Essays on Propensity Score Methods for Causal Inference in Observational Studies

by

Nghi Le Phuong Nguyen

Department of Statistical Science
Duke University

Date: _____________________

Approved:

__________________________
Fan Li, Supervisor

__________________________
Surya Tokdar

__________________________
Jerry Reiter

__________________________
V. Joseph Hotz

Dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Statistical Science in the Graduate School of Duke University 2018
Abstract

Essays on Propensity Score Methods for Causal Inference in Observational Studies

by

Nghi Le Phuong Nguyen

Department of Statistical Science
Duke University

Date: ______________________

Approved:

__________________________
Fan Li, Supervisor

__________________________
Surya Tokdar

__________________________
Jerry Reiter

__________________________
V. Joseph Hotz

An abstract of a dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Statistical Science in the Graduate School of Duke University 2018
Abstract

In this dissertation, I present three essays from three different research projects and they involve different usages of propensity scores in drawing causal inferences in observational studies.

Chapter 1 talks about the general idea of causal inference as well as the concept of randomized experiments and observational studies. It introduces the three different projects and their contributions to the literature.

Chapter 2 gives a critical review and an extensive discussion of several commonly-used propensity score methods when the data have a multilevel structure, including matching, weighting, stratification, and methods that combine these with regression. The usage of these methods is illustrated using a data set about endoscopic vein-graft harvesting in coronary artery bypass graft (CABG) surgeries. We discuss important aspects of the implementation of these methods such as model specification and standard error calculations. Based on the comparison, we provide general guidelines for using propensity score methods with multilevel data in practice. We also provide the relevant code in the form of an R package, available on GitHub.

In observational studies, subjects are no longer assigned to treatment at random as in randomized experiments, and thus the association between the treatment and outcome can be due to some unmeasured variable that affects both the treatment and the outcome. Chapter 3 focuses on conducting sensitivity analysis to assess the robustness of the estimated quantity when the unconfoundedness assumption is
violated. Two approaches to sensitivity analysis are presented, both are extensions from previous works to accommodate for a count outcome. One method is based on the subclassification estimator and it relies on maximum likelihood estimation. The second method is more flexible on the estimation method and is based on simulations. We illustrate both methods using a data set from a traffic safety research study about the safety effectiveness (measured in crash counts reduction) of the combined application of center line rumble strips and shoulder rumble strips on two-lane rural roads in Pennsylvania.

Chapter 4 proposes a method for estimating heterogeneous causal effects in observational studies by augmenting additive-interactive Gaussian process regression using the propensity scores, yielding a flexible yet robust way to predict the potential outcome surface from which the conditional treatment effects can be calculated. We show that our method works well even in presence of strong confounding and illustrate this by comparing with commonly-used methods in different settings using simulated data.

Finally, chapter 5 concludes this dissertation and discusses possible future works for each of the projects.
To my dearest family.
Contents

Abstract iv
List of Tables x
List of Figures xii
List of Abbreviations and Symbols xiii
Acknowledgements xv

1 Introduction 1
  1.1 Propensity score methods for multilevel data . . . . . . . . . . . . . . 2
  1.2 Sensitivity analysis in an observational study with count outcome . . 3
  1.3 Propensity score augmented additive-interactive Gaussian process re-
      gression . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 4

2 Propensity Score Methods for Multilevel Data 6
  2.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 6
  2.2 Motivation example . . . . . . . . . . . . . . . . . . . . . . . . . . . . 9
  2.3 Basics of propensity score methods . . . . . . . . . . . . . . . . . . . 12
  2.4 Stage 1: Estimating propensity score with multilevel data . . . . . . 17
  2.5 Stage 2: Estimating treatment effects with multilevel data . . . . . 19
    2.5.1 Matching . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 19
    2.5.2 Mixed matching-regression . . . . . . . . . . . . . . . . . . . . . 21
    2.5.3 Weighting . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 23
3 Sensitivity Analysis in an Observational Study with Count Outcome

3.1 Introduction .......................................................... 38

3.2 Methods ................................................................. 42
   3.2.1 Notations and settings ......................................... 42
   3.2.2 Sensitivity Analysis 1: MLE-based ......................... 45
   3.2.3 Sensitivity analysis 2: simulation-based .................. 50

3.3 Application to Pennsylvania rumble strips treatment on rural highways 54
   3.3.1 Data ............................................................... 54
   3.3.2 Sensitivity analysis 1: MLE-based ......................... 57
   3.3.3 Sensitivity analysis 2: simulation-based .................. 60

3.4 Conclusion and future work ........................................ 63

4 Propensity Score Augmented Additive-Interactive Gaussian Process Regression

4.1 Introduction .......................................................... 72

4.2 Proposed method ...................................................... 76

4.3 Simulation experiments .............................................. 80
   4.3.1 Homogeneous treatment effect ............................ 80
   4.3.2 Heterogeneous treatment effect ......................... 84
4.3.3 Individual treatment effects ........................................ 91

5 Conclusion ........................................................................ 96

5.1 Propensity score methods for multilevel data ................... 96
5.2 Sensitivity analysis in an observational study with count outcome ...... 97
5.3 Propensity score augmented additive-interactive Gaussian process regression ........................................ 98

A Sensitivity analysis .............................................................. 99

A.1 Sensitivity parameters assumption ...................................... 99
A.2 Simulation-based: killer confounder characteristics ............... 100

Bibliography ....................................................................... 102

Biography ......................................................................... 108
# List of Tables

2.1 Covariates included in both the propensity score and outcome models. 11  
2.2 Summary of the estimated propensity score by treatment group. 12  
2.3 Distribution of the cluster size for each year. 30  
2.4 Estimators for ATE and their SE ($\times 10^{-2}$) in the endoscopic data, by year and overall. 33  
3.3 CMF baseline estimates and their 95% confidence intervals for total crash count. 57  
3.5 Sensitivity analysis of CMF estimates to the effect of a binary confounding factor calibrated to mimic the distribution of some observed covariates. 62  
3.1 Mean and standard deviation for several characteristic covariates for treatment and reference sites. 67  
3.2 Covariates included in the propensity score and the outcome models. 68  
3.6 Proportion of sites by the transformed outcome $K$ and treatment status $Z$. 68  
3.7 “False negative” threats: Sensitivity parameters $p_{ij}$ for each combination of $d = p_{00} - p_{01} = p_{01} - p_{02}$ and $s = p_{1} - p_{0}$. 69  
3.8 “False negative” scenario: Sensitivity analysis of CMF estimates to the effect of a binary confounding factor, varying its effect on the outcome and the treatment selection. 70  
3.9 “False positive” scenario: Sensitivity analysis of CMF estimates to the effect of a binary confounding factor, varying its effect on the outcome and the treatment selection. 71
4.1 Homogeneous treatment effects: Models run with (top) estimated and (bottom) true propensity scores. . . . . . . . . . . . . . . . . . . . . . 82

4.2 Heterogeneous treatment effects: Models run with (top) true and (bottom) estimated propensity scores. . . . . . . . . . . . . . . . 90
## List of Figures

2.1 Histogram of endoscopic proportion by center ........................................ 13
2.2 Histogram of the estimated propensity scores by each technique. . . . 14
2.3 Standardized differences for observed covariates at baseline as well as after matching and weighting ................................................................. 32
2.4 STS Data: Estimators and error bars of different ATE estimators, for each year and the combined data. ................................................................. 35
3.1 Estimated propensity scores 2012-only data. ........................................ 56
4.1 Homogeneous treatment effect: True propensity scores for simulated data with \( n = 200 \) and \( p = 100 \). .................................................. 81
4.2 Homogeneous treatment effects: Average mean, 90\% CI and coverage of the ATT estimators for simulated data ............................................. 83
4.3 Heterogeneous treatment effects: Density plot of true propensity scores for each treatment group ................................................................. 85
4.4 Heterogeneous treatment effects: potential outcomes and individual treatment effects for our simulated data with \( n = 300 \) and \( p = 100 \). . 86
4.5 Heterogeneous treatment effects: models run with true propensity scores: Average mean, 90\% CI and coverage of the ATT estimators for simulated data with \( n = 200 \) and \( p = 100 \) ......................... 88
4.6 Heterogeneous treatment effects: models run with estimated propensity scores: Average mean, 90\% CI and coverage of the ATT estimators for simulated data with \( n = 200 \) and \( p = 100 \) ......................... 89
4.7 Bias, RMSE and coverage for each propensity score decile of several methods for the estimated individual treatment effects of 50 simulated data sets with \( n = 300 \) and \( p = 100 \). ................................. 93
List of Abbreviations and Symbols

Symbols

These symbols are used throughout this dissertation

\[ \mathbb{E}(X) \] Expectation of a random variable \( X \).

\[ GP(\mu, K) \] A Gaussian process with mean \( \mu \) and covariance function \( K \).

\[ C^{SE}(x, \tilde{x}|\gamma, \lambda) \] A squared exponential covariance function, \( = exp[-\lambda^2\|x_\gamma - \tilde{x}_\gamma\|^2] \)

Abbreviations

airGP Additive-interactive regression with Gaussian processes

AADT Average annual daily traffic

ACSD Adult cardiac surgery database

ATE Average treatment effect

ATT Average treatment effect for the treated

BART Bayesian Additive Regression Tree

CABG Coronary artery bypass graft

CI Confidence interval

CMF Crash modification factor

CLRS Centerline rumble strips

DR Doubly robust

EB Empirical Bayes
<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>GLM</td>
<td>Generalized linear model</td>
</tr>
<tr>
<td>GLMM</td>
<td>Generalized linear mixed model</td>
</tr>
<tr>
<td>HT</td>
<td>Horvitz-Thompson</td>
</tr>
<tr>
<td>ITE</td>
<td>Individual treatment effect</td>
</tr>
<tr>
<td>LASSO</td>
<td>Least absolute shrinkage and selection operator</td>
</tr>
<tr>
<td>MCMC</td>
<td>Markov chain Monte Carlo</td>
</tr>
<tr>
<td>MITSS</td>
<td>Multiple imputation with two splines in subclasses</td>
</tr>
<tr>
<td>MLE</td>
<td>Maximum-likelihood estimation</td>
</tr>
<tr>
<td>PennDOT</td>
<td>Pennsylvania Department of Transportation</td>
</tr>
<tr>
<td>PS</td>
<td>Propensity score</td>
</tr>
<tr>
<td>$AGP_{ps}$</td>
<td>Propensity score augmented additive interactive Gaussian process regression</td>
</tr>
<tr>
<td>RCM</td>
<td>Rubin causal model</td>
</tr>
<tr>
<td>RF</td>
<td>Random forest</td>
</tr>
<tr>
<td>RMSE</td>
<td>Root-mean-square error</td>
</tr>
<tr>
<td>SDf</td>
<td>Standardized difference</td>
</tr>
<tr>
<td>SRS</td>
<td>Shoulder rumble strips</td>
</tr>
<tr>
<td>SSOD</td>
<td>Side-swiped opposite direction crashes</td>
</tr>
<tr>
<td>STS</td>
<td>Society of thoracic surgeons</td>
</tr>
<tr>
<td>SUTVA</td>
<td>Stable unit treatment value assumption</td>
</tr>
</tbody>
</table>
Acknowledgements

First and foremost, I would like to thank my thesis advisor, Fan Li, for all her guidance, patience, and support. During the past five years, she guided me towards my current research area - causal inference, gave me lots of insightful conversations, and helped push me towards completing this dissertation. I would also like to thank her for introducing and supporting me financially to attend several conferences as well as internship opportunities. Additionally, I would like to thank the other members on my preliminary exam and thesis committee: Surya Tokdar, Jerry Reiter, and Joe Hotz for their time and valuable feedback.

I’ve been very fortunate to have a chance to collaborate with several people. I’m very grateful for Surya’s time spent to show me how to use his R-package efficiently. I thank Sean O’Brien at the Duke Clinical Research Institute for all his help with the data set during our collaboration. I would also like to thank Kimberly Eccles and Scott Himes and Vanasse Hangen Brustlin, Inc. for their valuable mentorship during my internship there. This dissertation was supported collectively by the grant NSF-SES 1424688 and the Duke Clinical Research Institute.

I’m thankful for the friendship of my classmates. The past five years wouldn’t have been the same without our study groups and parties. You brightened and lightened up my time in the program, and I’m very glad we met. I’m also grateful to the Vietnamese students I’ve met at Duke for the gatherings and delicious potlucks that helped with my homesickness. Thank you to other loved ones near and far:
Eric, Thanh, Tu, Stanley, An McMaru, Mai, and the Orndorff family, for the cheers, hugs, and laughs we shared.

Last but not least, I would like to express my deepest gratitude to my family for their unconditional love and support. My parents, Sang and Lan, have always been my pillars of strength. My dad was the one who instilled and nourished in me a passion for science and exploration. My mom, despite having doubts at first about my pursuit of a Ph.D., never stopped cheering me on and letting me know she would always have my back. Thank you to the best sisters one can hope for, Thao, Vy, and Tran, for the gossips, the care packages, and the much-needed random conversations.
Introduction

Causal inference has become a central goal for several research studies in education, sociology, public policy, and medical science, aiming to measure the impact of an intervention or exposure on an outcome of interest. In fact, several research questions are causal-based. Do listening to music while being pregnant make babies smarter? What is the effect of forced busing for diversity on math scores in Wake County, NC? What is the relationship between obesity and cancer? Does longer maternity leave lead to better skills for the child? Drawing causal inference is always tricky because just as the philosopher David Lewis put it, causation is “[...]something that makes a difference, and the difference it makes must be a difference from what would have happened without it.” Essentially, we are chasing after the answer to the question “what would have happened had the subjects had a different treatment?”.

It has long been agreed that randomized controlled trials (RCT) are ideal and have been considered the gold standard. In RCT, the treatment assignment is randomly given to the sample, which ensures that the distributions of background characteristics for the two treatment groups are as similar as possible. Therefore, the treatment effect can be estimated by comparing the outcomes between the two
groups. However due to practical or ethical reasons, RCT are not always feasible and instead we need to work with observational studies, where the treatment assignment is no longer random, and confounding is usually an issue. Nonetheless, there have been a lot of work done to develop methods to improve our ability to draw strong causal inferences from observational studies. Historically, multivariate regression is among the most straightforward and commonly used approach, building regression models for the outcome which include the treatment status and relevant background characteristics. Along the way, other methods have been developed, including matching, stratification, instrumental variables, propensity score methods, and machine-learning methods.

Each of the three next chapters in this dissertation makes an independent contribution to causal inference in observational studies, coming from different angles and filling specific gaps in the literature. All three essays involve the use of propensity scores - the conditional probability of being assigned the treatment, given the background covariates.

1.1 Propensity score methods for multilevel data

The first essay concerns causal inference in multilevel or clustered data setting, which has been increasingly common in practice. In these settings, we are interested in evaluating the treatment effect of an intervention on some outcome of interest, using observations that are clustered by design - for example, patients in hospitals or of the same physician, students in the same schools or counties, policy holders of the same health insurance plans. The way units are clustered can be relevant to the analysis and should not be dismissed. Suppose we are interested in the effect of a new teaching technique on students’ scores, and the collected data included students from several schools. There are several student characteristics that might be associated with how students are assigned to school: household average income, parents’ education,
how their parents value education, etc. These factors may or may not be observed, and they can create school-level variation in math scores. There have been several developments and proposals on how to apply propensity score methods to draw causal inference under multilevel settings, including but not limited to matching, weighting, stratification, mixed matching-regression, and doubly-robust estimators.

This chapter provide a critical review, an extensive discussion and general guidelines of these methods using a real data set about endoscopic vein-graft harvest in coronary artery bypass graft surgeries as an illustration. An R package is also provided and is available on GitHub. This package implements a variety of algorithms for calculating propensity score weighting estimators and their standard errors for treatment effects in a clustered data setting.

1.2 Sensitivity analysis in an observational study with count outcome

Upon working with the potential outcome framework, one of the assumptions that are typically made is the strong ignorability assumption. The first part of this assumption is unconfoundedness: we can remove the dependence between the treatment assignment and the potential outcomes after accounting for all factors that can affect both the treatment and the potential outcomes:

\[ Y^{p_1} - Y^{p_0} \perp Z \mid X. \]

If this conditional independence holds, and if we also have \( 0 < P(Z = 1 \mid X) < 1 \), we can say that the treatment assignment is strongly ignorable given the covariates \( X \).

In randomized experiments, both of these assumptions are valid; while it is not at all guaranteed for observational studies. The second part of strong ignorability can be assessed by looking at the propensity score distributions of the treatment groups to check for lack of overlap. The first part - unconfoundedness - is however generally not testable. The second essay - chapter 3 focuses on conducting sensitiv-
ity analysis to see how biased the results are when the unconfoundedness assumption is violated, extending previous approaches proposed by Rosenbaum and Rubin (1983b) and Ichino et al. (2008). Rosenbaum and Rubin (1983b)’s method is based on maximum-likelihood estimation. The authors suggested specifying (and varying) sensitivity parameters that determine the distribution of the unmeasured confounder \( U \) as well as the relationship between \( U \) and the treatment \( Z \), the potential outcomes \( Y(1), Y(0) \), given the observed covariates. Ichino et al. (2008) proposed simulating the unmeasured confounder \( U \) for each observation based on pre-specified values that determine the distribution of a binary \( U \) conditional on the treatment assignment and the outcome. This simulated \( U \) is then included the set of matching covariates used to compute the propensity scores and/ or estimate the outcome. Both approaches have binary unmeasured confounder, treatment, and outcome. Our contribution in this essay is offering modifications to these approaches to accommodate for count outcomes. We illustrate these methods using a real study on the application of the combination of center line rumble strips and shoulder rumble strips in Pennsylvania rural highways.

1.3 Propensity score augmented additive-interactive Gaussian process regression

Confounding and selection bias can make treatment effect estimation in observational studies complicated, especially on the individual or subpopulation level. Precise estimation of the heterogeneity of treatment effects is desirable since it can reveal insights about the impact of the observations’ background characteristics on the effect of the treatment. This in turn would help optimizing treatment assignment, for example by identifying the subgroups for whom the treatment will be most effective and thus only administer the treatment accordingly.

Chapter 4 introduces a statistical method to draw valid and robust causal infer-
ence from observational studies that can accommodate high-dimensional covariates as well as heterogeneous treatment effects, essentially combining the strengths of regression and propensity score methods. We propose using Bayesian nonparametric regression with Gaussian process priors to predict the potential outcome function for each treatment condition. Our method employs a sparse additive-interactive model for the nonparametric regression (Qamar and Tokdar, 2014), splitting the regression function into $k$ components, with each component including a small number of interacting covariates. We also include the estimated propensity scores as an additional covariate. We show that the efficiency of the estimators is increased when the estimated propensity score is consistent; and even when the propensity score model is misspecified, the robustness of airGP still carry through. When confounding is an issue, we would have treatment units with covariates $X$ far away from those of the control group; and airGP offset this potential bias due to extrapolation by increasing the estimate for the standard error. We provide simulation results comparing our method against commonly used propensity score methods, Bayesian additive regression trees and random forests, considering both scenarios in which the treatment effect is homogeneous and heterogeneous.
2.1 Introduction

Unconfounded comparison between groups is a central goal in many observational studies. In comparative effectiveness studies, the goal is to estimate the causal effect of a treatment unconfounded by differences between characteristics of subjects assigned to alternative treatment conditions. In noncausal descriptive studies such as racial disparity research, a common goal is to provide an unconfounded comparison of two populations, such as comparing outcomes among populations of different races or of cohorts in different years while controlling for the differences in important characteristics between the populations. Regardless of the purpose of the study, comparisons between groups can be biased when the groups are unbalanced with respect to confounders.

Propensity score methods (Rosenbaum and Rubin, 1983a, 1984) have been increasingly used in many disciplines as a robust method to achieve covariate balance (e.g. D’Agostino, 1998; Rosenbaum, 2002, and references therein). The propensity score is the probability that a unit is assigned to one treatment condition; balance
of the propensity score leads to balance of the multivariate covariates. Compared to standard regression adjustment, propensity score methods are less sensitive to model misspecification (Rubin, 1979). Propensity score methods were developed and have been applied in settings with unstructured cross-sectional data. However, multilevel or clustered data have became increasingly common in practice, where sample units clustered in ways that may be relevant to the analysis, for example by geographical area, treatment center (hospital or physician) or service unit (schools or insurance plans). The unknown mechanism that assigns subjects to clusters may be associated with both measured and unmeasured subject characteristics. These measured and unmeasured factors may create cluster-level variation in treatment quality and/or outcomes, which can be a source of confounding if correlated with treatment assignment at the cluster level (e.g. Griswold et al., 2010). Moreover, standard error calculations that ignore clustering are usually inaccurate. Multilevel models that include fixed effects and/or random effects have been developed to incorporate the hierarchical data structure in regression adjustment (e.g. Gatsonis et al., 1993).

There has been a stream of recent work on the application of propensity score methods to multilevel data. For example, Arpino and Mealli (2011) focused on matching and compared the performance of different propensity score models (random effects and fixed effects models) in the presence of unmeasured confounders at the cluster level via extensive Monte Carlo simulations. Steiner et al. (2013) discussed two propensity score matching methods: within-cluster matching and across-cluster matching. Arpino and Cannas (2015) also compared these two methods and included a third novel preferential matching method (perform matching first within a cluster and if it fails to obtain a match then across clusters)—through simulations based on a real application. Thoemmes and West (2011) gave a comprehensive review of of corrective methods using propensity scores for nonrandomized designs with multilevel data. Keele and Zubizarreta (2014) investigated the use of an optimal multilevel
matching method in a case study of the school voucher system in Chile. Li et al. (2013) focused on weighting and compared the performance of different modeling and weighting strategies via both analytical derivations and simulations. Lingle (2009) examined the subclassification method. Hong and Raudenbush (2006) also used multilevel propensity score stratification in a case study about the policy of retaining low-achieving children in kindergarten rather than promoting them to first grade. Kelcey (2009) carried out three studies focusing on improving causal inferences for observational studies in common educational setting with multilevel settings. Stuart (2007) provided an overview of common methods for estimating causal effects with multilevel data, including randomized experiments, regression analysis, prepost studies, and nonexperimental comparison group designs. Mixed methods, such as mixed weighting-regression, i.e., double-robust estimators (Li et al., 2013) and mixed weighting-subclassification (Hong, 2010), have also been investigated. The consensus among these articles is that the multilevel structure must be properly adjusted for in propensity score analysis, either through modeling or nonparametric adjustment, to reduce bias and increase precision.

This chapter aims to provide a critical review and an extensive discussion of five common classes of propensity score methods in the context of multilevel data: matching, weighting, stratification, mixed matching-regression, and doubly-robust estimators. We illustrated the usage of these methods using a data set about endoscopic vein-graft harvest in coronary artery bypass graft (CABG) surgeries (Shahian et al., 2009), identifying the advantages and limitations of each method. Based on the comparison, we provide general guidelines for using propensity score methods with multilevel data in practice. Important issues of implementation, such as model specification, balance check and standard error calculations are also discussed. We also provide the relevant code in the form of an R package.

The rest of the chapter is organized as follows. We present a motivation example
in Section 2.2. Section 2.3 introduces the basics of propensity score. Section 2.4 presents the multilevel models for estimating propensity score, and Section 2.5 gives an overview of different methods for estimating treatment effects given the estimated propensity score in the context of multilevel data. Section 5 presents an illustration to examine and compare the performance of these methods. Section 6 concludes with general practical guides and a discussion.

2.2 Motivation example

Endoscopic vein-graft harvesting techniques have been used instead of incision-based techniques since the mid 1990s. With their perceived advantages such as reduced post-operative discomfort, shorter patients’ hospital length of stay, increased patients’ satisfaction, reduced wound-infection risk, and reduced incision-site complications, endoscopic vein-graft harvesting techniques were widely adopted and have been used in the majority of more than 400,000 CABG surgeries in the US each year. However, a large observational study using the database from the Project of Ex-vivo Vein Graft Engineering via Transfection IV trial (PREVENT IV; ClinicalTrials.gov number, NCT00042081) in 2009 questioned the safety of this technique (Lopes et al., 2009). Among the 3000 observed patients, those receiving endoscopic vein-graft harvesting had a higher risk of 1-year angiographic vein-graft failure, as well as higher 3-year mortality than those receiving open vein-graft harvesting technique. For more information about the use of endoscopic vein-graft harvesting techniques, see (Lopes et al., 2009) and (JB et al., 2012). We used a data set from the Society of Thoracic Surgeons (STS) Adult Cardiac Surgery Database (ACSD) to demonstrate the propensity score methods, focusing on comparing the long-term outcomes of endoscopic vs open vein-graft harvest in coronary artery bypass graft (CABG) surgeries.

The study cohort included patients age 65 to 100 years undergoing a first-time coronary artery bypass grafting (CABG) operation on an elective or urgent basis
with at least one vein graft and a left internal mammary graft at hospitals participating in the Society of Thoracic Surgeons (STS) Adult Cardiac Database during January 1st, 2008 - June 30th, 2011. Being excluded are patients with a previous CABG, those undergoing emergency CABG or CABG combined with another major procedure, those receiving a radial artery graft, and patients with various rare risk factors usually associated with an emergency preoperative condition (prior shock within 24 hours, prior percutaneous coronary intervention within 6 hours, preoperative intraaortic balloon pump or inotropes). From the 218,090 STS records meeting the above criteria, we kept all records \( N = 138,422 \) that had a matching record in the Medicare database (same hospital, age, sex, admission date, discharge date). From these, we excluded 1,167 records with missing data for vein harvesting technique and 10 records that were coded as first-time CABG operations in the STS database but had a prior CABG claim in the Medicare database. The final study cohort consisted of 137,245 patient operations from 1,044 participating centers.

Mortality status of patients were determined through December 31st, 2014 by the Medicare denominator file. Therefore, all patients were considered to have complete mortality follow-up for at least 3 years following their CABG. Our primary outcome of interest was mortality status at 3 years. The list of individual-level covariates included the patients’ demographic information, hospitalization record, risk factors, previous cardiovascular interventions, preoperative cardiac status and medications, and other operation-related measures. The cluster-level covariates included state, region, census region, and whether the participating site associated with a surgical residency training program. For a complete list and description of covariates, see Table 2.1. There were less than 1% of missing data for the majority of covariates. Missing values of categorical covariates were imputed to the most common category. Missing values of continuous covariates were imputed by stratifying on variables associated with the covariate and imputing to the stratum-specific median value.
The ‘treatment’ $Z$ was whether the patient received the endoscopic technique ($1 = \text{Yes}, 0 = \text{No}$). The outcome $Y$ was mortality status at 3 years ($1 = \text{dead}, 0 = \text{alive}$).

Table 2.1: Covariates included in both the propensity score and outcome models.

<table>
<thead>
<tr>
<th>Covariates</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>xage</td>
<td>Age (years)</td>
</tr>
<tr>
<td>xmale</td>
<td>Male sex, indicator</td>
</tr>
<tr>
<td>xhispanic</td>
<td>Race category: Hispanic</td>
</tr>
<tr>
<td>xasian</td>
<td>Race category: Asian</td>
</tr>
<tr>
<td>xheight</td>
<td>Height in cm</td>
</tr>
<tr>
<td>xbmi</td>
<td>Body mass index (BMI)</td>
</tr>
<tr>
<td>xbmi2</td>
<td>BMI squared</td>
</tr>
<tr>
<td>xpvd</td>
<td>Peripheral vascular disease, indicator</td>
</tr>
<tr>
<td>xcigsmoker</td>
<td>Current or recent smoker, indicator</td>
</tr>
<tr>
<td>xdyslip</td>
<td>Dyslipidemia, indicator</td>
</tr>
<tr>
<td>xhypertn</td>
<td>Hypertension, indicator</td>
</tr>
<tr>
<td>xdiabinsulin</td>
<td>Diabetes, insulin dependent, indicator</td>
</tr>
<tr>
<td>xdiabnoninsulin</td>
<td>Diabetes, non-insulin dependent, indicator</td>
</tr>
<tr>
<td>ximmunmod</td>
<td>Immunosuppressive therapy, indicator</td>
</tr>
<tr>
<td>xrecentafib</td>
<td>Recent atrial fibrillation, indicator</td>
</tr>
<tr>
<td>xcvdncva</td>
<td>Cerebrovascular disease without stroke, indicator</td>
</tr>
<tr>
<td>xcvdca</td>
<td>Cerebrovascular disease with stroke, indicator</td>
</tr>
<tr>
<td>xchnyha4</td>
<td>Heart failure NYHA class IV, indicator</td>
</tr>
<tr>
<td>xchnotnyha4</td>
<td>Heart failure NYHA class I-III, indicator</td>
</tr>
<tr>
<td>xdialysis</td>
<td>Dialysis, indicator</td>
</tr>
<tr>
<td>xcreat000</td>
<td>$\max(0.5, \min(\text{creat}, 5))$</td>
</tr>
<tr>
<td>xcreat100</td>
<td>$\max(xcreat000-1.0, 0)$</td>
</tr>
<tr>
<td>xcreat150</td>
<td>$\max(xcreat000-1.5, 0)$</td>
</tr>
<tr>
<td>xunstable</td>
<td>Unstable angina but no MI within 7 days, indicator</td>
</tr>
<tr>
<td>xms24hr</td>
<td>Prior MI within 24 hours, indicator [note that categories are overlapping]</td>
</tr>
<tr>
<td>xms21d</td>
<td>Prior MI within 21 days, indicator [note that categories are overlapping]</td>
</tr>
<tr>
<td>xmi</td>
<td>Prior MI, indicator [note that MI categories are overlapping]</td>
</tr>
<tr>
<td>xpriorpci</td>
<td>Prior percutaneous coronary intervention (PCI), indicator</td>
</tr>
<tr>
<td>xstatus</td>
<td>Urgent operation, indicator</td>
</tr>
<tr>
<td>xmaindis</td>
<td>Left main disease, indicator</td>
</tr>
<tr>
<td>xdisvtriple</td>
<td>Three-vessel disease, indicator [note that categories are overlapping]</td>
</tr>
<tr>
<td>xchrlungmod</td>
<td>Chronic lung disease, moderate, indicator</td>
</tr>
<tr>
<td>xchrlungsev</td>
<td>Chronic lung disease, severe, indicator</td>
</tr>
<tr>
<td>xef</td>
<td>$\min(\text{Ejection fraction } %, 50)$</td>
</tr>
<tr>
<td>xinsufmge3</td>
<td>Mitral insufficiency, indicator</td>
</tr>
<tr>
<td>xsurgeryyear</td>
<td>Year of operation</td>
</tr>
<tr>
<td>xacademic</td>
<td>Academic hospital, indicator</td>
</tr>
<tr>
<td>xregion1</td>
<td>Hospital region, Midwest, indicator</td>
</tr>
<tr>
<td>xregion2</td>
<td>Hospital region, Northeast, indicator</td>
</tr>
<tr>
<td>xregion3</td>
<td>Hospital region, South, indicator</td>
</tr>
</tbody>
</table>

We restrict the analysis to only participating centers with at least 6 patients.
for each technique, yielding 66779 patient operations from 394 participating centers. Among these patients, 49568 received the endoscopic technique and 17211 received incision-based open vein-graft harvesting technique, or a 74.2% overall proportion of endoscopic patients, with substantial variation across centers (1.7% - 99.2%). Figure 2.1 plots the histogram of the overall proportion of endoscopic patients among different centers. To assess the covariate balance between races, we present in Table 2.2 the percentiles of the propensity scores estimated from a fixed effects model with all covariates in Table 2.1. These covariates are chosen for consistency with the published STS CABG mortality model (Shahian et al., 2009) as well as a prior analysis of endoscopic versus open vein harvesting using STS data (JB et al., 2012). Even though the range of the estimated propensity scores are similar among the two groups, the endoscopic group has much higher propensity scores than the direct-vision group at all quantiles. Figure 2.2 plots the estimated propensity scores of the base data by each technique.

Table 2.2: Summary of the estimated propensity score by treatment group.

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Min</th>
<th>Q1</th>
<th>Median</th>
<th>Mean</th>
<th>Q3</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Direct vision</td>
<td>0.005</td>
<td>0.255</td>
<td>0.478</td>
<td>0.480</td>
<td>0.705</td>
<td>0.994</td>
</tr>
<tr>
<td>Endoscopic</td>
<td>0.012</td>
<td>0.772</td>
<td>0.908</td>
<td>0.834</td>
<td>0.958</td>
<td>0.995</td>
</tr>
</tbody>
</table>

Note: Centers with less than 6 patients for each technique are excluded. Propensity scores are estimated using a fixed-effect model.

2.3 Basics of propensity score methods

To simplify, we focus on two-level structures. Consider a sample or population of $N$ units, from cluster $h$ ($h = 1, ..., H$), each including $n_h$ units indexed by $k = 1, ..., n_h$, and $N = \sum_h n_h$. Each unit belongs to one of two groups for which covariate-balanced comparisons are of interest, possibly defined by a treatment; in either case we will
Figure 2.1: Histogram of endoscopic proportion at participating centers, by each technique. Included are 394 participating centers with more than 5 endoscopic and 5 open-heart cases.

use the terms “treatment” and “control” to refer to the groups, and let $N_0$ and $N_1$ denote the number of treated and control units. For unit $k$ in cluster $h$, let $Z_{hk}$ be the binary treatment variable, equal to 1 if the subject is assigned to the treatment and equal to 0 if the control. Let $Q_{hk} = Q_h$ be the cluster indicator for each unit. Also let $U_{hk}$ be a vector of unit-level covariates, $V_h$ be a vector of cluster-level covariates, and $X_{hk} = (U_{hk}, V_h)$. For each unit, an outcome $Y_{hk}$ is observed.

**Estimands.** We define the population average treatment effect (ATE) as:

$$
\tau_{ATE} = \mathbb{E}[Y(1) - Y(0)],
$$

the expected difference in units’ potential outcomes. Another estimand usually of
Figure 2.2: Histogram of the estimated propensity scores of the full data by each technique. Included are 394 participating centers with more than 5 endoscopic and 5 open-heart cases.

Interest is the population average treatment effect for the treated (ATT):

$$\tau^{\text{ATT}} = E[Y(1) - Y(0) | Z = 1]$$.

Assumptions. For causal comparisons, we adopt the Rubin Causal Model (RCM) (Rubin, 1974; Holland, 1986), which entails the standard Stable Unit Treatment Value Assumption (SUTVA) (Rubin, 1980). SUTVA has two components. The first component is that the potential outcomes for each unit are unaffected by the treatment assignments received by other units. In other words, there is no interference among units. For our endoscopic vein-harvesting data, this first component of SUTVA is easily satisfied: whether a patient receives endoscopic or open vein-graft harvesting technique would essentially have no effect on another patient’s 3-year
mortality status. The second component of SUTVA is that there are no different versions of either the control or the treatment group: All units in the treatment (or control) group receive the exact same treatment. This component of SUTVA might be more questionable in healthcare-cluster settings such as our example, since the implementation of each technique may vary considerably across healthcare centers and surgeons. Bigger, higher-ranked participating centers might be better equipped and hire better surgeons, which in turn can affect the treatment quality. Nevertheless, this information is usually hard to quantified as well as observed, so we assume that both techniques are common enough so that doctors at different healthcare centers perform them with fairly similar skills.

Under SUTVA, each unit has two potential outcomes $Y_{hk}(z)$ for $z = 0, 1$, corresponding to the two treatment conditions. For each unit, only the potential outcome corresponding to the observed treatment condition, $Y_{hk}$, is observed with $Y_{hk} = Y_{hk}(1)Z_{hk} + Y_{hk}(0)(1 - Z_{hk})$, and the other is missing.

Since only one potential outcome is observed for each unit, additional assumptions are necessary to estimate the causal effects from observed data. The standard assumption is unconfoundedness, also known as the assumption of no unmeasured confounders, which states that the treatment is effectively randomized within cells defined by the values of observed covariates. For multilevel data, there’s a cascade of unconfoundedness assumption which is tied to the assignment mechanism:

1. $\{Y_{hk}(1), Y_{hk}(0)\} \perp Z_{hk}$: potential outcomes are unconfounded (or completely randomized) within and across clusters;

2. $\{Y_{hk}(1), Y_{hk}(0)\} \perp Z_{hk}|(U_{hk}, V_h)$: potential outcomes are unconfounded conditional on the observed values of individual and cluster-level covariates;

3. $\{Y_{hk}(1), Y_{hk}(0)\} \perp Z_{hk}|(U_{hk}, V_h, Q_h)$: potential outcomes are unconfounded within the same cluster, conditional on the observed values of individual and
cluster-level covariates.

Both descriptive and causal comparisons also require the overlap or positivity assumption: \( 0 < \Pr(Z = 1|\mathbf{X}) < 1 \), for all units, which states that the study population is restricted to values of covariates for which there can be both control and treated units. Jointly overlap and unconfoundedness define the strong ignorability condition (Rosenbaum and Rubin, 1983a).

**Assignment mechanism.** For multilevel observational data like in our endoscopic vein-graft harvesting example, both cluster-level and individual-level covariates can have an impact on the treatment assignment mechanism. Moreover, the effects of the individual-level covariates on the treatment assignment may even vary from cluster to cluster. When there is a clustering structure in the data, every propensity score model assumes an underlying assignment mechanism.

Using the same terminology as in Li et al. (2013), assignment mechanism 2 corresponds to the marginal propensity score model where all clusters and individuals share the same intercept and only observed covariates \( \mathbf{X}_{ij} \) are included as predictors. Both the fixed effects and random effects propensity score models implicitly assume assignment mechanism 3. However, it is not immediately clear what is the explicit form of the assignment mechanism corresponding to the random effects model, where the cluster-specific effects (e.g. random effects) follow a distribution. Yang (2017) specifies the latent ignorability (LI) assumption:

\[
(Y_{hk}(1), Y_{hk}(0)) \perp Z_{hk}|(X_{hk}, U_h),
\]

where \( U_h \) is a “cluster-specific latent effect that summarizes the effect of unobserved cluster-level confounders.” LI is in fact the same as Assumption (iii), replacing \( U_h \) by the cluster indicator \( Q_h \). This does not solve the conceptual difficulty of which assignment mechanism corresponds to the random effects model. The explicit difference between random effects and fixed effects model is whether the cluster-specific
intercepts follow a distribution, i.e., they were sampled from a super population. This suggests that the above notation system for the assignment mechanism might not be adequate and we must introduce super population and parameters to correctly define the assignment mechanism.

**Propensity scores.** The propensity score is the conditional probability of being in treatment group $Z = 1$ given covariates $X$:

$$e(X) = \Pr(Z = 1 \mid X).$$

Rosenbaum and Rubin (1983a) proved that the propensity score has two important properties: (i) it is a balancing score (in fact, the coarsest balancing score), that is, balancing the distribution of $e(X)$ between the treatment groups would balance the distribution of all $X$; and (ii) if the treatment assignment is strongly ignorable given $X$, then it is also strongly ignorable given $e(X)$. Therefore, the propensity score can be viewed as a summary score of the multivariate covariates.

Propensity score methods usually involve two stages: in the first stage, one estimates the propensity score from the data; and in the second stage, one estimates the treatment effects based on the estimated propensity score, via matching, weighting, subclassification, regression adjustment or combination of these methods. For a review of propensity score methods in unstructured data, see Imbens (2004). To apply propensity score methods to multilevel data, the central question is how to adjust for the multilevel structure in each of the two stages, which will be discussed in the next two sections.

2.4 Stage 1: Estimating propensity score with multilevel data

In the context of multilevel data, previous studies (e.g. Arpino and Mealli, 2011; Li et al., 2013; Arpino and Cannas, 2015) show that using single-level models to estimate propensity score, even with cluster-level covariates being included, generally leads to
severe bias in estimation, while multilevel models that include fixed effects and/or random effects give better fit of the data.

A fixed effects logistic model for the propensity score fits a cluster-level main effect for each cluster via creating a dummy variable for being in each cluster:

\[
\logit(e_{hk}) = \alpha_{0h} + U_{hk}\alpha.
\]  

(2.4)

The cluster-specific intercept (main effects) \(\alpha_{0h}\) absorbs the effects of both observed and unobserved cluster-level covariates of cluster \(h\), and thus the cluster-level covariates \(V_h\) are not required in the model. Therefore the fixed effects model can protect against misspecification due to cluster-level confounders, but it might lead to larger variance than the propensity score estimated under a correct model with fully observed \(V_h\). When there are many small clusters, the fixed effects model may lead to inconsistent score estimates due to the large number of free parameters, an example of the Neyman-Scott problem (Neyman and Scott, 1948), and the possibility of separation (representation of only one group) in some clusters, which are required to be excluded from the analysis by the overlap assumption.

An alternative multilevel model is the random effects or mixed effects model (Laird and Ware, 1982), which augments (2.4) with a prior distribution on the cluster-specific main effects \(\alpha_{0h}\):

\[
\logit(e_{hk}) = \alpha_{0h} + X_{hk}\alpha, \quad \alpha_{0h} \sim N(\alpha_0, \sigma^2_\alpha),
\]  

(2.5)

where both unit-level \((U_{hk})\) and cluster-level covariates \((V_h)\) are required in the model. More generally, the random effects may also include random coefficients of some individual-level covariates by imposing cluster-specific slopes \(\alpha_{0h}\) with a prior distribution. One can use the posterior mode or mean to estimate propensity scores from (2.5). The prior distribution on \(\alpha_{0h}\) in the random effects model allows for "borrowing information" across clusters, and largely reduces the number of parameters
for cluster-specific effects compared to fixed effects models. However, the random effects model does not guarantee balance within each cluster, due to the shrinkage of random effects toward zero, and therefore is more reliant on inclusion of important cluster-level covariates $V_k$ as regressors. It would also produce a biased estimate if the cluster-specific random effects are correlated with any of the covariates (Mundlak, 1978). Results of the random effects model converge to those from a corresponding fixed effects model as the sample size per cluster increases.

2.5 Stage 2: Estimating treatment effects with multilevel data

There are a range of model-free, model-based and mixed approaches available for estimating the treatment effects based on estimated propensity scores. Below we present five popular classes of methods and discuss the implications of the multilevel structure to each method.

2.5.1 Matching

Matching is the most widely used method for balancing covariates in practice (Stuart, 2010). The general procedure of matching for estimating ATE is as follows: first, for each unit, find $M$ closest matched units in the opposite treatment group according to a distance metric defined in the space of the covariates; second, take the average of the observed outcome of the $M$ matches as the estimated counterfactual outcome of that unit; and finally estimate the ATE by the average of the estimated individual effects of all matched units, and unmatched units are discarded. One can estimate ATT by applying the same procedure to treated units only.

The distance metric in this chapter is the estimated propensity score. For multilevel data, the propensity scores are estimated from the models in Section 2.4, and there are three general strategies for finding matches:

1. **Within-cluster matching**: Matched samples of a unit are restricted to the same
cluster of that unit. This strategy automatically balances cluster-level (both measured and unmeasured) confounders. However, it may also result in a large proportion of sample units unmatched and thus discarded, especially for studies with small cluster sample sizes. Even for studies with medium to large cluster sample sizes, an example in Arpino and Cannas (2015) shows that up to 90% of units could still be unmatched using within-cluster matching. Besides leading to inflated standard errors of the estimate, this would also alter the target population where the estimand is defined on. Therefore, unless all of the cluster sample sizes are very large or there are practical reasons, we do not recommend to use within-cluster matching.

2. **Across-cluster matching**: Matched samples of a unit are searched across all clusters. This strategy usually results in a high percentage of units being matched and thus smaller standard errors of the estimates compared to within-cluster matching. However, across-cluster matching may lead to large imbalance in cluster-level covariates between treatment groups, which in turn would increase bias of the estimates.

3. **Preferential matching**: Matched samples of a unit are first searched within the same cluster, if unavailable, then the search is expanded to other clusters. Arpino and Cannas (2015) proposed this hybrid strategy to address the bias-variance tradeoff between within-cluster and across-cluster matching, and showed that it outperforms the other two strategies in simulations.

Once the matched sample is obtained, let $\mathcal{M}_{hk}$ be the set of the indices of the $M$ closest matches of unit $k$ in cluster $h$, and let

$$
\hat{Y}_{hk}(0) = \left\{ \begin{array}{ll}
\sum_{j \in \mathcal{M}_{hk}} Y_j / M, & Z_{hk} = 1, \\
Y_{hk}, & Z_{hk} = 0,
\end{array} \right.
\quad \text{and} \quad
\hat{Y}_{hk}(1) = \left\{ \begin{array}{ll}
\sum_{j \in \mathcal{M}_{hk}} Y_j / M, & Z_{hk} = 1, \\
Y_{hk}, & Z_{hk} = 0.
\end{array} \right.
$$
Then the matching estimators for the ATE and ATT, respectively, are:

\[
\hat{\tau}_{\text{mat}}^{\text{ATE}} = \sum_{h,k} \left( \hat{Y}_{hk}(1) - \hat{Y}_{hk}(0) \right) / N, \quad \text{and} \quad \hat{\tau}_{\text{mat}}^{\text{ATT}} = \sum_{h,k} \left( Y_{hk} - \hat{Y}_{hk}(0) \right) Z_{hk} / N. \quad (2.6)
\]

Matching methods are intuitive and can be readily implemented by several open-source computer packages, e.g., the Matching (Sekhon, 2015) and Matchit (Ho et al., 2015) packages in R. Matching is a “bottom-up” approach in the sense that it creates local balance by matching nearby cases, and balance at more aggregated levels is a by-product if the matching is successful. A benefit of propensity score based matching is that it also approximately balances distributions in directions orthogonal to the estimated propensity score (Rubin, 2001), protecting against misspecification of the propensity score model. However, matching algorithms require tuning on a number of parameters, for example, the number of matches, matching with or without replacement, and caliber of matching. Tuning can be particularly challenging in multilevel data. Moreover, if the number of matches is fixed and matching is done with replacement, Abadie and Imbens (2006) showed that matching estimators are asymptotically biased, due to the residual imbalance in covariates.

2.5.2 Mixed matching-regression

To reduce the bias induced by the residual imbalance in matching, Rubin (1973) proposed to perform bias correction via regression on the matched sample. Abadie and Imbens (2011) further advanced this proposal, providing theoretical basis and computer software. The mixed matching-regression approach has been shown to be advantageous to matching in unstructured data (Abadie and Imbens, 2011; Mercatanti and Li, 2014), but has not been explored in multilevel data.

Let \( \mu_z(x) = \mathbb{E}[Y(z)|X = x] \), and \( \hat{\mu}_z(X_{hk}) \) be a consistent estimator of \( \mu_z(X_{hk}) \), for \( z = 0, 1 \). A regression estimator uses \( \hat{\mu}_z(X_{hk}) \) to impute missing potential out-
comes $Y_{hk}(z)$. Specifically, let

$$
\hat{Y}_{hk}(0) = \left\{ \begin{array}{ll}
\frac{1}{Y_{hk}} \sum_{j \in \mathcal{M}_{hk}} [Y_j + \hat{\mu}_0(X_{hk}) - \hat{\mu}_0(X_j)]/M, & Z_{hk} = 1, \\
Y_{hk}, & Z_{hk} = 0,
\end{array} \right.
$$

and

$$
\hat{Y}_{hk}(1) = \left\{ \begin{array}{ll}
\frac{1}{Y_{hk}} \sum_{j \in \mathcal{M}_{hk}} [Y_j + \hat{\mu}_1(X_{hk}) - \hat{\mu}_1(X_j)]/M, & Z_{hk} = 1, \\
Y_{hk}, & Z_{hk} = 0.
\end{array} \right.
$$

Then the bias-corrected matching estimators for the ATE and the ATT are respectively:

$$
\hat{\tau}^{ATE}_{mix} = \frac{\sum_{h,k} \left( \hat{Y}_{hk}(1) - \hat{Y}_{hk}(0) \right) / N, \quad \text{and} \quad \hat{\tau}^{ATT}_{mix} = \frac{\sum_{h,k} \left( Y_{hk} - \hat{Y}_{hk}(0) \right) Z_{hk} / N_1.}{(2.7)}
$$

With multilevel data, we recommend to specify multilevel regression models for estimating the potential outcomes ($\hat{\mu}_z(X_{hk})$), analogous to the fixed effects or random/mixed effects models for the propensity score in Section 2.4. For example, for a continuous outcome, a fixed effects model adjusts for cluster-level main effects (via a dummy variable for being in each cluster) and individual covariates:

$$
Y_{hk}(z) = \beta_{0h} + z\gamma_1 + U_{hk}\beta + z \cdot U_{hk} \cdot \gamma_2 + \epsilon_{hk}, \quad (2.8)
$$

where $\beta_{0h}$ is the cluster-specific main effect, which absorbs all between-cluster information, and thus cluster-level covariates $V_h$ do not enter the model, analogous to Model (2.4). The corresponding random/mixed effects model is obtained by imposing a prior distribution on $\beta_{0h}$’s:

$$
Y_{hk}(z) = \beta_{0h} + z\gamma_1 + X_{hk}\beta + z \cdot X_{hk} \cdot \gamma_2 + \epsilon_{hk}, \quad \beta_{0h} \sim N(\beta_0, \sigma_\beta^2), \quad (2.9)
$$

where cluster-level covariates $V_h$ are required in the model. More generally, one can assume a random slope for some of the coefficients $\beta$ and/or random treatment effects, replacing the $\gamma$s by their cluster-specific counterparts. Analogous generalized
linear models (GLM) or generalized linear mixed models (GLMM) can be used for binary or ordinal outcomes. Note that these models for potential outcomes translate to models for observed outcomes under unconfoundedness, replacing $z$ by the observed treatment assignment $Z$ and $Y(z)$ by the observed outcome $Y$ for each unit.

2.5.3 Weighting

The basis for propensity score weighting is the following equations,

$$
\mathbb{E} \left[ \frac{ZY}{e(\mathbf{X})} - \frac{(1 - Z)Y}{1 - e(\mathbf{X})} \right] = \tau_{ATE}, \quad \text{and} \quad \mathbb{E} \left[ ZY - \frac{(1 - Z)e(\mathbf{X})Y}{1 - e(\mathbf{X})} \right] = \tau_{ATT}. \tag{2.10}
$$

Therefore, ATE and ATT can be estimated by comparing weighted averages of the observed outcomes using the Horvitz-Thompson (HT, also known as inverse-probability) and ATT weights, respectively:

\[
\begin{align*}
\text{w}_{ATE}^{hk} &= \begin{cases} 
1/\hat{e}_{hk}, & Z_{hk} = 1, \\
1/(1 - \hat{e}_{hk}), & Z_{hk} = 0,
\end{cases} \\
\text{w}_{ATT}^{hk} &= \begin{cases} 
1, & Z_{hk} = 1, \\
\hat{e}_{hk}/(1 - \hat{e}_{hk}), & Z_{hk} = 0.
\end{cases}
\end{align*}
\tag{2.11}
\]

Li et al. (0) show that both the HT and the ATT weights are special cases of the general class of balancing weights, which balance the weighted distribution of covariates between the two comparison groups of a target population.

With multilevel data, one can first estimate the target estimand (ATE or ATT) within each cluster:

\[
\hat{\tau}_h = \sum_{k \in h} Y_{hk} Z_{hk} w_{hk}/w_{h1} - \sum_{k \in h} Y_{hk} (1 - Z_{hk}) w_{hk}/w_{h0},
\]

where $w_{hz} = \sum_{k \in h} w_{hk}$ for $z = 0, 1$, and then take their mean weighted by the total weights in each cluster, $w_h = w_{h0} + w_{h1}$:

\[
\hat{\tau}_{wt} = \frac{\sum_h w_h \hat{\tau}_h}{\sum_h w_h}. \tag{2.12}
\]

23
This estimator is referred to as the “cluster weighting” estimator in Li et al. (2013). To implement the weighting estimator with propensity scores estimated from a fixed effects model, one needs to exclude the clusters where all units are assigned to the group because the sum of the weights of these clusters will be zero. This may change the target population from the combined population (ATE) or the treated population (ATT) to their counterparts restricted to the subpopulation with overlap in each cluster. Another estimator is the “marginal weighting” estimator where each unit is weighted in the same fashion across all clusters:

\[
\hat{\tau} = \frac{\sum_{h,k} Y_{hk}Z_{hk}w_{hk}}{w_1} - \frac{\sum_{h,k} Y_{hk}(1-Z_{hk})w_{hk}}{w_0},
\]

(2.13)

\[w_z = \sum_{h,k}^{Z_{hk}=z} w_{hk} \text{ for } z = 0, 1.\] Li et al. (2013) show that the cluster weighting estimator usually outperforms the marginal weighting estimator.

In contrast to matching, weighting is a “top-down” approach in the sense that it is designed to create global balance for the target population, and balance at more detailed levels is a by-product of more complex models. Weighting methods require minimal tuning and are easy to extend to more complex settings such as complex surveys and longitudinal data. However, inverse-probability based weighting methods (including HT and ATT weights) are sensitive to probabilities close to 0 or 1, which would introduce extremely large weights and consequently dominate the estimates, resulting in poor balance and very large variance. To avoid such a situation, a common strategy in practice is to trim the weights, removing units with propensity scores outside a pre-fixed threshold, e.g. (0.1, 0.9) (Crump et al., 2009).

2.5.4 Doubly-robust (DR) estimators

The DR estimator (e.g. Robins et al., 1995; Robins and Rotnitzky, 1995; Bang and Robins, 2005) combines weighting and regression, augmenting the inverse-probability
weighting estimators. For the ATE, one common DR estimator is proposed in Lunceford and Davidian (2004) as follows:

$$\hat{\tau}_{ATE}^{dr} = \sum_{h,k} \left\{ \left[ \frac{Z_{hk}Y_{hk}}{\hat{e}_{hk}} - \frac{(Z_{hk} - \hat{e}_{hk})\hat{\mu}_1(X_{hk})}{\hat{e}_{hk}} \right] \right. $$

$$\left. - \left[ \frac{(1 - Z_{hk})Y_{hk}}{1 - \hat{e}_{hk}} + \frac{(Z_{hk} - \hat{e}_{hk})\hat{\mu}_0(X_{hk})}{1 - \hat{e}_{hk}} \right] \right\} / N. \quad (2.14)$$

For the ATT, Mercatanti and Li (2014) proposed the following DR estimator:

$$\hat{\tau}_{ATT}^{dr} = \sum_{h,k} \left[ Y_{hk}Z_{hk} - \frac{Y_{hk}(1 - Z_{hk})\hat{e}_{hk} + \hat{\mu}_0(X_{hk})(Z_{hk} - \hat{e}_{hk})}{1 - \hat{e}_{hk}} \right] / N_1, \quad (2.15)$$

where $\hat{\mu}_z(X_{hk})$ is the fitted potential outcome from an outcome model in group $z$.

The term ‘doubly-robust’ refers to the large sample property that $\hat{\tau}_{dr}$ is a consistent estimator of $\tau$ if either the propensity score model or the potential outcome model is correctly specified, but not necessarily both. The price for the robustness is efficiency: if one (and only one) model is misspecified, the DR estimate will have a larger standard error than the correctly specified direct regression or weighting estimates. The DR estimator can also be used as a diagnostic tool for modeling (Mercatanti and Li, 2014). All three estimates: weighting, regression, and DR can be obtained and compared to one another. A DR estimate close to the weighting but different from the regression estimate suggests that the outcome model might be misspecified. On the other hand, if the DR estimate is close to the regression estimate and far from the weighting estimate, then the propensity score model might be misspecified.

With multilevel data, similar to the mixed matching-regression approach, we recommend to specify multilevel models such as Model (2.8) and (2.9) for estimating the outcomes.
2.5.5 Stratification

Stratification or subclassification is another popular propensity score method (Rosenbaum and Rubin, 1984). The sample is partitioned into subclasses based on the quantiles of the estimated propensity scores, so that the estimated propensity scores are similar within the subclasses. Specifically, partition the range of the propensity score (i.e., the unit interval) into $L$ subclasses with boundary values equal to $l/L$ for $l = 1, \ldots, L-1$. Let $B_{hk,l}$ denote the indicator of unit $hk$ being in block $l$ ($l = 1, \ldots, L$) with $B_{hk,l} = 1 \{ (l-1)/L < \hat{e}_{hk} \leq l/L \}$, and let the number of sample units in block $l$ with $Z = z$ be $N_{(l),z}$ and $N_{(l)} = N_{(l),0} + N_{(l),1}$. For the ATE, one first obtains the within-block estimate of ATE:

$$
\hat{\tau}_l = \sum_{hk} B_{hk,l}Y_{hk}Z_{hk}/N_{(l),1} - \sum_{hk} B_{hk,l}Y_{hk}(1 - Z_{hk})/N_{(l),0},
$$

and then take their average weighted by the block sample size:

$$
\hat{\tau}_{ATE}^{str} = \frac{1}{L} \sum_{l=1}^{L} \hat{\tau}_l \cdot N_{(l)}/N. \quad (2.16)
$$

For the ATT, one will weight the within-block ATE by the number of treated units (Imbens and Rubin, 2015, Section 17):

$$
\hat{\tau}_{ATT}^{str} = \frac{1}{L} \sum_{l=1}^{L} \hat{\tau}_l \cdot N_{(l),1}/N_1. \quad (2.17)
$$

Stratification can also be mixed with regression for further covariate adjustment, that is, first estimating causal effects using regression within each propensity score subclass and then averaging the within-subclass estimates. Stratification can be viewed as a coarsened version of weighting, mitigating the sensitivity issue due to extreme weights at the cost of efficiency. With multilevel data, it is not necessary and in fact often infeasible to first calculate cluster-specific effects and then average.
2.5.6 Goodness-of-fit

The goodness of fit of some of these models can be assessed by standard diagnostics of covariate balance (e.g. Imbens and Rubin, 2015, Chapter 13), checking the overall and cluster-specific balance of covariates. Specifically, the standardized difference (SDF) can be used to check the balance of continuous and binary covariates between the two treatment groups:

\[
SDF = \left( \frac{\sum_{h,k} X_{hk} Z_{hk} t_{hk} - \sum_{h,k} X_{hk}(1 - Z_{hk}) t_{hk}}{\sqrt{s_1^2 + s_0^2}/2} \right),
\]

(2.18)

where \( s_z \) is the standard deviation of the unweighted covariate in group \( z \) (\( z = 0, 1 \)). For the original data, \( t_{hk} = 1 \) for each unit and SDF is the standard two-sample t-statistic. For the weighting estimators, \( t_{hk} \) are the weights. For the matching estimators, \( t_{hk} \) equal the number of times that unit appears in the matched sample - equivalently, the numerator in equation (2.18) is the difference in average of covariate \( X \) between treatment groups in the matched sample. Clearly, this method of assessing the goodness-of-fit doesn’t apply to clustered weighting, stratification, or the mixed methods; however, it is still a useful diagnostic tool for other methods. There is no theoretically backed or universally accepted criterion for the cut-off point of the SDF that indicates significant imbalance; however, an SDF with absolute value less than 0.1 has commonly been used to indicate negligible difference.

2.5.7 Standard errors calculation

Standard errors of the above estimators can be obtained using the delta method or bootstrapping. The delta method, when available, is straightforward and computationally efficient, but it is not always available, particularly for multilevel data. Lunceford and Davidian (2004) derived the approximate sampling variances for the marginal weighting and DR ATE estimators when the propensity score model is
of the form \( \{1 + \exp(-W^T \beta)\}^{-1} \) where \( W \) is a function of elements in \( X \). The approximate sampling variances are computed as \( n^{-2} \sum_{h,k} \hat{I}_{hk}^2 \), where:

\[
\hat{I}_{mwt,hk} = \frac{Z_{hk} Y_{hk}}{\hat{e}_{hk}} - \frac{(1 - Z_{hk}) Y_{hk}}{1 - \hat{e}_{hk}} - \hat{\tau}_{mwt} - (Z_{hk} - \hat{e}_{hk}) \hat{H}_\beta \hat{E}_{\beta\beta}^{-1} W_{hk};
\]

\[
\hat{I}_{DR,hk} = \frac{Z_{hk} Y_{hk} - (Z_{hk} - \hat{e}_{hk}) \hat{\mu}_1(X_{hk})}{\hat{e}_{hk}} - \frac{(1 - Z_{hk}) Y_{hk} + (Z_{hk} - \hat{e}_{hk}) \hat{\mu}_0(X_{hk})}{1 - \hat{e}_{hk}} - \hat{\tau}_{DR} \tag{2.19}
\]

\[
\hat{E}_{\beta\beta}^{-1} = n^{-1} \sum_{h,k} \hat{e}_{hk} (1 - \hat{e}_{h,k}) W_{hk} W_{hk}^T;
\]

\[
\hat{H}_\beta = n^{-1} \sum_{h,k} \left\{ \frac{Z_{hk} Y_{hk} (1 - \hat{e}_{hk})}{\hat{e}_{hk}} + \frac{(1 - Z_{hk}) Y_{hk} \hat{e}_{hk}}{1 - \hat{e}_{hk}} \right\} W_{hk}. \tag{2.20}
\]

Shinozaki and Matsuyama (2015) derived their ATT counterparts when both the propensity score and outcome are modeled with logistic regression (i.e. the ATT DR variance formula is only available for binary outcomes). The full formula is more complicated than their ATE counterparts and can be found at the paper cited above. However, it should be noted that the ATT standard error formula involve inverting matrices, which is not always feasible. Bootstrapping is versatile but requires substantial computational effort. For bootstrapping, one should resample the clusters rather than the individual units with multilevel data, that is, each bootstrap sample consists of \( H \) clusters (re)sampled with replacement from the clusters of the original data. Resampling clusters preserves the between- and within-cluster variance. The total sample size of each bootstrap sample usually differs. A second approach is to first resample the clusters followed by a second resampling step of individuals within each sampled cluster. Our experience suggests that there is little difference in the results between the two strategies and thus we recommend the first.

Bootstrapping is theoretically valid for weighting and stratification based methods, but may be invalid for matching methods with a fixed number of matches. Abadie and Imbens (2008) showed that the bootstrap standard errors are not valid
for matching estimators, as the process introduces discreteness in the distribution that will lead to ties in the matching algorithm (Abadie and Imbens, 2006). Instead one can use the estimator proposed by Abadie and Imbens (2011) to calculate the standard errors for both simple and bias-corrected matching estimators. The R Package \texttt{Matching} can be used to obtain this Abadie-Imbens standard error, which has correct coverage if the propensity score is known. In the case the propensity scores are estimated, the standard errors only account for the uncertainty of the matching procedure itself and not in estimating the propensity scores.

2.6 R Package: \texttt{psmultilevel}

We provide computer code for a subset of methods discussed above in the form of an R package, \texttt{psmultilevel}. \texttt{psmultilevel} implements a variety of algorithms for calculating propensity score weighting estimators and their standard errors for treatment effects in a clustered data setting. The treatment is assigned on an individual level, similar to our endoscopic vein-graft harvesting technique example. We include an optional threshold for the minimum number of units in each treatment group of each cluster - clusters don’t meet this requirement are removed before the propensity scores are estimated. This package offers four weighting estimators: marginal weighting, clustered weighting, stratification, and doubly-robust. Standard errors include closed-form formula when available, as well as bootstrapped errors.

We include in this package a simulated data set, \texttt{simdata}, where the treatment is randomly assigned to 30% of the observations. This is a data frame with 1000 observations and 5 variables named \texttt{X1,X2,cluster,Z,Y}. The first two are independent observed variables, \texttt{cluster} identifies the cluster to which the observation belongs. \texttt{Z} denotes the treatment assignment, and \texttt{Y} is the outcome of interest. To calculate the ATE for this data set using marginal weighting and a random-effects propensity score model, the syntax is as follows:
ps.formula.re<-function()
{
    return(Z~ X1+X2+ (1|cluster))
}
data<-simdata[,c(1:2)]

psmultilevel(data=data, Y=simdata$Y, Z=simdata$Z,
cluster=simdata$cluster, y.formula=NULL, pscore.formula=ps.formula.re(),
estimand='ATE', method='marwt',modeltype='random',y.type=NULL,
se.report=FALSE, ps.cut=c(0.05,0.95),cl.cutoff=0)

This package is available at: https://github.com/nln6/psmultilevel.

2.7 Application to the STS Data

We apply the previously discussed propensity-score methods on our motivating example about the effect of endoscopic vs open vein-graft harvesting technique on patients’ 3-year mortality status, having ATE - the sample average treatment effect - as the estimand of interest. We look at the data by each year, as well as the combined data set. Table 2.3 presents the distribution of the cluster size for each year. We remove clusters with less than 4 observations in each group for the yearly estimates, and remove clusters with less than 6 observations in each group for the overall estimates.

<table>
<thead>
<tr>
<th>Year</th>
<th>N_obs</th>
<th># of clusters</th>
<th>Min</th>
<th>Q1</th>
<th>Median</th>
<th>Mean</th>
<th>Q3</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>2008</td>
<td>15182</td>
<td>249</td>
<td>9.0</td>
<td>31.0</td>
<td>46.0</td>
<td>61.0</td>
<td>71.0</td>
<td>392.0</td>
</tr>
<tr>
<td>2009</td>
<td>11983</td>
<td>213</td>
<td>8.0</td>
<td>28.0</td>
<td>46.0</td>
<td>56.3</td>
<td>69.0</td>
<td>279.0</td>
</tr>
<tr>
<td>2010</td>
<td>10384</td>
<td>199</td>
<td>9.0</td>
<td>25.0</td>
<td>42.0</td>
<td>52.2</td>
<td>63.0</td>
<td>283.0</td>
</tr>
<tr>
<td>2011</td>
<td>3178</td>
<td>98</td>
<td>8.0</td>
<td>17.5</td>
<td>24.5</td>
<td>32.4</td>
<td>39.8</td>
<td>124.0</td>
</tr>
</tbody>
</table>

Note: Only clusters with at least 4 endoscopic and 4 direct-vision cases are included.

For each data set, we first estimate the propensity score from a fixed effect and a random effect model. We then estimate the ATE by four model-free methods (across-cluster matching, marginal weighting, cluster weighting, stratification) as
well as two mixed methods (matching-regression and Doubly-Robust) based on the estimated propensity score. We match the outcome model in the regression-adjusted methods with the propensity score model in the sense that a random (or fixed) effects outcome model would be used if a random (or fixed) effects propensity score model was used. The random effects models are fitted using the \texttt{lmer} and \texttt{glmer} commands in the \texttt{lme4} package in R.

For the weighting methods (model-free and mixed), we remove observations with estimated propensity scores outside the range of $[0.05,0.95]$. Stratification is performed using six subclasses, the default option of the \texttt{MatchIt} package in R. For across-cluster matching, we choose nearest neighbor 1-1 matching with the estimated propensity score as the distance, using the \texttt{Match} command in the \texttt{Matching} package in R. For the bias-corrected matching estimator, adjustment coefficients were obtained by estimating a regression function on the pooled matched sample. Imbens and Rubins (2015) discussed two other bias-adjustment approaches. The first approach regresses the within-pair outcome difference on the matching discrepancy. The second approach estimates the regression function for $\hat{\mu}_0(X_{hk}) (\hat{\mu}_1(X_{hk}))$ using all control (treated) units within the matched sample (instead of the entire matched sample). In practice, choosing among the three approaches is less important than whether or not to use a bias-adjustment method at all.

We include standard errors for each ATE estimator. The doubly-robust estimator has a simple closed-form formula (Lunceford and Davidian, 2004). Both across-cluster matching and mixed matching regression have the same Abadie-Imbens standard errors, calculated from the built-in option of package \texttt{Matching}. We use bootstrapping to get the standard errors of other methods, with 75 bootstrap samples. We present in Figure 2.3 a forest plot of the SDf for the baseline covariates as well as their SDfs after weighting and matching, with the propensity scores estimated from a fixed effects model. The ATE estimators and their standard errors for the STS data are presented in Table 2.4.
Figure 2.3: Standardized differences for observed covariates at baseline as well as after matching and weighting. Propensity scores are estimated from a fixed effects model.
Table 2.4: Estimators for ATE and their SE ($\times 10^{-2}$) in the endoscopic data, by year and overall.

<table>
<thead>
<tr>
<th>Year</th>
<th>Data $N_1/N$</th>
<th>Model</th>
<th>model-free(^a)</th>
<th>mixed with regression</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>across matching</td>
<td>matching</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>marginal weighting</td>
<td>weighting (DR)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>cluster weighting</td>
<td>pure regression</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>stratification</td>
<td></td>
</tr>
<tr>
<td>2008</td>
<td>11208 / 15182 fixed</td>
<td>-1.35 (1.11)</td>
<td>-1.52 (0.75)</td>
<td>-1.00 (0.77)</td>
</tr>
<tr>
<td></td>
<td>$N_{clus} = 249$ random</td>
<td>-1.00 (1.11)</td>
<td>-1.67 (0.71)</td>
<td>-1.55 (0.73)</td>
</tr>
<tr>
<td>2009</td>
<td>8357 / 11983 fixed</td>
<td>-1.17 (0.99)</td>
<td>-0.90 (0.87)</td>
<td>-1.37 (0.86)</td>
</tr>
<tr>
<td></td>
<td>$N_{clus} = 213$ random</td>
<td>-1.27 (1.02)</td>
<td>-0.52 (0.81)</td>
<td>-1.16 (0.83)</td>
</tr>
<tr>
<td>2010</td>
<td>6983 / 10834 fixed</td>
<td>0.40 (1.17)</td>
<td>-0.41 (0.72)</td>
<td>-0.29 (0.71)</td>
</tr>
<tr>
<td></td>
<td>$N_{clus} = 199$ random</td>
<td>-0.19 (1.15)</td>
<td>-0.04 (0.67)</td>
<td>-0.13 (0.69)</td>
</tr>
<tr>
<td>2011</td>
<td>2037 / 3178 fixed</td>
<td>1.69 (1.70)</td>
<td>0.18 (1.73)</td>
<td>0.18 (1.70)</td>
</tr>
<tr>
<td></td>
<td>$N_{clus} = 98$ random</td>
<td>0.25 (1.72)</td>
<td>0.07 (1.51)</td>
<td>0.26 (1.58)</td>
</tr>
<tr>
<td>Overall</td>
<td>49568 / 66779 fixed</td>
<td>-0.71 (0.58)</td>
<td>-0.26 (0.42)</td>
<td>-1.23 (0.42)</td>
</tr>
<tr>
<td></td>
<td>$N_{clus} = 394$ random</td>
<td>-0.49 (0.58)</td>
<td>-0.43 (0.43)</td>
<td>-1.43 (0.43)</td>
</tr>
</tbody>
</table>

\(^a\) “Model-free” means direct comparison of observed outcomes based on estimated propensity score, without additional regression adjustment.

For the yearly estimates, all clusters with less than 4 controls and 4 treatments are removed. For the overall estimates, all clusters with less than 6 controls and 6 treatments are removed. The fixed effect and random effect models are fitted with all observed individual and cluster-level covariates. $N$ denotes the size of the subset of data, $N_1$ the number of endoscopic cases, and $N_{clus}$ the number of clusters.
For the most part, the point estimators are negative, indicating a lower 3-year mortality rate in patients receiving endoscoping vein-graft harvesting technique compared to those receiving the traditional open technique. Year 2011 is the exception, when the majority of methods give positive estimators. However, this year only has data collected for the first half of the year instead of a full 12-months, so the results from this year’s data might be less reliable. Despite the negative point estimators, the standard errors calculated from either the close-form formula or bootstrapping are quite large, making the 90\% confidence intervals for these estimators include zero and thus the treatment effect on the outcome is not significant. We plot the point estimators and their CIs (calculated by \( \hat{\tau} \pm 2 \times se(\hat{\tau}) \)) in Figure 2.4.
Figure 2.4: STS Data: Estimators and error bars of different ATE estimators, for each year and the combined data.
2.8 Conclusion

Propensity score methods are widely used to achieve covariate balance between comparison groups in both causal and unconfounded descriptive studies. Recently propensity score methods have been increasingly applied to multilevel data. Though the consensus in the literature is that the multilevel structure must be properly adjusted for in at least one and preferably both stages of propensity score analysis, there has been little research on the comparative performance between different propensity score methods in realistic multilevel settings. In this paper, we fill this gap by providing a critical review and extensive discussion of several common classes of propensity score methods in the context of multilevel data. We demonstrate these methods using a real data set about endoscopic vein-graft harvesting techniques and provided the related software.

Though comparisons between methods heavily depend on estimand and data structure, there are several consistent patterns emerged, based on which we give the following practical guides for using propensity score methods with multilevel data in real applications. First, before applying any method to the data, it is prudent to check the overall and cluster-specific balance of covariates and proportions of treated. Second, in general, we recommend to use mixed methods (mixed matching-regression or DR) with multilevel regression models, which provide robust and efficient inferences. This is especially true when previous work has been done to build a reliable outcome model, similar to our case. Third, when using a model-free method, even with the propensity score estimated from multilevel models, it appears necessary to adjust for the clustering structure in the second (estimation) stage of the analysis, e.g., using preferential matching or cluster weighting. Fourth, when there may be unmeasured cluster-level confounders, we suggest to use fixed effects models; when there are a large number of (small) clusters, we suggest to use random effects models. Finally, one may apply more than one method (e.g., matching and weighting, or mixed matching-regression and DR) to the data and compare the results.

In this chapter we did not consider misspecification of the functional form of the
propensity score model. Simulations in unstructured data suggest that semipara-
metric specification based on power series improves robustness against deficiencies of
the propensity score model (Mercatanti and Li, 2014); looking into the performance
of such models when the data are structured is an interesting topic to investigate.
Sensitivity Analysis in an Observational Study with Count Outcome

3.1 Introduction

For any project whose goal is to measure a causal effect of an intervention or treatment, randomized experiments are generally considered the gold standard. Under this scheme, each subject is assigned treatment randomly based on a randomized process; therefore ensures that the distribution of background characteristics of the two treatment groups are as similar as possible, differing only by chance. However, under several circumstances and due to ethical or practical reasons, treatment cannot be assigned randomly and researchers need to work with observational studies. In observational studies, the treatment and control groups are not comparable at all in their baseline characteristics - this is called confounding. There have been several methods to control for observed confounders, such as propensity score methods (matching, weighting, stratification) or regression-based methods. However, when there are unobserved confounders, observations in the two groups that seem comparable might not indeed be so. Therefore, it is important and crucial that researchers
perform sensitivity analysis to see how robust the effect estimate is under the presence of potential unobserved confounders, which would in turn undoubtedly help the decision making process.

Several methods have been proposed to carry out sensitivity analysis. Among the first ones is a method introduced by Cornfield et al. (1959), where the authors looked into the relationship between cigarette smoking and lung cancer. Cornfield et al. showed that if there were an unmeasured confounder, the exposure to the unmeasured confounder needs to be considerably higher among smokers than non-smokers – specifically more than nine-fold – in order to completely account for the apparent effect of smoking on lung cancer. However, this paper only adjusted for a single unmeasured binary covariate $U$, completely ignoring the measured covariates.

Rosenbaum and colleagues (Rosenbaum and Rubin (1983b), Gastwirth et al. (1998)) proposed a class of methods whose core idea is to first balance the observed covariates (for example, with the estimated propensity scores) and then conduct sensitivity analysis on one single binary unmeasured confounder. Rosenbaum and Rubin (1983b) proposed an approach based on subclasses and maximum-likelihood estimation. Gastwirth et al. (1998) extended this to matched pairs by adjusting the strength of the relationships between $U$ and the treatment assignment as well as $U$ and the outcome, finding values of sensitivity parameters that can drive the true treatment effect to nonsignificant. The authors proposed three methods: primal, dual, and simultaneous, corresponding to varying the association between the unmeasured confounder and the treatment, the unmeasured confounder and the outcome, or both. A modified McNemar’s test was used to compute the upper-bound p-value, and from here determine when this upper bound would be $> 0.05$ (or a pre-selected threshold). This method is narrowed to only matched pairs - it is unclear how to implement this for $k - 1$ matching sample where $k > 1$.

VanderWeele and Arah (2011) derived a general formula to calculate the bias,
accommodating continuous or categorical treatment assignments as well as outcomes. For each level \( j \) of the observed covariates, one needs to specify (1) the relationship the outcome \( Y \) and \( U \) for each treatment level: 
\[
\mathbb{E}(Y \mid Z = 1, U = u, X = j) - \mathbb{E}(Y \mid Z = 1, U = u', X = j)
\]
\[
\mathbb{E}(Y \mid Z = 0, U = u, X = j) - \mathbb{E}(Y \mid Z = 0, U = u', X = j)
\]
where \( u' \) is a reference value, and (2) the distribution of \( U \) conditional on each treatment level compared with the overall distribution of \( U \): 
\[
P(U \mid Z = 1, X = j) - P(U \mid X = j).
\]
Despite its ability to accommodate several outcome, treatment and unmeasured confounder types, this approach requires the specification of many sensitivity parameters. The authors gave suggestions for simplification under specific settings, see VanderWeele and Arah (2011) for the full formula and details.

Ichino et al. (2008) approached sensitivity analysis from a different angle by simulating the unmeasured confounder \( U \) for each observation based on pre-specified values that determine the distribution of a binary \( U \) conditional on the treatment assignment and the outcome. This simulated \( U \) is then included in the set of matching covariates used to compute the propensity scores and/or estimate the outcome. The process can be repeated several times and the estimand of interest is obtained by averaging over the distribution of the simulated \( U \). This approach was presented for binary unmeasured confounder, treatment, and outcome, and the authors demonstrated this using a data set to estimate the effect of temporary work agency assignment on the probability of finding a stable job in two Italian regions.

Harding (2009) proposed a method which involved estimating the omitted variable bias in an ordinary least square regression setting. Harding applied this to a data set about the effect of neighborhood violence on two adolescent outcome: high-school graduation rate and teenage pregnancy. Note that here the “treatment” is not binary but a continuous “violence index”. Both the treatment and the unmeasured confounder are standardized (variance =1). The true treatment effect is calculated for hypothetical values of \( r_{ZU} \) (correlation between \( Z \) and \( U \)) and \( \beta_{UY} \) (the coeffi-
cient representing the effect of $U$ on $Y$ in a linear regression model). Clearly this approach depends on an explicit specification for the outcome model, controlled for all observed covariates and $U$. Imbens (2003) followed the MLE-based method of Rosenbaum and Rubin (1983b), but converted the sensitivity parameters into a partial $R^2$ to make interpretation easier. The authors constructed partial $R^2$ such that the implied average treatment effect is changed by a pre-fixed amount. The estimated treatment effect is considered robust if the set of $R^2$ don’t include reasonable values.

Ding and VanderWeele (2016); VanderWeele and Ding (2017) introduced a new bounding factor without having to make the typical assumptions (having only one binary confounder, or no interactions between the treatment and the unmeasured confounder). This method only requires specification of two sensitivity parameters, $RR_{UD}$ for the effect of $U$ on the outcome and $RR_{EU}$ the effect of $U$ on the treatment. Given these values, one can calculate the maximum relative amount that a (or a set) of unmeasured confounder can reduce an observed risk ratio as well as its confidence interval. This can reveal whether unmeasured confounder(s) of this strength can completely determine the treatment effect estimate. The authors extended this idea to instead computing the E-value, which is “the minimum strength of association, on the risk ratio scale, that an unmeasured confounder would need to have with both the treatment and outcome, conditional on the measured covariates, to explain away a treatment - outcome association” (VanderWeele and Ding, 2017). Liu et al. (2013) briefly gave an overview of several other sensitivity analysis methods.

This paper aims to extend two previous works by Rosenbaum and Rubin (1983b) and Ichino et al. (2008), modifying their approaches to accommodate count outcomes (instead of binary). With a count outcome, we need to specify additional sensitivity parameters and they would have different interpretations - this is the contribution this chapter makes. In both approaches, the propensity score is used as part of
the estimating process (and not of the sensitivity analysis itself). The first approach uses the propensity scores to stratify observations in subclasses, while the second one builds a matched sample based on the estimated propensity scores. We illustrated these methods using a real study on the application of rumble strips in Pennsylvania. In traffic safety research, a central goal is to evaluate the effectiveness of traffic safety programs and countermeasures. For example, one commonly used low-cost safety strategy is the application of shoulder rumble strips (SRS) and centerline rumble strips (CLRS) in combination. This application intends to reduce crash frequency by alerting drivers when they are about to depart from the travelled lane. There have been studies on the effectiveness of SRS or CLRS being used in isolation, but no research has been published on the safety effectiveness of the combined treatment.

3.2 Methods

3.2.1 Notations and settings

In our setting, we have a binary treatment status, a set of observed covariates, a binary unmeasured confounder, and a count outcome. There are $N$ observations total, with $X_i$ be the set of observed covariates, and $Z_i$ be the treatment indicator (=1 if the observation was treated, = 0 if the observation was a control). The $i^{th}$ observation have two potential outcomes: $Y_i(1)$ and $Y_i(0)$, corresponding to the potential outcomes that would have resulted if it had received treatment status $Z_i = 1$ and $Z_i = 0$, respectively. We only observe the potential outcome corresponding to the observed treatment condition, $Y_i = Y_i(1)Z_i + Y_i(0)(1 - Z_i)$, and the other is missing - this is the fundamental problem of causal inference. We make the assumption that given only the observed covariates, the treatment assignment is not unconfounded, but this would hold conditional on $X$ and $U$, that is,

$$Y(1), Y(0) \perp\!
\perp Z \mid X$$
is untrue, but this is true:

\[ Y(1), Y(0) \perp Z \mid X, U. \]

We also assume \( 0 < P(Z = 1 \mid X, U) < 1 \) (the overlap assumption). These jointly
together are called the strong ignorability condition.

**Treatment effect estimands.** A commonly used estimand is the *average treatment effect* (ATE). Treatment effects can either be additive or multiplicative. An
additive average treatment effect is the difference between the potential treated and the potential control outcome, \( \mathbb{E}(Y(1) - Y(0)) \). A multiplicative effect is the ratio between the potentially treated’s and potentially control’s outcomes, \( \mathbb{E}[Y(1)]/\mathbb{E}[Y(0)] \).

Another estimand that is usually desirable is the *average treatment effect for the treated* (ATT). Without randomization, ATT is generally different from ATE.

In traffic safety, it is the norm that people use the multiplicative treatment effect. For example, crash modification factors (CMFs) are multiplicative effects commonly used in the safety literature to compute the expected number of crashes after implementing a given countermeasure at a specific site (Gross et al., 2010); and this is the quantity we are interested in. It can be thought of as an index of how much crash counts are expected to change after applying the countermeasure. A CMF greater (less) than 1 indicates an expected increase (decrease) in crash frequency after implementing the safety countermeasure of interest. The traffic research literature is unclear on whether the CMF is an ATE or an ATT. However, often the intervention is given to sites with high crash counts. In order to estimate the treatment effect, reference sites would be picked subsequently to match characteristics of the sites with treatment. Therefore, we argue that the focus of the treatment effect is on the reduction in crash count at the treated sites, and thus CMF is essentially a
multiplicative ATT. We define the CMF as:

$$\tau = \frac{E[Y(1) \mid Z = 1]}{E[Y(0) \mid Z = 1]} = \frac{\tau_1}{\tau_0}$$

It should be noted that this is only one possible definition for the CMF. Depending on the purpose of the CMF, other definitions might be desirable, e.g. using expectation of the ratio instead of ratio of expectations, allowing heterogeneous CMF. This is beyond the scope of this chapter.

An outcome model might be required for certain estimation methods. Since the outcome of interest belongs to the count data family, the outcome model should be chosen accordingly. Two popular count data models are Poisson and negative binomial regression models. We choose to use a negative binomial regression since this is the most commonly used method in the transportation safety literature. The negative binomial distribution is similar to the Poisson distribution, but it allows for overdispersion. The probability mass function is:

$$P(Y = y) = \frac{\Gamma(y + \theta)}{y!\Gamma(\theta)} \cdot \left(\frac{\theta}{\mu + \theta}\right)^{\theta} \cdot \left(\frac{\mu}{\mu + \theta}\right)^y,$$

with $\mu$ be the mean of the distribution and $\theta$ the overdispersion parameter. The mean of the outcome can be modelled to relate to the independent variables via a log link:

$$\mu = \exp(X\beta).$$

For each observation, the estimated propensity score is needed in the analysis. This is the probability of the observation being treated, conditional on the observed covariates:

$$e(X) = Pr(Z = 1 \mid X).$$
This can be estimated from any commonly used method for dichotomous dependent variable, such as logistic or probit regression. Typically, this model would include all the observed covariates without nonlinear or interaction terms, especially when the number of covariates is large.

The general idea of this type of sensitivity analysis is to first specify some sensitivity parameters governing the association of $U$ and $Z$ as well as $U$ and $Y$. For each set of these sensitivity parameters, we estimate the estimand of interest (here, the CMF). If for a range of sensitivity parameters, the estimated CMF remains similar to the baseline estimate and the conclusion isn’t changed, we can conclude that the estimated CMF is robust to violation of unconfoundedness.

**Estimation methods.** There are several propensity-score methods that have been developed to estimate the average treatment effect, such as matching, weighting, subclassification, bias-corrected matching estimator (combining propensity scores and regression), and doubly-robust estimators. See Stuart (2010) and Imbens (2004), for example, for a review of these methods. We use two different estimators for the two sensitivity analyses and will discuss each of them in the corresponding subsection.

### 3.2.2 Sensitivity Analysis 1: MLE-based

This approach is an extension from Rosenbaum and Rubin (1983b). For this approach, each observation is stratified into one of $J$ subclasses. The stratification can be done with the observed covariates (if there are only a few), or with the estimated propensity scores in the treated group. We make an assumption of strong ignorability conditional on not just the subclass $s$ but also the unobserved covariate $u$; that is:

$$(Y_1, Y_0) \indep Z | s, u,$$

and that $0 < P(Z = 1|s, u) < 1$.

**Estimation method: subclassification.** After the propensity scores $\hat{e}(X)$ are
estimated, form \( J \) subclasses based on the sample quantiles of \( \hat{e}(X) \). Let \( S \) be the variable indicating the subclass. The estimator for the CMF is:

\[
\hat{\tau}_{\text{sub}} = \frac{\hat{\tau}_1}{\hat{\tau}_0} = \frac{\sum_{s=1}^{J} \phi_{s1} \cdot \mathbb{E}(Y(1) \mid S = s, X)}{\sum_{s=1}^{J} \phi_{s1} \cdot \mathbb{E}(Y(0) \mid S = s, X)},
\]

where \( \phi_{s1} \) represent the relative proportion of treated sites by subclass.

**Framework for the sensitivity analysis.** The joint distribution of \((Y, Z, U, S)\) can be written as:

\[
P(Y, Z, U, S) = P(Y \mid U, S, Z) \cdot P(Z \mid U, S) \cdot P(U \mid S) \cdot P(S)
= P(Y(z) \mid U, S) \cdot (Z \mid U, S) \cdot P(U \mid S) \cdot P(S). \tag{3.1}
\]

Where we have:

\[
P(S = s) = \phi_j, \quad \sum_{j=1}^{J} \phi_j = 1,
\]

and within each subclass \( s \), we have:

- Unmeasured confounder \( U \sim \text{Ber}(1 - \pi) \).
- Treatment model: \( \text{logit}[P(Z = 1 \mid U = u)] = \gamma + \alpha \cdot u \).
- Outcome model: a negative binomial with mean \( \mu \) and overdispersion parameter \( \theta \) such that:

\[
\mu_{zu} = \exp(\beta_z + \delta_z u)
\]

\[
\theta_{z1} = c_z \cdot \theta_{z0}
\]

Let’s consider the interpretation of the aforementioned parameters; \( \phi_s \) is the probabilities of being in subclass \( s \); \( \pi_s \) is the probability that the unmeasured confounder \( U \) is zero in subclass \( s \); \( \gamma_s \) is the log odds of assigning to the treatment \( Z = 1 \) for
subclass $s$ when there is no unmeasured confounder, and $\exp(\alpha)$ is the multiplicative
effect of $U = 1$ vs $U = 0$ on the treatment assignment. Similarly, $\beta_{sz}$ is the log of the
expected crash counts in subclass $s$ for treatment group $z$ when $U = 0$, and $\exp(\delta_0 s)\exp(\delta_1 s)$
is the multiplicative effect of $U = 1$ vs $U = 0$ on the mean crash count of
the sites without (and with) CLRS/ SRS applications in subclass $s$. We can make
the assumption that $(\pi, \alpha, \delta_0, \delta_1)$ are constant across subclasses for parsimonious rea-
sons, and these are our sensitivity parameters. Since $U$ is not observed, the joint
distribution in equation 3.1 cannot be calculated directly. However, the treatment
effect can still be estimated by fixing $(\pi, \alpha, \delta_0, \delta_1)$, integrating out $U$ and maximize
the observed likelihood to find the MLEs for the rest of the parameters.

We are interested in the CMF:

$$
\tau = \frac{\mathbb{E}[Y(1) \mid Z = 1]}{\mathbb{E}[Y(0) \mid Z = 1]}
= \frac{\tau_1}{\tau_0},
$$

where

$$
\tau_z = E(Y(z) \mid Z = 1) = \sum_S P(S = s \mid Z = 1)E(Y(z) \mid S = s, Z = 1)
= \sum_S P(S = s \mid Z = 1) \sum_{k=0,1} E(Y(z) \mid U, S) \cdot P(U = k \mid S = s, Z = 1)
= \sum_S \phi_{s1} [w_{s1} \cdot \exp(\beta_{sz}) + (1 - w_{s1}) \cdot \exp(\beta_{sz} + \delta_{sz})].
$$

with $\phi_{s1}$ being the proportion of treated sites in subclass $s$, and $w_{sz} = P(U = 0 \mid
S = s, Z = z)$ having the expression:
\[ P(U_i = 0|Z_i, S_i) = w_s \]

\[ = \frac{P(Z_i|U_i = 0, S_i) \cdot P(U_i = 0|S_i)}{P(Z_i|U_i = 0, S_i) \cdot P(U_i = 0|S_i) + P(Z_i|U_i = 1, S_i) \cdot P(U_i = 1|S_i)} \]

\[ = \frac{\pi_s \exp(z \gamma_s)}{1 + \exp(\gamma_s)} \left[ \frac{\pi_s \exp(z \gamma_s)}{1 + \exp(\gamma_s)} + \frac{(1 - \pi_s) \exp[z(\gamma_s + \alpha_s)]}{1 + \exp(\gamma_s + \alpha_s)} \right]^{-1} \]

\[ = \pi_s \left[ \frac{\pi_s + (1 - \pi_s) \exp(z \alpha_s)[1 + \exp(\gamma_s)]}{1 + \exp(\gamma_s + \alpha_s)} \right]^{-1} \quad (3.3) \]

**Maximum Likelihood Estimation.** We briefly walk through the estimation procedure proposed by Rosenbaum and Rubin (1983b), with necessary changes in the outcome model. The likelihood of the parameters given the observed data \((Y_i, Z_i, S_i)\), \(i = 1, \ldots, N\) is:

\[
\prod_{i=1}^{N} P(Y_i(z) = y|Z_i, S_i) \cdot P(Z_i|S_i) \cdot P(S_i),
\]

where:

\[
P(S_i = s) = \phi_s
\]

\[
P(Z_i = 0|S_i) = P(Z_i|S_i, U_i = 0) \cdot P(U_i = 0|S_i) + P(Z_i|S_i, U_i = 1) \cdot P(U_i = 1|S_i)
\]

\[= \frac{\pi_s}{1 + \exp(\gamma_s)} + \frac{1 - \pi_s}{1 + \exp(\gamma_s + \alpha_s)} \quad (3.4)\]

\[
P(Y_i(z)|Z, S) = P(Y_i(z)|Z_i, S_i, U_i = 0) \cdot P(U_i = 0|Z_i, S_i)
\]

\[+ P(Y_i(z)|Z_i, S_i, U_i = 1) \cdot P(U_i = 1|Z_i, S_i) \]

\[= \frac{\Gamma(y + \theta_{s20})}{y!\Gamma(\theta_{s20})} \left( \frac{\theta_{s20}}{\mu_{s20} + \theta_{s20}} \right)^{\theta_{s20}} \left( \frac{\mu_{s20}}{\mu_{s20} + \theta_{s20}} \right)^{y} \cdot P(U_i = 0|Z_i, S_i) \]

\[+ \frac{\Gamma(y + \theta_{s21})}{y!\Gamma(\theta_{s21})} \left( \frac{\theta_{s21}}{\mu_{s21} + \theta_{s21}} \right)^{\theta_{s21}} \left( \frac{\mu_{s21}}{\mu_{s21} + \theta_{s21}} \right)^{y} \cdot P(U_i = 1|Z_i, S_i) \quad (3.5)\]
with \( \mu_{szu} = \exp(\beta_{sz} + \delta_z u) \), \( \theta_{sz1} = c_z \cdot \theta_{sz0} \), and the expression for \( P(U_i = 0|Z_i, S_i) \) specified in equation 3.3.

The procedure can be listed as follows.

1. Fix values of sensitivity parameters \( \pi_s, \alpha_s, \delta_{s1}, \delta_{s0} \).

2. Fit a suitable model to estimate the propensity scores such as a logistic regression model while controlling for relevant covariates. The observations are stratified into \( J \) subclasses based on their estimated propensity scores.

3. For each subclass, estimate \( P(S = s) = \phi_s \) by the proportion of sites in subclass \( s \), and estimate \( P(Z = 0|S) = \phi_{s1} \) by the observed proportion of reference sites in subclass \( s \).

4. For fixed \( \pi_s, \alpha_s \), let \( \omega_s = \exp(\gamma_s) \). Equation 3.4 can be rewritten as a quadratic equation in \( \omega_s \), which can be solved for \( \hat{\gamma}_s \).

5. Calculate \( w_{sz} = P(U = 0|Z, S) \) by equation 3.3.

6. Fit a negative binomial model to the outcome to estimate the overdispersion parameters \( \theta_z \) and use these as the initial values for \( \theta_{sz0} \).

7. Find the values of \( \beta_{sz} \) which maximize:

\[
\prod_{i=1}^{N} P(Y_i = y|Z_i, S_i) = \prod_{i=1}^{N} \left[ \frac{\Gamma(y_i + \theta_{sz0})}{y_i! \Gamma(\theta_{sz0})} \left( \frac{\theta_{sz0}}{\exp(\beta_{sz}) + \theta_{sz0}} \right)^{\theta_{sz0}} \left( \frac{\exp(\beta_{sz})}{\exp(\beta_{sz}) + \theta_{sz0}} \right)^{y_i} \cdot w_{sz} \right. \\
+ \left. \frac{\Gamma(y + c\theta_{sz0})}{y! \Gamma(c\theta_{sz0})} \left( \frac{c\theta_{sz0}}{\exp(\beta_{sz} + \delta_z) + c\theta_{sz0}} \right)^{c\theta_{sz0}} \left( \frac{\exp(\beta_{sz} + \delta_z)}{\exp(\beta_{sz} + \delta_z) + c\theta_{sz0}} \right)^{y_i} \cdot (1 - w_{sz}) \right],
\]

(3.6)
using any commonly used optimization methods in R, such as package `optim`. Finally, \( \hat{\tau}_1 \) and \( \hat{\tau}_0 \) can be calculated using equation 3.2, and from here the ATT CMF can be estimated by taking the ratio of \( \hat{\tau}_1 \) and \( \hat{\tau}_0 \).

### 3.2.3 Sensitivity analysis 2: simulation-based

In this section, we introduce a different approach to sensitivity analysis which does not require any parametric model for the outcome, nor does it utilize maximum likelihood estimation. Instead, we choose a range of parameters that define the distribution of the confounding factor \( U \), used this to generate a value of \( U \) for each observation and estimated the estimand of interest using our method of choice, including \( U \) in the observed covariates. This approach was proposed by Ichino et al. (2008) but the authors only considered the case of binary potential outcomes. In this section we will extend it to a scenario with count outcomes.

This approach has several advantages compared to the MLE-based approach as discussed in Ichino et al. (2008). First, it is more flexible in terms of the method used to obtain the estimand of interest. After generating the confounding factor \( U \), any algorithm can be used to estimate the ATT CMF, as long as \( U \) is included as an additional covariate. Second, \( U \) can be made to have similar distribution to that of observed binary covariates, revealing the robustness of the chosen algorithm when we fail to observe a factor that is similar in distribution to one of the observed covariates. Finally, the relationship between \( U \), \( Y \) and \( Z \) are defined by the proportion of \( U \) conditional on \( Y \) and \( Z \), bypassing the need to assume a model for the outcome conditional on \( U \), \( Z \), and \( X \).

First we perform a transformation to convert the count outcome \( Y \) to an outcome \( K \) with three values: \( K_i = Y_i \) if \( Y_i < 2 \) and \( K_i = 2 \) for \( Y_i \geq 2 \). This transformation is suitable for our data, where we have 9505 sites with no crash, 2939 sites with exactly one crash, and 1171 sites with two or more crashes. Therefore this transformation
still leaves an adequate number of sites in each category of $K$. See Table 3.6 for a proportion breakdown of site counts by $K$ and $Z$. Consideration should be given to other types of data to achieve an appropriate transformation.

We make an assumption of unconfoundedness conditional on both the observed covariates and an unobserved binary covariate $U$:

$$K(1), K(0) \perp Z \mid X, U$$

(3.7)

and also that the unobserved confounder is completely independent of the observed covariates, conditional on the treatment and outcome. We then specified the following parameters to define the distribution of $U$:

$$p_{ij} = P(U = 1 \mid Z = i, K = j, X) = P(U = 1 \mid Z = i, K = j),$$

(3.8)

$i \in \{0, 1\}, j \in \{0, 1, 2\}$. These parameters $p_{ij}$ in turn would determine the proportion of sites with $U = 1$ for each treatment group:

$$p_i \equiv P(U = 1 \mid Z = i) = \sum_{j=0}^{2} p_{ij} \cdot P(K = j \mid Z = i).$$

Therefore, we can create situations where there are more sites with $U = 1$ in the treated sites vs. the reference sites and vice versa by adjusting $p_{ij}$. The key idea of this sensitivity analysis is that setting $p_{00} > p_{01} > p_{02}$ would imply:

$$\mathbb{E}(K \mid Z = 0, U = 1, X) < \mathbb{E}(K \mid Z = 0, U = 0, X),$$

and with the assumption from equation 3.7, we have:

$$\mathbb{E}(K(0) \mid Z = 0, U = 1, X) < \mathbb{E}(K(0) \mid Z = 0, U = 0, X).$$

For a full proof, see Appendix A. The same reasoning can be applied to the setting $p_{1} > p_{0}$. The relationship among the sensitivity parameters can be specified to
focus on the “threat” of interest to the estimate. For example, the assumptions $p_{00} > p_{01} > p_{02}$ and $p_1 > p_0$ correspond to confounders that have a negative effect on the potential outcomes and a positive effect on the treatment assignment probability when no treatment is applied. This means reference sites with $U = 1$ would have a lower total crash count on average. This setting is appropriate if we are interested in confounders that without observing them, we might incorrectly obtain a significant CMF estimate less than 1 even when there is no true effect of the treatment on the potential outcomes - the “false positive” case.

On the other hand, if we are concerned about “false negatives”, focus should be put on confounders with a positive effect on both the treatment assignment probability and the potential outcomes when no treatment is applied. This translates to $p_{00} < p_{01} < p_{02}$ and $p_1 > p_0$. An example of this type of confounder could be “bad weather” - $U = 1$ if the site is in an area of frequent bad weather, and $U = 0$ otherwise. Sites at bad weather areas are probably more likely to be given CLRS/SRS, at the same time this factor can certainly increase crash counts at reference sites. Without controlling for this factor, it is possible that we would come to a CMF estimate of near 1 when there is indeed a true, significant treatment effect.

The effect of $U$ on the untreated outcome and the treatment assignment can be quantified conditional on the observed covariates by calculating the “outcome effect” of $U$:

$$
\Gamma = \frac{\mathbb{E}(K \mid Z = 0, U = 1, X)}{\mathbb{E}(K \mid Z = 0, U = 0, X)},
$$

and the “treatment selection effect” of $U$:

$$
\Lambda = \frac{P(Z = 1 \mid U = 1, X)}{P(Z = 1 \mid U = 0, X)}.
$$

For simplicity, we assume a constant difference: $d = p_{00} - p_{01} = p_{01} - p_{02}$, and also let $s = p_1 - p_0$. In our first scenario concerning “false positives”, our assumption of
$d > 0$ and $s > 0$ would entail a decreasing outcome effect ($\Gamma < 1$) and an increasing treatment effect ($\Lambda > 1$). For the second scenario of “false negatives”, we would need to assume $d < 0$ and $s > 0$, which would lead to $\Gamma > 1$ and $\Lambda > 1$.

**Estimation method: Propensity score matching and regression.** This was proposed by Wood et al. (2015). After the propensity scores are estimated, some matching method can be used with the estimated propensity scores as the distance metric, searching for $k$ reference sites for each treated site. Guo and Fraser (2014) discussed several matching methods as well as their limitations. Note that selecting the number of matches is essentially a bias-variance tradeoff. The more reference sites are matched with each treatment sites, the larger the bias tends to be since the later matches are by definition not as close to the treated site as the first match. However, the variance will decrease because of the larger number of matched units. Covariate balance is checked by the absolute standardized difference of each covariate. If severe imbalance is detected for some covariates, then the propensity score model is refit with interaction and/or higher order terms of these covariates. After the match sample has been obtained, an appropriate outcome regression model can be fitted on the matched data, including the treatment status indicator. For example, in our crash count data we can use a negative binomial model as follows:

$$\mu_i = \exp(\beta_0 + \beta_z Z_i + X_i \beta),$$

where $\mu_i$ is the expected crash count, $Z$ is the treatment status and $X$ is a vector of relevant covariates for road segment (observation) $i$. The matching-based ATT CMF estimator is obtained by:

$$\hat{\tau}_{match} = \exp(\hat{\beta}_z).$$

The 95% confidence interval is estimated by $\exp(\hat{\beta}_z \pm 1.96\hat{\sigma}_z)$, where $\hat{\beta}_z$ is the estimated coefficient for the treatment variable, and $\hat{\sigma}_z$ its estimated standard error. Note that this is not the same as the typical approach of combining matching and re-
gression commonly used in causal inference, proposed by Abadie and Imbens (2011). Abadie and Imbens’ method estimates the individual average potential outcome, then obtain the matched sample, and within the matched pairs, the difference is adjusted for the difference in covariate values.

3.3 Application to Pennsylvania rumble strips treatment on rural highways

3.3.1 Data

We base our illustration on a data set that was part of a research study about the safety evaluation of centerline rumble strips (CLRS) and shoulder rumble strips (SRS) applications. CLRS/SRS is one commonly used low-cost safety strategy that intends to reduce crash frequency by alerting drivers when they are about to depart from the travel lane. This particular dataset focuses on sites in Pennsylvania that contained both CLRS and SRS. The Pennsylvania Department of Transportation (PennDOT) provided a total of 218 miles where both CLRS and SRS were installed. The reference group includes two-lane rural highway segments without rumble strips. We narrow the reference sites using the following characteristics:

- No access control.
- Divisor (none, painted divided, man-made barrier, earth divided).
- Divided width equals to 0 ft.
- Speed limit ranged from 20- 55mph.
- Having two lanes (one in each direction).
- Documented average annual daily traffic (AADT) volume data.

From 17,891 two-lane rural highway miles in Pennsylvania, this process narrows down to 17,931 miles. The data for treatment and reference sites have 466 and 39,360
segments, respectively. We further remove all sites with missing values for covariates, and our final data set include 334 treatment sites as well as 13,286 reference sites. In the final analysis database, the variables that were available included the following:

- **Roadway data**: surface type, pavement width, speed limit, number of lanes, resurfacing year, shoulder type, shoulder width, area type (urban/rural).
- **Traffic data**: AADT from 2003 - 2011 and percentage of trucks in the traffic stream.
- **Crash data**: Data from 2003 to 2012 were obtained for five types of crashes, excluding intersection-related and animal-related crashes.
  - Total: identified as a midblock crash and not “deer” or “animal”.
  - Fatal plus injury: if number of fatal or injured persons is greater than zero.
  - Run-off-road: crash occurred outside the trafficway in an area not intended for vehicles.
  - Head-on: opposite direction collision type.
  - Sideswipe-opposite-direction: collision type.

Our outcome of interest here is the total crash. Table 3.1 provides mean and standard deviation of several site characteristic covariates for both treatment and reference sites.

AADT values are only available from 2003 to 2011, but crash counts measurements are available through 2012. We choose to use the average of the previous three years as an estimate for AADT2012. Different sites have different starting years for the after period (2010, 2011, or 2012). For simplicity, we only considered data for the
year 2012. The propensity scores in the analysis were estimated by fitting a logistic regression model using this subset of data:

$$\text{logit}(p) = \mathbf{X}\beta,$$

where $p$ is the probability of receiving the treatment conditioned on the covariates $\mathbf{X}$, and $\mathbf{X}$ includes a subset of the observed covariates (see Table 3.2). Histograms of the estimated propensity scores by treatment group estimated from the year 2012 of the data are displayed in Figure 3.1. Note that this data set actually shows high balance between the covariates distributions of the two treatment groups, which is not usually observed in observational studies.

Before showing the results of the two sensitivity analysis approaches, we present a baseline estimated CMF obtained by subclassification as well as propensity score
matching combined with regression (Wood et al., 2015). Table 3.3 presents the estimates as well as the 95% confidence intervals for the ATT CMF of total crash. For subclassification, we use 5 subclasses based on the quantiles of the estimated propensity score of the treated sites. The 95% confidence interval is obtained from the bootstrap quantiles, with 300 bootstrap samples generated. For the other method, we used 1-1 nearest neighbor propensity score matching to obtain the matched sample before performing regression. Subclassification gives a lower estimate and a narrower confidence interval compared to propensity score matching. Both estimates are less than 1, however both confidence intervals include the value of 1. Therefore, we cannot conclude that applying the combination of centerline and shoulder rumble strips significantly reduces total crash counts at a 95% confidence level. Our question is, can this insignificance be due to some unmeasured confounder, that would have led to a significant reduction in crash counts had it been observed?

Table 3.3: CMF baseline estimates and their 95% confidence intervals for total crash count.

<table>
<thead>
<tr>
<th>Method</th>
<th>CMF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subclassification</td>
<td>Est 0.85 CI (0.70,1.06)</td>
</tr>
<tr>
<td>Pscore 1-1</td>
<td>Est 0.93 CI (0.70,1.20)</td>
</tr>
</tbody>
</table>

Subclassification: ATT CMF estimation with 5 subclasses. CI was obtained by bootstrapping with 350 bootstrap samples. Pscore 1-1: 1-1 propensity score matching combined with regression.

3.3.2 Sensitivity analysis 1: MLE-based

We stratify the data into five subclasses based on the estimated propensity scores of the treated sites. The reference sites with estimated propensity scores outside the range of the treated sites’ estimated propensity scores are removed from further
analysis. To simplify the analysis, we make the assumption that \((\pi_s, \delta_{s1}, \delta_{s0}, \alpha_s) = (\pi, \delta_1, \delta_0, \alpha_s)\), that is, the sensitivity parameters are the same across all subclasses. We assume a simple relationship between \(\theta_{sz1}\) and \(\theta_{sz0}\), for example let \(\theta_{sz1} = c_{sz} \cdot \theta_{sz0}\).

The covariates included in the propensity score and outcome models can be found in Table 3.2. The MLEs of \(\beta_{sz}\) are obtained using package \texttt{optim} in R. Table 3.4 gives the sensitivity of the ATT estimate to 24 sets of assumption about the binary unmeasured confounder \(U\). We consider doubling and tripling the odds of receiving the combined CLRS/ SRS treatment, varying the multiplicative effect of \(U\) on the mean total crash count of the reference and treatment sites, and a range of proportion of sites with \(U = 1\). For each set of sensitivity parameters, we obtained the confidence interval for the CMF by bootstrapping with approximately 60 bootstrap samples. The low number of samples is due to computing time.
Table 3.4: Effects of an unobserved binary covariate on the average total crash count of the reference and treatment sites, \( \tau_0 \) and \( \tau_1 \), and on their ratio \( k = \tau_1/\tau_0 \).

<table>
<thead>
<tr>
<th>Effect of ( U=1 ) on ( Z )</th>
<th>Effect of ( U=1 ) on ( Y, Z = 0 )</th>
<th>Effect of ( U=1 ) on ( Y, Z = 1 )</th>
<th>( 1 - \pi )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \exp(\delta_1) = 0.5 )</td>
<td>( \tau_1 )</td>
<td>0.38</td>
<td>0.37</td>
</tr>
<tr>
<td>( \exp(\delta_0) = 0.5 )</td>
<td>( \tau_0 )</td>
<td>0.40</td>
<td>0.42</td>
</tr>
<tr>
<td>( \exp(\alpha) = 2 )</td>
<td>( k )</td>
<td>0.93</td>
<td>0.87</td>
</tr>
<tr>
<td>( CI )</td>
<td>(0.67, 1.46)</td>
<td>(0.71, 1.30)</td>
<td>(0.71, 1.23)</td>
</tr>
<tr>
<td>( \exp(\delta_1) = 2 )</td>
<td>( \tau_1 )</td>
<td>0.52</td>
<td>0.58</td>
</tr>
<tr>
<td>( \exp(\delta_0) = 2 )</td>
<td>( \tau_0 )</td>
<td>0.39</td>
<td>0.42</td>
</tr>
<tr>
<td>( k )</td>
<td>1.29</td>
<td>1.38</td>
<td>0.83</td>
</tr>
<tr>
<td>( CI )</td>
<td>(0.92, 3.57)</td>
<td>(0.88, 1.51)</td>
<td>(0.71, 1.04)</td>
</tr>
<tr>
<td>( \exp(\delta_1) = 0.5 )</td>
<td>( \tau_1 )</td>
<td>0.35</td>
<td>0.36</td>
</tr>
<tr>
<td>( \exp(\delta_0) = 2 )</td>
<td>( \tau_0 )</td>
<td>0.64</td>
<td>0.52</td>
</tr>
<tr>
<td>( k )</td>
<td>0.54</td>
<td>0.69</td>
<td>0.88</td>
</tr>
<tr>
<td>( CI )</td>
<td>(0.44, 0.67)</td>
<td>(0.52, 0.83)</td>
<td>(0.72, 1.12)</td>
</tr>
<tr>
<td>( \exp(\delta_1) = 2 )</td>
<td>( \tau_1 )</td>
<td>0.54</td>
<td>0.44</td>
</tr>
<tr>
<td>( \exp(\delta_0) = 2 )</td>
<td>( \tau_0 )</td>
<td>0.64</td>
<td>0.52</td>
</tr>
<tr>
<td>( k )</td>
<td>0.85</td>
<td>0.85</td>
<td>0.94</td>
</tr>
<tr>
<td>( CI )</td>
<td>(0.62, 1.09)</td>
<td>(0.62, 1.16)</td>
<td>(0.69, 1.22)</td>
</tr>
<tr>
<td>( \exp(\delta_1) = 1/3 )</td>
<td>( \tau_1 )</td>
<td>0.34</td>
<td>0.37</td>
</tr>
<tr>
<td>( \exp(\delta_0) = 1/3 )</td>
<td>( \tau_0 )</td>
<td>0.41</td>
<td>0.42</td>
</tr>
<tr>
<td>( k )</td>
<td>0.83</td>
<td>0.87</td>
<td>0.85</td>
</tr>
<tr>
<td>( CI )</td>
<td>(0.69, 1.62)</td>
<td>(0.69, 1.16)</td>
<td>(0.71, 1.18)</td>
</tr>
<tr>
<td>( \exp(\delta_1) = 3 )</td>
<td>( \tau_1 )</td>
<td>0.68</td>
<td>0.68</td>
</tr>
<tr>
<td>( \exp(\delta_0) = 3 )</td>
<td>( \tau_0 )</td>
<td>0.40</td>
<td>0.42</td>
</tr>
<tr>
<td>( k )</td>
<td>1.67</td>
<td>1.62</td>
<td>1.01</td>
</tr>
<tr>
<td>( CI )</td>
<td>(1.33, 2.42)</td>
<td>(1.07, 2.30)</td>
<td>(0.78, 1.44)</td>
</tr>
<tr>
<td>( \exp(\delta_1) = 1/3 )</td>
<td>( \tau_1 )</td>
<td>0.31</td>
<td>0.35</td>
</tr>
<tr>
<td>( \exp(\delta_0) = 3 )</td>
<td>( \tau_0 )</td>
<td>0.91</td>
<td>0.62</td>
</tr>
<tr>
<td>( k )</td>
<td>0.34</td>
<td>0.57</td>
<td>0.88</td>
</tr>
<tr>
<td>( CI )</td>
<td>(0.31, 0.45)</td>
<td>(0.42, 0.68)</td>
<td>(0.75, 1.26)</td>
</tr>
<tr>
<td>( \exp(\delta_1) = 3 )</td>
<td>( \tau_1 )</td>
<td>0.82</td>
<td>0.68</td>
</tr>
<tr>
<td>( \exp(\delta_0) = 3 )</td>
<td>( \tau_0 )</td>
<td>0.91</td>
<td>0.62</td>
</tr>
<tr>
<td>( k )</td>
<td>0.90</td>
<td>1.10</td>
<td>1.08</td>
</tr>
<tr>
<td>( CI )</td>
<td>(0.64, 1.26)</td>
<td>(0.77, 1.50)</td>
<td>(0.80, 1.64)</td>
</tr>
</tbody>
</table>

In the table above:

\( 1 - \pi \) is the percentage of sites with \( U = 1 \).

\( \exp(\alpha) \) is the multiplicative effect of \( U = 1 \) vs \( U = 0 \) on the treatment assignment: \( \exp(\alpha) = 2 \) means doubling the odds of receiving the CLRS/ SRS treatment.

\( \exp(\delta_0) \) (\( \exp(\delta_1) \)) is the multiplicative effect of \( U = 1 \) vs \( U = 0 \) on the mean total crash count of the sites without (and with) CLRS/ SRS applications.
From Table 3.4, we can see that the estimates for the average crash counts range from 0.39 - 0.91 for reference sites and from 0.31 - 0.82 for sites with applications of CLRS/ SRS, leading to a range of 0.34 - 1.67 for the estimates of the CMF. This range is much wider than the range estimated by the subclassification method (0.70, 1.03). Two-third of these combinations yield a CMF estimate less than 1. The four scenarios with CMF greater than 1 correspond to a high percentage of sites with $U = 1$, and generally, high effect of $U$ on $Z$, with increasing mean total crash counts on the treated sites while lowering counts on reference sites. However, only four sets of sensitivity parameters have CIs not including 1, corresponding to high ($\geq 50\%$) percentage of $U = 1$, and high effect of $U$ on $Y(0)$ (doubling/ tripling) as well as reducing effect of $U$ on $Y(1)$ (halving/ cutting in third). Whether these conditions are realistic for any confounder is another discussion that should take place with traffic engineers as well as safety researchers.

There are a few things to note about this approach. First, the effects presented in Table 3.4 (effect of $U$ on $Z,Y(0),Y(1)$) are completely hypothetical. It would be useful to be able to mimic the effects from real covariates. Second, confidence intervals were obtained by bootstrapping with only approximately 60-70 bootstrap samples. A larger number of bootstrap samples is desirable, but this is compute intensive. Third, we need to make several explicit assumptions about the relationship of $U$ with the outcome and treatment assignment. Ideally, we would prefer a more flexible approach in which different estimation methods can be applied (and compared). In the next subsection we show the results for the second, more flexible approach adapted from Ichino et al. (2008).

3.3.3 Sensitivity analysis 2: simulation-based

We first demonstrate one of the aforementioned advantages of this approach by making the unobserved confounder $U$ similar to some observed covariates. Since
all of these observed covariates are continuous, we first create new binary variables indicating whether the observed value is greater than the mean of that variable in the sample, $I(X > \bar{X})$. For example, pavement width has an average of 22.62 ft, so the corresponding variable is 1 if the site has a pavement width greater than 22.62, and 0 otherwise. We use propensity score matching combined with regression to estimate the ATT CMF.

Table 3.5 presents this first set of results. As a comparison, recall that 1-1 propensity score matching on the total crash count yielded 0.93 (0.70, 1.20) (Table 3.3). We let $U$ mimic the (binary transformed) distribution of six covariates: segment length, pavement width, average shoulder width, truck percentage, number of intersections, and the log of average annual daily traffic. Upon mimicking these covariates, we see a range of mild negative to moderately positive effect on both the outcome and the treatment ($\Gamma$ ranges from 1.03 to 2.11, and $\Lambda$ ranges from 0.74 to 3.88). However, the CMF estimate fluctuates by a small amount, ranging from 0.87 to 0.93 (compared to the original estimate of 0.93), and all the estimates remain insignificant with the confidence intervals including 1. This means that propensity score matching combined with regression is fairly robust to an unmeasured confounder that is loosely similar in distribution to the six covariates included in Table 3.5. Perhaps only with a much stronger outcome effect $\Gamma$ and treatment effect $\Lambda$ would an unmeasured confounder $U$ significantly change the conclusion of our estimate.
Table 3.5: Sensitivity analysis of CMF estimates to the effect of a binary confounding factor calibrated to mimic the distribution of some observed covariates.

<table>
<thead>
<tr>
<th>Confounder</th>
<th>Fraction of $U = 1$ by $Z$ and $K$</th>
<th>CMF</th>
<th>CI</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$p_{00}$</td>
<td>$p_{01}$</td>
<td>$p_{02}$</td>
</tr>
<tr>
<td>Segment length</td>
<td>0.49</td>
<td>0.56</td>
<td>0.69</td>
</tr>
<tr>
<td>Pavement width</td>
<td>0.28</td>
<td>0.29</td>
<td>0.37</td>
</tr>
<tr>
<td>Avg shoulder width</td>
<td>0.40</td>
<td>0.43</td>
<td>0.48</td>
</tr>
<tr>
<td>Truck percentage</td>
<td>0.40</td>
<td>0.41</td>
<td>0.41</td>
</tr>
<tr>
<td>Number of intersections</td>
<td>0.27</td>
<td>0.29</td>
<td>0.34</td>
</tr>
<tr>
<td>Log(AADT)</td>
<td>0.30</td>
<td>0.45</td>
<td>0.64</td>
</tr>
<tr>
<td>neutral</td>
<td>0.50</td>
<td>0.50</td>
<td>0.50</td>
</tr>
</tbody>
</table>

In the table above, $U$ is a binary unmeasured confounding covariate, and $p_{ij} = P(U = 1|Z = i, K = j)$. The outcome effect $\Gamma$ is the ratio of $U = 1$ vs $U = 0$ on the average estimated outcome fitted with a negative binomial model: $E(K|U = 1, X) / E(K|U = 0, X)$. Similarly, the selection effect $\Lambda$ is the average estimated risk ratio $P(Z = 1|U = 1, X) / P(Z = 1|U = 0, X)$. Since the observed covariates are not binary, for each confounder of interest, we classify the observation into high-low categories based on whether it is greater than the average of all the observed values, calculate the $p_{ij}$ and use these parameters to impute $U$ for each individual. The CMF is estimated using 1-1 propensity score matching combined with regression (with $U$ included). We repeat this process 1000 times. The reported CMF, outcome effect $\Gamma$, selection effect $\Lambda$ are averaged over these 1000 simulations.

Following Ichino et al. (2008), we look into this aspect by varying both effects $\Gamma$ and $\Lambda$ to a large range of value and recorded when the CMF estimated by propensity score matching combined with regression is driven away from, or close to 1, depending on the threat of interest. Suppose we’re concerned about the “false negative” scenario. We fix $p_u = P(U = 1)$, let $m = p_{10} - p_{11} = p_{11} - p_{12}$ and vary the differences $d = p_{00} - p_{01}$, $p_{01} - p_{02}$ and $s = p_{1} - p_{0}$. This reduces to a problem of solving a system of linear equations for the parameters $p_{ij}$ that fully determine the distribution of $U$. See the appendix for a detailed example for this scenario. We assume that 50% the sites in our sample have bad weather, and for simplicity, also
assumed that bad weather doesn’t really affect total crash counts at treated sites: \( m = p_{10} - p_{11} = p_{11} - p_{12} = 0 \). Table 3.7 shows the parameters \( p_{ij} \) for each of the 25 combinations of \( d \) and \( s \) we are considering. Table 3.6 gives the proportions of sites grouped by values of the transformed outcome \( K \) and the treatment status \( Z \).

For each set of \( p_{ij} \), we generate 350 simulations and presented the results Table 3.8 for this “false negative” case. The results for the “false positive” case is included in Table 3.9 for comparison.

Let’s consider our “bad weather” confounder example results in Table 3.8. Across all combinations of \( d \) and \( s \), these sensitivity parameters are fairly high with a range of 0.60 – 0.99 for \( p_{1j} \) and 0.39 – 0.90 for \( p_{0j} \). As \( d \) varies from 0.05 to 0.25, the range of \( \Gamma \) goes from 1.12 to 3.87. Similarly, a range of \( s \) from 0.1 - 0.5 translates to an effect on selection to the treatment that varies from 1.01 to \( \Delta_2 \). Moving from the left to the right means bad weather has a bigger role in the site being selected to the CLRS/ SRS treatment, while moving from the top to the bottom means bad weather influences more on the untreated outcome. Across all \( (d, s) \) combinations, the CMF estimates remain less than 1, but the 95% confidence intervals include 1 for the most part. As these influences increase (as \( \Gamma \) and \( \Lambda \) get larger), the estimates decrease. Both the effect of \( U \) on the outcome and the treatment assignment have to be fairly large in order to have a significant CMF effect with a 95% CI not including 1. For example, when \( d = 0.1 \) (\( \Gamma \in (1.42, 1.75) \)), the CMF point estimate only became significant when \( \Lambda \) is questionably high, more than 17.34. For a more plausible range of \( \Lambda \), e.g \( \in (1.55, 3.50) \), \( U \)’s effect on the outcome needs be to fairly large, \( \Gamma > 3.08 \).

### 3.4 Conclusion and future work

When randomized controlled trials are not feasible due to ethical or practical reasons, confounding factors should be accounted for to ensure that the two treatment groups have as much balance in the characteristic distributions as possible. Sev-
eral methods are readily available to control for observed confounders, but unmeasured confounders can lead to biased effect estimates. Unconfoundedness is generally untestable; however, it is possible to see how sensitive the analysis results are under violation of unconfoundedness.

We present two approaches to sensitivity analysis, both are extensions from previous works to accommodate for count outcome. We illustrate both methods using a data set from a study whose goal is to evaluate the safety effectiveness (measured in crash counts) of the combined application of CLRS and SRS two-lane rural roads in Pennsylvania. The baseline results using subclassification is 0.85 with a 95% CI of (0.70, 1.03) and the corresponding results using propensity score 1-1 matching combined with regression is 0.93 (0.70, 1.20).

The first approach is a Rosenbaum and Rubin type and is based on maximum likelihood estimation, showing that the CMF point estimates appear robust for the most part except for a few scenarios that are fairly extreme. However, bootstrapping for the confidence intervals is compute intensive. This method also requires a specification of the relationship between the outcome model and the unmeasured confounder. The second approach relaxes this assumption and is more flexible in the choice of the method being used to estimate the CMF itself. The unmeasured confounder is first generated to mimic some observed covariates, and then generated with increasing effects on the outcome and treatment selection to see whether it would push the CMF estimate to a significant value. This method can also be compute intensive as the negative binomial model for the outcome becomes more complicated. Careful consideration should also go into how many bins to stratify the outcome into.

In both cases, the estimates appear robust, deviating little from the baseline estimate, and only became significantly less than 1 under very extreme scenarios of the unmeasured confounder - at least several times more extreme than the effect
of the observed confounders on $Z$ and $Y$, as pointed out in the second approach’s results. One possible explanation for this nonsignificant effect is that for many years, Pennsylvania had been installing rumble strips as a blanket safety measure for their road systems. Therefore, it is likely that the high crash count sites had been already treated; the sites selected for treatment didn’t truly have a high accident issue and thus wouldn’t show as high of a benefit from the CLRS/SRS combined application.

Each sensitivity analysis approach has its own pros and cons and thus the researchers should choose accordingly. If the research study is using a subclassification estimator and the researchers have a strong confidence in the outcome model, the MLE-based approach offers a quick solution since there is no simulation involved. On the other hand, the simulation-based approach would be suitable when other estimating methods are involved or when the researchers want to relax the assumption about the relationship of $U$ with the outcome and treatment.

There is one direction that I would like to explore had time permitted and is left for further exploration: extending the simulation-based sensitivity analysis to address other estimating methods (besides propensity score matching). Specifically, in traffic safety research, the state-of-the-art gold standard for estimating a CMF is the Empirical Bayes (EB) approach (Hauer, 1997). This approach relies on a before-after design; it focuses on precisely estimating the number of crashes that would have occurred at an individual treated site in the after period had a treatment not been implemented, and then the CMF is estimated from the change in crash frequency from before until after the implementation of the treatment. EB is capable of accounting for observed (reported) changes in crash counts before and after the treatment that may be due to regression to the mean. It is well-established and is easy to implement. However, the EB approach requires a before-after study design, which is not always feasible because the implementation data of a countermeasure may be unknown. Moreover, the EB approach hinges on an underlying Gamma-
Poisson assumption on the outcome model as well as the assumption of a constant time trend for the reference and treatment sites. It would be beneficial to provide an approach that can accommodate for this method since it is widely used in the industry. The simulation-based method is more promising since once the sensitivity parameters are chosen, the unmeasured confounder can be generated and any method can be used as long as this confounder is included in the models. However, thoughts should be given to the before-after design, whether and how it would affect the distribution of $U$. 
Table 3.1: Mean and standard deviation for several characteristic covariates for treatment and reference sites.

<table>
<thead>
<tr>
<th></th>
<th>Reference</th>
<th>Treatment</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
</tr>
<tr>
<td>Segment length (miles)</td>
<td>0.47</td>
<td>0.13</td>
</tr>
<tr>
<td>Pavement width (ft)</td>
<td>22.62</td>
<td>3.55</td>
</tr>
<tr>
<td>Average shoulder width</td>
<td>3.00</td>
<td>2.17</td>
</tr>
<tr>
<td>AADT 2003</td>
<td>3484.34</td>
<td>2798.71</td>
</tr>
<tr>
<td>AADT 2004</td>
<td>3502.73</td>
<td>2817.72</td>
</tr>
<tr>
<td>AADT 2005</td>
<td>3510.08</td>
<td>2807.82</td>
</tr>
<tr>
<td>AADT 2006</td>
<td>3484.51</td>
<td>2801.47</td>
</tr>
<tr>
<td>AADT 2007</td>
<td>3482.50</td>
<td>2809.78</td>
</tr>
<tr>
<td>AADT 2008</td>
<td>3415.61</td>
<td>2767.16</td>
</tr>
<tr>
<td>AADT 2009</td>
<td>3371.22</td>
<td>2749.67</td>
</tr>
<tr>
<td>AADT 2010</td>
<td>3336.93</td>
<td>2714.88</td>
</tr>
<tr>
<td>AADT 2011</td>
<td>3290.86</td>
<td>2701.40</td>
</tr>
<tr>
<td>Average AADT</td>
<td>3430.97</td>
<td>2739.37</td>
</tr>
<tr>
<td>Sum of total crashes</td>
<td>4.57</td>
<td>4.78</td>
</tr>
<tr>
<td>Sum of injury-related crashes</td>
<td>2.49</td>
<td>2.78</td>
</tr>
<tr>
<td>Sum of run-off-road crashes</td>
<td>0.75</td>
<td>1.24</td>
</tr>
<tr>
<td>Sum of head-on crashes</td>
<td>0.19</td>
<td>0.48</td>
</tr>
<tr>
<td>Sum of side-swiped opposite direction crashes</td>
<td>0.12</td>
<td>0.37</td>
</tr>
<tr>
<td>Roadside hazard rating</td>
<td>4.84</td>
<td>0.80</td>
</tr>
<tr>
<td>Number of driveways</td>
<td>7.81</td>
<td>6.70</td>
</tr>
<tr>
<td>Number of intersections</td>
<td>0.33</td>
<td>0.58</td>
</tr>
<tr>
<td>Number of horizontal curves</td>
<td>0.97</td>
<td>1.08</td>
</tr>
<tr>
<td>Number of horizontal curves/mi</td>
<td>2.08</td>
<td>2.36</td>
</tr>
<tr>
<td>Average length of horizontal curves</td>
<td>298.68</td>
<td>416.34</td>
</tr>
<tr>
<td>Length of curve per mile</td>
<td>930.57</td>
<td>1185.27</td>
</tr>
<tr>
<td>Average degree of curvature</td>
<td>3.95</td>
<td>6.96</td>
</tr>
<tr>
<td>Number of intersections and driveways per mile</td>
<td>18.01</td>
<td>15.59</td>
</tr>
</tbody>
</table>

Note: The average AADT is over the period 2003-2011. The sum of crashes for each type is over the period 2003 - 2012.
Table 3.2: Covariates included in the propensity score and the outcome models.

<table>
<thead>
<tr>
<th>Pscore model</th>
<th>Outcome model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Segment length, pavement width, speed limit (factor), U_R code, average shoulder width, truck percentage, roadside hazard rating, number of driveways, number of intersections, number of horizontal curves, number of horizontal curves per mile, average length of horizontal curves, length of curves per mile, average degree of curvature, number of intersections and driveways per mile, log (AADT)</td>
<td>Log(segment length), pavement width, speed limit (factor), average shoulder width, truck percentage, roadside hazard rating, number of driveways, number of intersections, number of horizontal curves per mile, average degree of curvature, number of intersections and driveways per mile, log (AADT)</td>
</tr>
</tbody>
</table>

Table 3.6: Proportion of sites by the transformed outcome \( K \) and treatment status \( Z \).

<table>
<thead>
<tr>
<th>( Z )</th>
<th>( K )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.680</td>
</tr>
<tr>
<td>1</td>
<td>0.018</td>
</tr>
<tr>
<td>2</td>
<td>0.085</td>
</tr>
<tr>
<td>0</td>
<td>0.211</td>
</tr>
<tr>
<td>1</td>
<td>0.005</td>
</tr>
<tr>
<td>2</td>
<td>0.001</td>
</tr>
</tbody>
</table>
Table 3.7: “False negative” threats: Sensitivity parameters $p_{ij}$ for each combination of $d = p_{00} - p_{01} = p_{01} - p_{02}$ and $s = p_{1} - p_{0}$.

<table>
<thead>
<tr>
<th>d</th>
<th>s</th>
<th>0.1</th>
<th>0.2</th>
<th>0.3</th>
<th>0.4</th>
<th>0.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0.48</td>
<td>0.48</td>
<td>0.47</td>
<td>0.47</td>
<td>0.47</td>
</tr>
<tr>
<td>0.05</td>
<td>1</td>
<td>0.53</td>
<td>0.53</td>
<td>0.52</td>
<td>0.52</td>
<td>0.52</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0.58</td>
<td>0.58</td>
<td>0.57</td>
<td>0.57</td>
<td>0.57</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0.46</td>
<td>0.46</td>
<td>0.45</td>
<td>0.45</td>
<td>0.45</td>
</tr>
<tr>
<td>0.1</td>
<td>1</td>
<td>0.56</td>
<td>0.56</td>
<td>0.55</td>
<td>0.55</td>
<td>0.55</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0.66</td>
<td>0.66</td>
<td>0.65</td>
<td>0.65</td>
<td>0.65</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0.44</td>
<td>0.44</td>
<td>0.43</td>
<td>0.43</td>
<td>0.43</td>
</tr>
<tr>
<td>0.15</td>
<td>1</td>
<td>0.59</td>
<td>0.60</td>
<td>0.59</td>
<td>0.58</td>
<td>0.58</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.74</td>
<td>0.74</td>
<td>0.73</td>
<td>0.73</td>
<td>0.73</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0.42</td>
<td>0.42</td>
<td>0.41</td>
<td>0.41</td>
<td>0.41</td>
</tr>
<tr>
<td>0.2</td>
<td>1</td>
<td>0.62</td>
<td>0.62</td>
<td>0.61</td>
<td>0.61</td>
<td>0.61</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.82</td>
<td>0.82</td>
<td>0.81</td>
<td>0.81</td>
<td>0.81</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0.40</td>
<td>0.40</td>
<td>0.39</td>
<td>0.39</td>
<td>0.39</td>
</tr>
<tr>
<td>0.25</td>
<td>1</td>
<td>0.65</td>
<td>0.65</td>
<td>0.65</td>
<td>0.65</td>
<td>0.65</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.90</td>
<td>0.90</td>
<td>0.89</td>
<td>0.89</td>
<td>0.89</td>
</tr>
</tbody>
</table>

In the table above, $p_{ij} = P(U = 1|Z = i, K = j)$, and $p_1 - (p_0)$ denotes the proportion of sites with $U = 1$ in the treated (control) group. We assume that 50% the sites in our sample have bad weather, and for simplicity, also assume that bad weather doesn’t really affect total crash counts at treated sites: $m = p_{10} - p_{11} = p_{11} - p_{12} = 0$. 

---

69
Table 3.8: “False negative” scenario: Sensitivity analysis of CMF estimates to the effect of a binary confounding factor, varying its effect on the outcome and the treatment selection.

<table>
<thead>
<tr>
<th>d</th>
<th>Γ</th>
<th>s</th>
<th>A</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.05</td>
<td>(1.12, 1.40)</td>
<td>0.1</td>
<td>0.89 (0.68, 1.17)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.2</td>
<td>0.89 (0.68, 1.16)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.3</td>
<td>0.86 (0.66, 1.12)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.4</td>
<td>0.84 (0.64, 1.10)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.5</td>
<td>0.84 (0.64, 1.09)</td>
</tr>
<tr>
<td>0.1</td>
<td>(1.42, 1.75)</td>
<td>0.1</td>
<td>0.88 (0.67, 1.15)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.2</td>
<td>0.85 (0.65, 1.11)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.3</td>
<td>0.82 (0.63, 1.06)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.4</td>
<td>0.79 (0.61, 1.02)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.5</td>
<td>0.75</td>
</tr>
<tr>
<td>0.15</td>
<td>(1.80, 2.22)</td>
<td>0.1</td>
<td>0.87 (0.67, 1.13)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.2</td>
<td>0.82 (0.65, 1.11)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.3</td>
<td>0.78 (0.63, 1.07)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.4</td>
<td>0.73 (0.60, 1.01)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.5</td>
<td>0.69</td>
</tr>
<tr>
<td>0.2</td>
<td>(2.33, 2.93)</td>
<td>0.1</td>
<td>0.85 (0.66, 1.11)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.2</td>
<td>0.79 (0.61, 1.03)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.3</td>
<td>0.74 (0.57, 0.95)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.4</td>
<td>0.68 (0.53, 0.88)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.5</td>
<td>0.64</td>
</tr>
<tr>
<td>0.25</td>
<td>(3.08, 3.87)</td>
<td>0.1</td>
<td>0.85 (0.65, 1.10)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.2</td>
<td>0.77 (0.60, 0.99)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.3</td>
<td>0.70 (0.54, 0.89)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.4</td>
<td>0.65 (0.51, 0.82)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.5</td>
<td>0.60</td>
</tr>
</tbody>
</table>

**Note:** We set $p_u = P(U = 1) = 0.5$ and $m = p_{10} - p_{11} = p_{11} - p_{12} = 0$. Here $p_{00} < p_{01} < p_{02}$ and $d = p_{02} - p_{01} = p_{01} - p_{00}$. Each CMF estimate is averaged over 350 simulations. The original estimate for the CMF is 0.93 (0.70, 1.20). The outcome effect $\Gamma$ is the ratio of $U = 1$ vs $U = 0$ on the average estimated outcome fitted with a negative binomial model: $E(K|U = 1, X)/E(K|U = 0, X)$. Similarly, the selection effect $\Lambda$ is the average estimated risk ratio $P(Z = 1|U = 1, X)/P(Z = 1|U = 0, X)$. 
Table 3.9: “False positive” scenario: Sensitivity analysis of CMF estimates to the effect of a binary confounding factor, varying its effect on the outcome and the treatment selection.

<table>
<thead>
<tr>
<th>d</th>
<th>Γ</th>
<th>s</th>
<th>0.1</th>
<th>0.2</th>
<th>0.3</th>
<th>0.4</th>
<th>0.5</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(1.09, 2.17)</td>
<td>(1.60, 3.44)</td>
<td>(2.53, 6.35)</td>
<td>(4.62, 15.37)</td>
<td>(0.08, 2e14)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(0.71, 1.23)</td>
<td>(0.73, 1.26)</td>
<td>(0.75, 1.30)</td>
<td>(0.77, 1.34)</td>
<td>(0.78, 1.37)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(0.73, 1.26)</td>
<td>(0.77, 1.34)</td>
<td>(0.80, 1.40)</td>
<td>(0.83, 1.46)</td>
<td>(0.87, 1.55)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(0.74, 1.29)</td>
<td>(0.80, 1.41)</td>
<td>(0.85, 1.51)</td>
<td>(0.93, 1.67)</td>
<td>(1.01, 1.83)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(0.77, 1.34)</td>
<td>(0.85, 1.50)</td>
<td>(0.94, 1.69)</td>
<td>(1.04, 1.90)</td>
<td>(1.16, 2.16)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(0.81, 1.40)</td>
<td>(0.91, 1.61)</td>
<td>(1.05, 1.90)</td>
<td>(1.21, 2.26)</td>
<td>(1.42, 2.74)</td>
</tr>
</tbody>
</table>

Note: We set $p_a = P(U = 1) = 0.5$ and $m = p_{10} - p_{11} = p_{11} - p_{12} = 0$. Here $p_{00} > p_{01} > p_{02}$ and $d = p_{00} - p_{01} = p_{01} - p_{02}$. Each CMF estimate is averaged over 350 simulations. The original estimate for the CMF is 0.93 (0.70, 1.20). The outcome effect $\Gamma$ is the ratio of $U = 1$ vs $U = 0$ on the average estimated outcome fitted with a negative binomial model: $E(K|U = 1, X)/E(K|U = 0, X)$. Similarly, the selection effect $\Lambda$ is the average estimated risk ratio $P(Z = 1|U = 1, X)/P(Z = 1|U = 0, X)$. 

71
Propensity Score Augmented Additive-Interactive Gaussian Process Regression

4.1 Introduction

In observational studies, one of the central goals is to estimate the causal effect of a treatment, unconfounded by the difference in the subjects’ characteristics across the treatment groups. When randomized experiments are not an option, researchers need to rely on observational studies. Inference is trickier in observational studies because it is usually the case that the confounders are unbalanced in the treatment groups. Furthermore, there has been growing interest in estimating the heterogeneous treatment effects for individuals, or across subpopulations instead of the overall average treatment effects. Being able to estimate precisely the heterogeneity of treatment effects can offer valuable insights, since it will be possible to identify the subgroups for whom the treatment will be most effective and thus only administer the treatment accordingly.

There have been several methods proposed to draw causal inference for observational studies, focusing on modeling either or both the treatment assignment and the
outcome. Propensity score methods (Rosenbaum and Rubin, 1983a, 1984) are robust to achieve covariate balance and have been increasingly used in many disciplines due to its ease of use as well as its being less sensitive to model misspecification (Rubin, 1979). The propensity score is the conditional probability of being given the treatment, given the covariates: \( e(X) = P(Z = 1|X) \). It has two important properties: (1) if the distribution of \( e(X) \) is balanced between treatment groups, the distribution of the covariates \( X \) themselves would also be balanced, and (2) strongly ignorability conditional on the covariates \( X \) implies strongly ignorability conditional on just the propensity scores. To apply propensity score methods, in general one first has to estimate the propensity scores, and then the estimands of interest can be found via weighting, matching, or stratification. See Stuart (2010) and Imbens (2004), for example, for a review of these methods. The propensity score can also be used either as the sole predictor or as an additional predictor in the outcome model (Berk and Newton (1985), Muller et al. (1986)). All of these approaches rely on having the correct model for the propensity scores itself, which can lead to substantial bias when this condition isn’t satisfied.

Mixed approaches have also been proposed, combining propensity scores with regression models. Abadie and Imbens (2006, 2011) developed bias-corrected matching estimators, using regression to adjust for the residual difference in the covariates of the matched sample. Regression can also be combined with weighting, leading to double-robustness (DR) methods. Lunceford and Davidian (2004), and Mercatanti and Li (2014) proposed DR estimators for the ATE and ATT, respectively. In both mixed approaches, the outcome model has larger weight that the propensity score model in the sense that if the outcome model is misspecified, it could lead to much worse estimates (in terms of bias or RMSE) compared with a misspecified propensity score model. One advantage of propensity scores methods is that it’s fairly easy and straightforward to implement. Matching-only methods are available from open-
packages, for example, Matching (Sekhon and Grieve, 2012) or MatchIt (Ho et al., 2011) in R.

Without using the propensity scores, a straight-forward method to estimate these causal estimands is to fit a parametric model for the expected value of the outcome, conditional on the treatment and covariates: \( f(x, z) = \mathbb{E}(Y|X = x, Z = z). \) However, there are two main problems with this approach. First, it heavily relies on the correct specification of the model and can lead to significantly biased estimates if the wrong model is used. Secondly, if there is severe imbalance (or lack of overlap) between the observed characteristics of the treatment and the control groups, regression-based estimates rely on extrapolation and thus are sensitive to model misspecification. To provide flexibility in estimating the outcome surface function, semi-parametric or non-parametric models can be used. Hill (2011) proposed a causal inference extension based on the Bayesian Additive Regression Tree (BART) (Chipman et al., 2010). BART’s essential idea is to use a sum-of-trees model with a prior such that the contribution of each tree is small. Hill (2011) showed that BART performs fairly well compared to commonly used causal inference methods such as linear regression, propensity score matching and weighting, when the response surfaces are nonlinear. However, when there is a lack of overlap between the treatment groups, BART still has to rely on extrapolation, similar to model-based estimators, or to strategies where these units with low or zero overlap aren’t discarded. Hill and Su (2013) proposed using the posterior variance of the causal effects to identify areas with low overlap and discard units accordingly. Moreover, the uncertainty for estimates in these areas will automatically increase when there is lack of overlap; however there is a limit to this amount of increase placed by the regularization prior, and thus can lead to low coverage (Hill, 2011).

On the other hand, many machine learning techniques have been developed to build a relationship between an observation’s predictors and their observed outcome.
Random forest (RF) algorithms were first introduced by Breiman (2001) with the idea of exhaustingly search for split points on many subsets of data, growing the trees to maximal depth. Foster et al. (2011) utilized the counterfactual outcome idea and proposed a "virtual twins" approach, essentially first predicting both counterfactual outcomes for each observation, then used the difference in the predicted outcomes to build a regression tree. Meinshausen generalized the RF algorithm to quantile regression forests which give way to estimate the conditional quantile and thus build prediction intervals (Meinshausen, 2006). Athey and Imbens (2015) focused on directly predicting the causal effects of a treatment (instead of the counterfactual outcome) by building causal trees. They also proposed cross-validation criteria (in-sample and out-of-sample goodness-of-fit measures) to evaluate the predictions' accuracy. Wager and Athey (2017) extended the random forest algorithm to a causal forest and also provided conditions under which consistency and asymptotic normality can be achieved.

We propose using Bayesian nonparametric regression with Gaussian process priors to predict the potential outcome function for each treatment condition. First, instead of the conventional Gaussian process regression priors, we employ the model developed by Qamar and Tokdar (2014), a sparse additive-interactive model for the nonparametric regression (airGP), splitting the regression function into $k$ components, with each component including a small number of interacting covariates. With its built-in variable selection feature choosing component elements, this method is demonstrated to provide very competitive prediction accuracy especially in high-dimensional settings. Second, we propose adding the estimated propensity scores as an additional covariate. When the estimated propensity score is consistent, combining propensity scores with airGP will increase the efficiency. In the case that the propensity score model is misspecified, this approach would still inherit the robustness property of airGP. Moreover, in presence of confounding, airGP would be
sensitive to this lack of overlap by increasing the standard error estimate. We show
the advantages of this approach via simulations, comparing it to propensity score
methods in estimating the ATE, and to machine-learning methods in estimating the
individual treatment effects.

4.2 Proposed method

Notations. Let us first introduce the notations and assumptions we’ll be mak-
ing throughout. Let our sample size be $n$, with the data for observation $i$ include
$\{X_i, Z_i, Y_i\}$ where $X_i$ is the covariate vector, $Z_i$ the treatment assignment ($Z_i = 0$ for
the control group and $Z_i = 1$ for the treatment group), and $Y_i$ the observed outcome.
We adopt the traditional Rubin Causal Model (Rubin, 1974): each observation has
two potential outcomes, $Y_i(0)$ and $Y_i(1)$, corresponding to the outcome that obser-
vation $i$ would have received with and without the treatment. However, only the
potential outcome corresponding to the assigned treatment condition is observed
and the other is missing. In other words, $Y_i = Y_i(1)Z_i + Y_i(0)(1 - Z_i)$.

There are several causal estimands that can be of interest, out of which the two
most common are the population average treatment effect (ATE) and the popula-
tion average treatment effect for the treated (ATT). When the treatment effects differ
across subgroups of the population, another quantity of interest is the individual
treatment effect (ITE), conditional on the observed covariates.

$$\tau^{ATE} = \mathbb{E}[Y(1) - Y(0)],$$

$$\tau^{ATT} = \mathbb{E}[Y(1) - Y(0)|Z = 1],$$

$$\tau^{ITE}(x) = \mathbb{E}[Y(1) - Y(0)|X_i = x].$$

In order for the potential outcomes to be well defined and the causal estimands
possible to estimate, we make the usual assumption of Stable Unit Treatment Value
Assumption (SUTVA), which has two components: (1) the outcome of one unit does
not depend on another unit’s assigned treatment and (2) there is only one “version” for each treatment level. Furthermore, we make the assumption of unconfoundedness, also known as the assumption of no unmeasured confounders. This assumption means that the treatment is effectively randomized given the variables, and thus is independent of the potential outcomes: \( \{Y(1), Y(0) \perp Z \mid X \} \), which implies that \( P(Y(z) \mid X) = P(Y(z) \mid X, Z = z) \), and thus we can estimate the ATE and ITE as:

\[
\tau_{ATE} = \mathbb{E}_X[\mathbb{E}(Y \mid X, Z = 1) - \mathbb{E}(Y \mid X, Z = 0)],
\]
\[
\tau_{ITE} = \mathbb{E}[Y \mid X_i = x_i, Z = 1] - \mathbb{E}[Y \mid X_i = x_i, Z = 0].
\]

Last but not least, we make the overlap assumption: \( 0 < P(Z = 1 \mid X) < 1 \), meaning that we restrict the analysis to observations for which either treatment can be assigned. Together with unconfoundedness, this forms the strong ignorability condition.

Our goal is to estimate the potential outcome function \( f(z, x) \) using a regression model:

\[
Y_i = f(Z_i, X_i) + \varepsilon_i, \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2), i = 1, \ldots, n. \tag{4.4}
\]

Let \( GP(\mu, K) \) be a Gaussian process with mean \( \mu \) and covariance function \( K \). The additive-interactive Gaussian process regression model is specified (Qamar and Tokdar, 2014) as follows:

\[
f = f_1 + \cdots + f_k
\]

\[
f_j(\rho_j, \lambda_j, \gamma_j, \sigma) \sim GP(\mu(.), \sigma^2 \rho_j^2 C^{SE}(., | \gamma_j, \lambda_j))
\]

\[
\rho_j \overset{iid}{\sim} \pi_\rho \sigma, \lambda_j \overset{iid}{\sim} \pi_\lambda, \gamma_j \overset{iid}{\sim} \pi_\gamma, \sigma \overset{iid}{\sim} \pi_\sigma
\]

where \( C^{SE}(., | \gamma_j, \lambda_j) \) is the squared exponential covariance function. The idea is to separate the regression function \( f \) into \( k \) additive components, each component includes a small number of interacting predictors. The parameter \( \rho_j \) indicates the
signal-to-noise ratio, while the parameters $\gamma_j$ and $\lambda_j$ help tuning the predictor inclusion and smoothing of each component. This method is more robust in recovering the interactions compared to using a single Gaussian process when the dimension of the covariates gets large (Qamar and Tokdar, 2014). The authors showed that this additive-interactive regression with Gaussian process- airGP method provides more accurate prediction than several other often-used regression techniques via several simulated and real data. See Qamar and Tokdar (2014) for the prior specification as well as the full MCMC sampling scheme.

Another advantage of using a Gaussian process -based method surfaces when confounding is an issue. Suppose we have a case when the distribution of the covariate is very different between the treatment and the control group, i.e. very low overlap. Linear models can yield good fits for each group but would most likely fail at the extrapolating areas. Similarly, BART would have narrow error bars due to the limit placed by the regularization prior on the uncertainly increase, and would also have low coverage. On the other hand, in respond to this airGP would trade off the potential bias with variance by increasing the estimated prediction standard error at these observations. A package implementing airGP is available at https://github.com/tokdar/airGP.

We propose to augment this approach by adding the propensity score as an additional predictor, and show that this increases the efficiency in estimating causal estimands. Specifically, we replace the regression function $f$ in Equation 4.4 with:

$$ Y_i(0) = f_0(\hat{e}(X_i), X_i) + \epsilon_i $$

$$ Y_i(1) = f_1(\hat{e}(X_i), X_i) + \epsilon_i, \epsilon_i \overset{iid}{\sim} N(0, \sigma^2), i = 1, \ldots, n $$

where $\hat{e}(X_i)$ is the propensity score estimated from the observed data. From here on, we refer to this approach as airGP_{ps}.

Gutman and Rubin (2015) had a similar idea of incorporating the propensity
scores into the regressing function, approaching the problem under the view of a missing data problem. Their method - multiple imputation with two splines in subclasses (MITSS) splits each treatment group’s response surface into two additive components: a spline function based on the estimated propensity score and a linear model which includes covariates orthogonalized to the propensity score. However, because the second component is a parametric model, when there is lack of overlap between the treatment this approach becomes sensitive to model misspecification. Little and An (2004) also proposed a method of missing data imputation - penalized spline of propensity prediction - and showed that it yields robust inference. First, the propensity scores need to be estimated, \( \hat{e}(X) \). Then the imputation can be done by drawing from the marginal mean of the outcome \( Y \) which is modelled with two components, a penalized spline of the estimated propensity score and a parametric model including other covariates. This method gives a consistent estimator for the outcome marginal mean if either the outcome model is correctly specified, or both the propensity score model and the spline model relating the outcome with the propensity score are correctly specified. In the latter case, the parametric part of the model does not need to be correctly specified.

Our approach is different from the aforementioned methods in the sense that we specify the outcome as one single Bayesian nonparametric model with both the (estimated) propensity scores and all the observed covariates, instead of splitting into a nonparametric and a parametric component which remains vulnerable to model misspecification. We show through simulations that this new approach provides both efficiency and robustness: (1) when the estimated propensity score \( \hat{e}(x) \) is consistent, combining it with airGP will increase the efficiency and (2) in the case that the propensity score model is misspecified, our augmented airGP_{ps} is still robust.

In our simulations, we use completely simulated data and consider two cases: when the treatment effects are homogeneous and heterogeneous. In the first scenario
with homogeneity, we compare airGP_{ps} against several commonly-used propensity score methods as well as BART in estimating the ATT. In the second scenario with heterogeneity, we compare airGP_{ps} (1) against several commonly-used propensity score methods as well as BART in estimating the ATT and (2) with the regular GP method as well as with BART and random forest in estimating the individual treatment effects.

4.3 Simulation experiments

4.3.1 Homogeneous treatment effect

Let our data have \( n = 200 \) observations, with the number of observed covariates being \( p = 100 \). The covariates, treatment assignment and outcome are generated as:

\[
X_k \sim N(0, 1) \quad k = 1 - 25 \\
X_k \sim U(0, 1) \quad k = 26 - 50 \\
X_k \sim \text{Bin}(\text{logit}^{-1}(X_{k-50} - X_{k-25})), \quad k = 51 - 75 \\
X_k \sim \text{Poisson}(5 + 0.75X_{k-75} \times (X_{k-50} + X_{k-25})), \quad k = 76 - 100 \\
\]

\[
Z \sim \text{Bernoulli}(p_i) \\
\text{logit}(p_i) = 0.25 \sum_{k=1}^{5} X_k + 0.035 \sum_{k=26}^{30} X_k X_{k+50} - 0.5 \sum_{k=51}^{55} X_k \\
Y_i = 2Z + \sum_{k=1}^{5} X_k + 0.5 \sum_{k=26}^{30} X_k X_{k+50} - 5 \sum_{k=51}^{55} X_k + \epsilon_i \\
\epsilon_i \sim N(0, 1)
\]

With this data-generating scheme, \( X_1, \ldots, X_{50} \) are independent covariates, \( X_{51} \ldots X_{75} \) depends on pairs of covariates in \( X_1, \ldots, X_{50} \), and \( X_{76} \ldots X_{100} \) depends on \( X_1 \ldots X_{75} \) by groups of three. Both the treatment and the outcomes are driven by the same set of covariates: the first five of every group of covariates, by distribution (so, the first
Figure 4.1: Homogeneous treatment effect: True propensity scores for simulated data with $n = 200$ and $p = 100$.

five of the normal group, the uniform group, the Bernoulli group, and the Poisson group). The treatment assignment is fixed at 2. Over 50 simulated data sets, the naive difference $E(Y|Z = 1) - E(Y|Z = 0)$ ranges from 3.009 to 7.087, with an average of 4.960. Figure 4.1 plots the true propensity scores used to generate the treatment assignment for one simulated data set.

Our estimand of interest is the average treatment effect for the treated group (ATT). We compare airGP’s performance - with and without the estimated propensity scores as a covariate - against six propensity score based methods and also against BART. With the high dimension of the propensity scores, we use LASSO regression for both the outcome model and the propensity scores, including only linear terms of all the covariates and thus missing all the interactions. The regularization penalty parameter is determined through 10-fold cross-validation and is adjusted to ensure that at least 20% of the covariates are in the final model.
Performance are compared based on (1) Estimate, (2) Absolute bias, (3) RMSE, and (4) Coverage of 90% confidence interval, averaging over 50 simulated data sets. We compute the Abadie-Imbens standard error for matching and mixed-matching and regression, and use bootstrapping with 75 bootstrap samples to obtain the standard error of other propensity-score methods. The 90% confidence interval is computed as $Est \pm 1.65 \times SE$ for each of the 50 generated data set. For BART and airGP, the 90% CI is obtained from the posterior draws of the ATT estimates. We run two sets of simulations: one using the true propensity scores, and the other one the estimated propensity scores. Summary of the two cases are displayed in Table 4.1. Figure 4.2 plots the estimates and 90% confidence intervals for these estimators under the two sets of simulation.

Table 4.1: Homogeneous treatment effects: Models run with (top) estimated and (bottom) true propensity scores.

<table>
<thead>
<tr>
<th></th>
<th>Reg1</th>
<th>Wt</th>
<th>Sub</th>
<th>Match</th>
<th>DR1</th>
<th>Mix1</th>
<th>AGP</th>
<th>AGP$_{ps}$</th>
<th>BART</th>
<th>BART$_{ps}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estimate</td>
<td>2.26</td>
<td>4.52</td>
<td>3.72</td>
<td>3.78</td>
<td>2.24</td>
<td>3.30</td>
<td>2.20</td>
<td>2.17</td>
<td>3.00</td>
<td>2.93</td>
</tr>
<tr>
<td>AA Bias</td>
<td>0.26</td>
<td>2.52</td>
<td>1.72</td>
<td>1.86</td>
<td>0.25</td>
<td>1.32</td>
<td>0.23</td>
<td>0.22</td>
<td>1.00</td>
<td>0.94</td>
</tr>
<tr>
<td>RMSE</td>
<td>0.30</td>
<td>2.59</td>
<td>1.98</td>
<td>2.17</td>
<td>0.29</td>
<td>1.50</td>
<td>0.37</td>
<td>0.29</td>
<td>1.06</td>
<td>0.98</td>
</tr>
<tr>
<td>Coverage</td>
<td>0.82</td>
<td>0.02</td>
<td>0.92</td>
<td>0.65</td>
<td>1.00</td>
<td>0.86</td>
<td>0.88</td>
<td>0.16</td>
<td>0.16</td>
<td>0.16</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Reg1</th>
<th>Wt</th>
<th>Sub</th>
<th>Match</th>
<th>DR1</th>
<th>Mix1</th>
<th>AGP</th>
<th>AGP$_{ps}$</th>
<th>BART</th>
<th>BART$_{ps}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estimate</td>
<td>2.29</td>
<td>2.17</td>
<td>2.02</td>
<td>2.21</td>
<td>2.14</td>
<td>2.57</td>
<td>2.17</td>
<td>2.04</td>
<td>3.12</td>
<td>2.45</td>
</tr>
<tr>
<td>AA Bias</td>
<td>0.30</td>
<td>0.75</td>
<td>0.40</td>
<td>0.52</td>
<td>0.18</td>
<td>0.66</td>
<td>0.19</td>
<td>0.13</td>
<td>1.13</td>
<td>0.45</td>
</tr>
<tr>
<td>RMSE</td>
<td>0.33</td>
<td>0.92</td>
<td>0.50</td>
<td>0.67</td>
<td>0.22</td>
<td>0.78</td>
<td>0.23</td>
<td>0.16</td>
<td>1.18</td>
<td>0.53</td>
</tr>
<tr>
<td>Coverage</td>
<td>0.82</td>
<td>0.88</td>
<td>0.90</td>
<td>0.90</td>
<td>0.94</td>
<td>0.84</td>
<td>0.90</td>
<td>0.92</td>
<td>0.08</td>
<td>0.56</td>
</tr>
</tbody>
</table>

True ATT is fixed at 2. Reg1: regression estimator with linear terms only. Wt: weighting estimator. Sub: Subclassification estimator with 6 subclasses. Match: 1-1 nearest neighbor matching estimator. DR1: doubly-robust estimator with outcome model including only linear terms. Mix1: mixed matching-regression estimator, with regression model including only linear terms. AGP: Additive GP, no propensity scores included. AGP$_{ps}$: Additive GP with propensity scores included as one additional variable. BART: Bayesian Additive Regression Trees, no propensity scores included. BART$_{ps}$: BART with propensity scores included as one additional covariate.
Figure 4.2: Homogeneous treatment effects: Average mean, 90% CI and coverage of the ATT estimators for simulated data with $n = 200$ and $p = 100$. (Left): true propensity scores and (Right) estimated propensity scores were used. True treatment effect is 2, indicated by the horizontal line.

Overall, using the true propensity scores shows an improvement across most methods in terms of bias, RMSE and coverage. This improvement is most apparent in propensity-score-only methods, where RMSE can be decreased by as much as 75% (subclassification). The linear regression model actually performs well overall, having the second smallest RMSE and a decent coverage (86%) when the propensity scores are estimated. This is most likely due to the fact that the simulated data is “easy” enough for a LASSO regression model to capture. Regression’s strength can also be seen from the performance increase by adding the regression adjustment to either weighting or matching. Using estimated propensity scores, the RMSE of DR is 10% that of the weighting-only method, and this ratio is 20% when the true propensity scores are known. GP-based methods perform relatively well compared to the propensity-score based methods and BART. $\text{airGP}_{ps}$ and BART yield similar
confidence intervals, but airGP’s estimators are closer to the true ATT. When the propensity scores are unknown and must be estimated (in this case, with a misspecified model), adding the propensity scores as a covariate does not appear to improve airGP\_ps’s and BART’s performance significantly. BART’s narrow confidence intervals lead to its low coverage - under 20% when the propensity scores are estimated.

4.3.2 Heterogeneous treatment effect

Next, we consider the case when the treatment effect is no longer homogeneous. Our simulated data has $n \approx 300$ with 100 covariates, and there are interactions between the treatment assignment and the covariates. For each observation $i$, we generate our data as follows:

$$X_k \sim N(0, 1) \quad k = 1 - 25$$

$$X_k \sim U(0, 1) \quad k = 26 - 50$$

$$X_k \sim Bin(\logit^{-1}(X_{k-50} - X_{k-25})), \quad k = 51 - 75$$

$$X_k \sim Poisson(5 + 0.75X_{k-75} \times (X_{k-50} + X_{k-25})), \quad k = 76 - 100.$$

$$Z \sim Bernoulli(p_i)$$

$$\logit(p_i) = 0.3 \sum_{1}^{5} X_k - 2.75 \sum_{26}^{30} X_k X_{k+25} + 0.055 \sum_{76}^{80} X_k$$

$$f(X) = \left[ \sum_{26}^{30} X_{k+25} \exp(X_k) \right] / \left[ 1 + \sum_{26}^{30} X_{k+25} \exp(X_k) \right]$$

$$Y(0) = 0.15 \sum_{1}^{5} X_k + 1.5 \exp(1 + 1.5f(X)) + \epsilon_i$$

$$Y(1) = \sum_{1}^{5} \left[ 0.15X_k + 0.5X_k^2 + 0.1X_k^3 \right] + 2.25 \sqrt{20 + 1.5 \sum_{76}^{80} X_k + \epsilon_i}$$

$$Y(Z) = ZY(1) + (1 - Z)Y(0); \quad \epsilon_i \sim N(0, 0.25)$$

84
The independent covariates are generated in the same fashion as in the homogeneous treatment effect scenario; and similarly, both the treatment and the outcomes are driven by the same set of covariates: the first five of every group of covariates, by distribution. To escalate the lack of overlap, we identify and remove all control units with the true propensity scores greater than the 55th percentile of the treated group’s propensity scores, as well as the treated units with propensity scores less than the 45th percentile of the control group’s. We adjust the simulated sample size such that the number of observations after this removal process averages to 300. The density plots of the true propensity scores for each treatment group as well as their range (minimum, median, maximum) under this simulating scheme is presented in Figure 4.3.

Figure 4.3: Heterogeneous treatment effects: Density plot of true propensity scores for each treatment group. Points on the plots denote the minimum, median and maximum propensity scores. Our simulated data have $n = 300$ and $p = 100$. 

85
The same set of covariates are involved in both the propensity score model and the outcome models. A plot of the potential outcomes by group as well as individual treatment effects by group can be seen in Figure 4.4. From the left panel, we can see that the potential outcomes under the treatment condition for both groups are generally similar, but the control outcome surface $Y(0)$ is negatively associated with the true propensity scores, leading to higher outcome values for the control group and lower for the treatment group. The right panel in Figure 4.4 confirms this: treated observations have higher treatment effects than control observations.

Similar to the case with homogeneous treatment effects, we use LASSO to perform variable selection for both the regression and propensity scores models, and also run two sets of simulations using the estimated propensity scores and the true propensity scores.
scores for comparison. In the analysis stage, LASSO variable selection was done with all two-way interaction terms for the propensity scores model, and only linear terms for the outcome model. The summary of the two cases are displayed in Table 4.2. Figures 4.5 and 4.6 plot the estimates, 90% confidence intervals and coverages of all estimators for both simulation sets. As before, each set includes 50 simulated data sets.
Figure 4.5: Heterogeneous treatment effects: models run with true propensity scores: Average mean, 90% CI and coverage of the ATT estimators for simulated data with \( n = 200 \) and \( p = 100 \). The x-axis is the sample ATT in each of the 50 simulated datasets, and this “true value” is also represented by the straight line across each graph.
Figure 4.6: Heterogeneous treatment effects: models run with estimated propensity scores: Average mean, 90% CI and coverage of the ATT estimators for simulated data with $n = 200$ and $p = 100$. The x-axis is the sample ATT in each of the 50 simulated datasets, and this “true value” is also represented by the straight line across each graph.
Table 4.2: Heterogeneous treatment effects: Models run with (top) true and (bottom) estimated propensity scores.

<table>
<thead>
<tr>
<th></th>
<th>Reg1</th>
<th>Wt</th>
<th>Match</th>
<th>DR1</th>
<th>Mix1</th>
<th>GP</th>
<th>GP_ps</th>
<th>AGP</th>
<th>AGP_ps</th>
<th>BART</th>
<th>BART_ps</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA Bias</td>
<td>1.66</td>
<td>1.47</td>
<td>1.33</td>
<td>1.33</td>
<td>1.56</td>
<td>1.75</td>
<td>1.19</td>
<td>1.12</td>
<td>0.64</td>
<td>1.52</td>
<td>0.96</td>
</tr>
<tr>
<td>RMSE</td>
<td>1.68</td>
<td>1.56</td>
<td>1.51</td>
<td>1.38</td>
<td>1.67</td>
<td>1.83</td>
<td>1.28</td>
<td>1.19</td>
<td>0.76</td>
<td>1.55</td>
<td>1.03</td>
</tr>
<tr>
<td>Coverage</td>
<td>0.00</td>
<td>0.13</td>
<td>0.51</td>
<td>0.04</td>
<td>0.31</td>
<td>0.00</td>
<td>0.01</td>
<td>0.00</td>
<td>0.23</td>
<td>0.00</td>
<td>0.08</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Reg1</th>
<th>Wt</th>
<th>Match</th>
<th>DR1</th>
<th>Mix1</th>
<th>GP</th>
<th>GP_ps</th>
<th>AGP</th>
<th>AGP_ps</th>
<th>BART</th>
<th>BART_ps</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA Bias</td>
<td>1.59</td>
<td>2.03</td>
<td>1.48</td>
<td>1.48</td>
<td>1.46</td>
<td>1.78</td>
<td>1.18</td>
<td>1.08</td>
<td>0.75</td>
<td>1.46</td>
<td>0.93</td>
</tr>
<tr>
<td>RMSE</td>
<td>1.62</td>
<td>2.05</td>
<td>1.96</td>
<td>1.51</td>
<td>2.03</td>
<td>1.84</td>
<td>1.31</td>
<td>1.15</td>
<td>0.93</td>
<td>1.49</td>
<td>1.02</td>
</tr>
<tr>
<td>Coverage</td>
<td>0.00</td>
<td>0.00</td>
<td>0.65</td>
<td>0.00</td>
<td>0.67</td>
<td>0.00</td>
<td>0.14</td>
<td>0.02</td>
<td>0.29</td>
<td>0.00</td>
<td>0.32</td>
</tr>
</tbody>
</table>

*Reg1*: regression estimator, linear terms only.  
*Wt*: weighting estimator.  
*Match*: 1-1 nearest neighbor matching.  
*DR1*: doubly-robust estimator, outcome model including only linear terms.  
*Mix1*: mixed matching-regression, regression model including only linear terms.  
*GP*: Regular GP, no propensity scores included.  
*GP_ps*: Regular GP with pscores included as one additional variable.  
*AGP*: Additive GP, no propensity scores included.  
*AGP_ps*: Additive GP with pscores included as one additional variable.  
*BART*: Bayesian Additive Regression Trees, pscores not included.  
*BART_ps*: BART, pscores included.

In figures 4.5 and 4.6, the x-axis represents the sample ATT in each of the 50 simulated data sets. This “true ATT” is also shown by the $y = x$ solid black line. We exclude subclassification because with the extreme imbalance among the two groups, most of the times many subclasses can’t be formed and the estimates suffer greatly. As expected, all methods suffer from the more severe lack of overlap as well as the more complicated outcome models, even when the true propensity scores are known and used. The regression method no longer works well due to extrapolation. When the true propensity scores are known, airGP_ps yields the smallest RMSE, with the second winner (BARTps) having RMSE 35.5% higher. The true propensity scores improve BART’s performance (reducing RMSE by 35.8%), yet the narrow confidence intervals lead to low coverage (8%). Even with the true propensity scores known, weighting and matching methods struggle to capture the true ATT, either resulting in high bias and low coverage (weighting) or very wide confidence interval (matching).
When the propensity scores have to be estimated, all methods involving propensity scores suffer, however airGP$_{ps}$ remains the method with the lowest RMSE, but the coverage is low (29%). 1-1 propensity score matching and bias-adjusted matching have the highest coverage (57 - 65%), yet their RMSE are high which can be seen by the wide confidence intervals.

4.3.3 Individual treatment effects

We move onto subgroup prediction, demonstrating the performance of airGP$_{ps}$ by comparing it to regular Gaussian process regression, BART, and quantile regression forest, considering both cases when the true propensity scores are known and when they are estimated. For each method considered, we train separately on each treatment group and use this to predict the counterfactual outcome for the other group. The individual treatment effect is defined as $\hat{\tau}(x) = Y_{obs} - \hat{Y}(0)$ for the treated and $\hat{\tau}(x) = \hat{Y}(1) - Y_{obs}$ for the control units.

Quantile regression forest computations are implemented using the quantregForest R-package (Meinshausen, 2017), based on 1000 trees with $mtry$ (number of variables available for splitting at each tree node) being the optimal value with respect to the out-of-bag error estimate. For BART, we use the BayesTrees R-package with 500 burn-in, 1000 posterior samples and the rest of the parameters are set at default. We set the maximum rank for both GP and airGP$_{ps}$ at 75, burn 20% of the chain and keep 1000 posterior samples. For airGP$_{ps}$, we also limit the maximum number of components at $\lfloor \log(1 + p) \rfloor$. When the true propensity scores are unknown, we use LASSO with all two-way interaction terms to estimate the propensity scores. The regularization penalty parameter is determined through 10-fold cross-validation and also is adjusted to ensure that the final model include at least 20% of the covariates. We repeat our simulation 50 times.

We compare the methods’ performance based on bias, MSE and coverage, con-
ditioning on the true propensity scores. We first separate the observations in to 10 blocks \( B = \{B_1, ..., B_{10}\} \) based on deciles of their true propensity scores \( e(X) \). Let \( K \) be the number of simulations we run, here \( K = 50 \). We estimate the conditional bias and MSE for block \( B_s \) by:

\[
\hat{\text{Bias}}(s) = \frac{1}{K} \sum_{k=1}^{K} \frac{1}{n_{s,k}} \sum_{x_i \in B_{s,k}} [\hat{\tau}(x_i) - \tau(x_i)]
\]

\[
\hat{\text{MSE}}(s) = \frac{1}{K} \sum_{k=1}^{K} \frac{1}{n_{s,k}} \sum_{x_i \in B_{s,k}} [\hat{\tau}(x_i) - \tau(x_i)]^2
\]

\[
\hat{\text{Coverage}}(s) = \frac{1}{K} \sum_{k=1}^{K} \frac{1}{n_{s,k}} \sum_{x_i \in B_{s,k}} I_s(x_i)
\]

where \( n_{s,k} \) denotes the number of observations being in block \( B_s \) at simulation \( k \), and \( I_k(x_i) \) is an indicator that equals 1 if the 90% confidence interval at simulation \( k \) covers the true individual treatment effect.
Figure 4.7: Bias, RMSE and coverage for each propensity score decile of several methods for the estimated individual treatment effects of 50 simulated data sets with $n = 300$ and $p = 100$. Models are run with true (top row) and estimated (bottom row) propensity scores. Solid lines and dashed lines denote estimators without and with added propensity scores, respectively.
Figure 4.7 plots the average absolute bias, RMSE and coverage at each propensity score decile for all methods considered, with the true propensity scores known (top row) or estimated from a LASSO model (bottom row). Solid lines and dashed lines denote estimators without and with the propensity score as an additional covariate, respectively. There does not appear to be a significant difference between using the true propensity scores or the estimated ones. However, it should be noted that in this simulation, the true propensity scores model involves only linear and two-way interactions between the covariates which LASSO was able to pick out from all the possible linear and two way interactions. The discrepancy would likely be larger when the true propensity scores model is more complex.

We observe a similar trend among all methods: their performance gets worse for larger true propensity scores, essentially upon estimating the treatment effect for the treated units. This is not unexpected since our outcome generating functions are such that the treated potential outcomes are fairly similar among the groups and thus the heterogeneity in the treatment effects is driven primarily from the outcome surface for the control condition. When the true propensity scores are used, adding them to the airGP model increases the performance - the dashed line with triangles is below the solid line with triangles in the bias and RMSE plots, whereas it is above the solid line in the coverage plot. This indicates lower bias, RMSE and higher coverage. On the second row, when the propensity scores have to be estimated, the bias-variance tradeoff of airGP_{ps} can be seen in the middle plot where RMSE is increased when the propensity scores are included. Among the methods considered, airGP (triangles) yields the smallest average absolute bias and MSE across all propensity score deciles. Regular GP (circles) performs similarly to BART (upside-down triangles) in terms of bias/ RMSE, but it has the lowest coverage - under 30% - for all propensity score deciles. Adding the propensity scores to the model helps not only airGP but also all methods in reducing bias/ RMSE as well as increasing coverage, regardless of
the propensity scores status (true or estimated). This suggests a potential easy to implement performance-boost for BART, RF, and possibly other machine-learning methods as well.
5

Conclusion

Here we give a summary of this dissertation and also lay out some possible ideas for future works.

5.1 Propensity score methods for multilevel data

Chapter 2 gives a critical review and an extensive discussion of several common classes of propensity score methods in the context of multilevel data, as well as a practical guide for using them. It is important that the multilevel structure should be taken into account in the estimation; yet the choice of the method used, or of the model type (fixed effects vs random effects) depends on several factors: the data structure, the estimand of interest, and whether there is a reliable model for the outcome based on previous studies.

We also provide an R package implementing a variety of algorithms for calculating propensity score weighting estimators and their standard errors for treatment effects. Currently, `psmultilevel` offers four weighting estimators: marginal weighting, clustered weighting, stratification, and doubly-robust. Closed-form formula is available for the standard errors of the ATE DR method (Lunceford and Davidian, 2004), and we use
bootstrapping to calculate SEs for the rest of the estimators. It would be useful to extend the package to include other methods: $k - 1$ matching, preferential matching (Arpino and Cannas, 2015), matching with regression. Another area for future work is developing and incorporate into the package closed-form standard errors for other methods. Some methods already have SEs developed but without consideration towards to multi-level structure.

5.2 Sensitivity analysis in an observational study with count outcome

Unconfoundedness is a necessary assumption to draw causal inference using the potential outcome framework, but in practice, for observational studies, this is a real concern due to the potential presence of hidden bias from unmeasured confounders. Chapter 3 focuses on assessing the robustness of methods to degrees of violation of this assumption. Several researchers have studied and developed methods to conduct sensitivity analysis. We provide extensions to two previous methods to accommodate count outcomes instead of the commonly considered case of binary outcome. With this new type of outcome, more sensitivity parameters are required and they bear different interpretations. We demonstrate these two approaches in the context of traffic safety research.

The main takeaway messages from this chapter are: (1) it is important to conduct sensitivity analysis and (2) choosing a method to use depends on several factors, e.g. the estimating method being considered, the computer potential, the outcome type. It should be noted that in order to assess the unconfoundedness assumption, we actually make more assumptions along the way. This is indeed one (among other) limitation of this type of sensitivity analysis. Potential further work includes developing sensitivity analysis approaches that can be used with the empirical Bayes method, which is the golden standard and usually the method of choice in the field. Another area to explore is that real crash count data sets usually include counts of
several years for each site, and the treated sites often have different year of treatment implementation. It would be beneficial to consider the impact it has on the unmeasured confounder’s association with the treatment and outcome.

5.3 Propensity score augmented additive-interactive Gaussian process regression

In chapter 4, we introduce a new Bayesian nonparametric regression approach with Gaussian priors to estimate the potential outcome function, augmenting a previously developed method - airGP- by adding the propensity scores. We argue that this joint approach provides consistency under “good” conditions and remains robust to mis-specification of the propensity score model. We show the comparative performance of this method to common propensity score methods, BART, and random forest methods in different scenarios, including homogeneous and heterogeneous treatment effects. Adding the propensity scores appear to also boost the performance of BART and RF and should be considered when these methods are used. Another area for investigation is looking at this additional boost in other machine-learning methods.
Appendix A

Sensitivity analysis

A.1 Sensitivity parameters assumption

Here we would show that:

\[ p_{02} < p_{01} < p_{00} \implies \mathbb{E}(K \mid Z = 0, U = 1, \mathbf{X}) < \mathbb{E}(K \mid Z = 0, U = 1, \mathbf{X}) \]  \quad (A.1)

We have:

\[
P(K = 1 \mid Z = 0, U = 1, \mathbf{X}) = \frac{P(K = 1, U = 1 \mid Z = 0, \mathbf{X})}{P(U = 1 \mid Z = 0, \mathbf{X})}
\]

\[
= \frac{P(U = 1, K = 1 \mid Z = 0, \mathbf{X})}{P(U = 1, K = 0 \mid Z = 0, \mathbf{X}) + P(U = 1, K = 1 \mid Z = 0, \mathbf{X}) + P(U = 1, K = 2 \mid Z = 0, \mathbf{X})}
\]

\[
= \frac{P(U = 1 \mid K = 1, Z = 0) \gamma_1}{P(U = 1 \mid K = 0, Z = 0) \gamma_0 + P(U = 1 \mid K = 1, Z = 0) \gamma_1 + P(U = 1 \mid K = 2, Z = 0) \gamma_2}
\]

\[
= \frac{p_{01} \gamma_1}{p_{00} \gamma_0 + p_{01} \gamma_1 + p_{02} \gamma_2}, \gamma_j = P(K = j \mid Z = 0, \mathbf{X})
\]

Therefore:

\[
\mathbb{E}(K \mid Z = 0, U = 1, \mathbf{X}) = P(K = 1 \mid Z = 0, U = 1, \mathbf{X}) + 2P(K = 2 \mid Z = 0, U = 1, \mathbf{X})
\]

\[
= \frac{p_{01} \gamma_1 + 2p_{02} \gamma_2}{p_{00} \gamma_0 + p_{01} \gamma_1 + p_{02} \gamma_2}
\]
The right hand side of Equation A.1 becomes:

\[
E(K \mid Z = 0, U = 1, X) < E(K \mid Z = 0, U = 1, X)
\]

\[
\iff \frac{p_0 \gamma_1 + 2p_0 \gamma_2}{p_0 + p_1 + p_2} < \frac{(1 - p_0) \gamma_1 + 2(1 - p_0) \gamma_2}{(1 - p_0) + (1 - p_0) \gamma_1 + (1 - p_0) \gamma_2}
\]

\[
\iff [p_0 \gamma_1 + 2p_0 \gamma_2] \cdot [(1 - p_0) \gamma_0 + (1 - p_0) \gamma_1 + (1 - p_0) \gamma_2]
\]

\[
< [(1 - p_0) \gamma_1 + 2(1 - p_0) \gamma_2] \cdot [p_0 \gamma_0 + p_1 \gamma_1 + p_2 \gamma_2]
\]

\[
\iff \gamma_0 \gamma_1 p_0 (1 - p_0) + \gamma_1^2 p_0 (1 - p_0) + \gamma_1 \gamma_2 p_0 (1 - p_0)
\]

\[
+ 2 \gamma_0 \gamma_2 p_0 (1 - p_0) + 2 \gamma_1 \gamma_2 p_0 (1 - p_0) + 2 \gamma_2 p_0 (1 - p_0)
\]

\[
< \gamma_0 \gamma_1 p_0 (1 - p_0) + \gamma_1^2 p_0 (1 - p_0) + \gamma_1 \gamma_2 p_0 (1 - p_0)
\]

\[
+ 2 \gamma_0 \gamma_2 p_0 (1 - p_0) + 2 \gamma_1 \gamma_2 p_0 (1 - p_0) + 2 \gamma_2 p_0 (1 - p_0)
\]

\[
\iff \gamma_0 \gamma_1 p_0 (1 - p_0) + \gamma_1 \gamma_2 p_0 (1 - p_0) + 2 \gamma_0 \gamma_2 p_0 (1 - p_0) + 2 \gamma_1 \gamma_2 p_0 (1 - p_0)
\]

\[
< \gamma_0 \gamma_1 p_0 (1 - p_0) + \gamma_1 \gamma_2 p_0 (1 - p_0) + 2 \gamma_0 \gamma_2 p_0 (1 - p_0) + 2 \gamma_1 \gamma_2 p_0 (1 - p_0)
\]

\[
\iff \gamma_0 \gamma_1 p_0 + \gamma_1 \gamma_2 p_0 + 2 \gamma_0 \gamma_2 p_0 + 2 \gamma_1 \gamma_2 p_0
\]

\[
< \gamma_0 \gamma_1 p_0 + \gamma_1 \gamma_2 p_0 + 2 \gamma_0 \gamma_2 p_0 + 2 \gamma_1 \gamma_2 p_0
\]

\[
\iff \gamma_0 \gamma_1 p_0 + 2 \gamma_0 \gamma_2 p_0 + \gamma_1 \gamma_2 p_0 < \gamma_0 \gamma_1 p_0 + 2 \gamma_0 \gamma_2 p_0 + \gamma_1 \gamma_2 p_0
\]

(A.2)

Clearly, equation A.2 is satisfied when \(p_0 < p_1 < p_0\).

### A.2 Simulation-based: killer confounder characteristics

We fixed \(p_u = P(U = 1)\) and \(m = p_{10} - p_{11} = p_{12} - p_{10}\), and varied \(d = p_{00} - p_{01} = p_{01} - p_{02}\) as well as \(s = p_1 - p_0\) over a range of values. Let \(p_{xzy} = P(K = x | Z = y)\).
We have:

\[
p_u = [p_{10} \cdot p_{k0z1} + p_{11} \cdot p_{k1z1} + p_{12} \cdot p_{k2z1}] \cdot P(Z = 1)
  + [p_{00} \cdot p_{k0z0} + p_{01} \cdot p_{k1z0} + p_{02} \cdot p_{k2z0}] \cdot P(Z = 0)
\]

\[
= [p_{10} \cdot p_{k0z1} + (p_{10} - m) \cdot p_{k1z1} + (p_{10} - 2m) \cdot p_{k2z1}] \cdot P(Z = 1)
  + [p_{00} \cdot p_{k0z0} + (p_{00} - d) \cdot p_{k1z0} + (p_{00} - 2d) \cdot p_{k2z0}] \cdot P(Z = 0)
\]

\[
= [p_{10} - m(p_{k1z1} + 2p_{k1z1})] \cdot P(Z = 1) + [p_{00} - d(p_{k1z0} + 2p_{k1z0})] \cdot P(Z = 0)
\]  \hspace{1cm} (A.3)

Similarly we have:

\[
s = p_{11} - p_{00}
\]

\[
= p_{10} \cdot p_{k0z1} + p_{11} \cdot p_{k1z1} + p_{12} \cdot p_{k2z1} - [p_{00} \cdot p_{k0z0} + p_{01} \cdot p_{k1z0} + p_{02} \cdot p_{k2z0}]
\]

\[
= p_{10} - m(p_{k1z1} + 2p_{k1z1}) - p_{00} + d(p_{k1z0} + 2p_{k1z0})
\]  \hspace{1cm} (A.4)

Equations A.3 and A.4 form a system of two equations that can be solved for \(p_{00}\) and \(p_{10}\), and the rest of the \(p_{ij}\) can be calculated. Here we assume that \(d > 0\), but the other case \(d < 0\) is just a matter of flipping the sign.
Bibliography


103


107
Biography

Nghi Le Phuong Nguyen was born on January 20, 1988 in Saigon, Vietnam. She attended Sacramento City College in 2006 and transferred to Mount Holyoke College in 2008. She received a Bachelor of Arts in Mathematics, Magna Cum Laude from Mount Holyoke College in December 2010 and an M.S. in Applied Mathematics from University of Massachusetts, Amherst in May 2013. In 2016 she received an M.S. in Statistical Science en route to her Ph.D. in Statistical Science. She plans to graduate with her Ph.D. in September 2018, under the supervision of Dr. Fan Li. Starting in 2018, she will be joining the Duke Molecular Physiology Institute as a Biostatistician.