ROBUST AND EFFICIENT ESTIMATION OF TREATMENT EFFECTS IN EXPERIMENTAL AND NON-EXPERIMENTAL SETTINGS

By

Akanksha Negi

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Economics – Doctor of Philosophy

2020

ABSTRACT

ROBUST AND EFFICIENT ESTIMATION OF TREATMENT EFFECTS IN EXPERIMENTAL AND NON-EXPERIMENTAL SETTINGS

By

Akanksha Negi

Broadly, this dissertation identifies and addresses issues that arise in experimental and observational data contexts for estimating causal effects. In particular, the three chapters in this dissertation focus on issues of consistent and efficient estimation of causal effects using methods that are robust to misspecification of a conditional model of interest.

Chapter 1: Revisiting regression adjustment in experiments with heterogeneous treatment effects

Regression adjustment with covariates in experiments is intended to improve precision over a simple difference in means between the treated and control outcomes. The efficiency argument in favor of regression adjustment has come under criticism lately, where papers like Freedman (2008a,b) find no systematic gain in asymptotic efficiency of the covariate adjusted estimator. This chapter shows that, like in Lin (2013), estimating separate regressions for the control and treated groups is guaranteed to do no worse than both the simple difference-in-means estimator and just including the covariates in additive fashion. This result appears to be new, and simulations show that the efficiency gains can be substantial. This chapter also talks about some important cases – applicable to binary, fractional, count, and other nonnegative responses – where nonlinear regression adjustment is consistent without any restrictions on the conditional mean functions.

Chapter 2: Robust and efficient estimation of potential outcome means under random assignment

This chapter studies improvements in efficiency for estimating the entire vector of potential outcome means using linear regression adjustment with two or more assignment levels. This chapter shows that using separate regression adjustments for each assignment level is never worse asymptotically than using the subsample averages and that separate regression adjustment generally improves over pooled regression adjustment, except in the obvious case where slope parameters in the linear projections are identical across the different assignment levels. An especially promising direction is to use certain nonlinear regression adjustment methods, which we show to be robust to misspecification in the conditional means. We apply this general potential outcomes framework to a contingent valuation study which seeks to estimate the lower bound mean willingness to pay (WTP) for an oil spill prevention program in California.

Chapter 3: Doubly weighted M-estimation for nonrandom assignment and missing outcomes

This chapter studies the problems of nonrandom assignment and missing outcomes, which together, undermine the validity of standard causal inference. While the econometrics literature has used weighting to address each issue in isolation, empirical analysis is often complicated by the presence of both. This chapter proposes a new class of inverse probability weighted M-estimators that deal with the two issues by combining propensity score weighting with weighting for missing data. This chapter also discusses applications of the proposed method for robust estimation of the two prominent causal parameters, namely, the average treatment effect and quantile treatment effects, under misspecification the framework's parametric components. This chapter also demonstrates the proposed estimator's viability in empirical settings by applying it to the sample of Aid to Families with Dependent Children (AFDC) women from the National Supported Work program compiled by Calónico and Smith (2017).

ACKNOWLEDGEMENTS

I would like to extend a sincere thanks to my advisor, Jeffrey M. Wooldridge, for his unrelenting support, guidance, and insights that has been invaluable for the development of this project and which have helped me grow into an independent researcher. I would also like to thank committee members, Steven Haider, Ben Zou, and Kenneth Frank for discussions from which this work has benefited immensely. A special thanks to Timothy Vogelsang, Jon X. Eguia, Wendun Wang, Mike Conlin, Todd Elder, Alyssa Carlson, for their helpful comments, advice, and encouragement in the years leading up to the job market.

This work has evolved significantly thanks to comments and suggestions received from participants at the Econometrics seminar series, Econometrics reading group meetings, and graduate student seminar series at Michigan State University. I am also grateful for the discussions at Midwest Econometric Group Meetings, International Association for Applied Econometrics, and International Econometrics PhD Conference in Rotterdam.

I would like to acknowledge financial support from the Department of Economics, Graduate School, College of Social Sciences, and Council of Graduate Students, along with the numerous fellowships; Dissertation Completion Fellowship, Kelly Research Fellowship and Supplemental Research Fellowship received through the duration of these PhD years that have been instrumental in successful completion of this work. The administrative help and support received from Lori Jean Nichols and Jay Feight has made navigating this PhD program easier than it would otherwise have been.

My warmest gratitude to family and friends who continue to inspire and motivate me to take up new adventures both personally and professionally. Finally, a note of thanks to my dear friend, and colleague, Christian Cox, with whom I have shared the vicissitudes of PhD life and who has supported me throughout this journey.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

Many questions in health, labor, development, and other areas of applied microeconomics are questions of causal inference. Establishing causality not only helps to quantify cause-and-effect relationships but also helps to formulate counterfactual scenarios which can be useful for informing policy and contributing to policy debates. Randomized experiments are generally considered to be a reasoned basis for conducting causal inference since the experimental design creates groups that are comparable on average and do not differ systematically in measured and unmeasured dimensions. On the other hand, observational or non-experimental data make it difficult to isolate causal mechanisms due to the absence of random assignment which introduces overt and hidden biases between treatment-control comparisons. This dissertation studies issues of consistency and efficiency in estimating causal effects under experimental and non-experimental data settings. Given this objective, the focus of this dissertation is on methods that do not rely on correct functional form assumptions of some conditional feature of interest to achieve consistent or efficient estimation of treatment effects.

Chapter 1 of this dissertation revisits the argument of regression adjustment which is routinely employed for improving precision on the estimated average treatment effect in experiments over a simple difference in means. As has been noted in the statistics literature by Freedman (2008b), Freedman (2008b), and Lin (2013), such an argument may be misguided. This is because additively controlling for covariates in a regression is not guaranteed to produce more precise estimates than a simple difference in means estimate. A similar result, however, is formally lacking in the economics literature where random sampling from a infinite population is the mainstream asymptotic paradigm. Chapter 2 builds on this idea of regression (or covariate) adjustment in experiments and extends it to the case of more than two treatment levels. This is a more general setting which is able to encompass a variety of applications such as experiments with multiple treatment levels, difference-in-difference de-

signs, and willingness to pay studies. Chapter 3 relaxes the experimental setting to consider treatment effects estimation with observational data when some of the observed outcomes are missing. This is a widely encountered problem in most micro-econometric empirical analyses. In this setting, consistent estimation of treatment effects is complicated unless the researcher imposes structure on the treatment assignment and missing data mechanisms.

Chapter 1 in this dissertation acknowledges the criticism leveled against simple regression adjustment for improving precision on the estimated average treatment effect. This chapter, then, proposes a regression estimator, which allows for separate estimation of the slopes in the treatment and control groups, that is guaranteed to be more precise than the simple difference in means and pooled regression adjustment estimators. This result is easily observed in the simulations where we consider different data generating processes and a range of assignment probabilities. Chapter 2 extends this efficiency result of the separate slopes regression estimator to the case of G arbitrary treatment levels and provides an empirical illustration and simulation results which provide favorable evidence for the separate slopes estimator. Finally, the third chapter proposes an inverse probability weighted estimator that double weights for the problems of nonrandom assignment and missing outcomes.

The methodology used in each of the chapters does not rely on correct conditional feature assumptions to propose efficient or consistent estimators for the causal parameters in question. In the first two chapters, efficiency gains with the separate slopes estimator are not based on correct conditional mean specification, but on linear projections which are consistently estimated under mild assumptions. Similarly, the double weighted IPW estimator proposed in the third chapter is consistent even when the outcome model is misspecified as long as the weights are correct. This dissertation aims to provide empiricists with a wide-range of treatment effect settings where the methods studied here may prove useful for conducting sound and robust causal analysis.

CHAPTER 1

REVISITING REGRESSION ADJUSTMENT IN EXPERIMENTS WITH HETEROGENEOUS TREATMENT EFFECTS[†]

1.1 Introduction

The role of covariates in randomized experiments has been studied since the early 1930s [Fisher (1935)][1]. When compared with the simple difference-in-means (SDM) estimator, the main benefit of adjusting for covariates is that the precision of the estimated average treatment effect (ATE) can be improved if the covariates are sufficiently predictive of the outcome [Cochran (1957); Lin (2013)]. Nevertheless, regression adjustment is not uniformly accepted as being preferred over the SDM estimator. For example, Freedman (2008b,a) argues against using regression adjustment because it is not guaranteed to be unbiased unless one makes the strong assumption that the conditional expectation functions are correctly specified (and linear in parameters).

It is important to understand that there are two different, potentially valid criticisms of regression adjustment (RA). The first is the issue of bias just mentioned: unlike the SDM estimator, which is unbiased conditional on having some units in both the control and treatment groups, RA estimators are only guaranteed to be consistent, not unbiased. Therefore, in experiments with small sample sizes, one might be willing to forego potential efficiency gains in order to ensure an unbiased estimator of the average treatment effect. As Bruhn and McKenzie (2009) points out, samples of 100 to 500 individuals or 20 to 100 schools or health clinics is fairly common in experiments conducted in development economics. In situations where unbiased estimation is the highest priority, the current paper has little

[†]This work is joint with Jeffrey M. Wooldridge and is unpublished.

[1]In the statistics and biomedical literature, such variables are also known as prognostic factors or concomitant variables. In the econometrics of program evaluation literature they are known as pre-treatment covariates. As the name suggests, they are ideally measured before the treatment is administered.

to add, other than to provide simulation evidence that using our preferred RA procedure often results in small bias. Henceforth, we are not interested in small-sample problems as a criticism of RA. More and more economic experiments, especially those conducted online, include enough units to make consistency and asymptotic efficiency a relevant criteria for evaluating estimators of ATEs. In cases where effect sizes are small (but important in the aggregate), improving precision can be important even when sample sizes seem fairly large.

A second criticism of RA methods, and the one most relevant for this paper, is that RA methods may not improve over the SDM estimator even if we focus on asymptotic efficiency. Freedman (2008b,a) and Lin (2013) level this criticism of RA when the covariates are simply added as controls along with a treatment indicator in a linear regression analysis. Freedman (2008b), for example, finds no systematic efficiency gain from using covariates. Lin (2013) provides an in-depth discussion about how simply adding covariates will not necessarily produce efficiency gains when treatment effects are heterogeneous. Both Freedman and Lin operate under a finite population paradigm where all population units are observed in the sample. Therefore, uncertainty in the estimators is due to the assignment into treatment and control, and not due to sampling from a population [Abadie et al. (2017b), Abadie et al. (2017a) and Rosenbaum (2002) discuss similar settings].

Random sampling from a population is still an important setting for empirical work, and the findings in Freedman (2008b) and Lin (2013) do not extend to the random sampling scenario. This involves accounting for both types of uncertainties, sampling-based and design-based. Our paper will not have more to say about the differences between these two types of uncertainties. For a deeper discussion of sampling-based and design-based uncertainty, see Abadie et al. (2017b). Imbens and Rubin (2015) study linear regression adjustment in the same sampling setting that we use here: independent and identically distributed (i.i.d.) draws from a population. However, they state efficiency results only in the case that the population means of the covariates are known, even though only a random sample is available. In addition, Imbens and Rubin only consider *linear* regression adjustment.

Regression adjustment in experiments has also been studied in the statistics literature by papers like Yang and Tsiatis (2001), Tsiatis et al. (2008), Ansel et al. (2018), and Berk et al. (2013) for the case of linear adjustment and by Rosenblum and Van Der Laan (2010) and Bartlett (2018) for the case of nonlinear regression adjustment. While some of the results derived in these papers overlap with what we show, the expressions and exposition is not as transparent and simple as ours. For nonlinear regression adjustment, we establish consistency by distinguishing between pooled and separate regression adjustment, which is missing from the discussion in Rosenblum and Van Der Laan (2010) and Bartlett (2018).

In this paper, we revisit regression adjustment and resolve some outstanding issues in the literature. We cover the standard case of i.i.d. draws from an underlying population, so that randomness comes from sampling error as well as assignment into control and treatment groups. Further, unlike Imbens and Rubin (2015), we consider the realistic case where the population means of the covariates must be estimated using the sample averages.

In the case of linear regression adjustment, we study four estimators: the SDM estimator, the pooled regression adjusted (PRA) estimator, the full regression adjusted (FRA) estimator – which uses separate regressions for the control and treatment groups – and what we call the infeasible regression adjusted (IRA) estimator, which is like the FRA estimator but assumes the population means of the covariates are known. We include IRA for completeness, as it is studied in Imbens and Rubin (2015), and doing so allows us to characterize the lost efficiency due to having to estimate the population means.

Our most important results in the linear regression case can be easily summarized. First, even when accounting for the sampling error of the covariates in estimating the population means, using separate linear regressions for the control and treatment groups leads to an ATE estimator that is never less precise (asymptotically) than the SDM estimator and the PRA estimator. Unless small sample bias is a concern, there is no reason not to use full regression adjustment. Further, there are two interesting cases when there will be no precision gain when using full RA compared with pooled RA. The first is when there is no heterogeneity

in the slopes of the linear projections of the potential outcomes (although there could be in the unobservables). In this case, it is not surprising that using pooled RA is sufficient to capture the efficiency gains of using covariates. The second important case where there is no additional gain from FRA is when the design is balanced: the probability of being in the treatment group is equal to 0.5. Therefore, if one has imposed a balanced design and is considering only linear regression adjustment, the pooled method is probably preferred (due to conserving degrees of freedom). A final result, which is pretty obvious, is that there is no efficiency gain when the covariates are not predictive of the potential outcomes; then, SDM is asymptotically efficient. We want to emphasize that there is no (asymptotic) cost in doing the regression adjustment, whether PRA or FRA: each estimator has the same asymptotic variance. In situations where one has good predictors of the outcome, regression adjustment can be attractive. Our simulation study illustrates the special cases derived from our theortical results.

Another important contribution of our paper is to characterize situations where nonlinear regression adjustment preserves consistency of average treatment effect estimators without imposing additional assumptions. In particular, we show that when the response is binary, fractional, count, or other nonnegative outcomes, certain kinds of nonlinear regression adjustment consistently estimates the average treatment effect. Our simulations for the case of binary and non-negative response suggest that nonlinear RA, especially the full version, can produce sampling variances that are smaller than SDM and also linear regression adjustment. In terms of bias, nonlinear FRA (NFRA) is comparable to SDM, which we know is unbiased.

The rest of the paper is organized as follows. Section 3.2 briefly introduces the potential outcomes setting and defines the population average treatment effect – the parameter of interest in this paper. Section 3.3 discusses the random assignment mechanism and the random sampling assumption. Section 3.4 is important and describes linear regression adjustment in the population in terms of linear projections, which are consistently estimated

by ordinary least squares (OLS) given a random sample. Importantly, we need not impose any assumptions on the conditional mean functions of the potential outcomes. Section 3.5 presents the asymptotic variances of the four linear estimators and ranks them on the basis of asymptotic efficiency. We also characterize the cases under which estimating two separate regressions does not improve efficiency over SDM or PRA (or both). Section 3.6 presents Monte Carlo simulations that compare the bias and root mean squared error (RMSE) of the estimators for eight different data generating processes. In section 3.7 we draw on results from the doubly robust estimation literature and characterize the nonlinear RA estimators – both pooled and full – that produce consistent estimators of the ATE. Our simulations in this section show that nonlinear methods have modest bias while considerably improving efficiency compared with both SDM and linear RA methods. Section 3.8 concludes the paper with a discussion of some future research topics. All proofs are included in the appendix.

1.2 Potential outcomes and parameter of interest

Our framework is the standard Neyman-Rubin causal model, involving potential (or counterfactual) outcomes. Let $\{Y(0), Y(1)\}$ be the two potential outcomes corresponding to the control and treatment states, respectively, where $\{Y(0), Y(1)\}$ has a joint distribution in the population. The setup is nonparametric in that we make no assumptions about the distribution of $\{Y(0), Y(1)\}$ other than finite moment conditions needed to apply standard asymptotic theory. In particular, $\{Y(0), Y(1)\}$ may be discrete, continuous, or mixed random variables. For example, $Y(0)$ and $Y(1)$ can be binary employment indicators for nonparticipation and participation in a job training program. Or, they could be the fraction of assets held in the stock market, or counts of the number of hospital visits taken by a patient.

Define the means of the potential outcomes as

$$\mu_0 = \mathbb{E}\left[Y(0)\right]$$

$$\mu_1 = \mathbb{E}\left[Y(1)\right]$$

The parameter of interest is the population average treatment effect (PATE),

$$\tau = \mathbb{E}\left[Y(1) - Y(0)\right] = \mu_1 - \mu_0$$

As has been often noted in the literature, the problem of causal inference is essentially a missing data problem. We only observe one of the the outcomes, $Y(0)$ or $Y(1)$, once the treatment, represented by the Bernoulli random variable W, is determined. Specifically, the observed Y is defined by

$$Y = \begin{cases} Y(0), \text{ if } W = 0 \\ Y(1), \text{ if } W = 1 \end{cases} \tag{1.1}$$

It is also useful to write Y as

$$Y = (1 - W) \cdot Y(0) + W \cdot Y(1). \tag{1.2}$$

1.3 Random assignment and random sampling

In determining an appropriate method to estimate τ, we need to know how the treatment, W, is assigned. In this paper, we assume that W is independent of the potential outcomes as well as observed covariates, which we write as $\mathbf{X} = (X_1, X_2, \ldots, X_K)$. Formally, the random assignment assumption is as follows.

Assumption 1.3.1. *The binary assignment indicator, W, is a Bernoulli trial and is independent of $\{Y(0), Y(1), \mathbf{X}\}$, where $\mathbf{X} = (X_1, X_2, \ldots, X_K)$. Mathematically,*

$$W \perp \{Y(0), Y(1), \mathbf{X}\}.$$

Letting $\rho = \mathbb{P}(W = 1)$ be the probability of being assigned into treatment, assume that $0 < \rho < 1$.

The assumption of random assignment implies that there are many consistent estimators of τ. The goal in this paper is to rank, as much as possible, commonly used estimators of τ in terms of asymptotic efficiency.

As mentioned in the introduction, both early [Neyman (1923) and Fisher (1935)] and recent [Freedman (2008b) Freedman (2008a) and Lin (2013)] approaches to estimating ATEs assume that the entire population is the sample. Therefore, the only stochastic element of the setup is the assignment of the treatment, which is randomized. Such a perspective rules out any uncertainty stemming from unobservability of the entire population (also termed sampling-based uncertainty) and only allows uncertainty that arises due to the experimental design (also known as design-based uncertainty). Here, we adopt the assumption commonly used in studying various estimators in statistics and econometrics.

Assumption 1.3.2. *For a nonrandom integer N, $\{\big(Y_i(0), Y_i(1), W_i, \mathbf{X}_i\big) : i = 1, 2, \ldots, N\}$ are independent and identically distributed draws from the population.*

Given the random sampling assumption, standard asymptotic theory for i.i.d. sequences of random vectors can be applied, where N tends to infinity. We assume in what follows that at least second moments of the potential outcomes and covariates are finite so that, when we use regression methods, we can apply the law of large numbers and central limit theorem. We do not state these moment assumptions explicitly as they cannot be checked, anyway.

For each unit i drawn from the population, the treatment effect is

$$Y_i(1) - Y_i(0),$$

which we can write as

$$Y_i(1) - Y_i(0) = \tau + \big[V_i(1) - V_i(0)\big]$$

where $V_i\left(w\right) = Y_i(w) - \mu_w$ for $w \in \{0, 1\}$. The treatment effects are homogeneous when the unit-specific components, $V_i(1) - V_i(0)$, are identically zero for any random draw i.

1.4 Estimators

We now carefully describe the estimators that we use in the linear regression context.

1.4.1 Simple difference in means (SDM)

Random assignment provides the luxury of using an estimator available from basic statistics. This estimator dates back to Neyman (1923) in the context of causal inference using potential outcomes. Let W_i be the treatment indicator for unit i. Then $N_0 = \sum_{i=1}^{N}(1 - W_i)$ and $N_1 = \sum_{i=1}^{N} W_i$ are the number of control and treated units in the sample, respectively. These are random variables. When N_0, $N_1 > 0$ we can define the sample averages for the control and treated units:

$$\bar{Y}_0 = N_0^{-1} \sum_{i=1}^{N} (1 - W_i) Y_i \tag{1.3}$$

$$\bar{Y}_1 = N_1^{-1} \sum_{i=1}^{N} W_i Y_i, \tag{1.4}$$

where Y_i is the observed outcome for unit i. The simple difference-in-means estimator is

$$\hat{\tau}_{SDM} = \bar{Y}_1 - \bar{Y}_0. \tag{1.5}$$

Under random assignment and conditional on N_0, $N_1 > 0$, $\hat{\tau}_{SDM}$ is unbiased for τ – see, for example, Imbens and Rubin (2015). Further, $\hat{\tau}_{SDM}$ is consistent as $N \to \infty$ for τ when $0 < \rho < 1$, as we assume. As is well know, $\hat{\tau}_{SDM}$ can be obtained as the coefficient on W_i in the simple regression

$$Y_i \text{ on } 1, W_i, \; i = 1, \ldots, N. \tag{1.6}$$

See, for example, Imbens and Rubin (2015).

1.4.2 Pooled regression adjustment (PRA)

When we have covariates that (hopefully) predict the outcome Y, the simplest way to use those covariates is to add them to the simple regression in (1.6). As documented in Słoczyński

(2018), adding covariates along with a binary treatment indicator is still very common in estimating treatment effects, whether one has randomized assignment or assumes unconfoundedness conditional on the covariates. Specifically, the regression is

$$Y_i \text{ on } 1, W_i, \mathbf{X}_i, \, i = 1, 2, \ldots, N.$$

The coefficient on W_i is the estimator of τ, and we called this estimator "pooled regression adjustment" (PRA) and denote it $\hat{\tau}_{PRA}$. The name "pooled" emphasizes that we are pooling across the control and treatment groups in imposing common coefficients on the vector of covariates, $\mathbf{X}_i$. In other words, the slopes are the same for $W = 0$ and $W = 1$. It is important to understand that we are making no assumption about whether the coefficients in an underlying linear model in the population are the same, or even whether there is an underlying linear model representing a conditional expectation. This will become clear in the next subsection when we formally describe linear projections.

As is well known, adding the variables $\mathbf{X}_i$ to the simple regression does not change the probability limit provided W_i and $\mathbf{X}_i$ are uncorrelated, which follows under random assignment. However, it is not always the case that adding $\mathbf{X}_i$ is a good idea, even if we focus on asymptotic efficiency: it may or may not improve asymptotic efficiency compared with the SDM estimator.

1.4.3 Linear projections and infeasible regression adjustment (IRA)

The SDM estimator is an example of an estimator that can be written as

$$\hat{\tau} = \hat{\mu}_1 - \hat{\mu}_0,$$

where, in the case of SDM, $\hat{\mu}_0$ and $\hat{\mu}_1$ are the sample averages of the control and treated groups, respectively. But there are other ways to consistently estimate μ_0 and μ_1 when we have covariates $\mathbf{X}$, represented as a $1 \times K$ vector. In particular, define the linear projections

of the potential outcomes on the vector of covariates as

$$\mathbb{L}\left[Y(0)|1,\mathbf{X}\right] = \alpha_0 + \mathbf{X}\boldsymbol{\beta}_0 \tag{1.7}$$

$$\mathbb{L}\left[Y(1)|1,\mathbf{X}\right] = \alpha_1 + \mathbf{X}\boldsymbol{\beta}_1, \tag{1.8}$$

where the expressions for α_0, α_1, $\boldsymbol{\beta}_0$, and $\boldsymbol{\beta}_1$ can be found in Wooldridge (2010) Section 2.3. As discussed in Wooldridge, the linear projections always exist and are well defined provided $Y(w)$ and the elements of $\mathbf{X}$ have finite second moments. Any of the random variables can be discrete, continuous, or mixed. The elements of $\mathbf{X}$ can include the usual functional forms, such as logarithms, squares, and interactions. The requirement for the coefficients in the linear projections (LPs) to be unique is simply that the variance-covariance matrix of $\mathbf{X}$, $\boldsymbol{\Omega}_{\mathbf{X}}$, is nonsingular, an assumption that rules out perfect collinearity in the population.

It is often helpful to slightly rewrite the LPs. Define the $1 \times K$ vector of population means of $\mathbf{X}$ as $\boldsymbol{\mu}_{\mathbf{X}} = \mathbb{E}(\mathbf{X})$, and let $\dot{\mathbf{X}} = \mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}}$ be the deviations from the population mean. Then we can write the linear projection in terms of the means μ_0 and μ_1 as

$$\mathbb{L}\left[Y(0)|1,\mathbf{X}\right] = \mu_0 + \dot{\mathbf{X}}\boldsymbol{\beta}_0 \tag{1.9}$$

$$\mathbb{L}\left[Y(1)|1,\mathbf{X}\right] = \mu_1 + \dot{\mathbf{X}}\boldsymbol{\beta}_1 \tag{1.10}$$

The two representations make it clear that the PATE, τ, can be expressed as

$$\tau = \mu_1 - \mu_0 = (\alpha_1 - \alpha_0) + \boldsymbol{\mu}_{\mathbf{X}}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0)$$

Therefore, if we have consistent estimators of α_0, α_1, $\boldsymbol{\beta}_0$, $\boldsymbol{\beta}_1$, and $\boldsymbol{\mu}_{\mathbf{X}}$, then we can consistently estimate τ. Importantly, as discussed in Wooldridge (2010) Chapter 4, ordinary least squares estimation using a random sample always consistently estimates the parameters in a population linear projection (subject to the mild finite second moment assumptions and the non-singularity of $\boldsymbol{\Omega}_{\mathbf{X}}$). This is true regardless of the nature of $Y(w)$ or $\mathbf{X}$. This insight is critical for understanding why regression adjustment produces consistent estimators of τ, and for the asymptotic efficiency arguments later on. Unlike in Imbens and Wooldridge

(2009), we do not assume that the linear projection is the same as the conditional mean. We are silent on the conditional mean functions $\mathbb{E}\left[Y(0)|\mathbf{X}\right]$ and $\mathbb{E}\left[Y(1)|\mathbf{X}\right]$.

Given random assignment, consistent estimators of the LP coefficients are obtained from the separate regressions

$$Y_i \text{ on } 1, \mathbf{X}_i \text{ using } W_i = 0 \tag{1.11}$$

$$Y_i \text{ on } 1, \mathbf{X}_i \text{ using } W_i = 1 \tag{1.12}$$

Wooldridge (2010) Chapter 19 formally shows that the linear projections are consistently estimated under the assumption that selection – in this case, W_i – is independent of $\left[\mathbf{X}_i, Y_i(w)\right]$. This is sometimes called the "missing completely at random" (MCAR) assumption in the missing data literature [for example, Little and Rubin (2002)].

If we assume that the vector of population means $\boldsymbol{\mu}_{\mathbf{X}}$ is known, a consistent estimator of τ is

$$\hat{\tau}_{IRA} = (\hat{\alpha}_1 - \hat{\alpha}_0) + \boldsymbol{\mu}_{\mathbf{X}}(\hat{\boldsymbol{\beta}}_1 - \hat{\boldsymbol{\beta}}_0),$$

where $\hat{\alpha}_0$, $\hat{\alpha}_1$, $\hat{\boldsymbol{\beta}}_0$, and $\hat{\boldsymbol{\beta}}_1$ are the OLS estimators from the separate regressions. We call this the "infeasible regression adjustment" (IRA) estimator because it depends on $\boldsymbol{\mu}_{\mathbf{X}}$, which is likely to be unknown in our context with random sampling.

From the algebra of ordinary least squares (OLS) it is easily shown that $\hat{\tau}_{IRA}$ can be obtained as the coefficient on W_i in the regression that includes a full set of interactions between W_i and $\dot{\mathbf{X}}_i$, namely,

$$Y_i \text{ on } 1,\ W_i,\ \mathbf{X}_i,\ W_i \cdot \dot{\mathbf{X}}_i,\ i = 1, \ldots, N.$$

The demeaning of the covariates ensures that the coefficient on W_i is $\hat{\tau}_{IRA}$. This regression is also convenient for obtaining a valid standard error for $\hat{\tau}_{IRA}$, as the usual Eicker-Huber-White heteroskedasticity-robust standard error is asymptotically valid.

In the case where the linear projections are also the conditional expectations – that is, $\mathbb{E}\left[Y(w)|\mathbf{X}\right] = \mathbb{L}\left[Y(w)|1, \mathbf{X}\right]$, $w \in \{0, 1\}$ – $\hat{\alpha}_0$, $\hat{\alpha}_1$, $\hat{\boldsymbol{\beta}}_0$, and $\hat{\boldsymbol{\beta}}_1$ are unbiased conditional on

$\{\mathbf{X}_i : i = 1, 2, \ldots, N\}$, provided we rule out perfect collinearity in the control and treated subsamples. Then, $\hat{\tau}_{IRA}$ would also be unbiased conditional on $\{\mathbf{X}_i : i = 1, 2, \ldots, N\}$, and unbiased unconditionally if its expectation exists. But linearity of the conditional expectations is much too strong an assumption, and it is clearly not needed for unbiasedness or consistency of the SDM estimator. Therefore, in what follows, we make no assumptions about $\mathbb{E}\left[Y(w)|\mathbf{X}\right]$. We simply assume enough moments are finite and rule out perfect collinearity in $\mathbf{X}$ in order for the linear projections to exist.

1.4.4 Full regression adjustment (FRA)

We can easily make the IRA estimator feasible by replacing $\boldsymbol{\mu}_{\mathbf{X}}$ with the sample average, $\bar{\mathbf{X}} = N^{-1}\sum_{i=1}^{N}\mathbf{X}_i$. This leads to what we will call the "full regression adjustment" (FRA) estimator:

$$\hat{\tau}_{FRA} = (\hat{\alpha}_1 - \hat{\alpha}_0) + \bar{\mathbf{X}}(\hat{\boldsymbol{\beta}}_1 - \hat{\boldsymbol{\beta}}_0).$$

This estimator can also be obtained as the OLS coefficient on W_i but from the regression

$$Y_i \text{ on } 1,\ W_i,\ \mathbf{X}_i,\ W_i \cdot \ddot{\mathbf{X}}_i,\ i = 1, 2, \ldots, N,$$

where

$$\ddot{\mathbf{X}}_i = \mathbf{X}_i - \bar{\mathbf{X}},\ i = 1, 2, \ldots, N$$

are the demeaned covariates using the sample average. This estimator is always available given a sample $\left\{(Y_i, W_i, \mathbf{X}_i) : i = 1, 2, \ldots, N\right\}$. Generally, $\hat{\tau}_{FRA} \neq \hat{\tau}_{IRA}$. Like $\hat{\tau}_{IRA}$, we can only conclude that $\hat{\tau}_{FRA}$ is consistent, although it will be unbiased under essentially the same assumptions discussed for $\hat{\tau}_{IRA}$. In the next section, we will rank the four estimators, to the extent possible, in terms of asymptotic efficiency.

1.5 Asymptotic variances and efficiency comparisons

We first derive the asymptotic variances of the SDM, PRA, IRA and FRA estimators in the general case of heterogeneous treatment effects. Naturally, the formulas include homo-

geneous treatment effects as a special case. We then compare the asymptotic variances in general and in special cases.

In order to obtain the asymptotic variances, we need to study the linear projections of the potential outcomes on the covariates more closely. Recall that we can write the potential outcomes as

$$Y(0) = \mu_0 + V(0) \tag{1.13}$$

$$Y(1) = \mu_1 + V(1), \tag{1.14}$$

where $V(0)$ and $V(1)$ have zero means, by construction. Following the discussion in Section 3.4, we linearly project each of $V(0)$ and $V(1)$ onto the population demeaned covariates, $\dot{\mathbf{X}}$:

$$V(0) = \dot{\mathbf{X}}\boldsymbol{\beta}_0 + U(0) \tag{1.15}$$

$$V(1) = \dot{\mathbf{X}}\boldsymbol{\beta}_1 + U(1) \tag{1.16}$$

where the intercepts are necessarily zero. Then

$$Y(0) = \mu_0 + \dot{\mathbf{X}}\boldsymbol{\beta}_0 + U(0) \tag{1.17}$$

$$Y(1) = \mu_1 + \dot{\mathbf{X}}\boldsymbol{\beta}_1 + U(1) \tag{1.18}$$

By definition of the linear projection,

$$\mathbb{E}\left[U(0)\right] = \mathbb{E}\left[U(1)\right] = 0$$
$$\mathbb{E}\left[\dot{\mathbf{X}}'U(0)\right] = \mathbb{E}\left[\dot{\mathbf{X}}'U(1)\right] = \mathbf{0}$$

It follows that

$$\mathbb{V}\left[Y(0)\right] = \boldsymbol{\beta}_0'\boldsymbol{\Omega}_{\mathbf{X}}\boldsymbol{\beta}_0 + \sigma_0^2$$
$$\mathbb{V}\left[Y(1)\right] = \boldsymbol{\beta}_1'\boldsymbol{\Omega}_{\mathbf{X}}\boldsymbol{\beta}_1 + \sigma_1^2$$

where $\sigma_0^2 = \mathbb{V}\left[U(0)\right]$ and $\sigma_1^2 = \mathbb{V}\left[U(1)\right]$.

We can write the observed outcome, Y, as

$$Y = (1 - W)\left[\mu_0 + \dot{\mathbf{X}}\boldsymbol{\beta}_0 + U(0)\right] + W\left[\mu_1 + \dot{\mathbf{X}}\boldsymbol{\beta}_1 + U(1)\right] \tag{1.19}$$

$$= \mu_0 + \dot{\mathbf{X}}\boldsymbol{\beta}_0 + U(0) + \tau W + \left(W \cdot \dot{\mathbf{X}}\right)\boldsymbol{\delta} + W \cdot \left[U(1) - U(0)\right] \tag{1.20}$$

where $\boldsymbol{\delta} = \boldsymbol{\beta}_1 - \boldsymbol{\beta}_0$. The following lemma is a precursor to the efficiency comparisons.

Lemma 1.5.1. *Under the assumptions of random assignment given in 1.3.1, random sampling 3.2.4, and finite moment assumptions, the following asymptotic distributions hold:*

$$\sqrt{N}\left(\hat{\tau}_{SDM} - \tau\right) \xrightarrow{d} \mathcal{N}\left(0, \omega^2_{SDM}\right) \tag{1.21}$$

$$\omega^2_{SDM} = \frac{\boldsymbol{\beta}_1'\boldsymbol{\Omega}_{\mathbf{X}}\boldsymbol{\beta}_1}{\rho} + \frac{\boldsymbol{\beta}_0'\boldsymbol{\Omega}_{\mathbf{X}}\boldsymbol{\beta}_0}{(1-\rho)} + \frac{\sigma^2_1}{\rho} + \frac{\sigma^2_0}{(1-\rho)} \tag{1.22}$$

$$\sqrt{N}\left(\hat{\tau}_{PRA} - \tau\right) \xrightarrow{d} \mathcal{N}\left(0, \omega^2_{PRA}\right) \tag{1.23}$$

$$\omega^2_{PRA} = \left(\frac{(1-\rho)^2}{\rho} + \frac{\rho^2}{(1-\rho)}\right)\left(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0\right)'\boldsymbol{\Omega}_{\mathbf{X}}\left(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0\right) + \frac{\sigma^2_1}{\rho} + \frac{\sigma^2_0}{(1-\rho)} \tag{1.24}$$

$$\sqrt{N}\left(\hat{\tau}_{FRA} - \tau\right) \xrightarrow{d} \mathcal{N}\left(0, \omega^2_{FRA}\right) \tag{1.25}$$

$$\omega^2_{FRA} = \left(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0\right)'\boldsymbol{\Omega}_{\mathbf{X}}\left(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0\right) + \frac{\sigma^2_1}{\rho} + \frac{\sigma^2_0}{(1-\rho)} \tag{1.26}$$

$$\sqrt{N}\left(\hat{\tau}_{IRA} - \tau\right) \xrightarrow{d} \mathcal{N}\left(0, \omega^2_{IRA}\right) \tag{1.27}$$

$$\omega^2_{IRA} = \frac{\sigma^2_1}{\rho} + \frac{\sigma^2_0}{(1-\rho)}. \; \square \tag{1.28}$$

The asymptotic variance expressions allow us to determine asymptotic efficiency under various scenarios. Not surprisingly, all four asymptotic variances depend on the error variances, σ^2_0 and σ^2_1. Generally, the asymptotic variances of the three feasible estimators can depend on $\boldsymbol{\Omega}_{\mathbf{X}}$, $\boldsymbol{\beta}_0$, and $\boldsymbol{\beta}_1$.

By comparing the formulas in Lemma 1.5.1 we have the following result, which ranks the asymptotic variances of the four different estimators in the general case of heterogeneous treatments and $\rho \in (0,1)$.

Theorem 1.5.2. *Under the assumptions of Lemma 1.5.1,*

(i)

$$\omega^2_{FRA} \leq \omega^2_{SDM} \tag{1.29}$$

$$\omega^2_{FRA} \leq \omega^2_{PRA} \tag{1.30}$$

$$\omega^2_{IRA} \leq \omega^2_{FRA} \tag{1.31}$$

(ii) If $\boldsymbol{\beta}_0 = \boldsymbol{\beta}_1 = \mathbf{0}$ then all asymptotic variances are the same.

(iii) If $\boldsymbol{\beta}_0 = \boldsymbol{\beta}_1 = \boldsymbol{\beta}$ then $\omega^2_{PRA} = \omega^2_{FRA} = \omega^2_{IRA}$ and if $\boldsymbol{\beta} \neq \mathbf{0}$ then ω^2_{SDM} is strictly larger.

(iv) If $\rho = 1/2$ then $\omega^2_{PRA} = \omega^2_{FRA} \leq \omega^2_{SDM}$, with strict inequality in the latter case unless $\boldsymbol{\beta}_1 = -\boldsymbol{\beta}_0$.

Many of the results in Theorem 1.5.2 follow from inspection of the asymptotic variance formulas, although some are more subtle. For example, (1.31) is immediate because the first term in (1.26) is nonnegative. Part (iii) is also immediate because all asymptotic variances equal $\sigma_1^2/\rho + \sigma_0^2/(1-\rho)$ when $\boldsymbol{\beta}_0 = \boldsymbol{\beta}_1$. For part (iv), the function

$$g(\rho) \equiv \frac{(1-\rho)^2}{\rho} + \frac{\rho^2}{(1-\rho)}, \rho \in (0,1)$$

can be shown to have a minimum value of unity, uniquely achieved when $\rho = 1/2$.

The most difficult inequality to establish, and the one that is most important, is (1.29). Straightforward matrix multiplication shows that

$$\omega^2_{SDM} - \omega^2_{FRA} = \frac{\boldsymbol{\beta}_1' \boldsymbol{\Omega}_{\mathbf{X}} \boldsymbol{\beta}_1}{\rho} + \frac{\boldsymbol{\beta}_0' \boldsymbol{\Omega}_{\mathbf{X}} \boldsymbol{\beta}_0}{(1-\rho)} - (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0)' \boldsymbol{\Omega}_{\mathbf{X}} (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0) = \boldsymbol{\lambda}' \boldsymbol{\Omega}_{\mathbf{X}} \boldsymbol{\lambda},$$

where

$$\boldsymbol{\lambda} = \sqrt{\left(\frac{1-\rho}{\rho}\right)}\boldsymbol{\beta}_1 + \sqrt{\left(\frac{\rho}{1-\rho}\right)}\boldsymbol{\beta}_0.$$

Because $\boldsymbol{\Omega}_{\mathbf{X}}$ is assumed positive definite, $\omega^2_{SDM} = \omega^2_{FRA}$ if and only if $\boldsymbol{\lambda} = \mathbf{0}$. One case where $\boldsymbol{\lambda} = \mathbf{0}$ is $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_0 = \mathbf{0}$, in which case the covariates do not predict the potential outcomes. It can happen in other cases but all of the slope coefficients would have to have opposite signs in the linear projections of the two potential outcomes. For example, if $\rho = 1/2$, we would need $\boldsymbol{\beta}_1 = -\boldsymbol{\beta}_0$, which means the slopes in the linear projection of $Y(1)$ on 1, $\mathbf{X}$ would be the opposite signs of the slope coefficients in the linear projection of $Y(0)$ on 1, $\mathbf{X}$. This seems highly unlikely. For example, we would expect pre-training education to have a positive effect on earnings whether or not someone participates in a job training program. In the homogenous case $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_0 \neq \mathbf{0}$, $\omega^2_{SDM} > \omega^2_{FRA} = \omega^2_{PRA}$.

We can never know for sure whether $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_0$, but we should know whether $\rho = 1/2$ based on the design of the experiment. If $\rho = 1/2$ then 1.5.2 suggests that the pooled estimator is probably preferred: it is as asymptotically efficient as the full RA estimator and conserves on degrees of freedom, which may be important if N is not large and the potential K (number of covariates) is somewhat large. For $\rho \neq 1/2$, Theorem 1.5.2 shows that the full RA estimator is attractive provided small-sample issues are not important. In particular, $\hat{\tau}_{FRA}$ is always more asymptotically efficient that both $\hat{\tau}_{SDM}$ and $\hat{\tau}_{PRA}$ in the presence of heterogenous slopes, and there is no (asymptotic) price to pay if $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_0$ or even if $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_0 = \mathbf{0}$. Estimating the $2K$ parameters is, asymptotically, harmless when it comes to the precision in estimating τ.

It may be helpful to provide intuition as to why $\hat{\tau}_{FRA}$ is more efficient than $\hat{\tau}_{SDM}$. Consider estimating the mean of the potential outcome in the treated state, μ_1. The FRA estimator is

$$\hat{\mu}_{1,FRA} = \hat{\alpha}_1 + \bar{\mathbf{X}}\hat{\boldsymbol{\beta}}_1,$$

where $\bar{\mathbf{X}}$ is the sample average across the entire sample. By the simple mechanics of OLS,

$$\bar{Y}_1 = \hat{\alpha}_1 + \bar{\mathbf{X}}_1\hat{\boldsymbol{\beta}}_1$$

where

$$\bar{\mathbf{X}}_1 = N_1^{-1}\sum_{i=1}^{N} W_i\mathbf{X}_i$$

is the sample average of the $\mathbf{X}_i$ over the treated units. Because of random assignment, $\bar{\mathbf{X}}_1$ is also a $\sqrt{N}$-consistent, asymptotically normal estimator of $\boldsymbol{\mu}_{\mathbf{X}}$. But it is inefficient compared with $\bar{\mathbf{X}}$ because the latter uses the entire sample. The same is true of $\hat{\mu}_{0,FRA}$ and $\bar{Y}_0$. This does not quite prove that $\hat{\tau}_{FRA}$ is asymptotically more efficient because $\hat{\mu}_{1,FRA}$ and $\hat{\mu}_{0,FRA}$ are not (asymptotically) uncorrelated, but it does provide some intuition. The same sort of intuition indicates why $\hat{\tau}_{IRA}$ is asymptotically more efficient than $\hat{\tau}_{FRA}$: $\hat{\tau}_{IRA}$ is not subject to the sampling error in estimating $\boldsymbol{\mu}_{\mathbf{X}}$.

1.6 Simulations

In this section we study the finite sample properties of the four estimators just discussed. We evaluate the estimators primarily in terms of root mean squared error (RMSE), since this accounts for bias as well as sampling variance. Since bias has been cited as a concern with covariate adjustment estimators, especially in small-scale experiments, looking at the trade offs between bias and efficiency through RMSE is key to studying the finite sample performance of these estimators. In order to compute the RMSE, we first generate a population of 10,000 observations to approximate an "infinite" population setting. We then draw random samples of sizes 100, 500 and 1,000 repeatedly from this population a thousand times. For a comprehensive assessment, we report the RMSE across these different sample sizes and treatment probability combinations where the treatment probabilities range from 0.1 to 0.9. To keep the tables simple, we report results only for the odd treatment probabilities even though the graphs are plotted for all values. The reported simulation results are for the case of heterogeneous treatment effects in the population, both in terms of the slopes on the linear projections and in the distribution of the projection errors, $U(0)$ and $U(1)$.

1.6.1 Design details

The treatment, W, is a binary variable, and so it has a Bernoulli distribution with

$$\mathbb{P}(W = 1) = \rho,$$

and we vary the value of ρ. For the potential outcomes, we consider continuous and discrete responses. In the first, the potential outcomes are conditionally normally distributed, with means linear in a quadratic in two covariates, X_1 and X_2. Specifically,

$$\begin{aligned} Y(0) &= \gamma_{00} + \gamma_{01}X_1 + \gamma_{02}X_2 + \gamma_{03}X_1^2 + \gamma_{04}X_2^2 + \gamma_{05}X_1X_2 + R(0) \equiv \mathbf{Z}\boldsymbol{\gamma}_0 + R(0) \\ Y(1) &= \gamma_{10} + \gamma_{11}X_1 + \gamma_{12}X_2 + \gamma_{13}X_1^2 + \gamma_{14}X_2^2 + \gamma_{15}X_1X_2 + R(1) \equiv \mathbf{Z}\boldsymbol{\gamma}_1 + R(1), \end{aligned}$$

where

$$\mathbf{Z} = \begin{pmatrix} 1 & X_1 & X_2 & X_1^2 & X_2^2 & X_1X_2 \end{pmatrix}$$

and

$$\begin{aligned} R(0)|\,(X_1, X_2) &\sim \mathcal{N}(0, \sigma_0^2) \\ R(1)|\,(X_1, X_2) &\sim \mathcal{N}(0, \sigma_1^2) \end{aligned}$$

We allow the γ_{wj} to differ across $w \in \{0, 1\}$, and so there is heterogeneity in the treatment effects in terms of the observables, $\mathbf{X}$, and the unobservables, $R(0)$ and $R(1)$ which are allowed to be correlated.[2] We also allow σ_0^2 and σ_1^2 to differ.

It is important to understand that, in order to be realistic, we do not assume that the quadratic conditional mean function is known. Instead, the researcher uses only linear regression on a constant, X_1, and X_2. In a traditional view of econometrics, these regressions would be "misspecified." Of course, to ensure we have the best mean squared error predictors of $Y(0)$ and $Y(1)$, we would use the correct specifications of $\mathbb{E}\left[Y(0)|\mathbf{X}\right]$ and $\mathbb{E}\left[Y(1)|\mathbf{X}\right]$. But it would be unusual for us to know the exact specification of the conditional mean

[2] $R(0)$ and $R(1)$ are generated to be affine transformations of the same standard normal variable.

functions. One can argue that most empirical researchers would include simple functions, such as squares as interactions. But then the true mean function could depend on higher order polynomials, or other more exotic functions. In fact, the mean might not even be linear in parameters. We take our setup as reflecting the realistic case that the researcher uses a linear regression that does not correspond to the correct conditional mean.

Our second design generates the potential outcomes as binary variables. Remember, when W is randomly assigned, we can use any kind of linear regression adjustment to improve asymptotic efficiency, regardless of the nature of $Y(0)$, $Y(1)$. Specifically, for $\mathbf{Z}$ defined above,

$$Y(0) = 1[\mathbf{Z}\boldsymbol{\gamma}_0 + R(0) > 0]$$
$$Y(1) = 1[\mathbf{Z}\boldsymbol{\gamma}_1 + R(1) > 0],$$

where $R(0)$ and $R(1)$ are again independent of (X_1, X_2) and normally distributed, this time each with unit variances. As before, $R(0)$ and $R(1)$ are allowed to be correlated. In the binary response case, one might traditionally think of two forms of "misspecification" in using linear regression adjustment on $(1, X_1, X_2)$. First, we are using what is traditionally called a "linear probability model" rather than the correct probit model. Second, we are omitting the terms X_1^2, X_2^2, and X_1X_2. Thus, there are two kinds of functional form "misspecification." Our view is that, to make a case for linear regression adjustment, it should produce notable efficiency gains even when the potential outcomes are discrete (although we return to this issue in the next section).

We consider two different designs for generating the covariates. Both are based on an underlying bivariate normal distribution:

$$\mathbf{X}^{*\prime} = \begin{pmatrix} X_1^* \\ X_2^* \end{pmatrix} \sim \mathcal{N}\left[\begin{pmatrix} 1 \\ 2 \end{pmatrix}, \begin{pmatrix} 2 & 0.5 \\ 0.5 & 3 \end{pmatrix} \right].$$

In the first design, $\mathbf{X} = \mathbf{X}^*$. In the second design, $X_1 = X_1^*$ and

$$X_2 = 1[X_2^* > 0],$$

so that X_2 is binary (in which case X_2^2 is redundant in the mechanism generating the potential outcomes).

With the linear and probit data we consider two different levels of heterogeneity across the coefficients $\boldsymbol{\gamma}_0$ and $\boldsymbol{\gamma}_1$, which we label "mild" and "strong." For the mild heterogeneity with continuous covariates

$$\boldsymbol{\gamma}_0' = \begin{pmatrix} 2 & 2 & -2 & -0.05 & -0.02 & 0.3 \end{pmatrix}, \quad \boldsymbol{\gamma}_1' = \begin{pmatrix} 3 & 1 & -1 & -0.05 & -0.03 & 0.6 \end{pmatrix}$$

and for the strong heterogeneity

$$\boldsymbol{\gamma}_0' = \begin{pmatrix} 0 & 1 & -1 & -0.05 & 0.02 & 0.6 \end{pmatrix}, \quad \boldsymbol{\gamma}_1' = \begin{pmatrix} 1 & -1 & 1.5 & 0.03 & -0.02 & -0.6 \end{pmatrix}$$

With the binary regressor, for mild heterogeneity we have

$$\boldsymbol{\gamma}_0' = \begin{pmatrix} 0 & 1 & -1 & 0.05 & 0.2 \end{pmatrix}, \quad \boldsymbol{\gamma}_1' = \begin{pmatrix} 3 & 3 & 1 & 0.05 & 0.9 \end{pmatrix}$$

and for strong heterogeneity we have

$$\boldsymbol{\gamma}_0' = \begin{pmatrix} 0 & 1 & -2 & -0.05 & 0.2 \end{pmatrix}, \quad \boldsymbol{\gamma}_1' = \begin{pmatrix} 3 & -1 & 1 & 0.05 & -0.9 \end{pmatrix}$$

Combining the linear and probit designs for the potential outcomes with two different levels of heterogeneity and two different covariate compositions leads to a total of eight scenarios. We allow the treatment probability, ρ, to range between 0.1 and 0.9 in increments of 0.1 . We consider three sample sizes, 100, 500, and 1,000. Note that when $N = 100$ and $\rho = 0.1$, we expect only 10 treated units and 90 control units. The covariates are generated to ensure that they are predictive of the potential outcomes, with the population R-squared ranging between $(0.1, 0.6)$. The variances σ_0^2 and σ_1^2 are allowed to be different for the two potential outcomes and across four of the eight different data generation processes.

We assess the relative finite sample performance of the four estimators under each such scenario, which we term a DGP. The Table below describes each of the DGP's in detail.

Table 1.1: Description of the data generating processes

DGP	Design	Covariates	Heterogeneity	R_0^2	R_1^2	σ_0^2	σ_1^2	PATE
1	Quadratic	$\mathbf{X}$	Mild	0.52	0.44	16	9	2.68
2	Quadratic	$\mathbf{X}$	Strong	0.31	0.46	16	9	0.93
3	Quadratic	$X_2 = 1[X_2^* > 0]$	Mild	0.59	0.33	1	4	7.46
4	Quadratic	$X_2 = 1[X_2^* > 0]$	Strong	0.27	0.34	9	4	2.92
5	Probit	$\mathbf{X}$	Mild	0.59	0.38	1	1	0.28
6	Probit	$\mathbf{X}$	Strong	0.51	0.45	1	1	0.09
7	Probit	$X_2 = 1[X_2^* > 0]$	Mild	0.45	0.28	1	1	0.35
8	Probit	$X_2 = 1[X_2^* > 0]$	Strong	0.38	0.40	1	1	0.43

1.6.2 Discussion of simulation findings

In the eight different DGPs, we see that FRA performs better than SDM and PRA in terms of RMSE. This behavior seems to be more pronounced at larger sample sizes as seen from the figures. Two things are worth pointing. One, the difference in IRA and FRA is less prominent for cases of mild heterogeneity. In such cases, PRA also performs comparably. This makes sense since pooling slopes in the treatment and control groups when the slopes are not very different should produces estimates that are close to the ones estimated by the separate slopes regression. Second, as was clear from Theorem 1.5.2, PRA and FRA have approximately the same RMSE at $\rho = 0.5$. This is not surprising to see in the graph because at larger sample sizes, biases in these estimators are negligible which means that RMSE is approximately the same as the variance.

Overall we see that the finite sample performance of FRA is superior to SDM and PRA for a variety of data generating processes (see figures A.1 and A.2 for quadratic design with mild and strong levels of heterogeneity, A.3 and A.4 for quadratic design with one binary covariate with mild and strong levels of heterogeneity, A.5 and A.6 for a probit design with mild and strong levels of heterogeneity and finally A.7 and A.8 for a probit design with one binary covariate with mild and strong levels of heterogeneity. For tables, see **??**, **??** and **??**).

1.7 Nonlinear regression adjustment

If the outcome Y – more precisely, the potential outcomes, $Y(0)$ and $Y(1)$ – are discrete, or have limited support, using nonlinear conditional mean functions, chosen to ensure fitted values are logically consistent with $\mathbb{E}\left[Y|\mathbf{X}\right]$, have considerable appeal. Intuitively, getting better approximations to $\mathbb{E}\left[Y(0)|\mathbf{X}\right]$ and $\mathbb{E}\left[Y(1)|\mathbf{X}\right]$ can yield estimators with smaller asymptotic variances when compared with the SDM estimator and linear regression adjustment. However, as cautioned by Imbens and Rubin (2015) page 128, one should not sacrifice consistency in order to obtain an asymptotically more efficient estimator. Imbens and Rubin (2015) leave the impression that all nonlinear models should be avoided because consistency cannot be ensured. In this section, we use the features of the linear exponential family class of distributions, combined with particular conditional mean models, to show that if one is careful in choosing the combination of conditional mean function and quasi-log likelihood (QLL) function, one can preserve consistency. Unfortunately, we cannot formally show that using this particular set of nonlinear models is more efficient than the SDM estimator, but our simulations suggest the efficiency gains can be substantial. (And we have found no cases where it is worse to do nonlinear RA.)

In deciding on nonlinear RA methods, the key is to remember is that

$$\tau_{ate} = \mu_1 - \mu_0,$$

and so we need to consistently estimate μ_1 and μ_0 without imposing additional assumptions. Earlier we showed how linear regression adjustment does just that. And, linear RA, when done separately to estimate μ_0 and μ_1, is asymptotically more efficient than the SDM estimator. Our goal here is to summarize the nonlinear methods that produce consistent estimators of τ_{ate} without additional assumptions (except for standard regularity conditions). We start with pooled methods.

1.7.1 Pooled nonlinear regression adjustment

In the generalized linear models (GLM) literature, it is well known that certain combinations of QLLs in the linear exponential family (LEF) and link functions lead to first order conditions where, in the sample, the residuals average to zero and are uncorrelated with every explanatory variable. To state the precise results, let $g(\cdot)$ be a strictly increasing function on $\mathbb{R}$, with range that can be a subset of $\mathbb{R}$. The inverse, $g^{-1}(\cdot)$, is known as the "link function" in the GLM literature. In the context of treatment effect estimation with mean function $g(\alpha + \mathbf{x}\boldsymbol{\beta} + \gamma w)$, when using the so-called canonical link function [McCullagh and Nelder (1989)] the first order conditions (FOCs) are of the form

$$\sum_{i=1}^{N}\left[Y_i - g\left(\hat{\alpha} + \mathbf{X}_i\hat{\boldsymbol{\beta}} + \hat{\gamma}W_i\right)\right] = 0$$

$$\sum_{i=1}^{N} W_i\left[Y_i - g\left(\hat{\alpha} + \mathbf{X}_i\hat{\boldsymbol{\beta}} + \hat{\gamma}W_i\right)\right] = 0$$

$$\sum_{i=1}^{N} \mathbf{X}_i'\left[Y_i - g\left(\hat{\alpha} + \mathbf{X}_i\hat{\boldsymbol{\beta}} + \hat{\gamma}W_i\right)\right] = \mathbf{0}$$

When $g(z) = z$, these equations produce the first order conditions for the pooled OLS estimator. The leading cases where these conditions hold for nonlinear estimation are for the Bernoulli QLL when $g(z) = \Lambda(z) = \exp(z)/[1 + \exp(z)]$ is the logistic function and for the Poisson QLL when $g(z) = \exp(z)$.

Under random sampling and weak regularity conditions, the probability limits of the estimators solve the population versions of the sample moment conditions. Let α^*, $\boldsymbol{\beta}^*$, and γ^* denote the probability limits of $\hat{\alpha}$, $\hat{\boldsymbol{\beta}}$, and $\hat{\gamma}$, respectively. Importantly, as argued in White (1982), these plims exist very generally without assuming that mean function is correctly specified – just as the parameters in the linear projection exist under very weak assumptions. The first two FOCs in the population are

$$\mathbb{E}\left[Y - g\left(\alpha^* + \mathbf{X}\boldsymbol{\beta}^* + \gamma^* W\right)\right] = 0 \tag{1.32}$$

$$\mathbb{E}\left\{W\left[Y - g\left(\alpha^* + \mathbf{X}\boldsymbol{\beta}^* + \gamma^* W\right)\right]\right\} = 0; \tag{1.33}$$

we will not need the last set of conditions obtained from the gradient with respect to $\boldsymbol{\beta}$. As before, we assume that $\rho = \mathbb{P}(W = 1)$ satisfies $0 < \rho < 1$.

Now, recall that $Y = (1 - W)Y(0) + WY(1)$. Then, by random assignment,

$$\mathbb{E}(Y) = \mathbb{E}(1 - W)\,\mathbb{E}\left[Y(0)\right] + \mathbb{E}(W)\,\mathbb{E}\left[Y(1)\right] = (1 - \rho)\mu_0 + \rho\mu_1.$$

Therefore, we can write (1.32) as

$$(1 - \rho)\mu_0 + \rho\mu_1 = \mathbb{E}\left[g\left(\alpha^* + \mathbf{X}\boldsymbol{\beta}^* + \gamma^* W\right)\right]$$

By iterated expectations,

$$\mathbb{E}\left[g\left(\alpha^* + \mathbf{X}\boldsymbol{\beta}^* + \gamma^* W\right)\right] = \mathbb{E}\left\{\mathbb{E}\left[g\left(\alpha^* + \mathbf{X}\boldsymbol{\beta}^* + \gamma^* W\right)|\mathbf{X}\right]\right\}$$

and, because W is independent of $\mathbf{X}$ with $\mathbb{P}(W = 1) = \rho$,

$$\mathbb{E}\left[g\left(\alpha^* + \mathbf{X}\boldsymbol{\beta}^* + \gamma^* W\right)|\mathbf{X}\right] = (1 - \rho)g\left(\alpha^* + \mathbf{X}\boldsymbol{\beta}^*\right) + \rho g\left(\alpha^* + \mathbf{X}\boldsymbol{\beta}^* + \gamma^*\right).$$

Therefore,

$$(1 - \rho)\mu_0 + \rho\mu_1 = (1 - \rho)\mathbb{E}\left[g\left(\alpha^* + \mathbf{X}\boldsymbol{\beta}^*\right)\right] + \rho\mathbb{E}\left[g\left(\alpha^* + \mathbf{X}\boldsymbol{\beta}^* + \gamma^*\right)\right]. \tag{1.34}$$

Further, using and $WY = WY(1)$, from (1.33),

$$\mathbb{E}\left[WY(1)\right] = \mathbb{E}\left[Wg\left(\alpha^* + \mathbf{X}\boldsymbol{\beta}^* + \gamma^* W\right)\right].$$

Again using random assignment and iterated expectations,

$$\rho\mu_1 = (1 - \rho) \cdot 0 + \rho\mathbb{E}\left[g\left(\alpha^* + \mathbf{X}\boldsymbol{\beta}^* + \gamma^*\right)\right] = \rho\mathbb{E}\left[g\left(\alpha^* + \mathbf{X}\boldsymbol{\beta}^* + \gamma^*\right)\right].$$

Because $\rho > 0$, we have

$$\mu_1 = \mathbb{E}\left[g\left(\alpha^* + \mathbf{X}\boldsymbol{\beta}^* + \gamma^*\right)\right]. \tag{1.35}$$

Also, because $\rho < 1$, (1.34) now implies

$$\mu_0 = \mathbb{E}\left[g\left(\alpha^* + \mathbf{X}\boldsymbol{\beta}^*\right)\right] \tag{1.36}$$

It follows that

$$\tau_{ate} = \mathbb{E}\left[g\left(\alpha^* + \mathbf{X}\boldsymbol{\beta}^* + \gamma^*\right)\right] - \mathbb{E}\left[g\left(\alpha^* + \mathbf{X}\boldsymbol{\beta}^*\right)\right]. \tag{1.37}$$

This equation essentially is the basis for proving that pooled regression adjustment, where we use a QLL in the linear exponential family and the conditional mean implied by the canonical link function, is consistent. Consistency follows because the estimated ATE using the pooled method is

$$\hat{\tau}_{ate,pooled} = N^{-1}\sum_{i=1}^{N}\left[g\left(\hat{\alpha} + \mathbf{X}_i\hat{\boldsymbol{\beta}} + \hat{\gamma}\right) - g\left(\hat{\alpha} + \mathbf{X}_i\hat{\boldsymbol{\beta}}\right)\right]. \tag{1.38}$$

By Wooldridge (2010) question 12.17, this converges in probability to (1.37). As a practical matter, it is convenient to note that (1.38) is the exact quantity reported by standard software packages when one requests the average "marginal" (or "partial") effect of the binary variable W after a standard GLM estimation. Packages that have this pre-programmed also provide a valid standard error, although one must be sure to use a "robust" option during the GLM estimation so that a sandwich estimator is used for the asymptotic variance of the parameter estimators $\left(\hat{\alpha}, \hat{\boldsymbol{\beta}}', \hat{\gamma}\right)'$.

Table B.1 summarizes the useful combinations of the QLL and mean functions that lead to consistent estimation of τ_{ate} without additional assumptions.

The Bernoulli/logistic case applies to binary or fractional outcomes, without change. When Y is fractional, it can have probability mass at zero, one, or anywhere else. See Papke and Wooldridge (1996) for further discussion. In any case, we treat the problem as

quasi-MLE because we do not wish to assume either that the distribution or mean function is correct.

The Poisson/exponential combination is very useful for nonnegative outcomes without a natural upper bound, although it can be applied to any nonnegative outcome. This includes count outcomes but also continuous outcomes and outcomes with corner solutions at zero (or other focal points). In the latter case, it is important to understand that commonly used models, such as Tobit, do not provide any known robustness to misspecification of the Tobit model. By contrast, the Poisson QMLE with an exponential mean provides full robustness. Remember, we are not trying to estimate the conditional mean functions; we are trying to consistently estimate the unconditional means, μ_0 and μ_1. Other than linear regression, the Poisson QMLE with an exponential mean is the only sensible choice for nonnegative, unbounded responses.

If the outcome has a natural, known upper bound, say B_i, which may vary by unit i, the binomial QMLE can be used in conjunction with the mean function

$$m(b, \mathbf{x}, w, \boldsymbol{\theta}) = b \left[\frac{\exp(\alpha + \mathbf{x}\boldsymbol{\beta} + \gamma w)}{1 + \exp(\alpha + \mathbf{x}\boldsymbol{\beta} + \gamma w)} \right],$$

as this is known to be the mean associated with the canonical link for the binomial distribution. The data then consists of $(Y_i, B_i, \mathbf{X}_i, W_i)$. Again, it does not matter whether Y_i is an integer response or is continuous, or even has a corner at zero, B_i, or both: using the binomial QMLE with logistic mean is simply a way to possibly improve over SDM or linear RA. The estimated ATE is

$$N^{-1} \sum_{i=1}^{N} B_i \left[\frac{\exp(\hat{\alpha} + \mathbf{X}_i\hat{\boldsymbol{\beta}} + \hat{\gamma})}{1 + \exp(\hat{\alpha} + \mathbf{X}_i\hat{\boldsymbol{\beta}} + \hat{\gamma})} - \frac{\exp(\hat{\alpha} + \mathbf{X}_i\hat{\boldsymbol{\beta}})}{1 + \exp(\hat{\alpha} + \mathbf{X}_i\hat{\boldsymbol{\beta}})} \right];$$

again, this is simple the average partial effect with respect to the binary variable W.

The last entry in Table B.1 extends the Bernoulli QLL/logistic mean and is relevant in two general situations. The first is when the support of the response is finite (and greater than two; otherwise one would use the logistic mean function with the Bernoulli QLL). For

example, $Y_g(w)$ could be an ordered response, such as a measure of health on a Lichert scale, or an unordered response, such as the choice of a health plan. A second situation is when the response consists of fractions that sum to unity, such as proportions of wealth in different investment categories, in which case the model has been called "multinomial fractional logit" [Mullahy (2015)]. If there are $G+1$ possible outcomes then there are $G+1$ means each for the control and treated states. The conditional mean functions for a pooled estimation would be

$$m_g(\mathbf{x}, w, \theta) = \frac{\exp(\alpha_g + \mathbf{x}\boldsymbol{\beta}_g + \gamma_g w)}{\left[1 + \sum_{h=1}^{G} \exp(\alpha_h + \mathbf{x}\boldsymbol{\beta}_h + \gamma_h w)\right]}, \; g = 0, 1, ..., G$$

with $\alpha_0 = 0$, $\boldsymbol{\beta}_0 = \mathbf{0}$. Then the estimated means are

$$\hat{\mu}_{wg} = N^{-1} \sum_{i=1}^{N} \frac{\exp(\hat{\alpha}_g + \mathbf{X}_i\hat{\boldsymbol{\beta}}_g + \hat{\gamma}_g w)}{\left[1 + \sum_{h=1}^{G} \exp(\hat{\alpha}_h + \mathbf{X}_i\hat{\boldsymbol{\beta}}_h + \hat{\gamma}_h w)\right]}, \; w \in \{0, 1\}, \; g \in \{0, 1, ..., G\}$$

and the estimated average treatment effect for each g is

$$\hat{\tau}_{ate,g} = \hat{\mu}_{1g} - \hat{\mu}_{0g}.$$

Because for each w, $\sum_{g=0}^{G} \hat{\mu}_{wg} = 1$, the sum over g of the $\hat{\tau}_{ate,g}$ is necessarily zero.

1.7.2 Full nonlinear regression adjustment

As in the linear case, consistency is preserved if we estimate two separate regression functions for the control and treatment cases. This follows from Wooldridge (2007) results on doubly robust estimators, where, in the current setting, the propensity score, $\mathbb{P}\left(W = 1|\mathbf{X} = \mathbf{x}\right) = \rho$, is not a function of $\mathbf{x}$. But a direct argument is easier to follow. For example, consider using a QLL with the canonical link function using only the treatments. The FOC for $\hat{\alpha}_1$, the intercept inside the conditional mean function, is simply

$$\sum_{i=1}^{N} W_i \left[Y_i - g\left(\hat{\alpha}_1 + \mathbf{X}_i \hat{\boldsymbol{\beta}}_1 \right) \right] = 0.$$

Notice again how the treatment indicator serves to select the subsample of treated units. The population analog is

$$\mathbb{E}\left[WY(1) \right] = \mathbb{E}\left[W g\left(\alpha_1^* + \mathbf{X}\boldsymbol{\beta}_1^* \right) \right]$$

or, because of random assignment,

$$\rho\mu_1 = \rho\mathbb{E}\left[g\left(\alpha_1^* + \mathbf{X}\boldsymbol{\beta}_1^* \right) \right].$$

It follows that

$$\mu_1 = \mathbb{E}\left[g\left(\alpha_1^* + \mathbf{X}\boldsymbol{\beta}_1^* \right) \right]$$

The same argument works for the untreated case, where W_i is replaced with $(1 - W_i)$, and $\left(\hat{\alpha}_1, \hat{\boldsymbol{\beta}}_1' \right)'$ are replaced with $\left(\hat{\alpha}_0, \hat{\boldsymbol{\beta}}_0' \right)'$. The conclusion is

$$\mu_0 = \mathbb{E}\left[g\left(\alpha_0^* + \mathbf{X}\boldsymbol{\beta}_0^* \right) \right].$$

Remember, the parameters with a "*" now indicate the probability limits from the two separate estimations, rather than there being the same parameters as in the pooled estimation. It follows under general regularity conditions that a consistent and asymptotically normal estimator of τ_{ate} is

$$\hat{\tau}_{ate,full} = N^{-1} \sum_{i=1}^{N} \left[g\left(\hat{\alpha}_1 + \mathbf{X}_i \hat{\boldsymbol{\beta}}_1 \right) - g\left(\hat{\alpha}_0 + \mathbf{X}_i \hat{\boldsymbol{\beta}}_0 \right) \right].$$

As a practical matter, some packages, such as Stata, have built-in commands for some full RA estimators, including the Bernoulli/logistic and the Poisson/exponential combinations, and so a standard error is computed along with the estimate. Again, one must be sure to

use a robust variance matrix estimator for the parameters. Alternatively, using a bootstrap routine is very efficient for these kinds of estimators.

In deciding on a procedure to use – linear versus nonlinear, pooled versus full – it is important to understand that all methods studied in this paper produce consistent estimators of τ_{ate}. In the linear case, we have the result that full RA is asymptotically efficient compared with SDM and pooled RA. As mentioned earlier, a proof that full nonlinear RA is asymptotically more efficient than the pooled version is elusive. Also, we have not proven that full nonlinear RA is always at least as asymptotically efficient as SDM. We now report representative simulations that show the nonlinear methods can improve precision substantially in some cases without introducing bias, even in pretty small sample sizes.

1.7.3 Simulations

For non-linear simulations we only consider continuous covariates which means that for both binary and non-negative data generating processes, $\mathbf{X} = \mathbf{X}^*$ where,

$$\mathbf{X}^{*\prime} = \begin{pmatrix} X_1^* \\ X_2^* \end{pmatrix} \sim \mathcal{N}\left[\begin{pmatrix} 1 \\ 2 \end{pmatrix}, \begin{pmatrix} 2 & 0.5 \\ 0.5 & 3 \end{pmatrix} \right].$$

As with the linear simulations,

$$\mathbf{Z} = \begin{pmatrix} 1 & X_1 & X_2 & X_1^2 & X_2^2 & X_1 X_2 \end{pmatrix}.$$

The tables report bias and standard deviation for sample sizes of $N = 500$ and $N = 1,000$. To keep the tables simple, we only report values for treatment probabilities $\rho = 0.1, 0.3, 0.5, 0.7, 0.9$ even though the graphs are plotted for ρ ranging between 0.1 to 0.9.

1.7.3.1 Binary response

For the binary response case, the outcomes have been generated using a probit mean function as given below

$$Y(0) = 1[\mathbf{Z}\boldsymbol{\gamma}_0 + R(0) > 0]$$
$$Y(1) = 1[\mathbf{Z}\boldsymbol{\gamma}_1 + R(1) > 0],$$

and

$$\boldsymbol{\gamma}_0' = \begin{pmatrix} -2 & 1 & 2 & 0.05 & 0.02 & 0.1 \end{pmatrix}, \quad \boldsymbol{\gamma}_1' = \begin{pmatrix} 0 & 3 & 1 & -0.05 & 0.03 & 0.9 \end{pmatrix}$$

and

$$R(0)|\,(X_1, X_2) \sim \mathcal{N}(0, 1)$$
$$R(1)|\,(X_1, X_2) \sim \mathcal{N}(0, 1)$$

where $R(0)$ and $R(1)$ are allowed to be correlated. While estimating the nonlinear pooled and separate slopes estimators we use case (ii) in Table B.1.

We find that separate slopes nonlinear estimator (NFRA) has the lowest root mean squared error compared with the linear estimators and the pooled nonlinear estimator (NPRA) for all treatment probabilities (see figure A.9). The tables show that nonlinear estimators have bias that is comparable to the SDM estimator (see table **??**).

1.7.3.2 Nonnegative response

For the non-negative response, the outcomes have been generated using a log normal distribution as given below

$$Y(0) = \exp\left(\frac{\mathbf{Z}\boldsymbol{\gamma_0} + R(0)}{10} + 0.3 \cdot \mathcal{N}(0, 1)\right)$$
$$Y(1) = \exp\left(\frac{\mathbf{Z}\boldsymbol{\gamma_1} + R(1)}{10} + 0.4 \cdot \mathcal{N}(0, 1)\right),$$

where

$$\boldsymbol{\gamma}_0' = \begin{pmatrix} 0 & 2 & -1 & -0.05 & 0.02 & 0.6 \end{pmatrix}, \quad \boldsymbol{\gamma}_1' = \begin{pmatrix} 1 & -1 & 1.5 & 0.03 & -0.02 & -0.6 \end{pmatrix}$$

and

$$R(0) | (X_1, X_2) \sim \mathcal{N}(0, 4)$$
$$R(1) | (X_1, X_2) \sim \mathcal{N}(0, 9)$$

where $R(0)$ and $R(1)$ are allowed to be correlated. While estimating the nonlinear pooled and separate slopes estimators we use case (iv) in Table B.1.

Similar to the binary response simulations, we see that that NFRA again has the lowest root mean squared error compared with both linear and pooled nonlinear estimators across all treatment probabilities. In fact, NFRA peforms better than both SDM and FRA. The NPRA and linear PRA are very similar in terms of RMSE; see table **??** and figure A.10.

1.8 Concluding remarks

We have studied linear and nonlinear regression adjustment estimators of the average treatment effect in an experimental framework. For linear regression adjustment, this paper makes some key contributions to the econometrics literature on randomized experiments. First, by considering a previously ignored aspect of the separate slopes estimator, this paper is able to fill a clear gap in the literature by showing the full RA estimator is always the most efficient even when the population means of the covariates is estimated using the sample sample. Second, in obtaining our results, we rely only on linear projections, and so no extra assumptions are used in establishing asymptotic efficiency. Third, the setup allows us to determine when using full RA, or RA at all, is unecessary to achieve efficiency. Our simulation findings support the theory and show that substantial efficiency gains are possible when we have good predictors of the response. Obtaining the correct standard errors for the full RA estimator is particularly simple in commonly used software packages. For example,

Stata®, with its built-in "teffects" command, provides the correct standard errors for the FRA estimator

As an interesting complement to our work, Słoczyński (2018) studies pooled versus full RA when assignment is unconfounded conditional on covariates. Assuming that the conditional means are linear in parameters, Słoczyński (2018) shows that using pooled RA when the treatment effects are heterogeneous is inconsistent for the ATE in a way that is particularly troublesome in designs that are heavily unbalanced. In particular, the pooled RA estimator consistently estimates the weighted average $(1-\rho)\cdot\tau_{ATT}+\rho\cdot\tau_{ATU}$, where τ_{ATT} is the average treatment effect on the treated ($W=1$) and τ_{ATU} is the ATE on the untreated ($W=0$). The ATE can be expressed as $\tau_{ATE}=\rho\cdot\tau_{ATT}+(1-\rho)\cdot\tau_{ATU}$, and so the PRA estimator, in the limit, gets the weights reversed. Under random assignment, there is no difference between τ_{ATE}, τ_{ATT}, and τ_{ATU}, and so consistency of PRA for τ_{ATE} is not the issue. But as we showed, the pooled RA estimator is generally inefficient when treatment effects are heterogeneous. Also, when $\rho=1/2$, there is no inconsistency in the pooled RA estimator when unconfoundedness holds. As we have shown in this paper, in the random assignment case $\rho=1/2$ is precisely the condition that implies no efficiency gain from full RA even when there is arbitrary heterogeneity in the treatment effects. Our findings mesh well with those of Słoczyński (2018), with the conclusion that in moderate samples, FRA should be used unless ρ is known to be close to $1/2$.

In addition to the linear estimators, we also propose nonlinear regression adjustment estimators, characterizing the combinations of quasi-log-likelihood functions and conditional means functions that ensure consistency regardless of misspecification. We believe this paper is the first to do so. We do not have theoretical results to show when the nonlinear RA methods unambiguously improve asymptotic efficiency, and this is a good topic for future research. However, our simulations suggest that nonlinear adjustment estimators can have bias comparable to that of simple difference in means (SDM) and can produce sampling variances that are considerably smaller than that of SDM and, in majority of cases, substantially

smaller than linear feasible regression adjustment.

Going forward, there are a lot of natural extensions. One is to study an assignment scheme that is different from the one considered here. This paper assumes independence across treatment assignments but a more common design, known as the completely randomized experiment, induces dependence across units by fixing the number of treated units before sampling from the population. Also, because most randomized experiments in economics are plagued with issues of nonparticipation or nonrandom attrition, it is also fruitful to study regression adjustment in conjunction with an Instrumental Variables (IV). Comparing the efficiency of standard regression adjustment estimators under random assignment to estimators based on stratified assignment schemes is also a good area for future research.

CHAPTER 2

ROBUST AND EFFICIENT ESTIMATION OF POTENTIAL OUTCOME MEANS UNDER RANDOM ASSIGNMENT[†]

2.1 Introduction

In the past several decades, the potential outcomes framework has become a staple of causal inference in statistics, econometrics, and related fields. Envisioning each unit in a population under different states of intervention or treatment allows one to define treatment or causal effects without referencing a model. One merely needs the means of the potential outcomes, or perhaps the potential outcome (PO) means for subpopulations.

When interventions are randomized – whether the assignment is to control and treatment groups in a clinical trial (Hirano and Imbens (2001)), assignment to participate in a job training program (Calónico and Smith (2017)), receiving a school voucher when studying the effects of private schooling on educational outcomes (Angrist et al. (2006a)), or contingent valuation studies, where different bid values are randomized among people (Carson et al. (2004)) – one can simply use the subsample means for each treatment level in order to obtain unbiased and consistent estimators of the PO means. In some cases, the precisions of the subsample means will be sufficient. Nevertheless, with the availability of good predictors of the outcome or response, it is appealing to think that the precision can be improved, thereby shrinking confidence intervals and making conclusions about interventions more reliable.

In this paper we build on Negi and Wooldridge (2019), who studied the problem of estimating the average treatment effect under random assignment with one control group and one treatment group. In the context of random sampling, we showed that performing separate linear regressions for the control and treatment groups in estimating the average treatment effect never does worse, asymptotically, than the simple difference in means esti-

[†]This work is joint with Jeffrey M. Wooldridge and is unpublished.

mator or a pooled regression adjustment estimator. In addition, we characterized the class of nonlinear regression adjustment methods that produce consistent estimators of the ATE without any additional assumptions (except regularity conditions). The simulation findings for both the linear and nonlinear cases are quite promising when covariates are available that predict the outcomes.

In the current paper, we consider any number of "treatment" levels and consider the problem of joint estimation of the vector of potential outcome means. We assume that the assignment to the treatment is random – that is, independent of both potential outcomes and observed predictors of the POs. Importantly, other than standard regularity conditions (such as finite second moments of the covariates), we impose no additional assumptions. In other words, the full RA estimators are consistent under essentially the same assumptions as the subsample means with, generally, smaller asymptotic variance. Interestingly, even if the predictors are unhelpful, or the slopes in the linear projections are the same across all groups, no asymptotic efficiency is lost by using the most general RA method.

We also extend the nonlinear RA results in Negi and Wooldridge (2019) to the general case of G assignment levels. We show that for particular kinds of responses such as binary, fractional or nonnegative, it is possible to consistently estimate PO means using pooled and separate regression adjustment. Unlike the linear regression adjustment case, we do not have any general asymptotic results to compare full nonlinear RA with pooled nonlinear RA.

Finally, we apply the full RA estimator to data from a contingent valuation study obtained from Carson et al. (2004). This data is used to elicit a lower bound on mean willingness to pay (WTP) for a program that would prevent future oil spills along California's central coast. Our results show that the PO means for the five different bid amounts that were randomly assigned to California residents are estimated more efficiently using separate regression adjustment than just using subsample averages. This efficiency result is preserved for estimating the lower bound since it is a linear combination of PO means. Hence, using separate RA also delivers a more precise lower bound mean WTP for the oil spill prevention

program than the ABERS estimator which uses subsample averages for constructing the estimate.

A monte carlo excercise substantiates the theoretical results across three kinds of data generating processes. We generate the outcomes to be either generated to be continuous non-negative values or multinomial responses. In addition, we consider four different configurations of the assignment vector. In each setting, we find FRA to be atleast as precise as SM across three different sample sizes. The performance of PRA relative to SM is less decisive since some of the subsample means are estimated more noisily than their SM counterparts.

The rest of the paper is organized as follows: Section 2.2 discusses the potential outcomes framework extended to the case of G treatment levels along with a discussion of the crucial random sampling and random assignment assumptions. Section 2.3 derives the asymptotic variances of the different linear regression adjustment estimators, namely, subsample means, pooled regression adjustment and full regression adjustment. Section 2.4 compares the asymptotic variances of the entire vector of subsample means, pooled and full regression adjustment. Section 2.5 considers a class of nonlinear regression adjustment estimators that ensure consistency of the subsample means without imposing additional assumptions. Section 2.6 discusses applications of this framework to randomized experiments, differences in differences settings and contingent valuation studies. This section also applies full regression adjustment estimator for estimating the lower bound mean WTP for the California Oil Spill study using data from Carson et al. (2004). Section 2.7 constructs a monte carlo study for studying and comparing the finite sample behavior of the linear regression adjustment estimators and Section 2.8 discusses the results of this study. Section 2.9 concludes.

2.2 Potential outcomes, random assignment, and random sampling

We use the standard potential outcomes framework, also known as the Neyman-Rubin causal model. The goal is to estimate the population means of G potential (counterfactual)

outcomes, $Y(g)$, $g = 1, ..., G$. Define

$$\mu_g = \mathbb{E}\left[Y(g)\right], \, g = 1, ..., G.$$

The vector of assignment indicators is

$$\mathbf{W} = (W_1, ..., W_G),$$

where each W_g is binary and

$$W_1 + W_2 + \cdots + W_G = 1.$$

In other words, the groups are exhaustive and mutually exclusive. The setup applies to many situations, including the standard treatment-control group setup, with $G = 2$, multiple treatment levels (with $g = 1$ the control group), the basic difference-in-differences setup with $G = 4$, and in contingent valuation studies where subjects are presented with a set of G prices or bid values.

We assume that each group, g, has a positive probability of being assigned:

$$\rho_g \equiv \mathbb{P}(W_g = 1) > 0, \, g = 1, ..., G$$

$$\rho_1 + \rho_2 + \cdots + \rho_G = 1$$

Next, let

$$\mathbf{X} = (X_1, X_2, ..., X_K)$$

be a vector of observed covariates.

Assumption 2.2.1 (**Random Assignment**). *Assignment is independent of the potential outcomes and observed covariates:*

$$\mathbf{W} \perp \left[Y(1), Y(2), ..., Y(G), \mathbf{X}\right].$$

Further, the assignment probabilities are all strictly positive. □

Assumption 1 is what puts us in the framework of experimental interventions. It would be much too strong for an observational study.

Assumption 2.2.2 (**Random Sampling**). *For a nonrandom integer N,*

$$\left\{ \left[\mathbf{W}_i, Y_i(1), Y_i(2), ..., Y_i(G), \mathbf{X}_i \right] : i = 1, 2, ..., N \right\}$$

is independent and identically distributed. □

The IID assumption is not the only one we can make. For example, we could allow for a sampling-without-replacement scheme given a fixed sample size N. This would complicate the analysis because it generates slight correlation within draws. As discussed in Negi and Wooldridge (2019), Assumption 2 is traditional in studying the asymptotic properties of estimators and is realistic as an approximation. Importantly, it forces us to account for the sampling error in the sample average, $\bar{\mathbf{X}}$, as an estimator of $\boldsymbol{\mu}_{\mathbf{X}} = \mathbb{E}(\mathbf{X})$.

For each draw i from the population, we only observe

$$Y_i = W_{i1} Y_i(1) + W_{i2} Y_i(2) + \cdots + W_{iG} Y_i(G),$$

and so the data we have to work with is

$$\left\{ (\mathbf{W}_i, Y_i, \mathbf{X}_i) : i = 1, 2, ..., N \right\}.$$

Definition of population quantities only requires us to use the random vector $(\mathbf{W}, Y, \mathbf{X})$, which represents the population.

Assumptions 1 and 2 are the only substantive restrictions used in this paper. Subsequently, we assume that linear projections exist and that the central limit theorem holds for properly standardized sample averages of IID random vectors. Therefore, we are implicitly imposing at least finite second moment assumptions on the $Y(g)$ and the X_j. We do not make this explicit in what follows.

2.3 Subsample means and linear regression adjustment

In this section we derive the asymptotic variances of three estimators: the subsample means, full (separate) regression adjustment, and pooled regression adjustment.

2.3.1 Subsample means

The simplest estimator of μ_g is the sample average within treatment group g:

$$\bar{Y}_g = N_g^{-1} \sum_{i=1}^{N} W_{ig} Y_i = N_g^{-1} \sum_{i=1}^{N} W_{ig} Y_i(g),$$

where

$$N_g = \sum_{i=1}^{N} W_{ig}$$

is a random variable in our setting. In expressing $\bar{Y}_g$ as a function of the $Y_i(g)$ we use $W_{ih}W_{ig} = 0$ for $h \neq g$. Under random assignment and random sampling,

$$\begin{aligned}
\mathbb{E}\left(\bar{Y}_g | W_{1g}, ..., W_{Ng}, N_g > 0\right) &= N_g^{-1} \sum_{i=1}^{N} W_{ig} \mathbb{E}\left[Y_i(g) | W_{1g}, ..., W_{Ng}, N_g > 0\right] \\
&= N_g^{-1} \sum_{i=1}^{N} W_{ig} \mathbb{E}\left[Y_i(g)\right] \\
&= N_g^{-1} \sum_{i=1}^{N} W_{ig} \mu_g = \mu_g,
\end{aligned}$$

and so $\bar{Y}_g$ is unbiased conditional on observing a positive number of units in group g.

By the law of large numbers, a consistent estimator of ρ_g is

$$\hat{\rho}_g = N_g / N,$$

the sample share of units in group g. Therefore, by the law of large numbers and Slutsky's Theorem,

$$\begin{aligned}
\bar{Y}_g &= \left(\frac{N}{N_g}\right) N^{-1} \sum_{i=1}^{N} W_{ig} Y_i(g) \xrightarrow{p} \rho_g^{-1} \mathbb{E}\left[W_g Y(g)\right] \\
&= \rho_g^{-1} \mathbb{E}\left(W_g\right) \mathbb{E}\left[Y(g)\right] = \mu_g,
\end{aligned}$$

and so $\bar{Y}_g$ is consistent for μ_g.

By the central limit theorem, $\sqrt{N}\left(\bar{Y}_g - \mu_g\right)$ is asymptotically normal. We need an asymptotic representation of $\sqrt{N}\left(\bar{Y}_g - \mu_g\right)$ that allows us to compare its asymptotic variance with those from regression adjustment estimators. To this end, write

$$\begin{aligned} Y(g) &= \mu_g + V(g) \\ \dot{\mathbf{X}} &= \mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}}, \end{aligned}$$

where $\dot{\mathbf{X}}$ is $\mathbf{X}$ demeaned using the population mean, $\boldsymbol{\mu}_{\mathbf{X}}$. Now project each $V(g)$ linearly onto $\dot{\mathbf{X}}$:

$$V(g) = \dot{\mathbf{X}}\boldsymbol{\beta}_{\boldsymbol{g}} + U(g),\ g = 1, ..., G.$$

By construction, the population projection errors $U(g)$ have the properties

$$\begin{aligned} \mathbb{E}\left[U(g)\right] &= 0,\ g = 1, ..., G \\ \mathbb{E}\left[\dot{\mathbf{X}}' U(g)\right] &= \mathbf{0},\ g = 1, ..., G. \end{aligned}$$

Plugging in gives

$$Y(g) = \mu_g + \dot{\mathbf{X}}\boldsymbol{\beta}_{\boldsymbol{g}} + U(g),\ g = 1, ..., G$$

Importantly, by random assignment, $\mathbf{W}$ is independent of $\left[U(1), ..., U(G), \dot{\mathbf{X}}\right]$. The observed outcome can be written as

$$Y = \sum_{g=1}^{G} W_g \left[\mu_g + \dot{\mathbf{X}}\boldsymbol{\beta}_{\boldsymbol{g}} + U(g)\right]$$

Theorem 2.3.1 (Asymptotic variance of Subsample means estimator of PO means)**.** *Under Assumptions 1, 2, and finite second moments,*

$$\begin{aligned} \sqrt{N}\left(\bar{\mathbf{Y}} - \boldsymbol{\mu}\right) &= \begin{pmatrix} N^{-1/2}\sum_{i=1}^{N}\left[W_{i1}\dot{\mathbf{X}}_i\boldsymbol{\beta}_{\mathbf{1}}/\rho_1 + W_{i1}U_i(1)/\rho_1\right] \\ N^{-1/2}\sum_{i=1}^{N}\left[W_{i2}\dot{\mathbf{X}}_i\boldsymbol{\beta}_{\mathbf{2}}/\rho_2 + W_{i2}U_i(2)/\rho_2\right] \\ \vdots \\ N^{-1/2}\sum_{i=1}^{N}\left[W_{iG}\dot{\mathbf{X}}_i\boldsymbol{\beta}_{\boldsymbol{G}}/\rho_G + W_{iG}U_i(G)/\rho_G\right] \end{pmatrix} + o_p(1) \\ &\equiv N^{-1/2}\sum_{i=1}^{N}\left(\mathbf{L}_i + \mathbf{Q}_i\right) + o_p(1) \end{aligned}$$

where

$$\mathbf{L}_i \equiv \begin{pmatrix} W_{i1}\dot{\mathbf{X}}_i\boldsymbol{\beta_1}/\rho_1 \\ W_{i2}\dot{\mathbf{X}}_i\boldsymbol{\beta_2}/\rho_2 \\ \vdots \\ W_{iG}\dot{\mathbf{X}}_i\boldsymbol{\beta_G}/\rho_G \end{pmatrix} \tag{2.1}$$

and

$$\mathbf{Q}_i \equiv \begin{pmatrix} W_{i1}U_i(1)/\rho_1 \\ W_{i2}U_i(2)/\rho_2 \\ \vdots \\ W_{iG}U_i(G)/\rho_G \end{pmatrix} \tag{2.2}$$

By random assignment and the linear projection property, $\mathbb{E}(\mathbf{L}_i) = \mathbb{E}(\mathbf{Q}_i) = \mathbf{0}$, and $\mathbb{E}\left(\mathbf{L}_i\mathbf{Q}_i'\right) = \mathbf{0}$. Also, because $W_{ig}W_{ih} = 0$, $g \neq h$, the elements of $\mathbf{L}_i$ are pairwise uncorrelated; the same is true of the elements of $\mathbf{Q}_i$.

2.3.2 Full regression adjustment

To motivate full regression adjustment, write the linear projection for each g as

$$\begin{aligned} Y(g) &= \alpha_g + \mathbf{X}\boldsymbol{\beta_g} + U(g) \\ \mathbb{E}\left[U(g)\right] &= 0 \\ \mathbb{E}\left[\mathbf{X}'U(g)\right] &= \mathbf{0} \end{aligned}$$

It follows immediately that

$$\mu_g = \alpha_g + \boldsymbol{\mu_X}\boldsymbol{\beta_g}.$$

Theorem 2.3.2 (Asymptotic variance of Full regression adjustment estimator of PO means)**.**

Under assumptions 1 and 2, and finite second moments,

$$\begin{aligned}\sqrt{N}\left(\hat{\boldsymbol{\mu}}-\boldsymbol{\mu}\right) &= \begin{pmatrix} N^{-1/2}\sum_{i=1}^{N}\left[\dot{\mathbf{X}}_i\boldsymbol{\beta_1}+W_{i1}U_i(1)/\rho_1\right] \\ N^{-1/2}\sum_{i=1}^{N}\left[\dot{\mathbf{X}}_i\boldsymbol{\beta_2}+W_{i2}U_i(2)/\rho_2\right] \\ \vdots \\ N^{-1/2}\sum_{i=1}^{N}\left[\dot{\mathbf{X}}_i\boldsymbol{\beta_G}+W_{iG}U_i(G)/\rho_G\right] \end{pmatrix} + o_p(1) \\ &\equiv N^{-1/2}\sum_{i=1}^{N}\left(\mathbf{K}_i+\mathbf{Q}_i\right)+o_p(1)\end{aligned}$$

where

$$\mathbf{K}_i = \begin{pmatrix} \dot{\mathbf{X}}_i\boldsymbol{\beta_1} \\ \dot{\mathbf{X}}_i\boldsymbol{\beta_2} \\ \vdots \\ \dot{\mathbf{X}}_i\boldsymbol{\beta_G} \end{pmatrix} \tag{2.3}$$

and $\mathbf{Q}_i$ *is given in (2.2).*

Both $\mathbf{K}_i$ and $\mathbf{Q}_i$ have zero means, the latter by random assignment. Further, by random assignment and the linear projection property, $\mathbb{E}\left(\mathbf{K}_i\mathbf{Q}_i'\right)=\mathbf{0}$ because

$$\mathbb{E}\left[\dot{\mathbf{X}}_i'W_{ig}U_i(g)\right]=\mathbb{E}(W_{ig})\mathbb{E}\left[\dot{\mathbf{X}}_i'U_i(g)\right]=\mathbf{0}.$$

However, unlike the elements of $\mathbf{L}_i$, we must recognize that the elements of $\mathbf{K}_i$ are correlated except in the trivial case that all but one of the $\boldsymbol{\beta_g}$ are zero.

2.3.3 Pooled regression adjustment

Now consider the pooled RA estimator, $\check{\mu}$, which can be obtained as the vector of coefficients on $\mathbf{W}_i=(W_{i1},W_{i2},...,W_{iG})$ from the regression

$$Y_i \text{ on } \mathbf{W}_i,\ \ddot{\mathbf{X}}_i,\ i=1,2,...,N.$$

We refer to this as a pooled method because the coefficients on $\ddot{\mathbf{X}}_i$, say, $\check{\boldsymbol{\beta}}$, are assumed to be the same for all groups. Compared with subsample means, we add the controls $\ddot{\mathbf{X}}_i$, but

unlike FRA, the pooled method imposes the same coefficients across all g.

Theorem 2.3.3 (Asymptotic variance of Pooled regression adjustment estimator of PO means)**.** *Under assumptions (1) and (2), along with finite second moments*

$$\begin{aligned}\sqrt{N}\left(\tilde{\boldsymbol{\mu}}-\boldsymbol{\mu}\right) &= \begin{pmatrix} N^{-1/2}\sum_{i=1}^{N}\left[\rho_1^{-1}W_{i1}\dot{\mathbf{X}}_i\left(\boldsymbol{\beta}_1-\boldsymbol{\beta}\right)+\dot{\mathbf{X}}_i\boldsymbol{\beta}+W_{i1}U_i(1)/\rho_1\right] \\ N^{-1/2}\sum_{i=1}^{N}\left[\rho_2^{-1}W_{i2}\dot{\mathbf{X}}_i\left(\boldsymbol{\beta}_2-\boldsymbol{\beta}\right)+\dot{\mathbf{X}}_i\boldsymbol{\beta}+W_{i2}U_i(2)/\rho_2\right] \\ \vdots \\ N^{-1/2}\sum_{i=1}^{N}\left[\rho_G^{-1}W_{iG}\dot{\mathbf{X}}_i\left(\boldsymbol{\beta}_G-\boldsymbol{\beta}\right)+\dot{\mathbf{X}}_i\boldsymbol{\beta}+W_{iG}U_i(G)/\rho_G\right] \end{pmatrix} + o_p(1) \\ &\equiv N^{-1/2}\sum_{i=1}^{N}\left(\mathbf{F}_i+\mathbf{K}_i+\mathbf{Q}_i\right)+o_p(1)\end{aligned}$$

where $\mathbf{K}_i$ *and* $\mathbf{Q}_i$ *are as before and, with* $\boldsymbol{\delta}_g=\boldsymbol{\beta}_g-\boldsymbol{\beta}$*,*

$$\mathbf{F}_i = \begin{pmatrix} \rho_1^{-1}\left(W_{i1}-\rho_1\right)\dot{\mathbf{X}}_i\boldsymbol{\delta}_1 \\ \rho_2^{-1}\left(W_{i2}-\rho_2\right)\dot{\mathbf{X}}_i\boldsymbol{\delta}_2 \\ \\ \rho_G^{-1}\left(W_{iG}-\rho_G\right)\dot{\mathbf{X}}_i\boldsymbol{\delta}_G \end{pmatrix} \tag{2.4}$$

Notice that, again by random assignment and the linear projection property,

$$\mathbb{E}\left(\mathbf{F}_i\mathbf{K}_i'\right)=\mathbb{E}\left(\mathbf{F}_i\mathbf{Q}_i'\right)=\mathbf{0}$$

2.4 Comparing the asymptotic variances

We now take the representations derived in Section 3 and use them to compare the asymptotic variances of the three estimators. For notational clarity, it is helpful summarize

the conclusions reached in Section 3:

$$\begin{aligned}
\sqrt{N}\left(\hat{\boldsymbol{\mu}}_{SM}-\boldsymbol{\mu}\right) &= N^{-1/2}\sum_{i=1}^{N}\left(\boldsymbol{L}_i+\mathbf{Q}_i\right)+o_p(1)\\
\sqrt{N}\left(\hat{\boldsymbol{\mu}}_{FRA}-\boldsymbol{\mu}\right) &= N^{-1/2}\sum_{i=1}^{N}\left(\mathbf{K}_i+\mathbf{Q}_i\right)+o_p(1)\\
\sqrt{N}\left(\hat{\boldsymbol{\mu}}_{PRA}-\boldsymbol{\mu}\right) &= N^{-1/2}\sum_{i=1}^{N}\left(\mathbf{F}_i+\mathbf{K}_i+\mathbf{Q}_i\right)+o_p(1),
\end{aligned}$$

where $\mathbf{L}_i$, $\mathbf{Q}_i$, $\mathbf{K}_i$, and $\mathbf{F}_i$ are defined in 2.1, 2.2, 2.3 and 2.4 respectively.

2.4.1 Comparing FRA to subsample means

Theorem 2.4.1. *Under assumptions of theorems 2.3.1 and 2.3.2,*

$$Avar\left[\sqrt{N}\left(\hat{\boldsymbol{\mu}}_{SM}-\boldsymbol{\mu}\right)\right]-Avar\left[\sqrt{N}\left(\hat{\boldsymbol{\mu}}_{FRA}-\boldsymbol{\mu}\right)\right]=\boldsymbol{\Omega}_{\boldsymbol{L}}-\boldsymbol{\Omega}_{\boldsymbol{K}} \tag{2.5}$$

is PSD.

The one case where there is no gain in asymptotic efficiency in using FRA is when $\boldsymbol{\beta}_g=\mathbf{0}$, $g=1,...,G$, in which case $\mathbf{X}$ does not help predict any of the potential outcomes. Importantly, there is no gain in asymptotic efficiency in imposing $\boldsymbol{\beta}_g=\mathbf{0}$, which is what the subsample means estimator does. From an asymptotic perspective, it is harmless to separately estimate the $\boldsymbol{\beta}_g$ even when they are zero. When they are not all zero, estimating them leads to asymptotic efficiency gains.

Theorem 2.4.1 implies that any smooth nonlinear function of $\boldsymbol{\mu}$ is estimated more efficiently using $\hat{\boldsymbol{\mu}}_{FRA}$. For example, in estimating a percentage difference in means, we would be interested in μ_2/μ_1, and using the FRA estimators is asymptotically more efficient than using the SM estimators.

2.4.2 Full RA versus pooled RA

The comparision between FRA and PRA is simple given the expressions in (m2) and (m3) because, as stated earlier, $\mathbf{F}_i$, $\mathbf{K}_i$, and $\mathbf{Q}_i$ are pairwise uncorrelated.

Theorem 2.4.2. *Under the assumptions of theorem 2.3.2 and 2.3.3,*

$$\text{Avar}\left[\sqrt{N}\left(\hat{\boldsymbol{\mu}}_{PRA}-\boldsymbol{\mu}\right)\right]-\text{Avar}\left[\sqrt{N}\left(\hat{\boldsymbol{\mu}}_{FRA}-\boldsymbol{\mu}\right)\right]=\boldsymbol{\Omega}_{\boldsymbol{F}}$$

which is PSD.

Therefore, $\hat{\boldsymbol{\mu}}_{FRA}$ is never less asymptotically efficient than $\hat{\boldsymbol{\mu}}_{PRA}$. There are some special cases where the estimators achieve the same asymptotic variance, the most obvious being when the slopes in the linear projections are homogeneous:

$$\boldsymbol{\beta}_1=\boldsymbol{\beta}_2=\cdots=\boldsymbol{\beta}_G$$

As with comparing FRA with subsample means, there is no gain in efficiency from imposing this restriction when it is true. This is another fact that makes FRA attractive if the sample size is not small.

Other situations where there is no asymptotic efficiency gain in using FRA are more subtle. In general, suppose we are interested in linear combinations $\tau=\mathbf{a}'\boldsymbol{\mu}$ for a given $G\times 1$ vector $\mathbf{a}$. If

$$\mathbf{a}'\boldsymbol{\Omega}_{\mathbf{F}}\mathbf{a}=0$$

then $\mathbf{a}'\hat{\boldsymbol{\mu}}_{PRA}$ is asymptotically as efficient as $\mathbf{a}'\hat{\boldsymbol{\mu}}_{FRA}$ for estimating τ. Generally, the diagonal elements of

$$\boldsymbol{\Omega}_F=\text{E}\left(\mathbf{F}_i\mathbf{F}_i'\right)$$

are

$$\frac{(1-\rho_g)}{\rho_g}\boldsymbol{\delta}_g'\boldsymbol{\Omega}_{\mathbf{X}}\boldsymbol{\delta}_g$$

because $E\left[\left(W_{ig}-\rho_g\right)^2\right]=\rho_g(1-\rho_g)$. The off diagonal terms of $\boldsymbol{\Omega}_{\mathbf{F}}$ are

$$-\boldsymbol{\delta}_g'\boldsymbol{\Omega}_{\mathbf{X}}\boldsymbol{\delta}_h$$

because $E\left[\left(W_{ig}-\rho_g\right)\left(W_{ih}-\rho_h\right)\right]=-\rho_g\rho_h$. Now consider the case covered in Negi and Wooldridge (2019), where $G=2$ and $\mathbf{a}'=(-1,1)$, so the parameter of interest is $\tau=\mu_2-\mu_1$ (the average treatment effect). If $\rho_1=\rho_2=1/2$ then

$$\mathbf{\Omega_F} = \begin{pmatrix} \boldsymbol{\delta}_1' \mathbf{\Omega_X} \boldsymbol{\delta}_1 & -\boldsymbol{\delta}_1' \mathbf{\Omega_X} \boldsymbol{\delta}_2 \\ -\boldsymbol{\delta}_2' \mathbf{\Omega_X} \boldsymbol{\delta}_1 & \boldsymbol{\delta}_2' \mathbf{\Omega_X} \boldsymbol{\delta}_2 \end{pmatrix}.$$

Now $\boldsymbol{\delta}_2 = -\boldsymbol{\delta}_1$ because $\boldsymbol{\delta}_1 = \boldsymbol{\beta}_1 - (\boldsymbol{\beta}_1 + \boldsymbol{\beta}_2)/2 = (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2)/2 = -\boldsymbol{\delta}_2$. Therefore,

$$\mathbf{\Omega_F} = \begin{pmatrix} \boldsymbol{\delta}_1' \mathbf{\Omega_X} \boldsymbol{\delta}_1 & \boldsymbol{\delta}_1' \mathbf{\Omega_X} \boldsymbol{\delta}_1 \\ \boldsymbol{\delta}_1' \mathbf{\Omega_X} \boldsymbol{\delta}_1 & \boldsymbol{\delta}_1' \mathbf{\Omega_X} \boldsymbol{\delta}_1 \end{pmatrix}$$

and

$$\begin{pmatrix} -1 & 1 \end{pmatrix} \begin{pmatrix} \boldsymbol{\delta}_1' \mathbf{\Omega_X} \boldsymbol{\delta}_1 & \boldsymbol{\delta}_1' \mathbf{\Omega_X} \boldsymbol{\delta}_1 \\ \boldsymbol{\delta}_1' \mathbf{\Omega_X} \boldsymbol{\delta}_1 & \boldsymbol{\delta}_1' \mathbf{\Omega_X} \boldsymbol{\delta}_1 \end{pmatrix} \begin{pmatrix} -1 \\ 1 \end{pmatrix} = 0.$$

This finding does not extend to the $G \geq 3$ case when interestingly, it is not true that for estimating each mean separately that PRA is asymptotically equivalent to FRA. So, for example, with lower bound WTP, it might require that bid values have the same frequency. But it is not clear that even that is sufficient.

What about general G with $\rho_g = 1/G$ for all g? Then

$$1 - \rho_g = 1 - \frac{1}{G} = \frac{(G-1)}{G}$$

and so

$$\frac{1 - \rho_g}{\rho_g} = G - 1.$$

Note that

$$\boldsymbol{\delta}_g = \boldsymbol{\beta}_g - \left(\boldsymbol{\beta}_1 + \boldsymbol{\beta}_2 + \cdots + \boldsymbol{\beta}_G\right)/G$$

and it is less clear when there is a degeneracy. Seems very likely for estimating pairwise differences.

2.5 Nonlinear regression adjustment

We now discuss a class of nonlinear regression adjustment methods that preserve consistency without adding additional assumptions (other than weak regularity conditions). In

particular, we extend the setup in Negi and Wooldridge (2019) to allow for more than two treatment levels.

We show that both separate and pooled methods are consistent provided we choose the mean functions and objective functions appropriately. Not surprisingly, using a canonical link function in the context of quasi-maximum likelihood in the linear exponential family plays a key role.

Unlike in the linear case, we can only show that full RA improves over the subsample means estimator when the conditional mean is correctly specified. Whether one can prove efficiency more general is an interesting topic for future research.

2.5.1 Full regression adjustment

We model the conditional means, $E\left[Y(g)|\mathbf{X}\right]$, for each $g = 1, 2, ..., G$. Importantly, we will not assume that the means are correctly specified. As it turns out, to ensure consistency, the mean should have the index form common in the generalized linear models literature. In particular, we use mean functions

$$m(\alpha_g + \mathbf{x}\boldsymbol{\beta_g}),$$

where $m(\cdot)$ is a smooth function defined on $\mathbb{R}$. The range of $m(\cdot)$ is chosen to reflect the nature of $Y(g)$. Given that the nature of $Y(g)$ does not change across g, we choose a common function $m(\cdot)$ across all g. Also, as usual, the vector $\mathbf{X}$ can include nonlinear functions (typically squares, interactions, and so on) of underlying covariates.

As discussed in Negi and Wooldridge (2019) in the binary treatment case, the function $m(\cdot)$ is tied to a specific quasi-log-likelihood function in the linear exponential family (LEF). Table 1 gives the pairs of mean function and quasi-log-likelihood function that ensure consistent estimation. Consistent estimation follows from the results on doubly-robust estimation in the context of missing data in Wooldridge (2007). Each quasi-LLF is tied to the mean function associated with the canonical link function.

Table 2.1: Combinations of means and QLLFs to ensure consistency

Support Restrictions	**Mean Function**	**Quasi-LLF**
None	Linear	Gaussian (Normal)
$Y(g) \in [0,1]$ (binary, fractional)	Logistic	Bernoulli
$Y(g) \in [0,B]$ (count, corners)	Logistic	Binomial
$Y(g) \geq 0$ (count, continuous, corner)	Exponential	Poisson
$Y_j(g) \geq 0$, $\sum_{j=0}^{J} Y_j(g) = 1$	Logistic	Multinomial

The binomial QMLE is rarely applied, but is a good choice for counts with a known upper bound, even if it is individual-specific (so B_i is a positive integer for each i). It can also be applied to corner solution outcomes in the interval $[0, B_i]$ where the outcome is continuous on $(0, B_i)$ but perhaps has mass at zero or B_i. The leading case is $B_i = B$ for all i. Note that we do not recommend a Tobit model in such cases because Tobit is not generally robust to distributional or mean failure. Combining the multinomial QLL and the logistic mean functions is attractive when the outcome is either a multinomial response or more than two shares that necessarily sum to unity.

As discussed in Wooldridge (2007), the key feature of the single outcome combinations in Table 1 is that it is always true that

$$E\left[Y(g)\right] = E\left[m(\alpha_g^* + \mathbf{X}\boldsymbol{\beta}_g^*)\right],$$

where α_g^* and $\boldsymbol{\beta}_g^*$ are the probability limits of the QMLEs whether or not the conditional mean function is correctly specified. The analog also holds for the multinomial logit objective function.

Applying nonlinear RA with multiple treatment levels is straightforward. For treatment level g, after obtaining $\hat{\alpha}_g$, $\hat{\boldsymbol{\beta}}_g$ by quasi-MLE using only units from treatment level g, the mean, μ_g, is estimated as

$$\hat{\mu}_g = N^{-1} \sum_{i=1}^{N} m(\hat{\alpha}_g + \mathbf{X}_i \hat{\boldsymbol{\beta}}_g),$$

which includes linear RA as a special case. This estimator is consistent by a standard application of the uniform law of large numbers; see, for example, Wooldridge (2010) (Chapter 12, question 12.17).

As in the linear case, and of the mean/QLL combinations in Table 1 allow us to write the subsample average as

$$\bar{Y}_g = N_g^{-1} \sum_{i=1}^{N} W_{ig} m(\hat{\alpha}_g + \mathbf{X}_i \hat{\boldsymbol{\beta}}_g).$$

It seems that $\hat{\mu}_g$ should be asymptotically more efficient than $\bar{Y}_g$ because $\hat{\mu}_g$ averages across all of the data rather than just the units at treatment level g. Unfortunately, the proof used in the linear case does not go through in the nonlinear case. At this point, we must be satisfied with consistent estimators of the POs that impose the logical restrictions on $E\left[Y(g)|\mathbf{X}\right]$. In the binary treatment case, Negi and Wooldridge (2019) find nontrivial efficiency gains in using logit, fractional logit, and Poisson regression even compared with full linear RA.

2.5.2 Pooled regression adjustment

In cases where N is not especially large, one might, just as in the linear case, resort to pooled RA. Provided the mean/QLL combinations are chosen as in Table 1, the pooled RA estimator is still consistent under arbitrary misspecification of the mean function. To see why, write the mean function, without an intercept in the index, as

$$m(\gamma_1 w_1 + \gamma_2 w_2 + \cdots + \gamma_G w_G + \boldsymbol{x}\boldsymbol{\beta}).$$

The first-order conditions of the pooled QMLE include the G conditions

$$N^{-1} \sum_{i=1}^{N} W_{ig} \left[Y_i - m(\hat{\gamma}_1 W_{i1} + \hat{\gamma}_2 W_{i2} + \cdots + \hat{\gamma}_G W_{iG} + \mathbf{X}_i \hat{\boldsymbol{\beta}}) \right] = 0, \; g = 1, ..., G.$$

Therefore, assuming no degeneracies, the probability limits of the estimators, denoted with a $*$, solve the population analogs:

$$\mathrm{E}\left(W_g Y\right) = \mathrm{E}\left[W_g Y\left(g\right)\right] = \mathrm{E}\left[W_g m(\mathbf{W}\gamma^* + \boldsymbol{X}\boldsymbol{\beta}^*)\right],$$

where $\mathbf{W} = (W_1, W_2, ..., W_G)$. By random assignment, $\mathrm{E}\left[W_g Y\left(g\right)\right] = \rho_g \mu_g$. By iterated expectations and random assignment,

$$\mathrm{E}\left[W_g m(\mathbf{W}\gamma^* + \mathbf{X}\boldsymbol{\beta}^*)\right] = \mathrm{E}\left\{\mathrm{E}\left[W_g m(\mathbf{W}\gamma^* + \mathbf{X}\boldsymbol{\beta}^*)|\mathbf{X}\right]\right\}$$

and

$$\mathrm{E}\left[W_g m(\mathbf{W}\gamma^* + \mathbf{X}\boldsymbol{\beta}^*)|\mathbf{X}\right] = \mathrm{P}(W_g = 1|\mathbf{X})m(\gamma_g^* + \mathbf{X}\boldsymbol{\beta}^*) = \rho_g m(\gamma_g^* + \mathbf{X}\boldsymbol{\beta}^*).$$

Therefore,

$$\mathrm{E}\left[W_g m(\mathbf{W}\gamma^* + \mathbf{X}\boldsymbol{\beta}^*)\right] = \rho_g \mathrm{E}\left[m(\gamma_g^* + \boldsymbol{X}\boldsymbol{\beta}^*)\right]$$

and, using $\rho_g > 0$, we have shown

$$\mu_g = \mathrm{E}\left[m(\gamma_g^* + \mathbf{X}\boldsymbol{\beta}^*)\right]$$

By definition, $\hat{\gamma}_g$ is consistent for γ_g^* and $\hat{\boldsymbol{\beta}}$ is consistent for $\boldsymbol{\beta}^*$. Therefore, after the pooled QMLE estimation, we obtain the estimated means as

$$\check{\mu}_g = N^{-1}\sum_{i=1}^{N} m(\hat{\gamma}_g + \mathbf{X}_i\hat{\boldsymbol{\beta}}),$$

and these are consistent by application of the uniform law of large numbers.

As in the case of comparing full nonlinear RA to the subsample averages, we have no general asymptotic efficiency results comparing full nonlinear RA to pooled nonlinear RA. As shown in Section 4.2, in the linear case it is never worse, asymptotically, to use full RA.

2.6 Applications

2.6.1 Treatment effects with multiple treatment levels

The most direct application of the previous results is in the context of a randomized intervention with more than two treatment levels. Regression adjustment can be used for any kind of response variable. With a reasonable sample size per treatment level, full regression adjustment is preferred to pooled regression adjustment.

If the outcome $Y(g)$ is restricted in some substantive way, a nonlinear RA method of the kind described in Section 5 can be used to exploit the logical restrictions on $\mathrm{E}\left[Y(g)|\mathbf{X}\right]$. While we cannot show this guarantees efficiency gains compared with using subsample averages, the simulation findings in Negi and Wooldridge (2019) suggest the gains can be nontrivial – even compared with full linear RA.

2.6.2 Difference-in-Differences designs

Difference-in-differences applications can be viewed as a special case of multiple treatment levels. For illustration, consider the standard setting where there is a single before period and a single post treatment period. Let C be the control group and T the treatment group. Label B the before period and A the after period. The standard DID treatment effect is a particular linear combination of the means from the four groups:

$$\tau = (\mu_{TA} - \mu_{TB}) - (\mu_{CA} - \mu_{CB})$$

Estimating the means by separate regression adjustment is generally better than not controlling for covariates, or putting them in additively.

2.6.3 Estimating lower bound mean willingness-to-pay

In the context of contingent valuation, individuals are randomly presented with the price of a new good or tax for a new project. They are asked whether they would purchase the good at the given price, or be in favor of the project at the given tax. Generally, the price or tax is called the "bid value." The outcome for each individual is a binary "vote" (yes = 1, no = 0).

A common approach in CV studies is to estimate a lower bound on the mean willingness-to-pay (WTP). The common estimators are based on the area under the WTP survival function:

$$\mathrm{E}(WTP) = \int_0^\infty S(a)da$$

When a population of individuals is presented with a small number of bid values, it is not possible to identify $\mathrm{E}(WTP)$, but only a lower bound. Specifically, let b_1, b_2, ..., b_G be G bid values and define the binary potential outcomes as

$$Y(g) = 1[WTP > b_g],\ g = 1, ..., G.$$

In other words, if a person is presented with bid value b_g, $Y(g)$ is the binary response, which is assumed to be unity if WTP exceeds the bid value. The connection with the survivor

function is

$$\mu_g \equiv \mathrm{E}\left[Y(g)\right] = P(WTP > b_g) = S(b_g)$$

Notice that μ_g is the proportion of people in the population who have a WTP exceeding b_g. This fits into the potential outcomes setting because each person is presented with only one bid value. Standard consumer theory implies that $\mu_{g+1} \leq \mu_g$, which simply means the demand curve is weakly declining in price.

It can be shown that, with $b_0 \equiv 0$ for notational ease,

$$\tau \equiv \sum_{g=1}^{G}(b_g - b_{g-1})\mu_g \leq \mathrm{E}(WTP),$$

and it is this particular linear combination of $\{\mu_g : g = 1, 2, ..., G\}$ that we are interested in estimating. The so-called ABERS (1955) estimator introduced by Ayer et al. (1955), without a downward sloping survival function imposed, replaces μ_g with its sample analog:

$$\hat{\tau}_{ABERS} = \sum_{g=1}^{G}(b_g - b_{g-1})\bar{Y}_g$$

where

$$\bar{Y}_g = N_g^{-1}\sum_{i=1}^{N} Y_i 1[B_i = b_g]$$

is the fraction of yes votes at bid value b_g. Of course, the $\bar{Y}_g$ can also be obtained as the coefficients from the regression

$$Y_i \text{ on } Bid1_i,\ Bid2_i,\ ...,\ BidG_i,\ i = 1, ..., N$$

Lewbel (2000) and Watanabe (2010) allows for covariates in order to see how WTP changes with individual or family characteristics and attitudes, but here we are interested in estimating τ.

We can apply the previous results on efficiency because τ is a linear combination of the μ_g. Therefore, using separate linear RA to estimate each μ_g, and then forming

$$\hat{\tau}_{FRA} = \sum_{g=1}^{G}(b_g - b_{g-1})\hat{\mu}_g$$

is generally asymptotically more efficient than $\hat{\tau}_{ABERS}$. Moreover, because Y is a binary outcome, we might improve efficiency further by using logit models at each bid value to obtain the $\hat{\mu}_g$.

2.6.4 Application to california oil data

This section applies the linear RA estimators discussed in section 2.3 to survey data from the California Oil Spill study from Carson et al. (2004). The study implemented a CV survey to assess the value of damages to natural resources from future oil spills along California's Central Coast. This was achieved by estimating a lower bound mean WTP measure of the cost of such spills to California's residents. The survey provided respondents with the choice of voting for or against a governmental program that would prevent natural resource injuries to shorelines and wildlife along California's central coast over the next decade. In return, the public would be asked to pay a one time lump sum income tax surcharge for setting up the program.

The main sample survey which was used to elicit the yes or no votes was conducted by Westat, Inc. The data was a random sample of 1,085 interviews conducted with English speaking Californian households where the respondent was 18 years or older, and lived in private residences that were either owned or rented. To address issues of non-representativeness of the interviewed sample from the total initially chosen sample, weights were used. Each respondent was randomly assigned one of five tax amounts: \$5, \$25, \$65, \$120, or \$220 and the binary choice of "yes" or "no" for the oil spill prevention program was recorded at the randomly assigned tax amount.

Apart from the choice at different bid amounts, data was also collected on demographics for the respondent and the respondent's household such as total income, prior knowledge of the spill site, distance to the site, environmental attitudes, attitudes towards big businesses, understanding of the program and the task of voting, beliefs about the oil spill scenario etc.

Table D.1 provides a summary of yes votes at the different bid or tax amounts presented

to the respondents. Table D.2 provides estimates for the PO means as well as the lower bound mean WTP estimate. We see that the FRA estimator delivers more precise estimates for the vector of PO means. Since the treatment effect, which in this case is the lower bound mean willingness to pay for the oil prevention program, is a smooth function of the estimated PO means, we see that the FRA estimate leads to a more precise lower bound mean WTP than the ABERS estimator.

2.7 Monte-carlo

This section reports the finite sample behavior of the three different linear regression adjustment estimators, namely, subsample means (SM), pooled regression adjustment (PRA) and separate slopes (or feasible) regression estimator (FRA) for the vector of PO means. For this monte-carlo study, we generate a population of 1 million observations and mimic the asymptotic setting of random sampling from an "infinite" population. The empirical distributions of the RA estimators are simulated for sample sizes $N \in \{500, 1000, 5000\}$ by randomly drawing the data vector $\{(Y_i, \mathbf{X}_i, \mathbf{W}_i); i = 1, 2, \cdots, N\}$ a thousand times from the above mentioned population. For a comprehensive assessment of the linear RA estimators, we consider three different populations along with four configurations of the treatment assignment vector. Tables D.3, D.5, and D.4 provide bias and standard deviation measures for the vector of PO means estimated using the different estimators for these unique combinations of population models, assignment vector, and sample sizes.

To simulate multiple treatments, we consider potential outcomes, $Y(g)$, corresponding to three different treatment states, $g = 1, 2, 3$. Hence, $G = 3$ for all the simulations. In each of the populations, the treatment vector $\mathbf{W} = (W_1, W_2, W_3)$ is generated with probability mass function defined in the following manner:

$Z = g$	$\mathbb{P}(W_g = 1)$
1	ρ_1
2	ρ_2
3	ρ_3

To generate the vector of assignments from the above distribution, we first draw a uniform random variable $U = \text{Uniform}(0, 1)$ and partition the unit interval $(0, 1)$ into subintervals

$$(0, \rho_1),\ (\rho_1, \rho_1 + \rho_2),\ (\rho_1 + \rho_2, \rho_1 + \rho_2 + \rho_3)$$

and record the interval in which the uniform draw falls. For a particular draw, if $U \in (0, \rho_1)$, then $\mathbf{W} = (1, 0, 0)$. If $U \in (\rho_1, \rho_1 + \rho_2)$, then $\mathbf{W} = (0, 1, 0)$ and finally, if $U \in (\rho_1 + \rho_2, \rho_1 +$

$\rho_2 + \rho_3)$, then $\mathbf{W} = (0, 0, 1)$. This would ensure that the number of treated observations in each treatment group g, on average, will be close to the true assignment probabilities and that each observation (or draw) will belong to only one treatment state i.e., $W_{i1} + W_{i2} + W_{i3} = 1$ for all i. In all the simulations, we consider the following configurations of the assignment vector

$$\boldsymbol{\rho} \in \left\{ \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right), \left(\frac{1}{3}, \frac{1}{6}, \frac{1}{6}\right), \left(\frac{1}{6}, \frac{2}{3}, \frac{1}{6}\right), \left(\frac{1}{5}, \frac{2}{5}, \frac{2}{5}\right) \right\}$$

2.7.1 Population models

To compare the empirical distributions of these linear RA estimators, we consider three different population models. Each model, which we term a data generating process (DGP), assumes that the potential outcomes follow a particular distribution, whether continuous or discrete. In the first two DGP's, $Y(g)'s$ are simulated to be continuous non-negative outcomes. The first model uses an exponential distribution whereas the second uses a a mixture of an exponential and log-normal distribution. The third DGP takes $Y(g)$ to be categorical responses which take four discrete values. Each DGP's is described in detail below.

For the first two DGP's, we consider two covariates, $\mathbf{X} = (X_1, X_2)$, which are drawn from a bivariate normal distribution as follows

$$\mathbf{X}' = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N \left[\begin{pmatrix} 1 \\ 2 \end{pmatrix}, \begin{pmatrix} 2 & 0.5 \\ 0.5 & 3 \end{pmatrix} \right]$$

For each DGP, we choose parameters such that covariates have some predictive power in explaining the potential outcomes, so that the benefits of regression adjustment can be reaped.

Population 1: For each g,

$$Y(g) \sim \text{Exponential}(\lambda_g)$$
$$\lambda_g = \exp\Big(\gamma_{0g} + \gamma_{1g} \cdot X_1 + \gamma_{2g} \cdot X_2 + R(g)\Big)$$

where $R(g)|X_1, X_2 \sim N(0, \sigma_g^2)$ and $R(1), R(2)$, and $R(3)$ are allowed to be correlated.[1] The parameter vector, $\boldsymbol{\gamma}_g = \Big(\gamma_{0g}, \gamma_{1g}, \gamma_{2g}\Big)'$, and variance σ_g^2 for g=1,2,3 are parameterized as follows

$$\boldsymbol{\gamma}_1 = (0, -1, 1)' \ , \sigma_1^2 = 0.04$$
$$\boldsymbol{\gamma}_2 = (1, 1.62, -0.5)' \ , \sigma_2^2 = 1$$
$$\boldsymbol{\gamma}_3 = (2, -2, 0.6)' \ , \sigma_3^2 = 0.01$$

For this configuration of parameters, the covariates are only mildly predictive of the outcomes in the three treatment groups, $R_1^2 = 0.04, R_2^2 = 0.02$, and $R_3^2 = 0.01$.

Population 2: In this case, we generate the outcomes to be a mixture between exponential and lognormal distributions,

$$Y(g) \sim \begin{cases} \text{Exponential}(\lambda_g) \text{ if } 0 \leq V < \delta_g \\ \text{Lognormal}(\eta_g, \nu_g^2) \text{ if } \delta_g \leq V \leq 1 \end{cases}$$
$$\eta_g = \alpha_{0g} + \alpha_{1g} \cdot X_1 + \alpha_{2g} \cdot X_2 + \alpha_{3g} \cdot X_1^2 + \alpha_{4g} \cdot X_2^2 + \alpha_{5g} \cdot X_1 \cdot X_2 + K(g)$$

where the mean λ_g is defined exactly as above. Also, $K(g)|X_1, X_2 \sim N(0, \kappa_g^2)$ and $K(1), K(2)$, and $K(3)$ are also allowed to be correlated.

[1] These are simulated to be affine transformations of the same standard normal random variable.

The other parameters $\boldsymbol{\alpha}_g$, δ_g, κ^2, and ν_g^2 are chosen as follows

$$\boldsymbol{\alpha}_1 = \left(\frac{1}{15}, \frac{3}{15}, \frac{-1}{15}, \frac{0.05}{15}, \frac{-0.02}{15}, \frac{0.1}{15}\right) \text{ , } \delta_1 = 0.7 \text{ , } \kappa_1^2 = \frac{0.04}{225} \text{ , } \nu_1^2 = 0.01$$
$$\boldsymbol{\alpha}_2 = \left(\frac{1.2}{15}, \frac{2}{15}, \frac{2}{15}, \frac{0.03}{15}, \frac{-0.02}{15}, \frac{0.5}{15}\right) \text{ , } \delta_2 = 0.5 \text{ , } \kappa_2^2 = \frac{0.09}{225} \text{ , } \nu_2^2 = 0.16$$
$$\boldsymbol{\alpha}_3 = \left(\frac{0.3}{10}, \frac{1}{10}, \frac{1.5}{10}, \frac{0.13}{10}, 0, \frac{0.15}{10}\right) \text{ , } \delta_3 = 0.3 \text{ , } \kappa_3^2 = \frac{0.16}{100} \text{ , } \nu_3^2 = 0.36$$

For this DGP, the population R-squared for the three treatment groups are $R_1^2 = 0.119$, $R_2^2 = 0.1570$, and $R_3^2 = 0.1177$ respectively.

Finally, for the third population model, we consider each potential outcome to be categorical response which is generated using a multinomial logit model. For this setting we only consider a single covariate, X, which is assumed to be distributed Poisson

$$X \sim Poisson(14)$$

As an example, one could imagine the treatment to be three different political advertisements that are shown to a voter and the response (or outcome) indicates the voter's preferred candidate amongst four potential choices with X denoting the voter's years of schooling.

Population 3: Let $Y(g)$ take one of four discrete values,, $j \in \{1, 2, 3, 4\}$, for each $g = \{1, 2, 3\}$, say,

$$\mathbb{P}\{Y(g) = j\} = \frac{\exp\left(\omega_{1gj} \cdot X + \omega_{2gj} \cdot X^2 + R_j(g)\right)}{\sum_{j=1}^{4} \exp\left(\omega_{1gj} \cdot X + \omega_{2gj} \cdot X^2 + R_j(g)\right)}$$

where $R_j(g)|X \sim U(0, \sigma_g^2)$. For notational simplicity, we collect all the index parameters in $\boldsymbol{\omega}_{1g} = (\omega_{1g1}, \omega_{1g2}, \omega_{1g3}, \omega_{1g4})'$ and $\boldsymbol{\omega}_{2g} = (\omega_{2g1}, \omega_{2g2}, \omega_{2g3}, \omega_{2g4})'$. For these we picked the

following values,

$$\boldsymbol{\omega}_{11} = (-0.1291, -0.1014, -0.7041, -0.7798)' \,, \quad \boldsymbol{\omega}_{21} = (-0.0108, -0.0234, -0.0376, -0.0192)'$$

$$\boldsymbol{\omega}_{12} = (0.7866, 0.1804, 0.6310, 0.9695)' \,, \quad \boldsymbol{\omega}_{22} = (0, 0, 0, 0)'$$

$$\boldsymbol{\omega}_{13} = (0.3271, 0.2005, 0.4048, 0.3930)' \,, \quad \boldsymbol{\omega}_{23} = (0.308, 0.0411, 0.0301, 0.0475)'$$

$$\sigma_1^2 = 1, \sigma_2^2 = 0.04, \sigma_3^2 = 0.1\bar{1}$$

Given these choices, the population R-squared's are $R_1^2 = 0.0859$, $R_2^2 = 0.0319$, and $R_3^2 = 0.1048$.

In all the three cases, while estimating the PO means, we assume that the above functional forms are unknown and simply run the regression of the observed outcome on a constant, and the covariates. This is meant to reflect the uncertainty that researchers often have about the underlying outcome distributions and how they are generated. Considering three different environments in which to compare the performance of the linear RA estimators also helps to mimic the variety of experimental settings that researchers may encounter where separate slopes regression adjustment produces substantial precision gains.

2.8 Discussion

Tables D.3, D.5, and D.4 below report the bias and standard deviation of SM, PRA, and FRA estimators for the three different DGP's respectively. Each table reports these measures across four assignment vectors that were chosen in the manner described above. Note that in most cases, the bias of FRA and PRA estimates is comparable and sometimes even smaller than its SM counterpart. However, one may be willing to forego the bias in RA estimates in favor of efficiency, in which case we turn our attention to the standard deviation measures on these estimates.

Across all four configurations, we see that the standard deviation of the separate slopes estimator is weakly smaller than that of the subsample means and pooled regression esti-

mators. The comparison of PRA and SM estimators is less unequivocal since in almost all cases and for all sample sizes, the PRA estimator for the first PO mean is almost always less precise than the SM counterpart.

For DGP 2 and 3, we see a similar pattern as for DGP 1. In all cases, PRA produces estimates that may or may not be more precise than the subsample means estimator. Note that some of the means are estimated more precisely with PRA than the others. However, the comparison between SM and FRA is less ambiguous. We always find all means estimated through FRA to be weakly more precise than those estimated using just the difference in subsample means.

2.9 Conclusion

In this paper, we build on the work of Negi and Wooldridge (2019) to study efficiency improvements in linear regression adjustment estimators when there are more than two treatment levels. In particular, we consider any arbitrary 'G' number of treatments when these treatments have been randomly assigned. We show that jointly estimating the vector of potential outcome means using linear RA that allows for separate slopes for the different assignment levels is asymptotically never worse than just using subsample averages. One case when there is no gain in asymptotic efficiency from using FRA is when the slopes are all zero. In other words, when the covariates are not predictive of the potential outcomes, then using separate slopes does not produce more precise estimates compared to just estimating the subsample averages. We also show that separate slopes RA is generally more efficient compared to pooled RA, unless the true linear projection slopes are homogeneous. In this case using FRA to estimate the vector of PO means is harmless. In other words, using FRA under this scenario does not hurt.

In addition, this paper also extends the discussion around nonlinear regression adjustment made in Negi and Wooldridge (2019) to more than two treatment levels. In particular, we show that pooled and separate nonlinear RA estimators in the quasi maximum likelihood

family are consistent if one chooses the mean and objective functions appropriately from the linear exponential family of distributions.

As an illustration of these efficiency arguments, we apply the different linear RA estimators for estimating the lower bound mean WTP using data from a contingent valuation study undertaken to provide an ex-ante measure of damages to natural resources from future oil spills along California's central coast. We find that the lower bound mean WTP is estimated more efficiently when we allow the slopes on the different bid values to be estimated separately as opposed to the ABERS estimator, which uses subsample averages for the PO means. A comprehensive simulation study also offers finite sample evidence on efficiency improvements with FRA over SM under three different empirical settings. We find FRA estimator of PO means to be unequivocally more precise than PRA and weakly better than SM for all data generating processes despite the covariates being only mildly predictive of the potential outcomes.

CHAPTER 3

DOUBLY WEIGHTED M-ESTIMATION FOR NONRANDOM ASSIGNMENT AND MISSING OUTCOMES[†]

3.1 Introduction

Much of the applied literature in economics is interested in questions of causal inference, such as measuring the impact of job training on labor market outcomes (Calónico and Smith (2017), Ba et al. (2017), Card et al. (2011)), determining the efficacy of school voucher programs on student achievement (Muralidharan and Sundararaman (2015)), and even, estimating the effects of firm competition on prices (Busso and Galiani (2019)). A key concern with causal effects estimation is that, typically, the units under comparison are different even before the treatment is assigned, rendering the task of drawing causal claims difficult. This task is made even more challenging when there is missing data on the outcome of interest, such as earnings, test scores, or prices.

The econometrics literature has proposed weighting to deal with non-random assignment (Hahn (1998), Hirano and Imbens (2001), Hirano et al. (2003), Firpo (2007)) and missing data (Robins and Rotnitzky (1995), Robins et al. (1994), Wooldridge (2002), Wooldridge (2007)).[1] However, the two weighting procedures have typically been studied in isolation.[2] This paper proposes a double inverse probability weighted (IPW) estimator that addresses these twin identification issues in a general M-estimation framework. Specific examples

[†]This work is unpublished.

[1]See Li et al. (2013) for a review of IPW approaches to deal with missing data under a variety of missing data patterns.

[2]Huber (2014b) studies treatment effects in the presence of the double selection problem using a nested weighting procedure. He considers the traditional problem of sample selection based on unobservables and uses a nested weighting structure, which includes the first stage sample selection probability as a covariate in the second stage propensity score model. Other papers that point or set identify causal parameters in the presence of the double selection problem include Fricke et al. (2015), Frölich and Huber (2014), Vossmeyer (2016), Mattei et al. (2014) and Huber and Mellace (2015).

include linear regression, maximum likelihood (MLE), and quantile regression (QR).

In particular, consider a prototypical training program. Learning about the effects of such an intervention on (say) earnings, necessitates comparing individuals based on their participation status. If these individuals are not randomly assigned to the program, such a comparison will confound the true training effect with factors that simultaneously determine selection into the program and future earnings. For instance, individuals with poor labor market histories may be more likely to participate, and contemporaneously, have lower earnings than nonparticipants. Hence, the true effect of the training program is not identified in the presence of nonrandom participation. This identification problem is compounded, if say, individuals who participate in the program are also likely to drop out, introducing the additional problem of missing outcomes.

Even in randomized experiments, the problem of missing outcomes can arise due to attrition, no-shows, dropouts, or non-response (see Bloom (1984), Heckman et al. (1998b), and Hausman and Wise (1979) for a discussion). A specific example is the National Supported Work (NSW) program, where 19 percent of the randomized sample attrited between the baseline and first round of follow-up interviews. In this case, the standard simple difference in means estimator will no longer produce an unbiased training effect estimate (see Huber (2012), Huber (2014a), Behaghel et al. (2015), Frumento et al. (2012) for alternative approaches of dealing with various post-randomization complications).

A common empirical strategy for dealing with missing data is to drop individuals with incomplete information, and treat the observed units as a random sample from the population of interest.[3] In a setting with only missing outcomes, such a strategy will not only waste potentially useful information on covariates, but more importantly create a non-random sample for estimation. In turn, this can generally lead to inconsistent treatment effect

[3]For example, Chen et al. (2018) drop observations with missing labor market outcomes for week 208 after random assignment using the National Job Corps Study data to derive bounds on the Average Treatment Effect as well as Average Treatment Effect on the Treated. Drange and Havnes (2018) also report excluding children with missing data on the outcomes to study the effect of early child care on cognitive development in Norway.

estimates.

One of the main contributions of this paper is to propose a new class of consistent and asymptotically normal estimators that combine propensity score weighting with weighting for missing data, to address the problems of nonrandom assignment and missing outcomes. Traditionally, the weighting literature has studied each problem individually. By studying them together, this paper builds upon and extends the existing weighting literature to incorporate both issues simultaneously. A second contribution is to consider a general M-estimation problem, which is permitted to be non-smooth in the underlying parameters. Therefore, the identification and estimation arguments made in this paper encompass both average treatment effect (ATE) and quantile treatment effect (QTE) parameters. Finally, a key feature of the proposed estimator is its *robustness* to parametric misspecification of a conditional model of interest (such as a conditional mean or conditional quantile) and the two weighting functions.

To obtain consistent estimation of causal parameters, this paper assumes that selection into treatment is based on observed covariates.[4] Put differently, this restriction implies that the treatment is as good as randomly assigned after conditioning on pre-treatment covariates.

Previous studies have found several situations where such an assumption is tenable, especially when pre-treatment values of the outcome variable are available. For example, LaLonde (1986) and Hotz et al. (2006) have shown that controlling for pre-training earnings alone reduces significant bias between non-experimental and experimental estimates. The literature assessing teacher impact on student achievement has reported similar findings with pre-test scores (Chetty et al. (2014), Kane and Staiger (2008) and Shadish et al. (2008)), indicating the plausibility of unconfoundedness in these settings.

This paper also assumes that the missing outcomes mechanism is ignorable once we con-

[4]This is a widely used assumption in the treatment effects literature (Imbens and Wooldridge (2009)) and is known by a variety of names such as unconfoundedness, exogenous assignment (exogeneity), ignorability of assignment, selection on observables and conditional independence assumption (CIA).

dition on covariates and the treatment status.[5] In other words, covariates and the treatment are sufficient for predicting observation into the sample (see Wooldridge (2007) for a similar ignorability assumption). This mechanism falls under the "Missing at Random" (MAR) or the "selection on observables" label used in the econometrics literature (for example, Moffit et al. (1999) use it to model attrition) and allows for differential non-response, attrition, and even non-compliance to the extent that conditioning variables predict it.[6]

Under unconfoundedness and ignorability, the proposed strategy leads to an estimation method that follows in two steps; the first step estimates the treatment and missing outcome probabilities using binary response MLE.[7] The second step uses these estimated probabilities as weights to minimize (or maximize) a general objective function. Given the parametric nature of the first and second steps, this paper highlights a *robustness* property which allows the estimator to remain consistent for a parameter of interest, under misspecification of either the conditional model or the two probabilities. Consequently, the asymptotic theory in this paper distinguishes between these two important halves. The first half focuses on misspecification of either a conditional expectation function (CEF) or a conditional quantile function (CQF), whereas the second half considers arbitrary misspecification of the weighting functions. Delineating the two cases helps to clarify the interpretation on causal estimands in different misspecification scenarios. This property also nests the well known result of '*double robustness*' (Słoczyński and Wooldridge (2018)) as a special case.

As illustrative examples, the paper discusses robust estimation of two specific causal parameters, namely, the ATE and QTEs. Consistent estimation of the ATE is achievable under both misspecification scenarios. Of particular interest is the case when the conditional mean function is misspecified. In this case, consistent estimation of ATE relies on double

[5]Typically, covariates also include pre-treatment outcomes like pre-training earnings or pre-test scores.

[6]Attrition in a two period panel is allowed as long as it is a function of key time-invariant characteristics and the assigned treatment status.

[7]As a practical matter, researchers typically follow the convention of estimating these probabilities as flexible logit functions.

weighting and results from the generalized linear model literature. For estimation of quantile treatment effects, the paper considers three different parameters, namely, conditional quantile treatment effect (CQTE), a linear approximation to CQTE, and the unconditional quantile treatment effect (UQTE), each of which may be of interest to the researcher depending on whether features of the conditional or unconditional outcomes distribution are of particular interest. In the event that the underlying CQF is assumed to be correct, the double-weighted estimator is shown to be consistent for the true CQTE, otherwise, delivers a consistent weighted linear approximation to the true CQTE (using results from Angrist et al. (2006b)). In addition, the paper underscores the importance of double weighting for a parameter like UQTE, where covariates, which serve to remove biases due to nonrandom assignment and missing outcomes, enter the estimating equation only through the two probability models. Simulations show that the doubly weighted ATE and QTE estimates have the lowest finite sample bias compared to alternatives that ignore one or both problems (such as the unweighted estimator that drops data, or the propensity score weighted estimator which weights only by the treatment probability).

Finally, the proposed method is applied to estimate average and distributional impacts of the NSW training program on earnings for the Aid to Families with Dependent Children (AFDC) target group. This sample is obtained from Calónico and Smith (2017), who recreate Lalonde's within-study analysis for the AFDC women. To have missing cases, these data are augmented to include women with missing earnings information in 1979 that were originally dropped from Calónico and Smith's analysis. This empirical application helps to quantify the estimated bias in the unweighted and propensity score weighted estimates, relative to the doubly weighted estimates, through the presence of an experimental benchmark. Results show that the doubly-weighted estimates have an estimated bias which is smaller than that computed for the unweighted estimates, but comparable in magnitude to the bias estimated for the single (propensity score) weighted estimates. This finding indicates that, for this particular application, the missing outcomes problem seems to be much less conse-

quential than the nonrandom assignment problem in obtaining estimates close to the true experimental benchmark.

The rest of this paper is structured as follows. Section 3.2 describes the framework and provides a short description of the population models with an introduction to the naive unweighted estimator. Section 3.3 discusses estimation of the probability weights which is a necessary first step in solving the doubly weighted problem. Section 3.4 develops the first half of the asymptotic theory which is explicitly focused on misspecification of a conditional model of interest. In contrast, section 3.5 discusses the second half which considers cases where the conditional model of interest is correctly specified. Section 3.6 studies the specifics of the robustness property for estimating ATE and QTEs in rigorous detail. It also provides supporting Monte Carlo evidence under different cases of misspecification. Section 3.7 illustrates the performance of double weighting using Calónico and Smith (2017) data. Section 3.8 concludes with a direction for future research. Tables, figures, proofs, and some auxiliary results are provided in the appendix.

3.2 Doubly weighted framework

3.2.1 Potential outcomes and the population models

Let $y(g)$ denote the potential outcome for $g = 0, 1$ and let w_g be an indicator variable for treatment level g where $w_0 + w_1 = 1$, implying that the two treatment groups are mutually exclusive and exhaustive. Then,

$$y = \ y(0) \cdot w_0 + y(1) \cdot w_1 \tag{3.1}$$

Let $(y(g), \mathbf{x})$ denote a $M \times 1$ random vector taking values in $\mathfrak{R}^M$ where $\mathbf{x}$ is the vector of pre-treatment characteristics.[8] Some feature of the distribution of $(y(g), \mathbf{x})$ is assumed

[8]For instance, in the NSW program, $y(1)$ and $y(0)$ will denote potential earnings in the event of participation and non-participation in the training program respectively. The covariates on which information was collected in the baseline period included individual's age, ethnicity, high-school dropout status, real earnings along with other socio-economic and demographic characteristics.

to depend on a finite $P_g \times 1$ vector $\boldsymbol{\theta_g}$, contained in a parameter space $\boldsymbol{\Theta_g} \subset \Re^{Pg}$.[9] Let $\mathbb{D}(y(g)|\mathbf{x})$ denote the conditional distribution of $y(g)$ and $\mathbf{x}$ and let $q(y(g), \mathbf{x}, \boldsymbol{\theta_g})$ be an objective function depending on $y(g)$, $\mathbf{x}$ and $\boldsymbol{\theta_g}$. This paper allows $q(\cdot)$ to be a smooth or a non-smooth function of the underlying parameter, $\boldsymbol{\theta_g}$. The parameter of interest, denoted by $\boldsymbol{\theta_g^0}$, is defined to be the solution to the following population problem

Assumption 3.2.1 (Identification of $\boldsymbol{\theta_g^0}$)**.** *The parameter vector $\boldsymbol{\theta_g^0} \in \boldsymbol{\Theta_g}$ is the unique solution to the population minimization problem*

$$\min_{\boldsymbol{\theta_g} \in \boldsymbol{\Theta_g}} \mathbb{E}\left[q(y(g), \mathbf{x}, \boldsymbol{\theta_g})\right], \ g = 0, 1 \tag{3.2}$$

Notice that assumption 3.2.1 describes a general M-estimation framework where the interest lies in minimizing some objective function, $q(y(g), \mathbf{x}, \boldsymbol{\theta_g})$. Specific examples include the smooth ordinary least squares objective function, $q(y(g), \mathbf{x}, \boldsymbol{\theta_g}) = (y(g) - \alpha_g - \mathbf{x}\boldsymbol{\beta_g})^2$ or the non-smooth check function of Koenker and Bassett (1978), $q(y(g), \mathbf{x}, \boldsymbol{\theta_g}) = c_\tau(y(g) - \alpha_g - \mathbf{x}\boldsymbol{\beta_g})$ where $\boldsymbol{\theta_g} \equiv (\alpha_g, \boldsymbol{\beta_g})'$.[10] Other examples of $q(\cdot)$ include log-likelihood and quasi-log likelihood (QLL) functions.

An implicit point made in the assumption above is that $\boldsymbol{\theta_g^0}$ is not assumed to be correctly specified for a conditional feature like a conditional mean, conditional variance or even the full conditional distribution. Assumption 3.2.1 simply requires $\boldsymbol{\theta_g^0}$ to *uniquely* minimize the population problem in (3.2). If $\boldsymbol{\theta_g^0}$ is correctly specified for any of the above mentioned quantities, then the parameter is of direct interest to researchers. However, if $\boldsymbol{\theta_g^0}$ is misspecified for any of these distributional features, assumption 3.2.1 guarantees a unique pseudo true solution, $\boldsymbol{\theta_g^*}$ (White (1982)). In the case of misspecification, determining whether $\boldsymbol{\theta_g^*}$ is meaningful will depend on the conditional feature being studied and the estimation method

[9] For generality, the dimension of $\boldsymbol{\theta_g}$ is allowed to be different for the treatment and control group problems and is also different than the dimension of $\mathbf{x}$, where $\mathbf{x} \in \mathfrak{X} \subset \Re^{\dim(\mathfrak{X})}$

[10] For a random variable u, $c_\tau(u) = (\tau - \mathbb{1}_{[u<0]})u$, is the asymmetric loss function for estimating quantiles.

used. For example, in the OLS case, $\boldsymbol{\theta_g^0}$ will still index a linear projection if one is agnostic about linearity of the CEF. Angrist et al. (2006b) establish analogous approximation properties for quantiles, where a misspecified CQF can still provide the best weighted mean square approximation to the true τ-CQF. In other words, they show that $\boldsymbol{\theta_g^0}$ solves the following weighted mean square error loss function

$$\min_{\boldsymbol{\theta_g} \in \boldsymbol{\Theta_g}} \left\{ \mathbb{E}\left[\omega_\tau(\mathbf{x}, \boldsymbol{\theta_g}) \cdot (\alpha_g(\tau) + \mathbf{x}\boldsymbol{\beta_g}(\tau) - Quant_\tau(y(g)|\mathbf{x}))^2 \right] \right\}$$

where $\omega_\tau(\mathbf{x}, \boldsymbol{\theta_g}) = \int_0^1 (1-u) f_{y(g)}(u \cdot \mathbf{x}\boldsymbol{\theta_g} + (1-u) \cdot Quant_\tau(y(g)|\mathbf{x})|\mathbf{x})$ is the weighting function given in Angrist et al. (2006b) adapted to the potential outcomes framework, $Quant_\tau(y(g)|\mathbf{x})$ is the true CQF and $\alpha_g^0(\tau) + \mathbf{x}\boldsymbol{\beta_g^0}(\tau)$ represents a weighted linear approximation. Hence, in this case, $\boldsymbol{\theta_g^0} \equiv (\alpha_g^0, \boldsymbol{\beta_g^0})'$ provides an interesting interpretation that can be of practical interest to researchers.

Note that assumption 3.2.1 only requires the parameter to solve an unconditional problem. A sufficient condition for the same is that the parameter additionally solves the conditional problem. However, this latter condition will not be required to derive the asymptotic theory in section 3.4. For the reader, an effective way to separate the results in section 3.4 from the ones discussed in section 3.5 is to consider the current section as allowing potential misspecification of the conditional feature being studied in the sense of assumption 3.2.1. Section 3.5 will require $\boldsymbol{\theta_g^0}$ to be identified in the stronger conditional sense. Together, results developed in section 3.4 and section 3.5 can then be used to characterize the robustness property of the proposed estimator.

3.2.2 The unweighted M-estimator

In this paper, the objective is to consistently estimate $\boldsymbol{\theta_g^0}$. If one obtains a random sample on $\{(y_i(0), y_i(1), w_{ig}, \mathbf{x}_i) : i = 1, 2, \ldots, N\}$ from the population of interest, then one can solve

$$\min_{\boldsymbol{\theta_g} \in \boldsymbol{\Theta_g}} \sum_{i=1}^{N} w_{ig} \cdot q(y_i(g), \mathbf{x}_i, \boldsymbol{\theta_g}) \tag{3.3}$$

For the estimator, which solves (3.3), to consistently estimate $\boldsymbol{\theta_g^0}$, the reverse analogy principle dictates that $\boldsymbol{\theta_g^0}$ must also solve,

$$\min_{\boldsymbol{\theta_g} \in \boldsymbol{\Theta_g}} \mathbb{E}\left[w_g \cdot q(y(g), \mathbf{x}, \boldsymbol{\theta_g})\right]; \; g = 0, 1 \tag{3.4}$$

However, this argument may not necessarily hold. For example, consider the linear model

$$\begin{aligned} &y(g) = \alpha_g + \mathbf{x}\boldsymbol{\beta_g} + u(g); \quad g = 0, 1 \\ &\mathbb{E}\left(u(g)\right) = 0, \mathbb{E}\left(\mathbf{x}'u(g)\right) = \mathbf{0} \end{aligned} \tag{3.5}$$

If the treatment (say, job training) is correlated with baseline characteristics, as can be expected when the program is non-randomly assigned, then $\mathbb{E}\left[w_g \cdot \mathbf{x}'u(g)\right] \neq \mathbf{0}$.[11] In addition, suppose there is missing data on the outcome of interest. To formalize this, let 's' be a binary indicator for missing outcomes, then

$$y = \begin{cases} y(0), \text{ if } g = 0, \; s = 1 \\ y(1), \text{ if } g = 1, \; s = 1 \\ \text{missing, if } \; s = 0 \end{cases} \tag{3.6}$$

where $s = 1$ if the outcome is observed and $s = 0$ if it is missing.[12] In this case, a common empirical strategy is to solve

$$\min_{\boldsymbol{\theta_g} \in \boldsymbol{\Theta_g}} \sum_{i=1}^{N} s_i \cdot w_{ig} \cdot q(y_i(g), \mathbf{x}_i, \boldsymbol{\theta_g}) \tag{3.7}$$

which only uses observed data to estimate $\boldsymbol{\theta_g^0}$. Let us refer to the estimator that solves 3.7 as the unweighted M-estimator, and denote it as $\boldsymbol{\hat{\theta}_g^u}$. In this case, even if the treatment is randomly assigned, the missing outcomes may still be correlated with the treatment, observable factors or both, which implies that $\mathbb{E}\left[s \cdot w_g \cdot \mathbf{x}'u(g)\right] \neq \mathbf{0}$. Hence, identification of $\boldsymbol{\theta_g^0}$ is now confounded on two grounds; non-random assignment which renders the treatment and

[11] When the treatment is randomized, as in the case of NSW, or as studied in Negi and Wooldridge (2019), one will necessarily have $\mathbb{E}\left[w_g \cdot \mathbf{x}'u(g)\right] = \mathbf{0}$, due to the experimental design.

[12] For an illustration of the observed sample, see figure I.1.

control groups incomparable and missing outcomes which leads to violation of the 'random sampling' assumption. The next section discusses the identification approach taken in this paper.

3.2.3 Ignorable missingness and unconfoundedness

Without imposing any structure on the assignment and missingness mechanism in the population, identifying and estimating $\boldsymbol{\theta_g^0}$ remains difficult because of the argument outlined in the previous section. To proceed with identification, I assume that the treatment is unconfounded on covariates (Rosenbaum and Rubin (1983)).[13] Formally, consider the following

Assumption 3.2.2. *(Strong ignorability) Assume,*

$$\{y(0), y(1) \perp\!\!\!\perp w_g\} / \mathbf{x} \tag{3.8}$$

1. *(3.8) implies that* $\mathbb{P}(w_g = 1|y(0), y(1), \mathbf{x}) = \mathbb{P}(w_g = 1|\mathbf{x}) \equiv p_g(\mathbf{x})$ *for* $g = 0, 1$*, where*

$$p_0(\mathbf{x}) + p_1(\mathbf{x}) = 1$$

2. *The vector of pre-treatment covariates,* $\mathbf{x}$*, is always observed for the entire sample.*

3. *For all* $\mathbf{x} \in \mathfrak{X} \subset \mathfrak{R}^{\dim(\mathfrak{X})}$*,* $p_g(\mathbf{x}) > \kappa_g > 0$

Assumption 3.2.2 part (1) says that conditioning on covariates is enough to parse out any systematic differences that may exist between the treatment and control groups. This is a widely used assumption in the treatment effects literature, and is known as unconfoundedness.[14] One advantage of unconfoundedness is that, intuitively, it has a better chance

[13] Like most other assumptions, unconfoundedness is non-refutable. For methods that indirectly test for its validity, see Huber and Melly (2015), de Luna and Johansson (2014), Rosenbaum (1987) and Heckman and Hotz (1989).

[14] Imbens and Wooldridge (2009) attribute the popularity of unconfoundedness, as an identifying restriction, to the paucity of general methods for estimating treatment effects.

of holding once we control for a rich set of variables in $\mathbf{x}$.[15] Note that unconfoundedness not only includes cases where the treatment is a deterministic function of the covariates, for example stratified (or block) experiments, but also cases where the treatment is a stochastic function of covariates.[16] Part (2) requires that we observe these covariates for all individuals. Part (3) is an overlap assumption which ensures that for all values of $\mathbf{x}$ in the support of the distribution, there is a chance of observing units in both treatment and control states.[17]

With respect to the missing outcomes mechanism, I assume ignorability conditional on covariates and the treatment status. Formally, consider

Assumption 3.2.3. *(Ignorability of missing outcomes) Assume,*

$$\{y(0), y(1) \perp\!\!\!\perp s\} / \mathbf{x}, w_g \tag{3.9}$$

1. *(3.9) implies that* $\mathbb{P}(s = 1|y(0), y(1), \mathbf{x}, w_g) = \mathbb{P}(s = 1|\mathbf{x}, w_g) \equiv r(\mathbf{x}, w_g)$

2. *In addition to* $\mathbf{x}$, w_g *is always observed for the entire sample.*

3. *For each* $(\mathbf{x}, w_g) \in \mathfrak{R}^{\dim(\mathfrak{X})+1}$, $r(\mathbf{x}, w_g) > \eta > 0$

Part (1) states that conditional on covariates and the treatment status, the individuals whose outcomes are missing do not differ systematically from those who are observed. This implies that adjusting for $\mathbf{x}$ and w_g renders the outcomes are as good as randomly missing. In the econometrics literature, this assumption falls under the "selection on observables" tag. In the statistics literature, this is also known as MAR and represents a scenario where missingness only depends on observables and not on the missing values of the variable (Little and Rubin (2002)). Special cases covered under this mechanism are patterns such as missing

[15] For example, Hirano and Imbens (2001) control for a rich set of prognostic factors to justify unconfoundedness while estimating the effects of right heart catheterization (RHC) on survival rates of patients.

[16] The appendix discusses the case of a stratified experiment where unconfoundedness is satisfied by design if one additionally assumes the missing outcome pattern to be ignorable.

[17] Methods for checking overlap involve calculating normalized sample average differences for each covariate and checking the empirical distribution of propensity scores.

completely at random (MCAR) and exogenous missingness, as considered in Wooldridge (2007), with $\mathbf{z} = \mathbf{x}$. Allowing the missingness probability to be a function of the treatment indicator is particularly useful in cases of differential nonresponse. For instance, in NSW, people assigned to the treatment group were less likely to drop out of the program compared to the control group. In such cases, covariates alone may not be sufficient for predicting missingness. To the extent that being observed in the sample is predicted by $\mathbf{x}$ and w_g, assumption 3.2.3 can accommodate non-observability due to sampling design, item non-response and attrition in a two period panel.[18]

Part (2) of the above assumption ensures that $\mathbf{x}$ and w_g are fully observed. Finally, part (3) imposes an overlap condition, where the probability of being observed in bounded away from zero. This implies that there is a positive probability of observing people in the sample with a given value of $\mathbf{x}$ and w_g in the population.

To study the estimation method in terms of the selected sample, I consider random sampling in the following sense,

Assumption 3.2.4. *(Sampling) Assume that* $\{(y_i(0), y_i(1), \mathbf{x}_i, w_{ig}, s_i);\ i = 1, 2, \ldots, N\}$ *are independent and identical random draws from the population where in the population*

1. w_{ig} *is unconfounded with respect to* $\{y_i(0), y_i(1)\}$ *given* $\mathbf{x}_i$

2. s_i *is ignorable with respect to* $\{y_i(0), y_i(1)\}$ *given* $(\mathbf{x}_i, w_{ig})$

The next section discusses identification and estimation of $\boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}}$ using a double inverse probability weighted procedure.

3.2.4 Population problem with double weighting

Consider the following population problem

[18]For the case of attrition, one must assume that second period missingness is ignorable conditional on initial period covariates and the treatment status.

$$\min_{\boldsymbol{\theta_g}\in\boldsymbol{\Theta_g}} \mathbb{E}\left[\frac{s}{r(\mathbf{x}, w_g)}\cdot\frac{w_g}{p_g(\mathbf{x})}\cdot q(y(g),\mathbf{x},\boldsymbol{\theta_g})\right]; \quad g=0,1 \tag{3.10}$$

then under unconfoundedness and ignorability, solving this doubly weighted population problem is the same as solving 3.2. The following lemma establishes this equivalence

Lemma 3.2.5. *Given assumptions 3.2.1, 3.2.2, 3.2.3 and 3.2.4, if* $q(y(g),\mathbf{x},\boldsymbol{\theta_g})$ *is a real valued function for all* $(y(g),\mathbf{x})\subset\mathbb{R}^M$ *and for all* $\boldsymbol{\theta_g}\in\boldsymbol{\Theta_g}$ *such that* $\mathbb{E}\left[|\frac{q(y(g),\mathbf{x},\boldsymbol{\theta_g})}{r(\mathbf{x},w_g)\cdot p_g(\mathbf{x})}|\right] < \infty$ *for* $g=0,1$*, then we have*

$$\mathbb{E}\left[\frac{s}{r(\mathbf{x}, w_g)}\cdot\frac{w_g}{p_g(\mathbf{x})}\cdot q\left(y(g),\mathbf{x},\boldsymbol{\theta_g}\right)\right] = \mathbb{E}\left[q\left(y(g),\mathbf{x},\boldsymbol{\theta_g}\right)\right]$$

The proof uses two applications of the law of iterated expectations with unconfoundedness and ignorability to arrive at the above result. This equivalence implies that one can now address the identification issue by solving the doubly weighted population problem. Consequently, one can obtain a consistent estimator of $\boldsymbol{\theta_g^0}$ by solving the sample analogue of (3.10) as follows

$$\min_{\boldsymbol{\theta_g}\in\boldsymbol{\Theta_g}} \sum_{i=1}^{N}\frac{s_i}{r(\mathbf{x}_i, w_{ig})}\cdot\frac{w_{ig}}{p_g(\mathbf{x}_i)}\cdot q(y_i(g),\mathbf{x}_i,\boldsymbol{\theta_g}); \quad g=0,1 \tag{3.11}$$

Let the estimator which solves eq (3.11) be denoted as $\tilde{\boldsymbol{\theta}}_{\boldsymbol{g}}$. Note, however, that this estimator is infeasible as it depends on unknown probabilities, $r(\cdot)$ and $p_g(\cdot)$. The next section discusses the first-step of estimating these probabilities.

3.3 Estimation

As mentioned above, one problem with the formulation of $\tilde{\boldsymbol{\theta}}_{\boldsymbol{g}}$ is that the treatment and missing outcome propensity scores are unknown. Therefore, in its current form $\tilde{\boldsymbol{\theta}}_{\boldsymbol{g}}$ cannot be implemented, unless the true probabilities are known. The following assumptions posit that I have a correctly specified model for the two probabilities which help me formulate consistent estimators of $p_g(\mathbf{x})$ and $r(\mathbf{x}, w_g)$.

Assumption 3.3.1. *(Correct parametric specification of propensity score) Assume that*

1. *There exists a known parametric function* $G(\mathbf{x}, \boldsymbol{\gamma})$ *for* $p_1(\mathbf{x})$ *where* $\boldsymbol{\gamma} \in \boldsymbol{\Gamma} \subset \mathfrak{R}^I$ *and* $0 < G(\mathbf{x}, \boldsymbol{\gamma}) < 1$ *for all* $\mathbf{x} \in \mathcal{X}$, $\boldsymbol{\gamma} \in \boldsymbol{\Gamma}$.

2. *There exists* $\boldsymbol{\gamma_0} \in \boldsymbol{\Gamma}$ *s.t.* $p_1(\mathbf{x}) = G(\mathbf{x}, \boldsymbol{\gamma_0})$

Part 1) postulates the existence of a parametric model for the propensity score that is known to the researcher and part 2) assumes that, for some true value of $\boldsymbol{\gamma}$, say $\boldsymbol{\gamma_0}$, this model is correctly specified for the true assignment probability. Similarly, in order to estimate the missing outcome propensity score, I assume that $R(\mathbf{x}, w_g, \boldsymbol{\delta})$ is a correctly specified parametric model for $r(\mathbf{x}, w_g)$. Formally,

Assumption 3.3.2. *(Correct parametric specification of missing outcomes probability) Assume that*

1. *There exists a known parametric function* $R(\mathbf{x}, w_g, \boldsymbol{\delta})$ *for* $r(\mathbf{x}, w_g)$ *where* $\boldsymbol{\delta} \in \boldsymbol{\Delta} \subset \mathfrak{R}^K$ *and* $R(\mathbf{x}, w_g, \boldsymbol{\delta}) > 0$ *for all* $\mathbf{x} \in \mathcal{X}$, $\boldsymbol{\delta} \in \boldsymbol{\Delta}$.

2. *There exists* $\boldsymbol{\delta_0} \in \boldsymbol{\Delta}$ *s.t.* $r(\mathbf{x}, w_g) \equiv R(\mathbf{x}, w_g, \boldsymbol{\delta_0})$

3.3.1 Estimated weights using binary response MLE

To estimate the probability functions $G(\mathbf{x}, \cdot)$ and $R(\mathbf{x}, w_g, \cdot)$, this paper uses binary response conditional maximum likelihood. Since both w_g and s are binary responses, estimation of $\boldsymbol{\gamma_0}$ and $\boldsymbol{\delta_0}$ using MLE will be asymptotically efficient under correct specification of these functions, as assumed in 3.3.1 and 3.3.2. The following two lemmas provide formal consistency and asymptotic normality conditions for MLE estimation of the two probability models. The conditions are adapted from theorems 2.5 and 3.3 of Newey and McFadden (1994).

Lemma 3.3.3. *(Consistency of maximum likelihood) Assume 3.2.4 so that* s_i *and* w_{i1} *are i.i.d with pdf's given as* $f(s_i|w_{i1}, \mathbf{x}_i, \boldsymbol{\delta}) = R(\mathbf{x}_i, w_{i1}, \boldsymbol{\delta})^{s_i} \cdot (1 - R(\mathbf{x}_i, w_{i1}, \boldsymbol{\delta}))^{1-s_i}$ *and* $f(w_{i1}|\mathbf{x}_i, \boldsymbol{\gamma}) = G(\mathbf{x}_i, \boldsymbol{\gamma})^{w_{i1}} \cdot (1 - G(\mathbf{x}_i, \boldsymbol{\gamma}))^{1-w_{i1}}$. *Additionally, assume that*

1. $\boldsymbol{\gamma_0} \in \boldsymbol{\Gamma}$ *and* $\boldsymbol{\delta_0} \in \boldsymbol{\Delta}$, *where* $\boldsymbol{\Gamma}$, $\boldsymbol{\Delta}$ *are compact sets.*

2. *If* $\boldsymbol{\gamma} \neq \boldsymbol{\gamma_0}$, *then* $f(w_{i1}|\mathbf{x}_i, \boldsymbol{\gamma}) \neq f(w_{i1}|\mathbf{x}_i, \boldsymbol{\gamma_0})$ *and if* $\boldsymbol{\delta} \neq \boldsymbol{\delta_0}$, *then* $f(s_i|w_{i1}, \mathbf{x}_i, \boldsymbol{\delta}) \neq f(s_i|w_{i1}, \mathbf{x}_i, \boldsymbol{\delta_0})$.

3. $\ln f(w_{i1}|\mathbf{x}_i, \boldsymbol{\gamma})$ *and* $\ln f(s_i|w_{i1}, \mathbf{x}_i, \boldsymbol{\delta})$ *is continuous at each* $\boldsymbol{\gamma} \in \boldsymbol{\Gamma}$ *and* $\boldsymbol{\delta} \in \boldsymbol{\Delta}$ *respectively with probability one.*

4. $\mathbb{E}\left[sup_{\boldsymbol{\gamma}\in\boldsymbol{\Gamma}}|\ln f(w_{i1}|\mathbf{x}_i, \boldsymbol{\gamma})|\right] < \infty$ *and* $\mathbb{E}\left[sup_{\boldsymbol{\delta}\in\boldsymbol{\Delta}}|\ln f(s_i|w_{i1}, \mathbf{x}, \boldsymbol{\delta})|\right] < \infty$. *Then*

$$\hat{\boldsymbol{\gamma}} \xrightarrow{p} \boldsymbol{\gamma_0} \text{ and } \hat{\boldsymbol{\delta}} \xrightarrow{p} \boldsymbol{\delta_0}$$

The proof of the lemma is given in the appendix. For asymptotic normality, consider the following

Lemma 3.3.4. *(Asymptotic normality for MLE) Assume that conditions of lemma 3.3.3 are satisfied and*

1. $\boldsymbol{\gamma} \in$ *interior* $(\boldsymbol{\Gamma})$ *and* $\boldsymbol{\delta} \in$ *interior* $(\boldsymbol{\Delta})$.

2. $f(s_i|w_{i1}, \mathbf{x}_i, \boldsymbol{\delta})$ *and* $f(w_{i1}|\mathbf{x}_i, \boldsymbol{\gamma})$ *are both twice continuously differentiable and* $f(s_i|w_{i1}, \mathbf{x}_i, \boldsymbol{\delta}) > 0$ *and* $f(w_{i1}|\mathbf{x}_i, \boldsymbol{\gamma}) > 0$ *in a neighborhood* $\mathscr{N}$ *of* $\boldsymbol{\gamma_0}$ *and* $\boldsymbol{\delta_0}$ *respectively.*

3. $\int sup_{\boldsymbol{\gamma}\in\mathscr{N}}||\nabla_{\boldsymbol{\gamma}} f(w_{i1}|\mathbf{x}_i, \boldsymbol{\gamma})||dw_1 < \infty$, $\int sup_{\boldsymbol{\gamma}\in\mathscr{N}}||\nabla_{\boldsymbol{\gamma}\boldsymbol{\gamma}'} f(w_{i1}|\mathbf{x}_i, \boldsymbol{\gamma})||dw_1 < \infty$. *Similarly,*
$\int sup_{\boldsymbol{\delta}\in\mathscr{N}}||\nabla_{\boldsymbol{\delta}} f(s_i|w_{i1}, \mathbf{x}_i, \boldsymbol{\delta})||ds < \infty$ *and* $\int sup_{\boldsymbol{\gamma}\in\mathscr{N}}||\nabla_{\boldsymbol{\delta}\boldsymbol{\delta}'} f(s_i|w_{i1}, \mathbf{x}_i, \boldsymbol{\delta})||ds < \infty$.

4. $\mathbb{E}\left[\nabla_{\boldsymbol{\gamma}} \ln f(w_{i1}|\mathbf{x}_i, \boldsymbol{\gamma_0})\{\nabla_{\boldsymbol{\gamma}} \ln f(w_{i1}|\mathbf{x}_i, \boldsymbol{\gamma_0})\}'\right]$ *exists and is non-singular. Similarly,*
$\mathbb{E}\left[\nabla_{\boldsymbol{\delta}} \ln f(s_i|w_{i1}, \mathbf{x}_i, \boldsymbol{\delta_0})\{\nabla_{\boldsymbol{\delta}} \ln f(s_i|w_{i1}, \mathbf{x}_i, \boldsymbol{\delta_0})\}'\right]$ *exists and is non-singular.*

5. $\mathbb{E}\left[sup_{\boldsymbol{\gamma}\in\mathscr{N}}||\nabla_{\boldsymbol{\gamma}\boldsymbol{\gamma}'} \ln f(w_{i1}|\mathbf{x}_i, \boldsymbol{\gamma})||\right] < \infty$ *and* $\mathbb{E}\left[sup_{\boldsymbol{\delta}\in\mathscr{N}}||\nabla_{\boldsymbol{\delta}\boldsymbol{\delta}'} \ln f(s_i|w_{i1}, \mathbf{x}_i, \boldsymbol{\delta})||\right] < \infty$.

Then, the MLE estimator, $\hat{\boldsymbol{\gamma}}$, and $\hat{\boldsymbol{\delta}}$ solve

$$\max_{\boldsymbol{\gamma}\in\boldsymbol{\Gamma}} \sum_{i=1}^{N} \left\{ w_{i1} \log G(\mathbf{x}_i, \boldsymbol{\gamma}) + (1 - w_{i1}) \log(1 - G(\mathbf{x}_i, \boldsymbol{\gamma})) \right\}$$

$$\max_{\boldsymbol{\delta}\in\boldsymbol{\Delta}} \sum_{i=1}^{N} \left\{ s_i \log\{R(\mathbf{x}_i, w_{i1}, \boldsymbol{\delta})\} + (1 - s_i) \log\{1 - R(\mathbf{x}_i, w_{i1}, \boldsymbol{\delta})\} \right\}$$

respectively.

For a proof, see appendix H. Given estimators $\hat{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\delta}}$, one can estimate the assignment and missing outcome propensity scores by $G(\cdot, \hat{\boldsymbol{\gamma}})$ and $R(\cdot, \hat{\boldsymbol{\delta}})$ respectively. Consistency and asymptotic normality follow from applying the continuous mapping theorem and the delta method given that $G(\cdot, \hat{\boldsymbol{\gamma}})$ and $R(\cdot, \hat{\boldsymbol{\delta}})$ are assumed to be continuously differentiable, which is implicit in Lemma 3.3.3 and 3.3.4. In practice, this paper follows the convention of estimating these probabilities as flexible logits where the above requirements of continuity and differentiability are easily satisfied.

3.3.2 Doubly weighted M-estimator

Once the probability weights have been estimated, let $\hat{\boldsymbol{\theta}}_{\mathbf{1}}$ denote the *doubly weighted* estimator which solves the treatment group problem,

$$\min_{\boldsymbol{\theta}_{\mathbf{1}}\in\boldsymbol{\Theta}_{\mathbf{1}}} \sum_{i=1}^{N} \frac{s_i}{R(\mathbf{x}_i, w_{i1}, \hat{\boldsymbol{\delta}})} \cdot \frac{w_{i1}}{G(\mathbf{x}_i, \hat{\boldsymbol{\gamma}})} \cdot q(y_i(1), \mathbf{x}_i, \boldsymbol{\theta}_{\mathbf{1}}) \tag{3.12}$$

with weights given by $G(\mathbf{x}, \hat{\boldsymbol{\gamma}})$ and $R(\mathbf{x}, w_1, \hat{\boldsymbol{\delta}})$, and let $\hat{\boldsymbol{\theta}}_{\mathbf{0}}$ be the estimator which solves the control group problem,

$$\min_{\boldsymbol{\theta}_{\mathbf{0}}\in\boldsymbol{\Theta}_{\mathbf{0}}} \sum_{i=1}^{N} \frac{s_i}{R(\mathbf{x}_i, w_{i0}, \hat{\boldsymbol{\delta}})} \cdot \frac{w_{i0}}{(1 - G(\mathbf{x}_i, \hat{\boldsymbol{\gamma}}))} \cdot q(y_i(0), \mathbf{x}_i, \boldsymbol{\theta}_{\mathbf{0}}) \tag{3.13}$$

using weights $(1 - G(\mathbf{x}, \hat{\boldsymbol{\gamma}}))$ and $R(\mathbf{x}, w_0, \hat{\boldsymbol{\delta}})$. Henceforth, this estimator will be denoted as $\hat{\boldsymbol{\theta}}_{\boldsymbol{g}}$ for $g = 0, 1$.

Example 1 (Ordinary least squares): In the case of a misspecified conditional mean function, $\hat{\boldsymbol{\theta}}_{\mathbf{1}} \equiv (\hat{\alpha}_1, \hat{\boldsymbol{\beta}}_{\mathbf{1}})'$ will solve a double weighted version of the OLS problem i.e.

$$\hat{\boldsymbol{\theta}}_{\mathbf{1}} = \underset{\boldsymbol{\theta}_{\mathbf{1}} \in \boldsymbol{\Theta}_{\mathbf{1}}}{\operatorname{argmin}} \sum_{i=1}^{N} \frac{s_i}{R(\mathbf{x}_i, w_{i1}, \hat{\boldsymbol{\delta}})} \cdot \frac{w_{i1}}{G(\mathbf{x}_i, \hat{\boldsymbol{\gamma}})} \cdot (y_i(1) - \alpha_1 - \mathbf{x}_i \boldsymbol{\beta}_{\mathbf{1}})^2$$

Similarly,

$$\hat{\boldsymbol{\theta}}_{\mathbf{0}} = \underset{\boldsymbol{\theta}_{\mathbf{0}} \in \boldsymbol{\Theta}_{\mathbf{0}}}{\operatorname{argmin}} \sum_{i=1}^{N} \frac{s_i}{R(\mathbf{x}_i, w_{i0}, \hat{\boldsymbol{\delta}})} \cdot \frac{w_{i0}}{(1 - G(\mathbf{x}_i, \hat{\boldsymbol{\gamma}}))} \cdot (y_i(0) - \alpha_0 - \mathbf{x}_i \boldsymbol{\beta}_{\mathbf{0}})^2$$

where $(\hat{\alpha}_g, \hat{\boldsymbol{\beta}}_{\boldsymbol{g}})'$ will be consistent for the linear projection of $y(g)|\mathbf{x}$.

Example 2 (Quantile regression): Similarly, in the case of a misspecified conditional quantile function, $\hat{\boldsymbol{\theta}}^{\mathbf{0}}_{\boldsymbol{g}}(\tau) \equiv (\hat{\alpha}^0_g(\tau), \hat{\boldsymbol{\beta}}^{\mathbf{0}}_{\boldsymbol{g}}(\tau))'$ will solve the following weighted mean square error loss functions (Angrist et al. (2006b)), i.e.

$$\hat{\boldsymbol{\theta}}_{\mathbf{1}}(\tau) = \underset{\boldsymbol{\theta}_{\mathbf{1}} \in \boldsymbol{\Theta}_{\mathbf{1}}}{\operatorname{argmin}} \sum_{i=1}^{N} \frac{s_i}{R(\mathbf{x}_i, w_{i1}, \hat{\boldsymbol{\delta}})} \cdot \frac{w_{i1}}{G(\mathbf{x}_i, \hat{\boldsymbol{\gamma}})} \cdot \omega_\tau(\mathbf{x}_i, \boldsymbol{\theta}_{\mathbf{1}}) \cdot \left[\text{Quant}_\tau(y_i(1)|\mathbf{x}_i) - \alpha_1(\tau) - \mathbf{x}_i \boldsymbol{\beta}_{\mathbf{1}}(\tau)\right]^2$$

$$\hat{\boldsymbol{\theta}}_{\mathbf{0}}(\tau) = \underset{\boldsymbol{\theta}_{\mathbf{0}} \in \boldsymbol{\Theta}_{\mathbf{0}}}{\operatorname{argmin}} \sum_{i=1}^{N} \frac{s_i}{R(\mathbf{x}_{i0}, w_{i0}, \hat{\boldsymbol{\delta}})} \cdot \frac{w_{i0}}{(1 - G(\mathbf{x}_i, \hat{\boldsymbol{\gamma}}))} \cdot \omega_\tau(\mathbf{x}_i, \boldsymbol{\theta}_{\mathbf{0}}) \cdot \left[\text{Quant}_\tau(y_i(0)|\mathbf{x}_i) - \alpha_0(\tau) - \mathbf{x}_i \boldsymbol{\beta}_{\mathbf{0}}(\tau)\right]^2$$

where $\hat{\boldsymbol{\theta}}_{\boldsymbol{g}}(\tau)$ will now be consistent for a weighted linear projection to the true CQF of $y(g)|\mathbf{x}$. Using the doubly weighted estimator, one can now consistently estimate causal parameters like the average treatment effect and different quantile treatment effects. Section 3.6 discusses each of these examples in detail. The next section develops and discusses the large sample theory of the proposed estimator.

3.4 Asymptotic theory

This paper implements the proposed estimator in a two-step procedure. The first step uses binary response MLE for the estimation of the probability weights and the second step uses the first-step weights to estimate the parameter of interest, $\boldsymbol{\theta}^{\mathbf{0}}_{\boldsymbol{g}}$. The asymptotic theory utilizes results for two-step estimators with a non-smooth objective function in the second step, to establish consistency and asymptotic normality of $\hat{\boldsymbol{\theta}}_{\boldsymbol{g}}$. Therefore, the usual

regularity conditions assuming continuity and twice differentiability with respect to $\boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}}$ are now relaxed.

3.4.1 Consistency

Using the conditions in Lemma 2.4 of Newey and McFadden (1994), it is easy to establish consistency of the doubly weighted M-estimator, $\hat{\boldsymbol{\theta}}_{\boldsymbol{g}}$. The conditions of the lemma are quite weak with continuity and a data dependent upper bound with a finite expectation being the only substantive requirements. The following theorem fills in the primitive regularity conditions for applying the uniform law of large numbers.

Theorem 3.4.1. *(Consistency) Assume 3.2.1, 3.2.2, 3.2.3, 3.2.4, 3.3.1 and 3.3.2 hold. Further, let*

1) $\boldsymbol{\Theta}_{\boldsymbol{g}}$ *is compact for* $g = 0, 1$

2) $q(y(g), \mathbf{x}, \boldsymbol{\theta}_{\boldsymbol{g}})$ *is continuous at each* $\boldsymbol{\theta}_{\boldsymbol{g}} \in \boldsymbol{\Theta}_{\boldsymbol{g}}$ *with probability one.*

3) For all $\boldsymbol{\theta}_{\boldsymbol{g}} \in \boldsymbol{\Theta}_{\boldsymbol{g}}$, $|q(y(g), \mathbf{x}, \boldsymbol{\theta}_{\boldsymbol{g}})| \leq b(y(g), \mathbf{x})$ *for some function* $b(\cdot)$ *such that* $\mathbb{E}\left[b(y(g), \mathbf{x})\right] < \infty$

Then,

$$\hat{\boldsymbol{\theta}}_{\boldsymbol{g}} \xrightarrow{p} \boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}} \text{ as } N \to \infty$$

The proof of the theorem can be found in the appendix. The conditions of the above theorem allow the objective function to not be continuous on all of $\boldsymbol{\theta}_{\boldsymbol{g}}$ for a given $\mathbf{x}$. This is useful for cases where $q(\cdot)$ is allowed to be non-smooth. Under the dominance condition given in (3), uniform convergence of sample averages holds quite generally. Compactness of the parameter space and identification, as given in Assumption 3.2.1, are both more primitive that can be relaxed without affecting consistency.

3.4.2 Asymptotic normality

For establishing asymptotic normality, I provide conditions for the general case of non-smooth objective functions since the conditions in this case can accommodate the smooth case as well. The main condition needed for establishing asymptotic normality of the doubly weighted estimator is stochastic equicontinuity that will be sufficient to guarantee uniform convergence of the objective function to its population counterpart. Before I state the conditions of the normality proof, let the population problem be denoted as

$$Q_0(\boldsymbol{\theta_0}) = \mathbb{E}\left[\frac{s_i \cdot w_{i0}}{R(\mathbf{x}_i, w_{i0}, \boldsymbol{\delta_0}) \cdot (1 - G(\mathbf{x}_i, \boldsymbol{\gamma_0}))} \cdot q(y_i(0), \mathbf{x}_i, \boldsymbol{\theta_0})\right]$$
$$Q_0(\boldsymbol{\theta_1}) = \mathbb{E}\left[\frac{s_i \cdot w_{i1}}{R(\mathbf{x}_i, w_{i1}, \boldsymbol{\delta_0}) \cdot G(\mathbf{x}_i, \boldsymbol{\gamma_0})} \cdot q(y_i(1), \mathbf{x}_i, \boldsymbol{\theta_1})\right]$$

and the sample analogue be given as

$$Q_N(\boldsymbol{\theta_0}) = \frac{1}{N_0}\sum_{i=1}^{N} \frac{s_i \cdot w_{i0}}{R(\mathbf{x}_i, w_{i0}, \hat{\boldsymbol{\delta}}) \cdot (1 - G(\mathbf{x}_i, \hat{\boldsymbol{\gamma}}))} \cdot q(y_i(0), \mathbf{x}_i, \boldsymbol{\theta_0})$$
$$Q_N(\boldsymbol{\theta_1}) = \frac{1}{N_1}\sum_{i=1}^{N} \frac{s_i \cdot w_{i1}}{R(\mathbf{x}_i, w_{i1}, \hat{\boldsymbol{\delta}}) \cdot G(\mathbf{x}_i, \hat{\boldsymbol{\gamma}})} \cdot q(y_i(1), \mathbf{x}_i, \boldsymbol{\theta_1})$$

where $N_1 = \sum_{i=1}^{N} s_i \cdot w_{i1}$ and $N_0 = \sum_{i=1}^{N} s_i \cdot w_{i0}$. Then, I have the following theorem for asymptotic normality which is taken from Newey and McFadden (1994) section 7 along with primitive conditions taken from Andrews (1994).

Theorem 3.4.2. *(Asymptotic Normality of the Doubly Weighted Estimator) Given assumptions 3.2.1, 3.2.2, 3.2.3, 3.2.4*

(1) Suppose that $\hat{\boldsymbol{\theta}}_{\boldsymbol{g}}$ *is an approximate minimum i.e*

$$Q_N(\hat{\boldsymbol{\theta}}_{\boldsymbol{g}}) \leq \inf_{\boldsymbol{\theta_g} \in \boldsymbol{\Theta_g}} Q_N(\boldsymbol{\theta_g}) + o_p(N^{-1})$$

(2) $\hat{\boldsymbol{\theta}}_{\boldsymbol{g}} \xrightarrow{p} \boldsymbol{\theta}^{\mathbf{0}}_{\boldsymbol{g}}$, $\boldsymbol{\theta}^{\mathbf{0}}_{\boldsymbol{g}} \in int(\boldsymbol{\Theta_g})$

(3) $Q_0(\boldsymbol{\theta_g})$ *is minimized on* $\boldsymbol{\Theta_g}$ *at* $\boldsymbol{\theta}^{\mathbf{0}}_{\boldsymbol{g}}$

(4) $Q_0(\boldsymbol{\theta_g})$ *is twice differentiable at* $\boldsymbol{\theta_g^0}$ *with a nonsingular Hessian,* $\mathbf{H_g}$

(5) Suppose $\boldsymbol{\nabla_{\theta_g}} Q_N(\boldsymbol{\theta_g^0})$ *exists with probability one and* $\sqrt{N}\ \boldsymbol{\nabla_{\theta_g}} Q_N(\boldsymbol{\theta_g^0}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{\Omega_g})$

(6) Let,

$$\mathbf{l} = \nabla_{\boldsymbol{\theta_1}} \left\{ \frac{s \cdot w_1}{R(\mathbf{x}, w_1, \boldsymbol{\delta^*}) \cdot G(\mathbf{x}, \boldsymbol{\gamma^*})} \cdot q(y(1), \mathbf{x}, \boldsymbol{\theta_1}) \right\}'$$

$$\mathbf{k} = \nabla_{\boldsymbol{\theta_0}} \left\{ \frac{s \cdot w_0}{R(\mathbf{x}, w_0, \boldsymbol{\delta^*}) \cdot (1 - G(\mathbf{x}, \boldsymbol{\gamma^*}))} \cdot q(y(0), \mathbf{x}, \boldsymbol{\theta_0}) \right\}'$$

and let the class

$$\mathscr{F} = \left\{ f; f(y(g), \mathbf{x}) = \begin{cases} \mathbf{l}, \ \textit{for } g = 1 \\ \\ \mathbf{k}, \ \textit{for } g = 0 \end{cases} : \boldsymbol{\theta_g} \in \boldsymbol{\Theta_g}, \ \forall (y(g), \mathbf{x}) \subset \Re^M \right\}$$

satisfy Pollard's entropy condition with envelope given by

$$F = 1 \vee \sup_{f \in \mathscr{F}} |f(\cdot)|$$

for Type-I classes ***or*** *satisfies Ossiander's* L^p *entropy condition with* $p = 2$ *with envelope given by*

$$F = \sup_{f \in \mathscr{F}} |f(\cdot)|$$

for Type II-VI classes, where these are defined in Andrews (1994).

(7) $\limsup_{N \to \infty} \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}(F)^{2+\zeta} < \infty$ *for some* $\zeta > 0$ *and* F *given above.*

(8) The conditions of Lemma 3.4 are satisfied allowing the first order influence function representation for $\hat{\boldsymbol{\gamma}}$

$$\sqrt{N}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma_0}) = \left[\mathbb{E}\left(\mathbf{d}_i \mathbf{d}_i' \right) \right]^{-1} \left\{ N^{-1/2} \sum_{i=1}^{N} \mathbf{d}_i \right\} + o_p(1) \tag{3.14}$$

where

$$\mathbf{d}_i = w_{i1} \cdot \left[\frac{\nabla_{\boldsymbol{\gamma}} G(\mathbf{x}_i, \boldsymbol{\gamma_0})'}{G(\mathbf{x}_i, \boldsymbol{\gamma_0})} \right] - (1 - w_{i1}) \cdot \left[\frac{\nabla_{\boldsymbol{\gamma}} G(\mathbf{x}_i, \boldsymbol{\gamma_0})'}{1 - G(\mathbf{x}_i, \boldsymbol{\gamma_0})} \right] \tag{3.15}$$

is the $I \times 1$ *score of the maximum likelihood binary response log-likelihood evaluated at the true parameter value* $\boldsymbol{\gamma_0}$.

Similarly, $\hat{\boldsymbol{\delta}}$ *has the following first order influence function representation*

$$\sqrt{N}\left(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta_0}\right) = \left[\mathbb{E}\left(\mathbf{b}_i\mathbf{b}_i'\right)\right]^{-1}\left\{N^{-1/2}\sum_{i=1}^{N}\mathbf{b}_i\right\} + o_p(1) \tag{3.16}$$

where

$$\mathbf{b}_i = s_i \cdot \left[\frac{\nabla_{\boldsymbol{\delta}} R(\mathbf{x}_i, w_{i1}, \boldsymbol{\delta_0})'}{R(\mathbf{x}_i, w_{i1}, \boldsymbol{\delta_0})}\right] - (1 - s_i) \cdot \left[\frac{\nabla_{\boldsymbol{\delta}} R(\mathbf{x}_i, w_{i1}, \boldsymbol{\delta_0})'}{1 - R(\mathbf{x}_i, w_{i1}, \boldsymbol{\delta_0})}\right] \tag{3.17}$$

is the $K \times 1$ *score of the maximum likelihood binary response log-likelihood evaluated at the true parameter value* $\boldsymbol{\delta_0}$.

Then,

$$\sqrt{N}(\hat{\boldsymbol{\theta}}_{\boldsymbol{g}} - \boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}}) \xrightarrow{d} N\left(\mathbf{0}, \mathbf{H}_{\boldsymbol{g}}^{-1}\boldsymbol{\Omega}_{\boldsymbol{g}}\mathbf{H}_{\boldsymbol{g}}^{-1}\right)$$

where

$$\boldsymbol{\Omega_1} = \mathbb{E}\left(\mathbf{l}_i\mathbf{l}_i'\right) - \mathbb{E}\left(\mathbf{l}_i\mathbf{b}_i'\right)\mathbb{E}\left(\mathbf{b}_i\mathbf{b}_i'\right)^{-1}\mathbb{E}\left(\mathbf{b}_i\mathbf{l}_i'\right) - \mathbb{E}\left(\mathbf{l}_i\mathbf{d}_i'\right)\mathbb{E}\left(\mathbf{d}_i\mathbf{d}_i'\right)^{-1}\mathbb{E}\left(\mathbf{d}_i\mathbf{l}_i'\right)$$
$$\boldsymbol{\Omega_0} = \mathbb{E}\left(\mathbf{k}_i\mathbf{k}_i'\right) - \mathbb{E}\left(\mathbf{k}_i\mathbf{b}_i'\right)\mathbb{E}\left(\mathbf{b}_i\mathbf{b}_i'\right)^{-1}\mathbb{E}\left(\mathbf{b}_i\mathbf{k}_i'\right) - \mathbb{E}\left(\mathbf{k}_i\mathbf{d}_i'\right)\mathbb{E}\left(\mathbf{d}_i\mathbf{d}_i'\right)^{-1}\mathbb{E}\left(\mathbf{d}_i\mathbf{k}_i'\right)$$

The primitive conditions for stochastic equicontinuity hold for classes of functions of type I-VI as defined in Andrews. Conditions (1)-(5) are standard for the case of non-smooth objective functions. Condition (5) requires that the score of the objective function exists with probability one and is normally distributed. This condition is important for establishing distributional convergence of $\hat{\boldsymbol{\theta}}_{\boldsymbol{g}}$.

Condition (6) and (7) together with random sampling are sufficient for stochastic equicontinuity of the remainder term in Newey and McFadden (1994).[19] Checking these conditions in a particular application would entail showing that $f(\cdot)$ belongs to one these classes. For instance, both linear and quantile regression examples considered in this paper belong to

[19]Directly verifying stochastic equicontinuity as mentioned in theorem 7.2 of Newey and McFadden (1994) is difficult and hence primitive conditions like (6) and (7) can be useful. Pollard (1985) also provides primitive conditions that are sufficient for stochastic differentiability which is quite similar to the condition of stochastic equicontinuity.

type-I class of functions. Consequently, stochastic equicontinuity follows from Theorem 1 and 4 in Andrews (1994) for type-I class and type II-VI classes respectively. Conditions (8) and (9) are simply imposing regularity conditions on $R(\cdot)$ and $G(\cdot)$ so that influence function representations given in 3.14 and 3.16 are possible. For a proof of the theorem, see appendix.

3.4.3 Efficiency gain with estimated weights

The asymptotic variance expression derived in the previous section offers some interesting insights. First, the middle term, $\boldsymbol{\Omega_g}$, represents the variance of the residual from the population regression of the weighted score on the two binary response scores, $\mathbf{b}_i$ and $\mathbf{d}_i$. Note that even though $\boldsymbol{\Omega_g}$ should involve a fourth term for the covariance between the two scores, this term is zero in the present case, on account of the two scores being conditionally independent.[20]

Second, the expression for $\boldsymbol{\Omega_g}$, as derived here, is different from what I obtain in section 3.5 under the stronger identification assumption. This difference has an interesting efficiency implication. In the case when a researcher is only willing to assume identification in the weaker sense of 3.2.1, it is potentially more efficient to estimate the two probabilities in a first step. Note though that this result is asymptotic in nature. In order to see that, let us assume that we know $G(\mathbf{x}_i, \boldsymbol{\gamma_0})$ and $R(\mathbf{x}_i, w_{ig}, \boldsymbol{\delta_0})$. Then, the asymptotic variance of the estimator, say $\tilde{\boldsymbol{\theta}}_{\boldsymbol{g}}$ which uses the known probabilities is:

$$\text{Avar}\left[\sqrt{N}\left(\tilde{\boldsymbol{\theta}}_{\boldsymbol{g}} - \boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}}\right)\right] = \mathbf{H}_g^{-1}\boldsymbol{\Sigma_g}\mathbf{H}_g^{-1}$$

where $\boldsymbol{\Sigma_1} = \mathbf{Var}(\mathbf{l}_i) = \mathbb{E}\left(\mathbf{l}_i\mathbf{l}_i'\right)$ for the treatment group and $\boldsymbol{\Sigma_0} = \mathbf{Var}(\mathbf{k}_i) = \mathbb{E}\left(\mathbf{k}_i\mathbf{k}_i'\right)$ for the control group. I formalize this result in the next theorem

Theorem 3.4.3. *(Efficiency gain with estimated weights) Under the assumptions of theorem*

[20] For the proof, see appendix.

3.4.2 we obtain

$$\text{Avar}\left[\sqrt{N}\left(\tilde{\boldsymbol{\theta}}_{\boldsymbol{g}}-\boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}}\right)\right]-\text{Avar}\left[\sqrt{N}\left(\hat{\boldsymbol{\theta}}_{\boldsymbol{g}}-\boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}}\right)\right]=\mathbf{H}_{g}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{g}}\mathbf{H}_{g}^{-1}-\mathbf{H}_{g}^{-1}\boldsymbol{\Omega}_{\boldsymbol{g}}\mathbf{H}_{g}^{-1}$$
$$=\mathbf{H}_{g}^{-1}\left(\boldsymbol{\Sigma}_{\boldsymbol{g}}-\boldsymbol{\Omega}_{\boldsymbol{g}}\right)\mathbf{H}_{g}^{-1}\text{ is PSD.}$$

The proof is given in the appendix. In other words, asymptotically, we do no worse by estimating the probabilities than when we actually know them. Due to the asymptotic nature of the results, there may not be any gain in finite samples. This result can be seen an extension of Wooldridge (2007) to the case when one has two sets of probability weights being estimated in the first stage.

In the missing data literature, this result has also been called the "efficiency puzzle". Prokhorov and Schmidt (2009) study this puzzle in a GMM framework using an augmented set of moment conditions, where the first subset of moments correspond to the weighted objective function and the second subset belong to the missing outcomes (or selection) problem. An interesting explanation that emerges from their framework is that the second set of moment conditions are useful even when selection probability parameters are known. Therefore, inefficiency of the known probability estimator (as seen above) is due to its failure to exploit the correlation between the first and second set of moments. Hence, knowledge of the selection parameters do not play a role in efficient estimation.

3.5 Some feature of interest is correctly specified

The results in the previous section were derived under the assumption that the parameter vector solves an unconditional M-estimation problem. Even though it can handle cases where the conditional feature of interest is correctly specified, the explicit focus was on examples of model misspecification such as estimating a misspecified linear model for either the true conditional mean or the true conditional quantile function. In contrast, this section focuses on situations where $\boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}}$ indexes a true conditional feature of interest. This could be a mean, quantile or the entire conditional distribution of $y(g)|\mathbf{x}$. In this case, $\boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}}$ can be said to be

identified in a stronger sense which is reflected in an improvisation of the basic identification assumption given in eq (3.2) to the following,

Assumption 3.5.1. *(Strong identification of $\boldsymbol{\theta_g^0}$)*

The parameter vector $\boldsymbol{\theta_g^0} \in \boldsymbol{\Theta_g}$ is the unique solution to the population minimization problem

$$\min_{\boldsymbol{\theta_g} \in \boldsymbol{\Theta_g}} \mathbb{E}\left[q(y(g), \mathbf{x}, \boldsymbol{\theta_g})|\mathbf{x}, w_g, s\right]; \ g = 0, 1 \tag{3.18}$$

for each $(\mathbf{x}, w_g, s) \in \mathcal{V} \subset \mathfrak{R}^{dim(\mathfrak{X})+2}$. In other words, under ignorability (as defined in 3.2.3) and unconfoundedness (defined in 3.2.2), $\boldsymbol{\theta_g^0}$ solves

$$\min_{\boldsymbol{\theta_g} \in \boldsymbol{\Theta_g}} \mathbb{E}\left[q(y(g), \mathbf{x}, \boldsymbol{\theta_g})|\mathbf{x}\right]; \ g = 0, 1 \tag{3.19}$$

for each $\mathbf{x} \in \mathfrak{X} \subset \mathfrak{R}^{dim(\mathfrak{X})}$.

The above assumption can be seen as a strengthening of the identification assumption in 3.2.1. The basic identification assumption simply defines $\boldsymbol{\theta_g^0}$ to be the solution to the unconditional M-estimation problem, irrespective of whether it is correctly specified for an underlying model or not. Assumption 3.5.1 is additionally requiring $\boldsymbol{\theta_g^0}$ to solve the stronger conditional M-estimation problem. For instance, assumption 3.5.1 will be satisfied for a correctly specified CEF given by

$$\begin{aligned} &y(g) = \alpha_g^0 + \mathbf{x}\boldsymbol{\beta_g^0} + u(g); \quad g = 0, 1 \\ &\mathbb{E}\left(u(g)|\mathbf{x}\right) = 0 \end{aligned} \tag{3.20}$$

with either OLS or QMLE in the linear exponential family as the chosen estimation method. This would also hold for a correctly specified CQF estimated either using quantile regression or QMLE in the tick exponential family (Komunjer (2005)). Requirement for the parameter, $\boldsymbol{\theta_g^0}$, to solve the stronger ID problem is an important distinction which will ultimately help me characterizing the robustness properties of the doubly weighted estimator. I will illustrate this property through two main examples; the first will study estimation of ATE and the

second will study estimation of quantile effects. Both these examples are studied in detail in section 3.6.

Until now I have not said anything about the parametric specifications of functions $R(\cdot)$ and $G(\cdot)$. In fact, under assumption 3.5.1, correct functional form assumptions on these two probabilities can be dispensed with. This is a second important distinction between the results characterized under assumption 3.2.1 and the results characterized in this section under 3.5.1. Therefore, the requirement that $\boldsymbol{\theta_g^0}$ solves the objective function for each $(\mathbf{x}, w_g, s) \in \mathcal{V}$ is much stronger than the requirement in assumption 3.2.1 since assumption 3.5.1 implies assumption 3.2.1 but not the other way around. Formally, I will show that the estimator of $\boldsymbol{\theta_g^0}$ that solves the sample equivalent of eq (3.19) with potentially misspecified treatment and missing outcomes probabilities will still consistently estimate $\boldsymbol{\theta_g^0}$. Before that, the following assumptions formalize possible misspecification of these probability models

Assumption 3.5.2. *(Parametric specification of propensity score) Assume that conditions (1) and (3) of 3.3.1 hold where condition (2) is defined for some* $\boldsymbol{\gamma}^* \in \boldsymbol{\Gamma}$ *such that plim*$(\hat{\boldsymbol{\gamma}}) = \boldsymbol{\gamma}^*$

Assumption 3.5.2 says that we have a known parametric function for the propensity score but there is no requirement for this model to be correctly specified. I continue to assume that the estimator of $\boldsymbol{\gamma}^*$ solves a binary response maximum likelihood problem and $G(\mathbf{x}, \boldsymbol{\gamma}^*)$ is the model evaluated at the pseudo true value. In the event that *the model is correctly specified* for the propensity score, $G(\mathbf{x}, \boldsymbol{\gamma}^*) = p_1(\mathbf{x})$. I make a similar assumption for the missing outcomes model.

Assumption 3.5.3. *(Parametric specification of missingness probability) Assume that conditions (1) and (3) of 3.3.2 hold where condition (2) is defined for some* $\boldsymbol{\delta}^* \in \boldsymbol{\Gamma}$ *such that plim*$(\hat{\boldsymbol{\delta}}) = \boldsymbol{\delta}^*$

Again, assumption 3.5.3 says that we have a known parametric function for the missing outcome probability given by $R(\mathbf{x}, w_g, \boldsymbol{\delta})$ and I do not impose any requirement for this model

to be correctly specified. However, when this model is *correctly specified*, $R(\mathbf{x}, w_g, \boldsymbol{\delta^*}) = r(\mathbf{x}, w_g)$.

Given assumptions 3.5.1, 3.5.2 and 3.5.3, its easy to show that $\boldsymbol{\theta_g^0}$ solves the doubly weighted problem in the population where the weights are constructed using potentially misspecified probabilities. I provide a sketch of the argument for the treatment group parameter $\boldsymbol{\theta_1^0}$ and the proof for $\boldsymbol{\theta_0^0}$ follows in a similar manner. Consider,

$$\mathbb{E}\left[\frac{s}{R(\mathbf{x}, w_1, \boldsymbol{\delta^*})} \cdot \frac{w_1}{G(\mathbf{x}, \boldsymbol{\gamma^*})} \cdot q(y(1), \mathbf{x}, \boldsymbol{\theta_1})\right] \tag{3.21}$$

Using three applications of LIE along with ignorability and unconfoundedness, I can rewrite the above expectation as

$$\mathbb{E}\left[\frac{r(\mathbf{x}, w_1)}{R(\mathbf{x}, w_1, \boldsymbol{\delta^*})} \cdot \frac{p_1(\mathbf{x})}{G(\mathbf{x}, \boldsymbol{\gamma^*})} \cdot \mathbb{E}[q(y(1), \mathbf{x}, \boldsymbol{\theta_1})|\mathbf{x}]\right]$$

Assumption 3.5.1 along with positive weights i.e. $\frac{r(\mathbf{x},w_1)}{R(\mathbf{x},w_1,\boldsymbol{\delta^*})} \geq 0$ and $\frac{p_1(\mathbf{x})}{G(\mathbf{x},\boldsymbol{\gamma^*})} \geq 0$ for all $(\mathbf{x}, w_1)$, implies

$$\frac{r(\mathbf{x}, w_1)}{R(\mathbf{x}, w_1, \boldsymbol{\delta^*})} \cdot \frac{p_1(\mathbf{x})}{G(\mathbf{x}, \boldsymbol{\gamma^*})} \cdot \mathbb{E}[\, q(y(1), \mathbf{x}, \boldsymbol{\theta_1^0})|\mathbf{x}] \leq \frac{r(\mathbf{x}, w_1)}{R(\mathbf{x}, w_1, \boldsymbol{\delta^*})} \cdot \frac{p_1(\mathbf{x})}{G(\mathbf{x}, \boldsymbol{\gamma^*})} \cdot \mathbb{E}[\, q(y(1), \mathbf{x}, \boldsymbol{\theta_1})|\mathbf{x}],\ \forall\ \boldsymbol{\theta_1} \in \boldsymbol{\Theta_1}$$

where the inequality is strict when $\boldsymbol{\theta_1} \neq \boldsymbol{\theta_1^0}$. Therefore, solving 3.21 identifies the parameter even if the weights are misspecified. In general, the parameter that solves 3.21 will be different from the one that solves 3.2. But as long as $\boldsymbol{\theta_g^0}$ is a unique solution, solving 3.21 will identify it.

When $R(\mathbf{x}, w_g, \boldsymbol{\delta^*}) = r(\mathbf{x}, w_g)$ and $G(\mathbf{x}, \boldsymbol{\gamma^*}) = p_1(\mathbf{x})$, then solving 3.21 will be the same as solving 3.12 for the treatment group and 3.13 for the control group. Estimation of $G(\cdot)$ and $R(\cdot)$ follows from Lemma 3.3.3 and 3.3.4 but with probability limits given by $\boldsymbol{\delta^*}$ and $\boldsymbol{\gamma^*}$ rather than $\boldsymbol{\delta_0}$ and $\boldsymbol{\gamma_0}$ respectively.

Since $R(\mathbf{x}, w_g, \boldsymbol{\delta^*})$ and $G(\mathbf{x}, \boldsymbol{\gamma^*})$ can be any positive functions of $\mathbf{x}$ and w_g, one special case corresponds to them being constants. Since weighting by fixed constants does not affect the minimization problem, this implies that the unweighted estimator, denoted by $\hat{\boldsymbol{\theta}}_{\boldsymbol{g}}^{\boldsymbol{u}}$, which is a special case of the doubly weighted estimator, is also be consistent for $\boldsymbol{\theta_g^0}$.

The following theorem establishes consistency of the doubly weighted estimator under strong identification.

Theorem 3.5.4. *(Consistency under strong identification)*

Under assumptions 3.2.2, 3.2.3, 3.2.4, 3.5.1, 3.5.2 and 3.5.3 and assume regularity conditions (1), (2) and (3) of Theorem 3.4.1. Then,

$$\hat{\boldsymbol{\theta}}_{\boldsymbol{g}} \xrightarrow{p} \boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}} \text{ as } N \to \infty$$

where $\hat{\boldsymbol{\theta}}_{\boldsymbol{g}}$ *is the doubly-weighted estimator that solves problem 3.21.*

The next theorem states asymptotic normality of the doubly weighted estimator that solves the conditional M-estimation problem with misspecified probabilities.

Theorem 3.5.5. *(Asymptotic Normality)*

Under the assumptions of theorem 3.5.4 and the regularity conditions of theorem 3.4.2, we obtain

$$\sqrt{N}(\hat{\boldsymbol{\theta}}_{\boldsymbol{g}} - \boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}}) \xrightarrow{d} N\left(\mathbf{0}, \mathbf{H}_{\boldsymbol{g}}^{-1}\boldsymbol{\Omega}_{\boldsymbol{g}}\mathbf{H}_{\boldsymbol{g}}^{-1}\right)$$

where

$$\boldsymbol{\Omega}_{\mathbf{1}} = \mathbb{E}\left(\mathbf{l}_i\mathbf{l}_i'\right) \text{ and } \boldsymbol{\Omega}_{\mathbf{0}} = \mathbb{E}\left(\mathbf{k}_i\mathbf{k}_i'\right)$$

with $\mathbf{H}_{\boldsymbol{g}}$ *as defined in condition (4) of Theorem 3.4.2 and* $\mathbf{l}_i$ *and* $\mathbf{k}_i$ *defined as in condition 6) of Theorem 3.4.2 but with weights given by* $G(\mathbf{x}, \boldsymbol{\gamma}^*)$ *and* $R(\mathbf{x}, w_g, \boldsymbol{\delta}^*)$.

Substantively, there is no real difference in the proof of the above theorem except that now $\hat{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\delta}}$ are converging to probability limits that could be potentially different from those indexing the true treatment and missing outcome probabilities. A consequence of the objective function solving the conditional problem is reflected in the asymptotic variance expression above. Compared to section 3.4, $\boldsymbol{\Omega}_{\boldsymbol{g}}$ now is just the variance of the weighted score without the first stage adjustment. Since the conditional score of weighted problem is zero i.e. $\mathbb{E}\left[\boldsymbol{\nabla}_{\boldsymbol{\theta}_{\boldsymbol{g}}} q(y(g), \mathbf{x}, \boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}})' | \mathbf{x}\right] = \mathbf{0}$, this implies that the correlation between the

weighted score and the two MLE scores will be zero, giving us the familiar expression above. A consequence of this simpler expression for $\boldsymbol{\Omega_g}$ is that now estimating the probabilities in a first step is not any superior than using known weights. This is formalized in the following corollary.

Corollary 3.5.6. *(No gain with estimated weights under strong identification) Under the assumptions of theorem 3.5.5 we obtain*

$$\text{Avar}\left[\sqrt{N}\left(\tilde{\boldsymbol{\theta}}_{\boldsymbol{g}}-\boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}}\right)\right]=\text{Avar}\left[\sqrt{N}\left(\hat{\boldsymbol{\theta}}_{\boldsymbol{g}}-\boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}}\right)\right]=\mathbf{H}_{\boldsymbol{g}}^{-1}\boldsymbol{\Omega}_{\boldsymbol{g}}\mathbf{H}_{\boldsymbol{g}}^{-1}$$

where $\tilde{\boldsymbol{\theta}}_{\boldsymbol{g}}$ is the estimator that uses known (potentially misspecified) probabilities and $\hat{\boldsymbol{\theta}}_{\boldsymbol{g}}$ is the estimator that uses estimated probabilities.

This too is attributable to the conditional score of the weighted problem being zero, namely, $\mathbb{E}\left[\boldsymbol{\nabla}_{\boldsymbol{\theta_g}}q(y(g),\mathbf{x},\boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}})'|\mathbf{x}\right]=\mathbf{0}$.

A second interesting question concerns the role of weighting in this scenario. As I mentioned earlier, the unweighted estimator or in fact any weighted estimator with possibly misspecified probabilities will be consistent for $\boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}}$ (In fact the estimator that only weights by the propensity score will also be consistent in this case). Interestingly, if the objective function satisfies the generalized conditional information matrix equality (GCIME) defined below, the unweighted estimator is asymptotically more efficient than any weighted estimator. The following theorem formalizes this efficiency result

Theorem 3.5.7. *(Efficiency gain with unweighted estimator under GCIME)*

Under assumptions of theorem 3.5.5 if we additionally assume that the objective function satisfies the generalized conditional information matrix equality (GCIME) in the population which is defined as:

$$\mathbb{E}\left[\boldsymbol{\nabla}_{\boldsymbol{\theta_g}}q(y(g),\mathbf{x},\boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}})'\boldsymbol{\nabla}_{\boldsymbol{\theta_g}}q(y(g),\mathbf{x},\boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}})|\mathbf{x}\right]=\sigma_{0g}^{2}\cdot\boldsymbol{\nabla}_{\boldsymbol{\theta_g}}^{2}\mathbb{E}[q(y(g),\mathbf{x},\boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}})|\mathbf{x}]=\sigma_{0g}^{2}\cdot\mathbf{A}(\mathbf{x}_i,\boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}}) \tag{3.22}$$

where

$$\boldsymbol{\nabla}_{\boldsymbol{\theta_g}}^{2}\mathbb{E}[q(y(g),\mathbf{x},\boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}})|\mathbf{x}]=\mathbf{A}(\mathbf{x}_i,\boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}})$$

Then,

$$\text{Avar}\left[\sqrt{N}\left(\hat{\boldsymbol{\theta}}_{\boldsymbol{g}}-\boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}}\right)\right]=\mathbf{H}_{\boldsymbol{g}}^{-1}\boldsymbol{\Omega}_{\boldsymbol{g}}\mathbf{H}_{\boldsymbol{g}}^{-1}$$

where

$$\mathbf{H_1}=\mathbb{E}\left[\frac{r(\mathbf{x}_i,w_{i1})\cdot p_1(\mathbf{x}_i)}{R(\mathbf{x}_i,w_{i1},\boldsymbol{\delta}^*)\cdot G(\mathbf{x}_i,\boldsymbol{\gamma}^*)}\mathbf{A}(\mathbf{x}_i,\boldsymbol{\theta}_{\mathbf{1}}^{\mathbf{0}})\right],\mathbf{H_0}=\mathbb{E}\left[\frac{r(\mathbf{x}_i,w_{i0})\cdot p_0(\mathbf{x}_i)}{R(\mathbf{x}_i,w_{i0},\boldsymbol{\delta}^*)\cdot(1-G(\mathbf{x}_i,\boldsymbol{\gamma}^*))}\mathbf{A}(\mathbf{x}_i,\boldsymbol{\theta}_{\mathbf{0}}^{\mathbf{0}})\right]$$

$$\boldsymbol{\Omega_1}=\sigma_{01}^2\cdot\mathbb{E}\left[\frac{r(\mathbf{x}_i,w_{i1})\cdot p_1(\mathbf{x}_i)}{R^2(\mathbf{x}_i,w_{i1},\boldsymbol{\delta}^*)\cdot G^2(\mathbf{x}_i,\boldsymbol{\gamma}^*)}\mathbf{A}(\mathbf{x}_i,\boldsymbol{\theta}_{\mathbf{1}}^{\mathbf{0}})\right]$$

$$\boldsymbol{\Omega_0}=\sigma_{00}^2\cdot\mathbb{E}\left[\frac{r(\mathbf{x}_i,w_{i0})\cdot p_0(\mathbf{x}_i)}{R^2(\mathbf{x}_i,w_{i0},\boldsymbol{\delta}^*)\cdot(1-G(\mathbf{x}_i,\boldsymbol{\gamma}^*))^2}\mathbf{A}(\mathbf{x}_i,\boldsymbol{\theta}_{\mathbf{0}}^{\mathbf{0}})\right]$$

and

$$\text{Avar}\left[\sqrt{N}\left(\hat{\boldsymbol{\theta}}_{\boldsymbol{g}}^{\boldsymbol{u}}-\boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}}\right)\right]=(\mathbf{H}_{\boldsymbol{g}}^{\mathbf{u}})^{-1}\boldsymbol{\Omega}_{\boldsymbol{g}}^{\mathbf{u}}(\mathbf{H}_{\boldsymbol{g}}^{\mathbf{u}})^{-1}$$

where

$$\mathbf{H}_{\mathbf{1}}^{\mathbf{u}}=\mathbb{E}\left[r(\mathbf{x}_i,w_{i1})\cdot p_1(\mathbf{x}_i)\cdot\mathbf{A}(\mathbf{x}_i,\boldsymbol{\theta}_{\mathbf{1}}^{\mathbf{0}})\right],\mathbf{H}_{\mathbf{0}}^{\mathbf{u}}=\mathbb{E}\left[r(\mathbf{x}_i,w_{i0})\cdot p_0(\mathbf{x}_i)\cdot\mathbf{A}(\mathbf{x}_i,\boldsymbol{\theta}_{\mathbf{0}}^{\mathbf{0}})\right]$$

$$\boldsymbol{\Omega}_{\mathbf{1}}^{\mathbf{u}}=\sigma_{01}^2\cdot\mathbb{E}\left[r(\mathbf{x}_i,w_{i1})\cdot p_1(\mathbf{x}_i)\cdot\mathbf{A}(\mathbf{x}_i,\boldsymbol{\theta}_{\mathbf{1}}^{\mathbf{0}})\right],\boldsymbol{\Omega}_{\mathbf{0}}^{\mathbf{u}}=\sigma_{00}^2\cdot\mathbb{E}\left[r(\mathbf{x}_i,w_{i0})\cdot p_0(\mathbf{x}_i)\cdot\mathbf{A}(\mathbf{x}_i,\boldsymbol{\theta}_{\mathbf{0}}^{\mathbf{0}})\right]$$

Given the above,

$$\text{Avar}\left[\sqrt{N}\left(\hat{\boldsymbol{\theta}}_{\boldsymbol{g}}-\boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}}\right)\right]-\text{Avar}\left[\sqrt{N}\left(\hat{\boldsymbol{\theta}}_{\boldsymbol{g}}^{\boldsymbol{u}}-\boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}}\right)\right]\ \textit{is positive semi-definite}$$

The proof of the above theorems is easy to establish and can be found in the appendix. The GCIME assumption is known in a variety of estimation contexts. In the case of full maximum likelihood, GCIME holds for $q(y(g),\mathbf{x},\boldsymbol{\theta}_{\boldsymbol{g}})=-\ln f(y(g)|\mathbf{x},\boldsymbol{\theta}_{\boldsymbol{g}})$ where $f(\cdot)$ is the true conditional density of $y(g)$ with $\sigma_{0g}^2=1$. For the case of quasi maximum likelihood in the linear exponential family for estimating the true conditional mean parameters, GCIME holds for the same $q(\cdot)$ but $f(\cdot)$ now denotes a density from the linear exponential family with $Var(y(g)|\mathbf{x})=\sigma_{0g}^2\cdot v[m_g(\mathbf{x},\boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}})]$. In other words, GCIME will be satisfied for the QMLE case when $Var(y(g)|\mathbf{x})$ satisfies the generalized linear model assumption, irrespective of whether the higher order moments of the conditional distribution of $y(g)$ correspond to the chosen LEF density or not. For estimation using NLS, GCIME will hold for $q(y(g),\mathbf{x},\boldsymbol{\theta}_{\boldsymbol{g}})=(y(g)-m_g(\mathbf{x},\boldsymbol{\theta}_{\boldsymbol{g}}))^2$ with the homoskedasticity assumption. Hence in all these cases the

unweighted estimator will be more efficient than its weighted counterpart. But when GCIME is not satisfied, the two cannot be ranked efficiency wise.

The next section uses the results discussed in this section and section 3.4 to characterize the nature of this robustness property with explicit focus on two important causal parameters; the ATE and QTEs. Before this, a flowchart on the next page outlines the different cases of misspecification that are possible under the doubly weighted framework.

3.6 Robust estimation

The asymptotic theory developed in sections 3.4 and 3.5 can now be used to characterize the robustness property of the doubly weighted estimator. Delineating the asymptotic theory using the weak and strong identification assumptions helps me to be precise about the nature of this robustness and its constituents.

3.6.1 Average treatment effect

The most common parameter of interest in applied work is the ATE, defined for an underlying population of interest.[21] Given the importance of this parameter in applied work, I discuss how the current framework allows robust estimation of the ATE. Depending on the component of the doubly weighted framework that is allowed to be misspecified, I will utilize the asymptotic results from sections 3.4 and 3.5 along with certain estimation methods to establish consistent estimation of the ATE.

In the presence of covariates, $\mathbf{x}$, that are predictive of the potential outcomes, it is helpful to define the average treatment effect (ATE) as

$$\tau_{\text{ate}} = \mathbb{E}\left[\mu_1(\mathbf{x})\right] - \mathbb{E}\left[\mu_0(\mathbf{x})\right]$$

where $\mu_g(\mathbf{x})$ denotes the true conditional mean (or regression function) of $y(g)$. Let $m_g(\mathbf{x}, \boldsymbol{\theta_g})$

[21] For instance, in the NSW program which is the main empirical application in this paper, I define ATE to be the expectation over the population of all eligible participants.

be a parametric model for $\mathbb{E}\left[y|\mathbf{x}, w_g = 1\right]$.[22] Then this model is said to be correctly specified if

$$\mu_g(\mathbf{x}) = m_g(\mathbf{x}, \boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}}), \text{ for some } \boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}} \in \boldsymbol{\Theta_g} \tag{3.23}$$

Given the parametric nature of this framework, I acknowledge and tackle misspecification of the conditional mean model, $m_g(\mathbf{x}, \boldsymbol{\theta_g})$, the propensity score model, $G(\mathbf{x}, \boldsymbol{\gamma})$, and the missing outcomes probability model, $R(\mathbf{x}, w_g, \boldsymbol{\delta})$. While the discussion in this section focuses on consistent estimation of $\boldsymbol{\theta}_{\mathbf{1}}^{\mathbf{0}}$, an analogous argument can be made for estimating $\boldsymbol{\theta}_{\mathbf{0}}^{\mathbf{0}}$. The first case considers correct specification of the missing outcomes probability model,

Case 1: Correct missing probability model, $\boldsymbol{R(\mathbf{x}, w_g, \delta)}$

In the current framework, when $R(\cdot)$ is correctly specified, one obtains the usual double robustness (DR) result of causal inference. DR ensures that $\boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}}$ is estimated consistently despite having either the propensity score or the conditional mean model being misspecified, but not both. Naturally, what $\boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}}$ represents in this case will depend on what is being assumed about the conditional mean model. However, I will show under each of these cases, a consistent estimate of ATE can always be obtained.

a. First half of DR: Correct conditional mean, $\mathbb{E}\left(\boldsymbol{y(g)}|\mathbf{x}\right)$

Having a correctly specified mean model implies that I can decompose the potential outcomes into their true means as follows

$$\begin{aligned} y(g) &= m_g(\mathbf{x}, \boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}}) + u(g) \\ \mathbb{E}\left[u(g)|\mathbf{x}\right] &= 0 \end{aligned} \tag{3.24}$$

for both $g = 0, 1$. In this case we know there are many estimation methods that can consistently estimate $\boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}}$ such as Nonlinear least squares and QMLE in the linear exponential family. The question that remains to be addressed is whether any of these procedures require

[22] Under unconfoundedness, the regression function can be identified as $\mathbb{E}\left[\mu_g(\mathbf{x})\right] = \mathbb{E}\left[y|\mathbf{x}, w_g = 1\right], \forall\ g \in \{0, 1\}$

weighting to obtain consistent estimates of $\boldsymbol{\theta_g^0}$. To answer this, I look at these two estimation methods in detail and tie them to the theoretical results developed in earlier sections.

Solving for $\boldsymbol{\theta_g^0}$ using NLS means minimizing the expected squared error between $y(g)$ and $m_g(\mathbf{x}, \boldsymbol{\theta_g})$. In fact, under 3.24, $\boldsymbol{\theta_g^0}$ is identified in the stronger sense that it solves the conditional NLS problem,

$$\boldsymbol{\theta_g^0} = \underset{\boldsymbol{\theta_g} \in \boldsymbol{\Theta_g}}{\operatorname{argmin}} \, \mathbb{E}\left[(y(g) - m_g(\mathbf{x}, \boldsymbol{\theta_g}))^2 | \mathbf{x}\right] \tag{3.25}$$

Similarly, for estimation of $\boldsymbol{\theta_g^0}$ using QMLE in the linear exponential family (Gourieroux et al. (1984), Wooldridge (2010) chapter 13), if one chooses the range of the conditional mean function, $m_g(\mathbf{x}, \boldsymbol{\theta_g})$, to correspond with the range of the quasi-log likelihood for a given linear exponential density, $\boldsymbol{\theta_g^0}$ is again identified in the conditional sense,

$$\boldsymbol{\theta_g^0} = \underset{\boldsymbol{\theta_g} \in \boldsymbol{\Theta_g}}{\operatorname{argmin}} \, \mathbb{E}\left[\ln f(y(g), m_g(\mathbf{x}, \boldsymbol{\theta_g})) | \mathbf{x}\right]$$

where $f(\cdot)$ is the density associated with the chosen linear exponential distribution.[23] For both these examples, results from section 3.5 dictate that weighting by either correct or misspecified probabilities is not needed for consistency. The fact that one could weight by the misspecified propensity score model and still obtain this result is what forms the '*first part*' of the DR result with propensity score weighting.

Once $\hat{\boldsymbol{\theta}}_{\boldsymbol{g}}$ has been estimated by solving the sample version of the NLS or QMLE problem, ATE can be estimated as follows

$$\hat{\tau}_{\text{ate}} = \frac{1}{N}\sum_{i=1}^{N} m_1(\mathbf{x}_i, \hat{\boldsymbol{\theta}}_{\mathbf{1}}) - \frac{1}{N}\sum_{i=1}^{N} m_0(\mathbf{x}_i, \hat{\boldsymbol{\theta}}_{\mathbf{0}})$$

If in addition to having a correct conditional mean, I also assume the error variance of the outcomes is homoskedastic ($\mathbb{E}\left(u^2(g)|\mathbf{x}\right) = \sigma_{0g}^2$), then the estimator that does not weight at all may be the preferred estimator from an efficiency perspective. This result is due to GCIME being satisfied under homoskedasticity with NLS.

[23] For example, if $m_g(\mathbf{x}, \boldsymbol{\theta_g}) \in (0, 1)$, one would typically use the Bernoulli density, $f(y(g), m_g(\mathbf{x}, \boldsymbol{\theta_g})) = m_g(\mathbf{x}, \boldsymbol{\theta_g})^{y(g)} \cdot (1 - m_g(\mathbf{x}, \boldsymbol{\theta_g}))^{(1-y(g))}$.

b. Second part of DR: Correct propensity score model, $G(\mathbf{x}, \boldsymbol{\gamma})$

If one acknowledges misspecification of the conditional mean model, then this brings us to the second case of DR where only the propensity score model is assumed to be correct. In this case, there is no general way of consistently estimating the ATE. A very useful mean fitting property of QMLEs in the linear exponential family can be utilized here to obtain consistent estimators of the unconditional means, $\mathbb{E}\left[y(g)\right]$, despite misspecification of $m_g(\mathbf{x}, \boldsymbol{\theta_g})$.[24] The estimation strategy is to choose $m_g(\mathbf{x}, \boldsymbol{\theta_g})$, to be the inverse canonical link function, $h(\cdot)$, with the QLL corresponding to a choice of LEF density. In the generalized linear model (GLM) literature, the link function, $h^{-1}(\cdot)$, relates the mean of the distribution to a linear index

$$h^{-1}(\mu_g(\mathbf{x})) = \mathbf{x}\boldsymbol{\theta_g} \tag{3.26}$$

Then the first order conditions of such a QMLE problem give us

$$\mathbb{E}\left[\frac{\boldsymbol{\nabla_\theta} m_g(\mathbf{x}, \boldsymbol{\theta_g^*})' \cdot (y(g) - m_g(\mathbf{x}, \boldsymbol{\theta_g^*}))}{v(m_g(\mathbf{x}, \boldsymbol{\theta_g^*}))}\right] = \mathbf{0} \tag{3.27}$$

where $\boldsymbol{\theta_g^*}$ denotes the pseudo true parameter indexing the misspecified conditional mean model (White (1982)). By choosing the canonical link as the mean model of choice, the gradient in the numerator of 3.27 cancels with the variance term in the denominator. Note that this only occurs only when one uses the canonical link associated with the chosen LEF density and not with any other choice of link function. This in turn ensures that if one includes an intercept in $\mathbf{x}$, the model fits overall mean of the distribution (see Wooldridge (2010) chapter 13 for more detail) i.e.

$$\mathbb{E}\left[y(g)\right] = \mathbb{E}\left[m_g(\mathbf{x}, \boldsymbol{\theta_g^*})\right]$$

Under '*i.i.d*' sampling, solving the sample analogue of the population FOC given in 3.27 would have been sufficient to obtain consistent estimates of $\boldsymbol{\theta_g^*}$. However, in the presence of

[24] Słoczyński and Wooldridge (2018) use this mean fitting property for developing doubly robust estimators of various ATEs.

non-random assignment and missing outcomes, one needs to weight the first order conditions in 3.27 to ensure that $\boldsymbol{\theta}_{\boldsymbol{g}}^{*}$ is estimated consistently. In other words, one would solve the following moment conditions

$$\begin{aligned}\sum_{i=1}^{N} \frac{s_i \cdot w_{i1}}{R(\mathbf{x}_i, w_{i1}, \hat{\boldsymbol{\delta}}) \cdot G(\mathbf{x}_i, \hat{\boldsymbol{\gamma}})} \cdot \mathbf{x}_i' \cdot \left[y_i - h(\hat{\alpha}_1 + \mathbf{x}_i \hat{\boldsymbol{\beta}}_{\mathbf{1}}) \right] &= \mathbf{0} \\ \sum_{i=1}^{N} \frac{s_i \cdot w_{i0}}{R(\mathbf{x}_i, w_{i0}, \hat{\boldsymbol{\delta}}) \cdot (1 - G(\mathbf{x}_i, \hat{\boldsymbol{\gamma}}))} \cdot \mathbf{x}_i' \cdot \left[y_i - h(\hat{\alpha}_0 + \mathbf{x}_i \hat{\boldsymbol{\beta}}_{\mathbf{0}}) \right] &= \mathbf{0}\end{aligned} \tag{3.28}$$

The choice of the LEF density would have to be consistent with the range and nature of the outcome, $y(g)$.[25]

Estimation summary under second part of DR:

Estimation of the average treatment effect in the case of a misspecified mean model but correct propensity score and missing probability models follows in two steps:

1. Depending upon the range and nature of the outcome variable, $y(g)$, choose an appropriate LEF density. Choose the mean function, $m_g(\mathbf{x}, \boldsymbol{\theta}_{\boldsymbol{g}}) = h(\mathbf{x}\boldsymbol{\theta}_{\boldsymbol{g}})$, where $h(\cdot)$

[25] The following combinations of QLL and link functions produce the mean fitting property.

1. Normal log-likelihood with identity link function when there are no restrictions on the range of $y(g)$

$$\mathbb{E}\left[\mathbf{x}' \cdot (y(g) - \mathbf{x}\boldsymbol{\theta}_{\boldsymbol{g}}^{*}) \right] = \mathbf{0}$$

This is the first order condition for OLS which ensures that $\mathbb{E}[y(g)] = \mathbb{E}\left[\mathbf{x}\boldsymbol{\theta}_{\boldsymbol{g}}^{*} \right]$ if an intercept is included in the linear projection.

2. Poisson log-likelihood with log link function when the range of $y(g)$ is restricted to be non-negative ($y(g) \geq 0$)

$$\mathbb{E}\left[\frac{\exp(\mathbf{x}\boldsymbol{\theta}_{\boldsymbol{g}}^{*}) \cdot \mathbf{x}' \cdot (y(g) - \exp(\mathbf{x}\boldsymbol{\theta}_{\boldsymbol{g}}^{*}))}{\exp(\mathbf{x}\boldsymbol{\theta}_{\boldsymbol{g}}^{*})} \right] = \mathbf{0}$$

3. Bernoulli log likelihood with logit link function when $y(g)$ is restricted to be in the unit interval, ($y(g) \in [0, 1]$)

$$\mathbb{E}\left[\frac{\frac{\exp(\mathbf{x}\boldsymbol{\theta}_{\boldsymbol{g}}^{*})}{(1+\exp(\mathbf{x}\boldsymbol{\theta}_{\boldsymbol{g}}^{*}))^2} \cdot \mathbf{x}' \cdot \left(y(g) - \frac{\exp(\mathbf{x}\boldsymbol{\theta}_{\boldsymbol{g}}^{*})}{1+\exp(\mathbf{x}\boldsymbol{\theta}_{\boldsymbol{g}}^{*})} \right)}{\frac{\exp(\mathbf{x}\boldsymbol{\theta}_{\boldsymbol{g}}^{*})}{(1+\exp(\mathbf{x}\boldsymbol{\theta}_{\boldsymbol{g}}^{*}))^2}} \right] = \mathbf{0}$$

is the inverse canonical link function associated with this chosen density. Using this combination of mean function and quasi-log-likelihood, use the moment conditions in 3.28 to obtain consistent estimates, $\hat{\boldsymbol{\theta}}_{\boldsymbol{g}}$.

2. Using estimates that solve problem 3.28, one can then obtain consistent estimates of the average treatment effect as follows

$$\hat{\tau}_{ate} = \frac{1}{N}\sum_{i=1}^{N} h(\hat{\alpha}_1 + \mathbf{x}_i\hat{\boldsymbol{\beta}}_{\mathbf{1}}) - \frac{1}{N}\sum_{i=1}^{N} h(\hat{\alpha}_0 + \mathbf{x}_i\hat{\boldsymbol{\beta}}_{\mathbf{0}}) \tag{3.29}$$

where $\hat{\alpha}_g$ and $\hat{\boldsymbol{\beta}}_{\boldsymbol{g}}$ are the solutions to 3.28. The formal proof of consistency for $\hat{\tau}_{\text{ate}}$ in this case is given in the appendix J, and follows in a manner similar to Negi and Wooldridge (2019).

Case 2: Misspecified missing outcomes probability model, $\boldsymbol{R(\mathbf{x}, w_g, \boldsymbol{\delta})}$

If the missing outcomes model is misspecified, then sufficient for consistent estimation of ATE is a strengthening of the identification assumption from 3.2.1 to 3.5.1. In other words, $\boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}}$ would index the true conditional mean function i.e. $\mathbb{E}\left(y(g)|\mathbf{x}\right) = m_g(\mathbf{x}, \boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}})$. Hence, misspecification in $R(\mathbf{x}, w_g, \boldsymbol{\delta})$ can be allowed in exchange for identification of $\boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}}$ in the conditional sense. For instance, estimation via NLS would imply that $\boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}}$ solves $\min_{\boldsymbol{\theta_g}\in\boldsymbol{\Theta_g}} \mathbb{E}\left[(y(g) - m_g(\mathbf{x}, \boldsymbol{\theta_g}))^2|\mathbf{x}\right]$ and similarly for the QMLE example
$\boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}} = \underset{\boldsymbol{\theta_g}\in\boldsymbol{\Theta_g}}{\text{argmin}}\ \mathbb{E}\left[\ln f(y(g), m_g(\mathbf{x}, \boldsymbol{\theta_g}))|\mathbf{x}\right]$.

To conclude, robust estimation of ATE under the doubly weighted framework can be achieved as follows: If the missing outcomes probability model $R(\cdot, \boldsymbol{\delta})$ is misspecified, then one can consistently estimate ATE when the conditional mean model is correct. However, if $R(\cdot, \boldsymbol{\delta})$ is correct, then one can estimate the ATE in the usual double robust manner i.e. misspecification may be allowed either in the propensity score model or the conditional mean model, but not both. Finally, if the conditional mean model is misspecified, then both the probability models, $G(\cdot, \boldsymbol{\gamma})$ (for propensity score) and $R(\cdot, \boldsymbol{\delta})$ (for missing outcomes proba-

bility) would need to be correct.

To illustrate robust estimation of the ATE using the proposed doubly weighted estimator and to study its finite sample behavior, the next section discusses a simulation study which considers the different cases of misspecification mentioned above.

3.6.2 Monte carlo evidence

To allow for possible misspecification of the regression functions $\mu_g(\mathbf{x})$, I simulate two binary potential outcomes generated using a probit as follows

$$y(g) = \begin{cases} 1, \ y^*(g) > 0 \\ 0, \ y^*(g) \leq 0 \end{cases}$$

$$y^*(g) = \ \mathbf{x}\boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}} + u(g)$$

Note that $\mathbf{x}$ includes an intercept. The linear index, $\mathbf{x}\boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}}$, is parameterized to have covariates be only mildly predictive of the potential outcomes with $R_0^2 = 0.19$ and $R_1^2 = 0.14$ in the population.[26] The two covariates and the two latent errors are drawn from two independent bivariate normal distributions as follows,

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \sim N\left(\begin{pmatrix} 1 \\ 2 \end{pmatrix}, \begin{pmatrix} 3 & 0.2 \\ 0.2 & 2 \end{pmatrix} \right) \text{ and } \begin{pmatrix} u(0) \\ u(1) \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.2 \\ 0.2 & 1 \end{pmatrix} \right) \tag{3.30}$$

The assignment and missing outcome mechanisms have been simulated to ensure that unconfoundedness and ignorability are satisfied

$$w_1 = \begin{cases} 1, \ w_1^* > 0 \\ 0, \ w_1^* \leq 0 \end{cases} \quad \text{and} \quad s = \begin{cases} 1, \ s^* > 0 \\ 0, \ s^* \leq 0 \end{cases} \tag{3.31}$$

[26] Here $\boldsymbol{\theta}_{\mathbf{0}}^{\mathbf{0}} = (0, 1, 1)'$ and $\boldsymbol{\theta}_{\mathbf{1}}^{\mathbf{0}} = (-1, 1, 1)'$. With cross sectional data, covariates are typically seen to be mildly predictive of the outcome. For example, in the National Supported Work dataset from Calónico and Smith (2017), baseline factors explain about 26-50 percent of the variation in the non-experimental sample and about .04-2 percent in the experimental sample depending upon the included subset of covariates.

where

$$w_1^* = \mathbf{x}\boldsymbol{\gamma_0} + \xi \qquad\qquad s^* = \mathbf{z}\boldsymbol{\delta_0} + \zeta$$

with the errors ξ and ζ drawn from two independent standard logistic distributions.[27] Misspecification in these models is allowed in both the functional form and linear index dimension where for the misspecified cases, I estimate a probit with x_1 omitted from the linear index. For scenarios where the conditional mean is misspecified, I estimate a linear model with a correct index. The parameters, $\boldsymbol{\gamma_0}$ and $\boldsymbol{\delta_0}$, indexing the assignment and missingness mechanisms have been chosen to ensure average propensity of assignment to be 0.41 and average propensity of being observed to be 0.38.[28] The missing data have been simulated to imitate empirical settings where a significant portion of the outcomes are missing. Table **??** gives an estimation summary for the eight different cases of misspecification that are considered here

Results

I discuss results for cases (4) and (5) as these two scenarios are highlighted in sections 3.4 and 3.5. Finally, I also discuss case (8). Even though the theory developed in this paper is silent when all three components of the framework are misspecified, the simulation results look promising. All other cases are given in the appendix. Case (4) depicts the possibility that the conditional mean model is correct but both probability models are misspecified. For this case, one can see that weighting does not have any added bite in resolving the identification problem, beyond that already achieved from having a correct mean function. In figure H.2 d), the empirical distributions of the estimated ATE for the unweighted, propensity score wighted and double weighted estimators all coincide. Moreover, all are centered on the true ATE. In terms of root mean squared error, all three perform the same for a sample size of 5000 but PS-weighting performs better when the sample size is 1000. This suggests

[27] This implies that $p(w_1 = 1|\mathbf{x}) = p_1(\mathbf{x}) = \Lambda(\mathbf{x}\boldsymbol{\gamma_0})$ and $p(s = 1|w_g, \mathbf{x}) = r(w_g, \mathbf{x}) = \Lambda(\mathbf{z}\boldsymbol{\delta_0})$ where $\Lambda(\cdot)$ is the standard logistic CDF.

[28] Here $\boldsymbol{\gamma_0} = (0.05, \text{-}0.2, \text{-}0.11)'$, $\boldsymbol{\delta_0} = (0.01, 0.03, 0.05, \text{-}0.28)'$ and $\mathbf{z} = (1, w_g, x_1, x_2)$

that PS-weighting could be beneficial in terms of RMSE, at-least for small sample sizes. Estimating the propensity score reduces the variance of the weighted score of the problem which will not necessarily be the case when estimating both probability models. So, it might be better to use the propensity score weighted estimator in the case when the conditional mean function is correctly specified.

Case (5) considers a scenario where the mean function is misspecified but the two probabilities models are correct. This is the principal case covered in section 3.4 where weighting has a crucial role to play. As one can see, the average bias in the unweighted estimator of ATE is higher than for the doubly weighted estimator. In fact, the empirical distribution of the unweighted estimator is shifted to the right whereas for the doubly weighted estimator is centered on the truth (refer figure **??**). In this case, the doubly weighted estimator has both; the smallest Bias and Rmse amongst all three estimators. Under this case, I also consider the doubly weighted estimator which uses known weights (see table **??** for reference). In finite samples, estimation of the weights could result in conservative variance estimates. While estimating the weights would result in a smaller residual of the weighted score ($\mathbf{l}_i$), the residual variance could be larger compared to the known weights estimator because of non-zero cross correlations between the probability scores.

Finally, case (8) considers the scenario where all components of the framework are misspecified. The theory in this paper does not address this case. However, this is an interesting possibility given that misspecification of all components is a valid concern. The simulation results do offer some insight here. The doubly weighted estimator seems to be the only estimator that delivers the true ATE on average whereas the others are away from the truth (see table **??**).

3.6.3 Quantile effects

Under treatment effect heterogeneity, distributional impacts beyond the ATE are of increasing interest to researchers, especially in program evaluation studies. However, unlike the case

of ATE, it is generally not possible to obtain robust estimation of UQTE$_\tau$.[29] In this section, I employ the double weighting framework to focus attention on estimating three different quantile effects, namely, UQTE$_\tau$, CQTE$_\tau$, and a weighted linear approximation (LP) to the true CQTE$_\tau$. Whether $\boldsymbol{\theta_g^0}$ indexes the true CQF or an approximation will depend on what is being assumed about the conditional quantile model and the estimation method used.

Let us assume that the two potential outcomes are continuous in $\mathfrak{R}$ and that the unconditional quantiles of $y(g)$ are unique and do not have any flat spots at the τ^{th} quantile. Then, the conditional quantiles of $y(0)$ and $y(1)$ given covariates, $\mathbf{x}$, are defined as:

$$Quant_\tau\left(y(g)|\mathbf{x}\right) = \inf\{y : F_{y(g)}(y|\mathbf{x}) \geq \tau\};\ \text{where } 0 < \tau < 1$$

where $F_{y(g)}(y(g)|\mathbf{x})$ is the distribution function of $y(g)$ conditional on $\mathbf{x}$ and is assumed to have density $f(y(g)|\mathbf{x})$. Then, the CQTE$_\tau$ at $\mathbf{x} = \mathbf{x_0}$ for the τ^{th} quantile is defined as the difference in the conditional quantiles of the two outcome distributions i.e.

$$CQTE_\tau(\mathbf{x_0}) = Quant_\tau(y(1)|\mathbf{x_0}) - Quant_\tau(y(0)|\mathbf{x_0})$$

Similarly, UQTE$_\tau$ is defined as the difference in the τ^{th} unconditional quantiles of the two outcome distributions.

$$UQTE_\tau = Quant_\tau(y(1)) - Quant_\tau(y(0))$$

Let $quant_{g,\tau}(\mathbf{x}, \boldsymbol{\theta_g})$ be a model for the τ^{th} conditional quantile of $y(g)$. This is said to be correctly specified for $Quant_\tau(y(g)|\mathbf{x})$ if

$$Quant_\tau(y(g)|\mathbf{x}) = quant_{g,\tau}(\mathbf{x}, \boldsymbol{\theta_g^0}) \text{ for some } \boldsymbol{\theta_g^0}(\tau) \in \boldsymbol{\Theta_g},\ g = 0, 1 \tag{3.32}$$

The next section discusses estimation under the first case when we have $R(\cdot)$ correctly specified.

Case 1: Correct missing probability model, $\boldsymbol{R(\mathbf{x}, w_g, \delta)}$

Similar to the ATE case, when $R(\cdot)$ is correctly specified, one obtains the nested DR result

[29] This is because averaging the CQTE$_\tau$ does not give us the UQTE$_\tau$.

of causal inference. However, the parameter estimable in each case depends on what is being assumed about the CQF. To consider each of these scenarios in detail, consider the first half of DR when we have a correct CQF.

a. First half of DR: Correct conditional quantiles, $\boldsymbol{quant_{g,\tau}}\left(\mathbf{x}, \boldsymbol{\theta_g}\right)$

If the CQF is correctly specified, as defined in 3.32, then one can decompose the potential outcomes as,

$$\begin{aligned} y(g) &= quant_{g,\tau}(\mathbf{x}, \boldsymbol{\theta_g}) + u_\tau(g) \\ Quant_\tau\left(u_\tau(g)|\mathbf{x}\right) &= 0 \end{aligned} \tag{3.33}$$

In this case there are two estimation methods that will ensure consistent estimation of the correct CQF parameters, $\boldsymbol{\theta_g^0}(\tau)$. The first is quantile regression (QR) of Koenker and Bassett (1978). The second is a class of quasi maximum likelihood estimators in a special 'tick-exponential' family of distributions proposed by Komunjer (2005). This method is analogous to estimation of correctly specified conditional mean parameters using QMLE in the linear exponential family. The '*first part*' of this double robustness result implies that any inverse propensity score weighted version of the QR or QML objective functions, irrespective of whether those weights are correct, will also deliver a consistent and $\sqrt{N}$-asymptotically normal estimator of $\boldsymbol{\theta_g}(\tau)$.

For estimation that uses QR, correct specification as given in 3.33 implies that $\boldsymbol{\theta_g}(\tau)$ will actually solve the stronger conditional problem,

$$\boldsymbol{\theta_g^0}(\tau) = \underset{\boldsymbol{\theta_g} \in \boldsymbol{\Theta_g}}{\operatorname{argmin}}\ \mathbb{E}\left[c_\tau(y(g) - quant_{g,\tau}(\mathbf{x}, \boldsymbol{\theta_g}))|\mathbf{x}\right] \tag{3.34}$$

where $c_\tau(u) = u \cdot \left(\tau - \mathbb{1}[u < 0]\right)$ is the check function defined for some random variable u. Since, $\boldsymbol{\theta_g^0}(\tau)$ satisfies the stronger identification condition, results from section 3.5 can be applied. This means that weighting is not needed for consistent estimation of $\boldsymbol{\theta_g}(\tau)$, irrespective of whether the weighting functions are correct or not.

In a similar vein, estimation via QML using the tick exponential family implies that as long as CQF is correct,

$$\boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}}(\tau) = \underset{\boldsymbol{\theta}_{\boldsymbol{g}} \in \boldsymbol{\Theta}_{\boldsymbol{g}}}{\operatorname{argmin}} \mathbb{E}\left[\ln\left(\phi^{\tau}(y(g), quant_{\tau,g}(\mathbf{x}, \boldsymbol{\theta}_{\boldsymbol{g}}))\right)|\mathbf{x}\right] \tag{3.35}$$

where $\phi^{\tau}(\cdot,\cdot)$ is the density that belongs to the '*tick-exponential*' family characterized by:

$$\phi^{\tau}(y,\eta) = exp\left[-(1-\tau)[a(\eta)-b(y)]\mathbb{1}\{y \leq \eta\} + \tau[a(\eta)-c(y)]\mathbb{1}\{y > \eta\}\right]$$

and $\tau \in (0,1)$, $a(\cdot)$ is continuously differentiable and $b(\cdot)$ and $c(\cdot)$ are continuous functions such that $\eta \in M \subset \mathbb{R}$.[30]. Once we have obtained $\hat{\boldsymbol{\theta}}_{\boldsymbol{g}}$ either by solving the QR or QML problem, the conditional quantile treatment effect for the subgroup defined by $\mathbf{x}_i$ can be estimated as $\hat{CQTE}_{\tau}(\mathbf{x}_i) = quant_{1,\tau}(\mathbf{x}_i, \hat{\boldsymbol{\theta}}_{\mathbf{1}}) - quant_{0,\tau}(\mathbf{x}_i, \hat{\boldsymbol{\theta}}_{\mathbf{0}})$.

b. Second half of DR: Correct propensity score model, $\mathbf{G}(\mathbf{x}, \boldsymbol{\gamma})$

Suppose now we have a correctly specified propensity score model or equivalently a misspecified conditional quantile model. Traditionally, the theory of quantile estimation has not dealt with this case of misspecification.[31] However, Angrist et al. (2006b) establish an approximation property of QR with a misspecified linear CQF that is analogous to the approximation property of linear regression.[32] Hence, solving the QR objective function with $quant_{\tau,g}(\mathbf{x}, \boldsymbol{\theta}_{\boldsymbol{g}}) = \mathbf{x}\boldsymbol{\theta}_{\boldsymbol{g}}$ would still identify a weighted approximation to the CQF.

[30] $\phi^{\tau}(y,\eta)$ is a probability density and η is the τ-quantile of ϕ^{τ} such that $\int_{-\infty}^{\eta} \phi^{\tau}(y,\eta)dy = \tau$. Komunjer (2005) shows that if one chooses $a(\eta) = \frac{1}{\tau(1-\tau)} \cdot \eta$ and $b(y) = c(y) = \frac{1}{\tau(1-\tau)} \cdot y$, then the quasi log likelihood function is proportional to the check function that was originally introduced by Koenker and Bassett (1978)

[31] Kim and White (2003) establish consistency and asymptotic normality of the QR estimator for a pseudo true value in the case of a misspecified linear conditional quantile model.

[32] Adapting Angrist et al. (2006b)'s notation to the potential outcomes framework, the parameters that solve the QR problem solve a weighted mean square approximation to the true CQF,

$$\boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}}(\tau) = \underset{\boldsymbol{\theta}_{\boldsymbol{g}} \in \boldsymbol{\Theta}_{\boldsymbol{g}}}{\operatorname{argmin}} \mathbb{E}\left[\omega_{\tau}(\mathbf{x}, \boldsymbol{\theta}_{\boldsymbol{g}}) \cdot (Quant_{\tau}(y(g)|\mathbf{x}) - \mathbf{x}\boldsymbol{\theta}_{\boldsymbol{g}})^2\right]$$

where

$$\omega_{\tau}(\mathbf{x}, \boldsymbol{\theta}_{\boldsymbol{g}}) = \int_0^1 (1-u) f_{y(g)}(u \cdot \mathbf{x}\boldsymbol{\theta}_{\boldsymbol{g}} + (1-u) \cdot Quant_{\tau}(y(g)|\mathbf{x})|\mathbf{x})du$$

Under '*i.i.d*' sampling, solving the sample QR objective function is sufficient to obtain consistent estimates of $\boldsymbol{\theta_g^*}$. However, as in the case of ATE, weighting becomes crucial in the presence of non-random assignment and missing outcomes. In other words, one would need to weight the QR estimator with the correct propensity score and missing outcomes probability models to consistently estimate $\boldsymbol{\theta_g^*}$. For instance, one would now solve the following treatment and control group problems,

$$\begin{aligned} \min_{\boldsymbol{\theta_1}\in\boldsymbol{\Theta_1}} \sum_{i=1}^{N} \left[\frac{s_i \cdot w_{i1}}{R(\mathbf{x}_i, w_{i1}, \hat{\boldsymbol{\delta}}) \cdot G(\mathbf{x}_i, \hat{\boldsymbol{\gamma}})} \cdot c_\tau(y(1) - \mathbf{x}_i\boldsymbol{\theta_1}) \right] \\ \min_{\boldsymbol{\theta_0}\in\boldsymbol{\Theta_0}} \sum_{i=1}^{N} \left[\frac{s_i \cdot w_{i0}}{R(\mathbf{x}_i, w_{i0}, \hat{\boldsymbol{\delta}}) \cdot (1 - G(\mathbf{x}_i, \hat{\boldsymbol{\gamma}}))} \cdot c_\tau(y(0) - \mathbf{x}\boldsymbol{\theta_0}) \right] \end{aligned} \tag{3.36}$$

and the solution to these sample problems, $\hat{\boldsymbol{\theta}}_{\boldsymbol{g}}$, is interpretable as providing a weighted LP to the true CQTE$_\tau$.

Case 2: Misspecified missing outcomes probability model, $\boldsymbol{R(\mathbf{x}, w_g, \delta)}$

If the missing outcomes model is misspecified, then sufficient for consistent estimation of $\boldsymbol{\theta_g^0}$ is a strengthening of the identification condition from 3.2.1 to 3.5.1. For estimation via quantile regression, this means that $\boldsymbol{\theta_g^0}$ solves the conditional QR problem

$$\boldsymbol{\theta_g^0} = \underset{\boldsymbol{\theta_g}\in\boldsymbol{\Theta_g}}{\operatorname{argmin}} \; \mathbb{E}\left[c_\tau(y(g) - quant_{g,\tau}(\mathbf{x}, \boldsymbol{\theta_g}))|\mathbf{x}\right]$$

which will hold only when the conditional score of the check function is zero i.e.

$$\mathbb{E}\left[-\mathbf{x}'\left\{\tau \cdot \mathbb{1}[y(g) - quant_{g,\tau}(\mathbf{x}, \boldsymbol{\theta_g^0}) \geq 0] - (1-\tau) \cdot \mathbb{1}[y(g) - quant_{g,\tau}(\mathbf{x}, \boldsymbol{\theta_g^0}) < 0]\right\} \middle| \mathbf{x}\right] = \mathbf{0}$$

and this will be true only when $Quant_\tau(y(g)|\mathbf{x}) = quant_{g,\tau}(\mathbf{x}, \boldsymbol{\theta_g^0})$. So, misspecification in $R(\mathbf{x}, w_g, \boldsymbol{\delta})$ can be allowed in exchange for having a correctly specified conditional quantile model.

---

is the weighting function that determines the importance given by the minimizer, $\boldsymbol{\theta_g^0}$, to points in the support of $\mathbf{x}$.

Direct estimation of UQTE$_\tau$

As was mentioned earlier in this section, estimating UQTE$_\tau$ from CQTE$_\tau(\mathbf{x})$ is generally not possible even if we assume a correct model for the conditional quantiles of the outcomes. This is because the mean of the quantiles is not equal to the quantiles of the mean. Hence, one cannot obtain unconditional quantiles from averaging conditional quantiles over $\mathbf{x}$. In this case, one can directly estimate the marginal quantiles by running a quantile regression of $y(g)$ on an intercept (as shown in Firpo (2007)).[33] In the present case, one would weight the objective function by the two probabilities in the following manner,

$$\theta_1^0(\tau) = \underset{\theta_1 \in \Theta_1}{\text{argmin}}\ \mathbb{E}\left[\frac{s_i \cdot w_{i1}}{R(\mathbf{x}_i, w_{i1}, \hat{\boldsymbol{\delta}}) \cdot G(\mathbf{x}_i, \hat{\boldsymbol{\gamma}})} \cdot c_\tau(y(1) - \theta_1)\right]$$
$$\theta_0^0(\tau) = \underset{\theta_0 \in \Theta_0}{\text{argmin}}\ \mathbb{E}\left[\frac{s_i \cdot w_{i0}}{R(\mathbf{x}_i, w_{i0}, \hat{\boldsymbol{\delta}}) \cdot (1 - G(\mathbf{x}_i, \hat{\boldsymbol{\gamma}}))} \cdot c_\tau(y(0) - \theta_0)\right]$$

Weighting by $G(\cdot)$ and $R(\cdot)$ is crucial here since these primarily serve to remove the selection biases due to non-random assignment and missing data. Then, one can obtain the unconditional quantile treatment effect as,

$$UQTE_\tau = \theta_1^0(\tau) - \theta_0^0(\tau)$$

The next section explores estimation of these three quantile estimands using a Monte Carlo experiment where I allow misspecification of the weighting functions and the conditional quantile model.

3.6.4 Monte carlo evidence

To ensure that the marginal quantiles of the potential outcome distributions are unique with no flat spots, I simulate two continuous non-negative outcomes as follows,

$$y(g) = \exp(\mathbf{x}\boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}} + u(g)),\ \text{for } g = 0, 1$$

[33] Firpo (2007) uses propensity score weighting to directly estimate unconditional quantiles in the presence of non-random assignment.

where $\boldsymbol{\theta_1^0} = (0.1, -0.36, -0.1)'$ and $\boldsymbol{\theta_0^0} = (0.2, 0.24, -0.45)'$ are parameterized to ensure $R_0^2 = 0.15$ and $R_1^2 = 0.13$ in the population. The two covariates and the two latent errors are drawn from two independent normal distributions following eq (3.30). The missing outcomes and the treatment assignment mechanisms are also generated according to eq (3.31). Since $\exp(\cdot)$ is an increasing continuous function, the equivariance property of quantiles implies that

$$\begin{aligned} Quant_\tau(y(g)|\mathbf{x}) &= Quant_\tau\left[\exp(\mathbf{x}\boldsymbol{\theta_g^0} + u(g))|\mathbf{x}\right] \\ &= \exp\left[Quant_\tau(\mathbf{x}\boldsymbol{\theta_g^0} + u(g)|\mathbf{x})\right] \\ &= \exp\left[\mathbf{x}\boldsymbol{\theta_g^0} + Quant_\tau(u(g)|\mathbf{x})\right] \\ &= \exp\left[\mathbf{x}\boldsymbol{\theta_g^0} + \Phi^{-1}(\tau)\right] \end{aligned}$$

where $\Phi^{-1}(\tau)$ is the inverse standard normal CDF evaluated at τ. This equivariance property helps to characterize and estimate CQTE for cases when the CQF is correct. For brevity, I study the behavior of the unweighted, propensity score weighted and the double weighted estimators for only five of the eight cases of misspecification. These are enumerated in table **??** below. Out of these, I discuss cases 4, 5 and 8 in the main text and results for the rest can be found in the appendix. Cases 4 and 5 correspond to the scenarios for which results are derived in section 3.5 and 3.4 respectively. The last case corresponds to the scenario where all components of the doubly weighted framework are misspecified. Even though the theory in this paper does not address that specific case, the simulation results show that the proposed estimator has the lowest bias among all three.

Results

For the first case when the CQF is correctly specified, figure H.3 plots CQTE as a function of x_1 for the 25th quantile of the outcome distribution. Results for the 50th and 75th quantiles are given in the appendix. One can see that the estimated function coincides with the true CQTE.[34] To make this case interesting, I consider misspecification of both probability models. As the results in section 3.5 dictate, all three estimators; unweighted,

[34]For plotting these functions, I first collect the QR estimates that solve the unweighted, ps-weighted and double-weighted check function (defined 3.34) corresponding to a particular

ps-weighted and double-weighted will be consistent for the true CQTE because the CQF is correctly specified. Hence, misspecification of the two probability models does not affect consistent estimation of the estimand. In fact weighting by any positive function would deliver this result, including just the ps-weighted estimator.

Next, I consider the case when the CQF is misspecified. Using the results in Angrist et al. (2006b), I interpret the solution to the double-weighted problem given in eq 3.36 as providing a consistent weighted-linear projection to the true CQF. I use these linear projections to estimate an LP to the true CQTE. Figure H.4 plots the bias in the estimated LP relative to the true LP as a function of x_1 for the three estimators. In panel A) where both probability models are correct, the relative bias from the double-weighted estimator is the lowest and coincides with the line of no bias. Panel D) considers the case where all three parametric specifications are wrong. Again, we see that the double weighted estimator is performing the best in terms of bias. Even though the theory does not guide us here, double weighting seems to be the least biased procedure.

Finally, I consider direct estimation of the unconditional quantile treatment effect (UQTE) at the 25th quantile. Again, results for the 50th and 75th quantiles can be found in the appendix. Notice that estimation of UQTE does not require parametric specification of the CQF since it is the difference in marginal quantiles. Hence, the two probability models are the only relevant components of the framework that affect consistent estimation of UQTE. In the first case, when both probability models are correct, unweighted and double-weighted estimators are both close to the true quantile effect. For the second case where both probability models are misspecified, double weighting does a little worse than not weighting at all. However, the results at other quantile levels reflect more favorably upon double weighting. Propensity score weighting performs the worst in both cases suggesting instances where it

quantile level, $\tau \in \{0.25, 0.50, 0.75\}$ across 1000 Monte Carlo repetitions. I then draw a linearly spaced x_1 vector and simulate the CQTE using the 1000 estimated QR coefficients. Averaging these 1000 functions at each point on the x_1 vector gives me the estimated average CQTE function. I plot this along with the 1000 individual functions and the true CQTE, which is calculated using the population QR parameters, $\boldsymbol{\theta_g^0}$.

might not be better to just correct for nonrandom assignment. Tables below report the bias and Rmse of the three estimators along with the double weighted estimator that uses known probability weights. When the two probability models are correct, the double-weighted estimator has the lowest Rmse. This, however, ceases to be true when the two probabilities are misspecified.

3.7 Application to Calónico and Smith (2017)

In this section, I apply the proposed estimator to the Aid to Families with Dependent Children (AFDC) sample of women from the National Supported Work program compiled by Calónico and Smith (2017).[35] NSW was a transitional and subsidized work experience program which was implemented as a randomized experiment in the United States between 1975-1979. CS replicate LaLonde (1986)'s within-study analysis for the AFDC women in the program, where the purpose of such an analysis is to evaluate how training estimates obtained from using non-experimental identification strategies (for example, CIA) compare to experimental estimates. To compute the non-experimental estimates, CS combine the NSW experimental sample with two non-experimental comparison groups drawn from PSID, called PSID-1 and PSID-2.[36] In this paper, I utilize the within-study feature of this empirical application to estimate bias in the unweighted and propensity-score weighted estimates, relative to the proposed double weighting procedure.

To construct these measures, I augment the CS sample to allow for women who had missing earnings information in 1979. This renders 26% of the experimental and 11% of the PSID samples missing. I then combine the experimental treatment group of NSW with three distinct comparison groups present in the CS dataset, namely, the experimental control group, and the two PSID samples, to compute the unweighted, single-weighted and double-

[35]Henceforth, Calónico and Smith (2017) is referred as CS.

[36]The PSID-1 sample constructed by CS involves keeping all female household heads continuously from 1975-1979 who were between 20 and 55 years of age in 1975 and were not retired in 1975. The sample labeled PSID-2 further restricts PSID-1 to include only those women who received AFDC welfare in 1975.

weighted training estimates.[37] The difference in the non-experimental estimate, obtained from using the doubly weighted estimator, and the experimental estimate provides the first measure of estimated bias associated with the proposed strategy. Combining the experimental control group with the non-experimental comparison group gives a second measure of estimated bias (Heckman et al. (1998a)). Much like CS, I report both these measures across a range of regression specifications for the average training estimates.

Given the growing importance of estimating distributional impacts of training programs, I also estimate marginal quantile treatment effects at every 10th quantile of the 1979 earnings distribution. The role of double weighting for ensuring consistency of the estimates is highlighted for the case of estimating marginal quantiles where covariates, which primarily serve to remove biases arising from non-random assignment and missing outcomes, enter the estimating equation only through the two probability models.

3.7.1 Results

First, to evaluate whether women with missing earnings in 1979 were significantly different than those who were observed, Table I.17 reports the mean and standard deviation of the woman's age, years of schooling, pre-training earnings and other characteristics across the observed and missing samples. In terms of age, the women who were observed in the experimentally treated group of NSW and the PSID-1 sample were, on average, older than those who were missing. The observed women in PSID-1 were also more likely to be married. For the PSID-2 sample, women who were observed had, on average, more kids with higher pre-training earnings. Apart from these minor differences, the observed women did not appear to be systematically different that those who were missing, as measured through observable characteristics.

The presence of non-experimental control groups implies that assignment was nonrandom

[37]For details regarding sample construction, and other aspects of this application, see appendix G

and therefore an issue in the sample. This is because the comparison groups were drawn from PSID after imposing only a partial version of the full NSW eligibility criteria. Table I.16 provides descriptive statistics for the covariates, by the treatment status. As can be expected, the treatment and control groups of NSW are not observably different, indicating the strong role that randomization plays in producing comparable groups. In contrast, the women in PSID-1 and PSID-2 groups are statistically different than the treatment group members, implying substantial scope for nonrandom assignment.

3.7.1.1 Estimated bias for average and unconditional quantile training effects

Table I.18 reports the doubly-weighted, ps-weighted and unweighted average training estimates which use the three different comparison groups; NSW control, PSID-1 and PSID-2. The unweighted (unadjusted and adjusted) experimental estimates given in row 1, are same as the estimates reported by CS in Table 3 of their paper. Overall, one can see that the double weighted experimental estimates are more stable than the single weighted or unweighted estimates across the different regression specifications, with a range between \$824-\$828.

For computing the ps-weighted and double-weighted non-experimental estimates, I first trim the sample to ensure common support between the treatment and comparison groups.[38] This reduces the sample size from 1,248 to 1,016 observations for the PSID-1 estimates and from 782 to 720 observations for the PSID-2 estimates. A pattern that is consistent across the two sets of non-experimental estimates is that weighting gets us much closer to the benchmark relative to not weighting at all. For instance, the unweighted simple difference in means estimate of training, which uses the PSID-1 comparison group, is -\$799 whereas the weighted estimates are \$827 and \$803. For the PSID-2 comparison group, the unweighted estimate which controls for all covariates is \$335 whereas the weighted estimates are \$905 and \$904.

[38] Appendix G describes estimation of the two probability weights along with the sample trimming criteria.

The second panel of Table I.18 reports the bias in training estimates from combining the experimental control group with the PSID comparison groups. A similar pattern is seen here with weighted bias estimates being much closer to zero than the unweighted estimates. For instance, the double-weighted estimate that adjusts for all covariates using the PSID-1 comparison group is -$21 whereas the unweighted estimates is -$568. These results suggest that the argument for weighting is strong when using a non-experimental comparison group where nonrandom assignment and missing outcomes are significant problems.[39]

Figure H.1 plots the relative bias in UQTE estimates at every 10th quantile of the 1979 earnings distribution. Much like the average training estimates, we see that the weighted estimates consistently lie below the unweighted estimates for most quantiles, irrespective of whether we use the PSID-1 or PSID-2 non-experimental group. Note that I do not plot the UQTE estimates for quantiles less than 0.46, since these are all zero.[40]

This empirical application illustrates the role of proposed estimator in both experimental and observational data contexts. The comparison involving the treatment and control group of NSW demonstrates its use in an experiment with missing outcomes, whereas the non-experimental sample demonstrates its use in the more realistic observational data setting.

3.8 Conclusion

In empirical research, the problems of nonrandom assignment and missing outcomes threaten identification of causal parameters. This paper proposes a new class of consistent and asymptotically-normal estimators that address these two issues using a double inverse probability weighted procedure. The method combines propensity score weighting with weighting for missing data in a general M-estimation framework, which can be utilized to study a range of problems, such as ordinary least-squares, quasi Maximum likelihood, and

[39]Note that the large standard errors for the non-experimental estimates can be attributed to the small sample sizes and to the large residual variance of earnings in the PSID-1 and PSID-2 populations.

[40]There are a lot of women in the experimental and PSID samples with zero real earnings in 1979.

quantile regression. In addition, the proposed class is characterized by a *robustness* property, which makes it resilient to parametric misspecification of a conditional model of interest (CEF or CQF) and the two weighting functions.

As leading applications of this framework, the paper discusses robust estimation of ATE and QTEs. A Monte Carlo study indicates that the doubly weighted estimates of average and quantile effects have the lowest bias, compared to naive alternatives (unweighted or propensity-score weighted estimators), for interesting cases of misspecification. Finally, the estimator is applied to the data on AFDC women from the NSW program compiled by Calónico and Smith (2017). The presence of experimental and non-experimental comparison groups in this application help to quantify the estimated bias in the double-weighted training estimates. Results suggest that the argument for weighting is strong whenever nonrandom assignment and (or) missing outcomes are significant concerns. Since the severity and magnitude of bias introduced from ignoring either problem cannot be assessed ex-ante, a safe bet from the practitioner's perspective is to provide both weighted and unweighted causal effect estimates.

Practically, the doubly weighted estimator is easy to implement. Appendix F.3 provides an example code that uses Stata `gmm` command for implementing the double-weighted estimator of ATE. Computation of analytically correct standard errors, however, requires additional coding and is still a work in progress. Alternatively, one can use bootstrapped standard errors which will provide asymptotically correct inference.

Even though missing outcomes are a common concern in empirical analysis, it is equally common to encounter missing data on the covariates. A particularly important future extension will be to allow for missing data on both. In this case, using a generalized method of moments framework which incorporates information on complete and incomplete cases could provide efficiency gains over just using the observed data.

APPENDICES

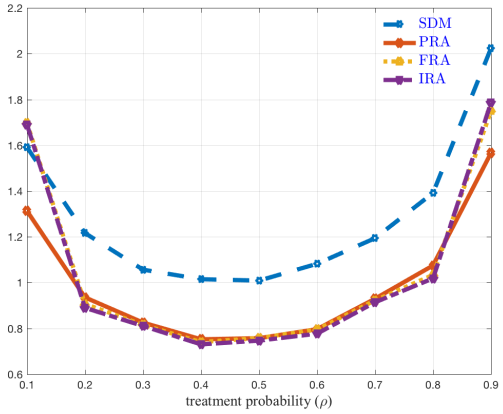
APPENDIX A

FIGURES FOR CHAPTER 1

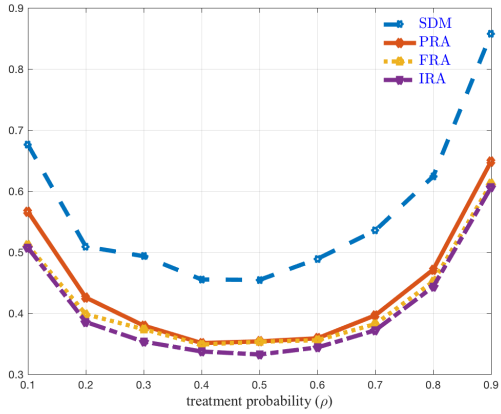
A.1 Root mean squared error across different sample sizes

Figure A.1: Quadratic design, continuous covariates (mild heterogeneity)

N=100



N=500



N=1000

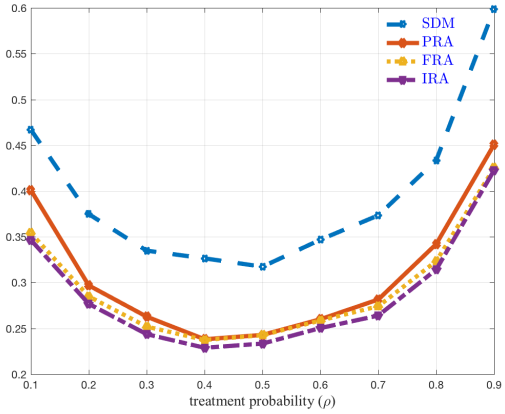
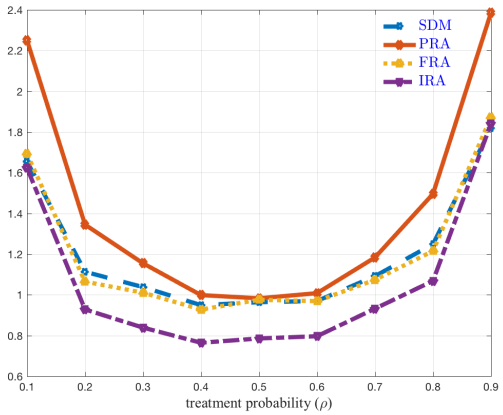
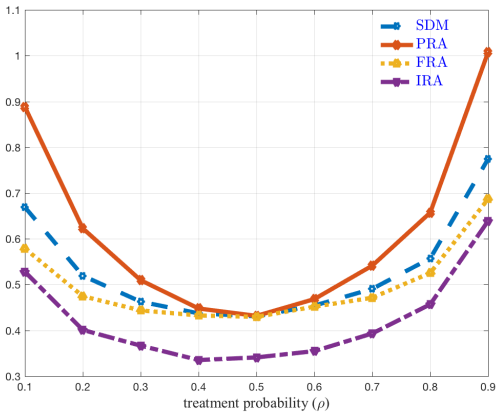


Figure A.2: Quadratic design, continuous covariates (strong heterogeneity)

N=100



N=500



N=1000

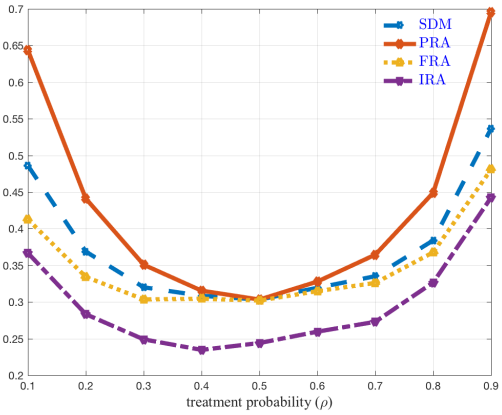
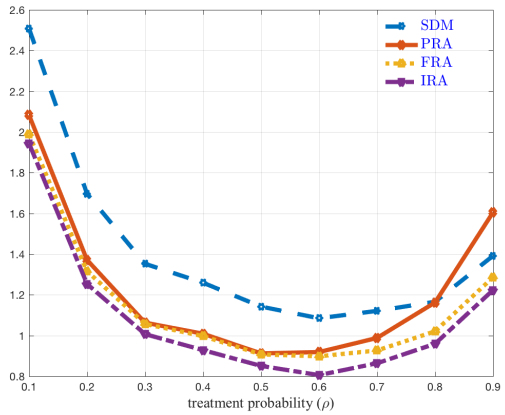
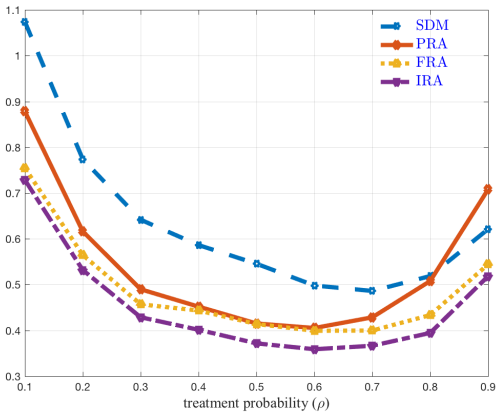


Figure A.3: Quadratic design, one binary covariate (mild heterogeneity)

N=100



N=500



N=1000

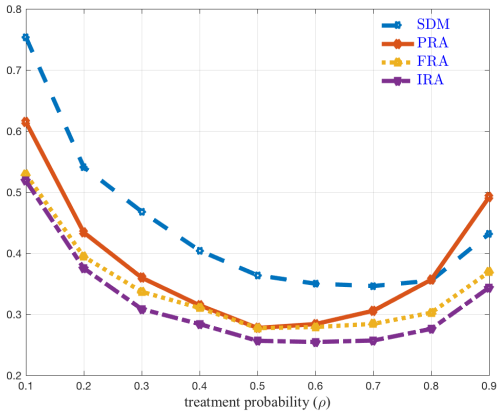
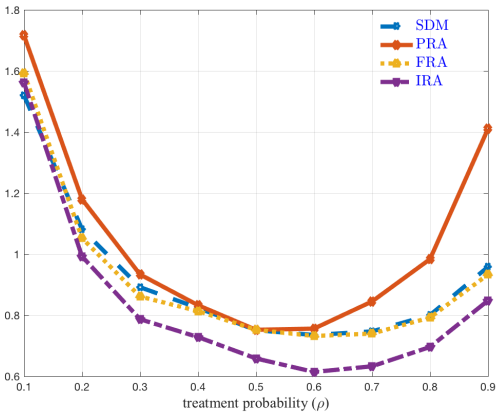
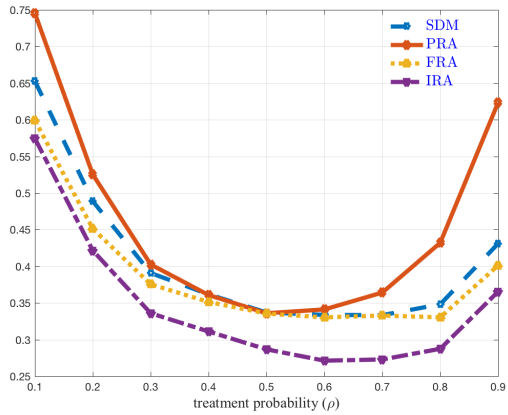


Figure A.4: Quadratic design, one binary covariate (strong heterogeneity)

N=100



N=500



N=1000

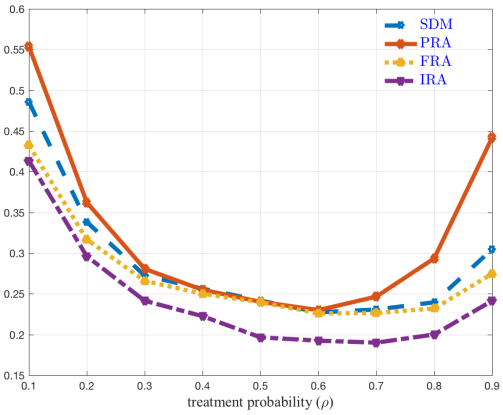
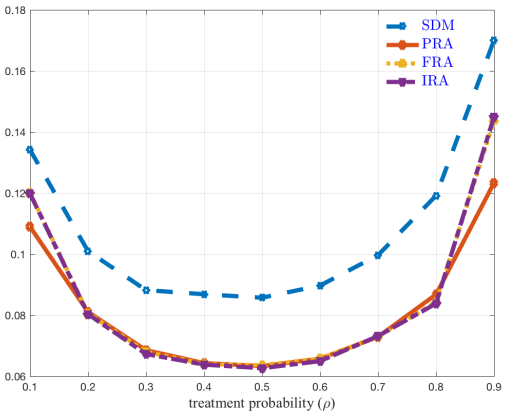
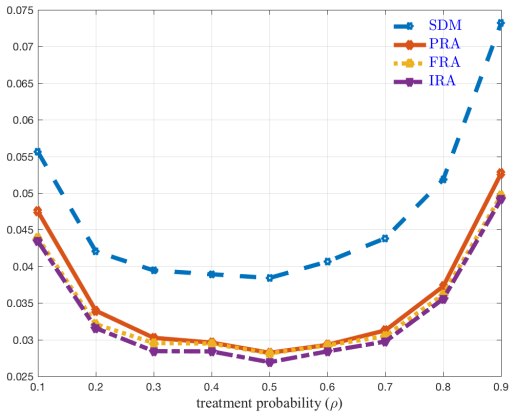


Figure A.5: Probit design, continuous covariates (mild heterogeneity)

N=100



N=500



N=1000

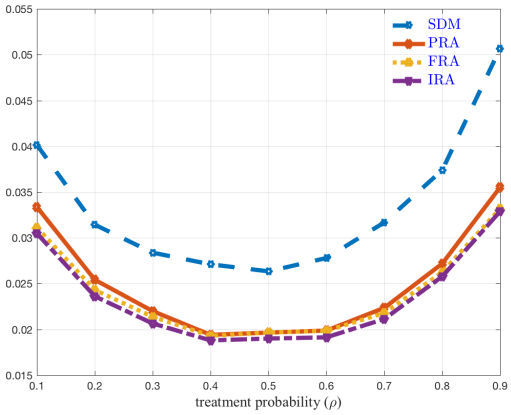
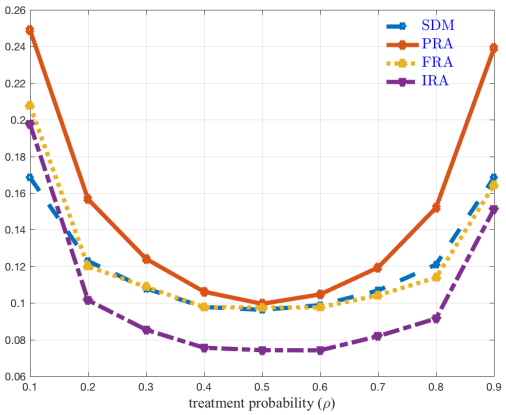
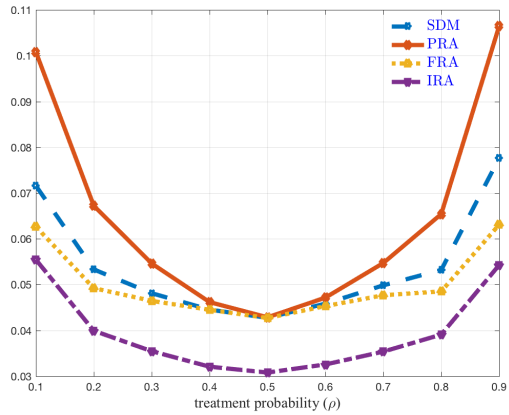


Figure A.6: Probit design, continuous covariates (strong heterogeneity)

N=100



N=500



N=1000

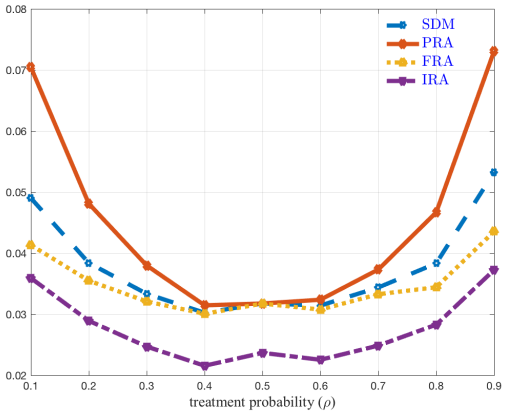
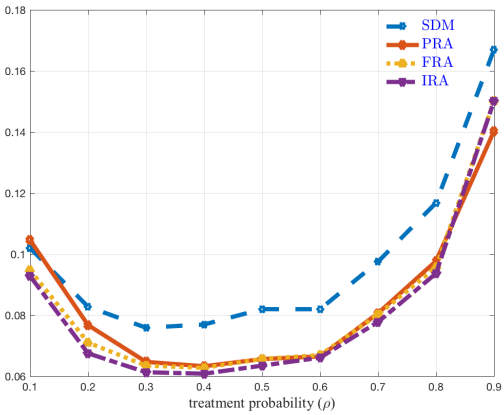
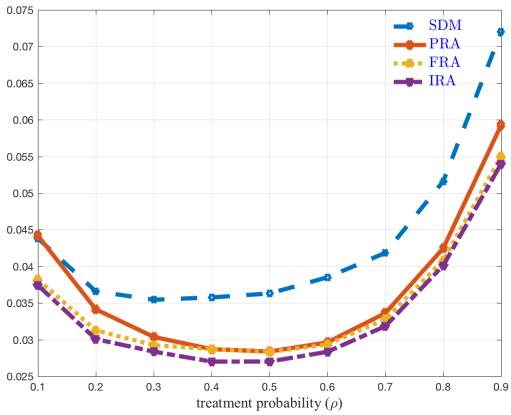


Figure A.7: Probit design, one binary covariate (mild heterogeneity)

N=100



N=500



N=1000

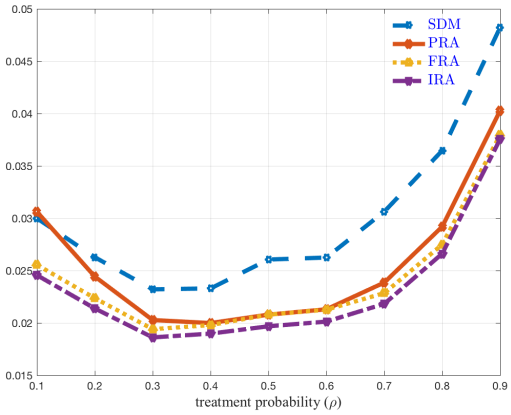
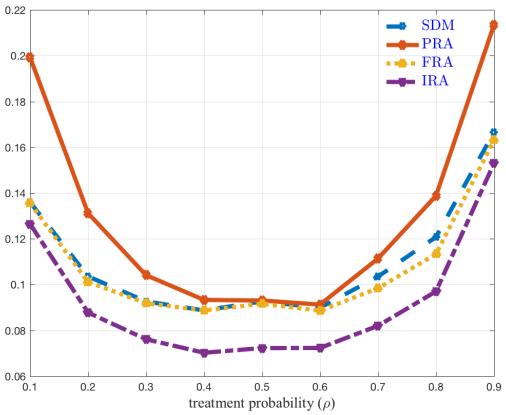
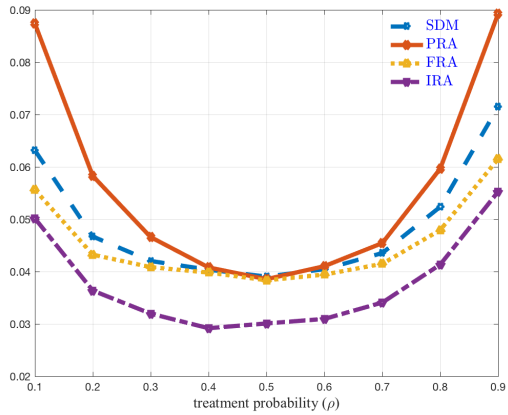


Figure A.8: Probit design, one binary covariate (strong heterogeneity)

N=100



N=500



N=1000

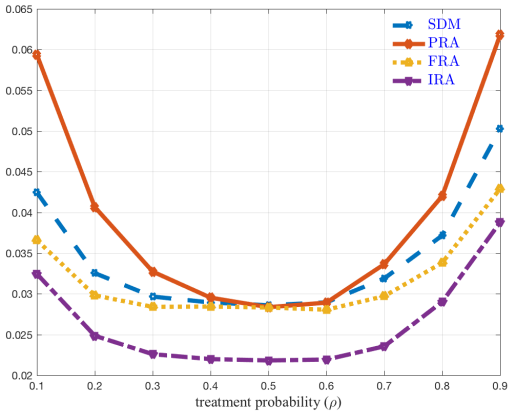
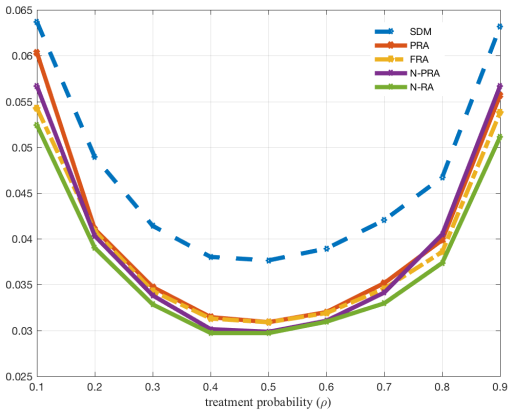


Figure A.9: Binary outcome, bernoulli QLL with logistic mean

N=500



N=1000

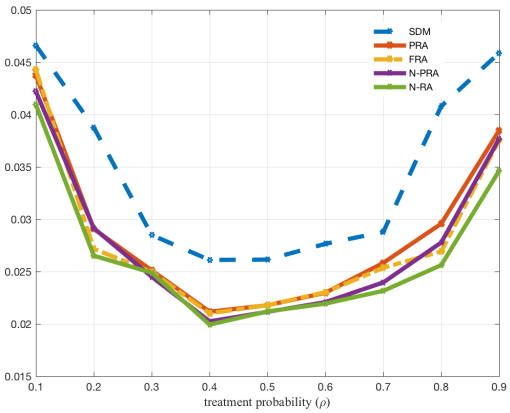
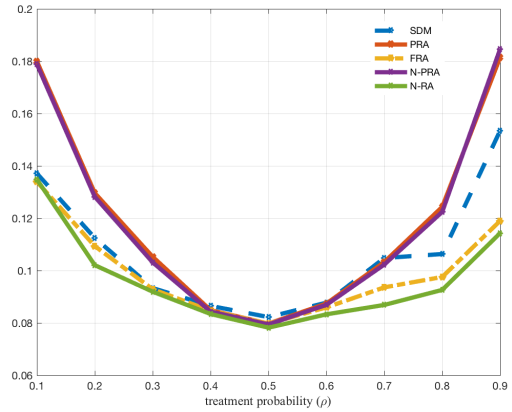
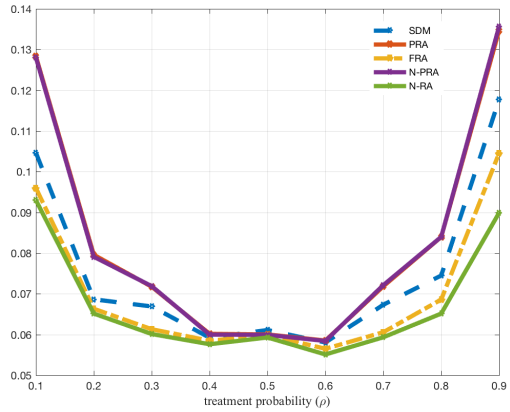


Figure A.10: Non-negative outcome, poisson QLL with exponential mean

N=500



N=1000



APPENDIX B

TABLES FOR CHAPTER 1

Table B.1: QLL and mean function combinations

Restrictions on support of response	**Quasi-Log Likelihood Function**	**Conditional Mean Function**
None	Gaussian (Normal)	Linear
$Y(w) \in [0, 1]$	Bernoulli	Logistic
$Y(w) \in [0, B]$	Binomial	Logistic
$Y(w) \geq 0$	Poisson	Exponential
$Y_g(w) \geq 0$, $\sum_{g=0}^{G} Y_g(w) = 1$	Multinomial	Logistic

Table B.2: Bias and standard deviation for N=100

					DGP1					
	0.1		0.3		0.5		0.7		0.9	
Estimator	**bias**	**std**	**bias**	**std**	**bias**	**std**	**bias**	**std**	**bias**	**std**
SDM	0.045	1.590	0.025	1.056	0.034	1.008	-0.035	1.194	0.070	2.023
PRA	0.047	1.312	-0.022	0.825	0.039	0.756	-0.031	0.929	0.039	1.566
FRA	0.017	1.697	-0.023	0.815	0.042	0.757	-0.022	0.922	-0.021	1.750
IRA	0.004	1.690	-0.014	0.810	0.026	0.746	-0.025	0.914	-0.022	1.786
					DGP2					
SDM	0.045	1.590	0.025	1.056	0.034	1.008	-0.035	1.194	0.070	2.023
PRA	0.047	1.312	-0.022	0.825	0.039	0.756	-0.031	0.929	0.039	1.566
FRA	0.017	1.697	-0.023	0.815	0.042	0.757	-0.022	0.922	-0.021	1.750
IRA	0.004	1.690	-0.014	0.810	0.026	0.746	-0.025	0.914	-0.022	1.786
					DGP3					
SDM	-0.058	2.508	-0.085	1.350	0.041	1.141	-0.038	1.120	-0.039	1.391
PRA	-0.045	2.083	-0.069	1.061	0.054	0.910	0.038	0.987	-0.094	1.602
FRA	0.051	1.988	-0.100	1.052	0.030	0.907	0.003	0.926	0.043	1.286
IRA	0.046	1.944	-0.085	1.003	0.019	0.850	0.014	0.864	0.072	1.221
					DGP4					
SDM	0.094	1.517	-0.040	0.891	0.005	0.751	0.014	0.747	-0.031	0.958
PRA	0.013	1.716	-0.047	0.932	0.007	0.752	0.004	0.845	-0.034	1.410
FRA	0.042	1.593	-0.050	0.860	0.002	0.752	0.015	0.739	0.058	0.931
IRA	0.022	1.561	-0.072	0.783	0.003	0.658	0.003	0.632	0.019	0.848
					DGP5					
SDM	0.002	0.134	0.002	0.088	-0.002	0.086	-0.002	0.100	0.003	0.170
PRA	0.003	0.109	-0.001	0.069	0.000	0.063	0.000	0.073	0.003	0.123
FRA	0.025	0.117	0.003	0.068	0.000	0.064	0.000	0.073	-0.001	0.144
IRA	0.026	0.117	0.005	0.067	0.001	0.063	0.002	0.073	-0.001	0.145
					DGP6					
SDM	-0.002	0.169	0.000	0.108	0.000	0.096	0.001	0.107	0.003	0.168
PRA	0.000	0.249	0.001	0.124	0.003	0.099	0.004	0.119	0.007	0.239
FRA	0.028	0.206	0.009	0.108	0.004	0.097	0.004	0.104	0.004	0.164
IRA	0.030	0.195	0.013	0.084	0.008	0.074	0.008	0.081	0.004	0.151
					DGP7					
SDM	0.000	0.102	0.000	0.076	-0.003	0.082	-0.005	0.097	-0.008	0.167
PRA	-0.001	0.105	0.002	0.065	0.001	0.066	-0.001	0.081	-0.005	0.140
FRA	0.019	0.093	0.005	0.063	0.000	0.066	-0.004	0.080	-0.004	0.150
IRA	0.020	0.091	0.005	0.061	0.001	0.063	-0.002	0.078	-0.003	0.150
					DGP8					
SDM	0.004	0.136	0.006	0.092	0.004	0.093	-0.003	0.103	-0.022	0.165
PRA	-0.005	0.199	0.009	0.104	0.008	0.093	0.002	0.111	-0.015	0.213
FRA	0.020	0.134	0.011	0.091	0.007	0.091	0.002	0.098	0.010	0.163
IRA	0.022	0.125	0.011	0.075	0.010	0.072	0.006	0.082	0.013	0.152

[a] Here SDM refers to simple difference in means, PRA refers to pooled regression adjustment, IRA is the infeasible regression adjustment estimator and FRA is the feasible regression adjustment estimator.
[b] Simulation across 1000 replications.

Table B.3: Bias and standard deviation for N=500

					DGP1					
	0.1		0.3		0.5		0.7		0.9	
Estimator	**bias**	**std**	**bias**	**std**	**bias**	**std**	**bias**	**std**	**bias**	**std**
SDM	0.035	0.675	0.023	0.493	-0.029	0.453	-0.005	0.535	0.049	0.856
PRA	0.025	0.566	0.009	0.379	-0.019	0.353	0.007	0.396	0.029	0.648
FRA	0.023	0.511	0.008	0.374	-0.019	0.353	0.008	0.382	0.011	0.612
IRA	0.021	0.507	0.007	0.353	-0.019	0.332	0.009	0.372	0.013	0.606
					DGP2					
SDM	0.035	0.675	0.023	0.493	-0.029	0.453	-0.005	0.535	0.049	0.856
PRA	0.025	0.566	0.009	0.379	-0.019	0.353	0.007	0.396	0.029	0.648
FRA	0.023	0.511	0.008	0.374	-0.019	0.353	0.008	0.382	0.011	0.612
IRA	0.021	0.507	0.007	0.353	-0.019	0.332	0.009	0.372	0.013	0.606
					DGP3					
SDM	0.054	1.073	-0.003	0.642	-0.009	0.546	-0.013	0.486	0.001	0.621
PRA	0.031	0.878	0.002	0.490	0.003	0.415	0.010	0.428	0.025	0.707
FRA	-0.014	0.755	-0.004	0.457	-0.003	0.414	0.000	0.400	0.007	0.544
IRA	-0.011	0.729	0.003	0.429	-0.005	0.372	0.004	0.366	0.011	0.518
					DGP4					
SDM	-0.034	0.652	0.012	0.391	-0.003	0.337	0.006	0.333	0.006	0.431
PRA	-0.051	0.744	0.013	0.402	-0.004	0.336	-0.001	0.364	-0.001	0.624
FRA	-0.007	0.599	0.012	0.375	-0.004	0.335	0.007	0.333	0.001	0.401
IRA	-0.010	0.574	0.001	0.336	-0.005	0.287	0.000	0.273	0.000	0.365
					DGP5					
SDM	0.001	0.056	0.001	0.039	0.000	0.038	0.000	0.044	0.006	0.073
PRA	0.000	0.047	0.000	0.030	0.001	0.028	0.001	0.031	0.004	0.053
FRA	0.003	0.044	0.000	0.030	0.001	0.028	0.001	0.030	0.002	0.050
IRA	0.003	0.043	0.001	0.028	0.001	0.027	0.001	0.030	0.002	0.049
					DGP6					
SDM	0.000	0.072	0.001	0.048	-0.001	0.043	-0.001	0.050	0.006	0.077
PRA	0.001	0.101	0.001	0.055	-0.001	0.043	-0.001	0.055	0.008	0.106
FRA	0.006	0.062	0.003	0.046	0.000	0.043	0.001	0.048	0.005	0.063
IRA	0.006	0.055	0.004	0.035	0.000	0.031	0.001	0.035	0.004	0.054
					DGP7					
SDM	-0.001	0.044	-0.001	0.035	0.000	0.036	0.001	0.042	0.001	0.072
PRA	-0.001	0.044	0.000	0.030	0.001	0.028	0.001	0.034	0.002	0.059
FRA	0.003	0.038	0.000	0.029	0.001	0.028	0.001	0.033	0.001	0.055
IRA	0.004	0.037	0.001	0.028	0.001	0.027	0.000	0.032	0.001	0.054
					DGP8					
SDM	-0.002	0.063	-0.002	0.042	0.001	0.039	0.001	0.044	0.002	0.071
PRA	-0.002	0.087	-0.001	0.047	0.001	0.039	0.002	0.045	0.003	0.089
FRA	0.002	0.056	-0.001	0.041	0.001	0.038	0.002	0.041	0.005	0.061
IRA	0.004	0.050	0.000	0.032	0.002	0.030	0.001	0.034	0.005	0.055

[a] Here SDM refers to simple difference in means, PRA refers to pooled regression adjustment, IRA is the infeasible regression adjustment estimator and FRA is the feasible regression adjustment estimator.
[b] Simulation across 1000 replications.

Table B.4: Bias and standard deviation for N=1000

					DGP1					
	0.1		0.3		0.5		0.7		0.9	
Estimator	**bias**	**std**	**bias**	**std**	**bias**	**std**	**bias**	**std**	**bias**	**std**
SDM	0.016	0.467	-0.006	0.335	0.008	0.317	-0.009	0.373	0.018	0.599
PRA	-0.002	0.401	-0.009	0.263	-0.001	0.243	-0.008	0.281	0.013	0.451
FRA	-0.001	0.354	-0.010	0.252	-0.001	0.243	-0.009	0.274	0.011	0.425
IRA	0.000	0.347	-0.009	0.244	-0.001	0.233	-0.007	0.264	0.011	0.423
					DGP2					
SDM	0.016	0.467	-0.006	0.335	0.008	0.317	-0.009	0.373	0.018	0.599
PRA	-0.002	0.401	-0.009	0.263	-0.001	0.243	-0.008	0.281	0.013	0.451
FRA	-0.001	0.354	-0.010	0.252	-0.001	0.243	-0.009	0.274	0.011	0.425
IRA	0.000	0.347	-0.009	0.244	-0.001	0.233	-0.007	0.264	0.011	0.423
					DGP3					
SDM	0.019	0.753	0.001	0.468	0.010	0.363	-0.002	0.346	0.006	0.432
PRA	0.015	0.615	0.000	0.360	0.006	0.277	0.004	0.306	0.001	0.492
FRA	-0.001	0.529	-0.006	0.337	0.003	0.277	0.000	0.284	0.001	0.369
IRA	0.001	0.519	-0.004	0.308	0.003	0.256	0.002	0.257	0.000	0.344
					DGP4					
SDM	-0.007	0.486	-0.006	0.272	0.002	0.242	0.004	0.231	0.004	0.305
PRA	-0.010	0.554	-0.006	0.281	0.002	0.240	0.003	0.247	0.007	0.442
FRA	0.002	0.432	-0.004	0.266	0.001	0.240	0.003	0.226	-0.003	0.275
IRA	-0.002	0.413	-0.006	0.241	0.002	0.196	0.001	0.190	-0.004	0.241
					DGP5					
SDM	0.001	0.040	0.001	0.028	0.001	0.026	-0.001	0.032	0.001	0.051
PRA	-0.001	0.033	0.000	0.022	0.000	0.020	0.000	0.022	0.001	0.036
FRA	0.001	0.031	0.001	0.021	0.000	0.020	-0.001	0.022	0.000	0.033
IRA	0.001	0.030	0.001	0.021	0.000	0.019	0.000	0.021	0.001	0.033
					DGP6					
SDM	-0.001	0.049	0.001	0.033	0.000	0.032	-0.001	0.034	0.002	0.053
PRA	-0.002	0.070	0.001	0.038	0.001	0.032	-0.001	0.037	0.002	0.073
FRA	0.003	0.041	0.001	0.032	0.001	0.032	0.000	0.033	0.001	0.044
IRA	0.003	0.036	0.001	0.025	0.001	0.024	0.000	0.025	0.002	0.037
					DGP7					
SDM	0.002	0.030	0.001	0.023	-0.001	0.026	-0.001	0.031	-0.001	0.048
PRA	0.000	0.031	0.000	0.020	-0.001	0.021	0.000	0.024	0.000	0.040
FRA	0.003	0.025	0.001	0.019	-0.001	0.021	0.000	0.023	-0.001	0.038
IRA	0.003	0.024	0.001	0.019	-0.001	0.020	0.000	0.022	-0.001	0.038
					DGP8					
SDM	0.001	0.042	0.000	0.030	0.000	0.029	-0.001	0.032	0.001	0.050
PRA	0.000	0.059	0.000	0.033	0.000	0.028	-0.001	0.034	0.000	0.062
FRA	0.004	0.036	0.001	0.028	0.000	0.028	-0.001	0.030	0.002	0.043
IRA	0.004	0.032	0.001	0.023	0.001	0.022	0.000	0.024	0.002	0.039

[a] Here SDM refers to simple difference in means, PRA refers to pooled regression adjustment, IRA is the infeasible regression adjustment estimator and FRA is the feasible regression adjustment estimator.
[b] Simulation across 1000 replications.

Table B.5: Bias and standard deviation for binary outcome

					N=500					
	0.1		**0.3**		**0.5**		**0.7**		**0.9**	
Estimator	**bias**	**std**	**bias**	**std**	**bias**	**std**	**bias**	**std**	**bias**	**std**
SDM	0.014	0.062	0.002	0.041	-0.007	0.037	0.003	0.042	0.001	0.063
PRA	0.023	0.056	0.000	0.035	-0.002	0.031	0.003	0.035	0.015	0.054
FRA	0.018	0.051	0.000	0.034	-0.002	0.031	0.002	0.035	0.009	0.053
N-PRA	0.013	0.055	0.001	0.034	-0.001	0.030	0.004	0.034	0.014	0.055
N-RA	0.006	0.052	0.001	0.033	-0.002	0.030	0.004	0.033	0.007	0.051
					N=1000					
SDM	-0.017	0.044	0.008	0.027	0.000	0.026	0.003	0.029	-0.016	0.043
PRA	-0.021	0.038	0.010	0.023	0.000	0.022	0.009	0.024	-0.006	0.038
FRA	-0.024	0.037	0.010	0.023	0.000	0.022	0.009	0.024	-0.010	0.036
N-PRA	-0.019	0.038	0.010	0.022	-0.001	0.021	0.006	0.023	-0.003	0.038
N-RA	-0.020	0.036	0.012	0.022	-0.001	0.021	0.006	0.022	-0.008	0.034

[a] Here SDM refers to simple difference in means, PRA refers to pooled regression adjustment, IRA is the infeasible regression adjustment estimator and FRA is the feasible regression adjustment estimator, N-PRA refers to pooled non-linear regression adjustment and N-RA refers to separate nonlinear regression adjustment.
[b] Simulation across 1000 replications.
[c.] True ATE is 0.037, $R_0^2 = 0.491$ and $R_1^2 = 0.457$.

Table B.6: Bias and standard deviation for non-negative outcome

					N=500					
	0.1		**0.3**		**0.5**		**0.7**		**0.9**	
Estimator	**bias**	**std**	**bias**	**std**	**bias**	**std**	**bias**	**std**	**bias**	**std**
SDM	0.010	0.137	-0.005	0.093	0.017	0.080	-0.027	0.101	-0.067	0.138
PRA	-0.003	0.180	-0.024	0.103	0.015	0.078	-0.023	0.101	-0.074	0.166
FRA	0.024	0.132	-0.006	0.093	0.015	0.078	-0.013	0.093	-0.041	0.112
N-PRA	0.000	0.179	-0.022	0.101	0.015	0.078	-0.024	0.100	-0.078	0.168
N-RA	0.027	0.132	-0.006	0.092	0.011	0.077	-0.013	0.086	-0.039	0.107
					N=1000					
SDM	-0.055	0.089	0.020	0.064	0.006	0.061	-0.014	0.066	-0.022	0.116
PRA	-0.059	0.114	0.028	0.066	0.004	0.060	-0.023	0.068	-0.023	0.133
FRA	-0.044	0.086	0.008	0.061	0.003	0.060	-0.002	0.061	-0.022	0.102
N-PRA	-0.056	0.115	0.028	0.066	0.004	0.060	-0.024	0.068	-0.025	0.133
N-RA	-0.040	0.084	0.006	0.060	0.006	0.059	-0.001	0.059	-0.013	0.089

[a] Here SDM refers to simple difference in means, PRA refers to pooled regression adjustment, IRA is the infeasible regression adjustment estimator and FRA is the feasible regression adjustment estimator, N-PRA refers to pooled non-linear regression adjustment and N-RA refers to separate nonlinear regression adjustment.
[b] Simulation across 1000 replications.
[c.] True ATE is 0.012, $R_0^2 = 0.435$ and $R_1^2 = 0.233$.

APPENDIX C

PROOFS FOR CHAPTER 1

Proof of Lemma 5.1

Proof. Asymptotic variance of SDM

Consider the difference-in-means estimator. We can write the sample average for the treated as

$$\begin{aligned}\bar{Y}_1 &= N_1^{-1}\sum_{i=1}^{N} W_i Y_i = N_1^{-1}\sum_{i=1}^{N} W_i\left[\mu_1 + \dot{\mathbf{X}}_i\boldsymbol{\beta}_1 + U_i(1)\right] \\ &= \mu_1 + N_1^{-1}\sum_{i=1}^{N} W_i\left[\dot{\mathbf{X}}_i\boldsymbol{\beta}_1 + U_i(1)\right]\end{aligned}$$

Therefore,

$$\begin{aligned}\sqrt{N}\left(\bar{Y}_1 - \mu_1\right) &= (N/N_1)N^{-1/2}\sum_{i=1}^{N} W_i\left[\dot{\mathbf{X}}_i\boldsymbol{\beta}_1 + U_i(1)\right] \\ &= (1/\rho)N^{-1/2}\sum_{i=1}^{N} W_i\left[\dot{\mathbf{X}}_i\boldsymbol{\beta}_1 + U_i(1)\right] + o_p(1)\end{aligned}$$

because $N_1/N \xrightarrow{p} \rho$. By the CLT,

$$N^{-1/2}\sum_{i=1}^{N} W_i\left[\dot{\mathbf{X}}_i\boldsymbol{\beta}_1 + U_i(1)\right] \xrightarrow{d} Normal(0, c_1^2)$$

$$\begin{aligned}\text{where, } c_1^2 &= E\left\{W_i\left[\dot{\mathbf{X}}_i\boldsymbol{\beta}_1 + U_i(1)\right]^2\right\} \\ &= \rho\left(\boldsymbol{\beta}_1'\boldsymbol{\Omega}_{\mathbf{X}}\boldsymbol{\beta}_1 + \sigma_1^2\right)\end{aligned}$$

where W_i independent of $\left(\mathbf{X}_i, U_i(1)\right)$ is used. It follows that

$$Avar\left[\sqrt{N}\left(\bar{Y}_1 - \mu_1\right)\right] = (1/\rho)^2\rho\left(\boldsymbol{\beta}_1'\boldsymbol{\Omega}_{\mathbf{X}}\boldsymbol{\beta}_1 + \sigma_1^2\right) = \left(\boldsymbol{\beta}_1'\boldsymbol{\Omega}_{\mathbf{X}}\boldsymbol{\beta}_1 + \sigma_1^2\right)/\rho \tag{C.1}$$

Similarly,

$$Avar\left[\sqrt{N}\left(\bar{Y}_0 - \mu_0\right)\right] = \left(\boldsymbol{\beta}_0'\boldsymbol{\Omega}_{\mathbf{X}}\boldsymbol{\beta}_0 + \sigma_0^2\right)/(1-\rho). \tag{C.2}$$

Combining results from eq(39) and eq(40), we have:

$$\sqrt{N}\left(\hat{\tau}_{SDM}-\tau\right) \xrightarrow{d} \mathcal{N}\left(0, \omega^2_{SDM}\right)$$

Since the sample averages are asymptotically uncorrelated, therefore

$$\omega^2_{SDM} = \boldsymbol{\beta}_1'\boldsymbol{\Omega}_{\mathbf{X}}\boldsymbol{\beta}_1/\rho + \boldsymbol{\beta}_0'\boldsymbol{\Omega}_{\mathbf{X}}\boldsymbol{\beta}_0/(1-\rho) + \sigma_1^2/\rho + \sigma_0^2/(1-\rho)$$

■

Proof. Asymptotic variance of P-RA

To find the asymptotic variance of $\check{\tau}$, note that it can be obtained from

$$Y_i \text{ on } 1,\, W_i,\, \dot{\mathbf{X}}_i.$$

Note that $\dot{\mathbf{X}}_i$ is orthogonal to $(1, W_i)$ because $E(\dot{\mathbf{X}}_i) = \mathbf{0}$ and W_i is independent of $\mathbf{X}_i$. We know that

$$L(Y_i|1, W_i) = \mu_0 + \tau W_i$$

because $\tau = E(Y_i|W_i = 1) - E(Y_i|W_i = 0)$. Therefore,

$$L(Y_i|1, W_i, \dot{\mathbf{X}}_i) = \mu_0 + \tau W_i + \dot{\mathbf{X}}_i\boldsymbol{\beta}$$

By orthogonality,

$$\boldsymbol{\beta} = \left[E\left(\dot{\mathbf{X}}_i'\dot{\mathbf{X}}_i\right)\right]^{-1} E\left(\dot{\mathbf{X}}_i'Y_i\right)$$

Now

$$\begin{aligned} Y_i = {} & (1-W_i)\mu_0 + (1-W_i)\dot{\mathbf{X}}_i\boldsymbol{\beta}_0 + (1-W_i)U_i(0) \\ & + W_i\mu_1 + W_i\dot{\mathbf{X}}_i\boldsymbol{\beta}_1 + W_iU_i(1) \end{aligned}$$

Therefore,

$$\begin{aligned} E\left(\dot{\mathbf{X}}_i'Y_i\right) &= E\left[(1-W_i)\dot{\mathbf{X}}_i'\dot{\mathbf{X}}_i\boldsymbol{\beta}_0\right] + E\left[W_i\dot{\mathbf{X}}_i'\dot{\mathbf{X}}_i\boldsymbol{\beta}_1\right] \\ &= (1-\rho)E\left(\dot{\mathbf{X}}_i'\dot{\mathbf{X}}_i\right)\boldsymbol{\beta}_0 + \rho E\left(\dot{\mathbf{X}}_i'\dot{\mathbf{X}}_i\right)\boldsymbol{\beta}_1 \end{aligned}$$

where we use the linear projection properties of the errors, $E\left(\dot{\mathbf{X}}_i\right) = \mathbf{0}$, and independence of W_i and $\left[\mathbf{X}_i, U_i(0), U_i(1)\right]$. Plugging in gives

$$\boldsymbol{\beta} = (1-\rho)\boldsymbol{\beta}_0 + \rho\boldsymbol{\beta}_1$$

Now we can write the projection error as

$$\begin{aligned}
U_i &= (1-W_i)\mu_0 + (1-W_i)\dot{\mathbf{X}}_i\boldsymbol{\beta}_0 + (1-W_i)U_i(0) \\
&\quad + W_i\mu_1 + W_i\dot{\mathbf{X}}_i\boldsymbol{\beta}_1 + W_iU_i(1) \\
&\quad - \mu_0 - (\mu_1-\mu_0)W_i - \dot{\mathbf{X}}_i\left[(1-\rho)\boldsymbol{\beta}_0 + \rho\boldsymbol{\beta}_1\right] \\
&= -(W_i-\rho)\dot{\mathbf{X}}_i\boldsymbol{\beta}_0 + (1-W_i)U_i(0) \\
&\quad + (W_i-\rho)\dot{\mathbf{X}}_i\boldsymbol{\beta}_1 + W_iU_i(1).
\end{aligned}$$

Because $(1, W_i)$ is orthogonal to $\dot{\mathbf{X}}_i$, it follows as in the previous section that

$$\begin{aligned}
\sqrt{N}(\hat{\tau}_{PRA} - \tau) &= \left[E(\dot{W}_i^2)\right]^{-1}\left(N^{-1/2}\sum_{i=1}^{N}(W_i-\rho)U_i\right) + o_p(1) \\
&= \left[\rho(1-\rho)\right]^{-1}\left(N^{-1/2}\sum_{i=1}^{N}(W_i-\rho)U_i\right).
\end{aligned}$$

Then using asymptotic equivalence lemma and CLT, we have:

$$\sqrt{N}(\hat{\tau}_{PRA} - \tau) \xrightarrow{d} \mathcal{N}\left(0, \omega^2_{PRA}\right)$$

where $\omega^2_{PRA} = Var\left((W_i-\rho)U_i\right) / \left[\rho(1-\rho)\right]^2$.

Now we need to find the asymptotic variance of $N^{-1/2}\sum_{i=1}^{N}(W_i-\rho)U_i$. The term $(W_i-\rho)U_i$ has zero mean by the linear projection property. Further,

$$\begin{aligned}
(W_i-\rho)U_i &= -(W_i-\rho)^2\dot{\mathbf{X}}_i\boldsymbol{\beta}_0 + (W_i-\rho)^2\dot{\mathbf{X}}_i\boldsymbol{\beta}_1 \\
&\quad + (W_i-\rho)(1-W_i)U_i(0) + (W_i-\rho)W_iU_i(1)
\end{aligned}$$

The covariance between the last two terms is zero as $(1-W_i)W_i = 0$. The last two terms can be written as

$$-\rho(1-W_i)U_i(0) + (1-\rho)W_iU_i(1)$$

and so

$$Var\left[-\rho(1-W_i)U_i(0)+(W_i-\rho)W_iU_i(1)\right]=\rho^2(1-\rho)\sigma_0^2+(1-\rho)^2\rho\sigma_1^2.$$

Write the first two terms as

$$(W_i-\rho)^2\dot{\mathbf{X}}_i\left(\boldsymbol{\beta}_1-\boldsymbol{\beta}_0\right).$$

The variance is

$$E\left[(W_i-\rho)^4\right]\left(\boldsymbol{\beta}_1-\boldsymbol{\beta}_0\right)'\boldsymbol{\Omega}_{\mathbf{X}}\left(\boldsymbol{\beta}_1-\boldsymbol{\beta}_0\right).$$

Combining all of the terms gives

$$\begin{aligned}
&Avar\left[\sqrt{N}(\hat{\tau}_{PRA}-\tau)\right]\\
&=\left[\rho(1-\rho)\right]^{-2}\left\{E\left[(W_i-\rho)^4\right]\left(\boldsymbol{\beta}_1-\boldsymbol{\beta}_0\right)'\boldsymbol{\Omega}_{\mathbf{X}}\left(\boldsymbol{\beta}_1-\boldsymbol{\beta}_0\right)+\rho^2(1-\rho)\sigma_0^2+(1-\rho)^2\rho\sigma_1^2\right\}\\
&=\frac{E\left[(W_i-\rho)^4\right]}{\left[\rho(1-\rho)\right]^2}\left(\boldsymbol{\beta}_1-\boldsymbol{\beta}_0\right)'\boldsymbol{\Omega}_{\mathbf{X}}\left(\boldsymbol{\beta}_1-\boldsymbol{\beta}_0\right)+\frac{\sigma_0^2}{(1-\rho)}+\frac{\sigma_1^2}{\rho}
\end{aligned}$$

Note that we can write

$$\frac{E\left[(W_i-\rho)^4\right]}{\left[\rho(1-\rho)\right]^2}=\frac{E\left[(W_i-\rho)^4\right]}{\left[Var(W_i)\right]^2}$$

and Jensen's inequality tells us this is greater than unity: take $Z_i=(W_i-\rho)^2$. We can also show

$$E\left[(W_i-\rho)^4\right]=(1-\rho)^4\rho+\rho^4(1-\rho)$$

and so the scale factor is

$$\frac{(1-\rho)^4\rho+\rho^4(1-\rho)}{\left[\rho(1-\rho)\right]^2}=\frac{(1-\rho)^2}{\rho}+\frac{\rho^2}{(1-\rho)}.$$

Hence,

$$Avar\left[\sqrt{N}\left(\hat{\tau}_{PRA}-\tau\right)\right]=\left(\frac{(1-\rho)^2}{\rho}+\frac{\rho^2}{(1-\rho)}\right)\left(\boldsymbol{\beta}_1-\boldsymbol{\beta}_0\right)'\boldsymbol{\Omega}_{\mathbf{X}}\left(\boldsymbol{\beta}_1-\boldsymbol{\beta}_0\right)+\frac{\sigma_0^2}{(1-\rho)}+\frac{\sigma_1^2}{\rho}$$

■

Proof. Asymptotic variance of F-RA

Now consider the full regression adjustment estimator. Let $\hat{\alpha}_1$ and $\hat{\boldsymbol{\beta}}_1$ be the OLS estimates from the $W_i = 1$ sample:

$$Y_i \text{ on } 1,\ \mathbf{X}_i \quad W_i = 1$$

and then

$$\hat{\mu}_{1,FRA} = \hat{\alpha}_1 + \bar{\mathbf{X}}\hat{\boldsymbol{\beta}}_1$$

where $\bar{\mathbf{X}}$ is the sample average over the entire sample. (For intuition, it is useful to note that $\bar{Y}_1 = \hat{\alpha}_1 + \bar{\mathbf{X}}_1\hat{\boldsymbol{\beta}}_1$, and so $\hat{\mu}_1$ uses a more efficient estimator of $\boldsymbol{\mu}_{\mathbf{X}}$.) By least squares mechanics, $\hat{\mu}_1$ is the intercept in the regression

$$Y_i \text{ on } 1,\ \mathbf{X}_i - \bar{\mathbf{X}},\ W_i = 1.$$

Let $\ddot{\mathbf{X}}_i = \mathbf{X}_i - \bar{\mathbf{X}}$ and

$$\ddot{\mathbf{R}}_i = (1, \ddot{\mathbf{X}}_i).$$

Define

$$\begin{aligned}\hat{\boldsymbol{\gamma}}_1 = \begin{pmatrix}\hat{\mu}_1 \\ \hat{\boldsymbol{\beta}}_1\end{pmatrix} &= \left(\sum_{i=1}^N W_i \ddot{\mathbf{R}}_i'\ddot{\mathbf{R}}_i\right)^{-1}\left(\sum_{i=1}^N W_i \ddot{\mathbf{R}}_i' Y_i\right) \\ &= \left(N^{-1}\sum_{i=1}^N W_i \ddot{\mathbf{R}}_i'\ddot{\mathbf{R}}_i\right)^{-1}\left(N^{-1}\sum_{i=1}^N W_i \ddot{\mathbf{R}}_i' Y_i(1)\right).\end{aligned}$$

Now write

$$\begin{aligned}Y_i(1) &= \mu_1 + \dot{\mathbf{X}}_i\boldsymbol{\beta}_1 + U_i(1) = \mu_1 + \ddot{\mathbf{X}}_i\boldsymbol{\beta}_1 + (\dot{\mathbf{X}}_i - \ddot{\mathbf{X}}_i)\boldsymbol{\beta}_1 + U_i(1) \\ &= \mu_1 + \ddot{\mathbf{X}}_i\boldsymbol{\beta}_1 + (\bar{\mathbf{X}} - \boldsymbol{\mu}_{\mathbf{X}})\boldsymbol{\beta}_1 + U_i(1) = \ddot{\mathbf{R}}_i\boldsymbol{\gamma}_1 + (\bar{\mathbf{X}} - \boldsymbol{\mu}_{\mathbf{X}})\boldsymbol{\beta}_1 + U_i(1)\end{aligned}$$

Plugging in gives

$$\begin{aligned}N^{-1}\sum_{i=1}^N W_i \ddot{\mathbf{R}}_i' Y_i(1) = &\left(N^{-1}\sum_{i=1}^N W_i \ddot{\mathbf{R}}_i'\ddot{\mathbf{R}}_i\right)\boldsymbol{\gamma}_1 + \left(N^{-1}\sum_{i=1}^N W_i \ddot{\mathbf{R}}_i'\right)(\bar{\mathbf{X}} - \boldsymbol{\mu}_{\mathbf{X}})\boldsymbol{\beta}_1 \\ &+ N^{-1}\sum_{i=1}^N W_i \ddot{\mathbf{R}}_i' U_i(1)\end{aligned}$$

Now we can write

$$\hat{\boldsymbol{\gamma}}_1 = \boldsymbol{\gamma}_1 + \left(N^{-1} \sum_{i=1}^{N} W_i \ddot{\mathbf{R}}_i' \ddot{\mathbf{R}}_i \right)^{-1} \left[\left(N^{-1} \sum_{i=1}^{N} W_i \ddot{\mathbf{R}}_i' \right) (\bar{\mathbf{X}} - \boldsymbol{\mu}_{\mathbf{X}}) \boldsymbol{\beta}_1 + N^{-1} \sum_{i=1}^{N} W_i \ddot{\mathbf{R}}_i' U_i(1) \right]$$

and so

$$\begin{aligned} &\sqrt{N} \left(\hat{\boldsymbol{\gamma}}_1 - \boldsymbol{\gamma}_1 \right) \\ &= \left(N^{-1} \sum_{i=1}^{N} W_i \ddot{\mathbf{R}}_i' \ddot{\mathbf{R}}_i \right)^{-1} \left[\left(N^{-1} \sum_{i=1}^{N} W_i \ddot{\mathbf{R}}_i' \right) \sqrt{N} (\bar{\mathbf{X}} - \boldsymbol{\mu}_{\mathbf{X}}) \boldsymbol{\beta}_1 + N^{-1/2} \sum_{i=1}^{N} W_i \ddot{\mathbf{R}}_i' U_i(1) \right] \end{aligned}$$

Next, because $\bar{\mathbf{X}} \xrightarrow{p} \boldsymbol{\mu}_{\boldsymbol{X}}$, the law of large numbers and Slutsky's Theorem imply

$$N^{-1} \sum_{i=1}^{N} W_i \ddot{\mathbf{R}}_i' \ddot{\mathbf{R}}_i = N^{-1} \sum_{i=1}^{N} W_i \dot{\mathbf{R}}_i' \dot{\mathbf{R}}_i + o_p(1)$$

where

$$\dot{\mathbf{R}}_i = (1, \dot{\mathbf{X}}_i) = (1, \mathbf{X}_i - \boldsymbol{\mu}_{\mathbf{X}})$$

Further,

$$N^{-1} \sum_{i=1}^{N} W_i \dot{\mathbf{R}}_i' \dot{\mathbf{R}}_i \xrightarrow{p} E \left(W_i \dot{\mathbf{R}}_i' \dot{\mathbf{R}}_i \right) = \rho E \left(\dot{\mathbf{R}}_i' \dot{\mathbf{R}}_i \right).$$

Note that

$$\mathbf{A} \equiv \begin{pmatrix} 1 & \mathbf{0} \\ \mathbf{0} & E \left(\dot{\mathbf{X}}_i' \dot{\mathbf{X}}_i \right) \end{pmatrix}$$

The terms $\sqrt{N}(\bar{\mathbf{X}} - \boldsymbol{\mu}_{\mathbf{X}})\boldsymbol{\beta}_1$ and $N^{-1/2} \sum_{i=1}^{N} W_i \ddot{\mathbf{R}}_i' U_i(1)$ are $O_p(1)$, and so

$$\begin{aligned} \sqrt{N} \left(\hat{\boldsymbol{\gamma}}_1 - \boldsymbol{\gamma}_1 \right) = (1/\rho) \mathbf{A}^{-1} &\left[\left(N^{-1} \sum_{i=1}^{N} W_i \ddot{\mathbf{R}}_i' \right) \sqrt{N} (\bar{\mathbf{X}} - \boldsymbol{\mu}_{\mathbf{X}}) \boldsymbol{\beta}_1 + N^{-1/2} \sum_{i=1}^{N} W_i \ddot{\mathbf{R}}_i' U_i(1) \right] \\ &+ o_p(1). \end{aligned}$$

Consider the first element of $N^{-1} \sum_{i=1}^{N} W_i \ddot{\mathbf{R}}_i'$:

$$N^{-1} \sum_{i=1}^{N} W_i \ddot{\mathbf{R}}_i' = N^{-1} \sum_{i=1}^{N} W_i \begin{pmatrix} 1 \\ \ddot{\mathbf{X}}_i \end{pmatrix}$$

and so the first element is

$$N^{-1}\sum_{i=1}^{N} W_i = N_1/N = \hat{\rho} \xrightarrow{p} \rho.$$

Also,

$$N^{-1/2}\sum_{i=1}^{N} W_i \ddot{\mathbf{R}}_i' U_i(1) = N^{-1/2}\sum_{i=1}^{N} W_i \begin{pmatrix} 1 \\ \ddot{\mathbf{X}}_i \end{pmatrix} U_i(1)$$

and so the first element is

$$N^{-1/2}\sum_{i=1}^{N} W_i U_i(1).$$

Because of the block diagonality of $\mathbf{A}$, the first element of, $\sqrt{N}\left(\hat{\boldsymbol{\gamma}}_1 - \boldsymbol{\gamma}_1\right)$, $\sqrt{N}\left(\hat{\mu}_1 - \mu_1\right)$ satisfies

$$\begin{aligned}\sqrt{N}\left(\hat{\mu}_{1,FRA} - \mu_1\right) &= (1/\rho)\rho\sqrt{N}(\bar{\mathbf{X}} - \boldsymbol{\mu}_{\mathbf{X}})\boldsymbol{\beta}_1 + (1/\rho)N^{-1/2}\sum_{i=1}^{N} W_i U_i(1) + o_p(1) \\ &= \sqrt{N}(\bar{\mathbf{X}} - \boldsymbol{\mu}_{\mathbf{X}})\boldsymbol{\beta}_1 + (1/\rho)N^{-1/2}\sum_{i=1}^{N} W_i U_i(1) + o_p(1).\end{aligned}$$

We can also write

$$\sqrt{N}\left(\hat{\mu}_{1,FRA} - \mu_1\right) = N^{-1/2}\sum_{i=1}^{N}\left[\left(\mathbf{X}_i - \boldsymbol{\mu}_{\mathbf{X}}\right)\boldsymbol{\beta}_1 + W_i U_i(1)/\rho\right] + o_p(1)$$

A similar argument gives

$$\sqrt{N}\left(\hat{\mu}_{0,FRA} - \mu_0\right) = N^{-1/2}\sum_{i=1}^{N}\left[\left(\mathbf{X}_i - \boldsymbol{\mu}_{\mathbf{X}}\right)\boldsymbol{\beta}_0 + (1 - W_i)U_i(0)/(1-\rho)\right] + o_p(1)$$

and so

$$\sqrt{N}\left(\hat{\tau}_{FRA} - \tau\right) = N^{-1/2}\sum_{i=1}^{N}\left[\dot{\mathbf{X}}_i\left(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0\right) + W_i U_i(1)/\rho - (1 - W_i)U_i(0)/(1-\rho)\right] + o_p(1)$$

Again, by asymptotic equivalence lemma and CLT, we have:

$$\sqrt{N}\left(\hat{\tau}_{FRA} - \tau\right) \xrightarrow{d} \mathcal{N}\left(0, \omega^2_{FRA}\right)$$

where $\omega^2_{FRA} = Var\left(\dot{\mathbf{X}}_i\left(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0\right) + W_i U_i(1)/\rho - (1 - W_i)U_i(0)/(1-\rho)\right)$

Now consider the above expression inside the variance. The three terms are pairwise uncorrelated, the second and third because $W_i(1 - W_i) = 0$, and the first with the other two because, for example,

$$E\left[\left(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0\right)'\dot{\mathbf{X}}_i' W_i U_i(1)\right] = E(W_i)\left(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0\right)' E\left[\dot{\mathbf{X}}_i' U_i(1)\right] = 0$$

because $E\left[\dot{\mathbf{X}}_i' U_i(1)\right] = \mathbf{0}$ by linear projection properties. It follows that

$$Avar\left[\sqrt{N}\left(\hat{\tau}_{FRA} - \tau\right)\right] = \left(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0\right)' \boldsymbol{\Omega}_{\mathbf{X}} \left(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0\right) + (1/\rho^2) E(W_i) E\left[U_i^2(1)\right] + (1/(1-\rho)^2) E(1-W_i) E\left[U_i^2(0)\right]$$

$$= \left(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0\right)' \boldsymbol{\Omega}_{\mathbf{X}} \left(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0\right) + \sigma_1^2/\rho + \sigma_0^2/(1-\rho).$$

■

Proof. Asymptotic variance of I-RA

The derivation for τ^* follows closely that for $\hat{\tau}$, with the important difference that $\ddot{\mathbf{X}}_i = \mathbf{X}_i - \bar{\mathbf{X}}$ is replaced with $\dot{\mathbf{X}}_i = \mathbf{X}_i - \boldsymbol{\mu}_{\mathbf{X}}$. This means that the terms $\sqrt{N}(\bar{\mathbf{X}} - \boldsymbol{\mu}_{\mathbf{X}})\boldsymbol{\beta}_1$ and $\sqrt{N}(\bar{\mathbf{X}} - \boldsymbol{\mu}_{\mathbf{X}})\boldsymbol{\beta}_0$ terms will not appear. Therefore,

$$Avar\left[\sqrt{N}\left(\hat{\tau}_{IRA} - \tau\right)\right] = \sigma_1^2/\rho + \sigma_0^2/(1-\rho).$$

■

Proof of Theorem 5.2

Proof. **CLAIM 1** : $\omega_{FRA}^2 \leq \omega_{SDM}^2$

For this consider consider the left hand side,

$$Avar\left[\sqrt{N}(\hat{\tau}_{SDM} - \tau)\right] - Avar\left[\sqrt{N}(\hat{\tau}_{FRA} - \tau)\right]$$
$$= \boldsymbol{\beta}_1' \boldsymbol{\Omega}_{\mathbf{X}} \boldsymbol{\beta}_1/\rho + \boldsymbol{\beta}_0' \boldsymbol{\Omega}_{\mathbf{X}} \boldsymbol{\beta}_0/(1-\rho) - \left(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0\right)' \boldsymbol{\Omega}_{\mathbf{X}} \left(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0\right)$$

The last term in the above expression can be written as:

$$\boldsymbol{\beta}_1' \boldsymbol{\Omega}_{\mathbf{X}} \boldsymbol{\beta}_1/\rho + \boldsymbol{\beta}_0' \boldsymbol{\Omega}_{\mathbf{X}} \boldsymbol{\beta}_0/(1-\rho) - \left[\boldsymbol{\beta}_1' \boldsymbol{\Omega}_{\mathbf{X}} \boldsymbol{\beta}_1 + \boldsymbol{\beta}_0' \boldsymbol{\Omega}_{\mathbf{X}} \boldsymbol{\beta}_0 - 2\boldsymbol{\beta}_0' \boldsymbol{\Omega}_{\mathbf{X}} \boldsymbol{\beta}_1\right]$$
$$= \left(\frac{1-\rho}{\rho}\right) \boldsymbol{\beta}_1' \boldsymbol{\Omega}_{\mathbf{X}} \boldsymbol{\beta}_1 + \left(\frac{\rho}{1-\rho}\right) \boldsymbol{\beta}_0' \boldsymbol{\Omega}_{\mathbf{X}} \boldsymbol{\beta}_0 + 2\boldsymbol{\beta}_0' \boldsymbol{\Omega}_{\mathbf{X}} \boldsymbol{\beta}_1$$
$$\equiv \boldsymbol{\delta}' \boldsymbol{\Omega}_{\mathbf{X}} \boldsymbol{\delta}$$

where

$$\boldsymbol{\delta} = \sqrt{\left(\frac{1-\rho}{\rho}\right)}\boldsymbol{\beta}_1 + \sqrt{\left(\frac{\rho}{1-\rho}\right)}\boldsymbol{\beta}_0.$$

Because $\boldsymbol{\Omega}_{\mathbf{X}}$ is positive definite, this proves the claim. One case where there is no efficiency gain is when $\rho = 1/2$ and $\boldsymbol{\beta}_1 = -\boldsymbol{\beta}_0$. The second condition seems unrealistic unless both vectors are zero.

CLAIM 2 : $\omega^2_{FRA} \leq \omega^2_{PRA}$

For this consider the left hand side of the expression above,

$$\begin{aligned}
&Avar\left[\sqrt{N}(\hat{\tau}_{PRA} - \tau)\right] - Avar\left[\sqrt{N}(\hat{\tau}_{FRA} - \tau)\right] \\
&= \left[\frac{(1-\rho)^2}{\rho} + \frac{\rho^2}{(1-\rho)} - 1\right](\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0)' \boldsymbol{\Omega}_{\mathbf{X}} (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0) \\
&\geq 0
\end{aligned}$$

CLAIM 3 : $\omega^2_{IRA} \leq \omega^2_{FRA}$

It is easy to see why this holds true since the L.H.S just equals

$$Avar\left[\sqrt{N}(\hat{\tau}_{FRA} - \tau)\right] - Avar\left[\sqrt{N}(\hat{\tau}_{IRA} - \tau)\right] = (\boldsymbol{\beta_1} - \boldsymbol{\beta_0})' \boldsymbol{\Omega}_{\boldsymbol{X}} (\boldsymbol{\beta_1} - \boldsymbol{\beta_0})$$

Because $\boldsymbol{\Omega}_{\boldsymbol{X}}$ is psd and the above is just a quadratic form which will be greater than or equal to zero.

Combing the results from CLAIM 1, 2 and 3 we have the result. ■

APPENDIX D

TABLES FOR CHAPTER 2

Table D.1: Summary of yes votes at different bid amounts

Bid	**Yes-votes**	**%**
\$5	219	20
\$25	216	20
\$65	241	22
\$120	181	17
\$220	228	21
Total	**1085**	**100**

Table D.2: Lower bound mean willingness to pay estimate using ABERS and FRA estimators

	PO means	
Bids	**SM**	**FRA**
\$5	0.689	0.685
	(0.0313)	(0.0288)
\$25	0.569	0.597
	(0.0338)	(0.0307)
\$65	0.485	0.489
	(0.0323)	(0.0294)
\$120	0.403	0.378
	(0.0365)	(0.0332)
\$220	0.289	0.290
	(0.0301)	(0.0286)
	ABERS	**FRA**
$\hat{\tau}$	85.39	84.67
	(3.905)	(3.792)
Obs	**1085**	**1085**

Table D.3: Bias and standard deviation of RA estimators for DGP 1 across four assignment vectors

	PO means\N	SM			PRA			FRA		
		500	**1000**	**5000**	**500**	**1000**	**5000**	**500**	**1000**	**5000**
$\boldsymbol{\rho = (1/3, 1/3, 1/3)}$										
BIAS	μ_1	-0.0003	0.0018	0.0008	0.0021	0.0030	0.0027	-0.0007	0.0014	0.0008
	μ_2	0.0226	0.0191	0.0170	0.0179	0.0161	0.0136	0.0136	0.0130	0.0110
	μ_3	-0.0036	-0.0136	-0.0058	-0.0014	-0.0117	-0.0043	-0.0019	-0.0117	-0.0040
SD	μ_1	0.0107	0.0077	0.0036	0.0110	0.0078	0.0036	0.0107	0.0077	0.0036
	μ_2	0.0113	0.0083	0.0037	0.0107	0.0080	0.0035	0.0105	0.0079	0.0035
	μ_3	0.0133	0.0094	0.0044	0.0128	0.0091	0.0043	0.0128	0.0091	0.0043
$\boldsymbol{\rho = (2/3, 1/6, 1/6)}$										
BIAS	μ_1	-0.0058	-0.0043	-0.0048	-0.0063	-0.0047	-0.0052	-0.0058	-0.0043	-0.0048
	μ_2	-0.0173	-0.0181	-0.0165	-0.0169	-0.0184	-0.0165	-0.0216	-0.0210	-0.0180
	μ_3	0.0078	-0.0075	0.0016	0.0092	-0.0053	0.0030	0.0101	-0.0035	0.0051
SD	μ_1	0.0074	0.0052	0.0025	0.0075	0.0052	0.0025	0.0074	0.0052	0.0025
	μ_2	0.0170	0.0109	0.0054	0.0164	0.0106	0.0052	0.0157	0.0104	0.0050
	μ_3	0.0191	0.0137	0.0063	0.0186	0.0132	0.0061	0.0186	0.0131	0.0060
$\boldsymbol{\rho = (1/6, 2/3, 1/6)}$										
BIAS	μ_1	-0.0085	-0.0065	-0.0082	-0.0010	-0.0025	-0.0014	-0.0088	-0.0065	-0.0074
	μ_2	0.0075	0.0013	0.0032	0.0041	-0.0013	0.0003	0.0031	-0.0021	-0.0004
	μ_3	0.0078	-0.0075	0.0016	0.0137	-0.0009	0.0064	0.0101	-0.0035	0.0051
SD	μ_1	0.0151	0.0108	0.0050	0.0160	0.0110	0.0052	0.0151	0.0108	0.0050
	μ_2	0.0078	0.0056	0.0026	0.0076	0.0055	0.0026	0.0076	0.0055	0.0026
	μ_3	0.0191	0.0137	0.0063	0.0184	0.0131	0.0060	0.0186	0.0131	0.0060
$\boldsymbol{\rho = (1/5, 2/5, 2/5)}$										
BIAS	μ_1	-0.0100	-0.0090	-0.0098	-0.0031	-0.0059	-0.0039	-0.0104	-0.0089	-0.0090
	μ_2	0.0173	0.0110	0.0135	0.0099	0.0054	0.0071	0.0071	0.0036	0.0057
	μ_3	-0.0075	-0.0153	-0.0087	-0.0035	-0.0112	-0.0052	-0.0036	-0.0110	-0.0045
SD	μ_1	0.0139	0.0098	0.0045	0.0145	0.0101	0.0046	0.0139	0.0099	0.0045
	μ_2	0.0103	0.0076	0.0035	0.0098	0.0071	0.0033	0.0096	0.0070	0.0033
	μ_3	0.0120	0.0086	0.0040	0.0116	0.0084	0.0039	0.0116	0.0084	0.0039

[a] Here SM refers to subsample means, PRA refers to pooled regression adjustment, and FRA is the feasible regression adjustment estimator.
[b] Empirical distributions generated with 1000 monte-carlo repetitions.
[c] The true population mean vector is: $\boldsymbol{\mu} = (1.4437, 1.6662, 1.8718)$

Table D.4: Bias and standard deviation of RA estimators for DGP 2 across four assignment vectors

	PO means\N	SM			PRA			FRA		
		500	**1000**	**5000**	**500**	**1000**	**5000**	**500**	**1000**	**5000**
$\boldsymbol{\rho = (1/3, 1/3, 1/3)}$										
BIAS	μ_1	-0.0009	-0.0028	-0.0063	0.0010	-0.0015	-0.0053	-0.0004	-0.0031	-0.0061
	μ_2	0.0039	0.0110	0.0108	0.0067	0.0123	0.0126	0.0070	0.0124	0.0135
	μ_3	0.0111	0.0049	0.0054	0.0065	0.0023	0.0027	0.0033	0.0006	0.0015
SD	μ_1	0.0109	0.0076	0.0034	0.0111	0.0077	0.0034	0.0109	0.0076	0.0034
	μ_2	0.0116	0.0084	0.0036	0.0112	0.0079	0.0035	0.0113	0.0079	0.0035
	μ_3	0.0139	0.0094	0.0043	0.0133	0.0090	0.0042	0.0132	0.0090	0.0041
$\boldsymbol{\rho = (2/3, 1/6, 1/6)}$										
BIAS	μ_1	0.0022	0.0003	-0.0035	0.0038	0.0010	-0.0026	0.0027	0.0002	-0.0032
	μ_2	-0.0236	-0.0315	-0.0266	-0.0241	-0.0303	-0.0254	-0.0272	-0.0317	-0.0248
	μ_3	0.0321	0.0316	0.0331	0.0262	0.0275	0.0287	0.0160	0.0194	0.0219
SD	μ_1	0.0078	0.0054	0.0023	0.0078	0.0054	0.0023	0.0078	0.0054	0.0023
	μ_2	0.0170	0.0109	0.0049	0.0164	0.0105	0.0048	0.0159	0.0103	0.0046
	μ_3	0.0195	0.0142	0.0063	0.0189	0.0137	0.0060	0.0186	0.0132	0.0059
$\boldsymbol{\rho = (1/5, 2/5, 2/5)}$										
BIAS	μ_1	0.0102	0.0087	0.0018	0.0079	0.0073	-0.0017	0.0092	0.0078	0.0017
	μ_2	-0.0062	-0.0032	-0.0043	-0.0030	-0.0008	-0.0012	-0.0029	-0.0006	-0.0007
	μ_3	0.0321	0.0316	0.0331	0.0219	0.0233	0.0243	0.0160	0.0194	0.0219
SD	μ_1	0.0158	0.0106	0.0048	0.0162	0.0109	0.0050	0.0158	0.0106	0.0048
	μ_2	0.0081	0.0059	0.0025	0.0080	0.0058	0.0025	0.0080	0.0058	0.0025
	μ_3	0.0195	0.0142	0.0063	0.0187	0.0134	0.0060	0.0186	0.0132	0.0059
$\boldsymbol{\rho = (1/5, 2/5, 2/5)}$										
BIAS	μ_1	0.0079	0.0080	0.0009	0.0071	0.0072	-0.0011	0.0070	0.0066	0.0006
	μ_2	-0.0024	0.0058	0.0034	0.0023	0.0087	0.0065	0.0024	0.0090	0.0075
	μ_3	0.0041	0.0020	-0.0001	0.0000	-0.0004	-0.0021	-0.0017	-0.0012	-0.0024
SD	μ_1	0.0142	0.0098	0.0044	0.0146	0.0101	0.0045	0.0142	0.0098	0.0043
	μ_2	0.0103	0.0077	0.0033	0.0099	0.0072	0.0032	0.0100	0.0072	0.0032
	μ_3	0.0127	0.0087	0.0040	0.0122	0.0084	0.0038	0.0121	0.0084	0.0038

[a] Here SM refers to subsample means, PRA refers to pooled regression adjustment, and FRA is the feasible regression adjustment estimator.
[b] Empirical distributions generated with 1000 monte-carlo repetitions.
[c] The true population mean vector is : $\boldsymbol{\mu} = (1.4439, 1.6665, 1.8722)$

Table D.5: Bias and standard deviation of RA estimators for DGP 3 across four assignment vectors

	PO means\N	SM			PRA			FRA		
		500	**1000**	**5000**	**500**	**1000**	**5000**	**500**	**1000**	**5000**
$\boldsymbol{\rho = (1/3, 1/3, 1/3)}$										
BIAS	μ_1	0.0215	0.0085	-0.0019	0.0214	0.0081	-0.0027	0.0215	0.0093	-0.0007
	μ_2	-0.0083	-0.0056	0.0001	-0.0079	-0.0055	0.0004	-0.0071	-0.0052	0.0006
	μ_3	0.0017	0.0002	-0.0005	0.0014	0.0005	0.0001	0.0012	0.0013	0.0010
SD	μ_1	0.0030	0.0021	0.0010	0.0032	0.0022	0.0010	0.0030	0.0021	0.0009
	μ_2	0.0063	0.0045	0.0021	0.0062	0.0045	0.0021	0.0062	0.0045	0.0020
	μ_3	0.0052	0.0036	0.0017	0.0051	0.0036	0.0016	0.0051	0.0035	0.0016
$\boldsymbol{\rho = (2/3, 1/6, 1/6)}$										
BIAS	μ_1	0.0119	0.0063	0.0003	0.0120	0.0063	0.0003	0.0127	0.0068	0.0007
	μ_2	-0.0257	-0.0120	-0.0015	-0.0263	-0.0121	-0.0015	-0.0202	-0.0100	-0.0012
	μ_3	-0.0141	-0.0061	-0.0058	-0.0140	-0.0063	-0.0061	-0.0134	-0.0033	-0.0022
SD	μ_1	0.0022	0.0015	0.0007	0.0022	0.0015	0.0007	0.0022	0.0015	0.0007
	μ_2	0.0091	0.0063	0.0030	0.0091	0.0063	0.0031	0.0089	0.0062	0.0030
	μ_3	0.0071	0.0052	0.0023	0.0072	0.0053	0.0023	0.0069	0.0050	0.0022
$\boldsymbol{\rho = (1/5, 2/5, 2/5)}$										
BIAS	μ_1	0.0466	0.0216	-0.0018	0.0485	0.0218	-0.0032	0.0425	0.0211	-0.0004
	μ_2	-0.0075	-0.0034	-0.0008	-0.0082	-0.0037	-0.0009	-0.0085	-0.0038	-0.0010
	μ_3	-0.0184	-0.0069	-0.0061	-0.0175	-0.0058	-0.0041	-0.0152	-0.0040	-0.0022
SD	μ_1	0.0042	0.0030	0.0013	0.0046	0.0033	0.0014	0.0041	0.0030	0.0013
	μ_2	0.0046	0.0032	0.0015	0.0046	0.0032	0.0015	0.0046	0.0032	0.0015
	μ_3	0.0071	0.0051	0.0024	0.0069	0.0049	0.0023	0.0068	0.0048	0.0023
$\boldsymbol{\rho = (1/5, 2/5, 2/5)}$										
BIAS	μ_1	0.0389	0.0162	-0.0017	0.0410	0.0167	-0.0026	0.0351	0.0153	-0.0009
	μ_2	-0.0112	-0.0027	-0.0008	-0.0116	-0.0031	-0.0011	-0.0115	-0.0031	-0.0012
	μ_3	0.0028	0.0020	0.0017	0.0022	0.0022	0.0025	0.0020	0.0026	0.0031
SD	μ_1	0.0039	0.0028	0.0012	0.0042	0.0030	0.0013	0.0038	0.0027	0.0012
	μ_2	0.0057	0.0043	0.0018	0.0056	0.0042	0.0018	0.0056	0.0042	0.0018
	μ_3	0.0048	0.0033	0.0015	0.0047	0.0032	0.0014	0.0047	0.0032	0.0014

[a] Here SM refers to subsample means, PRA refers to pooled regression adjustment, and FRA is the feasible regression adjustment estimator.
[b] Empirical distributions generated with 1000 monte-carlo repetitions.
[c] The true population mean vector is: $\boldsymbol{\mu} = (1.1897, 3.7310, 3.7752)$

APPENDIX E

PROOFS FOR CHAPTER 2

Proof of Theorem 3

Proof. Using the expression, $Y = \mu_g + \dot{\mathbf{X}}\boldsymbol{\beta}_g + U(g)$, write $\bar{Y}_g$ as

$$\begin{aligned}
\bar{Y}_g &= N_g^{-1}\sum_{i=1}^{N} W_{ig}Y_i(g) = N_g^{-1}\sum_{i=1}^{N} W_{ig}\left[\mu_g + \dot{\mathbf{X}}_i\boldsymbol{\beta_g} + U_i(g)\right] \\
&= \mu_g + N_g^{-1}\sum_{i=1}^{N} W_{ig}\left[\dot{\mathbf{X}}_i\boldsymbol{\beta_g} + U_i(g)\right] = \mu_g + (N/N_g)N^{-1}\sum_{i=1}^{N} W_{ig}\left[\dot{\mathbf{X}}_i\boldsymbol{\beta_g} + U_i(g)\right] \\
&= \mu_g + \left(\frac{1}{\hat{\rho}_g}\right)N^{-1}\sum_{i=1}^{N} W_{ig}\left[\dot{\mathbf{X}}_i\boldsymbol{\beta_g} + U_i(g)\right]
\end{aligned}$$

Therefore,

$$\sqrt{N}\left(\bar{Y}_g - \mu_g\right) = \hat{\rho}_g^{-1}\left\{N^{-1/2}\sum_{i=1}^{N} W_{ig}\left[\dot{\mathbf{X}}_i\boldsymbol{\beta_g} + U_i(g)\right]\right\} \tag{E.1}$$

By random assignment,

$$E\left(W_{ig}\dot{\mathbf{X}}_i\right) = E(W_{ig})E\left(\dot{\mathbf{X}}_i\right) = \mathbf{0}$$
$$E\left[W_{ig}U_i(g)\right] = E\left(W_{ig}\right)E\left[U_i(g)\right] = \rho_g E\left[U_i(g)\right] = 0,$$

and so the CLT applies to the standardized average in (E.1). Now use $\hat{\rho}_g = \rho_g + o_p(1)$ to obtain the following first-order representation:

$$\sqrt{N}\left(\bar{Y}_g - \mu_g\right) = \rho_g^{-1}\left\{N^{-1/2}\sum_{i=1}^{N} W_{ig}\left[\dot{\mathbf{X}}_i\boldsymbol{\beta_g} + U_i(g)\right]\right\} + o_p(1).$$

Our goal is to be able to make efficiency statements about both linear and nonlinear functions of the vector of means $\boldsymbol{\mu} = (\mu_1, \mu_2, ..., \mu_G)'$, and so we stack the subsample means into the $G \times 1$ vector $\bar{\mathbf{Y}}$. For later comparison, it is helpful to remember that $\bar{\mathbf{Y}}$ is the vector of OLS coefficients in the regression

$$Y_i \text{ on } W_{i1},\ W_{i2},\ ...,\ W_{iG},\ i = 1, 2, ..., N.$$

We have proven the following result. ■

Proof of Theorem 4

Proof. Consistent estimators of α_g and $\boldsymbol{\beta}_g$ are obtained from the regression

$$Y_i \text{ on } 1,\ \mathbf{X}_i, \text{ if } W_{ig} = 1,$$

which produces intercept and slopes $\hat{\alpha}_g$ and $\hat{\boldsymbol{\beta}}_{\boldsymbol{g}}$. Letting $\breve{\mathbf{X}}_i = (1, \mathbf{X}_i)$, the probability limit of $\left(\hat{\alpha}_g, \hat{\boldsymbol{\beta}}_g'\right)'$ is

$$\begin{aligned}
\left[\mathbb{E}\left(W_{ig}\breve{\mathbf{X}}_i'\breve{\mathbf{X}}_i\right)\right]^{-1}\left[\mathbb{E}\left(W_{ig}\breve{\mathbf{X}}_i'Y_i\right)\right] &= \rho_g^{-1}\left[\mathbb{E}\left(\breve{\mathbf{X}}_i'\breve{\mathbf{X}}_i\right)\right]^{-1}\left[\mathbb{E}\left(W_{ig}\breve{\mathbf{X}}_i'Y_i(g)\right)\right] \\
&= \rho_g^{-1}\left[\mathbb{E}\left(\breve{\mathbf{X}}_i'\breve{\mathbf{X}}_i\right)\right]^{-1}\left[\rho_g\mathbb{E}\left(\breve{\mathbf{X}}_i'Y_i(g)\right)\right] \\
&= \left[\mathbb{E}\left(\breve{\mathbf{X}}_i'\breve{\mathbf{X}}_i\right)\right]^{-1}\left[\mathbb{E}\left(\breve{\mathbf{X}}_i'Y_i(g)\right)\right] = \begin{pmatrix}\alpha_g \\ \boldsymbol{\beta}_{\boldsymbol{g}}\end{pmatrix}
\end{aligned}$$

where random assignment is used so that W_{ig} is independent of $\left[\mathbf{X}_i, Y_i(g)\right]$. It follows that $\left(\hat{\alpha}_g, \hat{\boldsymbol{\beta}}_{\boldsymbol{g}}'\right)'$ is consistent for $\left(\alpha_g, \boldsymbol{\beta}_{\boldsymbol{g}}'\right)$, and so a consistent estimator of μ_g is

$$\hat{\mu}_g = \hat{\alpha}_g + \bar{\mathbf{X}}\hat{\boldsymbol{\beta}}_{\boldsymbol{g}}.$$

Note that this estimator, which we refer to as full (or separate) regression adjustment (FRA), is the same as an imputation procedure. Given $\hat{\alpha}_g$ and $\hat{\boldsymbol{\beta}}_{\boldsymbol{g}}$, impute a value of $Y_i(g)$ for each i in the sample, whether or not i is assigned to group g:

$$\hat{Y}_i(g) = \hat{\alpha}_g + \mathbf{X}_i\hat{\boldsymbol{\beta}}_{\boldsymbol{g}},\ i = 1, 2, ..., N.$$

Averaging these imputed values across all i produces $\hat{\mu}_g$. In order to derive the asymptotic variance of $\hat{\mu}_g$, it is helpful to obtain it as the intercept from the regression

$$Y_i \text{ on } 1,\ \mathbf{X}_i - \bar{\mathbf{X}},\ \ W_{ig} = 1.$$

Let $\ddot{\mathbf{X}}_i = \mathbf{X}_i - \bar{\mathbf{X}}$ and

$$\ddot{\mathbf{R}}_i = (1, \ddot{\mathbf{X}}_i).$$

Define

$$\begin{aligned}\hat{\boldsymbol{\gamma}}_g &= \begin{pmatrix}\hat{\mu}_g \\ \hat{\boldsymbol{\beta}}_{\boldsymbol{g}}\end{pmatrix} = \left(\sum_{i=1}^{N} W_{ig}\ddot{\mathbf{R}}_i'\ddot{\mathbf{R}}_i\right)^{-1}\left(\sum_{i=1}^{N} W_{ig}\ddot{\mathbf{R}}_i' Y_i\right) \\ &= \left(N^{-1}\sum_{i=1}^{N} W_{ig}\ddot{\mathbf{R}}_i'\ddot{\mathbf{R}}_i\right)^{-1}\left(N^{-1}\sum_{i=1}^{N} W_{ig}\ddot{\mathbf{R}}_i' Y_i(g)\right).\end{aligned}$$

Now write

$$\begin{aligned}Y_i(g) &= \mu_g + \dot{\mathbf{X}}_i\boldsymbol{\beta}_{\boldsymbol{g}} + U_i(g) = \mu_g + \ddot{\mathbf{X}}_i\boldsymbol{\beta}_{\boldsymbol{g}} + (\dot{\mathbf{X}}_i - \ddot{\mathbf{X}}_i)\boldsymbol{\beta}_{\boldsymbol{g}} + U_i(g) \\ &= \mu_g + \ddot{\mathbf{X}}_i\boldsymbol{\beta}_{\boldsymbol{g}} + (\bar{\mathbf{X}} - \boldsymbol{\mu}_{\mathbf{X}})\boldsymbol{\beta}_{\boldsymbol{g}} + U_i(g) = \ddot{\mathbf{R}}_i\boldsymbol{\gamma}_g + (\bar{\mathbf{X}} - \boldsymbol{\mu}_{\mathbf{X}})\boldsymbol{\beta}_{\boldsymbol{g}} + U_i(g)\end{aligned}$$

Plugging in for $Y_i(g)$ gives

$$\hat{\boldsymbol{\gamma}}_g = \boldsymbol{\gamma}_g + \left(N^{-1}\sum_{i=1}^{N} W_{ig}\ddot{\mathbf{R}}_i'\ddot{\mathbf{R}}_i\right)^{-1}\left[\left(N^{-1}\sum_{i=1}^{N} W_{ig}\ddot{\mathbf{R}}_i'\right)(\bar{\mathbf{X}} - \boldsymbol{\mu}_{\mathbf{X}})\boldsymbol{\beta}_{\boldsymbol{g}} + N^{-1}\sum_{i=1}^{N} W_{ig}\ddot{\mathbf{R}}_i' U_i(g)\right]$$

and so

$$\begin{aligned}&\sqrt{N}\left(\hat{\boldsymbol{\gamma}}_g - \boldsymbol{\gamma}_g\right) \\ &= \left(N^{-1}\sum_{i=1}^{N} W_{ig}\ddot{\mathbf{R}}_i'\ddot{\mathbf{R}}_i\right)^{-1}\left[\left(N^{-1}\sum_{i=1}^{N} W_{ig}\ddot{\mathbf{R}}_i'\right)\sqrt{N}(\bar{\mathbf{X}} - \boldsymbol{\mu}_{\mathbf{X}})\boldsymbol{\beta}_{\boldsymbol{g}} + N^{-1/2}\sum_{i=1}^{N} W_{ig}\ddot{\mathbf{R}}_i' U_i(g)\right]\end{aligned}$$

Next, because $\bar{\mathbf{X}} \xrightarrow{p} \boldsymbol{\mu}_{\boldsymbol{X}}$, the law of large numbers and Slutsky's Theorem imply

$$N^{-1}\sum_{i=1}^{N} W_{ig}\ddot{\mathbf{R}}_i'\ddot{\mathbf{R}}_i = N^{-1}\sum_{i=1}^{N} W_{ig}\dot{\mathbf{R}}_i'\dot{\mathbf{R}}_i + o_p(1)$$

where

$$\dot{\mathbf{R}}_i = (1, \dot{\mathbf{X}}_i).$$

Further, by random assignment,

$$N^{-1}\sum_{i=1}^{N} W_{ig}\dot{\mathbf{R}}_i'\dot{\mathbf{R}}_i \xrightarrow{p} \mathbb{E}\left(W_{ig}\dot{\mathbf{R}}_i'\dot{\mathbf{R}}_i\right) = \rho_g\mathbb{E}\left(\dot{\mathbf{R}}_i'\dot{\mathbf{R}}_i\right) = \rho_g\mathbf{A},$$

where

$$\mathbf{A} \equiv \begin{pmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \mathbb{E}\left(\dot{\mathbf{X}}_i'\dot{\mathbf{X}}_i\right) \end{pmatrix}$$

The terms $\sqrt{N}(\bar{\mathbf{X}} - \boldsymbol{\mu}_{\mathbf{X}})\boldsymbol{\beta}_{\boldsymbol{g}}$ and $N^{-1/2}\sum_{i=1}^{N} W_{ig}\ddot{\mathbf{R}}_i'U_i(g)$ are $O_p(1)$ by the CLT, and so

$$\sqrt{N}\left(\hat{\boldsymbol{\gamma}}_g - \boldsymbol{\gamma}_g\right) = (1/\rho_g)\mathbf{A}^{-1}\left[\left(N^{-1}\sum_{i=1}^{N} W_{ig}\ddot{\mathbf{R}}_i'\right)\sqrt{N}(\bar{\mathbf{X}} - \boldsymbol{\mu}_{\mathbf{X}})\boldsymbol{\beta}_{\boldsymbol{g}} + N^{-1/2}\sum_{i=1}^{N} W_{ig}\ddot{\mathbf{R}}_i'U_i(g)\right].$$

Consider the first element of $N^{-1}\sum_{i=1}^{N} W_{ig}\ddot{\mathbf{R}}_i'$:

$$N^{-1}\sum_{i=1}^{N} W_{ig}\ddot{\mathbf{R}}_i' = N^{-1}\sum_{i=1}^{N} W_{ig}\begin{pmatrix} 1 \\ \ddot{\mathbf{X}}_i' \end{pmatrix}$$

and so the first element is

$$N^{-1}\sum_{i=1}^{N} W_{ig} = N_g/N = \hat{\rho}_g \xrightarrow{p} \rho_g.$$

Also,

$$N^{-1/2}\sum_{i=1}^{N} W_{ig}\ddot{\mathbf{R}}_i'U_i(g) = N^{-1/2}\sum_{i=1}^{N} W_{ig}\begin{pmatrix} 1 \\ \ddot{\mathbf{X}}_i' \end{pmatrix}U_i(g)$$

and so the first element is

$$N^{-1/2}\sum_{i=1}^{N} W_{ig}U_i(g).$$

Because of the block diagonality of $\mathbf{A}$, the first element of, $\sqrt{N}\left(\hat{\boldsymbol{\gamma}}_g - \boldsymbol{\gamma}_g\right)$, $\sqrt{N}\left(\hat{\mu}_g - \mu_g\right)$, satisfies

$$\begin{aligned} \sqrt{N}\left(\hat{\mu}_g - \mu_g\right) &= (1/\rho_g)\rho_g\sqrt{N}(\bar{\mathbf{X}} - \boldsymbol{\mu}_{\mathbf{X}})\boldsymbol{\beta}_{\boldsymbol{g}} + (1/\rho_g)N^{-1/2}\sum_{i=1}^{N} W_{ig}U_i(g) + o_p(1) \\ &= \sqrt{N}(\bar{\mathbf{X}} - \boldsymbol{\mu}_{\mathbf{X}})\boldsymbol{\beta}_{\boldsymbol{g}} + (1/\rho_g)N^{-1/2}\sum_{i=1}^{N} W_{ig}U_i(g) + o_p(1). \end{aligned}$$

We can also write

$$\sqrt{N}\left(\hat{\mu}_g - \mu_g\right) = N^{-1/2}\sum_{i=1}^{N}\left[\left(\mathbf{X}_i - \boldsymbol{\mu}_{\mathbf{X}}\right)\boldsymbol{\beta}_{\boldsymbol{g}} + W_{ig}U_i(g)/\rho_g\right] + o_p(1)$$

The above representation holds for all g. Then, stacking the RA estimates gives us theorem 4. ■

Proof of Theorem 5

Proof. In order to find a useful first order representation of $\sqrt{N}\left(\check{\boldsymbol{\mu}}-\boldsymbol{\mu}\right)$, we first characterize the probability limit of $\check{\boldsymbol{\beta}}$. Under random assignment,

$$\mathbb{E}\left(\mathbf{W}'\dot{\mathbf{X}}\right)=\mathbb{E}\left(\mathbf{W}\right)'\mathbb{E}\left(\dot{\mathbf{X}}\right)=\mathbf{0},$$

which means that the coefficients on $\mathbf{W}$ in the linear projections $\mathbb{L}\left(Y|\mathbf{W}\right)$ and $\mathbb{L}\left(Y|\mathbf{W},\dot{\mathbf{X}}\right)$ are the same and equal to $\boldsymbol{\mu}$. This essentially proves that adding the demeaned covariates still consistently estimates $\boldsymbol{\mu}$. Moreover, we can find the coefficients on $\dot{\mathbf{X}}$ in $\mathbb{L}\left(Y|\mathbf{W},\dot{\mathbf{X}}\right)$ by finding $\mathbb{L}\left(Y|\dot{\mathbf{X}}\right)$. Let $\boldsymbol{\beta}$ be the the linear projection of Y on $\dot{\mathbf{X}}$. Then

$$\boldsymbol{\beta}=\left[\mathbb{E}\left(\dot{\mathbf{X}}'\dot{\mathbf{X}}\right)\right]^{-1}\mathbb{E}\left(\dot{\mathbf{X}}'Y\right)=\boldsymbol{\Omega}_{\mathbf{X}}^{-1}\mathbb{E}\left(\dot{\mathbf{X}}'Y\right)$$

Now use

$$Y=\sum_{g=1}^{G}W_g\left[\mu_g+\dot{\mathbf{X}}\boldsymbol{\beta_g}+U(g)\right]$$

so that

$$\begin{aligned}\mathbb{E}\left(\dot{\mathbf{X}}'Y\right) &= \sum_{g=1}^{G}\left\{\mathbb{E}\left(\dot{\mathbf{X}}'W_g\mu_g\right)+\mathbb{E}\left(\dot{\mathbf{X}}'W_g\dot{\mathbf{X}}\right)\boldsymbol{\beta_g}+\mathbb{E}\left[\dot{\mathbf{X}}'W_gU(g)\right]\right\}\\ &= \sum_{g=1}^{G}\left\{\mathbf{0}+\rho_g\boldsymbol{\Omega}_{\mathbf{X}}\boldsymbol{\beta_g}+\mathbf{0}\right\}=\boldsymbol{\Omega}_{\mathbf{X}}\left(\sum_{g=1}^{G}\rho_g\boldsymbol{\beta_g}\right),\end{aligned}$$

where we again use random assignment, $\mathbb{E}\left(\dot{\mathbf{X}}\right)=\mathbf{0}$, and $\mathbb{E}\left[\dot{\mathbf{X}}'U(g)\right]=\mathbf{0}$. It follows that

$$\boldsymbol{\beta}=\boldsymbol{\Omega}_{\mathbf{X}}^{-1}\boldsymbol{\Omega}_{\mathbf{X}}\left(\sum_{g=1}^{G}\rho_g\boldsymbol{\beta_g}\right)=\left(\sum_{g=1}^{G}\rho_g\boldsymbol{\beta_g}\right).$$

Therefore, the $\boldsymbol{\beta}$ in the linear projection $\mathbb{L}\left(Y|\dot{\mathbf{X}}\right)$ is simply a weighted average of the coefficients from the separate linear projections using the potential outcomes.

Now we can write

$$Y_i=\mathbf{W}_i\boldsymbol{\mu}+\dot{\mathbf{X}}_i\boldsymbol{\beta}+U_i$$

where the linear projection error U_i is

$$\begin{aligned} U_i &= \sum_{g=1}^{G} W_{ig}\left[\mu_{ig} + \dot{\mathbf{X}}_i\boldsymbol{\beta}_{\boldsymbol{g}} + U_i(g)\right] - \mathbf{W}_i\boldsymbol{\mu} - \dot{\mathbf{X}}_i\left(\sum_{g=1}^{G}\rho_g\boldsymbol{\beta}_{\boldsymbol{g}}\right) \\ &= \sum_{g=1}^{G}(W_{ig}-\rho_g)\dot{\mathbf{X}}_i\boldsymbol{\beta}_{\boldsymbol{g}} + \sum_{g=1}^{G} W_{ig}U_i(g) \end{aligned}$$

We can now obtain the asymptotic representation for $\sqrt{N}\left(\check{\boldsymbol{\mu}} - \boldsymbol{\mu}\right)$. Write $\boldsymbol{\theta} = \left(\boldsymbol{\mu}', \boldsymbol{\beta}'\right)'$, $\dot{\mathbf{R}}_i = \left(\mathbf{W}_i, \dot{\mathbf{X}}_i\right)$, $\ddot{\mathbf{R}}_i = \left(\mathbf{W}_i, \ddot{\mathbf{X}}_i\right)$, and $\check{\boldsymbol{\theta}} = \left(\check{\boldsymbol{\mu}}', \check{\boldsymbol{\beta}}'\right)'$ as the OLS estimators. The asymptotic variance of $\sqrt{N}\left(\check{\boldsymbol{\mu}} - \boldsymbol{\mu}\right)$ is not the same as replacing $\ddot{\mathbf{X}}_i$ with $\dot{\mathbf{X}}_i$ (even though for $\check{\boldsymbol{\beta}}$ it is). Write

$$\begin{aligned} Y_i &= \mathbf{W}_i\boldsymbol{\mu} + \ddot{\mathbf{X}}_i\boldsymbol{\beta} + \left(\dot{\mathbf{X}}_i - \ddot{\mathbf{X}}_i\right)\boldsymbol{\beta} + U_i = \mathbf{W}_i\boldsymbol{\mu} + \ddot{\mathbf{X}}_i\boldsymbol{\beta} + \left(\bar{\mathbf{X}} - \boldsymbol{\mu}_{\mathbf{X}}\right)\boldsymbol{\beta} + U_i \\ &= \ddot{\mathbf{R}}_i\boldsymbol{\theta} + \left(\bar{\mathbf{X}} - \boldsymbol{\mu}_{\mathbf{X}}\right)\boldsymbol{\beta} + U_i. \end{aligned}$$

Now

$$\begin{aligned} \check{\boldsymbol{\theta}} &= \left(N^{-1}\sum_{i=1}^{N}\ddot{\mathbf{R}}_i'\ddot{\mathbf{R}}_i\right)^{-1}\left(N^{-1}\sum_{i=1}^{N}\ddot{\mathbf{R}}_i'Y_i\right) \\ &= \boldsymbol{\theta} + \left(N^{-1}\sum_{i=1}^{N}\ddot{\mathbf{R}}_i'\ddot{\mathbf{R}}_i\right)^{-1}\left[\left(N^{-1}\sum_{i=1}^{N}\ddot{\mathbf{R}}_i\right)'\left(\bar{\mathbf{X}} - \boldsymbol{\mu}_{\mathbf{X}}\right)\boldsymbol{\beta} + N^{-1}\sum_{i=1}^{N}\ddot{\mathbf{R}}_i'U_i\right] \end{aligned}$$

and so

$$\begin{aligned} &\sqrt{N}\left(\check{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) \\ &= \left(N^{-1}\sum_{i=1}^{N}\ddot{\mathbf{R}}_i'\ddot{\mathbf{R}}_i\right)^{-1}\left[\left(N^{-1}\sum_{i=1}^{N}\ddot{\mathbf{R}}_i\right)'\left[\sqrt{N}\left(\bar{\mathbf{X}} - \boldsymbol{\mu}_{\mathbf{X}}\right)\right]\boldsymbol{\beta} + N^{-1/2}\sum_{i=1}^{N}\ddot{\mathbf{R}}_i'U_i\right] \\ &= \left(N^{-1}\sum_{i=1}^{N}\ddot{\mathbf{R}}_i'\ddot{\mathbf{R}}_i\right)^{-1}\left[\begin{pmatrix} N^{-1}\sum_{i=1}^{N}\mathbf{W}_i' \\ \mathbf{0} \end{pmatrix}\left[\sqrt{N}\left(\bar{\mathbf{X}} - \boldsymbol{\mu}_{\mathbf{X}}\right)\right]\boldsymbol{\beta} + N^{-1/2}\sum_{i=1}^{N}\begin{pmatrix}\mathbf{W}_i \\ \ddot{\mathbf{X}}_i\end{pmatrix}'U_i\right] \end{aligned}$$

because $N^{-1}\sum_{i=1}^{N}\ddot{\mathbf{X}}_i' = \mathbf{0}$. Further, the terms in $[\cdot]$ are $O_p(1)$ and

$$N^{-1}\sum_{i=1}^{N}\ddot{\mathbf{R}}_i'\ddot{\mathbf{R}}_i \xrightarrow{p} \begin{pmatrix}\mathbb{E}\left(\mathbf{W}_i'\mathbf{W}_i\right) & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Omega}_{\mathbf{X}}\end{pmatrix}$$

by random assignment and $E\left(\dot{\mathbf{X}}_i\right) = \mathbf{0}$. Therefore,

$$\begin{aligned}
&\sqrt{N}\left(\check{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) \\
&= \begin{pmatrix} \left[\mathbb{E}\left(\mathbf{W}_i'\mathbf{W}_i\right)\right]^{-1} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Omega}_{\mathbf{X}}^{-1} \end{pmatrix} \left[\begin{pmatrix} N^{-1}\sum_{i=1}^{N}\mathbf{W}_i' \\ \mathbf{0} \end{pmatrix} \left[\sqrt{N}\left(\bar{\mathbf{X}} - \boldsymbol{\mu}_{\mathbf{X}}\right)\right]\boldsymbol{\beta} + N^{-1/2}\sum_{i=1}^{N}\ddot{\mathbf{R}}_i'U_i \right]
\end{aligned}$$

We can now look at $\sqrt{N}\left(\check{\boldsymbol{\mu}} - \boldsymbol{\mu}\right)$, the first G elements of $\sqrt{N}\left(\check{\boldsymbol{\theta}} - \boldsymbol{\theta}\right)$. But

$$N^{-1}\sum_{i=1}^{N}\mathbf{W}_i' \xrightarrow{p} \begin{pmatrix} \rho_1 \\ \rho_2 \\ \vdots \\ \rho_G \end{pmatrix}$$

and so

$$\sqrt{N}\left(\check{\boldsymbol{\mu}} - \boldsymbol{\mu}\right) = \left[\mathbb{E}\left(\mathbf{W}_i'\mathbf{W}_i\right)\right]^{-1} \left[\begin{pmatrix} \rho_1 \\ \rho_2 \\ \vdots \\ \rho_G \end{pmatrix} N^{-1/2}\sum_{i=1}^{N}\dot{\mathbf{X}}_i\boldsymbol{\beta} + N^{-1/2}\sum_{i=1}^{N}\mathbf{W}_i'U_i \right] + o_p(1).$$

Note that

$$\mathbf{W}_i'\mathbf{W}_i = \begin{pmatrix} W_{i1} & 0 & \cdots & 0 \\ 0 & W_{i2} & \ddots & \vdots \\ \vdots & \ddots & & 0 \\ 0 & \cdots & 0 & W_{iG} \end{pmatrix}$$

and so

$$\mathbb{E}\left(\mathbf{W}_i'\mathbf{W}_i\right) = \begin{pmatrix} \rho_1 & 0 & \cdots & 0 \\ 0 & \rho_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \rho_G \end{pmatrix}.$$

Therefore,

$$\sqrt{N}\left(\check{\boldsymbol{\mu}}-\boldsymbol{\mu}\right)=\mathbf{j}_G N^{-1/2}\sum_{i=1}^{N}\dot{\mathbf{X}}_i\boldsymbol{\beta}+\left[\mathbb{E}\left(\mathbf{W}_i'\mathbf{W}_i\right)\right]^{-1}N^{-1/2}\sum_{i=1}^{N}\mathbf{W}_i'U_i+o_p(1)\qquad((\text{k1}))$$

where $\mathbf{j}_G=(1,1,...,1)'$. Now write

$$\sqrt{N}\left(\check{\boldsymbol{\mu}}-\boldsymbol{\mu}\right)=\mathbf{j}_G N^{-1/2}\sum_{i=1}^{N}\dot{\mathbf{X}}_i\boldsymbol{\beta}+\begin{pmatrix}\rho_1 & 0 & \cdots & 0\\ 0 & \rho_2 & \ddots & \vdots\\ \vdots & \ddots & \ddots & 0\\ 0 & \cdots & 0 & \rho_G\end{pmatrix}^{-1}\left[N^{-1/2}\sum_{i=1}^{N}\mathbf{W}_i'U_i\right]+o_p(1)$$

$$\begin{aligned}\mathbf{W}_i'U_i &= \begin{pmatrix}W_{i1}\\ W_{i2}\\ \vdots\\ W_{iG}\end{pmatrix}\left[\sum_{h=1}^{G}(W_{ih}-\rho_g)\dot{\mathbf{X}}_i\boldsymbol{\beta}_g+\sum_{h=1}^{G}W_{ih}U_i(h)\right]\\ &= \begin{pmatrix}W_{i1}\sum_{h=1}^{G}(W_{ih}-\rho_h)\dot{\mathbf{X}}_i\boldsymbol{\beta}_h\\ W_{i2}\sum_{h=1}^{G}(W_{ih}-\rho_h)\dot{\mathbf{X}}_i\boldsymbol{\beta}_h\\ \vdots\\ W_{iG}\sum_{h=1}^{G}(W_{ih}-\rho_h)\dot{\mathbf{X}}_i\boldsymbol{\beta}_h\end{pmatrix}+\begin{pmatrix}W_{i1}U_i(1)\\ W_{i2}U_i(2)\\ \vdots\\ W_{iG}U_i(G)\end{pmatrix}\end{aligned}$$

and so

$$\sqrt{N}\left(\check{\boldsymbol{\mu}}-\boldsymbol{\mu}\right)\qquad(\text{E.2})$$

$$=N^{-1/2}\sum_{i=1}^{N}\left[\begin{pmatrix}\dot{\mathbf{X}}_i\boldsymbol{\beta}\\ \dot{\mathbf{X}}_i\boldsymbol{\beta}\\ \vdots\\ \dot{\mathbf{X}}_i\boldsymbol{\beta}\end{pmatrix}+\begin{pmatrix}\sum_{h=1}^{G}W_{i1}(W_{ih}-\rho_h)\dot{\mathbf{X}}_i\boldsymbol{\beta}_h/\rho_1\\ \sum_{h=1}^{G}W_{i2}(W_{ih}-\rho_h)\dot{\mathbf{X}}_i\boldsymbol{\beta}_h/\rho_1\\ \vdots\\ \sum_{h=1}^{G}W_{iG}(W_{ih}-\rho_h)\dot{\mathbf{X}}_i\boldsymbol{\beta}_h/\rho_1\end{pmatrix}+\begin{pmatrix}W_{i1}U_i(1)/\rho_1\\ W_{i2}U_i(2)/\rho_2\\ \vdots\\ W_{iG}U_i(G)/\rho_G\end{pmatrix}\right]\qquad(\text{k4})$$

For each g, we can write the second term in bracket as follows. Then, combine the first and second parts and simplify using the expression for $\boldsymbol{\beta}$. For example,

$$\begin{aligned}\sum_{g=1}^{G} W_{i1}(W_{ig}-\rho_g)\dot{\mathbf{X}}_i\boldsymbol{\beta}_g/\rho_1 &= \rho_1^{-1}\left[W_{i1}(W_{i1}-\rho_1)\dot{\mathbf{X}}_i\boldsymbol{\beta}_1 - W_{i1}\rho_2\dot{\mathbf{X}}_i\boldsymbol{\beta}_2 - \cdots - W_{i1}\rho_G\dot{\mathbf{X}}_i\boldsymbol{\beta}_G\right] \\ &= \rho_1^{-1}\dot{\mathbf{X}}_i\left[W_{i1}(1-\rho_1)\boldsymbol{\beta}_1 - W_{i1}\rho_2\boldsymbol{\beta}_2 - \cdots - W_{i1}\rho_G\boldsymbol{\beta}_G\right] \\ &= \rho_1^{-1}W_{i1}\dot{\mathbf{X}}_i\left[\boldsymbol{\beta}_1 - (\rho_1\boldsymbol{\beta}_1 + \rho_2\boldsymbol{\beta}_2 + \cdots + \rho_G\boldsymbol{\beta}_G)\right] \\ &= \rho_1^{-1}W_{i1}\dot{\mathbf{X}}_i\left(\boldsymbol{\beta}_1 - \boldsymbol{\beta}\right).\end{aligned}$$

Using (k4) and adding $\dot{\mathbf{X}}_i\boldsymbol{\beta}$ and rearranging, we obtain the following theorem, ■

Proof of Theorem 6

Proof. We now show that, asymptotically, $\hat{\boldsymbol{\mu}}_{FRA}$ is no worse than $\hat{\boldsymbol{\mu}}_{SM}$. From (m1), (m2), $\mathbb{E}\left(\mathbf{L}_i\mathbf{Q}_i'\right)=\mathbf{0}$, and $\mathbb{E}\left(\mathbf{K}_i\mathbf{Q}_i'\right)=\mathbf{0}$, it follows that

$$\text{Avar}\left[\sqrt{N}\left(\hat{\boldsymbol{\mu}}_{SM}-\boldsymbol{\mu}\right)\right] = \boldsymbol{\Omega}_{\mathbf{L}}+\boldsymbol{\Omega}_{\mathbf{Q}}$$

$$\text{Avar}\left[\sqrt{N}\left(\hat{\boldsymbol{\mu}}_{FRA}-\boldsymbol{\mu}\right)\right] = \boldsymbol{\Omega}_{\mathbf{K}}+\boldsymbol{\Omega}_{\mathbf{Q}}$$

where $\boldsymbol{\Omega}_{\mathbf{L}}=\mathbb{E}\left(\mathbf{L}_i\mathbf{L}_i'\right)$ and so on. Therefore, to show that Avar $\left[\sqrt{N}\left(\hat{\boldsymbol{\mu}}_{FRA}-\boldsymbol{\mu}\right)\right]$ is smaller (in the matrix sense), we must show

$$\boldsymbol{\Omega}_{\mathbf{L}}-\boldsymbol{\Omega}_{\mathbf{K}}$$

is PSD, where

$$\mathbf{K}_i=\begin{pmatrix}\dot{\mathbf{X}}_i\boldsymbol{\beta}_1\\ \dot{\mathbf{X}}_i\boldsymbol{\beta}_2\\ \vdots\\ \dot{\mathbf{X}}_i\boldsymbol{\beta}_G\end{pmatrix} \text{ and } \mathbf{L}_i=\begin{pmatrix}W_{i1}\dot{\mathbf{X}}_i\boldsymbol{\beta}_1/\rho_1\\ W_{i2}\dot{\mathbf{X}}_i\boldsymbol{\beta}_2/\rho_2\\ \vdots\\ W_{iG}\dot{\mathbf{X}}_i\boldsymbol{\beta}_G/\rho_G\end{pmatrix}$$

The elements of $\mathbf{L}_i$ are uncorrelated because $W_{ig}W_{ih}=0$ for $g\neq h$. The variance of the g^{th} element is

$$\mathbb{E}\left[\left(W_{ig}\dot{\mathbf{X}}_i\boldsymbol{\beta}_g/\rho_g\right)^2\right]=\mathbb{E}\left(W_{ig}\right)\rho_g^{-2}E\left[\left(\dot{\mathbf{X}}_i\boldsymbol{\beta}_g\right)^2\right]=\rho_g^{-1}\mathbb{E}\left[\left(\dot{\mathbf{X}}_i\boldsymbol{\beta}_g\right)^2\right]=\rho_g^{-1}\boldsymbol{\beta}_g'\boldsymbol{\Omega}_{\mathbf{X}}\boldsymbol{\beta}_g.$$

Therefore,

$$\mathbb{E}\left(\mathbf{L}_i\mathbf{L}_i'\right) = \begin{pmatrix} \boldsymbol{\beta}_1'\boldsymbol{\Omega}_{\mathbf{X}}\boldsymbol{\beta}_1/\rho_1 & 0 & \cdots & 0 \\ 0 & \boldsymbol{\beta}_2'\boldsymbol{\Omega}_{\mathbf{X}}\boldsymbol{\beta}_2/\rho_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \boldsymbol{\beta}_G'\boldsymbol{\Omega}_{\mathbf{X}}\boldsymbol{\beta}_G/\rho_G \end{pmatrix}$$

$$= \mathbf{B}' \begin{pmatrix} \boldsymbol{\Omega}_{\mathbf{X}}/\rho_1 & 0 & \cdots & 0 \\ 0 & \boldsymbol{\Omega}_{\mathbf{X}}/\rho_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \boldsymbol{\Omega}_{\mathbf{X}}/\rho_G \end{pmatrix} \mathbf{B} = \mathbf{B}' \left[\begin{pmatrix} \rho_1^{-1} & 0 & \cdots & 0 \\ 0 & \rho_2^{-1} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \rho_G^{-1} \end{pmatrix} \otimes \boldsymbol{\Omega}_{\mathbf{X}} \right] \mathbf{B}$$

where

$$\mathbf{B} = \begin{pmatrix} \boldsymbol{\beta}_1 & 0 & \cdots & 0 \\ 0 & \boldsymbol{\beta}_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \boldsymbol{\beta}_G \end{pmatrix}$$

For the variance matrix of $\mathbf{K}_i$,

$$\begin{aligned} \mathbb{V}\left(\dot{\mathbf{X}}_i\boldsymbol{\beta}_g\right) &= \boldsymbol{\beta}_g'\boldsymbol{\Omega}_{\mathbf{X}}\boldsymbol{\beta}_g \\ \mathbb{C}(\dot{\mathbf{X}}_i\boldsymbol{\beta}_g, \dot{\mathbf{X}}_i\boldsymbol{\beta}_h) &= \boldsymbol{\beta}_g'\boldsymbol{\Omega}_{\mathbf{X}}\boldsymbol{\beta}_h \end{aligned}$$

Therefore,

$$\mathbb{E}\left(\mathbf{K}_i\mathbf{K}_i'\right) = \mathbf{B}' \begin{pmatrix} \boldsymbol{\Omega}_{\mathbf{X}} & \boldsymbol{\Omega}_{\mathbf{X}} & \cdots & \boldsymbol{\Omega}_{\mathbf{X}} \\ \boldsymbol{\Omega}_{\mathbf{X}} & \boldsymbol{\Omega}_{\mathbf{X}} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \boldsymbol{\Omega}_{\mathbf{X}} \\ \boldsymbol{\Omega}_{\mathbf{X}} & \cdots & \boldsymbol{\Omega}_{\mathbf{X}} & \boldsymbol{\Omega}_{\mathbf{X}} \end{pmatrix} \mathbf{B} = \mathbf{B}' \left[\left(\mathbf{j}_G\mathbf{j}_G'\right) \otimes \boldsymbol{\Omega}_{\mathbf{X}} \right] \mathbf{B}$$

where $\mathbf{j}_G' = (1, 1, ..., 1)$. Therefore, the comparison we need to make is

$$\begin{pmatrix} \rho_1^{-1} & 0 & & 0 \\ 0 & \rho_2^{-1} & & \\ & & & 0 \\ 0 & & 0 & \rho_G^{-1} \end{pmatrix} \otimes \mathbf{\Omega_X} \text{ versus } \left(\mathbf{j}_G \mathbf{j}_G'\right) \otimes \mathbf{\Omega_X}$$

That is, we need to show

$$\left[\begin{pmatrix} \rho_1^{-1} & 0 & \cdots & 0 \\ 0 & \rho_2^{-1} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \rho_G^{-1} \end{pmatrix} - \left(\mathbf{j}_G \mathbf{j}_G'\right) \right] \otimes \mathbf{\Omega_X}$$

is PSD. The Kronecker product of two PSD matrices is also PSD, so it suffices to show

$$\begin{pmatrix} \rho_1^{-1} & 0 & \cdots & 0 \\ 0 & \rho_2^{-1} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \rho_G^{-1} \end{pmatrix} - \left(\mathbf{j}_G \mathbf{j}_G'\right)$$

is PSD when the ρ_g add to unity. Let $\mathbf{a}$ be any $G \times 1$ vector. Then

$$\mathbf{a}' \begin{pmatrix} \rho_1^{-1} & 0 & \cdots & 0 \\ 0 & \rho_2^{-1} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \rho_G^{-1} \end{pmatrix} \mathbf{a} = \sum_{g=1}^{G} a_g^2/\rho_g$$

$$\mathbf{a}' \left(\mathbf{j}_G \mathbf{j}_G'\right) \mathbf{a} = \left(\mathbf{a}' \mathbf{j}_G\right)^2 = \left(\sum_{g=1}^{G} a_g\right)^2$$

So we have to show

$$\sum_{g=1}^{G} a_g^2/\rho_g \geq \left(\sum_{g=1}^{G} a_g\right)^2.$$

Define vectors $\mathbf{b} = \left(a_1/\sqrt{\rho_1}, a_2/\sqrt{\rho_2}, ..., a_G/\sqrt{\rho_G}\right)'$ and $\mathbf{c} = \left(\sqrt{\rho_1}, \sqrt{\rho_2}, ..., \sqrt{\rho_G}\right)'$ and apply the Cauchy-Schwarz inequality:

$$\begin{aligned}\left(\sum_{g=1}^{G} a_g\right)^2 &= \left(\mathbf{b}'\mathbf{c}\right)^2 \leq \left(\mathbf{b}'\mathbf{b}\right)\left(\mathbf{c}'\mathbf{c}\right) = \left(\sum_{g=1}^{G} a_g^2/\rho_g\right)\left(\sum_{g=1}^{G} \rho_g\right) \\ &= \left(\sum_{g=1}^{G} a_g^2/\rho_g\right)\end{aligned}$$

because $\sum_{g=1}^{G} \rho_g = 1$. ■

Proof of Theorem 7

Proof. By random assignment and the linear projection property, $\mathbb{E}(\mathbf{F}_i\mathbf{K}_i') = \mathbb{E}(\mathbf{K}_i\mathbf{Q}_i') = \mathbb{E}(\mathbf{F}_i\mathbf{Q}_i') = \mathbf{0}$. Hence, $\mathbf{F}_i$, $\mathbf{K}_i$, and $\mathbf{Q}_i$ are pairwise uncorrelated. ■

APPENDIX F

AUXILIARY RESULTS FOR CHAPTER 3

F.1 Stratified (or block) experiment with missing outcome

Consider a stratified experiment where the population is partitioned into J strata, or blocks, based on the covariates $\mathbf{x} \in \mathcal{X} \subset \mathcal{R}^J$, given by $\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_J$, such that these sets are mutually exclusive and exhaustive.

Then draw a sample of size N_j from stratum j where $j = 1, 2, \ldots, J$ with $N = \sum_{j=1}^{J} N_j$. Let w_{ijg} be a binary indicator for treatment level $g = 0, 1$ for unit i in stratum j. Then, by construction, the probability of unit i getting treated in stratum j is a function of the covariates that have been used to define the strata. In other words,

$$\mathbb{P}(w_{ijg} = 1 | y_{ij}(0), y_{ij}(1), \mathbf{x}_{ij}) = \mathbb{P}(w_{ijg} = 1 | \mathbf{x}_{ij}) \equiv p_g(\mathbf{x}_{ij}) \text{ for } j = 1, 2, \ldots, J; \quad g = 0, 1$$

Hence, in a stratified experiment, the treatment assignment satisfies unconfoundedness by design, where this probability is constant for all units in a particular stratum j, but varies across the different strata.

Let s_{ij} be a missing data indicator for unit i belonging to stratum j, such that

$$s_{ij} = \begin{cases} 1; \ y_{ij} \text{ is observed} \\ 0; \ y_{ij} \text{ is missing} \end{cases}$$

Then, one can characterize a stratified sample from stratum 'j' as $\{(y_{ij}, \mathbf{x}_{ij}, w_{ijg}, s_{ij}); \ i = 1, \ldots, N_j\}$. Now, suppose that the missing outcomes are ignorable, i.e.

$$\mathbb{P}(s_{ij} = 1 | y_{ij}(0), y_{ij}(1), \mathbf{x}_{ij}, w_{ijg}) = \mathbb{P}(s_{ij} = 1 | \mathbf{x}_{ij}, w_{ijg}) \equiv r(\mathbf{x}_{ij}, w_{ijg})$$

which implies that missingness is sufficiently well predicted by the covariates and the treatment indicator. Given this setup, one can use the doubly weighted estimator to consistently

estimate $\boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}}$ as follows

$$\hat{\boldsymbol{\theta}}_{\mathbf{1}} = \underset{\boldsymbol{\theta}_{\mathbf{1}} \in \boldsymbol{\Theta}_{\mathbf{1}}}{\operatorname{argmin}} \sum_{j=1}^{J} \sum_{i=1}^{N_j} \frac{s_{ij} \cdot w_{ij1}}{r(\mathbf{x}_{ij}, w_{ij1}) \cdot p_1(\mathbf{x}_{ij})} \cdot q(y_{ij}(1), \mathbf{x}_{ij}, \boldsymbol{\theta}_{\mathbf{1}})$$

and

$$\hat{\boldsymbol{\theta}}_{\mathbf{0}} = \underset{\boldsymbol{\theta}_{\mathbf{0}} \in \boldsymbol{\Theta}_{\mathbf{0}}}{\operatorname{argmin}} \sum_{j=1}^{J} \sum_{i=1}^{N_j} \frac{s_{ij} \cdot w_{ij0}}{r(\mathbf{x}_{ij}, w_{ij0}) \cdot p_0(\mathbf{x}_{ij})} \cdot q(y_{ij}(0), \mathbf{x}_{ij}, \boldsymbol{\theta}_{\mathbf{0}})$$

where $r(\mathbf{x}_{ij}, w_{ijg})$ and $p_g(\mathbf{x}_{ij})$ can be replaced by consistent estimators without changing the result. Note that even though the assignment probabilities are typically known in a stratified experiment, it can be asymptotically more efficient to estimate them using binary response MLE.

F.2 Consistent variance estimation

In order to construct asymptotic confidence intervals and obtain valid inference with the doubly weighted estimator, it is important to find a consistent estimator of its asymptotic variance. For smooth objective functions like OLS, NLS, MLE, this task is simple as one can replace the population Hessian and Jacobian functions by their sample counterparts. This involves substituting the sample average in place of the population expectations. However, for non-smooth objective functions, the task of obtaining a consistent variance estimator is not straightforward. The first order or second order derivatives of the objective function may not exist. In such situations, numerical derivatives of the objective functions can be used to approximate the true derivatives. Following Newey and McFadden (1994), let $\boldsymbol{e}_i$ denote the i^{th} unit vector and ε_N denote a small positive constant that depends on the sample size.

For the doubly weighted estimator that solves the treatment problem, $\hat{\boldsymbol{\theta}}_{\boldsymbol{g}}$, the asymptotic variance expression is given as $\mathbf{H}_{\mathbf{g}}^{-\mathbf{1}} \boldsymbol{\Omega}_{\boldsymbol{g}} \mathbf{H}_{\mathbf{g}}^{-\mathbf{1}}$, where $\mathbf{H}_{\mathbf{g}}$ can now be estimated using a second order numerical derivative of the objective function by $\hat{\mathbf{H}}_{\mathbf{g}}$, where the $(j,k)^{th}$ element of

matrix $\hat{\mathbf{H}}_{\mathbf{g}}$ is given as,

$$\begin{aligned}
&\hat{\mathbf{H}}_{\mathbf{g}jk} \\
&= \left[\frac{Q_N(\hat{\boldsymbol{\theta}}_{\boldsymbol{g}} + \boldsymbol{e}_j\varepsilon_N + \boldsymbol{e}_k\varepsilon_N) - Q_N(\hat{\boldsymbol{\theta}}_{\boldsymbol{g}} - \boldsymbol{e}_j\varepsilon_N + \boldsymbol{e}_k\varepsilon_N) - Q_N(\hat{\boldsymbol{\theta}}_{\boldsymbol{g}} + \boldsymbol{e}_j\varepsilon_N - \boldsymbol{e}_k\varepsilon_N)}{4\varepsilon_N^2}\right] \\
&+ \left[\frac{+Q_N(\hat{\boldsymbol{\theta}}_{\boldsymbol{g}} - \boldsymbol{e}_j\varepsilon_N - \boldsymbol{e}_k\varepsilon_N)}{4\varepsilon_N^2}\right]
\end{aligned}$$

For the middle term of the asymptotic variance expression which is $\boldsymbol{\Omega}_{\boldsymbol{g}} = \mathbb{E}\left(\mathbf{u}_{ig}\mathbf{u}_{ig}'\right)$, we can approximate it as $\hat{\boldsymbol{\Omega}}_{\boldsymbol{g}} = \frac{1}{N}\sum_{i=1}^{N}\hat{\mathbf{u}}_{ig}\hat{\mathbf{u}}_{ig}'$, where $\mathbf{u}_{ig}$ exists with probability one. Hence, $\hat{\mathbf{H}}_{\mathbf{g}}^{-\mathbf{1}}\hat{\boldsymbol{\Omega}}_{\boldsymbol{g}}\hat{\mathbf{H}}_{\mathbf{g}}^{-\mathbf{1}}$ will be consistent under the conditions of the following theorem.

Theorem F.2.1. *(Consistency of asymptotic variance) Suppose that* $\varepsilon_N \to 0$ *and* $\varepsilon_N\sqrt{N} \to \infty$*, then under conditions of theorem 3.4.2,* $\hat{\mathbf{H}}_{\boldsymbol{g}} \xrightarrow{p} \mathbf{H}_{\boldsymbol{g}}$ *and* $\hat{\boldsymbol{\Omega}}_{\boldsymbol{g}} \xrightarrow{p} \boldsymbol{\Omega}_{\boldsymbol{g}}$*.*

The proof of this theorem is given in appendix J and follows from Theorem 7.4 in Newey and McFadden (1994).

Table **??** characterizes cases when the weighted and unweighted estimator will be consistent for the true parameter, $\boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}}$. Table I.4 talks about situations when the unweighted estimator is more efficient than the weighted estimator.

F.3 Asymptotic variance for ATE

Given $\sqrt{N}$ consistent and asymptotically normal estimators, $\hat{\boldsymbol{\theta}}_{\mathbf{1}}$ and $\hat{\boldsymbol{\theta}}_{\mathbf{0}}$, the estimated average treatment effect

$$\hat{\tau}_{\text{ate}} = \frac{1}{N}\sum_{i=1}^{N} m_1(\mathbf{x}_i, \hat{\boldsymbol{\theta}}_{\mathbf{1}}) - \frac{1}{N}\sum_{i=1}^{N} m_0(\mathbf{x}_i, \hat{\boldsymbol{\theta}}_{\mathbf{0}})$$

is easily shown to also be $\sqrt{N}$-consistent and asymptotically normal (Wooldridge (2010) chapter 21). Regularity conditions for such an asymptotic result would require that the parametric model, $m_g(\mathbf{x}, \boldsymbol{\theta}_{\boldsymbol{g}})$, is continuously differentiable on the parameter space $\boldsymbol{\Theta}_{\boldsymbol{g}} \subset \mathfrak{R}^{Pg}$ and $\boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}}$ is in the interior of $\boldsymbol{\Theta}_{\boldsymbol{g}}$. Then, by the continuous mapping theorem and slutsky's

theorem,

$$\sqrt{N}\left(\hat{\tau}_{\text{ate}} - \tau_{\text{ate}}\right) \xrightarrow{d} N(0, \text{V})$$

where $\text{V} = \mathbb{E}\left[\psi(\mathbf{x}_i)\psi(\mathbf{x}_i)'\right]$. Let's denote $\mathbb{E}\left[\boldsymbol{\nabla}_{\boldsymbol{\theta_g}} m_g(\mathbf{x}_i, \boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}})\right] \equiv \mathbf{G}_{\boldsymbol{g}}^{\mathbf{0}}$, then

$$\psi(\mathbf{x}_i) = \{m_1(\mathbf{x}_i, \boldsymbol{\theta}_{\mathbf{1}}^{\mathbf{0}}) - m_0(\mathbf{x}_i, \boldsymbol{\theta}_{\mathbf{0}}^{\mathbf{0}}) - \tau_{\text{ate}}\} - \mathbf{G}_{\mathbf{1}}^{\mathbf{0}} \cdot \mathbf{H}_{\mathbf{1}}^{-\mathbf{1}}\mathbf{u}_{i\mathbf{1}} + \mathbf{G}_{\mathbf{0}}^{\mathbf{0}} \cdot \mathbf{H}_{\mathbf{0}}^{-\mathbf{1}}\mathbf{u}_{i\mathbf{0}}$$

where $\mathbf{H}_{\boldsymbol{g}}$ is the Hessian for the treatment group g, and $\mathbf{u}_{ig}$ is the residual from the regression of the weighted score on the scores of two probability models. For the case when the conditional mean model is correctly specified, the variance expression simplifies to

$$\text{V} = \mathbb{E}\left[(m_1(\mathbf{x}_i, \boldsymbol{\theta}_{\mathbf{1}}^{\mathbf{0}}) - m_0(\mathbf{x}_i, \boldsymbol{\theta}_{\mathbf{0}}^{\mathbf{0}})) - \tau_{\text{ate}}\right]^2 + \mathbf{G}_{\mathbf{1}}^{\mathbf{0}} \cdot \mathbf{V}_{\mathbf{1}} \cdot \mathbf{G}_{\mathbf{1}}^{\mathbf{0}\prime} + \mathbf{G}_{\mathbf{0}}^{\mathbf{0}} \cdot \mathbf{V}_{\mathbf{0}} \cdot \mathbf{G}_{\mathbf{0}}^{\mathbf{0}\prime} \tag{F.1}$$

Here $\mathbf{V_1}$ and $\mathbf{V_0}$ are the asymptotic variances of the doubly weighted estimator that solve the treatment and control group problems respectively. The above formula makes it clear that it better to use more efficient estimators of $\hat{\boldsymbol{\theta}}_{\boldsymbol{g}}$. But we know from the results in section 3.5 that when the conditional mean model is correctly specified, using estimated weights is as efficient as using known weights. Another alternative in this case is to use unweighted estimators of $\boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}}$ since under GCIME, unweighted estimators can be potentially more efficient than the doubly weighted estimators of $\boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}}$.

For the case when the mean model is misspecified, the asymptotic variance of the ATE is given as follows

$$\begin{aligned}\text{V} = {} & \mathbb{E}\left[(m_1(\mathbf{x}_i, \boldsymbol{\theta}_{\mathbf{1}}^{\mathbf{0}}) - m_0(\mathbf{x}_i, \boldsymbol{\theta}_{\mathbf{0}}^{\mathbf{0}})) - \tau_{\text{ate}}\right]^2 + \mathbf{G}_{\mathbf{1}}^{\mathbf{0}} \cdot \mathbf{V}_{\mathbf{1}} \cdot \mathbf{G}_{\mathbf{1}}^{\mathbf{0}\prime} + \mathbf{G}_{\mathbf{0}}^{\mathbf{0}} \cdot \mathbf{V}_{\mathbf{0}} \cdot \mathbf{G}_{\mathbf{0}}^{\mathbf{0}\prime} \\ & - 2\mathbb{E}\left[\{m_1(\mathbf{x}_i, \boldsymbol{\theta}_{\mathbf{1}}^{\mathbf{0}}) - m_0(\mathbf{x}_i, \boldsymbol{\theta}_{\mathbf{0}}^{\mathbf{0}}) - \tau_{\text{ate}}\}\mathbf{u}_{i1}'\right] \mathbf{H}_{\mathbf{1}}^{-\mathbf{1}}\mathbf{G}_{\mathbf{1}}^{\mathbf{0}\prime} \\ & + 2\mathbb{E}\left[\{m_1(\mathbf{x}_i, \boldsymbol{\theta}_{\mathbf{1}}^{\mathbf{0}}) - m_0(\mathbf{x}_i, \boldsymbol{\theta}_{\mathbf{0}}^{\mathbf{0}}) - \tau_{\text{ate}}\}\mathbf{u}_{i0}'\right] \mathbf{H}_{\mathbf{0}}^{-\mathbf{1}}\mathbf{G}_{\mathbf{0}}^{\mathbf{0}\prime}\end{aligned} \tag{F.2}$$

In this case, the variance expression is a bit more complicated than the previous case. Even though it is better to have more efficient estimators of $\boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}}$ in this case as well, it is not obvious whether that would help obtain a smaller variance for the ATE since we now have cross correlation terms in the variance expression. The proof of the asymptotic variances is provided in appendix .....

F.4 Practical advice for obtaining double-weighted ATE estimates

An easy way to obtain the doubly weighted estimates, $\hat{\boldsymbol{\theta}}_{\boldsymbol{g}}$, for estimating ATE, is to combine the treatment and control group problems into a one-step GMM procedure. Essentially, this means that one would stack the moment conditions from the first and second steps, which can then be solved jointly via GMM. Since there are no over-identifying restrictions in the double weighted framework, one-step estimation of $\boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}}$ is equivalent to two-step estimation (Negi (2019)). For ease of notation, let $w_{i1} = w_i$ and $w_{i0} = (1 - w_i)$. Then, consider the following set of moment conditions:

$$\bar{\mathbf{m}}(\boldsymbol{\theta_0}, \boldsymbol{\theta_1}, \boldsymbol{\gamma}, \boldsymbol{\delta}) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{m}_i(\boldsymbol{\theta_0}, \boldsymbol{\theta_1}, \boldsymbol{\gamma}, \boldsymbol{\delta}) = N^{-1} \begin{pmatrix} \frac{N}{N_0} \cdot \sum_{i=1}^{N} \mathbf{m}_{i\mathbf{0}}(\boldsymbol{\theta_0}, \boldsymbol{\gamma}, \boldsymbol{\delta}) \\ \frac{N}{N_1} \cdot \sum_{i=1}^{N} \mathbf{m}_{i\mathbf{1}}(\boldsymbol{\theta_1}, \boldsymbol{\gamma}, \boldsymbol{\delta}) \\ \sum_{i=1}^{N} \mathbf{m}_{i\mathbf{2}}(\boldsymbol{\gamma}) \\ \sum_{i=1}^{N} \mathbf{m}_{i\mathbf{3}}(\boldsymbol{\delta}) \end{pmatrix}$$

where,

$$\begin{aligned} \mathbf{m}_{i\mathbf{0}}(\boldsymbol{\theta_0}, \boldsymbol{\gamma}, \boldsymbol{\delta}) &= \frac{s_i \cdot (1 - w_i)}{R(\mathbf{x}_i, w_i, \hat{\boldsymbol{\delta}}) \cdot (1 - G(\mathbf{x}_i, \hat{\boldsymbol{\gamma}}))} \cdot \boldsymbol{\nabla}_{\boldsymbol{\theta_0}} q(y_i(0), \mathbf{x}_i, \boldsymbol{\theta_0})' \\ \mathbf{m}_{i\mathbf{1}}(\boldsymbol{\theta_1}, \boldsymbol{\gamma}, \boldsymbol{\delta}) &= \frac{s_i \cdot w_i}{R(\mathbf{x}_i, w_i, \hat{\boldsymbol{\delta}}) \cdot G(\mathbf{x}_i, \hat{\boldsymbol{\gamma}})} \cdot \boldsymbol{\nabla}_{\boldsymbol{\theta_1}} q(y_i(1), \mathbf{x}_i, \boldsymbol{\theta_1})' \\ \mathbf{m}_{i\mathbf{2}}(\boldsymbol{\gamma}) &= \boldsymbol{\nabla}_{\boldsymbol{\gamma}} G(\mathbf{x}_i, \boldsymbol{\gamma})' \cdot \frac{w_i - G(\mathbf{x}_i, \boldsymbol{\gamma})}{G(\mathbf{x}_i, \boldsymbol{\gamma}) \cdot (1 - G(\mathbf{x}_i, \boldsymbol{\gamma}))} \\ \mathbf{m}_{i\mathbf{3}}(\boldsymbol{\delta}) &= \boldsymbol{\nabla}_{\boldsymbol{\delta}} R(\mathbf{x}_i, w_i, \boldsymbol{\delta})' \cdot \frac{s_i - R(\mathbf{x}_i, w_i, \boldsymbol{\delta})}{R(\mathbf{x}_i, w_i, \boldsymbol{\delta}) \cdot (1 - R(\mathbf{x}_i, w_i, \boldsymbol{\delta}))} \end{aligned}$$

The example code below uses STATA's `gmm` command to estimate two weighted linear regressions for estimating ATE.

Example code using STATA's gmm

```
local Rhat="exp(b31+b32*w+b33*x1+b34*x2)/(1+exp(b31+b32*w+b33*x1+b34*x2))"
local Ghat="exp(b21+b22*x1+b23*x2)/(1+exp(b21+b22*x1+b23*x2))"

gmm ((-2*s*(1-w)/(`Rhat'*(1-`Ghat')))*(y-b00-b01*x1-b02*x2)*(n/nc)) ///
((-2*s*w/(`Rhat'*`Ghat'))*(y-b10-b11*x1-b12*x2)*(n/nt)) ///
(w-exp(b21+b22*x1+b23*x2)/(1+exp(b21+b22*x1+b23*x2))) ///
(s-exp(b31+b32*w+b33*x1+b34*x2)/(1+exp(b31+b32*w+b33*x1+b34*x2))), ///
instruments(1 2 3:  x1 x2) instruments(4:  w x1 x2) winitial(identity) ///
nocommonesample onestep from(b00 0.1 b01 0.1 b02 0.1 b10 0.1 b11 0.1 b12
///
0.1 b21 0.1 b22 0.1 b23 0.1 b31 0.1 b32 0.1 b33 0.1 b34 0.1)
```

Then using the GMM estimates, one can estimate the average treatment effect as

```
gen y0hat = _b[b00:  _cons]+_b[b01:  _cons]*x1+_b[b02:  _cons]*x2
gen y1hat = _b[b10:  _cons]+_b[b11:  _cons]*x1+_b[b12:  _cons]*x2
egen ate = mean(y1hat-y0hat)
```

Since I am estimating the two probability models as logits (as is the convention in applied work), the third and fourth moments simplify to

$$\mathbf{m}_{i\mathbf{2}}(\boldsymbol{\gamma}) = \mathbf{x}_i' \cdot (w_i - \Lambda(\mathbf{x}_i\boldsymbol{\gamma}))$$

$$\mathbf{m}_{i\mathbf{3}}(\boldsymbol{\delta}) = \mathbf{z}_i' \cdot (s_i - \Lambda(\mathbf{z}_i\boldsymbol{\delta}))$$

Even though this one-step estimation allows us to obtain variance estimates $\hat{\mathbf{V}}_1$ and $\hat{\mathbf{V}}_0$ for $\hat{\boldsymbol{\theta}}_\mathbf{1}$ and $\hat{\boldsymbol{\theta}}_\mathbf{0}$ respectively, obtaining analytically correct standard errors for estimated ATE requires additional work. A command that implements the correct standard errors is still in the works. Meanwhile, one can use bootstrapped standard errors, which provide asymptotically correct inference.

F.5 Asymptotic variance for QTEs

Given that $\hat{\boldsymbol{\theta}}_{\boldsymbol{g}}$ are $\sqrt{N}$-consistent and asymptotically normal for the CQF parameters, $\boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}}$ (conditions for QR & Komunjer type QMLE estimators can be easily verified and can be found in basic textbooks), the estimated CQTE$_\tau$ will also be $\sqrt{N}$-consistent and asymptotically normal under the condition that $quant_{g,\tau}(\mathbf{x}, \boldsymbol{\theta}_{\boldsymbol{g}})$ is continuously differentiable on the parameter space $\boldsymbol{\Theta}_{\boldsymbol{g}} \subset \Re^{Pg}$ and $\boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}}$ is an interior point in $\boldsymbol{\Theta}_{\boldsymbol{g}}$ (just like for the case of ATE). Then, again, by the continuous mapping theorem, we obtain

$$\sqrt{N}\left(\hat{CQTE}_\tau(\mathbf{x}_i) - CQTE_\tau(\mathbf{x}_i)\right) \xrightarrow{d} N(0, \mathrm{V}(\mathbf{x}_i))$$

where

$$\begin{aligned}\mathrm{V}(\mathbf{x}_i) = &\boldsymbol{\nabla}_{\boldsymbol{\theta_1}} quant_{1,\tau}(\mathbf{x}_i, \boldsymbol{\theta}_{\mathbf{1}}^{\mathbf{0}}) \mathbf{V_1} \boldsymbol{\nabla}_{\boldsymbol{\theta_1}} quant_{1,\tau}(\mathbf{x}_i, \boldsymbol{\theta}_{\mathbf{1}}^{\mathbf{0}})' \\ &+ \boldsymbol{\nabla}_{\boldsymbol{\theta_1}} quant_{0,\tau}(\mathbf{x}_i, \boldsymbol{\theta}_{\mathbf{0}}^{\mathbf{0}}) \mathbf{V_0} \boldsymbol{\nabla}_{\boldsymbol{\theta_0}} quant_{0,\tau}(\mathbf{x}_i, \boldsymbol{\theta}_{\mathbf{0}}^{\mathbf{0}})'\end{aligned}$$

In the case when we are able to consistently estimate the CQTE, the researcher may be interested in a quantity which I call the average quantile effect (AQE). This is defined to be the average difference in the CQTE function at a given quantile. Using the weak law of large numbers, one can also establish that

$$\sqrt{N}(\hat{AQE}_\tau - AQE_\tau) \xrightarrow{d} N(0, \mathrm{V})$$

where

$$\begin{aligned}\mathrm{V} = &\mathbb{E}\left[\left\{\left(quant_{1,\tau}(\mathbf{x}_i, \boldsymbol{\theta}_{\mathbf{1}}^{\mathbf{0}}) - quant_{0,\tau}(\mathbf{x}_i, \boldsymbol{\theta}_{\mathbf{0}}^{\mathbf{0}})\right) - AQE_\tau\right\}^2\right] + \mathbb{E}\left[\boldsymbol{\nabla}_{\boldsymbol{\theta_1}} quant_{1,\tau}(\mathbf{x}_i, \boldsymbol{\theta}_{\mathbf{1}}^{\mathbf{0}})\right] \cdot \mathbf{V_1} \cdot \\ &\mathbb{E}\left[\boldsymbol{\nabla}_{\boldsymbol{\theta_1}} quant_{1,\tau}(\mathbf{x}_i, \boldsymbol{\theta}_{\mathbf{1}}^{\mathbf{0}})\right]' + \mathbb{E}\left[\boldsymbol{\nabla}_{\boldsymbol{\theta_1}} quant_{0,\tau}(\mathbf{x}_i, \boldsymbol{\theta}_{\mathbf{0}}^{\mathbf{0}})\right] \cdot \mathbf{V_0} \cdot \mathbb{E}\left[\boldsymbol{\nabla}_{\boldsymbol{\theta_0}} quant_{0,\tau}(\mathbf{x}_i, \boldsymbol{\theta}_{\mathbf{0}}^{\mathbf{0}})\right]'\end{aligned}$$

and $\mathbf{V_1}$ and $\mathbf{V_0}$ are the asymptotic variances of the doubly weighted estimator that solves the QR or QMLE problem for the treatment and control groups respectively. The derivation of the two asymptotic variances is provided appendix H. For average quantile effect, the derivation proceeds in a similar manner to the case of ATE. Since the above results hinge on

correct quantile specification, one may use the usual robust asymptotic variance form for $\mathbf{V_1}$ and $\mathbf{V_0}$. However, one might be able to obtain a smaller finite sample variance from using estimated weights even though weighting would not have any bite in establishing consistency here.

As discussed in the examples section, when the conditional quantile model is misspecified, $\boldsymbol{\theta_g^0}$ can still be interpreted as a weighted linear approximation parameter to the true τ-CQF of $y(g)$. Since linear projections can be used as linear operators, the difference in the two linear projections of the two potential outcomes will give us a linear projection to the true CQTE. Formally,

$$\text{LP}\left[\text{CQTE}_\tau\right] = \text{LP}[quant_{1,\tau}(\mathbf{x}_i, \boldsymbol{\theta_1})] - \text{LP}[quant_{0,\tau}(\mathbf{x}_i, \boldsymbol{\theta_0})]$$

Therefore, one can use $\boldsymbol{\theta_g^0}$ in the case of a misspecified CQF to define a linear projection to the true CQTE.

APPENDIX G

APPLICATION APPENDIX FOR CHAPTER 3

G.1 National Supported Work Program

The NSW was a transitional and subsidized work experience program that was mainly intended to target four sub-populations; ex-offenders, former drug addicts, women on AFDC welfare and high school dropouts.[1] The program became operational in 1975 and continued until 1979 at fifteen locations in the United States. In ten of these sites, the program operated as a randomized experiment where individuals who qualified for the training program were randomly assigned to either the treatment or control group.[2] At the time of enrollment in April 1975, individuals were given a retrospective baseline survey which was then followed by four follow-up interviews conducted at nine month intervals each. The survey data was collected using these baseline and follow-up interviews over a period of four years. The data included measurement on baseline covariates like age, years of education, number of children in 1975, high school dropout status, marital status, two race indicators for black and Hispanic sub-populations and other demographic and socio-economic information. The main outcome of interest was real earnings for the post-training year of 1979.

G.2 Augmenting the Calónico and Smith (2017) sample to account for missing earnings in 1979

I obtain the data from Calónico and Smith (2017)'s supplementary data files in the Journal of Labor Economics where the authors recreate the experimental sample on AFDC

[1]The AFDC program is administered and funded by the federal and state governments and is meant to provide financial assistance to needy families. *Source*: US Census Bureau. Beyond the main eligibility criteria that was applied to all four target populations, the AFDC group was subjected to two additional criteria which were, a) no child below 6 years of age and b) on AFDC welfare for at least 30 of the last 36 months.

[2]Out of the 10 sites, 7 served AFDC women with random assignment at one or more of these sites in operation from Feb 1976-Aug 1977 (Calónico and Smith (2017)).

women using the raw public use data files maintained by the Inter-University Consortium for Political and Social Research (ICPSR). Then, I use the PSIDcross file provided by CS along with other supplementary data files to add back the individuals whom CS originally dropped from the analysis for not having valid earnings information between 1975-1979. For this, I apply the same filters applied by CS who use them to match their PSID samples to the ones used by LaLonde (1986). These filters involve keeping all female household heads continuously from 1975-1979 who were between 20 and 55 years of age in 1975 and were not retired in 1975.[3] This constitutes the first non-experimental sample that CS use in their analysis, which they call the PSID-1 sample. The second PSID sample, which they label PSID-2 further restricts the PSID-1 sample to include only those women who received AFDC welfare in 1975.[4] In order to compare my sample with the original sample used by CS, I first apply all the above mentioned filters and create a dummy variable which I call "cs". Next, I remove the filter which requires the women to be continuous household heads and instead only impose that filter for 1975 and 1976. The reason this filter is imposed for both years 1975 and 1976 but not for any other years is because in the PSID datasets, the income information in a particular year corresponds to the previous calendar year. This ensures that merging the cross-file with the separate single-year files for 1975 and 1976 guarantee that only those women are included who do not have any missing earnings information for the pre-training year of 1974 and 1975. This is important since pre-training earnings are treated as any other baseline covariate in this paper, on which I do not allow any missing information.

After merging cross year individual file with the single year family files, I then merge this PSID dataset with the NSW dataset using CS's .do files and generate the various sample

[3]For the additional filters that CS impose, see the Calónico and Smith (2017) supplementary material provided in JLE.

[4]Even though the two PSID comparison groups are not perfectly representative of women who would have proven eligible for NSW, there is no clear alternative since the PSID data lacks detailed covariate information that would be needed to impose the full eligibility criteria on the PSID sample.

dummies essentially in the same manner as they do. After this, I further restrict the sample to include only those women who have valid earnings information in 1975, which is the pre-training year for AFDC women. I also drop the cases where the measured age or education is less than zero. In order to make sure that any observations not used by CS only correspond to the ones that have missing post-program earnings, I also drop observations that do not satisfy the CS criteria but have observed earnings in 1979.

G.3 Treatment and missing outcome probability specifications and sample trimming

In this application, I estimate three sets of treatment assignment and missing outcomes probability models depending upon which comparison group is used for obtaining the estimates. For the experimental estimates, I use the experimental treatment and control groups to estimate the propensity score model. For the PSID-1 estimates, I consider the NSW experimental observations to be the treatment group and use PSID-1 as the control group. For estimating the PSID-2 propensity score model, I switch to PSID-2 as being the comparison control group. For estimating the missing outcome probability models, I include the treatment indicator depending upon the comparison group as mentioned above. The probability models are estimated as logits and include the following covariates in their specification. For the treatment probability, I include the real earnings in 1974 and 1975 along with an indicator variable for whether the individual had any zero earnings in 1974 and 1975. Beyond these, I also include Age, Age-squared, Education, High school dropout status, the race indicators of black and Hispanic along as well as the number of children in 1975. CS also add some interaction terms in their propensity score specification which I do not. I noticed that allowing for those terms in my specifications drove the final weights for many women in the sample too close to a 0 or 1. For the missing outcomes probability, I include the treatment indicator along with the same covariates. I kept the specifications to be the same for the three sets of probabilities I estimated. However, my regression specifications

include the same covariates as CS to allow for some comparison across the analyses. These comparisons should be made with some caution. Except the estimates that use the NSW control group, all other estimates are obtained using samples that are different than the CS samples.

The final sample used to obtain estimates for the PSID-1 comparison group is trimmed in order to ensure common support for the weights in the treatment and comparison groups. For the PSID-1 group, this meant dropping observations with final weight either less than 0.03 or greater than 0.8. For the PSID-2 sample, this meant dropping observations with final weight that was either less than 0.1 or greater than 0.86. These final weights are the weights that are specified in the regression commands in STATA and are constructed as follows:

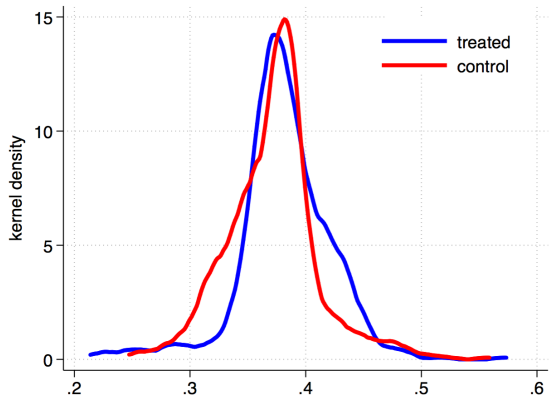
```
weight = (w/Ghat+(1-w)/(1-Ghat))*(s/Rhat)
```

The trimming threshold for PS-weighted estimates is kept the same as for computing the double weighted estimates since the overlap problem was relatively more severe when using the composite weights than when using propensity scores only. The graphs below plot the kernel density for the probabilities `Rhat*Ghat` for the treatment group and `Rhat*(1-Ghat)` for the control group. The common support problem due to which the samples were appropriately trimmed can be seen in the graphs below.

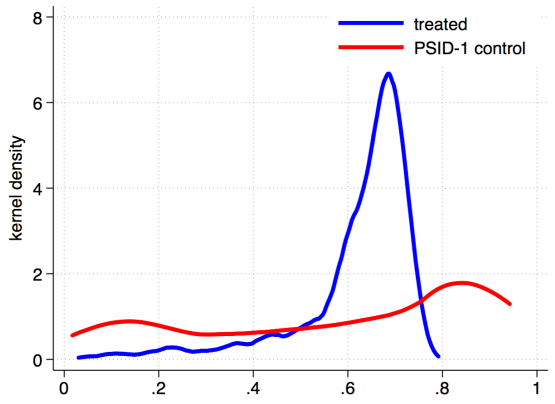
Additionally, figures G.2 and G.3 plot the estimated distributions for the propensity score and missing outcomes probability, where panel (a)-(c) display these for the three treatment and comparison group combinations. A couple of points emerge from the estimated graphs. For figure G.2, panel (a), we see that the treatment and control distributions appear very similar, confirming the strong role of randomization in producing groups that are balanced in terms of covariates. For panel (b), we see that the experimental observations have a relatively high probability of being treated whereas the control group have low probabilities. Note, however, that the common support condition holds quite strongly for the PSID-1 group. In panel (c), while the estimated distribution for the treated units still has a higher mean, the PSID-2 comparison group distribution is relatively similar than PSID-1 in panel (b).

Figure G.1: Kernel density plots for the composite probability

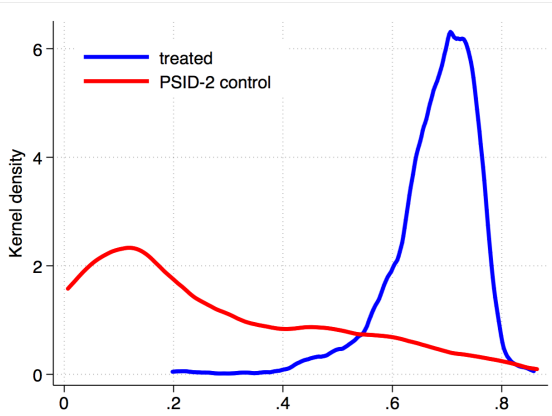
a) Experimental treatment and control



b) Experimental treatment and PSID-1



c) Experimental treatment and PSID-2

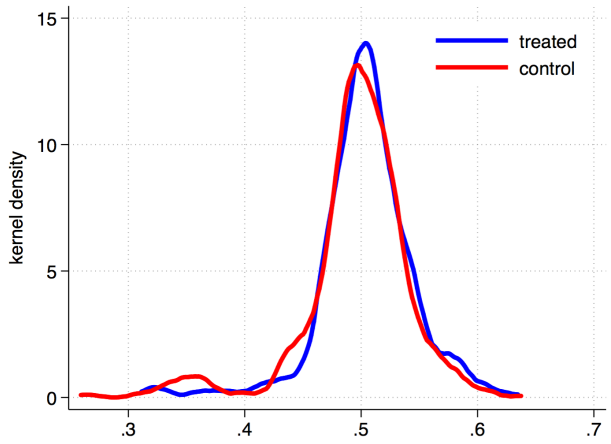


Notes: The weights here correspond to the product of the estimated assignment and missing outcomes probabilities. Following Calónico and Smith (2017), I exploit the efficiency gain from combining the experimental treatment and control groups for estimating the treatment and missing outcome probability models. For the PSID-1 group, this means using the full experimental group to be the treatment group and the PSID-1 as the control group. Similarly, to construct weights for the PSID-2 group, this means using the full experimental group along with the PSID-2 as the control group.

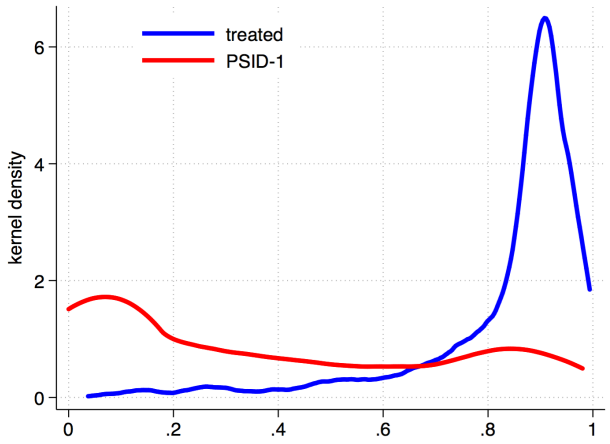
These findings suggest that nonrandom assignment is predicted well by the covariates in the propensity score distributions. The same cannot be said for the estimated missing outcomes probabilities where panel (b) and (c) reveal a strong overlap problem. Moreover, we see that the treated units are less likely to be missing outcomes compared to the comparison groups.

Figure G.2: Kernel density plots for the estimated propensity score

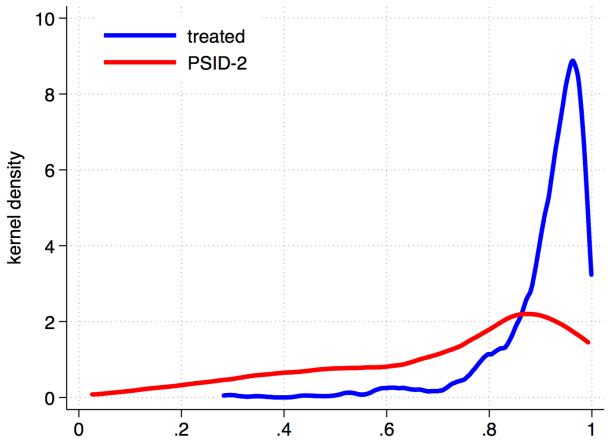
a) Experimental treatment and control



b) Experimental treatment and PSID-1



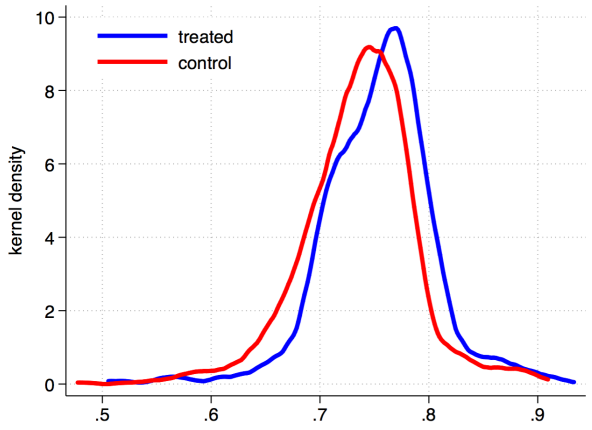
c) Experimental treatment and PSID-2



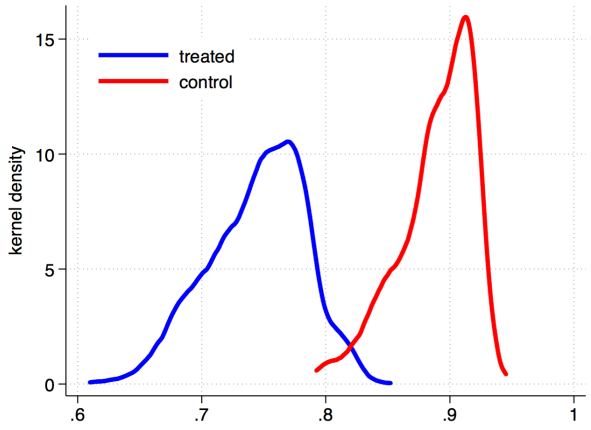
Notes: Following Calónico and Smith (2017), I exploit the efficiency gains from combining the experimental treatment and control groups for estimating the propensity scores. For the PSID-1 group, this means using the full experimental group to be the treatment group and the PSID-1 as the control group. Similarly, to construct weights for the PSID-2 group, this means using the full experimental group along with the PSID-2 as the control group.

Figure G.3: Kernel density plots for the estimated missing outcomes probability

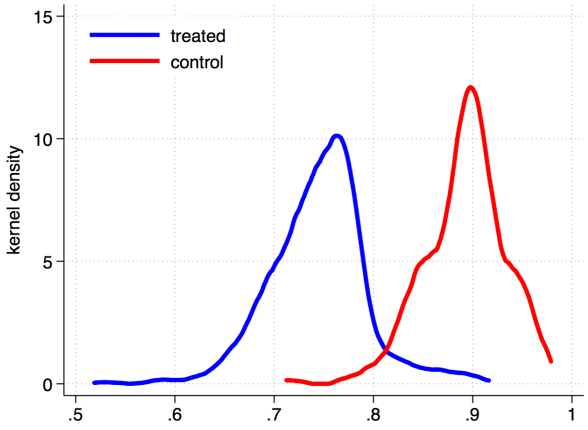
a) Experimental treatment and control



b) Experimental treatment and PSID-1



c) Experimental treatment and PSID-2



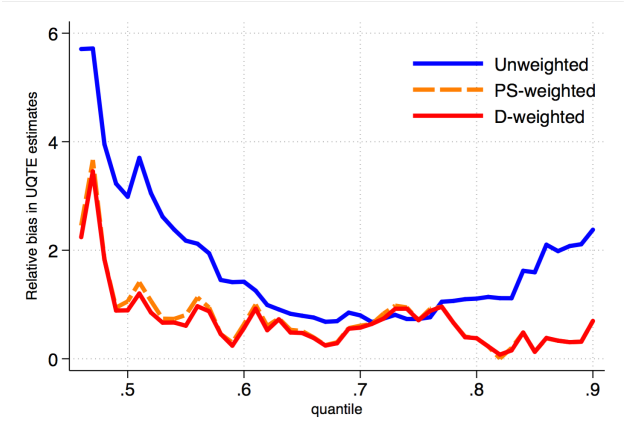
Notes: Following Calónico and Smith (2017), I exploit the efficiency gains from combining the experimental treatment and control groups for estimating the missing outcome probability. For the PSID-1 group, this means using the full experimental group to be the treatment group and the PSID-1 as the control group. Similarly, to construct weights for the PSID-2 group, this means using the full experimental group along with the PSID-2 as the control group.

APPENDIX H

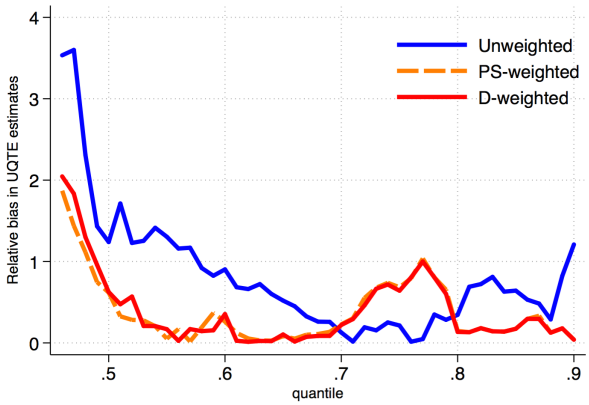
FIGURES FOR CHAPTER 3

Figure H.1: Relative estimated bias in UQTE estimates at different quantiles of the 1979 earnings distribution

a) PSID-1 control group



b) PSID-2 control group

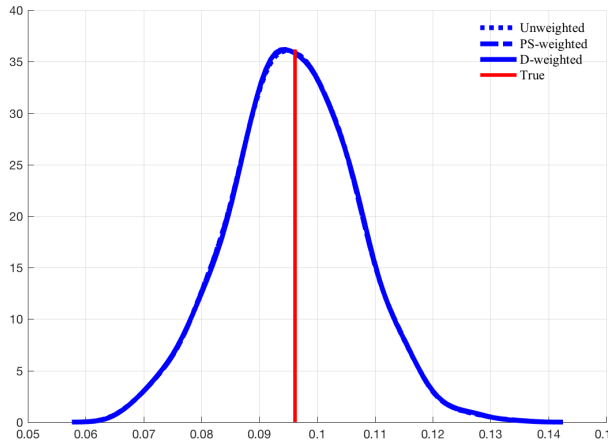


Notes: This graph plots the bias in the unweighted, PS-weighted and doubly-weighted UQTE estimates relative to the true experimental estimates across different quantiles of the 1979 earnings distribution. Panel (a) plots the relative bias estimates using the PSID-1 comparison group and Panel (b) plots the same using the PSID-2 comparison group. The treatment and missing outcome propensity score models have been estimated as flexible logits and the samples used for constructing these estimates have been trimmed to ensure common support across the two groups. The treatment propensity score has been estimated using the full experimental sample along with either PSID-1 or PSID-2 comparison group. The UQTE estimates for $\tau < 0.46$ are omitted from the graph since these are zero.

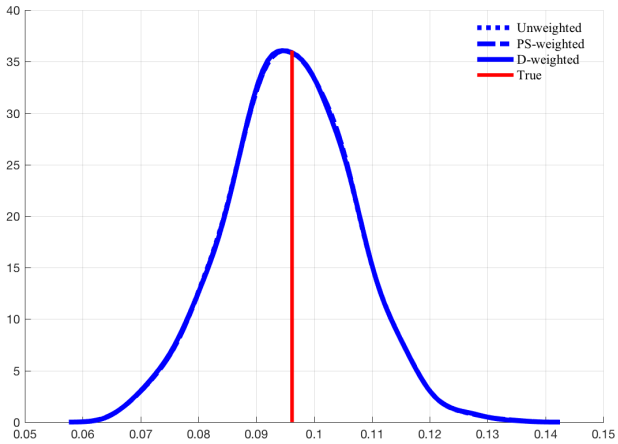
Figure H.2: Empirical distribution of estimated ATE for N=5000

Case 1: When conditional mean model is correct

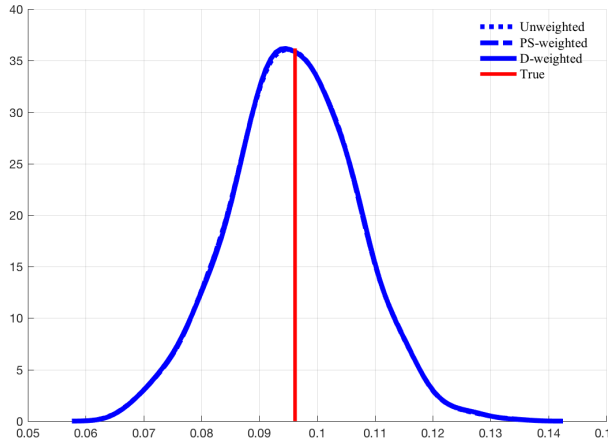
a) Both probability models are correct



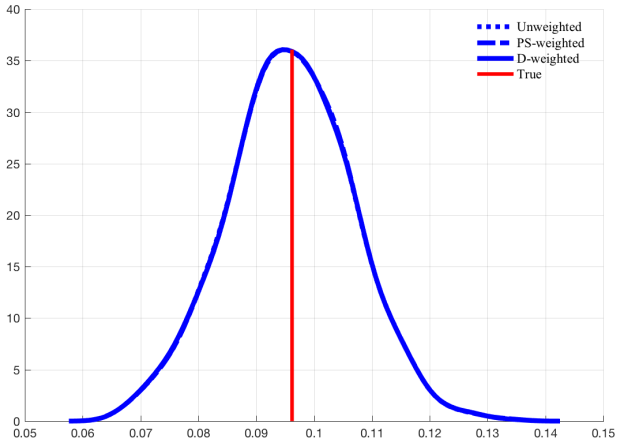
b) Misspecified propensity score model



c) Misspecified missingness model



d) Both probability models are misspecified

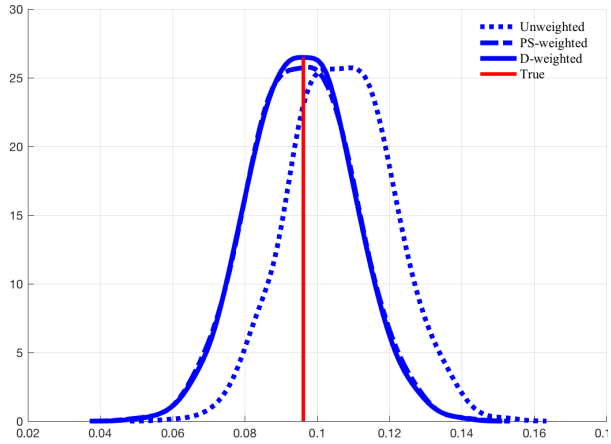


Notes: The empirical distribution is obtained from 1000 simulation draws of sample size 5000. However, the effective sample sizes are much smaller. Since the average propensity of treatment is a 0.41 and average propensity of being observed as 0.38, the average treated sample is $N_1 = 5000 \times 0.41 \times 0.38 = 779$ and average control sample is $N_0 = 5000 \times (1 - 0.41) \times 0.38 = 1,121$. The true ATE $= 0.096$. The graphs display the empirical distribution of the estimated ATE with correct mean specification under three different cases of misspecification of the probability models. For the fourth case, see the main text. The graphs communicate the theoretical findings of this paper which state that under correct specification of the conditional model (conditional mean for these simulations), unweighted and weighted estimators will all be consistent for the true average treatment effect. Hence, correct specification of the probabilities does not have any added bite here in terms of achieving consistency.

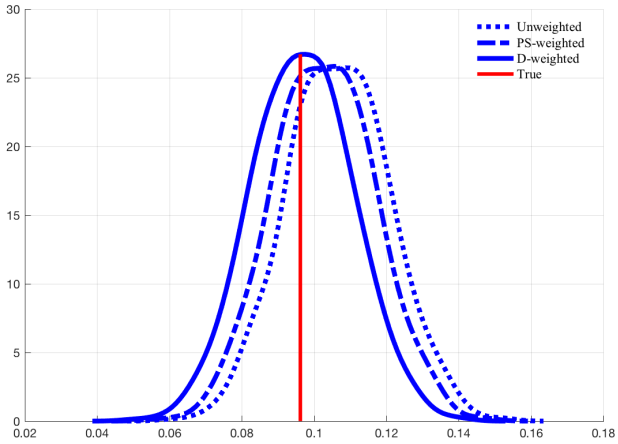
Figure H.2 (cont'd)

Case 2: When the conditional mean model is misspecified

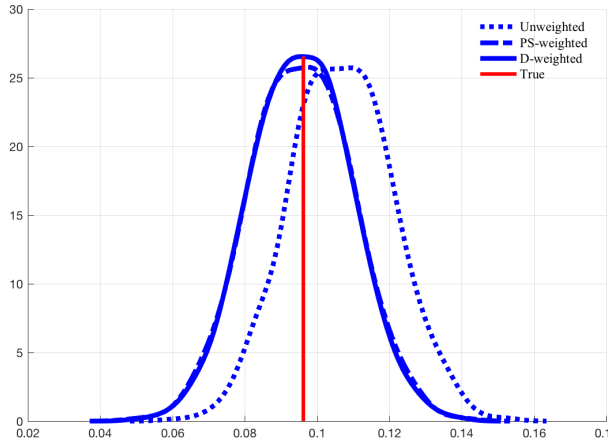
a) Both probability models are correct



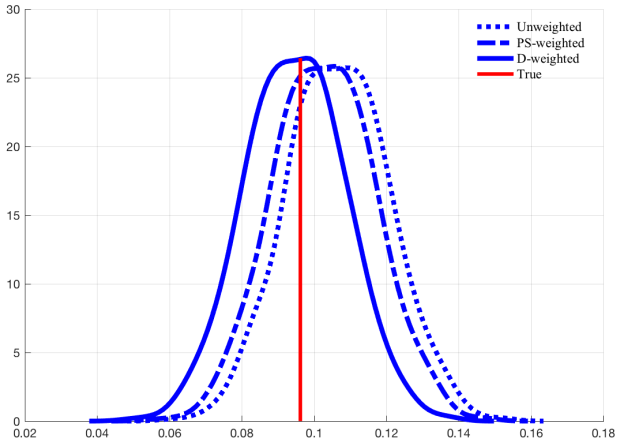
b) Misspecified propensity score model



c) Misspecified missingness model



d) Both probability models are misspecified

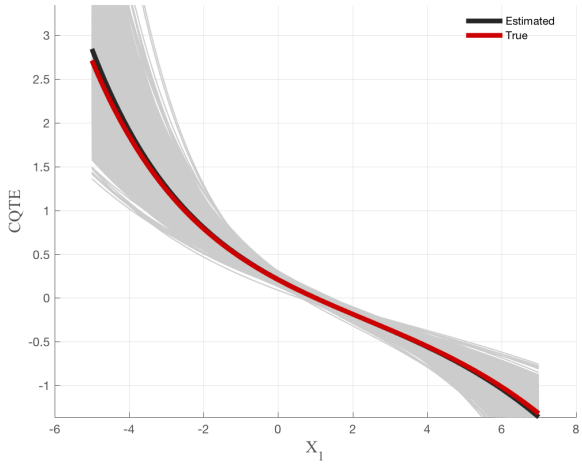


Notes: The empirical distribution is obtained from 1000 simulation draws of sample size 5000. However, the effective sample sizes are much smaller. Since the average propensity of treatment is a 0.41 and average propensity of being observed as 0.38, the average treated sample is $N_1 = 5000 \times 0.41 \times 0.38 = 779$ and average control sample is $N_0 = 5000 \times (1 - 0.41) \times 0.38 = 1,121$. The true ATE $= 0.096$. The graphs display the empirical distribution of the estimated ATE with misspecified mean model under two different cases of misspecification of the probability models. For the other two, see the main text. In each of these graphs we can the doubly weighted estimator is consistent for the true ATE whereas the unweighted and PS-weighted are away from the truth.

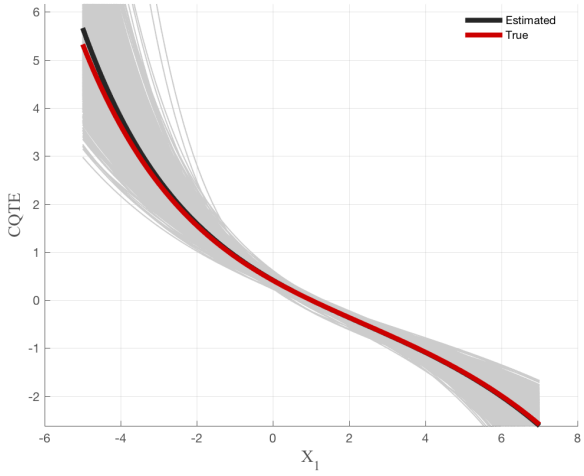
Figure H.3: Estimated CQTE with true CQTE as a function of x_1, N=5000

D) Both probability models are misspecified

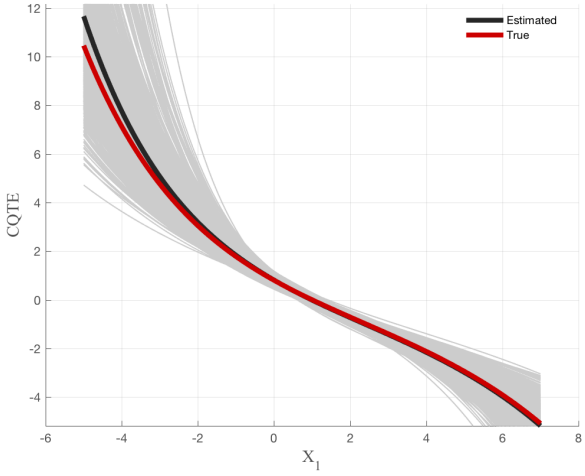
a) $\tau = 0.25$



b) $\tau = 0.50$



c) $\tau = 0.75$

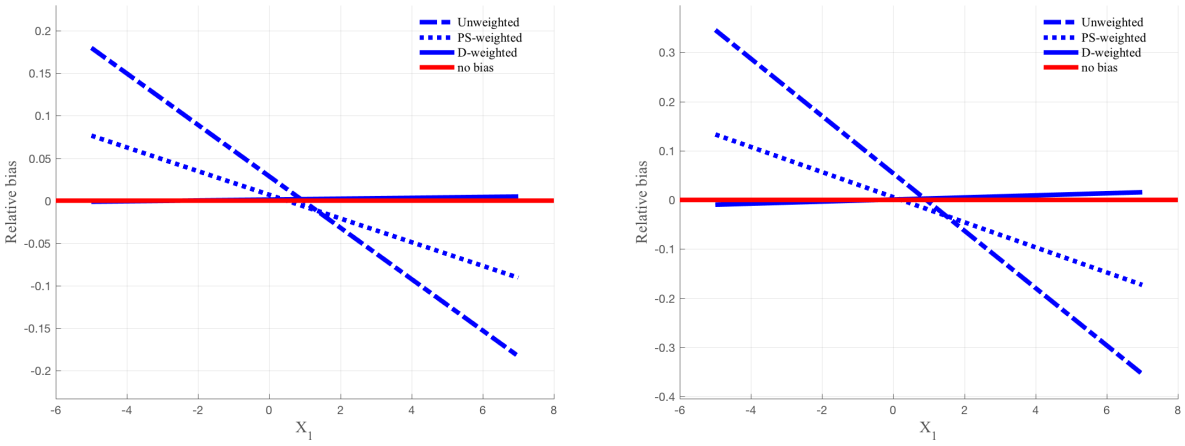


Notes: This figure plots the estimated CQTE along with the true CQTE as a function of x_1. The figure corresponds to the scenario when the conditional quantile functions for the treated and control groups are correctly specified but the two probability models are misspecified. Along with these two graphs, the figure also plots the function across the 1000 simulations (reps). The other three cases for when the propensity score or the missing data probability is allowed to be misspecified are not considered since under correct CQF specification, all these graphs look identical. For, $N = 5000$, the average treated sample is $N_1 = 5000 \times 0.41 \times 0.38 = 779$ and average control sample is $N_0 = 5000 \times (1 - 0.41) \times 0.38 = 1,121$.

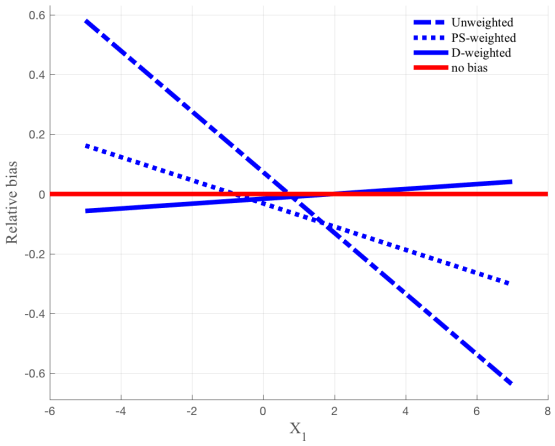
Figure H.4: Bias in estimated linear projection relative to true linear projection as a function of x_1 using Angrist et al. (2006b) methodology, N=5000

A) Both probability models are correct

a) $\tau = 0.25$ b) $\tau = 0.50$



c) $\tau = 0.75$

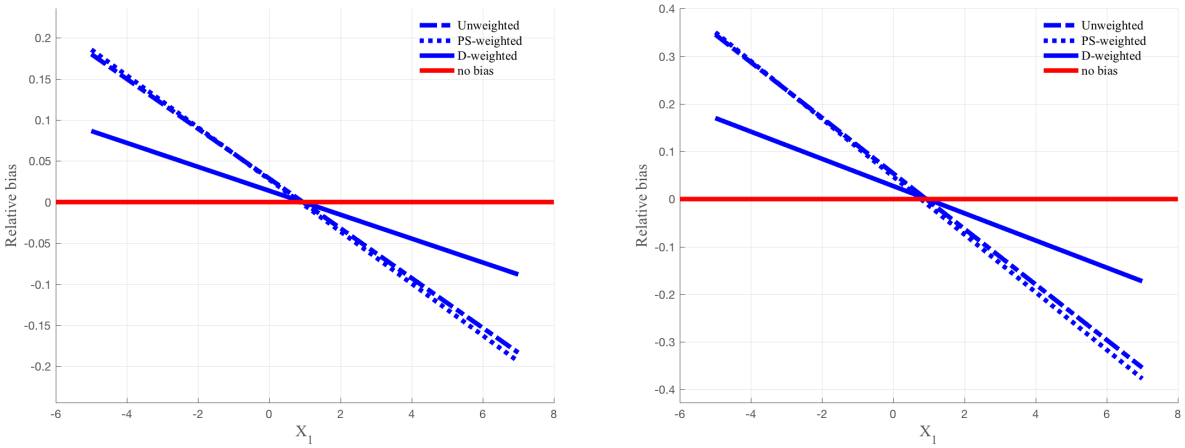


Notes: Angrist et al. (2006b) show that under the case of misspecification of the true CQF, the check function can still estimate a weighted linear projection to the true CQF. Since this particular case corresponds to misspecification of the CQF (where I estimate it to be linear), the solutions to problem 3.36 will consistently estimate the LP's to the two CQFs under the problems of non-random assignment and missing outcomes. Therefore, one can characterize an LP to the true CQTE using these two objects. This figure plots the bias in the doubly-weighted, PS-weighted and unweighted linear projection of the true CQTE relative to the true population LP of CQTE. For, $N = 5000$, the average treated sample is $N_1 = 5000 \times 0.41 \times 0.38 = 779$ and average control sample is $N_0 = 5000 \times (1 - 0.41) \times 0.38 = 1,121$. For a description of how these functions were estimated, see the simulation appendix.

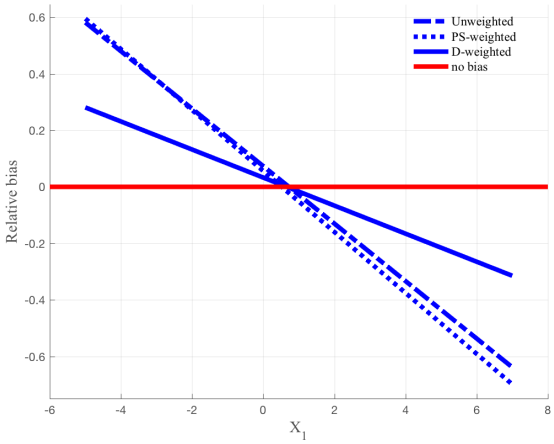
Figure H.4 (cont'd)

D) Both probability models are misspecified

a) $\tau = 0.25$ b) $\tau = 0.50$



c) $\tau = 0.75$

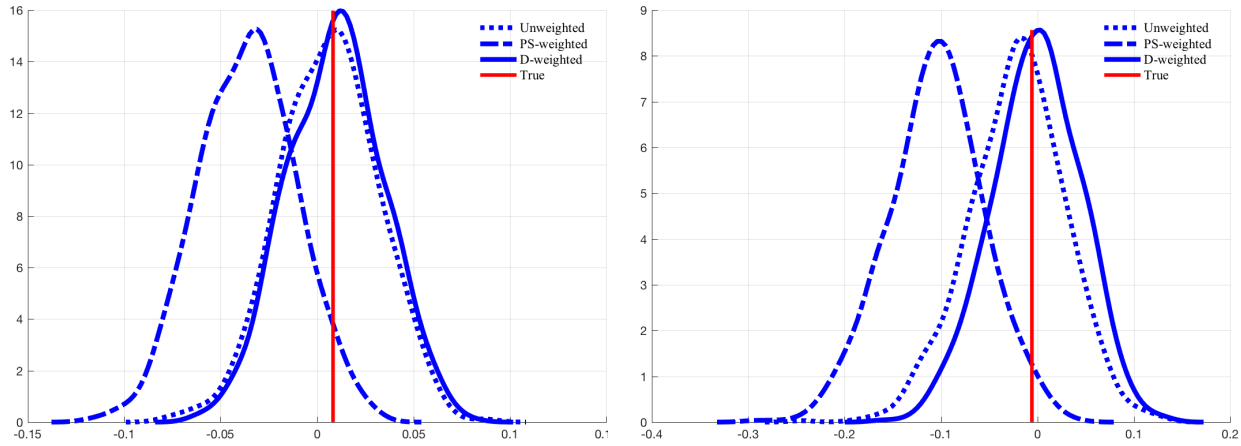


Notes: Angrist et al. (2006b) show that under the case of misspecification of the true CQF, the check function can still estimate a weighted linear projection to the true CQF. Since this particular case corresponds to misspecification of all three components of the doubly weighted framework, the solutions to problem 3.36 will not consistently estimate the LP's to the two CQFs under the problems of non-random assignment and missing outcomes. This figure plots the bias in the doubly-weighted, PS-weighted and unweighted linear projection of the true CQTE relative to the true population LP of CQTE. For, $N = 5000$, the average treated sample is $N_1 = 5000 \times 0.41 \times 0.38 = 779$ and average control sample is $N_0 = 5000 \times (1 - 0.41) \times 0.38 = 1,121$. For a description of how these functions were estimated, see the simulation appendix. The unweighted estimator does not weight the observed data by anything. The PS-weighted estimator weights to correct only for non-random assignment and the doubly weighted estimator weights by both the treatment and missing outcomes propensity score models to deal with non-random assignment and missing outcome problems.

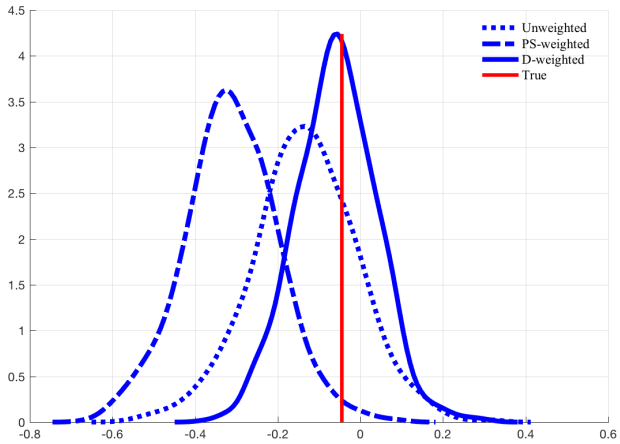
Figure H.5: Empirical distribution of estimated UQTE for N=5000

A) Both probability models are correct

a) $\tau = 0.25$ b) $\tau = 0.50$



c) $\tau = 0.75$

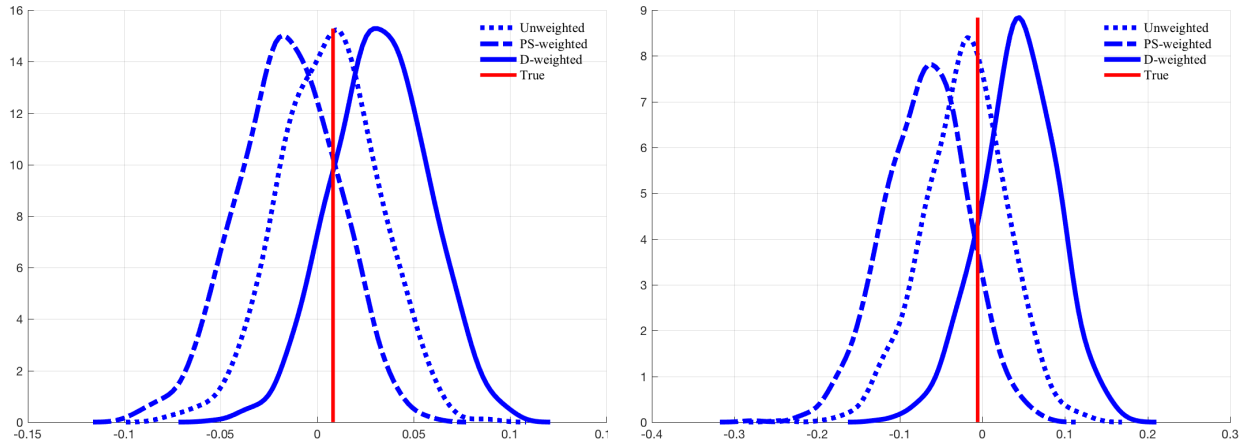


Notes: The empirical distribution is obtained from 1000 simulation draws for $N = 5000$. Since the average propensity of treatment is 0.41 and average propensity of being observed is 0.38, this implies the average treated sample is $N_1 = 5000 \times 0.41 \times 0.38 = 779$ and average control sample is $N_0 = 5000 \times (1-0.41) \times 0.38 = 1,121$. The unweighted estimator does not weight the observed data by anything. The PS-weighted estimator weights to correct only for non-random assignment and the doubly weighted estimator weights by both the treatment and missing outcomes propensity score models to deal with non-random assignment and missing outcome problems.

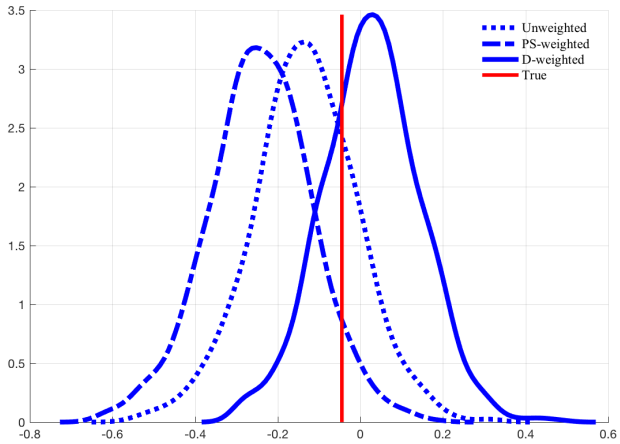
Figure H.5 (cont'd)

B) Correct missing outcomes probability but misspecified propensity score model

a) $\tau = 0.25$ b) $\tau = 0.50$



c) $\tau = 0.75$

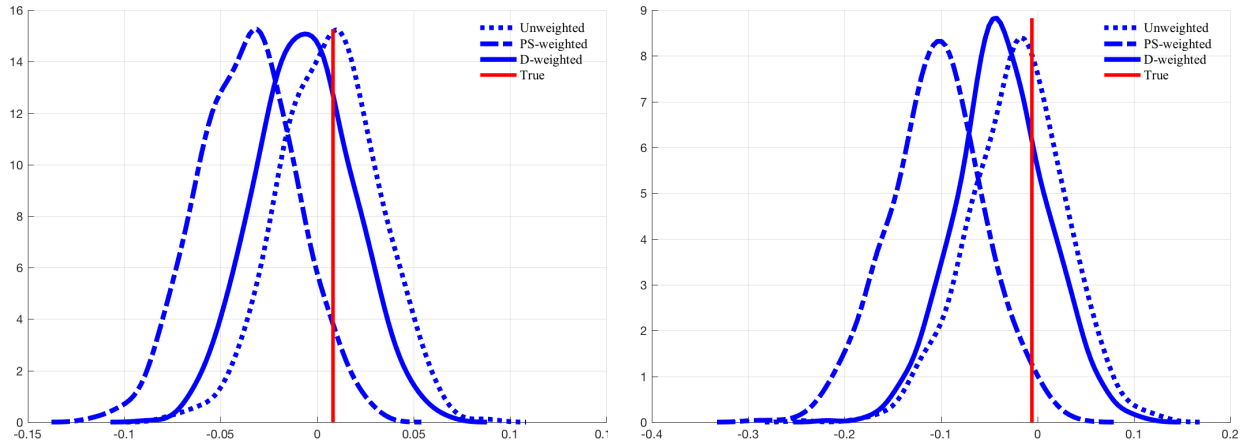


Notes: The empirical distribution is obtained from 1000 simulation draws for $N = 5000$. Since the average propensity of treatment is 0.41 and average propensity of being observed is 0.38, this implies the average treated sample is $N_1 = 5000 \times 0.41 \times 0.38 = 779$ and average control sample is $N_0 = 5000 \times (1 - 0.41) \times 0.38 = 1,121$. The unweighted estimator does not weight the observed data by anything. The PS-weighted estimator weights to correct only for non-random assignment and the doubly weighted estimator weights by both the treatment and missing outcomes propensity score models to deal with non-random assignment and missing outcome problems.

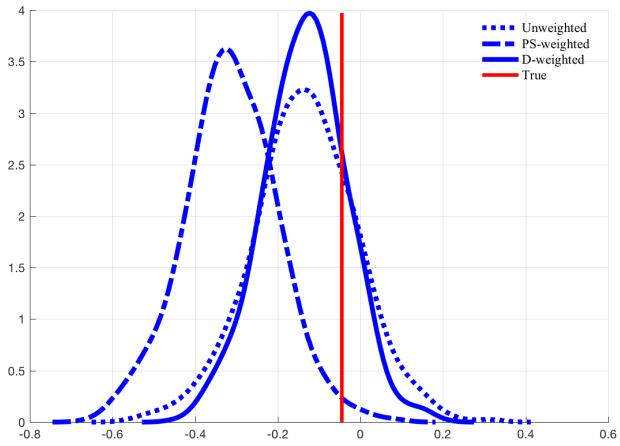
Figure H.5 (cont'd)

C) Misspecified missing outcomes probability but correct propensity score model

a) $\tau = 0.25$ b) $\tau = 0.50$



c) $\tau = 0.75$

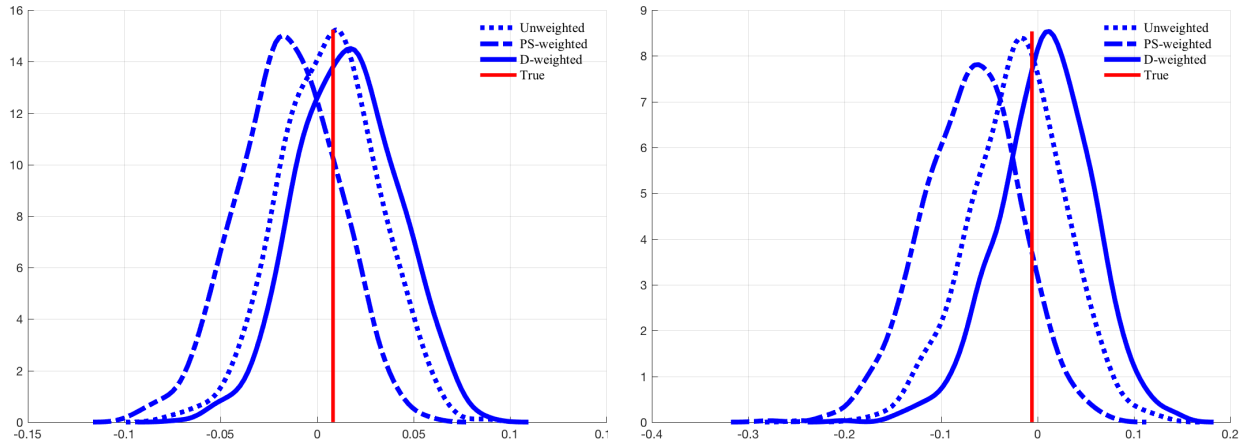


Notes: The empirical distribution is obtained from 1000 simulation draws for $N = 5000$. Since the average propensity of treatment is 0.41 and average propensity of being observed is 0.38, this implies the average treated sample is $N_1 = 5000 \times 0.41 \times 0.38 = 779$ and average control sample is $N_0 = 5000 \times (1 - 0.41) \times 0.38 = 1,121$. The unweighted estimator does not weight the observed data by anything. The PS-weighted estimator weights to correct only for non-random assignment and the doubly weighted estimator weights by both the treatment and missing outcomes propensity score models to deal with non-random assignment and missing outcome problems.

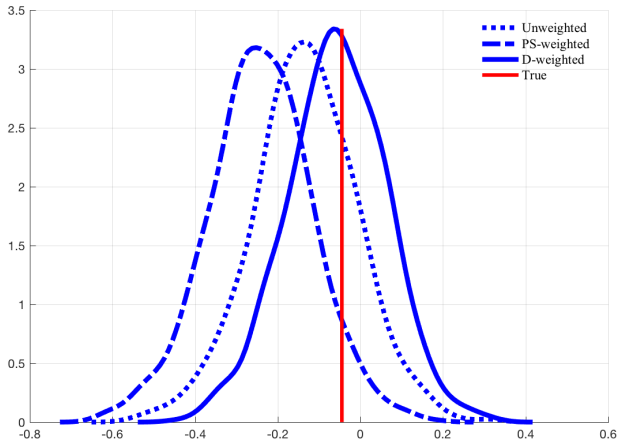
Figure H.5 (cont'd)

D) Both probability models are misspecified

a) $\tau = 0.25$ b) $\tau = 0.50$



c) $\tau = 0.75$



Notes: The empirical distribution is obtained from 1000 simulation draws for $N = 5000$. Since the average propensity of treatment is 0.41 and average propensity of being observed is 0.38, the average treated sample is $N_1 = 5000 \times 0.41 \times 0.38 = 779$ and average control sample is $N_0 = 5000 \times (1 - 0.41) \times 0.38 = 1,121$. The unweighted estimator does not weight the observed data by anything. The PS-weighted estimator weights to correct only for non-random assignment and the doubly weighted estimator weights by both the treatment and missing outcomes propensity score models to deal with non-random assignment and missing outcome problems.

APPENDIX I

TABLES FOR CHAPTER 3

Table I.1: An illustration of the observed sample (✓means observed, ? means missing)

i	y	x	w	s
1	?	✓	1	0
2	y(1)	✓	0	1
3	y(0)	✓	1	1
4	?	✓	1	0
⋮	⋮	⋮	⋮	⋮
N	y(1)	✓	0	1

Table I.2: Different scenarios under ignorability and unconfoundedness

Situation	$\mathbb{D}(y(g)\|\mathbf{x})$	$\mathbb{P}(s=1\|y(g),\mathbf{x},w_g)=\mathbb{P}(s=1\|\mathbf{x},w_g)$?	$\mathbb{P}(s=1\|\mathbf{x},w_g)$ correct?	$\mathbb{P}(w_g=1\|y(g),\mathbf{x})=\mathbb{P}(w=1\|\mathbf{x})$?	$\mathbb{P}(w=1\|\mathbf{x})$ correct?	Unweigted for $\mathbb{D}(y(g)\|\mathbf{x})$?	Weighted for $\mathbb{D}(y(g)\|\mathbf{x})$?	Weighted for 2.1?
1	Correctly specified	No	Either	Either	Either	No	No	No
2	Correctly specified	Yes	Either	Yes	Either	Yes	Yes	Yes
3	Correctly specified	Yes	Either	No	Either	No	No	No
5	Misspecified	No	Either	Either	Either	No	No	No
6	Misspecified	Yes	Either	Yes	No	No	No	No
7	Misspecified	Yes	Yes	Yes	Yes	No	No	Yes
8	Misspecified	Yes	Yes	No	Either	No	No	No
9	Misspecified	Yes	No	Either	Either	No	No	No

[a.] Notice that if the missingness mechanism is not ignorable or for that matter the assignment mechanism is not unconfounded, then nothing can be consistently estimated whether or not other components of the framework are correctly specified. This can be seen in cases (1) and (3) in the table above. Situations (2) and (7) together forms what is called robust estimation that has been described in the sections above. Remember that under unconfoundedness and ignorability, $\mathbb{D}(y(g)|\mathbf{x})$ is the same as $\mathbb{D}(y(g)|\mathbf{x}, w_g, s)$.

Table I.3: Different scenarios under exogeneity of missingness and unconfoundedness

Situation	$\mathbb{D}(\boldsymbol{y}(\boldsymbol{g})\|\mathbf{x})$	$\mathbb{P}(\boldsymbol{s}=1\|\boldsymbol{y}(\boldsymbol{g}),\mathbf{x},w_g)=\mathbb{P}(\boldsymbol{s}=1\|\mathbf{x})$?	$\mathbb{P}(\boldsymbol{s}=1\|\mathbf{x})$ correct?	$\mathbb{P}(w_g=1\|\boldsymbol{y}(\boldsymbol{g}),\mathbf{x})=\mathbb{P}(w_g=1\|\mathbf{x})$?	$\mathbb{P}(w_g=1\|\mathbf{x})$ correct?	Unweigted for $\mathbb{D}(\boldsymbol{y}(\boldsymbol{g})\|\mathbf{x})$?	Weighted for $\mathbb{D}(\boldsymbol{y}(\boldsymbol{g})\|\mathbf{x})$?	Weighted for 2.1?
1	Correctly specified	Yes	Either	Yes	Either	Yes	Yes	Yes
2	Correctly specified	No	Either	Either	Either	No	No	No
3	Correctly specified	Yes	Either	No	Either	No	No	No
4	Misspecified	Yes	Yes	Yes	Yes	No	No	Yes
5	Misspecified	Yes	No	Yes	No	No	No	No
6	Misspecified	Yes	No	Yes	Yes	No	No	No
7	Misspecified	Yes	Yes	Yes	No	No	No	No
8	Misspecified	No	Either	Either	Either	No	No	No
9	Misspecified	Yes	Yes	No	Either	No	No	No

[a.] Situations (1) and (4) combine to give you the double robustness result which says that either the conditional feature of interest needs to be correctly specified or the treatment and missing probabilities both need to be correctly specified. Again, just like the previous case, if the missingness mechanism is not exogenous or if the assignment mechanism is not unconfounded, then even correctly specifying either of these features will not consistently estimate the parameter of interest. This is illustrated in cases (2) and (3). If one looks at case (2) in the tabel above, under both these situations the unweighted estimator works to deliver a consistent estimator of $\boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}}$. In such a scenario where both the unweighted and weighted estimators are consistent, how can we choose amongst them? The following table enumerates situations where not weighting is better than weighting.

Table I.4: When is unweighted more efficient than weighted assuming ignorability and unconfoundedness and $\mathbb{D}(y(g)|\mathbf{x})$ correctly specified?

Situation	$\mathbb{P}(s=1\|\mathrm{x}, w_g)$ correct?	$\mathbb{P}(s=1\|y(g), \mathrm{x}, w_g) = \mathbb{P}(s=1\|\mathrm{x})$?	$\mathbb{P}(s=1\|\mathrm{x})$ correct?	$\mathbb{P}(w_g=1\|x)$ correct?	GCIME holds?	Unweighted more efficient	Weighted with estimated probabilities more efficient?
1	Either	No	Doesn't apply	Either	Yes	Yes	No
2	Either	Yes	Either	Either	Yes	Yes	No
3	Either	Either	Either	Either	No	Can't say	Can't say

I.0.1 Bias and root-mean squared error for ATE simulations

Table I.5: When the conditional mean model is correctly specified

A) Both probability models are correct

	N=1000				N=5000			
Estimator	**Unweighted**	**PS-weighted**	**D-weighted**		**Unweighted**	**PS-weighted**	**D-weighted**	
			Estimated	**Known**			**Estimated**	**Known**
BIAS	-0.00082	-0.00067	-0.00065	-0.00066	-0.00039	-0.00037	-0.00034	-0.00034
RMSE	0.02372	0.02370	0.02374	0.02376	0.01074	0.01074	0.01075	0.01075

Notes: The unweighted estimator does not weight the observed data. The PS-weighted estimator weights to correct only for non-random assignment and the doubly weighted estimator weights by both the propensity score and the missingness probabilities to deal with the assignment and missing data problems. The two columns under the doubly weighted estimator report the Bias and RMSE of the estimators that use estimated and known probability weights respectively. The efficiency results in section 3.5 dictate no asymptotic efficiency gains in the case when we have the conditional model correctly specified. However, in finite samples, one could obtain smaller or larger variance estimates. For, $N = 1000$, the average treated sample is $N_1 = 1000 \times 0.41 \times 0.38 = 156$ and average control sample is $N_0 = 1000 \times (1 - 0.41) \times 0.38 = 224$. For, $N = 5000$, $N_1 = 5000 \times 0.41 \times 0.38 = 779$ and $N_0 = 5000 \times (1 - 0.41) \times 0.38 = 1,121$. The Bias and RMSE are reported across 1000 simulations.

B) Correct missingness model but misspecified propensity score model

	N=1000			N=5000		
Estimator	**Unweighted**	**PS-weighted**	**D-weighted**	**Unweighted**	**PS-weighted**	**D-weighted**
BIAS	-0.00082	-0.00069	-0.00081	-0.00039	-0.00035	-0.00040
RMSE	0.02372	0.02369	0.02376	0.01074	0.01074	0.01076

Notes: The unweighted estimator does not weight the observed data. The PS-weighted estimator weights to correct only for non-random assignment and the doubly weighted estimator weights by both the treatment and missing outcomes propensity score models to deal with non-random assignment and missing outcomes. For, $N = 1000$, the average treated sample is $N_1 = 1000 \times 0.41 \times 0.38 = 156$ and average control sample is $N_0 = 1000 \times (1-0.41) \times 0.38 = 224$. For, $N = 5000$, $N_1 = 5000 \times 0.41 \times 0.38 = 779$ and $N_0 = 5000 \times (1 - 0.41) \times 0.38 = 1,121$. The Bias and RMSE are reported across 1000 simulations.

Table I.5 (cont'd)

C) Misspecified missingness model but correct propensity score model

	N=1000			N=5000		
Estimator	**Unweighted**	**PS-weighted**	**D-weighted**	**Unweighted**	**PS-weighted**	**D-weighted**
BIAS	-0.00082	-0.00067	-0.00067	-0.00039	-0.00037	-0.00035
RMSE	0.02372	0.02370	0.02372	0.01074	0.01074	0.01075

Notes: The unweighted estimator does not weight the observed data by anything. The PS-weighted estimator weights to correct only for non-random assignment and the doubly weighted estimator weights by both the treatment and missing outcomes propensity score models to deal with non-random assignment and missing outcome problems. For, $N = 1000$, the average treated sample is $N_1 = 1000 \times 0.41 \times 0.38 = 156$ and average control sample is $N_0 = 1000 \times (1-0.41) \times 0.38 = 224$. For, $N = 5000$, $N_1 = 5000 \times 0.41 \times 0.38 = 779$ and $N_0 = 5000 \times (1 - 0.41) \times 0.38 = 1,121$. The Bias and Rmse are reported across 1000 simulations.

D) Both probability models are misspecified

	N=1000			N=5000		
Estimator	**Unweighted**	**PS-weighted**	**D-weighted**	**Unweighted**	**PS-weighted**	**D-weighted**
BIAS	-0.00082	-0.00069	-0.00080	-0.00039	-0.00035	-0.00039
RMSE	0.02372	0.02369	0.02373	0.01074	0.01074	0.01075

Notes: The unweighted estimator does not weight the observed data. The PS-weighted estimator weights to correct only for non-random assignment and the doubly weighted estimator weights by both the treatment and missing outcomes propensity scores to deal with non-random assignment and missing outcomes. For, $N = 1000$, the average treated sample size is $N_1 = 1000 \times 0.41 \times 0.38 = 156$ and average control sample size is $N_0 = 1000 \times (1 - 0.41) \times 0.38 = 224$. For, $N = 5000$, $N_1 = 5000 \times 0.41 \times 0.38 = 779$ and $N_0 = 5000 \times (1 - 0.41) \times 0.38 = 1,121$. The Bias and Rmse are reported across 1000 Monte Carlo repetitions.

Table I.6: Misspecified conditional mean model

A) Both probability models are correct

	N=1000				N=5000			
Estimator	**Unweighted**	**PS-weighted**	**D-weighted**		**Unweighted**	**PS-weighted**	**D-weighted**	
			estimated	**known**			**estimated**	**known**
BIAS	0.01087	0.00008	-0.00002	0.00003	0.01052	-0.00058	-0.00064	-0.00064
RMSE	0.03250	0.03038	0.02979	0.02986	0.01744	0.01396	0.01376	0.01375

Notes: The unweighted estimator does not weight the observed data. The PS-weighted estimator weights to correct only for non-random assignment and the doubly weighted estimator weights by both the treatment and missing outcomes propensity scores to deal with non-random assignment and missing outcomes. The two columns under the doubly weighted estimator report the Bias and Rmse of the estimators that use estimated and known probability weights respectively. For, $N = 1000$, the average treated sample size is $N_1 = 1000 \times 0.41 \times 0.38 = 156$ and average control sample size is $N_0 = 1000 \times (1 - 0.41) \times 0.38 = 224$. For, $N = 5000$, $N_1 = 5000 \times 0.41 \times 0.38 = 779$ and $N_0 = 5000 \times (1 - 0.41) \times 0.38 = 1,121$. The Bias and Rmse are reported across 1000 Monte Carlo repetitions.

B) Correct missingness model but misspecified propensity score model

	N=1000			N=5000		
Estimator	**Unweighted**	**PS-weighted**	**D-weighted**	**Unweighted**	**PS-weighted**	**D-weighted**
BIAS	0.01087	0.00699	0.00102	0.01052	0.00651	0.00049
RMSE	0.03250	0.03117	0.02984	0.01744	0.01532	0.01378

Notes: The unweighted estimator does not weight the observed data by anything. The PS-weighted estimator weights to correct only for non-random assignment and the doubly weighted estimator weights by both the treatment and missing outcomes propensity score models to deal with non-random assignment and missing outcome problems. For, $N = 1000$, the average treated sample is $N_1 = 1000 \times 0.41 \times 0.38 = 156$ and average control sample is $N_0 = 1000 \times (1 - 0.41) \times 0.38 = 224$. For, $N = 5000$, $N_1 = 5000 \times 0.41 \times 0.38 = 779$ and $N_0 = 5000 \times (1 - 0.41) \times 0.38 = 1,121$. The Bias and Rmse are reported across 1000 simulations.

Table I.6 (cont'd)

C) Misspecified missingness model but correct propensity score model

	N=1000			N=5000		
Estimator	**Unweighted**	**PS-weighted**	**D-weighted**	**Unweighted**	**PS-weighted**	**D-weighted**
BIAS	0.01087	0.00008	0.00001	0.01052	-0.00058	-0.00063
RMSE	0.03250	0.03038	0.02970	0.01744	0.01396	0.01371

Notes: The unweighted estimator does not weight the observed data by anything. The PS-weighted estimator weights to correct only for non-random assignment and the doubly weighted estimator weights by both the treatment and missing outcomes propensity score models to deal with non-random assignment and missing outcome problems. For, $N = 1000$, the average treated sample is $N_1 = 1000 \times 0.41 \times 0.38 = 156$ and average control sample is $N_0 = 1000 \times (1 - 0.41) \times 0.38 = 224$. For, $N = 5000$, $N_1 = 5000 \times 0.41 \times 0.38 = 779$ and $N_0 = 5000 \times (1 - 0.41) \times 0.38 = 1,121$. The Bias and Rmse are reported across 1000 simulations.

D) Both probability models are misspecified

	N=1000			N=5000		
Estimator	**Unweighted**	**PS-weighted**	**D-weighted**	**Unweighted**	**PS-weighted**	**D-weighted**
BIAS	0.01087	0.00699	-0.00093	0.01052	0.00651	-0.00150
RMSE	0.03250	0.03117	0.02992	0.01744	0.01532	0.01380

Notes: The unweighted estimator does not weight the observed data. The PS-weighted estimator weights to correct only for non-random assignment and the doubly weighted estimator weights by both the treatment and missing outcomes propensity scores to deal with non-random assignment and missing outcomes. For, $N = 1000$, the average treated sample size is $N_1 = 1000 \times 0.41 \times 0.38 = 156$ and average control sample size is $N_0 = 1000 \times (1 - 0.41) \times 0.38 = 224$. For, $N = 5000$, $N_1 = 5000 \times 0.41 \times 0.38 = 779$ and $N_0 = 5000 \times (1 - 0.41) \times 0.38 = 1,121$. The Bias and Rmse are reported across 1000 Monte Carlo repetitions.

I.0.2 Bias and root-mean squared error for UQTE simulations

Table I.7: A) Both probability models are correct

For $\tau = 0.25$ (25th quantile)

	N=1000				N=5000			
Estimator	**Unweighted**	**PS-weighted**	**D-weighted**		**Unweighted**	**PS-weighted**	**D-weighted**	
			Estimated	**Known**			**Estimated**	**Known**
BIAS	-0.0014	-0.0424	0.0046	0.0038	-0.0022	-0.0446	0.0012	0.0012
RMSE	0.0554	0.0690	0.0532	0.0549	0.0254	0.0512	0.0247	0.0255

Notes: The unweighted estimator does not weight the observed data. The ps-weighted estimator weights to correct only for non-random assignment and the doubly weighted estimator weights by both the propensity score model and the missingness model to correct for non-random assignment and missing outcomes. For, $N = 1000$, the average treated sample is $N_1 = 1000 \times 0.41 \times 0.38 = 156$ and average control sample is $N_0 = 1000 \times (1-0.41) \times 0.38 = 224$. For, $N = 5000$, $N_1 = 5000 \times 0.41 \times 0.38 = 779$ and $N_0 = 5000 \times (1 - 0.41) \times 0.38 = 1,121$. The Bias and Rmse are reported across 1000 Monte Carlo repetitions.

For $\tau = 0.50$ (50th quantile)

	N=1000				N=5000			
Estimator	**Unweighted**	**PS-weighted**	**D-weighted**		**Unweighted**	**PS-weighted**	**D-weighted**	
			Estimated	**Known**			**Estimated**	**Known**
BIAS	-0.0206	-0.1072	0.0000	0.0007	-0.0157	-0.0998	0.0043	0.0044
RMSE	0.1181	0.1543	0.1028	0.1068	0.0522	0.1114	0.0462	0.0483

Notes: The unweighted estimator does not weight the observed data by anything. The PS-weighted estimator weights to correct only for non-random assignment and the doubly weighted estimator weights by both the treatment and missing outcomes propensity score models to deal with non-random assignment and missing outcome problems. For, $N = 1000$, the average treated sample is $N_1 = 1000 \times 0.41 \times 0.38 = 156$ and average control sample is $N_0 = 1000 \times (1 - 0.41) \times 0.38 = 224$. For, $N = 5000$, $N_1 = 5000 \times 0.41 \times 0.38 = 779$ and $N_0 = 5000 \times (1 - 0.41) \times 0.38 = 1,121$. The Bias and Rmse are reported across 1000 simulations.

Table I.7 (cont'd)

For $\tau = 0.75$ (75th quantile)

	N=1000				N=5000			
Estimator	**Unweighted**	**PS-weighted**	**D-weighted**		**Unweighted**	**PS-weighted**	**D-weighted**	
			Estimated	**Known**			**Estimated**	**Known**
BIAS	-0.0899	-0.2742	-0.0210	-0.0217	-0.0896	-0.2687	-0.0145	-0.0147
RMSE	0.3097	0.3803	0.2399	0.2523	0.1550	0.2917	0.0983	0.1036

Notes: The unweighted estimator does not weight the observed data by anything. The PS-weighted estimator weights to correct only for non-random assignment and the doubly weighted estimator weights by both the treatment and missing outcomes propensity score models to deal with non-random assignment and missing outcome problems. For, $N = 1000$, the average treated sample is $N_1 = 1000 \times 0.41 \times 0.38 = 156$ and average control sample is $N_0 = 1000 \times (1 - 0.41) \times 0.38 = 224$. For, $N = 5000$, $N_1 = 5000 \times 0.41 \times 0.38 = 779$ and $N_0 = 5000 \times (1 - 0.41) \times 0.38 = 1,121$. The Bias and Rmse are reported across 1000 simulations.

Table I.8: B) When missing data probability is misspecified and propensity score is correct

For $\tau = 0.25$ (25th quantile)

	N=1000			N=5000		
Estimator	**Unweighted**	**PS-weighted**	**D-weighted**	**Unweighted**	**PS-weighted**	**D-weighted**
BIAS	-0.0014	-0.0424	-0.0116	-0.0022	-0.0446	-0.0150
RMSE	0.0554	0.0690	0.0557	0.0254	0.0512	0.0291

Notes: The unweighted estimator does not weight the observed data by anything. The PS-weighted estimator weights to correct only for non-random assignment and the doubly weighted estimator weights by both the treatment and missing outcomes propensity score models to deal with non-random assignment and missing outcome problems. For, $N = 1000$, the average treated sample is $N_1 = 1000 \times 0.41 \times 0.38 = 156$ and average control sample is $N_0 = 1000 \times (1 - 0.41) \times 0.38 = 224$. For, $N = 5000$, $N_1 = 5000 \times 0.41 \times 0.38 = 779$ and $N_0 = 5000 \times (1 - 0.41) \times 0.38 = 1,121$. The Bias and Rmse are reported across 1000 simulations.

Table I.8 (cont'd)

For $\tau = 0.50$ (50th quantile)

	N=1000			**N=5000**		
Estimator	**Unweighted**	**PS-weighted**	**D-weighted**	**Unweighted**	**PS-weighted**	**D-weighted**
BIAS	-0.0206	-0.1072	-0.0355	-0.0157	-0.0998	-0.0319
RMSE	0.1181	0.1543	0.1119	0.0522	0.1114	0.0571

Notes: The unweighted estimator does not weight the observed data by anything. The PS-weighted estimator weights to correct only for non-random assignment and the doubly weighted estimator weights by both the treatment and missing outcomes propensity score models to deal with non-random assignment and missing outcome problems. For, $N = 1000$, the average treated sample is $N_1 = 1000 \times 0.41 \times 0.38 = 156$ and average control sample is $N_0 = 1000 \times (1 - 0.41) \times 0.38 = 224$. For, $N = 5000$, $N_1 = 5000 \times 0.41 \times 0.38 = 779$ and $N_0 = 5000 \times (1 - 0.41) \times 0.38 = 1,121$. The Bias and Rmse are reported across 1000 simulations.

For $\tau = 0.75$ (75th quantile)

	N=1000			**N=5000**		
Estimator	**Unweighted**	**PS-weighted**	**D-weighted**	**Unweighted**	**PS-weighted**	**D-weighted**
BIAS	-0.0899	-0.2742	-0.0989	-0.0896	-0.2687	-0.0896
RMSE	0.3097	0.3803	0.2648	0.1550	0.2917	0.1348

Notes: The unweighted estimator does not weight the observed data by anything. The PS-weighted estimator weights to correct only for non-random assignment and the doubly weighted estimator weights by both the treatment and missing outcomes propensity score models to deal with non-random assignment and missing outcome problems. For, $N = 1000$, the average treated sample is $N_1 = 1000 \times 0.41 \times 0.38 = 156$ and average control sample is $N_0 = 1000 \times (1 - 0.41) \times 0.38 = 224$. For, $N = 5000$, $N_1 = 5000 \times 0.41 \times 0.38 = 779$ and $N_0 = 5000 \times (1 - 0.41) \times 0.38 = 1,121$. The Bias and Rmse are reported across 1000 simulations.

Table I.9: C) When missing data probability is correct and propensity score is misspecified

For $\tau = 0.25$ (25th quantile)

	N=1000			**N=5000**		
Estimator	**Unweighted**	**PS-weighted**	**D-weighted**	**Unweighted**	**PS-weighted**	**D-weighted**
BIAS	-0.0014	-0.0227	0.0265	-0.0022	-0.0239	0.0243
RMSE	0.0554	0.0598	0.0614	0.0254	0.0352	0.0348

Notes: The unweighted estimator does not weight the observed data by anything. The PS-weighted estimator weights to correct only for non-random assignment and the doubly weighted estimator weights by both the treatment and missing outcomes propensity score models to deal with non-random assignment and missing outcome problems. For, $N = 1000$, the average treated sample is $N_1 = 1000 \times 0.41 \times 0.38 = 156$ and average control sample is $N_0 = 1000 \times (1 - 0.41) \times 0.38 = 224$. For, $N = 5000$, $N_1 = 5000 \times 0.41 \times 0.38 = 779$ and $N_0 = 5000 \times (1 - 0.41) \times 0.38 = 1,121$. The Bias and Rmse are reported across 1000 simulations.

For $\tau = 0.50$ (50th quantile)

	N=1000			**N=5000**		
Estimator	**Unweighted**	**PS-weighted**	**D-weighted**	**Unweighted**	**PS-weighted**	**D-weighted**
BIAS	-0.0206	-0.0681	0.0456	-0.0157	-0.0637	0.0488
RMSE	0.1181	0.1349	0.1168	0.0522	0.0809	0.0673

Notes: The unweighted estimator does not weight the observed data by anything. The PS-weighted estimator weights to correct only for non-random assignment and the doubly weighted estimator weights by both the treatment and missing outcomes propensity score models to deal with non-random assignment and missing outcome problems. For, $N = 1000$, the average treated sample is $N_1 = 1000 \times 0.41 \times 0.38 = 156$ and average control sample is $N_0 = 1000 \times (1 - 0.41) \times 0.38 = 224$. For, $N = 5000$, $N_1 = 5000 \times 0.41 \times 0.38 = 779$ and $N_0 = 5000 \times (1 - 0.41) \times 0.38 = 1,121$. The Bias and Rmse are reported across 1000 simulations.

Table I.9 (cont'd)

For $\tau = 0.75$ (75th quantile)

	N=1000			N=5000		
Estimator	**Unweighted**	**PS-weighted**	**D-weighted**	**Unweighted**	**PS-weighted**	**D-weighted**
BIAS	-0.0899	-0.2005	0.0691	-0.0896	-0.1978	0.0709
RMSE	0.3097	0.3611	0.2894	0.1550	0.2346	0.1377

Notes: The unweighted estimator does not weight the observed data by anything. The PS-weighted estimator weights to correct only for non-random assignment and the doubly weighted estimator weights by both the treatment and missing outcomes propensity score models to deal with non-random assignment and missing outcome problems. For, $N = 1000$, the average treated sample is $N_1 = 1000 \times 0.41 \times 0.38 = 156$ and average control sample is $N_0 = 1000 \times (1 - 0.41) \times 0.38 = 224$. For, $N = 5000$, $N_1 = 5000 \times 0.41 \times 0.38 = 779$ and $N_0 = 5000 \times (1 - 0.41) \times 0.38 = 1,121$. The Bias and Rmse are reported across 1000 simulations.

Table I.10: D) Both probability models are misspecified

For $\tau = 0.25$ (25th quantile)

	N=1000			N=5000		
Estimator	**Unweighted**	**PS-weighted**	**D-weighted**	**Unweighted**	**PS-weighted**	**D-weighted**
BIAS	-0.0014	-0.0227	0.0095	-0.0022	-0.0239	0.0074
RMSE	0.0554	0.0598	0.0571	0.0254	0.0352	0.0268

Notes: The unweighted estimator does not weight the observed data by anything. The PS-weighted estimator weights to correct only for non-random assignment and the doubly weighted estimator weights by both the treatment and missing outcomes propensity score models to correct for non-random assignment and missing outcome. For, $N = 1000$, the average treated sample is $N_1 = 1000 \times 0.41 \times 0.38 = 156$ and average control sample is $N_0 = 1000 \times (1 - 0.41) \times 0.38 = 224$. For, $N = 5000$, $N_1 = 5000 \times 0.41 \times 0.38 = 779$ and $N_0 = 5000 \times (1 - 0.41) \times 0.38 = 1,121$. The Bias and Rmse are reported across 1000 Monte Carlo repetitions.

Table I.10 (cont'd)

For $\tau = 0.50$ (50th quantile)

	N=1000			N=5000		
Estimator	**Unweighted**	**PS-weighted**	**D-weighted**	**Unweighted**	**PS-weighted**	**D-weighted**
BIAS	-0.0206	-0.0681	0.0070	-0.0157	-0.0637	0.0136
RMSE	0.1181	0.1349	0.1108	0.0522	0.0809	0.0503

Notes: The unweighted estimator does not weight the observed data by anything. The PS-weighted estimator weights to correct only for non-random assignment and the doubly weighted estimator weights by both the treatment and missing outcomes propensity score models to deal with non-random assignment and missing outcome problems. For, $N = 1000$, the average treated sample is $N_1 = 1000 \times 0.41 \times 0.38 = 156$ and average control sample is $N_0 = 1000 \times (1 - 0.41) \times 0.38 = 224$. For, $N = 5000$, $N_1 = 5000 \times 0.41 \times 0.38 = 779$ and average control sample size, $N_0 = 5000 \times (1 - 0.41) \times 0.38 = 1,121$. The Bias and Rmse are reported across 1000 simulations.

For $\tau = 0.75$ (75th quantile)

	N=1000			N=5000		
Estimator	**Unweighted**	**PS-weighted**	**D-weighted**	**Unweighted**	**PS-weighted**	**D-weighted**
BIAS	-0.0899	-0.2005	-0.0216	-0.0896	-0.1978	-0.0131
RMSE	0.3097	0.3611	0.2806	0.1550	0.2346	0.1202

Notes: The unweighted estimator does not weight the observed data by anything. The PS-weighted estimator weights to correct only for non-random assignment and the doubly weighted estimator weights by both the treatment and missing outcomes propensity score models to deal with non-random assignment and missing outcome problems. For, $N = 1000$, the average treated sample is $N_1 = 1000 \times 0.41 \times 0.38 = 156$ and average control sample is $N_0 = 1000 \times (1 - 0.41) \times 0.38 = 224$. For, $N = 5000$, $N_1 = 5000 \times 0.41 \times 0.38 = 779$ and average control sample size, $N_0 = 5000 \times (1 - 0.41) \times 0.38 = 1,121$. The Bias and Rmse are reported across 1000 simulations.

I.0.3 Calonico and Smith Application

Table I.11: Proportion of missing earnings in the experimental sample

Earnings in 1979	**Treated**	**Control**	**Total**
Missing	196	210	406
Observed	600	585	1185
Total	796	795	1591

Table I.12: Proportion of missing data in the PSID samples

Earnings in 1979	**PSID-1**	**PSID-2**
Missing	81	22
Observed	648	182
Total	729	204

Table I.13: Unweighted and weighted pre-training earnings comparisons using NSW and PSID comparison groups

Comparison group	Pre-training estimates					
	Unadjusted			Adjusted		
	Unweighted	PS-weighted	D-weighted	Unweighted	PS-weighted	D-weighted
NSW	-18	-9	1	-22	-10	-1
N=1,185	(123.45)	(51.07)	(48.76)	(124.70)	(51.34)	(48.97)
PSID-1	-2,534	-222	-255	-2,804	-199	-222
N=1,016	(283.95)	(213.57)	(205.59)	(281.49)	(212.55)	(205.45)
PSID-2	-2,080	-1,371	-1,357	-2,181	-1,505	-1,467
N=720	(411.23)	(331.41)	(317.41)	(427.24)	(359.98)	(342.16)
			Bias using NSW control			
PSID-1	-2,517	289	236	-2,760	334	287
N=1,001	(279.38)	(256.93)	(247.18)	(283.09)	(257.50)	(248.20)
PSID-2	-2,063	-1,249	-1,255	-2,144	-1,306	-1,297
N=705	(416.53)	(323.36)	(310.59)	(435.74)	(354.12)	(337.68)
Adjusted covariates						
Pre-training earnings (1975)				✓	✓	✓
Age				✓	✓	✓
Age2				✓	✓	✓
Education				✓	✓	✓
High school droput				✓	✓	✓
Black				✓	✓	✓
Hispanic				✓	✓	✓
Marital status				✓	✓	✓

Notes: This table reports unadjusted and adjusted pre-training earnings differences where the first row reports the experimental estimates. The second and third row reports non-experimental estimates computed using the PSID-1 and PSID-2 control groups respectively. The second panel of the table reports bias estimates computed from combining the NSW control and PSID-1 and PSID-2 comparison groups respectively. Both the pre-training estimates and the bias estimates should be compared to zero. Bootstrapped standard errors (in parentheses) have been constructed using 10,000 replications. All values are in 1982 dollars. The samples used for estimating the training and bias estimates using PSID-1 and PSID-2 comparison groups have been trimmed to ensure common support in the distribution of weights for the NSW-treatment and comparison groups. For more detail, see appendix G.

Table I.14: Estimation summary for ATE under different cases of misspecification

Scenario	CEF		$G(\cdot)$		$R(\cdot)$	
	Model	**Estimation**	**Model**	**Estimation**	**Model**	**Estimation**
1	C	$\Phi(\mathbf{x}\boldsymbol{\theta_g})$	C	$\Lambda(\mathbf{x}\boldsymbol{\gamma})$	C	$\Lambda(\mathbf{z}\boldsymbol{\gamma})$
2	C	$\Phi(\mathbf{x}\boldsymbol{\theta_g})$	C	$\Lambda(\mathbf{x}\boldsymbol{\gamma})$	M	$\Phi(\mathbf{z}^{(1)}\boldsymbol{\gamma}^{(1)})$
3	C	$\Phi(\mathbf{x}\boldsymbol{\theta_g})$	M	$\Phi(\mathbf{x}^{(1)}\boldsymbol{\gamma}^{(1)})$	C	$\Lambda(\mathbf{z}\boldsymbol{\gamma})$
4	C	$\Phi(\mathbf{x}\boldsymbol{\theta_g})$	M	$\Phi(\mathbf{x}^{(1)}\boldsymbol{\gamma}^{(1)})$	M	$\Phi(\mathbf{z}^{(1)}\boldsymbol{\gamma}^{(1)})$
5	M	$\mathbf{x}\boldsymbol{\theta_g}$	C	$\Lambda(\mathbf{x}\boldsymbol{\gamma})$	C	$\Lambda(\mathbf{z}\boldsymbol{\gamma})$
6	M	$\mathbf{x}\boldsymbol{\theta_g}$	C	$\Lambda(\mathbf{x}\boldsymbol{\gamma})$	M	$\Phi(\mathbf{z}^{(1)}\boldsymbol{\gamma}^{(1)})$
7	M	$\mathbf{x}\boldsymbol{\theta_g}$	M	$\Phi(\mathbf{x}^{(1)}\boldsymbol{\gamma}^{(1)})$	C	$\Lambda(\mathbf{z}\boldsymbol{\gamma})$
8	M	$\mathbf{x}\boldsymbol{\theta_g}$	M	$\Phi(\mathbf{x}^{(1)}\boldsymbol{\gamma}^{(1)})$	M	$\Phi(\mathbf{z}^{(1)}\boldsymbol{\gamma}^{(1)})$

Notes: C and M correspond to whether the estimated model is correctly specified or misspecified. $\mathbf{x}$ and $\mathbf{z}$ both include an intercept. $\mathbf{x}^{(1)}$ and $\mathbf{z}^{(1)}$ are the subsets of $\mathbf{x}$ and $\mathbf{z}$ left after omitting x_1. Therefore, the probability models have been misspecified in both the functional form and linear index dimension. $G(\cdot)$ refers to the propensity score model and $R(\cdot)$ refers to the missing outcomes probability model.

Table I.15: Estimation summary for quantile effects under different cases of misspecification

Scenario	CQF		$G(\cdot)$		$R(\cdot)$	
	Model	**Estimation**	**Model**	**Estimation**	**Model**	**Estimation**
4	C	$\exp(\mathbf{x}\boldsymbol{\theta}_{\boldsymbol{g}}(\tau))$	M	$\Phi(\mathbf{x}^{(1)}\boldsymbol{\gamma}^{(1)})$	M	$\Phi(\mathbf{x}^{(1)}\boldsymbol{\gamma}^{(1)})$
5	M	$\mathbf{x}\boldsymbol{\theta}_{\boldsymbol{g}}(\tau)$	C	$\Lambda(\mathbf{x}\boldsymbol{\gamma})$	C	$\Lambda(\mathbf{z}\boldsymbol{\gamma})$
6	M	$\mathbf{x}\boldsymbol{\theta}_{\boldsymbol{g}}(\tau)$	C	$\Lambda(\mathbf{x}\boldsymbol{\gamma})$	M	$\Phi(\mathbf{x}^{(1)}\boldsymbol{\gamma}^{(1)})$
7	M	$\mathbf{x}\boldsymbol{\theta}_{\boldsymbol{g}}(\tau)$	M	$\Phi(\mathbf{x}^{(1)}\boldsymbol{\gamma}^{(1)})$	C	$\Lambda(\mathbf{z}\boldsymbol{\gamma})$
8	M	$\mathbf{x}\boldsymbol{\theta}_{\boldsymbol{g}}(\tau)$	M	$\Phi(\mathbf{x}^{(1)}\boldsymbol{\gamma}^{(1)})$	M	$\Phi(\mathbf{x}^{(1)}\boldsymbol{\gamma}^{(1)})$

Notes: C and M denote whether the estimated model is correctly specified or misspecified. $\mathbf{x}$ and $\mathbf{z}$ both include an intercept. $\mathbf{x}^{(1)}$ and $\mathbf{z}^{(1)}$ are the subsets of $\mathbf{x}$ and $\mathbf{z}$ left after omitting x_1. Therefore, the probability models have been misspecified in both the functional form and the linear index dimension. $G(\cdot)$ refers to the propensity score model and $R(\cdot)$ refers to the missing outcomes probability model.

Table I.16: Covariate means and p-values from the test of equality of two means, by treatment status

Covariates	**Treatment**	**Control**	**P(|T| > |t|)**	**PSID-1**	**P(|T| > |t|)**	**PSID-2**	**P(|T| > |t|)**
Age in years	33.37	33.64	0.46	36.73	0.00	34.41	0.11
	(7.42)	(7.19)		(10.60)		(9.48)	
Years of education	10.30	10.27	0.72	11.32	0.00	10.55	0.07
	(1.92)	(2.00)		(2.71)		(2.09)	
Proportion of high school dropouts	0.70	0.69	0.73	0.45	0.00	0.59	0.00
	(0.46)	(0.46)		(0.50)		(0.49)	
Proportion Married	0.02	0.04	0.03	0.02	0.05	0.01	0.08
	(0.15)	(0.20)		(0.13)		(0.10)	
Proportion Black	0.84	0.82	0.29	0.66	0.00	0.87	0.13
	(0.37)	(0.39)		(0.47)		(0.34)	
Proportion Hispanic	0.12	0.13	0.59	0.02	0.00	0.02	0.00
	(0.32)	(0.33)		(0.12)		(0.16)	
Number of children in 1975	2.17	2.26	0.21	1.70	0.00	2.91	0.00
	(1.30)	(1.32)		(1.75)		(1.73)	
Real earnings in 1975	799.88	811.19	0.91	7446.15	0.00	2069.65	0.00
	(1931.92)	(2041.32)		(7515.59)		(3474.10)	
Observations	796	795		729		204	

Notes: Along with the covariate means and standard deviation (in parentheses), the table also reports p-values from the test of equality for two means. Column 4 tests for differences between the NSW treatment and control groups, column 6 and 8 report the same using PSID-1 and PSID-2 comparison groups respectively. Real earnings in 1975 are expressed in terms of 1982 dollars.

Table I.17: Covariate means and p-values from the test of equality of two means for the observed and missing samples

	Control			Treatment			PSID-1			PSID-2		
Covariates	Missing	Observed	$P(\|T\| > \|t\|)$	Missing	Observed	$P(\|T\| > \|t\|)$	Missing	Observed	$P(\|T\| > \|t\|)$	Missing	Observed	$P(\|T\| > \|t\|)$
Age	33.36	33.74	0.51	32.15	33.77	0.01	34.00	37.07	0.01	33.32	34.54	0.62
	(7.30)	(7.15)		(7.39)	(7.40)		(10.50)	(10.57)		(10.81)	(9.34)	
Years of education	10.29	10.26	0.85	10.29	10.31	0.89	11.44	11.30	0.60	11.05	10.49	0.18
	(1.93)	(2.03)		(2.05)	(1.88)		(2.17)	(2.77)		(1.73)	(2.13)	
Proportion of high school dropouts	0.70	0.68	0.57	0.69	0.70	0.77	0.43	0.45	0.73	0.55	0.59	0.68
	(0.46)	(0.47)		(0.46)	(0.46)		(0.50)	(0.50)		(0.51)	(0.49)	
Proportion married	0.05	0.04	0.61	0.03	0.02	0.75	0.00	0.02	0.00	0.00	0.01	0.16
	(0.21)	(0.19)		(0.16)	(0.15)		(0.00)	(0.14)		(0.00)	(0.10)	
Proportion black	0.81	0.82	0.81	0.83	0.84	0.87	0.74	0.65	0.10	0.91	0.86	0.50
	(0.39)	(0.39)		(0.38)	(0.37)		(0.44)	(0.48)		(0.29)	(0.35)	
Proportion hispanic	0.12	0.13	0.87	0.13	0.12	0.64	0.01	0.02	0.82	0.05	0.02	0.62
	(0.33)	(0.33)		(0.33)	(0.32)		(0.11)	(0.12)		(0.21)	(0.15)	
Number of children in 1975	2.33	2.23	0.34	2.14	2.19	0.69	1.54	1.71	0.33	2.41	2.97	0.05
	(1.29)	(1.34)		(1.32)	(1.29)		(1.45)	(1.78)		(1.14)	(1.79)	
Real earnings in 1975	621.54	879.28	0.12	610.77	861.65	0.11	6927.95	7510.92	0.50	896.56	2211.45	0.02
	(1,523.00)	(2,194.93)		(1,677.36)	(2,005.53)		(7,330.74)	(7,541.41)		(2,315.12)	(3,567.50)	
Observations	795	795		796	796		729	729		204	204	

Notes: Along with the covariate means and standard deviation (in parentheses), the table also reports p-values from the test of equality for two means between the observed and missing samples. Real earnings in 1975 are expressed in terms of 1982 dollars.

Table I.18: Unweighted and weighted earnings comparisons and estimated training effects using NSW and PSID comparison groups

Comparison group	Post-training earnings estimates								
	Unadjusted		Adjusted				Adjusted		
	Unweighted	PS-weighted	D-weighted	Unweighted	PS-weighted	D-weighted	Unweighted	PS-weighted	D-weighted
NSW	821	848	824	845	852	828	864	850	826
N=1,185	(307.22)	(304.04)	(304.61)	(303.60)	(302.94)	(303.53)	(303.47)	(302.96)	(303.58)
PSID-1	-799	827	803	298	909	907	335	905	904
N=1,016	(444.84)	(503.00)	(503.26)	(428.60)	(497.76)	(501.54)	(440.18)	(518.54)	(522.97)
PSID-2	-31	569	566	492	1,040	996	698	1,082	1,049
N=720	(713.88)	(1041.81)	(1027.12)	(664.46)	(961.74)	(953.80)	(784.28)	(1264.18)	(1217.46)
			Bias estimates using NSW control						
PSID-1	-1,620	169	156	-493	-40	-21	-568	-38	-21
N=1,001	(431.75)	(561.74)	(553.07)	(427.93)	(499.91)	(501.44)	(434.59)	(504.19)	(507.02)
PSID-2	-853	-228	-212	-109	207	200	-378	-17	-24
N=705	(707.87)	(1041.44)	(1025.87)	(663.80)	(962.85)	(954.61)	(759.75)	(1195.47)	(1156.39)
Adjusted covariates									
Pre-training earnings (1975)				✓	✓	✓	✓	✓	✓
Age				✓	✓	✓	✓	✓	✓
Age^2				✓	✓	✓	✓	✓	✓
Education				✓	✓	✓	✓	✓	✓
High school droput				✓	✓	✓	✓	✓	✓
Black				✓	✓	✓	✓	✓	✓
Hispanic				✓	✓	✓	✓	✓	✓
Marital status				✓	✓	✓	✓	✓	✓
Number of Children (1975)							✓	✓	✓

Notes: This table reports unadjusted and adjusted post-training earnings differences between the NSW treatment and three different comparison groups, namely, NSW control, PSID-1 and PSID-2. The first row reports experimental training estimates which combines the NSW treatment and control group whereas the second and third rows report non-experimental estimates computed from using the PSID-1 and PSID-2 groups respectively. Each of the non-experimental estimates should be compared to the experimental benchmark. The second panel of the table reports bias estimates computed from combining the NSW control with PSID-1 and PSID-2 comparison groups respectively. These represent a second measure of bias which should be compared to zero. Bootstrapped standard errors are given in parentheses and have been constructed using 10,000 replications. All values are in 1982 dollars. The samples used for estimating the training and bias estimates have been trimmed to ensure common support in the distribution of weights for the treatment and comparison groups. For more detail, see the application appendix.

Table I.19: Unconditional quantile treatment effect (UQTE) using PSID-1 comparison group

Quantile	Experimental	Unweighted	PS-weighted	D-weighted
0.1	0	0	0	0
	(0)	(0)	(0)	(0)
0.2	0	0	0	0
	(0)	(0)	(0)	(0)
0.3	0	0	0	0
	(0)	(12.91)	(0)	(0)
0.4	0	-1124.61	0	0
	(11.17)	(552.97)	(207.14)	(174.89)
0.5	993.52	-2227.26	2076.58	1847.04
	(695.93)	(983.43)	(851.09)	(829.42)
0.6	2004.40	-860.55	3602.76	3535.85
	(1112.82)	(964.97)	(1299.08)	(1284.64)
0.7	2129.93	428.01	3415.47	3340.84
	(716.04)	(728.22)	(988.24)	(992.95)
0.8	1753.27	-190.60	2019.44	2019.44
	(372.37)	(519.63)	(984.59)	(999.47)
0.9	1134.21	-1563.27	-385.45	-385.45
	(449.86)	(952.85)	(1059.43)	(1056.09)

Notes: This table reports unweighted, PS-weighted and double-weighted UQTE estimates for three different comparison groups, namely, NSW control, PSID-1 and PSID-2. The estimates are reported at every 10th quantile of the 1979 earnings distribution. The experimental and PSID-1 estimates have been constructed using N=1,185 and N=1,016 observations respectively. Bootstrapped standard errors are given in parentheses and have been constructed using 1,000 replications. All values are in 1982 dollars. The samples used for constructing these estimates have been trimmed to ensure common support across the treatment and comparison groups.

Table I.20: Unconditional quantile treatment effect (UQTE) using PSID-2 comparison group

Quantile	Experimental	Unweighted	PS-weighted	D-weighted
0.1	0	0	0	0
	(0)	(0)	(0)	(0)
0.2	0	0	0	0
	(0)	(0)	(10.07)	(10.07)
0.3	0	0	0	0
	(0)	(111.74)	(136.31)	(129.77)
0.4	0	-795.71	0	0
	(13.25)	(672.87)	(573.22)	(546.78)
0.5	993.52	-237.98	378.98	372.07
	(693.73)	(1232.63)	(1312.93)	(1267.28)
0.6	2004.40	193.77	1480.47	1294.77
	(1114.65)	(1426.40)	(1647.31)	(1659.69)
0.7	2129.93	1857.64	2616.22	2599.73
	(710.26)	(943.38)	(1217.80)	(1209.60)
0.8	1753.27	1148.85	2010.87	1990.37
	(371.73)	(1152.92)	(1541.14)	(1553.67)
0.9	1134.21	-237.08	1089.10	1089.10
	(452.08)	(1888.06)	(3321.56)	(3246.78)

Notes: This table reports unweighted, PS-weighted and double-weighted UQTE estimates for three different comparison groups, namely, NSW control, PSID-1 and PSID-2. The estimates are reported at every 10th quantile of the 1979 earnings distribution. The experimental and PSID-2 estimates have been computed using N=1,185 and N=720 observations respectively. Bootstrapped standard errors are given in parentheses and have been constructed using 1,000 replications. All values are in 1982 dollars. The samples used for constructing these estimates have been trimmed to ensure common support across the treatment and comparison groups.

APPENDIX J

PROOFS FOR CHAPTER 3

Proof of Lemma 3.2.5

Proof. By the law of iterated expectations (LIE)

$$
\begin{aligned}
&\mathbb{E}\left[\frac{s}{r(\mathbf{x}, w_g)} \cdot \frac{w_g}{p_g(\mathbf{x})} \cdot q\left(y(g), \mathbf{x}, \boldsymbol{\theta_g}\right)\right] \\
&= \mathbb{E}\left[\mathbb{E}\left(\frac{s}{r(\mathbf{x}, w_g)} \cdot \frac{w_g}{p_g(\mathbf{x})} \cdot q\left(y(g), \mathbf{x}, \boldsymbol{\theta_g}\right) \middle| y(g), \mathbf{x}, w_g\right)\right] \\
&= \mathbb{E}\left[\frac{w_g}{r(\mathbf{x}, w_g) \cdot p_g(\mathbf{x})} \cdot q\left(y(g), \mathbf{x}, \boldsymbol{\theta_g}\right) \cdot \mathbb{P}\left(s = 1 | y(g), \mathbf{x}, w_g\right)\right] \\
&= \mathbb{E}\left[\frac{w_g}{r(\mathbf{x}, w_g) \cdot p_g(\mathbf{x})} \cdot q\left(y(g), \mathbf{x}, \boldsymbol{\theta_g}\right) \cdot \mathbb{P}\left(s = 1 | \mathbf{x}, w_g\right)\right] \\
&= \mathbb{E}\left[\frac{w_g}{r(\mathbf{x}, w_g) \cdot p_g(\mathbf{x})} \cdot q\left(y(g), \mathbf{x}, \boldsymbol{\theta_g}\right) \cdot r(\mathbf{x}, w_g)\right] \\
&= \mathbb{E}\left[\frac{w_g}{p_g(\mathbf{x})} \cdot q\left(y(g), \mathbf{x}, \boldsymbol{\theta_g}\right)\right]
\end{aligned}
$$

Using another application of LIE, rewrite the above expectation as

$$
\begin{aligned}
&= \mathbb{E}\left[\mathbb{E}\left(\frac{w_g}{p_g(\mathbf{x})} \cdot q\left(y(g), \mathbf{x}, \boldsymbol{\theta_g}\right) \middle| y(g), \mathbf{x}\right)\right] \\
&= \mathbb{E}\left[\frac{q\left(y(g), \mathbf{x}, \boldsymbol{\theta_g}\right)}{p_g(\mathbf{x})} \cdot P\left(w_g = 1 | y(g), \mathbf{x}\right)\right] \\
&= \mathbb{E}\left[\frac{q\left(y(g), \mathbf{x}, \boldsymbol{\theta_g}\right)}{p_g(\mathbf{x})} \cdot P\left(w_g = 1 | \mathbf{x}\right)\right] \\
&= \mathbb{E}\left[\frac{q\left(y(g), \mathbf{x}, \boldsymbol{\theta_g}\right)}{p_g(\mathbf{x})} \cdot p_g(\mathbf{x})\right] \\
&= \mathbb{E}\left[q\left(y(g), \mathbf{x}, \boldsymbol{\theta_g}\right)\right]
\end{aligned}
$$

where the third equality follows from ignorability and the third last equality follows from unconfoundedness. Hence, $\boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}}$ solves the weighted population problem. ■

Proof of Lemma 3.3.3

Proof. Proving consistency of $\hat{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\delta}}$ follows directly after verifying the conditions of Theorem 2.1 in Newey and McFadden (1994). Condition 2.1(i), which implies unique solution to the maximization problem, is satisfied using 3.3.3 (1), (4) and Lemma 2.2 of Newey and McFadden (1994). Condition 2.1(ii), which implies compactness of the parameter space, holds due to 3.3.3(i). Conditions 2.1(iii) and (iv) follow from Lemma 2.4 in Newey and McFadden (1994). ■

Proof of Lemma 3.3.4

Proof. Again, the proof of asymptotic normality follows from verifying the conditions of Theorem 3.1 in Newey and McFadden (1994), which is the basic asymptotic normality proof for extremum estimators. I will then use the arguments as laid out in Newey and McFadden (1994) to prove asymptotic normality of $\sqrt{N}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma_0})$. The asymptotic normality for $\sqrt{N}(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta_0})$ follows in a similar manner.
By lemma 3.3.3, we have $\hat{\boldsymbol{\gamma}} \xrightarrow{p} \boldsymbol{\gamma}$. Theorem 3.1(i) and (ii) hold because of condition 3.4(i) and (ii). 3.1(iii) holds with $\boldsymbol{\Sigma} = -\mathbb{E}\left[\nabla_{\boldsymbol{\gamma}\boldsymbol{\gamma}'} \ln f(w_1|\mathbf{x}, \boldsymbol{\gamma_0})\right]$ by the information matrix equality, $\mathbb{E}\left[\nabla_{\boldsymbol{\gamma}} \ln f(w_1|\mathbf{x}, \boldsymbol{\gamma_0})\right] = \mathbf{0}$ (condition 3.4(iii)), existence of $\mathbb{E}\left[\nabla_{\boldsymbol{\gamma}\boldsymbol{\gamma}'} \ln f(w_1|\mathbf{x}, \boldsymbol{\gamma_0})\right]$ (condition 3.4(iii)), and the Lindberg-Levy central limit theorem. Condition 3.1(iv) follows from results of Lemma 2.4 in Newey and McFadden (1994) which require compactness of $\boldsymbol{\Gamma}$, $\boldsymbol{\gamma}$ being an interior point in $\boldsymbol{\Gamma}$, with $a(\mathbf{z}, \boldsymbol{\theta}) = \nabla_{\boldsymbol{\gamma}\boldsymbol{\gamma}'} \ln f(w_1|\mathbf{x}, \boldsymbol{\gamma})$ using conditions (ii) and (v). Condition 3.1(v) follows from non-singularity of $\mathbb{E}\left[\nabla_{\boldsymbol{\gamma}\boldsymbol{\gamma}'} \ln f(w_1|\mathbf{x}, \boldsymbol{\gamma_0})\right]$ using condition 3.4(iv). Then, asymptotic normality follows from the conclusion of Theorem 3.1 in Newey and McFadden (1994). ■

Proof of Theorem 3.4.1

Proof. I have already shown that

$$\mathbb{E}\left[\frac{s}{r(\mathbf{x}, w_g)} \cdot \frac{w_g}{p_g(\mathbf{x})} \cdot q(y(g), \mathbf{x}, \boldsymbol{\theta_g})\right] = \mathbb{E}\left[q(y(g), \mathbf{x}, \boldsymbol{\theta_g})\right]$$

for both $g = 0, 1$. Now, one needs to prove uniform convergence of the weighted sample objective function to its population expectation. Formally, I need to show

$$\sup_{\boldsymbol{\theta_g} \in \boldsymbol{\Theta_g}} \left\| \frac{1}{N_g} \sum_{i=1}^{N} \frac{s_i \cdot w_{ig}}{r(\mathbf{x}_i, w_{ig}) \cdot p_g(\mathbf{x}_i)} \cdot q(y_i(g), \mathbf{x}_i, \boldsymbol{\theta_g}) - \mathbb{E}\left[\frac{s}{r(\mathbf{x}, w_g)} \cdot \frac{w_g}{p_g(\mathbf{x})} \cdot q(y(g), \mathbf{x}, \boldsymbol{\theta_g})\right] \right\| \xrightarrow{p} 0$$

Then consider,

$$\left| \frac{s}{r(\mathbf{x}, w_g)} \cdot \frac{w_g}{p_g(\mathbf{x})} \cdot q\left(y(g), \mathbf{x}, \boldsymbol{\theta_g}\right) \right| \leq \frac{|q\left(y(g), \mathbf{x}, \boldsymbol{\theta_g}\right)|}{r(\mathbf{x}, w_g) \cdot p_g(\mathbf{x})} \tag{J.1}$$

$$\leq \frac{b(y(g), \mathbf{x})}{\eta \cdot \kappa_g} \tag{J.2}$$

Inequality J.2 holds due to part (3) of Assumptions 3.2.2 and 3.2.3. Now, $\mathbb{E}\left[b(y(g), \mathbf{x})\right] < \infty$ by condition (3) in this theorem. Therefore, uniform convergence is established by Lemma 2.4 of Newey and McFadden (1994). Hence,

$$\hat{\boldsymbol{\theta}}_{\boldsymbol{g}} \xrightarrow{p} \boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}}$$

Replacing the true probabilities, $r(\cdot)$, and, $p_g(\cdot)$, by their consistent estimates does not change the above result. ■

Proof of Theorem 3.4.2

Proof. Following Newey and McFadden (1994), with minor modifications, I will first show that $\sqrt{N}\left\|\hat{\boldsymbol{\theta}}_{\boldsymbol{g}} - \boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}}\right\| = O_p(1)$ or in other words, $\hat{\boldsymbol{\theta}}_{\boldsymbol{g}}$ is $\sqrt{N}-$consistent.

$$\begin{aligned} Q_0(\boldsymbol{\theta_g}) &= Q_0(\boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}}) + \nabla_{\boldsymbol{\theta}_g} Q_0(\boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}})' + (\boldsymbol{\theta_g} - \boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}})' \mathbf{H_g} (\boldsymbol{\theta_g} - \boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}})/2 + o\left(\left\|\boldsymbol{\theta_g} - \boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}}\right\|^2\right) \\ &= Q_0(\boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}}) + (\boldsymbol{\theta_g} - \boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}})' \mathbf{H_g} (\boldsymbol{\theta_g} - \boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}})/2 + o\left(\left\|\boldsymbol{\theta_g} - \boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}}\right\|^2\right) \end{aligned} \tag{J.3}$$

where the first equality follows from the second order Taylor series approximation. For the second equality, since $Q_N(\boldsymbol{\theta_g})$ has a local minimum at $\boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}}$ at $\boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}}$, the first derivative will be

zero. Since $\mathbf{H_g}$ is positive definite and non-signular, there exists a constant $C \geq 0$ and a small enough neighborhood of $\boldsymbol{\theta_g^0}$ such that

$$(\boldsymbol{\theta_g} - \boldsymbol{\theta_g^0})' \mathbf{H_g} (\boldsymbol{\theta_g} - \boldsymbol{\theta_g^0})/2 + o\left(\left\|\boldsymbol{\theta_g} - \boldsymbol{\theta_g^0}\right\|\right)^2 \geq C\left\|\boldsymbol{\theta_g} - \boldsymbol{\theta_g^0}\right\|^2$$

Therefore, since $\boldsymbol{\theta_g} \xrightarrow{p} \boldsymbol{\theta_g^0}$ with probability approaching one we can rewrite eq (J.3) as

$$Q_0(\boldsymbol{\theta_g}) \geq Q_0(\boldsymbol{\theta_g^0}) + C\left\|\boldsymbol{\theta_g} - \boldsymbol{\theta_g^0}\right\|^2$$

Let us define,

$$R_N(\boldsymbol{\theta_g}) = Q_N(\boldsymbol{\theta_g}) - Q_N(\boldsymbol{\theta_g^0}) - (Q_0(\boldsymbol{\theta_g}) - Q_0(\boldsymbol{\theta_g^0})) - \nabla_{\boldsymbol{\theta_g}} Q_N(\boldsymbol{\theta_g^0})'(\boldsymbol{\theta_g} - \boldsymbol{\theta_g^0})$$

then using Ossiander's entropy conditions given in 4.2(6), 4.2(7) along with $i.i.d$ sampling as given in assumption (2.4), one obtains stochastic equicontinuity using Theorem 4 and Theorem 5 (with $p = 2$) of Andrews (1994). Hence, for any sequence, $\beta_N \to 0$

$$\sup_{\left\|\boldsymbol{\theta_g} - \boldsymbol{\theta_g^0}\right\| \leq \beta_N} \frac{\sqrt{N} \cdot R_N(\boldsymbol{\theta_g})}{\left\|\boldsymbol{\theta_g} - \boldsymbol{\theta_g^0}\right\| (1 + \sqrt{N}\left\|\boldsymbol{\theta_g} - \boldsymbol{\theta_g^0}\right\|)} = o_p(1)$$

In other words, the above implies that with probability approaching one, for all $\boldsymbol{\theta_g}$,

$$\sqrt{N} \cdot R_N(\boldsymbol{\theta_g}) \leq \left\|\boldsymbol{\theta_g} - \boldsymbol{\theta_g^0}\right\| \left(1 + \sqrt{N}\left\|\boldsymbol{\theta_g} - \boldsymbol{\theta_g^0}\right\|\right) o_p(1) \tag{J.4}$$

Choose U_N so that $\hat{\boldsymbol{\theta}}_{\boldsymbol{g}} \in U_N$ with probability approaching one, so that eq (J.4) holds. Again since $\hat{\boldsymbol{\theta}}_{\boldsymbol{g}}$ is consistent for $\boldsymbol{\theta_g^0}$, we can write

$$\begin{aligned}
0 &\geq Q_N(\hat{\boldsymbol{\theta}}_{\boldsymbol{g}}) - Q_N(\boldsymbol{\theta_g^0}) - o_p(N^{-1}) \\
&= Q_0(\hat{\boldsymbol{\theta}}_{\boldsymbol{g}}) - Q_0(\boldsymbol{\theta_g^0}) + \nabla_{\boldsymbol{\theta}_g} Q_N(\boldsymbol{\theta_g^0})'(\hat{\boldsymbol{\theta}}_{\boldsymbol{g}} - \boldsymbol{\theta_g^0}) + R_N(\hat{\boldsymbol{\theta}}_{\boldsymbol{g}}) - o_p(N^{-1}) \\
&\geq C\left\|\hat{\boldsymbol{\theta}}_{\boldsymbol{g}} - \boldsymbol{\theta_g^0}\right\|^2 + \left\|\nabla_{\boldsymbol{\theta}_g} Q_N(\boldsymbol{\theta_g^0})'\right\|\left\|\hat{\boldsymbol{\theta}}_{\boldsymbol{g}} - \boldsymbol{\theta_g^0}\right\| + \left\|\hat{\boldsymbol{\theta}}_{\boldsymbol{g}} - \boldsymbol{\theta_g^0}\right\| \left(1 + \sqrt{N}\left\|\hat{\boldsymbol{\theta}}_{\boldsymbol{g}} - \boldsymbol{\theta_g^0}\right\|\right) o_p(N^{-1/2}) \\
&\quad - o_p(N^{-1}) \\
&\geq \left[C + o_p(1)\right]\left\|\hat{\boldsymbol{\theta}}_{\boldsymbol{g}} - \boldsymbol{\theta_g^0}\right\|^2 + O_p(N^{-1/2})\left\|\hat{\boldsymbol{\theta}}_{\boldsymbol{g}} - \boldsymbol{\theta_g^0}\right\| - o_p(N^{-1})
\end{aligned} \tag{J.5}$$

We obtain the above simplification because,

$$
\begin{aligned}
& C \cdot \left\| \hat{\boldsymbol{\theta}}_{\boldsymbol{g}} - \boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}} \right\|^2 + \left\| \nabla_{\boldsymbol{\theta}_g} Q_N(\boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}})' \right\| \left\| \hat{\boldsymbol{\theta}}_{\boldsymbol{g}} - \boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}} \right\| + \left\| \hat{\boldsymbol{\theta}}_{\boldsymbol{g}} - \boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}} \right\| \left(1 + \sqrt{N} \left\| \hat{\boldsymbol{\theta}}_{\boldsymbol{g}} - \boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}} \right\| \right) o_p(N^{-1/2}) \\
& - o_p(N^{-1}) \\
& = C \cdot \left\| \hat{\boldsymbol{\theta}}_{\boldsymbol{g}} - \boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}} \right\|^2 + O_p(N^{-1/2}) \left\| \hat{\boldsymbol{\theta}}_{\boldsymbol{g}} - \boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}} \right\| + \left\| \hat{\boldsymbol{\theta}}_{\boldsymbol{g}} - \boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}} \right\| \cdot o_p(N^{-1/2}) + \left\| \hat{\boldsymbol{\theta}}_{\boldsymbol{g}} - \boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}} \right\|^2 \cdot o_p(1) \\
& - o_p(N^{-1}) \\
& = \left[C + o_p(1) \right] \cdot \left\| \hat{\boldsymbol{\theta}}_{\boldsymbol{g}} - \boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}} \right\|^2 + \left(O_p(N^{-1/2}) + o_p(N^{-1/2}) \right) \cdot \left\| \hat{\boldsymbol{\theta}}_{\boldsymbol{g}} - \boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}} \right\| - o_p(N^{-1}) \\
& = \left[C + o_p(1) \right] \cdot \left\| \hat{\boldsymbol{\theta}}_{\boldsymbol{g}} - \boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}} \right\|^2 + O_p(N^{-1/2}) \cdot \left\| \hat{\boldsymbol{\theta}}_{\boldsymbol{g}} - \boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}} \right\| - o_p(N^{-1})
\end{aligned}
$$

Then we can write the the inequality in (J.5) as

$$
\left[C + o_p(1) \right] \left\| \hat{\boldsymbol{\theta}}_{\boldsymbol{g}} - \boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}} \right\|^2 \leq -O_p(N^{-1/2}) \left\| \hat{\boldsymbol{\theta}}_{\boldsymbol{g}} - \boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}} \right\| + o_p(N^{-1})
$$

Since $C + o_p(1)$ is bounded away from zero with probability approaching one,

$$
\left\| \hat{\boldsymbol{\theta}}_{\boldsymbol{g}} - \boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}} \right\|^2 \leq -O_p(N^{-1/2}) \left\| \hat{\boldsymbol{\theta}}_{\boldsymbol{g}} - \boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}} \right\| + o_p(N^{-1})
$$

We now use completing the square trick with $x = \left\| \hat{\boldsymbol{\theta}}_{\boldsymbol{g}} - \boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}} \right\|$, $b = O_p(N^{-1/2})$ and $c = -o_p(N^{-1})$ to obtain

$$
\left(\left\| \hat{\boldsymbol{\theta}}_{\boldsymbol{g}} - \boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}} \right\| + \frac{O_p(N^{-1/2})}{2}) \right)^2 - \left(o_p(N^{-1}) + \frac{O_p(N^{-1/2}) \cdot O_p(N^{-1/2})}{4}) \right) \leq 0
$$

By the rules of the asymptotic notation, $O_p(N^{-1/2}) \cdot O_p(N^{-1/2}) = O_p(N^{-1})$. Therefore, we obtain,

$$
\begin{aligned}
\left(\left\| \hat{\boldsymbol{\theta}}_{\boldsymbol{g}} - \boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}} \right\| + O_p(N^{-1/2}) \right)^2 & \leq \left(o_p(N^{-1}) + O_p(N^{-1})) \right) \\
\left(\left\| \hat{\boldsymbol{\theta}}_{\boldsymbol{g}} - \boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}} \right\| + O_p(N^{-1/2}) \right)^2 & \leq O_p(N^{-1})
\end{aligned}
$$

Taking a square root on both sides,

$$
\left| \left\| \hat{\boldsymbol{\theta}}_{\boldsymbol{g}} - \boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}} \right\| + O_p(N^{-1/2}) \right| \leq O_p(N^{-1/2}) \tag{J.6}
$$

Now, by triangle inequality

$$\left\|\hat{\boldsymbol{\theta}}_{\boldsymbol{g}} - \boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}}\right\| = \left\|\hat{\boldsymbol{\theta}}_{\boldsymbol{g}} - \boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}}\right\| + O_p(N^{-1/2}) - O_p(N^{-1/2}) \leq \left|\left\|\hat{\boldsymbol{\theta}}_{\boldsymbol{g}} - \boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}}\right\| + O_p(N^{-1/2})\right| + \left|-O_p(N^{-1/2})\right|$$
$$\leq O_p(N^{-1/2}) \qquad \text{(By equation J.6)}$$

Hence we have established that $\hat{\boldsymbol{\theta}}_{\boldsymbol{g}}$ is $\sqrt{N}$ consistent. Now let, $\ddot{\boldsymbol{\theta}}_{\boldsymbol{g}} = \boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}} - \mathbf{H}_{\boldsymbol{g}}^{-1}\nabla_{\boldsymbol{\theta}_g} Q_N(\boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}})$, then $\ddot{\boldsymbol{\theta}}_{\boldsymbol{g}}$ is $\sqrt{N}-$consistent almost by construction since $\nabla_{\boldsymbol{\theta}_g} Q_N(\boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}})$ is $O_p(N^{-1/2})$. Now consider,

$$Q_N(\hat{\boldsymbol{\theta}}_{\boldsymbol{g}}) - Q_N(\boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}}) = Q_0(\hat{\boldsymbol{\theta}}_{\boldsymbol{g}}) - Q_0(\boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}}) + \nabla_{\boldsymbol{\theta}_g} Q_N(\boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}})' \cdot (\hat{\boldsymbol{\theta}}_{\boldsymbol{g}} - \boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}}) + R_N(\hat{\boldsymbol{\theta}}_{\boldsymbol{g}}) + o_p(N^{-1})$$

Using J.3 gives me,

$$Q_N(\hat{\boldsymbol{\theta}}_{\boldsymbol{g}}) - Q_N(\boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}}) = (\hat{\boldsymbol{\theta}}_{\boldsymbol{g}} - \boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}})'\mathbf{H}_{\boldsymbol{g}}(\hat{\boldsymbol{\theta}}_{\boldsymbol{g}} - \boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}})/2 + o\left(\left\|\hat{\boldsymbol{\theta}}_{\boldsymbol{g}} - \boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}}\right\|^2\right) +$$
$$\nabla_{\boldsymbol{\theta}_g} Q_N(\boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}})' \cdot (\hat{\boldsymbol{\theta}}_{\boldsymbol{g}} - \boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}}) + R_N(\hat{\boldsymbol{\theta}}_{\boldsymbol{g}}) + o_p(N^{-1})$$

Therefore, using the fact that $\nabla_{\boldsymbol{\theta}_g} Q_N(\boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}}) = -\mathbf{H}_{\boldsymbol{g}}(\ddot{\boldsymbol{\theta}}_{\boldsymbol{g}} - \boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}})$ we get

$$2\left[Q_N(\hat{\boldsymbol{\theta}}_{\boldsymbol{g}}) - Q_N(\boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}})\right] = (\hat{\boldsymbol{\theta}}_{\boldsymbol{g}} - \boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}})'\mathbf{H}_{\boldsymbol{g}}(\hat{\boldsymbol{\theta}}_{\boldsymbol{g}} - \boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}}) + 2\nabla_{\boldsymbol{\theta}_g} Q_N(\boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}})' \cdot (\hat{\boldsymbol{\theta}}_{\boldsymbol{g}} - \boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}}) + o_p(N^{-1})$$
$$= (\hat{\boldsymbol{\theta}}_{\boldsymbol{g}} - \boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}})'\mathbf{H}_{\boldsymbol{g}}(\hat{\boldsymbol{\theta}}_{\boldsymbol{g}} - \boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}}) - 2(\ddot{\boldsymbol{\theta}}_{\boldsymbol{g}} - \boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}})'\mathbf{H}_{\boldsymbol{g}}(\hat{\boldsymbol{\theta}}_{\boldsymbol{g}} - \boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}}) + o_p(N^{-1})$$

To show that the remaining terms are of order $o_p(N^{-1})$, observe that the order of magnitude

$$o\left(\left\|\hat{\boldsymbol{\theta}}_{\boldsymbol{g}} - \boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}}\right\|^2\right) + R_N(\hat{\boldsymbol{\theta}}_{\boldsymbol{g}}) + o_p(N^{-1}) = o(O_p(N^{-1/2}) \cdot O_p(N^{-1/2})) + O_p(N^{-1/2})$$
$$\cdot\, o_p(N^{-1/2} + O_p(N^{-1/2}))$$
$$= o_p(N^{-1})$$

In a similar manner, we can write,

$$2\left[Q_N(\ddot{\boldsymbol{\theta}}_{\boldsymbol{g}}) - Q_N(\boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}})\right] = (\ddot{\boldsymbol{\theta}}_{\boldsymbol{g}} - \boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}})'\mathbf{H}_{\boldsymbol{g}}(\ddot{\boldsymbol{\theta}}_{\boldsymbol{g}} - \boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}}) + 2\nabla_{\boldsymbol{\theta}_g} Q_N(\boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}})' \cdot (\ddot{\boldsymbol{\theta}}_{\boldsymbol{g}} - \boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}}) + o_p(N^{-1})$$
$$= (\ddot{\boldsymbol{\theta}}_{\boldsymbol{g}} - \boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}})'\mathbf{H}_{\boldsymbol{g}}(\ddot{\boldsymbol{\theta}}_{\boldsymbol{g}} - \boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}}) - 2(\ddot{\boldsymbol{\theta}}_{\boldsymbol{g}} - \boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}})'\mathbf{H}_{\boldsymbol{g}}(\ddot{\boldsymbol{\theta}}_{\boldsymbol{g}} - \boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}}) + o_p(N^{-1})$$
$$= -(\ddot{\boldsymbol{\theta}}_{\boldsymbol{g}} - \boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}})'\mathbf{H}_{\boldsymbol{g}}(\ddot{\boldsymbol{\theta}}_{\boldsymbol{g}} - \boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}}) + o_p(N^{-1})$$

Then,

$$
\begin{aligned}
&2\left[Q_N(\hat{\boldsymbol{\theta}}_{\boldsymbol{g}}) - Q_N(\boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}})\right] - 2\left[Q_N(\ddot{\boldsymbol{\theta}}_{\boldsymbol{g}}) - Q_N(\boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}})\right] \\
=&(\hat{\boldsymbol{\theta}}_{\boldsymbol{g}} - \boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}})'\mathbf{H}_{\boldsymbol{g}}(\hat{\boldsymbol{\theta}}_{\boldsymbol{g}} - \boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}}) - 2(\ddot{\boldsymbol{\theta}}_{\boldsymbol{g}} - \boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}})'\mathbf{H}_{\boldsymbol{g}}(\hat{\boldsymbol{\theta}}_{\boldsymbol{g}} - \boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}}) + (\ddot{\boldsymbol{\theta}}_{\boldsymbol{g}} - \boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}})'\mathbf{H}_{\boldsymbol{g}}(\ddot{\boldsymbol{\theta}}_{\boldsymbol{g}} - \boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}}) + o_p(N^{-1})
\end{aligned}
$$

where

$$
2\left[Q_N(\hat{\boldsymbol{\theta}}_{\boldsymbol{g}}) - Q_N(\boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}})\right] - 2\left[Q_N(\ddot{\boldsymbol{\theta}}_{\boldsymbol{g}}) - Q_N(\boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}})\right] \leq o_p(N^{-1})
$$

and

$$
\begin{aligned}
o_p(N^{-1}) &\geq (\hat{\boldsymbol{\theta}}_{\boldsymbol{g}} - \boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}})'\mathbf{H}_{\boldsymbol{g}}(\hat{\boldsymbol{\theta}}_{\boldsymbol{g}} - \boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}}) - 2(\ddot{\boldsymbol{\theta}}_{\boldsymbol{g}} - \boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}})'\mathbf{H}_{\boldsymbol{g}}(\hat{\boldsymbol{\theta}}_{\boldsymbol{g}} - \boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}}) + (\ddot{\boldsymbol{\theta}}_{\boldsymbol{g}} - \boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}})'\mathbf{H}_{\boldsymbol{g}}(\ddot{\boldsymbol{\theta}}_{\boldsymbol{g}} - \boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}}) \\
&= (\hat{\boldsymbol{\theta}}_{\boldsymbol{g}} - \ddot{\boldsymbol{\theta}}_{\boldsymbol{g}})'\mathbf{H}_{\boldsymbol{g}}(\hat{\boldsymbol{\theta}}_{\boldsymbol{g}} - \ddot{\boldsymbol{\theta}}_{\boldsymbol{g}}) \geq C\left\|\hat{\boldsymbol{\theta}}_{\boldsymbol{g}} - \ddot{\boldsymbol{\theta}}_{\boldsymbol{g}}\right\|^2
\end{aligned}
$$

Hence,

$$
\left\|\sqrt{N}(\hat{\boldsymbol{\theta}}_{\boldsymbol{g}} - \boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}}) - (-\mathbf{H}_{\boldsymbol{g}}^{-1}\sqrt{N}\nabla_{\boldsymbol{\theta}_{\boldsymbol{g}}}Q_N(\boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}}))\right\| = \sqrt{N}\left\|\hat{\boldsymbol{\theta}}_{\boldsymbol{g}} - \ddot{\boldsymbol{\theta}}_{\boldsymbol{g}}\right\| \xrightarrow{p} 0
$$

Therefore, the conclusion follows from the fact that

$$
-\mathbf{H}_{\boldsymbol{g}}^{-1}\sqrt{N}\nabla_{\boldsymbol{\theta}_{\boldsymbol{g}}}Q_N(\boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}}) \xrightarrow{d} N(\mathbf{0}, \mathbf{H}_{\boldsymbol{g}}^{-1}\boldsymbol{\Omega}_{\boldsymbol{g}}\mathbf{H}_{\boldsymbol{g}}^{-1})
$$

■

Proof of Theorem 3.4.3

Proof. Consider,

$$
\begin{aligned}
&\boldsymbol{\Sigma}_{\mathbf{1}} - \boldsymbol{\Omega}_{\mathbf{1}} \\
&= \mathbb{E}\left(\mathbf{l}_i\mathbf{l}_i'\right) - \{\mathbb{E}\left(\mathbf{l}_i\mathbf{l}_i'\right) - \mathbb{E}\left(\mathbf{l}_i\mathbf{b}_i'\right)\mathbb{E}\left(\mathbf{b}_i\mathbf{b}_i'\right)^{-1}\mathbb{E}(\mathbf{b}_i\mathbf{l}_i') - \mathbb{E}\left(\mathbf{l}_i\mathbf{d}_i'\right)\mathbb{E}\left(\mathbf{d}_i\mathbf{d}_i'\right)^{-1}\mathbb{E}(\mathbf{d}_i\mathbf{l}_i')\} \\
&= \mathbb{E}\left(\mathbf{l}_i\mathbf{b}_i'\right)\mathbb{E}\left(\mathbf{b}_i\mathbf{b}_i'\right)^{-1}\mathbb{E}(\mathbf{b}_i\mathbf{l}_i') + \mathbb{E}\left(\mathbf{l}_i\mathbf{d}_i'\right)\mathbb{E}\left(\mathbf{d}_i\mathbf{d}_i'\right)^{-1}\mathbb{E}(\mathbf{d}_i\mathbf{l}_i')
\end{aligned}
$$

since each component matrix in the above expression is positive semi-definite, therefore the sum of the two matrices is also positive semi-definite. The proof for the control group follows analogously. ■

Proof of Theorem 3.5.4

Proof. I have shown that

$$\mathbb{E}\left[\frac{s}{R(\mathbf{x}, w_1, \boldsymbol{\delta}^*)} \cdot \frac{w_1}{G(\mathbf{x}, \boldsymbol{\gamma}^*)} \cdot q(y(1), \mathbf{x}, \boldsymbol{\theta_1})\right]$$

will identify the parameter of interest, $\boldsymbol{\theta_1^0}$, under the strong identification condition given in 3.5.1. In order to prove consistency of $\hat{\boldsymbol{\theta}}_\mathbf{1}$ for $\boldsymbol{\theta_1^0}$, we need to prove uniform convergence of the weighted sample objective function to its population expectation. Formally, we need to show

$$\sup_{\boldsymbol{\theta_1} \in \boldsymbol{\Theta_1}} \left\| \frac{1}{N_1} \sum_{i=1}^{N} \frac{s_i \cdot w_{i1}}{R(\mathbf{x}_i, w_{i1}, \boldsymbol{\delta}^*) \cdot G(\mathbf{x}_i, \boldsymbol{\gamma}^*)} \cdot q(y_i(1), \mathbf{x}_i, \boldsymbol{\theta_1}) \right.$$
$$\left. - \mathbb{E}\left[\frac{s \cdot w_1}{R(\mathbf{x}, w_1, \boldsymbol{\delta}^*) \cdot G(\mathbf{x}, \boldsymbol{\gamma}^*)} \cdot q(y(1), \mathbf{x}, \boldsymbol{\theta_1})\right] \right\| \xrightarrow{p} 0$$

Replacing $r(\mathbf{x}, w_1)$ and $p_1(\mathbf{x})$ in the proof of theorem 3.4.1 by $R(\mathbf{x}, w_1, \boldsymbol{\delta}^*)$ and $G(\mathbf{x}, \boldsymbol{\gamma}^*)$ gives us the desired result. Consistency of $\hat{\boldsymbol{\theta}}_\mathbf{0}$ for $\boldsymbol{\theta_0^0}$ can be established analogously by replacing w_1, $G(\mathbf{x}, \boldsymbol{\gamma}^*)$ and $R(\mathbf{x}, w_1, \boldsymbol{\delta}^*)$ above by w_0, $(1 - G(\cdot, \boldsymbol{\gamma}^*))$ and $R(\mathbf{x}, w_0, \boldsymbol{\delta}^*)$ respectively. ■

Proof of Theorem 3.5.5

Proof. The proof of this theorem follows in the manner of Theorem 3.4.2 but where $\mathbf{H_g}$ now denotes the non-singular Hessian, with weights given by $G(\mathbf{x}, \boldsymbol{\gamma}^*)$ and $R(\mathbf{x}, w_g, \boldsymbol{\delta}^*)$. Also, $\boldsymbol{\Omega_g}$ now denotes the variance of the doubly weighted scores, $\mathbf{l}_i$ and $\mathbf{k}_i$ for the treatment and control group problems respectively. ■

Proof of corollary 3.5.6

Proof. This proof follows from the proof of the above theorem, 3.5.5, and the asymptotic variance of the estimator that uses known weights which is

$$\text{Avar}\sqrt{N}\left(\tilde{\boldsymbol{\theta}}_\mathbf{g} - \boldsymbol{\theta_g^0}\right) = \mathbf{H_g}^{-1}\boldsymbol{\Omega_g}\mathbf{H_g}^{-1}$$

where $\boldsymbol{\Omega_1} = \mathbb{E}\left(\mathbf{l}_i\mathbf{l}_i'\right)$ and $\boldsymbol{\Omega_0} = \mathbb{E}\left(\mathbf{k}_i\mathbf{k}_i'\right)$. The result follows immediately. ■

Proof of theorem 3.5.7

Proof. Using two applications of LIE and invoking ignorability and unconfoundedness, I can rewrite

$$\begin{aligned}
&\mathbb{E}\left[\frac{s_i \cdot w_{i1}}{R(\mathbf{x}_i, w_{i1}, \boldsymbol{\delta}^*) \cdot G(\mathbf{x}_i, \boldsymbol{\gamma}^*)} \cdot q(y_i(1), \mathbf{x}_i, \boldsymbol{\theta}_\mathbf{1}^\mathbf{0})\right] \\
&= \mathbb{E}\left[\frac{r(\mathbf{x}_i, w_{i1}) \cdot p_1(\mathbf{x}_i)}{R(\mathbf{x}_i, w_{i1}, \boldsymbol{\delta}^*) \cdot G(\mathbf{x}_i, \boldsymbol{\gamma}^*)} \cdot q(y_i(1), \mathbf{x}_i, \boldsymbol{\theta}_\mathbf{1}^\mathbf{0})\right]
\end{aligned}$$

Using another application of LIE, I can rewrite the above as

$$= \mathbb{E}\left[\frac{r(\mathbf{x}_i, w_{i1}) \cdot p_1(\mathbf{x}_i)}{R(\mathbf{x}_i, w_{i1}, \boldsymbol{\delta}^*) \cdot G(\mathbf{x}_i, \boldsymbol{\gamma}^*)} \cdot \mathbb{E}\left(q(y_i(1), \mathbf{x}_i, \boldsymbol{\theta}_\mathbf{1}^\mathbf{0})|\mathbf{x}_i\right)\right]$$

Then,

$$\begin{aligned}
\mathbf{H}_1 &= \boldsymbol{\nabla}^2_{\boldsymbol{\theta}_\mathbf{1}} \mathbb{E}\left[\frac{r(\mathbf{x}_i, w_{i1}) \cdot p_1(\mathbf{x}_i)}{R(\mathbf{x}_i, w_{i1}, \boldsymbol{\delta}^*) \cdot G(\mathbf{x}_i, \boldsymbol{\gamma}^*)} \cdot \mathbb{E}\left(q(y_i(1), \mathbf{x}_i, \boldsymbol{\theta}_\mathbf{1}^\mathbf{0})|\mathbf{x}_i\right)\right] \\
&= \mathbb{E}\left[\frac{r(\mathbf{x}_i, w_{i1}) \cdot p_1(\mathbf{x}_i)}{R(\mathbf{x}_i, w_{i1}, \boldsymbol{\delta}^*) \cdot G(\mathbf{x}_i, \boldsymbol{\gamma}^*)} \cdot \mathbf{A}(\mathbf{x}_i, \boldsymbol{\theta}_\mathbf{1}^\mathbf{0})\right]
\end{aligned}$$

Similarly, I use LIE to express $\boldsymbol{\Omega}_\mathbf{1}$ as

$$\begin{aligned}
&\boldsymbol{\Omega}_\mathbf{1} \\
&= \mathbb{E}\Bigg[\mathbb{E}\Bigg(\frac{r(\mathbf{x}_i, w_{i1}) \cdot p_1(\mathbf{x}_i)}{R^2(\mathbf{x}_i, w_{i1}, \boldsymbol{\delta}^*) \cdot G^2(\mathbf{x}_i, \boldsymbol{\gamma}^*)} \\
&\cdot \boldsymbol{\nabla}_{\boldsymbol{\theta}_\mathbf{1}} q(y_i(1), \mathbf{x}_i, \boldsymbol{\theta}_\mathbf{1}^\mathbf{0})' \boldsymbol{\nabla}_{\boldsymbol{\theta}_\mathbf{1}} \cdot q(y_i(1), \mathbf{x}_i, \boldsymbol{\theta}_\mathbf{1}^\mathbf{0}) \Big| \mathbf{x}_i, w_{i1}, s_i\Bigg)\Bigg] \\
&= \mathbb{E}\left[\frac{r(\mathbf{x}_i, w_{i1}) \cdot p_1(\mathbf{x}_i)}{R^2(\mathbf{x}_i, w_{i1}, \boldsymbol{\delta}^*) \cdot G^2(\mathbf{x}_i, \boldsymbol{\gamma}^*)} \cdot \mathbb{E}\left(\boldsymbol{\nabla}_{\boldsymbol{\theta}_\mathbf{1}} q(y_i(1), \mathbf{x}_i, \boldsymbol{\theta}_\mathbf{1}^\mathbf{0})' \boldsymbol{\nabla}_{\boldsymbol{\theta}_\mathbf{1}} q(y_i(1), \mathbf{x}_i, \boldsymbol{\theta}_\mathbf{1}^\mathbf{0}) \Big| \mathbf{x}_i, w_{i1}, s_i\right)\right] \\
&= \mathbb{E}\left[\frac{r(\mathbf{x}_i, w_i) \cdot p(\mathbf{x}_i)}{R^2(\mathbf{x}_i, w_i, \boldsymbol{\delta}^*) \cdot G^2(\mathbf{x}_i, \boldsymbol{\gamma}^*)} \cdot \mathbb{E}\left(\boldsymbol{\nabla}_{\boldsymbol{\theta}_\mathbf{1}} q(y_i(1), \mathbf{x}_i, \boldsymbol{\theta}_\mathbf{1}^\mathbf{0})' \boldsymbol{\nabla}_{\boldsymbol{\theta}_\mathbf{1}} q(y_i(1), \mathbf{x}_i, \boldsymbol{\theta}_\mathbf{1}^\mathbf{0}) \Big| \mathbf{x}_i\right)\right] \\
&= \sigma^2_{01} \cdot \mathbb{E}\left[\frac{r(\mathbf{x}_i, w_i) \cdot p(\mathbf{x}_i)}{R^2(\mathbf{x}_i, w_i, \boldsymbol{\delta}^*) \cdot G^2(\mathbf{x}_i, \boldsymbol{\gamma}^*)} \cdot \mathbf{A}(\mathbf{x}_i, \boldsymbol{\theta}_\mathbf{1}^\mathbf{0})\right]
\end{aligned}$$

For the unweighted estimator, the variance simplifies, and this happens precisely due to the GCIME. To see this, consider $\mathbf{H_1^u}$. Then using LIE, I can rewrite

$$\begin{aligned}\mathbf{H_1^u} &= \boldsymbol{\nabla}^2_{\boldsymbol{\theta_1}} \mathbb{E}\left[r(\mathbf{x}_i, w_{i1}) \cdot p_1(\mathbf{x}_i) \cdot \mathbb{E}\left(q(y_i(1), \mathbf{x}_i, \boldsymbol{\theta_1^0}) | \mathbf{x}_i\right)\right] \\ &= \mathbb{E}\left[r(\mathbf{x}_i, w_{i1}) \cdot p_1(\mathbf{x}_i) \cdot \mathbf{A}(\mathbf{x}_i, \boldsymbol{\theta_1^0})\right]\end{aligned}$$

and similarly we can rewrite $\mathbf{\Omega_1^u}$ using LIE as

$$\begin{aligned}\mathbf{\Omega_1^u} &= \mathbb{E}\left[r(\mathbf{x}_i, w_i) \cdot p(\mathbf{x}_i) \cdot \mathbb{E}\left(\boldsymbol{\nabla}_{\boldsymbol{\theta_1}} q(y_i(1), \mathbf{x}_i, \boldsymbol{\theta_1^0})' \boldsymbol{\nabla}_{\boldsymbol{\theta_1}} q(y_i(1), \mathbf{x}_i, \boldsymbol{\theta_1^0}) | \mathbf{x}_i, w_i, s_i\right)\right] \\ &= \mathbb{E}\left[r(\mathbf{x}_i, w_i) \cdot p(\mathbf{x}_i) \cdot \mathbb{E}\left(\boldsymbol{\nabla}_{\boldsymbol{\theta_1}} q(y_i(1), \mathbf{x}_i, \boldsymbol{\theta_1^0})' \boldsymbol{\nabla}_{\boldsymbol{\theta_1}} q(y_i(1), \mathbf{x}_i, \boldsymbol{\theta_1^0}) | \mathbf{x}_i\right)\right] \\ &= \sigma^2_{01} \cdot \mathbb{E}\left[r(\mathbf{x}_i, w_i) \cdot p(\mathbf{x}_i) \cdot \mathbf{A}(\mathbf{x}_i, \boldsymbol{\theta_1^0})\right]\end{aligned}$$

Therefore, the asymptotic variance simplifies to simply

$$\text{Avar}\sqrt{N}\left(\hat{\boldsymbol{\theta}}^{\boldsymbol{u}}_{\mathbf{1}} - \boldsymbol{\theta_1^0}\right) = \sigma^2_{01} \cdot \left(\mathbb{E}\left[r(\mathbf{x}_i, w_i) \cdot p(\mathbf{x}_i) \cdot \mathbf{A}(\mathbf{x}_i, \boldsymbol{\theta_1^0})\right]\right)^{-1}$$

For showing that the two variances are positive semi-definite consider the following

$$\begin{aligned}&\left(\text{Avar}\sqrt{N}\left(\hat{\boldsymbol{\theta}}^{\boldsymbol{u}}_{\mathbf{1}} - \boldsymbol{\theta_1^0}\right)\right)^{-1} - \left(\text{Avar}\sqrt{N}\left(\hat{\boldsymbol{\theta}}_{\mathbf{1}} - \boldsymbol{\theta_1^0}\right)\right)^{-1} \\ =&\frac{1}{\sigma^2_{01}} \cdot \left\{\mathbb{E}\left(r_i \cdot p_i \cdot \mathbf{A}_i\right) - \mathbb{E}\left(\frac{r_i \cdot p_i}{R_i \cdot G_i} \cdot \mathbf{A}_i\right) \cdot \mathbb{E}\left(\frac{r_i \cdot p_i}{R_i^2 \cdot G_i^2} \cdot \mathbf{A}_i\right)^{-1} \cdot \mathbb{E}\left(\frac{r_i \cdot p_i}{R_i \cdot G_i} \cdot \mathbf{A}_i\right)\right\} \\ &\text{Let } \mathbf{B}_i = r_i^{1/2} \cdot p_i^{1/2} \cdot \mathbf{A}_i^{1/2} \text{ and } \mathbf{D}_i = \left(r_i^{1/2}/R_i\right) \cdot \left(p_i^{1/2}/G_i\right) \cdot \mathbf{A}_i^{1/2} \\ =&\frac{1}{\sigma^2_{01}} \left\{\mathbb{E}\left(\mathbf{B}_i'\mathbf{B}_i\right) - \mathbb{E}\left(\mathbf{B}_i'\mathbf{D}_i\right) \cdot \mathbb{E}\left(\mathbf{D}_i'\mathbf{D}_i\right)^{-1} \cdot \mathbb{E}\left(\mathbf{D}_i'\mathbf{B}_i\right)\right\}\end{aligned}$$

where the quantity inside the brackets is nothing but the variance of the residuals from the population regression of $\mathbf{B}_i$ on $\mathbf{D}_i$. Hence, the difference is positive semi-definite. The results for the control group can be proven analogously. ■

Proof of theorem F.2.1

Proof. Consider a constant vector $\boldsymbol{a}$. Then by the conclusion of theorem 3.4.2, we know that $\left\|\hat{\boldsymbol{\theta}}_{\boldsymbol{g}} + \boldsymbol{a}\varepsilon_N - \boldsymbol{\theta}_{\boldsymbol{g}}^{\boldsymbol{0}}\right\| = O_p(\varepsilon_N)$. Consider,

$$|Q_N(\hat{\boldsymbol{\theta}}_{\boldsymbol{g}} + \boldsymbol{a}\varepsilon_N) - Q_N(\boldsymbol{\theta}_{\boldsymbol{g}}^{\boldsymbol{0}}) - Q_0(\hat{\boldsymbol{\theta}}_{\boldsymbol{g}} + \boldsymbol{a}\varepsilon_N) + Q_0(\boldsymbol{\theta}_{\boldsymbol{g}}^{\boldsymbol{0}})| \tag{J.7}$$

Then we know that,

$$R_N(\boldsymbol{\theta}_{\boldsymbol{g}} + \boldsymbol{a}\varepsilon_N) + \boldsymbol{\nabla}_{\boldsymbol{\theta}_{\boldsymbol{g}}} Q_N(\boldsymbol{\theta}_{\boldsymbol{g}}^{\boldsymbol{0}})'(\boldsymbol{\theta}_{\boldsymbol{g}} + \boldsymbol{a}\varepsilon_N - \boldsymbol{\theta}_{\boldsymbol{g}}^{\boldsymbol{0}}) = \tag{J.8}$$

$$Q_N(\boldsymbol{\theta}_{\boldsymbol{g}} + \boldsymbol{a}\varepsilon_N) - Q_N(\boldsymbol{\theta}_{\boldsymbol{g}}^{\boldsymbol{0}}) - Q_0(\boldsymbol{\theta}_{\boldsymbol{g}} + \boldsymbol{a}\varepsilon_N) + Q_0(\boldsymbol{\theta}_{\boldsymbol{g}}^{\boldsymbol{0}}) \tag{J.9}$$

Using eq(J.7) with eq(J.8) we obtain,

$$\begin{aligned}
&|Q_N(\hat{\boldsymbol{\theta}}_{\boldsymbol{g}} + \boldsymbol{a}\varepsilon_N) - Q_N(\boldsymbol{\theta}_{\boldsymbol{g}}^{\boldsymbol{0}}) - Q_0(\hat{\boldsymbol{\theta}}_{\boldsymbol{g}} + \boldsymbol{a}\varepsilon_N) + Q_0(\boldsymbol{\theta}_{\boldsymbol{g}}^{\boldsymbol{0}})| \\
&= |R_N(\hat{\boldsymbol{\theta}}_{\boldsymbol{g}} + \boldsymbol{a}\varepsilon_N) + \boldsymbol{\nabla}_{\boldsymbol{\theta}_{\boldsymbol{g}}} Q_N(\boldsymbol{\theta}_{\boldsymbol{g}}^{\boldsymbol{0}})'(\hat{\boldsymbol{\theta}}_{\boldsymbol{g}} + \boldsymbol{a}\varepsilon_N - \boldsymbol{\theta}_{\boldsymbol{g}}^{\boldsymbol{0}})|
\end{aligned}$$

Then using Triangle and Cauchy-Schwartz inequality,

$$\begin{aligned}
&|Q_N(\hat{\boldsymbol{\theta}}_{\boldsymbol{g}} + \boldsymbol{a}\varepsilon_N) - Q_N(\boldsymbol{\theta}_{\boldsymbol{g}}^{\boldsymbol{0}}) - Q_0(\hat{\boldsymbol{\theta}}_{\boldsymbol{g}} + \boldsymbol{a}\varepsilon_N) + Q_0(\boldsymbol{\theta}_{\boldsymbol{g}}^{\boldsymbol{0}})| \\
&\leq |R_N(\hat{\boldsymbol{\theta}}_{\boldsymbol{g}} + \boldsymbol{a}\varepsilon_N)| + \left\|\boldsymbol{\nabla}_{\boldsymbol{\theta}_{\boldsymbol{g}}} Q_N(\boldsymbol{\theta}_{\boldsymbol{g}}^{\boldsymbol{0}})'\right\| \left\|\hat{\boldsymbol{\theta}}_{\boldsymbol{g}} + \boldsymbol{a}\varepsilon_N - \boldsymbol{\theta}_{\boldsymbol{g}}^{\boldsymbol{0}}\right\|
\end{aligned}$$

Now, using stochastic equicontinuity condition,

$$R_N(\boldsymbol{\theta}_{\boldsymbol{g}}) \leq \left\|\boldsymbol{\theta}_{\boldsymbol{g}} + \boldsymbol{a}\varepsilon_N - \boldsymbol{\theta}_{\boldsymbol{g}}^{\boldsymbol{0}}\right\| \left(1 + \sqrt{N}\left\|\boldsymbol{\theta}_{\boldsymbol{g}} + \boldsymbol{a}\varepsilon_N - \boldsymbol{\theta}_{\boldsymbol{g}}^{\boldsymbol{0}}\right\|\right) o_p(1/\sqrt{N})$$

Then,

$$\begin{aligned}
&|Q_N(\hat{\boldsymbol{\theta}}_{\boldsymbol{g}} + \boldsymbol{a}\varepsilon_N) - Q_N(\boldsymbol{\theta}_{\boldsymbol{g}}^{\boldsymbol{0}}) - Q_0(\hat{\boldsymbol{\theta}}_{\boldsymbol{g}} + \boldsymbol{a}\varepsilon_N) + Q_0(\boldsymbol{\theta}_{\boldsymbol{g}}^{\boldsymbol{0}})| \\
&\leq \left\|\boldsymbol{\theta}_{\boldsymbol{g}} + \boldsymbol{a}\varepsilon_N - \boldsymbol{\theta}_{\boldsymbol{g}}^{\boldsymbol{0}}\right\| \left(1 + \sqrt{N}\left\|\boldsymbol{\theta}_{\boldsymbol{g}} + \boldsymbol{a}\varepsilon_N - \boldsymbol{\theta}_{\boldsymbol{g}}^{\boldsymbol{0}}\right\|\right) o_p(1/\sqrt{N}) \\
&+ O_p(N^{-1/2}) \cdot O_p(\varepsilon_N) \\
&= O_p(\varepsilon_N)\left[1 + \sqrt{N} O_p(\varepsilon_N)\right] o_p(1/\sqrt{N}) + O_p(\varepsilon_N/\sqrt{N}) \\
&= o_p(\varepsilon_N^2)
\end{aligned}$$

Hence,

$$\frac{|Q_N(\hat{\boldsymbol{\theta}}_{\boldsymbol{g}} + \boldsymbol{a}\varepsilon_N) - Q_N(\boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}}) - (Q_0(\hat{\boldsymbol{\theta}}_{\boldsymbol{g}} + \boldsymbol{a}\varepsilon_N) - Q_0(\boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}}))|}{\varepsilon_N^2} = o_p(1) \tag{J.10}$$

Since, $Q_0(\boldsymbol{\theta}_{\boldsymbol{g}})$ is twice differentiable in $\boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}}$,

$$\begin{aligned}
&\left| \frac{\left[Q_0(\hat{\boldsymbol{\theta}}_{\boldsymbol{g}} + \boldsymbol{a}\varepsilon_N) - Q_0(\boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}})\right]}{\varepsilon_N^2} - \frac{\boldsymbol{a}'\mathbf{H}_{\mathbf{g}}\boldsymbol{a}}{2} \right| \\
&= \left| \frac{1}{\varepsilon_N^2} \left[\frac{(\hat{\boldsymbol{\theta}}_{\boldsymbol{g}} + \varepsilon_N\boldsymbol{a} - \boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}})'\mathbf{H}_{\mathbf{g}}(\hat{\boldsymbol{\theta}}_{\boldsymbol{g}} + \varepsilon_N\boldsymbol{a} - \boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}})}{2} + o\left(\left\| \hat{\boldsymbol{\theta}}_{\boldsymbol{g}} + \varepsilon_N\boldsymbol{a} - \boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}} \right\|^2 \right) \right] - \frac{\boldsymbol{a}'\mathbf{H}_{\mathbf{g}}\boldsymbol{a}}{2} \right| \\
&\leq \left| \frac{1}{\varepsilon_N}(\hat{\boldsymbol{\theta}}_{\boldsymbol{g}} - \boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}})'\mathbf{H}_{\mathbf{g}}\boldsymbol{a} \right| + \left| \frac{1}{\varepsilon_N^2}(\hat{\boldsymbol{\theta}}_{\boldsymbol{g}} - \boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}})'\mathbf{H}_{\mathbf{g}}(\hat{\boldsymbol{\theta}}_{\boldsymbol{g}} - \boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}}) \right| + o_p(1) = o_p(1)
\end{aligned} \tag{J.11}$$

Then using J.10, J.11 and triangle inequality,

$$\begin{aligned}
&\left| \frac{1}{\varepsilon_N^2} \left[Q_N(\hat{\boldsymbol{\theta}}_{\boldsymbol{g}} + \boldsymbol{a}\varepsilon_N) - Q_N(\boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}}) \right] - \frac{\boldsymbol{a}'\mathbf{H}_{\mathbf{g}}\boldsymbol{a}}{2} \right| \\
\leq &\left| \frac{1}{\varepsilon_N^2} \left[Q_N(\hat{\boldsymbol{\theta}}_{\boldsymbol{g}} + \boldsymbol{a}\varepsilon_N) - Q_N(\boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}}) - Q_0(\hat{\boldsymbol{\theta}}_{\boldsymbol{g}} + \boldsymbol{a}\varepsilon_N) + Q_0(\boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}}) \right] \right| \\
&+ \left| \frac{1}{\varepsilon_N^2} \left[Q_0(\hat{\boldsymbol{\theta}}_{\boldsymbol{g}} + \boldsymbol{a}\varepsilon_N) - Q_0(\boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}}) \right] - \frac{\boldsymbol{a}'\mathbf{H}_{\mathbf{g}}\boldsymbol{a}}{2} \right| \\
\leq &o_p(1) + o_p(1) = o_p(1)
\end{aligned}$$

It follows that,

$$\begin{aligned}
&\hat{\mathbf{H}}_{\boldsymbol{g}jk} \\
&= \left[\frac{Q_N(\hat{\boldsymbol{\theta}}_{\boldsymbol{g}} + \boldsymbol{e}_j\varepsilon_N + \boldsymbol{e}_k\varepsilon_N) - Q_N(\hat{\boldsymbol{\theta}}_{\boldsymbol{g}} - \boldsymbol{e}_j\varepsilon_N + \boldsymbol{e}_k\varepsilon_N) - Q_N(\hat{\boldsymbol{\theta}}_{\boldsymbol{g}} + \boldsymbol{e}_j\varepsilon_N - \boldsymbol{e}_k\varepsilon_N)}{4\varepsilon_N^2} \right] \\
&+ \left[\frac{Q_N(\hat{\boldsymbol{\theta}}_{\boldsymbol{g}} - \boldsymbol{e}_j\varepsilon_N - \boldsymbol{e}_k\varepsilon_N)}{4\varepsilon_N^2} \right] \\
&\xrightarrow{p} \left[2(e_j + e_k)'\mathbf{H}_{\boldsymbol{g}jk}(e_j + e_k) - (e_j - e_k)'\mathbf{H}_{\boldsymbol{g}jk}(e_j - e_k) - (e_k - e_j)'\mathbf{H}_{\boldsymbol{g}jk}(e_k - e_j) \right] /8 \\
&= 2\left[e_j'\mathbf{H}_{\boldsymbol{g}jk}e_j + e_k'\mathbf{H}_{\boldsymbol{g}jk}e_k - e_i'\mathbf{H}_{\boldsymbol{g}jk}e_i - e_k'\mathbf{H}_{\boldsymbol{g}jk}e_k \right] /8 + e_j'\mathbf{H}_{\boldsymbol{g}jk}e_k \\
&= e_j'\mathbf{H}_{\boldsymbol{g}jk}e_k = \mathbf{H}_{\boldsymbol{g}jk}
\end{aligned}$$

■

Pooled slopes

Proof. Let us assume that $m(\mathbf{x}, \boldsymbol{\theta}) = h(\alpha + \mathbf{x}\boldsymbol{\beta} + \eta w_1)$ is the chosen mean function for $\mu(\mathbf{x})$. Then, in the presence of non-random sampling, we have the following first order conditions

$$\sum_{i=1}^{N} s_i \cdot \left(\frac{w_{i1}}{\hat{R} \cdot \hat{G}} + \frac{w_{i0}}{\hat{R} \cdot (1 - \hat{G})} \right) \cdot \left[y_i - h(\hat{\alpha} + \mathbf{x}_i \hat{\boldsymbol{\beta}} + \hat{\eta} w_{i1}) \right] = 0$$

$$\sum_{i=1}^{N} \frac{s_i \cdot w_{i1}}{\hat{R} \cdot \hat{G}} \cdot \left[y_i - h(\hat{\alpha} + \mathbf{x}_i \hat{\boldsymbol{\beta}} + \hat{\eta} w_{i1}) \right] = 0$$

$$\sum_{i=1}^{N} s_i \cdot \left(\frac{w_{i1}}{\hat{R} \cdot \hat{G}} + \frac{w_{i0}}{\hat{R} \cdot (1 - \hat{G})} \right) \cdot \mathbf{x}_i' \left[y_i - h(\hat{\alpha} + \mathbf{x}_i \hat{\boldsymbol{\beta}} + \hat{\eta} w_{i1}) \right] = 0$$

where $\hat{R} = R(\mathbf{x}, w, \hat{\boldsymbol{\delta}})$ and $\hat{G} = G(\mathbf{x}, \hat{\boldsymbol{\gamma}})$. Ignoring the set of conditions corresponding to the slope parameter $\boldsymbol{\beta}$, the population counterparts to the above FOC are

$$\mathbb{E}\left[s \cdot \left(\frac{w_1}{R \cdot G} + \frac{w_0}{R \cdot (1 - G)} \right) \cdot \left[y - h(\alpha^* + \mathbf{x}\boldsymbol{\beta}^* + \eta^* w_1) \right] \right] = 0 \tag{J.12}$$

$$\mathbb{E}\left[\frac{s \cdot w_1}{R \cdot G} \cdot \left[y - h(\alpha^* + \mathbf{x}\boldsymbol{\beta}^* + \eta^* w_1) \right] \right] = 0 \tag{J.13}$$

where α^*, $\boldsymbol{\beta}^*$ and γ^* are the probability limits of QMLE estimators $\hat{\alpha}$, $\hat{\boldsymbol{\beta}}$ and $\hat{\gamma}$.

Rearranging J.12 and J.13 gives us

$$\mathbb{E}\left[s \cdot \left(\frac{w_1}{R \cdot G} + \frac{w_0}{R \cdot (1 - G)} \right) \cdot y \right] = \mathbb{E}\left[s \cdot \left(\frac{w_1}{R \cdot G} + \frac{w_0}{R \cdot (1 - G)} \right) \cdot h(\alpha^* + \mathbf{x}\boldsymbol{\beta}^* + \eta^* w_1) \right] \tag{J.14}$$

$$\mathbb{E}\left[\frac{s \cdot w_1}{R \cdot G} \cdot y \right] = \mathbb{E}\left[\frac{s \cdot w_1}{R \cdot G} \cdot h(\alpha^* + \mathbf{x}\boldsymbol{\beta}^* + \eta^* w_1) \right] \tag{J.15}$$

Now, $y = y(0) \cdot w_0 + y(1) \cdot w_1$ which implies that we can replace y in the above two equations to obtain the LHs of J.14 equal to

$$\mathbb{E}\left[s \cdot \left(\frac{w_1 \cdot y(1)}{R \cdot G} + \frac{w_0 \cdot y(0)}{R \cdot (1 - G)} \right) \right]$$

By using iterated expectations we can rewrite the above equation as

$$\mathbb{E}\left[\frac{w_1}{G \cdot R} \cdot \mathbb{E}(s \cdot y(1)|\mathbf{x}, w_1) + \frac{w_0}{(1-G) \cdot R} \cdot \mathbb{E}(s \cdot y(0)|\mathbf{x}, w_0)\right]$$

Due to ignorability of sample selection, we can split the conditional expectation into parts.

$$\mathbb{E}\left[\frac{w_1}{G \cdot R} \cdot \mathbb{E}(s|\mathbf{x}, w_1) \cdot \mathbb{E}(y(1)|\mathbf{x}, w_1) + \frac{w_0}{(1-G) \cdot R} \cdot \mathbb{E}(s|\mathbf{x}, w_0) \cdot \mathbb{E}(y(0)|\mathbf{x}, w_0)\right]$$

Note that, $w_1 \cdot \mathbb{E}(s|\mathbf{x}, w_1) = w_1 \cdot R$. similarly, $w_0 \cdot \mathbb{E}(s|\mathbf{x}, w_0) = w_0 \cdot R$ and due to unconfoundedness we have, $\mathbb{E}\left(y(1)|\mathbf{x}, w_1\right) = \mathbb{E}\left(y(1)|\mathbf{x}\right)$ and $\mathbb{E}\left(y(0)|\mathbf{x}, w_0\right) = \mathbb{E}\left(y(0)|\mathbf{x}\right)$. Therefore, we can simplify the above expression into

$$\mathbb{E}\left[\frac{w_1 \cdot R}{G \cdot R} \cdot \mathbb{E}(y(1)|\mathbf{x}) + \frac{w_0 \cdot R}{(1-G) \cdot R} \cdot \mathbb{E}(y(0)|\mathbf{x})\right]$$

Another application of iterated expectation gives us

$$\begin{aligned}
&\mathbb{E}\left[\frac{\mu_1(\mathbf{x})}{G} \cdot \mathbb{E}(w_1|\mathbf{x}) + \frac{\mu_0(\mathbf{x})}{(1-G)} \cdot \mathbb{E}(w_0|\mathbf{x})\right] \\
&= \mathbb{E}\left[\mu_1(\mathbf{x}) + \mu_0(\mathbf{x})\right] \\
&= \mathbb{E}[y(1)] + \mathbb{E}[y(0)]
\end{aligned}$$

Manipulating the RHS of J.14 using iterated expectations gives us

$$\begin{aligned}
&\mathbb{E}\left[h(\alpha^* + \mathbf{x}\boldsymbol{\beta}^* + \eta^* w_1) \cdot \left\{\frac{w_1}{G} \cdot \frac{1}{R} \cdot \mathbb{E}(s|\mathbf{x}, w_1) + \frac{w_0}{(1-G)} \cdot \frac{1}{R} \cdot \mathbb{E}(s|\mathbf{x}, w_0)\right\}\right] \\
&= \mathbb{E}\left[h(\alpha^* + \mathbf{x}\boldsymbol{\beta}^* + \eta^* w_1) \cdot \left\{\frac{w_1}{G} + \frac{w_0}{(1-G)}\right\}\right] \\
&= \mathbb{E}\left[h(\alpha^* + \mathbf{x}\boldsymbol{\beta}^* + \eta^* w_1) \cdot \frac{w_1}{G}\right] + \mathbb{E}\left[h(\alpha^* + \mathbf{x}\boldsymbol{\beta}^* + \eta^* w_1) \cdot \frac{w_0}{(1-G)}\right]
\end{aligned}$$

Therefore, combining the LHS and RHS give the result

$$\mathbb{E}[y(1)] + \mathbb{E}[y(0)] = \mathbb{E}\left[h(\alpha^* + \mathbf{x}\boldsymbol{\beta}^* + \eta^* w_1) \cdot \frac{w_1}{G}\right] + \mathbb{E}\left[h(\alpha^* + \mathbf{x}\boldsymbol{\beta}^* + \eta^* w_1) \cdot \frac{w_0}{(1-G)}\right] \tag{J.16}$$

Now, consider the LHS of J.15.

$$\begin{aligned}\mathbb{E}\left[\frac{s \cdot w_1}{R \cdot G} \cdot y\right] &= \mathbb{E}\left[\frac{s \cdot w_1}{R \cdot G} \cdot y(1)\right] \\ &= \mathbb{E}[y(1)] \qquad \text{by LIE}\end{aligned}$$

Similarly using LIE, the RHS of J.15 can be re-written as

$$\begin{aligned}\mathbb{E}\left[\frac{s \cdot w_1}{R \cdot G} \cdot h(\alpha^* + \mathbf{x}\boldsymbol{\beta}^* + \eta^* w_1)\right] &= \mathbb{E}\left[h(\alpha^* + \mathbf{x}\boldsymbol{\beta}^* + \eta^* w_1) \cdot \frac{w_1}{G} \cdot \frac{1}{R} \cdot \mathbb{E}(s|\mathbf{x}, w_1)\right] \\ &= \mathbb{E}\left[h(\alpha^* + \mathbf{x}\boldsymbol{\beta}^* + \eta^* w_1) \cdot \frac{w_1}{G}\right]\end{aligned}$$

Therefore combining the LHS and RHS give us

$$\mathbb{E}[y(1)] = \mathbb{E}\left[h(\alpha^* + \mathbf{x}\boldsymbol{\beta}^* + \eta^* w_1) \cdot \frac{w_1}{G}\right] \tag{J.17}$$

Then using J.17 along with J.16 implies that

$$\mathbb{E}[y(0)] = \mathbb{E}\left[h(\alpha^* + \mathbf{x}\boldsymbol{\beta}^* + \eta^* w_1) \cdot \frac{w_0}{(1-G)}\right] \tag{J.18}$$

Consider

$$\begin{aligned}&\mathbb{E}\left[h(\alpha^* + \mathbf{x}\boldsymbol{\beta}^* + \eta^* w_1) \cdot w_1|\mathbf{x}\right] \\ &= \mathbb{E}\left[h(\alpha^* + \mathbf{x}\boldsymbol{\beta}^* + \eta^*)\right] \cdot P(w_1 = 1|\mathbf{x})\end{aligned}$$

Therefore, $\mathbb{E}\left[h(\alpha^* + \mathbf{x}\boldsymbol{\beta}^* + \eta^* w_1) \cdot \frac{w_1}{G}\right] = \mathbb{E}\left[h(\alpha^* + \mathbf{x}\boldsymbol{\beta}^* + \eta^*)\right]$
Similarly, we can also show that $\mathbb{E}\left[h(\alpha^* + \mathbf{x}\boldsymbol{\beta}^* + \eta^* w_1) \cdot \frac{w_0}{(1-G)}\right] = \mathbb{E}\left[h(\alpha^* + \mathbf{x}\boldsymbol{\beta}^*)\right]$
Hence, the pooled regression adjustment estimator can be written as

$$\tau_{PRA} = \mathbb{E}\left[h(\alpha^* + \mathbf{x}\boldsymbol{\beta}^* + \eta^*)\right] - \mathbb{E}\left[h(\alpha^* + \mathbf{x}\boldsymbol{\beta}^*)\right]$$

so a consistent estimator of the QMLE pooled regression adjustment estimator can be obtained by replacing the population expectation by the sample average in the above expression and weighting by the appropriate probabilities to recover the balance of the random sample which gives us

$$\hat{\tau}_{PRA} = \frac{1}{N}\sum_{i=1}^{N}\left[h(\hat{\alpha} + \mathbf{x}_i\hat{\boldsymbol{\beta}} + \hat{\eta})\right] - \frac{1}{N}\sum_{i=1}^{N}\left[h(\hat{\alpha} + \mathbf{x}_i\hat{\boldsymbol{\beta}})\right]$$

■

Separate slopes

Proof. Let us assume that $m_g(\mathbf{x}, \boldsymbol{\theta_g}) = h(\alpha_g + \mathbf{x}\boldsymbol{\beta_g})$ is the chosen mean function for $\mu_g(\mathbf{x})$. Then the population FOC's are

$$\mathbb{E}\left[\frac{s \cdot w_1}{R \cdot G} \cdot \left[y - h(\alpha_1^* + \mathbf{x}\boldsymbol{\beta_1^*})\right]\right] = 0 \tag{J.19}$$

$$\mathbb{E}\left[\frac{s \cdot w_0}{R \cdot (1-G)} \cdot \left[y - h(\alpha_0^* + \mathbf{x}\boldsymbol{\beta_0^*})\right]\right] = 0 \tag{J.20}$$

where where α_g^*, $\boldsymbol{\beta}_g^*$ are the probability limits of QMLE estimators $\hat{\alpha}_g$, $\hat{\boldsymbol{\beta}}_g$. Rearranging J.19 and J.20 just like in the pooled case gives us the following equalities.

$$\mathbb{E}\left[\frac{s \cdot w_1}{R \cdot G} \cdot y\right] = \mathbb{E}\left[\frac{s \cdot w_1}{R \cdot G} \cdot h(\alpha_1^* + \mathbf{x}\boldsymbol{\beta_1^*})\right] \tag{J.21}$$

$$\mathbb{E}\left[\frac{s \cdot w_0}{R \cdot (1-G)} \cdot y\right] = \mathbb{E}\left[\frac{s \cdot w_0}{R \cdot (1-G)} \cdot h(\alpha_0^* + \mathbf{x}\boldsymbol{\beta_0^*})\right] \tag{J.22}$$

Proceeding with the above two equations in the same way as in the pooled case gives us the results

$$\mathbb{E}[y(1)] = \mathbb{E}\left[h(\alpha_1^* + \mathbf{x}\boldsymbol{\beta_1^*})\right]$$

$$\mathbb{E}[y(0)] = \mathbb{E}\left[h(\alpha_0^* + \mathbf{x}\boldsymbol{\beta_0^*})\right]$$

Therefore, $\tau_{FRA} = \mathbb{E}\left[h(\alpha_1^* + \mathbf{x}\boldsymbol{\beta}_1^*)\right] - \mathbb{E}\left[h(\alpha_0^* + \mathbf{x}\boldsymbol{\beta}_0^*)\right]$ and so a consistent estimator of the QMLE separate regression adjustment estimator can be obtained as

$$\hat{\tau}_{FRA} = \frac{1}{N}\sum_{i=1}^{N}\left[h(\hat{\alpha}_1 + \mathbf{x}_i\hat{\boldsymbol{\beta}}_1)\right] - \frac{1}{N}\sum_{i=1}^{N}\left[h(\hat{\alpha}_0 + \mathbf{x}_i\hat{\boldsymbol{\beta}}_0)\right]$$

Consistency of $\hat{\tau}_{PRA}$ for τ_{PRA} and $\hat{\tau}_{FRA}$ for τ_{FRA} follows from the results on double weighting and generalized linear model properties. Remember that the framework of this paper does not rely on the correct specification of some conditional mean of the distribution. I have allowed for both; when the mean function is correctly specified but everything else about the distribution is misspecified as well as when everything is allowed to be misspecified including the mean. In both cases, results from quasi maximum likelihood in the linear exponential family have been instrumental in guaranteeing consistency of pooled and separate slopes methods. ■

Asymptotic variance expression for ATE: Correct CEF

Proof. Assuming continuous differentiability of $m_g(\mathbf{x}_i, \boldsymbol{\theta_g})$ on $\boldsymbol{\Theta_g}$, mean value expansion around $\boldsymbol{\theta_g^0}$ gives

$$\frac{1}{N}\sum_{i=1}^{N} m_g(\mathbf{x}_i, \hat{\boldsymbol{\theta}}_{\boldsymbol{g}}) = \frac{1}{N}\sum_{i=1}^{N} m_g(\mathbf{x}_i, \boldsymbol{\theta_g^0}) + \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{\nabla}_{\boldsymbol{\theta_g}} m_g(\mathbf{x}_i, \ddot{\boldsymbol{\theta}}_{\boldsymbol{g}}) \cdot (\hat{\boldsymbol{\theta}}_{\boldsymbol{g}} - \boldsymbol{\theta_g^0}) + \textit{Remainder}$$

where $\ddot{\boldsymbol{\theta}}_{\boldsymbol{g}}$ lies between $\hat{\boldsymbol{\theta}}_{\boldsymbol{g}}$ and $\boldsymbol{\theta_g^0}$. Since $\hat{\boldsymbol{\theta}}_{\boldsymbol{g}} \xrightarrow{p} \boldsymbol{\theta_g^0}$, so does $\ddot{\boldsymbol{\theta}}_{\boldsymbol{g}}$. Hence, using the weak law of large numbers, we obtain

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{N} m_g(\mathbf{x}_i, \hat{\boldsymbol{\theta}}_{\boldsymbol{g}}) = \frac{1}{\sqrt{N}}\sum_{i=1}^{N} m_g(\mathbf{x}_i, \boldsymbol{\theta_g^0}) + \mathbb{E}\left[\boldsymbol{\nabla}_{\boldsymbol{\theta_g}} m_1(\mathbf{x}_i, \boldsymbol{\theta_g^0})\right] \cdot \sqrt{N}(\hat{\boldsymbol{\theta}}_{\boldsymbol{g}} - \boldsymbol{\theta_g^0}) + o_p(1)$$

Adding and subtracting $\sqrt{N} \cdot \mathbb{E}\left(m_g(\mathbf{x}_i, \boldsymbol{\theta_g^0})\right)$ on both sides gives us

$$\begin{aligned}\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\{m_g(\mathbf{x}_i, \hat{\boldsymbol{\theta}}_{\boldsymbol{g}}) - \mathbb{E}\left(m_g(\mathbf{x}_i, \boldsymbol{\theta_g^0})\right)\} = \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\{m_g(\mathbf{x}_i, \boldsymbol{\theta_g^0}) - \mathbb{E}\left(m_g(\mathbf{x}_i, \boldsymbol{\theta_g^0})\right)\} \\ + \mathbb{E}\left[\boldsymbol{\nabla}_{\boldsymbol{\theta_g}} m_g(\mathbf{x}_i, \boldsymbol{\theta_g^0})\right] \cdot \sqrt{N}(\hat{\boldsymbol{\theta}}_{\boldsymbol{g}} - \boldsymbol{\theta_g^0}) + o_p(1)\end{aligned}$$

Let $\mathbb{E}\left[\boldsymbol{\nabla}_{\boldsymbol{\theta_g}} m_g(\mathbf{x}_i, \boldsymbol{\theta_g^0})\right] \equiv \mathbf{G_g^0}$. Then, using the asymptotic results from section 3.5, where we posit that the conditional feature of interest is correctly specified, we have

$$\sqrt{N}\left(\hat{\boldsymbol{\theta}}_{\mathbf{1}} - \boldsymbol{\theta_1^0}\right) = -\mathbf{H_1^{-1}}\left\{\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\mathbf{l}_i\right\} + o_p(1)$$

$$\sqrt{N}\left(\hat{\boldsymbol{\theta}}_{\mathbf{0}} - \boldsymbol{\theta_0^0}\right) = -\mathbf{H_0^{-1}}\left\{\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\mathbf{k}_i\right\} + o_p(1)$$

Therefore,

$$\begin{aligned}&\sqrt{N}\left(\hat{\tau}_{\text{ate}} - \tau_{\text{ate}}\right) \\ =&\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left(\{m_1(\mathbf{x}_i, \boldsymbol{\theta_1^0}) - m_0(\mathbf{x}_i, \boldsymbol{\theta_0^0}) - \tau_{\text{ate}}\} - \mathbf{G_1^0}\cdot\mathbf{H_1^{-1}}\mathbf{l}_i + \mathbf{G_0^0}\cdot\mathbf{H_0^{-1}}\mathbf{k}_i\right) + o_p(1)\end{aligned}$$

We may rewrite the above using the influence function representation as

$$\sqrt{N}\left(\hat{\tau}_{\text{ate}} - \tau_{\text{ate}}\right) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \psi(\mathbf{x}_i) + o_p(1) \text{ where } \mathbb{E}\left[\psi(\mathbf{x}_i)\right] = 0$$

Then, provided that $\mathbb{E}\left[\psi(\mathbf{x}_i)\psi(\mathbf{x}_i)'\right]$ exists,

$$\begin{aligned} Avar\left[\sqrt{N}\left(\hat{\tau}_{\text{ate}} - \tau_{\text{ate}}\right)\right] = \mathbb{E}\left[(m_1(\mathbf{x}_i, \boldsymbol{\theta}_{\mathbf{1}}^{\mathbf{0}}) - m_0(\mathbf{x}_i, \boldsymbol{\theta}_{\mathbf{0}}^{\mathbf{0}})) - \tau_{\text{ate}}\right]^2 + \mathbf{G}_{\mathbf{1}}^{\mathbf{0}} \cdot \mathbf{V}_{\mathbf{1}} \cdot \mathbf{G}_{\mathbf{1}}^{\mathbf{0}\prime} \\ + \mathbf{G}_{\mathbf{0}}^{\mathbf{0}} \cdot \mathbf{V}_{\mathbf{0}} \cdot \mathbf{G}_{\mathbf{0}}^{\mathbf{0}\prime} \end{aligned}$$

Note that the covariance term involving $\mathbf{l}_i$ and $\mathbf{k}_i$ is zero since the two denote scores for the treatment and control group problems. The covariance terms involving $(m_1(\mathbf{x}_i, \boldsymbol{\theta}_{\mathbf{1}}^{\mathbf{0}}) - m_0(\mathbf{x}_i, \boldsymbol{\theta}_{\mathbf{0}}^{\mathbf{0}}) - \tau_{\text{ate}})$ and $\mathbf{l}_i$ and $(m_1(\mathbf{x}_i, \boldsymbol{\theta}_{\mathbf{1}}^{\mathbf{0}}) - m_0(\mathbf{x}_i, \boldsymbol{\theta}_{\mathbf{0}}^{\mathbf{0}}) - \tau_{\text{ate}})$ and $\mathbf{k}_i$ will also be zero. This is because $\boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}}$ solves the conditional problem, which implies that $\mathbb{E}\left[\nabla_{\boldsymbol{\theta}_{\boldsymbol{g}}} q(y_i(g), \mathbf{x}_i, \boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}})' | \mathbf{x}_i\right] = \mathbf{0}$ (i.e. for $g = 1$, $\mathbb{E}\left(\mathbf{l}_i | \mathbf{x}_i\right) = \mathbf{0}$ and for $g = 0$, $\mathbb{E}\left(\mathbf{k}_i | \mathbf{x}_i\right) = \mathbf{0}$). Then, using LIE, those covariance terms can be shown to be zero.

Misspecified mean model In the case of a misspecified mean model, we still have

$$\begin{aligned} &\frac{1}{\sqrt{N}} \sum_{i=1}^{N} \{m_g(\mathbf{x}_i, \hat{\boldsymbol{\theta}}_{\boldsymbol{g}}) - \mathbb{E}\left(m_g(\mathbf{x}_i, \boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}})\right)\} \\ =&\frac{1}{\sqrt{N}} \sum_{i=1}^{N} \{m_g(\mathbf{x}_i, \boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}}) - \mathbb{E}\left(m_g(\mathbf{x}_i, \boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}})\right)\} + \mathbb{E}\left[\boldsymbol{\nabla}_{\boldsymbol{\theta}_{\boldsymbol{g}}} m_g(\mathbf{x}_i, \boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}})\right] \cdot \\ &\sqrt{N}(\hat{\boldsymbol{\theta}}_{\boldsymbol{g}} - \boldsymbol{\theta}_{\boldsymbol{g}}^{\mathbf{0}}) + o_p(1) \end{aligned}$$

However now, using results from section 3.4

$$\begin{aligned}
\sqrt{N}\left(\hat{\boldsymbol{\theta}}_{\mathbf{1}}-\boldsymbol{\theta}_{\mathbf{1}}^{\mathbf{0}}\right) =& -\mathbf{H}_{\mathbf{1}}^{-1}\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left\{\mathbf{l}_i-\mathbb{E}\left(\mathbf{l}_i\mathbf{b}_i'\right)\mathbb{E}\left(\mathbf{b}_i\mathbf{b}_i'\right)^{-1}\mathbf{b}_i-\mathbb{E}(\mathbf{l}_i\mathbf{d}_i')\mathbb{E}(\mathbf{d}_i\mathbf{d}_i')^{-1}\mathbf{d}_i\right\}\\
&+o_p(1)\\
=& -\mathbf{H}_{\mathbf{1}}^{-1}\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\mathbf{u}_{i1}+o_p(1)\\
\sqrt{N}\left(\hat{\boldsymbol{\theta}}_{\mathbf{0}}-\boldsymbol{\theta}_{\mathbf{0}}^{\mathbf{0}}\right) =& -\mathbf{H}_{\mathbf{0}}^{-1}\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left\{\mathbf{k}_i-\mathbb{E}\left(\mathbf{k}_i\mathbf{b}_i'\right)\mathbb{E}\left(\mathbf{b}_i\mathbf{b}_i'\right)^{-1}\mathbf{b}_i-\mathbb{E}(\mathbf{k}_i\mathbf{d}_i')\mathbb{E}(\mathbf{d}_i\mathbf{d}_i')^{-1}\mathbf{d}_i\right\}\\
&+o_p(1)\\
=& -\mathbf{H}_{\mathbf{0}}^{-1}\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\mathbf{u}_{i0}+o_p(1)
\end{aligned}$$

Then,

$$\begin{aligned}
&\sqrt{N}\left(\hat{\tau}_{\text{ate}}-\tau_{\text{ate}}\right)\\
&=\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left(\left\{m_1(\mathbf{x}_i,\boldsymbol{\theta}_{\mathbf{1}}^{\mathbf{0}})-m_0(\mathbf{x}_i,\boldsymbol{\theta}_{\mathbf{0}}^{\mathbf{0}})-\tau_{\text{ate}}\right\}-\mathbf{G}_{\mathbf{1}}^{\mathbf{0}}\cdot\mathbf{H}_{\mathbf{1}}^{-\mathbf{1}}\mathbf{u}_{i\mathbf{1}}+\mathbf{G}_{\mathbf{0}}^{\mathbf{0}}\cdot\mathbf{H}_{\mathbf{0}}^{-\mathbf{1}}\mathbf{u}_{i\mathbf{0}}\right)+o_p(1)\\
&=\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\psi(\mathbf{x}_i)+o_p(1)
\end{aligned}$$

Then,

$$\begin{aligned}
Avar\left[\sqrt{N}\left(\hat{\tau}_{\text{ate}}-\tau_{\text{ate}}\right)\right] =& \mathbb{E}\left[(m_1(\mathbf{x}_i,\boldsymbol{\theta}_{\mathbf{1}}^{\mathbf{0}})-m_0(\mathbf{x}_i,\boldsymbol{\theta}_{\mathbf{0}}^{\mathbf{0}}))-\tau_{\text{ate}}\right]^2+\mathbf{G}_{\mathbf{1}}^{\mathbf{0}}\cdot\mathbf{V}_{\mathbf{1}}\cdot\mathbf{G}_{\mathbf{1}}^{\mathbf{0}\prime}\\
&+\mathbf{G}_{\mathbf{0}}^{\mathbf{0}}\cdot\mathbf{V}_{\mathbf{0}}\cdot\mathbf{G}_{\mathbf{0}}^{\mathbf{0}\prime}\\
&-2\mathbb{E}\left[\left\{m_1(\mathbf{x}_i,\boldsymbol{\theta}_{\mathbf{1}}^{\mathbf{0}})-m_0(\mathbf{x}_i,\boldsymbol{\theta}_{\mathbf{0}}^{\mathbf{0}})-\tau_{\text{ate}}\right\}\mathbf{u}_{i1}'\right]\mathbf{H}_{\mathbf{1}}^{-\mathbf{1}}\mathbf{G}_{\mathbf{1}}^{\mathbf{0}\prime}\\
&+2\mathbb{E}\left[\left\{m_1(\mathbf{x}_i,\boldsymbol{\theta}_{\mathbf{1}}^{\mathbf{0}})-m_0(\mathbf{x}_i,\boldsymbol{\theta}_{\mathbf{0}}^{\mathbf{0}})-\tau_{\text{ate}}\right\}\mathbf{u}_{i0}'\right]\mathbf{H}_{\mathbf{0}}^{-\mathbf{1}}\mathbf{G}_{\mathbf{0}}^{\mathbf{0}\prime}
\end{aligned}$$

■

BIBLIOGRAPHY

BIBLIOGRAPHY

Abadie, A., S. Athey, G. W. Imbens, and J. Wooldridge (2017a): "When Should You Adjust Standard Errors for Clustering?" Tech. rep., National Bureau of Economic Research.

Abadie, A., S. Athey, G. W. Imbens, and J. M. Wooldridge (2017b): "Sampling-based vs. Design-based Uncertainty in Regression Analysis," *Working Paper*.

Andrews, D. W. (1994): "Empirical process methods in econometrics," *Handbook of econometrics*, 4, 2247–2294.

Angrist, J., E. Bettinger, and M. Kremer (2006a): "Long-term educational consequences of secondary school vouchers: Evidence from administrative records in Colombia," *American economic review*, 96, 847–862.

Angrist, J., V. Chernozhukov, and I. Fernández-Val (2006b): "Quantile Regression under Misspecification, with an Application to the U.S. Wage Structure," *Econometrica*, 74, 539–563.

Ansel, J., H. Hong, and J. Li (2018): "OLS and 2SLS in Randomized and Conditionally Randomized Experiments," *Jahrbücher für Nationalökonomie und Statistik*, 238, 243–293.

Ayer, M., H. D. Brunk, G. M. Ewing, W. T. Reid, and E. Silverman (1955): "An empirical distribution function for sampling with incomplete information," *The annals of mathematical statistics*, 641–647.

Ba, B. A., J. C. Ham, R. J. LaLonde, and X. Li (2017): "Estimating (easily interpreted) dynamic training effects from experimental data," *Journal of Labor Economics*, 35, S149–S200.

Bartlett, J. W. (2018): "Covariate adjustment and estimation of mean response in randomised trials," *Pharmaceutical statistics*, 17, 648–666.

Behaghel, L., B. Crépon, M. Gurgand, and T. Le Barbanchon (2015): "Please call again: Correcting nonresponse bias in treatment effect models," *Review of Economics and Statistics*, 97, 1070–1080.

Berk, R., E. Pitkin, L. Brown, A. Buja, E. George, and L. Zhao (2013): "Covariance adjustments for the analysis of randomized field experiments," *Evaluation review*, 37, 170–196.

Bloom, H. S. (1984): "Accounting for no-shows in experimental evaluation designs," *Evaluation review*, 8, 225–246.

Bruhn, M. and D. McKenzie (2009): "In pursuit of balance: Randomization in practice in development field experiments," *American economic journal: applied economics*, 1, 200–232.

Busso, M. and S. Galiani (2019): "The causal effect of competition on prices and quality: Evidence from a field experiment," *American Economic Journal: Applied Economics*, 11, 33–56.

Calónico, S. and J. Smith (2017): "The women of the national supported work demonstration," *Journal of Labor Economics*, 35, S65–S97.

Card, D., P. Ibarrarán, F. Regalia, D. Rosas-Shady, and Y. Soares (2011): "The labor market impacts of youth training in the Dominican Republic," *Journal of Labor Economics*, 29, 267–300.

Carson, R. T., M. B. Conaway, W. M. Hanemann, J. A. Krosnick, R. C. Mitchell, and S. Presser (2004): "Valuing oil spill prevention," .

Chen, X., C. A. Flores, and A. Flores-Lagunes (2018): "Going beyond LATE Bounding Average Treatment Effects of Job Corps Training," *Journal of Human Resources*, 53, 1050–1099.

Chetty, R., J. N. Friedman, and J. E. Rockoff (2014): "Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates," *American Economic Review*, 104, 2593–2632.

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

de Luna, X. and P. Johansson (2014): "Testing for the unconfoundedness assumption using an instrumental assumption," *Journal of Causal Inference*, 2, 187–199.

Drange, N. and T. Havnes (2018): "Early child care and cognitive development: Evidence from an assignment lottery," *Journal of Labor Economics*, 0.

Firpo, S. (2007): "Efficient semiparametric estimation of quantile treatment effects," *Econometrica*, 75, 259–276.

Fisher, R. A. (1935): "The design of experiments. 1935," *Oliver and Boyd, Edinburgh*.

Freedman, D. A. (2008a): "On regression adjustments in experiments with several treatments," *The annals of applied statistics*, 176–196.

——— (2008b): "On Regression adjustments to experimental data," *Advances in Applied Mathematics*, 40, 180–193.

Fricke, H., M. Frölich, M. Huber, and M. Lechner (2015): "Endogeneity and nonresponse bias in treatment evaluation: Nonparametric identification of causal effects by instruments," .

Frölich, M. and M. Huber (2014): "Treatment evaluation with multiple outcome periods under endogeneity and attrition," *Journal of the American Statistical Association*, 109, 1697–1711.

Frumento, P., F. Mealli, B. Pacini, and D. B. Rubin (2012): "Evaluating the effect of training on wages in the presence of noncompliance, nonemployment, and missing outcome data," *Journal of the American Statistical Association*, 107, 450–466.

Gourieroux, C., A. Monfort, and A. Trognon (1984): "Pseudo maximum likelihood methods: Theory," *Econometrica: Journal of the Econometric Society*, 681–700.

Hahn, J. (1998): "On the role of the propensity score in efficient semiparametric estimation of average treatment effects," *Econometrica*, 315–331.

Hausman, J. A. and D. A. Wise (1979): "Attrition bias in experimental and panel data: the Gary income maintenance experiment," *Econometrica: Journal of the Econometric Society*, 455–473.

Heckman, J., H. Ichimura, J. Smith, and P. Todd (1998a): "Characterizing Selection Bias Using Experimental Data," *Econometrica*, 66, 1017–1098.

Heckman, J., J. Smith, and C. Taber (1998b): "Accounting for dropouts in evaluations of social programs," *Review of Economics and Statistics*, 80, 1–14.

Heckman, J. J. and V. J. Hotz (1989): "Choosing among alternative nonexperimental methods for estimating the impact of social programs: The case of manpower training," *Journal of the American statistical Association*, 84, 862–874.

Hirano, K. and G. W. Imbens (2001): "Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization," *Health Services and Outcomes research methodology*, 2, 259–278.

Hirano, K., G. W. Imbens, and G. Ridder (2003): "Efficient estimation of average treatment effects using the estimated propensity score," *Econometrica*, 71, 1161–1189.

Hotz, V. J., G. W. Imbens, and J. A. Klerman (2006): "Evaluating the differential effects of alternative welfare-to-work training components: A reanalysis of the California GAIN program," *Journal of Labor Economics*, 24, 521–566.

Huber, M. (2012): "Identification of average treatment effects in social experiments under alternative forms of attrition," *Journal of Educational and Behavioral Statistics*, 37, 443–474.

——— (2014a): "Identifying causal mechanisms (primarily) based on inverse probability weighting," *Journal of Applied Econometrics*, 29, 920–943.

——— (2014b): "Treatment evaluation in the presence of sample selection," *Econometric Reviews*, 33, 869–905.

Huber, M. and G. Mellace (2015): "Sharp bounds on causal effects under sample selection," *Oxford bulletin of economics and statistics*, 77, 129–151.

Huber, M. and B. Melly (2015): "A test of the conditional independence assumption in sample selection models," *Journal of Applied Econometrics*, 30, 1144–1168.

Imbens, G. W. and D. B. Rubin (2015): *Causal inference in statistics, social, and biomedical sciences*, Cambridge University Press.

Imbens, G. W. and J. M. Wooldridge (2009): "Recent developments in the econometrics of program evaluation," *Journal of economic literature*, 47, 5–86.

Kane, T. J. and D. O. Staiger (2008): "Estimating teacher impacts on student achievement: An experimental evaluation," Tech. rep., National Bureau of Economic Research.

Kim, T.-H. and H. White (2003): "Estimation, inference, and specification testing for possibly misspecified quantile regression," in *Maximum likelihood estimation of misspecified models: twenty years later*, Emerald Group Publishing Limited, 107–132.

Koenker, R. and G. Bassett (1978): "Regression Quantiles," *Econometrica*, 46, 33–50.

Komunjer, I. (2005): "Quasi-maximum likelihood estimation for conditional quantiles," *Journal of Econometrics*, 128, 137 – 164.

LaLonde, R. J. (1986): "Evaluating the econometric evaluations of training programs with experimental data," *The American economic review*, 604–620.

Lewbel, A. (2000): "Semiparametric qualitative response model estimation with unknown heteroscedasticity or instrumental variables," *Journal of Econometrics*, 97, 145–177.

Li, L., C. Shen, X. Li, and J. M. Robins (2013): "On weighting approaches for missing data," *Statistical methods in medical research*, 22, 14–30.

Lin, W. (2013): "Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique," *The Annals of Applied Statistics*, 7, 295–318.

Little, R. J. and D. B. Rubin (2002): "Statistical analysis with missing data: Wiley series in probability and statistics," .

Mattei, A., F. Mealli, and B. Pacini (2014): "Identification of Local Causal Effects with Missing Outcome Values and an Instrument for Non Response," *Communications in Statistics-Theory and Methods*, 43, 815–825.

McCullagh, P. and J. Nelder (1989): *Generalized Linear Models*, London, Chapman and Hall.

Moffit, R., J. Fitzgerald, and P. Gottschalk (1999): "Sample attrition in panel data: The role of selection on observables," *Annales d'Economie et de Statistique*, 129–152.

Mullahy, J. (2015): "Multivariate fractional regression estimation of econometric share models," *Journal of econometric methods*, 4, 71–100.

Muralidharan, K. and V. Sundararaman (2015): "The aggregate effect of school choice: Evidence from a two-stage experiment in India," *The Quarterly Journal of Economics*, 130, 1011–1066.

Negi, A. (2019): "GMM characterization of the Doubly weighted M-estimator," *Working paper*.

Negi, A. and J. M. Wooldridge (2019): "Regression adjustment in experiments with heterogeneous treatment effects," *Working paper*.

Newey, W. K. and D. McFadden (1994): "Large sample estimation and hypothesis testing," *Handbook of econometrics*, 4, 2111–2245.

Neyman, J. (1923): "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9.(Tlanslated and edited by DM Dabrowska and TP Speed, Statistical Science (1990), 5, 465-480)," *Annals of Agricultural Sciences*, 10, 1–51.

Papke, L. E. and J. M. Wooldridge (1996): "Econometric methods for fractional response variables with an application to 401 (k) plan participation rates," *Journal of applied econometrics*, 11, 619–632.

Pollard, D. (1985): "New ways to prove central limit theorems," *Econometric Theory*, 1, 295–313.

Prokhorov, A. and P. Schmidt (2009): "GMM redundancy results for general missing data problems," *Journal of Econometrics*, 151, 47–55.

Robins, J. M. and A. Rotnitzky (1995): "Semiparametric efficiency in multivariate regression models with missing data," *Journal of the American Statistical Association*, 90, 122–129.

Robins, J. M., A. Rotnitzky, and L. P. Zhao (1994): "Estimation of regression coefficients when some regressors are not always observed," *Journal of the American statistical Association*, 89, 846–866.

Rosenbaum, P. R. (1987): "The role of a second control group in an observational study," *Statistical Science*, 2, 292–306.

——— (2002): "Observational studies," in *Observational studies*, Springer, 1–17.

Rosenbaum, P. R. and D. B. Rubin (1983): "The central role of the propensity score in observational studies for causal effects," *Biometrika*, 70, 41–55.

Rosenblum, M. and M. J. Van Der Laan (2010): "Simple, efficient estimators of treatment effects in randomized trials using generalized linear models to leverage baseline variables," *The international journal of biostatistics*, 6.

Shadish, W. R., M. H. Clark, and P. M. Steiner (2008): "Can nonrandomized experiments yield accurate answers? A randomized experiment comparing random and nonrandom assignments," *Journal of the American statistical association*, 103, 1334–1344.

Słoczyński, T. (2018): "A general weighted average representation of the ordinary and two-stage least squares estimands," *arXiv preprint arXiv:1810.01576*.

Słoczyński, T. and J. M. Wooldridge (2018): "A general double robustness result for estimating average treatment effects," *Econometric Theory*, 34, 112–133.

Tsiatis, A. A., M. Davidian, M. Zhang, and X. Lu (2008): "Covariate adjustment for two-sample treatment comparisons in randomized clinical trials: a principled yet flexible approach," *Statistics in medicine*, 27, 4658–4677.

Vossmeyer, A. (2016): "Sample Selection and Treatment Effect Estimation of Lender of Last Resort Policies," *Journal of Business & Economic Statistics*, 34, 197–212.

Watanabe, M. (2010): "Nonparametric estimation of mean willingness to pay from discrete response valuation data," *American Journal of Agricultural Economics*, 92, 1114–1135.

White, H. (1982): "Maximum likelihood estimation of misspecified models," *Econometrica: Journal of the Econometric Society*, 1–25.

Wooldridge, J. M. (2002): "Inverse probability weighted M-estimators for sample selection, attrition, and stratification," *Portuguese Economic Journal*, 1, 117–139.

——— (2007): "Inverse probability weighted estimation for general missing data problems," *Journal of Econometrics*, 141, 1281–1301.

——— (2010): *Econometric analysis of cross section and panel data*, MIT press.

Yang, L. and A. A. Tsiatis (2001): "Efficiency study of estimators for a treatment effect in a pretest–posttest trial," *The American Statistician*, 55, 314–321.