Topics and Applications of Weighting Methods in Case-Control and Observational Studies

by

Fan (Frank) Li

Department of Biostatistics and Bioinformatics
Duke University

Date: _____

Approved:

_____
Andrew S. Allen, Chair

_____
Fan Li, Co-Chair

_____
Kouros Owzar

_____
Elizabeth L. Turner

_____
Elizabeth R. DeLong

Dissertation submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in the Department of Biostatistics and Bioinformatics
in the Graduate School of Duke University
2019

Abstract

Topics and Applications of Weighting Methods in Case-Control and Observational Studies

by

Fan (Frank) Li

Department of Biostatistics and Bioinformatics
Duke University

Date: _____

Approved:

_____
Andrew S. Allen, Chair

_____
Fan Li, Co-Chair

_____
Kouros Owzar

_____
Elizabeth L. Turner

_____
Elizabeth R. DeLong

An abstract of a dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biostatistics and Bioinformatics
in the Graduate School of Duke University
2019

Abstract

Weighting methods have been widely used in statistics and related applications. For example, the inverse probability weighting is a standard approach to correct for survey non-response. The case-control design, frequently seen in epidemiologic or genetic studies, can be regarded as a special type of survey design; analogous inverse probability weighting approaches have been explored when the interest is the association between exposures and the disease (primary analysis) as well as when the interest is the association among exposures (secondary analysis). Meanwhile, in observational comparative effectiveness research, inverse probability weighting has been suggested as a valid approach to correct for confounding bias. This dissertation develops and extends weighting methods for case-control and observational studies.

The first part of this dissertation extends the inverse probability weighting approach for secondary analysis of case-control data. We revisit an inverse probability weighting estimator to offer new insights and extensions. Specifically, we construct its more general form by generalized least squares (GLS). Such a construction allows us to connect the GLS estimator with the generalized method of moments and motivates a new specification test designed to assess the adequacy of the inverse probability weights. The specification test statistic measures the weighted discrepancy between the case and control subsample estimators, and asymptotically follows a Chi-squared distribution under correct model specification. We illustrate the GLS estimator and specification test using a case-control sample of peripheral arterial

disease, and use simulations to shed light on the operating characteristics of the specification test. The second part develops a robust difference-in-differences (DID) estimator for estimating causal effect with observational before-after data. Within the DID framework, two common estimation strategies are outcome regression and propensity score weighting. Motivated by a real application in traffic safety research, we propose a new double-robust DID estimator that hybridizes outcome regression and propensity score weighting. We show that the proposed estimator possesses the desirable large-sample robustness property, namely the consistency only requires either one of the outcome model or the propensity score model to be correctly specified. We illustrate the new estimator to study the causal effect of rumble strips in reducing vehicle crashes, and conduct a simulation study to examine its finite-sample performance. The third part discusses a unified framework, the balancing weights, for estimating causal effects in observational studies with multiple treatments. These weights incorporate the generalized propensity scores to balance the weighted covariate distribution of each treatment group, all weighted toward a common pre-specified target population. Within this framework, we further develop the generalized overlap weights, constructed as the product of the inverse probability weights and the harmonic mean of the generalized propensity scores. The generalized overlap weights corresponds to the target population with the most overlap in covariates between treatments, similar to the population in equipoise in clinical trials. We show that the generalized overlap weights minimize the total asymptotic variance of the nonparametric estimators for the pairwise contrasts within the class of balancing weights. We apply the new weighting method to study the racial disparities in medical expenditure and further examine its operating characteristics by simulations.

To my family

Contents

List of Tables

List of Figures

List of Abbreviations and Symbols

Symbols

$\boldsymbol{X}$	Covariates
Y	Outcome
D	Disease status
G, Z	Treatment status
n	Sample size
$\boldsymbol{\beta}$, $\boldsymbol{\gamma}$	Regression coefficients
τ	Target estimand

Abbreviations

GMM	Generalized Method of Moments
GLS	Generalized Least Squares
PAD	Peripheral Arterial Disease
SPREG	Semiparametric Retrospective Likelihood
PS	Propensity Score
DID	Different-in-Differences
DR	Double-Robust
IPW	Inverse Probability Weighting
GPSM	Generalized Propensity Score Matching
GPS	Generalized Propensity Score
G-MW	Generalized Matching Weights

G-OW Generalized Overlap Weights

Acknowledgements

I would like to thank my dissertation advisors Andrew Allen and Fan Li, for their guidance and instruction, all of which made this dissertation possible. Fan Li's patience and generosity of her time have been invaluable for my graduate study, and her understanding of both statistics and writing taught me an incredible amount over the last two years. I am also indebted to Andrew Allen for providing me the opportunity to discover my passion and research interest, and for putting up with me while I did. I also thank Liz DeLong, Kouros Owzar and Liz Turner for their advice and support while serving on my committee. A special thanks to Liz DeLong, who offered me invaluable opportunities to work within the NIH Collaboratory of Pragmatic Clinical Trials and guided me through my first statistical publication, to Liz Turner, who encouraged me to work in cluster randomized trials and always wanted the best for me, and to Kouros Owzar, who generously shared with me his advanced computing facilities during my study at Duke. Last but not the least, I want to express my gratitude to Laine Thomas, who has funded me for my last year of graduate study and provided valuable suggestions to Chapter 4 of this dissertation.

The work in Chapter 2 was partially supported by the National Institutes of Health (NIH) grant P01CA142538. The content is solely the responsibility of the author and does not necessarily represent the official views of the NIH. I also thank Mayo Clinic and National Human Genome Research Institute for making the case-control data of peripheral arterial disease in Chapter 2 available through dbGaP.

1

Introduction

Weighting methods have embraced wide applications in statistics and econometrics. For example, the inverse probability weighting is routinely used to correct for survey non-response (Horvitz and Thompson, 1952). These weights adjust each individual observation by the reciprocal of its selection probability, and hence recover the population information from the biased sample. The case-control design, frequently seen in epidemiologic and genetic studies, can be regarded as a special type of survey design; analogous inverse probability weighting approaches have been extensively explored in the statistical literature when the interest is the association between exposures and the disease (primary analysis) and when the interest is the association among exposures (secondary analysis); see, for example, Scott and Wild (2002); Richardson et al. (2007); Monsees et al. (2009); Li and Gail (2012). Meanwhile, in observational comparative effectiveness research, the inverse probability weighting has been suggested as a valid approach to correct for confounding bias (Imbens and Rubin, 2015). In this context, Rosenbaum and Rubin (1983) first defined the propensity score—the conditional probability of receiving the treatment given the pre-treatment covariates—and justified the use of the reciprocal of the propensity score as weights.

Weighting by the inverse propensity score creates a pseudo-population where the pre-treatment covariates are balanced across the treatment groups; the weighted pseudo-population then mimics a randomized study and henceforth the comparison between the weighted outcomes leads to an unbiased causal effect (Hernán and Robins, 2019).

This dissertation develops and extends the inverse probability weighting methods for retrospective case-control studies and other prospective observational studies. The main contributions of this dissertation are (1) developing a new specification test to assess the adequacy of the estimated inverse probability weights for secondary analysis of case-control data; (2) providing a double-robust construction of the difference-in-differences estimator in before-after observational studies; the new estimator is consistent as long as the either the propensity score model or the outcome model is correctly specified and therefore provides two chances to correctly estimate the target parameter; (3) providing a unified framework—the balancing weights—for estimating causal effects with multiple treatments, and further developing a new weighting scheme, the generalized overlap weights, to emphasize the overlapped population at clinical equipoise. In the remainder of this chapter, we provide some background information that is important to the subsequent Chapters, and explicitly state our contributions to the literature.

1.1 Generalized Method of Moments

The generalized method of moments (GMM) is a framework for deriving estimators based moment conditions (Hansen, 1982). The method of moments (MM) could be regarded as a special type of GMM, in which case we have the equal number of moment conditions and parameters. GMM extends the MM to accommodate the scenario where we have more moment conditions than the number of parameters; i.e., when the moment conditions over-identify the parameters. Specifically, we write

$\boldsymbol{\theta}$ as a $p \times 1$ vector of parameters such that the following moment conditions hold

$$E[m(\boldsymbol{X}_i, \boldsymbol{\theta}_0)] = \mathbf{0}, \tag{1.1}$$

where $m(\cdot)$ is a $q \times 1$ $(q \geqslant p)$ vector of known estimating functions, θ_0 is the true value of the target parameter, and $\boldsymbol{X}_i$ is the ith $(i = 1, \ldots, n)$ observed data record. Further define the sample moment conditions as

$$\bar{m}_n(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} m(\boldsymbol{X}_i, \boldsymbol{\theta}), \tag{1.2}$$

the GMM estimator is then the minimizer of the following quadratic distance

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \bar{m}_n^T(\boldsymbol{\theta}) \boldsymbol{W}_n \bar{m}_n(\boldsymbol{\theta}),$$

where $\boldsymbol{W}_n$ is a weight matrix that converges in probability to a nonnegative definite matrix $\boldsymbol{W}$. Since the first-order condition of the GMM estimator could be expressed as

$$2 \left(\frac{\partial \bar{m}_n(\hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}^T} \right)^T \boldsymbol{W}_n \bar{m}_n(\hat{\boldsymbol{\theta}}) = \mathbf{0},$$

the GMM estimator is essentially obtained as the solution of a weighted combination of the estimating equations corresponding to each moment condition. We next review the key asymptotic properties of the GMM estimator; the detailed regularity conditions and proofs can be found in Hansen (1982) and Hall (2004).

Theorem 1.1.1. *(Hansen, 1982)*
(i) Under regularity conditions, for any given weighting matrix $\boldsymbol{W}_n$ that converges in probability to $\boldsymbol{W}$, the GMM estimator is consistent and asymptotically normal, namely, $\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_0$ and

$$\sqrt{n} \left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 \right) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Sigma}(\boldsymbol{W})),$$

where the asymptotic variance $\boldsymbol{\Sigma}(\boldsymbol{W}) = \left(\boldsymbol{D}^T\boldsymbol{W}\boldsymbol{D}\right)^{-1}\boldsymbol{D}^T\boldsymbol{W}\boldsymbol{V}\boldsymbol{W}\boldsymbol{D}\left(\boldsymbol{D}^T\boldsymbol{W}\boldsymbol{D}\right)^{-1}$, $\boldsymbol{D} = E\left(\frac{\partial \bar{m}_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_0^T}\right)$, *and* $\boldsymbol{V} = E\left(\bar{m}_n(\boldsymbol{\theta}_0)\bar{m}_n(\boldsymbol{\theta}_0)^T\right)$.

(ii) The asymptotic variance $\boldsymbol{\Sigma}(\boldsymbol{W})$ *is minimized when* $\boldsymbol{W} = \boldsymbol{V}^{-1}$. *In other words, the efficient GMM estimator* $\tilde{\boldsymbol{\theta}}$ *is obtained by setting the weighting matrix* $\boldsymbol{W}_n$ *as the sample precision matrix of the mean moment conditions* (1.2). *In this case, the resulting GMM estimator* $\tilde{\boldsymbol{\theta}}$ *attains the semiparametric efficiency bound formed by the collection of moment conditions (Chamberlain, 1987).*

The consistency of the GMM estimator relies on the population moment conditions (1.1), and misspecification of the estimating function $m(\cdot)$ may lead to biased estimates. The test of over-identifying restrictions, or sometimes named the specification test, has been developed to examine whether the estimating functions are correctly specified. The following theorem introduces the test statistic and identifies its asymptotic behaviour.

Theorem 1.1.2. *(Hansen, 1982)*
Under the null hypothesis that the estimating functions are unbiased, namely moment condition (1.1) *holds, the specification test statistic corresponding to the efficient GMM estimator* $\tilde{\boldsymbol{\theta}}$ *is,*

$$T = n\bar{m}_n^T(\tilde{\boldsymbol{\theta}})\boldsymbol{V}^{-1}(\tilde{\boldsymbol{\theta}})\bar{m}_n(\tilde{\boldsymbol{\theta}}) \xrightarrow{d} \chi^2_{q-p},$$

where $\boldsymbol{V}^{-1}(\tilde{\boldsymbol{\theta}})$ *is a consistent, plug-in estimator of the true precision matrix* $\boldsymbol{V}^{-1}$.

The specification test is powered against departure from the population moment condition (1.1). Specifically, when the estimating functions are misspecified such that the moment conditions no longer holds, the observed specification test statistic will follow a Chi-squared distribution with a non-centrality parameter, and usually lead to a rejection of the null hypothesis (Hall, 2004).

In Chapter 2, we revisit a recent estimator from Xing et al. (2016) for the analysis of secondary traits in case-control data to offer new insights and extensions. Specifically, the Xing et al. (2016) estimator is based on two sets of weighted moment conditions, with the weights estimated from the case-control data to correct for the sample ascertainment bias (because case-control data is a biased representation of the target population). As will be seen in due course, each set of the weighted moment conditions just-identifies the parameters, but they jointly over-identify the parameters. For estimation and inference, Xing et al. (2016) solved each set of weighted estimating equations and then optimally combined the solutions by inverse variance weighting. However, as we pointed out just now, a natural strategy to accommodate over-identifying moment conditions is GMM. Given such considerations, our first contribution is to establish the connection between the Xing et al. (2016) estimator and the GMM estimator, with the additional complexity arising from exploiting the weighted moment conditions under case-control sampling. We prove in Theorem 2.3.1 that these two estimators are asymptotically equivalent. Intuitively, the Xing et al. (2016) estimator can be regarded as a "solve-and-combine" approach, as it exploits the best linear combination of solutions to separate estimating equations which just-identifies the parameters. On the other hand, the GMM estimator can be regarded as a "combine-and-solve" approach, as it solves the just-identified estimating equations defined by the best linear combination of over-identifying estimating equations. Theorem 2.3.1 states that these two types of approaches are asymptotically equivalent. Based on this connection and the fact that the specification test has been used to diagnose moment misspecification, our second contribution is to develop an analogous specification test appropriate for the Xing et al. (2016) estimator. The asymptotic behaviour and the finite-sample performance of the new specification test is thoroughly discussed.

1.2 Difference-in-Differences

Difference-in-differences (DID) is a statistical technique developed in econometrics for the evaluation of policy effects using observational data involving multiple time periods (Card and Krueger, 1994; Heckman et al., 1997; Heckman, 1998). Even though popularized in econometrics, DID has recently received increasing attention in the medical and epidemiological applications; see, for instance, Branas et al. (2011); Dimick and Ryan (2014); Stuart et al. (2014); Sommers et al. (2014); Grabich et al. (2015). The simplest and most common setting involves two periods. The first period is a pre-treatment period where all units are not exposed to the new policy. The new policy (treatment) is introduced after the first period such that a certain fraction of units is exposed to the treatment during the second period. The DID estimator is formed by contrasting two differences, one calculated as the difference between the observed outcomes across periods among the treated group while the other among the control group. Under the identification assumption of the parallel trend, that is, the treated group would have had the same average time trend as the control group had it not been treated, an unbiased nonparametric estimator of the policy effect among the treated could be identified as the difference between these two differences (Imbens and Wooldridge, 2009). The presentation related to this dissertation will mainly cover this common design with two periods.

The standard DID estimator is developed from a linear fixed-effects model adjusting for both time and treatment status (Ashenfelter and Card, 1985). The coefficient of the treatment-time interaction term is then intepreted as the DID estimator. Further, the identification condition for this fixed-effects model dictates that the time trend in the average outcomes is identical between the treated and control group (Abadie, 2005). Indeed, when the treated and control groups are balanced in pre-treatment covariates (confounders) that are potentially associated with both treat-

ment assignment and the outcome, such a parallel trend assumption is reasonable. In practical applications where the confounders are likely unbalanced between groups, an naive application of this fixed-effects model may lead to biased inference. Outcome regression is introduced as a viable strategy to avoid such bias issues. By incorporating covariates into the fixed-effects model, one may identify the conditional average treatment effect by estimating the regression parameters, and obtain interpretable results by integrating over the a pre-specified covariate distribution. However, it is also known in the causal inference literature that regression adjustment is sensitive to model misspecification when the covariate distributions differ greatly between the treatment groups (Rubin, 1979). Further, in the presence of high-dimensional covariates, the estimation of the regression coefficients becomes unstable and numeric integration poses additional computational challenges.

An alternative nonparametric inference strategy to balance covariates is through propensity score weighting. The propensity score is defined by Rosenbaum and Rubin (1983) as follows:

Definition 1.2.1. *(Rosenbaum and Rubin, 1983)*
The propensity score is the conditional probability of receiving the treatment given the set of pre-treatment covariates, $e(\boldsymbol{X}) = Pr(G = 1|\boldsymbol{X})$*, where* $G \in \{1, 0\}$ *is the treatment group indicator, and* $\boldsymbol{X}$ *is the set of pre-treatment covariates.*

Inverse probability weighting (IPW) is a standard technique to control for baseline differences between treatment groups. The general idea is to weight each subject by the inverse probability of the observed treatment status, such that the potential confounders are approximately balanced between groups in the weighted pseudo-population (Rosenbaum and Rubin, 1983; Rosenbaum, 1987; D'Agostino Jr, 1998). Because of its practical convenience, the weighting approach has been adopted in various applications; see, for instance, Imbens (2000); Robins et al. (2000); Hirano

and Imbens (2001); Sato and Matsuyama (2003); Hirano et al. (2003); Joffe et al. (2004); Lunceford and Davidian (2004); Cole and Hernán (2008). Within the DID framework, Abadie (2005) studied nonparametric identification conditions and proposed a weighting strategy for estimating the average treatment effect among the treated (ATT). A key identification restriction for weighting is the following:

Assumption 1.2.2. *(Abadie, 2005)*
Given a vector of pre-treatment covariates $\boldsymbol{X}_i$, the treatment assignment is mean independent of the time trend of potential outcomes absent treatment:

$$E[Y_{i,t+1}(0) - Y_{it}(0)|\boldsymbol{X}_i, G_i = 1] = E[Y_{i,t+1}(0) - Y_{it}(0)|\boldsymbol{X}_i, G_i = 0] \quad \forall\ i,$$

where $Y_{i,t+1}(0)$, and $Y_{it}(0)$ are the potential outcomes in the absence of the treatment at period $t+1$ and t, and G_i is the treatment group indicator.

This assumption is a conditional version of the parallel trend, which specifies that conditional on a set of pre-treatment covariates, the treated group would have had the same average time trend as the control group had it not been treated. In other words, it is assumed that conditional on the set of covariates, treatment assignment is mean independent of the underlying time trend if all units are not exposed to the treatment. With the conditional parallel trend assumption, Abadie (2005) showed the following nonparametric identification results:

Theorem 1.2.3. *(Abadie, 2005)*
Under Assumption 1.2.2, the ATT estimand can be identified by

$$E[Y_{i,t+1}(1) - Y_{i,t+1}(0)|G_i = 1] = E\Big[\frac{Y_{i,t+1} - Y_{it}}{Pr(G_i = 1)}\left\{\frac{G_i - e(\boldsymbol{X}_i)}{1 - e(\boldsymbol{X}_i)}\right\}\Big]$$

$$= E[Y_{i,t+1} - Y_{it}|G_i = 1] - \frac{1 - Pr(G_i = 1)}{Pr(G_i = 1)} E\Big[\frac{(Y_{i,t+1} - Y_{it})e(\boldsymbol{X}_i)}{1 - e(\boldsymbol{X}_i)}\Big|G_i = 0\Big],$$

where $Y_{i,t+1}$, Y_{it} are the observed outcomes for the ith unit at period $t+1$ and t, respectively.

The above theorem indicates that the conventional propensity score weighting scheme for estimating ATT could be applied to the observed individual differences across the two time periods for the construction of a weighting estimator. However, the consistency of such an estimator critically depends on the correct specification of the propensity score model. If the propensity score is misspecified, the corresponding DID inference may produce spurious results.

In Chapter 3, our contribution is to develop a double-robust (DR) DID estimator by hybridizing outcome regression and inverse probability weighting. The development is motivated by a traffic safety before-after study evaluating the causal effect of rumble strip installation on reducing vehicle crashes, based on data obtained from the Pennsylvania Department of Transportation (PennDOT). In the traffic safety studies, the ATT estimand is considered appropriate since rumble strips were installed only for the selected pilot sites due to their roadway features, and there is an interest in considering the causal effect on these selected pilot sites before rolling out such interventions to a larger scale. To infer this causal quantity, we show that the proposed DR DID estimator improves upon the outcome regression alone or weighting alone in the sense that it is consistent to the target estimand as long as one of the regression or propensity score models are correctly specified, without distinguishing which one is the correct model. Both simulations and empirical data analysis are presented in Chapter 3 to illustrate the new estimator.

1.3 Overlap Weights for Binary Treatments

Sources of data for observational treatment comparisons are expanding to include billing claims, large registries and electronic health records. These resources have the potential to answer questions about the safety and effectiveness of treatments, although statistical methods must account for the lack of randomization. For cross-sectional comparisons between two treatments (treatment and control), the inverse

probability weighting (IPW) is a popular approach to adjust for confounding due to differences between comparator groups that arise in such observational data (Austin and Stuart, 2015; Bouillon et al., 2018; Brown et al., 2017; Jones et al., 2018). Assuming that all of the important confounders are measured, this approach is appealing for its simplicity and alignment with a potential experiment: what if the entire sample had instead been randomized to the intervention of interest? In practice, IPW may perform poorly when the treatment groups are initially very different and some patients have extreme propensity scores near 1 or 0, i.e. almost always and never receive treatment, respectively (Stuart, 2010; Lee et al., 2011; Hirano and Imbens, 2001). Extreme propensities are particularly common in the setting of big data, where inclusion criteria can be defined broadly. The increasing prevalence of large data sources precipitates the need to clarify best practice in weighting-based inference with regard to the handling of extreme propensity scores.

When propensity scores are close to 0 or 1, IPW has limitations that include: (1) large weights for individual patients; (2) bias and (3) large variability in the estimated treatment effect (Stuart, 2010; Lee et al., 2011; Hirano and Imbens, 2001). To address these problems, trimming methods have been proposed that exclude individuals from the sample who have very high predicted probabilities of being on treatment or control (Stürmer et al., 2010; Crump et al., 2009). Despite the potential gains from trimming, the decision regarding how many patients to exclude is ad hoc and can result in substantial loss of sample size. These problems are mitigated by the newly developed overlap weighting method, in which each patient's weight is the probability of that patient being assigned to the opposite group (Li et al., 2018). The properties of overlap weights have been demonstrated theoretically, and include improvements in balance and precision relative to IPW. In addition, these weights are bounded and smoothly reduce the influence of patients at the tails of the propensity distribution without making any exclusions. Specifically, the overlap weights are also

shown to be the optimal member in the family of balancing weights that provide the smallest asymptotic variance of the sample weighting estimator. We now provide a brief overview on the definition of balancing weights and the construction of the overlap weights for binary treatments.

We adopt the notations in the Rubin Causal Model and assume $Y(1)$, $Y(0)$ are the pair of potential outcomes under the treatment condition and the control condition. We define $\tau(\boldsymbol{X}) = E[Y(1) - Y(0)|\boldsymbol{X}]$ as the conditional average treatment effect (Li et al., 2013). Assume the marginal density of the pre-treatment covariates, $f(\boldsymbol{X})$, exists, with respect to a base measure μ. In causal studies, the interest is on the average effects of units in a target population, whose density (up to a normalizing constant) we represent by $f(\boldsymbol{X})h(\boldsymbol{X})$, with $h(\boldsymbol{X})$ being a pre-specified function of covariates, or equivalently a tilting function. We could first define the expectation of the potential outcomes over the target population $f(\boldsymbol{X})h(\boldsymbol{X})$:

$$\tau^h \equiv \frac{\int \tau(\boldsymbol{X}) f(\boldsymbol{X}) h(\boldsymbol{X}) \mu(d\boldsymbol{X})}{\int f(\boldsymbol{X}) h(\boldsymbol{X}) \mu(d\boldsymbol{X})}.$$

Notice that this is a general characterization of the target estimand, but once we set $h(\boldsymbol{X}) = 1$, the general target estimand reduces to the usual average treatment effect (ATE), namely, $\tau^{h=1} = \mathrm{E}[Y(1) - Y(0)]$.

To estimate the general class of estimand, Li et al. (2018) formalized the concept of balancing weights as follows. Define Z as the treatment variable, taking 1 or 0 to label treatment or control. Let $f_1(\boldsymbol{X}) = f(\boldsymbol{X}|Z = 1)$ be the density of $\boldsymbol{X}$ in the treatment group, and $f_0(\boldsymbol{X}) = f(\boldsymbol{X}|Z = 0)$ be the density of $\boldsymbol{X}$ in the control group. It is easy to see that $f_1(\boldsymbol{X}) \propto f(\boldsymbol{X})e(\boldsymbol{X})$, and $f_0(\boldsymbol{X}) \propto f(\boldsymbol{X})[1 - e(\boldsymbol{X})]$, where $e(\boldsymbol{X}) = Pr(Z = 1|\boldsymbol{X})$ is the previously-defined propensity score (Rosenbaum and Rubin, 1983). Given any pre-specified function h, we can weight the group-specific density $f_1(\boldsymbol{X})$ and $f_0(\boldsymbol{X})$ to the target population using the following weights, pro-

portional up to a normalizing constant: $w_1(\boldsymbol{X}) \propto \frac{h(\boldsymbol{X})}{e(\boldsymbol{X})}$ for the treatment group, and $w_0(\boldsymbol{X}) \propto \frac{h(\boldsymbol{X})}{1-e(\boldsymbol{X})}$ for the control group. Then the sample weighting estimator can be written as

$$\hat{\tau}^h = \frac{\sum_{i=1}^n Z_i Y_i w_1(\boldsymbol{X}_i)}{\sum_{i=1}^n Z_i w_1(\boldsymbol{X}_i)} - \frac{\sum_{i=1}^n (1-Z_i) Y_i w_0(\boldsymbol{X}_i)}{\sum_{i=1}^n (1-Z_i) w_0(\boldsymbol{X}_i)}.$$

The asymptotic properties of the sample weighting estimator is established in Li et al. (2018) and reviewed as follows.

Theorem 1.3.1. *(Li et al., 2018)*

(i) Under the unconfoundedness assumption such that $\{Y(1), Y(0)\} \perp Z|\boldsymbol{X}$*, the sample weighting* $\hat{\tau}^h$ *is consistent to the target estimand* τ^h *for all choices of* h*.*

(ii) As $n \to \infty$*, the expectation over possible samples of covariate values of the conditional variance of the estimator* $\hat{\tau}^h$ *converges*

$$nE_{\boldsymbol{X}}\{V[\hat{\tau}^h|\boldsymbol{X}]\} \to \int f(\boldsymbol{X}) h^2(\boldsymbol{X}) [v_1(\boldsymbol{X})/e(\boldsymbol{X}) + v_0(\boldsymbol{X})/(1-e(\boldsymbol{X}))] \mu(d\boldsymbol{X})/C_h^2,$$

where $v_1(\boldsymbol{X}) = V[Y(1)|\boldsymbol{X}]$*,* $v_0(\boldsymbol{X}) = V[Y(0)|\boldsymbol{X}]$ *are residual variances of the potential outcomes, and* $C_h^2 = \int h(\boldsymbol{X}) f(\boldsymbol{X}) d\mu(\boldsymbol{X})$ *is a normalizing constant.*

(iii) The function $h(\boldsymbol{X}) \propto e(\boldsymbol{X})(1-e(\boldsymbol{X}))$ *gives the smallest asymptotic variance for the sample weighting estimator* $\hat{\tau}^h$ *among all* h*'s under homoscedasticity in the residual variance. Therefore the optimal (overlap) weights are proportional to the probability to be assigned to opposite group:* $w_1(\boldsymbol{X}) \propto (1-e(\boldsymbol{X}))$ *and* $w_0(\boldsymbol{X}) \propto e(\boldsymbol{X})$*.*

Since the overlap weights are defined through the optimal function $h(\boldsymbol{X}) \propto e(\boldsymbol{X})(1-e(\boldsymbol{X}))$, which is maximized at $e(\boldsymbol{X}) = 1/2$, the overlap weights emphasize patients with substantial probability to receive each treatment, namely patients at clinical equipoise. This scientifically relevant interpretation exemplifies the principles of observational studies analyzed like randomized trials.

With the expansions of sources for observational treatment comparisons, the comparative effectiveness evidence among multiple treatments is often of substantial interest. For this reason, our contribution in Chapter 4 is to generalize the overlap weights to accommodate non-randomized comparisons of multiple treatments. We first extend the unified framework, the balancing weights, for estimating causal effects in the context of multiple treatments. These weights incorporate the generalized propensity score to balance the weighted covariate distribution of each treatment group, all weighted toward a common pre-specified target population. Within this framework, we then define a class of target estimands based on linear contrasts and their corresponding nonparametric weighting estimators. We further develop the generalized overlap weights, constructed as the product of the inverse probability weights and the harmonic mean of the generalized propensity scores. The generalized overlap weights corresponds to the target population with the most overlap in covariates between treatments, similar to the population in equipoise in clinical trials. We show that the generalized overlap weights minimize the total asymptotic variance of the nonparametric estimators for the pairwise contrasts within the class of balancing weights. We apply these methods to study the racial disparities in medical expenditure and further examine their operating characteristics by simulations.

2

Secondary Analysis of Case-Control Association Studies: Insights on Weighting-Based Inference Motivate a New Specification Test

2.1 Introduction

The case-control design is a biased sampling design in which two separate samples, one of diseased individuals (cases) and one of disease-free individuals (controls), are sampled separately from a population. Because cases and controls are sampled separately, cases can be over-sampled so that the number of cases and controls are more closely balanced in the resulting sample. This balance minimizes the number of exposures that need to be assessed to attain a given level of statistical power. When the disease is rare or exposures are expensive to collect, the case-control study represents a cost-effective strategy relative to a prospective cohort study. For example, although sequencing costs continue to fall, measuring genetic exposures using whole-genome sequencing technology is still expensive. Because of this, most genetic association studies adopt the case-control design.

Although the primary interest in such studies is in identifying genetic associations

with disease, there remains considerable interest in making the most out of these datasets. Most studies collect information about phenotypes beyond disease status, either because they are thought to be relevant to the disease process or simply because they are readily available. This abundance of information allows researchers to explore associations between genetic exposures and the secondary phenotypes without additional, data-generation related costs. Identifying such associations could be of interest in itself, or it could help elucidate the underlying biologic processes involved in the primary disease. Examples of secondary phenotypes in case-control association analyses include body mass index (Willer et al., 2009) and human height (Weedon et al., 2007). Such analyses, however, are complicated by the fact that the case-control sample is a biased representation of the population. As a result, if the secondary phenotypes are associated with disease status (which is often the case), any population-level secondary trait associations will likely be distorted in the sample, and direct analysis of the combined case-control data that ignores the retrospective design can lead to severely biased results.

Appropriate methods for the analysis of secondary phenotypes in case-control studies can be broadly categorized as retrospective likelihood and weighting approaches. Retrospective likelihood approaches explicitly account for case-control sampling by formulating a retrospective likelihood (Jiang et al., 2006; Wang and Shete, 2012; Chen et al., 2013; Tchetgen Tchetgen, 2014). Lin and Zeng (2009) studied the retrospective likelihood for both continuous and binary secondary traits, and provided software (SPREG) for implementation. It is, however, recognized that likelihood-based methods critically depend on correctly modeling the secondary trait distribution. Relaxing these distributional assumptions, Wei et al. (2013) developed a semiparametric efficient estimator for continuous secondary phenotypes under rare disease and homoscedastic variance assumptions. Ma and Carroll (2016) extended this semiparametric approach by relaxing the rare disease and homoscedasticity as-

sumptions. However, since this improved procedure involves non-parametric regression conditional on the covariates, it may not be easily implemented with many covariates. Although these semiparametric estimators are locally efficient, the additional computational complexity involved in their implementation may limit their utility in genome-wide analyses.

A computational convenient approach for secondary analysis uses weighting to correct for sample ascertainment bias. The survey-weighted estimator, originated from the survey sampling literature, weights each individual observation by the reciprocal of its selection probability (Richardson et al., 2007; Monsees et al., 2009; Li and Gail, 2012). This approach produces consistent estimates if the selection probability and hence the weighted estimating functions are correctly specified. The selection probabilities are readily available when the case-control sample is obtained from a large cohort or when the disease prevalence is known. However, simply weighting by the inverse selection probability gives inefficient estimates (Robins et al., 1994). To address this limitation, we have previously proposed an alternative weighting-based estimator that uses the inverse conditional disease probability as weights (Xing et al., 2016), and shown improved power over the survey-weighted approach and robustness to distributional assumptions of the secondary traits. However, the statistical properties of the our weighting estimator have not been thoroughly explored. In particular, the consistency of our weighting estimator depends on the retrospective unbiasedness of the estimating equations, but diagnostic tools for misspecification of the disease model have not yet been studied.

In this Chapter, we revisit our previous weighting-based estimator for the analysis of secondary case-control data to offer new insights and methodological extensions. Specifically, our contributions to the literature on secondary analysis are two-fold. First, we extend our previous scalar estimator to the multivariate case, and introduce a more general estimator based on generalized least squares (GLS). Such an

extension allows us to connect the GLS estimator with the well-known generalized method of moments (GMM) and therefore establish the efficiency property of the GLS estimator. Second, the connection between GLS and GMM motivates a new specification test designed to assess the adequacy of the disease model or the weights. By construction, the specification test statistic measures the weighted distance between the case and control subsample estimators, and asymptotically follows a central Chi-squared distribution under correct model specification. We illustrate the GLS estimator and specification test using a case-control sample of peripheral arterial disease, and use simulations to further shed light on the operating characteristics of the specification test. We found that the proposed specification test is particularly powerful when misspecification of the disease model likely leads to biased estimates and under-coverage, and thus recommend their use as a routine check for weighting-based secondary analysis.

2.2 Preliminaries

We briefly introduce the basic setup for secondary analysis and review our previous estimation strategy as follows. Consider a population-level restricted moment model $E(Y|\boldsymbol{X}) = \mu(\boldsymbol{X}, \boldsymbol{\beta})$, where Y is a scalar secondary phenotype, $\boldsymbol{X}$ is a vector of explanatory variables including the genetic exposures of interest (e.g. minor allele count at a given locus taking values from $\{0, 1, 2\}$) and possibly additional covariates (e.g. principal components adjusting for population stratification), μ is a known smooth function and $\boldsymbol{\beta}$ is a p-dimensional parameter including the target secondary association of interest. Under prospective random sampling, the natural estimating function could be the efficient score $g(Y, \boldsymbol{X}, \boldsymbol{\beta}) = [\partial\mu(\boldsymbol{X}, \boldsymbol{\beta})/\partial\boldsymbol{\beta}^T]^T V^{-1}(Y|\boldsymbol{X})[Y - \mu(\boldsymbol{X}, \boldsymbol{\beta})]$, where $V(Y|\boldsymbol{X})$ denotes the conditional variance of Y given $\boldsymbol{X}$. Clearly, the efficient score function is prospectively unbiased, namely, the following moment

condition holds under the true value $\boldsymbol{\beta}_0$,

$$E[g(Y, \boldsymbol{X}, \boldsymbol{\beta}_0)] = \mathbf{0}. \tag{2.1}$$

The secondary analysis could be conceptualized as estimating the population-level parameter $\boldsymbol{\beta}$ based on a case-control sample with n_0 controls ($D = 0$) and n_1 cases ($D = 1$). Define the triplet $\boldsymbol{Z} = (Y, \boldsymbol{X}, D)$, and denote its observed version by $\boldsymbol{z}_i = (y_i, \boldsymbol{x}_i, d_i)$, $i = 1, \ldots, n$, where $n = n_0 + n_1$ is the total sample size, $d_i = 0$ for $i = 1, \ldots, n_0$ for the controls and $d_i = 1$ for $i = n_0 + 1, \ldots, n$ for the cases. We assume there exists positive constants ρ_0 and ρ_1 such that $\lim_{n\to\infty} n_0/n = \rho_0$, $\lim_{n\to\infty} n_1/n = \rho_1$ with $\rho_0 + \rho_1 = 1$. Usually the disease status D is related to both $\boldsymbol{X}$ and Y, so the association between Y and $\boldsymbol{X}$ can be different in the cases relative to the controls. Therefore, although the estimating function g is prospectively unbiased, it is generally not retrospectively unbiased, i.e., $E[g(Y, \boldsymbol{X}, \boldsymbol{\beta}_0)|D = d] \neq \mathbf{0}$ for both $d = 0, 1$, suggesting that solutions to the prospectively derived estimating equations are inconsistent with case-control data.

To construct the weighted estimating functions that are retrospectively unbiased, we assume that the true disease probability can be uniquely characterized by a parametric model $Pr(D = 1|Y, \boldsymbol{X}) = H(Y, \boldsymbol{X}, \boldsymbol{\gamma}_0)$, where $\boldsymbol{\gamma}_0$ is a q-dimensional parameter vector. If $\boldsymbol{\gamma}_0$ is known, we could define the weighted estimating function for the control sample as $\Psi_0(\boldsymbol{Z}, \boldsymbol{\beta}, \boldsymbol{\gamma}_0) = \frac{(1-D)g(Y,\boldsymbol{X},\boldsymbol{\beta})}{1-H(Y,\boldsymbol{X},\boldsymbol{\gamma}_0)}$. Similarly, the weighted estimating function for the case sample is $\Psi_1(\boldsymbol{Z}, \boldsymbol{\beta}, \boldsymbol{\gamma}_0) = \frac{Dg(Y,\boldsymbol{X},\boldsymbol{\beta})}{H(Y,\boldsymbol{X},\boldsymbol{\gamma}_0)}$. We have previously shown that the prospective moment condition (2.1) holds if and only if estimating functions Ψ_0, Ψ_1 are both retrospectively unbiased. In other words, the retrospective unbiasedness property translates to the following set of retrospective moment conditions,

$$E[\Psi_0(\boldsymbol{Z}, \boldsymbol{\beta}_0, \boldsymbol{\gamma}_0)|D = d] = \mathbf{0}, \tag{2.2}$$

$$E[\Psi_1(\boldsymbol{Z}, \boldsymbol{\beta}_0, \boldsymbol{\gamma}_0)|D = d] = \mathbf{0}, \qquad d = 0, 1. \tag{2.3}$$

According to the above moment conditions, if $\boldsymbol{\gamma}_0$ is known, we can simply construct estimating equations based on Ψ_0 or Ψ_1 to consistently estimate $\boldsymbol{\beta}$. In practice, $\boldsymbol{\gamma}_0$ is not known but could be estimated from case-control data. We assume the true disease model takes a general logistic form, with the prospective parameter vector $\boldsymbol{\gamma} = (\gamma_1, \boldsymbol{\gamma}_2^T)^T$, as $\text{logit}[Pr(D = 1|Y, \boldsymbol{X})] = \gamma_1 + m(Y, \boldsymbol{X}, \boldsymbol{\gamma}_2)$, where m is a known smooth function. Note that this general model includes as special cases the linear logistic model and the multiplicative-intercept risk model (Weinberg and Wacholder, 1993). For estimating $\boldsymbol{\gamma}$, we could consider two complimentary scenarios: the disease is common with known prevalence $\lambda = Pr(D = 1)$, and the disease is rare with unknown prevalence.

When the disease is common, accurate estimates of λ are usually available from existing studies. In this case, it is possible to estimate $\boldsymbol{\gamma}$ using a case-control sample with a modified logistic model (Scott and Wild, 1986; Prentice and Pyke, 1979; Carroll et al., 1995)

$$\text{logit}[Pr(D = 1|Y, \boldsymbol{X})] = \gamma_1 + o + m(Y, \boldsymbol{X}, \boldsymbol{\gamma}_2), \tag{2.4}$$

where the fixed offset $o = \log(n_1/n_0) - \log[\lambda/(1-\lambda)]$. Denote the score function for model (2.4) by $\Phi(\boldsymbol{Z}, \boldsymbol{\gamma})$, it is known that even though the score function has nonzero expectation under retrospective sampling, the estimating equations are retrospectively unbiased (Carroll et al., 1995), i.e., $\sum_{d=0}^{1} \sum_{i=0}^{n_d} E\left[\Phi(\boldsymbol{z}_i, \boldsymbol{\gamma}_0)|D_i = d_i\right] = \mathbf{0}$ holds, where $\boldsymbol{\gamma}_0 = (\gamma_{10}, \boldsymbol{\gamma}_{20})^T$ is the truth. The above result implies that the following retrospective moment condition holds:

$$\sum_{d=0}^{1} \rho_d E\left[\Phi(\boldsymbol{Z}, \boldsymbol{\gamma}_0)|D = d\right] = \mathbf{0}. \tag{2.5}$$

Thus, fitting model (2.4) to the case-control data allows consistent estimation for $\boldsymbol{\gamma}$ provided the disease prevalence is known.

When the disease is rare, accurate prevalence estimates are usually not available, but such estimates are in fact no longer required. Following Weinberg and Wacholder Weinberg and Wacholder (1993), we may use $H(Y, \boldsymbol{X}, \boldsymbol{\gamma}_0) \approx e^{\gamma_{10}+m(Y,\boldsymbol{X},\boldsymbol{\gamma}_{20})} \overset{Y,\boldsymbol{X}}{\propto} e^{m(Y,\boldsymbol{X},\boldsymbol{\gamma}_{20})}$ and $1 - H(Y, \boldsymbol{X}, \boldsymbol{\gamma}_0) \approx 1$, and the approximate weighted estimating functions become $\Psi_0(\boldsymbol{Z}, \boldsymbol{\beta}, \boldsymbol{\gamma}_0) \approx (1-D)g(Y, \boldsymbol{X}, \boldsymbol{\beta})$ and $\Psi_1(\boldsymbol{Z}, \boldsymbol{\beta}, \boldsymbol{\gamma}_0) \approx Dg(Y, \boldsymbol{X}, \boldsymbol{\beta})e^{-m(Y,\boldsymbol{X},\boldsymbol{\gamma}_{20})}$. Note that $e^{\gamma_{10}}$ does not involve $\boldsymbol{X}$ or Y and factors out in the estimating function Ψ_1 without affecting its approximate unbiasedness. These approximate estimating functions suggest that the rare disease assumption dispenses with the knowledge of disease prevalence, and only $\boldsymbol{\gamma}_2$ need to be consistently estimated. This can be achieved by directly fitting model $\text{logit}[Pr(D = 1|Y, \boldsymbol{X})] = \gamma_1 + m(Y, \boldsymbol{X}, \boldsymbol{\gamma}_2)$ to the case-control sample as if it is obtained prospectively (Prentice and Pyke, 1979). For brevity, we focus on the common disease scenario in subsequent derivations and revisit the rare disease approximation in numerical studies.

2.3 Estimation and Inference via Generalized Least Squares

2.3.1 Estimation

Our previous weighting-based estimator Xing et al. (2016) is only presented for estimating a scalar element of $\boldsymbol{\beta}$. Here we extend the method to simultaneously estimate all elements of $\boldsymbol{\beta}$ for the purpose of obtaining additional insights. To proceed, we first estimate $\hat{\boldsymbol{\gamma}}$ to obtain the plug-in weighted estimating functions $\Psi_0(\boldsymbol{Z}, \boldsymbol{\beta}, \hat{\boldsymbol{\gamma}})$ and $\Psi_1(\boldsymbol{Z}, \boldsymbol{\beta}, \hat{\boldsymbol{\gamma}})$. We then solve

$$\sum_{i=1}^{n} \Psi_0(\boldsymbol{z}_i, \hat{\boldsymbol{\beta}}^0, \hat{\boldsymbol{\gamma}}) = \mathbf{0},$$

$$\sum_{i=1}^{n} \Psi_1(\boldsymbol{z}_i, \hat{\boldsymbol{\beta}}^1, \hat{\boldsymbol{\gamma}}) = \mathbf{0}$$

separately for $\hat{\boldsymbol{\beta}}^0$ and $\hat{\boldsymbol{\beta}}^1$, both of which are consistent for $\boldsymbol{\beta}_0$ based on case or control subsample. Notice that these estimating equations are of dimension p, so they each can be conveniently solved by standard software routines permitting the use of weights. We write the stacked vector $\hat{\boldsymbol{\beta}}^{01} = (\hat{\boldsymbol{\beta}}^{0T}, \hat{\boldsymbol{\beta}}^{1T})^T$, and denote the asymptotic variance of $\hat{\boldsymbol{\beta}}^{01}$ by $\boldsymbol{\Omega}$ (the explicit form is provided in Appendix A). If $\hat{\boldsymbol{\Omega}}$ is a consistent estimator for $\boldsymbol{\Omega}$ under case-control sampling, we can formulate the following quadratic distance function

$$M_n(\boldsymbol{\beta}) = (\hat{\boldsymbol{\beta}}^{01} - \boldsymbol{R}\boldsymbol{\beta})^T \hat{\boldsymbol{\Omega}}^{-1} (\hat{\boldsymbol{\beta}}^{01} - \boldsymbol{R}\boldsymbol{\beta}), \tag{2.6}$$

where $\boldsymbol{R} = (\boldsymbol{I}^{p\times p}, \boldsymbol{I}^{p\times p})^T$ is a $2p \times p$ constant matrix. The final estimator for $\boldsymbol{\beta}$ is defined as the minimizer of the above distance and is given in closed-form by the GLS solution

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{R}^T \hat{\boldsymbol{\Omega}}^{-1} \boldsymbol{R})^{-1} \boldsymbol{R}^T \hat{\boldsymbol{\Omega}}^{-1} \hat{\boldsymbol{\beta}}^{01}. \tag{2.7}$$

The GLS formulation implies that our estimator is a weighted average of the two subsample estimators, $\hat{\boldsymbol{\beta}}^0$ and $\hat{\boldsymbol{\beta}}^1$, i.e. it is a member in the class of estimators with the form $\boldsymbol{A}_0\hat{\boldsymbol{\beta}}^0 + \boldsymbol{A}_1\hat{\boldsymbol{\beta}}^1$, subject to the sum-to-unity constraint $\boldsymbol{A}_0 + \boldsymbol{A}_1 = \boldsymbol{I}^{p\times p}$. Other members in this class can be obtained if we replace $\hat{\boldsymbol{\Omega}}$ in (2.6) by some other positive definite matrix. Given that $\hat{\boldsymbol{\Omega}} \xrightarrow{p} \boldsymbol{\Omega}$, we can follow the arguments of the Gauss-Markov theorem and conclude that $\hat{\boldsymbol{\beta}}$ is the asymptotically best linear unbiased estimator in the sense that it has the minimum asymptotic variance among all estimators of the form $\boldsymbol{A}_0\hat{\boldsymbol{\beta}}^0 + \boldsymbol{A}_1\hat{\boldsymbol{\beta}}^1$ and subject to the sum-to-unity constraint. It is also straightforward to verify that each element of the GLS estimator $\hat{\boldsymbol{\beta}}$ reduces to our previous scalar estimator (Xing et al., 2016).

The formulation of quadratic distance (2.6) motivates the connection between the GLS estimator and the GMM estimator. GMM was introduced by Hansen in the econometrics literature (Hansen, 1982); GMM bases inference on a set of moment

conditions and is particularly attractive when the set of estimating functions over-identifies the parameter of interest (Hall, 2004). It is clear from moment conditions (2.2) and (2.3) that the dimension of the weighted estimating functions exceeds the dimension of $\boldsymbol{\beta}$, and hence a natural option for making inference is through the GMM that simultaneously estimates $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \boldsymbol{\gamma}^T)^T$. More specifically, we write $\Psi(\boldsymbol{Z}, \boldsymbol{\theta}) = (\Psi_0^T(\boldsymbol{Z}, \boldsymbol{\theta}), \Psi_1^T(\boldsymbol{Z}, \boldsymbol{\theta}))^T$ and the joint estimating functions $\Upsilon(\boldsymbol{Z}, \boldsymbol{\theta}) = (\Psi^T(\boldsymbol{Z}, \boldsymbol{\theta}), \Phi^T(\boldsymbol{Z}, \boldsymbol{\gamma}))^T$. Suppose that $\mathcal{V}$ is the variance matrix of $\Upsilon(\boldsymbol{Z}, \boldsymbol{\theta})$ under case-control sampling evaluated at the truth $\boldsymbol{\theta}_0$, and $\boldsymbol{K}_n$ is a positive semi-definite matrix such that $\boldsymbol{K}_n \xrightarrow{p} \mathcal{V}^{-1}$, the GMM estimator $\hat{\boldsymbol{\theta}}_{\text{GMM}}$ minimizes the following objective function

$$J_n(\boldsymbol{\theta}) = \Big[n^{-1}\sum_{i=1}^{n} \Upsilon(\boldsymbol{z}_i, \boldsymbol{\theta})\Big]^T \boldsymbol{K}_n \Big[n^{-1}\sum_{i=1}^{n} \Upsilon(\boldsymbol{z}_i, \boldsymbol{\theta})\Big]. \tag{2.8}$$

The GMM estimator is an appealing choice since the resulting estimator $\hat{\boldsymbol{\theta}}_{\text{GMM}}$ is guaranteed to be asymptotically efficient in the sense that its asymptotic variance achieves the variance bound formed by the retrospective moment conditions (2.2), (2.3) and (2.5). In other words, $\hat{\boldsymbol{\theta}}_{\text{GMM}}$ (and hence $\hat{\boldsymbol{\beta}}_{\text{GMM}}$) is the asymptotically minimum variance estimator among the class of estimators derived solely from these moment conditions (Chamberlain, 1987; Hall, 2004). However, since GMM identifies $\boldsymbol{\gamma}$ through the joint estimating functions (while $\boldsymbol{\gamma}$ is directly identifiable by logistic regression), this approach is generally computationally slow. In our experience, GMM also tends to be numerically unstable by iteratively updating $\boldsymbol{K}_n$ for optimizing $J_n(\boldsymbol{\theta})$. These issues may limit its application to the genome-wide scan for secondary trait associations. Nevertheless, we prove in Theorem 2.3.1 that the $\hat{\boldsymbol{\beta}}$ is asymptotically equivalent to $\hat{\boldsymbol{\beta}}_{\text{GMM}}$, thus establishing the asymptotic efficiency optimality of the computational convenient alternative based on GLS.

Theorem 2.3.1. *Under regularity conditions listed in Appendix, the GLS estimator*

is consistent and asymptotically normal with

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{d} N\left(\mathbf{0}, (\mathcal{J}_{11}^T \mathcal{E}^{-1} \mathcal{J}_{11})^{-1}\right),$$

where $\mathcal{J}_{11}$ and $\mathcal{E}$ are defined in Appendix A. Further, the GLS estimator is asymptotically equivalent to the GMM estimator $\hat{\boldsymbol{\beta}}_{GMM}$, and attains the efficiency bound for $\boldsymbol{\beta}$ imposed by the moment conditions (2.2), (2.3) and (2.5).

Proof. See Section 2.7.2. □

By construction, the GLS estimator can be regarded as a "solve-and-combine" approach, as it exploits the best linear combination of subsample solutions to separate estimating equations which just-identifies $\boldsymbol{\beta}$. On the other hand, the GMM estimator can be regarded as a "combine-and-solve" approach, as it solves the just-identified estimating equations defined by the best linear combination of over-identifying estimating equations. Theorem 2.3.1 indicates that these two types of approaches are asymptotically equivalent. In fact, this statistical result is more general, and a related result for the optimal analysis of longitudinal data will appear elsewhere. Finally, in our experience with simulation studies and real data analysis, the GLS estimator is computationally much more efficient compared to the GMM estimator with case-control data, and therefore may be preferred in applications.

2.3.2 Testing for General Linear Hypothesis

Before introducing the specification test for model diagnostics, we provide a special result of hypothesis testing involving regression coefficients $\boldsymbol{\beta}$, due to the simplicity of the quadratic distance function (2.6). Testing for secondary trait association usually concerns the general linear hypotheses, i.e. $H_0 : \boldsymbol{C}\boldsymbol{\beta} = \boldsymbol{t}$, where $\boldsymbol{C}$ is a $k \times p$ contrast matrix and $\boldsymbol{t}$ is a k-vector of constants (e.g. the null vector). Following standard conventions, we require that the contrast matrix $\boldsymbol{C}$ is of

full row rank k ($k \leqslant p$) so that H_0 is testable. Since $\hat{\boldsymbol{\beta}}$ minimizes the quadratic distance function, we can treat $-M_n(\boldsymbol{\beta})/2$ as a pseudo-likelihood function, where the GLS estimator $\hat{\boldsymbol{\beta}}$ is the maximum pseudo-likelihood estimator (MPLE). Therefore, analogous to the tests developed for maximum likelihood, we define Wald, likelihood-ratio and score tests and denote these by S_{Wald}, S_{LR} and S_{score}, respectively. The Wald test examines whether $\hat{\boldsymbol{\beta}}$ satisfies H_0, while taking into consideration the retrospective sampling variability. With the consistent estimator for the asymptotic covariance matrix of $\hat{\boldsymbol{\beta}}$ as $(\boldsymbol{R}^T\hat{\boldsymbol{\Omega}}^{-1}\boldsymbol{R})^{-1}$, the Wald statistic is $S_{\text{Wald}} = n(\boldsymbol{C}\hat{\boldsymbol{\beta}} - \boldsymbol{t})^T[\boldsymbol{C}(\boldsymbol{R}^T\hat{\boldsymbol{\Omega}}^{-1}\boldsymbol{R})^{-1}\boldsymbol{C}^T]^{-1}(\boldsymbol{C}\hat{\boldsymbol{\beta}} - \boldsymbol{t})$. To form the other two test statistics, we require solving $\tilde{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}}[-M_n(\boldsymbol{\beta})/2]$ subject to H_0, which can be obtained analytically by introducing Lagrange multipliers as

$$\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}} - (\boldsymbol{R}^T\hat{\boldsymbol{\Omega}}^{-1}\boldsymbol{R})^{-1}\boldsymbol{C}^T[\boldsymbol{C}(\boldsymbol{R}^T\hat{\boldsymbol{\Omega}}^{-1}\boldsymbol{R})^{-1}\boldsymbol{C}^T]^{-1}(\boldsymbol{C}\hat{\boldsymbol{\beta}} - \boldsymbol{t}).$$

The likelihood-ratio statistic is then defined to be the difference between the distance function evaluated at the constrained estimator $\tilde{\boldsymbol{\beta}}$ and unconstrained estimator $\hat{\boldsymbol{\beta}}$, i.e., $S_{\text{LR}} = n[M_n(\tilde{\boldsymbol{\beta}}) - M_n(\hat{\boldsymbol{\beta}})]$. The score statistic assesses whether the constrained estimator $\tilde{\boldsymbol{\beta}}$ satisfies the gradient condition from the unconstrained minimization of the distance function, and is written as

$$S_{\text{score}} = n(\hat{\boldsymbol{\beta}}^{01} - \boldsymbol{R}\tilde{\boldsymbol{\beta}})^T\hat{\boldsymbol{\Omega}}^{-1}\boldsymbol{R}(\boldsymbol{R}^T\hat{\boldsymbol{\Omega}}^{-1}\boldsymbol{R})^{-1}\boldsymbol{R}^T\hat{\boldsymbol{\Omega}}^{-1}(\hat{\boldsymbol{\beta}}^{01} - \boldsymbol{R}\tilde{\boldsymbol{\beta}}).$$

Although these three tests are defined through different mechanisms, we show in the following theorem that they are in fact equivalent under case-control sampling, due to the simple structure of the quadratic distance function (2.6).

Theorem 2.3.2. *The test statistics, S_{Wald}, S_{LR} and S_{score} are numerically equivalent. Under H_0, they converge in distribution to $\chi^2(k)$. Under a sequence of local alternatives $H_{1,n} : \boldsymbol{C}\boldsymbol{\beta} = \boldsymbol{t} + \boldsymbol{\delta}/\sqrt{n}$, they converge in distribution to $\chi^2(k, \varpi)$ with non-centrality parameter $\varpi = \boldsymbol{\delta}^T\left[\boldsymbol{C}(\mathcal{J}_{11}^T\mathcal{E}^{-1}\mathcal{J}_{11})^{-1}\boldsymbol{C}^T\right]^{-1}\boldsymbol{\delta}$.*

Proof. See Section 2.7.3. □

2.4 A Specification Test

To develop the specification test, we assume that the population-level mean structure $E(Y|\boldsymbol{X})$ is correctly specified. This assumption helps clarify a target association estimand of interest, and allows us to focus on studying the misspecification of the disease model that generates the inverse probability weights. Specifically, the retrospective unbiasedness of the weighted estimating function critically depends on the specification of the disease probability $H(Y, \boldsymbol{X}, \boldsymbol{\gamma}_0)$, which in practice may be incorrectly specified. For example, if one misspecifies the disease model by $\tilde{H}(Y, \boldsymbol{X}, \tilde{\boldsymbol{\gamma}}_0)$ for some $\tilde{\boldsymbol{\gamma}}_0$, then the weighted estimating function may be biased, i.e., the retrospective moment conditions (2.2) and (2.3) no longer hold. A possible consequence of such misspecification is that there may not exist $\boldsymbol{\beta}$ in the parameter space satisfying both

$$E\left[\frac{(1-D)g(Y,\boldsymbol{X},\boldsymbol{\beta})}{1-\tilde{H}(Y,\boldsymbol{X},\tilde{\boldsymbol{\gamma}}_0)}\Big|D=0\right]=\mathbf{0}$$

$$E\left[\frac{Dg(Y,\boldsymbol{X},\boldsymbol{\beta})}{\tilde{H}(Y,\boldsymbol{X},\tilde{\boldsymbol{\gamma}}_0)}\Big|D=1\right]=\mathbf{0}.$$

Therefore, the GLS estimator will subject to bias and the absolute bias may depend on the degree of misspecification. With this in mind, we construct a specification test that shares the same spirit with the conventional specification test used in GMM inference. If the prospective estimating function $g(\cdot)$ is valid, and the weights are well approximated, we expect the estimates $\hat{\boldsymbol{\beta}}^0$ and $\hat{\boldsymbol{\beta}}^1$ to be close to each other and hence the minimum quadratic distance function should be close to zero. This suggests the use of $nM_n(\hat{\boldsymbol{\beta}})$ as the specification test statistic, which simplifies to,

$$nM_n(\hat{\boldsymbol{\beta}}) = n\hat{\boldsymbol{\beta}}^{01T}[\hat{\boldsymbol{\Omega}}^{-1} - \hat{\boldsymbol{\Omega}}^{-1}\boldsymbol{R}(\boldsymbol{R}^T\hat{\boldsymbol{\Omega}}^{-1}\boldsymbol{R})^{-1}\boldsymbol{R}^T\hat{\boldsymbol{\Omega}}^{-1}]\hat{\boldsymbol{\beta}}^{01}.$$

Accounting for retrospective sampling variability, it is straightforward to show that $nM_n(\hat{\boldsymbol{\beta}})$ converges in distribution to $\chi^2(p)$ if the weighted estimation functions are valid, i.e., under the null that the models are correctly specified. Intuitively, the test statistic $nM_n(\hat{\boldsymbol{\beta}})$ measures the multivariate discrepancy between $\hat{\boldsymbol{\beta}}^0$ and $\hat{\boldsymbol{\beta}}^1$, and will likely be extreme under incorrect specification of the disease model relative to what would be expected under correct specification.

On the other hand, we note that $\hat{\boldsymbol{\beta}}^0$ and $\hat{\boldsymbol{\beta}}^1$ are both consistent subsample estimators, hence an alternative specification test would be to directly test for equality between $\hat{\boldsymbol{\beta}}^0$ and $\hat{\boldsymbol{\beta}}^1$. Since $\sqrt{n}(\hat{\boldsymbol{\beta}}^{01} - \boldsymbol{R}\boldsymbol{\beta}_0)$ converges in distribution to $N(\mathbf{0}, \boldsymbol{\Omega})$, we must have $\sqrt{n}\bar{\boldsymbol{R}}^T\hat{\boldsymbol{\beta}}^{01}$ converges in distribution to $N(\mathbf{0}, \bar{\boldsymbol{R}}^T\boldsymbol{\Omega}\bar{\boldsymbol{R}})$ by the Delta method, where $\bar{\boldsymbol{R}} = (\boldsymbol{I}^{p\times p}, -\boldsymbol{I}^{p\times p})^T$. Then the desired test proceeds with the Wald statistic

$$T_n = n(\bar{\boldsymbol{R}}^T\hat{\boldsymbol{\beta}}^{01})^T(\bar{\boldsymbol{R}}^T\hat{\boldsymbol{\Omega}}\bar{\boldsymbol{R}})^{-1}(\bar{\boldsymbol{R}}^T\hat{\boldsymbol{\beta}}^{01}).$$

Though constructed differently, the above two specification tests are in fact numerically equivalent. We formally state this equivalence result in Theorem 2.4.1, whose proof is given in Appendix C.

Theorem 2.4.1. *The specification test statistics $nM_n(\hat{\boldsymbol{\beta}})$ and T_n are numerically equivalent. Further, the specification test statistic $nM_n(\hat{\boldsymbol{\beta}})$ is asymptotically independent of both the GLS estimator $\hat{\boldsymbol{\beta}}$ and the corresponding test statistics for general linear hypothesis: S_{Wald}, S_{LR} and S_{score}.*

Proof. See Section 2.7.4. □

Theorem 2.4.1 indicates that testing for model specification under the GLS inference framework is numerically equivalent to assessing the concordance (in terms of multivariate distance) between estimators derived from separate weighted estimating equations, hence there is no difference in choosing $nM_n(\hat{\boldsymbol{\beta}})$ or T_n. Additionally, the specification test statistic $nM_n(\boldsymbol{\beta})$ captures the residual information contained in the

sample that is unused for estimating $\boldsymbol{\beta}$ (Hall, 2004), and the fact that this residual information is asymptotically orthogonal to $\hat{\boldsymbol{\beta}}$ indicates an interesting perspective on the efficiency optimality of the GLS estimator (2.7). That is, if we replace $\hat{\boldsymbol{\Omega}}$ in (2.6) by some other arbitrary positive definite matrix and perform a similar asymptotic analysis, the resulting GLS estimator will be asymptotic correlated with the residual information captured by the corresponding specification test statistic, in which case we have not extracted all possible information about the parameter contained in the weighted estimating equations, leading to an inefficient GLS estimator.

2.5 Numerical Illustrations

2.5.1 Secondary Analysis of Peripheral Arterial Disease Case-Control Data

We apply the GLS estimator to a case-control sample of peripheral arterial disease (PAD), which was originally designed to identify the genetic variants associated with the risk of PAD. PAD is a common disease with a prevalence estimate of 12% in the population, and the prevalence is about equal in men and women Hiatt (2001). PAD cases are defined as having a post-exercise ankle-brachial index no larger than 0.9, a history of lower extremity re-vascularization, or having poorly-compressible leg arteries. Controls are identified as having no history of PAD or having an ankle-brachial index between 1.0 and 1.3. We obtain a final case-control sample of 3110 independent individuals with 1554 PAD cases and 1556 disease-free controls after data pre-processing (details omitted for brevity). Two SNPs, rs7025486 (DAB2IP gene, chromosome 9) and rs1051730 (CHRNA3 gene, chromosome 15) were found to be significantly associated with PAD in a study described by Thorgeirsson et al. Thorgeirsson et al. (2008). Since body mass index (BMI) was also collected in the case-control sample and has been shown to be correlated with PAD Ix et al. (2011), we consider examining the associations between these two SNPs with BMI as an illustrative secondary analysis. As previously noted, appropriate analysis strategy

should take into account the case-control sampling because of the association between the secondary trait BMI and the case-control status.

Table 2.1: Association analyses between rs7025486 and rs1051730 and BMI in the PAD case-control sample.

SNP	Method	$\hat{\beta}_1$	95% CI	Specification Test
rs7025486	GLS	0.079	$(-0.216, 0.374)$	0.837
	SPREG	0.071	$(-0.231, 0.373)$	–
rs1051730	GLS	0.076	$(-0.195, 0.348)$	0.567
	SPREG	0.093	$(-0.187, 0.372)$	–

In this illustrative data analysis, we implement both the GLS and commonly-used retrospective likelihood (SPREG) approaches to estimate the secondary associations. The SPREG approach explicitly accounts for the case-control sampling by formulating a retrospective likelihood, and is implemented in the `spreg` software (`http://dlin.web.unc.edu/software/spreg-2/`). Of note, both our GLS and the SPREG approaches assume a disease model $Pr(D = 1|Y, X)$ in order to make inference for $\boldsymbol{\beta}$. While GLS extract information from the estimating equations weighted by the inverse disease probability, SPREG additionally assumes normality for the secondary trait and relies on a likelihood involving the disease probability. For the GLS approach, we specify the prospective restricted moment model as $E(Y|X_G, X_{PC}) = \alpha + \beta_1 X_G + \beta_2 X_{PC}$, where Y is BMI, X_G is the minor allele count for each SNP, and X_{PC} is the first principal component estimated during the data pre-processing step. Based on the assessment of the scree plot, the first principal component is used to control for potential confounding by population substructure Price et al. (2006). The SPREG approach further assumes that Y is normally distributed conditional on X_G and X_{PC}. For both the GLS and SPREG approaches, we assume the prospective disease probability to be $\text{logit}[Pr(D = 1|Y, X_G, X_{PC})] = \gamma_1 + \gamma_2 X_G + \gamma_3 X_{PC} + \gamma_4 Y$. We present the association estimates, 95% confidence intervals and the p-values of the

specification test in Table 2.1. From Table 2.1, we conclude that the SNPs rs7025486 and rs1051730 do not have significant associations with BMI, without regard to the choices of methods. However, it is interesting to note that the GLS approach has slightly narrower confidence interval than SPREG. This optimality is not surprising since the distribution of BMI is known to be non-normal and skewed in existing literature Penman and Johnson (2006), which supports the use of the GLS approach because of its robustness to departure from the normality assumption Xing et al. (2016). However, we also observe that the lengths of intervals are close between the GLS and SPREG, so it is likely that the skewness of BMI is only slight. The specification tests for GLS estimators give non-significant p-values, and we infer that the assumption on the disease model are adequate (assuming the prospective restricted moment model is correctly formulated), an evidence further supporting the results from the GLS approaches for this secondary analysis.

2.5.2 Simulations

In this Section, we investigate the finite-sample operating characteristics of the specification test by using simulated case-control data. Specifically, we illustrate in a simulation study that the specification test developed in Section 2.4 has power against moderate misspecification of the disease model, but is relatively insensitive to mild misspecification of the disease model. We consider a single biallelic genotype X with an additive mode of inheritance and minor allele frequency 0.3 (hence X equals the number of minor alleles and takes values from 0, 1 or 2); the genotype X is related to a continuous secondary phenotype Y through $Y = \alpha + \beta X + \epsilon$, where $\alpha = 1$ is a fixed intercept, β is the genetic effect and ϵ is the error term. We fix $\beta = -0.12$ as the association between the genotype and the secondary phenotype. We generate ϵ from $N(0, 1)$ and, to model departure from normality, also consider a gamma distributions with shape parameter $4/\nu^2$, where the skewness parameter $\nu = 2$. We assume the

true disease model as $\text{logit}(Pr(D = 1|Y, X)) = \gamma_1 + \gamma_2 X + \gamma_3 Y + \gamma_4 XY$, where XY term represents an interaction effect between the genotype and the secondary phenotype on the disease. We fix $\gamma_2 = \log(1.2)$ and $\gamma_3 = \log(2)/2$, but remark that other choices of coefficients yield qualitatively similar results. We choose γ_1 so that the disease prevalence approximately 15% for the common disease scenario, and the disease prevalence is around 0.15% for the rare disease scenario. We also choose γ_4 so that the disease prevalence is the same across different data generating processes. For each scenario, we consider $n_1 = 1000$ cases and $n_0 = 1000$ controls, and simulate $10,000$ data replicates. When fitting the disease model, we assume $Pr(D = 1|Y, X)$ by omitting the interaction term. It is worth noting that when $\gamma_4 = 0$, we have used the correctly-specified disease model and the specification test should have a close-to-nominal type I error rate (which assesses the validity of the proposed test). When $\gamma_4 \neq 0$, the disease model is misspecified and hence it is desirable that the specification test has power when γ_4 deviates substantially from zero. As a comparison, we also implemented the SPREG for each simulated data replicate assuming the same disease model as used in GLS to assess its bias due to model misspecification. Throughout, we assume that the secondary outcome regression model is correctly formulated so that our discussion centers around a correctly-defined association estimand. The nominal level of the test is fixed at 0.05.

Table 2.2 presents the absolute bias of $\hat{\beta}$, coverage probability of the associated 95% confidence interval and power of the specification test when the X, Y interaction is omitted in the fitted disease model with a normal secondary phenotype. In other words, each scenario represented by a value of $\gamma_4 \neq 0$ describes the results under a misspecified disease model including only the main effects of X and Y. For either the common or rare disease scenario, with increasing magnitude of the X, Y interaction in the true disease model, the absolute bias of $\hat{\beta}$ increases, but the standard error estimates are all valid for all estimators (not shown), and so the coverage probabil-

Table 2.2: Simulation results when the X, Y interaction is omitted in the disease model with a normal secondary trait.

		Common Disease			Rare Disease		
γ_4	Method	Abs Bias	Coverage	Power	Abs Bias	Coverage	Power
-log(1.3)	GLS	0.086	0.306	0.919	0.124	0.062	0.919
	SPREG	0.089	0.265	–	0.129	0.038	–
-log(1.2)	GLS	0.060	0.595	0.617	0.086	0.323	0.617
	SPREG	0.062	0.558	–	0.090	0.261	–
-log(1.1)	GLS	0.030	0.866	0.203	0.044	0.762	0.198
	SPREG	0.032	0.843	–	0.048	0.715	–
0	GLS	0.000	0.952	0.048	0.000	0.947	0.049
	SPREG	0.000	0.956	–	0.001	0.951	–
log(1.1)	GLS	0.028	0.865	0.188	0.041	0.791	0.199
	SPREG	0.032	0.844	–	0.049	0.698	–
log(1.2)	GLS	0.050	0.702	0.564	0.076	0.470	0.561
	SPREG	0.058	0.590	–	0.096	0.185	–
log(1.3)	GLS	0.068	0.531	0.860	0.102	0.255	0.854
	SPREG	0.081	0.327	–	0.139	0.014	–

ity of the associated confidence interval declines. It is apparent that omitting such an interaction in the disease model has deleterious consequences for both GLS and SPREG approaches since they all present relatively large absolute bias (frequently larger absolute bias is observed with the SPREG estimator). However, the specification test is powerful to detect such misspecification, with the larger power given by increasing degrees of misspecification. Further, when $\gamma_4 = 0$ and so the disease model is correctly specified, the empirical type I error rate of the proposed specification test is close to nominal. The results for the gamma secondary phenotype with $\nu = 2$ are similar and presented in Table 2.3.

Since our methodology requires fitting the logistic disease model to case-control data, we further evaluate the proposed estimators when the true disease model falls outside the logistic family. The findings may help us assess the applicability of

Table 2.3: Simulation results when the X, Y interaction is omitted in the disease model with a gamma secondary trait ($\nu = 2$)

		Common Disease			Rare Disease		
γ_4	Method	Abs Bias	Coverage	Power	Abs Bias	Coverage	Power
-log(1.3)	GLS	0.096	0.142	0.917	0.136	0.007	0.931
	SPREG	0.104	0.192	–	0.187	0.002	–
-log(1.2)	GLS	0.068	0.417	0.633	0.100	0.104	0.664
	SPREG	0.078	0.451	–	0.148	0.041	–
-log(1.1)	GLS	0.035	0.803	0.185	0.054	0.579	0.238
	SPREG	0.043	0.799	–	0.090	0.431	–
0	GLS	0.000	0.945	0.047	0.001	0.948	0.052
	SPREG	0.002	0.949	–	0.001	0.951	–
log(1.1)	GLS	0.030	0.835	0.257	0.055	0.588	0.260
	SPREG	0.039	0.827	–	0.142	0.152	–
log(1.2)	GLS	0.052	0.611	0.675	0.101	0.131	0.744
	SPREG	0.071	0.566	–	0.334	0.000	–
log(1.3)	GLS	0.070	0.400	0.921	0.134	0.025	0.966
	SPREG	0.096	0.327	–	0.526	0.000	–

Table 2.4: Simulation results when the link function is misspecified in the disease model with a normal secondary trait.

		Common Disease			Rare Disease		
Link	Method	Abs Bias	Coverage	Power	Abs Bias	Coverage	Power
$t(1)$	GLS	0.004	0.945	0.036	0.000	0.949	0.016
	SPREG	0.005	0.946	–	0.000	0.951	–
$t(2)$	GLS	0.005	0.942	0.060	0.001	0.949	0.018
	SPREG	0.006	0.948	–	0.001	0.951	–
$t(4)$	GLS	0.001	0.949	0.050	0.002	0.946	0.035
	SPREG	0.002	0.950	–	0.002	0.947	–
$t(8)$	GLS	0.000	0.945	0.060	0.003	0.948	0.048
	SPREG	0.001	0.951	–	0.003	0.946	–
$t(16)$	GLS	0.003	0.943	0.086	0.000	0.941	0.068
	SPREG	0.003	0.949	–	0.001	0.950	–
Probit	GLS	0.003	0.937	0.135	0.008	0.904	0.345
	SPREG	0.005	0.947	–	0.023	0.892	–

Table 2.5: Simulation results when the link function is misspecified in the disease model with a gamma secondary trait ($\nu = 2$).

		Common Disease			Rare Disease		
Link	Method	Abs Bias	Coverage	Power	Abs Bias	Coverage	Power
$t(1)$	GLS	0.004	0.947	0.063	0.001	0.949	0.017
	SPREG	0.007	0.945	–	0.000	0.949	–
$t(2)$	GLS	0.005	0.946	0.052	0.001	0.946	0.022
	SPREG	0.004	0.949	–	0.000	0.950	–
$t(4)$	GLS	0.002	0.946	0.104	0.002	0.947	0.064
	SPREG	0.004	0.947	–	0.004	0.947	–
$t(8)$	GLS	0.001	0.938	0.141	0.003	0.947	0.075
	SPREG	0.009	0.944	–	0.008	0.945	–
$t(16)$	GLS	0.003	0.940	0.155	0.001	0.948	0.047
	SPREG	0.013	0.939	–	0.032	0.918	–
Probit	GLS	0.005	0.932	0.167	0.007	0.950	0.222
	SPREG	0.015	0.930	–	0.135	0.421	–

the proposed methods to more general settings. We now consider a true disease model with only X, Y main effects, but let the link function be the inverse t CDF with varying degrees of freedom and, in the limit, the inverse normal CDF (Probit). Results similar to Table 2.2 are presented in Table 2.4 with a normal secondary phenotype. Within the family of link functions considered, the absolute bias of the two estimators remain relatively small, with coverage probability close to the nominal level. Overall, the SPREG estimator appears to be more biased than our proposed estimators, especially when the true disease model involves a Probit link. Further, the specification tests are not sensitive to misspecification of the link function except in one case when the disease is rare and the true link is Probit (perhaps due to the inadequacy of the rare disease approximation under that setting). Overall, the logistic disease model reasonably approximates various kinds of disease models for the set of t links we have considered, leading to estimates of low bias. We note that the low power of the specification tests further motivates their use as they are insensitive to slight imperfections in estimating the inverse disease probability weights (which

leads to small bias in association estimates). Similar results for a gamma secondary trait with $\nu = 2$ are presented in Table 2.5.

2.6 Discussion

In this Chapter, we have introduced a GLS formulation of our previous weighting-based estimator for secondary analysis of case control data. Our weighting-based estimator has two major advantages. First, similar to the retrospective likelihood estimator (SPREG), the GLS estimator exploits additional information by formulating a disease probability model, and improves efficiency over the frequently used survey-weighted estimator, as explained in our previous study (Xing et al., 2016). Further, as the GLS estimator is derived from the unbiased estimating functions and dispenses with the full likelihood formulation, it has also been shown to be robust against violation of distributional assumptions on the secondary phenotypes (Xing et al., 2016). Second, a major concern of the weighting-based procedure is the adequacy of the weights, estimated from an assumed disease model. Motivated by connection between GLS and GMM, we propose to use the minimum distance function as a test statistic to assess the adequacy of disease model. Using simulated data, we illustrated that the specification test has power against moderate misspecification in a range of scenarios and remains insensitive to mild departure from the true disease model. Finally, as the correct specification of disease model is also critical for the consistency of the SPREG estimator, we note that our specification test could be used in conjunction with the SPREG to validate moment assumptions if SPREG is the preferred analytical approach. However, unlike SPREG (Lin and Zeng, 2009), the GLS estimator is capable of incorporating high-dimensional covariates in the restricted moment model without demanding computation, and may be preferred in the genome-wide analysis of secondary associations.

Although we illustrate our general methodology using a continuous secondary

phenotype, we note that our approach is well applicable to a range of settings. With a binary secondary phenotype, we can choose g to be the score function of the logistic model. With a count phenotype, we can let g be the score function of a Poisson log-linear model, and we expect GLS estimators to demonstrate robustness to over-dispersion, similar to our previous simulations for the continuous secondary phenotype (Xing et al., 2016). As current methodology on analyzing count secondary phenotype is lacking, the GLS approach may stand out as a computationally convenient method. Applications to quantile regression are also possible. To obtain the τ-th quantile association between the genotype and a continuous secondary phenotype, we may use $g(Y, \boldsymbol{X}, \boldsymbol{\beta}) = [\tau - I(Y \leqslant \boldsymbol{X}^T\boldsymbol{\beta})]\boldsymbol{X}$, which has potential to serve as a computational convenient alternative to the estimators discussed by Wei et al. Wei et al. (2016). Such a topic opens up a promising avenue for future research. Specifically, since this choice of estimating function is no longer smooth in $\boldsymbol{\beta}$, an asymptotic analysis requires modified regularity conditions and practical implementation also merits additional considerations.

2.7 Technical Proofs of the Theorems

2.7.1 Assumptions

The following regularity assumptions are required for proving the theorems.

Assumption 2.7.1 (Retrospective sampling)**.** *Suppose the disease prevalence,* $\lambda = Pr(D = 1)$*, is strictly between* 0 *and* 1*. The control subsample,* $\boldsymbol{Z}_{1:n_0} = \{(Y_i, \boldsymbol{X}_i, D_i = 0)\}_{i=1}^{n_0}$*, represents a random sample among the disease-free subpopulation, and the case subsample,* $\boldsymbol{Z}_{(n_0+1):n} = \{(Y_i, \boldsymbol{X}_i, D_i = 1)\}_{i=n_0+1}^{n}$*, is a random sample from the diseases subpopulation. Further, there exists positive constants* ρ_0 *and* ρ_1 *such that* $\lim_{n\to\infty} n_0/n = \rho_0$*,* $\lim_{n\to\infty} n_1/n = \rho_1$ *with* $\rho_0 + \rho_1 = 1$ *and* $n_0 + n_1 = n$*.*

Assumption 2.7.2 (Disease probability)**.** *There exists a unique* $q \times 1$ *vector* $\boldsymbol{\gamma}_0 \in$

int($\boldsymbol{\Theta}_{\boldsymbol{\gamma}}$) *with* $\boldsymbol{\Theta}_{\boldsymbol{\gamma}} \in \mathbb{R}^q$ *and compact, and a positive constant* ϵ *such that* $Pr(D = 1|Y, \boldsymbol{X}) = H(Y, \boldsymbol{X}, \boldsymbol{\gamma}_0)$*; the function* $H(Y, \boldsymbol{X}, \boldsymbol{\gamma}_0) > \epsilon$ *for all* $Y, \boldsymbol{X}$ *such that* $f(Y, \boldsymbol{X}|D = 1) > 0$ *and* $H(Y, \boldsymbol{X}, \boldsymbol{\gamma}_0) < 1 - \epsilon$ *for all* $Y, \boldsymbol{X}$ *satisfying* $f(Y, \boldsymbol{X}|D = 0) > 0$.

To present the subsequent conditions, we define a gradient matrix $\mathcal{J}$ and a covariance matrix (assumed positive definite) $\mathcal{V}$ as

$$\mathcal{J} = \sum_{d=0}^{1} \rho_d E\left[\frac{\partial \Upsilon(\boldsymbol{Z}, \boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}^T}\Big| D = d\right] = \begin{pmatrix} \mathcal{J}_{11} & \mathcal{J}_{12} \\ \mathbf{0} & \mathcal{J}_{22} \end{pmatrix},$$

$$\mathcal{V} = \sum_{d=0}^{1} \rho_d E\left[\Upsilon(\boldsymbol{Z}, \boldsymbol{\theta}_0)^{\otimes 2}|D = d\right] = \begin{pmatrix} \mathcal{V}_{11} & \mathcal{V}_{12} \\ \mathcal{V}_{21} & \mathcal{V}_{22} \end{pmatrix},$$

with the respective component matrices defined by,

$$\mathcal{J}_{11} = \sum_{d=0}^{1} \rho_d E\left[\frac{\partial \Psi(\boldsymbol{Z}, \boldsymbol{\beta}_0, \boldsymbol{\gamma}_0)}{\partial \boldsymbol{\beta}^T}\Big| D = d\right] \qquad \mathcal{J}_{22} = \sum_{d=0}^{1} \rho_d E\left[\frac{\partial \Phi(\boldsymbol{Z}, \boldsymbol{\gamma}_0)}{\partial \boldsymbol{\gamma}^T}\Big| D = d\right]$$

$$\mathcal{J}_{12} = \sum_{d=0}^{1} \rho_d E\left[\frac{\partial \Psi(\boldsymbol{Z}, \boldsymbol{\beta}_0, \boldsymbol{\gamma}_0)}{\partial \boldsymbol{\gamma}^T}\Big| D = d\right] \qquad \mathcal{V}_{11} = \sum_{d=0}^{1} \rho_d E\left[\Psi(\boldsymbol{Z}, \boldsymbol{\beta}_0, \boldsymbol{\gamma}_0)^{\otimes 2}|D = d\right]$$

$$\mathcal{V}_{12} = \sum_{d=0}^{1} \rho_d E\left[\Psi(\boldsymbol{Z}, \boldsymbol{\beta}_0, \boldsymbol{\gamma}_0)\Phi(\boldsymbol{Z}, \boldsymbol{\gamma}_0)^T|D = d\right] = \mathcal{V}_{21}^T$$

$$\mathcal{V}_{22} = \sum_{d=0}^{1} \rho_d E\left[\Phi(\boldsymbol{Z}, \boldsymbol{\beta}_0, \boldsymbol{\gamma}_0)^{\otimes 2}|D = d\right] - \sum_{d=0}^{1} \rho_d E\left[\Phi(\boldsymbol{Z}, \boldsymbol{\beta}_0, \boldsymbol{\gamma}_0)|D = d\right]^{\otimes 2}$$

Assumption 2.7.3 (Identifiability)**.** *There exists a unique* $p \times 1$ *vector* $\boldsymbol{\beta}_0 \in$ *int*($\boldsymbol{\Theta}_{\boldsymbol{\beta}}$) *with* $\boldsymbol{\Theta}_{\boldsymbol{\beta}} \in \mathbb{R}^p$ *and compact, such that* $E[\Psi_d(\boldsymbol{Z}, \boldsymbol{\beta}_0, \boldsymbol{\gamma}_0)|D = d] = \mathbf{0}$ *for both* $d = 0, 1$ *and*

$$\mathcal{V}_{11}^{-1}\left[\sum_{d=0}^{1} \rho_d E[\Psi(\boldsymbol{Z}, \boldsymbol{\beta}, \boldsymbol{\gamma}_0)|D = d]\right] \neq \mathbf{0}$$

for $\boldsymbol{\beta} \neq \boldsymbol{\beta}_0$.

Assumption 2.7.4 (Regularity conditions)**.**

1. *With probability* 1, $\Upsilon(\boldsymbol{Z}, \boldsymbol{\theta})$ *is continuous for all* $\boldsymbol{\theta} \in \boldsymbol{\Theta} = \boldsymbol{\Theta}_{\boldsymbol{\beta}} \times \boldsymbol{\Theta}_{\boldsymbol{\gamma}}$ *and is once continuously differentiable in an* ϵ*-neighborhood of* $\boldsymbol{\theta}_0$, $\mathcal{N}(\boldsymbol{\theta}_0, \epsilon)$.

2. *Denote* $\|\cdot\|$ *as the Frobenius norm, for both* $d = 0, 1$,

$$E\left[\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \|\Upsilon(\boldsymbol{Z}, \boldsymbol{\theta})\| \,\middle|\, D = d\right] < \infty, \qquad E\left[\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \|\Upsilon(\boldsymbol{Z}, \boldsymbol{\theta})\|^2 \,\middle|\, D = d\right] < \infty,$$

$$E\left[\sup_{\boldsymbol{\theta}\in N(\boldsymbol{\theta}_0,\epsilon)} \left\|\partial\Upsilon(\boldsymbol{Z}, \boldsymbol{\theta})/\partial\boldsymbol{\theta}^T\right\| \,\middle|\, D = d\right] < \infty.$$

2.7.2 Proof of Theorem 2.3.1

Using standard theory of estimating equations (van der Vaart, 1998), we can easily show that $\hat{\boldsymbol{\beta}}^0 \xrightarrow{p} \boldsymbol{\beta}_0$ and $\hat{\boldsymbol{\beta}}^1 \xrightarrow{p} \boldsymbol{\beta}_0$, and so it follows that $\hat{\boldsymbol{\beta}}^{01} \xrightarrow{p} \boldsymbol{R}\boldsymbol{\beta}_0$, where the rectangular matrix $\boldsymbol{R} = (\boldsymbol{I}^{p\times p}, \boldsymbol{I}^{p\times p})^T$ is defined in the main manuscript. By assumption $\hat{\boldsymbol{\Omega}} \xrightarrow{p} \boldsymbol{\Omega}$, so we conclude by continuous mapping theorem that $\hat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}_0$ from the GLS formulation $\hat{\boldsymbol{\beta}} = (\boldsymbol{R}^T\hat{\boldsymbol{\Omega}}^{-1}\boldsymbol{R})^{-1}\boldsymbol{R}^T\hat{\boldsymbol{\Omega}}^{-1}\hat{\boldsymbol{\beta}}^{01}$.

For asymptotic normality of $\hat{\boldsymbol{\beta}}$, we define the $2p \times 2p$ block diagonal matrix

$$\mathcal{Q} = \begin{pmatrix} \rho_0 E\left[\partial\Psi_0(\boldsymbol{Z}, \boldsymbol{\beta}_0, \boldsymbol{\gamma}_0)/\partial\boldsymbol{\beta}^T \middle| D = 0\right] & \mathbf{0} \\ \mathbf{0} & \rho_1 E\left[\partial\Psi_1(\boldsymbol{Z}, \boldsymbol{\beta}_0, \boldsymbol{\gamma}_0)/\partial\boldsymbol{\beta}^T \middle| D = 1\right] \end{pmatrix}.$$

We assume that $\mathcal{Q}$ is non-singular so that $\mathcal{Q}^{-1}$ exists. Observe that $\mathcal{J}_{11} = \mathcal{Q}\boldsymbol{R}$. Assuming the regularity conditions hold, we expand (note that Ψ_0 and Ψ_1 are has zero covariance)

$$\mathbf{0} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \begin{pmatrix} \Psi_0(\boldsymbol{Z}_i, \hat{\boldsymbol{\beta}}^0, \hat{\boldsymbol{\gamma}}) \\ \Psi_1(\boldsymbol{Z}_i, \hat{\boldsymbol{\beta}}^1, \hat{\boldsymbol{\gamma}}) \end{pmatrix}$$

around $\boldsymbol{\theta}_0$ and rearrange the terms to get

$$\sqrt{n}(\hat{\boldsymbol{\beta}}^{01} - \boldsymbol{R}\boldsymbol{\beta}_0) = -\mathcal{Q}^{-1}\left[\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Psi(\boldsymbol{Z}_i, \boldsymbol{\beta}_0, \boldsymbol{\gamma}_0)\right] - \mathcal{Q}^{-1}\mathcal{J}_{12}\left[\sqrt{n}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0)\right] + o_p(1). \tag{2.9}$$

Meanwhile, we have the following stochastic expansion for the logistic disease model

$$\sqrt{n}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0) = -\mathcal{J}_{22}^{-1}\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Phi(\boldsymbol{Z}_i, \boldsymbol{\gamma}_0)\right] + o_p(1). \tag{2.10}$$

Inserting (2.10) into (2.9), we obtain

$$\sqrt{n}(\hat{\boldsymbol{\beta}}^{01} - \boldsymbol{R}\boldsymbol{\beta}_0) = -\mathcal{Q}^{-1}\mathcal{L}\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Upsilon(\boldsymbol{Z}_i, \boldsymbol{\theta}_0)\right] + o_p(1) \xrightarrow{d} N(\mathbf{0}, \mathcal{Q}^{-1}\mathcal{E}\mathcal{Q}^{-T}), \tag{2.11}$$

where $\mathcal{E} = \mathcal{L}\mathcal{V}\mathcal{L}^T$, $\mathcal{L} = (\boldsymbol{I}^{2p\times 2p}, -\mathcal{J}_{12}\mathcal{J}_{22}^{-1})$, and the last result follows from Lindeberg-Feller central limit theorem and it is immediate that $\boldsymbol{\Omega} = \mathcal{Q}^{-1}\mathcal{E}\mathcal{Q}^{-T}$. Now since $\hat{\boldsymbol{\Omega}} \xrightarrow{p} \boldsymbol{\Omega}$, we have $\hat{\boldsymbol{\Omega}}^{-1} \xrightarrow{p} \boldsymbol{\Omega}^{-1}$ and $(\boldsymbol{R}^T\hat{\boldsymbol{\Omega}}^{-1}\boldsymbol{R})^{-1}\boldsymbol{R}^T\hat{\boldsymbol{\Omega}}^{-1} \xrightarrow{p} \left(\boldsymbol{R}^T\boldsymbol{\Omega}^{-1}\boldsymbol{R}\right)^{-1}\boldsymbol{R}^T\boldsymbol{\Omega}^{-1}$, a finite quantity, and so we can write

$$\begin{aligned}\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) &= -\left(\boldsymbol{R}^T\boldsymbol{\Omega}^{-1}\boldsymbol{R}\right)^{-1}\boldsymbol{R}^T\boldsymbol{\Omega}^{-1}\mathcal{Q}^{-1}\mathcal{L}\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Upsilon(\boldsymbol{Z}_i, \boldsymbol{\theta}_0)\right] + o_p(1) \\ &\xrightarrow{d} N(\mathbf{0}, \left[\boldsymbol{R}^T(\mathcal{Q}^{-1}\mathcal{E}\mathcal{Q}^{-T})^{-1}\boldsymbol{R}\right]^{-1}) = N\left(\mathbf{0}, (\mathcal{J}_{11}^T\mathcal{E}^{-1}\mathcal{J}_{11})^{-1}\right).\end{aligned} \tag{2.12}$$

To consistently estimate the above asymptotic variance matrix in case-control samples, we use the sample counterpart to replace the retrospective expectations.

From Hansen (1982) and Hall (2004), the asymptotic variance for $\hat{\boldsymbol{\theta}}_{\text{GMM}}$ is given by $(\mathcal{J}^T\mathcal{V}^{-1}\mathcal{J})^{-1}$. We now re-express the asymptotic variance for $\hat{\boldsymbol{\beta}}_{\text{GMM}}$ by a sequence of block matrix inversion arguments and show that it is identical to the asymptotic variance of the GLS estimator even after accounting for the uncertainty in estimating the weights from the disease model. Since the covariance matrix $\mathcal{V}$ is assumed finite and positive definite, $\mathcal{V}^{-1}$ exists and by block matrix inversion

$$\mathcal{V}^{-1} = \begin{pmatrix} \mathcal{V}^{11} & \mathcal{V}^{12} \\ \mathcal{V}^{21} & \mathcal{V}^{22} \end{pmatrix} = \begin{pmatrix} (\mathcal{V}_{11} - \mathcal{V}_{12}\mathcal{V}_{22}^{-1}\mathcal{V}_{21})^{-1} & -\mathcal{V}_{11}^{-1}\mathcal{V}_{12}(\mathcal{V}_{22} - \mathcal{V}_{21}\mathcal{V}_{11}^{-1}\mathcal{V}_{12})^{-1} \\ -\mathcal{V}_{22}^{-1}\mathcal{V}_{21}(\mathcal{V}_{11} - \mathcal{V}_{12}\mathcal{V}_{22}^{-1}\mathcal{V}_{21})^{-1} & (\mathcal{V}_{22} - \mathcal{V}_{21}\mathcal{V}_{11}^{-1}\mathcal{V}_{12})^{-1} \end{pmatrix}.$$

By matrix multiplication, we have

$$\mathcal{J}^T\mathcal{V}^{-1}\mathcal{J} = \begin{pmatrix} \mathcal{K}_{11} & \mathcal{K}_{12} \\ \mathcal{K}_{21} & \mathcal{K}_{22} \end{pmatrix},$$

where $\mathcal{K}_{11} = \mathcal{J}_{11}^T\mathcal{V}^{11}\mathcal{J}_{11}$, $\mathcal{K}_{12} = \mathcal{K}_{21}^T = \mathcal{J}_{11}^T(\mathcal{V}^{11}\mathcal{J}_{12} + \mathcal{V}^{12}\mathcal{J}_{22})$ and $\mathcal{K}_{22} = \mathcal{J}_{12}^T\mathcal{V}^{11}\mathcal{J}_{12} + \mathcal{J}_{22}^T\mathcal{V}^{21}\mathcal{J}_{12} + \mathcal{J}_{12}^T\mathcal{V}^{12}\mathcal{J}_{22} + \mathcal{J}_{22}^T\mathcal{V}^{22}\mathcal{J}_{22}$. The asymptotic variance of $\hat{\boldsymbol{\beta}}_{\text{GMM}}$ is the upper-left $p \times p$ sub-matrix of $(\mathcal{J}^T\mathcal{V}^{-1}\mathcal{J})^{-1}$, which by block matrix inversion is

$$\begin{aligned} \text{avar}(\hat{\boldsymbol{\beta}}_{\text{GMM}}) &= (\mathcal{J}_{11}^T\mathcal{V}^{11}\mathcal{J}_{11} - \mathcal{K}_{12}\mathcal{K}_{22}^{-1}\mathcal{K}_{21})^{-1} \\ &= \left[\mathcal{J}_{11}^T\left[\mathcal{V}^{11} - (\mathcal{V}^{11}\mathcal{J}_{12} + \mathcal{V}^{12}\mathcal{J}_{22})\mathcal{K}_{22}^{-1}(\mathcal{V}^{11}\mathcal{J}_{12} + \mathcal{V}^{12}\mathcal{J}_{22})\right]\mathcal{J}_{11}\right]^{-1} \\ &= \left[\mathcal{J}_{11}^T\mathcal{F}^{-1}\mathcal{J}_{11}\right]^{-1} \end{aligned}$$

Applying the Sherman-Morrison-Woodbury inversion formula (Golub and Van Loan, 1996),

$$\begin{aligned} \mathcal{F} &= \left[\mathcal{V}^{11} - (\mathcal{V}^{11}\mathcal{J}_{12} + \mathcal{V}^{12}\mathcal{J}_{22})\mathcal{K}_{22}^{-1}(\mathcal{V}^{11}\mathcal{J}_{12} + \mathcal{V}^{12}\mathcal{J}_{22})\right]^{-1} \\ &= \left(\mathcal{V}^{11}\right)^{-1} + \left[\mathcal{J}_{12} + \left(\mathcal{V}^{11}\right)^{-1}\mathcal{V}^{12}\mathcal{J}_{22}\right] \times \\ &\quad \left[\mathcal{K}_{22} - (\mathcal{V}^{11}\mathcal{J}_{12} + \mathcal{V}^{12}\mathcal{J}_{22})^T\left(\mathcal{V}^{11}\right)^{-1}(\mathcal{V}^{11}\mathcal{J}_{12} + \mathcal{V}^{12}\mathcal{J}_{22})\right]^{-1}\left[\mathcal{J}_{12} + \left(\mathcal{V}^{11}\right)^{-1}\mathcal{V}^{12}\mathcal{J}_{22}\right]^T \end{aligned}$$

Observe that

$$\begin{aligned} & \left[\mathcal{K}_{22} - (\mathcal{V}^{11}\mathcal{J}_{12} + \mathcal{V}^{12}\mathcal{J}_{22})^T\left(\mathcal{V}^{11}\right)^{-1}(\mathcal{V}^{11}\mathcal{J}_{12} + \mathcal{V}^{12}\mathcal{J}_{22})\right] \\ = \;& \mathcal{J}_{12}^T\mathcal{V}^{11}\mathcal{J}_{12} + \mathcal{J}_{22}^T\mathcal{V}^{21}\mathcal{J}_{12} + \mathcal{J}_{12}^T\mathcal{V}^{12}\mathcal{J}_{22} + \mathcal{J}_{22}^T\mathcal{V}^{22}\mathcal{J}_{22} - \\ & \left(\mathcal{J}_{22}^T\mathcal{V}^{21}\mathcal{J}_{12} + \mathcal{J}_{12}^T\mathcal{V}^{11}\mathcal{J}_{12} + \mathcal{J}_{22}^T\mathcal{V}^{21}\left(\mathcal{V}^{11}\right)^{-1}\mathcal{V}^{12}\mathcal{J}_{22} + \mathcal{J}_{12}^T\mathcal{V}^{12}\mathcal{J}_{22}\right) \\ = \;& \mathcal{J}_{22}^T\mathcal{V}^{22}\mathcal{J}_{22} - \mathcal{J}_{22}^T\mathcal{V}^{21}\left(\mathcal{V}^{11}\right)^{-1}\mathcal{V}^{12}\mathcal{J}_{22} = \mathcal{J}_{22}^T\mathcal{V}_{22}^{-1}\mathcal{J}_{22}, \end{aligned}$$

where the last equality comes from the fact that $\mathcal{V}_{22} = \left(\mathcal{V}^{22} - \mathcal{V}^{21}\left(\mathcal{V}^{11}\right)^{-1}\mathcal{V}^{12}\right)^{-1}$ by inverting $\mathcal{V}^{-1}$. Next, observe the following,

1. $\left(\mathcal{J}_{22}^T \mathcal{V}_{22}^{-1} \mathcal{J}_{22}\right)^{-1} = \mathcal{J}_{22}^{-1} \mathcal{V}_{22} \mathcal{J}_{22}^{-T}$,

2. The inverse of $\mathcal{V}$ can also be written as

$$\mathcal{V}^{-1} = \begin{pmatrix} \mathcal{V}^{11} & \mathcal{V}^{12} \\ \mathcal{V}^{21} & \mathcal{V}^{22} \end{pmatrix} = \begin{pmatrix} (\mathcal{V}_{11} - \mathcal{V}_{12}\mathcal{V}_{22}^{-1}\mathcal{V}_{21})^{-1} & -(\mathcal{V}_{11} - \mathcal{V}_{12}\mathcal{V}_{22}^{-1}\mathcal{V}_{21})^{-1}\mathcal{V}_{12}\mathcal{V}_{22}^{-1} \\ (\mathcal{V}_{22} - \mathcal{V}_{21}\mathcal{V}_{11}^{-1}\mathcal{V}_{12})^{-1}\mathcal{V}_{21}\mathcal{V}_{11}^{-1} & (\mathcal{V}_{22} - \mathcal{V}_{21}\mathcal{V}_{11}^{-1}\mathcal{V}_{12})^{-1} \end{pmatrix},$$

3. Since the aforementioned two representations are equal, we can write

$$\mathcal{V}^{21} = -\mathcal{V}_{22}^{-1}\mathcal{V}_{21}(\mathcal{V}_{11} - \mathcal{V}_{12}\mathcal{V}_{22}^{-1}\mathcal{V}_{21})^{-1} = -\mathcal{V}_{22}^{-1}\mathcal{V}_{21}\mathcal{V}^{11} \tag{2.13}$$

$$\mathcal{V}^{12} = -(\mathcal{V}_{11} - \mathcal{V}_{12}\mathcal{V}_{22}^{-1}\mathcal{V}_{21})^{-1}\mathcal{V}_{12}\mathcal{V}_{22}^{-1} = -\mathcal{V}^{11}\mathcal{V}_{12}\mathcal{V}_{22}^{-1} \tag{2.14}$$

   Note that (2.13) and (2.14) can be obtained from each other from the fact that $\mathcal{V}^{-1}$ is a symmetric (positive definite) matrix.

Using the above intermediate results, we have

$$\begin{aligned}
\mathcal{F} &= \left(\mathcal{V}^{11}\right)^{-1} + \left[\mathcal{J}_{12}\mathcal{J}_{22}^{-1} + \left(\mathcal{V}^{11}\right)^{-1}\mathcal{V}^{12}\right]\mathcal{V}_{22}\left[\mathcal{V}^{21}\left(\mathcal{V}^{11}\right)^{-1} + \mathcal{J}_{22}^{-1}\mathcal{J}_{12}^T\right] \\
&= \left(\mathcal{V}^{11}\right)^{-1} + \left[\mathcal{J}_{12}\mathcal{J}_{22}^{-1} - \mathcal{V}_{12}\mathcal{V}_{22}^{-1}\right]\mathcal{V}_{22}\left[-\mathcal{V}_{22}^{-1}\mathcal{V}_{21} + \mathcal{J}_{22}^{-1}\mathcal{J}_{12}^T\right] \\
&= \left(\mathcal{V}^{11}\right)^{-1} - \mathcal{J}_{12}\mathcal{J}_{22}^{-1}\mathcal{V}_{21} + \mathcal{V}_{12}\mathcal{V}_{22}^{-1}\mathcal{V}_{21} + \mathcal{J}_{12}\mathcal{J}_{22}^{-1}\mathcal{V}_{22}\mathcal{J}_{22}^{-1}\mathcal{J}_{12}^T - \mathcal{V}_{12}\mathcal{J}_{22}^{-1}\mathcal{J}_{12}^T \\
&= \mathcal{V}_{11} - \mathcal{J}_{12}\mathcal{J}_{22}^{-1}\mathcal{V}_{21} + \mathcal{J}_{12}\mathcal{J}_{22}^{-1}\mathcal{V}_{22}\mathcal{J}_{22}^{-1}\mathcal{J}_{12}^T - \mathcal{V}_{12}\mathcal{J}_{22}^{-1}\mathcal{J}_{12}^T = \mathcal{L}\mathcal{V}\mathcal{L}^T = \mathcal{E}.
\end{aligned}$$

Therefore, $\hat{\boldsymbol{\beta}}$ is asymptotically equivalent to $\hat{\boldsymbol{\beta}}_{\mathrm{GMM}}$, and also achieves the asymptotic efficiency bound imposed by the retrospective moment conditions.

2.7.3 Proof of Theorem 2.3.2

Observe that the likelihood-ratio statistic can be simplified to

$$\begin{aligned}
S_{\mathrm{LR}} &= 2n(\hat{\boldsymbol{\beta}}^{01} - \boldsymbol{R}\hat{\boldsymbol{\beta}})^T\hat{\boldsymbol{\Omega}}^{-1}\boldsymbol{R}(\boldsymbol{R}^T\hat{\boldsymbol{\Omega}}^{-1}\boldsymbol{R})^{-1}\boldsymbol{C}^T[\boldsymbol{C}(\boldsymbol{R}^T\hat{\boldsymbol{\Omega}}^{-1}\boldsymbol{R})^{-1}\boldsymbol{C}^T]^{-1}(\boldsymbol{C}\hat{\boldsymbol{\beta}} - \boldsymbol{t}) + S_{\mathrm{Wald}} \\
&= S_{\mathrm{Wald}}
\end{aligned}$$

since the first-order condition for minimizing the distance function is

$$\boldsymbol{R}^T\hat{\boldsymbol{\Omega}}^{-1}(\hat{\boldsymbol{\beta}}^{01} - \boldsymbol{R}\hat{\boldsymbol{\beta}}) = \boldsymbol{0}. \tag{2.15}$$

Similarly we can verify $S_{\text{score}} = S_{\text{Wald}}$ by using condition (2.15), thus the three tests statistics are numerically equivalent.

Under the null, $\boldsymbol{C}\boldsymbol{\beta}_0 - \boldsymbol{t} = \boldsymbol{0}$. By expansion (2.12), we observe that

$$\begin{aligned}\sqrt{n}(\boldsymbol{C}\hat{\boldsymbol{\beta}} - \boldsymbol{t}) =& \left[\sqrt{n}(\boldsymbol{C}\hat{\boldsymbol{\beta}} - \boldsymbol{t})\right] - \left[\sqrt{n}(\boldsymbol{C}\boldsymbol{\beta}_0 - \boldsymbol{t})\right] = \boldsymbol{C}\left[\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\right] \\ =& -\boldsymbol{C}\left(\boldsymbol{R}^T\boldsymbol{\Omega}^{-1}\boldsymbol{R}\right)^{-1}\boldsymbol{R}^T\boldsymbol{\Omega}^{-1}\mathcal{Q}^{-1}\mathcal{L}\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Upsilon(\boldsymbol{Z}_i, \boldsymbol{\theta}_0)\right] + o_p(1).\end{aligned} \tag{2.16}$$

Since $\hat{\boldsymbol{\Omega}}^{-1} \xrightarrow{p} \boldsymbol{\Omega}^{-1} = (\mathcal{Q}^T\mathcal{E}^{-1}\mathcal{Q})^{-1}$, the Wald statistic can be written as

$$\begin{aligned}S_{\text{Wald}} =& \boldsymbol{u}^T\mathcal{V}^{1/2}\mathcal{L}^T\mathcal{Q}^{-T}\boldsymbol{\Omega}^{-1}\boldsymbol{R}(\boldsymbol{R}^T\boldsymbol{\Omega}^{-1}\boldsymbol{R})^{-1}\boldsymbol{C}^T\left[\boldsymbol{C}(\boldsymbol{R}^T\hat{\boldsymbol{\Omega}}^{-1}\boldsymbol{R})^{-1}\boldsymbol{C}^T\right]^{-1} \times \\ & \boldsymbol{C}\left(\boldsymbol{R}^T\boldsymbol{\Omega}^{-1}\boldsymbol{R}\right)^{-1}\boldsymbol{R}^T\boldsymbol{\Omega}^{-1}\mathcal{Q}^{-1}\mathcal{L}\mathcal{V}^{1/2}\boldsymbol{u} + o_p(1),\end{aligned} \tag{2.17}$$

where $\boldsymbol{u}$ is a $2p+q$ dimensional, mean-zero and unit-variance Gaussian random vector. Note that the matrix sandwiched between $\boldsymbol{u}^T$ and $\boldsymbol{u}$ is a projection matrix with rank k since we could write $\boldsymbol{C}(\boldsymbol{R}^T\boldsymbol{\Omega}^{-1}\boldsymbol{R})^{-1}\boldsymbol{C}^T = \boldsymbol{\Lambda}^T\boldsymbol{\Lambda}$ with $\boldsymbol{\Lambda} = \boldsymbol{C}\left(\boldsymbol{R}^T\boldsymbol{\Omega}R\right)^{-1}\boldsymbol{R}^T\boldsymbol{\Omega}^{-1}\mathcal{Q}^{-1}\mathcal{L}\mathcal{V}^{1/2}$. Hence $S_{\text{Wald}} \xrightarrow{d} \chi^2(k)$. Further, under the sequence of Pitman local alternatives $H_{1,n}$, we could write

$$\begin{aligned}\sqrt{n}(\boldsymbol{C}\hat{\boldsymbol{\beta}} - \boldsymbol{t}) &= \left[\sqrt{n}(\boldsymbol{C}\hat{\boldsymbol{\beta}} - \boldsymbol{t})\right] - \left[\sqrt{n}(\boldsymbol{C}\boldsymbol{\beta}_0 - \boldsymbol{t} - \boldsymbol{\delta}/\sqrt{n})\right] = \boldsymbol{C}\left[\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\right] + \boldsymbol{\delta} \\ &= -\boldsymbol{C}\left(\boldsymbol{R}^T\boldsymbol{\Omega}^{-1}\boldsymbol{R}\right)^{-1}\boldsymbol{R}^T\boldsymbol{\Omega}^{-1}\mathcal{Q}^{-1}\mathcal{L}\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Upsilon(\boldsymbol{Z}_i, \boldsymbol{\theta}_0)\right] + \boldsymbol{\delta} + o_p(1),\end{aligned}$$

and the non-central Chi-squared distributional results in Theorem 2.3.2 follow readily.

2.7.4 Proof of Theorem 2.4.1

For equivalence between $nM_n(\hat{\boldsymbol{\beta}})$ and T_n, it suffices to check $\hat{\boldsymbol{\Omega}}^{-1} - \hat{\boldsymbol{\Omega}}^{-1}\boldsymbol{R}(\boldsymbol{R}^T\hat{\boldsymbol{\Omega}}^{-1}\boldsymbol{R})^{-1}\boldsymbol{R}^T\hat{\boldsymbol{\Omega}}^{-1} = \bar{\boldsymbol{R}}\left(\bar{\boldsymbol{R}}^T\hat{\boldsymbol{\Omega}}\bar{\boldsymbol{R}}\right)^{-1}\bar{\boldsymbol{R}}^T$, which is equivalent to

$$\hat{\boldsymbol{\Omega}} - \boldsymbol{R}(\boldsymbol{R}^T\hat{\boldsymbol{\Omega}}^{-1}\boldsymbol{R})^{-1}\boldsymbol{R}^T = \hat{\boldsymbol{\Omega}}\bar{\boldsymbol{R}}\left(\bar{\boldsymbol{R}}^T\hat{\boldsymbol{\Omega}}\bar{\boldsymbol{R}}\right)^{-1}\bar{\boldsymbol{R}}^T\hat{\boldsymbol{\Omega}}. \tag{2.18}$$

Since $\hat{\boldsymbol{\Omega}}$ is block diagonal as

$$\hat{\boldsymbol{\Omega}} = \begin{pmatrix} \hat{\boldsymbol{\Omega}}^0 & \mathbf{0} \\ \mathbf{0} & \hat{\boldsymbol{\Omega}}^1 \end{pmatrix},$$

we can multiply out the LHS of (2.18) and obtain

$$\begin{pmatrix} \hat{\boldsymbol{\Omega}}^0 - \left(\left(\hat{\boldsymbol{\Omega}}^0\right)^{-1} + \left(\hat{\boldsymbol{\Omega}}^1\right)^{-1}\right)^{-1} & -\left(\left(\hat{\boldsymbol{\Omega}}^0\right)^{-1} + \left(\hat{\boldsymbol{\Omega}}^1\right)^{-1}\right)^{-1} \\ -\left(\left(\hat{\boldsymbol{\Omega}}^0\right)^{-1} + \left(\hat{\boldsymbol{\Omega}}^1\right)^{-1}\right)^{-1} & \hat{\boldsymbol{\Omega}}^1 - \left(\left(\hat{\boldsymbol{\Omega}}^0\right)^{-1} + \left(\hat{\boldsymbol{\Omega}}^1\right)^{-1}\right)^{-1} \end{pmatrix}.$$

Further multiplying out the RHS of (2.18), we obtain

$$\begin{pmatrix} \hat{\boldsymbol{\Omega}}^0\left(\hat{\boldsymbol{\Omega}}^0 + \hat{\boldsymbol{\Omega}}^1\right)^{-1}\hat{\boldsymbol{\Omega}}^0 & -\hat{\boldsymbol{\Omega}}^0\left(\hat{\boldsymbol{\Omega}}^0 + \hat{\boldsymbol{\Omega}}^1\right)^{-1}\hat{\boldsymbol{\Omega}}^1 \\ -\hat{\boldsymbol{\Omega}}^1\left(\hat{\boldsymbol{\Omega}}^0 + \hat{\boldsymbol{\Omega}}^1\right)^{-1}\hat{\boldsymbol{\Omega}}^0 & \hat{\boldsymbol{\Omega}}^1\left(\hat{\boldsymbol{\Omega}}^0 + \hat{\boldsymbol{\Omega}}^1\right)^{-1}\hat{\boldsymbol{\Omega}}^1 \end{pmatrix}.$$

By the Sherman-Morrison-Woodbury inversion formula (Golub and Van Loan, 1996),

$$\left[\left(\hat{\boldsymbol{\Omega}}^0\right)^{-1} + \left(\hat{\boldsymbol{\Omega}}^1\right)^{-1}\right]^{-1} = \hat{\boldsymbol{\Omega}}^0 - \hat{\boldsymbol{\Omega}}^0\left(\hat{\boldsymbol{\Omega}}^0 + \hat{\boldsymbol{\Omega}}^1\right)^{-1}\hat{\boldsymbol{\Omega}}^0 = \hat{\boldsymbol{\Omega}}^1 - \hat{\boldsymbol{\Omega}}^1\left(\hat{\boldsymbol{\Omega}}^0 + \hat{\boldsymbol{\Omega}}^1\right)^{-1}\hat{\boldsymbol{\Omega}}^1.$$

Therefore the LHS and RHS of (2.18) are equal.

For asymptotic independence between $nM_n(\hat{\boldsymbol{\beta}})$ and $\hat{\boldsymbol{\beta}}$, note that we could write $nM_n(\hat{\boldsymbol{\beta}}) = \boldsymbol{u}^T\boldsymbol{N}_1\boldsymbol{u} + o_p(1)$, where

$$\boldsymbol{N}_1 = \mathcal{V}^{1/2}\mathcal{L}^T\mathcal{Q}^{-T}\left(\boldsymbol{\Omega}^{-1} - \boldsymbol{\Omega}^{-1}\boldsymbol{R}(\boldsymbol{R}^T\boldsymbol{\Omega}^{-1}\boldsymbol{R})^{-1}\boldsymbol{R}^T\boldsymbol{\Omega}^{-1}\right)\mathcal{Q}^{-1}\mathcal{L}\mathcal{V}^{1/2}$$

with $\boldsymbol{u}$ a $2p+q$ dimensional, mean zero and unit-variance Gaussian random vector. From (2.12), we have $\sqrt{n}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0) = \boldsymbol{N}_2\boldsymbol{u}+o_p(1)$, where $\boldsymbol{N}_2 = -\left(\boldsymbol{R}^T\boldsymbol{\Omega}^{-1}\boldsymbol{R}\right)^{-1}\boldsymbol{R}^T\boldsymbol{\Omega}^{-1}\mathcal{Q}^{-1}\mathcal{L}\mathcal{V}^{1/2}$. Asymptotic independence holds since $\boldsymbol{N}_2\boldsymbol{N}_1 = \boldsymbol{0}$. Further, it should be noted that $nM_n(\hat{\boldsymbol{\beta}})$ and the test statistic S_{Wald} are also asymptotically independent. From (2.17), $S_{\text{Wald}} = \boldsymbol{u}^T\boldsymbol{N}_3\boldsymbol{u} + o_p(1)$, where

$$\begin{aligned}\boldsymbol{N}_3 &= \mathcal{V}^{1/2}\mathcal{L}^T\mathcal{Q}^{-T}\boldsymbol{\Omega}^{-1}\boldsymbol{R}(\boldsymbol{R}^T\boldsymbol{\Omega}^{-1}\boldsymbol{R})^{-1}\boldsymbol{C}^T\left[\boldsymbol{C}(\boldsymbol{R}^T\hat{\boldsymbol{\Omega}}^{-1}\boldsymbol{R})^{-1}\boldsymbol{C}^T\right]^{-1}\times \\ &\quad \boldsymbol{C}\left(\boldsymbol{R}^T\boldsymbol{\Omega}^{-1}\boldsymbol{R}\right)^{-1}\boldsymbol{R}^T\boldsymbol{\Omega}^{-1}\mathcal{Q}^{-1}\mathcal{L}\mathcal{V}^{1/2}.\end{aligned}$$

The asymptotic independence hold since $\boldsymbol{N}_3\boldsymbol{N}_1 = \boldsymbol{0}$.

3

Double-Robust Estimation in Difference-in-Differences with an Application to Traffic Safety Evaluation

3.1 Introduction

3.1.1 Traffic Safety Evaluation

A central component in transportation research is the evaluation of safety impact from traffic safety countermeasures. From a statistical point of view, evaluating the effectiveness of traffic safety countermeasures is a causal inference problem, which refers to designs and methods for assessing intervention effect. Due to ethical and practical constraints with roadway safety experimentation, observational data are routinely used for safety evaluations. An example of traffic safety countermeasure is the application of rumble strips, which are longitudinal safety features installed along the center and the shoulder of roadway segments to prevent vehicle crashes associated with drowsy or sleepy driving. In 2009, the Federal Highway Administration (FHWA, `https://safety.fhwa.dot.gov/`) reported that more than half of the fatal crashes in the United States occurred after a driver crossed the centerline or

edge line of a roadway, with around two-thirds of these crashes located in rural areas (Federal Highway Administration, 2014). As a low-cost safety solution, rumble strips have been installed at many state-owned highways over the past decade. The rumble strips are engineered to alert the inattentive drivers through audible and vibratory warning as the vehicle strays across the center or edge line, and hence to improve the chance for a safe return to the current driveway (Anund et al., 2008). A number of empirical traffic studies have used observational data to seek statistical evidence on the effectiveness of rumble strips for mitigating crashes (Griffith, 1999; Persaud, Retting, and Lyon, 2004; Gårder and Davies, 2006; El-Basyouny and Sayed, 2012; Khan, Abdel-Rahim, and Williams, 2015). However, these existing statistical evidence depends critically on the model specifications and their ability to address selection bias, which represents a major challenge in non-experimental program evaluation studies (Heckman, 1998).

The before-after design is commonly used in traffic safety evaluation (Hauer, 1997). For instance, empirical studies collect crash outcomes for roadway segments with and without rumble strips before and after the installation period and synthesize these crash information to make a probabilistic statement on the safety impact. In observational before-after studies, the state-of-the-art practice is the empirical Bayes regression-based approach (Hauer, 1997; Persaud and Lyon, 2007), which can be implemented by following two steps. First, a crash frequency model is estimated among the reference sites without rumble strips. In this model, the crash counts are assumed to follow a negative binomial distribution with the mean parameter specified as a function of site-specific characteristics (AASHTO, 2010). Since the crash outcomes in the absence of rumble strips are unobserved among the sites receiving rumble strips in the after period, the empirical Bayes approach imputes these unobserved counts by a posterior mean, defined as a weighted average between the observed counts in the before period and the predicted counts by the fitted crash

frequency model. Although the empirical Bayes approach is considered the gold-standard in before-after designs, it does not provide a formal causal interpretation within the potential outcome framework (Rubin, 1974, 1978) and the effectiveness estimates may be subject to model misspecification (Elvik, 2002, 2008).

Within the potential outcome framework, the propensity score approach is a popular alternative to outcome regression (Rosenbaum and Rubin, 1983). The propensity score is defined as the conditional probability of receiving treatment given a collection of pre-treatment covariates, and is often regarded as a scalar summary of the multi-dimensional information. Rosenbaum and Rubin (1983) showed that balancing the propensity score leads to balancing all the covariates, under the unconfoundedness assumption and correct specification of the propensity score model. Using cross-sectional data obtained from the after period, a number of authors (Karwa et al., 2011; Wood et al., 2015; Wood and Donnell, 2016) considered propensity score weighting and matching estimators for safety evaluation problems and demonstrated their potential in removing selection bias. In this Chapter, we consider methods of difference-in-differences (DID) to evaluate the average causal effect of rumble strip installation on crash outcomes. Similar to previous developments, we use propensity score weighting and outcome regression to address selection bias; different from previous developments, our DID framework is particularly suitable to before-after designs and entails a natural causal interpretation.

3.1.2 Difference-in-Differences

Originated from the program evaluation literature, DID methods use data with a time dimension to control for unobserved but fixed confounding, and identify causal effects by contrasting the change in outcomes pre- and post-intervention, among the treated and control groups (Card and Krueger, 1994; Ashenfelter, 1978; Ashenfelter and Card, 1985). The central assumption underpinning the DID methods is the

parallel trend, that is, the counterfactual trend behavior of treatment and control groups, in the absence of treatment, is the same, possibly conditioning on some observed covariates (Heckman, Ichimura, and Todd, 1997). Under this assumption, the counterfactual outcome absent treatment can be identified among the treated, leading to estimators of the causal effect for the treatment group.

DID methods are attractive for our traffic safety application for the following three reasons. First, the DID design shares the same feature with the traffic safety before-after design, in which crash outcomes are collected pre- and post-intervention for both the treated and control sites. Second, the rumble strips were installed for selected pilot sites, and the scientific question of interest lies in whether the installation benefited those pilot sites in mitigating vehicle crashes. This suggests the average treatment effect among the treated (ATT) as the target estimand, which corresponds well with the target of inference granted by DID. Third, the causal inference framework provided by DID requires that the treated and control traffic sites are genuinely comparable (Heckman, 1998), thus avoiding undue extrapolation to irrelevant sites.

In the traffic safety application, we first consider a regression-based DID approach. Our outcome regression approach is operationally similar to the empirical Bayes before-after analysis, but it is formulated within the DID framework which enables a causal interpretation. As an alternative, we also consider the propensity score weighting DID estimator due to Abadie (2005). Since DID targets ATT as the causal estimand, we construct the balancing weights (Li, Morgan, and Zaslavsky, 2018) appropriate for ATT to re-weight the covariate distributions among the control sites to inform valid group comparisons. It is worth noting that correct specifications of the outcome model and the propensity score model are required for the corresponding regression and weighting methods to consistently estimate ATT. To improve upon these two estimators, we contribute to the literature by further developing an

augmented weighting estimator that hybridizes regression and weighting. This augmented estimator extends the work of Mercatanti and Li (2014) for cross-sectional data, and is doubly robust in that consistency requires either the outcome model or the propensity score model to be correctly specified, but not necessarily both, thus offering two chances to correctly quantify the safety impact of the countermeasures.

The remainder of this Chapter is organized as follows. Section 3.2 defines the causal estimands, and introduces DID estimators: outcome regression, propensity score weighting and the proposed double-robust estimators. Section 3.3 illustrates the DID estimators through simulations mimicking the traffic safety study. Section 3.4 presents the application to highway crash data collected by the Pennsylvania Department of Transportation. Section 3.5 concludes.

3.2 Causal Inference via Difference-in-Differences

3.2.1 Causal Estimands

We introduce the notation in the context of the evaluation of rumble strips (i.e., treatment). We consider the basic two-period two-group DID design. Assume a sample of traffic sites—units of analysis—indexed by $i = 1, \ldots, N$, belong to one of the two groups, with $G_i = 1$ indicating that rumble strips were applied in the after period, i.e. the treatment group, and $G_i = 0$ indicating that rumble strips were not applied in either period, i.e., the control group. Units in both groups are followed in two periods of time, with $T = t$ and $T = t + 1$ denoting the before and after period, respectively. For each unit i, let D_{iT} be the observed treatment status at period T. Since none of the traffic sites received treatment in the before period, we have $D_{it} = 0$ for all i. Because the treatment is only administered to one group ($G_i = 1$) in the after period, $D_{i,t+1} = 1$ for all units in group $G_i = 1$ and $G_i = D_{i,t+1}$ for all i. Similar to prior traffic safety evaluation studies (Karwa et al., 2011; Wood and Donnell, 2017), we make the Stable Unit Treatment Value

Assumption (SUTVA, Imbens and Rubin, 2015), meaning no interference between units and no different versions of the treatment. This assumption is more reasonable when the traffic sites are far apart from one another, but may be questionable when the sites are in close proximity. We will proceed with this assumption and discuss in Section 3.5 the implications when SUTVA is violated. Under SUTVA, each unit has two potential crash counts in each period, $Y_{iT}(0)$ and $Y_{iT}(1)$, and only the one corresponding to the observed treatment status, $Y_{iT} = Y_{iT}(D_{iT})$, is observed. The DID design allows us to write $Y_{it} = Y_{it}(0)$ and $Y_{i,t+1} = (1 - G_i)Y_{i,t+1}(0) + G_i Y_{i,t+1}(1)$. A vector of p pre-treatment variables, $\boldsymbol{X}_i$, is also observed for each unit. We denote the collection of observed data by $\mathcal{Z}_i = \{Y_{it}, Y_{i,t+1}, G_i, \boldsymbol{X}_i\}$, and assume that the $\mathcal{Z}_i$'s are independent and identically distributed from some common distribution $\mathbb{F}(\mathcal{Z})$.

In traffic safety studies, safety countermeasures are usually applied only to selected pilot sites before rolling out to a larger scale. The safety effectiveness is usually evaluated in a multiplicative fashion using the Crash Modification Factor (CMF, AASHTO, 2010), which can then be used to understand the expected change in crash frequency after a traffic safety countermeasure is implemented. In our traffic application, the rumble strip installation is implemented as part of a national safety improvement program, and the interest lies in quantifying its potential effectiveness among the sites where the rumble strips were installed. Similar to Wood and Donnell (2017), we formally define the CMF as a causal estimand that characterizes the ratio between the expected observed outcome after the installation and the expected counterfactual outcome had the countermeasure not been installed in the pilot sites. Using the potential outcome notation, we define the CMF

$$\tau_{\text{CMF}} \equiv \frac{E\big[Y_{i,t+1}(1)|G_i = 1\big]}{E\big[Y_{i,t+1}(0)|G_i = 1\big]} = \theta_1/\theta_0, \tag{3.1}$$

where we denote $\theta_1 = E\big[Y_{i,t+1}(1)|G_i = 1\big]$ and $\theta_0 = E\big[Y_{i,t+1}(1)|G_i = 0\big]$. Because the crash outcomes are count data, τ_{CMF} is a causal rate ratio that quantifies the relative

average change in crash counts due to rumble strip installation among the treated. The scale-free τ_{CMF} is a ratio version of the usual average treatment effect among the treated (ATT) estimand. Here, to characterize the causal rate difference in the absolute scale, we also define an additive version—the Crash Frequency Difference (CFD):

$$\tau_{\text{CFD}} \equiv E[Y_{i,t+1}(1) - Y_{i,t+1}(0)|G_i = 1] = \theta_1 - \theta_0. \tag{3.2}$$

We argue that using the pair of parameters $(\tau_{\text{CFD}}, \tau_{\text{CMF}})$ instead of τ_{CMF} alone presents a more complete picture of the effectiveness of safety countermeasure.

3.2.2 Assumptions

Estimands τ_{CFD} and τ_{CMF} are functions of θ_1 and θ_0. Under SUTVA, θ_1 is nonparametrically identified: $\theta_1 = E[Y_{i,t+1}|G_i = 1]$, with a consistent moment estimator

$$\hat{\theta}_1 = \sum_{i=1}^{n} G_i Y_{i,t+1} \Big/ \sum_{i=1}^{n} G_i. \tag{3.3}$$

In contrast, θ_0—the expected counterfactual outcome in the absence of treatment at time $T = t + 1$—must rely on additional restrictions to identify. Following the convention in DID design, we impose the parallel trend assumption,

Assumption 3.2.1. *(Parallel Trend) For each unit $i = 1, \ldots, N$,*

$$E[Y_{i,t+1}(0) - Y_{it}(0)|\boldsymbol{X}_i, G_i = 1] = E[Y_{i,t+1}(0) - Y_{it}(0)|\boldsymbol{X}_i, G_i = 0].$$

Assumption 3.2.1 imposes that, conditional on the pre-treatment covariates $\boldsymbol{X}_i$, the average outcomes in the treated and control groups, in the absence of treatment, would have followed a parallel path over time. The quantity θ_0 is therefore identified

under Assumption 3.2.1 as

$$\begin{aligned}\theta_0 =& E_{\boldsymbol{X}}\{E[Y_{i,t+1}(0)|\boldsymbol{X}_i, G_i = 1]|G_i = 1\} \\ =& E_{\boldsymbol{X}}\{E[Y_{it}(0)|\boldsymbol{X}_i, G_i = 1] + E[Y_{i,t+1}(0) - Y_{it}(0)|\boldsymbol{X}_i, G_i = 0]|G_i = 1\} \\ =& E[Y_{it}|G_i = 1] + E_{\boldsymbol{X}}\{E[Y_{i,t+1} - Y_{it}|\boldsymbol{X}_i, G_i = 0]|G_i = 1\},\end{aligned} \tag{3.4}$$

where both terms of the right hand side of the equation involve only expectations of observed data and are identified.

It is important to note that a direct DID estimator that uses

$$\hat{\theta}_0^{\text{direct}} = \frac{\sum_{i=1}^n G_i Y_{it}}{\sum_{i=1}^n G_i} + \frac{\sum_{i=1}^n (1 - G_i)(Y_{i,t+1} - Y_{it})}{\sum_{i=1}^n (1 - G_i)}. \tag{3.5}$$

to estimate θ_0 neglects the pre-treatment covariate information and is subject to selection bias. In fact, $\hat{\theta}_0^{\text{direct}}$ is only consistent to θ_0 under the unconditional version of the parallel trend assumption, i.e., $E[Y_{i,t+1}(0) - Y_{it}(0)|G_i = 1] = E[Y_{i,t+1}(0) - Y_{it}(0)|G_i = 0]$, which is arguably stronger than Assumption 3.2.1. On the other hand, unlike the standard unconfoundedness condition usually assumed for the cross-sectional data, Assumption 3.2.1 does not necessarily assume that $\boldsymbol{X}$ controls for all sources of confounding. Indeed, DID allows for unobserved confounders to affect treatment assignment as long as their impact on the potential outcomes is both separable and time-invariant (Lechner, 2011). Assumption 3.2.1 is generally untestable and may be questionable in practice. As an indirect way to assess the plausibility of parallel trend, in this application, we will conduct a "no treatment" evaluation by performing DID analyses for crash outcomes from two pre-treatment periods ($T = t - 1$ and $T = t$). Specifically, if the parallel trend assumption is plausible, that is,

$$E[Y_{it}(0) - Y_{i,t-1}(0)|\boldsymbol{X}_i, G_i = 1] = E[Y_{it}(0) - Y_{i,t-1}(0)|\boldsymbol{X}_i, G_i = 0],$$

then the estimated CFD and CMF based on time $T = t-1, t$ should be close to 0 and 1, respectively, because in reality rumble strips were not applied until after time t

and should have no causal effect for the pre-treatment outcomes. This idea is similar to the falsification endpoints or negative control idea in assessing unconfoundedness (Rosenbaum, 2002).

As in most ATT estimation, we also assume *weak overlap*, that is, each unit has a nonzero probability of receiving the control, $e(\boldsymbol{X}_i) \equiv \Pr(G_i = 1|\boldsymbol{X}_i) < 1$, where $e(\boldsymbol{X}_i)$ is the propensity score. The weak overlap assumption is directly testable by visually comparing the estimated propensity score distributions between the treatment groups.

3.2.3 Extant Methods: Regression and Weighting

Two main classes of existing estimating methods of DID are outcome regression and propensity score weighting. We first introduce a regression-based estimator specifically for count outcomes. To identify θ_0, we need to identify all components on the right hand side of equation (3.4). Similar to $\hat{\theta}_1$, the first term $E[Y_{it}|G_i = 1]$ in θ_0 can be consistently estimated by a moment estimator, $\sum_{i=1}^n G_i Y_{it} / \sum_{i=1}^n G_i$. The second term in θ_0 requires a regression model for the difference in crash counts $Y_{i,t+1} - Y_{it}$ given $\boldsymbol{X}_i$ among the control sites. Given that a regression model for the difference in counts is difficult to obtain, we separately assume a negative binomial model for each of the cross-sectional counts

$$\begin{aligned} \{Y_{it}(0)|\boldsymbol{X}_i, G_i = 0\} &\sim \text{NB}(\mu(\boldsymbol{X}_i;\boldsymbol{\beta}), \phi), \\ \{Y_{i,t+1}(0)|\boldsymbol{X}_i, G_i = 0\} &\sim \text{NB}(\nu(\boldsymbol{X}_i;\boldsymbol{\gamma}), \psi), \end{aligned} \tag{3.6}$$

where μ, ν are known smooth mean functions with parameter $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, and the variances are $\mathbb{V}(Y_{it}(0)|\boldsymbol{X}_i, G_i = 0) = \mu(\boldsymbol{X}_i;\boldsymbol{\beta}) + \mu^2(\boldsymbol{X}_i;\boldsymbol{\beta})/\phi$ and $\mathbb{V}(Y_{i,t+1}(0)|\boldsymbol{X}_i, G_i = 0) = \nu(\boldsymbol{X}_i;\boldsymbol{\gamma}) + \nu^2(\boldsymbol{X}_i;\boldsymbol{\gamma})/\psi$, with potentially different dispersion parameters ϕ and ψ. Model (3.6) is called the crash frequency model in traffic safety research (AASHTO, 2010). When the dispersion parameters approach infinity, model (3.6)

reduces to Poisson regression. As is evident from equation (3.4), the crash frequency model is only required for the control group, but not the treatment group. We obtain the maximum likelihood estimates of the parameters $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\gamma}}$ using the control sites data. Under SUTVA and Assumption 3.2.1, equation (3.4) suggests the following estimator for θ_0,

$$\hat{\theta}_0^{\text{reg}} = \frac{\sum_{i=1}^n G_i Y_{it}}{\sum_{i=1}^n G_i} + \frac{\sum_{i=1}^n G_i\{\nu(\boldsymbol{X}_i;\hat{\boldsymbol{\gamma}}) - \mu(\boldsymbol{X}_i;\hat{\boldsymbol{\beta}})\}}{\sum_{i=1}^n G_i}. \tag{3.7}$$

When the crash frequency model (3.6) is correctly specified, $\hat{\theta}_0^{\text{reg}}$ is a consistent estimator of θ_0, and thus $\hat{\tau}_{\text{CFD}}^{\text{reg}} = \hat{\theta}_1 - \hat{\theta}_0^{\text{reg}}$ and $\hat{\tau}_{\text{CMF}}^{\text{reg}} = \hat{\theta}_1/\hat{\theta}_0^{\text{reg}}$ are consistent for τ_{CFD} and τ_{CMF}, respectively.

The second estimator is based on weighting. Specifically, Abadie (2005) showed that under Assumptions 3.2.1 and weak overlap,

$$\theta_0 = \frac{1}{\pi} E\Big\{G_i Y_{it} + \frac{(1-G_i)(Y_{i,t+1} - Y_{it})e(\boldsymbol{X}_i))}{1 - e(\boldsymbol{X}_i)}\Big\}. \tag{3.8}$$

where $\pi = Pr(G_i = 1)$.

If the propensity score is correctly estimated by $e(\boldsymbol{X}_i;\hat{\boldsymbol{\alpha}})$, where $\boldsymbol{\alpha}$ is the parameter of the propensity score model, equation (3.8) suggests the following weighting estimator for θ_0:

$$\hat{\theta}_0^{\text{wt}} = \frac{\sum_{i=1}^n G_i Y_{it} w_i}{\sum_{i=1}^n G_i} + \frac{\sum_{i=1}^n (1-G_i)(Y_{i,t+1} - Y_{it})w_i}{\sum_{i=1}^n G_i}, \tag{3.9}$$

where $w_i = 1$ for the treated group and $w_i = e(\boldsymbol{X}_i;\hat{\boldsymbol{\alpha}})/[1 - e(\boldsymbol{X}_i;\hat{\boldsymbol{\alpha}})]$ for the control group. This further gives $\hat{\tau}_{\text{CFD}}^{\text{wt}} = \hat{\theta}_1 - \hat{\theta}_0^{\text{wt}}$ and $\hat{\tau}_{\text{CMF}}^{\text{wt}} = \hat{\theta}_1/\hat{\theta}_0^{\text{wt}}$. Re-weighting the observed crash counts by these ATT weights, we create a pseudo-population in which the covariates are balanced between treatment groups (**?**); the covariate balance consists the basis of valid group comparison. The weighting estimator avoids specifying

the distributions of outcomes, but is in general not as efficient as outcome regression if the outcome model is correctly specified.

3.2.4 Double-Robust Estimation

The consistency of the regression estimator and the weighting estimator depends on the correct specification of the outcome model and propensity score model, respectively. Here, we propose a new hybrid DID estimator that augments weighting with regression:

$$\hat{\theta}_0^{\mathrm{dr}} = \hat{\theta}_0^{\mathrm{wt}} + \frac{1}{\sum_{i=1}^n G_i} \sum_{i=1}^n \frac{(G_i - e(\boldsymbol{X}_i; \hat{\boldsymbol{\alpha}}))\{\nu(\boldsymbol{X}_i; \hat{\boldsymbol{\gamma}}) - \mu(\boldsymbol{X}_i; \hat{\boldsymbol{\beta}})\}}{1 - e(\boldsymbol{X}_i; \hat{\boldsymbol{\alpha}})}. \tag{3.10}$$

This estimator can alternatively be written as a regression estimator augmented with weighting as

$$\hat{\theta}_0^{\mathrm{dr}} = \hat{\theta}_0^{\mathrm{reg}} + \frac{\sum_{i=1}^n (1 - G_i)(\hat{R}_{i,t+1} - \hat{R}_{it}) w_i}{\sum_{i=1}^n G_i}, \tag{3.11}$$

where the residuals are defined as $\hat{R}_{i,t+1} = Y_{i,t+1} - \nu(\boldsymbol{X}_i; \hat{\boldsymbol{\gamma}})$, $\hat{R}_{it} = Y_{it} - \mu(\boldsymbol{X}_i; \hat{\boldsymbol{\beta}})$. Based on these two equivalent formulations, we establish the following large-sample robustness property in Proposition 3.2.2, and include the proof in the Appendix.

Proposition 3.2.2. *As the sample size $n \to \infty$, the proposed estimator $\hat{\theta}_0^{\mathrm{dr}}$ converges in probability to θ_0 if either $e(\boldsymbol{X}_i; \hat{\boldsymbol{\alpha}})$ is consistent to the true propensity score or both $\nu(\boldsymbol{X}_i; \hat{\boldsymbol{\gamma}})$ and $\mu(\boldsymbol{X}_i; \hat{\boldsymbol{\beta}})$ are consistent for the true mean functions.*

By proposition 3.2.2, we can obtain the DR estimators for τ_{CFD} and τ_{CMF} by $\hat{\tau}_{\mathrm{CFD}}^{\mathrm{dr}} = \hat{\theta}_1 - \hat{\theta}_0^{\mathrm{dr}}$ and $\hat{\tau}_{\mathrm{CMF}}^{\mathrm{dr}} = \hat{\theta}_1 / \hat{\theta}_0^{\mathrm{dr}}$. In fact, estimator (3.11) extends the DR estimator for ATT in the cross-sectional setting by Mercatanti and Li (2014), who point out that the DR estimator may serve as a diagnostic tool in practical applications. Specifically, if the DR estimate differs from the regression estimate but is similar to the weighting

estimate, it may suggest a potential misspecification of the regression function or lack of covariate overlap; if the DR estimate is close to the regression estimate but differs from the weighting estimate, it may suggest a potential misspecification of the propensity score model. We will exploit this diagnostic property of the DR estimate in the traffic safety study in Section 3.4.

Although our presentation of the DR estimator is centered around the traffic safety application, the DR estimator should apply equally well in conventional program evaluation studies, such as estimating the causal effect of job training program on earnings (Heckman et al., 1997; Heckman, 1998). In that case, the causal estimand is usually defined on the additive scale similar to (3.2). However, since the earning outcomes are treated as continuous variables, the predicted mean functions ν and μ in the DR estimator could simply be obtained from the two-way fixed-effects model used in Ashenfelter and Card (1985) and Abadie (2005) rather than from (3.6).

For estimating the additive causal estimand, τ_{CFD}, in the traffic study, the DR estimator $\hat{\tau}^{\text{dr}}_{\text{CFD}}$ differs from the existing double-robust estimator for ATE, in the sense that $\hat{\theta}^{\text{dr}}_0$ only requires estimating the outcome model among the control group but not the treated group. Further, the proposed estimator $\hat{\tau}^{\text{dr}}_{\text{CFD}}$ is indeed a member of the augmented inverse probability weighting (AIPW) estimator (Robins et al., 1994), but is distinct from the most efficient member, which necessarily requires an outcome model for the treated group. To obtain the most efficient AIPW-DID estimator, one could adapt the corresponding efficient AIPW estimator for estimating ATT designed for cross-sectional data (Yang and Ding, 2018), by essentially replacing their cross-sectional outcome with the before-after difference. Despite the efficiency advantage, we caution that such an estimator is not double-robust since it fails to be consistent to the target estimand once the propensity score model is misspecified. In traffic safety applications where the treated group often includes only a small number of pilot sites, the efficient DID estimator is less attractive because a count regression

(e.g. negative binomial regression is routinely used in traffic safety studies) model tends to be unstable with non-convergence issues. For these reasons, we focus on the double-robust DID estimator $\hat{\theta}_0^{\text{dr}}$ instead.

Since our estimator $\hat{\tau}_{\text{CFD}}^{\text{dr}}$ differs from the most efficient AIPW-DID estimator, we suspect that $\hat{\tau}_{\text{CFD}}^{\text{dr}}$ may not guarantee to be asymptotically more efficient than the weighting estimator when all models are correct. This is formalized in Proposition 3.2.3 with an additive causal estimand and when the true propensity score is known. Proposition 3.2.3 is not directly useful for inference as it assumes known propensity scores, but may provide insights for efficiency comparisons of the DID estimators in simulation experiments. Its proof is given in the Appendix.

Proposition 3.2.3. *For estimating the additive causal estimand, assuming the true propensity score is known, the ith influence function of the weighting estimator is*

$$\varphi_i^{wt} = \frac{(G_i - e(\boldsymbol{X}_i))(Y_{i,t+1} - Y_{it})}{\pi(1 - e(\boldsymbol{X}_i))} - \tau_{CFD}.$$

Further assuming the regression functions are known, the ith influence function of the double-robust estimator is

$$\varphi_i^{\text{dr}} = \frac{(G_i - e(\boldsymbol{X}_i))(Y_{i,t+1} - Y_{it} - \{\nu(\boldsymbol{X}_i) - \mu(\boldsymbol{X}_i)\})}{\pi(1 - e(\boldsymbol{X}_i))} - \tau_{CFD}.$$

The double-robust estimator is asymptotically at least as efficient as the weighting estimator only when $\mathbb{V}(\varphi_i^{wt}) - \mathbb{V}(\varphi_i^{dr}) \geqslant 0$. *However, this inequality does not always hold. The full expression of* $\mathbb{V}(\varphi_i^{wt}) - \mathbb{V}(\varphi_i^{dr})$ *is provided in the Appendix.*

Even though the DR estimator is more robust to model misspecification in the DID design, Proposition 3.2.3 suggests that it may be asymptotically less efficient than the weighting estimator even if all models are correctly specified. This is in sharp contrast to the existing results developed for estimating the average treatment

effect (ATE). In the latter case, it is well-known that the double-robust estimator is asymptotically at least as efficient as the propensity score weighting estimator when all models are correctly specified (Bang and Robins, 2005; Tsiatis, 2006).

In the traffic safety application, we use a logistic regression to estimate the propensity scores. Since both the logistic and negative binomial models are smooth parametric models, we use the nonparametric bootstrap (Efron and Tibshirani, 1993) to obtain the associated $(1-\alpha)$ confidence interval (CI) and hence account for the uncertainty in estimating the nuisance parameters. For example, the following two steps are carried out to arrive at the CI estimator for $\hat{\tau}^{\mathrm{dr}}$. First, we re-sample with replacement from the empirical distribution $\hat{\mathbb{F}}_N(\mathcal{Z})$ to obtain the bth $(b = 1, \ldots, B)$ bootstrap replicate, $\{\mathcal{Z}_j^b, j = 1, \ldots, N\}$, from which we compute $\hat{\tau}^{\mathrm{dr},b}$. We then estimate the $\alpha/2$th and $(1-\alpha/2)$th quantiles of the collection of the bootstrap estimates, $\{\hat{\tau}^{\mathrm{dr},b}, b = 1, \ldots, B\}$, to form the lower and upper confidence limits for $\hat{\tau}^{\mathrm{dr}}$. Since Y_{it} and $Y_{i,t+1}$ are repeated measurements from the same site in the before and after periods, there may be non-zero residual correlation between these crash counts. An advantage of the bootstrap procedure is that the correlation between repeated measurements are automatically taken into account by re-sampling the entire observed data vector $\mathcal{Z}_i$.

3.3 Simulations

To illustrate the performance of different DID estimators, we conduct a small simulation study that mimics the real rumble strip application. Specifically, we simulate under a two-period two-group design. Each simulation has $N = 2000$ units. Each unit has a binary covariate X_1 and a continuous covariate X_2, generated as follows:

$$X_1 \sim \text{Bernoulli}(0.25), \quad X_2|X_1 \sim \text{Normal}(2 + 6X_1, 2^2).$$

We simulate the treatment group label G_i independently from a Bernoulli distribution with success probability being the propensity score:

$$\text{logit}\{e(\boldsymbol{X})\} = -2.0 + X_1 - 0.2X_2 + 0.04X_2^2. \tag{3.12}$$

Under the true propensity score model, the marginal treatment prevalence is approximately 20%, resembling our real application.

We generate the potential crash counts from negative binomial models, with different mean functions but same dispersion parameter $\phi = 2.5$. Specifically, we assume

$$\begin{aligned} Y_t(0)|\boldsymbol{X}, G = 0 \sim \text{NB}(\mu_{00}(\boldsymbol{X}), \phi), \quad Y_t(0)|\boldsymbol{X}, G = 1 \sim \text{NB}(\mu_{01}(\boldsymbol{X}), \phi), \\ Y_{t+1}(0)|\boldsymbol{X}, G = 0 \sim \text{NB}(\nu_{00}(\boldsymbol{X}), \phi), \quad Y_{t+1}(1)|\boldsymbol{X}, G = 1 \sim \text{NB}(\nu_{11}(\boldsymbol{X}), \phi), \end{aligned}$$

with

$$\begin{aligned} \mu_{00}(\boldsymbol{X}) &= \exp(-2.0 + 0.4X_1 + 0.43X_2 - 0.022X_2^2), \\ \mu_{01}(\boldsymbol{X}) &= \exp(-3.0 + 0.3X_1 + 0.43X_2 - 0.022X_2^2), \\ \nu_{00}(\boldsymbol{X}) &= \exp(-1.9 + 0.5X_1 + 0.43X_2 - 0.022X_2^2), \\ \nu_{11}(\boldsymbol{X}) &= \exp(-2.5 + 0.1X_1 + 0.43X_2 - 0.022X_2^2). \end{aligned} \tag{3.13}$$

Under the parallel trend, the mean function of the counterfactual crash outcome for the treated sites is $\nu_{01}(\boldsymbol{X}) = \nu_{00}(\boldsymbol{X}) + \mu_{01}(\boldsymbol{X}) - \mu_{00}(\boldsymbol{X})$. The coefficients of the mean functions in (3.13) are informed by regression fit from analyzing the total crashes from the traffic safety application, and ensure that $\nu_{01}(\boldsymbol{X})$ is positive over the support of $\boldsymbol{X}$. The true values of CFD and CMF, evaluated in large samples, are -0.078 and 0.862, respectively.

We simulate 500 replicates based on the models specified above. For each replicate, we use $\hat{\theta}_1$ to estimate θ_1, but use different estimators for θ_0. We first use the direct moment estimator based on the observed sample averages given in equation

(3.5). This estimator ignores pre-treatment covariate information and is only valid when there is no selection bias, namely, when the parallel trend holds unconditionally on the pre-treatment covariates. It is used here to quantify the selection bias in the data generation process. Further, the following estimators are compared.

Outcome regression: we adopt the regression estimator in Equation (3.7) with correctly specified mean functions for $\mu(\boldsymbol{X})$ and $\nu(\boldsymbol{X})$. We also study the regression estimator with incorrectly specified mean functions that omit the linear term X_1 and the quadratic term X_2^2 in $\mu(\boldsymbol{X})$ and $\nu(\boldsymbol{X})$. These two estimators are labeled by REG and REG-mis, respectively.

Propensity score weighting: we consider the weighting estimator in Equation (3.9) with the correctly specified propensity score model, as well as the weighting estimator with an incorrectly specified propensity score model that omits X_1 and X_2^2 in Model (3.12). These two estimators are labeled by WT and WT-mis, respectively.

Double-Robust methods: we compare the DR estimator in Equation (3.10) with correctly specified propensity score and outcome models (DR), the DR estimator with correctly specified outcome regression model but incorrectly specified propensity score model that omits X_1 and X_2^2 (DR-po), the DR estimator with correctly specified propensity score model but incorrectly specified outcome regression model that omits X_1 and X_2^2 (DR-ps), and the DR estimator with propensity score and outcome regression models being both incorrectly specified (DR-mis).

Table 3.1 presents the absolute bias, root mean squared error (RMSE) of each point estimator and the coverage of the corresponding 95% bootstrap confidence intervals. Among all the estimators, the direct estimator is associated with the largest bias and RMSE and the lowest coverage in estimating both τ_{CFD} and $\log(\tau_{\text{CMF}})$. This is as expected because X_1 and X_2 affect both the treatment assignment and the potential outcomes, and induce selection bias. The DID regression, weighting, and DR estimators all present small and comparable bias when the corresponding

Table 3.1: Simulation results comparing DID estimators.

	τ_{CFD}			$\log(\tau_{\text{CMF}})$		
	Bias	RMSE	Coverage	Bias	RMSE	Coverage
Direct	13.4	14.5	33.4	27.6	30.5	38.4
REG	0.4	13.4	94.8	1.9	26.6	94.8
REG-mis	10.6	20.0	90.0	14.3	31.3	90.4
WT	0.2	14.1	95.6	2.6	27.7	95.6
WT-mis	4.7	10.0	90.8	9.8	20.7	91.0
DR	0.5	14.5	95.4	2.2	28.6	95.4
DR-po	0.4	13.4	94.6	2.0	26.6	94.8
DR-ps	2.6	15.8	95.8	1.1	30.0	95.6
DR-mis	7.0	16.7	91.8	9.2	27.6	92.0

* The results include absolute bias (Bias $\times 10^2$), root mean squared error (RMSE $\times 10^2$) and coverage of the 95% confidence interval (Coverage) associated with each estimator for estimating τ_{CFD} and $\log(\tau_{\text{CMF}})$ in the simulations. The confidence intervals are computed based on 500 bootstrap samples from each simulated data set.

models are correctly specified. When the outcome regression functions $\mu(\boldsymbol{X})$ and $\nu(\boldsymbol{X})$ are misspecified, the regression estimator shows inflated bias and RMSE, with reduced coverage. Similarly, misspecification of the propensity score model also leads to increased bias and sub-nominal coverage for the DID weighting estimator. In this simulation, the substantial reduction in the variance of the weights from a misspecified propensity score model appears to outweigh the inflation in bias, which explains the decreased RMSE associated with WT-mis relative to WT.

The simulation also demonstrates the double robustness property of the DID-DR estimator: when either the propensity score model or outcome model is misspecified, the DR estimator (DR-po and DR-ps) leads to small bias and nominal coverage for both estimands. Interestingly, in estimating the additive effect τ_{CFD}, the outcome model appears have a bigger impact on the DR estimator than the propensity score model. Specifically, when only the outcome model is correctly specified, the DR estimator performs very close to the DR estimator with both models being correctly

specified, but the DR estimator under-performs much if only the propensity score is correct. Similar phenomenon was previously observed in the DR estimation of ATT and ATE in the cross-sectional setting (e.g. Li, Zaslavsky, and Landrum, 2013). This pattern is not obvious for ratio estimand $\log(\tau_{\text{CMF}})$, likely because that the bias—an additive and scaled quantity by definition—of a scale-free ratio quantity does not fully capture the true discrepancy in estimating θ_1 and θ_0. Lastly, when both the propensity score and outcome models are misspecified, the DR estimator (DR-mis) results in inflated bias and under-coverage; nonetheless, even under this scenario, the misspecified DR estimator still outperforms the corresponding misspecified regression estimator with 46% and 6% reduction in relative bias for estimating the additive and ratio estimands, respectively. In addition, we observe in the simulations that the Monte Carlo variance of the DR estimator, when both models are correctly specified, is very close to that of the weighting estimator with a correct propensity score model. This phenomenon may be partially explained by Proposition 3.2.3. Specifically, under the current data generating process mimicking the traffic safety application, we found that the Monte Carlo estimate of $N^{-1}\{\mathbb{V}(\varphi_i^{\text{wt}}) - \mathbb{V}(\varphi_i^{\text{dr}})\} < 0$ is negative and close to zero (averaged across simulation iterations).

3.4 Application to the Pennsylvania Rumble Strip Data

3.4.1 The Data

Our application is based on the Federal Highway Administration Evaluation of Low-Cost Safety Improvements Pooled Fund Study (Lyon, Persaud, and Eccles, 2015). The study embraced a broader scope and focused on quantifying the safety effectiveness of the combined application of centerline and shoulder rumble strips in mitigating crash outcomes among two-lane rural road locations in Kentucky, Missouri, and Pennsylvania. We obtained the subset of traffic safety records from the Pennsylvania Department of Transportation (PennDOT; `http://www.penndot.gov/`), which

includes vehicle crash counts for traffic sites within the state of Pennsylvania up to 2012. Since each traffic site is a roadway segment, we use these two terms interchangeably. From 2009 to 2011, centerline and shoulder rumble strips were installed in 331 rural, undivided two-lane roadway segments for a total of over 200 miles. The control group consists of five times more sites that did not receive rumble strips before 2012 but had similar traffic volume. Therefore, the data we analyze consist of around 2000 rural highway segments, approximately 17% of which received the treatment. We define year 2008 as the before period and year 2012 as the after period.

We consider four types of crash outcomes: (1) fatal-plus-injury (FI)—crashes that involve at least one fatal or injured person; (2) property-damage-only (PDO)—crashes where no occupant was injured; (3) run-off-the-road (ROR)—crashes where a vehicle travels outside the trafficway and collides with a natural or artificial object in an area not intended for vehicles; this is a subset of the first two crash types; (4) total number of fatal-plus-injury and property-damage-only crashes (TOT). Table 3.2 presents the aggregated crash counts for each type among both treated and control sites in the before and after periods.

Table 3.2: Crash counts by type for both treated and control sites in the before and after periods.

	Treated ($N_1 = 331$)		Control ($N_0 = 1,655$)	
	Before	After	Before	After
FI[a]	78	77	441	436
PDO[b]	61	41	350	321
ROR[c]	22	21	123	143
TOT[d]	139	118	791	757

[a] FI: fatal-plus-injury;
[b] PDO: property-damage-only;
[c] ROR: run-off-the-road;
[d] TOT: total.

The pre-treatment covariates we consider are site-specific characteristics often suggested in constructing crash frequency models (AASHTO, 2010). These variables include the operational characteristic of a roadway segment, the speed limit (high speed if the posted limit is above 45 mph and low speed otherwise), as well as geometric features of a roadway: segment length in miles, pavement width (three categories), average shoulder width (three categories), number of driveways (three categories), existence of intersections (two categories), existence of curves (two categories) and average degree of curvature. An important covariate is AADT—the average annual daily traffic volume. Although strictly speaking AADT is a time-varying covariate, we found that in this application the AADT of the before and after periods are very similar across all sites; thus we assume AADT is time-invariant and take the before period value as the covariate. Table 3.3 presents the descriptive statistics of the covariates.

3.4.2 Model Specification

We estimate the propensity score by logistic regression including all the pre-treatment site characteristics. We adopt the power series specification for the continuous variables (AADT, Length and Curvature) with the optimal order of terms $1 \leqslant l \leqslant 5$ selected by leave-one-out-cross-validation. We choose to include up to the third-order terms of the continuous variables in the propensity score model since $l = 3$ corresponds to the lowest mean squared error for predicting treatment. The fitted propensity score model suggests that road segments with wider pavement and shoulder, low speed limit, at least one driveway, no intersections nor curves are more likely to receive rumble strip installation.

For both the regression and DR estimators, we model the cross-sectional means of potential outcomes among the reference sites during each period. AADT and segment length are transformed to the log scale, as is common practice in developing

Table 3.3: Definition of variables and their descriptive statistics by treatment group.

Variable	Definition	Treated	Control
AADT	Annual average daily traffic volume; vehicles per day	3,520 (2,628) [818, 15,033]	3,636 (2,495) [678, 15,379]
Length	Roadway segment length in miles	0.47 (0.16) [0.01, 0.75]	0.48 (0.13) [0.03, 0.76]
Width	Pavement width in feet		
	width $\leqslant 20$	20 (6.0)	346 (20.9)
	$20 <$ width $\leqslant 23$	169 (51.1)	828 (50.0)
	otherwise	142 (42.9)	481 (29.1)
Speed	Posted speed limit		
	low if limit $\leqslant 45$ mph	216 (65.3)	956 (57.9)
	high otherwise	115 (34.7)	696 (42.1)
Shoulder	Average shoulder width in feet		
	width $\leqslant 3$	88 (26.6)	867 (52.4)
	$3 <$ width $\leqslant 6$	191 (57.7)	643 (38.8)
	otherwise	52 (15.7)	145 (8.8)
Driveways	Number of driveways		
	no driveway	24 (7.2)	101 (6.1)
	$1 \leqslant$ number of driveways $\leqslant 10$	235 (71.0)	1,100 (66.5)
	otherwise	72 (21.8)	454 (27.4)
Intersections	Inclusion of intersections		
	No intersections	257 (77.6)	1,162 (70.2)
	At least 1 intersection	74 (22.4)	493 (29.8)
Curves	Existence of horizontal curves		
	No curves	143 (43.2)	701 (42.4)
	At least 1 curve	188 (56.8)	954 (57.6)
Curvature	Average degree of curvature	2.81 (3.59) [0, 25.28]	3.92 (7.59) [0, 132.30]

* Mean (st. dev), [Min, Max] values are given for each continuous variable and the number of traffic sites (percentages) are given for each level of the categorical variables. Total sample size $N = 1,986$.

crash frequency models in traffic safety research (AASHTO, 2010). To allow for over-dispersion, we use negative binomial regression to estimate model parameters. Specifically, we assume the conditional distributions of $Y_{it}(0)$ and $Y_{i,t+1}(0)$ given $\boldsymbol{X}_i$

in the reference group as in (3.6), where

$$\begin{aligned}\mu(\boldsymbol{X}) &= L^{\beta_L} \cdot \mathrm{AADT}^{\beta_{\mathrm{AADT}}} \cdot \exp\Big(\beta_0 + \sum_{j=1}^{J} \beta_j X_j\Big), \\ \nu(\boldsymbol{X}) &= L^{\gamma_L} \cdot \mathrm{AADT}^{\gamma_{\mathrm{AADT}}} \cdot \exp\Big(\gamma_0 + \sum_{j=1}^{J} \gamma_j X_j\Big).\end{aligned} \tag{3.14}$$

In (3.14), L denotes the segment length, AADT is the traffic volume and J is the number of remaining covariates (including dummy variables). We adopt the log-linear specification for the outcome model since it performs as well as its power series counterpart regarding mean squared error estimated by leave-one-out-cross-validation, and yet is computationally convenient without convergence issues.

3.4.3 Assessment of Overlap, Balance and Parallel Trend

We assess the weak overlap assumption by visually checking the overlap in the histograms of the estimated propensity scores for the treated and control sites. The left panel of Figure 3.1 includes a histogram of the estimated propensity score for the treated sites (blue) and the control sites (red). The histogram suggests satisfactory overlap between the two groups.

We further check the covariate balance in the original and weighted sample by calculating the absolute standardized difference (ASD) of each covariate (including up to the third-order term for each continuous variable) between the two treatment groups, defined as

$$\mathrm{ASD} = \left| \frac{\sum_{i=1}^{n} G_i X_i w_i}{\sum_{i=1}^{n} G_i w_i} - \frac{\sum_{i=1}^{n} (1-G_i) X_i w_i}{\sum_{i=1}^{n} (1-G_i) w_i} \right| \Big/ \sqrt{\frac{s_1^2}{N_1} + \frac{s_0^2}{N_0}}, \tag{3.15}$$

where N_1, N_0 are the number of treated and reference sites, s_1^2, s_0^2 are the variances of the unweighted covariate in the treated and control group, respectively. The

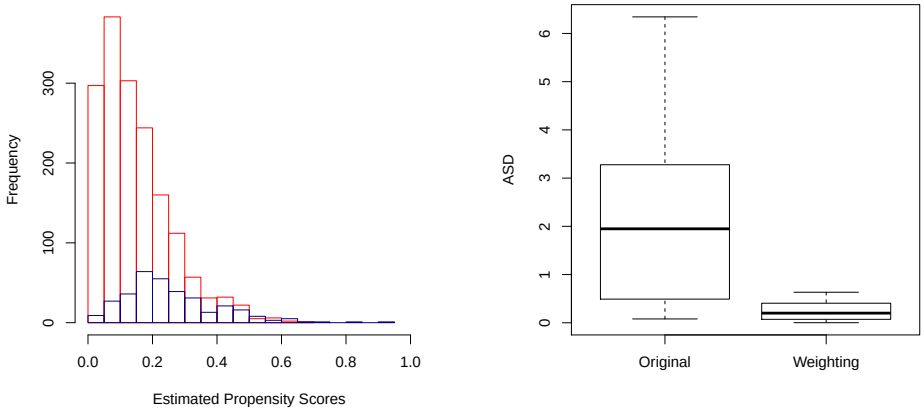


FIGURE 3.1: Assessment of overlap and covariate balance.

weight $w_i = 1$ for all sites in the original data and the ASD is the standard two-sample t-statistic. For the weighted data, w_i is the ATT weight introduced in Section 3.2.3. The right panel in Figure 3.1 presents the boxplots of the ASD; it shows that propensity score weighting substantially improves the covariate balance, with the largest ASD value equal to 0.63 in the weighted sample compared to 6.34 in the unweighted sample (the standard threshold for significant difference is 1.96). The good covariate balance supports that the propensity scores are well estimated.

To indirectly assess the key parallel trend assumption, we perform a DID analysis of the crash outcomes for two pre-treatment periods. Specifically, we obtain the crash outcome, $Y_{i,t-1}$, during the year of 2004 for each traffic site and treat it as the proxy-before observation; the crash outcome, Y_{it}, during the year of 2008 are then regarded as the proxy-after data. As discussed in Section 3.2.2, if the parallel trend assumption is plausible, then the estimated CFD and CMF based on the proxy-before-after observations should be close to 0 and 1, respectively, because in reality rumble strips were not applied until after 2008.

Table 3.4 presents the results of this "no treatment" analysis. For all crash types, DR and weighting estimators produce similar estimates for CFD and CMF. Overall,

Table 3.4: Estimated CFD ($\hat{\tau}_{\text{CFD}}$) and CMF ($\hat{\tau}_{\text{CMF}}$) and the 95% confidence intervals in the "no treatment" evaluation.

		Direct	REG	WT	DR
FI	$\hat{\tau}_{\text{CFD}}$	-0.060	-0.026	-0.029	-0.028
		(-0.144,0.026)	(-0.104,0.062)	(-0.121,0.073)	(-0.123,0.067)
	$\hat{\tau}_{\text{CMF}}$	0.798	0.902	0.891	0.892
		(0.587,1.107)	(0.651,1.331)	(0.621,1.370)	(0.623,1.341)
PDO	$\hat{\tau}_{\text{CFD}}$	0.008	0.010	0.026	0.027
		(-0.066,0.078)	(-0.066,0.086)	(-0.052,0.108)	(-0.052,0.110)
	$\hat{\tau}_{\text{CMF}}$	1.045	1.059	1.164	1.169
		(0.692,1.594)	(0.693,1.686)	(0.738,2.054)	(0.744,2.106)
ROR	$\hat{\tau}_{\text{CFD}}$	-0.016	0.001	0.007	0.009
		(-0.064,0.029)	(-0.051,0.046)	(-0.043,0.053)	(-0.044,0.054)
	$\hat{\tau}_{\text{CMF}}$	0.809	1.014	1.121	1.150
		(0.431,1.522)	(0.491,2.565)	(0.524,3.096)	(0.521,3.234)
TOT	$\hat{\tau}_{\text{CFD}}$	-0.052	-0.015	-0.003	-0.002
		(-0.146,0.073)	(-0.117,0.109)	(-0.119,0.140)	(-0.115,0.138)
	$\hat{\tau}_{\text{CMF}}$	0.890	0.965	0.993	0.996
		(0.714,1.192)	(0.761,1.333)	(0.764,1.427)	(0.762,1.425)

the confidence intervals for CFD include 0 for all types of crashes regardless of the choice of DID estimator. However, it is worth noting that the CFD estimates from DR and weighting for the ROR and total crashes are close to 0, which further support the plausibility of the parallel trend. By contrast, there is a potential for violation of parallel trend regarding FI and PDO crashes since the DR estimates for CFD tend to deviate from the null. Nevertheless, the lack of statistical significance may still permit the subsequent DID analyses.

3.4.4 Results

We analyzed crash outcomes in 2008 and 2012 using different DID estimators for all crash types and present the results in Table 3.5. As observed in the simulations, the direct estimator (3.5) is subject to selection bias and tends to give different results from the rest. Across all four crash types, the DR estimator produces CFD

Table 3.5: Estimated CFD ($\hat{\tau}_{\text{CFD}}$) and CMF ($\hat{\tau}_{\text{CMF}}$) and the 95% confidence intervals for all crash types with before and after data.

		Direct	REG	WT	DR
FI	$\hat{\tau}_{\text{CFD}}$	0.000	-0.022	-0.009	-0.008
		(-0.087,0.078)	(-0.118,0.060)	(-0.110,0.077)	(-0.108,0.079)
	$\hat{\tau}_{\text{CMF}}$	1.000	0.912	0.963	0.966
		(0.706,1.470)	(0.629,1.318)	(0.657,1.458)	(0.660,1.474)
PDO	$\hat{\tau}_{\text{CFD}}$	-0.043	-0.037	-0.056	-0.058
		(-0.113,0.026)	(-0.106,0.036)	(-0.134,0.022)	(-0.137,0.018)
	$\hat{\tau}_{\text{CMF}}$	0.743	0.770	0.687	0.681
		(0.455,1.223)	(0.462,1.384)	(0.409,1.196)	(0.404,1.180)
ROR	$\hat{\tau}_{\text{CFD}}$	-0.015	-0.022	-0.039	-0.039
		(-0.060,0.029)	(-0.066,0.022)	(-0.099,0.006)	(-0.096,0.006)
	$\hat{\tau}_{\text{CMF}}$	0.808	0.746	0.617	0.619
		(0.437,1.592)	(0.417,1.382)	(0.328,1.085)	(0.331,1.090)
TOT	$\hat{\tau}_{\text{CFD}}$	-0.043	-0.060	-0.065	-0.066
		(-0.145,0.063)	(-0.174,0.049)	(-0.188,0.052)	(-0.191,0.052)
	$\hat{\tau}_{\text{CMF}}$	0.893	0.856	0.845	0.844
		(0.684,1.189)	(0.651,1.154)	(0.627,1.166)	(0.626,1.154)

and CMF results similar to the weighting estimator, but sometimes different from the regression estimator. Given the satisfactory overlap indicated in Figure 3.1, the difference in estimates suggests that the outcome regression model may be mildly misspecified. The CFD and CMF for FI crashes estimated by the DR approach are both close to the null values, implying negligible effect from rumble strips on mitigating FI crashes. The application of rumble strips seems to reduce the PDO crashes with CFD and CMF estimated to be $\hat{\tau}^{\text{dr}}_{\text{CFD}} = -0.058$ and $\hat{\tau}^{\text{dr}}_{\text{CMF}} = 0.681$ using the DR approach. However, cautions need to be exercised to interpret these estimates since the parallel trend assumption may be questionable, as discussed previously. The parallel trend is deemed plausible for the total crashes, and we find that rumble strips have a potentially beneficial effect on total crash frequency ($\hat{\tau}^{\text{dr}}_{\text{CFD}} = -0.066$ and $\hat{\tau}^{\text{dr}}_{\text{CMF}} = 0.844$), but the 95% CIs include the null values. Additionally, the application

of rumble strips suggests a potential causal effect on mitigating the ROR crashes, with $\hat{\tau}^{\text{dr}}_{\text{CFD}} = -0.039$ and $\hat{\tau}^{\text{dr}}_{\text{CMF}} = 0.619$ estimated by the DR approach, but the CIs cover 0 and 1. Overall, our analysis only finds beneficial but statistically insignificant effects of rumble strips on reducing crashes. This agrees with the empirical findings of several other traffic safety studies based on alternative data sources and modeling strategies (Griffith, 1999; Gårder and Davies, 2006; Khan et al., 2015).

3.5 Discussion

In this Chapter, we draw causal inference in traffic safety before-after studies within the DID framework and propose a new double-robust DID estimator. The primary concern for observational traffic safety data is related to bias, which may be due to confounding, site selection or model misspecification, among others. Our DR estimator grants two chances for consistent estimation of the causal effect and has been demonstrated to have small bias from misspecification of either the propensity score model or the outcome model. Applying the DR method and several alternative methods to a real data, we find that rumble strips have a moderate but statistically insignificant beneficial effect in reducing vehicle crashes. These insignificant findings may be partially due to the limited number of crash events over a one-year period, a limitation of our available data. It would be of interest to update the CFD and CMF estimates with longer before and after periods.

Though the causal rate ratio estimand, CMF, dominates the traffic safety studies, we recommend assessing alternative estimand such as the causal rate difference, CFD to offer a more comprehensive picture of the effectiveness. This is because that CMF is scale-free and does not inform the absolute change in the expected crash frequency. For example, in our application, the CFD estimate suggests a modest absolute change in crash frequency ($\hat{\tau}^{\text{dr}}_{\text{CFD}} = -0.039$) for the ROR crashes, which can be translated into an average reduction of 4 crashes per 100 road segments due to rumble strips. On the

other hand, the CMF estimate is $\hat{\tau}^{\text{dr}}_{\text{CMF}} = 0.617$, which indicate a large proportional change. This slight discrepancy comes from the fact that the ROR crashes constitute a small fraction of the total crashes.

There are several limitations of this study. First, a limitation of the DID framework is that the parallel trend assumption is scale-dependent. For example, it may hold for the original Y but not for a nonlinear monotone transformation of Y, such as $\log(Y)$. A common alternative scale-free identification condition for the before-after design is the ignorability assumption conditional on the lagged outcomes. In the context of linear models, Angrist and Pischke (2009) show that the DID estimate and lagged-outcome regression estimates have a bracketing relationship. Namely, if ignorability is correct, then mistakenly assuming parallel trend will overestimate a true positive effect; in contrast, if parallel trend is correct, then mistakenly assuming ignorability will underestimate a true positive effect. Thus, one can treat the estimate under each assumption as the upper and lower bounds of the true effect in practice. It is particularly relevant to traffic safety studies—where the outcome is usually counts—to evaluate whether such a bracketing relationship holds more generally beyond the linear setting.

Second, the SUTVA may be violated and such a violation could lead to a biased average causal effect estimate. The violation is more likely if the traffic sites are adjacent to each other allowing for a potential spillover effect. For example, it is possible that a drowsy and fatigued driver was alerted in a roadway segment with shoulder rumble strips and hence was less likely to have a run-off-the-road crash in a nearby reference site, thus biasing the causal estimate towards the null. It is also likely that *crash migration* leads to violation of SUTVA. For instance, a vehicle travelling through a reference site with low visibility may end up in a crash in a consecutive site with rumble strips. However, the reporting officer usually traced the location where the crash was initiated by a careful analysis of the available evidence

at the crash site, and may mitigate such concerns. In any case, the extension of the DID analysis that accounts for interference between roadway segments in the spirit of Hudgens and Halloran (2008) would be of interest.

We have adopted smooth parametric models to estimate the propensity score and the crash counts. In this case, the resulting DID estimators are all asymptotically linear, and the nonparametric bootstrap enables valid inference (Shao and Tu, 2012). This also underlies why the bootstrap CI maintains nominal coverage for the DR-po and DR-ps estimators in our simulation study. In practice, since well-estimated propensity score and outcome models are critical for the consistency of the DR estimator, an appealing strategy is to leverage data-adaptive machine learning techniques for estimating the propensity scores and for predicting the crash counts. Specifically, one could use boosting to estimate the propensity score (McCaffrey et al., 2004, 2013), which has been demonstrated to maintain adequate weighted covariate balance (Lee, Lessler, and Stuart, 2009), or use random forest to better predict the counterfactual safety outcomes (Breiman, 2001; Liaw and Wiener, 2002). However, the nonparametric bootstrap CI may not guarantee to carry nominal coverage in those cases since the resulting estimator may no longer be asymptotically linear.

Finally, we have only developed double-robust estimation within the canonical two-period DID design with panel data. More complicated before-after data structure may arise in other policy evaluation contexts. For example, the treatment (policy) could be administered to a small number of states, and repeated cross sections or surveys are taken at the household level or individual level to measure the before and after outcomes. When repeated cross sections or random surveys are taken in both the before and after periods (rather than panel observations for the same group of units), the proposed double-robust DID estimator may not directly apply since the identification condition differs from equation (3.8). Abadie (2005) provided the revised identification condition and suggested the corresponding weighting estimator

(Section 3.2.1 of the paper). Therefore, an appropriate double-robust estimator is obtained by modifying the propensity score weighting component along the lines of Abadie (2005). Additionally, one must address the within-state correlations among households or individuals when the treatment is applied at the state level. In particular, valid bootstrap should proceed by resampling the states so that the within-state correlation structure is preserved (Li et al., 2013). In other program evaluation applications with staggered adoption and multiple time periods, the two-way fixed-effects model represents a standard regression estimator for causal inference. Callaway and Sant'Anna (2018) recently defined several new average causal estimands appropriate for the multiple-period DID design, and studied a propensity score weighting estimator. An important avenue for future research is to provide a double-robust DID extension by combining the weighting estimator of Callaway and Sant'Anna (2018) and the two-way fixed-effects outcome model for improved inference with multiple-period data.

3.6 Technical Proofs of the Theorems

3.6.1 Proof of Theorem 3.2.2

The DR estimators are constructed as $\hat{\tau}^{\text{dr}}_{\text{CFD}} = \hat{\theta}_1 - \hat{\theta}^{\text{dr}}_0$ and $\hat{\tau}^{\text{dr}}_{\text{CMF}} = \hat{\theta}_1/\hat{\theta}^{\text{dr}}_0$; the moment-based estimator

$$\hat{\theta}_1 = \frac{\sum_{i=1}^n G_i Y_{i,t+1}}{\sum_{i=1}^n G_i} \xrightarrow{p} \frac{E[G_i Y_{i,t+1}]}{\pi} = E[Y_{i,t+1}(1)|G_i = 1] = \theta_1, \tag{3.16}$$

where $\pi = Pr(G_i = 1) > 0$. To show that $\hat{\tau}^{\text{dr}}_{\text{CFD}}$ and $\hat{\tau}^{\text{dr}}_{\text{CMF}}$ are double-robust for estimating τ_{CFD} and τ_{CMF}, it suffices to show that $\hat{\theta}^{\text{dr}}_0$ is double-robust for estimating θ_0.

We first assume that the propensity score model $e(\boldsymbol{X}; \boldsymbol{\alpha})$ is correctly specified while the outcome model may be subject to misspecification. We assume certain regularity conditions hold (e.g., smooth regression functions and bounded moments for

all covariates), and denote the maximum likelihood estimators for model parameters by $\hat{\boldsymbol{\alpha}}$, $\hat{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\beta}}$. Under these assumptions, $\hat{\boldsymbol{\alpha}} \xrightarrow{p} \boldsymbol{\alpha}_0$, $\hat{\boldsymbol{\gamma}} \xrightarrow{p} \boldsymbol{\gamma}^*$, $\hat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}^*$, where $\boldsymbol{\alpha}_0$ is the true value for the correct propensity score model but $\boldsymbol{\gamma}^*$, $\boldsymbol{\beta}^*$ may be different from the true values $\boldsymbol{\gamma}_0$, $\boldsymbol{\beta}_0$. By the results of White (1982), $\boldsymbol{\gamma}^*$ and $\boldsymbol{\beta}^*$ are the least false values that minimize the Kullback-Leibler distance between the probability distribution based on the postulated models and the true data generating models. We first observe that the last term on the right hand side of equation (3.10) converges in probability to zero. To see why, we write

$$
\begin{aligned}
&\frac{1}{\sum_{i=1}^n G_i} \sum_{i=1}^n \frac{(G_i - e(\boldsymbol{X}; \hat{\boldsymbol{\alpha}}))\{\nu(\boldsymbol{X}_i; \hat{\boldsymbol{\gamma}}) - \mu(\boldsymbol{X}_i; \hat{\boldsymbol{\beta}})\}}{1 - e(\boldsymbol{X}; \hat{\boldsymbol{\alpha}})} \\
\xrightarrow{p} &\frac{1}{\pi} E\Big[\frac{(G_i - e(\boldsymbol{X}; \boldsymbol{\alpha}_0))\{\nu(\boldsymbol{X}_i; \boldsymbol{\gamma}^*) - \mu(\boldsymbol{X}_i; \boldsymbol{\beta}^*)\}}{1 - e(\boldsymbol{X}; \boldsymbol{\alpha}_0)} \Big] \\
= &\frac{1}{\pi} E\Big[\frac{\{E(G_i | \boldsymbol{X}_i) - e(\boldsymbol{X}; \boldsymbol{\alpha}_0)\}\{\nu(\boldsymbol{X}_i; \boldsymbol{\gamma}^*) - \mu(\boldsymbol{X}_i; \boldsymbol{\beta}^*)\}}{1 - e(\boldsymbol{X}; \boldsymbol{\alpha}_0)} \Big] = 0,
\end{aligned}
$$

where the second to last equation is an application of the Law of Iterated Expectation. Therefore by (3.10), it is immediate that $\hat{\theta}_0^{\mathrm{dr}}$ shares the same probability limit with $\hat{\theta}_0^{\mathrm{wt}}$, which is consistent to θ_0 when the propensity score model is correctly specified (Abadie, 2005). This is why $\hat{\theta}_0^{\mathrm{dr}} \xrightarrow{p} \theta_0$.

Alternatively, suppose the outcome model is correctly specified but the propensity score model may subject to misspecification. In this case, $\hat{\boldsymbol{\gamma}} \xrightarrow{p} \boldsymbol{\gamma}_0$, $\hat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}_0$, $\hat{\boldsymbol{\alpha}} \xrightarrow{p} \boldsymbol{\alpha}^*$, where $\boldsymbol{\alpha}^*$ minimizes the Kullback-Leibler distance between the probability distribution based on the postulated model and the true data generating model (White, 1982) and thus may differ from truth data generating model parameter $\boldsymbol{\alpha}_0$. Then the last term on the right hand side of (3.11) converges in probability to zero.

This is because

$$\frac{\sum_{i=1}^{n}(1-G_i)(\hat{R}_{i,t+1}-\hat{R}_{it})w_i}{\sum_{i=1}^{n}G_i}$$

$$=\frac{\sum_{i=1}^{n}(1-G_i)}{\sum_{i=1}^{n}G_i}\frac{1}{\sum_{i=1}^{n}(1-G_i)}\sum_{i=1}^{n}\frac{(1-G_i)(\hat{R}_{i,t+1}-\hat{R}_{it})e(\boldsymbol{X};\hat{\boldsymbol{\alpha}})}{1-e(\boldsymbol{X};\hat{\boldsymbol{\alpha}})}$$

$$\xrightarrow{p}\frac{1-\pi}{\pi}E\Big[\frac{(Y_{i,t+1}-Y_{it})e(\boldsymbol{X};\boldsymbol{\alpha}_0)}{1-e(\boldsymbol{X};\boldsymbol{\alpha}_0)}\Big|G_i=0\Big]-\frac{1-\pi}{\pi}E\Big[\frac{\{\nu(\boldsymbol{X}_i;\boldsymbol{\gamma}_0)-\mu(\boldsymbol{X}_i;\boldsymbol{\beta}_0)\}e(\boldsymbol{X}_i)}{1-e(\boldsymbol{X};\boldsymbol{\alpha}_0)}\Big|G_i=0\Big].$$

and

$$\Big[\frac{(Y_{i,t+1}-Y_{it})e(\boldsymbol{X};\boldsymbol{\alpha}_0)}{1-e(\boldsymbol{X};\boldsymbol{\alpha}_0)}\Big|G_i=0\Big]$$

$$=E\Big[\frac{\{Y_{i,t+1}(0)-Y_{it}(0)\}e(\boldsymbol{X};\boldsymbol{\alpha}_0)}{1-e(\boldsymbol{X};\boldsymbol{\alpha}_0)}\Big|G_i=0\Big]$$

$$=E\Big[\frac{e(\boldsymbol{X};\boldsymbol{\alpha}_0)}{1-e(\boldsymbol{X};\boldsymbol{\alpha}_0)}E[Y_{i,t+1}(0)-Y_{it}(0)|\boldsymbol{X}_i,G_i=0]\Big|G_i=0\Big]$$

$$=E\Big[\frac{e(\boldsymbol{X};\boldsymbol{\alpha}_0)}{1-e(\boldsymbol{X};\boldsymbol{\alpha}_0)}\{\nu(\boldsymbol{X}_i;\boldsymbol{\gamma}_0)-\mu(\boldsymbol{X}_i;\boldsymbol{\beta}_0)\}\Big|G_i=0\Big],$$

where the second to last equality is granted by the Law of Iterated Expectation and the last equality comes from the definition of the regression function. By (3.11), it follows that $\hat{\theta}_0^{\mathrm{dr}}$ shares the same probability limit with $\hat{\theta}_0^{\mathrm{reg}}$, which is consistent to θ_0 when the cross-sectional crash frequency model is correctly specified. Therefore $\hat{\theta}_0^{\mathrm{dr}} \xrightarrow{p} \theta_0$, and the double-robust property holds.

3.6.2 *Proof of Theorem 3.2.3*

Denote the ith known true propensity score as $e(\boldsymbol{X}_i)$, then the weighting estimator is

$$\begin{aligned}\tau_{\text{CFD}}^{\text{wt}} &= \frac{\sum_{i=1}^n G_i(Y_{i,t+1}-Y_{it})}{\sum_{i=1}^n G_i} - \frac{1}{\sum_{i=1}^n G_i}\sum_{i=1}^n \frac{(1-G_i)(Y_{i,t+1}-Y_{it})e(\boldsymbol{X}_i)}{1-e(\boldsymbol{X}_i)} \\ &= \left(\frac{N}{\sum_{i=1}^n G_i}\right)\left\{\frac{1}{N}\sum_{i=1}^n G_i(Y_{i,t+1}-Y_{it}) - \frac{1}{N}\sum_{i=1}^n \frac{(1-G_i)(Y_{i,t+1}-Y_{it})e(\boldsymbol{X}_i)}{1-e(\boldsymbol{X}_i)}\right\}.\end{aligned}$$

Further observe that

$$\begin{aligned}&\sqrt{N}(\tau_{\text{CFD}}^{\text{wt}} - \tau_{\text{CFD}}) \\ =&\frac{1}{\pi}\frac{1}{\sqrt{N}}\sum_{i=1}^n\left\{G_i(Y_{i,t+1}-Y_{it}) - \frac{(1-G_i)(Y_{i,t+1}-Y_{it})e(\boldsymbol{X}_i)}{1-e(\boldsymbol{X}_i)} - \tau_{\text{CFD}}\right\} + o_p(1) \\ =&\frac{1}{\pi}\frac{1}{\sqrt{N}}\sum_{i=1}^n\left\{\frac{(G_i-e(\boldsymbol{X}_i))(Y_{i,t+1}-Y_{it})}{\pi(1-e(\boldsymbol{X}_i))} - \tau_{\text{CFD}}\right\} + o_p(1),\end{aligned}$$

where $o_p(1)$ is a residual term that converges in probability to zero. We then obtain φ_i^{wt} as the individual summand in the bracket (Tsiatis, 2006). A similar reasoning is used to obtain φ_i^{dr} in Proposition 2. Notice that

$$\begin{aligned}\varphi_i^{\text{wt}} + \varphi_i^{\text{dr}} &= \frac{1}{\pi}\frac{G_i - e(\boldsymbol{X}_i)}{1-e(\boldsymbol{X}_i)}\{2(Y_{i,t+1}-Y_{it}) - (\nu(\boldsymbol{X}_i)-\mu(\boldsymbol{X}_i))\} - 2\tau_{\text{CFD}}, \\ \varphi_i^{\text{wt}} - \varphi_i^{\text{dr}} &= \frac{1}{\pi}\frac{G_i - e(\boldsymbol{X}_i)}{1-e(\boldsymbol{X}_i)}\{\nu(\boldsymbol{X}_i)-\mu(\boldsymbol{X}_i)\},\end{aligned}$$

and the difference in asymptotic variance

$$\begin{aligned}&\mathbb{V}(\varphi_i^{\text{wt}}) - \mathbb{V}(\varphi_i^{\text{dr}}) = E[(\varphi_i^{\text{wt}} + \varphi_i^{\text{dr}})(\varphi_i^{\text{wt}} - \varphi_i^{\text{dr}})] \\ =&\frac{2}{\pi^2}E\Big[\Big(\frac{G_i - e(\boldsymbol{X}_i)}{1-e(\boldsymbol{X}_i)}\Big)^2(Y_{i,t+1}-Y_{it})(\nu(\boldsymbol{X}_i)-\mu(\boldsymbol{X}_i))\Big] \\ &- \frac{1}{\pi^2}E\Big[\Big(\frac{G_i - e(\boldsymbol{X}_i)}{1-e(\boldsymbol{X}_i)}\Big)^2(\nu(\boldsymbol{X}_i)-\mu(\boldsymbol{X}_i))^2\Big],\end{aligned}$$

since

$$\frac{2\tau_{\text{CFD}}}{\pi}E\Big[\Big(\frac{G_i - e(\boldsymbol{X}_i)}{1-e(\boldsymbol{X}_i)}\Big)\Big(\nu(\boldsymbol{X}_i) - \mu(\boldsymbol{X}_i)\Big)\Big]$$

$$=\frac{2\tau_{\text{CFD}}}{\pi}E\Big[\Big(\frac{E(G_i|\boldsymbol{X}_i)_i - e(\boldsymbol{X}_i)}{1-e(\boldsymbol{X}_i)}\Big)\Big(\nu(\boldsymbol{X}_i) - \mu(\boldsymbol{X}_i)\Big)\Big|\boldsymbol{X}_i\Big] = 0.$$

Unfortunately, the expression for $\mathbb{V}(\varphi_i^{\text{wt}}) - \mathbb{V}(\varphi_i^{\text{dr}})$ does not further simplify to more elegant forms. But it is evident that there is no guarantee that this difference is non-negative since it could not be simplified to the expectation of a quadratic form (this is in sharp contrast to the analogous results developed for estimating the average treatment effect, or ATE). Hence even if all models are correct, the double-robust estimator is not necessarily asymptotically more efficient than the weighting estimator.

4

Propensity Score Weighting for Causal Inference with Multiple Treatments

4.1 Introduction

Unconfounded comparisons between groups in observational studies usually require adjustment for differences in pre-treatment covariates. Standard parametric regression adjustment may be sensitive to model misspecification when there is limited overlap in covariate distributions between groups. The propensity score plays a central role in estimating the causal effects in such settings (Rosenbaum and Rubin, 1983). With binary treatments, the propensity score summarizes the multi-dimensional covariates into a scalar score and balancing this score balances all covariates. Unconfounded comparisons of more than two groups are increasingly common in practice. For example, it is of much policy interest to study disparities in health services between more than two races, such as Whites, Asians, Blacks and Hispanics (e.g., Zaslavsky and Ayanian, 2005), but most existing studies have focused on separate comparison of each White-minority pair (e.g., Cook et al., 2009, 2010). Such separate pairwise comparisons may lead to conclusions that are not transitive across groups

because each may target at a different target population (McCaffrey et al., 2013).

For multiple-group comparisons, Imbens (2000) has developed the generalized propensity score method; the key insight is that the scalar generalized propensity score of each treatment level can be exploited to separately estimate the average potential outcomes at that level. With the generalized propensity score device, matching and subclassification have been discussed extensively; see, for instance, Lechner (2002); Zanutto et al. (2005); Rassen et al. (2013); Yang et al. (2016); Lopez and Gutman (2017a,b). Another popular strategy, propensity score weighting, has also been generalized to multiple treatments (Feng et al., 2012; Cattaneo, 2010; McCaffrey et al., 2013), with much focus on the pairwise average treatment effect (ATE) based on the inverse probability weighting (IPW) estimator. However, observational studies often rely on convenience samples, which does not represent a population of scientific meaning. In such cases, the automatic focus on ATE may be questionable because it is not clear what target population the causal conclusion is applicable to. Meanwhile, multiple treatments exacerbate the overlap issues as different treatments may be applicable only to certain subpopulations, and the ATE may correspond to an infeasible intervention. Regardless of the number of treatment options, extreme propensity scores close to zero or one will likely result in bias and excessive variance of the IPW estimators (Li et al., 2018). Crump et al. (2009) proposed an optimal trimming procedure that focuses on regions with good overlap and thus improves the efficiency of the IPW estimator for binary treatments; Yang et al. (2016) extended the trimming rule to more than two treatments. Though easy to implement, propensity trimming often corresponds to an ambiguous target population and may discard a large number of units.

In this Chapter, motivated by a racial disparity study in health services research, we present a unified framework for weighting-based causal inference applicable to multiple treatments. Specifically, we generalize the balancing weights framework for

binary treatments (Li et al., 2018) to balance the distribution of covariates from multiple treatment groups according to a pre-specified target population. Within this framework, we propose a class of target estimands based on linear contrasts and their corresponding nonparametric weighting estimators. We derive several asymptotic results on these nonparametric estimators, based on which we develop the generalized overlap weights, constructed as the product of the inverse probability weights and the harmonic mean of the generalized propensity scores. The generalized overlap weights focus on the overlap population with substantial probabilities to be assigned to all treatments. This target population aligns with the spirit of randomized clinical trials by emphasizing patients at clinical equipoise, and is thus of natural relevance to medical and policy studies. Under mild conditions, we further show that the generalized overlap weights minimize the total asymptotic variance of the nonparametric estimators for the pairwise contrasts within the class of balancing weights. These weights are strictly bounded between zero and one, and thus automatically bypass the issue of extreme weights.

The remainder of this Chapter is organized as follows. Section 4.2 introduces the general framework of balancing weights. In Section 4.3, we propose the generalized overlap weights for pairwise comparisons with multiple treatments, discuss balance check criteria and variance estimation. In Section 4.4, we reanalyze the Medical Expenditure Panel Survey data and study the racial disparities in medical expenditures among Whites, Asians, Blacks and Hispanics using the proposed generalized overlap weights. Additional simulations to examine the benefits of the proposed weighting method are carried out in Section 4.5. Section 4.6 concludes.

4.2 Balancing Weights for Multiple Treatments

4.2.1 Basic Setup, definitions and assumptions

We consider a sample of n units, each belonging to one of $J \geqslant 3$ groups for which covariate-balanced comparisons are of interest. Let $Z_i \in \mathbb{Z} = \{1, \ldots, J\}$ denote the treatment group membership, and $D_{ij} = \mathbb{1}\{Z_i = j\}$ the indicator of receiving treatment level j. For each unit, we observe an outcome Y_i and a set of p pre-treatment covariates $\boldsymbol{X}_i = (X_{i1}, ..., X_{ip})'$. The concept of the propensity score has been extended by Imbens (2000) to $J \geqslant 3$ treatments:

Definition 4.2.1. *(Generalized Propensity Scores) The generalized propensity score is the conditional probability of being assigned to each group given the covariates:*

$$e_j(\boldsymbol{X}) = \Pr(Z = j|\boldsymbol{X}), \quad j \in \mathbb{Z}.$$

By definition, the sum-to-unity restriction $\sum_{j=1}^{J} e_j(\boldsymbol{X}) = 1$ holds for all $\boldsymbol{X}$ in support $\mathbb{X}$, and hence each unit's propensity can be uniquely characterized by $J-1$ scalar scores. Under the Stable Unit Treatment Value Assumption (SUTVA, Imbens and Rubin, 2015), each unit has a potential outcome $Y_i(j)$ mapped to each treatment level $j \in \mathbb{Z}$, among which, only the one corresponding to the received treatment, $Y_i = Y_i(Z_i)$, is observed. To proceed, we make the following two identification assumptions.

Assumption 4.2.2. *(Weak Unconfoundedness) The assignment is weakly unconfounded if*

$$Y(j) \perp \mathbb{1}\{Z = j\}|\boldsymbol{X}, \quad \forall\ j \in \mathbb{Z}.$$

Assumption 4.2.3. *(Overlap) For all values of $\boldsymbol{X} \in \mathbb{X}$ and all treatment group j, the probability of being assignment to any treatment group is strictly bounded away from zero:*

$$e_j(\boldsymbol{X}) > 0, \quad \forall\ \boldsymbol{X} \in \mathbb{X},\ j \in \mathbb{Z}.$$

Assumption 4.2.2 imposes unconfoundedness separately for each level of the treatment, and is sufficient for identification of the population-level estimand (Imbens, 2000). An implication from this assumption is that the potential outcome $Y(j)$ is independent of the assignment indicator $\mathbb{1}\{Z = j\}$ conditional on the scalar generalized propensity score $e_j(\boldsymbol{X})$. In other words, adjusting for the scalar score is sufficient to remove the bias in estimating the average value of $Y(j)$ over the target population. Assumption 4.2.3 restricts the study population to the covariate space where each unit has non-zero probability to receive any treatment, and therefore causal comparisons are feasible without relying on unwarranted extrapolation.

To further elaborate, we define the conditional expected potential outcomes in group j as $m_j(\boldsymbol{X}) = \mathrm{E}[Y(j)|\boldsymbol{X}]$. Under Assumption 4.2.2, we have $m_j(\boldsymbol{X}) = \mathrm{E}[Y|Z = j, \boldsymbol{X}]$, which is estimable from observed data. As previously mentioned, the propensity score methods are also applicable to unconfounded descriptive comparisons where the group membership is a non-manipulable state, such as different races. In these cases, the common objective is to compare the expected observed outcomes, $m_j(\boldsymbol{X}) = \mathrm{E}[Y|Z = j, \boldsymbol{X}]$; for example, when $J = 2$, Li et al. (2013) defined the contrast between $m_1(\boldsymbol{X})$ and $m_2(\boldsymbol{X})$ averaged over a population the average controlled difference (ACD). We use the potential outcome notations throughout for technical discussions, but the methodology is directly applicable to non-causal descriptive comparisons.

4.2.2 Balancing Weights

Assume the marginal density of the covariates, $f(\boldsymbol{X})$, exists, with respect to a base measure μ. In causal studies, the interest is on the average effects of units in a target population, whose density (up to a normalizing constant) we represent by $f(\boldsymbol{X})h(\boldsymbol{X})$, with $h(\boldsymbol{X})$ being a pre-specified function of covariates or equivalently a tilting function. We first define the expectation of the potential outcomes over the

target population $f(\boldsymbol{X})h(\boldsymbol{X})$:

$$m_j^h \equiv \frac{\int_{\mathbb{X}} m_j(\boldsymbol{X}) f(\boldsymbol{X}) h(\boldsymbol{X}) \mu(d\boldsymbol{X})}{\int_{\mathbb{X}} f(\boldsymbol{X}) h(\boldsymbol{X}) \mu(d\boldsymbol{X})}.$$

Then we can define a class of estimands as a linear combination of the above expectations, with coefficients $\boldsymbol{a} = (a_1, \cdots, a_J)'$:

$$\tau^h(\boldsymbol{a}) \equiv \sum_{j=1}^{J} a_j m_j^h. \tag{4.1}$$

The causal estimand $\tau^h(\boldsymbol{a})$ generalizes the definition of weighted average treatment effect (WATE) in binary treatments (Hirano et al., 2003) where $J = 2$ and $\boldsymbol{a} = (1, -1)$. As will be seen in due course, $\tau^h(\boldsymbol{a})$ includes several existing causal estimands for multiple treatments as special cases.

We next define the class of balancing weights. Let $f_j(\boldsymbol{X}) = f(\boldsymbol{X}|Z = j)$ be the density of $\boldsymbol{X}$ in the jth group over its support $\mathbb{X}_j$, we have $f_j(\boldsymbol{X}) \propto f(\boldsymbol{X}) e_j(\boldsymbol{X})$. Given any pre-specified function h, we can weight the group-specific density $f_j(\boldsymbol{X})$ to the target population using the following weights, proportional up to a normalizing constant:

$$w_j(\boldsymbol{X}) \propto \frac{f(\boldsymbol{X}) h(\boldsymbol{X})}{f(\boldsymbol{X}) e_j(\boldsymbol{X})} = \frac{h(\boldsymbol{X})}{e_j(\boldsymbol{X})}, \quad \forall\ j \in \mathbb{Z}. \tag{4.2}$$

It is straightforward to show that the class of weights defined in (4.2) balance the weighted distributions of the covariates across J comparison groups:

$$f_j(\boldsymbol{X}) w_j(\boldsymbol{X}) = f(\boldsymbol{X}) h(\boldsymbol{X}), \quad \forall\ j \in \mathbb{Z}. \tag{4.3}$$

To apply the above framework, a key is to specify the coefficients $\boldsymbol{a}$ and the tilting function h, with the former defining the causal contrast and the latter representing the target population. We focus on the case of multiple nominal treatments, where

the scientific interest usually lies in pairwise comparisons. More specifically, the choice of $\boldsymbol{a}$ is contained in the finite set $\mathbb{S} = \{\boldsymbol{\lambda}_{j,j'} = \boldsymbol{\lambda}_j - \boldsymbol{\lambda}_{j'} : j < j'\}$, where $\boldsymbol{\lambda}_j$ is the $J \times 1$ unit vector with one at the jth position and zero everywhere else. In principle, the tilting function h can take any form, each leading to a unique type of balancing weights; statistical, scientific and policy considerations all play into the specification of h. We illustrate specifications of $\boldsymbol{a}$ and h by connecting the general definition (4.1) with existing estimands in the literature.

When $h(\boldsymbol{X}) = 1$, the corresponding target population $f(\boldsymbol{X})$ is the combined population from all groups and the weights become the standard inverse probability weights, $\{1/e_j(\boldsymbol{X}),\ j \in \mathbb{Z}\}$; the target estimand is the pairwise ATE as in Feng et al. (2012). When $h(\boldsymbol{X}) = e_{j'}(\boldsymbol{X})$, the target population is the subpopulation receiving treatment $Z = j'$, and the weights, $\{e_{j'}(\boldsymbol{X})/e_j(\boldsymbol{X}),\ j \in \mathbb{Z}\}$, are designed to estimate the average treatment effect for the treated (ATT). Define

$$\underline{e}_j = \max_{1 \leqslant l \leqslant J} \{\min_{\boldsymbol{X} \in \mathbb{X}_l} \{e_j(\boldsymbol{X})\}\}, \quad \bar{e}_j = \min_{1 \leqslant l \leqslant J} \{\max_{\boldsymbol{X} \in \mathbb{X}_l} \{e_j(\boldsymbol{X})\}\},$$

and an eligibility function $E_j(\boldsymbol{X}) = \mathbb{1}\{\underline{e}_j \leqslant e_j(\boldsymbol{X}) \leqslant \bar{e}_j\}$ for all $j \in \mathbb{Z}$. When $h(\boldsymbol{X}) = e_{j'}(\boldsymbol{X}) \prod_{j=1}^{J} E_j(\boldsymbol{X})$, the target population is the subpopulation receiving treatment $Z = j'$ but remaining eligible for all other treatments (Lopez and Gutman, 2017b). Define a threshold α as the largest value such that

$$\alpha \leqslant \frac{2\,\mathrm{E}\left[\sum_{j=1}^{J} 1/e_j(\boldsymbol{X}) | \sum_{j=1}^{J} 1/e_j(\boldsymbol{X}) \leqslant \alpha\right]}{\Pr\left(\sum_{j=1}^{J} 1/e_j(\boldsymbol{X}) \leqslant \alpha\right)}. \tag{4.4}$$

When $h(\boldsymbol{X}) = \mathbb{1}\{\boldsymbol{X} \in \mathbb{C}\}$ with $\mathbb{C} = \{\boldsymbol{X} \in \mathbb{X} | \sum_{j=1}^{J} 1/e_j(\boldsymbol{X}) \leqslant \alpha\}$, the target population is characterized by the subpopulation $\mathbb{C}$, and the inverse probability weights are formulated after applying the optimal trimming rule (Yang et al., 2016). Finally, when $h(\boldsymbol{X}) = \min_{1 \leqslant j \leqslant J}\{e_j(\boldsymbol{X})\}$, one arrives at the generalized matching weights

(Yoshida et al., 2017)—an extension of the matching weights of Li and Greene (2013) to multiple treatments. The matching weights is a weighting analogue to exact matching and the causal comparisons are made for the matched population. Moreover, one could choose indicator functions for h that directly involves covariates of a subpopulation of interest, such as a specific gender or a range of age. Table 4.1 summarizes the above definitions.

Table 4.1: Examples of balancing weights for pairwise comparisons with multiple treatments and different tilting functions.

Target population	Tilting function $h(\boldsymbol{X})$	Weights $\{w_j(\boldsymbol{X}),\ j \in \mathbb{Z}\}$
Combined	1	$\{1/e_j(\boldsymbol{X}),\ j \in \mathbb{Z}\}$
Treated (j'th group)	$e_{j'}(\boldsymbol{X})$	$\{e_{j'}(\boldsymbol{X})/e_j(\boldsymbol{X}),\ j \in \mathbb{Z}\}$
Treated (restricted)	$e_{j'}(\boldsymbol{X}) \prod_{j=1}^{J} E_j(\boldsymbol{X})$	$\{e_{j'}(\boldsymbol{X}) \prod_{j=1}^{J} E_j(\boldsymbol{X})/e_j(\boldsymbol{X}),\ j \in \mathbb{Z}\}$
Trimmed combined	$\mathbb{1}\{\boldsymbol{X} \in \mathbb{C}\}$	$\{\mathbb{1}\{\boldsymbol{X} \in \mathbb{C}\}/e_j(\boldsymbol{X}),\ j \in \mathbb{Z}\}$
Matching	$\min_{1 \leqslant j \leqslant J}\{e_j(\boldsymbol{X})\}$	$\{\min_{1 \leqslant j \leqslant J}\{e_j(\boldsymbol{X})\}/e_j(\boldsymbol{X}),\ j \in \mathbb{Z}\}$
Overlap	$(\sum_{k=1}^{J} 1/e_k(\boldsymbol{X}))^{-1}$	$\{(\sum_{k=1}^{J} 1/e_k(\boldsymbol{X}))^{-1}/e_j(\boldsymbol{X}),\ j \in \mathbb{Z}\}$

For ordinal treatments where the treatment levels are ordered categories, target estimands may differ from the pairwise comparisons and require different choice of $\boldsymbol{a}$. For instance, one may be interested in the quadratic contrasts between unit increases in the treatment level, namely $(m^h_{j+1} - m^h_j) - (m^h_j - m^h_{j-1})$. In other cases, one may be interested in the weighted average of unit increase in the treatment level, $\sum_{j=1}^{J-1} \pi_j (m^h_{j+1} - m^h_j)$, or the accumulative effect of the maximum treatment, $m^h_J - m^h_1$. In this Chapter, we focus on nominal treatments, but the general framework of balancing weights is applicable to ordinal treatments.

4.2.3 Transitivity

With multiple treatments, a desirable property of a given class of estimands is *transitivity*. For pairwise comparisons, lack of transitivity often implies that comparisons of treatments are based on different populations. As a result, non-transitivity may

lead to incompatible pairwise contrasts; for example, it is possible that treatment A is favored over treatment B, treatment B is favored over treatment C, but treatment C is found to better than treatment A at the same time. Below we provide a formal definition of transitivity and offer two related remarks.

Definition 4.2.4. *The class of causal estimands* $\mathbb{T}(h, \mathbb{A}) = \{\tau^h(\boldsymbol{a}) : \boldsymbol{a} \in \mathbb{A} \subset \mathbb{R}^J\}$ *is transitive if the following equivariance relationship holds:* $\tau^h(\boldsymbol{a}) + \tau^h(\boldsymbol{a}') = \tau^h(\boldsymbol{a}'')$ *whenever* $\boldsymbol{a}, \boldsymbol{a}', \boldsymbol{a}'' \in \mathbb{A}$ *and* $\boldsymbol{a} + \boldsymbol{a}' = \boldsymbol{a}''$.

Remark **1.** *Fixing a tilting function* h*, the class of estimands specifying all pairwise contrasts, namely,* $\mathbb{T}(h, \mathbb{S})$ *is transitive. For example, with* $h(\boldsymbol{X}) = 1$*, the class of pairwise ATE estimands is transitive; with* $h(\boldsymbol{X}) = e_j(\boldsymbol{X}) \prod_{l=1}^{J} E_l(\boldsymbol{X})$*, the class of ATT estimands in Lopez and Gutman (2017b) is also transitive.*

Remark **2.** *The union of* $\mathbb{T}(h_1, \mathbb{S})$ *and* $\mathbb{T}(h_2, \mathbb{S})$ *or that of their subsets is generally non-transitive for* $h_1 \neq h_2$*. This explains why several existing classes of estimands are non-transitive, including the class of ATT estimands of Lechner (2001),* $\{\mathrm{E}[Y_i(j) - Y_i(j') | Z_i = j \text{ or } Z_i = j'] : j < j'\}$*. The reason is that each individual estimand corresponds to a distinct tilting function* $h_{j,j'}(\boldsymbol{X}) = (e_j(\boldsymbol{X}) + e_{j'}(\boldsymbol{X}))/e_1(\boldsymbol{X})$*, and therefore this class of estimands is the union of* $\binom{J}{2}$ *elements, each of which is contained in* $\mathbb{T}(h_{j,j'}, \mathbb{S})$ *for some* $j < j'$.

4.2.4 Large-sample Properties of Nonparametric Estimators

For any pre-specified vector $\boldsymbol{a}$ and tilting function h, we could first use the plug-in sample estimator to obtain the expectation of the potential outcomes among the target population

$$\hat{m}_j^h = \frac{\sum_{i=1}^n D_{ij} Y_i w_j(\boldsymbol{X}_i)}{\sum_{i=1}^n D_{ij} w_j(\boldsymbol{X}_i)}, \tag{4.5}$$

and then estimate $\tau^h(\boldsymbol{a})$ by a linear combination, $\hat{\tau}^h(\boldsymbol{a}) = \sum_{j=1}^{J} a_j \hat{m}_j^h$, where the sum is over a sample drawn from density $f(\boldsymbol{X})$. Below we establish three large-sample results of $\hat{\tau}^h(\boldsymbol{a})$.

Theorem 4.2.5. *Given any h and $\boldsymbol{a}$, $\hat{\tau}^h(\boldsymbol{a})$ is a consistent estimator of $\tau^h(\boldsymbol{a})$.*

Proof. See Section 4.7.1. □

Denote the collection of treatment assignment $\underline{\boldsymbol{Z}} = \{Z_1, \ldots, Z_n\}$ and covariate design points $\underline{\boldsymbol{X}} = \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\}$. The next two results concern the variance of the sample estimator, which is decomposed as

$$\mathrm{V}[\hat{\tau}^h(\boldsymbol{a})] = \mathrm{E}_{\underline{\boldsymbol{Z}},\underline{\boldsymbol{X}}}\, \mathrm{V}[\hat{\tau}^h(\boldsymbol{a})|\underline{\boldsymbol{Z}}, \underline{\boldsymbol{X}}] + \mathrm{V}_{\underline{\boldsymbol{Z}},\underline{\boldsymbol{X}}}\, \mathrm{E}[\hat{\tau}^h(\boldsymbol{a})|\underline{\boldsymbol{Z}}, \underline{\boldsymbol{X}}].$$

The first term is the variation due to residual variance in $\hat{\tau}^h(\boldsymbol{a})$ conditional on the design points. The second term arises from the dependence of the expectation of the plug-in estimator on the sample, and estimating it involves the outcome model (associations between $Y(j)$ and $\boldsymbol{X}$). As individual variation is typically much larger than conditional mean variation, the benefit of further optimizing the weights by a preliminary look at the outcomes, which mixes the design and analysis, would usually not justify the risk of biasing model specification to attain desired results (Imbens, 2004). Hence, we focus on the first term.

Theorem 4.2.6. *Given $\boldsymbol{a}$, suppose the family of residual variances $\{\mathrm{V}[\hat{\tau}^h(\boldsymbol{a})|\underline{\boldsymbol{Z}}, \underline{\boldsymbol{X}}], n \geqslant 1\}$ is uniformly integrable. Then the expectation of the conditional variance converges*

$$n \cdot \mathrm{E}_{\underline{\boldsymbol{Z}},\underline{\boldsymbol{X}}}\, \mathrm{V}[\hat{\tau}^h(\boldsymbol{a})|\underline{\boldsymbol{Z}}, \underline{\boldsymbol{X}}] \to Q(\boldsymbol{a}, h) \equiv \int_{\mathbb{X}} \Big(\sum_{j=1}^{J} a_j^2 v_j(\boldsymbol{X})/e_j(\boldsymbol{X}) \Big) h^2(\boldsymbol{X}) f(\boldsymbol{X}) \mu(d\boldsymbol{X})/C_h^2,$$

where $v_j(\boldsymbol{X}) = \mathrm{V}[Y(j)|\boldsymbol{X}]$ and $C_h \equiv \int_{\mathbb{X}} h(\boldsymbol{X}) f(\boldsymbol{X}) \mu(d\boldsymbol{X})$ is a constant.

Proof. See Section 4.7.2. □

When the residual variance of the potential outcome is homoscedastic across all groups such that $v_j(\boldsymbol{X}) = v$, then the limit $Q(\boldsymbol{a}, h)$ can further simplify and the following result holds.

Theorem 4.2.7. *Under homoscedasticity, the function*

$$\tilde{h}(\boldsymbol{X}) \propto \frac{1}{\sum_{j=1}^{J} a_j^2 / e_j(\boldsymbol{X})}$$

gives the smallest asymptotic variance for the weighted estimator $\hat{\tau}^h(\boldsymbol{a})$ *among all* h*'s, and* $\min_h Q(\boldsymbol{a}, h) = v/C_{\tilde{h}}$.

Proof. See Section 4.7.3. □

A more general result of Theorem 4.2.7 can be obtained under heteroscedasticity. In that case, the optimal tilting function,

$$\tilde{h}(\boldsymbol{X}) \propto \frac{1}{\sum_{j=1}^{J} a_j^2 v_j(\boldsymbol{X}) / e_j(\boldsymbol{X})},$$

explicitly depends on the residual variances of the potential outcomes. Although estimates of $v_j(\boldsymbol{X})$ can be obtained by outcome regression modeling in the analysis stage, it is rarely the case that accurate prior information is available in the design stage. Therefore, such a tilting function is difficult to specify for design purposes and may find limited use without peaking at the outcomes (mixing the design and analysis). For such considerations, we motivate the generalized overlap weights in Section 4.3 under homoscedasticity. Overall, these asymptotic results generalize those for binary treatments in Li et al. (2018); they also extend the asymptotic results on propensity score trimming in Crump et al. (2009) and Yang et al. (2016), who have similarly assumed homoscedasticity but restricted the class of tilting functions to indicator functions.

4.3 Generalized Overlap Weighting for Pairwise Comparisons

4.3.1 The Generalized Overlap Weights

For nominal treatments, scientific interest often lies in comparing outcomes between each pair of treatment groups in a common target population. In this case, as $\boldsymbol{a} \in \mathbb{S}$, we propose to choose the tilting function h that minimizes the total asymptotic variance of the sample estimators for all pairwise comparisons; in other words, the objective function is

$$\sum_{j<j'} Q(\boldsymbol{\lambda}_{j,j'}, h) \propto Q(\mathbf{1}_{J\times 1}, h).$$

According to Theorem 4.2.7, the function $h(\boldsymbol{X}) = (\sum_{j=1}^{J} 1/e_j(\boldsymbol{X}))^{-1}$—the harmonic mean of the generalized propensity scores—minimizes $Q(\mathbf{1}_{J\times 1}, h)$ among all h. Based on this optimal tilting function h, we define the generalized overlap weights for $j = 1, ..., J$:

$$w_j(\boldsymbol{X}) \propto \frac{1/e_j(\boldsymbol{X})}{\sum_{k=1}^{J} 1/e_k(\boldsymbol{X})}.$$

For binary treatments ($J = 2$), the generalized overlap weights reduce to the overlap weights in Li et al. (2018), namely the propensity of assignment to the other group: $w_1(\boldsymbol{X}) \propto 1 - e_1(\boldsymbol{X}) = e_2(\boldsymbol{X})$, $w_2(\boldsymbol{X}) \propto 1 - e_2(\boldsymbol{X}) = e_1(\boldsymbol{X})$.

The maximum of the harmonic mean function h is attained when $e_j(\boldsymbol{X}) = 1/J$ for all j, that is, when the units have the same propensity to each of the treatments. Heuristically, the tilting function h gives the most relative weight to the covariate regions in which none of the propensities are close to zero. While it is generally difficult to visualize the value of h in higher dimensions, we could certainly do so with $J = 3$ treatments. Figure 4.1 provides a ternary plot of the value of h (up to a proportionality constant) when $J = 3$. Each point in the triangular plane represents a unit with certain values of the generalized propensity scores. The value

of each generalized propensity score is proportional to the orthogonal distance from that point to each edge. It is evident that the weighting scheme emphasizes the centroid region with good overlap and smoothly down-weights the edges. In other words, the optimal tilting function gives the most relative weight to the covariate regions in which none of the propensities are close to zero, and down-weights the region where there is lack of overlap in at least one dimension. Therefore, we can interpret the corresponding target population to be the subpopulation with the most overlap in covariates among all groups, and call the target estimand as the pairwise average treatment effect among the overlap population (ATO). As the overlap population tilts $f(\boldsymbol{X})$ most heavily towards equipoise, it is naturally of policy and clinical relevance. For clinical practice, this target population aligns with the spirit of randomized studies and emphasizes patients with most clinical equipoise. The treatment decisions for these patients remain unclear and thus comparative information is most needed. Analogously, in descriptive studies for racial disparities, the overlap population represents individuals with most similarity in observed demographic and health-related characteristics, based on whom subsequent policy interventions on health care utilization become most meaningful.

Besides its optimality in asymptotic efficiency, the generalized overlap weights have several additional attractive features. First, the harmonic mean function h is strictly bounded

$$0 < \min_{1 \leqslant l \leqslant J} \{e_l(\boldsymbol{X})\}/J \leqslant h(\boldsymbol{X}) \leqslant \min_{1 \leqslant l \leqslant J} \{e_l(\boldsymbol{X})\} < 1,$$

and thus the weighting scheme is robust to extreme weights, in contrast to the standard IPW scheme. Second, the target population defined by the generalized overlap weights is adaptive to the covariate distributions among the J comparison groups. For example, when the propensity of assignment to treatment j is small

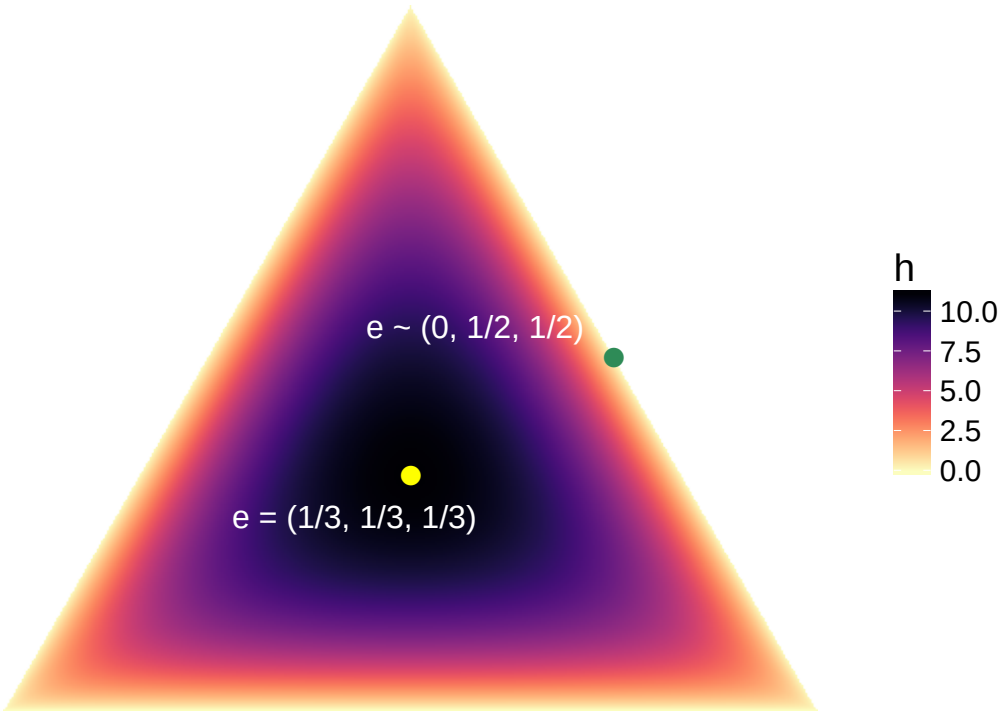


FIGURE 4.1: Ternary plot for optimal h as a function of the generalized propensity score with $J = 3$ treatments.

compared to others so that $e_j(\boldsymbol{X}) \approx 0$, the tilting function

$$h(\boldsymbol{X}) \propto \prod_{l=1}^{J} e_l(\boldsymbol{X}) / \sum_{k=1}^{J} \prod_{l \neq k} e_l(\boldsymbol{X}) \approx \prod_{l=1}^{J} e_l(\boldsymbol{X}) / \prod_{l \neq j} e_l(\boldsymbol{X}) = e_j(\boldsymbol{X}),$$

suggesting that the target population is similar to the jth treatment group and the associated estimand approximates the ATT. On the other hand, if the treatment groups are almost balanced in size and covariate distribution so that $e_j(\boldsymbol{X}) \approx 1/J$ for all j, we have $h(\boldsymbol{X}) \propto 1$ and the target estimand approximates the pairwise ATE. Ar-

guably this adaptiveness enables the generalized overlap weighting scheme to define a scientific question that may be best answered nonparametrically by the available data. Finally, the generalized matching weights (Yoshida et al., 2017)—defined by $h(\boldsymbol{X}) = \min_{1 \leqslant j \leqslant J}\{e_j(\boldsymbol{X})\}$—share some of the above advantages, but these weights are not asymptotically efficient and are non-smooth in $\boldsymbol{X}$, which renders closed-form variance calculation more complex.

4.3.2 Estimate Generalized Propensity Scores and Balance Check

In practice, usually the propensity scores are not known and must be estimated from the data. For multiple nominal treatments, the generalized propensity scores are frequently modeled by a multinomial logistic regression,

$$\begin{aligned} e_1(\boldsymbol{X}_i) &= \frac{1}{1 + \sum_{k=2}^{J} \exp(\alpha_k + \boldsymbol{X}_i^T \boldsymbol{\beta}_k)}, \\ e_j(\boldsymbol{X}_i) &= \frac{\exp(\alpha_j + \boldsymbol{X}_i^T \boldsymbol{\beta}_j)}{1 + \sum_{k=2}^{J} \exp(\alpha_k + \boldsymbol{X}_i^T \boldsymbol{\beta}_k)}, \quad j = 2, \ldots, J, \end{aligned} \tag{4.6}$$

where the covariate vector $\boldsymbol{X}$ are allowed to contain higher-order moments, splines and interactions. Model parameters $\boldsymbol{\theta} = (\alpha_2, \ldots, \alpha_J, \boldsymbol{\beta}_2^T, \ldots, \boldsymbol{\beta}_J^T)^T$ can be estimated by standard maximum likelihood, from which we obtain the estimated propensity scores. To assess the fit of the propensity score model, we check the weighted covariate balance in the target population. We consider two ways for balance check motivated by the population balancing constraint (4.3). First, constraint (4.3) implies the weighted covariate balance between each group and the target population. Therefore, we inspect, for each treatment level, the weighted covariate mean deviation from that of the target population. Specifically, we define

$$\bar{X}_j = \sum_{i=1}^{n} D_{ij} X_i w_j(\boldsymbol{X}_i) \Big/ \sum_{i=1}^{n} D_{ij} w_j(\boldsymbol{X}_i),$$

as the weighted mean of covariate X from the jth group and $S^2_{X,j}$ as the unweighted variance. Further, we define

$$\bar{X}_p = \sum_{i=1}^{n}\sum_{l=1}^{J} D_{il}X_i h(\boldsymbol{X}_i) \Big/ \sum_{i=1}^{n}\sum_{l=1}^{J} D_{il}h(\boldsymbol{X}_i),$$

as the average value of covariate X in the target population and $S^2_X = J^{-1}\sum_{j=1}^{J} S^2_{X,j}$ as the averaged unweighted variance. The population standardized difference (PSD) is then defined for each covariate and each treatment level as

$$\text{PSD}_j = |\bar{X}_j - \bar{X}_p|/S_X.$$

Similar to McCaffrey et al. (2013), we then use $\max_j |\text{PSD}_j|$ as the balance metric for each covariate X and inspect the adequacy of the propensity score model. If a covariate is not well balanced in one group, interaction terms of that variable with other variables can be added to the model, and the new model is re-fit and re-evaluated until balance is deemed satisfactory. On the other hand, the population balance constraint (4.3) also implies pairwise balance $f_j(\boldsymbol{X})w_j(\boldsymbol{X}) = f_{j'}(\boldsymbol{X})w_{j'}(\boldsymbol{X})$ for all $j \neq j'$, and so we could alternatively assess balance by checking the pairwise absolute standardized differences (ASD),

$$\text{ASD}_{j,j'} = |\bar{X}_j - \bar{X}_{j'}|/S_X.$$

The balance metric for each covariate can then be similarly specified as $\max_{j<j'} |\text{ASD}_{j,j'}|$. In practice, the two balance criteria perform similarly and we will report both in Section 4.4.

In real applications, a rich propensity score model is often preferable because the ultimate goal of weighting-based causal inference is to remove bias through balancing the covariates in the target population, rather than to maximize the predictive utility of the propensity score model. Therefore, the above balance check constitutes a

crucial step for diagnosing propensity score models and the traditional goodness-of-fit diagnostics are minimally relevant. However, the richness of the propensity score model will be capped by the bias-variance tradeoff: when the propensity score model becomes saturated and discovers a separating plane in the data, the variance of the weights will likely increase and reduce the precision of the effect estimates.

Finally, a special property of the overlap weights with binary treatments is exact balance, that is, when the propensities scores are estimated from a logistic model, the standardized difference of all the covariates entering the propensity model is zero, i.e, $\text{ASD}_{1,2} = 0$ for $J = 2$ (Li et al., 2018, Theorem 3). However, this exact balance property is due to the happenstance that the logistic score equations exploit the covariate-balancing moment conditions, and does not directly extend to the generalized overlap weights with $J \geqslant 3$ when the propensity score is estimated by a multinomial logistic model. Therefore, we still recommend to use the conventional iterative fitting-checking procedure to improve the propensity model.

4.3.3 Variance Estimation

The asymptotic variance results in Section 4.2.4 are not directly useful for calculating the sample variance of $\hat{\tau}^h(\boldsymbol{\lambda}_{j,j'})$ in practice because the $v_j(\boldsymbol{X})$'s are not known. Moreover, one has to account for the additional uncertainty in estimating the propensities in the variance estimation. Here we derive an empirical sandwich variance estimator (Stefanski and Boos, 2002) that accounts for the uncertainty in estimating the generalized overlap weights from the multinomial logistic model (4.6). We provide the following theorem to motivate the closed-variance calculation for the pairwise ATO estimates.

Theorem 4.3.1. *Under standard regularity conditions, when the generalized propensity scores are estimated by multinomial logistic regression (4.6), the resulting ATO*

estimator between groups j and j' is asymptotically normal

$$\sqrt{n}\{\hat{\tau}^h(\boldsymbol{\lambda}_{j,j'}) - \tau^h(\boldsymbol{\lambda}_{j,j'})\} \xrightarrow{d} \mathcal{N}\left(0, \mathrm{E}\left\{\psi_{ij} - \psi_{ij'}\right\}^2 / [\mathrm{E}\{h(\boldsymbol{X})\}]^2\right),$$

where

$$\psi_{ij} = D_{ij}(Y_i - m_j^h)w_j(\boldsymbol{X}_i) + \mathrm{E}\left\{D_{ij}(Y_i - m_j^h)\frac{\partial}{\partial \boldsymbol{\theta}^T}w_j(\boldsymbol{X}_i)\right\}\boldsymbol{I}_{\boldsymbol{\theta}\boldsymbol{\theta}}^{-1}\boldsymbol{S}_{\boldsymbol{\theta},i},$$

and $\boldsymbol{S}_{\boldsymbol{\theta},i}$, $\boldsymbol{I}_{\boldsymbol{\theta}\boldsymbol{\theta}}$ are the individual score and information matrix of $\boldsymbol{\theta}$, respectively.

Proof. See Section 4.7.4. □

Theorem 4.3.1 suggests the following consistent variance estimator. Denote $\hat{\boldsymbol{\theta}}$, $\hat{\boldsymbol{S}}_{\boldsymbol{\theta},i}$, $\hat{\boldsymbol{I}}_{\boldsymbol{\theta}\boldsymbol{\theta}}$ as the maximum likelihood estimator of $\boldsymbol{\theta}$, the plug-in consistent estimators for the individual score and information matrix, the variance estimator for the estimated ATO is expressed by

$$\hat{\mathrm{V}}[\hat{\tau}^h(\boldsymbol{\lambda}_{j,j'})] = \frac{\sum_{i=1}^n \left(\hat{\psi}_{ij} - \hat{\psi}_{ij'}\right)^2}{\left[\sum_{i=1}^n \{\sum_{k=1}^J 1/\hat{e}_k(\boldsymbol{X}_i)\}^{-1}\right]^2}, \tag{4.7}$$

where

$$\hat{\psi}_{ij} = D_{ij}(Y_i - \hat{m}_j^h)w_j(\boldsymbol{X}_i;\hat{\boldsymbol{\theta}}) + \left\{\frac{1}{n}\sum_{i=1}^n D_{ij}(Y_i - \hat{m}_j^h)\frac{\partial}{\partial \boldsymbol{\theta}^T}w_j(\boldsymbol{X}_i;\hat{\boldsymbol{\theta}})\right\}\hat{\boldsymbol{I}}_{\boldsymbol{\theta}\boldsymbol{\theta}}^{-1}\hat{\boldsymbol{S}}_{\boldsymbol{\theta},i}.$$

The true generalized propensity score is generally unknown in applications and will be substituted by its sample analogue. Hirano et al. (2003) suggested that a consistent estimator of the propensity score leads to more efficient estimation of the WATE with binary treatments than the true propensity score. Our derivation of the variance estimator re-interprets their findings in the context of multiple treatments. Specifically, with a consistent estimator for the generalized propensity score, the influence function for estimating m_j^h, $\psi_{ij}/\mathrm{E}\{h(\boldsymbol{X})\}$, can be viewed as the residual

of $D_{ij}(Y_i - m_j^h)w_j(\boldsymbol{X}_i)/\,\mathrm{E}\{h(\boldsymbol{X})\}$—the influence function for estimating m_j^h using the true propensity score—after projecting it onto the nuisance tangent space of $\boldsymbol{\theta}$ (Tsiatis, 2006). Therefore, the efficiency implications from Hirano et al. (2003) carry over to our pairwise comparisons emphasizing the overlap population.

4.4 Application to the Racial Disparities in Medical Expenditure

4.4.1 The MEPS Data

Our application is based on the 2009 Medical Expenditure Panel Survey (MEPS) data. The data sample resembles the one examined by Cook et al. (2010) and contains demographic, health information and health care expenditures for four racial groups: 9830 non-Hispanic Whites, 1446 Asians, 4020 Blacks, 5150 Hispanics. The goal of our analysis is to estimate the racial disparities in total medical expenditures for adult respondents aged 18 years and older after balancing key confounding variables. Since race is non-manipulable, these comparisons are descriptive but share the same nature with causal comparisons with respect to confounding control. Both Cook et al. (2010) and Li et al. (2018) have carried out similar unconfounded descriptive comparisons to estimate racial disparities, but restricted to a series of separate binary comparisons for each White-minority pair. As mentioned in Section 4.2.3, one possible limitation of separate binary comparisons is the non-transitivity among the pairwise estimands, as the comparisons may be made for different target populations. Here, we focus on the simultaneous multiple-group comparisons facilitated by defining a common target population.

The potential confounders we consider include demographics and variables describing individual health status. Specifically, the list of potential confounders include five continuous variables: SF-12 physical component summary, SF-12 mental component summary, age, body mass index, time since last general checkup. A total of 20 categorical variables are also included: gender, marital status, self-

reported physical health status (five categories), self-reported mental health status (five categories), any limitation of activity, any limitation of social participation, any limitation of cognitive functions, exercise, history of high blood pressure, coronary heart disease, emphysema, high cholesterol, cancer, stroke, diabetes, angina, arthritis, asthma, myocardial infarction and smoking status. As indicated by the first column (unweighted) of the boxplots in Figure 4.2, there are substantial differences in the covariate distributions among the four racial groups, suggesting the necessity of confounding adjustment for unbiased estimation of the racial disparities.

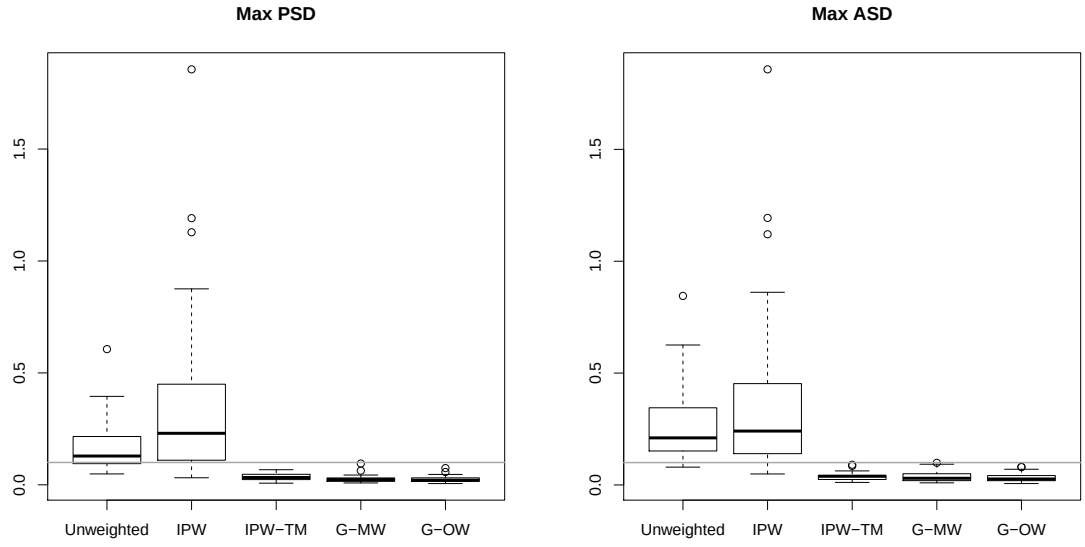


FIGURE 4.2: Boxplots for PSD and ASD corresponding to each weighting method.

In this application, we focus on the proposed generalized overlap weighting (G-OW) estimator. In addition, the crude difference-in-means (DIF) estimator is used as a benchmark to quantify the confounding bias. For comparison purposes, we also consider three additional weighting approaches: IPW based on the entire sample, IPW with optimal trimming rule (4.4) (IPW-TM) and the generalized matching weights (G-MW). As explained in Section 4.2.2, each of these weighting approach corresponds to a specific choice of h and a target population, based on which the

comparisons are made. We further apply a recent propensity score matching estimator proposed by Yang et al. (2016), both with and without the optimal trimming step (GPSM and GPSM-TM). GPSM separately exploits each scalar propensity score for estimating the average potential outcomes and thus resolves the issue of matching on high-dimensional propensity score vector.

4.4.2 Generalized Propensity Score Model and Balance Check

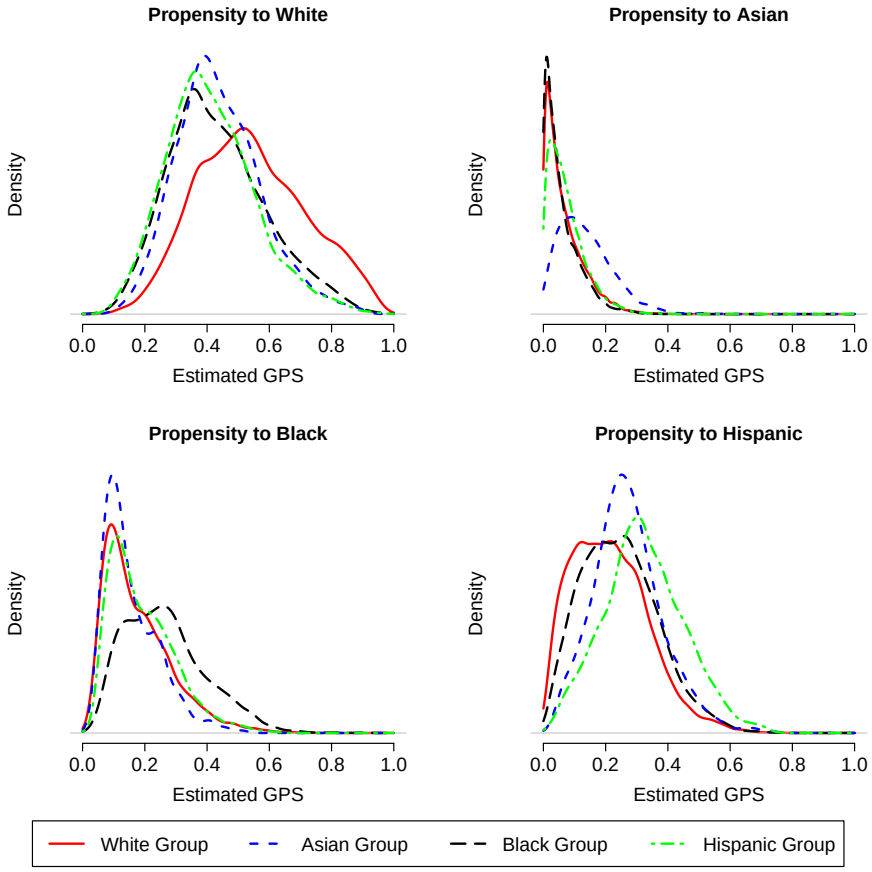


FIGURE 4.3: Distribution of the estimated generalized propensity scores for all racial groups in the MEPS data.

We estimate the generalized propensity scores using a multinomial logistic regression including the main effects of each covariate. The distributions of the estimated scores are presented in Figure 4.3. There is a moderate lack of overlap especially

regarding the Asian group. In this case, focusing on estimating the disparities with the inverse probability weights would force one to consider a hypothetical target population defined by the combined population from all four racial groups. This combined population may not only be an infeasible target of inference due to the lack of overlap, but also lack policy relevance for tracking disparities since it may emphasize individuals atypical for their own racial group. Meanwhile, the largest normalized inverse probability weight is equal to 0.32, accounting for almost one third of the total weights out of 1446 Asians. This weight corresponds to a 78-year-old Asian female with a very high body mass index of 55.4 (atypical value among Asians) and consequently an extreme generalized propensity score close to zero, leading to potential bias and excessive variance in the resulting disparity estimates. On the other hand, we could also use the generalized overlap weights to emphasize the subpopulation with the most overlap across all racial group, namely the individuals who, based on their observed characteristics, could easily be either White or from other minority groups. This overlap population has policy relevance since it targets covariate profiles at the intersection of the White and minority populations covariates distributions. The generalized overlap weighting scheme also bypasses the extremity issue of IPW scheme, evidenced by the fact that the largest generalized overlap weight is normalized to be only 0.0017. Finally, the lack of overlap is also apparent when applying the optimal trimming rule (4.4), which excludes 18.5% of the sample (2125 Whites, 44 Asians, 1001 Blacks and 603 Hispanics) or equivalently 3773 individuals in total.

Based on the estimated generalized propensity scores, we calculate for each covariate the values of $\max_j |\text{PSD}_j|$ and $\max_{j<j'} |\text{ASD}_{j,j'}|$, which are two criteria defined in Section 4.3.2 to examine covariate mean balance in the corresponding weighted populations. Figure 4.2 presents the boxplot of PSD and ASD corresponding to each weighting method. The gray horizontal line indicates adequate balance at 0.1,

and the labels on the a-axis indicate balance for unweighted original sample (Unweighted), inverse probability weighting (IPW), inverse probability weighting combined with optimal trimming (IPW-TM), generalized matching weighting (G-MW) and generalized overlap weighting (G-OW). Due to the extreme propensities, inverse probability weights result in even worse covariate balance than no weighting at all, yielding substantial imbalance of several covariates. By contrast, optimal trimming, the generalized matching weights, and generalized overlap weights lead to satisfactory balance in their respective target populations, with the best balance achieved by generalized overlap weights. The two balance criteria perform similarly in this application.

4.4.3 Results

We estimate the weighted average controlled difference in total health care expenditure between all pairs of racial groups using the generalized overlap weights, and report the point estimates and 95% confidence intervals in the last row of Table 4.2. By emphasizing the subpopulation where each racial group have the most similar characteristics, we find Whites are associated with the highest expenditure ($4097), while Blacks have the second highest ($3212), followed by Asians ($2937) and Hispanics ($2876). The total health expenditure for Whites is $1160, $886 and $1221 higher than Asians, Blacks and Hispanics, and these differences are all statistically significant ($p < 0.001$ for all three comparisons). Among the minority groups, the total expenditure for Blacks is $335 and $274 higher than Hispanics and Asians, but these differences are not statistically significant. Overall, there is evidence that minority groups have spent considerably less health expenditures than Whites, given all other demographic and health-related information. One implication from the analysis is that health policy decision makers could potentially provide more resources and infrastructures for the minority groups to improve their access to medical facilities

as a means to minimize the White-minority disparities in health expenditures.

Table 4.2: Weighted average controlled difference in total health care expenditure (dollars).

	W-A	W-B	W-H	A-B	A-H	B-H
DIF	2764 (2317, 3216)	786 (346, 1234)	2651 (2288, 2997)	-1978 (-2499, -1461)	-113 (-566, 335)	1865 (1426, 2328)
IPW	2402 (530, 4274)	908 (505, 1311)	719 (129, 1309)	-1494 (-3385, 397)	-1683 (-3621, 255)	-189 (-836, 459)
IPW-TM	1335 (671, 1999)	1148 (781, 1515)	1257 (804, 1711)	-187 (-872, 499)	-77 (-812, 657)	109 (-375, 594)
GPSM	1818 (1091, 2545)	827 (334, 1320)	902 (439, 1364)	-991 (-1796, -185)	-916 (-1700, -132)	74 (-504, 652)
GPSM-TM	1402 (814, 1989)	1147 (689, 1605)	1392 (927, 1856)	-255 (-922, 411)	-10 (-680, 660)	245 (-314, 804)
G-MW	1112 (648, 1569)	839 (455, 1239)	1234 (813, 1623)	-273 (-737, 281)	122 (-385, 621)	395 (-100, 820)
G-OW	1160 (660, 1661)	886 (518, 1253)	1221 (849, 1593)	-274 (-813, 264)	61 (-479, 601)	335 (-82, 752)

* W: non-Hispanic Whites; A: Asians; B: Blacks; H: Hispanics.

We further consider the alternative weighting and matching estimators as mentioned in Section 4.4.1. The associated point estimates and 95% confidence intervals are also presented in Table 4.2. Specifically, the 95% confidence intervals were obtained using: 2.5 and 97.5 quantiles of 1000 bootstrap samples for DIF and G-MW; point estimates $\pm$ 1.96 $\times$ (empirical sandwich variance)$^{1/2}$ for IPW; point estimates $\pm$ 1.96 $\times$ (Abadie and Imbens variance)$^{1/2}$ for GPSM (Abadie and Imbens, 2012). Similar to the sandwich variance of G-OW, the empirical sandwich variance estimator for IPW takes into account the uncertainty in estimating the generalized propensity scores (see Section 4.7.4 for the derivation of the IPW sandwich variance, which extends the work of Lunceford and Davidian (2004)). By contrast, the weight function $w_j(\boldsymbol{X})$ for G-MW is not everywhere differentiable and fails to satisfy the regularity conditions for deriving a sandwich variance. Because it is generally difficult to smoothly approximate $w_j(\boldsymbol{X})$ around its infinite-many non-differentiable points, we

resort to the computationally-intensive bootstrap approach for interval estimation. The Abadie and Imbens variance for GPSM also takes into account the uncertainty involving in both the matching process and propensity score estimation. Finally, whenever trimming is used (IPW-TM and GPSM-TM), the generalized propensity scores are re-estimated based on the trimmed sample as refitting generally improves the finite-sample performance of the resulting estimators (Li et al., 2018); accordingly, variance calculation is carried out based on the trimmed sample.

We find disparity estimates differ between different methods. For example, the weighted average difference between Whites and Asians in health care expenditure is estimated as $2402 and $1818 from IPW and matching. The optimal trimming reduces the White-Asian disparity estimates for IPW and GPSM ($1335 and $1402) and tightens the confidence intervals, whereas the White-Asian disparity estimates are even smaller ($1112 and $ 1160) according to G-MW and G-OW. The same pattern is also observed for Asian-Black disparity estimates. On the other hand, IPW and GPSM report much higher expenditure for Hispanics compared to Asians (a statistically significant difference is reported by GPSM with $p = 0.022$), while G-OW reverses the sign and reports slightly higher ($p = 0.825$) expenditure for Asians that are most comparable to other racial groups. Overall, G-MW provides point and interval estimates close to G-OW in this application, but computing its bootstrap intervals takes much longer time than calculating the closed-form intervals of G-OW.

4.4.4 Effective Sample Size

To further compare the different weighting methods, we calculate the effective sample size (ESS) of each racial group according to each weighting scheme. Following McCaffrey et al. (2013), we define the ESS for group j as

$$\mathrm{ESS}_j^h = \frac{\left(\sum_{i=1}^n \sum_{j=1}^J D_{ij} w_j(\boldsymbol{X}_i)\right)^2}{\sum_{i=1}^n \sum_{j=1}^J D_{ij} w_j^2(\boldsymbol{X}_i)}.$$

As weighting generally increases the variance compared to the unweighted estimates based on the same sample, the ESS may serve as a conservative measure to characterize the variance inflation or precision loss due to weighting. Table 4.3 presents the ESS (estimated by plugging in the estimated weights in the definition) and the total ESS. It is evident that all weighting methods reduce the ESS compared to the original sample. However, IPW results in a very small value of ESS for Asians relative to the original group size, signaling the presence of extreme weights and lack of overlap. This observation further underlies the wide confidence intervals for pairwise differences associated with IPW. By contrast, G-MW, G-OW and trimming result in relatively balanced ESS across groups. Among them, G-OW reports the largest total ESS, supporting its asymptotic efficiency optimality among all balancing weighting schemes.

Table 4.3: Effective sample size (ESS) of each weighted group according to different weighting methods.

	Whites	Asians	Blacks	Hispanics	Total
Unweighted	9830	1446	4020	5150	20446
IPW	8371	10	2549	2482	13412
Unweighted (Trimmed)	7705	1402	3019	4547	16673
IPW-TM	6524	695	2183	3071	12473
G-MW	4937	1285	1875	3176	11273
G-OW	6015	1166	2234	3756	13171

4.5 Simulations

To further understand the results of the racial disparity application and, more importantly, to further shed light on the comparison between different weighting methods, we conduct a simulation study for estimating pairwise causal effects with multiple treatments. Our data generating process is similar to the one in Yang et al. (2016) except that we consider nonzero pairwise average treatment effect among the consid-

ered target populations. We generate covariates X_{i1}, X_{2i} and X_{3i} from a multivariate normal distribution with mean vector $(2, 1, 1)$ and covariances of $(1, -1, -0.5)$; $X_{4i} \sim \text{Uniform}[-3, 3]$; $X_{5i} \sim \chi^2_1$ and $X_{6i} \sim \text{Bernoulli}(0.5)$, with the covariate vector $\boldsymbol{X}_i^T = (X_{1i}, X_{2i}, X_{3i}, X_{4i}, X_{5i}, X_{6i})$. The assignment mechanism follows the multinomial logistic regression

$$(D_{i1}, \ldots, D_{iJ}) \sim \text{Multinom}(e_1(\boldsymbol{X}_i), \ldots, e_J(\boldsymbol{X}_i)),$$

where D_{ij} is the treatment indicator defined in Section 4.2.1 and $e_j(\boldsymbol{X}_i) = \exp(\alpha_j + \boldsymbol{X}_i^T\boldsymbol{\beta}_j)/\sum_{k=1}^{J}\exp(\alpha_k + \boldsymbol{X}_i^T\boldsymbol{\beta}_k)$ is the true generalized propensity score with $\alpha_1 = 0$, $\boldsymbol{\beta}_1^T = (0, 0, 0, 0, 0, 0)$. In the first simulation with $J = 3$ treatment groups, $\boldsymbol{\beta}_2^T = \kappa_2 \times (1, 1, 1, -1, -1, 1)$ and $\boldsymbol{\beta}_3^T = \kappa_3 \times (1, 1, 1, 1, 1, 1)$. We set $(\kappa_2, \kappa_3) = (0.2, 0.1)$ to simulate a scenario with adequate covariate overlap and $(\kappa_2, \kappa_3) = (0.8, 0.4)$ to induce lack of overlap with strong propensity tails, i.e., the propensity to receive certain treatment is close to zero for specific design values. We further choose α_2 and α_3 so that the overall treatment proportions are fixed at $(0.3, 0.4, 0.3)$. The potential outcomes are generated from $Y_i(j) = \left(1, \boldsymbol{X}_i^T\right)\boldsymbol{\gamma}_j + \epsilon_i$ with $\epsilon_i \sim N(0, 1)$, $\boldsymbol{\gamma}_1^T = (-1.5, 1, 1, 1, 1, 1, 1)$, $\boldsymbol{\gamma}_2^T = (-4, 2, 3, 1, 2, 2, 2)$ and $\boldsymbol{\gamma}_3^T = (3, 3, 1, 2, -1, -1, -1)$. In the second simulation with $J = 6$ groups, we specify $\boldsymbol{\beta}_2^T = \kappa_2 \times (1, 1, 2, 1, 1, 1)$, $\boldsymbol{\beta}_3^T = \kappa_3 \times (1, 1, 1, 1, 1, -5)$, $\boldsymbol{\beta}_4^T = \kappa_4 \times (1, 1, 1, 1, 1, 5)$, $\boldsymbol{\beta}_5^T = \kappa_5 \times (1, 1, 1, -2, 1, 1)$ and $\boldsymbol{\beta}_6^T = \kappa_6 \times (1, 1, 1, -2, -1, 1)$. We use $(\kappa_2, \kappa_3, \kappa_4, \kappa_5, \kappa_6) = (0.1, 0.15, 0.2, 0.25, 0.3)$ to simulate a scenario with adequate overlap and $(\kappa_2, \kappa_3, \kappa_4, \kappa_5, \kappa_6) = (0.4, 0.6, 0.8, 1, 1.2)$ to represent a challenging scenario with strong propensity tails. The intercepts are chosen so that the marginal treatment proportions are fixed around $(0.12, 0.16, 0.12, 0.25, 0.2, 0.15)$. Finally, the coefficients for the outcome model is specified as $\boldsymbol{\gamma}_1^T = (-1.5, 1, 1, 1, 1, 1, 1)$, $\boldsymbol{\gamma}_2^T = (-4, 2, 3, 1, 2, 2, 2)$, $\boldsymbol{\gamma}_3^T = (4, 3, 1, 2, -1, -1, -4)$, $\boldsymbol{\gamma}_4^T = (1, 4, 1, 2, -1, -1, -3)$, $\boldsymbol{\gamma}_5^T = (3.5, 5, 1, 2, -1, -1, -2)$ and $\boldsymbol{\gamma}_6^T = (3.5, 6, 1, 2, -1, -1, -1)$. The total sample size is fixed at $n = 1500$ for $J = 3$ and $n = 6000$ for $J = 6$.

Table 4.4: Simulation results with $J = 3$ treatment groups and adequate overlap.

	Absolute Bias			RMSE			Coverage of 95% CI		
	$\tau(\boldsymbol{\lambda}_{1,2})$	$\tau(\boldsymbol{\lambda}_{1,3})$	$\tau(\boldsymbol{\lambda}_{2,3})$	$\tau(\boldsymbol{\lambda}_{1,2})$	$\tau(\boldsymbol{\lambda}_{1,3})$	$\tau(\boldsymbol{\lambda}_{2,3})$	$\tau(\boldsymbol{\lambda}_{1,2})$	$\tau(\boldsymbol{\lambda}_{1,3})$	$\tau(\boldsymbol{\lambda}_{2,3})$
DIF	0.46	0.60	0.14	0.55	0.65	0.37	0.64	0.36	0.92
IPW	0.02	0.01	0.01	0.20	0.16	0.26	0.92	0.95	0.95
IPW-TM	0.01	0.002	0.01	0.16	0.16	0.23	0.94	0.94	0.94
GPSM	0.02	0.01	0.01	0.26	0.22	0.31	0.99	0.97	0.97
GPSM-TM	0.02	0.004	0.01	0.25	0.23	0.31	0.98	0.96	0.98
G-MW	0.02	0.01	0.02	0.17	0.18	0.27	0.95	0.96	0.94
G-OW	0.01	0.001	0.01	0.15	0.15	0.22	0.94	0.96	0.95

* In this case, the optimally trimming excludes at most 2% of the total sample.

Table 4.5: Simulation results with $J = 3$ treatment groups and lack of overlap.

	Absolute Bias			RMSE			Coverage of 95% CI		
	$\tau(\boldsymbol{\lambda}_{1,2})$	$\tau(\boldsymbol{\lambda}_{1,3})$	$\tau(\boldsymbol{\lambda}_{2,3})$	$\tau(\boldsymbol{\lambda}_{1,2})$	$\tau(\boldsymbol{\lambda}_{1,3})$	$\tau(\boldsymbol{\lambda}_{2,3})$	$\tau(\boldsymbol{\lambda}_{1,2})$	$\tau(\boldsymbol{\lambda}_{1,3})$	$\tau(\boldsymbol{\lambda}_{2,3})$
DIF	0.43	0.64	0.21	0.50	0.68	0.38	0.65	0.23	0.90
IPW	0.19	0.02	0.17	1.04	0.61	1.16	0.79	0.88	0.91
IPW-TM	0.03	0.01	0.01	0.38	0.28	0.47	0.93	0.90	0.91
GPSM	0.25	0.10	0.15	0.86	0.51	0.90	0.88	0.91	0.91
GPSM-TM	0.08	0.02	0.05	0.53	0.37	0.60	0.95	0.92	0.95
G-MW	0.001	0.01	0.01	0.29	0.24	0.36	0.95	0.95	0.95
G-OW	0.01	0.01	0.003	0.28	0.23	0.35	0.95	0.94	0.94

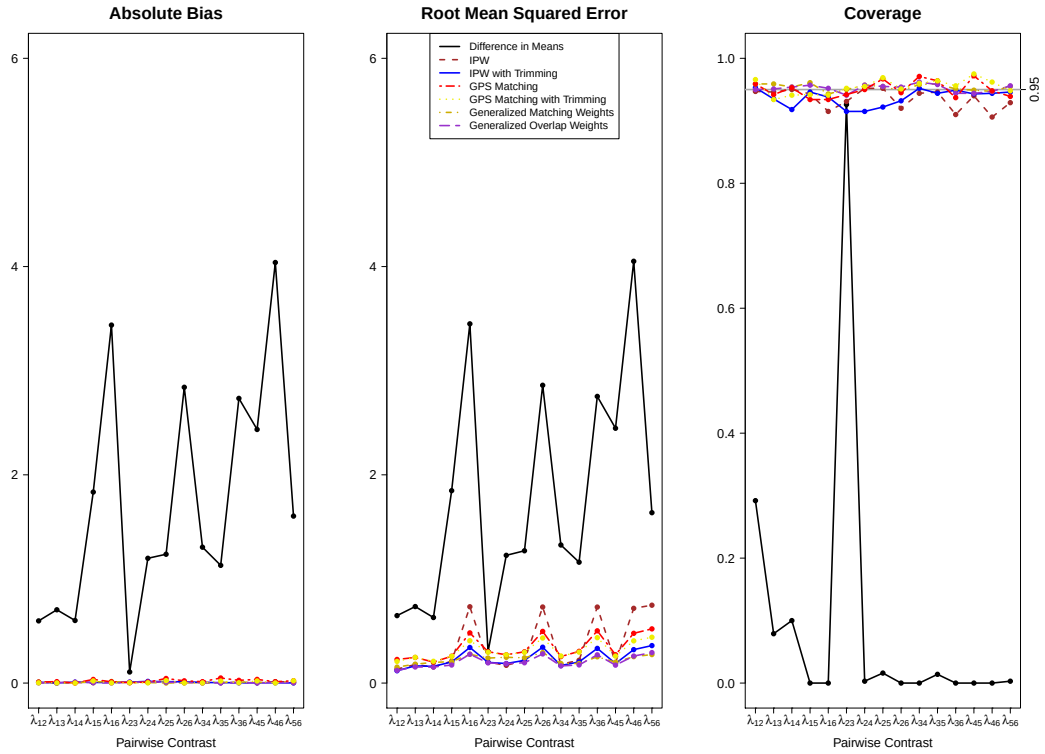
* In this case, the optimal trimming rule excludes 19% to 30% of the total sample.

For each scenario, we simulate 1000 datasets and estimate the pairwise causal effects using the seven estimators examined in Section 4.4. Because the target population may differ in different estimators, we assess the accuracy of estimators relative to their corresponding target estimands. Specifically, the target estimands of DIF, IPW and GPSM are pairwise ATE for the combined population and are analytically determined from the true potential outcome model, whereas the target estimands for G-MW, G-OW, IPW-TM and GPSM-TM are defined for specific subpopulations and evaluated numerically based on Monte Carlo integration. For each replicate, we estimate the generalized propensity scores based on the correct multinomial logistic regression model including all covariates. The 95% confidence intervals were obtained using the same methods described in Section 4.4.3.

Table 4.4 and 4.5 summarize the absolute bias, root mean squared error (RMSE) and coverage of each estimator with $J = 3$ groups. As expected, DIF shows substantial bias and under-coverage, characterizing the strong confounding in the simulations. All other approaches perform reasonably well when there is adequate overlap. However, with lack of overlap, IPW and GPSM are sensitive to extreme propensities and produce biased point estimates. The optimal trimming method excludes 19% to 30% of the total sample, reduces the bias and improves efficiency and coverage in estimating the subpopulation causal effects. By down-weighting extreme units, both G-MW and G-OW provide unbiased point estimates with nominal coverage. Overall, IPW-TM, G-MW and G-OW are associated with the smallest RMSE and are more efficient than the other methods. Among them, G-OW is the most efficient with the smallest RMSE, matching the theoretical predictions in Section 4.2.4.

The simulation results with $J = 6$ groups are presented in Figure 4.4 and Figure 4.5. With adequate overlap, all methods have good control of confounding bias, produce unbiased estimates and close to nominal coverage. G-MW and G-OW provide the lowest RMSE, with the latter demonstrating higher efficiency for estimating most of the causal contrasts (the ratio of total MSE is 1.18). With lack of overlap, the clear separation of covariate space makes it challenging to simultaneously remove all confounding for estimating the 15 pairwise contrasts. By discarding more than half of the sample, the optimal trimming method improves the bias, efficiency and coverage properties over IPW and GPSM, both of which are subject to bias and excessive variance with extreme propensities. G-MW and G-OW further improve the efficiency and coverage properties upon trimming by down-weighting the extreme units. As expected from the large-sample theory, G-OW produces more efficient estimates than G-MW for 12 out of 15 causal contrasts (the ratio of total MSE is 1.17). In this challenging scenario, the bootstrap CI for G-MW has slightly better finite-sample coverage than the closed-form CI for G-OW based on the empirical sandwich

variance, but the closed-form CI estimator for G-OW demonstrates the best coverage among all the considered closed-form CI estimators. However, another substantial gain of G-OW over G-MW is the computational time: for each simulation, the bootstrap interval estimates for G-MW with 1000 resamples require more than 80 times longer running time than that of the closed-form G-OW interval estimates, which can be very burdensome for large data sets.

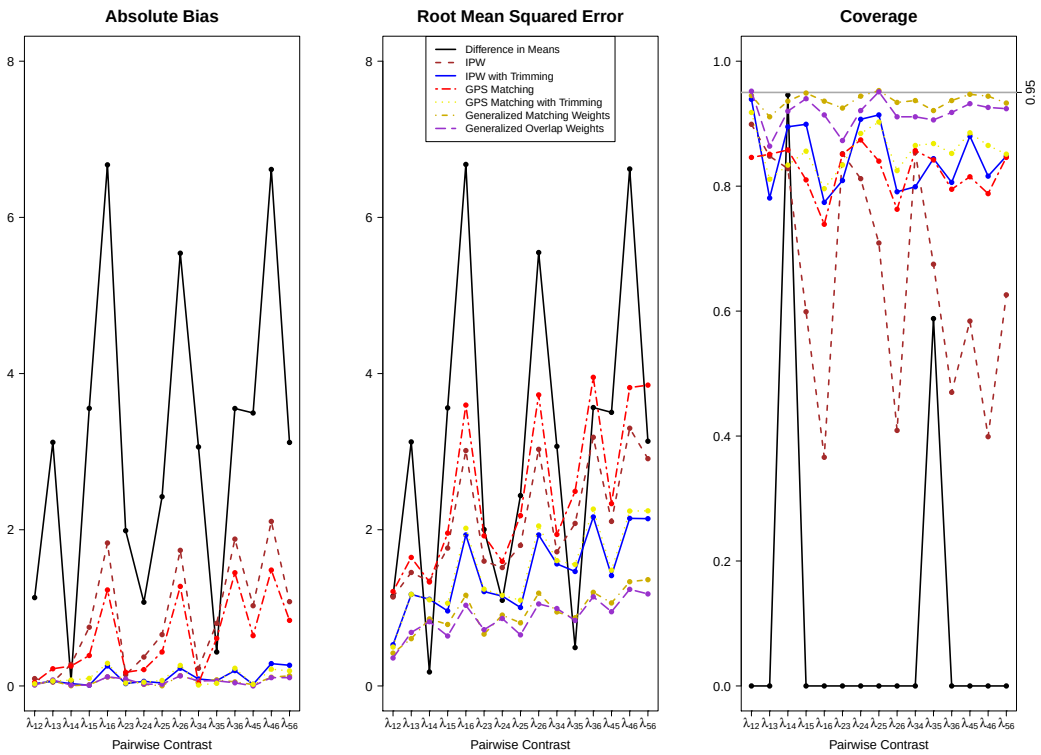


Optimal trimming excludes $3\% \sim 7\%$ *of the total sample. For a given approach, each one of the* 15 *causal comparisons is represented by the contrast* $\boldsymbol{\lambda}_{j,j'}$ *for notational simplicity.*

FIGURE 4.4: Simulation results with $J = 6$ treatment groups and $(\kappa_2, \kappa_3, \kappa_4, \kappa_5, \kappa_6) = (0.1, 0.15, 0.2, 0.25, 0.3)$, i.e., with adequate overlap.

4.6 Discussion

In this Chapter, we proposed a unified framework, the balancing weights, for estimating causal effects with multiple treatments. Within this framework, focusing on



Optimal trimming excludes 52% $\sim$ 74% *of the total sample. For a given approach, each one of the* 15 *causal comparisons is represented by the contrast* $\boldsymbol{\lambda}_{j,j'}$ *for notational simplicity.*

FIGURE 4.5: Simulation results with $J = 6$ treatment groups and $(\kappa_2, \kappa_3, \kappa_4, \kappa_5, \kappa_6) = (0.4, 0.6, 0.8, 1, 1.2)$, i.e., with strong propensity tails.

pairwise comparisons for nominal treatments, we proposed the generalized overlap weights to emphasize the target population with the most covariate overlap across all groups. We show that the generalized overlap weights minimize the total asymptotic variance of the nonparametric estimators for the pairwise contrasts within the class of balancing weights.

One arguable limitation of the generalized overlap weights is that its corresponding causal estimand—the pairwise ATO—is defined through the true propensity scores. An obvious concern is that is the estimand still meaningful if the propensity score is misspecified in the analysis? Below we argue that this concern is not specific to overlap weights; in fact, it is general to propensity score methods, and the core

issue is the difference between *sample* and *target population*. Specifically, take the popular IPW approach for example, IPW corresponds to the ATE estimand, the definition of which indeed does not rely on the propensity score. However, modern observational datasets are often convenience samples resulting from vague and debatable inclusion criteria and consequently do not represent a scientifically interpretable population. Consequently, applying IPW to such a sample does not lead to an ATE on a meaningful target population. Moreover, in practice IPW is usually implemented with trimming, which essentially modifies the causal estimand and the estimand indeed depends on the estimated propensity scores. The same phenomenon applies to propensity score matching. Specifically, different estimates of the propensity scores, tuning parameters and algorithms can lead to very different matched samples, which in turn might represent different target populations. The causal estimand is implicitly dependent on the matched sample and thus the propensity scores. Unless the target populations are well-defined *a priori* and then systematically and comprehensively sampled, effectively all propensity score methods lead to causal estimands that are implicitly or explicitly dependent on the propensity scores. The overlap weighting scheme attempts to address this *sample-population* discordance by moving the goalpost toward the causal comparison with the most internal validity, namely, on the subpopulation with the most overlap in observed characteristics.

Nonetheless, a well-estimated propensity score is crucial to the implementation of any generalized balancing weighting scheme. One important avenue for future research is to develop flexible semiparametric or nonparametric propensity score models as in Mercatanti and Li (2014) and McCaffrey et al. (2013). In particular, McCaffrey et al. (2004) and Ridgeway et al. (2006) have developed the Generalized Boosting Model (GBM), a multivariate nonparametric technique, for the estimation of the propensity scores with associated software (McCaffrey et al., 2013). Prior studies have shown that these flexible methods can offer considerable improvement

in balancing the covariate distribution towards the target population when estimating the ATE or ATT (e.g. Lee, Lessler, and Stuart, 2009). It is of interest to extend the GBM algorithm for the estimation of generalized overlap weights, by incorporating the balancing constraints discussed in Section 4.3.2.

Another potential improvement is to consider the class of augmented weighting estimators, as the current sample estimator (4.5) does not exploit smoothness of the outcome in each level of the treatment and thus may not achieve semiparametric efficiency. One could in fact construct, for each choice of the balancing weights, a regression-augmented estimator for the expectation of the potential outcomes among the target population as

$$\hat{m}_j^{h,\text{aug}} = \hat{m}_j^h - \frac{\sum_{i=1}^n (D_{ij} - e_j(\boldsymbol{X}_i)) w_j(\boldsymbol{X}_i) \hat{m}_j(\boldsymbol{X}_i)}{\sum_{i=1}^n h(\boldsymbol{X}_i)},$$

where $\hat{m}_j(\boldsymbol{X}_i) = \hat{\mathrm{E}}[Y(j)|\boldsymbol{X}]$ is the outcome regression function. It can be shown that, under weak unconfoundedness, $\hat{m}_j^{h,\text{aug}}$ achieves the semiparametric efficiency bound for estimating m_j^h when both the generalized propensity score model and the regression function are correctly specified. Of note, when the tilting function $h(\boldsymbol{X}_i) = 1$, $\hat{m}_j^{h,\text{aug}}$ has an additional double-robustness property such that it is consistent to $\mathrm{E}[Y(j)]$ when either the generalized propensity score model or the regression function is correctly specified, but not necessarily both. However, this robustness property does not hold in general for $\hat{m}_j^{h,\text{aug}}$ concerning other choices of h and balancing weights, including the generalized overlap weights. Nevertheless, additional work is warranted to study the efficiency property of the augmented generalized overlap weighting estimator in the context of multiple treatments; for example, one can consider more flexible semiparametric outcome regression models as in Mercatanti and Li (2014).

4.7 Technical Proofs of the Theorems

For proving the Theorems, we assume regularity conditions on $m_j(\boldsymbol{X}) = \mathrm{E}[Y(j)|\boldsymbol{X}]$ and $v_j(\boldsymbol{X}) = \mathrm{V}(Y(j)|\boldsymbol{X})$ necessary to ensure that the integrals are well defined.

4.7.1 Proof of Theorem 4.2.5

By definition of the generalized propensity score, we must have $\mathrm{E}[\mathbb{1}(Z = j)/e_j(\boldsymbol{X})|\boldsymbol{X}] = 1$ for all $j \in \mathbb{Z}$. Then the average of the potential outcomes in target population h

$$\begin{aligned} m_j^h &= \frac{\int_{\mathbb{X}} m_j(\boldsymbol{X}) f(\boldsymbol{X}) h(\boldsymbol{X}) \mu(d\boldsymbol{X})}{\int_{\mathbb{X}} f(\boldsymbol{X}) h(\boldsymbol{X}) \mu(d\boldsymbol{X})} \\ &= \frac{\int_{\mathbb{X}} \mathrm{E}[\mathbb{1}\{Z = j\} Y(j) (h(\boldsymbol{X})/e_j(\boldsymbol{X}))|\boldsymbol{X}] f(\boldsymbol{X}) \mu(d\boldsymbol{X})}{\int_{\mathbb{X}} \mathrm{E}[\mathbb{1}\{Z = j\} (h(\boldsymbol{X})/e_j(\boldsymbol{X}))|\boldsymbol{X}] f(\boldsymbol{X}) \mu(d\boldsymbol{X})} \\ &= \frac{\int_{\mathbb{X}} \mathrm{E}[\mathbb{1}\{Z = j\} Y(j) w_j(\boldsymbol{X})|\boldsymbol{X}] f(\boldsymbol{X}) \mu(d\boldsymbol{X})}{\int_{\mathbb{X}} \mathrm{E}[\mathbb{1}\{Z = j\} w_j(\boldsymbol{X})|\boldsymbol{X}] f(\boldsymbol{X}) \mu(d\boldsymbol{X})} \end{aligned} \tag{4.8}$$

where the second equation holds due to the weak unconfoundedness assumption, $Y(j) \perp \mathbb{1}\{Z = j\}|\boldsymbol{X}$ (Imbens, 2000). Because $D_{ij} = \mathbb{1}\{Z_i = j\}$, it follows that the estimators, $n^{-1} \sum_{i=1}^n D_{ij} Y_i w_j(\boldsymbol{X}_i)$ and $n^{-1} \sum_{i=1}^n D_{ij} w_j(\boldsymbol{X}_i)$, consistently estimate the numerator and denominator of (4.8). Therefore, $\hat{m}_j^h = \sum_{i=1}^n D_{ij} Y_i w_j(\boldsymbol{X}_i) \Big/ \sum_{i=1}^n D_{ij} w_j(\boldsymbol{X}_i)$ is consistent for m_j^h, and $\hat{\tau}^h(\boldsymbol{a}) = \sum_{j=1}^J a_j \hat{m}_j^h$ must be consistent for $\tau^h(\boldsymbol{a}) = \sum_{j=1}^J a_j m_j^h$.

4.7.2 Proof of Theorem 4.2.6

By SUTVA (Imbens and Rubin, 2015), we write

$$\hat{\tau}_{\boldsymbol{a}}^h = \sum_{j=1}^J a_j \frac{\sum_{i=1}^n D_{ij} Y_i w_j(\boldsymbol{X}_i)}{\sum_{i=1}^n D_{ij} w_j(\boldsymbol{X}_i)} = \sum_{j=1}^J a_j \frac{\sum_{i=1}^n D_{ij} Y_i(j) w_j(\boldsymbol{X}_i)}{\sum_{i=1}^n D_{ij} w_j(\boldsymbol{X}_i)}.$$

Conditional on the assignment $\underline{\boldsymbol{Z}}$ and sample design $\underline{\boldsymbol{X}}$, only the potential outcomes are random. Therefore the residual variance of $\hat{\tau}^h(\boldsymbol{a})$ is

$$\begin{aligned}
\mathrm{V}[\hat{\tau}^h(\boldsymbol{a})|\underline{\boldsymbol{Z}},\underline{\boldsymbol{X}}] &= \sum_{j=1}^{J} a_j^2 \frac{\sum_{i=1}^n v_j(\boldsymbol{X}_i) D_{ij} w_j^2(\boldsymbol{X}_i)}{[\sum_{i=1}^n D_{ij} w_j(\boldsymbol{X}_i)]^2} \\
&= \sum_{j=1}^{J} a_j^2 \frac{\sum_{i=1}^n \{v_j(\boldsymbol{X}_i)/e_j(\boldsymbol{X}_i)\}\{D_{ij}/e_j(\boldsymbol{X}_i)\} h^2(\boldsymbol{X}_i)}{[\sum_{i=1}^n \{D_{ij}/e_j(\boldsymbol{X}_i)\} h(\boldsymbol{X}_i)]^2}.
\end{aligned}$$

Averaging over the joint distribution of $\underline{\boldsymbol{Z}}$ and $\underline{\boldsymbol{X}}$, we observe by the Weak Law of Large Numbers that

$$\frac{1}{n}\sum_{i=1}^{n} \{D_{ij}/e_j(\boldsymbol{X}_i)\} h(\boldsymbol{X}_i) \xrightarrow{p} \int_{\mathbb{X}} \mathrm{E}[\mathbb{1}\{Z=j\}/e_j(\boldsymbol{X})|\boldsymbol{X}] h(\boldsymbol{X}) f(\boldsymbol{X}) \mu(d\boldsymbol{X}) = C_h,$$

and

$$\begin{aligned}
& \frac{1}{n}\sum_{i=1}^{n} \{v_j(\boldsymbol{X}_i)/e_j(\boldsymbol{X}_i)\}\{D_{ij}/e_j(\boldsymbol{X}_i)\} h^2(\boldsymbol{X}_i) \\
\xrightarrow{p} & \int_{\mathbb{X}} v_j(\boldsymbol{X})/e_j(\boldsymbol{X})\, \mathrm{E}[\mathbb{1}\{Z=j\}/e_j(\boldsymbol{X})|\boldsymbol{X}] h^2(\boldsymbol{X}) f(\boldsymbol{X}) \mu(d\boldsymbol{X}) \\
= & \int_{\mathbb{X}} \{v_j(\boldsymbol{X})/e_j(\boldsymbol{X})\} h^2(\boldsymbol{X}) f(\boldsymbol{X}) \mu(d\boldsymbol{X})
\end{aligned}$$

An application of the Slutsky's Theorem shows $n \cdot \mathrm{V}[\hat{\tau}^h(\boldsymbol{a})|\underline{\boldsymbol{Z}},\underline{\boldsymbol{X}}] \xrightarrow{p} Q(\boldsymbol{a},h)$, where $Q(\boldsymbol{a},h)$ is a constant defined in Theorem 2. The uniform integrability assumption for the family of random variables $\{\mathrm{V}[\hat{\tau}^h(\boldsymbol{a})|\underline{\boldsymbol{Z}},\underline{\boldsymbol{X}}], n \geqslant 1\}$ then gives the desired L_1 convergence result.

4.7.3 Proof of Theorem 4.2.7

For notational simplicity, we use the $\mathrm{E}[\cdot]$ operator to represent $\int_{\mathbb{X}} \cdot f(\boldsymbol{X})\mu(d\boldsymbol{X})$. Under homoscedasticity, $v_j(\boldsymbol{X}) = v$,

$$\begin{aligned} Q(\boldsymbol{a}, h) &= (v/C_h^2) \int_{\mathbb{X}} \Big(\sum_{j=1}^{J} a_j^2/e_j(\boldsymbol{X}) \Big) h^2(\boldsymbol{X}) f(\boldsymbol{X}) \mu(d\boldsymbol{X}) \\ &= (v/C_h^2)\, \mathrm{E}\Big\{ h^2(\boldsymbol{X}) \Big(\sum_{j=1}^{J} a_j^2/e_j(\boldsymbol{X}) \Big) \Big\}. \end{aligned}$$

Applying the Cauchy-Schwarz inequality, we have

$$\begin{aligned} C_h^2 = [\mathrm{E}\{h(\boldsymbol{X})\}]^2 &= \Big[\mathrm{E}\Big\{ h(\boldsymbol{X}) \Big(\sum_{j=1}^{J} a_j^2/e_j(\boldsymbol{X}) \Big)^{1/2} \Big(\sum_{j=1}^{J} a_j^2/e_j(\boldsymbol{X}) \Big)^{-1/2} \Big\} \Big]^2 \\ &\leqslant \mathrm{E}\Big\{ h^2(\boldsymbol{X}) \Big(\sum_{j=1}^{J} a_j^2/e_j(\boldsymbol{X}) \Big) \Big\}\, \mathrm{E}\Big\{ \Big(\sum_{j=1}^{J} a_j^2/e_j(\boldsymbol{X}) \Big)^{-1} \Big\}, \end{aligned}$$

and the equality is attained when $h = \tilde{h}(\boldsymbol{X}) \propto \Big(\sum_{j=1}^{J} a_j^2/e_j(\boldsymbol{X}) \Big)^{-1}$. This implies that

$$\mathrm{E}\Big\{ h^2(\boldsymbol{X}) \Big(\sum_{j=1}^{J} a_j^2/e_j(\boldsymbol{X}) \Big) \Big\} \Big/ C_h^2 \geqslant \Big[\mathrm{E}\Big\{ \Big(\sum_{j=1}^{J} a_j^2/e_j(\boldsymbol{X}) \Big)^{-1} \Big\} \Big]^{-1} = C_{\tilde{h}}^{-1},$$

which gives $Q(\boldsymbol{a}, \tilde{h}) = v/C_{\tilde{h}}$.

4.7.4 Proof of Theorem 4.3.1 and Related Remarks

From the multinomial logistic model, we have for $i = 1, \ldots, n$,

$$\begin{aligned} e_1(\boldsymbol{X}_i) &= Pr(Z_i = 1 | \boldsymbol{X}_i) = \frac{1}{1 + \sum_{k=2}^{J} \exp(\alpha_k + \boldsymbol{X}_i^T \boldsymbol{\beta}_k)}. \\ e_j(\boldsymbol{X}_i) &= Pr(Z_i = j | \boldsymbol{X}_i) = \frac{\exp(\alpha_j + \boldsymbol{X}_i^T \boldsymbol{\beta}_j)}{1 + \sum_{k=2}^{J} \exp(\alpha_k + \boldsymbol{X}_i^T \boldsymbol{\beta}_k)}, \quad j = 2, \ldots, J. \end{aligned}$$

Since $D_{ij} = \mathbb{1}\{Z_i = j\}$, it is straightforward to show that the log likelihood function

$$l(\boldsymbol{\theta}) = \sum_{i=1}^{n} l_i(\boldsymbol{\theta}) = \sum_{i=1}^{n} \Big[\sum_{j=2}^{J} \Big\{ D_{ij}(\alpha_j + \boldsymbol{X}_i^T \boldsymbol{\beta}_j) \Big\} - \log \Big\{ 1 + \sum_{k=2}^{J} \exp(\alpha_k + \boldsymbol{X}_i^T \boldsymbol{\beta}_k) \Big\} \Big]$$

When the estimation of model parameters is carried out by maximum likelihood, the first-order condition is obtained by differentiating the log likelihood with respect to $\boldsymbol{\theta}$,

$$\begin{aligned} \mathbf{0} &= \boldsymbol{S}_{\boldsymbol{\theta}} = \sum_{i=1}^{n} \boldsymbol{S}_{\boldsymbol{\theta},i} = \sum_{i=1}^{n} \frac{\partial}{\partial \boldsymbol{\theta}} l_i(\boldsymbol{\theta}) \\ &= \sum_{i=1}^{n} \left(\frac{\partial}{\partial \alpha_2} l_i(\boldsymbol{\theta}), \ldots, \frac{\partial}{\partial \alpha_J} l_i(\boldsymbol{\theta}), \frac{\partial}{\partial \boldsymbol{\beta}_2^T} l_i(\boldsymbol{\theta}), \ldots, \frac{\partial}{\partial \boldsymbol{\beta}_J^T} l_i(\boldsymbol{\theta}) \right)^T, \end{aligned} \tag{4.9}$$

where for $l = 2, \ldots, J$,

$$\frac{\partial}{\partial \boldsymbol{\beta}_l} l_i(\boldsymbol{\theta}) = \boldsymbol{X}_i \frac{\partial}{\partial \alpha_l} l_i(\boldsymbol{\theta}) = \boldsymbol{X}_i \{ D_{il} - e_l(\boldsymbol{X}_i) \}.$$

We further let $\boldsymbol{I}_{\boldsymbol{\theta}\boldsymbol{\theta}} = -\mathrm{E}[\frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} l_i(\boldsymbol{\theta})]$ be the information matrix, whose exact form can be expressed in a similar fashion but is omitted here for brevity. We denote a consistent estimator for this information by $\hat{\boldsymbol{I}}_{\boldsymbol{\theta}\boldsymbol{\theta}}$. Under standard regularity conditions (Lehmann, 1983), the stochastic expansion for the maximum likelihood estimator is

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) = \boldsymbol{I}_{\boldsymbol{\theta}\boldsymbol{\theta}}^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \boldsymbol{S}_{\boldsymbol{\theta},i} + o_p(1),$$

where $o_p(1)$ is asymptotically negligible as $n \to \infty$.

With the multinomial logistic model, the generalized overlap weights are expressed as functions of $\boldsymbol{\theta}$:

$$\begin{aligned} w_1(\boldsymbol{X}_i) &= w_1(\boldsymbol{X}_i; \boldsymbol{\theta}) = \frac{1}{1 + \sum_{k=2}^{J} \exp(-\alpha_k - \boldsymbol{X}_i^T \boldsymbol{\beta}_k)} \\ w_j(\boldsymbol{X}_i) &= w_j(\boldsymbol{X}_i; \boldsymbol{\theta}) = \frac{\exp(-\alpha_j - \boldsymbol{X}_i^T \boldsymbol{\beta}_j)}{1 + \sum_{k=2}^{J} \exp(-\alpha_k - \boldsymbol{X}_i^T \boldsymbol{\beta}_k)}, \qquad j = 2, \ldots, J, \end{aligned}$$

and the derivative of the weights takes the form

$$\dot{w}_j(\boldsymbol{X}_i) \equiv \frac{\partial}{\partial \boldsymbol{\theta}} w_j(\boldsymbol{X}_i) = \left(\frac{\partial}{\partial \alpha_2} w_j(\boldsymbol{X}_i), \ldots, \frac{\partial}{\partial \alpha_J} w_j(\boldsymbol{X}_i), \frac{\partial}{\partial \boldsymbol{\beta}_2^T} w_j(\boldsymbol{X}_i), \ldots, \frac{\partial}{\partial \boldsymbol{\beta}_J^T} w_j(\boldsymbol{X}_i) \right)^T,$$

where for $j = 1, \ldots, J$ and $l = 2, \ldots, J$,

$$\frac{\partial}{\partial \boldsymbol{\beta}_l} w_j(\boldsymbol{X}_i) = \boldsymbol{X}_i \frac{\partial}{\partial \alpha_l} w_j(\boldsymbol{X}_i) = \boldsymbol{X}_i \{ w_j(\boldsymbol{X}_i) w_l(\boldsymbol{X}_i) - \delta_{jl} w_l(\boldsymbol{X}_i) \},$$

and $\delta_{jl} = \mathbb{1}\{j = l\}$.

For $j = 1, \ldots, J$, the plug-in weighting estimator $\hat{m}_j^h$ can be regarded as the solution of the following estimating equation

$$\sum_{i=1}^{n} \boldsymbol{U}(\hat{m}_j^h, \hat{\boldsymbol{\theta}}) = \sum_{i=1}^{n} D_{ij}(Y_i - \hat{m}_j^h) w_j(\boldsymbol{X}_i; \hat{\boldsymbol{\theta}}) = \mathbf{0}.$$

Under standard regularity conditions (van der Vaart, 1998), a first-order Taylor expansion of the unbiased estimating equations around the truth leads to

$$\begin{aligned} \sqrt{n}(\hat{m}_j^h - m_j^h) &= \varpi^{-1} \Big\{ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \boldsymbol{U}(m_j^h, \boldsymbol{\theta}) + \boldsymbol{H}_j^T \sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \Big\} + o_p(1) \\ &= \varpi^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Big\{ D_{ij}(Y_i - m_j^h) w_j(\boldsymbol{X}_i) + \boldsymbol{H}_j^T \boldsymbol{I}_{\boldsymbol{\theta}\boldsymbol{\theta}}^{-1} \boldsymbol{S}_{\boldsymbol{\theta},i} \Big\} + o_p(1), \end{aligned} \tag{4.10}$$

where $\varpi = \mathrm{E}[D_{ij} w_j(\boldsymbol{X}_i)] = \mathrm{E}[h(\boldsymbol{X}_i)]$, and $\boldsymbol{H}_j = \mathrm{E}[D_{ij}(Y_i - m_j^h)\dot{w}_j(\boldsymbol{X}_i)] = \mathrm{E}[(Y_i - m_j^h) e_j(\boldsymbol{X}_i)\dot{w}_j(\boldsymbol{X}_i)]$. Therefore, given any fixed coefficient $\boldsymbol{a} = (a_1, \ldots, a_J)'$, we have

$$\sqrt{n}\{\hat{\tau}^h(\boldsymbol{a}) - \tau^h(\boldsymbol{a})\} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Big\{ \frac{1}{\varpi} \sum_{j=1}^{J} a_j \psi_{ij} \Big\} + o_p(1),$$

where we define $\psi_{ij} = D_{ij}(Y_i - m_j^h) w_j(\boldsymbol{X}_i) + \boldsymbol{H}_j^T \boldsymbol{I}_{\boldsymbol{\theta}\boldsymbol{\theta}}^{-1} \boldsymbol{S}_{\boldsymbol{\theta},i}$. Since the triplets $\{Y_i, \boldsymbol{X}_i, Z_i\}'s$ are assumed i.i.d., an application of the standard Central Limit Theorem gives,

$$\sqrt{n}\{\hat{\tau}^h(\boldsymbol{a}) - \tau^h(\boldsymbol{a})\} \xrightarrow{d} \mathcal{N}\left(0, \varpi^{-2} \,\mathrm{E} \Big\{ \sum_{j=1}^{J} a_j \psi_{ij} \Big\}^2 \right).$$

In practice, we use the empirical sandwich estimator to consistently estimate the large-sample variance (Stefanski and Boos, 2002); the variance of $\hat{\tau}^h(\boldsymbol{a})$ is estimated by

$$\frac{1}{(n\hat{\varpi})^2}\sum_{i=1}^{n}\left\{\sum_{j=1}^{J}a_j\hat{\psi}_{ij}\right\}^2,$$

where

$$\begin{aligned}
\hat{\psi}_{ij} &= D_{ij}(Y_i-\hat{m}_j^h)w_j(\boldsymbol{X}_i;\hat{\boldsymbol{\theta}})+\hat{\boldsymbol{H}}_j^T\hat{\boldsymbol{I}}_{\boldsymbol{\theta}\boldsymbol{\theta}}^{-1}\hat{\boldsymbol{S}}_{\boldsymbol{\theta},i},\\
\hat{\varpi} &= \frac{1}{n}\sum_{i=1}^{n}\Big\{\sum_{k=1}^{J}1/\hat{e}_k(\boldsymbol{X}_i)\Big\}^{-1},\\
\hat{\boldsymbol{H}}_j &= \frac{1}{n}\sum_{i=1}^{n}D_{ij}(Y_i-\hat{m}_j^h)\dot{w}_j(\boldsymbol{X}_i;\hat{\boldsymbol{\theta}}),
\end{aligned}$$

and $\hat{\boldsymbol{S}}_{\boldsymbol{\theta},i}$ is the estimated individual score function (4.9) from the propensity model. For pairwise comparisons, we substitute $\boldsymbol{a}$ with $\boldsymbol{\lambda}_{j,j'}$ to obtain the results in Theorem 1.

For completeness, we next offer three remarks regarding variance estimation.

REMARK **3.** *One could similarly characterize the asymptotic distribution of a collection of estimators specified by different contrast coefficients. Briefly, let the coefficient matrix* $\boldsymbol{A}_{J\times R}=(\boldsymbol{a}_1,\ldots,\boldsymbol{a}_R)$*, where the vector* $\boldsymbol{a}$*'s are distinct from one another. For pairwise comparisons, each vector* $\boldsymbol{a}$ *is a distinct element in the set* $\mathbb{S}$*. Write* $\boldsymbol{\tau}=(\tau^h(\boldsymbol{a}_1),\ldots,\tau^h(\boldsymbol{a}_R))'$*, and* $\hat{\boldsymbol{\tau}}=(\hat{\tau}^h(\boldsymbol{a}_1),\ldots,\hat{\tau}^h(\boldsymbol{a}_R))'$ *as the corresponding weighting estimators. Further denote* $\boldsymbol{\psi}_i=(\psi_{i1},\ldots,\psi_{iJ})'$*, and it can be shown using similar arguments that*

$$\sqrt{n}(\hat{\boldsymbol{\tau}}^h-\boldsymbol{\tau}^h)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varpi^{-1}\boldsymbol{A}^T\boldsymbol{\psi}_i+o_p(1)\xrightarrow{d}\mathcal{N}\left(\boldsymbol{0},\varpi^{-2}\boldsymbol{A}^T\,\mathrm{E}\{\boldsymbol{\psi}_i\boldsymbol{\psi}_i^T\}\boldsymbol{A}\right).$$

The covariance for $\hat{\boldsymbol{\tau}}^h$ can then be estimated by the empirical sandwich estimator

$$\hat{\mathrm{V}}(\hat{\boldsymbol{\tau}}^h) = (n\hat{\varpi})^{-2}\boldsymbol{A}^T\left\{\sum_{i=1}^{n}\hat{\boldsymbol{\psi}}_i\hat{\boldsymbol{\psi}}_i^T\right\}\boldsymbol{A}.$$

REMARK **4.** *Although the above derivation focuses on the generalized overlap weights, a more general presentation for other members of the balancing weights is possible, provided that the balancing weights is a differentiable function in the generalized propensity scores. This differentiability condition rules out the generalized matching weights, which is smooth but not everywhere differentiable and so closed-form variance requires parametric smooth approximation (Li and Greene, 2013) (such approximations could be challenging with multiple treatments since the weight function have infinite-many non-differentiable points). In particular, if we choose the balancing weights as the inverse probability weights, in which case $h(\boldsymbol{X}) = 1$ and the target population is the combined population from all groups, the above derivation can be repeated by substituting the correct forms of $w_j(\boldsymbol{X}_i)$ and $\dot{w}_j(\boldsymbol{X}_i)$. For example, the inverse probability weights are*

$$\begin{aligned} w_1(\boldsymbol{X}_i) &= 1/e_1(\boldsymbol{X}_i) = 1 + \sum_{k=2}^{J}\exp(\alpha_k + \boldsymbol{X}_i^T\boldsymbol{\beta}_k) \\ w_j(\boldsymbol{X}_i) &= 1/e_j(\boldsymbol{X}_i) = \frac{1 + \sum_{k=2}^{J}\exp(\alpha_k + \boldsymbol{X}_i^T\boldsymbol{\beta}_k)}{\exp(\alpha_j + \boldsymbol{X}_i^T\boldsymbol{\beta}_j)}, \qquad j = 2, \ldots, J, \end{aligned}$$

and the derivative of the weights takes the form

$$\frac{\partial}{\partial\boldsymbol{\beta}_l}w_j(\boldsymbol{X}_i) = \boldsymbol{X}_i\frac{\partial}{\partial\alpha_l}w_j(\boldsymbol{X}_i) = \boldsymbol{X}_i\{w_j(\boldsymbol{X}_i)/w_l(\boldsymbol{X}_i) - \delta_{jl}w_l(\boldsymbol{X}_i)\},$$

for $j = 1, \ldots, J$ and $l = 2, \ldots, J$. Of note, theis empirical sandwich variance for $h(\boldsymbol{X}) = 1$ extends the one proposed by Lunceford and Davidian (2004) for binary treatments, and is used to obtain the interval estimates for IPW in the main manuscript.

REMARK **5.** *We have focused on the case with a multinomial logistic propensity score model, but in fact the derivation can be made more general to accommodate other propensity score models that admit a regular and asymptotically linear estimator for the model parameters (Tsiatis, 2006). This condition permits a stochastic expansion for* $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$*, which can then be substituted into (4.10) to obtain the corresponding sandwich variance estimator. In particular, one could replace the multinomial logistic model with a multinomial Probit model, which is another commonly used regression model to accommodate categorical responses.*

5

Conclusions

This dissertation developed and extended weighting methods for case-control and observational studies. In Chapter 2, we developed a new specification test to assess the adequacy of the estimated inverse probability weights in the context of secondary analysis of case-control data. In Chapter 3, we provided a double-robust construction of the difference-in-differences estimator in before-after observational studies; the new estimator is consistent as long as the either the propensity score model or the outcome model is correctly specified and therefore provides two chances to correctly estimate the target parameter. In Chapter 4, we proposed a new weighting scheme, the generalized overlap weights, to emphasize the overlapped population at clinical equipoise for causal inference in observational studies with multiple treatments.

In case-control association studies, there is an increasing interest in the joint analysis of multiple secondary traits (Schifano et al., 2013). In this scenario, the secondary outcome for the ith patient could be written as a M-vector of continuous traits, $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{iM})^T$, and natural motivates trait-specific covariate and SNP effects. Roy et al. (2003) and Schifano et al. (2013) developed the scaled marginal model to describe such effects: $E(Y_{ij}|\boldsymbol{X}_i)/\sigma_j = \boldsymbol{X}_i^T\boldsymbol{\beta}_j$ for $j = 1, \ldots, M$, where σ_j is

the trait-specific dispersion. To account for the correlation among these secondary traits, the prospectively unbiased estimating equations could be the usual Generalized Estimating Equations (GEE), coupled with a working correlation structure (Liang and Zeger, 1986). It is interesting to note that our GLS formulation could be readily extended to estimate the regression coefficients $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_M^T)^T$ in the scaled marginal model, once we estimate the disease probabilities and appropriately formulate the diagonal entries of the weighting matrix in the weighted GEE. Additionally, a corresponding specification test could be developed by following similar arguments in Chapter 2; the specification test can be used to diagnose the adequacy of the weights in this context of multiple secondary traits.

The double-robust DID estimator in Chapter 3 is consistent to the target estimand as long as either one of the outcome model or the propensity score model is correctly specified. However, the proposed double-robust DID estimator belong to the family of the AIPW estimator (Robins et al., 1994), and may suffer from excessive variance when (i) the propensity scores are near 0 or 1, and/or (ii) both models are only mildly misspecified (Kang and Schafer, 2007). To prevent the deleterious consequences of the double-robust DID estimator in Chapter 3, one could extend the idea of Qin and Zhang (2008) and develop an empirical likelihood DID double-robust estimator. The empirical likelihood construction of the estimator is attractive because near-zero values of the propensity scores are prevented by maximizing the product of the empirical weights subject to certain balancing constraints. An additional benefit of adopting the empirical likelihood framework is that one could further develop a multiply robust (MR) DID estimator, adapting the arguments in Han and Wang (2013). The MR estimator allows for multiple outcome models and multiple propensity score models, and is consistent to the target estimand as long as one of propensity score models or outcome models is correctly specified, without distinguishing which one (which explains the name the 'multiply-robustness').

Finally, another promising direction to move forward is to investigate the application of the (generalized) overlap weights in (multi-arm) randomized trials as a flexible covariate-adjustment approach. We will now outline the rationale for overlap weights (Chapter 1) with binary treatments, but the justification for the generalized overlap weights (Chapter 4) in multi-arm randomized trials should follow. In randomized trials with binary treatments, although regression-based approaches such as ANCOVA have been routinely used for estimating the treatment effect with enhanced precision, the inverse probability weighting estimator has also been demonstrated to be a flexible and robust covariate-adjustment method, with advantages detailed in Williamson et al. (2014). Specifically, Shen et al. (2014) has proved that if the (known) propensity score is estimated by logistic regression, the IPW estimator for the treatment effect is asymptotically equivalent to the ANCOVA estimator, and is therefore semiparametrically efficient. In this scenario, because the true propensity scores are known, we could show that the target estimands for IPW and for overlap weighting coincide, and further that the IPW estimator and the overlap weighting estimator are asymptotically equivalent. It would be interesting to examine the finite-sample performance among these estimators, especially in Phase I or II trials with a limited sample size, in which case the overlap weights usually possess better finite-sample efficiency. Additionally, whether similar asymptotic results hold for the application of generalized overlap weights in multi-arm randomized trials represents an avenue for future research.

Bibliography

AASHTO (2010). *Highway Safety Manual*. Washington, D.C., American Association of State Highwayand Transportation Officials (AASHTO).

Abadie, A. (2005). Semiparametric difference-in-differences estimators. *Review of Economic Studies* **72,** 1–19.

Abadie, A. and Imbens, G. W. (2012). A martingale representation for matching estimators. *Journal of the American Statistical Association* **107,** 833–843.

Angrist, J. D. and Pischke, J.-S. (2009). *Mostly Harmless Econometrics: an Empiricists Companion*. Princeton University Press, Princeton, NJ.

Anund, A., Kecklund, G., Vadeby, A., Hjälmdahl, M., and Åkerstedt, T. (2008). The alerting effect of hitting a rumble strip-A simulator study with sleepy drivers. *Accident Analysis and Prevention* **40,** 1970–1976.

Ashenfelter, O. (1978). Estimating the effect of training programs on earnings. *The Review of Economics and Statistics* **60,** 47–57.

Ashenfelter, O. and Card, D. (1985). Using the longitudinal structure of earnings to estimate the effect of training programs. *The Review of Economics and Statistics* **67,** 648–660.

Austin, P. C. and Stuart, E. A. (2015). Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. *Statistics in medicine* **34,** 3661–3679.

Bang, H. and Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics* **61,** 962–972.

Bouillon, K., Bertrand, M., Bader, G., Lucot, J. P., Dray-Spira, R., and Zureik, M. (2018). Association of Hysteroscopic vs Laparoscopic Sterilization With Procedural, Gynecological, and Medical Outcomes. *Journal of the American Medical Association* **319,** 375–387.

Branas, C. C., Cheney, R. A., MacDonald, J. M., Tam, V. W., Jackson, T. D., and Ten Havey, T. R. (2011). A difference-in-differences analysis of health, safety, and greening vacant urban space. *American Journal of Epidemiology* **174,** 1296–1306.

Breiman, L. (2001). Random forests. *Machine Learning* **45,** 5–32.

Brown, H. K., Ray, J. G., Wilton, A. S., Lunsky, Y., Gomes, T., and Vigod, S. N. (2017). Association between serotonergic antidepressant use during pregnancy and autism spectrum disorder in children. *Journal of the American Medical Association* **317,** 1544–1552.

Callaway, B. and Sant'Anna, P. (2018). Difference-in-Differences With Multiple Time Periods and an Application on the Minimum Wage and Employment. *arXiv:1803.09015v2* pages 1–47.

Card, D. and Krueger, A. B. (1994). Minimum wages and employment: A case study of the fast-food industry in New Jersey and Pennsylvania. *American Economic Review* **82,** 772–793.

Carroll, R. J., Wang, S., and Wang, C. Y. (1995). Prospective analysis of logistic case-control studies. *Journal of the American Statistical Association* **90,** 157–169.

Cattaneo, M. D. (2010). Efficient semiparametric estimation of multi-valued treatment effects under ignorability. *Journal of Econometrics* **155,** 138–154.

Chamberlain, G. (1987). Asymptotic efficiency in estimation with conditional moment restrictions. *Journal of Econometrics* **34,** 305–334.

Chen, H. Y., Kittles, R., and Zhang, W. (2013). Bias correction to secondary trait analysis with case-control design. *Statistics in medicine* **32,** 1494–508.

Cole, S. R. and Hernán, M. A. (2008). Constructing inverse probability weights for marginal structural models. *American Journal of Epidemiology* **168,** 656–664.

Cook, B. L., Mcguire, T. G., Lock, K., and Zaslavsky, A. M. (2010). Comparing methods of racial and ethnic disparities measurement across different settings of mental health care. *Health Services Research* **45,** 825–847.

Cook, B. L., Mcguire, T. G., Meara, E., and Zaslavsky, A. M. (2009). Adjusting for health status in non-linear models of health care disparities. *Health Services and Outcomes Research Methodology* **9,** 1–21.

Crump, R. K., Hotz, V. J., Imbens, G. W., and Mitnik, O. A. (2009). Dealing with limited overlap in estimation of average treatment effects. *Biometrika* **96,** 187–199.

D'Agostino Jr, R. B. (1998). Tutorial in Biostatistics: Propensity score methods for bias reduction in the comparison of a treatment to a non randomized control group. *Statistics in Medicine* **17,** 2265–2281.

Dimick, J. B. and Ryan, A. M. (2014). Methods for evaluating changes in health care policy: the difference-in-differences approach. *JAMA : the journal of the American Medical Association* **312,** 2401–2402.

Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. Chapman & Hall, New York, NY.

El-Basyouny, K. and Sayed, T. (2012). Linear and nonlinear safety intervention models: Novel methods applied to evaluation of shoulder rumble strips. *Transportation Research Record* **2280,** 28–37.

Elvik, R. (2002). The importance of confounding in observational before-and-after studies of road safety measures. *Accident Analysis and Prevention* **34,** 631–635.

Elvik, R. (2008). The predictive validity of empirical Bayes estimates of road safety. *Accident Analysis and Prevention* **40,** 1964–1969.

Federal Highway Administration (2014). Roadway Departure Countermeasures. Available `https://safety.fhwa.dot.gov/roadway_dept/rdctrm.cfm`.

Feng, P., Zhou, X. H., Zou, Q. M., Fan, M. Y., and Li, X. S. (2012). Generalized propensity score for estimating the average treatment effect of multiple treatments. *Statistics in Medicine* **31,** 681–697.

Gårder, P. and Davies, M. (2006). Safety effect of continuous shoulder rumble strips on rural interstates in Maine. *Transportation Research Record* **1953,** 156–162.

Golub, G. H. and Van Loan, C. F. (1996). *Matrix Computations, Third Edition*. The John Hopkins University Press, Baltimore, MD.

Grabich, S. C., Robinson, W. R., Engel, S. M., Konrad, C. E., Richardson, D. B., and Horney, J. A. (2015). County-level hurricane exposure and birth rates: application of difference-in-differences analysis for confounding control. *Emerging Themes in Epidemiology* **12,** 19.

Griffith, M. (1999). Safety evaluation of rolled-in continuous shoulder rumble strips installed on freeways. *Transportation Research Record* **1665,** 28–34.

Hall, A. (2004). *Generalized Method of Moments*. Oxford University Press Inc., New York.

Han, P. and Wang, L. (2013). Estimation with missing data: beyond double robustness. *Biometrika* **100,** 417–430.

Hansen, L. P. (1982). Large sample properties of generalised method of moments estimators. *Econometrica* **50,** 1029–1054.

Hauer, E. (1997). *Observational Before-After Studies in Road Safety: Estimating the Effect of Highway and Traffic Engineering Measures on Road Safety*. Emerald Group Publishing Limited, Oxford, OX, U.K., Pergamon.

Heckman, J. J. (1998). Characterizing Selection Bias Using Experimental Data. *Econometrica* **66,** 1017–1098.

Heckman, J. J., Ichimura, H., and Todd, P. E. (1997). Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *The Review of Economic Studies* **64,** 605–654.

Hernán, M. and Robins, J. (2019). *Causal Inference*. Chapman & Hall/CRC, Boca Raton, FL.

Hiatt, W. R. (2001). Medical treatment of peripheral arterial disease and claudication. *New England Journal of Medicine* **344,** 1608–1621.

Hirano, K. and Imbens, G. W. (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., Imbiens, G. W., and Geert, R. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica* **71,** 1161–1189.

Horvitz, D. G. and Thompson, D. (1952). A Generalization of Sampling Without Replacement From a Finite Universe. *Journal of the American Statistical Association* **44,** 663–685.

Hudgens, M. G. and Halloran, M. E. (2008). Toward causal inference with interference. *Journal of the American Statistical Association* **103,** 832–842.

Imbens, G. W. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika* **87,** 706–710.

Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics* **86,** 4–29.

Imbens, G. W. and Rubin, D. B. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences*. Cambridge University Press, New York, NY.

Imbens, G. W. and Wooldridge, J. M. (2009). Recent Developments in the Econometrics of Program Evaluation. *Journal of Economic Literature* **47,** 5–86.

Ix, J. H., Biggs, M. L., Kizer, J. R., Mukamal, K. J., Djousse, L., Zieman, S. J., De Boer, I. H., Nelson, T. L., Newman, A. B., Criqui, M. H., and Siscovick, D. S. (2011). Association of body mass index with peripheral arterial disease in older adults. *American Journal of Epidemiology* **174,** 1036–1043.

Jiang, Y., Scott, A. J., and Wild, C. J. (2006). Secondary analysis of case-control data. *Statistics in Medicine* **25,** 1323–1339.

Joffe, M. M., Ten Have, T. R., Feldman, H. I., and Kimmel, S. E. (2004). Model selection, confounder control, and marginal structural models: review and new applications. *The American Statistician* **58,** 272.

Jones, P. M., Cherry, R. A., Allen, B. N., Bray Jenkyn, K. M., Shariff, S. Z., Flier, S., Vogt, K. N., and Wijeysundera, D. N. (2018). Association between handover of anesthesia care and adverse postoperative outcomes among patients undergoing major surgery. *Journal of the American Medical Association* **319,** 143–153.

Kang, J. D. Y. and Schafer, J. L. (2007). Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science* **22,** 523–539.

Karwa, V., Slavkovic, A. B., and Donnell, E. T. (2011). Causal inference in transportation safety studies: Comparison of potential outcomes and diagrams. *The Annals of Applied Statistics* **5,** 1428–1455.

Khan, M., Abdel-Rahim, A., and Williams, C. J. (2015). Potential crash reduction benefits of shoulder rumble strips in two-lane rural highways. *Accident Analysis and Prevention* **75,** 35–42.

Lechner, M. (2001). Identification and estimation of causal effects of multiple treatments under the conditional independence assumption. In Lechner, M. and Pfeiffer, F., editors, *Econometric Evaluations of Active Labor Market Policies in Europe*, pages 43–58. Heidelberg: Physica.

Lechner, M. (2002). Program heterogeneity and propensity score matching: An application to the evaluation of active labor market policies. *Review of Economics and Statistics* **84,** 205–220.

Lechner, M. (2011). The estimation of causal effects by difference-in-difference methods. *Foundations and Trends in Econometrics* **4,** 165–224.

Lee, B. K., Lessler, J., and Stuart, E. A. (2009). Improving propensity score weighting using machine learning. *Statistics in Medicine* pages 337–346.

Lee, B. K., Lessler, J., and Stuart, E. A. (2011). Weight trimming and propensity score weighting. *PLOS ONE* **6,** 1–6.

Lehmann, E. (1983). *Theory of Point Estimation*. Spinger, New York.

Li, F., Morgan, K. L., and Zaslavsky, A. M. (2018). Balancing covariates via propensity score weighting. *Journal of the American Statistical Association* **113,** 390–400.

Li, F., Thomas, L. E., and Li, F. (2018). Addressing extreme propensity scores via the overlap weights. *Forthcoming at American Journal of Epidemiology* .

Li, F., Zaslavsky, A. M., and Landrum, M. B. (2013). Propensity score weighting with multilevel data. *Statistics in Medicine* **32,** 3373–3387.

Li, H. and Gail, M. H. (2012). Efficient adaptively weighted analysis of secondary phenotypes in case-control genome-wide association studies. *Human heredity* **73,** 159–73.

Li, L. and Greene, T. (2013). A weighting analogue to pair matching in propensity score analysis. *International Journal of Biostatistics* **9,** 1–20.

Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73,** 13–22.

Liaw, A. and Wiener, M. (2002). Classification and regression by randomForest. *R News* **2,** 18–22.

Lin, D. and Zeng, D. (2009). Proper analysis of secondary phenotype data in case-control association studies. *Genetic epidemiology* **33,** 256–265.

Lopez, M. J. and Gutman, R. (2017a). Estimating the average treatment effects of nutritional label use using subclassification with regression adjustment. *Statistical Methods in Medical Research* **26,** 839–864.

Lopez, M. J. and Gutman, R. (2017b). Estimation of causal effects with multiple treatments: A review and new ideas. *Statistical Science* **32,** 432–454.

Lunceford, J. K. and Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in Medicine* **23,** 2937–2960.

Lyon, C., Persaud, B., and Eccles, K. (2015). Safety evaluation of centerline plus shoulder rumble strips. Technical Report Report No. FHWA-HRT-15-048, Federal Highway Administration, McLean, VA.

Ma, Y. and Carroll, R. J. (2016). Semiparametric estimation in the secondary analysis of case-control studies. *Journal of Royal Statistical Society, Series B: Statistical Methodology* **78,** 127–151.

McCaffrey, D. F., Griffin, B. A., Almirall, D., Slaughter, M. E., Ramchand, R., and Burgette, L. F. (2013). A tutorial on propensity score estimation for multiple treatments using generalized boosted models. *Statistics in Medicine* **32,** 3388–3414.

McCaffrey, D. F., Ridgeway, G., , and Morral, A. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological Methods* pages 403–425.

Mercatanti, A. and Li, F. (2014). Do bebit cards increase household spending? evidence from a semiparametric causal analysis of a survey. *Annals of Applied Statistics* **8,** 2405–2508.

Monsees, G., Tamimi, R., and Kraft, P. (2009). Genome-wide association scans for secondary traits using case-control samples. *Genetic epidemiology* **33,** 717–728.

Penman, A. D. and Johnson, W. D. (2006). The changing shape of the body mass index distribution curve in the population: implications for public health policy to reduce the prevalence of adult obesity. *Preventing chronic disease* **3,** 1–4.

Persaud, B. and Lyon, C. (2007). Empirical Bayes before-after safety studies: Lessons learned from two decades of experience and future directions. *Accident Analysis and Prevention* **39,** 546–555.

Persaud, B. N., Retting, R. A., and Lyon, C. A. (2004). Crash reduction following installation of centerline rumble strips on rural two-lane roads. *Accident Analysis and Prevention* **36,** 1073–1079.

Prentice, R. L. and Pyke, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika* **66,** 403–411.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics* **38,** 904–909.

Qin, J. and Zhang, B. (2008). Empirical-likelihood-based difference-in-differences estimators. *Journal of the Royal Statistical Society. Series B: Statistical Methodology* **70,** 329–349.

Rassen, J. A., Shelat, A. A., Franklin, J. M., Glynn, R. J., Solomon, D. H., and Schneeweiss, S. (2013). Matching by propensity score in cohort studies with three treatment groups. *Epidemiology* **24,** 401–409.

Richardson, D. B., Rzehak, P., Klenk, J., and Weiland, S. K. (2007). Analyses of case-control data for additional outcomes. *Epidemiology* **18,** 441–445.

Ridgeway, G., McCaffrey, D., Morral, A., Burgette, L. F., and Griffin, B. A. (2006). Toolkit for weighting and analysis of nonequivalent groups: A tutorial for the twang. Technical report, RAND Corporation, Santa Monica, CA.

Robins, J. M., Hernán, M. Á., and Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology* **11,** 550–560.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association* **89,** 846–866.

Rosenbaum, P. (2002). *Observational Studies*. Springer Series in Statistics. Springer, New York, NY.

Rosenbaum, P. R. (1987). Model-based direct adjustment. *Journal of the American Statistical Association* **82,** 387–394.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70,** 41–55.

Roy, J., Lin, X., and Ryan, L. M. (2003). Scaled marginal models for multiple continuous outcomes. *Biostatistics* **4,** 371–383.

Rubin, D. B. (1974). Estimating causal effects of treatment in randomized and nonrandomized studies. *Journal of Educational Psychology* **66,** 688–701.

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics* **6,** 34–58.

Rubin, D. B. (1979). Using Multivariate Matched Sampling and Regression Adjustment to Control Bias in Observational Studies. *Journal of the American Statistical Association* **74,** 318–328.

Sato, T. and Matsuyama, Y. (2003). Marginal Structural Models as a Tool for Standardization. *Epidemiology* **14,** 680–686.

Schifano, E. D., Li, L., Christiani, D. C., and Lin, X. (2013). Genome-wide association analysis for multiple continuous secondary phenotypes. *The American Journal of Human Genetics* **92,** 744–759.

Scott, A. J. and Wild, C. J. (1986). Fitting logistic models under case-control or choice based sampling. *Journal of the Royal Statistical Society, Series B: Statistical Methodology* **48,** 170–182.

Scott, A. J. and Wild, C. J. (2002). On the robustness of weighted methods for fitting models to case-control data. *Journal of the Royal Statistical Society, Series B: Statistical Methodology* **64,** 207–219.

Shao, J. and Tu, D. (2012). *The Jackknife and Bootstrap*. Springer, New York, NY.

Shen, C., Li, X., and Li, L. (2014). Inverse probability weighting for covariate adjustment in randomized studies. *Statistics in Medicine* **33,** 555–568.

Sommers, B. D., Long, S. K., and Baicker, K. (2014). Changes in mortality after Massachusetts health care reform : A quasi-experimental study. *Annals of Internal Medicine* **160,** 585–593.

Stefanski, L. A. and Boos, D. D. (2002). The calculus of M-estimation. *American Statistician* **56,** 29–38.

Stuart, E. A. (2010). Matching Methods for Causal Inference: A Review and a Look Forward. *Statistical Science* **25,** 1–21.

Stuart, E. A., Huskamp, H. A., Duckworth, K., Simmons, J., Song, Z., Chernew, M., and Barry, C. L. (2014). Using propensity scores in difference-in-differences models to estimate the effects of a policy change. *Health Service Outcomes Research Methodology* **14,** 166–182.

Stürmer, T., Rothman, K. J., Avorn, J., and Glynn, R. J. (2010). Treatment effects in the presence of unmeasured confounding: Dealing with observations in the tails of the propensity score distribution-A simulation study. *American Journal of Epidemiology* **172,** 843–854.

Tchetgen Tchetgen, E. J. (2014). A general regression framework for a secondary outcome in case-control studies. *Biostatistics* **15,** 117–128.

Thorgeirsson, T. E., Geller, F., Sulem, P., Rafnar, T., Wiste, A., Magnusson, K. P., Manolescu, A., Thorleifsson, G., Stefansson, H., Ingason, A., et al. (2008). A variant associated with nicotine dependence, lung cancer and peripheral arterial disease. *Nature* **3,** 638–642.

Tsiatis, A. A. (2006). *Semiparametric Theory and Missing Data*. Springer, New York.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Wang, J. and Shete, S. (2012). Analysis of secondary phenotype involving the interactive effect of the secondary phenotype and genetic variants on the primary disease. *Annals of Human Genetics* **76,** 484–499.

Weedon, M. N., Lettre, G., Freathy, R. M., Lindgren, C. M., Voight, B. F., Perry, J. R. B., Elliott, K. S., Hackett, R., Guiducci, C., Shields, B., et al. (2007). A common variant of HMGA2 is associated with adult and childhood height in the general population. *Nature Genetics* **39,** 1245–1250.

Wei, J., Carroll, R. J., and Müller, U. U. (2013). Robust estimation for homoscedastic regression in the secondary analysis of case-control data. *Journal of Royal Statistical Society, Series B: Statistical Methodology* **75,** 185–206.

Wei, Y., Song, X., Liu, M., Ionita-Laza, I., and Reibman, J. (2016). Quantile regression in the secondary analysis of case-control data. *Journal of the American Statistical Association* **111,** 344–354.

Weinberg, C. R. and Wacholder, S. (1993). Prospective analysis of case-control data under general multiplicative-intercept risk models. *Biometrika* **80,** 461–465.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica* **50,** 1–25.

Willer, C. J., Speliotes, E. K., Loos, R. J. F., Li, S., Lindgren, C. M., Heid, I. M., Berndt, S. I., Elliott, A. L., Jackson, A. U., Lamina, C., and et al. (2009). Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. *Nature Genetics* **41,** 25–34.

Williamson, E. J., Forbes, A., and White, I. R. (2014). Variance reduction in randomised trials by inverse probability weighting using the propensity score. *Statistics in Medicine* **33,** 721–737.

Wood, J. and Donnell, E. T. (2016). Safety evaluation of continuous green T intersections: A propensity scores-genetic matching-potential outcomes approach. *Accident Analysis and Prevention* **93,** 1–13.

Wood, J. S. and Donnell, E. T. (2017). Causal inference framework for generalizable safety effect estimates. *Accident Analysis and Prevention* **104,** 74–87.

Wood, J. S., Donnell, E. T., and Porter, R. J. (2015). Comparison of safety effect estimates obtained from empirical Bayes before-after study, propensity scores-potential outcomes framework, and regression model with cross-sectional data. *Accident Analysis and Prevention* **75,** 144–154.

Xing, C., Mccarthy, J. M., Cupples, L. A., Meigs, J. B., Lin, X., and Allen, A. S. (2016). Robust analysis of secondary phenotypes in case-control genetic association studies. *Statistics in Medicine* **35,** 4226–4237.

Yang, S. and Ding, P. (2018). Asymptotic inference of causal effects with observational studies trimmed by the estimated propensity scores. *Biometrika* **105,** 487–493.

Yang, S., Imbens, G. W., Cui, Z., Faries, D. E., and Kadziola, Z. (2016). Propensity score matching and subclassification in observational studies with multi-level treatments. *Biometrics* **72,** 1055–1065.

Yoshida, K., Hernández-Díaz, S., Solomon, D. H., Jackson, J. W., Gagne, J. J., Glynn, R. J., and Franklin, J. M. (2017). Matching weights to simultaneously compare three treatment groups comparison to three-way matching. *Epidemiology* **28,** 387–395.

Zanutto, E., Lu, B., and Hornik, R. (2005). Using propensity score subclassification for multiple treatment doses to evaluate a national antidrug media campaign. *Journal of Educational and Behavioral Statistics* **30,** 59–73.

Zaslavsky, A. M. and Ayanian, J. Z. (2005). Integrating research on racial and ethnic disparities in health care over place and time. *Medical Care* **43,** 303–307.

Biography

Fan (Frank) Li received his Master's in Biostatistics from Duke University School of Medicine. He is expecting his Ph.D. in Biostatistics from Duke University in May 2019. During his study at Duke University, he authored and co-authored 15 methodological and collaborative publications, which appeared in *Biometrics*, *Statistics in Medicine*, *American Journal of Epidemiology*, *Statistics and Probability Letters*, among others. He has also won a number of national student paper awards during his study at Duke, including the 2019 Student Travel Award from the Biometrics Section of American Statistical Association (ASA), 2018 NSF Student Travel Award from the Atlantic Causal Inference Conference (ACIC), 2018 Distinguished Student Paper Award from the Eastern Northern American Region (ENAR), International Biometric Society, and 2015 Student Travel Award from the 11th International Conference on Health Policy Statistics (ICHPS). In October 2018, he received the Chancellor's Award for Research Excellence, a prestigious award for his exceptional academic record in Duke University School of Medicine.