{\theta_j}) - 2\pi(\theta_h)(\delta^1_{\theta_h} - \delta^0_{\theta_h}) + o(1).$$

In particular, for $\varepsilon = \pi(\theta_h)(\delta^0_{\theta_h} - \delta^1_{\theta_h})$,

$$\liminf_{\ell \to \infty} \sum_{j=1}^{h-1} \underbrace{\pi(\theta_j) \frac{\alpha^{k_\ell}_{\theta_j}}{\alpha^{k_\ell}_{\theta_h}} (\delta^{k_\ell}_{\theta_j} - \delta^0_{\theta_j})}_{= a^\ell_j} \geq \varepsilon > 0. \tag{A.4}$$

Hence there must exists some h^* and a subsequence ℓ_s such that

$$a^{\ell_s}_{h^*} \to \nu \in (0, \infty], \qquad \limsup_{s \to \infty} \frac{a^{\ell_s}_j}{a^{\ell_s}_{h^*}} \leq 1 \ \forall j < h. \tag{A.5}$$

(That is, $a^{\ell_s}_{h^*}$ is a maximal sequence within that subsequence, for a suitable asymptotic notion of maximality; it is not unique, but an instance can be constructed from iterated subsequences.) For simplicity, I write $k_s = k_{\ell_s}$. I assume wlog that $\delta^{k_s}_{\theta_{h^*}} - \delta^0_{\theta_j} > 0$ for all s. By (A.3),

$$\sum_{j=1}^{h^*-1} |\delta^{k_s}_{\theta_j} - \delta^0_{\theta_j}| \geq f_{h^*}(\delta^{k_s}_{\theta_{h^*}} - \delta^0_{\theta_{h^*}}),$$

so there must exist some $j^* < h^*$ and a refinement of the subsequence along which $|\delta^{k_s}_{\theta_{j^*}} - \delta^0_{\theta_{j^*}}| \geq f_{h^*}(\delta^{k_s}_{\theta_{h^*}} - \delta^0_{\theta_{h^*}})/(h^* - 1)$. Note that

$$\frac{\pi(\theta_{j^*}) \frac{\alpha^{k_s}_{\theta_{j^*}}}{\alpha^{k_s}_{\theta_h}} |\delta^{k_s}_{\theta_{j^*}} - \delta^0_{\theta_{j^*}}|}{\pi(\theta_{h^*}) \frac{\alpha^{k_s}_{\theta_{h^*}}}{\alpha^{k_s}_{\theta_h}} (\delta^{k_s}_{\theta_{h^*}} - \delta^0_{\theta_{h^*}})} \geq \frac{\pi(\theta_{j^*})}{\pi(\theta_{h^*})(h^* - 1)} \frac{\alpha^{k_s}_{\theta_{j^*}}}{\alpha^{k_s}_{\theta_{h^*}}} \frac{f_{h^*}(\delta^{k_s}_{\theta_{h^*}} - \delta^0_{\theta_{h^*}})}{\delta^{k_s}_{\theta_{h^*}} - \delta^0_{\theta_{h^*}}}.$$

By (A.5) there exists some $\nu_0 \in (0, \infty)$ such that $a^{\ell_s}_{h^*} \geq \nu_0$ for all large s. By the definition of $a^\ell_{h^*}$ we find, again for large s, that

$$\delta^{k_s}_{\theta_{h^*}} - \delta^0_{\theta_{h^*}} = \frac{a^{\ell_s}_{h^*}}{\pi(\theta_{h^*})} \frac{\alpha^{k_s}_{\theta_h}}{\alpha^{k_s}_{\theta_{h^*}}} \geq \frac{\nu_0}{\pi(\theta_{h^*})} \frac{\alpha^{k_s}_{\theta_h}}{\alpha^{k_s}_{\theta_{h^*}}}.$$

By monotonicty of $\frac{f_{h^*}(\varepsilon)}{\varepsilon}$ therefore for large s

$$\frac{\pi(\theta_{j^*})\frac{\alpha^{k_s}_{\theta_{j^*}}}{\alpha^{k_s}_{\theta_h}}|\delta^{k_s}_{\theta_{j^*}}-\delta^0_{\theta_{j^*}}|}{\pi(\theta_{h^*})\frac{\alpha^{k_s}_{\theta_{h^*}}}{\alpha^{k_s}_{\theta_h}}(\delta^{k_s}_{\theta_{h^*}}-\delta^0_{\theta_{h^*}})} \geq \frac{\pi(\theta_{j^*})}{\pi(\theta_{h^*})(h^*-1)}\frac{\alpha^{k_s}_{\theta_{j^*}}}{\alpha^{k_s}_{\theta_{h^*}}}\frac{f_{h^*}\left(\frac{\nu_0}{\pi(\theta_{h^*})}\frac{\alpha^{k_s}_{\theta_h}}{\alpha^{k_s}_{\theta_{h^*}}}\right)}{\frac{\nu_0}{\pi(\theta_{h^*})}\frac{\alpha^{k_s}_{\theta_h}}{\alpha^{k_s}_{\theta_{h^*}}}}.$$

By construction of the rates α^k_θ, we have that for every triple $j^* < h^* < h$ and every constant $c > 0$ and all large k

$$\begin{aligned}\frac{\alpha^k_{\theta_{j^*}}}{\alpha^k_{\theta_h}}f_{h^*}\left(c\frac{\alpha^k_{\theta_h}}{\alpha^k_{\theta_{h^*}}}\right) &\geq \frac{\alpha^k_{\theta_{j^*}}}{\alpha^k_{\theta_J}}f_{h^*}\left(c\frac{\alpha^k_{\theta_J}}{\alpha^k_{\theta_{h^*}}}\right) = \frac{\alpha^k_{\theta_{j^*}}}{k}f_{h^*}\left(\frac{ck}{\alpha^k_{\theta_{h^*}}}\right)\\ &\geq c\alpha^k_{\theta_{j^*}}f_{h^*}\left(\frac{1}{\alpha^k_{\theta_{h^*}}}\right) \geq ck \to \infty.\end{aligned}$$

It follows that

$$\frac{\pi(\theta_{j^*})\frac{\alpha^{k_s}_{\theta_{j^*}}}{\alpha^{k_s}_{\theta_h}}|\delta^{k_s}_{\theta_{j^*}}-\delta^0_{\theta_{j^*}}|}{\pi(\theta_{h^*})\frac{\alpha^{k_s}_{\theta_{h^*}}}{\alpha^{k_s}_{\theta_h}}(\delta^{k_s}_{\theta_{h^*}}-\delta^0_{\theta_{h^*}})} \to \infty.$$

By (A.5), $\delta^{k_s}_{\theta_{j^*}} - \delta^0_{\theta_{j^*}} < 0$ for all but at most finitely many s. Hence $a^{\ell_s}_{j^*}/a^{\ell_s}_{h^*} \to -\infty$, and thus $\sum_{j=1}^{h-1} a^{\ell_s}_j \to -\infty$, contradicting (A.4). Therefore $\delta^1 = \delta^0$.

Consider now the case when $\mathcal{C}$ is unbounded. First, if $\mathcal{C}$ is unbounded but $\mathcal{B}$ is still bounded (and thus wlog compact by linearity of the expectation projection), then the same argument as above goes through since there is always an estimator with finite variance and expectation δ^0 available (and the investigator minimizes variance given expectation), so unbounded variance along the investigator path can only make the choice with expectation δ^0 more attractive.

Second, if $\mathcal{B}$ is also unbounded, then $\mathcal{C}$ cannot be minimax optimal. Since $\mathcal{B}$ is unbounded, it must contain a sequence $\delta^k \in \mathcal{B}$ with $\|\delta^k\|$ diverging. The projection of δ^k on the unit sphere towards the origin must contain a converging subsequence with limit v where $\|v\| = 1$. Consider a sequence of investigators with $\tilde{\tau}^k = v$ along the ray defined by the direction of this cluster point. One, if the average variance along the sequence of investigator choices is unbounded, then so is the average risk of the designer. Two, if the average variance along the sequence of investigator choices is bounded, then the bias diverges and average risk of the

designer is again unbounded. Indeed, I show that it is not possible that both average variance and average expectation remain bounded along the ray. If the expectation vector $\mathrm{E}_\theta[\hat{\tau}(z)]$ along that sequence of investigators remains bounded, pick a point arbitrarily close to the ray that falls outside that bound. (Such a point exists by construction of v.) As investigator preference moves along the ray, the gain in average investigator risk from moving to that point outweigh any cost in terms of variance since the marginal cost of being off the expectation target only increases, while the variance cost remains bounded. Hence, the bias cannot remain bounded and the average risk of the designer diverges.

We therefore have that for any $\pi \in \Delta^*(\Theta)$ the bias of investigator choices along the sequence r^{I_k} converges to $\bar{\beta}_\theta = \delta^0_\theta - \tau_\theta$ for all $\theta \in \Theta$.

Proof of minimax optimality. Given any mechanism, by richness there exists a sequence of investigator risk functions r^{I_k} in $\mathcal{R}^*$ and a bias vector $\bar{\beta}$ such that $\mathrm{E}_\theta[\hat{\tau}_\pi(r^{I_k})] - \tau_\theta \to \bar{\beta}_\theta$ for all $\pi \in \Delta^*(\Theta)$ and all $\theta \in \Theta$. The expected average designer's risk along this sequence is

$$\mathrm{E}_\eta[(\hat{\tau}_\pi(r^{I_k}) - \tau_\theta)^2] = \mathrm{E}_\eta \operatorname{Var}_\theta(\hat{\tau}_\pi(r^{I_k})) + \mathrm{E}_\eta \underbrace{(\mathrm{E}_\theta[\hat{\tau}_\pi(r^{I_k})] - \tau_\theta)^2}_{\to \bar{\beta}^2_\theta \forall \theta \in \Theta, \pi \in \Delta^*(\Theta)},$$

where I omit the argument z of the estimators. Since biases are bounded (since $\mathcal{D}$ is) and the support of η is in $\Delta^*(\Theta)$, by dominated convergence

$$\begin{aligned}\liminf_{k\to\infty} \mathrm{E}_\eta[(\hat{\tau}_\pi(r^{I_k}) - \tau_\theta)^2] &= \liminf_{k\to\infty} \mathrm{E}_\eta \operatorname{Var}_\theta(\hat{\tau}_\pi(r^{I_k})) + \mathrm{E}_\eta \bar{\beta}^2_\theta \\ &\geq \mathrm{E}_\eta \liminf_{k\to\infty} \mathrm{E}_\pi \operatorname{Var}_\theta(\hat{\tau}_\pi(r^{I_k})) + \mathrm{E}_\eta \bar{\beta}^2_\theta.\end{aligned}$$

For fixed $\pi \in \Delta^*(\Theta)$, $\liminf_{k\to\infty} \mathrm{E}_\pi \operatorname{Var}_\theta(\hat{\tau}_\pi(r^{I_k}))$ is at least the minimal asymptotic variance along a sequence $\hat{\tau}^k_\pi$ with bounded bias that converges to $\bar{\beta}$, and is otherwise unrestricted. Take such a sequence for which $\mathrm{E}_\pi \operatorname{Var}_\theta(\hat{\tau}^k_\pi)$ converges to its minimal limit. Along this sequence, $\hat{\tau}^k_\pi$ must be bounded, so it must have a convergent subsequence with some limit $\hat{\tau}^0_\pi$ in $\mathbb{R}^{\mathcal{Z}}$ for which by continuity also $\mathrm{E}_\theta[\hat{\tau}^0_\pi] - \tau_\theta = \bar{\beta}_\theta$. But then the variance of $\hat{\tau}^0_\pi$ must be at least the

variance of a variance-minimizing estimator subject to the bias constraint. Taken together,

$$\begin{aligned}\inf_{r^I \in \mathcal{R}^*} \mathrm{E}_\eta[r^D_\theta(\hat{\tau}_\pi(r^I))] &\geq \liminf_{k\to\infty} \mathrm{E}_\eta[(\hat{\tau}_\pi(r^{I_k}) - \tau_\theta)^2] \\ &\geq \mathrm{E}_\eta \min_{\hat{\tau}\in\mathcal{C}_{\bar{\beta}}} \mathrm{E}_\pi \operatorname{Var}_\theta(\hat{\tau}) + \mathrm{E}_\eta \bar{\beta}^2_\theta.\end{aligned}$$

Now, by aligned delegation,

$$\min_{\hat{\tau}\in\mathcal{C}_{\bar{\beta}}} \mathrm{E}_\pi(\operatorname{Var}_\theta(\hat{\tau}) + \bar{\beta}^2_\theta) = \min_{\hat{\tau}\in\mathcal{C}_{\bar{\beta}}} \mathrm{E}_\pi r^D_\theta(\hat{\tau}) = \mathrm{E}_\pi r^D_\theta(\hat{\tau}_\pi(r^I))$$

for every $r^I \in \mathcal{R}^*$ for choices from $\mathcal{C}_{\bar{\beta}}$. It follows that for every mechanism there is a set of biases such that the fixed-bias mechanisms has at least weakly better worst-case (over investigator types in $\mathcal{R}^*$) performance. Hence, at an optimal choice of biases β^η given the hyperprior η, the fixed-bias restriction $\mathcal{C}^\eta$ is minimax optimal. Such a minimizer exists because the set of biases is wlog compact (indeed, we can assume $\mathrm{E}_\eta \beta^2_\theta \leq \mathrm{E}_\eta r^D_\theta(z \mapsto 0) < \infty$) and the expected average risk continuous in the choice of bias. □

I conjecture that the restriction of the support to priors with full support is not necessary.

A.2 Representation of Unbiased Estimators

As in the main text, for fixed $n \geq 1$ and finite support $\mathcal{Y}$ I consider potential outcomes $\theta = (y(1), y(0)) \in \Theta = (\mathcal{Y}^2)^n$ from which for treatment $d \in \{0,1\}^n$ we observe $y = d \circ y(1) + (\mathbf{1} - d) \circ y(0) \in \mathcal{Y}^n$. (Here, $\circ$ denotes the Hadamard (entry-wise) product.) The estimate of interest is $\tau_\theta = \mathbf{1}'(y(1) - y(0))/n$.

Lemma 1.1 (Representation of unbiased estimators)**.** *The estimator $\hat{\tau}$ is unbiased, $\mathrm{E}_\theta[\hat{\tau}(z)] = \tau_\theta$ for all potential outcomes $\theta \in \Theta$, if and only if:*

1. *For a known treatment probability p, there exist leave-one-out regression adjustments $(\phi_i : (\mathcal{Y} \times \{0,1\})^{n-1} \to \mathbb{R})_{i=1}^n$ such that*

$$\hat{\tau}(z) = \frac{1}{n}\sum_{i=1}^n \frac{d_i - p}{p(1-p)}(y_i - \phi_i(z_{-i})).$$

2. *For a fixed number n_1 of treated units, there exist leave-two-out regression adjustments $(\phi_{ij} : (\mathcal{Y} \times \{0,1\})^{n-2} \to \mathbb{R})_{i<j}$ such that*

$$\hat{\tau}(z) = \frac{1}{n_1 n_0} \sum_{i<j} (d_i - d_j)(y_i - y_j - \phi_{ij}(z_{-ij})),$$

where $\phi_{ij}(z_{-ij})$ may be undefined outside $\mathbf{1}'d_{-ij} = n_1 - 1$.

I build up this general representation result in steps from simple estimators with binary outcomes to general estimators with finite support.

A.2.1 Known Treatment Probability, Binary Outcomes

I start with known treatment probability $p = \mathrm{E}_\theta[d_i]$ with d_i iid and binary support.

A natural class of admissible estimators are Bayes estimators, so a tempting starting point for the analysis of optimal unbiased estimators are (limits of) Bayes estimators that minimize average mean-squared error given the data and are also unbiased. However:

Remark A.1. *For $\mathcal{Y} = \{0,1\}$ and $p = .5$, the only unconstrained Bayes estimator (with respect to average mean-squared error) that is unbiased (conditional on $(y(1), y(0))$) is $\hat{\tau}(y,d) = \frac{1}{n}(2d - \mathbf{1})'(2y - \mathbf{1})$. For $\mathcal{Y} = \{0,1\}$ and $p \neq .5$, there are no unconstrained Bayes estimators that are also unbiased.*

Sketch of proof. For any prior, the unconstrained Bayes estimator with respect to average mean-squared error is the posterior expectation of τ_θ given the data. Any posterior expectation of τ_θ is bounded between the maximal treatment effect $+1$ and the minimal treatment effect -1. To achieve unbiasedness, any data that is consistent with either of the extremes must therefore yield an estimate of $+1$ or -1, respectively. Iterating this argument, the unique unconstrained Bayes estimator is the one achieved from a prior that puts full probability on $(y_i(1), y_i(0)) \in \{(1,0),(0,1)\}$ and zero probability on the configurations $\{(1,1),(0,0)\}$. This yields $\mathrm{E}_\theta[y_i(1) - y_i(0)|y_i, d_i] = (2d_i - 1)(2y_i - 1)$, which is unbiased for $p = .5$, but not for $p \neq .5$. □

The remark implies that searching for unbiased estimators among unconstrained Bayes

estimators to characterize the class of admissible unbiased estimators is futile, and I instead first characterize unbiased estimators before returning to optimality by solving for constrained Bayes estimators subject to the resulting representation.

Theorem A.1. *For $\mathcal{Y} = \{0,1\}$, assume that the estimators $\hat{\tau}^A, \hat{\tau}^B$ are unbiased τ_θ (conditional on $\theta = (y(1), y(0))$). Then,*

$$\hat{\tau}^B(y,d) - \hat{\tau}^A(y,d) = \frac{1}{n}\sum_{i=1}^{n} \frac{d_i - p}{p(1-p)}\phi_i(y_{-i}, d_{-i})$$

for a set of functions $\phi_i : (\mathcal{Y} \times \{0,1\})^{n-1} \to \mathbb{R}$.

For $n = 2$, the proof of Theorem A.1 can be made on a two-dimensional lattice folded into a torus. The general proof can similarly be understood as summing over hypercubes on the surface of an n-torus.

Proof. For $\hat{\delta}(y,d) = \hat{\tau}^B(y,d) - \hat{\tau}^A(y,d)$, take $\phi_i(y_{-i}, d_{-i})$ such that

$$\hat{\delta}(y,d) = \frac{1}{n}\sum_{i=1}^{n} \frac{d_i - p}{p(1-p)}\phi_i(y_{-i}, d_{-i}) \tag{A.6}$$

for all (y,d) with $y'd > 0$ (that is, all those that include some pair $(y_j, d_j) = (1,1)$). This is always feasible, say by the following inductive construction:

1. Set the $\phi_i(\mathbf{1}_{n-1}, \mathbf{1}_{n-1})$ in any way that has (A.6) hold for $\hat{\delta}(\mathbf{1}_n, \mathbf{1}_n)$.

2. Assuming that $\phi_i(y_{-i}, d_{-i})$ has been set for all i and (y,d) with $y'd \geq n-k$ such that (A.6) holds for such (y,d) (as is the case for $k = 0$ by the previous step), consider (y,d) with $y'd = n-(k+1)$. Among the terms $\phi_i(y_{-i}, d_{-i})$ in (A.6), those with $y'_{-i}d_{-i} = n-(k+1)$ have already been set by the induction assumption, and it remains to show that we can set conformable terms $\phi_i(y_{-i}, d_{-i})$ for $y'_{-i}d_{-i} = n-(k+2)$.

   Provided that $k < n-1$, note that any (y,d) with $y'd = n-(k+1)$ contains at least one (y_i, d_i) with $y'_i d_i = 1$, $\hat{\delta}(y,d)$ has the term $\phi_i(y_{-i}, d_{-i})$ appear on the right in (A.6), where thus $y'_{-i}d_{-i} = y'd - 1 = n-(k+2)$ (so it has not yet been set). But note that this specific $\phi_i(y_{-i}, d_{-i})$ also appears only for that (y,d) among all (y,d) with $y'd = n-(k+1)$ as

necessarily $y_i' d_i = 1$. Hence, we can set all previously undetermined $\phi_i(y_{-i}, d_{-i})$ for all i and $y'd$ with $y'd \geq n - (k+1)$ in a way that (A.6) holds for such (y, d).

By induction, we have set all $\phi_i(y_{-i}, d_{-i})$ for any i and $y'd \geq 1$ conformably with (A.6) for such (y, d). since this includes *all* terms of the form $\phi_i(y_{-i}, d_{-i})$, it remains to show that the unbiasedness assumption implies that (A.6) extends to (y, d) with $y'd = 0$.

Write $\hat{\delta}^\phi$ for the function defined by (A.6) for all (y, d). We have thus shown that $\hat{\delta}^\phi(y, d) = \hat{\delta}(y, d)$ for all (y, d) with $y'd > 0$. By assumption, $\mathrm{E}_\theta[\hat{\delta}(y, d)] = 0$ for all $\theta = (y(1), y(0))$, so

$$0 = \mathrm{E}_\theta[\hat{\delta}(y, d)] = \sum_{d \in \{0,1\}^n} \mathrm{P}(d)\, \hat{\delta}(d \circ y(1) + (1 - d) \circ y(0), d).$$

Fixing (y^*, d^*), it follows for any $\tilde{y}$ that

$$\hat{\delta}(y^*, d^*) = -\sum_{d \in \{0,1\}^n \setminus \{d^*\}} \mathrm{P}(d) / \mathrm{P}(d^*)\, \hat{\delta}((\mathbb{1}_{d_i = d_i^*})_{i=1}^n \circ y^* + (\mathbb{1}_{d_i \neq d_i^*})_{i=1}^n \circ \tilde{y}, d) \tag{A.7}$$

Since $\hat{\delta}^\phi$ is similarly zero-bias by construction, the same holds for $\hat{\delta}^\phi$. Thus, if for some (y^*, d^*) $\hat{\delta}$ and $\hat{\delta}^\phi$ agree on

$$\tilde{y}^*(d) = (\mathbb{1}_{d_i = d_i^*})_{i=1}^n \circ y^* + (\mathbb{1}_{d_i \neq d_i^*})_{i=1}^n \circ \tilde{y}, d)$$

for some $\tilde{y}$ and all $d \neq d^*$, then $\hat{\delta}(y^*, d^*) = \hat{\delta}^\phi(y^*, d^*)$.

We are ready to show (A.6) for all (y^*, d^*), by induction over $\mathbf{1}'d^*$. We let $\tilde{y} = \mathbf{1}$ throughout. At $k = 0$, $d^* = \mathbf{0}$. For any $d \neq d^*$, $\tilde{y}^*(d)'d \geq 1$, so $\hat{\delta}(\tilde{y}^*(d), d) = \hat{\delta}^\phi(\tilde{y}^*(d), d)$. By (A.7), $\hat{\delta}(y^*, d^*) = \hat{\delta}^\phi(y^*, d^*)$. Assume now that the claim holds for all (y^*, d^*) with $\mathbf{1}'d^* \leq k$, and consider some (y^*, d^*) with $\mathbf{1}'d^* = k + 1$. Then, for any $d \neq d^*$ with $\mathbf{1}'d \leq k$, $\hat{\delta}(y^*, d^*) = \hat{\delta}^\phi(y^*, d^*)$ by the induction assumption. For any $d \neq d^*$ with $\mathbf{1}'d \geq k+1$ there must be at least one dimension i with $d_i = 1, d_i^* = 0$, thus $\tilde{y}^*(d)'d \geq 1$ and $\hat{\delta}(y^*, d^*) = \hat{\delta}^\phi(y^*, d^*)$ follows by construction. We conclude that $\hat{\delta}(y^*, d^*) = \hat{\delta}^\phi(y^*, d^*)$ for all (y^*, d^*). □

Since $\hat{\tau}(y, d) = \frac{1}{n} \sum_{i=1}^n \frac{d_i - p}{p(1-p)} y_i$ is unbiased for τ_θ, the following characterization is immediate:

Corollary A.1. *For $\mathcal{Y} = \{0, 1\}$, any unbiased estimator $\hat{\tau}$ of τ_θ can be expressed as*

$$\hat{\tau} = \frac{1}{n} \sum_{i=1}^{n} \frac{d_i - p}{p(1-p)} (y_i - \phi_i(y_{-i}, d_{-i})).$$

The following result for the special case $n = 2$ shows that the reduction in degrees of freedom in the estimator implied by unbiasedness is substantial:

Remark A.2. *For $n = 2$, the $\phi_i(y_{-i}, d_{-i})$ are unique up to the one-dimensional equivalence class $\phi_i'(y_{-i}, d_{-i})) = \phi_i'(y_{-i}, d_{-i})) + (-1)^i(2d_{3-i} - 1)\Delta$, so unbiasedness reduces the degrees of freedom from $\hat{\tau} \in \mathbb{R}^{16}$ to $[\phi] \in \mathbb{R}^7$.*

A.2.2 Fixed Treatment Group Size, Binary Outcomes

Assume now that instead of the treatment probability, the number of treated is fixed at n_1, so that $d \sim \mathcal{U}(\mathcal{D}_{n_1})$ with $\mathcal{D}_{n_1} = \{t \in \{0, 1\}^n; t'n = n_1\}$. Effectively, we assume invariance to permutations in the assignment of treatment, but not more.

The natural, unbiased treatment-control-difference estimator can be written as

$$\hat{\tau}^*(y, d) = \frac{1}{n_1} \sum_{d_i=1} y_i - \frac{1}{n_0} \sum_{d_i=0} y_i = \frac{1}{n_1 n_0} \sum_{d_i=1, d_j=0} (y_i - y_j),$$

of which an unbiased extension is

$$\hat{\tau}^\phi(y, d) = \frac{1}{n_1 n_0} \sum_{d_i=1, d_j=0} (y_i - y_j - \phi_{ij}(y_{-ij}, d_{-ij}))$$

with $\phi_{ij} = -\phi_{ji}$. I claim that these are also *all* extensions.

Theorem A.2. *Let $\mathcal{Y} = \{0, 1\}$. Assume that $\hat{\tau}^A, \hat{\tau}^B$ are unbiased for τ_θ. Then,*

$$\hat{\tau}^B(y, d) - \hat{\tau}^A(y, d) = \frac{1}{n_1 n_0} \sum_{d_i=1, d_j=0} \phi_{ij}(y_{-ij}, d_{-ij}), \qquad \phi_{ij} = -\phi_{ji}$$

for functions $\phi_{ij} : (\mathcal{Y} \times \{0, 1\})^{n-2} \to \mathbb{R}$.

Note that we can alternatively write

$$\hat{\tau}^B(y, d) - \hat{\tau}^A(y, d) = \frac{1}{n_1 n_0} \sum_{i=1}^{n} \sum_{j=i+1}^{n} (d_i - d_j)\, \phi_{ij}(y_{-ij}, d_{-ij}),$$

where we sum over each pair once and ϕ_{ij} is only defined for $j > i$.

We first establish a lemma that adopts the proof strategy from Theorem A.1 to the setting at hand. To this end, for $(y(1), y(0)) \in (\mathcal{Y}^2)^n$ write

$$N(y(1), y(0)) = \{(d \circ y(1) + (1-d) \circ y(0), d); d \in \mathcal{D}_{n_1}\}$$

(the set of observations consistent with $y(1), y(0)$) and let

$$\mathcal{C} = \bigcup_{(y(1),y(0)) \in (\mathcal{Y}^2)^n} N(y(1), y(0)).$$

Let $c : \mathcal{C} \to \mathcal{C}^-$ be the surjective correspondence

$$(y, d) \mapsto \{(ij, (y_{-ij}, d_{-ij})); i < j, d_i \neq d_j\}.$$

Lemma A.1. *If there exists a partition $\mathcal{C} = \bigcup_{t=1}^{T} \mathcal{C}_t$ such that for some T^**

1. *for $\mathcal{C}_t^- = \bigcup_{(y,d) \in \mathcal{C}_t} c(y, d)$ and*

$$\mathcal{D}_t = \mathcal{C}_t^- \setminus \bigcup_{s<t} \mathcal{C}_s^-,$$

   there exists injections $b_t : \mathcal{C}_t \to \mathcal{D}_t$ for $t \leq T^*$ and

2. *for all $t > T^*$ and $(y, d) \in \mathcal{C}_t$, there exists some $(y(1), y(0)) \in (\mathcal{Y}^2)^n$ both $(y, d) \in N(y(1), y(0))$ and*

$$(N(y(1), y(0)) \setminus \{(y, d)\}) \cap \bigcup_{s \geq t} \mathcal{C}_s = \emptyset$$

then for any $\hat{\delta}$ that is mean-zero there exist a function $\phi : \mathcal{C}^- \to \mathbb{R}$ such that $\hat{\delta} = \hat{\delta}^\phi$ with

$$\hat{\delta}^\phi(y, d) = \frac{1}{n_1 n_0} \sum_{i=1}^{n} \sum_{j=i+1}^{n} (d_i - d_j)\, \phi_{ij}(y_{-ij}, d_{-ij}).$$

Proof. Given some $\hat{\delta}$, we first construct such a family ϕ with $\hat{\delta}^\phi(y, d) = \hat{\delta}$ for all $(y, d) \in \bigcup_{t \leq T^*} \mathcal{C}_t$, and then establish that this implies $\hat{\delta}^\phi(y, d) = \hat{\delta}$ also for $(y, d) \in \bigcup_{t > T^*} \mathcal{C}_t$.

For the first part, I argue inductively as follows: Take $t \leq T^*$ and assume ϕ has been set on $\bigcup_{s<t} \mathcal{C}_s^-$ such that $\hat{\delta}^\phi = \hat{\delta}$ on $\bigcup_{s<t} \mathcal{C}_s$ (which is given trivially for $t = 1$) then for every $(y, d) \in \mathcal{C}_t$

by the first assumption of the lemma there exists a unique term $\phi_{ij}(y_{-ij}, d_{-ij}) = \phi(b_t(y,d))$ with $b_t(y,d) \in \mathcal{D}_t$ that has not yet been set, so we can set the terms $\phi(\mathcal{D}_t)$ in a way that $\hat{\delta}^\phi = \hat{\delta}$ on $\mathcal{C}_t$ and thus on $\bigcup_{s \le t} \mathcal{C}_s$. This completes the proof of the first part.

For the second part, note that by assumption $\mathrm{E}_\theta[\hat{\delta}(y,d)] = 0$ for all $\theta = (y(1), y(0))$, so

$$0 = \mathrm{E}_\theta[\hat{\delta}(y,d)] = \sum_{(y,d) \in N(y(1),y(0))} \hat{\delta}(y,d).$$

Fixing (y^*, d^*) it follows for any $(y(1), y(0))$ with $(y^*, d^*) \in N(y(1), y(0))$ that

$$\hat{\delta}(y^*, d^*) = -\sum_{(y,d) \in N(y(1),y(0)) \setminus \{(y^*,d^*)\}} \hat{\delta}(y,d) \tag{A.8}$$

Since $\hat{\delta}^\phi$ is similarly zero-bias by construction, the same holds for $\hat{\delta}^\phi$. We are now ready to show that $\hat{\delta}^\phi = \hat{\delta}$ for all $(y,d) \in \mathcal{C}_t$, by induction over t. For some $t > T^*$, assuming $\hat{\delta}^\phi = \hat{\delta}$ holds for all $(y,d) \in \mathcal{C}_s$ with $s < t$ (as is the case for all $s \le T^*$), take any $(y^*, d^*) \in \mathcal{C}_t$. By the second part of the lemma, (A.8) and the induction assumption we must have $\hat{\delta}(y^*, d^*) = \hat{\delta}^\phi(y^*, d^*)$. This completes the proof. □

We are ready to prove the main result:

Proof of Theorem A.2. $\hat{\delta}(y,d) = \hat{\tau}^B(y,d) - \hat{\tau}^A(y,d)$ is a unbiased estimator of zero. Define $a, b : \mathcal{C} \to \mathbb{N}_0$ by

$$a(y,d) = y'd, \qquad\qquad b(y,d) = (\mathbf{1} - y)'(\mathbf{1} - d).$$

Note that $a(y,d) + b(y,d) \le n$.

First, set $T^* = n - 1$ and for every $t \le T$

$$\mathcal{C}_t = \{(y,d) \in \mathcal{C}; \min(a(y,d), b(y,d)) \ge 1, a(y,d) + b(y,d) = n + 1 - t\}.$$

Then the first assumption of Lemma A.1 is fulfilled, as for every $(y,d) \in \mathcal{C}_t$ there exists some $(ij, (y_{-ij}, d_{-ij}) \in \mathcal{C}_t$ with $y'_{-ij} d_{-ij} + (\mathbf{1} - y_{-ij})'(\mathbf{1} - d_{-ij}) = n - 1 - t = a(y,d) + b(y,d) - 2$, but (y,d) is also the unique element in $\mathcal{C}_t$ covering that element of $\mathcal{D}_t$ under the correspondence c (as indeed necessarily $y_i = d_i, y_j = d_j$, which pins down (y,d) from $(ij, (y_{-ij}, d_{-ij}))$.

Second, with $T = n + 1$ and

$$\mathcal{C}_n = \{(y, d) \in \mathcal{C}; a(y, d) = 0, b(y, d) \geq 1\},$$
$$\mathcal{C}_{n+1} = \{(y, d) \in \mathcal{C}; b(y, d) = 0\},$$

note that for each $(y^*, d^*) \in \mathcal{C}_n \cup \mathcal{C}_{n+1}$ we have that $(y(1), y(0)) = (y^* \circ d^* + \mathbf{1} \circ (1 - d^*), y^* \circ (1 - d^*))$ produces

$$N(y(1), y(0)) \cap \{(y, d) \in \mathcal{C}; \min(a(y, d), b(y, d)) = 0\} = \{(y^*, d^*)\}$$

for $(y^*, d^*) \in \mathcal{C}_n$ and

$$N(y(1), y(0)) \cap \{(y, d) \in \mathcal{C}; b(y, d) = 0\} = \{(y^*, d^*)\}$$

for $(y^*, d^*) \in \mathcal{C}_{n+1}$. This verifies the second assumption of Lemma A.1. □

Unbiased estimators (for binary outcomes) are thus fully characterized by leave-two-out adjustments. Note that leave-one-out adjustments as in the case of known treatment probability p would not generally be unbiased.

A.2.3 Extension to Finite Support

Take some distribution over the treatment assignment vector $d \in \{0, 1\}^n$, data $(y(1), y(0)) \in (\mathcal{Y}^2)^n$ as before where $\mathcal{Y} \subseteq \mathbb{R}$, and $y = d \circ y(1) + (\mathbf{1} - d) \circ y(0)$. Our goal now is to extend a representation for binary outcomes to one for finite (but arbitrarily large) support $\mathcal{Y}$.

Lemma A.2. *Assume that for $\mathcal{Y} = \{0, 1\}$ any $\hat{\delta}$ with $\mathrm{E}_\theta[\delta(y, d)] = 0$ for all $\theta = (y(1), y(0))$ permits a representation $\hat{\delta} = \hat{\delta}^\phi$ with*

$$\hat{\delta}^\phi(y, d) = \sum_{\iota \in \mathcal{I}} w_\iota(d_{S_\iota}) \phi_\iota(y_{-S_\iota}, d_{-S_\iota})$$

for fixed $\mathcal{I}, (w_\iota)_{\iota \in \mathcal{I}}, (S_\iota)_{\iota \in \mathcal{I}}$ (where $\mathcal{I}$ finite) and variable $(\phi_\iota)_{\iota \in \mathcal{I}}$ where

$$\phi_\iota : (\mathcal{Y} \times \{0, 1\})^{\{1,\ldots,n\} \setminus S_\iota} \to \mathbb{R}.$$

Then the representation result extends to any finite $\mathcal{Y} \subseteq \mathbb{R}$ (with the same $\mathcal{I}, (w_\iota)_{\iota\in\mathcal{I}}, (S_\iota)_{\iota\in\mathcal{I}}$).

Proof. Write $\mathcal{Y}_\ell = \{0, 1, \dots, \ell\}$ and define (for $\ell \geq 2, m \geq 0$)

$$\mathcal{Y}_{\ell,m} = \bigtimes_{i=1}^{m} \mathcal{Y}_{2\ell-1} \times \bigtimes_{i=m+1}^{n} \mathcal{Y}_\ell$$

We first establish the following intermediate result by induction over $t = ns+m$ from $t = 0$: For any $(s, m) \in (\mathbb{N}_0 \times \{1, \dots, n\}) \cup \{(0, 0)\}$ for $\ell = 2^s + 1$ any $\hat{\delta}$ with $\mathrm{E}_\theta[\hat{\delta}(y, d)] = 0$ for all $\theta = (y(1), y(0)) \in \mathcal{Y}^2_{\ell,m}$ permits a representation $\hat{\delta} = \hat{\delta}^\phi$ as above with $\phi_\iota : \bigtimes_{i\in\{1,\dots,n\}\setminus S_\iota} (\mathcal{Y}_{\ell,m})_i \to \mathbb{R}$

For $t = 0$, the statement holds by the assumption of the lemma. Assume now that its holds for t with such (s, m) such that $t = ns + m$ and $\ell = 2^s + 1$, and consider the $(s^+, m^+) \in \mathbb{N}_0 \times \{1, \dots, n\}$ with $ns^+ + m^+ = t + 1$, and write $\ell^+ = 2^{s^+} + 1$ Fix an estimator $\hat{\delta}$ with $\mathrm{E}_\theta[\hat{\delta}(y, d)] = 0$ for all $\theta = (y(1), y(0)) \in \mathcal{Y}^2_{\ell^+,m^+}$. Write For $(y, d) \in \mathcal{Y}_{\ell,m} \times \{0, 1\}^n$ define $y^+_{m^+} = \ell^+ + y_{m^+} - 1, y^+_{-m^+} = y_{-m^+}$ as well as $y^-_{m^+} = \ell^+, y^-_{-m^+} = y_{-m^+}$ to obtain $y^+, y^- \in \mathcal{Y}_{\ell^+,m^+}$, and define estimators by

$$\hat{\delta}_1(y, d) = \hat{\delta}(y^+, d) - \hat{\delta}(y^-, d) \qquad\qquad \hat{\delta}_2(y, d) = \hat{\delta}(y, d)$$

where thus $\hat{\delta}_2$ is merely a restriction of $\hat{\delta}$ to $\mathcal{Y}_{\ell,m} \times \{0, 1\}^n$. For $(y, d) \in \mathcal{Y}_{\ell^+,m^+} \times \{0, 1\}^n$ define $\bar{y}_{m^+} = \min(y_{m^+}, \ell^+), \bar{y}_{-m^+} = y_{-m^+}$ to obtain a $\bar{y} \in \mathcal{Y}^2_{\ell,m}$ for which

$$\begin{aligned}\hat{\delta}(y, d) &= \hat{\delta}(y, d) - \hat{\delta}(\bar{y}_{m^+}, d) + \hat{\delta}(\bar{y}_{m^+}, d)\\ &= \hat{\delta}_1(y - \bar{y}_{m^+}, d) + \hat{\delta}_2(\bar{y}_{m^+}, d).\end{aligned}$$

$\hat{\delta}_2$ is unbiased (for $\mathcal{Y}_{\ell,m}$) by construction. Note that

$$\mathrm{E}_\theta[\hat{\delta}_1(y, d)] = \mathrm{E}_\theta[\hat{\delta}(y^+, d)] - \mathrm{E}_\theta[\hat{\delta}(y^-, d)] = 0$$

for any $y(1), y(0) \in \mathcal{Y}_{\ell,m}$, as they generate $y^+(1), y^+(0) \in \mathcal{Y}_{\ell^+,m^+}$ for which $\hat{\delta}$ is unbiased by assumption, so $\hat{\delta}_1$ is likewise unbiased (for $y(1), y(0) \in \mathcal{Y}_{\ell,m}$). By the induction assumption, there are thus ϕ^1, ϕ^2 with

$$\hat{\delta}(y, d) = \sum_{\iota\in\mathcal{I}} w_\iota(d_{S_\iota})(\phi^1_\iota((y - \bar{y}_{m^+})_{-S_\iota}, d_{-S_\iota}) + \phi^2_\iota((\bar{y}_{m^+})_{-S_\iota}, d_{-S_\iota})$$

for any $(y,d) \in \mathcal{Y}_{\ell^+,m^+} \times \{0,1\}^n$. For

$$\phi_\iota(y_{-S_\iota}, d_{-S_\iota}) = \phi_\iota^1(y_{-S_\iota} - (\bar{y}_{m^+})_{-S_\iota}, d_{-S_\iota}) + \phi_\iota^2((\bar{y}_{m^+})_{-S_\iota}, d_{-S_\iota})$$

we therefore have $\hat{\delta} = \hat{\delta}^\phi$. This concludes the induction step and thus the proof of the intermediate result.

Setting $m = n$, it is immediate that the statement of the lemma holds for all $\mathcal{Y} = \mathcal{Y}_{2^s+1}$. Since it will always hold for subsets, it holds for all $\mathcal{Y} = \mathcal{Y}_\ell$. Now take arbitrary $\mathcal{Y} = \{z_1, \ldots, z_\ell\}$, and define for $(y,d) \in (\mathcal{Y}_\ell \times \{0,1\})^n$

$$\tilde{\delta}(y,d) = \hat{\delta}(z_y, d)$$

where $(z_y)_i = z_{y_i} \in \mathcal{Y}$. By the intermediate result there is some $\tilde{\phi}$ such that $\tilde{\delta} = \hat{\delta}^{\tilde{\phi}}$. Setting $\phi_\iota(y_{-S_\iota}, d_{-S_\iota}) = \tilde{\phi}(\tilde{y}_{-S_\iota}, d_{-S_\iota})$ with $\tilde{y}$ such that $z_{\tilde{y}} = y$ yields $\hat{\delta}(y,d) = \hat{\delta}^\phi(y,d)$. □

We are now ready to proof the representation result in the main chapter.

Proof of Lemma 1.1. The representation for general finite support follows from Lemma A.2 applied to the binary representation results in Theorem A.1 and Theorem A.2, respectively. □

A.3 Characterization of Optimal Unbiased Estimators

When is an estimator not just unbiased, but has also low average mean-squared error? I start with the representation

$$\hat{\tau}^\phi(y,d) = \frac{1}{n}\sum_{i=1}^n \frac{d_i - p}{p(1-p)}(y_i - \phi_i(y_{-i}, d_{-i}))$$

for known treatment probability p and consider the error

$$\begin{aligned}
\Delta_\theta^\phi(y,d) &= \hat{\tau}^\phi(y,d) - \tau_\theta \\
&= \frac{1}{n}\sum_{i=1}^n \left(\frac{d_i - p}{p(1-p)}(y_i - \phi_i(y_{-i}, d_{-i})) - (y(1)_i - y(0)_i) \right) \\
&= \frac{1}{n}\sum_{i=1}^n \frac{d_i - p}{p(1-p)}(\bar{y}_i - \phi_i(y_{-i}, d_{-i}))
\end{aligned}$$

for the adjustment oracle $\bar{y}_i = (1-p)y(1)_i + py(0)_i$, which would be the loss-minimizing choice for $\phi_i(y_{-i}, d_{-i})$.

Proposition A.1. *For some prior π over $\theta = (y(1), y(0))$, any ϕ^*_π with*

$$\phi^*_\pi(y_{-i}, d_{-i}) = \mathrm{E}_\pi\left[\bar{y}_i | y_{-i}, d_{-i}\right]$$

is a (global) minimizer of average loss $\mathrm{E}_\pi L_\theta(\phi)$, where $L_\theta(\phi) = \mathrm{E}_\theta(\Delta^\phi_\theta(y,d))^2$.

Proof. The restriction that adjustments $\phi_i(y_{-i}, d_{-i})$ are functions only of y_{-i}, d_{-i} (and of π) requires some care, as each such adjustments appears given multiple draws of (y, d). Write

$$M_i(y^*_{-i}, d^*_{-i}) = \{(y,d) \in (\mathcal{Y} \times \{0,1\})^n; (y_{-i}, d_{-i}) = (y^*_{-i}, d^*_{-i})\}$$

for the (y, d) for which $\hat{\tau}^\phi(y,d)$ (and thus $\Delta^\phi_\theta(y,d)$) includes the term $\phi_i(y^*_{-i}, d^*_{-i})$. Then,

$$\begin{aligned}\frac{\partial \mathrm{E}_\pi L_\theta(\phi)}{\partial \phi_i(y^*_{-i}, d^*_{-i})} &= \frac{\partial \mathrm{E}_\pi\left[\mathbb{1}_{(y,d)\in M(y^*_{-i}, d^*_{-i})}(\Delta^\phi_\theta(y,d))^2\right]}{\partial \phi_i(y^*_{-i}, d^*_{-i})} \\ &= \mathrm{E}_\pi\left[\mathbb{1}_{(y,d)\in M(y^*_{-i}, d^*_{-i})}\frac{\partial(\Delta^\phi_\theta(y,d))^2}{\partial \phi_i(y^*_{-i}, d^*_{-i})}\right],\end{aligned}$$

where we note that we can exchange differentiation and integration because all summands are bounded. I omit writing E_θ explicitly inside E_π and consider the joint distribution of θ and z. Here, for all $(y,d) \in M(y^*_{-i}, d^*_{-i})$,

$$\begin{aligned}&\frac{\partial(\Delta^\phi_\theta(y,d))^2}{\partial \phi_i(y^*_{-i}, d^*_{-i})} = -\frac{2}{n}\frac{d_i - p}{p(1-p)}\Delta^\phi_\theta(y,d) \\ &= -\frac{2}{n^2}\left(\frac{(d_i-p)^2}{(p(1-p))^2}(\bar{y}_i - \phi_i(y^*_{-i}, d^*_{-i})) + \sum_{j\neq i}\frac{(d_i-p)(d^*_j - p)}{(p(1-p))^2}(\bar{y}_j - \phi_j(y_{-j}, d_{-j}))\right).\end{aligned}$$

The first-order condition $\frac{\partial \mathrm{E}_\pi L_\theta(\phi)}{\partial \phi_i(y^*_{-i}, d^*_{-i})} = 0$ is therefore

$$\begin{aligned}&\mathrm{E}_\pi\left[\mathbb{1}_{(y,d)\in M(y^*_{-i}, d^*_{-i})}(d_i - p)^2(\phi_i(y^*_{-i}, d^*_{-i}) - \bar{y}_i)\right] \\ &= -\sum_{j\neq i}(d^*_j - p)\,\mathrm{E}_\pi\left[\mathbb{1}_{(y,d)\in M(y^*_{-i}, d^*_{-i})}(d_i - p)(\phi_j(y_{-j}, d_{-j}) - \bar{y}_j)\right].\end{aligned}$$

The condition is trivially fulfilled for $\mathrm{P}_\pi((y,d) \in M(y^*_{-i}, d^*_{-i})) = 0$. Otherwise, equivalently

$$\overbrace{\mathrm{E}[(d_i-p)^2]\phi_i(y^*_{-i},d^*_{-i}) - \mathrm{E}_\pi[(d_i-p)^2\bar{y}_i|(y_{-i},d_{-i})=(y^*_{-i},d^*_{-i})]}^{=p(1-p)(\phi_i(y^*_{-i},d^*_{-i})-\mathrm{E}_\pi[\bar{y}_i|(y_{-i},d_{-i})=(y^*_{-i},d^*_{-i})])}$$
$$= -\sum_{j\neq i}(2d^*_j-1)\,\mathrm{E}_\pi\left[(d_i-p)\phi_j(y_{-j},d_{-j})|(y_{-i},d_{-i})=(y^*_{-i},d^*_{-i})\right]$$

Note that this system of first-order conditions will generally have many solutions, as the ϕ-representation of $\hat{\tau}^\phi$ is not generally unique. I now show that the specific choice

$$\phi_i(y^*_{-i},d^*_{-i}) = \mathrm{E}_\pi[\bar{y}_i|(y_{-i},d_{-i})=(y^*_{-i},d^*_{-i})]$$

(for $\mathrm{E}_\pi \mathrm{P}_d((y,d) \in M(y^*_{-i},d^*_{-i})) > 0$, otherwise, say, zero) is a (global) posterior-loss minimizer. To that end, note that for $i \neq j$

$$\begin{aligned}
&\mathrm{E}_\pi\left[(d_i-p)\,\mathrm{E}_\pi[\bar{y}_j|y_{-j},d_{-j}]|y_{-i},d_{-i}\right]\\
&=\mathrm{E}_\pi\left[(d_i-p)\,\mathrm{E}_\pi[\bar{y}_j|y_i,d_i,y_{-ij},d_{-ij}]|y_j,d_j,y_{-ij},d_{-ij}\right]\\
&=\mathrm{E}_\pi\left[\mathrm{E}_\pi\left[(d_i-p)\,\mathrm{E}_\pi[\bar{y}_j|y_i,d_i,y_{-ij},d_{-ij}]|d_i,y_{-ij},d_{-ij}\right]|y_{-ij},d_{-ij}\right]\\
&=\mathrm{E}_\pi\left[(d_i-p)\,\mathrm{E}_\pi[\bar{y}_j|d_i,y_{-ij},d_{-ij}]|y_{-ij},d_{-ij}\right]\\
&=\mathrm{E}_\pi\left[(d_i-p)\,\mathrm{E}_\pi[\bar{y}_j|y_{-ij},d_{-ij}]|y_{-ij},d_{-ij}\right]=0.
\end{aligned}$$

The first-order condition follows. Also

$$\begin{aligned}
&\frac{\partial^2\,\mathrm{E}_\pi L_\theta(\phi)}{\partial\phi_i(y^A_{-i},d^A_{-i})\partial\phi_j(y^B_{-j},d^B_{-j})}\\
&=\frac{1}{(p(1-p)n)^2}\,\mathrm{E}_\pi\left[\mathbb{1}_{(y,d)\in M(y^A_{-i},d^A_{-i})\cap M(y^B_{-j},d^B_{-j})}(d^B_i-p)(d^A_j-p)\right]\\
&=\begin{cases}
\frac{1}{p(1-p)n^2}\,\mathrm{P}_\pi((y_{-i},d_{-i})=(y^A_{-i},d^A_{-i})), & (i,y^A_{-i},d^A_{-i})=(j,y^B_{-j},d^B_{-j})\\
\frac{(d^*_i-p)(d^*_j-p)}{(p(1-p)n)^2}\,\mathrm{P}_\pi(y^*,d^*), & i\neq j,(y^{A/B}_{-i/j},x^{A/B}_{-i/j})=(y^*_{-i/j},d^*_{-i/j})\\
0, & \text{otherwise.}
\end{cases}
\end{aligned}$$

Note that $\frac{\partial^2\,\mathrm{E}_\pi L_\theta(\phi)}{\partial\phi_i(y^A_{-i},d^A_{-i})\partial\phi_j(y^B_{-j},d^B_{-j})}$ is two times the variance-covariance matrix of the (mean-zero) random variables $\mathbb{1}_{(y,d)\in M(y^*_{-i},d^*_{-i})}\frac{d_i-p}{p(1-p)n}$, and therefore everywhere positive semi-definite. It follows that the first-order conditions locate a (global) minimum. □

The proposition directly yields the first part of the general characterization result in the main chapter. The second part follows analogously with oracle adjustments

$$\Delta\bar{y}_{ij} = \underbrace{\left(\frac{n_0}{n}y_i(1) + \frac{n_1}{n}y_i(0)\right)}_{\bar{\bar{y}})_i} - \left(\frac{n_0}{n}y_j(1) + \frac{n_1}{n}y_j(0)\right).$$

Theorem 1.2 (Choice of the investigator from unbiased estimators)**.** *An investigator with risk $r \in \mathcal{R}^*$ and prior π over Θ chooses the following unbiased Bayes estimators:*

1. *For a known treatment probability p,*

$$\hat{\tau}(z) = \frac{1}{n}\sum_{i=1}^{n}\frac{d_i - p}{p(1-p)}(y_i - \mathrm{E}_\pi[\bar{y}_i|z_{-i}]).$$

2. *For a fixed number n_1 of treated units,*

$$\hat{\tau}(z) = \frac{1}{n_1 n_0}\sum_{i<j}(d_i - d_j)(y_i - y_j - \mathrm{E}_\pi[\Delta\bar{y}_{ij}|z_{-ij}]).$$

Proof. The first part is immediate from Proposition A.1. For the second part, we can wlog consider adjustments

$$\phi_{i;j}(y_{-ij}, d_{-ij}) \tag{A.9}$$

for which we set $\phi_{ij}(y_{-ij}, d_{-ij}) = \phi_{i;j}(y_{-ij}, d_{-ij}) - \phi_{j;i}(y_{-ij}, d_{-ij})$ to find

$$\begin{aligned}
\Delta^{\phi}_{\theta}(y, d) &= \hat{\tau}^{\phi}(y, d) - \tau_\theta \\
&= \frac{1}{n_1 n_0}\sum_{i<j}(d_i - d_j)\left((\bar{\bar{y}}_i - \phi_{i;j}(y_{-ij}, d_{-ij})) - (\bar{\bar{y}}_j - \phi_{j;i}(y_{-ij}, d_{-ij}))\right) \\
&= \frac{1}{n_1 n_0}\sum_{i,j}(d_i - d_j)(\bar{\bar{y}}_i - \phi_{i;j}(y_{-ij}, d_{-ij})).
\end{aligned}$$

As in the proof of Proposition A.1, we can then verify that the choice

$$\phi_{i;j}(y_{-ij}, d_{-ij}) = \mathrm{E}_\pi[\bar{\bar{y}}_i|y_{-ij}, d_{-ij}]$$

fulfils the associated first-order condition. □

A.4 OLS is Biased

Consider a sample of n units (y_i, d_i, x_i), where $d_i \in \{0, 1\}$ are iid given $x_1, \ldots, x_n$ with $\mathrm{P}(d_i = 1) = p \in (0, 1)$.

A.4.1 Conditional on Covariates

Conditional on covariates $x_i = \mathbb{1}_{i=1}$ and for $y_i = x_i d_i$, the sample-averge treatment effect is $\tau = 1/n$ (one for the first unit, zero for all other units). The coefficient $\hat{\tau}^{\mathrm{OLS}}$ on d in a linear regression of y on d and x (with intercept) has expectation $\mathrm{E}[\hat{\tau}^{\mathrm{OLS}} | n_1] = 0$ conditional on any number $1 < n_1 < n - 1$ of treated units. Indeed, x perfectly explains y, so the coefficient on d will always be zero (by Frisch-Waugh or otherwise).

A.4.2 Over the Sampling Distribution

Assume that $x_i \in \mathbb{R}^{k_n+1}$ with $\mathrm{P}(x_{i0}) = q \in (0, 1)$ and

$$x_{i1}, \ldots, x_{ik} | x_{i0} \overset{\text{iid}}{\sim} (1 - x_{i0}) \cdot \mathcal{N}(0, 1)$$

(that is, $x_{ij} = 0$ for all $j > 0$ if $x_{i0} = 1$), x_i iid across units. (Alternatively, any non-degenerate distribution will do.) Let $y_i = x_{i0} d_i$. The average treatment effect of d_i on y_i is

$$\tau^{\mathrm{pop}} = \mathrm{E}[y_i | d_i = 1] - \mathrm{E}[y_i | d_i = 0] = q.$$

Let $\hat{\tau}^{\mathrm{OLS}}$ be the coefficient on d in a linear regression of y on d and x (with intercept). For $k_n/n \to \alpha \in (0, 1 - q)$ as $n \to \infty$ we also find

$$\hat{\tau}^{\mathrm{OLS}} \xrightarrow{\mathrm{P}} \frac{q}{1 - \alpha}.$$

Indeed, writing A_x for the annihilator matrix with respect to x and the intercept, by Frisch-Waugh $\hat{\tau}^{\mathrm{OLS}} = \frac{d' A_x y}{d' A_x d}$ with

$$\mathrm{E}[d' A_x y | x] = p(1 - p)(n_{x=1} - 1),$$
$$\mathrm{E}[d' A_x d | x] = p(1 - p) \operatorname{trace}(A_x) = p(1 - p)(n - k_n - 1).$$

By the law of large numbers (where variances are suitably bounded),

$$\frac{d' A_x y}{n} \xrightarrow{\text{P}} p(1-p)\,\text{E}[n_{x=1}/n] = p(1-p)q,$$
$$\frac{d' A_x y}{n} \xrightarrow{\text{P}} p(1-p)(1-\alpha).$$

A.5 Asymptotic Inference

In this section, I derive asymptotically valid inference of the average treatment effect. These results deviate from the approach in the main chapter in two notable, related ways. First, I assume that potential outcomes and controls themselves are sampled iid from a population distribution, and inference will not condition on their realizations. Second, in order to obtain valid inference, I take large-sample approximations. The estimator of interest is still unbiased in finite samples for the sample-average treatment effect. But for efficiency and inference I focus on the estimation of the population-average treatment effect in large samples.

Building up to a characterization of the variance of the treatment-effect estimator in terms of out-of-sample prediction quality, I first state an auxiliary remark that will simplify the proof of the main result.

Remark A.3 (K-fold variance bound)**.** *Consider n square-integrable, mean-zero random variables $a_1, \ldots, a_n$ and a partition $\bigcup_{k=1}^{K} \mathcal{J}_k = \{1, \ldots, n\}$ such that, for all k, $\text{E}[a_i a_j] = 0$ for all $i, j \in \mathcal{J}_k$. Then,*

$$\text{Var}\left(\textstyle\sum_{i=1}^{n} a_i\right) \leq K \displaystyle\sum_{i=1}^{n} \text{Var}(a_i).$$

Proof. By Cauchy-Schwarz, applied once per row, we find that

$$\begin{aligned}\text{Var}\left(\textstyle\sum_{i=1}^{n} a_i\right) = \text{Var}\left(\textstyle\sum_{k=1}^{K}\sum_{i\in\mathcal{J}_k} a_i\right) &\leq \left(\textstyle\sum_{k=1}^{K}\sqrt{\text{Var}\left(\sum_{i\in\mathcal{J}_k} a_i\right)}\right)^2 \\ &\leq K\displaystyle\sum_{k=1}^{K}\text{Var}\left(\textstyle\sum_{i\in\mathcal{J}_k} a_i\right) = K\displaystyle\sum_{k=1}^{K}\sum_{i\in\mathcal{J}_k}\text{Var}(a_i),\end{aligned}$$

where the last equality follows because increments are uncorrelated within folds. □

I assume that potential outcomes and control variables are drawn iid from a population

distribution

$$(y_i(1), y_i(0), x_i) \overset{\text{iid}}{\sim} \mathrm{P},$$

treatment is assigned according to a known treatment probability $\mathrm{P}(d_i = 1) = p \in (0, 1)$, and data (y_i, d_i, x_i) obtained from $y_i = y_i(d_i)$.

In this section, I focus on K-fold estimators similar to those in Remark 1.1. Specifically, I assume that a sample of size n is divided into K equally-sized folds

$$\bigcup_{k=1}^{K} \mathcal{J}_k = \{1, \ldots, n\}$$

(so I implicitly assume that K divides n). In this setting, I consider the asymptotic distribution of the estimator

$$\hat{\tau} = \frac{1}{n} \sum_{k=1}^{K} \sum_{i \in \mathcal{J}_k} \frac{d_i - p}{p(1-p)} (y_i - \hat{f}_k(x_i))$$

of the population-average treatment effect $\tau = \mathrm{E}[y(1) - y(0)]$, where each $\hat{f}_k : \mathcal{X} \to \mathbb{R}$ is fitted only on folds other than $\mathcal{J}_k$. My first result characterizes the asymptotic distribution of $\hat{\tau}$. Throughout, I use indices i and k outside sums for a representative draw from the respective distribution.

Theorem A.3 (Asymptotic distribution of K-fold estimator)**.** *Assume that*

1. $\mathrm{E}[\mathrm{Var}(\hat{f}_k(x_i)|x_i)] \to 0$ *as* $n \to \infty$,
2. $\mathrm{E}\left[\left(\frac{1-p}{p}\right)^{2d_i - 1} (y_i - \hat{f}_k(x_i))^2\right] \to L$ *(where* $i \in \mathcal{J}_k$*), and*
3. $\mathrm{E}[(\hat{f}_k(x_i) - y_i)^{2+\delta}] < C < \infty$ *for some* $\delta, C > 0$.

Then,

$$\sqrt{n}(\hat{\tau} - \tau) \xrightarrow{d} \mathcal{N}(0, s^2), \qquad s^2 = \frac{L}{p(1-p)} - \tau^2.$$

Note that the distribution of prediction functions $\hat{f}_k$ will depend on the sample size of the training sample, and thus on n. Furthermore, the result can be extended to the case where

the population distribution itself depends on n. While I assume that K is fixed here, the conclusion also holds with K growing provided that $K \,\mathrm{E}[\mathrm{Var}(\hat{f}_k(x_i)|x_i)] \to 0$.

The first condition expresses that the prediction variance vanishes and predictions stabilize in large samples. The second condition defines the asymptotic prediction loss of the algorithm. The third condition is a regularity assumption that will ensure asymptotic convergence. When this condition holds, I do not require the assumption of bounded support of potential outcomes from the main chapter. Importantly, I do not assume that the prediction functions approximate the best prediction of y given x or are risk-consistent, only that their variance vanishes.

Proof of Theorem A.3. Write $t_i = \frac{d_i - p}{p(1-p)}$. I decompose

$$\begin{aligned}
\sqrt{n}(\hat{\tau} - \tau) &= \frac{1}{\sqrt{n}} \sum_{k=1}^{K} \sum_{i \in \mathcal{J}_k} (t_i(y_i - \hat{f}_k(x_i)) - \tau) \\
&= \frac{1}{\sqrt{n}} \sum_{k=1}^{K} \sum_{i \in \mathcal{J}_k} (t_i(y_i - \underbrace{\mathrm{E}[\hat{f}_k(x_i)|x_i]}_{=g_n(x_i)}) + t_i(\mathrm{E}[\hat{f}_k(x_i)|x_i] - \hat{f}_k(x_i)) - \tau) \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (t_i(y_i - g_n(x_i)) - \tau) + \frac{1}{\sqrt{n}} \sum_{k=1}^{K} \sum_{i \in \mathcal{J}_k} t_i(\hat{f}_k(x_i) - g_n(x_i)).
\end{aligned}$$

For the first part, note that $\mathrm{E}[(t_i(y_i - g_n(x_i)) - \tau)^{2+\delta}]$ is bounded, uniformly in n. Its expectation is zero and its variance is

$$\begin{aligned}
s_n^2 &= \mathrm{Var}\left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (t_i(y_i - g_n(x_i)) - \tau) \right) = \mathrm{Var}\left(t_i(y_i - g_n(x_i))\right) \\
&= \mathrm{E}\big[\underbrace{t_i^2}_{=\left(\frac{d_i - p}{p(1-p)}\right)^2 = \frac{1}{p(1-p)}\left(\frac{1-p}{p}\right)^{2d_i - 1}} (y_i - g_n(x_i))^2 \big] - (\underbrace{\mathrm{E}[t_i(y_i - g_n(x_i))]}_{=\tau})^2 \\
&= \frac{\mathrm{E}\left[\left(\frac{1-p}{p}\right)^{2d_i - 1} (y_i - g_n(x_i))^2 \right]}{p(1-p)} - \tau^2.
\end{aligned}$$

Hence, by the Lyapunov CLT for triangular arrays,

$$\frac{1}{\sqrt{n s_n^2}} \sum_{i=1}^{n} (t_i(y_i - g_n(x_i)) - \tau) \xrightarrow{d} \mathcal{N}(0, 1).$$

Combining the first two assumptions,

$$\mathrm{E}\left[\left(\frac{1-p}{p}\right)^{2d_i-1}(y_i - g_n(x_i))^2\right] \to L,$$

so we obtain that $s_n^2 \to s^2 = \frac{L}{p(1-p)} - \tau^2$ and thus

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(t_i(y_i - g_n(x_i)) - \tau) \xrightarrow{d} \mathcal{N}(0, s^2).$$

For the second part, by Remark A.3,

$$\begin{aligned}
&\mathrm{Var}\left(\frac{1}{\sqrt{n}}\sum_{k=1}^{K}\sum_{i\in\mathcal{J}_k} t_i(\hat{f}_k(x_i) - g_n(x_i))\right) \\
&\leq \frac{K}{n}\sum_{k=1}^{K}\sum_{i\in\mathcal{J}_k}\mathrm{Var}\left(t_i(\hat{f}_k(x_i) - g_n(x_i))\right) = K\,\mathrm{E}\left[t_i^2(\hat{f}_k(x_i) - g_n(x_i))^2\right] \\
&= K\,\mathrm{E}\left[\left(\frac{d_i - p}{p(1-p)}\right)^2\right]\mathrm{E}\left[(\hat{f}_k(x_i) - g_n(x_i))^2\right] \\
&= \frac{K}{p(1-p)}\,\mathrm{E}\left[(\hat{f}_k(x_i) - \mathrm{E}[\hat{f}_k(x_i)|x_i])^2\right] = \frac{K}{p(1-p)}\,\mathrm{E}\left[\mathrm{Var}(\hat{f}_k(x_i)|x_i)\right] \longrightarrow 0
\end{aligned}$$

as $n \to \infty$. In particular,

$$\frac{1}{\sqrt{n}}\sum_{k=1}^{K}\sum_{i\in\mathcal{J}_k} t_i(\hat{f}_k(x_i) - g_n(x_i)) \xrightarrow{\mathrm{P}} 0.$$

The claim of the theorem follows. □

The asymptotic variance is a function of the expected prediction loss and the treatment effect, and can be estimated consistently from the sample analogs.

Remark A.4 (Asymptotically valid variance estimate)**.** *Under the assumptions of Theorem A.3, the asymptotic variance of $\hat{\tau}$ can be estimated consistently by*

$$\hat{s}^2 = \frac{1}{n-1}\sum_{k=1}^{K}\sum_{i\in\mathcal{J}_k}\left(\frac{d_i - p}{p(1-p)}(y_i - \hat{f}_k(x_i)) - \hat{\tau}\right)^2.$$

As a consequence, we can construct asymptotically valid standard errors and Normal-theory

confidence intervals from $\hat{s}^2$. To be more precise, $\frac{\hat{s}}{\sqrt{n}}$ is a valid standard error for $\hat{\tau}$, and

$$[\hat{\tau} - z_{1-\alpha/2}\frac{\hat{s}}{\sqrt{n}}, \hat{\tau} + z_{1-\alpha/2}\frac{\hat{s}}{\sqrt{n}}]$$

a $1-\alpha$ confidence interval for τ (where $z_{1-\alpha/2}$ is the $1-\alpha/2$-quantile of the standard Normal distribution).

The asymptotic results extend to the case of fixed n_1 (by setting $p = n_1/n$, provided that $\mathrm{E}[\hat{f}_k(x_i)] \to \mathrm{E}[\bar{y}_i]$), exact cross-fitting as in Remark 1.1 with balanced folds, and folds that are only approximately of the same size or only approximately balanced.

Now that we have established asymptotically valid inference, I am ready to return to preference alignment.

Remark A.5 (Alignment over precision)**.** *Assume the investigator chooses among unbiased estimators, that is, by Lemma 1.1 among regression adjustments. Assume further that she constructs regression adjustments in a K-fold procedure with (a sequence of) prediction functions that fulfill the regularity assumptions for asymptotically valid inference in Theorem A.3. Then, if the investigator wants to obtain small standard errors or tight confidence intervals, her choices are aligned with the designer's preference for low mean-squared error $\mathrm{E}[(\hat{\tau}-\tau)^2]$ among these unbiased estimators.*

Proof. The asymptotic distribution of $\hat{\tau}$ as well as the probability limit of $\hat{s}^2$ only depend on the asymptotic loss L, the treatment probability p, and the treatment effect τ. The investigator through her choice of adjustments can only control L, and for these preferences chooses a sequence of prediction functions that minimizes asymptotic prediction loss. This is also the variance-minimizing choice the designer prefers. (Since L is non-random, the specific utility function over the size of standard errors or confidence intervals does not matter here.) □

Note that unbiasedness is crucial to reduce the degrees of freedom over the asymptotic distribution to the variance, with respect to which designer and investigator are aligned. Conversely, designer and investigator may have different preferences over the bias-variance trade-off, so allowing for (asymptotic) bias would break alignment even when the estimator is asymptotically Normal.

By the same argument as in the proof of Remark A.5, choices are also aligned over the power of a test against some null hypothesis. Since the investigator cannot move the expectation of the estimator, the best she can do is to pick a sequence of prediction functions for which the asymptotic loss L is minimal.

Remark A.6 (Alignment over power)**.** *Consider a sequence of population distributions with* $\tau_n = \tau_0 + \frac{\delta}{\sqrt{n}}$*. Assume that the investigator constructs a one- or two-sided test against the null hypothesis* $\tau = \tau_0$ *by comparing the test statistic* $\frac{\sqrt{n}(\hat{\tau}-\tau_0)}{\hat{s}}$ *to the standard Normal distribution, and that the investigator's (sequence of) prediction functions fulfill the regularity assumptions in Theorem A.3. If the investigator has a preference for rejecting* $\tau = \tau_0$*, then her choices are aligned with the designer's goal of minimizing* $\mathrm{E}[(\hat{\tau} - \tau)^2]$.

Based on the asymptotic approximation from Theorem A.3, I am now ready to prove the result from the main chapter that distribution to two researchers attains asymptotic efficiency.

Remark 1.4 (Semi-parametric efficiency)**.** *If researchers use prediction algorithms* $(A_n : \mathcal{Z} \to \mathbb{R}^{\mathcal{X}}, z \mapsto \hat{f}_n)_{n=1}^{\infty}$ *with*

$$\mathrm{E}[(\hat{f}_n(x_i) - \mathrm{E}[\bar{y}_i|x_i])^2] \to 0$$

as $n \to \infty$*, then delegation to two researchers with risk functions in* $\mathcal{R}^*$ *(who each obtain access to half of the data, say) without further commitment achieves both finite-sample unbiased estimation of* τ_θ*, and large-sample semi-parametric efficient estimation of* τ *for the semi-parametric efficiency bound of Hahn (1998).*

Proof of Remark 1.4. Similar to the proof of Theorem A.3, again setting $t_i = \frac{d_i - p}{p(1-p)}$, I decomposem, with $K = 2$,

$$\begin{aligned}
\sqrt{n}(\hat{\tau} - \tau) &= \frac{1}{\sqrt{n}} \sum_{k=1}^{K} \sum_{i \in \mathcal{J}_k} (t_i(y_i - \hat{f}_k(x_i)) - \tau) \\
&= \frac{1}{\sqrt{n}} \sum_{k=1}^{K} \sum_{i \in \mathcal{J}_k} (t_i(y_i - \mathrm{E}[\bar{y}_i|x_i]) + t_i(\mathrm{E}[\bar{y}_i|x_i] - \hat{f}_k(x_i)) - \tau) \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (t_i(y_i - \mathrm{E}[\bar{y}_i|x_i])) - \tau) + \frac{1}{\sqrt{n}} \sum_{k=1}^{K} \sum_{i \in \mathcal{J}_k} t_i(\hat{f}_k(x_i) - \mathrm{E}[\bar{y}_i|x_i])).
\end{aligned}$$

The latter part converges to zero in probability by Remark A.3 as in the proof of Theorem A.3. Since the support of potential outcomes is bounded, the first part converges by the standard CLT to a mean-zero Normal distribution with asymptotic variance

$$\operatorname{Var}(t_i(y_i - \operatorname{E}[\bar{y}_i|x_i])) = \frac{\operatorname{E}\operatorname{Var}(y_i(1)|x_i)}{p} + \frac{\operatorname{E}\operatorname{Var}(y_i(0)|x_i)}{1-p} + \operatorname{Var}(\operatorname{E}[y_i(1) - y_i(0)|x_i]),$$

which is the efficiency bound of Hahn (1998). □

A.6 Hyperpriors and Optimal Biases

The minimax result in the main chapter establishes that *fixing* the bias is a minimax optimal solution to the designer's delegation problem that aligns the choices of the investigator with the goal of the designer. An optimal choice of biases, however, depends on the hyperprior of the designer, and zero as a choice is not an optimal solution in general.

In this section, I discuss one justifications for why I put special emphasis on unbiasedness coming from a specific notion of an uninformed designer. I then highlight that some hyperpriors, even on finite support, deliver zero bias as an exact solution. Finally, I discuss how the characterization of unbiased estimators extends to the case of other choices of the bias.

A.6.1 Uninformativeness and Zero Bias

Intuitively, a designer who has no systematic information about the location of the average treatment effect will set the biases to zero. If we assumed that the support of the outcome variables was continuous and unbounded, then one elegant formalization of this argument would capture uninformativeness about the location of the treatment effect by an invariance to translation actions in that direction, yielding an improper hyperprior that would deliver zero bias under an appropriate criterion. Since my chapter is, however, formulated for finite support, and since dealing with improper priors would bring with it additional technical complications, I propose here one way of obtaining (approximately) zero bias under a specific notion of (approximate) uninformativeness in order to highlight the connection between invariances and bias.

In order to illustrate one construction of an approximately uninformative hyperprior, I start with an arbitrary hyperprior η over priors with (full) support in the grid

$$\Theta_0 = \mathcal{Y}_k^{2n} \qquad\qquad \mathcal{Y}_k = \{-k, -k+1, \ldots, -1, 0, 1, \ldots, k-1, k\}$$

for some $k \in \mathbb{N}$. (I am choosing an equally-spaced grid for convenience.) From η I construct increasingly uninformative hyperpriors η^m over priors with support $\Theta_m = \mathcal{Y}_{k+m}^{2n}$ for all $m \in \mathbb{N}_0$.

In order to construct the hyperprior η^m for $m \geq 0$, consider

$$g = (r, t) \in \{-1, 1\}^{2n} \times \mathcal{Y}_m^{2n} = G_m$$

and define the action of G_m on $\mathbb{Z}^{2n}$ by $g \circ \theta = (r_i \cdot \theta_i + t_i)$. (Note that g maps Θ_0 to Θ_m.) The distribution η over priors π on Θ_0 implies a distribution $g \circ \eta$ over priors $g \circ \pi$ (defined by $(g \circ \pi)(g \circ \theta) = \pi(\theta)$) on $g \circ \Theta_0 \subseteq \Theta_m$ that extends to a distribution on Θ_m. The distribution η^m over priors with support in Θ_m is then given by the composition of Uniform(G_m) and η^m that first draws a random action $\tilde{g}$ and then independently draws a prior over Θ_m according to $\tilde{g} \circ \eta^m$.

This construction yields hyperpriors that are increasingly uninformative about the location of outcomes in that they exhibit more and more symmetries with respect to reflection and translation of the data. Writing $(\beta_\theta^m)_{\theta \in \Theta_m}$ for the biases chosen optimally by the designer according to the hyperior η^m constructed in this way, any bias β_θ^m will therefore approach zero as the support grows.

Remark A.7 (Approximate unbiasedness)**.** *For any fixed $\theta \in \mathbb{Z}^{2n}$ (with m_0 large enough such that $\theta \in \Theta_m$),* $\lim_{\substack{m \to \infty \\ m \geq m_0}} \beta_\theta^m = 0$.

Note that the priors drawn from η^m do not have full support, which could be rectified by taking appropriate approximations. Note also that in this example I am using invariances in the outcomes, and not just in the treatment effects, so a smaller class of invariances may suffice to obtain a similar result.

A similar approach would start with an (improper) invariant hyperpior over priors with support in $\mathbb{Z}^{2n}$ and then consider restrictions of that distribution to an increasing sequence of

finite support sets, showing similarly that as the support grows and the hyperprior approaches the uninformative hyperprior, the optimal biases shrink to zero.

Proof idea. By symmetry, the optimal bias at the origin $\mathbf{0}$ within any support Θ_m subject to the hyperprior η^m is zero, $\beta^m_{\mathbf{0}} = 0$. Similarly, at fixed $\theta \in \Theta_{m_0}$ and for $m \geq m_0 + k$, the optimal bias would be zero if we conditioned η^m on

$$\max_{i \in \{1,\ldots,2n\}} |\theta_i - \tilde{t}| \leq (m - m_0)$$

in $\tilde{g} = (\tilde{r}, \tilde{t})$, since this would make θ the center of symmetry. Since

$$\mathrm{P}_{\eta^m}\left(\max_{i \in \{1,\ldots,2n\}} |\theta_i - \tilde{t}| \leq (m - m_0)\right) \to 1$$

as $m \to \infty$, the argument extends to the unconditional optimization. □

A.6.2 Zero Bias as a Minimax Solution

In my main setup, the designer optimizes against a worst-case risk, and averages over a (hyper-)prior over the investigator's prior information. One approach of fixing the bias would replace the hyperprior with assuming a worst-case prior. However, without restrictions on the priors and for a fixed, finite support, such a minimax solution would be driven by priors that put full weight on extreme outcome values, which is econometrically unappealing.

Rather, I propose a minimax approach to fixing the biases that includes uncertainty about the location of the outcomes. For generality, I formulate this result on the level of uncertainty about hyperpriors, and then return to implications for uncertainty over priors. I follow the construction and nomenclature from the above discussion of uninformativeness.

Specifically, I start with a set H (which can be a singleton) of hyperpriors, where for each $\eta \in H$ the priors in the support of η have as support the grid $\mathcal{Y}_k^{2n}$ for

$$\mathcal{Y}_k = \{-k, -k+1, \ldots, -1, 0, 1, \ldots, k-1, k\}.$$

Consider

$$g = (r, t) \in \{-1, 1\}^{2n} \times \mathbb{Z}^{2n} = G$$

and, similar to the above, define the action of G on $\mathbb{Z}^{2n}$ by $g \circ \theta = (r_i \cdot \theta_i + t_i)$. The distribution η over priors π on $\mathcal{Y}_k^{2n}$ implies a distribution $g \circ \eta$ over priors $g \circ \pi$ (defined by $(g \circ \pi)(g \circ \theta) = \pi(\theta)$) on $g \circ \mathcal{Y}_k^{2n} \subseteq \mathbb{Z}^{2n}$. From that, I obtain the set of hyperpriors

$$H^* = G \circ H = \{g \circ \eta; g \in G, \eta \in H\}.$$

Following the logic of the proof of Remark A.7, the invariances of H^* imply that an investigator who optimizes against a worst-case hyperprior in H^* chooses zero bias as a minimax (in risk and hyperprior) optimal restriction of this form:

Remark A.8 (Minimax optimality of zero bias)**.** *The unbiased estimators are minimax optimal for the invariant set H^* of hyperpriors in the sense that the choice $\beta_\theta = 0$ for all $\theta \in \mathbb{Z}^{2n}$ minimizes (among fixed-bias restrictions)*

$$\sup_{\eta \in H^*} \sup_{r^I \in \mathcal{R}^*} \mathrm{E}_\eta \left[r_\theta^D \left(\arg\min_{\hat{\tau} \in \mathcal{C}} \mathrm{E}_\pi [r_\theta^I(\hat{\tau})] \right) \right].$$

As a special case, this result includes the case where all hyperpriors are singletons, and the designer thus optimizes against a worst-case prior directly. Hence, for any set of priors Π with support $\mathcal{Y}_k^{2n}$ and

$$\Pi^* = G \circ \Pi = \{g \circ \pi; g \in G, \pi \in \Pi\},$$

the minimax result yields minimization of

$$\sup_{\pi \in \Pi^*} \sup_{r^I \in \mathcal{R}^*} \mathrm{E}_\pi \left[r_\theta^D \left(\arg\min_{\hat{\tau} \in \mathcal{C}} \mathrm{E}_\pi [r_\theta^I(\hat{\tau})] \right) \right].$$

Note that the priors in Π^* now have varying support.

A.6.3 Hyperprios with Exactly Zero Bias

There are hyperpriors that trivially yield zero bias, namely those that by virtue of $\text{Var}_\pi(\bar{y}_i|z_{-i}) = 0$ (for known p) allow the investigator to pick an unbiased estimator with zero loss (such as, in the case of $p = .5$, if the investigator knows $y_i(1) + y_i(0)$ from x_i). While these constitute extreme examples, they point towards a general intuition: if the investigator has strong private information about the choice of optimal adjustments, then setting a non-zero bias would create a burden by imposing loss that cannot be avoided.

A.6.4 When Bias is Optimal

If the designer has non-trivial information about the distribution of treatment effects, the remaining results in the chapter formulated in terms of unbiased estimation extend at least partly.

First, assume that the hyperprior of the designer implies (approximately) that $\text{E}_\theta[\hat{\tau}(z)] = (1 - \lambda)\tau_\theta$ is an optimal restriction (for $\lambda \in (0, 1)$), expressing a fixed shrinkage factor that is set by the designer. Then, the investigator still faces the same unbiased estimation problem as in the main chapter since the optimal shrunk estimator is the optimal unbiased estimator multiplied by that factor ex-post.

Second, even if no such structure is available, the results still extend with modifications. Note that the designer's solution can equivalently be phrased as choosing a reference estimator $\hat{\tau}^D$ (with the desired biases) and letting the investigator choose mean-zero adjustments $\hat{\delta}^I$ to obtain an estimator

$$\hat{\tau}(z) = \hat{\tau}^D(z) + \hat{\delta}^I(z).$$

My characterization of unbiased regression adjustments directly yields a characterization of mean-zero adjustments that characterize the choice set of the investigator, only that the reference estimator has now changed. However, the optimal adjustments now take a different form. For example, in the case of $n = 1$ with known p, the optimal adjustment now takes the

general form

$$\phi_i = p(1-p)\,\mathrm{E}_\pi[\hat{\tau}^D(y = y(1), d = 1) - \hat{\tau}^D(y = y(0), d = 0)],$$

which precisely yields the familiar adjustment $\mathrm{E}_\pi[(1-p)y(1) + py(0)]$ when applied to the unbiased reference estimator $\hat{\tau}^D(z) = \frac{d-p}{p(1-p)}y$.

A.7 Additional Proofs

In this section, I restate and sketch the proofs of the remaining results, which largely follow from the main results proved earlier.

Theorem 1.3 (Complete-class theorem for unbiased estimators)**.** *For any unbiased estimator $\hat{\tau}$ of the sample-average treatment effect that is not dominated with respect to variance, there is a converging sequence of priors $(\pi_t)_{t=1}^{\infty}$ with full support such that $\hat{\tau}$ equals the limit of the respective estimators in Theorem 1.2. Conversely, for any converging sequence of priors $(\pi_t)_{t=1}^{\infty}$ that put positive weight on every state $\theta \in \Theta$, every converging subsequence of corresponding estimators is admissible among unbiased estimators.*

Proof. Note first that, for π with full support, the estimator that minimizes average variance among unbiased estimators is unique (even though the representation in Theorem 1.2 in general is not). Indeed, among unbiased estimators the investigator minimizes (conflating the distribution of θ)

$$\mathrm{E}_\pi[(\hat{\tau}(z) - \tau_\theta)^2] = \mathrm{E}_\pi[(\hat{\tau}(z) - \mathrm{E}_\pi[\tau_\theta|z])^2] + \mathrm{E}_\pi[(\mathrm{E}_\pi[\tau_\theta|z] - \tau_\theta)^2].$$

Hence, within the affine linear subspace of $\mathbb{R}^{\mathcal{Z}}$ given by the unbiased estimators, the investigator chooses the point $\hat{\tau}$ closest to $(z \mapsto \mathrm{E}_\pi[\tau_\theta|z])_{z\in\mathcal{Z}}$ according to the weighted (with positive weights) Euclidean distance

$$d(\hat{\tau}_1, \hat{\tau}_2) = \mathrm{E}_\pi[(\hat{\tau}_2(z) - \hat{\tau}_1(z))^2]$$

(with the distribution over $\mathcal{Z}$ implied by π through draws of θ). Hence, the investigator's

solution is unique when π has full support.

Since every limiting estimator is the limit of Bayes estimators with full support, with finite domain and bounded codomain, any such limiting estimator is admissible. Since the state space is finite, every admissible estimator is Bayes (e.g. Ferguson, 1967, Chapter 2). If the estimator is Bayes with respect to a prior with full support, it is unique and therefore has an adjustment representation of the claimed form. If the corresponding prior does not have full support, we can write it as a limit of admissible estimators that are Bayes with respect to priors with full support and thus unique, so the estimator is a limit of estimators of the claimed form. □

Corollary 1.1 (Characterization of fixed-bias K-fold distribution contracts)**.** *For K disjoint folds $\mathcal{I}_k \subseteq \{1, \ldots, n\}$ with projections $g_k : (y, d) = z \mapsto z_{-\mathcal{I}_k} = (y_i, d_i)_{i \neq \mathcal{I}_k}$, a K-distribution contract $\hat{\tau}^\Phi$ has given bias if and only if:*

1. *For a known treatment probability p, there exist a fixed estimator $\hat{\tau}_0(z)$ with the given bias and regression adjustment mappings $(\Phi_k)_{k=1}^K$ such that*

$$\hat{\tau}^\Phi((\hat{\phi}_k)_{k=1}^K; z) = \hat{\tau}_0(z) - \frac{1}{n} \sum_{k=1}^K \sum_{i \in \mathcal{I}_k} \frac{d_i - p}{p(1-p)} \phi_i^k(z_{-i})$$

   where $(\phi_i^k)_{i \in \mathcal{I}_k} = \Phi_k(\hat{\phi}_k(z_{-\mathcal{I}_k}))$.

2. *For a fixed number n_1 of treated units, there exist a fixed estimator $\hat{\tau}_0(z)$ with the given bias and regression adjustment mappings $(\Phi_k)_{k=1}^K$ such that*

$$\hat{\tau}^\Phi((\hat{\phi}_k)_{k=1}^K; z) = \hat{\tau}_0(z) - \frac{1}{n_1 n_0} \sum_{k=1}^K \sum_{\{i<j\} \subseteq \mathcal{I}_k} (d_i - d_j) \phi_{ij}^k(z_{-ij}),$$

   where $(\phi_i^k)_{i \in \mathcal{I}_k} = \Phi_k(\hat{\phi}_k(z_{-\mathcal{I}_k}))$.

Proof. The result is a special case of Lemma 1.3 for this specific choice of the functions g_k. □

Remark 1.1 (Exact K-fold cross-fitting)**.** *For a partition of the sample*

$$\{1, \ldots, n\} = \bigcup_{k=1}^K \mathcal{I}^{(k)}$$

into K folds with $n^{(k)} \geq 2$ units each of which $n_1^{(k)} > 0$ treated and $n_0^{(k)} > 0$ untreated, the estimator

$$\hat{\tau}(z) = \frac{1}{n} \sum_{k=1}^{K} n^{(k)} \sum_{i \in \mathcal{I}^{(k)}} \frac{d_i n^{(k)} - n_1^{(k)}}{n_1^{(k)} n_0^{(k)}} \left(y_i - \phi_i^{(k)}(z_{-\mathcal{I}^{(k)}}) \right)$$

is unbiased for the sample-average treatment effect τ conditional on $(\mathcal{I}^{(k)})_{k=1}^K$ and $(n_1^{(k)})_{k=1}^K$ under either randomization. The investigator obtains their constrained optimal (Bayes) $\hat{\tau}$ among these estimators at

$$\phi_i^{(k)}(z_{-\mathcal{I}^{(k)}}) = \mathrm{E}_\pi[n_0^{(k)} y_i(1) + n_1^{(k)} y_i(0) | z_{-\mathcal{I}^{(k)}}] / n^{(k)}.$$

Proof. Unbiasedness is immediate from Lemma 1.1. Optimality of this choice of adjustments follows as in the proof of Theorem 1.2. □

Lemma 1.3 (Characterization of unbiased K-distribution contracts)**.** *A K-distribution contract $\hat{\tau}^\Phi$ is unbiased for the sample-average treatment effect τ_θ for any conformable researcher input $(\hat{\phi}_k)_{k=1}^K$ if and only if:*

1. *For known treatment probability p, there exist regression adjustments $(\phi_i : (\bigtimes_{k \in C_i} B_k) \times (\mathcal{Y} \times \{0,1\})^{n-1} \to \mathbb{R})_{i=1}^n$ such that*

   $$\hat{\tau}^\Phi((\hat{\phi}_k)_{k=1}^K; z) = \frac{1}{n} \sum_{i=1}^{n} \frac{d_i - p}{p(1-p)} (y_i - \phi_i((\hat{\phi}_k(g_k(z))_{k \in C_i}; z_{-i}))$$

   for $C_i = \{k; g_k(z) = \tilde{g}(z_{-i})$ for some $\tilde{g}\}$.

2. *For fixed number n_1 of treated units, there exist regression adjustments $(\phi_{ij} : (\bigtimes_{k \in C_{ij}} B_k) \times (\mathcal{Y} \times \{0,1\})^{n-2} \to \mathbb{R})_{i<j}$ such that*

   $$\hat{\tau}^\Phi((\hat{\phi}_k)_{k=1}^K; z) = \frac{1}{n_1 n_0} \sum_{i<j} (d_i - d_j)(y_i - y_j - \phi_{ij}((\hat{\phi}_k(g_k(z))_{k \in C_{ij}}; z_{-ij})),$$

   for $C_{ij} = \{k; g_k(z) = \tilde{g}(z_{-ij})$ for some $\tilde{g}\}$.

Proof. Since the resulting estimator must be unbiased, and researcher choices are themselves functions of the data made available to them, the result follows directly from the general representation result of unbiased estimators (Lemma 1.1). □

Theorem 1.4 (Hybrid pre-analysis plan dominates rigid pre-analysis plan)**.** *Assume that investigator and researcher have risk functions in $\mathcal{R}^*$. The optimal unbiased pre-committed estimator $\hat{\tau}^{\text{pre}}$ is strictly dominated by an unbiased hybrid pre-analysis plan with respect to average variance, i.e. the hybrid plan is as least as precise on average over any ex-ante prior η^I and strictly better for many non-trivial ex-ante priors η^I.*

Proof. A researcher with risk function in $\mathcal{R}^*$ minimizes variance among unbiased estimators. Since the original adjustments corresponding to $\hat{\tau}^{\text{pre}}$ are available to the researcher, her choice can only reduce variance on average over her prior. Unless the ex-post changes are ineffectual, this will strictly improve variance averaged over the hyperprior. □

Remark 1.2 (Optimal hybrid pre-analysis plan)**.** *The dominating hybrid plan is:*

1. *For known treatment probability p, the researcher chooses regression adjustments $(\phi_i^{\text{post}} : (\mathcal{Y} \times \{0,1\})^{n-1} \to \mathbb{R})_{i \notin T} = \hat{\phi}(z_T)$ to obtain*

$$\hat{\tau}^{\text{hybrid}}(\hat{\phi}; z) = \hat{\tau}^{\text{pre}}(z) - \frac{1}{n} \sum_{i \notin T} \frac{d_i - p}{p(1-p)} \, \phi_i^{\text{post}}(z_{-i})$$

   where $1 \leq |T| \leq n-1$.

2. *For fixed number n_1 of treated units, the researcher chooses adjustments $(\phi_{ij}^{\text{post}} : (\mathcal{Y} \times \{0,1\})^{n-2} \to \mathbb{R})_{\{i<j\} \cap T = \emptyset} = \hat{\phi}(z_T)$ to obtain*

$$\hat{\tau}^{\text{hybrid}}(\hat{\phi}; z) = \hat{\tau}^{\text{pre}}(z) - \frac{1}{n_1 n_0} \sum_{\{i<j\} \cap T = \emptyset} (d_i - d_j) \, \phi_{ij}^{\text{post}}(z_{-ij})$$

   where $1 \leq |T| \leq n-2$.

In both cases, the investigator commits to the training sample $T \subseteq \{1, \ldots, n\}$ and the unbiased estimator $\hat{\tau}^{\text{pre}} : \mathcal{Z} \to \mathbb{R}$.

Proof. This is a special case of Corollary 1.1. □

Remark 1.3 (More researchers are better)**.** *Assume that the investigator and researchers all have risk functions in $\mathcal{R}^*$, and that the researchers all share the same (ex-post) prior π. Then*

an optimal unbiased K-distribution contract is dominated by an unbiased $K+1$-distribution contract in the sense of Theorem 1.4.

Proof. The result follows by the revealed-preference argument in the proof of Theorem 1.4. □

Note that to obtain this result I assume that all researchers have the same prior, which renders the proof trivial, but represents an unrealistic assumption. A more attractive result would assume that the K researchers each obtain a draw from the same hyperprior η (where draws are correlated between each other and with the true distribution, also drawn from η, of θ), and I conjecture that in this case more researchers still improve average estimation quality.

Appendix B

Appendix to Chapter 3

B.1 General Convergence of Matched Sums

In this section, we derive general convergence results for sums within the matched sample $\mathcal{S}^*$ that we will later use to establish consistency and asymptotic Normality of the various estimators in this article. The main tool behind these lemmas is a martingale representation similar to Abadie and Imbens (2012).

Let $F(Y, W, S)$ be a $(s \times t)$-matrix of real-valued measurable functions,

$$\Phi_1(x) = E[F(Y,W,S)|W=1, X=x], \qquad \Phi_0(x) = E[F(Y,W,S)|W=0, X=x],$$

$$\widehat{\Phi} = \frac{1}{n}\sum_{i=1}^{n} F(Y_{ni}, W_{ni}, S_{ni}),$$

and

$$\Phi = E^*[F(Y,T,S)] = E\left[\frac{1}{M+1}\Phi_1(X) + \frac{M}{M+1}\Phi_0(X)\Big|W=1\right].$$

Lemma B.1

Under Assumptions 3.1 to 3.3, and if

(a.1) $\Phi_0(\cdot)$ *is (component-wise) Lipschitz on* $\mathcal{X}_0$*,*

(a.2) $E[\|F(Y,W,S)\|^2|W=w, X=x]$ *is uniformly bounded on* $\mathcal{X}_w$*, for* $w \in \{0,1\}$*,*

then $\widehat{\Phi} \xrightarrow{p} \Phi$.

Proof. Because $\widehat{\Phi}$ converges in probability if and only if each of its components converges, we assume without loss of generality that $s = t = 1$. We decompose

$$\widehat{\Phi} = \frac{1}{n}\sum_{i=1}^{N} W_i\Big(\Phi_1(X_i) + M\Phi_0(X_i)\Big) + \frac{1}{n}\sum_{i=1}^{n}\Big(F(Y_{ni}, W_{ni}, S_{ni}) - \Phi_{W_{ni}}(X_{ni})\Big) + \frac{1}{n}\sum_{i=1}^{N} W_i \sum_{j\in\mathcal{J}(i)}\Big(\Phi_0(X_j) - \Phi_0(X_i)\Big).$$

The first term on the right-hand side of the last equation is a sum of iid random variables. Hence, by the weak law of large numbers, we have that

$$\frac{1}{n}\sum_{i=1}^{n} W_i\Big(\Phi_1(X_i) + M\Phi_0(X_i)\Big) \xrightarrow{p} E\left[\frac{1}{M+1}\Phi_1(X) + \frac{M}{M+1}\Phi_0(X)\Big|W = 1\right] = \Phi.$$

For the second sum, notice that,

$$\begin{aligned}&\mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n}\Big(F(Y_i, W_i, S_i) - \Phi_{W_i}(X_i)\Big)\,\middle|\, \begin{array}{l} X_1,\ldots,X_N, \\ W_1,\ldots,W_N \end{array}\right) \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(F(Y,W,S)|W = W_i, X = X_i)\end{aligned}$$

which (by Assumption *(a.2)* in the lemma) is bounded by a sequence that converges to zero. By the law of total variance, we obtain

$$\mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n}\Big(F(Y_i, W_i, S_i) - \Phi_{W_i}(X_i)\Big)\right) \longrightarrow 0.$$

For the third sum, Assumption *(a.1)* in the lemma implies

$$\begin{aligned}\left|\frac{1}{n}\sum_{i=1}^{n} W_i \sum_{j\in\mathcal{J}(i)}\Big(\Phi_0(X_j) - \Phi_0(X_i)\Big)\right| &\le \frac{1}{n}\sum_{i=1}^{n} W_i \sum_{j\in\mathcal{J}(i)}\Big|\Phi_0(X_j) - \Phi_0(X_i)\Big| \\ &\le \frac{L}{\sqrt{n}}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} W_i \sum_{j\in\mathcal{J}(i)} d(X_j, X_i)\right) \xrightarrow{p} 0,\end{aligned}$$

for some Lipschitz constant L. □

Lemma B.2

In the setup of Lemma B.1, let $t = 1$, and define

$$\Psi_1(x) = \mathrm{Var}(F(Y,W,S)|W=1, X=x), \quad \Psi_0(x) = \mathrm{Var}(F(Y,W,S)|W=0, X=x),$$

which are $(s \times s)$-matrices. Suppose that, in addition to the assumptions of Lemma B.1, we have

(a.3) $\Psi_0(\cdot)$ is (component-wise) Lipschitz on $\mathcal{X}_0$,

(a.4) $E[\|F(Y,W,S)\|^{2+\delta}|W=w, X=x]$ is uniformly bounded on $\mathcal{X}_w$ for all $w \in \{0,1\}$ and some $\delta > 0$.

Then,

$$\sqrt{n}(\widehat{\Phi} - \Phi) \xrightarrow{d} \mathcal{N}(0, V^*)$$

where

$$V^* = \frac{\mathrm{Var}\Big(\Phi_1(X) + M\Phi_0(X)\big|W=1\Big)}{M+1} + \frac{E\Big[\Psi_1(X) + M\Psi_0(X)\big|W=1\Big]}{M+1}.$$

Proof. Fix $\lambda \in \mathbb{R}^s$. We decompose

$$\begin{aligned}
&\sqrt{n}(\widehat{\Phi} - \Phi)'\lambda \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^n W_i\Big(\Phi_1(X_i) + M\Phi_0(X_i) - \Phi(X_i)\Big)'\lambda + \frac{1}{\sqrt{n}}\sum_{i=1}^n \Big(F(Y_i,W_i,S_i) - \Phi_{W_i}(X_i)\Big)'\lambda \\
&+ \frac{1}{\sqrt{n}}\sum_{i=1}^n W_i \sum_{j\in\mathcal{J}(i)} \Big(\Phi_0(X_j) - \Phi_0(X_i)\Big)'\lambda.
\end{aligned}$$

The last term on the right-hand side of last equation vanishes in probability:

$$\begin{aligned}
\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n W_i \sum_{j\in\mathcal{J}(i)} \Big(\Phi_0(X_j) - \Phi_0(X_i)\Big)'\lambda\right| &\le \frac{1}{\sqrt{n}}\sum_{i=1}^n W_i \sum_{j\in\mathcal{J}(i)} \big\|\Phi_0(X_j) - \Phi_0(X_i))\big\|\|\lambda\| \\
&\le \frac{\|\lambda\|L}{\sqrt{n}}\sum_{i=1}^n W_i \sum_{j\in\mathcal{J}(i)} d(X_j, X_i) \xrightarrow{p} 0
\end{aligned}$$

for an appropriate Lipschitz constant L.

The first two parts of the sum form a martingale. Consider the filtration

$$\mathcal{F}_i = \begin{cases} \sigma(W_1, \ldots, W_N, X_1, \ldots, X_i), & i \leq N_1, \\ \sigma(W_1, \ldots, W_N, X_1, \ldots, X_N, (Y_1, S_1), \ldots, (Y_{i-N}, S_{i-N})), & N_1{+}1 \leq i \leq N_1{+}n. \end{cases}$$

Then,

$$\xi_i = \begin{cases} \frac{1}{\sqrt{n}} W_i \Big(\Phi_1(X_i) + M\Phi_0(X_i) - \Phi \Big)' \lambda, & i \leq N_1, \\ \frac{1}{\sqrt{n}} \Big(F(Y_{i-N}, W_{i-N}, S_{i-N}) - \Phi_{W_{i-N}}(X_{i-N}) \Big)' \lambda, & N_1 + 1 \leq i \leq N_1 + n \end{cases}$$

is a martingale difference array with respect to the filtration $\mathcal{F}$. Also, notice that

$$\begin{aligned} \sum_{i=1}^{N_1+n} E[\xi_i^2 | \mathcal{F}_{i-1}] &= \frac{1}{n} \sum_{i=1}^{N_1} \operatorname{Var}\Big((\Phi_1(X) + M\Phi_0(X))' \lambda \Big| W = 1 \Big) \\ &\quad + \frac{1}{n} \sum_{i=1}^{n} \operatorname{Var}\Big(F(Y, W, S)' \lambda \Big| W = W_i, X = X_i \Big) \\ &= \frac{\lambda' \operatorname{Var}(\Phi_1(X) + M\Phi_0(X) | W = 1) \lambda}{1 + M} + \frac{1}{n} \sum_{i=1}^{n} \lambda' \Psi_{W_i}(X_i) \lambda, \end{aligned}$$

where the last term converges in probability to

$$\lambda' E \left[\frac{1}{M+1} \Psi_1(X) + \frac{M}{M+1} \Psi_0(X) \Big| W = 1 \right] \lambda,$$

by Lemma B.1. Hence,

$$\sum_{i=1}^{N_1+n} E[\xi_i^2 | \mathcal{F}_{i-1}] \xrightarrow{p} \lambda' V^* \lambda.$$

Next, note that

$$\begin{aligned} |\xi_i| &\leq \begin{cases} \frac{1}{\sqrt{n}} \|\Phi_1(X_i) + M\Phi_0(X_i) - \Phi\|_2 \|\lambda\|_2, & i \leq N_1, \\ \frac{1}{\sqrt{n}} \|F(Y_{i-N}, W_{i-N}, S_{i-N}) - \Phi_{W_{i-N}}(X_{i-N})\|_2 \|\lambda\|_2, & N_1 + 1 \leq i \leq N_1 + n \end{cases} \\ &\leq \begin{cases} \frac{1}{\sqrt{n}} (\|\Phi_1(X_i)\|_2 + M\|\Phi_0(X_i)\|_2 + \|\Phi\|_2) \|\lambda\|_2, & i \leq N_1, \\ \frac{1}{\sqrt{n}} (\|F(Y_{i-N}, W_{i-N}, S_{i-N})\|_2 + \|\Phi_{W_{i-N}}(X_{i-N})\|_2) \|\lambda\|_2, & N_1 + 1 \leq i \leq N_1 + n \end{cases} \end{aligned}$$

by the Cauchy-Schwarz and triangle inequalities. It follows that

$$E[|\xi_i|^{2+\delta}] \leq \begin{cases} \frac{\|\lambda\|_2^{2+\delta}}{n^{1+\delta/2}} E\left[(\|\Phi_1(X_i)\|_2 + M\|\Phi_0(X_i)\|_2 + \|\Phi\|_2)^{2+\delta}\right], & i \leq N_1, \\ \frac{\|\lambda\|_2^{2+\delta}}{n^{1+\delta/2}} E\left[(\|F(Y_{i-N}, W_{i-N}, S_{i-N})\|_2 + \|\Phi_{W_{i-N}}(X_{i-N})\|_2)^{2+\delta}\right], & i > N_1 \end{cases}$$

$$\leq \begin{cases} \frac{\|\lambda\|_2^{2+\delta}}{n^{1+\delta/2}} \left((E[\|\Phi_1(X_i)\|_2^{2+\delta}])^{1/(2+\delta)} + M(E[\|\Phi_0(X_i)\|_2^{2+\delta}])^{1/(2+\delta)} + (\|\Phi\|_2^{2+\delta})^{1/(2+\delta)}\right)^{2+\delta} \\ \frac{\|\lambda\|_2^{2+\delta}}{n^{1+\delta/2}} \left((E[\|F(Y_{i-N}, W_{i-N}, S_{i-N})\|_2^{2+\delta}])^{1/(2+\delta)} + (E[\|\Phi_{W_{i-N}}(X_{i-N})\|_2^{2+\delta}])^{1/(2+\delta)}\right)^{2+\delta} \end{cases}$$

where the latter inequality is implied by Minkowski's inequality. By assumption (a.3), note that for both $w \in \{0,1\}$ and $x \in \mathcal{X}_w$, by Jensen's inequality we have

$$\begin{aligned} \|\Phi_w(x)\|_2^{2+\delta} &= \|E[F(Y,W,S)|W=w, X=x]\|_2^{2+\delta}) \\ &\leq E[\|F(Y,W,S)\|_2^{2+\delta}|W=w, X=x] \leq C \end{aligned}$$

and hence $E[\|\Phi_w(X)\|_2^{2+\delta}] \leq C$, while also

$$\begin{aligned} \|\Phi\|_2^{2+\delta} &= \left\| E\left[\frac{1}{M+1}\Phi_1(X) + \frac{M}{M+1}\Phi_0(X)\Big| W=1\right]\right\|_2^{2+\delta} \\ &\leq E\left[\left\|\frac{1}{M+1}\Phi_1(X) + \frac{M}{M+1}\Phi_0(X)\right\|_2^{2+\delta}\Big| W=1\right] \\ &\leq E\left[\left(\frac{1}{M+1}C^{1/(2+\delta)} + \frac{M}{M+1}C^{1/(2+\delta)}\right)^{2+\delta}\Big| W=1\right] \leq C \end{aligned}$$

and

$$E[\|F(Y_{i-N}, W_{i-N}, S_{i-N})\|_2^{2+\delta}] = E[E[|\ F(Y_{i-N}, W_{i-N}, S_{i-N})\|_2^{2+\delta}|W_{i-N}, X_{i-N}]] \leq C$$

for some uniform constant C. Hence,

$$\begin{aligned} E[|\xi_i|^{2+\delta}] &\leq \begin{cases} \frac{\|\lambda\|_2^{2+\delta}}{n^{1+\delta/2}}(M+2)^{2+\delta}C, & i \leq N_1, \\ 2^{2+\delta}C, & N_1+1 \leq i \leq N_1+n \end{cases} \\ &\leq \frac{\|\lambda\|_2^{2+\delta}}{n^{1+\delta/2}}(M+2)^{2+\delta}C, \end{aligned}$$

from which we obtain Lyapounov's condition, namely that

$$\sum_{i=1}^{N_1+n} E[|\xi_i|^{2+\delta}] \leq \frac{N_1+n}{n} \frac{\|\lambda\|_2^{2+\delta}(M+2)^{2+\delta}C}{n^{\delta/2}} \to 0.$$

Hence, by the Lindeberg–Feller Martingale Central Limit Theorem,

$$\sqrt{n}(\widehat{\Phi}-\Phi)'\lambda = \sum_{i=1}^{N_1+n} \xi_i + o_P(1) \xrightarrow{d} \mathcal{N}(0, \lambda' V^* \lambda).$$

The assertion of the lemma follows now from the Cramér-Wold device. □

B.2 The Matched Bootstrap

In this section, we develop a general result for the coupled resampling of martingale increments that we then apply to the matched bootstrap.

Proposition B.1

Let $\lambda \geq 1$ be fixed. Assume we have a collated martingale difference array

$$\{\zeta_{n,1}^{(1)}, \ldots, \zeta_{n,n}^{(1)}, \zeta_{n,1}^{(2)}, \ldots, \zeta_{n,n}^{(2)}, \ldots, \zeta_{n,1}^{(\lambda)}, \ldots, \zeta_{n,n}^{(\lambda)}\}, n \geq 1,$$

with respect to the filtration array

$$\{\mathcal{F}_{n,1}^{(1)}, \ldots, \mathcal{F}_{n,n}^{(1)}, \mathcal{F}_{n,1}^{(2)}, \ldots, \mathcal{F}_{n,n}^{(2)}, \ldots, \mathcal{F}_{n,1}^{(\lambda)}, \ldots, \mathcal{F}_{n,n}^{(\lambda)}\}, n \geq 1,$$

and the following properties:

1. *For all $\ell \in \{1, \ldots, \lambda\}$,*

$$\sum_{i=1}^{n} E[(\zeta_{n,i}^{(\ell)})^2 | \mathcal{F}_{n,i-1}^{(\ell)}] \xrightarrow{p} \sigma_\ell^2,$$

   where $\mathcal{F}_{n,0}^{(\ell+1)} := \mathcal{F}_{n,n}^{(\ell)}$ for all $\ell \in \{1, \ldots, \lambda-1\}$.

2. *There exist some $C > 0$ and $\delta > 0$ such that for all i, n, ℓ,*

$$E[(\zeta_{n,i}^{(\ell)})^4] \leq \frac{C}{n^{1+\delta}}.$$

Consider the sum of increments

$$S_n := \sum_{\ell=1}^{\lambda} \sum_{i=1}^{n} \zeta_{n,i}^{(\ell)}$$

and the bootstrapped sum of coupled increments

$$T_n := \sum_{\ell=1}^{\lambda} \sum_{i=1}^{n} (\mathbf{w}_{n,i}^{(\ell)} - 1)\zeta_{n,i}^{(\ell)},$$

where $(\mathbf{w}_{n,1}^{(\lambda)}, \ldots, \mathbf{w}_{n,n}^{(\lambda)})$ *is multinomially distributed with parameters* $(n; n^{-1}, \ldots, n^{-1})$ *independent of the data, and*

$$\mathbf{w}_{n,\iota_n^{(\ell)}(i)}^{(\ell)} = \mathbf{w}_{n,i}^{(\lambda)}$$

for all $i \in \{1, \ldots, n\}, \ell \in \{1, \ldots, \lambda - 1\}$ *and* $\mathcal{F}_{n,n}^{(1)}$*-measurable bijections* $\iota_n^{(\ell)} : \{1, \ldots, n\} \to \{1, \ldots, n\}$.

Then, we have convergence of the sum,

$$S_n \xrightarrow{d} \mathcal{N}(0, \sigma^2), \tag{B.1}$$

where $\sigma^2 = \sum_{\ell=1}^{\lambda} \sigma_\ell^2$*, and conditional convergence of the bootstrapped sum,*

$$\sup_{x \in \mathbb{R}} \left| P\left(T_n \leq x \middle| \mathcal{F}_{n,n}^{(\lambda)}\right) - \Phi(x/\sigma) \right| \xrightarrow{p} 0, \tag{B.2}$$

as $n \to \infty$.

Note that from convergence of the bootstrapped sum conditional on the data (that is, (B.2)) follows unconditional convergence $T_n \xrightarrow{d} \mathcal{N}(0, \sigma^2)$.

Proof. Observe that

$$\sum_{\ell=1}^{\lambda} \sum_{i=1}^{n} E[(\zeta_{n,i}^{(\ell)})^2 | \mathcal{F}_{n,i-1}^{(\ell)}] \xrightarrow{p} \sum_{\ell} \sigma_\ell^2$$

as $n \to \infty$ by (1.) and Lyapounov's condition follows directly from (2.). Hence, (B.1) follows via the Martingale Central Limit Theorem.

For (B.2), our goal is to modify the proof of Theorem 2.1 in Pauly (2011) for the case of

coupled resampling. We do so by considering the coupled increments

$$Z_{n,i} := \sum_{\ell=1}^{\lambda} \zeta^{(\ell)}_{n,\iota_n^{(\ell)}(i)},$$

where $\iota_n^{(\lambda)}$ is the identity. For these increments,

$$S_n = \sum_{i=1}^{n} Z_{n,i}$$

and

$$T_n = \sum_{i=1}^{n} (\mathbf{w}_{n,i} - 1) Z_{n,i} = \sum_{i=1}^{n} \mathbf{w}_{n,i} (Z_{n,i} - \overline{Z}_n),$$

corresponding to weights $W_{n,i} = \mathbf{w}_{n,i}/\sqrt{n}$ that fulfil (2.3), (2.4) and (2.5) in Pauly (2011). Note, however, that $(Z_{n,i})_i$ is not a martingale difference array any more; hence, we cannot apply Theorem 2.1 directly, but instead invoke Theorem 4.1 in the appendix of Pauly (2011), which holds for more general triangular arrays of random variables.

(4.1) in Theorem 4.1 of Pauly (2011) follows from the boundedness condition (2.) by noting that

$$\max_{i \leq n, \ell \leq \lambda} |\zeta^{(\ell)}_{n,i}| \leq \sum_{\ell \leq \lambda} \max_{i \leq n} |\zeta^{(\ell)}_{n,i}|$$

and that

$$\max_{i \leq n} |\zeta^{(\ell)}_{n,i}| \xrightarrow{p} 0$$

is equivalent to the weak Lindeberg condition

$$\sum_{i=1}^{n} (\zeta^{(\ell)}_{n,i})^2 \mathbb{I}_{|\zeta^{(\ell)}_{n,i}| > \epsilon} \xrightarrow{p} 0 \; \forall \epsilon > 0,$$

which is implied by (2.) via Lyapounov's condition.

For (4.2), note that

$$\sum_{i=1}^{n}(Z_{n,i}-\overline{Z}_n)^2=\sum_{i=1}^{n}Z_{n,i}^2-\overline{Z}_n\sum_{i=1}^{n}Z_{n,i}=\sum_{i=1}^{n}\left(\sum_{\ell=1}^{\lambda}\zeta^{(\ell)}_{n,\iota_n^{(\ell)}(i)}\right)^2-\left(\sum_{i=1}^{n}Z_{n,i}\right)^2/n$$
$$=\sum_{\ell=1}^{\lambda}\sum_{i=1}^{n}\left(\zeta^{(\ell)}_{n,i}\right)^2+2\sum_{\overline{\ell}=2}^{\lambda}\sum_{\underline{\ell}=1}^{\overline{\ell}-1}\sum_{i=1}^{n}\zeta^{(\underline{\ell})}_{n,\iota^{\underline{\ell}}_n(i)}\zeta^{(\overline{\ell})}_{n,\iota^{\overline{\ell}}_n(i)}-\left(\frac{S_n}{\sqrt{n}}\right)^2.$$

Now,

$$A^{(\ell)}_{n,i}:=(\zeta^{(\ell)}_{n,i})^2-E[(\zeta^{(\ell)}_{n,i})^2|\mathcal{F}^{(\ell)}_{n,i-1}]$$

defines a martingale difference array with respect to the filtration array $\mathcal{F}^{(\ell)}_{n,i}$ for all $1\leq\ell\leq\lambda$, and

$$B^{\underline{\ell},\overline{\ell}}_{n,i}:=\zeta^{(\underline{\ell})}_{n,\iota^{\underline{\ell}}_n(i)}\zeta^{(\overline{\ell})}_{n,\iota^{\overline{\ell}}_n(i)},$$

defines a martingale difference array with respect to the filtration array $\mathcal{F}^{(\overline{\ell})}_{n,i}$ (where we have used $\mathcal{F}^{(1)}_{n,n}$-measurability of all ι^{ℓ}_n) for all $1\leq\underline{\ell}<\overline{\ell}\leq\lambda$, In both cases, the increments have second moments bounded by $\frac{C}{n^{1+\delta}}$ by (2.): Indeed,

$$E[(A^{(\ell)}_{n,i})^2]=E[(\zeta^{(\ell)}_{n,i})^4]-E[E[(\zeta^{(\ell)}_{n,i})^2|\mathcal{F}^{(\ell)}_{n,i-1}]^2]\leq E[(\zeta^{(\ell)}_{n,i})^4]\leq\frac{C}{n^{1+\delta}}$$

and

$$E[(B^{\underline{\ell},\overline{\ell}}_{n,i})^2]=E[(\zeta^{(\underline{\ell})}_{n,\iota^{\underline{\ell}}_n(i)}]^2(\zeta^{(\overline{\ell})}_{n,\iota^{\overline{\ell}}_n(i)})^2)\leq\sqrt{E[(\zeta^{(\underline{\ell})}_{n,\iota^{\underline{\ell}}_n(i)}]^4)}\sqrt{E[(\zeta^{(\overline{\ell})}_{n,\iota^{\overline{\ell}}_n(i)})^4]}\leq\frac{C}{n^{1+\delta}}$$

by the Cauchy-Schwarz inequality. Now, for any martingale difference array $(C_{i,n})_{i=1}^n$ with $EC_{i,n}^2\leq\frac{C}{n^{1+\delta}}$,

$$E\left(\sum_{i=1}^{n}C_{i,n}\right)^2=\sum_{i=1}^{n}EC_{i,n}^2\leq\frac{C}{n^{\delta}}\to 0$$

and hence

$$\sum_{i=1}^{n}C_{i,n}\xrightarrow{p}0$$

as $n \to \infty$. It follows that

$$\sum_{i=1}^{n}(Z_{n,i} - \overline{Z}_n)^2 = \sum_{\ell=1}^{\lambda} \underbrace{\sum_{i=1}^{n} E[(\zeta_{n,i}^{(\ell)})^2 | \mathcal{F}_{n,i-1}^{(\ell)}]}_{\xrightarrow{p} \sigma_\ell^2} + \sum_{\ell=1}^{\lambda} \underbrace{\sum_{i=1}^{n} A_{n,i}^{(\ell)}}_{\xrightarrow{p} 0} + 2 \sum_{\overline{\ell}=2}^{\lambda} \sum_{\underline{\ell}=1}^{\overline{\ell}-1} \underbrace{\sum_{i=1}^{n} B_{n,i}^{\underline{\ell},\overline{\ell}}}_{\xrightarrow{p} 0} - \underbrace{\left(\frac{S_n}{\sqrt{n}}\right)^2}_{\xrightarrow{p} 0}$$

$$\xrightarrow{p} \sum_{\ell=1}^{\lambda} \sigma_\ell^2 = \sigma^2,$$

where we have used (1.) and the unconditional convergence result (B.1).

Finally, note that the $Z_{n,i}$ are a sufficient statistic of $\mathcal{F}_{n,n}^{(\lambda)}$ for calculating T_n, incorporating sufficient information about both the $\zeta_{n,i}^{(\ell)}$ and ι_n^ℓ. Hence, (B.2) follows from Theorem 4.1 in the appendix of Pauly (2011). □

We now apply this result to our matching setting:

Proposition B.2

Under the setup and assumptions of Lemma B.2, and also

(a.5) $E[F_k^4(Y, W, S)|W = w, X = x]$ *uniformly bounded on* $\mathcal{X}$ *for all* k, $w \in \{0, 1\}$,

consider the bootstrapped sum

$$\widehat{\Phi}^* = \frac{1}{n} \sum_{W_i=1} \mathbf{w}_i \left(F(Y_i, W_i, S_i) + \sum_{j \in \mathcal{J}(i)} F(Y_j, W_j, S_j) \right),$$

where $\mathbf{w}$ *is multinomial with parameters* $(N_1; N_1^{-1}, \ldots, N_1^{-1})$ *independent of the data. Then,*

$$\sup_{r \in \mathbb{R}^s} \left| P_{\mathbf{w}} \left(\sqrt{n}(\widehat{\Phi}^* - \widehat{\Phi}) \leq r \middle| \mathcal{S} \right) - P(\mathcal{N}(\mathbf{0}, V^*) \leq r) \right| \xrightarrow{p} 0.$$

Proof. Fix $\lambda \in \mathbb{R}^s$. Similar to the proof of Lemma B.2, we decompose

$$\begin{aligned}
\sqrt{n}(\widehat{\Phi}^* - \widehat{\Phi})'\lambda &= \sqrt{n}((\widehat{\Phi}^* - \Phi) - (\widehat{\Phi} - \Phi))'\lambda \\
&= \frac{1}{\sqrt{n}} \Big(\sum_{W_i=1} (\mathbf{w}_i - 1)(\Phi_1(X_i) + M\Phi_0(X_i) - \Phi)'\lambda \\
&\qquad + \sum_{i \in \mathcal{S}^*} (\mathbf{w}_i - 1)(F(Y, W, S) - \Phi_{W_i}(X_i))'\lambda \Big) \\
&\quad + \frac{1}{\sqrt{n}} \sum_{W_i=1} (\mathbf{w}_i - 1) \sum_{j \in \mathcal{J}(i)} (\Phi_0(X_j) - \Phi_0(X_i))'\lambda.
\end{aligned}$$

The last part of the sum still vanishes in probability, as

$$\begin{aligned}
&E_{\mathbf{w}}\left(\left|\frac{1}{\sqrt{n}}\sum_{W_i=1}(\mathbf{w}_i-1)\sum_{j\in\mathcal{J}(i)}(\Phi_0(X_j)-\Phi_0(X_i))'\lambda\right|\middle|\mathcal{S}\right)\\
&\quad\leq\frac{1}{\sqrt{n}}\sum_{W_i=1}\sum_{j\in\mathcal{J}(i)}\underbrace{E_{\mathbf{w}}(|\mathbf{w}_i-1|)}_{\leq 2}\,|(\Phi_0(X_j)-\Phi_0(X_i))'\lambda|\\
&\quad\leq\frac{2}{\sqrt{n}}\sum_{W_i=1}\sum_{j\in\mathcal{J}(i)}|(\Phi_0(X_j)-\Phi_0(X_i))'\lambda|\\
&\quad\leq\frac{2L}{\sqrt{n}}\sum_{W_i=1}\sum_{j\in\mathcal{J}(i)}d(X_j,X_i)\xrightarrow{p}0
\end{aligned}$$

for an appropriate Lipschitz constant $L = L(\lambda)$, where we have used that

$$E_{\mathbf{w}}(|\mathbf{w}_i - 1|) \leq E_{\mathbf{w}}(\mathbf{w}_i + 1) = 2.$$

We can decompose the other parts into martingale increments as in the proof of Lemma B.2:

$$\sqrt{n}(\widehat{\Phi}^* - \widehat{\Phi})'\lambda = \sum_{i=1}^{N_1}(\mathbf{w}_i - 1)\xi_i + \sum_{i=N_1+1}^{(M+2)N_1}(\mathbf{w}_{i-N_1} - 1)\xi_i + o_P(1)$$

The result follows from Proposition B.1, which establishes a general result for the coupled resampling of martingale difference arrays. □

B.3 Proofs of Main Results

B.3.1 Asymptotic Behavior of Post-Matching OLS

Proof of Proposition 3.1. Let $E_{Q(\cdot|W=1)}$ and $E_{Q(\cdot|W=0)}$ be expectation operators for $Q(\cdot|W = 1)$ and $Q(\cdot|W = 0)$. Notice first that for any measurable function q,

$$E_{Q(\cdot|W=1)}[q(Y(1), S)] = E[q(Y, S)|W = 1] \tag{B.3}$$

The result holds also replacing $W = 1$ with $W = 0$, and after conditioning on X. In particular,

$$E_{Q(\cdot|W=0)}[q(Y(0), S)|X] = E[q(Y, S)|X, W = 0]. \tag{B.4}$$

The regression coefficient in the population defined by (a), (b) is the minimizer of

$$\frac{1}{M+1}E_{Q(\cdot|W=1)}[(Y(1)-g(1,S)'b)^2]+\frac{M}{M+1}E_{Q(\cdot|W=1)}[(Y(0)-g(0,S)'b)^2].$$

Notice that,

$$\begin{aligned}E_{Q(\cdot|W=1)}[(Y(1)-g(1,S)'b)^2] &= E[(Y-g(1,S)'b)^2|W=1]\\ &= E^*[(Y-Z'b)^2|W=1],\end{aligned}$$

where the first equality follows from Equation (B.3) and the second equality follows from the definitions of $P^*(\cdot|W=1)$ and Z. Similarly,

$$\begin{aligned}E_{Q(\cdot|W=1)}[(Y(0)-g(0,S)'b)^2] &= E_{Q(\cdot|W=1)}[E_{Q(\cdot|W=1)}[(Y(0)-g(0,S)'b)^2|X]]\\ &= E_{Q(\cdot|W=1)}[E_{Q(\cdot|W=0)}[(Y(0)-g(0,S)'b)^2|X]]\\ &= E[E[(Y-g(W,S)'b)^2|X,W=0]|W=1]\\ &= E^*[(Y-Z'b)^2|W=0].\end{aligned}$$

In the last equation, the first equality follows from the law of iterated expectations, the second equality follows from selection on observables, the third equality follows from (B.4) and (B.3), the last equation follows from the definition of $P^*(\cdot|W=0)$. Therefore, we obtain

$$\begin{aligned}&\frac{1}{M+1}E_{Q(\cdot|W=1)}[(Y(1)-g(1,S)'b)^2]+\frac{M}{M+1}E_{Q(\cdot|W=1)}[(Y(0)-g(0,S)'b)^2]\\ &\quad= \frac{1}{M+1}E^*[(Y-Z'b)^2|W=1]+\frac{M}{M+1}E^*[(Y-Z'b)^2|W=0]\\ &\quad= E^*[(Y-Z'b)^2],\end{aligned}$$

which implies the result of the proposition. □

Proof of Proposition 3.2. By Lemma B.1,

$$\frac{1}{n}\sum_{i\in\mathcal{S}^*}Z_iZ_i' \xrightarrow{p} H;$$

by Lemma B.2,

$$\widehat{H}\sqrt{n}\left(\widehat{\beta}-\beta\right)=\sqrt{n}\left(\frac{1}{n}\sum_{i\in\mathcal{S}^*}(Z_iY_i-Z_iZ_i'\beta)\right)\xrightarrow{d}\mathcal{N}(\mathbf{0},J),$$

where we note that

$$E[ZY-ZZ'\beta|W=0,X=x]$$

is Lipschitz. Hence,

$$\sqrt{n}\left(\widehat{\beta}-\beta\right)=\overbrace{\widehat{H}^{-1}}^{\xrightarrow{p}H^{-1}}\underbrace{\widehat{H}\sqrt{n}\left(\frac{1}{n}\sum_{i\in\mathcal{S}^*}(Z_iY_i-Z_iZ_i'\beta)\right)}_{\xrightarrow{d}\mathcal{N}(\mathbf{0},J)}\xrightarrow{d}\mathcal{N}(\mathbf{0},H^{-1}JH^{-1}).$$

□

B.3.2 Post-Matching Inference

Proof of Proposition 3.3. We have that

$$\begin{aligned}\widehat{J}_r&=\frac{1}{n}\sum_{i=1}^{n}Z_i(Y_i-Z_i'\widehat{\beta})^2Z_i'\\&=\frac{1}{n}\sum_{i=1}^{n}Z_i(Y_i-Z_i'\beta)^2Z_i'+\frac{1}{n}\sum_{i=1}^{n}Z_i\Big((Y_i-Z_i'\widehat{\beta})^2-(Y_i-Z_i'\beta)^2\Big)Z_i'.\end{aligned}$$

Notice that

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n}Z_i\Big((Y_i-Z_i'\widehat{\beta})^2-(Y_i-Z_i'\beta)^2\Big)Z_i'\\&\qquad=(\widehat{\beta}-\beta)'\left(\frac{1}{n}\sum_{i=1}^{n}Z_i(Z_i'Z_i)Z_i'(\widehat{\beta}+\beta)-2\frac{1}{n}\sum_{i=1}^{n}Z_i(Z_i'Z_i)Y_i\right).\end{aligned}$$

By assumption, the functions

$$E[\|Z\|^4|X=x,W=w]\quad\text{and}\quad E[|Y|^4|X=x,W=w]$$

are uniformly bounded on $\mathcal{X}_w$, for $w = 0, 1$. By Hölder's Inequality, this implies finiteness of

$$E\left[\left\|\frac{1}{n}\sum_{i=1}^{n} Z_i Z_i' Z_i Z_i'\right\|\right] \quad \text{and} \quad E\left[\left\|\frac{1}{n}\sum_{i=1}^{n} Z_i Z_i' Z_i Y_i'\right\|\right].$$

Then, for $\epsilon \in (0, 1/2)$, by Markov's Inequality, we obtain

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} Z_i((Y_i - Z_i'\widehat{\beta})^2 - (Y_i - Z_i'\beta)^2)Z_i' \\
&\quad = n^{1/2-\epsilon}(\widehat{\beta} - \beta)' \left(\frac{\sum_{i=1}^n Z_i(Z_i Z_i')Z_i'/n}{n^{1/2-\epsilon}}(\widehat{\beta} + \beta) - \frac{2\sum_{i=1}^n Z_i(Z_i Z_i')Y_i/n}{n^{1/2-\epsilon}} \right) \xrightarrow{p} 0.
\end{aligned}$$

As a result,

$$\widehat{J}_r = \frac{1}{n}\sum_{i=1}^{n} Z_i(Y_i - Z_i'\beta)^2 Z_i' + o_p(1),$$

and the claim follows from Lemma B.1. □

Proof of Corollary 3.4. Under correct specification, we find that

$$\begin{aligned}
\Gamma_W(X) &= E[Z(Y - Z'\beta)|W, X] = E[Z\epsilon|W, X] \\
&= E[E[Z\epsilon|Z, W, X]|W, X] \\
&= E[Z \underbrace{E[\epsilon|Z, W, X]}_{=0}|W, X] = 0.
\end{aligned}$$

□

Proof of Proposition 3.5. First, note that

$$\begin{aligned}
\widehat{J} =& \frac{1}{n}\sum_{W_i=1} \left(Z_i(Y_i - Z_i'\beta) + \textstyle\sum_{j\in\mathcal{J}(i)} Z_j(Y_j - Z_j'\beta) \right) \left(Z_i(Y_i - Z_i'\beta) + \textstyle\sum_{j\in\mathcal{J}(i)} Z_j(Y_j - Z_j'\beta) \right)' \\
&+ o_P(1),
\end{aligned}$$

where we replace $\widehat{\beta}$ by β analogous to the proof of Proposition 3.3.

Write

$$G := Z(Y - Z'\beta) \qquad \Gamma_w(x) := E[Z(Y - Z'\beta)|W = w, X = x].$$

Note that $\Gamma_0(x)$ is Lipschitz on $\mathcal{X}$, and that G_i has uniformly bounded fourth moments. We

decompose

$$
\begin{aligned}
\widehat{J} =& \frac{1}{n} \sum_{W_i=1} \left(G_i + \textstyle\sum_{j \in \mathcal{J}(i)} G_j\right) \left(G_i + \textstyle\sum_{j \in \mathcal{J}(i)} G_j\right)' + o_P(1) \\
=& \frac{1}{n} \sum_{W_i=1} \left(\Gamma_1(X_i) + M\Gamma_0(X_i)\right) \left(\Gamma_1(X_i) + M\Gamma_0(X_i)\right)' \\
&+ \frac{1}{n} \sum_{i \in \mathcal{S}^*} \left(G_i - \Gamma_{W_i}(X_i)\right) \left(G_i - \Gamma_{W_i}(X_i)\right)' \\
&+ \frac{1}{n} \sum_{W_i=1} \sum_{\ell \neq \ell' \in \mathcal{J}(i) \cup \{i\}} \left(G_\ell - \Gamma_{W_\ell}(X_\ell)\right) \left(G_{\ell'} - \Gamma_{W_{\ell'}}(X_{\ell'})\right)' \\
&+ \frac{1}{n} \sum_{W_i=1} \Big(\left(\Gamma_1(X_i) + M\Gamma_0(X_i)\right) \left(G_i - \Gamma_1(X_i) + \textstyle\sum_{j \in \mathcal{J}(i)} (G_j - \Gamma_0(X_j))\right)' \\
&\quad + \left(G_i - \Gamma_1(X_i) + \textstyle\sum_{j \in \mathcal{J}(i)} (G_j - \Gamma_0(X_i))\right) \left(\Gamma_1(X_i) + M\Gamma_0(X_j)\right)' \Big) + o_P(1).
\end{aligned}
$$

Here, the o_P terms absorb the deviation due to using $\widehat{\beta}$ instead of β, as well as the matching discrepancies in the conditional expectations.

The first sum is iid with

$$
\begin{aligned}
&\frac{1}{n} \sum_{W_i=1} \left(\Gamma_1(X_i) + M\Gamma_0(X_i)\right) \left(\Gamma_1(X_i) + M\Gamma_0(X_i)\right)' \\
&\xrightarrow{p} \frac{E\left[(\Gamma_1(X) + M\Gamma_0(X))(\Gamma_1(X) + M\Gamma_0(X))' | W = 1\right]}{1+M} \\
&= \frac{\mathrm{Var}(\overbrace{\Gamma_1(X) + M\Gamma_0(X)}^{E[\cdot|W=1]=\mathbf{0}} | W = 1)}{1+M},
\end{aligned}
$$

while the second is a martingale with

$$
\begin{aligned}
&\frac{1}{n} \sum_{i \in \mathcal{S}^*} \left(G_i - \Gamma_{W_i}(X_i)\right) \left(G_i - \Gamma_{W_i}(X_i)\right)' \\
&\xrightarrow{p} \frac{E[\mathrm{Var}(Z(Y - Z'\beta) | W = 1, X) + M \,\mathrm{Var}(Z(Y - Z'\beta) | W = 0, X) | W = 1]}{1+M}
\end{aligned}
$$

by Lemma B.1. Under appropriate reordering of the individual increments, all other sums can be represented as averages of mean-zero martingale increments; since the second moments of the increments are uniformly bounded, they vanish asymptotically. □

Proof of Proposition 3.6. Write

$$\widehat{H}_{\mathbf{w}} = \frac{1}{n} \sum_{i \in \mathcal{S}^*} \mathbf{w}_i Z_i Z_i'.$$

Note first that

$$\begin{aligned} H^{-1}\sqrt{n}(\widehat{H}_{\mathbf{w}}(\widehat{\beta}_{\mathbf{w}} - \beta) - \widehat{H}(\widehat{\beta} - \beta)) &= H^{-1}\sqrt{n}\left(\frac{1}{n}\sum_{i \in \mathcal{S}^*}(\mathbf{w}_i - 1) Z_i (Y_i - Z_i'\beta)\right) \\ &\xrightarrow{d} \mathcal{N}(\mathbf{0}, H^{-1} J H^{-1}), \end{aligned}$$

conditional on $\mathcal{S}$, by Proposition B.2. Now,

$$\begin{aligned} &\sqrt{n}(\widehat{\beta}_{\mathbf{w}} - \widehat{\beta}) \\ &= \widehat{H}_{\mathbf{w}}^{-1} H (H^{-1}\sqrt{n}(\widehat{H}_{\mathbf{w}}(\widehat{\beta}_{\mathbf{w}} - \beta) - \widehat{H}_{\mathbf{w}}(\widehat{\beta} - \beta)) \\ &= \underbrace{\widehat{H}_{\mathbf{w}}^{-1} H}_{\xrightarrow{p} \mathbb{I}} (H^{-1}\sqrt{n}(\widehat{H}_{\mathbf{w}}(\widehat{\beta}_{\mathbf{w}} - \beta) - \widehat{H}(\widehat{\beta} - \beta))) + \underbrace{(\widehat{H}_{\mathbf{w}}^{-1}\widehat{H} - \mathbb{I})}_{\xrightarrow{p} \mathbb{O}} \sqrt{n}(\widehat{\beta} - \beta) \\ &\xrightarrow{d} \mathcal{N}(\mathbf{0}, H^{-1} J H^{-1}), \end{aligned}$$

conditional on $\mathcal{S}$, where we have used that $\widehat{H}_{\mathbf{w}} - \widehat{H} \xrightarrow{p} \mathbb{O}$. □

B.4 Inference Conditional on Covariates

In the main part of this chapter, we analyze the variation of the post-matching estimator under resampling of units from a population distribution. In some applications, for example if the sample is the full population, inference conditional on the sample regressors may be more appropriate. In this section, we discuss standard errors conditional on the covariates X and treatment W; in particular, this implies that we condition on matches.

In a standard regression setting, Abadie *et al.* (2014) argue that Eicker–Huber–White standard error estimates (Eicker, 1967; Huber, 1967; White, 1980a,b, 1982), which are robust to misspecification for the unconditional standard errors, are not generally valid for conditional standard errors if the regression model is misspecified (in which case unconditional standard errors may overestimate the conditional variation). They propose an estimator of the conditional

variance that is based on nearest-neighbor matching, and show consistency for the variance conditional on the sample covariates even under misspecification.

Once we condition on the covariates X and treatment status W, the matching step is irrelevant for post-matching linear least-squares inference, and the analysis of Abadie *et al.* (2014) goes through. In particular, if the regression model is correctly specified, naive OLS standard error estimates are valid for the conditional variation (as they are for the unconditional variation by Corollary 3.4). If the regression model is not correctly specified, naive OLS standard errors are not generally valid for the conditional variation, but valid conditional standard error estimates that ignore the matching step – for example those proposed by Abadie *et al.* (2014) – are valid, as the matching step ceases to play a role.

Table B.1: *Simulation results for the main and alternative setups from 100,000 Monte Carlo iterations*

(a) *Target parameter: Coefficient τ_0 on treatment W*

		Post-matching		Average SE estimates		
Setup	Spec	$E[E[\widehat{\tau}_0\|(W_i, X_i)_{i\in\mathcal{S}^*}]]$	$E[\text{SE}(\widehat{\tau}_0\|(W_i, X_i)_{i\in\mathcal{S}^*})]$	Naive	Clust	Cond
Main	1	0.00	.115	.228	.110	.118
	2	0.00	.115	.109	.109	.113
Alt	1	1.53	.116	.425	.589	.135
	2	1.53	.116	.425	.589	.135

(b) *Target parameter: Coefficient τ_1 on the interaction WX of treatment W with covariate X*

		Post-matching		Average SE estimates		
Setup	Spec	$E[E[\widehat{\tau}_1\|(W_i, X_i)_{i\in\mathcal{S}^*}]]$	$E[\text{SE}(\widehat{\tau}_1\|(W_i, X_i)_{i\in\mathcal{S}^*})]$	Naive	Clust	Cond
Main	1	1.00	.200	.411	.191	.206
	2	1.00	.200	.189	.190	.197
Alt	1	1.12	0.20	0.76	1.06	0.24
	2	1.12	0.20	0.76	1.06	0.24

Table B.1 reports simulation results from the same two setups and two specifications discussed in Section 3.4. For the simulations, we construct conditional standard error estimates according to Abadie *et al.* (2014), using only the matched sample $\mathcal{S}^*$ and ignoring the matching step. Under correct specification (Specification 2 in the main setup), naive OLS standard errors are close to nominal on average, as predicted by the theory. Clustered standard errors,

which are valid for the *un*conditional variation, are not generally valid, as the alternative setup confirms; in the main setup, they produce appropriate estimates of the conditional standard errors because the variation in covariates does not contribute to the total variation in this specific case due to the matching. The conditional standard error estimates from Abadie *et al.* (2014) are close to nominal throughout.

In most applications, the covariates Z in the post-matching regression include only regressors from X and W. If Z contains additional regressors, the standard errors in Abadie *et al.* (2014) can be used to either estimate the variation of the the post-matching estimator conditional on X and W only, or conditional on the full set X, W, Z of covariates – as long as treatment W and covariates X are included, the matching step becomes irrelevant for conditional analysis.