Robust and Efficient Bayesian Inference for Large-scale Non-probability Samples

by

Ali Rafei

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Survey and Data Science)
in The University of Michigan
2021

Doctoral Committee:

Professor Michael R. Elliott, Chair
Research Associate Professor Philip S. Boonstra
Research Professor Carol A. C. Flannagan
Professor Roderick J. A. Little
Research Associate Professor Brady T. West

Ali Rafei

arafei@umich.edu

ORCID iD: 0000-0002-1436-5671

To my Mom who inspires me the most.

ACKNOWLEDGEMENTS

My sincerest gratitude goes to my advisor, Professor Michael R. Elliott, for his generous support and for granting me a unique opportunity to conduct research under his supervision. It was through this academic journey with him that I came across this exciting thesis topic and could take part in multiple academic conferences and student award competitions. He is a very motivating advisor personally and professionally who graciously shared his time with me to discuss any problem I faced during my masters and doctoral programs. Words fail to acknowledge his support over the past five years in a convenient way that actually deserves him.

I would also like to express my immense gratitude to Professor Carol A. C. Flannagan, the co-supervisor of my masters and doctoral research, who inspired the application parts of this thesis by various insightful suggestions and discussions. Undoubtedly, I would never be able to achieve this milestone without her constant encouragement and support. In addition, I would like to thank the other respected members of my dissertation committee, Professors Roderick J. Little, Brady T. West, and Philip S. Boonstra, who have continuously supported me with their excellent comments and critical feedback. Their contribution to the improvement of the thesis contents is truly beyond my expression.

My gratitude also goes to my wonderful instructors in the doctoral seminar course, Professors Brady T. West, Stanley Presser, Katharine G. Abraham, and Joseph Sedransk, and all my classmates who greatly helped me develop my initial ideas and draft my proposal by giving me valuable comments and advice and by correcting

my writings. Furthermore, I am very grateful to all the researchers and staff who have directly or indirectly engaged in collecting and processing the data used in this dissertation, including the Safety Pilot Model Deployment (SPMD), National Household Travel Survey (NHTS), Strategic Highway Research Program II (SHRP2), Crash Injury Research Engineering Network (CIREN), and Crashworthiness Data System (CDS). Last but certainly not least, I am thankful to my parents and my sisters for their comprehension and patience throughout the time I was away from home.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

ABSTRACT

The steady decline of response rates in probability surveys, in parallel with the fast emergence of large-scale unstructured data ("Big Data"), has led to a growing interest in the use of such data for finite population inference. However, the non-probabilistic nature of their data-generating process makes big-data-based findings prone to selection bias. When the sample is unbalanced with respect to the population composition, the larger data volume amplifies the relative contribution of selection bias to total error. Existing robust approaches assume that the models governing the population structure or selection mechanism have been correctly specified. Such methods are not well-developed for outcomes that are not normally distributed and may perform poorly when there is evidence of outlying weights. In addition, their variance estimator often lacks a unified framework and relies on asymptotic theory that might not have good small-sample performance.

This dissertation proposes novel Bayesian approaches for finite population inference based on a non-probability sample where a parallel probability sample is available as the external benchmark. Bayesian inference satisfies the likelihood principle and provides a unified framework for quantifying the uncertainty of the adjusted estimates by simulating the posterior predictive distribution of the unknown parameter of interest in the population. The main objective of this thesis is to draw robust inference by weakening the modeling assumptions because the true structure of the underlying models is always unknown to the analyst. This is achieved through either combining different classes of adjustment methods, i.e. quasi-randomization and prediction modeling, or using flexible non-parametric models including Bayesian Additive Regression Trees (BART) and Gaussian Process (GP) Regression.

More specifically, I modify the idea of augmented inverse propensity weighting such that BART can be used for predicting both propensity scores and outcome variables. This offers

additional shields against model misspecification beyond the double robustness. To eliminate the need for design-based estimators, I take one further step and develop a fully model-based approach where the outcome is imputed for all non-sampled units of the population via a partially linear GP regression model. It is demonstrated that GP behaves as an optimal kernel matching tool based on the estimated propensity scores. To retain double robustness with good repeated sampling properties, I estimate the outcome and propensity scores jointly under a unified Bayesian framework. Further developments are suggested for situations where the reference sample is complex in design, and particular attention is paid to the computational scalability of the proposed methods where the population or the non-probability sample is large in size. Throughout the thesis, I assess the repeated sampling properties of the proposed methods in simulation studies and apply them to real-world non-probability sampling inference.

Keywords: doubly-robust, pseudo-weighting, prediction modeling, Bayesian Additive Regression Trees, Gaussian Process Regression.

CHAPTER I

Introduction

1.1 Non-probability samples

The 21st century is witnessing a re-emergence of non-probability sampling in various domains (Murdoch and Detsky, 2013; Daas et al., 2015; Lane, 2016; Senthilkumar et al., 2018). On the one side, probability sampling, which has dominated the survey methodology realm for decades, is facing new challenges, mainly because of a steady drop in response rates and increased costs (Groves, 2011; Johnson and Smith, 2017; Miller, 2017). On the other side, new modes of data collection using sensors, web portals, and smart devices have emerged that routinely capture a variety of human activities. These automated processes have led to an ever-accumulating massive volume of unstructured information, so-called "Big Data" (Couper, 2013; Kreuter and Peng, 2014; Japec et al., 2015). While being cheaper, larger, faster, and more detailed make Big Data appealing as an alternative or supplement to probability surveys, the non-probabilistic nature of their data-generating process introduces new impediments to valid inference for such data.

Non-probability sampling has a long history of being a cheap and timely alternative to conducting a full census of a larger population. Matching by a set of known population auxiliary totals, also termed quota selection, constitutes one of the earliest strategies to achieve balance in sampling (Kiaer, 1897; Rao and Fuller, 2017). Ever

since the role of random selection was recognized by Neyman's seminal effort in the mid-1930s, purposive methods were quickly abandoned by scientific sampling. It was through his study that the design-based mode of inference found its way to the survey methodology domain, where estimation relies entirely on the randomization distribution (Neyman, 1934). Neyman demonstrated that equal probabilities of selection are not a necessary requirement to get unbiased estimates in stratified sampling.

As a result, any sampling design in which all population units are assigned a known non-zero chance of being selected became the standard definition for probability sampling (Särndal et al., 2003). This not only negates the influence of unobserved effect modifiers on the selection mechanism (Elliott, 2016) but allows for undoing the sampling procedure by the analyst using inverse probability weighting (Narain, 1951; Horvitz and Thompson, 1952; Hájek, 1971). The history of probability sampling is replete with attempts to improve the sampling efficiency statistically and with respect to budget, logistics, and sampling frame constraints, but at the expense of additional complexity in the sample design, such as stratification and multi-stage clustering. Theories for design-based inference and variance estimation under such sampling designs have been well-documented in Kish (1965); Cochran (1977); Särndal et al. (2003) and Fuller (2011).

Since non-response and imperfect sampling frames are two unavoidable obstacles to fully random selection, parallel efforts have been devoted to developing post-survey adjustment techniques to limit the resulting bias (Holt and Smith, 1979). Weighting class adjustment and post-stratification via raking are the most commonly used methods to compensate for unit non-response and undercoverage in the large-scale surveys conducted by federal statistical agencies, which usually appear as a factor applied to the base weights (Brick and Kalton, 1996; Kalton and Flores-Cervantes, 2003; Valliant et al., 2018). A more recent add-on to post-survey adjustment involves model-assisted methods, such as general regression estimator, that incorporate fea-

tures of matching based on known control totals into probability sampling (Deville and Särndal, 1992; Deville and Tillé, 2004).

Along with these developments, there has been a persisting decline in the response rate of probability samples (Miller, 2017; Groves, 2011). According to a report by Pew Research Center, the average response rate in telephone surveys has dropped by 75% over the past two decades (Keeter et al., 2017; Dutwin and Lavrakas, 2016), and the trend for in-person household surveys is similar (Meyer et al., 2015; Williams and Brick, 2018). Researchers speculate that multiple factors, including the rising response burden from a multitude of surveys with lengthy and sophisticated instruments, busier-than-ever lifestyles, and increased privacy concerns, contribute to this downward trend (Brick and Williams, 2013). It is perhaps because of this issue that pollsters increasingly fail to predict the outcomes of the political elections in the U.S. (Forsberg, 2020; Vittert et al., 2020).

This downward trend not only imposes excess implementational costs for refusal conversion but casts doubt on the external validity of probability survey-based findings (Presser and McCulloch, 2011; Groves, 2011). In the absence of accurate auxiliary information for non-respondents, post-survey adjustments will fail to correct for the non-response bias (West and Little, 2013). Although responsive and adaptive survey designs have increased the response propensities (Groves and Heeringa, 2006; Brick and Tourangeau, 2017; Tourangeau et al., 2017), these approaches may not remain effective forever as the cost of refusal conversion continues to rise (Luiten et al., 2020).

Parallel to this paradigm, large-scale unstructured data, which I collectively term "Big Data", are becoming increasingly available thanks to the recent advances in measurement technologies (Groves, 2011; Johnson and Smith, 2017). Examples include political views shared on social media, Google searches for particular terms, payment transactions recorded by online stores, electronic health records of the patients admitted to a group of hospitals, videos captured by traffic cameras, and mobile GPS

trajectory data by satellite. This broad range of data examples share several common characteristics in terms of volume, velocity, variety, and veracity (Couper, 2013).

Though usually not targeted to answer pre-specified research questions, Big Data can address questions that are not easily answered by traditional surveys. Surveys are often cross-sectional, whereas Big Data are often longitudinal, collecting by real-time sampling (Johnson and Smith, 2017). In addition, Big Data are more cost-efficient, and unlike probability surveys, the data collection cost is not a linear function of the sample size (Tam and Clarke, 2015). Its immense size makes Big Data a rich resource for rare event studies, predictive analysis as well as small area estimation (Rao, 2015; Lohr et al., 2017; Kim et al., 2018). Because of these features, there exist a growing hope that Big Data can be used for official statistics in the near future (Struijs et al., 2014; Kitchin, 2015; Beręsewicz et al., 2018).

However, concerns have been over the use of Big Data for finite population inference (Hargittai, 2015; Buelens et al., 2014), because the mechanism of selection of their elements is often unknown and beyond the control of researchers. Unlike probability surveys, such data may lack explicit definitions of the target population, sampling frame, and the mechanism by which data elements are selected. When the sample is unbalanced with respect to the target population composition, larger data volume even increases the relative contribution of selection bias to total error (Raghunathan, 2015). In the words of Meng (2016): "the bigger the data, the more certain we will miss our target". Meng et al. (2018) call this phenomenon a "Big Data Paradox", and show that the effective sample size compared with probability sampling drops dramatically when even trivial degrees of selection bias are present.

There may be situations where conducting a probability survey may not be practical. This is usually the case in many clinical and epidemiological studies, e.g. randomized clinical trials (RCTs), though the emphasis in such studies is on internal validity rather than on generalizability to a larger population. For finite population

inference, opt-in panels are now widely used for web surveys in social and public opinion research as a cheap and fast method of data collection. The potential selection bias in such methods of sampling prompted the American Association for Public Opinion Research (AAPOR) to issue two comprehensive task force reports on the use of non-probability samples. Neither recommends the use of online panels, and the authors emphasize that these samples tend to give less accurate estimates than probability samples and that checking the required modeling assumptions is difficult (Baker et al., 2010, 2013).

The majority of the inferential methods for non-probability samples borrow their core idea from the causal inference context, where the goal is obtaining internally valid associations between exposures and an outcome variable from observational studies by removing the effects of potential confounders. Analogies between the external validity in non-probability samples, i.e. generalizability to a larger population and the internal validity in causal inference for observational data have been well-recognized (Mercer et al., 2017; Stuart et al., 2018; Kohler et al., 2019). While RCTs deal with randomly assigning the sampled units to the treatment group, in probability sampling, it is the population units that are assigned randomly to the sample. Elliott (2016) emphasizes that RCTs negate the influence of unobserved confounders, whereas probability samples negate the influence of effect modifiers.

In contrast, both observational studies and non-probability samples suffer from the fact that the corresponding assignment mechanisms are unknown to the analyst. Two widely studied strategies in causal inference encompass imputing the "potential outcome" and balancing the sample across levels of the exposure given a set of observed confounders. Both strategies rely on a "strongly ignorable" condition, under which the assignment mechanism is assumed to be completely at random with adequate observed samples at each level of confounders. Matching, stratification, and weighting are three common technical approaches to achieve balance in the second

strategy (Rubin, 1976; Rosenbaum and Rubin, 1983). While these methods are discussed in detail elsewhere throughout this thesis, substantial issues may arise when adapting the causal inference methods to a non-probability sample setting.

Adjustments for non-random sampling require auxiliary information observed for the entire population, or a probability survey representing that population (Thompson, 2019). They rely on well-specified underlying models, while the true structure of these models is almost always unknown in reality. My goal is to develop robust and efficient methods that weaken the modeling assumptions and account for the sampling design of the probability survey. By efficiency, I mean not only computation scalability, but reduced variance of the adjusted estimators as well. Note that checking this is limited to empirical assessments by comparing the length of confidence intervals. Specifically, I employ non-parametric Bayesian modeling for the prediction that not only captures non-linear associations as well as high-order interactions automatically but also allows for direct quantification of the uncertainty of the proposed estimator by simulating the posterior predictive distributions.

The focus of this thesis is on a situation where a well-designed probability sample is available as an external benchmark. For the benchmark sample, a range of sampling designs is investigated from independent selection with unequal probabilities of selection to stratified multistage cluster sampling. Throughout the thesis, I assume a *strongly ignorable* selection mechanism, given the common set of observed auxiliary variables in the probability and non-probability samples. That is the non-probability sample is assumed to arise from a probability sample design but with unknown non-zero selection probabilities, and that the auxiliary variables fully determine the selection mechanism in the non-probability sample. It is also assumed that the measurement of auxiliary variables is error-free in both samples. In the following subsection, I define these assumptions more formally.

1.2 Notation, assumptions and a general framework

Denote by U a finite population of size N, which may be known or unknown. For $i = 1, ..., N$, let y_i be the realized value of a scalar outcome variable, Y, in U, and $x_i = [x_{i1}, x_{i2}, ..., x_{ip}]^T$ the values of a p-dimensional set of relevant auxiliary variables, X. Let S_A be a non-probability sample selected from U with (x_i^T, y_i) observed and n_A being the sample size. The main objective of descriptive inference is to learn about an unknown population quantity that is a function of Y, e.g. $Q(y)$. Throughout this thesis, I consider this quantity to be the finite population mean, i.e. $Q(y) = \bar{y}_U = \sum_{i=1}^{N} y_i/N$. Suppose $\delta_i^A = I(i \in S_A)$ represents the inclusion indicator variable of S_A for $i \in U$ whose distribution can be explained by x_i. Since the selection mechanism in S_A is unknown, valid inference about $Q(y)$ requires the following strong conditions:

C1. Positivity—The nonprobability sample S_A actually does have a probabilistic sampling mechanism, albeit unknown. That means $p(\delta_i^A = 1|x_i) > 0$ for all possible values of x_i in U.

C2. Ignorability—the selection mechanism of S_A is fully governed by x, which implies that $Y \perp\!\!\!\perp \delta^A | X$. Then, for $i \in U$, the pseudo-inclusion probability associated with S_A is defined as $\pi_i^A = p(\delta_i^A = 1|x_i)$.

C3. Independence—units in S_A are selected independently given x, i.e. $\delta_i^A \perp\!\!\!\perp \delta_j^A | x_i, x_j$ for $i \neq j \in U$. This assumption is made to avoid unnecessary complications; where required, I will relax this condition by considering S_A to be clustered.

Note that **C1-C2** are collectively called a *strongly ignorable* condition by Rosenbaum and Rubin (1983).

Now, suppose S_R is a parallel reference survey of size n_R, in which the same set of covariates, X, has been measured, but Y is unobserved. Also, let $\delta_i^R = I(i \in S_R)$ denote the inclusion indicator variable associated with S_R for $i \in U$. Units of S_R

may be selected independently or through a stratified multistage cluster sampling design. Being a full probability sample implies that the selection mechanism in S_R is ignorable given its design features, i.e. $p(\delta_i^R|y_i, d_i) = p(\delta_i^R|d_i)$ for $i \in U$, where $d_i = [d_{i1}, d_{i2}, ..., d_{iq}]^T$ denotes a q-dimensional set of associated design variables. The inclusion probabilities in S_R as $\pi_i^R = p(\delta_i^R = 1|d_i)$ for $i \in U$, are known, but typically only observed for $i \in S_R$. Most often, probability survey data are accompanied by a set of sampling weights that are supposed to be inversely proportional to the selection probabilities, i.e. $w_i^R \propto 1/\pi_i^R$. The sampling weights may comprise of post-survey adjustments for ineligibility, non-response, and non-coverage errors in addition to the sampling design (Korn and Graubard, 1999; Valliant et al., 2018).

Now, I combine the two samples and define $S_C = S_A \cup S_R$ with $n_C = n_A + n_R$ being the total sample size. While X and D may overlap or correlate, in addition to the aforementioned conditions, I also assume

C4. Independence of samples— conditional on $[X, D]$, S_R and S_A are selected independently, i.e. $\delta^A \perp\!\!\!\perp \delta^R|X, D$.

Considering **C1-C4**, the joint density of y_i, δ_i^A and δ_i^R based on a "selection model" can be factorized by

$$p(y_i, \delta_i^A, \delta_i^R|x_i, d_i; \theta, \beta) = p(y_i|x_i, d_i; \theta)p(\delta_i^A|x_i; \beta)p(\delta_i^R|d_i), \quad \forall i \in U \tag{1.1}$$

where $\eta = (\theta, \beta)$ are unknown parameters indexing the conditional distribution of $Y|X, D$ and $\delta^A|X$, respectively. The conditional density $p(y_i|x_i, d_i; \theta)$ denotes the underlying model that governs the response surface structure of a superpopulation from which U has been selected. Also, $p(\delta_i^A|x_i; \beta)$ and $p(\delta_i^R|d_i)$ denote the randomization distributions associated with the selection mechanisms of S_A and S_R, respectively. Note that the latter does not depend on any unknown parameter as S_R is a probability sample with a known sampling design.

Figure 1.1 depicts the data structure in both the finite population and the combined sample. Generally, in a non-probability sample setting $S_{obs} = \{x_i, d_i, y_j, \pi_k^R | i \in S_C, j \in S_A, k \in S_R\}$ is assumed to be observed (shaded area), while $S_{mis} = \{\pi_i^A, y_j | i \in S_A, j \in S_R\}$ is missing and has to be imputed. The following subsection gives an overview of the proposed methods across the next chapters.

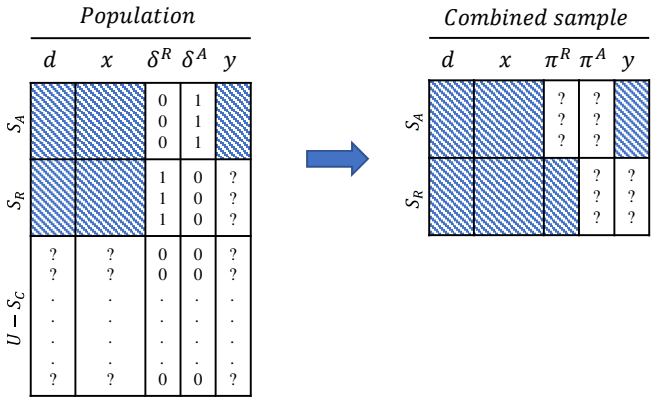

Figure 1.1: Data structure in the population and the combined sample.
Note: To simplify visualizing δ^R and δ^A, I have assumed that $S_R \cap S_A = \emptyset$.

1.3 Overview of the following chapters

In Chapter II, I develop a robust two-step Bayesian pseudo-weighting approach using Bayesian Additive Regression Trees (BART). The flexibility of BART as a predictive tool can offer a strong shield against misspecifying the selection propensity model. I compare the performance of the proposed method with those of two alternative pseudo-likelihood-based techniques in terms of repeated sampling properties in a simulation study. Under BART, point and variance estimators are obtained using Rubin's combining rules. I also demonstrate the consistency of the adjusted estimates and develop a sandwich-type variance estimator based on Generalized Linear Models (GLM) under the strongly ignorable condition. Finally, I apply the proposed method to the naturalistic driving data from the Safety Pilot Model Deployment using the 2009 National Household Travel Survey (NHTS) as a benchmark.

To further protect against model misspecification, Chapter III suggests combining the proposed pseudo-weighting method in Chapter II with a prediction model based on Augmented Inverse Propensity Weighting (AIPW). This yields double robustness that guarantees consistency in the adjusted estimator when the underlying model of either approach holds. As in Chapter II, I utilize Rubin's combining rules to obtain the point and variance estimators under BART. Under GLM, I also assess the asymptotic properties of the proposed method theoretically. The repeated sampling properties of the proposed estimator are then checked under different model specification scenarios. Considering the 2017 NHTS as a benchmark, I eventually apply the proposed method to the naturalistic driving data from the second phase of the Strategic Highway Research Program (SHRP2).

In Chapter IV, I develop a fully Bayesian approach by modeling the joint distribution of the outcome variable and the sample inclusion indicator, which allows for directly simulating the posterior predictive distribution of the unknown population quantity. This method utilizes a partially linear Gaussian process regression as the prediction model that non-parametrically links the estimated propensity scores to the response surface. I show that Gaussian process (GP) regression behaves as a non-parametric matching technique based on the estimated propensity scores, which yields double robustness and reduced sensitivity to outlying pseudo-weights. I assess the repeated sampling properties of the proposed method through Monte Carlo simulation studies and apply it to SHRP2/NHTS data to estimate police-reportable crash rates per distance unit driven in the U.S. As a second application, I estimate severe crash injury rates in different body regions in the U.S. based on the Crash Injury Research Engineering Network (CIREN) data as the non-probability sample and considering the Crashworthiness Data System (CDS) as a benchmark. Finally, in Chapter V, I highlight the main findings across these four chapters and close the dissertation by discussing limitations and potential areas for future research.

CHAPTER II

Robust Bayesian Quasi-random Inference for Non-probability Samples

2.1 Introduction

This chapter has been motivated by a desire for finite population inference based on naturalistic driving studies (NDS), which are one real-world example of Big Data for rare event investigations. Since traffic collisions are inherently rare events, measuring accurate pre-crash behaviors as well as exposure frequency in normal driving demands accurate long-term follow-up of the population of drivers. Thus, NDS are designed to continuously monitor drivers' behavior via in-vehicle sensors, cameras, and advanced wireless technologies (Guo et al., 2009). The detailed information collected by NDS are considered a rich resource for assessing various aspects of transportation such as traffic safety, crash causality, and travel patterns (Huisingh et al., 2018; Tan et al., 2017). However, because of the high administrative and technical costs, participants are usually recruited voluntarily via convenience samples from limited geographical areas. Therefore, inference based on the NDS data may suffer from selection bias.

It is apparent that classical design-based approaches cannot be applied to an NDS sample directly for making finite population inference, even though one could imagine that willingness to participate is quite random. The main reason is that

the probabilities of selection are missing in a non-probability sample and cannot be estimated from the sample itself (Chen et al., 2019). Thus, as recommended by the American Association for Public Opinion Research (AAPOR) task force on non-probability samples, adjustment methods should rely on models and external auxiliary information (Baker et al., 2013). One potential solution might be treating the NDS sample as a quasi-random sample but with unknown selection probabilities and then employing models to estimate the pseudo-inclusion probabilities for units of the NDS (Valliant and Dever, 2011; Elliott and Valliant, 2017).

Also known as quasi-randomization (QR), this method borrows the idea of propensity scores (PS) adjustment from Rosenbaum and Rubin (1983) for causal inference in observational studies. However, in a non-probability sample setting, estimating the propensity of being selected in the sample cannot be performed without the assistance of external data. In many situations, there can be found a well-designed probability sample properly representing the target population for inference. Such a sample is often termed a "reference survey". Combining the non-probability sample with a reference survey, Terhanian et al. (2000) expand the QR method to improve the potential selection bias in web surveys by fitting the propensity model on the combined sample. Since then, PS-based matching, subclassification, and inverse weighting have been widely used to adjust for the selection bias in such samples (Lee, 2006; Rivers, 2007; Lee and Valliant, 2009; Valliant and Dever, 2011; Brick, 2015).

To guarantee unbiasedness, it is critical to assume that the selection mechanism in the non-probability sample is *ignorable*, i.e. the set of common covariates is fully characterizing the selection mechanism. In the nonresponse adjustment context, Little and Vartivarian (2005) emphasize that adjustments are effective in bias reduction as long as the auxiliary variables are strongly associated with both the analytic variable of interest and nonresponse mechanism; otherwise, they will only inflate the variance without substantial reduction in bias. It is also essential to correctly specify

the underlying model governing the selection mechanism of population units in the sample. A big challenge arises from the fact that the true propensity model is almost always unknown to the analyst in a non-probability sample setting. In addition, estimating the PS requires auxiliary information to be available for the entire population units, while external data are often limited to a reference survey (Zhang, 2019).

When units in the reference survey are selected independently with unequal inclusion probabilities, to predict the PS, Valliant et al. (2018)[p.565-603] suggest using a weighted logistic regression with a pseudo-maximum likelihood estimation (PMLE) of parameters. Recently, an alternative expansion of the pseudo-likelihood function is given by Chen et al. (2019) for the PS model, where the population-level term is replaced by its Horvitz Thompson (HT)-estimator from the reference survey. However, PMLE is a design-based method where the population *log*-likelihood function is approximated by its weighted estimate from the sample. Despite being design-consistent, the solutions of the estimating equations may not be efficient when the sampling weights are highly variable (Little, 2004). More importantly, this parametric method is limited to the likelihood-based models with an exponential family, so it cannot be applied to a broader range of predictive methods. In situations where the true propensity model is unknown or variable selection is a hurdle because of high-dimensional covariates, one might hope to be able to use more flexible predictive methods such as modern tree-based methods, which automatically perform variable selection and take into account non-linear associations and high-order interactions.

Alternatively, Elliott et al. (2010) propose a two-step approach where pseudo-inclusion probabilities are directly derived by multiply applying the Bayes rule. This method, which I term as propensity-adjusted probability prediction (PAPP), computationally separates the sampling weights from the propensity model. This can be especially advantageous when the goal is applying a broader range of predictive methods, such as algorithmic tree-based methods, for the PS prediction. Being able to use

more flexible non-parametric predictive methods helps us further protect against misspecifying the QR model. This chapter aims to employ Bayesian Additive Regression Trees (BART) to predict the PS based on the PAPP method. BART provides a strong predictive tool by automatically handling complex associations as well as multi-way interactions (Chipman et al., 2007). The idea of BART is based on the sum-of-trees regression approximating the outcome variable as an unspecified function of predictors. However, to avoid trees from overfitting, a set of prior distributions is assigned to the trees' structure as well as parameters in the terminal nodes. Given the data, these priors are updated through a Bayesian backfitting Monte Carlo Markov Chain (MCMC) algorithm (Chipman et al., 2010).

BART is especially desirable for high-dimensional data where variable selection is a big challenge (Hill et al., 2011; Spertus and Normand, 2018). In addition, the posterior predictive distribution produced by BART makes it easier to quantify the uncertainty due to the pseudo-weights (Tan et al., 2019). BART has advantages in PS adjustment in the presence of heterogeneous treatment effects (Kern et al., 2016; Hahn et al., 2020; Wendling et al., 2018). Mercer (2018) recently employed BART to adjust for selection bias in non-probability samples, but in order to properly account for the sampling weights in the adjustments, the author employs a weighted Bayesian bootstrap technique to multiply impute the auxiliary variables for the non-sampled units of the population. When the non-probability sample or the finite population is large in size, such a method may not be tractable computationally as one has to fit BART repeatedly on the simulated synthetic populations. In addition, Tan et al. (2019) exploits BART to compare different adjustment methods including inverse propensity weighting in an item-missing imputation setting.

When BART is applied, I employ a two-step Bayesian framework where PS are multiply imputed as the first step using the posterior predictive draws simulated by BART, and then Rubin's combining rules are used to aggregate the pseudo-weighted

estimates for the construction of point and interval estimates (Kaplan and Chen, 2012; Rubin, 1976). For the application of this chapter, I am interested in generating a set of pseudo-weights for the Safety Pilot Model Deployment (SPMD) sample, which is a large-scale NDS. Participants in the SPMD have been selected through a combination of convenience and snowball sampling, and geographically, the sample is limited to the Ann Arbor area. Therefore, the SPMD sample may not be representative of the population of U.S. drivers. In particular, I use the 2009 National Household Travel Survey (NHTS) as the reference sample, which is a nationally representative telephone survey of the U.S. population. My goal is to develop a set of pseudo-weights that can be used for improving the generalizability of the sample in any SPMD-specific study. I evaluate the performance of estimated pseudo-weights by comparing the weighted estimates for some trip-related measures in both SPMD and NHTS, as well as in a Monte Carlo simulation study.

The rest of the chapter is organized as follows: In Section 2.2, I start by reviewing the theoretical background behind the QR approaches including PAPP and PMLE with additional proofs given in Appendix 2.6. Variance estimation under pseudo-weighting and *ad hoc* methods of weight trimming are also discussed in this section. In Section 2.3, I provide a simulation study to evaluate the proposed methods. Section 2.4 describes the data and auxiliary variables I used in the current chapter and presents the results of pseudo-weighting on SPMD data at the individual level. Finally, Section 2.5 reviews the strengths and weaknesses of the study in more detail and suggests some future research directions.

2.2 Methods

2.2.1 Quasi-randomization

Consider the assumptions **C1-C4** determined in Section 1.2. To simplify the notation, I define $x_i^* = [x_i, d_i]$, a $(p+q)$-dimensional vector of all auxiliary variables associated with S_A and S_R. Now, suppose S_A and S_R have trivial overlap, i.e. $p(\delta_i^A + \delta_i^R = 2) \approx 0$. This assumption is reasonable when the sampling fraction in both samples is small. Note that under the *ignorable* assumption, the propensity model for S_A depends on X observed for the entire population. Thus, given the combined sample, $S_C = S_A \cup S_R$, with $n_C = n_A + n_R$ being the sample size, it is reasonable to expect that the pseudo-inclusion probabilities, π_i^A's, are a function of both x_i and d_i for $i \in S_C$. Let $z_i = I(i \in S_A | \delta_i = 1)$ be the indicator of subject i belonging to the non-probability sample in the combined sample where $\delta_i = \delta_i^A + \delta_i^R$. Note that since $S_A \cap S_R = \emptyset$, δ_i can take values of either 0 or 1 as below:

$$\delta_i = \begin{cases} 0, & \text{if } \delta_i^R = 0 \quad and \quad \delta_i^A = 0 \\ 1, & \text{if } \delta_i^R = 1 \quad or \quad \delta_i^A = 1 \end{cases}$$

As discussed earlier, in quasi-randomization (QR), S_A is treated as if the self-selection mechanism of the population units mimics a stochastic process, but with unknown selection probabilities. Then, attempts are made to estimate these missing quantities in S_A based on the external auxiliary information, which involves modeling $f(\delta_i^A | x_i; \beta)$ in Eq. 1.1. However, this requires full knowledge about x_i for the entire population.

Suppose X is linearly associated with the *logit* of the unknown selection probabilities of S_A in U. I have

$$log\left\{\frac{\pi^A(x_i)}{1-\pi^A(x_i)}\right\} = \beta_0 + x_i^T \beta_1 \quad i \in U \tag{2.1}$$

where $\beta^T = [\beta_0, \beta_1^T]$ is a set of $p+1$ unknown parameters. Therefore, the true propensity scores (PS) in U are given by

$$\pi^A(x_i) = e(x_i;\beta) = \frac{exp\{\beta_0 + x_i^T\beta_1\}}{1 + exp\{\beta_0 + x_i^T\beta_1\}} \quad i \in U \tag{2.2}$$

Assuming that δ_i^A follows *Bernoulli* distribution with the success probability π_i^A, which is the case under a Poisson sampling design, the likelihood function of β given the selection indicators in U, $\delta_U^A = [\delta_1^A, \delta_2^A, ..., \delta_N^A]^T$, is given by

$$L(\beta; \delta_U^A | x_U) \prod_{i=1}^{N} e(x_i;\beta)^{\delta_i^A}[1 - e(x_i;\beta)]^{1-\delta_i^A} \tag{2.3}$$

where x_U is a design matrix defined across the units of U. The *log*-likelihood function is then obtained by taking the *log* transformation as below:

$$\begin{aligned} l(\beta; \delta_U^A | x_U) &= \sum_{i=1}^{N} \delta_i^A log\{e(x_i,\beta)\} + \sum_{i=1}^{N}(1-\delta_i^A) log\{1 - e(x_i,\beta)\} \\ &= \sum_{i=1}^{N} \delta_i^A log\left\{\frac{e(x_i,\beta)}{1-e(x_i,\beta)}\right\} + \sum_{i=1}^{N} log\{1 - e(x_i,\beta)\} \end{aligned} \tag{2.4}$$

As seen, the first term in Eq. 2.4 is reduced to sum over $i \in S_A$, because $\delta_i^A = 0$ for $i \in \bar{S}_A$, but the second term still depends on $i \in U$. Since π_i^R's are known for S_R, Chen et al. (2019) suggest replacing the second term with its HT-estimator, which is design-consistent, from S_R, i.e.

$$\begin{aligned} l^*(\beta; x) &= \sum_{i=1}^{n_A} log\left\{\frac{e(x_i,\beta)}{1-e(x_i,\beta)}\right\} + \sum_{i=1}^{n_R} log\{1 - e(x_i,\beta)\}/\pi_i^R \\ &= \sum_{i=1}^{n_A} x_i^T\beta - \sum_{i=1}^{n_R} log\{1 + exp(\beta_0 + x_i^T\beta_1)\}/\pi_i^R \end{aligned} \tag{2.5}$$

Taking the first-order derivative with respect to β yields a *score* function, and a pseudo-maximum likelihood estimation (PMLE) of model parameters is obtained by

solving the following estimating equations:

$$U(\beta) = \frac{\partial l^*}{\partial \beta} = \sum_{i=1}^{n_A} x_i - \sum_{i=1}^{n_R} e(x_i, \beta) x_i / \pi_i^R = 0 \tag{2.6}$$

Since Eq. 2.6 is non-linear with respect to β, a numerical method such as the Newton-Raphson iterative procedure is needed to solve such estimating equations. As illustrated, these equations depend only on $i \in S_C$. Therefore, the estimate of the π_i^A's is obtained by plugging the solution of Eq. 2.6, i.e. $\hat{\beta}$, into Eq. 2.2. I call this approach PMLE-C for short.

Alternatively, Valliant and Dever (2011) recommend modeling Z_i for $i \in S_C$ using a weighted logistic regression where weights are given by

$$w_i^* = \begin{cases} w_i^R, & \text{for } i \in S_R \\ 1, & \text{for } i \in S_A \end{cases} \tag{2.7}$$

The rationale behind this approach is that units in S_R should be weighted up such that S_R properly represents the non-sampled part of the population, i.e. $S_{\bar{B}}$. Thus, the sampling weights are suggested to be normalized such that $\sum_{i=1}^{n_R} w_i^R = N - n_A$ (Valliant et al., 2018, p. 574). As a notable advantage, such a model can be implemented by standard software that supports complex sample analysis. The pseudo-log-likelihood function is then given by

$$\begin{aligned} l^{**}(\beta; z|x) &= \sum_{i=1}^{n_C} w_i^* z_i log\{e(x_i, \beta)\} + \sum_{i=1}^{n_C} w_i^* (1 - z_i) log\{1 - e(x_i, \beta)\} \\ &= \sum_{i=1}^{n_A} log\left\{ \frac{e(x_i, \beta)}{1 - e(x_i, \beta)} \right\} + \sum_{i=1}^{n_A} log\{1 - e(x_i, \beta)\} + \sum_{i=1}^{n_R} log\{1 - e(x_i, \beta)\} / \pi_i^R \end{aligned} \tag{2.8}$$

which leads to solving a different system of estimating equations as below:

$$\begin{aligned} U^*(\beta) &= \sum_{i=1}^{n_C} w_i^* x_i [z_i - e(x_i, \beta)] \\ &= \sum_{i=1}^{n_A} x_i - \sum_{i=1}^{n_R} x_i e(x_i, \beta)/\pi_i^R - \sum_{i=1}^{n_A} x_i e(x_i, \beta) = 0 \end{aligned} \tag{2.9}$$

Once $\hat{\beta}$ is obtained, Wang et al. (2020c) suggest estimating the pseudo-inclusion probabilities in S_A by $\hat{\pi}_i^A \propto e(x_i; \hat{\beta})/[1 - e(x_i; \hat{\beta})]$. This is because the use of *odds* transformation eliminates the duplicated units of S_A in the pseudo-population created by the w_i^R's. I abbreviate this approach as PMLE-V. Note that both approaches lead to a standard logistic regression problem if S_R is a simple random sample (SRS). As long as the QR model is correctly specified, one can show that the inverse PS weighted (IPSW) mean from S_A yields a consistent and asymptotically unbiased estimate for the population mean under certain regularity conditions. Sandwich-type variance estimators under these PMLE approaches have been also developed in Chen et al. (2019) and Wang et al. (2020c).

With one additional assumption, which is mutual exclusiveness of the two samples, i.e. $S_A \cap S_R = \emptyset$, I show that estimating π_i^A's can be reduced to modeling Z_i for $i \in S_C$, but eliminates the need for constructing a pseudo-likelihood function. Intuitively, one can view the selection process of the i-th population unit in S_A as being initially selected in the joint sample ($\delta_i = 1$) and then being selected in S_A given the combined sample ($Z_i = 1$). By conditioning on x_i^*, the selection probabilities in S_A are factorized as

$$\begin{aligned} p(\delta_i^A = 1 | x_i^*) &= p(\delta_i^A = 1, \delta_i = 1 | x_i^*) \\ &= p(\delta_i^A = 1 | \delta_i = 1, x_i^*) p(\delta_i = 1 | x_i^*) \\ &= p(Z_i = 1 | x_i^*) p(\delta_i = 1 | x_i^*) \quad i \in S \end{aligned} \tag{2.10}$$

Note that the last expression in Eq. 2.2 results from the definition of Z_i given S_C. The same factorization can be derived for the selection probabilities in S_R. Thus, we

have

$$p(\delta_i^R = 1|x_i^*) = p(Z_i = 0|x_i^*)p(\delta_i = 1|x_i^*) \tag{2.11}$$

By dividing the two sides of the equations 2.2 and 2.3, one can get rid of $p(\delta_i = 1|x_i^*)$ and obtain the pseudo-selection probabilities in S_A as below:

$$p(\delta_i^A = 1|x_i^*) = p(\delta_i^R = 1|x_i^*)\frac{p(Z_i = 1|x_i^*)}{p(Z_i = 0|x_i^*)} \tag{2.12}$$

It is clear that $p(\delta_i^R = 1|x_i^*) = \pi_i^R$ as x_i^* contains d_i and the sampling design of S_R is known given d_i.

Note that Eq. 2.12 is identical to the pseudo-weighting formula Elliott and Valliant (2017) derive for a non-probability sample. Unlike the PMLE approach, modeling Z_i in S_C can be performed using the standard binary logistic regression or any alternative classification methods, such as supervised machine learning algorithms. Under a logistic regression model, I have

$$p(Z_i = 1|x_i^*) = \frac{\exp\{\beta_0 + \beta_1^T x_i^*\}}{1 + \exp\{\beta_0 + \beta_1^T x_i^*\}} \tag{2.13}$$

where β denotes the vector of model parameters being estimated via maximum likelihood estimation (MLE). Hence, in situations where π_i^R is known or can be calculated for $i \in S_A$, the estimate of π_i^A for $i \in S_A$ is given by

$$\hat{\pi}_i^A = \pi_i^R \exp\{\hat{\beta}_0 + \hat{\beta}_1^T x_i^*\} = \pi_i^R \frac{p_i(\hat{\beta})}{1 - p_i(\hat{\beta})} \tag{2.14}$$

where $\hat{\beta}$ denotes the MLE estimate of the logistic regression model parameters, and $p_i(\hat{\beta})$ is a shorthand of $p(Z_i = 1|x_i^*; \hat{\beta})$. Intuitively, one can envision that the first factor in 2.14 treats S_A as if it is selected under the design of S_R, and the second

factor attempts to balance the distribution of x in S_A with respect to that in S_R.

Having π_i^A estimated based on 2.14 for all $i \in S_A$, one can construct the Hájek-type pseudo-weighted estimator for the finite population mean as below:

$$\hat{\bar{y}}_{PAPW} = \frac{1}{\hat{N}_A} \sum_{i=1}^{n_A} \frac{y_i}{\hat{\pi}_i^A} \tag{2.15}$$

where $\hat{N}_A = \sum_{i=1}^{n_A} 1/\hat{\pi}_i^A$. Hereafter, I refer to the estimator in 2.8 as propensity-adjusted probability weighting (PAPW). Under mild regularity conditions, the ignorable assumption in S_A given x, the logistic regression model and the additional assumption of $S_A \cap S_R = \emptyset$, Appendix 2.6.1 shows that this estimator is consistent and asymptotically unbiased for $\bar{y}_U$. Further, when π_i^R is known, the sandwich-type variance estimator for $\hat{\bar{y}}_{PAPW}$ is given by

$$\begin{aligned} \widehat{Var}\left(\hat{\bar{y}}_{PAPW}\right) = & \frac{1}{N^2} \sum_{i=1}^{n_A} \left\{1 - \hat{\pi}_i^A\right\} \left(\frac{y_i - \hat{\bar{y}}_{PAPW}}{\hat{\pi}_i^A}\right)^2 \\ & - 2\frac{\hat{b}^T}{N^2} \sum_{i=1}^{n_A} \left\{1 - p_i(\hat{\beta})\right\} \left(\frac{y_i - \hat{\bar{y}}_{PAPW}}{\hat{\pi}_i^A}\right) x_i^* \\ & + \hat{b}^T \left[\frac{1}{N^2} \sum_{i=1}^{n} p_i(\hat{\beta}) x_i^* x_i^{*T}\right] \hat{b} \end{aligned} \tag{2.16}$$

where

$$\hat{b}^T = \left\{\frac{1}{N} \sum_{i=1}^{n_A} \left(\frac{y_i - \hat{\bar{y}}_{PAPW}}{\hat{\pi}_i^A}\right) x_i^{*T}\right\} \left\{\frac{1}{N} \sum_{i=1}^{n} p_i(\hat{\beta}) x_i^* x_i^{*T}\right\}^{-1} \tag{2.17}$$

and $\hat{\pi}_i^A$ is the estimated pseudo-selection probability based on Eq. 2.14 for $i \in S_A$. See Appendix 2.6.1 for the derivation.

In situations where π_i^R is unknown for $i \in S_A$, Elliott and Valliant (2017) suggest predicting this quantity for units of the non-probability sample. Note that, in this situation, it is no longer required to condition on d_i in addition to x_i. Treating π_i^R as a random variable for $i \in S_A$ conditional on x_i, one can obtain this quantity by regressing the π_i^R's on the x_i's in the reference survey. According to Pfeffermann and

Sverchkov (2009), I have

$$\begin{aligned} p(\delta_i^R = 1|x_i) &= \int_0^1 p(\delta_i^R = 1|\pi_i^R, x_i) p(\pi_i^R|x_i) d\pi_i^R \\ &= \int_0^1 \pi_i^R p(\pi_i^R|x_i) d\pi_i^R \\ &= E(\pi_i^R|x_i) \quad i \in S_R. \end{aligned} \tag{2.18}$$

However, since the outcome is continuous bounded taking values within $(0, 1)$, fitting a *Beta* regression model is recommended (Ferrari and Cribari-Neto, 2004). An alternative approach (not considered here) would use the equality $P(\delta_i^R|x_i) = E^{-1}(w_i^R|x_i, \delta_i^R = 1)$ for $w_i^R = 1/\pi_i^R$ to model w_i^R rather than π_i^R (Pfeffermann and Sverchkov, 1999a). Note that, π_i^R is fixed given d_i as S_R is a probability sample, but conditional on x_i, π_i^R can be regarded as a random variable.

I call this approach propensity-adjusted probability prediction (PAPP). This two-step derivation of pseudo-inclusion probabilities is especially useful, as it separates sampling weights in S_R from the propensity model computationally. When the true model is unknown, this feature enables us to fit a broader and more flexible range of models, such as algorithmic tree-based methods. It is worth noting that modeling $E(\pi_i^R|x_i)$ does not impose an additional ignorable assumption in S_R given x, because in the extreme case if $\delta_i^R \perp\!\!\!\perp x_i$, that means weighted and unweighted distributions of x are identical in S_R, and therefore the π_i^R's can be safely ignored in propensity modeling. when the distribution of x is identical in the two samples, $\hat{e}_i/(1 - \hat{e}_i)$ will become a constant for all the units in S_A. Then, I can assume the two samples are matched, so the only action to be taken is predicting selection probabilities for units in S_A. If S_R is drawn under SRS, then, $\hat{\pi}_i^R$ will be fixed for units in S_A, so all I need to do is balancing the distribution of x in S_A with respect to that in S_R by estimating the inverse of the odds of being in the probability sample.

In situations where π_i^R is incalculable for $i \in S_A$, deriving a sandwich-type vari-

ance estimator for the PAPP-based mean becomes complicated. Therefore, to incorporate the uncertainties due to both sampling and estimating the pseudo-weights into variance estimation, I use a delete-*one* jackknife repeated replication (JRR) method proposed by Elliott (2009). To this end, I initially treat the samples in S_C as two independent strata. At each replication i ($i = 1, 2, ..., n_C$), I then drop the i-th observation from either stratum and re-estimate the PAPP-based mean denoted by $\hat{\bar{y}}^{(i)}_{PAPP}$. The variance estimator is then given by

$$\widehat{Var}(\hat{\bar{y}}_{PAPP}) = \frac{n_R - 1}{n_R} \sum_{i=1}^{n_R} (\hat{\bar{y}}^{(i)}_{PAPP} - \hat{\bar{y}}_{PAPP})^2 + \frac{n_A - 1}{n_A} \sum_{i=1}^{n_A} (\hat{\bar{y}}^{(i)}_{PAPP} - \hat{\bar{y}}_{PAPP})^2 \quad (2.19)$$

where $\hat{\bar{y}}^{(i)}_{PAPP}$ is the pseudo-weighted sample mean after deleting observation i and $\hat{\bar{y}}_{PAPP} = \sum_{i=1}^{n} \hat{\bar{y}}^{(i)}_{PAPP}/n_C$.

2.2.2 Bayesian Additive Regression Trees

BART is a flexible ensemble of trees method, which allows handling non-linear relationships as well as multi-way interaction effects. The idea of BART is based on the sum-of-trees, where trees are sequentially modified on the basis of residuals from the other trees. In a tree-based method, the variation in the response variable is explained by hierarchically splitting the sample into more homogeneous subgroups (Green and Kern, 2012). As illustrated in Figure 2.1, a binary-structured tree consists of a root node, a set of interior nodes, a set of terminal nodes associated with parameters and decision rules that links these nodes (Abu-Nimeh et al., 2008).

2.2.2.1 BART for continuous outcomes

Suppose $y = f(x) + \epsilon$ as is the case in every statistical model, where $y \in \mathbb{R}$ is a continuous outcome, x denotes an $n \times p$ matrix of covariates, and $\epsilon \sim N(0, \sigma^2)$ is the

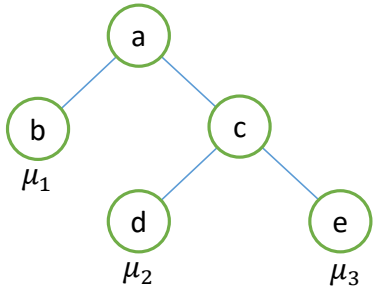

Figure 2.1: Example of a binary-structured trees model

error term. BART will then approximate the outcome as below:

$$y \approx \sum_{j=1}^{m} f(x, T_j, M_j) \tag{2.20}$$

where T_j is the j-th tree with b_j terminal nodes, and associated $M_j = (\mu_{1j}, \mu_{2j}, \ldots, \mu_{b_j j})^T$ parameters. BART is a Bayesian approach, since it assigns prior distributions to T, M, and σ (Chipman et al., 2010; Tan et al., 2016). Assuming an independence structure between trees, we can define the prior as follows:

$$p\left[(T_1, M_1), ..., (T_m, M_m), \sigma^{-2}\right] = \left[\prod_{j=1}^{m} p(T_j, M_j)\right] p(\sigma^{-2}) \tag{2.21}$$

Using the multiplication law of probability, the joint distribution of $p(T_j, M_j)$ can be written as:

$$\begin{aligned} p(T_j, M_j) &= p(M_j|T_j)p(T_j) \\ &= \prod_{i=1}^{b_j} p(\mu_{ij}|T_j)p(T_j) \end{aligned} \tag{2.22}$$

where $i = 1, ..., b_j$ denotes the terminal node parameters for tree j. Therefore, the joint distribution in 2.21 can be factored as below:

$$p\left[(T_1, M_1), ..., (T_m, M_m), \sigma^{-2}\right] = \left[\prod_{j=1}^{m} \left\{ \prod_{i=1}^{b_j} p(\mu_{ij}|T_j) \right\} p(T_j)\right] p(\sigma^{-2}) \tag{2.23}$$

Suggested by Chipman et al. (2007), the following distributions can be used for $\mu_{ij}|T_j$ and σ^{-2}:

$$\mu_{ij}|T_j \sim N(\mu_\mu, \sigma_\mu^2) \tag{2.24}$$

$$\sigma^{-2} \sim G(\frac{\nu}{2}, \frac{\nu\lambda}{2}) \tag{2.25}$$

The prior for T_j involves three components of the tree structure: length of the tree, decision rules, and the choice of covariate at a given node. However, prior specification for T_j depends on several factors, and detailed discussions can be found in Chipman et al. (2010). Given the data, these parameters are updated through a combination of the "Bayesian backfitting" and MCMC Gibbs sampler method. The trained trees are then summed up to approximate the outcome variable. Finally, m is typically assumed to be fixed but can be assessed by cross-validation.

2.2.2.2 BART for binary outcomes

For the binary outcome, a *probit* link function is usually employed in the sense that y is an indicator variable dichotomizing a normally distributed latent continuous outcome like y^* at a real value c so that:

$$y = \begin{cases} 1 & \mathrm{y}^* > \mathrm{c} \\ 0 & \mathrm{y}^* \leq \mathrm{c} \end{cases}, \quad y^* \sim N(0,1) \tag{2.26}$$

Therefore, the new model will be given by:

$$G(x) = \Phi^{-1}[p(y=1|x)] = \sum_{j=1}^{m} f(x, T_j, M_j) \tag{2.27}$$

where $\Phi^{-1}[.]$ is the inverse of standard normal CDF. Since we implicitly assumed $\sigma \equiv 1$, the only priors we need to specify are $p(\mu_{ij}|T_j)$ and $p(T_j)$. In order to be able

to draw the posterior distribution of T_j and μ_{ij}, we need to generate the latent continuous variable, y^*, given y_k. Chipman et al. (2010) recommends a data augmentation method based on the following algorithm:

$$y_k^* = \begin{cases} \max(\Phi(G(x_k)) & \text{if} \quad y_k = 1 \\ \min(\Phi(G(x_k)) & \text{if} \quad y_k = 0 \end{cases} \tag{2.28}$$

Since the structure of priors is very similar to BART for continuous outcomes (Tan et al., 2016), we update the estimates $G(x_k)$ after drawing samples from T_j's and μ_{ij}'s. To apply BART in this chapter, I utilize the '*BayesTree*' and '*BART*' packages in *R*.

2.2.3 A robust two-step Bayesian approach using BART

A two-step Bayesian approach views the problem as a multiple imputation scenario, which involves two sequential steps: (1) *design*—where the unknown pseudo-weights are multiply imputed, and (2) *analysis*—where the population unknown quantity is estimated given any set of imputed pseudo-weights. Rubin's combining rules are then employed to aggregate them for the construction of both point and variance estimates (Rubin, 1976). Although there is no explicit modeling for the outcome, this approach separates the QR model from the outcome model. This precludes the notorious feedback between the two models that occurs when jointly estimating the PS and outcome variable(s), which negatively impacts the estimate of model parameters under QR (Zigler, 2016; Zigler et al., 2013).

For the first step, one can use the posterior predictive distribution simulated by BART to multiply impute the pseudo-weights in S_A based on the PAPP method. This chapter considers the more complicated situation where π_i^R is not calculable for units of S_A. As will be seen in Section 2.4, this is the case in the empirical study. Under this setting, one can use BART for modeling both π_i^R and Z_i given the common auxiliary

variables x_i. Suppose BART approximates $p(Z_i = 1|x_i)$ by an arbitrary function h based on the data augmentation technique described in Section 2.2.2.2 as below:

$$\Phi^{-1}[p(Z_i = 1|x_i)] = h(x_i) \qquad \forall i \in S \tag{2.29}$$

where Φ^{-1} is the inverse CDF of the standard normal distribution.

For modeling π_i^R, I first employ a *logit* transformation to map the values of π_i^R from $(0, 1)$ to $\mathbb{R}$. By applying BART to a continuous outcome, I have

$$\log\left(\frac{\pi_i^R}{1-\pi_i^R}\right) = k(x_i) + \epsilon_i \qquad \forall i \in S_R \tag{2.30}$$

where k is a sum-of-trees function approximated by BART. I denote these functions by $h^{(m)}(.)$ and $k^{(m)}(.)$ for the m-th draw of the posterior distribution ($m = 1, 2, ..., M$), respectively. Then, for the m-th imputation, a posterior predictive draw for π_i^A is given by

$$\hat{\pi}_i^{A(m)} = \left\{\frac{\exp[k^{(m)}(x_i)]}{1+\exp[k^{(m)}(x_i)]}\right\}\left\{\frac{\Phi[h^{(m)}(x_i)]}{1-\Phi[h^{(m)}(x_i)]}\right\} \tag{2.31}$$

For a known N, the final estimate of the population mean is given by

$$\hat{\bar{y}}_{PAPP} = \frac{1}{M}\sum_{m=1}^{M}\hat{\bar{y}}_{PAPP}^{(m)} \tag{2.32}$$

where

$$\hat{\bar{y}}_{PAPP}^{(m)} = \frac{1}{\hat{N}}\sum_{i=1}^{n_A}\frac{y_i}{\hat{\pi}_i^{A(m)}} \tag{2.33}$$

When N is unknown to the analyst, it can be estimated by $\hat{N}_A^{(m)} = \sum_{i=1}^{n_A} 1/\hat{\pi}_i^{A(m)}$ for the m-th imputation in $\hat{\bar{y}}_{PAPP}^{(m)}$. This estimator is expected to be approximately unbiased for the population mean even when the true functional form of the QR model is unknown to the analyst.

The variance estimator for the estimator in 2.32 can be obtained using the Rubin's

combining rules for finite multiple imputation (Rubin, 2004) as below:

$$\widehat{Var}(\bar{y}_{PAPP}) = \bar{V}_W + (1 + \frac{1}{M})V_B \tag{2.34}$$

where $\bar{V}_W = \sum_{m=1}^{M} var(\bar{y}_{PAPP}^{(m)})/M$, $V_B = \sum_{m=1}^{M}(\bar{y}_{PAPP}^{(m)} - \hat{\bar{y}}_{PAPP})^2/(M-1)$, and $\hat{\bar{y}}_{PAPP}$ is given by Eq. 2.32. In the current article, this method was used for variance estimation for the PAPP under BART, and JRR was used for all other approaches.

2.2.4 Weight trimming

As discussed before, pseudo-weighting sometimes tends to produce highly extreme weights when either S_R or S_A lacks adequate observations for some levels of x. Trimming is a potential solution, in which "influential" weights are identified and modified. The most commonly used type of trimming is the *ad hoc* method, where weights above a pre-specified cut-off point are forced to that value, and the outstanding weights are redistributed across the rest of the units. The choice of the cut-off point is controversial, but the existing options are reviewed and evaluated in Chen et al. (2017b). Here I consider two methods to find such a cut-point: (1) the contribution to entropy procedure and (2) median plus multiple of the interquartile range. The entropy procedure ("Trim 1") compares the contribution of each weight to the sampling variance. This is performed by systematically comparing the individual weights with a constant value computed by the average of the square weights of the sample. In this method, the cut-point is defined as:

$$K_n = \sqrt{c\sum_{i=1}^{n_A} \hat{w}_i^2/n_A}, \quad i \in S_A \tag{2.35}$$

where c is an arbitrary constant and can be chosen empirically, and $\hat{w}_i = 1/\hat{\pi}_i^A$. In the present study, I set $c = 5$. The median plus multiple of the interquartile range

method ("Trim 2") detects outliers by assuming a symmetric distribution for the analytic variable. To construct the cut-point, I may use 4 or 5 times the interquartile range (IQR). Therefore, the cut-point of detecting extreme weights will be:

$$K_n = \hat{w}_i + c \times (Q_3 - Q_1) \tag{2.36}$$

where $\tilde{w}$, Q_1, and Q_3 are the median, 1st and 3rd quartiles of pseudo-weights, respectively, and c is again an arbitrary constant and can be set to 5 (Potter and Zheng, 2015). Alternative model-based weight trimming methods using random effects or variable selection methods can be considered as well (Elliott and Little, 2000), although I do not pursue such approaches here.

2.3 Simulation study

I designed a simulation study to evaluate the performance of the proposed PAPP in this article and to compare it with the pseudo-likelihood method proposed by Valliant et al. (2018) (PMLE-V) and its extension proposed by Chen et al. (2019) (PMLE-C) in terms of improvement rates in selection bias and other repeated sampling properties. For a better assessment of BART against the alternative models, non-linear associations including quadratic terms as well as interactions were taken into account in constructing the variables.

2.3.1 Simulation design

First, I generated a hypothetical population of size $N = 100,000$ with two sets of dependent covariates, $D = \{D_1, D_2\}$ and $X = \{X_1, X_2\}$, from a multivariate normal

distribution with the following correlation structure:

$$\begin{pmatrix} D_1 \\ D_2 \\ X_1 \\ X_2 \end{pmatrix} \sim MVN \left(\begin{pmatrix} 0 \\ 1 \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 & -\rho/2 & \rho & -\rho/2 \\ -\rho/2 & 1 & -\rho/2 & \rho \\ \rho & -\rho/2 & 1 & -\rho/2 \\ -\rho/2 & \rho & -\rho/2 & 1 \end{pmatrix} \right) \tag{2.37}$$

where X represents a set of observed common covariates that are associated with both the outcome variables of interest and the selection indicator in the non-probability sample, S_A, and D corresponds to a set of design features in the reference survey, S_R. Note that ρ characterizes how strongly the design variables of S_R are associated with those of S_A; I initially set $\rho = 0.8$, but later considered other values ranging from 0 to 0.9. Given X, I specified a continuous response variables, Y^c, and a binary response variable, Y^b, in the population as below:

$$Y_i^c | X_i = x_i \sim N(\mu = -2 + x_{1i} - 2x_{2i} + 3x_{1i}x_{2i}, \sigma^2 = 1) \tag{2.38}$$

$$Y_i^b | X_i = x_i \sim BER\left(p = \frac{e^{2-x_{1i}+2x_{2i}-3x_{1i}x_{2i}}}{1 + e^{2-x_{1i}+2x_{2i}-3x_{1i}x_{2i}}}\right) \tag{2.39}$$

To draw samples corresponding to NHTS and SPMD from the hypothetical population, I considered an informative sampling strategy with unequal probabilities of inclusion, where the selection mechanism is given through a *logistic* function as below:

$$p(\delta_i^R = 1 | D_i = d_i) = \frac{e^{-1-0.5d_{1i}^2-d_{2i}}}{4(1 + e^{-1-0.5d_{1i}^2-d_{2i}})} \tag{2.40}$$

$$p(\delta_i^A = 1 | X_i = x_i) = \frac{e^{-3-x_{1i}+x_2-0.5x_{1i}w_{2i}}}{2(1 + e^{-3-x_{1i}+x_{2i}-0.5x_{1i}x_{2i}})} \tag{2.41}$$

where δ_i^R and δ_i^A denote sample indicators of subject $i \in U$ being selected for S_R and S_A, respectively. This is a case where, given X, the sampling design is ignorable in S_A, but not in S_R. For simplicity, I ignored further complexity in the sample design

of the reference survey like clustering or stratification. In order to consider the second sample as a non-probability survey, it is assumed that inclusion probabilities for that sample are unknown, and X is the only set of observed covariates in both S_R and S_A. The population mean of the outcome variables were $\bar{Y}^c = 4.01$ and $\bar{Y}^b = 20.23\%$.

I then repeatedly drew samples of sizes $n_R = 200$ and $n_A = 1000$ with a systematic PPS design based the two sets of unequal probabilities of selections generated above. In each iteration, I applied PAPP, PMLE-V and PMLE-C approaches to estimate the pseudo-inclusion probabilities based on a GLM with two-way interactions. Under the PAPP method, BART and CART were also used to estimate pseudo-weights. Furthermore, the two trimming techniques described in Section 2.2.3 were assessed in mitigating the effect of outlying weights, while setting $c = 5$. The simulation was then replicated $K = 1,000$ times, where for each iteration, the pseudo-weighted mean and 95% CI of the response variables, Y^c and Y^b, were estimated. Relative bias (rBias), relative root mean square error (rMSE), the nominal coverage rate of 95% CIs (crCI) and standard error ratio (rSE) were calculated by

$$rbias(\hat{\bar{y}}_{PW}) = 100 \times \frac{1}{K}\sum_{k=1}^{K}\left(\hat{\bar{y}}_{PW}^{(k)} - \bar{y}_U\right)/\bar{y}_U \tag{2.42}$$

$$rMSE(\hat{\bar{y}}_{PW}) = 100 \times \sqrt{\frac{1}{K}\sum_{k=1}^{K}\left(\hat{\bar{y}}_{PW}^{(k)} - \bar{y}_U\right)^2}/\bar{y}_U \tag{2.43}$$

$$crCI(\hat{\bar{y}}_{PW}) = 100 \times \frac{1}{K}\sum_{k=1}^{K} I\left(\left|\hat{\bar{y}}_{PW}^{(k)} - \bar{y}_U\right| < z_{0.975}\sqrt{var(\hat{\bar{y}}_{PW}^{(k)})}\right) \tag{2.44}$$

$$rSE(\hat{\bar{y}}_{PW}) = \frac{1}{K}\sum_{k=1}^{K}\sqrt{var(\hat{\bar{y}}_{PW}^{(k)})}/\sqrt{\frac{1}{K-1}\sum_{k=1}^{K}\left(\hat{\bar{y}}_{PW}^{(k)} - \bar{\bar{y}}_{PW}\right)^2} \tag{2.45}$$

where $\hat{\bar{y}}_{PW}^{(k)}$ denotes the pseudo-weighted mean for the k-th iteration, $\bar{\bar{y}}_{PW} = \sum_{k=1}^{K}\bar{y}_{PW}^{(k)}/K$, and $\bar{y}_U$ is the finite population true mean. For the GLM and CART, I used a modified delete-one Jackknife repeated replication (JRR) method and for BART, a conditional variance method with $M = 500$ as described in Section 2.2.4 was applied. To eval-

uate the proposed variance estimation methods, I computed the ratio of estimated standard error (SE) over the true SE to evaluate my proposed method of variance estimation under pseudo-weighting based on BART.

2.3.2 Simulation results

Table 2.1 summarizes the simulation results based on the two different quasi-random methods with ρ fixed at 0.80. Overall, the simulation results suggest a better performance of PAPP compared to the PMLE methods in reducing selection bias. Under the GLM, PAPP gives smaller rBias and rMSE than both PMLE methods when no trimming is applied. However, when more complex associations come into play in describing the outcome variables and the selection mechanism of sample units, the use of classical modeling approaches may not be an appropriate solution to estimate the pseudo-weights. For both continuous and categorical outcome variables, I found that the rMSE value for the PAPP approach with BART was closest to the same quantity in fully weighted estimates. (By fully weighted I mean adjustment based on the (unknown) true weights). With no trimming, it seems PMLE-V outperforms PMLE-C with respect to rBias and rMSE. Compared to other modeling methods, PAPP with CART worked worst in terms of both bias and rMSE.

However, it seems that all the pseudo-weighting techniques tend to generate some influential pseudo-weights because trimming reduces the bias in almost all situations. My simulation reveals that trimming is an effective way to treat the outlying weights, and that, given $c = 5$, the method based on IQR performs a bit more efficiently than the entropy method. Under trimming the smallest value of rBias was associated with the PMLE-C method.

Regarding variance estimation, my primary simulation results showed that the previously proposed method based on JRR overestimates the variance substantially for BART and CART. It was not surprising as empirical findings show that JRR does

Table 2.1: Comparing the performance of pseudo-weighting approaches in the simulation study.

Method		**Continuous outcome (Y^c)** rBias	rMSE	crCI	rSE	**Binary outcome (Y^b)** rBias	rMSE	crCI	rSE
Unweighted		125.64	125.75	0.00	0.98	-70.38	70.47	0.00	1.01
Fully weighted		0.23	7.78	93.90	0.96	0.31	15.97	94.60	0.97
PMLE-V									
GLM	trim no	-16.98	43.78	87.00	0.98	7.85	49.22	95.00	1.16
	trim 1	11.98	30.58	91.80	0.89	-23.23	32.00	59.10	0.77
	trim 2	3.05	30.17	89.50	0.92	-12.09	29.94	77.90	0.88
PMLE-C									
GLM	trim no	-26.04	57.93	88.10	1.05	17.19	66.34	96.50	1.23
	trim 1	7.99	31.49	91.80	0.94	-20.30	32.05	63.50	0.78
	trim 2	-0.72	32.41	88.80	0.94	-8.86	32.54	78.30	0.86
PAPP									
GLM	trim no	-15.08	21.98	82.60	1.01	5.91	26.95	95.40	1.04
	trim 1	6.26	16.28	90.80	0.93	-20.54	25.76	52.60	0.75
	trim 2	-3.57	16.13	90.70	0.98	-8.28	20.48	82.90	0.85
BART	trim no	-3.93	14.56	96.90	1.04	-6.12	21.85	89.60	1.00
	trim 1	3.65	13.36	96.60	1.04	-15.82	22.84	78.20	1.01
	trim 2	1.20	12.44	96.50	1.03	-12.88	20.91	83.30	0.99
CART	trim no	69.04	70.92	30.90	2.08	-36.13	38.08	45.10	1.85
	trim 1	83.28	84.44	12.40	1.58	-41.98	42.86	4.40	1.17
	trim 2	74.46	75.87	17.50	1.70	-38.88	40.17	17.10	1.31

NOTE 1: PMLE-V: Pseudo-maximum likelihood estimation method by Valliant & Dever (2011); PMLE-V: Pseudo-maximum likelihood estimation method by Chen et al (2019); PAPP: Propensity-adjusted Probability Prediction; GLM: Generalized Linear Model; BART: Bayesian Additive Regression Trees; CART: Classification and Regression Trees.

NOTE 2: Variance estimation under BART is based on the conditional variance method and for the rest of models, a delete-one Jackknife repeated replication method is used.

not work well with non-linear estimators. Even under the GLM, the JRR tends to slightly overestimate the variance, though trimming mitigates this to some extent. For the PAPP based on BART, my proposed method performs very well for both continuous and binary outcomes as the values of SE ratio are very close to 1 for both continuous and binary outcomes, with the nominal 95% CI coverage rate tending to be closest to the nominal rate.

Finally, figure 2.2 depicts the effect of different values of ρ on rMSE and SE ratio through a heatmap. For the rMSE, darker colors indicate larger values of rMSE, but for SE-ratio darker colors show values closer to 1. I consider $\rho = 0, 0.1, ..., 0.9$, where $\rho = 0$ implies that design of S_R is non-ignorable given X and $\rho = 1$ implies the design of both S_R and S_A is ignorable given X. As illustrated, for both continuous and binary outcomes, smaller values of rMSE are associated with PAPP based on BART. This is while the performance of the PAPP method is almost robust across all different values of ρ. For the binary outcome, the worst situation is associated with the PMLE-C method. Regarding the variance estimation, it seems more accuracy is achieved by PAPP with GLM regardless of the type of outcome variable. Finally, an extension to the simulation for further comparing the PAPP method with alternative approaches has been given in Appendix 2.6.3.

2.4 Application

2.4.1 Safety Pilot Model Deployment

Launched in 2012 by the University of Michigan Transportation Research Institute, the SPMD is one of the world's largest ongoing NDS, collecting data from over 3,100 instrumented vehicles, including cars, vans, trucks, and buses. SPMD also characterizes a real-world implementation of connected vehicle safety systems with the primary aim of testing Dedicated Short Range Communications (DSRC)-based con-

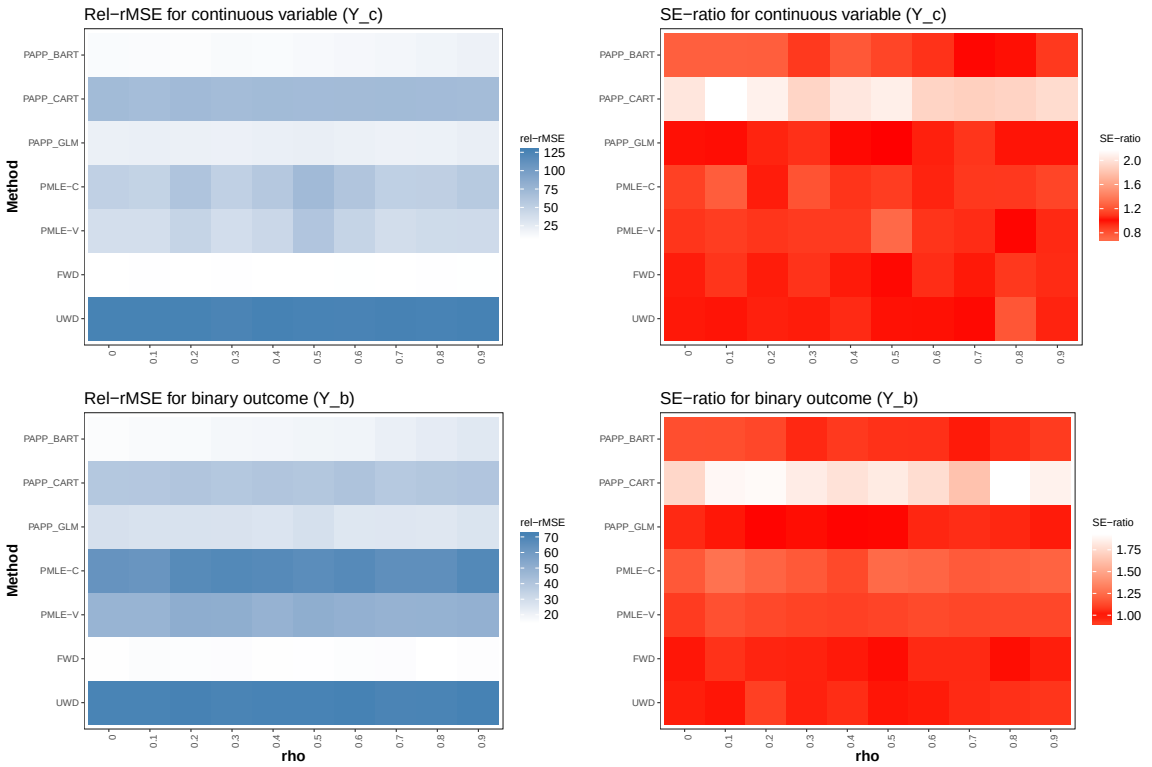

Figure 2.2: Evaluating the effects of the degree of design ignorability in the reference survey given common auxiliary variables in the simulation study. UWD=unweighted; FWD=fully weighted

nected vehicle communication technology (Narla, 2013). The smallest unit in the SPMD database is called a basic safety message (BSM), which includes vehicle trajectories, driver-vehicle interactions, and video records. The average time of participation in the study was approximately one person-year. Participants in SPMD were mostly volunteers from the southeast Michigan area, especially those in and around the city of Ann Arbor, recruited in a one-year period (August 2012 to August 2013) through a combination of snowball and convenience sampling techniques.

When a participant's vehicle is switched on, a data acquisition system (DAS), installed on the vehicles, starts recording GPS coordinates as well as corresponding timestamps 10 times per second and continues recording until it is switched off. In SPMD, a trip can be defined as the time interval during which the vehicle is on. Unique IDs are generated by the DAS for each participant and for each trip. Using these key features, I could identify trips in SPMD and measure several characteristics of trips, including trip distance, trip duration, trip average speed, and the start/end time of trips. I built a data set based on the trip summary information, where each record corresponds to a specific trip made by a specific vehicle. Over six million records of trips were available in the raw data, but after the data cleaning process, this number was reduced to 4,591,884. Detailed information about the vehicles including vehicle age, vehicle type, odometer reads, and vehicle make as well as gender and age of participants were recorded at the time of recruitment and joined to the trip summary dataset.

2.4.2 National Household Travel Survey

In the present chapter, I used data from the seventh round of the NHTS conducted from March 2008 through May 2009 as the reference survey. The NHTS is a nationally representative survey, repeated cross-sectionally approximately every seven years, that characterizes personal travel behaviors among the civilian, non-institutionalized

population of the United States. The 2009 NHTS was a telephone survey, in which participants were selected systematically using a list-assisted random digit dialing (RDD) technique. All eligible individuals aged $\geq$18 years within households were recruited for interviews conducted by landline. Proxy interviews were requested for younger household members who were <15 years old. Interviews were conducted in English or Spanish and data were collected using computer-assisted telephone interview (CATI) technology. Furthermore, a travel diary was mailed to the selected households to record trips made on a randomly assigned travel day by household members.

The initial sample size is approximately 25,000, representing all 50 States of the U.S. as well as the District of Columbia. However, an additional 125,000 households, including 20 states and Metropolitan Planning Organizations (MPO), were purchased for their respective regions (Santos et al., 2011), and these have been accounted for in the overall weighting scheme. The overall response rate of 19.8% is based on 'useable' households, defined as those in which at least 50% of household members completed an interview. Of those who completed the interview, 72% filled out the travel diary (Santos et al., 2011). In NHTS, a travel day is defined from 4:00 AM of the assigned day to 3:59 AM of the following day on a typical weekday. On weekends, it begins on Friday at 6:00 PM and ends on Sunday at midnight. A trip is defined as that made by one person in any mode of transportation. A total of 308,901 eligible individuals aged $\geq$5 took part in the study, for which 1,294,219 trips were recorded.

2.4.3 Auxiliary variables and analysis plan

As mentioned in the methods section, because of the ignorable assumption, the common auxiliary variables available in both the non-probability sample and the reference survey play a key role in the quasi-random approach. Therefore, particular attention was paid to identify and build as many common variables as possible that are ex-

pected to be predictors of pseudo-inclusion probabilities. However, since the SPMD sample is gathered from a limited geographical area, in order to be able to generalize the findings to the U.S. population of drivers, I have to assume that no other variable than those common covariates investigated in this study will contribute to the geographical dispersion of the outcome variables of interest. This assumption is in fact embedded in the ignorable condition in the SPMD given the observed set of common covariates.

In the current study, the ultimate goal was to generate and assign pseudo-weights to individuals in SPMD, so I used the individual-level data. Two distinct sets of covariates were considered: (1) demographic information of the drivers including gender, age, and population size of the residential areas, and (2) vehicle characteristics, including vehicle age, vehicle type, vehicle make, and odometer readings. Figure 2.3 compares the frequency distribution/kernel density of common auxiliary variables in SPMD with weighted ones in NHTS.

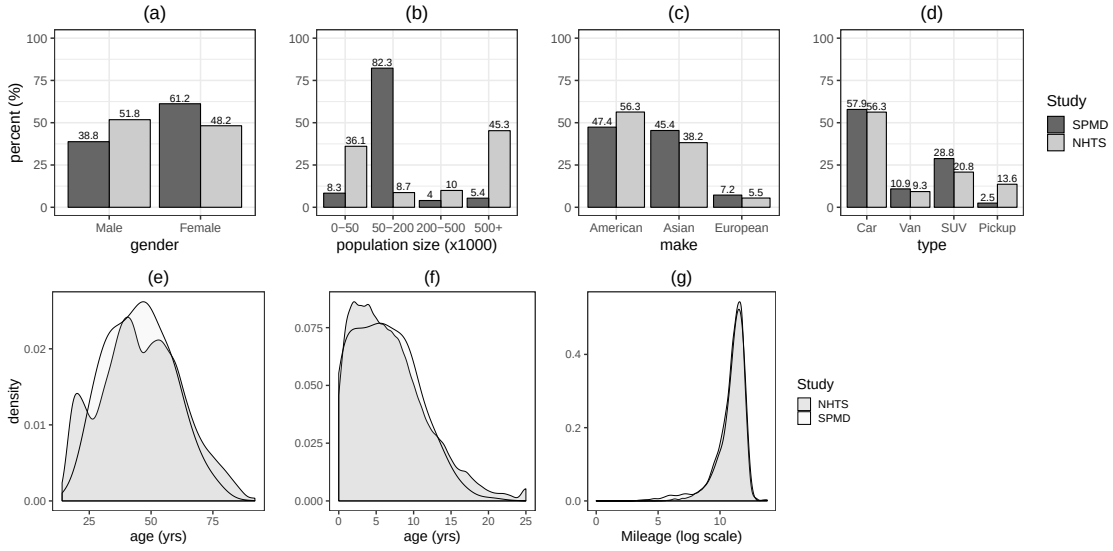

Figure 2.3: Comparison of frequency distributions of common auxiliary variables, including (a) gender, (b) population size of residential area, (c) vehicle make, (d) vehicle type, (e) participants age, (f) vehicles age and (g) odometer reads, between SPMD and NHTS (weighted)

Although demographic information as common auxiliary variables seems to be

essential in the calibration process, such data were limited to gender and age in SPMD. Moreover, the item-missing rate was 13.1% for gender and 30.5% for age. I assumed volunteers in SPMD were the drivers for the whole trips recorded throughout the participation time. In addition, there was no information available about participants' residential areas in the SPMD dataset. However, I could predict the residential area based on the mode of start location of trips made between 6:00 PM and 12:00 PM every day. Before performing any statistical analysis, several further attempts were built to make the two datasets comparable. I filtered out all the individuals in NHTS who were passengers of the vehicle, not the driver. Any trips for which public transportation was used were dropped. Furthermore, I only kept trips that involved passenger cars, SUVs, vans, or pickup trucks in the NHTS.

Detailed information in SPMD was collected on vehicles characteristics at the time of registration. Vehicle age (truncated at 25 years), vehicle type (passenger car, SUV, van, and pickup truck), vehicle make (American, Asian, and European), and odometer readings were the variables identified in common between the two datasets. Besides the auxiliary variables, I considered multiple trip-related measures as outcome variables for evaluating weighted estimates in terms of bias. One major structural difference between NHTS and SPMD is that trips made within one pre-specified day per individual, while in SPMD, individuals were followed up for several months and years. Therefore, in both studies, I computed trip characteristics on a daily basis, and then considered the mean over days of follow-up for each individual in SPMD. Using this approach, I constructed several outcome variables including the total duration of trips per day, the total distance of trips per day, mean average speed, and mean time-of-day of trips.

2.4.4 Results

I initially compared the discrepancies in the distribution of auxiliary variables between the two data sets. Figure 2.3 illustrates that the largest distributional discrepancies among the set of common auxiliary variables are the population size of participants' residential area. The SPMD sample is largely limited to the Southeast Michigan area, especially in and around Ann Arbor, so it was expected that SPMD underrepresents the most and least densely populated areas. Men are underrepresented in the SPMD compared to the population of U.S. drivers, while participants in SPMD (mean=45.4; sd=13.3) tend to be younger than NHTS (mean=54.4; sd=12.8). I found no substantial differences in the vehicle characteristic distributions.

As discussed in Section 2.2, estimating pseudo-weights based on the PAPP approach requires modeling two conditional probabilities, (1) $p(\delta_i^R = 1|X_i = x_i)$, and (2) $p(Z_i = 0|X_i = x_i)$. Since probabilities of selection in NHTS are bounded within $(0, 1)$, I modeled the *logit* transformation of the selection probabilities in NHTS, which maps the values to $(-\infty, \infty)$. Considering the first 100 iterations as the burn-in period, MCMC with 1,100 iterations was applied to train the model. the pseudo-R^2 values associated with BART was 17.1%. I also utilized cross-validation to evaluate the predictive accuracy of the fitted model. The average Pseudo-R^2 for training and test data was 17.6% and 17.8%, respectively. In addition, I compared the predictive power of BART with some regression-based models, including linear regression, Poisson regression, beta regression, and adaptive spline, and found that BART performs much better than all these alternatives (See Table 2.2). Then, the *logit* of inclusion probabilities were projected to individuals in SPMD using the common covariates.

To estimate the propensity scores, i.e. $P(Z_i = 1|X_i = x_i)$, I combined the two datasets, and created the Z_i $(i = 1, 2, ..., n)$ indicator. Again, BART was used to model the binary outcome Z on the common set of covariates. I then compared the classification power of BART against some regression-based models such as binary

logistic regression and algorithmic classification methods such as CART based on the area under the curve (AUC) of the receiver operating characteristic (ROC). AUC can be considered as a proper measure to test the predictive accuracy of propensity models. The largest value of AUC (93.2%) was pertaining to BART (the ROC curve is displayed in figure 2.4). In addition, I estimated the pseudo-weights for units in SPMD based on the pseudo-likelihood approach by Valliant et al. (2018) (PMLE-V) and its extension by Chen et al. (2019) (PMLE-C) as discussed in section 2.2. These two also involved modeling $P(Z_i = 1|X_i = x_i)$ but through a weighted logistic regression model, where parameters estimation is achieved by solving the equations 2.6 and 2.9, respectively.

Table 2.2: Comparing the goodness-of-fit of BART with other existing methods, I=main effects in the model; II=two-way interaction effects were included

Model	**RMSE**	R^2/**Pseudo-**R^2
Original scale of response		
Linear Regression I	0.02	5.94
Linear Regression II	0.02	6.00
Poisson Regression I	0.02	5.87
Poisson Regression II	0.02	6.27
Beta Regression I	0.02	5.01
Beta Regression II	0.02	5.20
Multivariate adaptive spline	0.02	5.66
Bayesian Additive Regression Trees	0.02	**6.93**
Log Scale of response		
Linear Reg I	1.35	14.95
Linear Reg II	1.35	15.27
Multivariate adaptive spline	1.35	14.97
Bayesian Additive Regression Trees	1.33	**17.07**

Final pseudo-weights were obtained for samples in SPMD after normalizing them so that the sum of the estimated pseudo-weights in SPMD is equal to the sum of the weights in NHTS. Figure 2.5 compares the kernel density of estimated propensity scores in the *log* scale based on BART between SPMD and NHTS. As illustrated, there is a lack of common support on the left tail of the PS distribution in SPMD.

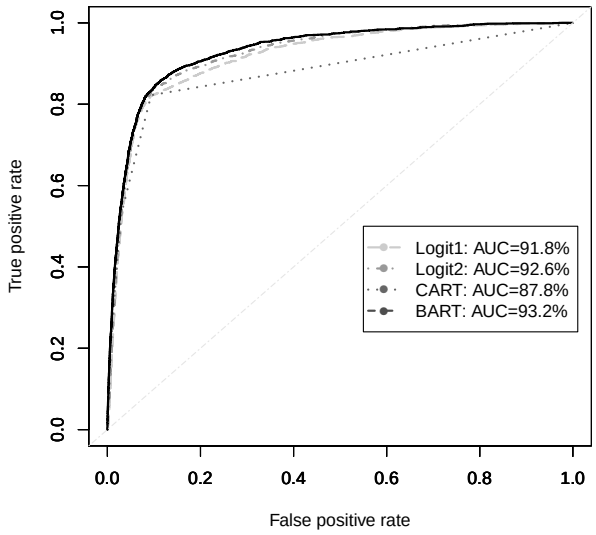

Figure 2.4: ROC curve analysis for comparing the prediction power of BART with other existing methods

Figure 2.6 displays how the PAPP approach corrects for the discrepancies in the distribution of auxiliary variables compared to unweighted distributions in figure 2.3.

I then contrast the performance of different quasi-random methods on the actual data by estimating pseudo-weighted estimates and associated 95% confidence intervals (CI) for several outcome variables in common between SPMD and NHTS. These variables include mean daily frequency of trips, mean daily total time of trips (Minutes), mean average speed (Km/h), mean daily start time of the trips, mean annual mileage (Km), mean daily percentages of trips started between 6-10 AM, mean stop duration per trip, mean daily percentage of trips using interstates, mean time spent on interstates per trip, and mean annual mileage. The point estimates and associated 95% CIs under different quasi-randomization approaches and different weight trimming methods are compared with weighted estimates in NHTS in figure 2.7.

In addition, I show point estimates and associated 95% CIs for some SPMD-specific outcomes in figure 2.8, including the mean daily percentage of trips started between 6-10 AM, mean daily percentage of trips used the interstate, mean percent-

age of trips duration spent on the interstate, and mean percentage of stop duration per trip. Detailed numerical comparisons for these two sets of outcome variables by demographics and vehicle characteristics are provided in Tables 2.3 through 2.8 in Appendix 2.6.2. The percentage of trimmed pseudo-weights, obtained by PAPP with BART, was 17.2% and 14.2% in the entropy and interquartile range methods, respectively, with $c = 5$.

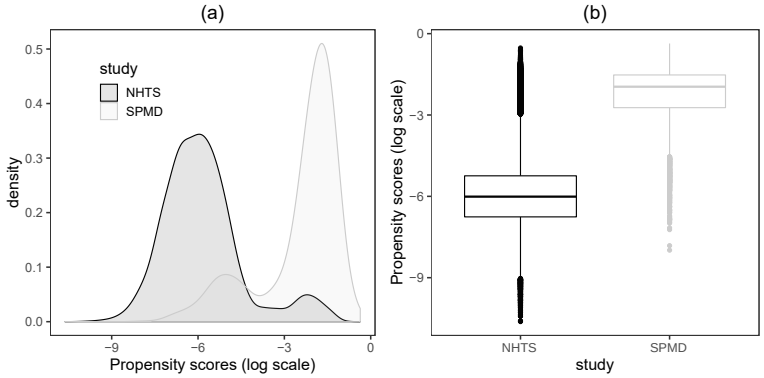

Figure 2.5: Comparing the distributions of estimated propensity scores between SPMD and NHTS (log scale)

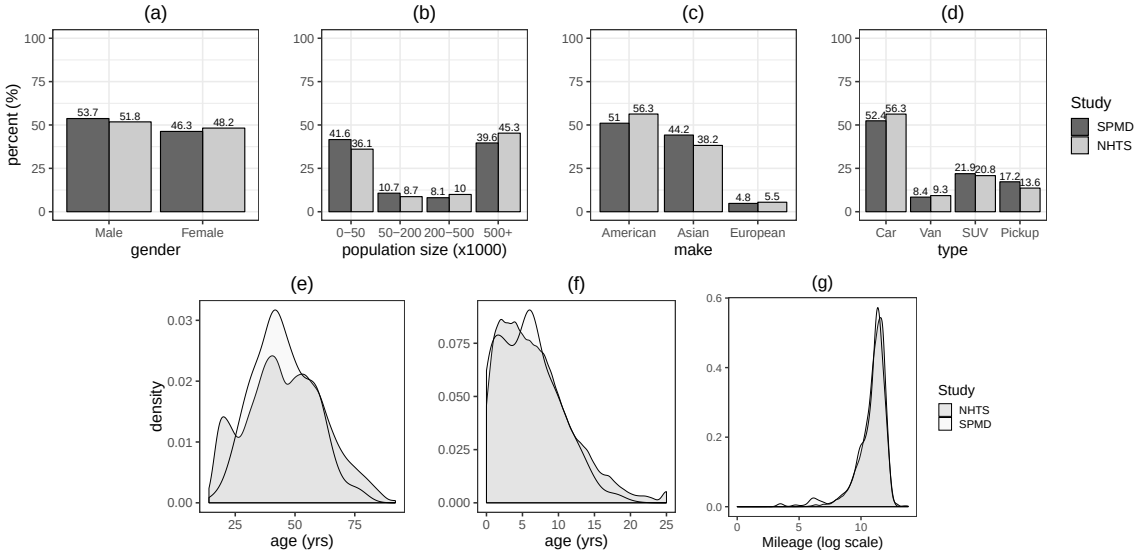

Figure 2.6: Comparison of frequency distributions of common auxiliary variables, including (a) gender, (b) population size of residential area, (c) vehicle make, (d) vehicle type, (e) participants age, (f) vehicles age and (g) odometer reads, between weighted SPMD using pseudo-weighting approach and weighted NHTS

Although the findings of actual data analysis vary across different outcome variables, it seems that none of the untrimmed pseudo-weights performs well in bias correction. Instead, for all the outcome variables with the exception of the mean daily start time of trips, trimming pseudo-weights appears to be significantly effective in reducing bias and improving stability. This might be evidence of outlying pseudo-weights generated by the QR approach. Given the constant $c = 5$, a smaller bias was obtained for the trimming based on the IQR than entropy. These findings on trimming were consistent with the simulation results.

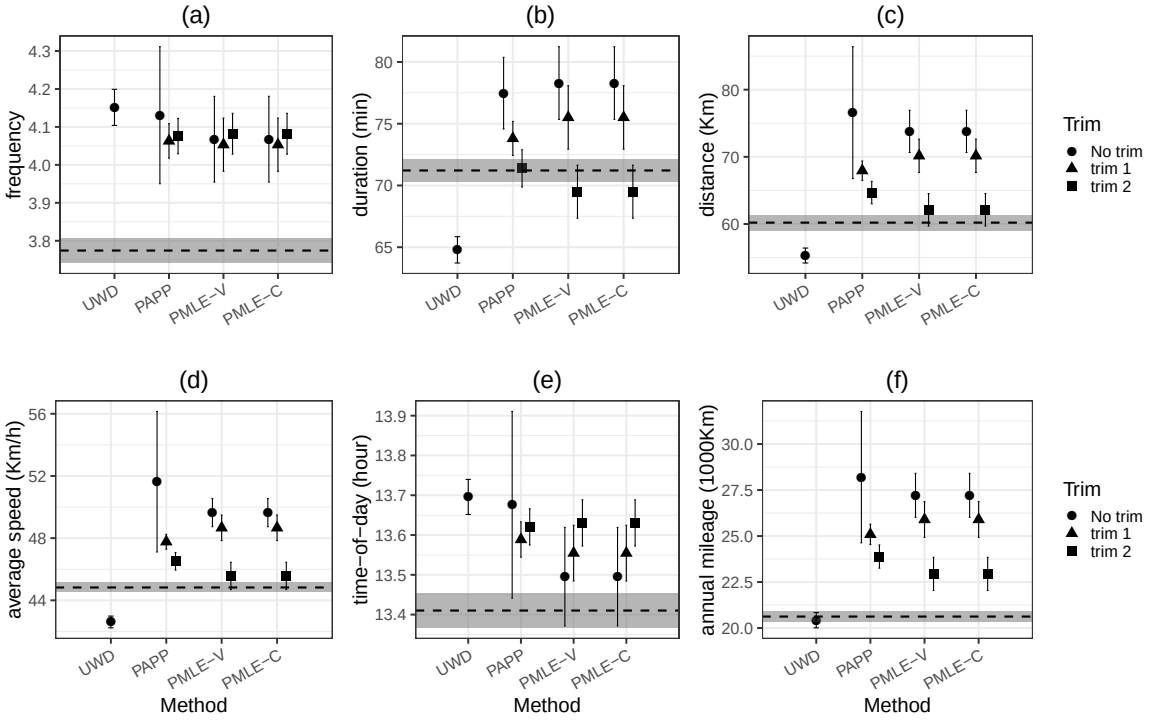

Figure 2.7: Evaluation of pseudo-weights by comparing weighted estimates of the daily frequency of trips between NHTS and SPMD: (a) Mean daily frequency of trips, (b) Mean daily total trip duration, (c) Mean daily total distance driven, (d) Mean trip average speed, (e) Mean daily start time of the trip, and (f) Mean annual mileage. The dashed line and surrounding shadowed area represent weighted estimates and 95% CIs in NHTS, respectively. UWD=unweighted; Trim 1=pseudo-weights trimmed based on the entropy method; Trim 2=pseudo-weights trimmed based on the IQR method

These interpretations can be generalized to those outcome variables in figure 2.8, though no benchmark is available. Overall, PAPP ends up with relatively wide 95%

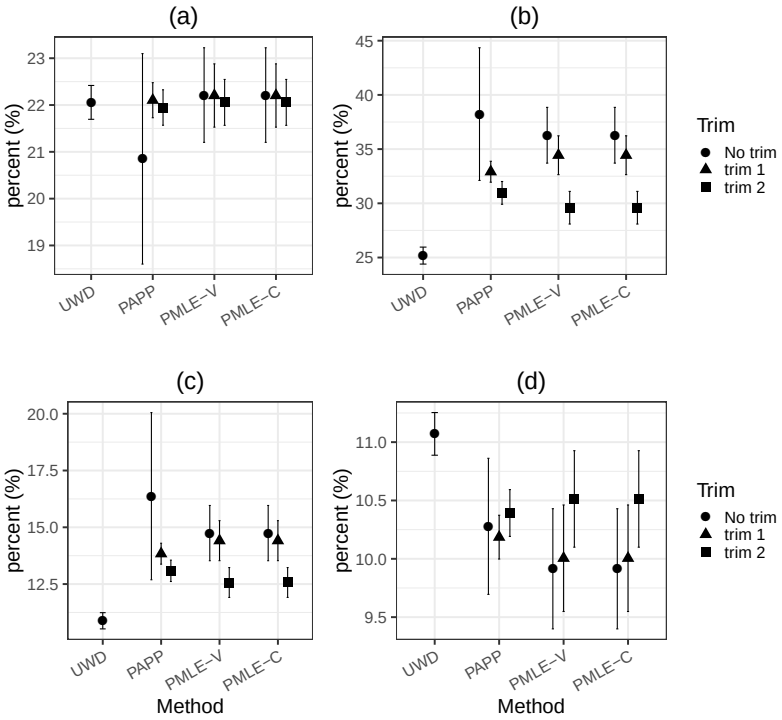

Figure 2.8: Weighted estimates of some SPMD-specific outcomes: (a) Mean daily frequency of trips used interstate, (b) Mean percentage of trip spent on interstate, (c) Mean percentage of stop duration per trip, and (d) Mean percentage of trips started between 6-10am. UWD=unweighted; Trim 1=pseudo-weights trimmed based on the entropy method; Trim 2=pseudo-weights trimmed based on the IQR

CIs, indicating the existence of influential pseudo-weights. However, it seems trimming works efficiently in treating such outlying pseudo-weights. According to Tables 2.3-2.8 in Appendix 2.6.2, adjusted values for the SPMD using pseudo-weighting are generally in the direction I would expect: interstate use might be expected to be for shorter trips in a suburban area that underrepresents very rural areas, and to include a large fraction in a sample with typical morning commutes as employed, white-collar workers. It is less clear why SPMD might underrepresent trips with more stops.

2.5 Discussion

In the present chapter, I sought to improve the representativeness of Big Data in SPMD, which is a large-scale naturalistic driving study. In particular, I was interested in using the quasi-randomization approach so that a single set of pseudo-weights are created and can be applied easily in any post hoc statistical analysis. To correct for selection bias in point estimates, two general quasi-randomization methods, PAPP and PMLE, were applied. For the earlier method, I was able to employ BART to predict components of PAPP, and results were compared with respect to reduction in bias.

The simulation findings generally reflect that the PAPP outperforms the traditional method of inverse propensity weighting. Generalizing this result to those I found from the application is somewhat limited as there was some evidence of non-ignorability given the available set of common covariates. However, in the absence of strongly predictive covariates, the use of advanced supervised machine learning techniques such as BART can produce less biased estimates for population inference than traditional linear and generalized linear model approaches. Furthermore, a partial lack of common support in the distribution of auxiliary variables led to increased bias and wide CIs, which demands more advanced trimming methods for identifying and

treating outlying pseudo-weights.

Note that BART seems to be more sensitive to the problem of inadequate common support compared to the GLM with only main effects included. This is mainly because BART automatically accounts for high-order interactions. Therefore, what matters to be checked for the positivity assumption under BART is the common support with respect to the joint distribution of the auxiliary variables. The problem becomes more severe if BART is able to accurately predict the selection mechanism but fails to appropriately model the outcome. I believe these are the two major reasons explaining why we observed relatively wide 95% CIs under the PAPP method with BART. Hill and Su (2013) propose a criterion based on BART for identifying the common support in the causal inference context. The authors highlight the fact that failure to properly detect the regions with inadequate common support leads to invalid inference because of imbalance with respect to covariates' joint distribution or inappropriate extrapolation by the propensity model.

One of the major challenges to the present research was that the definition of a trip in SPMD did not quite match that in NHTS. As discussed briefly in the introduction, trips in SPMD are captured by DAS instruments while in NHTS, trips are recorded through a travel diary, which relies on estimates and individuals' memories. This may cause a kind of differential measurement error in covariates. The other major challenge in this study arose from the structural differences in the design of the two studies, SPMD and NHTS. In NHTS, trip diaries were filled out for only one day for each survey participant while in SPMD, individuals were followed up for several months or even years.

The other weakness of this study was the limited set of auxiliary variables and especially demographics, which might question the design ignorability condition in SPMD. Having further relevant covariates, such as race and education level, observed could potentially improve the performance of adjustments. Finally, some studies

suggest that sampling weights and sampling design of the reference survey should be incorporated in modeling the propensity scores, i.e. $p(Z_i = 1|X_i = x_i)$, in the pseudo-weighting approach. However, I was unable to do so, because the current version of BART does not incorporate sample weights for a binary outcome variable.

An interesting issue arises in the construction of the propensity score models for sample membership. While accurate assessment of this propensity is important, over-fitting these models can lead to complete separation, yielding to highly variable and unstable weights. In the NHTS/SPMD application, the state of residence indicator would be such a variable. This is similar to the setting in causal inference where lack of overlap in the propensity of being treated can sometimes be obtained by including a sufficient number of covariates (Westreich et al., 2011). In such a setting one must make the assumption that these variables are not associated with the targets of inference. Here future work that borrows from the causal inference literature on propensity score construction (Griffin et al., 2017) may be fruitful.

When the design of the reference survey involves clustering and stratification, the current pseudo-weighting methods may not be appropriate as they assume independence among observations. One potential approach in such situations is to generate a synthetic population by undoing the sampling design through a finite population bootstrap method. Once auxiliary variables are imputed for the whole population, then, a simple propensity model can be used to estimate pseudo-weights. Another alternative approach to quasi-randomization is the super-population technique, in which models are fitted to predict the outcome variable for the non-sampled units of the population. However, unlike quasi-randomization, adjustments need to be repeated operationally for any analytic variable.

As a third approach, this method can be combined with quasi-randomization to construct a doubly robust estimator, where estimates are consistent if either model holds. This provides protection against model misspecification. One may be inter-

ested in applying the doubly robust approach on SPMD data to do calibration for a set of specific outcomes, and then compare it with the quasi-randomization approach in terms of reduction in selection bias. Finally, as I discussed earlier, the measurement error structure in organically collected data may appear differently than survey data. The authors would like to suggest research on how to adjust for such differential measurement errors when adjusting for selection bias, which is especially the case when combining Big Data with survey data.

Finally, I analyzed complete data only, which means all cases for whom at least one common covariate was missing were excluded from the analysis. While imputing missing data under the missing at random condition based on the same set of common covariates should not result in reductions in bias or variance, there may be settings where additional covariate information for item-level missingness is available above and beyond that associated with the differing sampling mechanism. Hence dealing with missing data is an important topic for future research.

2.6 Appendix

2.6.1 Theoretical proofs

Suppose there exists an infinite sequence of finite populations U_ν of sizes N_ν with $\nu = 1, 2, ..., \infty$. Corresponding to U_ν are a non-probability sample $S_{A,\nu}$ and a probability sample $S_{R,\nu}$ with $n_{A,\nu}$ and $n_{R,\nu}$ being the respective sample sizes. Also, let us assume that $N_\nu \to \infty$, $n_{A,\nu} \to \infty$ and $n_{R,\nu} \to \infty$ as $\nu \to \infty$, while $n_{A,\nu}/N_\nu \to f_A$, and $n_{R,\nu}/N_\nu \to f_R$ with $0 < f_R < 1$ and $0 < f_A < 1$. However, from now on, we suppress the subscript ν for rotational simplicity. In order to be able to make unbiased inference based on S_A, we consider the following conditions:

1. The set of observed auxiliary variables, X, fully governs the selection mechanism in S_A. This is called an *ignorable* condition, implying $p(\delta_i^A = 1|y_i, x_i) = p(\delta_i^A = 1|x_i)$ for $i \in U$.

2. The S_A actually does have a probability sampling mechanism, albeit unknown. This means $p(\delta_i^A = 1|x_i) > 0$ for all $i \in U$.

3. Units of S_R and S_A are selected independently from U given the observed auxiliary variables, X^*, i.e. $\delta_i^R \perp\!\!\!\perp \delta_j^A | X^*$ for $i \neq j$.

4. The sampling fractions, f_R and f_A, are small enough such that the possible overlap between S_R and S_A is negligible, i.e. $S_R \cap S_A = \emptyset$.

5. The true underling models for $Y|X^*$ and $\delta_A|X$ and $\delta^R|X$ are known.

In addition, to be able to drive the asymptotic properties of the proposed estimators, we consider the following regularity conditions according to Chen et al. (2019):

1. For any given x, $\partial m(x;\theta)/\partial\theta$ exists and is continuous with respect to θ, and $|\partial m(x;\theta)/\partial\theta| \leq h(x;\theta)$ for θ in the neighborhood of θ, and $\sum_{i=1}^N h(x_i;\theta) = O(1)$.

2. For any given x, $\partial^2 m(x;\theta)/\partial\theta^T$ exists and is continuous with respect to θ, and $\max_{j,l}|\partial^2 m(x;\theta)/\theta_j\partial\theta_l| \leq k(x;\theta)$ for θ in the neighborhood of θ, and $\sum_{i=1}^{N} k(x_i;\theta) = O(1)$.

3. For $u_i = \{x_i, y_i, m(x_i;\theta)\}$, the finite population and the sampling design in S_R satisfy $N^{-1}\sum_{i=1}^{n_R} u_i/\pi_i^R - N^{-1}\sum_{i=1}^{N} u_i = O_p(n_R^{-1/2})$.

4. There exist c_1 and c_2 such that $0 < c_1 \leq N\pi_i^A/n_A \leq c_2$ and $0 < c_1 \leq N\pi_i^R/n_R \leq c_2$ for all $i \in U$.

5. The finite population and the propensity scores satisfy $N^{-1}\sum_{i=1}^{N} y_i^2 = O(1)$, $N^{-1}\sum_{i=1}^{N}||x_i||^3 = O(1)$, and $N^{-1}\sum_{i=1}^{N}\pi_i^A(1-\pi_i^A)x_i x_i^T$ is a positive definite matrix.

Note that while we assume π_i^R is calculable for $i \in S_A$ throughout the proofs, extensions can be provided for situations where π_i^R need to be predicted for $i \in S_A$.

2.6.1.1 Asymptotic properties of PAPW estimator

Since $\hat{\beta}$ is the MLE estimate of β in the logistic regression of Z_i on x_i^*, it is clear that $\hat{\beta} \xrightarrow{p} \beta$. Two immediate result of this are that $\hat{\pi}_i^A \xrightarrow{p} \pi_i^A$ and $E(\hat{\pi}_i^A|x_i^*) = \pi_i^A$ where $\hat{\pi}_i^A$ is defined as in 2.6. Now, we prove the consistency and asymptotic unbiasedness of the PAPW estimator in 2.14. To this end, we show that $\hat{\bar{y}}_{PAPW} - \bar{y}_U = O_p(n_A^{-1/2})$. Consider the following set of estimating equations:

$$\Phi_n(\eta) = \begin{bmatrix} n^{-1}\sum_{i=1}^{n} Z_i(y_i - \bar{y}_U)/\pi_i^A \\ \\ n^{-1}\sum_{i=1}^{n}\{Z_i - p_i(\beta)\}x_i^* \end{bmatrix} = \begin{bmatrix} N^{-1}\sum_{i=1}^{N}\delta_i^A(y_i - \bar{y}_U)/\pi_i^A \\ \\ N^{-1}\sum_{i=1}^{N}\delta_i\{Z_i - p_i(\beta)\}x_i^* \end{bmatrix} = 0 \quad (2.46)$$

where $\eta = (\bar{y}_U, \beta)$.

In the following, we show that $E_{\delta^A}[\Phi_n(\hat{\eta})|x_i^*] = 0$. We start with the first component of $\Phi_n(\hat{\eta})$

$$\begin{aligned} E_{\delta^A}\left[\frac{1}{N}\sum_{i=1}^{N}\frac{E_{\delta^A}(\delta_i^A)(y_i-\bar{y}_U)}{\pi_i^A}\Big|x_i^*\right] &= \frac{1}{N}\sum_{i=1}^{N}\frac{E_{\delta^A}(\delta_i^A|x_i^*)(y_i-\bar{y}_U)}{\pi_i^A} \\ &= \frac{1}{N}\sum_{i=1}^{N}\frac{\pi_i^A(y_i-\bar{y}_U)}{\pi_i^A} \\ &= 0 \end{aligned} \tag{2.47}$$

Noting that $E_{\delta^A}[\Phi_n(\hat{\eta})] = E_\delta[E_Z\{\Phi_n(\hat{\eta})|\delta_i = 1\}]$, for the second component, we have

$$\begin{aligned} E_{\delta^A}\left[\frac{1}{N}\sum_{i=1}^{N}\delta_i\{Z_i - p_i(\beta)\}x_i\Big|x_i\right] &= E_\delta\left[E_Z\left\{\frac{1}{N}\sum_{i=1}^{N}\delta_i\{Z_i-p_i(\beta)\}x_i\big|\delta_i=1,x_i\right\}\right] \\ &= E_\delta\left[\frac{1}{N}\sum_{i=1}^{N}\delta_i\{E_Z(Z_i|\delta_i=1,x_i)-p_i(\beta)\}x_i^*\right] \\ &= E_\delta\left[\frac{1}{N}\sum_{i=1}^{N}\delta_i\{p_i(\beta)-p_i(\beta)\}x_i^*\right] \\ &= 0 \end{aligned} \tag{2.48}$$

Now, we apply the first-order Taylor approximation to $\Phi_n(\hat{\eta})$ around η_1 as below:

$$\hat{\eta} - \eta_1 = [E\{\phi_n(\eta_1)\}]^{-1}\Phi_n(\eta_1) + O_p(n_A^{-1/2}) \tag{2.49}$$

where $\phi_n(\eta) = \partial\Phi_n(\eta)/\partial\eta$.

$$\frac{\partial}{\partial\bar{y}_U}\left[\frac{1}{N}\sum_{i=1}^{N}\delta_i^A\frac{(y_i-\bar{y}_U)}{\pi_i^A}\right] = -\frac{1}{N}\sum_{i=1}^{N}\frac{\delta_i^A}{\pi_i^A} \tag{2.50}$$

$$\begin{aligned}\frac{\partial}{\partial\beta}\left[\frac{1}{N}\sum_{i=1}^{N}\delta_i^A\frac{(y_i-\bar{y}_U)}{\pi_i^A}\right] &= \frac{\partial}{\partial\beta}\left[\frac{1}{N}\sum_{i=1}^{N}\frac{\delta_i^A}{\pi_i^A}\left\{\frac{p_i(\beta)}{1-p_i(\beta)}\right\}(y_i-\bar{y}_U)\right] \\ &= -\frac{1}{N}\sum_{i=1}^{N}\frac{\delta_i^A}{\pi_i^A}(y_i-\bar{y}_U)x_i^{*T}\end{aligned} \tag{2.51}$$

$$\frac{\partial}{\partial\beta}\left[\frac{1}{N}\sum_{i=1}^{N}\delta_i\big\{Z_i-p_i(\beta)\big\}\right] = -\frac{1}{N}\sum_{i=1}^{N}\delta_i p_i(\beta)\left[1-p_i(\beta)\right]x_i^*x_i^{*T} \tag{2.52}$$

Therefore, we have

$$\phi_n(\eta_1) = \begin{pmatrix} -\frac{1}{N}\sum_{i=1}^{N}\frac{\delta_i^A}{\pi_i^A} & -\frac{1}{N}\sum_{i=1}^{N}\frac{\delta_i^A}{\pi_i^A}(y_i-\bar{y}_U)x_i^{*T} \\ 0 & -\frac{1}{N}\sum_{i=1}^{N}\delta_i p_i(\beta)\left[1-p_i(\beta)\right]x_i^*x_i^{*T}\end{pmatrix} \tag{2.53}$$

Thus, it follows that $\hat{\bar{y}}_{PM} = \bar{y}_U + O_p(n_A^{-1/2})$.

Now, we turn to deriving the asymptotic variance estimator for $\hat{\bar{y}}_{PM}$. According to the sandwich formula, we have

$$Var(\hat{\eta}_1) = \left[E\{\phi_n(\eta_1)\}\right]^{-1} Var\big\{\phi_n(\eta_1)\big\}\left[E\{\phi_n(\eta_1)\}^T\right]^{-1} + O_p(n_A^{-1}) \tag{2.54}$$

Given the fact that

$$E(\delta_i = 1|x_i^*) = \frac{p(\delta_i^A = 1|x_i^*)}{p(Z_i = 1|x_i^*)} = \frac{\pi_i^R}{1-p_i(\beta)} \tag{2.55}$$

It can be shown that

$$E\big\{\phi_n(\eta_1)\big\} = \begin{pmatrix} -1 & -\frac{1}{N}\sum_{i=1}^{N}(y_i-\bar{y}_U)x_i^{*T} \\ 0 & -\frac{1}{N}\sum_{i=1}^{N}\pi_i^A\left[1-p_i(\beta)\right]x_i^*x_i^{*T}\end{pmatrix} \tag{2.56}$$

And

$$\left[E\left\{\phi_n(\eta_1)\right\}\right]^{-1} = \begin{pmatrix} -1 & b^T \\ 0 & -\left[\frac{1}{N}\sum_{i=1}^{N}\pi_i^A\left[1-p_i(\beta)\right]x_i^*x_i^{*T}\right]^{-1} \end{pmatrix} \tag{2.57}$$

where

$$b^T = \left\{\frac{1}{N}\sum_{i=1}^{N}(y_i-\bar{y}_U)x_i^{*T}\right\}\left\{\frac{1}{N}\sum_{i=1}^{N}\pi_i^A\left\{1-p_i(\beta)\right\}x_i^*x_i^{*T}\right\}^{-1} \tag{2.58}$$

Now, the goal is to calculate $Var\left\{\phi_n(\eta_1)\right\}$. We know that

$$\begin{aligned} Var_{\delta^A}\left(\frac{1}{N}\sum_{i=1}^{N}\frac{\delta_i^A(y_i-\bar{y}_U)}{\pi_i^A}\middle| x_i\right) &= \frac{1}{N}\sum_{i=1}^{N}\frac{(y_i-\bar{y}_U)^2}{(\pi_i^A)^2}\pi_i^A(1-\pi_i^A) \\ &= \frac{1}{N}\sum_{i=1}^{N}\left\{\frac{1-\pi_i^A}{\pi_i^A}\right\}(y_i-\bar{y}_U)^2 \end{aligned} \tag{2.59}$$

$$\begin{aligned} Var_{\delta^A}\left(\frac{1}{N}\sum_{i=1}^{N}\delta_i\left\{Z_i-p_i(\beta)\right\}\middle| x_i^*\right) &= E_\delta\left[Var_Z\left(\frac{1}{N}\sum_{i=1}^{N}\delta_i\left\{Z_i-p_i(\beta)\right\}\middle|\delta_i=1,x_i^*\right)\right] \\ &\quad + Var_\delta\left[\delta_i E_Z\left(\frac{1}{N}\sum_{i=1}^{N}\delta_i\left\{Z_i-p_i(\beta)\right\}\middle|\delta_i=1,x_i^*\right)\right] \\ &= \frac{1}{N^2}E_\delta\left(\sum_{i=1}^{N}\delta_i^2 Var_Z(Z_i)x_i^*x_i^{*T}\middle| x_i^*\right)+0 \\ &= \frac{1}{N^2}\sum_{i=1}^{N}\pi_i^R p_i(\beta)x_i^*x_i^{*T} \end{aligned} \tag{2.60}$$

$$\begin{aligned} Cov\left(\frac{1}{N}\sum_{i=1}^{N}\frac{\delta_i^A(y_i-\bar{y}_U)}{\pi_i^A},\frac{1}{N}\sum_{i=1}^{N}\delta_i\left\{Z_i-p_i(\beta)\right\}\middle| x_i^*\right) &= E_\delta\left[E_Z\left(\frac{1}{N}\sum_{i=1}^{N}\delta_i\frac{Z_i(y_i-\bar{y}_U)}{\pi_i^A}\middle|\delta_i=1,x_i^*\right)\right] \\ &= \frac{1}{N^2}\sum_{i=1}^{N}\left\{1-p_i(\beta)\right\}(y_i-\bar{y}_U)x_i^* \end{aligned} \tag{2.61}$$

Therefore, we have

$$Var\{\Phi_n(\eta_1)\} = \begin{pmatrix} \frac{1}{N^2}\sum_{i=1}^{N}\{(1-\pi_i^A)/\pi_i^A\}(y_i-\bar{y}_U) & \frac{1}{N^2}\sum_{i=1}^{N}\{1-p_i(\beta)\}(y_i-\bar{y}_U)x_i^{*T} \\ \frac{1}{N^2}\sum_{i=1}^{N}\{1-p_i(\beta)\}(y_i-\bar{y}_U)x_i^{*} & \frac{1}{N^2}\sum_{i=1}^{N}\pi_i^A\{1-p_i(\beta)\}x_i^{*}x_i^{*T} \end{pmatrix} \tag{2.62}$$

The final asymptotic variance estimator of $\hat{\bar{y}}_{PAPW}$ is given by

$$Var\{\hat{\bar{y}}_{PAPW}\} = \frac{1}{N^2}\sum_{i=1}^{N}\left\{\frac{1-\pi_i^A}{\pi_i^A}\right\}(y_i-\bar{y}_U)^2 - 2\frac{b^T}{N^2}\sum_{i=1}^{N}\{1-p_i(\beta)\}(y_i-\bar{y}_U)x_i^* + b^T\left[\frac{1}{N^2}\sum_{i=1}^{N}\pi_i^A\{1-p_i(\beta)\}x_i^*x_i^{*T}\right]b \tag{2.63}$$

To obtain the variance estimate based on the observed samples of S_A and S_R, we substitute the population components with their estimates from the samples.

$$\begin{aligned}\widehat{Var}\{\hat{\bar{y}}_{PAPW}\} = {} & \frac{1}{N^2}\sum_{i=1}^{n_A}\{1-\hat{\pi}_i^A\}\left(\frac{y_i-\bar{y}_U}{\hat{\pi}_i^A}\right)^2 \\ & - 2\frac{\hat{b}^T}{N^2}\sum_{i=1}^{n_A}\{1-p_i(\hat{\beta})\}\left(\frac{y_i-\bar{y}_U}{\hat{\pi}_i^A}\right)x_i^* \\ & + \hat{b}^T\left[\frac{1}{N^2}\sum_{i=1}^{n}p_i(\hat{\beta})x_i^*x_i^{*T}\right]\hat{b}\end{aligned} \tag{2.64}$$

where

$$\hat{b}^T = \left\{\frac{1}{N}\sum_{i=1}^{n_A}\left(\frac{y_i-\bar{y}_U}{\hat{\pi}_i^A}\right)x_i^{*T}\right\}\left\{\frac{1}{N}\sum_{i=1}^{n}p_i(\hat{\beta})x_i^*x_i^{*T}\right\}^{-1} \tag{2.65}$$

2.6.2 Further extensions of the simulation study

I extend the simulation study to further assess the performance of pseudo-weighting approaches in terms of bias reduction and other repeated sampling properties. To show that my PAPP method can potentially work better than those based on a PMLE approach, comparisons are made under various scenarios as below:

1. The sampling design is ignorable for both S_R and S_A given the set of observed common auxiliary variables, X.

2. The sampling design of S_R is non-ignorable given X, but its design variable, D, is available for units in S_A as well.

3. The sampling design of S_R is non-ignorable given X, but D is not available for units in S_A.

4. I replicate the simulation studied by Chen et al. (2017b) to assess how my method performs compared to Chen's method.

For the first three scenarios, a hypothetical population of size $N = 100,000$ is initially generated, for which two covariates, $\{X, D\}$, are considered with the following distribution:

$$\begin{pmatrix} D \\ X \end{pmatrix} \sim MVN\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right) \tag{2.66}$$

where D denotes a design variable in S_R, and X describes the selection mechanism in S_A. Here, I set $\rho = 0.4$, because in reality one would expect some degrees of correlation between the identified set of common covariates and the design variables of S_R. Given X and D, two outcome variables, one continuous (Y^c) and one binary (Y^b), are constructed as below:

$$Y_i^c | X_i = x_i, D_i = d_i \sim N(\mu = 2 + d_i + x_i, \sigma^2 = 1) \tag{2.67}$$

$$Y_i^b | X_i = x_i, D_i = d_i \sim BER\left(p = \frac{e^{-1+d_i+x_i}}{1 + e^{-1+d_i+x_i}}\right) \tag{2.68}$$

In each scenario, I consider an informative sampling strategy with unequal probabilities of inclusion, where the selection mechanism of S_A and S_R depends on X and/or D, respectively. A Poisson sampling method is then employed to draw samples of S_R and S_A, in a way that the average sample sizes remain $n_R = 200$ and $n_A = 1000$, respectively. I choose a situation where $n_R << n_A$, which might be the case in a Big Data setting.

The simulation is iterated $K = 1,000$ times, where for each time, the adjusted mean, standard error (SE), and 95% confidence intervals (CIs) of the response variable are estimated. To make the quasi-random approaches fairly comparable, a GLM is fitted to estimate all three pseudo-weighting approaches, PAPP, PMLE-V, and PMLE-C. In addition, I use a modified delete-*one* JRR method discussed in section 2.2.4 to estimate the variance. To evaluate the repeated sampling properties of the comparing method, relative bias (rBias), relative root mean square error (rMSE), the nominal coverage rate of 95% CIs (crCI), and SE ratio (rSE) are calculated.

2.6.2.1 Scenario 1:

Under this scenario, I define the measures of size corresponding to S_R and S_A as below:

$$\pi^R(x_i) = P(\delta_i^R = 1 | X_i = x_i) = \frac{e^{-6.34+0.5x_i}}{1 + e^{-6.34+0.5x_i}} \tag{2.69}$$

$$\pi^A(x_i) = P(\delta_i^A = 1 | X_i = x_i) = \frac{e^{-4.67-0.4x_i}}{1 + e^{-4.67-0.4x_i}} \tag{2.70}$$

where δ_i^R and δ_i^A are the indicators of being selected in S_R and S_A, respectively. Dependence of both designs on the same set of covariates implies that given X, ignorable condition holds for both S_R and S_A. I later assume that π_i^A's are unknown, and estimate them based on the observed X using different quasi-random approaches.

The simulation results of this scenario have been illustrated in Table 2.3. For both binary and continuous outcome variables, the values of rBias and rMSE are both smaller for the proposed PAPP than those using the PMLE idea. This is not unexpected as PAPP relies on the ignorable condition for both samples. The two PMLE approaches perform quite similarly in terms of bias reduction. Furthermore, it seems the JRR variance estimator works accurately, as the values of SE-ratio and

95% CIs are close to 1 and 95%, respectively.

Table 2.3: Comparing the performance of adjustment methods in the scenario 1.

	Continuous outcome (Y^c)				**Binary outcome (Y^b)**			
Measure	rel-bias	rel-rMSE	Cov. rate	SE ratio	rel-bias	rel-rMSE	Cov. rate	SE ratio
Unweighted	-19.62	19.80	0.00	1.03	-19.05	19.55	0.90	0.98
Fully weighted	0.02	3.01	94.50	1.01	-0.04	5.25	94.90	1.00
PAPP	0.37	4.56	95.10	1.01	0.33	6.28	94.60	1.00
PMLE-V	2.05	8.27	94.80	1.01	2.12	9.60	95.30	1.01
PMLE-C	2.39	8.58	95.10	1.01	2.47	9.91	95.70	1.02

NOTE 1: GLM has been used to predict the pseudo-weights for all approaches.
NOTE 2: Variance estimation is based on a delete-one Jackknife repeated replication method.

2.6.2.2 Scenario 2

Now I consider a situation where the design of S_R is non-ignorable given X. To do so, I define the selection probabilities in the two samples as below:

$$\pi^R(x_i) = P(\delta_i^R = 1 | D_i = d_i) = \frac{e^{-6.34+0.5d_i}}{1+e^{-6.34+0.5d_i}} \tag{2.71}$$

$$\pi^A(x_i) = P(\delta_i^A = 1 | X_i = x_i) = \frac{e^{-4.67-0.4x_i}}{1+e^{-4.67-0.4x_i}} \tag{2.72}$$

Under this scenario, I assume that X and D are observed in both S_R and S_A. This is usually the case when full information about the design of the reference survey is provided to the analyst. PAPP pseudo-weights are then estimated based on both X and D, but estimates for PMLE pseudo-weights only require modeling X as the predictor.

As illustrated in Table 2.4, the PAPP still gives the smallest values of rBias and rMSE for both binary and continuous outcome variables compared to the PMLE methods. In addition, the JRR variance estimator continues to perform well with slight overestimation according to the quantities of SE-ratio and 95% CI rates.

Table 2.4: Comparing the performance of adjustment methods in the scenario 2.

	Continuous outcome (Y^c)				**Binary outcome (Y^b)**			
Measure	rel-bias	rel-rMSE	Cov. rate	SE ratio	rel-bias	rel-rMSE	Cov. rate	SE ratio
Unweighted	-27.32	27.50	0.00	1.00	-23.13	23.52	0.00	0.96
Fully weighted	0.07	3.47	94.90	1.00	0.13	5.25	93.90	0.97
PAPP	**-0.02**	**5.88**	**87.60**	**1.21**	**0.08**	**6.66**	**90.80**	**1.24**
PMLE-V	1.00	8.20	85.70	1.17	1.03	8.60	88.70	1.21
PMLE-C	1.38	8.43	85.90	1.18	1.38	8.82	88.60	1.21

NOTE 1: GLM has been used to predict the pseudo-weights for all approaches.
NOTE 2: Variance estimation is based on a delete-one Jackknife repeated replication method.

2.6.2.3 Scenario 3:

The third scenario is quite similar to the second scenario with a difference in that I assume this time that D is unobserved in the non-probability sample, S_A. Therefore, all I can rely on is to estimate PAPP pseudo-weights based on the observed X. Table 2.5 exhibits the results for bias and variance estimation. As displayed, in terms of bias improvement, PAPP still performs best among the whole applied methods for both outcome variables. The values of SE-ratio also indicate an accurate estimator of variance by the JRR method.

Table 2.5: Comparing the performance of adjustment methods in the scenario 3.

	Continuous outcome (Y^c)				**Binary outcome (Y^b)**			
Measure	rel-bias	rel-rMSE	Cov. rate	SE ratio	rel-bias	rel-rMSE	Cov. rate	SE ratio
Unweighted	-35.61	35.77	0.00	0.99	-27.73	28.02	0.00	0.98
Fully weighted	-0.03	4.00	94.00	0.97	-0.01	5.11	94.40	0.98
PAPP	**0.05**	**6.81**	**95.90**	**1.07**	**0.10**	**6.71**	**95.90**	**1.06**
PMLE-V	2.01	12.27	95.30	1.06	1.81	11.02	95.40	1.06
PMLE-C	2.55	12.72	95.20	1.06	2.28	11.41	95.60	1.06

NOTE 1: GLM has been used to predict the pseudo-weights for all approaches.
NOTE 2: Variance estimation is based on a delete-one Jackknife repeated replication method.

In addition, I replicate the simulation under this scenario for different values of ρ.

It is apparent that $\rho = 0$ means non-ignorability given X in the probability sample, and $\rho = 1$ is a situation identical to Scenario I, where I have ignorability given X in both probability and non-probability sample. As illustrated in Table 2.6, for all the values of ρ, smaller values of rBias and rMSE are associated with the PAPP. There is evidence of more stability of bias and rMSE for the PAPP method, while these quantities are consistently increasing with the increase in ρ value for both PMLE methods.

Table 2.6: Comparing the values of rBias and rMSE for different methods across different values of ρ.

	Continuous outcome (Y^c)					**Binary outcome (Y^b)**				
ρ	UWD	FWD	PAPP	PMLE-V	PMLE-C	UWD	FWD	PAPP	PMLE-V	PMLE-C
					rBias					
0	-19.9	-0.13	**-0.01**	0.43	0.69	-18.90	-0.32	**-0.18**	0.28	0.54
0.1	-21.88	-0.22	**-0.06**	0.66	0.96	-20.40	-0.33	**-0.16**	0.56	0.85
0.2	-23.72	-0.10	**-0.09**	0.45	0.77	-21.43	-0.27	**-0.23**	0.31	0.61
0.3	-25.78	-0.13	**-0.25**	0.59	0.94	-22.43	-0.04	**-0.10**	0.72	1.05
0.4	-27.48	0.09	**0.29**	1.30	1.69	-23.34	0.16	**0.38**	1.34	1.70
0.5	-29.53	-0.08	**0.23**	1.38	1.81	-24.66	-0.07	**0.24**	1.32	1.70
0.6	-31.66	-0.08	**-0.09**	1.43	1.89	-25.58	-0.17	**-0.14**	1.24	1.65
0.7	-33.61	-0.16	**0.34**	2.12	2.63	-26.89	-0.26	**0.22**	1.82	2.27
0.8	-35.61	-0.03	**0.05**	2.01	2.55	-27.73	-0.01	**0.10**	1.81	2.28
0.9	-38.06	-0.23	**0.21**	2.85	3.46	-28.52	0.16	**0.57**	2.84	3.36
					rMSE					
0	20.08	3.03	**4.65**	5.51	5.64	19.33	4.96	**6.00**	6.77	6.88
0.1	22.07	3.18	**4.88**	6.09	6.24	20.82	5.15	**6.29**	7.23	7.37
0.2	23.91	3.32	**5.14**	6.43	6.59	21.84	5.12	**6.34**	7.40	7.54
0.3	25.94	3.23	**5.58**	7.36	7.55	22.78	4.94	**6.34**	7.74	7.92
0.4	27.65	3.56	**5.90**	8.30	8.56	23.71	5.25	**6.78**	8.73	8.96
0.5	29.69	3.58	**6.59**	9.42	9.71	25.00	5.17	**6.96**	9.37	9.63
0.6	31.82	3.61	**6.54**	10.22	10.54	25.90	5.13	**6.80**	9.76	10.04
0.7	33.76	3.70	**6.93**	11.67	12.09	27.17	5.02	**7.17**	10.90	11.26
0.8	35.77	4.00	**6.81**	12.27	12.72	28.02	5.11	**6.71**	11.02	11.41
0.9	38.21	3.93	**7.42**	14.21	14.78	28.81	5.23	**6.94**	12.45	12.94

UWD=unweighted; FWD=fully weighted.
GLM has been used to predict the pseudo-weights for all approaches.

2.6.3 Supplemental results on SPMD/NHTS data

Tables 2.7 through 2.12 provide detailed numerical estimates of mean trip speed, mean daily start time of trips, mean percentage of stop duration per trip, daily percent of trips using interstate highways, daily percent of trip duration spent on interstate highways, and mean annual miles, by gender, age, urbanicity, and vehicle age, type, and make. (See Figures 2.7 and 2.8 for overall estimates.)

Table 2.7: Weighted mean trip average speed (Km/h) across demographics and vehicle characteristics

	Unweighted		**PAPP**		**PMLE-V**		**PMLE-C**	
Characteristic	n	Mean	Mean	95% CI	Mean	95% CI	Mean	95% CI
Gender								
Male	785	43.06	45.8	(45.01,46.58)	44.79	(43.79,45.8)	44.79	(43.79,45.8)
Female	1,239	42.32	46.53	(45.8,47.27)	45.34	(44.36,46.32)	45.34	(44.36,46.32)
Age group								
≤25	124	42.1	47.11	(44.42,49.79)	44.27	(41.62,46.92)	44.27	(41.62,46.92)
(25-45]	898	43.23	46.8	(46.03,47.58)	46.25	(45.19,47.31)	46.25	(45.19,47.31)
(45-65]	856	42.68	46.43	(45.69,47.17)	45.93	(44.93,46.92)	45.93	(44.93,46.92)
>65	146	38.83	41.12	(39.72,42.52)	40.51	(38.69,42.34)	40.51	(38.69,42.34)
Urban size								
≤50	169	51.84	51.83	(50.35,53.31)	52.13	(50.8,53.46)	52.13	(50.8,53.46)
(50-200]	1,665	40.79	40.76	(40.35,41.16)	40.92	(40.42,41.41)	40.92	(40.42,41.41)
(200-500]	81	52.15	51.95	(50.26,53.63)	51.9	(50,53.8)	51.9	(50,53.8)
>500	109	48.95	48.49	(47.04,49.94)	49.12	(47.69,50.56)	49.12	(47.69,50.56)
Vehicle age								
≤5	948	43.54	47.44	(46.73,48.15)	46.81	(45.69,47.93)	46.81	(45.69,47.93)
(5-10]	750	42.47	46.45	(45.63,47.27)	45.28	(43.9,46.66)	45.28	(43.9,46.66)
>10	326	40.23	43.4	(42.25,44.55)	42.15	(40.8,43.5)	42.15	(40.8,43.5)
Vehicle type								
Car	1,171	42.53	46.19	(45.31,47.07)	45.09	(44.08,46.1)	45.09	(44.08,46.1)
Pickup	50	43.56	47.47	(44.81,50.12)	45.36	(42.26,48.46)	45.36	(42.26,48.46)
SUV	583	43.43	47.47	(46.67,48.26)	45.95	(44.72,47.19)	45.95	(44.71,47.19)
Van	220	40.59	43.34	(42.17,44.52)	42.68	(41.24,44.12)	42.68	(41.24,44.12)
Vehicle make								
American	959	44.08	49.22	(48.59,49.86)	49.86	(48.42,51.3)	49.86	(48.42,51.3)
Asian	919	41.18	43.7	(42.95,44.45)	43.33	(42.26,44.41)	43.33	(42.26,44.41)
European	146	41.95	45.78	(43.82,47.74)	44.22	(41.58,46.86)	44.22	(41.58,46.86)
Total	2,024	42.61	46.51	(45.94,47.07)	45.57	(44.69,46.45)	45.57	(44.69,46.45)

NOTE 1: Weights are trimmed using the IQR method.
NOTE 2: Variance estimates for 95% CIs are based on the conditional variance method.

Table 2.8: Weighted daily percentage of trips started between 6AM-10AM across demographics and vehicle characteristics

	Unweighted		PAPP		PMLE-V		PMLE-C	
Characteristic	n	Mean	Mean	95% CI	Mean	95% CI	Mean	95% CI
Gender								
Male	783	22.24	22.12	(21.45,22.78)	22.3	(21.51,23.08)	22.3	(21.51,23.08)
Female	1,237	21.94	21.78	(21.34,22.22)	21.86	(21.29,22.44)	21.86	(21.29,22.44)
Age group								
≤25	123	17.79	16.89	(15.17,18.61)	17.58	(15.39,19.78)	17.58	(15.39,19.78)
(25-45]	896	23.57	23.52	(22.99,24.04)	23.45	(22.76,24.15)	23.45	(22.76,24.15)
(45-65]	855	21.66	21.8	(21.24,22.36)	21.71	(20.94,22.48)	21.71	(20.94,22.48)
>65	146	18.65	18.98	(17.58,20.39)	19.05	(17.43,20.67)	19.05	(17.43,20.67)
Urban size								
≤50	169	22.86	22.28	(20.91,23.65)	22.67	(21.35,23.99)	22.67	(21.35,23.99)
(50-200]	1,662	22.01	21.48	(21,21.95)	21.87	(21.39,22.36)	21.87	(21.39,22.36)
(200-500]	81	20.52	20.25	(18.26,22.25)	20.59	(18.73,22.46)	20.59	(18.73,22.46)
>500	108	22.69	22.79	(21.14,24.45)	22.42	(20.65,24.19)	22.42	(20.65,24.19)
Vehicle age								
≤5	947	22.35	22.29	(21.78,22.8)	22.39	(21.77,23.01)	22.39	(21.77,23.01)
(5-10]	749	22.22	22.09	(21.5,22.69)	22.27	(21.53,23.01)	22.27	(21.53,23.01)
>10	324	20.82	20.51	(19.37,21.64)	20.67	(19.27,22.06)	20.67	(19.27,22.06)
Vehicle type								
Car	1,170	22.07	21.89	(21.36,22.42)	22.12	(21.52,22.72)	22.12	(21.52,22.72)
Pickup	49	21.22	21.01	(18.55,23.47)	21.24	(18.56,23.92)	21.24	(18.56,23.92)
SUV	581	22.06	21.92	(21.3,22.53)	22	(21.33,22.67)	22	(21.33,22.67)
Van	220	22.16	22.19	(21.2,23.17)	22.07	(20.99,23.15)	22.07	(20.99,23.15)
Vehicle make								
American	956	21.8	22.08	(21.53,22.62)	22.16	(21.05,23.26)	22.16	(21.05,23.26)
Asian	918	22.22	21.91	(21.33,22.49)	22.16	(21.52,22.79)	22.16	(21.52,22.79)
European	146	22.69	22.71	(21.24,24.17)	22.77	(21.24,24.31)	22.77	(21.24,24.31)
Total	2,020	22.06	21.95	(21.57,22.33)	22.06	(21.57,22.55)	22.06	(21.57,22.55)

NOTE 1: Weights are trimmed using the IQR method.

NOTE 2: Variance estimates for 95% CIs are based on the conditional variance method.

Table 2.9: Weighted mean percentage of stop duration per trips across demographics and vehicle characteristics.

	Unweighted		PAPP		PMLE-V		PMLE-C	
Characteristic	n	Mean	Mean	95% CI	Mean	95% CI	Mean	95% CI
Gender								
Male	783	10.71	10.53	(10.19,10.86)	10.55	(10.01,11.1)	10.55	(10.01,11.1)
Female	1,237	11.3	10.4	(10.15,10.64)	10.64	(10.23,11.06)	10.64	(10.23,11.06)
Age group								
≤25	123	11.34	10.4	(9.52,11.28)	10.95	(9.81,12.1)	10.95	(9.81,12.1)
(25-45]	896	11.4	10.79	(10.52,11.06)	10.85	(10.39,11.31)	10.85	(10.39,11.31)
(45-65]	855	10.8	10.15	(9.84,10.45)	10.2	(9.69,10.7)	10.2	(9.69,10.7)
>65	146	10.41	9.77	(9.1,10.43)	10	(9.07,10.92)	10	(9.07,10.92)
Urban size								
≤50	169	9.46	9.46	(8.75,10.17)	9.43	(8.68,10.18)	9.43	(8.68,10.18)
(50-200]	1,662	11.39	11.35	(11.13,11.56)	11.29	(10.87,11.71)	11.29	(10.87,11.71)
(200-500]	81	8.71	8.44	(7.62,9.26)	8.65	(7.59,9.71)	8.65	(7.59,9.71)
>500	108	10.52	10.41	(9.72,11.1)	10.35	(9.63,11.07)	10.35	(9.63,11.06)
Vehicle age								
≤5	947	10.96	10.11	(9.83,10.39)	10.23	(9.7,10.77)	10.23	(9.7,10.77)
(5-10]	749	10.93	10.36	(10.05,10.67)	10.5	(10.02,10.98)	10.5	(10.02,10.98)
>10	324	11.72	11.39	(10.9,11.89)	11.41	(10.77,12.06)	11.41	(10.77,12.06)
Vehicle type								
Car	1,170	10.78	10.16	(9.87,10.44)	10.31	(9.82,10.8)	10.31	(9.82,10.8)
Pickup	49	10.75	10.05	(9.1,11)	10.54	(9.48,11.61)	10.54	(9.48,11.61)
SUV	581	11.45	10.83	(10.46,11.2)	11.04	(10.48,11.6)	11.04	(10.48,11.6)
Van	220	11.69	10.99	(10.45,11.53)	11.22	(10.48,11.96)	11.22	(10.48,11.96)
Vehicle make								
American	956	11.07	10.03	(9.74,10.32)	9.95	(8.71,11.19)	9.95	(8.71,11.19)
Asian	918	11.14	10.71	(10.44,10.98)	10.73	(10.26,11.2)	10.73	(10.26,11.2)
European	146	10.66	10.18	(9.55,10.81)	10.22	(9.41,11.02)	10.22	(9.41,11.02)
Total	2,020	11.07	10.39	(10.19,10.59)	10.51	(10.1,10.93)	10.51	(10.1,10.93)

NOTE 1: Weights are trimmed using the IQR method.

NOTE 2: Variance estimates for 95% CIs are based on the conditional variance method.

Table 2.10: Weighted daily percentage of trips used interstate across demographics and vehicle characteristics.

	Unweighted		PAPP		PMLE-V		PMLE-C	
Characteristic	n	Mean	Mean	95% CI	Mean	95% CI	Mean	95% CI
Gender								
Male	783	11.35	12.81	(12.11,13.5)	12.35	(11.47,13.24)	12.35	(11.47,13.24)
Female	1,237	10.59	13	(12.39,13.61)	12.27	(11.52,13.02)	12.27	(11.52,13.02)
Age group								
≤25	123	11.1	15.82	(13.21,18.43)	13.11	(10.44,15.78)	13.11	(10.44,15.78)
(25-45]	896	11.66	13.82	(13.14,14.5)	13.49	(12.61,14.37)	13.49	(12.61,14.37)
(45-65]	855	10.68	12.6	(11.93,13.26)	12.37	(11.5,13.23)	12.37	(11.5,13.23)
>65	146	7.15	7.32	(6.44,8.2)	7.49	(6.33,8.66)	7.49	(6.33,8.66)
Urban size								
≤50	169	14.99	15.37	(13.67,17.06)	14.96	(13.24,16.67)	14.96	(13.24,16.67)
(50-200]	1,662	9.82	9.64	(9.26,10.03)	9.84	(9.39,10.28)	9.84	(9.39,10.28)
(200-500]	81	18.35	18.33	(15.79,20.88)	18.22	(15.63,20.81)	18.22	(15.63,20.81)
>500	108	15.26	14.08	(12.07,16.09)	15.3	(13.27,17.33)	15.3	(13.27,17.33)
Vehicle age								
≤5	947	11.76	14.04	(13.38,14.7)	13.72	(12.78,14.65)	13.72	(12.78,14.65)
(5-10]	749	10.48	12.42	(11.72,13.12)	11.88	(10.88,12.88)	11.88	(10.88,12.88)
>10	324	9.25	11.53	(10.46,12.61)	10.63	(9.32,11.93)	10.63	(9.32,11.93)
Vehicle type								
Car	1,170	11.06	13.3	(12.58,14.02)	12.69	(11.85,13.54)	12.69	(11.85,13.54)
Pickup	49	9.2	10.45	(8.37,12.54)	9.79	(7.34,12.24)	9.79	(7.34,12.24)
SUV	581	11.23	13.39	(12.66,14.12)	12.52	(11.5,13.55)	12.52	(11.5,13.55)
Van	220	9.44	11	(10.03,11.98)	10.61	(9.47,11.75)	10.61	(9.47,11.75)
Vehicle make								
American	956	11.8	14.37	(13.66,15.08)	14.79	(12.43,17.16)	14.79	(12.43,17.16)
Asian	918	9.98	11.64	(10.97,12.31)	11.41	(10.48,12.34)	11.41	(10.48,12.34)
European	146	10.58	12.36	(10.82,13.9)	11.71	(9.96,13.46)	11.71	(9.96,13.46)
Total	2,020	10.88	13.08	(12.61,13.55)	12.57	(11.91,13.23)	12.57	(11.91,13.23)

NOTE 1: Weights are trimmed using the IQR method.

NOTE 2: Variance estimates for 95% CIs are based on the conditional variance method.

Table 2.11: Weighted mean trip duration spent on interstate by demographics and vehicle characteristics.

	Unweighted		PAPP		PMLE-V		PMLE-C	
Characteristic	n	Mean	Mean	95% CI	Mean	95% CI	Mean	95% CI
Gender								
Male	783	3.5	4.07	(3.77,4.37)	3.87	(3.49,4.26)	3.88	(3.49,4.26)
Female	1,237	3.32	4.2	(3.98,4.42)	3.93	(3.64,4.22)	3.93	(3.64,4.22)
Age group								
≤25	123	3.53	4.91	(4.13,5.68)	4.15	(3.28,5.03)	4.15	(3.28,5.03)
(25-45]	896	3.55	4.32	(4.07,4.57)	4.2	(3.86,4.53)	4.2	(3.86,4.53)
(45-65]	855	3.38	4.16	(3.87,4.45)	4.05	(3.64,4.45)	4.05	(3.64,4.45)
>65	146	2.35	2.33	(2.03,2.64)	2.42	(2.06,2.78)	2.42	(2.06,2.78)
Urban size								
≤50	169	5	5.01	(4.3,5.72)	4.97	(4.21,5.73)	4.97	(4.21,5.73)
(50-200]	1,662	3	2.92	(2.79,3.05)	2.98	(2.83,3.13)	2.98	(2.83,3.13)
(200-500]	81	5.68	5.6	(4.81,6.39)	5.57	(4.78,6.36)	5.57	(4.78,6.36)
>500	108	5.24	4.71	(3.99,5.44)	5.18	(4.44,5.92)	5.18	(4.44,5.92)
Vehicle age								
≤5	947	3.8	4.71	(4.45,4.96)	4.55	(4.19,4.9)	4.55	(4.19,4.9)
(5-10]	749	3.18	3.85	(3.55,4.14)	3.66	(3.24,4.09)	3.66	(3.24,4.09)
>10	324	2.69	3.44	(3.09,3.79)	3.15	(2.72,3.59)	3.15	(2.72,3.59)
Vehicle type								
Car	1,170	3.4	4.19	(3.93,4.44)	3.95	(3.66,4.24)	3.95	(3.66,4.24)
Pickup	49	2.73	3.2	(2.53,3.87)	2.95	(2.18,3.73)	2.95	(2.18,3.73)
SUV	581	3.54	4.43	(4.05,4.81)	4.09	(3.57,4.6)	4.09	(3.57,4.6)
Van	220	3.06	3.73	(3.33,4.12)	3.56	(3.07,4.05)	3.56	(3.07,4.05)
Vehicle make								
American	956	3.75	4.83	(4.52,5.14)	5	(4.28,5.72)	5	(4.28,5.72)
Asian	918	3.03	3.61	(3.38,3.84)	3.5	(3.18,3.82)	3.5	(3.18,3.82)
European	146	3.34	3.86	(3.33,4.38)	3.68	(3.07,4.28)	3.68	(3.07,4.28)
Total	2,020	3.39	4.21	(4.02,4.39)	4	(3.74,4.26)	4	(3.74,4.26)

NOTE 1: Weights are trimmed using the IQR method.

NOTE 2: Variance estimates for 95% CIs are based on the conditional variance method.

Table 2.12: Weighted mean annual mileage by demographics and vehicle characteristics.

	Unweighted		PAPP		PMLE-V		PMLE-C	
Characteristic	n	Mean	Mean	95% CI	Mean	95% CI	Mean	95% CI
Gender								
Male	785	20.29	23.29	(22.23,24.34)	22.05	(20.73,23.37)	22.05	(20.73,23.37)
Female	1,239	20.51	23.92	(23.22,24.63)	22.9	(21.97,23.83)	22.9	(21.97,23.83)
Age group								
≤25	124	18.73	22.19	(20.01,24.38)	20.01	(17.63,22.38)	20.01	(17.63,22.38)
(25-45]	898	20.74	23.91	(23.09,24.74)	23.24	(22.18,24.31)	23.24	(22.18,24.31)
(45-65]	856	20.93	24.43	(23.5,25.35)	23.84	(22.55,25.14)	23.84	(22.55,25.14)
>65	146	17.02	18.79	(17.23,20.34)	18.37	(16.24,20.51)	18.37	(16.24,20.51)
Urban size								
≤50	169	28.66	28.54	(26.43,30.64)	28.9	(26.81,30.99)	28.9	(26.81,30.99)
(50-200]	1,665	18.75	18.48	(18,18.97)	18.5	(18.03,18.98)	18.5	(18.03,18.98)
(200-500]	81	28.06	27.21	(25.59,28.84)	27.44	(25.9,28.98)	27.44	(25.9,28.98)
>500	109	27.57	26.76	(24.89,28.63)	27.73	(25.79,29.67)	27.73	(25.79,29.67)
Vehicle age								
≤5	948	22.16	25.61	(24.88,26.33)	24.94	(23.85,26.03)	24.94	(23.85,26.03)
(5-10]	750	20.02	23.53	(22.5,24.57)	22.49	(20.93,24.04)	22.49	(20.93,24.04)
>10	326	16.31	19.03	(17.86,20.2)	17.95	(16.62,19.28)	17.95	(16.62,19.28)
Vehicle type								
Car	1,171	19.47	22.63	(21.76,23.51)	21.62	(20.65,22.6)	21.63	(20.65,22.6)
Pickup	50	19.82	24.19	(21.25,27.13)	21.88	(18.52,25.24)	21.88	(18.52,25.24)
SUV	583	22	25.83	(24.71,26.95)	24.38	(22.9,25.86)	24.38	(22.9,25.86)
Van	220	21.49	24.2	(22.85,25.54)	23.55	(21.92,25.18)	23.55	(21.92,25.18)
Vehicle make								
American	959	21.99	26.93	(26.09,27.76)	27.59	(24.29,30.9)	27.59	(24.29,30.9)
Asian	919	19.04	21.21	(20.39,22.04)	20.67	(19.65,21.69)	20.67	(19.65,21.69)
European	146	18.88	22.16	(20.09,24.24)	20.89	(18.45,23.33)	20.89	(18.45,23.33)
Total	2,024	20.43	23.89	(23.26,24.52)	22.94	(22.04,23.84)	22.94	(22.04,23.85)

NOTE 1: Weights are trimmed using the IQR method.

NOTE 2: Variance estimates for 95% CIs are based on the conditional variance method.

CHAPTER III

Doubly Robust Two-step Bayesian Inference for Non-probability Samples

3.1 Introduction

Chapter II developed a robust quasi-random (QR) approach for finite population inference based on a non-probability sample. Assuming a random selection mechanism for the sample units, my goal was to estimate the missing selection probabilities non-parametrically using Bayesian Additive Regression Trees (BART). In the presence of a "reference survey" with a set of common auxiliary variables, I observed that adjusted estimates are consistent under a *strongly ignorable* condition (**C1-C2**). This assumption, however, may not hold necessarily for the propensity model. It is always likely that some of the key variables governing the selection mechanism in the non-probability sample are unobserved. Although the strong flexibility of BART, as a predictive tool, reduces the risk of model misspecification, I realized that BART performs poorly when the two samples lack common support for the joint distribution of auxiliary variables (Rafei et al., 2020; Hill and Su, 2013).

An alternative model-assisted approach involves *prediction* modeling (PM) where the outcome variable(s) is predicted for units of the reference survey (Rivers, 2007; Kim and Rao, 2012; Wang et al., 2015; Kim et al., 2021a). Therefore, unlike the QR

where the response indicator is modeled, it is the outcome variable that has to be modeled in PM. Note that this method requires one to fit a model separately for any given outcome variable, whereas the estimated set of pseudo-weights by QR could be applied to any outcome variable. For non-normal outcomes, attention should be paid to an appropriate choice of the *link* function as well. Once the outcome is imputed for all units of the reference survey, design-based approaches can be then utilized to compute point and interval estimates. In PM, however, model misspecification is an even bigger concern than in QR as the PM-based estimates rely on extrapolation.

To further protect against model misspecification, a third method can be applied by combining the QR approach with the PM method, in a way that the adjusted estimate of a population quantity, such as the population mean, is consistent if either model does hold. In this sense, adjustments by such a method are called doubly robust (DR). Proposed by Robins et al. (1994), augmented inverse propensity weighting (AIPW) is the earliest class of DR methods, which borrows the idea of a generalized difference estimator from Cassel et al. (1976). This prominence led the AIPW estimator to gain popularity quickly in the causal inference setting (Scharfstein et al., 1999; Bang and Robins, 2005; Tan, 2006; Kang et al., 2007; Tan et al., 2019). It was later expanded to adjust for non-response bias in the survey sampling context (Kott, 1994; Kim and Park, 2006; Kott, 2006; Haziza and Rao, 2006; Kott and Chang, 2010).

A further extension to multiple robustness has been developed by Han and Wang (2013), where multiple models are specified and consistency is obtained as long as at least one of the models are correctly specified. However, simulation results show that a DR estimator is always less efficient than either a correctly-specified QR or PM solely. In situations where both QR and PM are misspecified, even moderately, Kang et al. (2007) show that the AIPW method may not work that well. Cao et al. (2009) conclude that the performance of the QR model in the AIPW estimator depends on how close the inverse PS weighted (IPSW) mean of the selection indicator variable

is to the sample size. As a result, they recommended estimating the parameters of the QR model under the restriction that the sum of the quotients of the selection indicators by PS equals the sample size approximately.

Chen et al. (2019) offer further adjustments to adapt the AIPW estimator to a non-probability sampling setting where an external benchmark survey is available. While their method employs a modified pseudo-likelihood approach to estimate the selection probabilities for the non-probability sample, a parametric model is used to impute the outcome for units of the reference survey. Inspired by Kim and Haziza (2014), the authors propose to estimate the model parameters by simultaneously solving the estimating equations to maintain the DR property for the variance estimator. Wu and Sitter (2001) point out that the AIPW estimator resembles inverse propensity weighting followed by a GREG calibration based on the estimated auxiliary totals from the reference survey. This two-step method has been frequently used elsewhere (Lee and Valliant, 2009; Brick, 2015; Dutwin and Buskirk, 2017; Valliant, 2020). Combining a model-assisted method with pseudo-weighting, Valliant (2020) also proposes an equivalent DR weighting approach for inference in non-probability samples. An extension of AIPW to high-dimensional data using LASSO has also been suggested by Yang et al. (2019).

In reality, however, the functional structure of neither propensity nor prediction models is known to the analyst, and undoubtedly, a DR estimator is no longer consistent if both underlying models are incorrectly specified. To further weaken or relax the modeling assumptions, the current article aims to propose alternative model-assisted DR methods by incorporating more flexible prediction methods, such as supervised machine learning algorithms, into the AIPW estimator. A notable advantage of such predictive tools is automatic variable selection, which features the ability to capture complex non-linear relationships and high-order interactions. As a result, these algorithmic methods, e.g. tree ensembles, kernels, and neural networks, have been

widely used in the contexts of causal inference and incomplete data analysis (Mayer et al., 2020; McConnell and Lindner, 2019; Wendling et al., 2018). However, a major challenge with the use of them in a non-probability sample is how to handle the selection probabilities of the reference survey when fitting the propensity model. Under a Bayesian framework, incorporating the sampling weights into the regression models is an even bigger hurdle (Gelman et al., 2007). The method provided by Chen et al. (2019) relies on the pseudo-likelihood approach, which is generally limited to the parametric models from the exponential family.

To augment the PM estimator while avoiding the creation of synthetic populations, I propose to incorporate the pseudo-weighting approach in Chapter II into the AIPW estimator Elliott and Valliant (2017). As demonstrated, this two-step method computationally separates the propensity model from the sampling weights, allowing for a broader range of models such as algorithmic methods to be utilized for imputing the missing inclusion probabilities. Because of this feature, one can also perform Bayesian PS modeling or Bayesian AIPW under a non-probability sample setting through the well-known two-step method (Kaplan and Chen, 2012; Saarela et al., 2016). A *well-calibrated* Bayesian method can appropriately capture the uncertainty in the imputed PS or the outcome variable via Monte Carlo Markov Chain (MCMC) algorithms, meeting the desirable frequentist repeated sampling properties (Dawid, 1982).

As in Chapter II, in addition to the parametric Bayes, I apply BART to impute both the PS and outcome in the AIPW estimator (Chipman et al., 2007). Using BART, Mercer (2018) compared the AIPW estimator with a prediction model that uses the estimated PS as a predictor in the model and found that the AIPW estimator performed best in terms of both bias and efficiency. His method, however, simulated a synthetic population to cope with the unequal selection probabilities of the reference survey. In an item non-response setting, Tan et al. (2019) exploited BART to compare

the AIPW method with its competitor, the penalized spline of propensity prediction (PSPP) method, where the latter uses BART only for the PS model. According to their simulation study, BART outperformed the GLM when the true models are unknown, but with a slight loss of efficiency. However, PSPP proved to give a smaller root mean square error than AIPW, which contradicted the main finding in Mercer (2018).

To assess the performance of my proposed method under BART, I apply it to the sensor-based Big Data from the second phase of the Strategic Highway Research Program (SHRP2), which is the largest NDS conducted to date. The aim is to adjust for the potential selection bias in the sample mean of some trip-related variables (Antin et al., 2015). To this end, I employ the 2017 National Household Travel Survey (NHTS) as the reference survey, which can serve as a probability sample representing the population of American drivers (Santos et al., 2011). While daily trip measures in SHRP2 are recorded via sensors, NHTS asks respondents to self-report their trip measures through an online travel log. By analyzing the aggregated data at the day level, I compare the DR adjusted sensor-based estimates in SHRP2 with the self-reported weighted estimates in NHTS to assess the performance of my proposed methods in terms of bias and efficiency.

The rest of the article is organized as follows. In Section 3.2, I develop the theoretical background behind the proposed methods and associated variance estimators. A simulation study is designed in Section 3.3 to assess the repeated sampling properties of the proposed estimator, i.e. bias and efficiency. Section 3.4 describes the datasets and auxiliary variables I use in the current study and discusses the results of adjusted estimates based on the combined samples of SHRP2 and NHTS at the day level. Finally, Section 3.5 reviews the strengths and weaknesses of the study in more detail and suggests some future research directions.

3.2 Methods

3.2.1 Prediction modeling approach

Consider the notation and conditions **C1-C4** defined in Section 1.2 of Chapter I, and let $x_i^* = [x_i, d_i]$. As demonstrated in Chapter II, a quasi-random (QA) approach involved modeling $f(\delta^A|x)$. An alternative approach to deal with selectivity in Big Data is modeling $f(y|x^*)$ (Smith, 1983). In a fully model-based fashion, this essentially involves mass imputing y for the population non-sampled units, $U - S_A$. When x^* is unobserved for non-sampled units, it is recommended that a synthetic population is generated by undoing the selection mechanism of S_R through a non-parametric Bayesian bootstrap method using the design variables in S_R (Dong et al., 2014; Zangeneh and Little, 2015). In the non-probability sample context, Elliott and Valliant (2017) propose an extension of the General Regression Estimator (GREG) when only summary information about x^*, such as totals, is known regarding U. In situations where an external probability sample is available with x^* measured, an alternative is to limit the outcome prediction to the units in S_R, and then, use design-based approaches to estimate the population quantity (Rivers, 2007; Kim et al., 2021a; Yang and Kim, 2018).

However, to the best of my knowledge, none of the prior literature distinguish the role of D from X in the conditional mean structure of the outcome, while looking back at Chapter I, the likelihood factorization in Eq. 1.1 indicated that predicting y requires conditioning not only on x but also on d. Suppose U is a realization of a repeated random sampling process from a super-population under the following model:

$$E(y_i|x_i^*;\theta) = m(x_i^*;\theta) \quad \forall i \in U \tag{3.1}$$

where $m(x_i^*;\theta)$ can be either a parametric model with m being a continuous differentiable function or an unspecified non-parametric form. Under the *ignorable* condition

C1, one can show that

$$f(y_i|x_i^*, z_i = 1; \theta) = f(y_i|x_i^*, \delta_i^A = 1; \theta) = f(y_i|x_i, d_i; \theta) \tag{3.2}$$

Using Bayes' rule, I have

$$\begin{aligned} f(y_i|x_i^*, \delta_i^A = 1) &= \frac{f(\delta_i^A = 1|y_i, x_i^*)}{f(\delta_i^A = 1|x_i^*)} f(y_i|x_i^*) \\ &= f(y_i|x_i^*) \end{aligned} \tag{3.3}$$

which implies that a *consistent* estimate of the population parameter θ can be obtained by regressing Y on X^* given S_A.

Under a linear regression model, where $m(x_i^*; \hat{\theta}) = \hat{\theta}_0 + \hat{\theta}_1^T x_i^*$, the maximum likelihood estimation (MLE) of θ is given by

$$\hat{\theta} = X^*(X^{*T}X^*)^{-1}X^{*T}y \tag{3.4}$$

where $X^* = [1, X, D]$ is an $(n_A \times (p + q + 1))$-dimensional design matrix. Note that for y being non-normal, one has to use a generalized linear model (GLM) with an appropriate *link* function, where an MLE of θ is obtained by solving a set of estimating equations. The predictions for units in S_R are then given by

$$\hat{y}_i = E(y_i|x_i^*, z_i = 0; \hat{\theta}) = m(x_i^*; \hat{\theta}) \quad \forall i \in S_R \tag{3.5}$$

Once y is imputed for all units in the reference survey, the population mean can be estimated through the Hájek formula as below:

$$\hat{\bar{y}}_{PM} = \frac{1}{\hat{N}_R} \sum_{i=1}^{n_R} \frac{\hat{y}_i}{\pi_i^R} \tag{3.6}$$

where $\hat{y}_i = m(x_i^*; \hat{\theta})$ for $i \in S_R$, $\hat{N}_R = \sum_{i=1}^{n_R} w_i^R$ and π_i^R is the selection probability for

subject $i \in S$. One can replace $\hat{N}_R$ with N, if known, which yields a HT-estimator.

The asymptotic properties of the estimator in 3.6, including consistency and unbiasedness, have been investigated by Kim et al. (2021a). Note that in situations where π_i^R is available for $i \in S_A$, one can use w_i^R instead of the high-dimensional d_i as a predictor in $m(.)$. This method is known as linear-in-the-weight prediction (LWP) (Scharfstein et al., 1999; Bang and Robins, 2005; Zhang and Little, 2011). However, since outcome imputation relies fully on extrapolation, even minor misspecification of the underlying model can be seriously detrimental to bias correction.

3.2.2 Doubly robust adjustment approach

To reduce the sensitivity to model misspecification, Chen et al. (2019) reconcile the two aforementioned approaches, i.e. QR and PM, in a way that estimates remain consistent even if one of the two models is incorrectly specified. Their method involves an extension of the augmented inverse propensity weighting (AIPW) proposed by Robins et al. (1994). When N is known, the expanded AIPW estimator takes the following form:

$$\bar{y}_{DR} = \frac{1}{N}\sum_{i=1}^{n_A} \frac{\{y_i - m(x_i^*;\theta)\}}{\pi^A(x_i^*;\beta)} + \frac{1}{N}\sum_{j=1}^{n_R} \frac{m(x_i^*;\theta)}{\pi_j^R} \tag{3.7}$$

where given x^*, θ and β are estimated using the MLE method mentioned in the previous section and the modified pseudo-MLE method described in Section 2.1 of Chapter II, respectively. The theoretical proof of the asymptotic unbiasedness of $\bar{y}_{DR}$ under the correct modeling of $\pi^A(x_i^*;\beta)$ or $m(x_i^*;\theta)$ is reviewed in Appendix 3.6.1.

To avoid using π^R in modeling δ_i^A because of the PMLE restrictions I discussed in Section 2.1 of Chapter II, in this study, I suggest estimating π_i^A for $i \in S_A$ in Eq. 3.39 based on the PAPW/PAPP method depending on whether π_i^R is available for $i \in S_A$ or not. As a result, in situations where π_i^R is known for $i \in S_A$, my proposed DR

estimator is given by

$$\hat{\bar{y}}_{DR} = \frac{1}{N}\sum_{i=1}^{n_A}\frac{1}{\pi_i^R}\left[\frac{1-p_i(\beta)}{p_i(\beta)}\right]\{y_i - m(x_i^*;\theta)\} + \frac{1}{N}\sum_{j=1}^{n_R}\frac{m(x_i^*;\theta)}{\pi_j^R} \tag{3.8}$$

where $p_i(\beta) = p(Z_i = 1|x_i^*;\beta)$. I demonstrate that this form of the AIPW estimator is identical to that defined by Kim and Haziza (2014) in the non-response adjustment context under probability surveys. Assuming that y_i is fully observed for $i \in S_R$, let us define the following HT-estimator for the population mean:

$$\hat{\bar{y}}_U = \frac{1}{N}\sum_{i=1}^{n_R}\frac{y_i}{\pi_i^R} \tag{3.9}$$

Now, one can easily conclude that

$$\begin{aligned}\hat{\bar{y}}_{DR} &= \frac{1}{N}\sum_{i=1}^{n}\frac{1}{\pi_i^R}\left[Z_i\left(\frac{1-p_i(\beta)}{p_i(\beta)}\right)\{y_i - m(x_i^*;\theta)\} + (1-Z_i)m(x_i^*;\theta)\right] \\ &= \hat{\bar{y}}_U + \frac{1}{N}\sum_{i=1}^{n}\frac{1}{\pi_i^R}\left[\frac{Z_i}{p_i(\beta)} - 1\right]\{y_i - m(x_i^*;\theta)\}\end{aligned} \tag{3.10}$$

where $p_i(\beta) = p(Z_i = 1|x_i^*;\beta)$. The formula in 2.18 is similar to what is derived by Kim and Haziza (2014). Therefore, the rest of the theoretical proof of asymptotic unbiasedness, i.e. $\hat{\bar{y}}_{DR} - \bar{\hat{y}}_U = O_p(n_A^{-1/2})$, in Kim and Haziza (2014) should hold for the modified AIPW estimator in 3.40 as well.

To preserve the DR property for both the point and variance estimator of $\bar{y}_{DR}$, as suggested by Kim and Haziza (2014), one can solve the following estimating equations simultaneously given S_C to obtain the estimate of (β, θ). The aim is to cancel the first-order derivative terms in the Taylor-series expansion of $\hat{\bar{y}}_{DR} - \hat{\bar{y}}_U$ under QR and

PM. These estimating equations are given by

$$\begin{aligned}
\frac{\partial}{\partial\beta}\left[\hat{\bar{y}}_{DR}-\hat{\bar{y}}_{U}\right] &= \frac{1}{N}\sum_{i=1}^{n}\frac{Z_i}{\pi_i^R}\left[\frac{1}{p_i(\beta)}-1\right]\{y_i-m(x_i^*;\theta)\}x_i^* = 0 \\
\frac{\partial}{\partial\theta}\left[\hat{\bar{y}}_{DR}-\hat{\bar{y}}_{U}\right] &= \frac{1}{N}\sum_{i=1}^{n}\frac{1}{\pi_i^R}\left[\frac{Z_i}{p_i(\beta)}-1\right]\dot{m}(x_i^*;\theta) = 0
\end{aligned} \tag{3.11}$$

where $\dot{m}$ is the derivative of m with respect to β. Under a linear regression model, $\dot{m}(x_i^*) = x_i^*$. Therefore, given the same regularity conditions, ignorability in S_A, the logistic regression model as well as the additional imposed assumption of $S_A \cap S_R = \emptyset$, one can show that the proposed DR estimator is consistent and approximately unbiased given that either the QR or PM model holds.

It is important to note that the system of equations in 3.11 may not have unique solutions unless the dimension of covariates in QR and PM is identical. Therefore, the AIPW estimator by Chen et al. (2019) may not be applicable here, as my likelihood factorization suggests that conditioning on d_i is necessary at least for the PM. Furthermore, when π_i^R is known for $i \in S_A$, one can replace the q-dimensional d_i with the 1-dimensional w_i^R in modeling both QR and PM. Bang and Robins (2005) show that estimators based on a linear-in-weight prediction model remains consistent.

3.2.3 Extensions to a two-step Bayesian framework

A fully Bayesian approach specifies a model for the joint distribution of selection indicator, δ_i^A, and the outcome variable, y_i, for $i \in U$ (McCandless et al., 2009; An, 2010). This requires multiply generating synthetic populations and fitting the QR and PM models on each of them repeatedly (Little and Zheng, 2007; Zangeneh and Little, 2015), which can be computationally expensive under a Big Data setting. While joint modeling may result in good frequentist properties (Little, 2004), feedback occurs between the two models (Zigler et al., 2013). This can be controversial in the sense that PS estimates should not be informed by the outcome model (Rubin, 2007).

Here, I am interested in modeling the PS and the outcome separately through the two-step framework proposed by Kaplan and Chen (2012). The first step involves fitting Bayesian models to multiply impute the PS and the outcome by randomly subsampling the posterior predictive draws, and Rubin's combining rules are utilized as the second step to obtain the final point and interval estimates. This method not only is computationally efficient as it suffices to fit the models once and on the combined sample but also cuts the undesirable feedback between the models as they are fitted separately. Bayesian modeling can be performed either parametrically or non-parametrically.

3.2.3.1 Parametric Bayes

As the first step, I employ Bayesian Generalized Linear Models to handle multiple imputations of π_i^A and y_i for $i \in S$, and π_i^R if it is unknown for $i \in S_R$. Under a standard Bayesian framework, a set of independent prior distributions are assigned to the model parameters, and conditional on the observed data, the associated posterior distributions are simulated through an appropriate MCMC method, such as the Metropolis-Hastings algorithm. I propose the following steps:

$$
\begin{aligned}
Step1: &\quad (\gamma^T, \phi, \beta^T, \theta^T, \sigma) \sim p(\gamma)p(\phi)p(\beta)p(\theta)p(\sigma) \\
Step2: &\quad \pi_i^R | x_i, \gamma, \phi \sim Beta(\phi[logit^{-1}(\gamma^T x_i)], \phi[1 - logit^{-1}(\gamma^T x_i)]) \\
Step3: &\quad Z_i | x_i, \beta \sim Bernoulli(logit^{-1}\{\beta^T x_i\}) \\
Step4: &\quad Y_i | x_i, \theta, \sigma \sim Normal(\theta^T x_i, \sigma^2)
\end{aligned}
$$

where (γ^T, ϕ), β^T and (θ^T, σ) are the parameters associated with modeling π_i^R in a Beta regression ($Step2$), Z_i in a binary logistic regression ($Step3$) and Y_i is a linear regression ($Step4$), respectively, and $p(.)$ denotes a prior density function. Note that in situations where π_i^R is calculable for $i \in S_A$, $Step2$ should be skipped, and x_i should be replaced by x_i^*. It is understood that setting non-informative priors to

the model parameters can avoid overestimating the variance in a two-step Bayesian method (Kaplan and Chen, 2012). I also note that $Step2$, which will be required for the estimation of π_i^R when not provided directly or through the availability of d_i in S_B, relies on a reasonably strong association between the available x_i and π_i^R in order to accurately estimate π_i^R. I explore the effect of differing degrees of this association via simulation in Sections 3.3.2 and 3.3.3.

Suppose $\hat{\Theta}^{(m)T} = \left[(\hat{\gamma}^{(m)T}, \hat{\phi}^{(m)}, \hat{\beta}^{(m)T}, \hat{\theta}^{T(m)}, \hat{\sigma}^{(m)}\right]$ is the m-th unit of an M-sized random sample from the MCMC draws associated with the posterior distribution of the models parameters. Then, given that π_i^R is known for $i \in S_A$, one can obtain the m-th draw of the proposed AIPW estimator as below:

$$\hat{\bar{y}}_{DR}^{(m)} = \frac{1}{\hat{N}_A} \sum_{i=1}^{n_A} \frac{y_i - \hat{\theta}^{(m)T} x_i^*}{\pi_i^R exp[\hat{\beta}^{(m)T} x_i^*]} + \frac{1}{\hat{N}_R} \sum_{j=1}^{n_R} \frac{\hat{\theta}^{(m)T} x_i^*}{\pi_j^R} \tag{3.12}$$

In situations where π_i^R is unknown for $i \in S_A$, the m-th draw of the AIPW estimator is given by

$$\hat{\bar{y}}_{DR}^{(m)} = \frac{1}{\hat{N}_A} \sum_{i=1}^{n_A} \left\{ \frac{1 + exp[\hat{\gamma}^{(m)T} x_i]}{exp[\hat{\gamma}^{(m)T} x_i]} \right\} \left\{ \frac{y_i - \hat{\theta}^{(m)T} x_i^*}{exp[\hat{\beta}^{(m)T} x_i]} \right\} + \frac{1}{\hat{N}_R} \sum_{j=1}^{n_R} \frac{\hat{\theta}^{(m)T} x_i^*}{\pi_j^R} \tag{3.13}$$

Having $\hat{\bar{y}}_{DR}^{(m)}$ for all $m = 1, 2, ..., M$, then, Rubin's combining rule for the point estimate (Rubin, 2004) can be employed to obtain the final AIPW estimator as below:

$$\hat{\bar{y}}_{DR} = \frac{1}{M} \sum_{m=1}^{M} \hat{\bar{y}}_{DR}^{(m)} \tag{3.14}$$

If at least one of the underlying models is correctly specified, I would expect that this estimator is approximately unbiased. The variance estimation under the two-step Bayesian method is discussed in Section 2.5.

3.2.3.2 Non-parametric Bayes

Despite the prominent feature of double robustness, for a given non-probability sample, neither QR nor PM has a known modeling structure in practice. When both working models are invalid, the AIPW estimator will be biased and a non-robust estimator based on PM may produce a more efficient estimate than the AIPW (Kang et al., 2007). To show the advantage of my modified estimator in Eq. 3.40 over that proposed by Chen et al. (2019), I employ Bayesian Additive Regression Trees (BART) as a predictive tool for multiply imputing the π_i^A's as well as the y_i's in S. A brief introduction to BART was provided in Section 2.2.2 of Chapter II.

Suppose BART approximates a continuous outcome variable through an implicit function f as below:

$$y_i = f(x_i^*) + \epsilon_i \qquad \forall i \in S_A \tag{3.15}$$

where $\epsilon_i \sim N(0, \sigma^2)$. Accordingly, one can train BART in S_A and multiply impute y_i for $i \in S_R$ using the simulated posterior predictive distribution. Regarding QR, I consider two situations; first, π_i^R is known for $i \in S_A$. Under this circumstance, it suffices to model z_i on x_i^* in S to estimate π_i^A for $i \in S_A$. For a binary outcome variable, BART utilizes a data augmentation technique with respect to a given *link* function, to map $\{0, 1\}$ values to $\mathbb{R}$ via a *probit* link. Suppose

$$\Phi^{-1}[p(Z_i = 1|x_i^*)] = h(x_i^*) \qquad \forall i \in S \tag{3.16}$$

where Φ^{-1} is the inverse CDF of the standard normal distribution. Hence, using the posterior predictive draws generated by BART in Eq. 3.16, $p(Z_i = 1|x_i^*)$ and consequently π_i^A can be imputed multiply for $i \in S_A$. For a given imputation m

$(m = 1, 2, ..., M)$, one can expand the DR estimator in 2.16 as below:

$$\hat{\bar{y}}_{DR}^{(m)} = \frac{1}{\hat{N}_A} \sum_{i=1}^{n_A} \frac{1}{\pi_i^R} \left\{ \frac{1 - \Phi[\hat{h}^{(m)}(x_i^*)]}{\Phi[\hat{h}^{(m)}(x_i^*)]} \right\} \{ y_i - \hat{f}^{(m)}(x_i^*) \} + \frac{1}{\hat{N}_R} \sum_{j=1}^{n_R} \frac{\hat{f}^{(m)}(x_j^*)}{\pi_j^R} \tag{3.17}$$

where $\hat{f}^{(m)}(.)$ and $\hat{h}^{(m)}(.)$ are the constructed sum-of-trees associated with the m-th MCMC draw in Eq. 3.15 and Eq. 3.16, respectively, after training BART on the combined sample.

Secondly, in situations where π_i^R is not available for $i \in S_A$, I suggest applying BART to multiply impute the missing π_i^R's in S_A. Since the outcome is continuous bounded within $(0, 1)$, a *logit* transformation of the π_i^R's can be used as the outcome variable in BART to map the values to $\mathbb{R}$. Such a model is presented by

$$log \left(\frac{\pi_i^R}{1 - \pi_i^R} \right) = k(x_i) + \epsilon_i \qquad \forall i \in S_R \tag{3.18}$$

where k is a sum-of-trees function approximated by BART. Under this circumstance, $\hat{\bar{y}}_{DR}$ based on the m-th draw from the posterior distribution is given by

$$\begin{aligned} \hat{\bar{y}}_{DR}^{(m)} = \frac{1}{\hat{N}_A} \sum_{i=1}^{n_A} \left\{ \frac{1 + exp[\hat{k}^{(m)}(x_i)]}{exp[\hat{k}^{(m)}(x_i)]} \right\} \left\{ \frac{1 - \Phi[\hat{h}^{(m)}(x_i)]}{\Phi[\hat{h}^{(m)}(x_i)]} \right\} \{ y_i - \hat{f}^{(m)}(x_i^*) \} \\ + \frac{1}{\hat{N}_R} \sum_{j=1}^{n_R} \frac{\hat{f}^{(m)}(x_j^*)}{\pi_j^R} \end{aligned} \tag{3.19}$$

Having $\hat{\bar{y}}_{DR}^{(m)}$ estimated for $m = 1, 2, ..., M$, one can eventually use Rubin's combining rule (Rubin, 2004) to obtain the ultimate point estimate as in 3.14.

3.2.4 Variance estimation

To obtain an unbiased variance estimate for the proposed DR estimator, one needs to account for three sources of uncertainty: (i) the uncertainty due to estimated pseudo-weights in S_A, (ii) the uncertainty due to the predicted outcome in both S_A

and S_R, and (iii) the uncertainty due to sampling itself. In the following, I consider two scenarios:

3.2.4.1 Scenario I: π_i^R is known for $i \in S_A$

In this scenario, the derivation of the asymptotic variance estimator for $\hat{\bar{y}}_{DR}$ is then straightforward and follows Chen et al. (2019). It is given by

$$\widehat{Var}(\hat{\bar{y}}_{DR}) = \hat{V}_1 + \hat{V}_2 - \hat{B}(\hat{V}_2) \tag{3.20}$$

where $V_1 = Var(\hat{\bar{y}}_{PM})$ under the sampling design of S_R. V_2 is the variance of $\hat{\bar{y}}_{DR} - \hat{\bar{y}}_U$ under the joint sampling design of S_R and the PS model. This quantity can be estimated from S_R as below:

$$\hat{V}_2 = \frac{1}{N^2} \sum_{i=1}^{n_A} \left[\frac{1 - \hat{\pi}_i^A}{(\hat{\pi}_i^A)^2} \right] \{y_i - m(x_i^*; \hat{\theta})\}^2 \tag{3.21}$$

Finally, $B(\hat{V}_2)$ corrects for the bias in V_2 under the PM, and is given by

$$\hat{B}(\hat{V}_2) = \frac{1}{N^2} \sum_{i=1}^{n_C} \left[\frac{Z_i}{\hat{\pi}_i^A} - \frac{1 - Z_i}{\pi_i^R} \right] \hat{\sigma}_i^2 \tag{3.22}$$

where $\hat{\sigma}_i^2 = \widehat{Var}(y_i|x_i)$. Since the quantity in 3.22 tends to be *zero* asymptotically under the QR model, the derived variance estimator in 3.20 is DR. Note that such an asymptotic estimator needs N to be known.

3.2.4.2 Scenario II: π_i^R is unknown for $i \in S_A$

To estimate the variance of $\hat{\bar{y}}_{DR}$ in 3.40 under the GLM, I employ the bootstrap repeated replication method proposed by Rao and Wu (1988). For a given replication b ($b = 1, 2, ..., B$), I draw replicated bootstrap subsamples, $S_R^{(b)}$ and $S_A^{(b)}$, of sizes $n_R - 1$ and $n_A - 1$ from S_R and S_A, respectively. The sampling weights in $S_R^{(b)}$ are updated

as below:

$$w_i^{(b)} = w_i \frac{n_R}{n_R - 1} h_i \qquad \forall i \in S_R^{(b)} \tag{3.23}$$

where h_i is the number of times the i-th unit has been repeated in $S_A^{(b)}$. Let's assume $\hat{\bar{y}}_{DR}^{(b)}$ is the DR estimate based on the b-th combined bootstrap sample, $S^{(b)}$, using Eq. 3.40. The variance estimator is given by

$$\widehat{Var}(\hat{\bar{y}}_{DR}) = \frac{1}{B} \sum_{b=1}^{A} \left[\hat{\bar{y}}_{DR}^{(b)} - \bar{\bar{y}}_{DR} \right]^2 \tag{3.24}$$

where $\bar{\bar{y}}_{DR} = \sum_{b=1}^{A} \hat{\bar{y}}_{DR}^{(b)} / B$. Note that when S_R and S_A are clustered, which is the case in my application, bootstrap subsamples are selected from the primary sampling units (PSU), and n_R and n_A are replaced by their respective PSU sizes.

To estimate the variance of $\hat{\bar{y}}_{DR}$ under a Bayesian framework, whether parametric or non-parametric, I treat y_i for $i \in S_R$, and π_i^R and e_i for $i \in S_A$, as missing values in Eq. 3.40 and multiply impute these quantities using the Monte Carlo Markov Chain (MCMC) sequence of the posterior predictive distribution generated by BART. For M randomly selected MCMC draws, one can estimate $\hat{\bar{y}}_{DR}^{(m)}$ for $m = 1, 2, ..., M$ based on Eq. 3.40. Following Rubin's combining rule for variance estimation under multiple imputation, the final variance estimate of $\hat{\bar{y}}_{DR}$ is given as below:

$$\widehat{Var}(\hat{\bar{y}}_{DR}) = \bar{V}_W + \left(1 + \frac{1}{M}\right) V_B \tag{3.25}$$

where $\bar{V}_W = \sum_{m=1}^{M} \widehat{Var}(\hat{\bar{y}}_{DR}^{(m)})/M$, $V_B = \sum_{m=1}^{M} (\hat{\bar{y}}_{DR}^{(m)} - \bar{\bar{y}}_{DR})^2/(M-1)$ and $\bar{\bar{y}}_{DR} = \sum_{m=1}^{M} \hat{\bar{y}}_{DR}^{(m)}/M$. I have shown in the Appendix 3.6.2 that the within-imputation component can be approximated by

$$\widehat{Var}(\hat{\bar{y}}_{DR}^{(m)}) \approx \frac{1}{\hat{N}_A^2} \sum_{i=1}^{n_A} \frac{var(y_i)}{(\hat{\pi}_i^A)^2} + \frac{1}{\hat{N}_R^2} var\left(\frac{1}{\pi_i^R}\right) \left\{ \sum_{i=1}^{n_R} \left(\hat{y}_i^{(m)}\right)^2 + n_R \left(\frac{\hat{t}_R}{\hat{N}_R}\right)^2 - 2 \sum_{i=1}^{n_R} \hat{y}_i^{(m)} \right\} \tag{3.26}$$

where $t_R = \sum_{i=1}^{n_R} \hat{y}_i^{(m)}/\pi_i^R$. Note that when S_R or S_A is clustered, under a Bayesian framework, it is important to fit multilevel models to obtain unbiased variance (Zhou et al., 2020). In addition, one needs to account for the intraclass correlation across the sample units in $\widehat{Var}(\hat{\bar{y}}_{DR}^{(m)})$ for $m = 1, 2, ..., M$. Further, one may use the extension of BART with random intercept to properly specify the working models under a cluster sampling design (Tan et al., 2016).

3.3 Simulation study

Three simulations are studied in this section to assess the performance of my proposed methods and associated variance estimators in terms of bias magnitude and other repeated sampling properties. To this end, I consider various situations depending on whether π_i^R is available for $i \in S_A$ or not, and whether units of S_A are independent or not.

3.3.1 Simulation I

The design of my first simulation is inspired by the one implemented in Chen et al. (2019). For all three studies, the non-probability samples are given a random selection mechanism with unequal probabilities, but it is later assumed that these selection probabilities are unknown at the stage of analysis, and the goal is to adjust for the selection bias using a parallel probability sample whose sampling mechanism is known. I conduct the simulation under both asymptotic frequentist and two-step Bayesian frameworks. Consider a finite population of size $N = 10^6$ with $z = \{z_1, z_2, z_3, z_4\}$ being a set of auxiliary variables generated as follows:

$$z_1 \sim Ber(p = 0.5) \qquad z_2 \sim U(0, 2) \qquad z_3 \sim Exp(\mu = 1) \qquad z_4 \sim \chi^2_{(4)} \tag{3.27}$$

and $x = \{x_1, x_2, x_3, x_4\}$ is defined as a function of z as below:

$$x_1 = z_1 \qquad x_2 = z_2 + 0.3z_1 \qquad x_3 = z_3 + 0.2(x_1 + x_2) \qquad x_4 = z_4 + 0.1(x_1 + x_2 + x_3) \tag{3.28}$$

Given x, a continuous outcome variable y is defined by

$$y_i = 2 + x_{1i} + x_{2i} + x_{3i} + x_{4i} + \sigma\epsilon_i \tag{3.29}$$

where $\epsilon_i \sim N(0, 1)$, and σ is defined such that the correlation between y_i and $\sum_{k=1}^{4} x_{ki}$ equals ρ, which takes one of the values $\{0.2, 0.5, 0.8\}$. Further, associated with the design of S_A, a set of selection probabilities are assigned to the population units through the following logistic model:

$$log\left(\frac{\pi_i^A}{1 - \pi_i^A}\right) = \gamma_0 + 0.1x_{1i} + 0.2x_{2i} + 0.1x_{3i} + 0.2x_{4i} \tag{3.30}$$

where γ_0 is determined such that $\sum_{i=1}^{N} \pi_i^A = n_A$. For S_R, I assume $\pi_i^R \propto \gamma_1 + z_{3i}$ where γ_1 is obtained such that $max\{\pi_i^R\}/\min\{\pi_i^R\} = 50$. Hence, π_i^R is assumed to be known for $i \in S_A$ as long as z_3 is observed in S_A. Using these measures of size, I repeatedly draw pairs of samples of sizes $n_R = 100$ and $n_A = 1,000$ associated with S_R and S_A from U through a Poisson sampling method. Note that units in both S_R and S_A are independently selected, and $n_R << n_A$, which might be the case in a Big Data setting. Extensions with $n_A = 100$ and $n_A = 10,000$ for both frequentist and Bayesian methods are provided in Appendix 3.6.2.

Once S_A and S_R are drawn from U, I assume that the π_i^A's for $i \in S_A$ and y_i's for $i \in S_R$ are unobserved, and the aim is to adjust for the selection bias in S_A based on the combined sample, S. The simulation is then iterated $K = 5,000$ times, where the bias-adjusted mean, SE, and associated 95% confidence interval (CI) for the mean of y are estimated in each iteration. Under the frequentist approach, the AIPW point

estimates are obtained by simultaneously solving the estimating equations in 2.19. In addition, the proposed two-step method is used to derive the AIPW point estimates under the parametric Bayes. Also, to estimate the variance, I use the DR asymptotic method proposed by Chen et al. (2019), and the conditional variance formula in Eq. 3.25 under the frequentist and Bayesian approaches, respectively. For the latter, I set flat priors to the model parameters, and simulate the posteriors using $1,000$ MCMC draws after omitting an additional $1,000$ draws as the burn-in period. I then get a random sample of size $M = 200$ from the posterior draws to obtain the point and variance estimates.

To evaluate the repeated sampling properties of the competing method, relative bias (rBias), relative root mean square error (rMSE), the nominal coverage rate of 95% CIs (crCI) and SE ratio (rSE) are calculated as below:

$$rbias(\hat{\bar{y}}_{DR}) = 100 \times \frac{1}{K}\sum_{k=1}^{K}\left(\hat{\bar{y}}_{DR}^{(k)} - \bar{y}_U\right)/\bar{y}_U \tag{3.31}$$

$$rMSE(\hat{\bar{y}}_{DR}) = 100 \times \sqrt{\frac{1}{K}\sum_{k=1}^{K}\left(\hat{\bar{y}}_{DR}^{(k)} - \bar{y}_U\right)^2}/\bar{y}_U \tag{3.32}$$

$$crCI(\hat{\bar{y}}_{DR}) = 100 \times \frac{1}{K}\sum_{k=1}^{K} I\left(\left|\hat{\bar{y}}_{DR}^{(k)} - \bar{y}_U\right| < z_{0.975}\sqrt{var(\hat{\bar{y}}_{DR}^{(k)})}\right) \tag{3.33}$$

$$rSE(\hat{\bar{y}}_{DR}) = \frac{1}{K}\sum_{k=1}^{K}\sqrt{var(\hat{\bar{y}}_{DR}^{(k)})}/\sqrt{\frac{1}{K-1}\sum_{k=1}^{K}\left(\hat{\bar{y}}_{DR}^{(k)} - \bar{\bar{y}}_{DR}\right)^2} \tag{3.34}$$

where $\hat{\bar{y}}_{DR}^{(k)}$ denotes the DR adjusted sample mean from iteration k, $\bar{\bar{y}}_{DR} = \sum_{k=1}^{K}\hat{\bar{y}}_{DR}^{(k)}/K$, $\bar{y}_U$ is the finite population true mean, and $var(.)$ represents the variance estimate of the adjusted mean based on the sample. Finally, I investigate different scenarios of whether models are correctly specified or not to test if my proposed method is DR. In order to misspecify a model, I remove x_4 from the predictors of the working model.

Table 3.1 summarizes the results of the first simulation study under the frequentist approach. As illustrated, unweighted estimates of the population mean are biased in

Table 3.1: Comparing the performance of the bias adjustment methods and associated asymptotic variance estimator under the frequentist approach in the first simulation study for $\rho = \{0.2, 0.5, 0.8\}$

	$\rho = 0.2$				$\rho = 0.5$				$\rho = 0.8$			
Method	rBias	rMSE	crCI	rSE	rBias	rMSE	crCI	rSE	rBias	rMSE	crCI	rSE
Unweighted	8.528	19.248	92.580	1.009	8.647	11.065	77.400	1.018	8.682	9.719	50.880	1.020
Fully weighted	-0.029	20.276	94.740	1.001	0.006	8.035	95.080	1.010	0.015	5.008	94.880	1.008
Non-probability sample (S_A)												
Unweighted	31.742	32.230	0.000	1.009	31.937	32.035	0.000	1.012	31.996	32.049	0.000	1.013
Fully weighted	0.127	6.587	95.440	1.013	0.078	2.583	95.660	1.014	0.061	1.554	95.440	1.012
Non-robust adjustment												
Model specification: True												
PAPW	-1.780	8.088	96.960	1.107	-1.906	4.734	95.680	1.103	-1.947	4.186	94.040	1.100
IPSW	-3.054	10.934	97.240	1.305	-3.134	8.145	95.220	1.173	-3.160	7.778	92.380	1.067
PM	0.490	7.577	95.160	1.007	0.190	4.668	94.620	0.991	0.095	4.204	94.560	0.985
Model specification: False												
PAPW	26.338	27.089	3.140	1.112	26.434	26.618	0.000	1.123	26.461	26.580	0.000	1.128
IPSW	28.269	28.917	0.580	1.021	28.474	28.648	0.000	1.018	28.536	28.654	0.000	1.014
PM	28.093	28.750	0.640	1.022	28.315	28.494	0.000	1.022	28.382	28.505	0.000	1.021
Doubly robust adjustment												
Model specification: QR–True, PM–True												
AIPW–PAPW	0.238	8.070	95.160	1.017	0.100	4.787	95.020	0.996	0.056	4.235	94.640	0.987
AIPW–IPSW	0.105	7.861	95.100	1.019	0.053	4.737	94.760	0.996	0.036	4.222	94.600	0.987
Model specification: QR–True, PM–False												
AIPW–PAPW	0.311	8.197	95.420	1.021	0.172	4.988	95.000	1.013	0.127	4.460	95.180	1.011
AIPW–IPSW	0.222	7.962	95.460	1.024	0.170	4.901	95.420	1.019	0.152	4.405	95.300	1.018
Model specification: QR–False, PM–True												
AIPW–PAPW	0.877	13.362	96.900	1.028	0.327	6.089	95.820	1.027	0.154	4.523	95.240	1.006
AIPW–IPSW	0.609	12.532	96.580	1.025	0.232	5.842	95.500	1.022	0.113	4.464	95.340	1.003
Model specification: QR–False, PM–False												
AIPW–PAPW	28.301	28.995	1.000	1.024	28.392	28.579	0.000	1.021	28.419	28.546	0.000	1.018
AIPW–IPSW	28.104	28.762	0.720	1.024	28.313	28.493	0.000	1.023	28.376	28.500	0.000	1.022

PAPW: propensity-adjusted probability weighting; IPSW: Inverse propensity score weighting; QR: quasi-randomization; PM: prediction model; AIPW: augmented inverse propensity weighting.

Table 3.2: Comparing the performance of the bias adjustment methods and associated variance estimator under the two-step parametric Bayesian approach in the first simulation study for $\rho = \{0.2, 0.5, 0.8\}$

	$\rho = 0.2$				$\rho = 0.5$				$\rho = 0.8$			
Method	rBias	rMSE	crCI	rSE	rBias	rMSE	crCI	rSE	rBias	rMSE	crCI	rSE
Non-robust adjustment												
Model specification: True												
Unweighted	8.528	19.248	92.580	1.009	8.647	11.065	77.400	1.018	8.682	9.719	50.880	1.020
Fully weighted	-0.029	20.276	94.740	1.001	0.006	8.035	95.080	1.010	0.015	5.008	94.880	1.008
Non-probability sample (S_A)												
Unweighted	31.620	32.106	0.000	1.014	31.906	32.003	0.000	1.015	31.993	32.045	0.000	1.017
Fully weighted	0.026	6.615	95.260	1.010	0.052	2.604	95.240	1.007	0.059	1.564	95.240	1.006
Non-robust adjustment												
Model specification: True												
PAPW	-1.846	8.148	96.340	1.081	-1.890	4.749	96.860	1.163	-1.906	4.195	96.560	1.200
PAPP	0.113	7.566	96.500	1.076	0.117	4.302	97.700	1.140	0.117	3.759	97.900	1.164
PM	0.385	7.534	95.180	1.027	0.151	4.644	95.060	1.001	0.078	4.190	95.000	0.989
Model specification: False												
PAPW	26.290	27.041	2.280	1.051	26.499	26.687	0.000	1.071	26.562	26.684	0.000	1.083
PAPP	28.151	28.784	0.500	1.038	28.446	28.612	0.000	1.025	28.535	28.647	0.000	1.015
PM	27.981	28.641	0.840	1.040	28.291	28.472	0.000	1.025	28.384	28.510	0.000	1.015
Doubly robust adjustment												
Model specification: QR–True, PM–True												
AIPW–PAPW	0.115	8.093	96.940	1.097	0.057	4.764	97.120	1.121	0.037	4.219	97.200	1.130
AIPW–PAPP	0.009	7.803	96.600	1.083	0.019	4.704	96.980	1.106	0.020	4.206	96.960	1.114
Model specification: QR–True, PM–False												
AIPW–PAPW	-0.016	7.930	97.180	1.108	-0.080	4.444	97.940	1.166	-0.098	3.842	98.140	1.193
AIPW–PAPP	-0.079	7.648	96.820	1.095	-0.074	4.411	97.700	1.151	-0.069	3.867	97.900	1.175
Model specification: QR–False, PM–True												
AIPW–PAPW	0.557	7.693	96.380	1.086	0.214	4.669	96.760	1.092	0.105	4.195	96.600	1.090
AIPW–PAPP	0.392	7.526	95.980	1.067	0.155	4.637	96.340	1.077	0.080	4.189	96.420	1.078
Model specification: QR–False, PM–False												
AIPW–PAPW	28.167	28.864	1.360	1.096	28.359	28.549	0.000	1.082	28.416	28.548	0.000	1.068
AIPW–PAPP	27.990	28.647	0.980	1.069	28.289	28.471	0.000	1.059	28.379	28.506	0.000	1.049

PAPW: Propensity-adjusted probability weighting; PAPP: Propensity-adjusted probability prediction; QR: quasi-randomization; PM: prediction model; AIPW: augmented inverse propensity weighting.

both S_R and S_A. For the non-robust estimators, when the working model is valid, it seems that PM outperforms QR consistently in terms of bias correction across different ρ values. While PAPW works slightly better than IPSW with respect to bias, when the QR model is true, the latter tends to overestimate the variance slightly according to the values of rSE. In addition, the smaller value of rMSE indicates that PAPW is more efficient than IPSW. For the PM, both crCI and rSE reflect accurate estimates of the variance for all values of ρ. When the working model is incorrect, point estimates associated with both QR and PM are biased, but the variance remains unbiased. These findings hold across all three values of ρ.

For the DR methods, it is evident that estimates based on both PAPW and IPSW remain unbiased when at least one of the PM or QR models holds. Also, the values of crCI and rSE reveal that the asymptotic variance estimator is DR for both methods. Comparing the rMSE values, the AIPW estimate based on IPSW is slightly more efficient than the one based on PAPW. While the variance estimates remain unbiased under the false-false model specification status, point estimates are severely biased. Finally, the performance of both AIPW estimators improves with respect to bias reduction especially when the QR model is misspecified.

For the Bayesian approach, the simulation results are displayed in Table 3.2. Note that I no longer am able to use the PMLE approach. Instead, I apply the PAPP method assuming that π_i^R is unknown for $i \in S_A$. As illustrated, PAPP outperforms with respect to bias among all the non-robust methods. Surprisingly, the magnitude of the bias is even smaller in the Bayesian PAPP than the QR methods examined under the frequentist framework. In addition, it seems estimates under the Bayesian approach are slightly more efficient than those obtained under the frequentist methods. While the variance is approximately unbiased for $\rho = 0.2$, there is evidence that PM and QR increasingly underestimate and overestimate the true variance, respectively, as the value of ρ increases. Regarding the DR methods, it is evident that AIPW

estimates are even less biased and more efficient in the Bayesian approach compared to the alternative frequentist method, especially when the PM is misspecified, but the QR model holds. It is clear from the table that DR property holds for all values of ρ when at least one of the working models is correctly specified.

3.3.2 Simulation II

In the previous simulation study, I violated the ignorable assumption in order to misspecify the working model by dropping a key auxiliary variable. Now, I focus on a situation where models are misspecified with respect to the functional form of their conditional means. To this end, I consider non-linear associations and two-way interactions in constructing the outcome variables as well as the selection probabilities. This also allows us to examine the flexibility of BART as a non-parametric method when the true functional form of the underlying models is unknown. In addition, to simulate a more realistic situation, this time, two separate sets of auxiliary variables are generated, D associated with the design of S_A, and X associated with the design of S_R. However, I allow the two variables to be correlated through a bivariate Gaussian distribution as below:

$$\begin{pmatrix} d \\ x \end{pmatrix} \sim MVN \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right) \tag{3.35}$$

Note that ρ controls how strongly the sampling design of S_R is associated with that of S_A. In addition, the values of d_i can be either observed or unobserved for $i \in S_A$. In this simulation, I set $\rho = 0.2$, but later I check other values ranging from 0 to 0.9 as well.

To generate the outcome variable in U, I consider the following non-linear model:

$$y_i = 2f_k(x_i) - d_i^2 + 0.5x_i d_i + \sigma\epsilon_i \tag{3.36}$$

where $\epsilon_i \sim N(0,1)$, and σ is determined such that the correlation between y_i and $2f_k(x_i) - d_i^2 + 0.5x_i d_i$ equals 0.5 for $i \in U$. The function $f_k(.)$ is assumed to take one of the following forms:

$$SIN : f_1(x) = sin(x) \qquad EXP : f_2(x) = exp(x/2) \qquad SQR : f_3(x) = x^2/3 \tag{3.37}$$

I then consider an informative sampling strategy with unequal probabilities of inclusion, where the selection mechanism of S_A and S_R depends on x and d, respectively. Thus, each $i \in U$ is assigned two values corresponding to the probabilities of selection in S_R and S_A through a *logistic* function as below:

$$\begin{aligned} \pi^R(x_i) = P(\delta_i^R = 1|d_i) &= \frac{exp\{\gamma_0 + 0.2d_i^2\}}{1 + exp\{\gamma_0 + 0.2d_i^2\}} \\ \pi_k^A(x_i) = P_k(\delta_i^A = 1|x_i) &= \frac{exp\{\gamma_1 + f_k(x_i)\}}{1 + exp\{\gamma_1 + f_k(x_i)\}} \end{aligned} \tag{3.38}$$

where δ_i^R and δ_i^A are the indicators of being selected in S_R and S_A, respectively.

Associated with S_R and S_A, independent samples of size $n_R = 100$ and $n_A = 1,000$ were selected randomly from U with Poisson sampling at the first stage and simple random sampling at the second stage. The sample size per cluster, n_α, was 1 and 50 for S_R and S_A, respectively. The model intercepts, γ_0 and γ_1 in Eq. 3.38, are obtained such that $\sum_{i=1}^N \pi_i^R = n_R$ and $\sum_{i=1}^N \pi_i^A = n_A$. I restrict this simulation to Bayesian analysis based on the proposed PAPW and PAPP methods but focus on how well the non-parametric Bayes performs over the parametric Bayes in situations when the true structure of both underlying models are supposed to be unknown. The rest of the simulation design is similar to that defined in Simulation I, except for the way I specify a working model. This is done by including only the main and linear effects of X and D in the PM model, and the main and linear effect of X in the QR model. BART's performance is examined under the assumption that the true functional form

of both QR and PM models is unknown, and thus, only main effects are included in BART. Note that dropping a key auxiliary variable, which was the case in Simulation I, leads to a violation from the ignorable assumption, which may not be compensated by the use of more flexible approaches, such as BART.

The findings of this simulation for the two-step Bayesian approach with $1,000$ MCMC draws and $M = 200$, are exhibited numerically in Table 3.3. Regarding the non-robust methods, both QR and PM estimators show unbiased results across the three defined functions, i.e. SIN, EXP, and SQR, as long the working GLM is valid, with the minimum value of rBias associated with the PAPP method. According to the rSE values, there is evidence that PAPW and PAPP overestimate the variance, and PM underestimates the variance to some degrees, especially under the EXP and SQR scenarios. When the specified GLM is wrong, as seen, point estimates are biased for both QR and PM methods across all three functions. However, BART produces approximately unbiased results with smaller values of rMSE than GLM. In general, the PM method outperforms the QR methods Under BART with respect to bias, but results based on the PAPP method are more efficient. In addition, BART tends to overestimate the variance under both QR and PM methods.

When it comes to the DR adjustment, Bayesian GLM produces unbiased results across all the three defined functions if the working model of either QR or PM holds. However, the variance is slightly underestimated for the SIN function when the PM specified model is wrong, and it is overestimated for the EXP function under all model-specification scenarios. As expected, point estimates are biased when the GLM is misspecified for both QR and PM. However, BART tends to produce unbiased estimates consistently across all three functions, and the magnitude of both rBias and rMSE are smaller in the AIPW estimator based on PAPP compared to the AIPW estimator based on PAPW. Finally, as in the non-robust method, variance under BART is overestimated compared to the GLM. Extensions of the second simulation

to situations with $n_R = 100$ and $n_A = 100$, and $n_R = 100$ and $n_A = 10,000$ can be found in Appendix 3.6.2.

Table 3.3: Comparing the performance of the bias adjustment methods and associated variance estimator under the two-step parametric Bayesian approach in the second simulation study for $\rho = 0.2$

	SIN				EXP				SQR			
Model-method	rBias	rMSE	crCI	rSE	rBias	rMSE	crCI	rSE	rBias	rMSE	crCI	rSE
Probability sample (S_R)												
Unweighted	-17.210	23.109	80.000	0.999	-8.406	11.126	78.300	1.000	-17.302	20.563	65.800	1.002
Fully weighted	-0.623	17.027	94.440	0.987	-0.303	7.947	94.580	0.987	-0.675	13.219	94.000	0.975
Non-probability sample (S_A)												
Unweighted	33.063	33.379	0.000	1.003	40.307	40.409	0.000	1.079	49.356	49.570	0.000	1.016
Fully weighted	0.019	6.010	95.120	1.006	0.005	2.755	94.880	1.005	0.009	3.948	94.980	0.992
Non-robust adjustment												
Model specification: True												
GLM–PAPW	-0.425	9.257	96.300	1.072	-0.185	4.262	98.700	1.257	-0.325	6.649	98.360	1.213
GLM–PAPP	0.018	8.460	95.680	1.018	0.040	3.870	98.560	1.238	-0.037	5.914	98.760	1.222
GLM–PM	-0.411	9.899	94.680	0.982	-0.371	4.504	94.440	0.988	-0.762	8.115	92.520	0.947
Model specification: False												
GLM–PAPW	7.180	11.635	86.360	1.027	2.511	5.299	97.220	1.316	52.170	52.559	0.000	1.102
GLM–PAPP	7.647	11.265	78.000	0.954	3.025	5.425	96.180	1.277	53.095	53.397	0.000	1.122
BART–PAPW	4.035	10.078	96.980	1.217	2.811	5.129	98.440	1.472	8.356	11.082	97.180	1.468
BART–PAPP	1.098	8.530	96.660	1.121	1.108	4.120	98.880	1.391	4.482	7.479	98.020	1.401
GLM–PM	5.870	10.542	87.920	0.972	-6.589	9.264	82.520	0.976	48.993	49.409	0.000	0.994
BART–PM	0.577	9.635	96.960	1.115	0.087	4.501	97.540	1.155	0.249	8.276	96.080	1.062
Doubly robust adjustment												
Model specification: QR–True, PM–True												
GLM–AIPW–PAPW	-0.450	9.930	95.760	1.023	-0.165	4.593	98.180	1.200	-0.458	8.116	96.520	1.089
GLM–AIPW–PAPP	-0.452	9.925	95.780	1.020	-0.162	4.592	98.140	1.193	-0.453	8.106	96.500	1.086
Model specification: QR–True, PM–False												
GLM–AIPW–PAPW	-0.279	9.996	93.160	0.926	0.310	5.697	98.780	1.303	-0.338	7.128	97.480	1.154
GLM–AIPW–PAPP	-0.134	9.418	94.120	0.961	0.508	4.977	99.480	1.475	-0.275	7.376	97.580	1.152
Model specification: QR–False, PM–True												
GLM–AIPW–PAPW	-0.411	10.098	96.080	1.024	-0.176	4.715	98.460	1.234	-0.771	8.122	95.480	1.057
GLM–AIPW–PAPP	-0.417	10.101	96.020	1.021	-0.173	4.705	98.400	1.229	-0.778	8.119	95.420	1.057
Model specification: QR–False, PM–False												
GLM–AIPW–PAPW	9.015	13.176	84.140	1.000	6.735	8.693	94.100	1.456	50.835	51.288	0.000	1.019
GLM–AIPW–PAPP	9.191	12.717	84.860	1.082	6.787	8.181	96.660	1.761	51.667	52.131	0.000	1.047
BART–AIPW–PAPW	0.425	10.071	97.900	1.184	0.122	4.689	99.280	1.407	-0.259	8.349	97.960	1.231
BART–AIPW–PAPP	-0.144	9.794	97.820	1.184	-0.100	4.541	99.280	1.405	-0.245	8.329	97.740	1.203

PAPW: propensity-adjusted probability weighting; PAPP: propensity-adjusted probability prediction; QR: quasi-randomization; PM: prediction model; AIPW: augmented inverse propensity weighting.

3.3.3 Simulation III

Since the non-probability sample in the application of this study is clustered, I performed a third simulation study. To this end, the hypothetical population is assumed to be clustered with $A = 10^3$ clusters, each of size $n_\alpha = 10^3$ ($N = 10^6$). Then, three cluster-level covariates, $\{x_1, x_2, d\}$, are defined with the following distributions:

$$\begin{pmatrix} d_\alpha \\ x_{0\alpha} \\ x_{1\alpha} \end{pmatrix} \sim MVN\left(\begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 & -\rho/2 & \rho \\ -\rho/2 & 1 & -\rho/2 \\ \rho & -\rho/2 & 1 \end{pmatrix}\right) \qquad x_{2\alpha} = I(x_{0\alpha} > 0) \quad (3.39)$$

where d denotes a design variable in S_R, and $\{x_1, x_2\}$ describes the selection mechanism in S_A. Primarily, I set $\rho = 0.8$, but later I check other values ranging from 0 to 0.9 as well. Note that ρ controls how strongly the sampling design of S_R is associated with that of S_A. Furthermore, I assume that both d and x are observed for units of S.

Again, to be able to assess BART's performance, I consider non-linear associations with polynomial terms and two-way interactions in construction of the outcome variables as well as the selection probabilities. Two outcome variables are studied, one continuous (y_c) and one binary (y_b), which both depend on $\{x, d\}$ as below:

$$y^c_{\alpha i}|x_\alpha, d_\alpha \sim N(\mu = 1 + 0.5x^2_{1\alpha} + 0.4x^3_{1\alpha} - 0.3x_{2\alpha} - 0.2x_{1\alpha}x_{2\alpha} - 0.1d_\alpha + u_\alpha, \sigma^2 = 1) \tag{3.40}$$

$$y^A_{\alpha i}|x_\alpha, d_\alpha \sim Ber\left(p = \frac{exp\{-1 + 0.1x^2_{1\alpha} + 0.2x^3_{1\alpha} - 0.3x_{2\alpha} - 0.4x_{1\alpha}x_2 - 0.5d_\alpha + u_\alpha\}}{1 + exp\{-1 + 0.1x^2_{1\alpha} + 0.2x^3_{1\alpha} - 0.3x_{2\alpha} - 0.4x_{1\alpha}x_2 - 0.5d_\alpha + u_\alpha\}}\right) \tag{3.41}$$

where $u_\alpha \sim N(0, \sigma^2_u)$, and σ^2_u is determined such that the intraclass correlation equals 0.2 (Oman and Zucker, 2001; Hunsberger et al., 2008). For each $i \in U$, I then consider the following set of selection probabilities associated with the design of the S_R and S_A:

$$\begin{aligned}
\pi^R(x_\alpha) &= P(\delta^R_\alpha = 1|d_\alpha) = \frac{exp\{\gamma_0 + 0.5d_\alpha\}}{1 + exp\{\gamma_0 + 0.5d_\alpha\}} \\
\pi^A(x_\alpha) &= P(\delta^A_\alpha = 1|x_\alpha) = \frac{exp\{\gamma_1 - 0.1x_{1\alpha} + 0.2x^2_{1\alpha} + 0.3x_{2\alpha} - 0.4x_{1\alpha}x_{2\alpha}\}}{1 + exp\{\gamma_1 - 0.1x_{1\alpha} + 0.2x^2_{1\alpha} + 0.3x_{2\alpha} - 0.4x_{1\alpha}x_{2\alpha}\}}
\end{aligned} \tag{3.42}$$

where δ^R_i and δ^A_i are the indicators of being selected in S_R and S_A, respectively. Associated with S_R and S_A, two-stage cluster samples of size $n_R = 100$ and $n_A = 10,000$ were selected randomly from U with Poisson sampling at the first stage and simple random sampling at the second stage. The sample size per cluster, n_α, was 1 and 50 for S_R and S_A, respectively. The model intercepts, γ_0 and γ_1 in 3.42, are

obtained such that $\sum_{i=1}^{N} \pi_i^R = n_R$ and $\sum_{i=1}^{N} \pi_i^R = n_A/n_\alpha$.

The rest of the simulation design is similar to that defined in Simulation II, except for the methods I use for point and variance estimation. In addition to the situation where π_i^R is known for $i \in S_A$, I consider a situation where π^R is unobserved for $i \in S_A$ and draw the estimates based on PAPP. Furthermore, unlike Simulation I, DR estimates are achieved by separately fitting the QR and PM models, and to get the variance estimates, a bootstrap technique is applied with $B = 200$ based on Rao and Wu (1988). Finally, under BART, Rubin's combining rules are employed to derive the point and variance estimates based on the random draws of the posterior predictive distribution. As in Simulations II, I consider different scenarios of model specification. To misspecify a model, I only include the main effects in the working model. Also, under BART, no interaction or polynomial is included as input. Again, I misspecify the functional form of the working models, while in Simulation I, I assumed that auxiliary variables are partially observed when misspecifying a model.

The means of the synthesized U for the outcome variables were $\bar{y}_U^c = 3.39$ and $\bar{y}_U^b = 0.40$. Figure 3.1 compares the bias magnitude and efficiency across the non-robust methods. As illustrated, point estimates from both S_R and S_A are biased if the true sampling weights are ignored. After adjusting, for both continuous and binary outcomes, the bias is close to *zero* under both QR and PM methods when the working model is correct. However, the lengths of the error bars reveal that the proposed PAPW/PAPP method is more efficient than the IPSW. When only main effects are included in the model, all adjusted estimates are biased except for those based on BART. Note that BART cannot be applied under IPSW. Further details about the simulation results for the non-robust methods are displayed in Appendix 3.6.2. I see that IPSW tends to have slightly larger magnitudes of rBias and rMSE for both y^c and y^b. Also, the values of rSE close to 1 indicate that Rao & Wu's bootstrap method of variance estimation performs well under both QR and PM approaches. However,

95% coverage is achieved only when the working model is correct.

In Figure 3.2, I depict the findings of the simulation corresponding to the DR estimators under different permutations of model specification. One can immediately infer that for all the employed methods, AIPW produces unbiased results when either the PM or QR model holds. However, in situations where the true underlying models for both PM and QR are unknown, the point estimates based on BART remain unbiased under both PAPW and PAPP approaches. Furthermore, under the GLM, it is evident that AIPW estimates based on PAPW/PAPP are slightly less biased and more efficient than those based on IPSW when the PM is incorrect (TF) according to the lengths of the error bars. Details of the numerical results can be found in Appendix 3.6.2. The latter compares BART with GLM under a situation where both working models are wrong. Results showing the performance of the bootstrap variance estimator are provided in Figure 3.3. The crCI values are all close to the correct value unless both working models are incorrectly specified. While the same result I observed for the continuous variable under BART, there is evidence that BART widely underestimates the variance of the AIPW estimator for the binary outcome. Note that the estimation of variance under BART is based on the MCMC draws of the posterior prediction distribution using Rubin's combining rule. To conclude, I observe that when neither the PM nor QR model is known, BART based on PAPP produces unbiased and efficient estimates with accurate variance.

As the final step, I replicate the simulation for different values of ρ ranging from 0 to 0.9 to show how stable the competing methods perform in terms of rbias and rMSE. Figure 3.4 depicts changes in the values of rBias and rMSE for different adjustment methods as the value of ρ increases. Generally, it seems that the value of rMSE decreases for all competing methods as ρ increases, but for all values of ρ, PAPW and PAPP are less biased than IPSW. It is only when $\rho = 0$ for the continuous variable that IPSW outperforms the PPAW/PAPP in bias reduction. However, when

d is highly correlated with x, there is also evidence of better performance by PAPP than IPSW in terms of bias reduction. I believe this is mainly because the stronger association between x and d implies that the additional ignorable assumption under PAPP is better met, while this correlation causes a sort of collinearity in IPSW leading to a loss of efficiency. The rest of the methods did not show significant changes as the value of ρ increases. Numerical values associated with Figure 3.4 have been provided in Appendix 3.6.2.

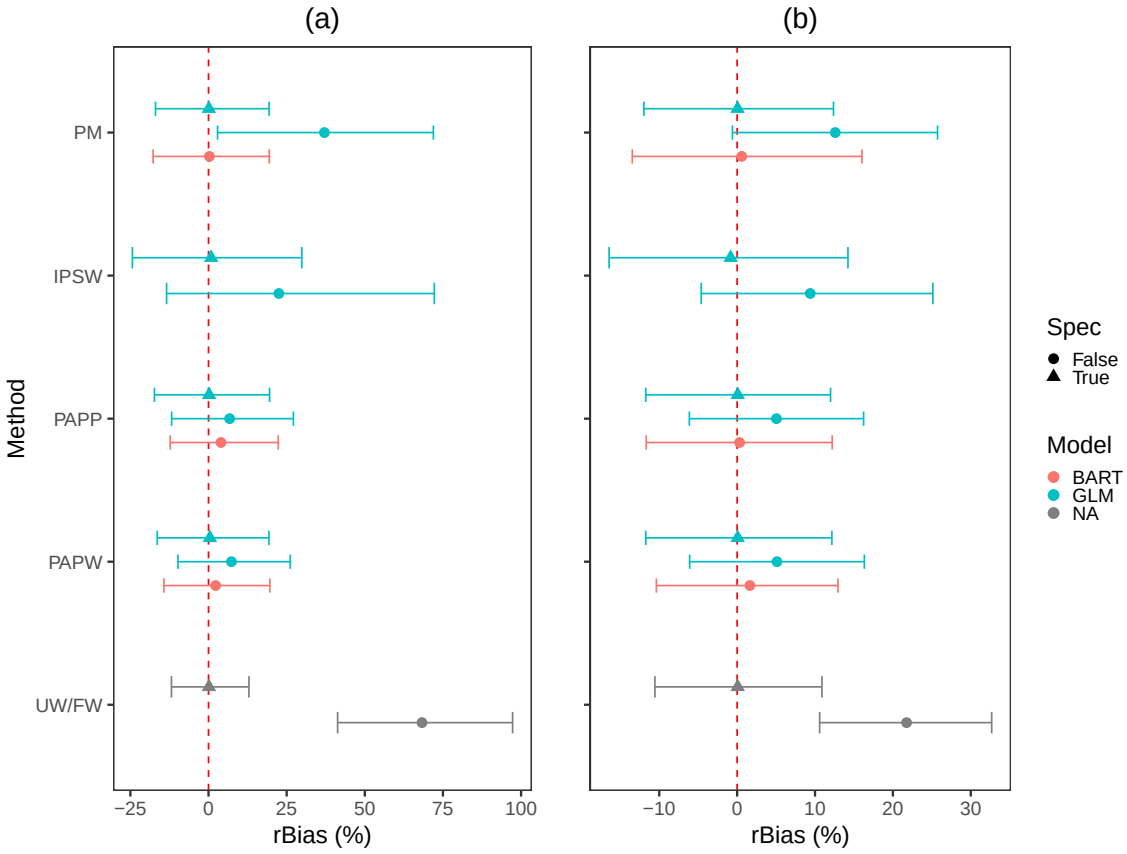

Figure 3.1: Comparing the performance of the non-robust approaches for (a) the *continuous* outcome (Y_c) and (b) the *binary* outcome (Y_b) when the model is correctly specified. Error bars represent the 2.5% and 97.5% percentiles of the empirical distribution of bias over the simulation iterations. UW: unweighted; FW: fully weighted; PM: prediction model; PAPP: propensity-adjusted probability prediction; IPSW: inverse propensity score weighting

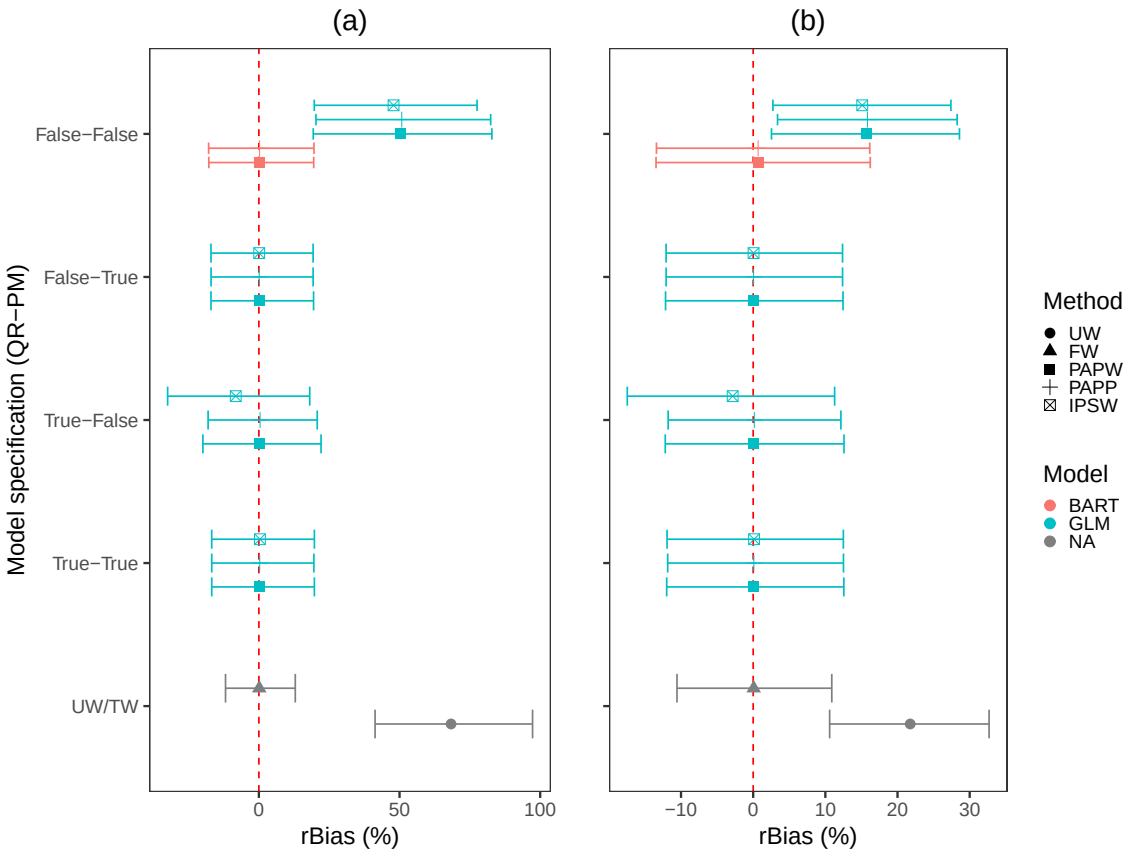

Figure 3.2: Comparing the performance of the doubly robust estimators under different model-specification scenarios for (a) the *continuous* outcome (Y_c) and (b) the *binary* outcome (Y_b). 95% CIs have been generated based on the 2.5% and 97.5% percentiles of the empirical distribution of bias over the simulation iterations. UW: unweighted; FW: fully weighted; PAPP: propensity-adjusted probability prediction; IPSW: inverse propensity score weighting

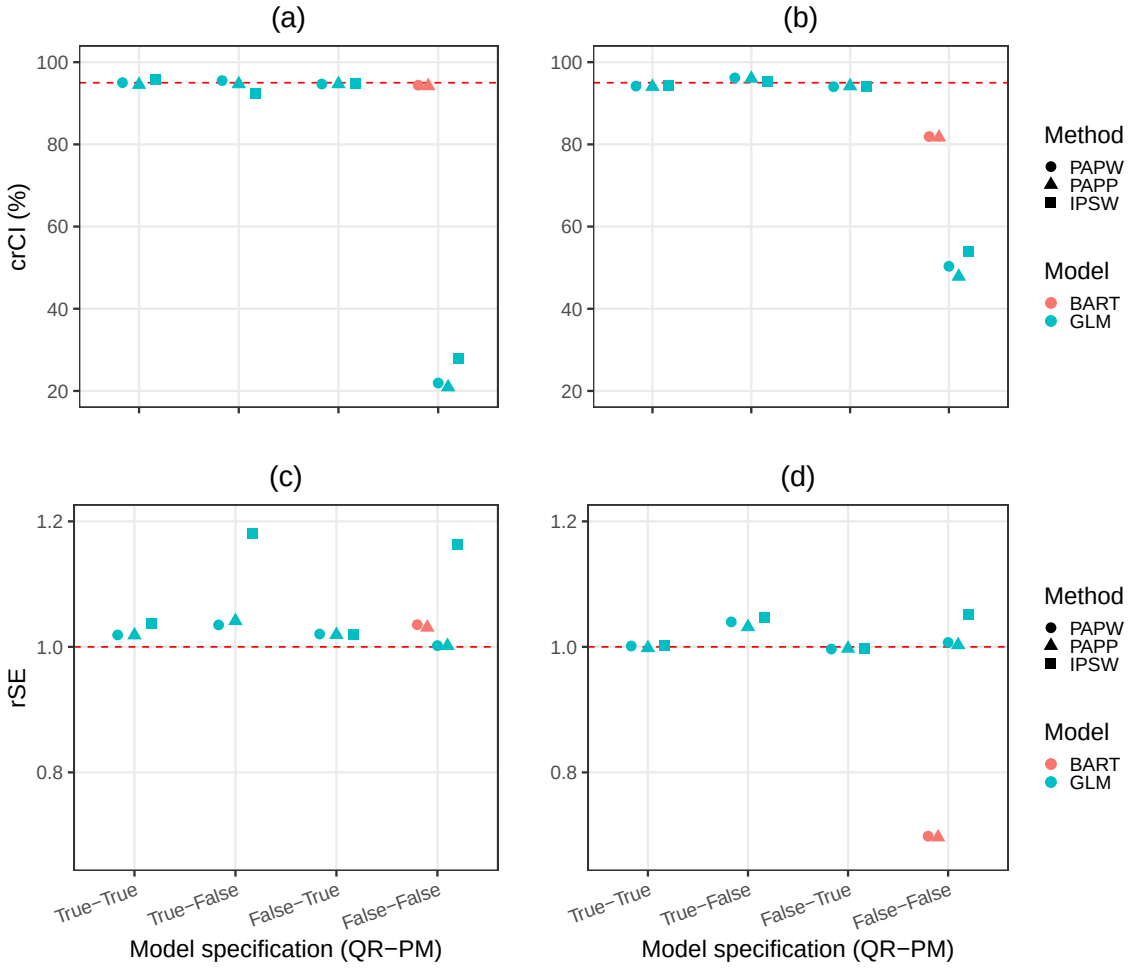

Figure 3.3: Comparing the 95% CI coverage rates for the means of (a) *continuous* outcome and (b) *binary* outcome and SE ratios for (c) *continuous* outcome and (d) *binary* outcome across different DR methods under different model specification scenarios. UW: unweighted; FW: fully weighted; PAPP: propensity-adjusted probability prediction; IPSW: inverse propensity score weighting

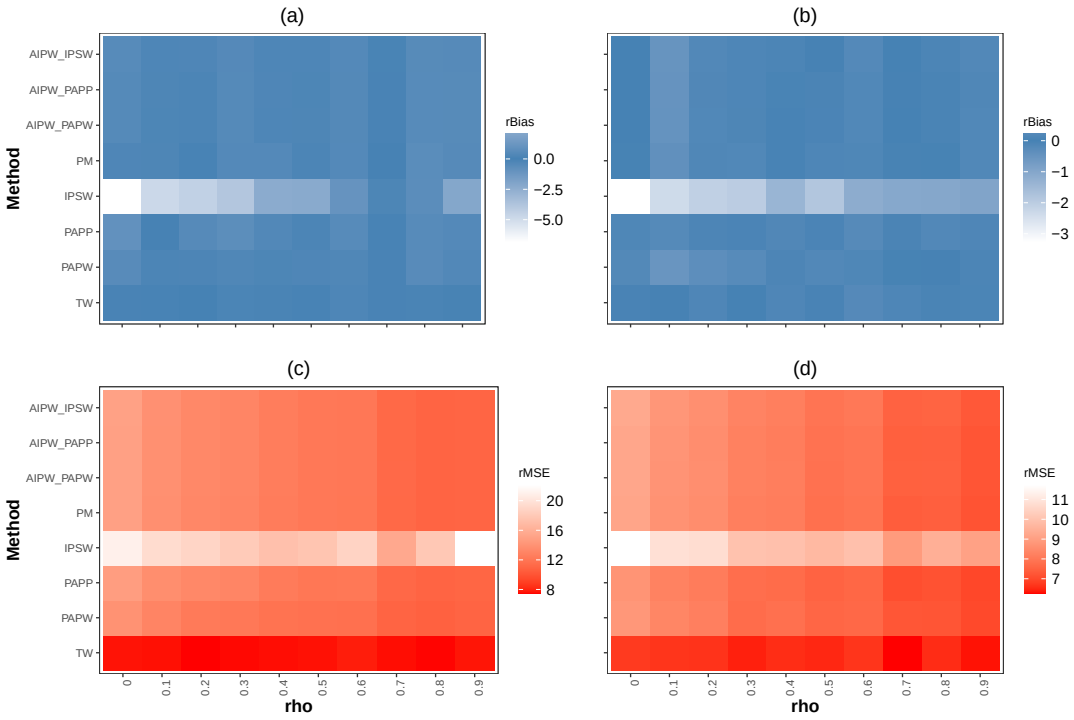

Figure 3.4: Comparing the rBias for the means of (a) *continuous* outcome and (b) *binary* outcome and rMSE for the means of (c) *continuous* outcome and (d) *binary* outcome across different adjustment methods and different values of ρ. UW: unweighted; FW: fully weighted; PAPP: propensity-adjusted probability prediction; IPSW: inverse propensity score weighting

3.4 Application

I briefly describe SHRP2, the non-probability sample, and the NHTS, the probability sample as well as the variables used for statistical adjustment.

3.4.1 Strategic Highway Research Program 2

SHRP2 is the largest naturalistic driving study conducted to date, with the primary aim to assess how people interact with their vehicle and traffic conditions while driving (SHRP2, 2013). About $A = 3,140$ drivers aged $16 - 95$ years were recruited from six geographically dispersed sites across the United States (Florida, Indiana, New York, North Carolina, Pennsylvania, and Washington), and over five million trips and 50 million driven miles have been recorded. The average follow-up time per person was $\bar{n}_\alpha = 440$ days. A quasi-random approach was initially employed to select samples by random cold calling from a pool of $17,000$ pre-registered volunteers. However, because of the low success rate along with budgetary constraints, the investigators later chose to pursue voluntary recruitment. Sites were assigned one of three pre-determined sample sizes according to their population density (Campbell, 2012). The youngest and oldest age groups were oversampled because of the higher crash risk among those subgroups. Thus, one can conclude that the selection mechanism in SHRP2 is a combination of convenience and quota sampling methods. Further description of the study design and recruitment process can be found in Antin et al. (2015).

SHRP2 data are collected in multiple stages. Selected participants are initially asked to complete multiple assessment tests, including executive function and cognition, visual perception, visual-cognitive, physical and psychomotor capabilities, personality factors, sleep-related factors, general medical condition, driving knowledge, etc. In addition, demographic information such as age, gender, household income, education level, and marital status as well as vehicle characteristics such as vehicle type, model year, manufacturer, and annual mileage are gathered at the screening stage.

A trip in SHRP2 is defined as the time interval during which the vehicle is operating. The in-vehicle sensors start recording kinematic information, the driver's behaviors, and traffic events continuously as soon as the vehicle is switched on. Encrypted data are stored in a removable hard drive, and participants are asked to provide access to the vehicle every four to six months, so that hard drives with accumulated data are removed and replaced. Then, Trip-related information such as average speed, duration, distance, and GPS trajectory coordinates are obtained by aggregating the sensor records at the trip level (Antin et al., 2019; Campbell, 2012).

3.4.2 National Household Travel Survey data

In the present study, I use data from the eighth round of the NHTS conducted from March 2016 through May 2017 as the reference survey. The NHTS is a nationally representative survey, repeated cross-sectionally approximately every seven years. It is aimed at characterizing personal travel behaviors among the civilian, non-institutionalized population of the United States. The 2017 NHTS was a mixed-mode survey, in which households were initially recruited by mailing through an address-based sampling (ABS) technique. Within the selected households, all eligible individuals aged ≥ 5 years were requested to report the trips they made on a randomly assigned weekday through a web-based travel log. Proxy interviews were requested for younger household members who were ≤ 15 years old.

The overall sample size was $129,696$, of which roughly 20% was used for national representativity and the remaining 80% was regarded as add-ons for the state-level analysis. The recruitment response rate was 30.4%, of which 51.4% reported their trips via the travel logs (Santos et al., 2011). In NHTS, a travel day is defined from 4:00 AM of the assigned day to 3:59 AM of the following day on a typical weekday. A trip is defined as that made by one person using any mode of transportation. While trip distance was measured by online geocoding, the rest of the trip-related

information was based on self-reporting. A total of 264,234 eligible individuals aged $\geq$5 took part in the study, for which $923,572$ trips were recorded (McGuckin and Fucci, 2018).

3.4.3 Auxiliary variables and analysis plan

Because of the critical role of auxiliary variables in maintaining the ignorable assumption for the selection mechanism of the SHRP2 sample, particular attention was paid to identify and build as many common variables as possible in the combined sample that are expected to govern both selection mechanism and outcome variables in SHRP2. However, since the SHRP2 sample is gathered from a limited geographical area, in order to be able to generalize the findings to the American population of drivers, I had to assume that no other auxiliary variable apart from those investigated in this study will define the distribution of the outcome variables. This assumption is in fact embedded in the ignorable condition in the SHRP2 given the common set of observed covariates. Three distinct sets of variables were considered: (i) demographic information of the drivers, (ii) vehicle characteristics, and (iii) day-level information. These variables and associated levels/ranges are listed in Table 3.4.

My focus was on inference at the day level, so SHRP2 data were aggregated. I constructed several trip-related outcome variables such as daily frequency of trips, daily total trip duration, daily total distance driven, mean daily trip average speed, and mean daily start time of trips that were available in both datasets as well as daily maximum speed, daily frequency of brakes per mile, and daily percentage of trips with a full stop, which was available in SHRP2 only. The final sample sizes of the complete day-level datasets were $n_A = 837,061$ and $n_R = 133,582$ in SHRP2 and NHTS, respectively.

In order to make the two datasets more comparable, I filtered out all the subjects in NHTS who were not drivers or were younger than 16 years old or used public

transportation or transportation modes other than cars, SUVs, vans, or light pickup trucks. One major structural difference between NHTS and SHRP2 was that in the NHTS, participants' trips were recorded for only one randomly assigned weekday, while in SHRP2, individuals were followed up for several months or years. Therefore, to properly account for the potential intraclass correlation across sample units in SHRP2, I treated SHRP2 participants as clusters for variance estimation. For BART, I fitted random intercept BART (Tan et al., 2016). In addition, since the π_i^R were not observed for units of SHRP2, I employed the PAPP and IPSW methods to estimate pseudo-weights, so variance estimation under the GLM was based on the Rao & Wu bootstrap method throughout the application section.

Table 3.4: List of auxiliary variables and associated levels/ranges that are used to adjust for selection bias in SHRP2

Auxiliary variables (scale)	**Levels/range**
Demographic information	
gender	(female, male)
age (yrs)	(16-24, 25-34, 35-44, 45-54, 55-64, 65-74, 75+)
race	(White, Black, other)
ethnicity	(Hispanic, non-Hispanic)
birth country	(citizen, alien)
education level	(≤HS, HS completed, associate, grad, post-grad)
household income (×$1,000)	(0-49k, 50-99k, 100-149k, 150k+)
household size	(1, 2, 3-5, 6-10, 10+)
job status	(part-time, full time)
home ownership	(owner, renter)
pop. size of resid. area ($\times 1,000$)	(0-49, 50-200, 200-500, 500+)
Vehicle characteristics	
age (yrs)	(0-4, 5-9, 10-14, 15-19, 20+)
type	(passenger car, Van, SUV, truck)
make	(American, European, Asian)
mileage (×1,000km)	(0-4, 5-9, 10, 10-19, 20-49, 50+)
fuel type	(gas, other)
Day-level information	
weekend indicator of trip day	{0,1}
season of trip day	(winter, spring, summer, fall)

3.4.4 Results

According to Figure 3.11 of Appendix 3.6.3, one can visually infer that the largest discrepancies between the sample distribution of auxiliary variables in SHRP2 and that in the population stem from participants' age, race, and population size of residential areas as well as vehicles' age and vehicles' type. The youngest and eldest age groups have been oversampled as are Whites and non-Hispanics. In addition, I found that the proportion of urban dwellers is higher in SHRP2 than that in the NHTS. In terms of vehicle characteristics, SHRP2 participants tend to own passenger cars more than the population average, whereas individuals with other vehicle types were underrepresented in SHRP2.

As the first step of QR, I checked if there is any evidence of a lack of common distributional support between the two studies for the auxiliary variables. Figure 3.5a compares the kernel density of the estimated PS using BART across the two samples. As illustrated, a notable lack appears on the left tail of the PS distribution in SHRP2. However, owing to the huge sample size in SHRP2, I believe this does not jeopardize the positivity assumption seriously. The estimated population size of drivers was $\hat{N} = 133,047,744$ based on the sampling weight in NHTS. The available auxiliary variables are strong predictors of the NHTS selection probabilities for SRHP2: the average pseudo-R^2 was for BART 73% in a 10-fold cross validation.

In Figure 3.5b, I compare the distribution of estimated pseudo-weights across the QR methods. It seems that PAPP based on BART is the only method that does not produce influential weights. Also, the highest variability in the estimated pseudo-weights belonged to the PAPP method under GLM. Figure 3.6 compares the predictive power of BART with GLM and also classification and regression trees (BART) in modeling Z and Y on X. As can be seen, the largest values of area under the ROC curve (AUC) and the largest values of (pseudo)-R^2 in the radar across different trip-related outcome variables are associated with BART. Additionally, Figure 3.12 in

Appendix 3.6.3 exhibits how pseudo-weighting based on PAPP-BART improves the imbalance in the distribution of X in SHRP2 with respect to the weighted distribution of NHTS.

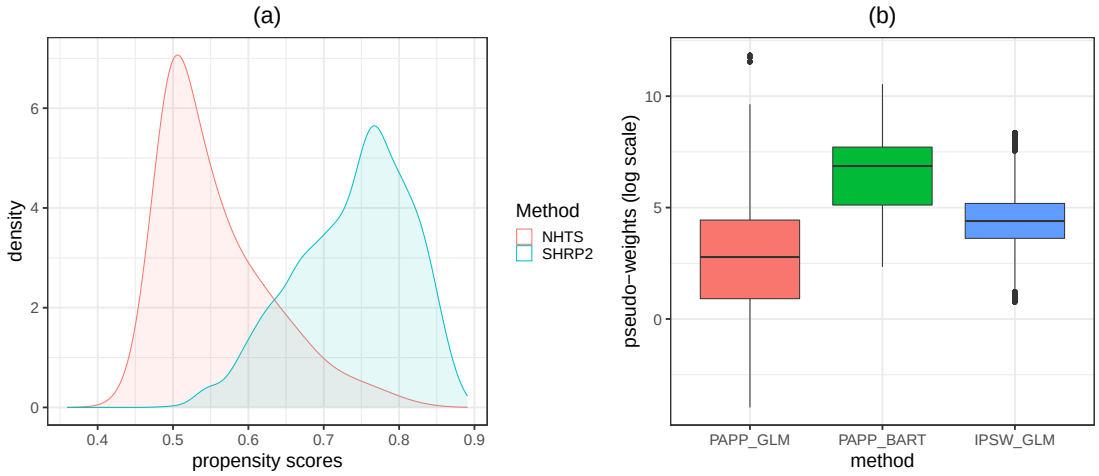

Figure 3.5: Comparing the distribution of (a) estimated propensity scores between SHRP2 and NHTS using BART and (b) estimated pseudo-weights in SHRP2 across the applied quasi-randomization methods

Figure 3.8 depicts the adjusted sample means for some trip-related measures that were available in both SHRP2 and NHTS. The methods I compare here encompass PAPP, IPSW, and PM as the non-robust approaches, and AIPW with PAPP and AIPW with IPSW as the DR approaches. Also, a comparison is made between GLM and BART for all the methods except those involving IPSW. My results suggest that, as expected, the oversampling of younger and older drivers leads to underestimating miles driven and length of trips, and overestimating the time of the first trip of the day; other factors may impact these variables, as well as the average speed of a given drive. For three of these four variables (total trip duration, total distance driven, and start hour of daily trip), there appeared to be improvements with respect to the bias considering the NHTS weighted estimates as the benchmark, although only trip duration appears to be fully corrected. In Figure 3.7, I display the posterior predictive density of mean daily total distance driven under PAPP, PM, and AIPW-PAPP. Note

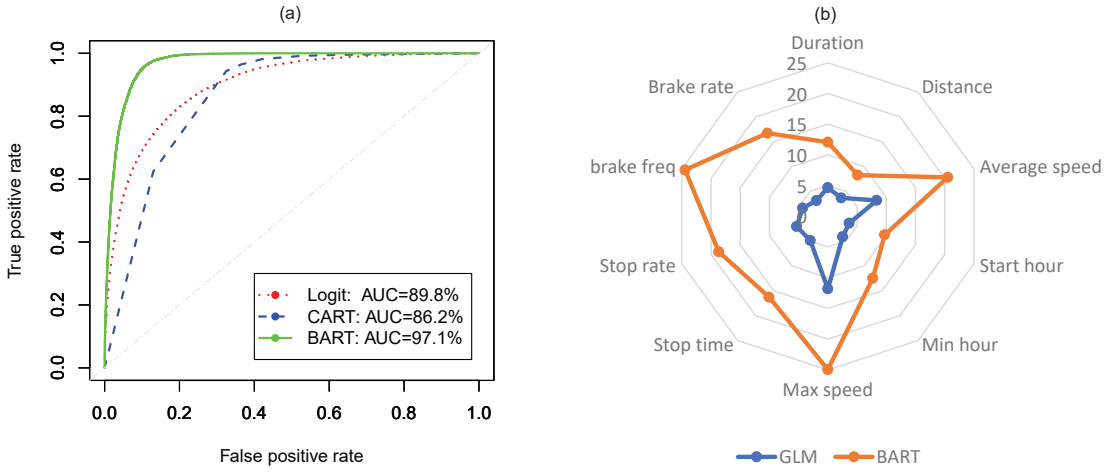

Figure 3.6: Comparing the performance of BART vs GLM in both estimating propensity scores and predicting some trip-related outcomes. The radar plot on the right side displays the values of (pseudo-)R^2 between BART and GLM. AUC: area under curve; CART: classification and regression trees

that the narrow variance associated with the PAPP approach is due to the fact that the posterior predictive distribution under pseudo-weighting does not account for the clustering effects in SHRP2. It is in fact $\bar{V}_W$ in Eq. 3.25 that is capturing this source of uncertainty in the variance estimation.

Among the QR methods, I observed that the PAPP based on BART gives the most accurate estimate with respect to bias for this variable. However, the relatively narrow 95% CI associated with BART may indicate that BART does not properly propagate the uncertainty in pseudo-weighting. Regarding the PM, it seems BART performs as well as GLM, but with wider uncertainty. As a consequence, the AIPW estimator performs the same in terms of bias across different QR methods. The AIPW estimator based on IPSW, on the other hand, seems to be is more efficient than the ones based on PAPP. However, these findings are not consistent across the outcome variables. For the daily total duration variable, which is displayed in plot (b) of Figure 3.8, it is only the PAPP-based estimator whose 95% CI covers the population mean. For the daily average speed depicted in (c) and the daily mean start time of the trip depicted in (d), I observed no reliable correction for bias.

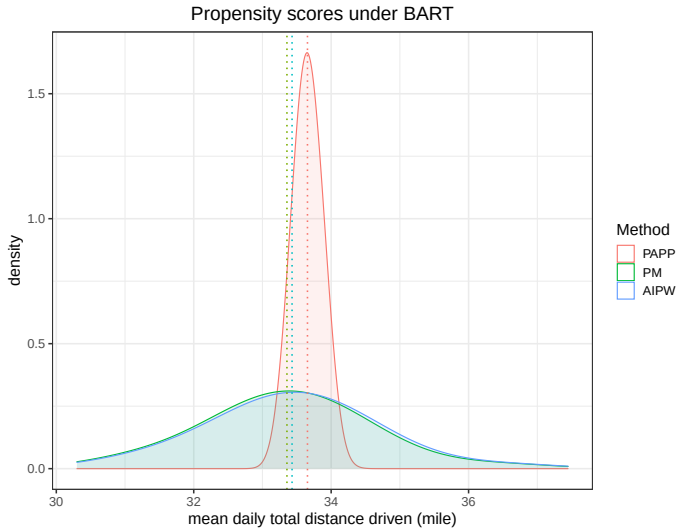

Figure 3.7: The posterior predictive distributions of the adjusted sample mean of daily total distance driven based on BART

Results related to the adjusted means for some SHRP2-specific outcome variables are summarized in Figure 3.9. These variables consist of (a) daily maximum speed, (b) frequency of brakes per mile, and (c) percentage of trip duration when the vehicle is fully stopped. For the daily maximum speed, I take one further step and present the DR adjusted mean based on the IPSW-GLM and PAPP-BART by some auxiliary variables in Figure 3.10. As illustrated, higher levels of mean daily maximum speed are associated with males, age group 35-44 years, Blacks, high school graduates, Asian cars, and weekends. According to the lengths of 95% CIs, one can see that the AIPW-PAPP-BART consistently produces more efficient estimates than the AIPW-IPSW-GLM. Further numerical details of these findings by the auxiliary variables have been provided in Tables 3.13-3.19 in Appendix 3.6.3.

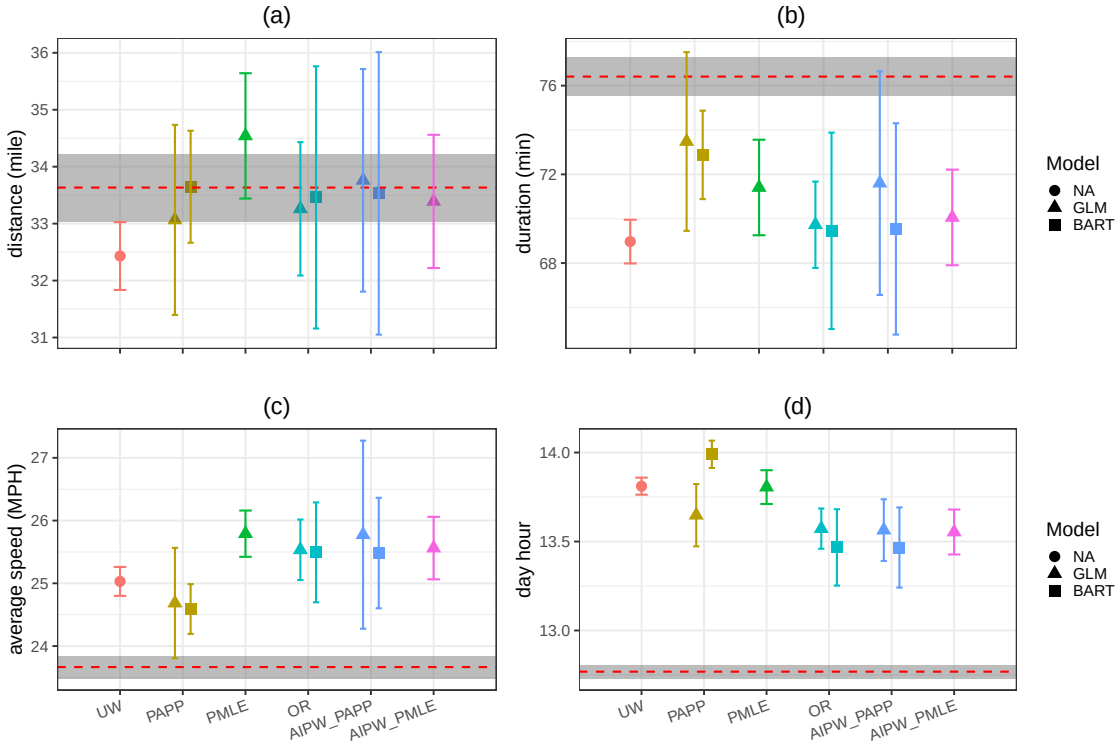

Figure 3.8: Evaluation of pseudo-weights by comparing weighted estimates of the daily frequency of trips between NHTS and SHRP2: (a) Mean daily total trip duration, (b) Mean daily total distance driven, (c) Mean trip average speed, and (d) Mean daily start hour of trips. The dashed line and surrounding shadowed area represent weighted estimates and 95% CIs in NHTS, respectively. UW: unweighted; PAPP: propensity-adjusted probability prediction; IPSW: inverse propensity score weighting; NA: not applicable

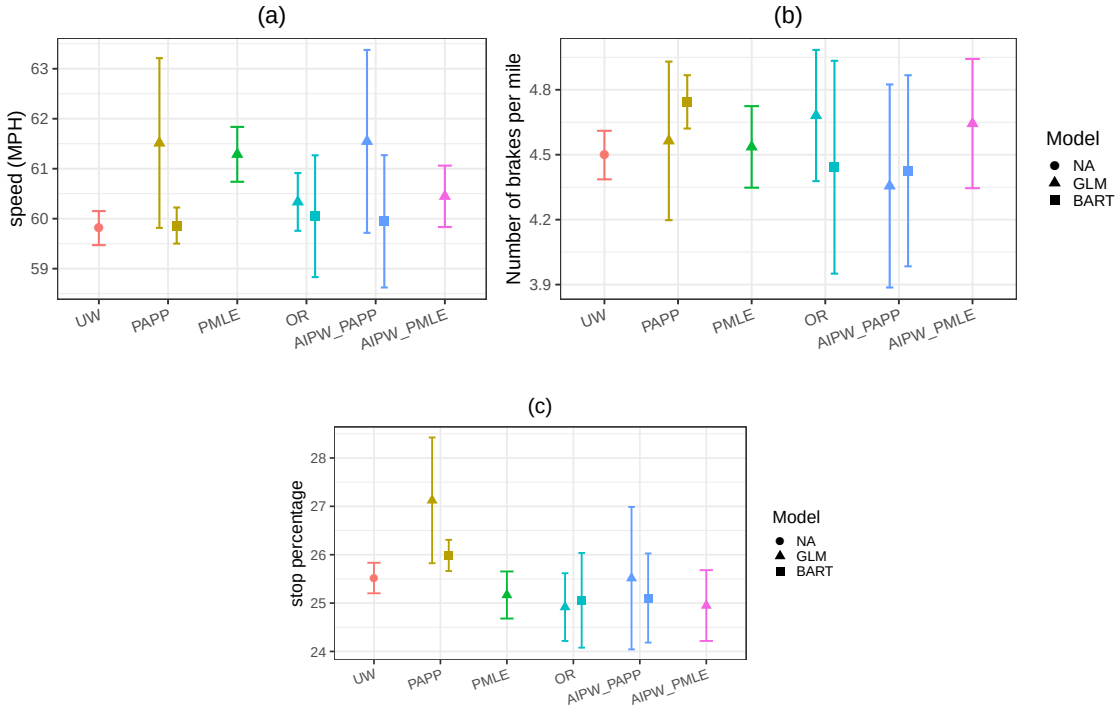

Figure 3.9: Adjusted estimates of some SHRP2-specific outcomes: (a) Mean daily maximum speed, (b) daily frequency of brakes per mile driven, and (c) daily percentage of stop time. UW: unweighted; PAPP: propensity-adjusted probability prediction; IPSW: inverse propensity score weighting

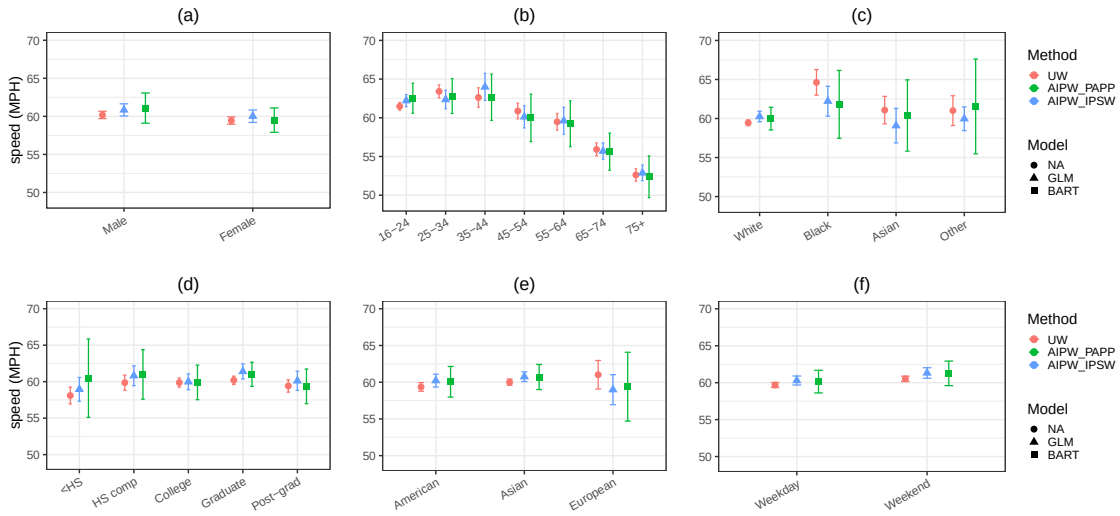

Figure 3.10: Bias-adjusted estimates of mean daily maximum speed (MPH) driven by (a) gender, (b) age groups, (c) race, (d) education, (e) vehicle manufacturer, and (f) weekend indicator. UW: unweighted; PAPP: propensity-adjusted probability prediction; IPSW: inverse propensity score weighting; NA: not applicable

3.5 Discussion

In this study, I proposed a doubly robust (DR) adjustment method for finite population inference in non-probability samples when a well-designed probability sample is available as a benchmark. Combining the ideas of pseudo-weighting with prediction modeling, my method involved a modified version of AIPW, which is DR in the sense that estimates are consistent if either underlying model holds. More importantly, the proposed method permitted us to apply a wider class of predictive tools, especially supervised algorithmic methods. To better address model misspecification, the present study employed BART to multiply impute both pseudo-inclusion probabilities and the outcome variable. I also proposed a method to estimate the variance of the DR estimator based on the posterior predictive draws simulated by BART. In a simulation study, I then assessed the repeated sampling properties of the proposed estimator. Finally, I apply it to real Big Data from naturalistic driving studies with the aim to improve the potential selection bias in the estimates of finite population mean.

Generally, the simulation findings revealed that the modified AIPW method produces less biased estimates than its competitors, especially when $n_R << n_A$. When at least one of the models, i.e. QR or PM, is correctly specified, all the DR methods generated unbiased results, though my estimator was substantially more efficient with narrower 95% CIs. However, when both working models are invalid, my findings suggest that DR estimates based on the GLM can be severely biased. However, under BART, it seems that estimates remain approximately unbiased if the true model structure associated with both QR and PM is unknown to the researcher. In contrast to the conventional IPSW estimator, I found that the new proposed estimator produces more stable results in terms of bias and efficiency across different sampling fractions and various degrees of association between the sampling designs of S_R and S_A.

Overall, the results of the application suggest near total removal of bias for only one of the four variables that can be estimated from the reference survey (daily total distance driven). I believe this failure originates from several sources. First and foremost, the bias observed in the final estimates is very likely to be mixed with measurement error because I compared the results of sensor data with self-reported data as a benchmark. Second, there was evidence of departure from the positivity assumption in SHRP2. Studies show that even a slight lack of common support in the distribution of auxiliary variables may lead to inflated variance and aggravated bias (Hill and Su, 2013). Part of this can be due to the fact that I attempted to generalize the results to the general population of American drivers, while SHRP2 data was restricted to six states. Another reason might be a deviation from the ignorable assumptions: The associations between the auxiliary variables and the outcome variables were relatively weak and varying across the variables.

This study was not without weaknesses. First, my approach assumes the ideal situation where the d_i are available in the non-probability sample since that is demanded by the general theory linking together the probability and non-probability samples. In practice, it can be difficult to fully meet this requirement, and indeed in many practical settings, it might be that only the available subset of x_i^* is required to fully model selection into the non-probability sample and the outcome variable, or alternatively, that the available components of x_i^* will provide a much better approximation to the true estimates than simply using the non-probability sample without correction. Second, my adjustment method assumes that the two samples are mutually exclusive. However, in many Big Data scenarios (though not the one I consider), the sampling fraction may be non-trivial, so the two samples may overlap substantially. In such a situation, it is important to check how sensitive my proposed pseudo-weighting approach is to this assumption. Extensions may be plausible to account for the duplicate units of the population in the pooled sample. Third, the multiple imputation variance

estimator (Eq. 3.25) ignores covariance between $\overline{V}_W$ and V_B induced by the weights (Kim et al., 2006). This covariance is typically negative and leads to conservative inference, as seen in the modest overestimation of variance in the BART estimations in Simulations 2 and 3. Use of a bootstrap procedure such as that described in the simulation study of Chen et al. (2019) may be an alternative, although impractical in my setting given the computational demands of fitting the BART models to each bootstrap sample.

Another drawback is that the combined dataset may be subject to differential measurement error in the variables. This issue is particularly acute in the SHPR2 analysis, because the definition of a *trip* may not be identical between the two studies: although trip measures in the SHRP2 are recorded by sensors, in the NHTS trip measures are memory and human estimation based, as they are self-reported. Having such error-prone information either as the outcome or as an auxiliary variable may lead to biased results. Finally, I failed to use the two-step Bayesian method under GLM for the application part, because SHRP2 data were clustered demanding for Bayesian generalized linear mixed effect models to properly estimate the variance of the DR estimators required computational resources beyond my reach. This prompted us to apply resampling techniques to the actual data instead of a fully Bayesian method.

There are a number of potential future directions for this research. First, I would like to expand the asymptotic variance estimator under PAPP when π_i^R cannot be computed for $i \in S_B$. Alternatively, one may be interested in developing a fully model-based approach, in which a synthetic population is created by undoing the sampling stages via a Bayesian bootstrap method, and attempts are made to impute the outcome for non-sampled units of the population (Dong et al., 2014; Zangeneh and Little, 2015; An and Little, 2008). The synthetic population idea makes it easier to incorporate the design features of the reference survey into adjustments, especially

when Bayesian inference is of interest. While correcting for selection bias, one can adjust for the potential measurement error in the outcome variables as well if there exists a validation dataset where both mismeasured and error-free values of the variables are observed (Kim et al., 2021b). When combining data from multiple sources, it is also likely that auxiliary variables are subject to differential measurement error. Hong et al. (2017) propose a Bayesian approach to adjust for a different type of measurement error in a causal inference context. Also, in a Big Data setting, fitting models can be computationally demanding. To address this issue, it might be worth expanding the divide-and-recombine techniques for the proposed DR methods. Finally, as noted by a reviewer, the basic structure of our problem (see Figure 1.1) approximates that tackled by "data fusion" methods, developed primarily in the computer science literature (Castanedo, 2013). While this literature does not appear to have directly addressed issues around sample design, it may be a useful vein of research to mine for future connections to non-probability sampling research.

3.6 Appendix

3.6.1 Theoretical proofs

3.6.1.1 Proof of doubly robustness

As discussed in Section 3.2.2, a doubly robust estimator should be consistent even if either model is misspecified. To prove the doubly robustness property of the AIPW estimator proposed here, let initially assume that $\hat{\theta} \xrightarrow{p} \theta$ if the prediction model (PM) is correctly specified, and $\hat{\phi} \xrightarrow{p} \phi$ and $\hat{\beta} \xrightarrow{p} \beta$ if the pseudo-weighting model is correctly specified. Given the true probabilities of selection in S_A, I know that the

HT-estimator is design-unbiased for the population total, i.e.

$$\begin{aligned} E\left(\sum_{i=1}^{n_A} y_i/\pi_i^A\right) &= E\left(\sum_{i=1}^{N} \delta_i^A y_i/\pi_i^A\right) \\ &= \sum_{i=1}^{N} E(\delta_i^A) y_i/\pi_i^A \\ &= \sum_{i=1}^{N} \pi_i^A y_i/\pi_i^A \\ &= \sum_{i=1}^{N} y_i \\ &= \hat{y}_U \end{aligned} \tag{3.43}$$

And the same result will be obtained for S_R. Therefore

$$\begin{aligned} E\left(\sum_{i=1}^{n_A} y_i/\pi_i^A\right) &= E\left(\sum_{i=1}^{n_R} y_i/\pi_i^R\right) \\ &= \hat{y}_U \end{aligned} \tag{3.44}$$

Now I have

$$\begin{aligned} \hat{y}_{DR} \xrightarrow{p} E(\hat{y}_{DR}) &= E\Big\{ \sum_{i=1}^{n_A} \frac{(y_i - \hat{y}_i)}{\hat{\pi}_i^A} + \sum_{i=1}^{n_R} \frac{\hat{y}_i}{\pi_i^R} \Big\} \\ &= E\Big\{ \sum_{i=1}^{n_A} \frac{(y_i - \hat{y}_i)}{\hat{\pi}_i^A} + \sum_{i=1}^{n_A} \frac{\hat{y}_i}{\pi_i^A} \Big\} \\ &= E\Big\{ \sum_{i=1}^{n_A} \frac{(y_i - \hat{y}_i)}{\hat{\pi}_i^A} + \frac{\hat{y}_i}{\pi_i^A} \Big\} \\ &= E\Big\{ \sum_{i=1}^{n_A} \frac{\hat{y}_i}{\pi_i^A} + \frac{(y_i - \hat{y}_i)}{\hat{\pi}_i^A} - \frac{(\hat{y}_i - \hat{y}_i)}{\pi_i^A} \Big\} \end{aligned} \tag{3.45}$$

$$\begin{aligned} \hat{y}_{DR} \xrightarrow{p} E(\hat{y}_{DR}) &= y_U + E\Big\{ \sum_{i=1}^{n_A} (y_i - \hat{y}_i)(\frac{1}{\hat{\pi}_i^A} - \frac{1}{\pi_i^A}) \Big\} \\ &= y_U + E\Big\{ \sum_{i=1}^{n_A} (y_i - \hat{y}_i)(\frac{\pi_i^A}{\hat{\pi}_i^A} - 1) \Big\} \end{aligned} \tag{3.46}$$

Under the ignorable assumption in S_A, I have $Y \perp\!\!\!\perp \pi^A | X, \pi^R$. Hence

$$\begin{aligned}
\hat{y}_{DR} \xrightarrow{p} E(\hat{y}_{DR}) &= y_U + E\left\{ \sum_{i=1}^{n_A} (y_i - \hat{y}_i)(\frac{\pi_i^A}{\hat{\pi}_i^A} - 1) \right\} \\
&= y_U + E\left\{ E\left\{ \sum_{i=1}^{n_A} (y_i - \hat{y}_i)(\frac{\pi_i^A}{\hat{\pi}_i^A} - 1) | x_i, \pi_i^R \right\} \right\} \\
&= y_U + E\left\{ \sum_{i=1}^{n_A} E(y_i - \hat{y}_i | x_i, \pi_i^R) E(\frac{\pi_i^A}{\hat{\pi}_i^A} - 1 | x_i, \pi_i^R) \right\}
\end{aligned} \tag{3.47}$$

If I assume the pseudo-weighting model is correctly specified, then I expect $\hat{\pi}_i^A \xrightarrow{p} \pi_i^A$ and

$$E\left(\frac{\pi_i^A}{\hat{\pi}_i^A} - 1 | x_i, \pi_i^R \right) \xrightarrow{p} \frac{\pi_i^A}{\pi_i^A} - 1 = 0 \tag{3.48}$$

which implies that $\hat{y}_{DR} \xrightarrow{p} y_U$ regardless of whether the PM is correctly specified or not. In situations where the mean model is correctly specified, then I expect that $\hat{y}_i \xrightarrow{p} y_i$. Hence

$$E\left(y_i - \hat{y}_i | x_i, \pi_i^R \right) \xrightarrow{p} E\left(y_i - y_i | x_i, \pi_i^R \right) = 0 \tag{3.49}$$

which means that $\hat{y}_{DR} \xrightarrow{p} y_U$ even if the PW model is incorrectly specified.

3.6.1.2 Variance estimation under the Bayesian approach

As discussed in Section 3.2.4, in this study, I use Rubin's combining rule to estimate the variance of the AIPW estimator under the two-step Bayesian approach. The idea stems from the conditional variance formula, which involves two parts: (1) within-imputation variance and between-imputation variance. The latter is straightforward and achieves by taking the variance of the $\hat{\hat{y}}_{DR}^{(m)}$ across the M MCMC draws. The within-imputation variance requires more attention as one needs to account for the intraclass correlations due to clustering and use linearization techniques when dealing with a ratio estimator.

It is clear that this component is calculated conditional on the observed $\hat{y}_i^{(m)}$ for

$i \in S$, $\hat{\pi}_i^{A(m)}$ for $i \in S_A$ and $\hat{p}i_i^A$.

$$
\begin{aligned}
var\left(\hat{\bar{y}}_{DR}^{(m)} \middle| \hat{\pi}_i^{A(m)}, \hat{y}_i^{(m)}\right) &= var\left(\frac{1}{\hat{N}_A}\sum_{i=1}^{n_A}\frac{\left(y_i - \hat{y}_i^{(m)}\right)}{\hat{\pi}_i^A} + \frac{1}{\hat{N}_R}\sum_{i=1}^{n_R}\frac{\hat{y}_i^{(m)}}{\pi_i^R}\middle| \hat{\pi}_i^{A(m)}, \hat{y}_i^{(m)}\right) \\
&= \frac{1}{\hat{N}_A^2}\sum_{i=1}^{n_A}\frac{var(y_i)}{\left(\hat{\pi}_i^A\right)^2} + var\left(\frac{1}{\hat{N}_R}\sum_{i=1}^{n_R}\frac{\hat{y}_i^{(m)}}{\pi_i^R}\middle| \hat{y}_i^{(m)}\right)
\end{aligned}
\tag{3.50}
$$

For the first component, which equals $var(y)\sum_{i=1}^{n_A}\left(\hat{\pi}_i^A\right)^{-2}/\hat{N}_A^2$, it suffices to estimate the variance of y. The second component, however, deals with the variance of a ratio estimator, which requires linearization techniques. Let's define $\hat{t}_R = \sum_{i=1}^{n_R}\hat{y}_i^{(m)}/\pi_i^R$, Taylor-series approximation of the variance is given by

$$
\begin{aligned}
var\left(\frac{1}{\hat{N}_R}\sum_{i=1}^{n_R}\frac{\hat{y}_i^{(m)}}{\pi_i^R}\middle| \hat{y}_i^{(m)}\right) &= var\left(\frac{\hat{t}_R}{\hat{N}_R}\middle| \hat{y}_i^{(m)}\right) \\
&\approx \frac{1}{\hat{N}_R^2}\Bigg\{var\left(\hat{t}_R\middle|\hat{y}_i^{(m)}\right) + \left(\frac{\hat{t}_R}{\hat{N}_R}\right)^2 var(\hat{N}_R) \\
&\quad - 2\left(\frac{\hat{t}_R}{\hat{N}_R}\right) cov\left(\hat{t}_R, \hat{N}_R\middle|\hat{y}_i^{(m)}\right)\Bigg\}
\end{aligned}
\tag{3.51}
$$

Since $\hat{t}_R$ depends on $\hat{y}_i^{(m)}$, I have

$$
var\left(\hat{t}_R\middle|\hat{y}_i^{(m)}\right) = \sum_{i=1}^{n_R}\left(\hat{y}_i^{(m)}\right)^2 var\left(\frac{1}{\pi_i^R}\right) \tag{3.52}
$$

$$
cov\left(\hat{t}_R, \hat{N}_R\middle|\hat{y}_i^{(m)}\right) = \sum_{i=1}^{n_R}\hat{y}_i^{(m)} var\left(\frac{1}{\pi_i^R}\right) \tag{3.53}
$$

Therefore, the variance of the ratio estimator is approximated by

$$
var\left(\frac{1}{\hat{N}_R}\sum_{i=1}^{n_R}\frac{\hat{y}_i^{(m)}}{\pi_i^R}\middle|\hat{y}_i^{(m)}\right) \approx \frac{1}{\hat{N}_R^2}var\left(\frac{1}{\pi_i^R}\right)\left\{\sum_{i=1}^{n_R}\left(\hat{y}_i^{(m)}\right)^2 + n_R\left(\frac{\hat{t}_R}{\hat{N}_R}\right)^2 - 2\sum_{i=1}^{n_R}\hat{y}_i^{(m)}\right\} \tag{3.54}
$$

And the final within-imputation variance can be given by

$$var\left(\hat{\bar{y}}_{DR}^{(m)}\middle|\hat{\pi}_i^{A(m)}, \hat{y}_i^{(m)}\right) \approx \frac{1}{\hat{N}_A^2}\sum_{i=1}^{n_A}\frac{var(y_i)}{\left(\hat{\pi}_i^A\right)^2} + \frac{1}{\hat{N}_R^2}var\left(\frac{1}{\pi_i^R}\right)\left\{\sum_{i=1}^{n_R}\left(\hat{y}_i^{(m)}\right)^2 + n_R\left(\frac{\hat{t}_R}{\hat{N}_R}\right)^2 - 2\sum_{i=1}^{n_R}\hat{y}_i^{(m)}\right\} \tag{3.55}$$

Note that in situations where either S_R or S_A is a clustered sample, the derivation of the within-imputation variance would remain the same, but y_i, π_i^R, $\hat{\pi}_i^{A(m)}$, and $\hat{y}_i^{(m)}$ will represent the total for cluster i, and n_R and n_A are the number of clusters in S_R and S_A, respectively.

3.6.2 Further extensions of the simulation study

3.6.2.1 Simulation study I

This subsection provides additional results associated with Simulation I. Table 3.5 and Table 3.6 summarize the findings of the simulation in 3.3.1 under the frequentist approach when $n_A = 100$, and $n_A = 10,000$. I report the corresponding results under the two-step Bayesian approach in Table 3.7 and Table 3.8, respectively.

Table 3.5: Comparing the performance of the bias adjustment methods and associated asymptotic variance estimator under the frequentist approach in the first simulation study for $n_R = 100$ and $n_A = 100$

	$\rho = 0.3$				$\rho = 0.5$				$\rho = 0.8$			
Method	rBias	rMSE	crCI	rSE	rBias	rMSE	crCI	rSE	rBias	rMSE	crCI	rSE
Probability sample (S_R)												
Unweighted	8.528	19.248	92.6	1.009	8.647	11.065	77.4	1.018	8.682	9.719	50.9	1.02
Fully weighted	-0.029	20.276	94.7	1.001	0.006	8.035	95.1	1.010	0.015	5.008	94.9	1.008
Non-probability sample (S_A)												
Unweighted	31.895	36.418	57.0	1.014	32.213	33.2	1.740	1.008	32.310	32.853	0.0	0.995
Fully weighted	0.171	21.078	94.8	0.996	0.247	8.265	94.9	0.999	0.268	4.994	94.2	0.995
Non-robust adjustment												
Model specification: True												
PAPW	-1.192	23.466	95.2	1.018	-1.205	9.452	95.3	1.015	-1.211	5.982	95.8	1.007
IPSW	-2.917	26.505	97.3	1.386	-3.036	12.700	97.0	1.355	-3.075	9.470	97.0	1.308
PM	0.372	20.989	94.6	0.994	0.148	8.351	94.9	0.995	0.077	5.160	95.0	0.992
Model specification: False												
PAPW	27.140	33.436	75.6	1.059	27.393	28.814	16.6	1.043	27.470	28.276	2.5	1.025
IPSW	28.372	33.972	67.9	1.012	28.711	29.951	8.3	1.002	28.815	29.515	0.5	0.99
PM	28.199	33.790	68.4	1.011	28.541	29.771	8.3	1.001	28.645	29.337	0.3	0.988
Doubly robust adjustment												
Model specification: QR–True, PM–True												
AIPW–PAPW	-0.084	22.973	96.4	1.047	-0.014	8.996	96.2	1.038	0.007	5.368	95.5	1.017
AIPW–IPSW	-0.184	22.449	96.3	1.046	-0.049	8.826	96.1	1.038	-0.009	5.314	95.9	1.016
Model specification: QR–True, PM–False												
AIPW–PAPW	-0.436	23.709	96.4	1.038	-0.286	9.866	96.6	1.062	-0.241	6.520	97.2	1.101
AIPW–IPSW	-0.427	23.083	96.4	1.039	-0.227	9.570	96.6	1.070	-0.166	6.298	97.5	1.119
Model specification: QR–False, PM–True												
AIPW–PAPW	-0.045	29.068	97.3	1.107	0.011	11.113	96.9	1.097	0.026	6.073	96.2	1.068
AIPW–IPSW	-0.194	28.208	97.5	1.104	-0.044	10.825	97.1	1.094	0.001	5.974	96.5	1.062
Model specification: QR–False, PM–False												
AIPW–PAPW	28.301	34.194	71.3	1.037	28.570	29.868	10.9	1.028	28.652	29.379	0.7	1.016
AIPW–IPSW	28.178	33.806	70.4	1.035	28.525	29.764	9.4	1.025	28.631	29.326	0.5	1.013

PAPW: propensity-adjusted probability weighting; IPSW: Inverse propensity score weighting; QR: quasi-randomization; PM: prediction model; AIPW: augmented inverse propensity weighting. Fully weighted implies the weighted means if the true sampling weights are known.

Table 3.6: Comparing the performance of the bias adjustment methods and associated asymptotic variance estimator under the frequentist approach in the first simulation study for $n_R = 100$ and $n_A = 10,000$

	$\rho = 0.2$				$\rho = 0.5$				$\rho = 0.8$			
Method	rBias	rMSE	crCI	rSE	rBias	rMSE	crCI	rSE	rBias	rMSE	crCI	rSE
Probability sample (S_R)												
Unweighted	8.528	19.248	92.6	1.009	8.647	11.065	77.4	1.018	8.682	9.719	50.9	1.02
Fully weighted	-0.029	20.276	94.7	1.001	0.006	8.035	95.1	1.010	0.015	5.008	94.9	1.008
Non-probability sample (S_A)												
Unweighted	30.014	30.066	0.0	1.008	30.197	30.207	0.0	1.019	30.252	30.257	0.0	1.033
Fully weighted	0.032	2.083	95.3	1.005	0.018	0.816	95.1	1.007	0.012	0.490	95.1	1.007
Non-robust adjustment												
Model specification: True												
PAPW	-2.067	4.582	94.9	1.108	-2.145	4.120	92.8	1.107	-2.170	4.072	92.2	1.107
PAPP	-2.618	7.717	94.5	0.958	-2.673	7.334	91.1	0.923	-2.692	7.308	90.6	0.979
PM	0.296	4.515	95.2	0.994	0.121	4.134	94.8	0.986	0.065	4.095	94.6	0.985
Model specification: False												
PAPW	24.493	24.616	0.0	1.126	24.592	24.651	0.0	1.153	24.621	24.673	0.0	1.161
PAPP	26.675	26.804	0.0	0.992	26.871	26.949	0.0	0.970	26.930	27.002	0.0	0.964
PM	26.509	26.645	0.0	1.003	26.717	26.800	0.0	0.989	26.779	26.856	0.0	0.986
Doubly robust adjustment												
Model specification: QR–True, PM–True												
AIPW–PAPW	0.180	4.633	95.1	0.994	0.080	4.162	94.8	0.986	0.047	4.104	94.7	0.985
AIPW–PAPP	0.052	4.582	95.2	0.995	0.035	4.152	94.6	0.987	0.028	4.101	94.5	0.985
Model specification: QR–True, PM–False												
AIPW–PAPW	0.262	4.719	95.1	1.000	0.163	4.250	94.9	0.997	0.130	4.191	94.7	0.996
AIPW–PAPP	0.188	4.652	95.4	1.002	0.171	4.225	95.0	0.998	0.164	4.174	94.8	0.998
Model specification: QR–False, PM–True												
AIPW–PAPW	1.376	8.569	94.5	0.953	0.503	4.829	95.1	0.995	0.231	4.215	95.2	0.992
AIPW–PAPP	0.864	7.648	94.7	0.948	0.322	4.643	95.3	0.990	0.152	4.182	95.0	0.989
Model specification: QR–False, PM–False												
AIPW–PAPW	26.696	26.835	0.0	0.998	26.779	26.862	0.0	0.987	26.803	26.880	0.0	0.985
AIPW–PAPP	26.520	26.655	0.0	1.001	26.718	26.801	0.0	0.989	26.777	26.854	0.0	0.986

PAPW: propensity-adjusted probability weighting; PAPP: propensity-adjusted probability prediction; QR: quasi-randomization; PM: prediction model; AIPW: augmented inverse propensity weighting. Fully weighted implies the weighted means if the true sampling weights are known.

Table 3.7: Comparing the performance of the bias adjustment methods and associated variance estimator under the two-step Bayesian approach in the first simulation study for $n_R = 100$ and $n_A = 100$

	$\rho = 0.3$				$\rho = 0.5$				$\rho = 0.8$			
Method	rBias	rMSE	crCI	rSE	rBias	rMSE	crCI	rSE	rBias	rMSE	crCI	rSE
Probability sample (S_R)												
Unweighted	8.528	19.248	92.6	1.009	8.647	11.065	77.4	1.018	8.682	9.719	50.9	1.020
Fully weighted	-0.029	20.276	94.7	1.001	0.006	8.035	95.1	1.010	0.015	5.008	95.0	1.008
Non-probability sample (S_A)												
Unweighted	32.238	36.815	56.3	1.003	32.303	33.3	1.620	1.003	32.322	32.865	0.0	0.996
Fully weighted	0.494	21.398	94.3	0.981	0.329	8.400	94.0	0.981	0.276	5.057	93.6	0.979
Non-robust adjustment												
Model specification: True												
PAPW	-0.589	24.195	97.4	1.117	-0.755	9.795	99.0	1.326	-0.801	6.178	99.8	1.653
PAPP	1.169	22.844	97.2	1.118	1.016	9.163	98.6	1.345	0.976	5.719	99.8	1.701
PM	0.709	21.489	95.280	1.029	0.272	8.545	95.580	1.020	0.140	5.245	94.640	1.000
Model specification: False												
PAPW	28.008	34.396	76.3	1.091	28.027	29.477	19.6	1.116	28.022	28.840	3.4	1.141
PAPP	29.763	35.215	70.2	1.083	29.827	31.032	10.0	1.106	29.841	30.519	0.8	1.125
PM	28.588	34.226	70.9	1.055	28.658	29.895	10.6	1.050	28.691	29.380	0.7	1.042
Doubly robust adjustment												
Model specification: QR–True, PM–True												
AIPW–PAPW	0.320	23.802	97.8	1.154	0.125	9.306	99.1	1.357	0.067	5.493	99.9	1.731
AIPW–PAPP	0.249	22.778	97.4	1.142	0.099	8.976	99.1	1.339	0.056	5.387	99.9	1.688
Model specification: QR–True, PM–False												
AIPW–PAPW	0.304	23.858	97.7	1.156	0.126	9.386	99.2	1.389	0.065	5.661	99.9	1.781
AIPW–PAPP	0.226	22.814	97.5	1.146	0.096	9.041	99.1	1.376	0.052	5.543	99.8	1.747
Model specification: QR–False, PM–True												
AIPW–PAPW	0.881	22.077	96.8	1.126	0.333	8.742	98.6	1.281	0.153	5.303	99.8	1.558
AIPW–PAPP	0.762	21.483	96.6	1.103	0.290	8.554	98.4	1.251	0.135	5.246	99.7	1.509
Model specification: QR–False, PM–False												
AIPW–PAPW	28.659	34.756	77.6	1.135	28.660	30.013	17.4	1.142	28.649	29.399	2.1	1.151
AIPW–PAPP	28.575	34.237	74.7	1.115	28.656	29.903	13.7	1.124	28.674	29.368	1.1	1.132

PAPW: propensity-adjusted probability weighting; PAPP: propensity-adjusted probability prediction; QR: quasi-randomization; PM: prediction model; AIPW: augmented inverse propensity weighting. Fully weighted implies the weighted means if the true sampling weights are known.

Table 3.8: Comparing the performance of the bias adjustment methods and associated variance estimator under the two-step Bayesian approach in the first simulation study for $n_R = 100$ and $n_A = 10,000$

	$\rho = 0.2$				$\rho = 0.5$				$\rho = 0.8$			
Method	rBias	rMSE	crCI	rSE	rBias	rMSE	crCI	rSE	rBias	rMSE	crCI	rSE
Probability sample (S_R)												
Unweighted	8.528	19.248	92.6	1.009	8.647	11.065	77.4	1.018	8.682	9.719	50.9	1.020
Fully weighted	-0.029	20.276	94.7	1.001	0.006	8.035	95.1	1.010	0.015	5.008	94.9	1.008
Non-probability sample (S_A)												
Unweighted	30.014	30.066	0.0	1.008	30.197	30.207	0.0	1.019	30.252	30.257	0.0	1.033
Fully weighted	0.032	2.083	95.3	1.005	0.018	0.816	95.1	1.007	0.012	0.490	95.1	1.007
Non-robust adjustment												
Model specification: True												
PAPW	-2.032	4.578	93.0	1.031	-2.106	4.111	90.9	1.032	-2.138	4.062	90.2	1.035
PAPP	-0.015	4.094	95.2	1.011	-0.036	3.605	95.1	1.004	-0.042	3.547	95.2	1.002
PM	0.297	4.517	81.6	0.679	0.120	4.136	75.3	0.579	0.065	4.094	73.1	0.563
Model specification: False												
PAPW	24.524	24.647	0.0	1.042	24.618	24.678	0.0	1.062	24.650	24.702	0.0	1.069
PAPP	26.406	26.518	0.0	0.982	26.602	26.662	0.0	0.940	26.663	26.717	0.0	0.931
PM	26.512	26.648	0.0	0.851	26.715	26.798	0.0	0.728	26.779	26.856	0.0	0.700
Doubly robust adjustment												
Model specification: QR–True, PM–True												
AIPW–PAPW	0.178	4.635	84.7	0.721	0.079	4.160	77.3	0.607	0.047	4.103	75.7	0.588
AIPW–PAPP	0.058	4.574	83.6	0.705	0.036	4.149	77.0	0.601	0.028	4.100	75.5	0.585
Model specification: QR–True, PM–False												
AIPW–PAPW	0.151	4.273	94.5	0.971	0.050	3.734	93.7	0.943	0.025	3.660	93.9	0.941
AIPW–PAPP	0.106	4.245	94.4	0.966	0.083	3.767	93.7	0.945	0.075	3.712	93.7	0.941
Model specification: QR–False, PM–True												
AIPW–PAPW	0.496	4.566	83.7	0.709	0.193	4.142	76.8	0.599	0.096	4.096	75.2	0.581
AIPW–PAPP	0.312	4.514	82.7	0.695	0.127	4.133	76.7	0.595	0.068	4.094	74.9	0.579
Model specification: QR–False, PM–False												
AIPW–PAPW	26.709	26.849	0.0	0.893	26.786	26.869	0.0	0.751	26.808	26.885	0.0	0.717
AIPW–PAPP	26.521	26.656	0.0	0.870	26.718	26.800	0.0	0.740	26.777	26.854	0.0	0.709

PAPW: propensity-adjusted probability weighting; PAPP: Inverse propensity score weighting; QR: quasi-randomization; PM: prediction model; AIPW: augmented inverse propensity weighting. Fully weighted implies the weighted means if the true sampling weights are known.

3.6.2.2 Simulation study II

In Table 3.9 and Table 3.10, I provide extensions of Simulation II in Section 3.3.2 for the situations where $n_R = 100$ and $n_A = 100$, and $n_R = 100$ and $n_A = 10,000$, respectively.

Table 3.9: Comparing the performance of the bias adjustment methods and associated variance estimator under the two-step parametric Bayesian approach in the second simulation study for $\rho = 0.2$ and $n_R = 100$ and $n_A = 100$

	SIN				EXP				SQR			
Model-method	rBias	rMSE	crCI	rSE	rBias	rMSE	crCI	rSE	rBias	rMSE	crCI	rSE
Probability sample (S_R)												
Unweighted	-17.210	23.109	80.000	0.999	-8.406	11.126	78.300	1.000	-17.302	20.563	65.800	1.002
Fully weighted	-0.623	17.027	94.440	0.987	-0.303	7.947	94.580	0.987	-0.675	13.219	94.000	0.975
Non-probability sample (S_A)												
Unweighted	32.676	35.764	38.540	0.995	49.188	50.302	0.280	1.083	51.432	53.609	5.200	0.996
Fully weighted	0.100	19.087	93.880	0.982	0.157	8.874	94.220	0.992	0.019	12.338	94.940	1.003
Non-robust adjustment												
Model specification: True												
GLM–PAPW	-1.694	19.888	97.880	1.146	-0.588	9.294	99.960	1.743	-0.905	13.213	99.540	1.517
GLM–PAPP	-0.698	19.524	96.520	1.062	-0.244	9.312	99.780	1.642	-0.151	13.030	99.260	1.43
GLM–PM	-0.705	18.942	95.960	1.022	-0.824	8.451	95.660	1.023	-0.945	13.184	95.480	1.016
Model specification: False												
GLM–PAPW	5.536	22.321	95.960	1.071	-0.225	11.582	99.920	1.699	54.588	57.238	11.760	1.071
GLM–PAPP	6.341	22.406	93.940	0.993	0.550	11.746	99.800	1.590	55.726	58.248	6.820	1.029
BART–PAPW	5.530	20.412	99.420	1.503	4.335	10.413	99.980	2.151	12.487	18.382	99.520	1.864
BART–PAPP	1.435	20.258	98.920	1.362	1.975	9.974	99.980	1.945	6.427	14.735	99.500	1.663
GLM–PM	5.256	19.164	93.400	0.983	-10.991	16.579	88.140	0.994	49.821	52.251	10.980	1.017
BART–PM	4.325	18.758	95.340	1.054	0.848	9.443	97.980	1.175	4.957	14.879	97.140	1.169
Doubly robust adjustment												
Model specification: QR–True, PM–True												
GLM–AIPW–PAPW	-0.773	19.230	97.800	1.144	-0.093	9.047	99.940	1.767	-0.594	13.545	99.480	1.461
GLM–AIPW–PAPP	-0.754	19.197	97.560	1.120	-0.121	9.033	99.900	1.729	-0.582	13.458	99.360	1.435
Model specification: QR–True, PM–False												
GLM–AIPW–PAPW	-0.964	19.745	96.860	1.077	0.107	11.385	99.780	1.539	-0.350	13.394	99.560	1.494
GLM–AIPW–PAPP	-0.590	19.262	96.660	1.067	0.886	11.038	99.780	1.538	0.033	13.420	99.420	1.456
Model specification: QR–False, PM–True												
GLM–AIPW–PAPW	-0.662	20.029	97.820	1.151	-0.044	10.447	99.880	1.831	-0.960	13.302	99.360	1.408
GLM–AIPW–PAPP	-0.671	20.008	97.840	1.134	-0.077	10.340	99.900	1.796	-0.960	13.307	99.240	1.388
Model specification: QR–False, PM–False												
GLM–AIPW–PAPW	7.461	23.271	95.720	1.018	11.977	19.012	99.480	1.432	54.692	57.230	11.000	1.094
GLM–AIPW–PAPP	7.761	22.970	95.260	1.014	11.915	18.421	99.520	1.461	55.257	57.780	9.840	1.084
BART–AIPW–PAPW	2.172	20.303	99.340	1.406	0.878	10.030	99.980	2.058	2.224	14.830	99.800	1.686
BART–AIPW–PAPP	0.965	19.919	99.220	1.389	0.263	9.830	99.980	2.003	1.632	14.527	99.760	1.618

PAPW: propensity-adjusted probability weighting; PAPP: propensity-adjusted probability prediction; QR: quasi-randomization; PM: prediction model; AIPW: augmented inverse propensity weighting.

Table 3.10: Comparing the performance of the bias adjustment methods and associated variance estimator under the two-step parametric Bayesian approach in the second simulation study for $\rho = 0.2$ and $n_R = 100$ and $n_A = 10,000$

	SIN				EXP				SQR			
Model-method	rBias	rMSE	crCI	rSE	rBias	rMSE	crCI	rSE	rBias	rMSE	crCI	rSE
Probability sample (S_R)												
Unweighted	-17.210	23.109	80.000	0.999	-8.406	11.126	78.300	1.000	-17.302	20.563	65.800	1.002
Fully weighted	-0.623	17.027	94.440	0.987	-0.303	7.947	94.580	0.987	-0.675	13.219	94.000	0.975
Non-probability sample (S_A)												
Unweighted	32.730	32.763	0.000	0.989	31.063	31.074	0.000	1.070	41.318	41.339	0.000	1.071
Fully weighted	0.009	1.915	95.140	0.996	0.017	0.855	95.020	1.001	0.000	1.208	95.060	1.011
Non-robust adjustment												
Model specification: True												
GLM–PAPW	-0.057	7.609	94.320	0.985	-0.026	3.427	95.340	1.010	-0.160	5.482	95.580	1.035
GLM–PAPP	0.148	6.467	94.420	0.977	0.151	2.873	94.540	0.982	-0.050	4.550	94.660	1.047
GLM–PM	-0.326	8.558	94.400	0.968	-0.199	3.928	93.680	0.971	-0.631	7.286	90.780	0.937
Model specification: False												
GLM–PAPW	8.030	10.415	74.860	0.994	2.785	4.470	90.560	1.104	43.528	43.741	0.000	1.108
GLM–PAPP	8.292	9.699	60.440	0.978	3.021	4.342	92.040	1.115	44.333	44.417	0.000	1.292
BART–PAPW	4.237	8.862	91.608	1.032	2.958	4.569	90.909	1.126	7.599	9.773	93.629	1.280
BART–PAPP	1.206	6.469	93.706	0.980	1.838	3.323	95.338	1.154	3.656	5.768	96.115	1.198
GLM–PM	5.863	9.223	86.020	0.973	-3.579	6.212	87.400	0.974	40.997	41.273	0.000	0.973
BART–PM	-0.037	8.703	93.939	0.987	0.024	4.011	94.328	0.991	-0.082	7.543	92.385	0.945
Doubly robust adjustment												
Model specification: QR–True, PM–True												
GLM–AIPW–PAPW	-0.354	8.557	94.320	0.974	-0.176	3.936	94.380	0.990	-0.478	7.264	91.920	0.954
GLM–AIPW–PAPP	-0.354	8.556	94.340	0.973	-0.177	3.937	94.320	0.989	-0.476	7.258	91.960	0.954
Model specification: QR–True, PM–False												
GLM–AIPW–PAPW	-0.126	8.515	90.000	0.848	0.128	4.553	97.480	1.168	-0.344	6.164	94.360	1.000
GLM–AIPW–PAPP	-0.070	7.894	91.660	0.895	0.194	3.880	98.760	1.350	-0.360	6.510	94.440	1.007
Model specification: QR–False, PM–True												
GLM–AIPW–PAPW	-0.303	8.572	94.480	0.973	-0.205	3.949	94.440	0.992	-0.635	7.292	91.280	0.948
GLM–AIPW–PAPP	-0.305	8.571	94.460	0.973	-0.207	3.950	94.540	0.992	-0.636	7.288	91.240	0.949
Model specification: QR–False, PM–False												
GLM–AIPW–PAPW	8.784	11.327	78.680	1.024	4.599	6.234	93.100	1.315	42.262	42.571	0.000	0.961
GLM–AIPW–PAPP	8.955	10.851	80.600	1.184	4.694	5.741	95.800	1.688	42.941	43.252	0.000	1.003
BART–AIPW–PAPW	-0.042	8.880	93.862	0.985	0.039	4.099	94.639	1.004	-0.267	7.589	92.618	0.962
BART–AIPW–PAPP	-0.287	8.758	94.017	0.992	-0.064	4.022	94.639	1.015	-0.274	7.571	92.230	0.959

PAPW: propensity-adjusted probability weighting; PAPP: propensity-adjusted probability prediction; QR: quasi-randomization; PM: prediction model; AIPW: augmented inverse propensity weighting.

3.6.2.3 Simulation study III

Table 3.11 and Table 3.12 exhibits the numerical results associated with the plots of Simulation III in Section 3.3.3.

Table 3.11: Comparing the performance of the bias adjustment methods in the third simulation study for $\rho = 0.8$

	Continuous outcome (Y_c)				**Binary outcome (Y_b)**			
Model-method	rBias	rMSE	crCI	rSE	rBias	rMSE	crCI	rSE
Probability sample (S_R)								
Unweighted	48.705	52.900	30.7	1.015	11.304	16.881	88.2	1.022
Fully weighted	0.080	15.400	96.2	1.025	0.131	13.858	95.3	1.026
Non-probability sample (S_A)								
Unweighted	68.309	70.415	0.0	0.156	21.763	22.794	0.5	0.181
Fully weighted	0.137	7.581	95.7	1.023	0.074	6.512	94.7	0.99
Non-robust adjustment								
Model specification: True								
GLM–PAPW	0.448	10.994	94.7	1.036	0.072	7.266	96.2	1.034
GLM–PAPP	0.204	11.192	93.9	1.037	0.080	7.188	96.2	1.031
GLM–IPSW	0.839	18.138	96.0	1.275	-0.838	9.458	97.3	1.116
GLM–PM	0.110	11.157	94.2	1.015	0.055	7.401	94.4	0.995
Model specification: False								
GLM–PAPW	7.337	13.187	94.2	1.033	5.115	8.502	90.4	1.02
GLM–PAPP	6.762	13.546	94.2	1.032	5.046	8.471	88.5	1.035
GLM–IPSW	22.513	35.600	99.5	1.155	9.390	13.098	89.5	1.099
BART–PAPW	2.272	10.468	100.0	2.487	1.633	7.391	99.5	1.436
BART–PAPP	3.990	11.469	100.0	2.299	0.313	7.243	99.3	1.342
GLM–PM	37.071	42.523	53.0	1.006	12.600	14.932	63.6	1.003
BART–PM	0.286	11.581	92.7	0.996	0.594	9.102	81.2	0.688
Doubly robust adjustment								
Model specification: QR–True, PM–True								
GLM–AIPW–PAPW	0.307	11.186	95.0	1.019	0.083	7.459	94.2	1.001
GLM–AIPW–PAPP	0.295	11.187	94.5	1.019	0.089	7.439	94.0	0.998
GLM–AIPW–IPSW	0.372	11.193	95.8	1.037	0.120	7.478	94.4	1.003
Model specification: QR–True, PM–False								
GLM–AIPW–PAPW	0.381	12.774	95.5	1.035	0.047	7.487	96.2	1.04
GLM–AIPW–PAPP	0.424	11.934	94.7	1.041	0.155	7.275	96.0	1.032
GLM–AIPW–IPSW	-8.223	17.625	92.3	1.181	-2.842	9.086	95.2	1.047
Model specification: QR–False, PM–True								
GLM–AIPW–PAPW	0.127	11.177	94.7	1.020	0.067	7.451	94.0	0.997
GLM–AIPW–PAPP	0.122	11.172	94.7	1.019	0.054	7.438	94.2	0.997
GLM–AIPW–IPSW	0.117	11.167	94.8	1.020	0.055	7.433	94.0	0.998
Model specification: QR–False, PM–False								
GLM–AIPW–PAPW	50.327	53.922	21.9	1.002	15.651	17.552	50.3	1.007
GLM–AIPW–PAPP	50.793	54.215	20.9	1.002	15.834	17.605	47.8	1.003
GLM–AIPW–IPSW	47.867	51.106	27.9	1.163	15.112	16.884	53.8	1.051
BART–AIPW–PAPW	0.276	11.593	94.4	1.035	0.701	9.186	81.9	0.698
BART–AIPW–PAPP	0.261	11.591	94.2	1.031	0.682	9.155	81.7	0.697

PAPW: propensity-adjusted probability weighting; PAPP: propensity-adjusted probability prediction; IPSW: Inverse propensity score weighting; QR: quasi-randomization; PM: prediction model; AIPW: augmented inverse propensity weighting.

Table 3.12: Comparing the values of rBias and rMSE for different methods across different values of ρ.

	Continuous outcome (Y_c)							**Binary outcome (Y_b)**						
	Non-robust				Doubly robust			Non-robust				Doubly robust		
ρ	PAPW	PAPP	IPSW	PM	PAPW	PAPP	IPSW	PAPW	PAPP	IPSW	PM	PAPW	PAPP	IPSW
							rBias							
0.0	0.545	0.870	-6.791	0.259	0.447	0.443	0.511	-0.186	0.128	-3.248	-0.016	-0.006	-0.005	0.004
0.1	-0.179	0.022	-4.772	-0.224	-0.215	-0.218	-0.235	-0.537	-0.220	-2.345	-0.399	-0.464	-0.475	-0.489
0.2	-0.195	0.493	-4.406	0.048	-0.161	-0.160	-0.250	-0.329	0.095	-2.071	-0.144	-0.159	-0.149	-0.172
0.3	0.288	0.668	-3.832	0.420	0.459	0.449	0.435	-0.244	0.069	-1.980	0.160	0.108	0.114	0.111
0.4	0.212	0.361	-2.332	0.425	0.237	0.254	0.233	-0.097	0.150	-1.372	-0.031	0.031	0.048	0.085
0.5	0.248	0.173	-2.257	0.175	0.227	0.216	0.239	-0.169	-0.067	-1.817	0.117	0.068	0.065	-0.010
0.6	0.286	0.516	-1.010	0.411	0.404	0.404	0.420	0.128	0.231	-1.146	0.133	0.162	0.156	0.198
0.7	0.072	-0.052	-0.217	-0.027	0.084	0.084	0.100	-0.019	0.062	-1.029	0.021	-0.001	0.009	-0.001
0.8	0.538	0.527	0.652	0.623	0.509	0.517	0.498	0.012	0.175	-1.017	0.015	0.053	0.048	0.078
0.9	0.343	0.424	2.090	0.469	0.496	0.495	0.466	0.079	0.122	-0.932	0.155	0.158	0.144	0.164
							rMSE							
0.0	13.916	14.724	20.949	14.934	14.964	14.934	14.994	8.702	8.702	11.773	9.214	9.214	9.214	9.214
0.1	12.979	13.640	19.378	13.760	13.790	13.790	13.850	8.443	8.188	10.746	8.699	8.699	8.699	8.699
0.2	12.297	13.220	18.877	13.161	13.220	13.220	13.250	8.237	8.237	10.811	8.494	8.494	8.494	8.494
0.3	12.187	13.049	18.132	13.019	13.049	13.019	13.049	7.859	7.859	9.887	8.113	8.113	8.113	8.366
0.4	11.823	12.392	17.330	12.452	12.511	12.511	12.511	7.910	7.654	9.951	8.165	8.165	8.165	8.165
0.5	11.745	12.101	17.647	12.190	12.190	12.190	12.190	7.576	7.576	9.849	7.829	7.829	7.829	7.829
0.6	11.691	12.145	18.748	12.085	12.115	12.115	12.115	7.673	7.673	9.975	7.929	7.929	7.929	7.929
0.7	10.927	11.166	15.567	11.196	11.226	11.226	11.226	7.332	7.080	8.849	7.332	7.585	7.332	7.585
0.8	10.769	10.918	17.888	10.918	10.918	10.918	10.948	7.359	7.105	9.389	7.359	7.612	7.359	7.612
0.9	10.951	11.042	22.084	10.981	11.012	11.012	11.012	6.935	6.935	8.990	7.192	7.192	7.192	7.192

NOTE: GLM has been used for prediction, and the underlying models in each method have been correctly specified.

3.6.3 Supplemental results on SHRP2/NHTS data

Table 3.13: Mean daily trip duration (min) and associated 95% CIS by different covariates across DR adjustment methods

Covariate	n	Unweighted (95%CI)	GLM-AIPW-PAPP (95%CI)	GLM-AIPW-PMLE (95%CI)	BART–AIPW-PAPP (95%CI)
Total	837,061	68.94 (67.955,69.925)	71.603 (66.565,76.641)	70.058 (67.902,72.214)	69.582 (66.117,73.047)
Gender					
Male	407,312	70.289 (68.809,71.77)	72.411 (63.583,81.238)	70.97 (67.971,73.97)	70.61 (66.131,75.088)
Female	429,749	67.662 (66.355,68.968)	70.79 (67.353,74.226)	69.107 (66.683,71.531)	68.522 (64.432,72.611)
Age group					
16-24	311,106	70 (68.514,71.485)	72.889 (69.435,76.342)	72.318 (69.636,74.999)	71.937 (66.79,77.085)
25-34	117,758	73.889 (71.099,76.679)	72.669 (67.713,77.625)	71.562 (67.688,75.435)	72.511 (66.132,78.889)
35-44	61,908	75.4 (71.304,79.496)	71.215 (64.668,77.762)	75.72 (69.882,81.559)	71.919 (63.874,79.964)
45-54	77,903	74.666 (71.734,77.599)	71.803 (61.432,82.175)	70.437 (66.525,74.349)	73.237 (67.727,78.747)
55-64	63,891	70.823 (67.027,74.62)	66.99 (60.85,73.13)	67.054 (62.252,71.855)	67.518 (60.885,74.152)
65-74	88,762	67.122 (64.13,70.113)	84.262 (52.155,116.369)	64.475 (59.374,69.576)	64.286 (59.779,68.794)
75+	115,733	54.103 (51.965,56.241)	49.358 (46.14,52.576)	51.359 (47.896,54.822)	51.442 (46.894,55.99)
Race					
White	745,596	67.845 (66.833,68.858)	71.687 (65.246,78.128)	68.183 (65.836,70.529)	67.861 (64.386,71.336)
Black	43,109	86.294 (80.759,91.83)	74.42 (66.374,82.466)	81.587 (75.046,88.127)	79.728 (68.019,91.437)
Asian	26,265	68.723 (63.684,73.761)	66.792 (58.089,75.495)	66.777 (60.785,72.769)	65.958 (53.748,78.169)
Other	22,091	72.284 (66.895,77.674)	79.723 (69.505,89.942)	75.924 (69.729,82.118)	75.314 (63.089,87.539)
Ethnicity					
Non-Hisp	808,098	68.699 (67.697,69.701)	71.999 (66.066,77.933)	69.337 (67.166,71.507)	68.555 (64.866,72.244)
Hispanic	28,963	75.681 (70.488,80.873)	72.068 (63.599,80.536)	74.545 (69.145,79.944)	75.449 (66.582,84.316)
Education					
<High school	50,943	61.108 (58.134,64.083)	67.647 (58.129,77.165)	67.32 (61.588,73.051)	68.246 (56.385,80.108)
HS completed	78,045	69.025 (65.979,72.071)	86.848 (58.569,115.128)	69.752 (64.868,74.637)	70.399 (61.472,79.325)
College	237,206	68.997 (67.153,70.841)	70.312 (64.184,76.44)	70.712 (66.638,74.785)	70.896 (65.722,76.069)
Graduate	326,860	70.859 (69.188,72.529)	71.314 (68.333,74.296)	71.313 (69.073,73.554)	69.984 (65.783,74.186)
Post-grad	144,007	67.218 (64.984,69.451)	64.26 (60.143,68.377)	68.713 (64.864,72.562)	66.496 (62.395,70.597)
HH income					
0-49	332,586	68.105 (66.553,69.658)	75.441 (62.136,88.745)	69.441 (65.872,73.009)	69.049 (65.13,72.968)
50-99	309,387	69.755 (68.089,71.421)	70.608 (63.639,77.578)	70.359 (67.276,73.442)	69.836 (66.552,73.12)
100-149	132,757	69.487 (66.999,71.975)	68.685 (63.743,73.626)	70.276 (66.911,73.642)	69.55 (60.835,78.265)
150+	62,331	68.187 (65.109,71.264)	69.772 (66.389,73.154)	69.9 (66.158,73.643)	70.352 (64.31,76.394)
HH size					
1	177,140	66.779 (64.452,69.106)	80.258 (54.973,105.544)	66.501 (62.817,70.186)	67.607 (63.28,71.934)
2	286,106	67.608 (65.994,69.223)	65.532 (61.489,69.574)	66.781 (63.894,69.667)	67.282 (63.371,71.193)
3	152,684	71.233 (68.836,73.631)	72.398 (66.412,78.384)	74.177 (69.507,78.848)	71.127 (67.04,75.214)
4	143,442	70.161 (67.969,72.352)	69.794 (65.273,74.315)	69.944 (66.494,73.395)	70.839 (65.417,76.261)
5+	77,689	72.012 (68.913,75.11)	74.664 (64.68,84.648)	76.567 (71.368,81.765)	73.321 (68.479,78.163)
Urban size					
<50k	34,987	67.602 (62.771,72.432)	79.22 (59.18,99.26)	65.75 (59.749,71.751)	66.109 (57.069,75.149)
50-200k	119,970	62.608 (60.337,64.879)	65.759 (61.25,70.268)	65.151 (62.164,68.138)	67.211 (61.409,73.014)
200-500k	44,578	68.576 (63.52,73.632)	87.248 (73.018,101.477)	68.884 (63.664,74.104)	69.636 (61.746,77.526)
500-1000k	276,629	68.017 (66.289,69.745)	66.524 (61.364,71.685)	68.123 (65.323,70.923)	70.338 (64.971,75.704)
1000k+	360,897	71.928 (70.451,73.404)	70.91 (67.926,73.894)	73.567 (71.441,75.693)	72.962 (68.493,77.43)
Vehicle make					
American	290,228	66.507 (64.905,68.108)	71.826 (59.917,83.734)	68.256 (65.302,71.21)	69.04 (63.968,74.113)
Asian	528,810	70.265 (69,71.53)	72.7 (69.653,75.747)	71.602 (69.436,73.768)	70.211 (66.415,74.007)
European	18,023	69.261 (63.898,74.624)	66.191 (59.703,72.679)	71.403 (65.95,76.855)	69.836 (60.506,79.166)
Vehicle type					
Car	610,245	68.686 (67.539,69.834)	73.853 (65.931,81.776)	69.706 (67.4,72.012)	70.236 (66.799,73.673)
Van	27,866	69.2 (64.432,73.968)	68.389 (61.064,75.714)	73.096 (66.388,79.804)	64.905 (54.298,75.512)
SUV	158,202	68.993 (66.851,71.134)	68.424 (62.145,74.704)	69.291 (66.318,72.263)	69.469 (64.351,74.587)
Pickup	40,748	72.361 (66.713,78.008)	69.934 (59.062,80.805)	74.495 (64.949,84.04)	70.256 (58.87,81.643)
Fuel type					
Gas/D	761,292	68.637 (67.61,69.664)	71.334 (66.221,76.446)	69.895 (67.66,72.131)	69.443 (65.954,72.931)
Other	75,769	71.986 (68.598,75.373)	82.674 (72.987,92.361)	77.039 (72.37,81.708)	75.696 (67.822,83.571)
Weekend					
Weekday	712,411	67.671 (66.701,68.64)	70.362 (65.734,74.991)	68.72 (66.616,70.824)	68.348 (64.806,71.89)
Weekend	124,650	76.196 (75.001,77.392)	78.646 (71.128,86.164)	77.649 (75.099,80.199)	76.577 (73.08,80.074)

Table 3.14: Mean daily trip distance (mile) and associated 95% CIS by different covariates across DR adjustment methods

Covariate	n	Unweighted (95%CI)	GLM-AIPW-PAPP (95%CI)	GLM-AIPW-PMLE (95%CI)	BART–AIPW-PAPP (95%CI)
Total	837,061	32.418 (31.823,33.013)	33.76 (31.806,35.715)	33.39 (32.22,34.56)	32.926 (31.185,34.667)
Gender					
Male	407,312	33.852 (32.963,34.741)	35.51 (32.247,38.773)	34.782 (33.146,36.418)	34.358 (32.254,36.461)
Female	429,749	31.06 (30.27,31.849)	31.901 (29.932,33.871)	31.947 (30.601,33.293)	31.428 (28.965,33.89)
Age group					
16-24	311,106	32.828 (32,33.657)	34.904 (33.085,36.723)	34.864 (33.358,36.369)	34.491 (32.804,36.178)
25-34	117,758	36.246 (34.603,37.888)	36.546 (33.364,39.728)	34.841 (32.742,36.94)	35.324 (32.837,37.81)
35-44	61,908	35.958 (33.318,38.597)	31.774 (28.585,34.962)	35.067 (32.321,37.813)	33.9 (30.173,37.627)
45-54	77,903	36.103 (34.231,37.976)	35.721 (31.46,39.981)	34.301 (31.843,36.759)	34.578 (30.108,39.049)
55-64	63,891	35.037 (32.735,37.34)	33.159 (29.352,36.966)	33.138 (30.097,36.178)	32.135 (29.896,34.375)
65-74	88,762	31.548 (29.552,33.544)	33.875 (24.893,42.856)	29.948 (26.934,32.962)	29.028 (26.593,31.464)
75+	115,733	22.269 (21.044,23.493)	20.184 (18.108,22.259)	21.037 (19.304,22.77)	21.421 (18.979,23.863)
Race					
White	745,596	32.189 (31.554,32.824)	33.173 (30.862,35.484)	33.426 (32.11,34.743)	32.85 (31.118,34.582)
Black	43,109	37.275 (34.577,39.973)	35.696 (31.364,40.029)	34.187 (31.146,37.228)	34.131 (28.834,39.427)
Asian	26,265	30.638 (28.095,33.181)	29.601 (25.527,33.675)	28.311 (25.238,31.383)	28.289 (22.62,33.958)
Other	22,091	32.789 (29.699,35.879)	37.919 (30.975,44.862)	35.518 (30.553,40.484)	35.058 (27.876,42.239)
Ethnicity					
Non-Hisp	808,098	32.328 (31.723,32.933)	32.852 (30.823,34.881)	33.217 (32.085,34.349)	32.362 (30.845,33.879)
Hispanic	28,963	34.935 (31.713,38.158)	36.882 (32.766,40.997)	35.126 (31.226,39.027)	36.344 (29.859,42.828)
Education					
<High school	50,943	25.659 (23.905,27.412)	28.248 (24.014,32.482)	28.977 (25.986,31.967)	30.351 (22.902,37.8)
HS completed	78,045	32.04 (30.14,33.939)	36.853 (29.19,44.515)	33.596 (30.989,36.203)	33.497 (30.336,36.658)
College	237,206	31.848 (30.812,32.885)	33.666 (30.509,36.824)	33.038 (31.177,34.899)	32.956 (30.622,35.29)
Graduate	326,860	33.879 (32.85,34.908)	35.56 (32.73,38.389)	35.093 (33.48,36.706)	34.091 (31.938,36.244)
Post-grad	144,007	32.637 (31.235,34.039)	29.935 (27.6,32.269)	32.801 (30.877,34.725)	31.414 (29.555,33.273)
HH income					
0-49	332,586	31.185 (30.273,32.097)	32.845 (28.788,36.901)	31.979 (30.333,33.626)	31.538 (28.932,34.144)
50-99	309,387	33.024 (32.004,34.043)	35.811 (32.755,38.866)	33.907 (32.201,35.613)	33.744 (32.011,35.476)
100-149	132,757	33.765 (32.235,35.295)	33.354 (30.228,36.479)	34.433 (32.472,36.394)	33.532 (30.736,36.329)
150+	62,331	33.124 (31.231,35.017)	31.693 (29.605,33.781)	33.795 (31.147,36.442)	33.428 (29.886,36.97)
HH size					
1	177,140	30.588 (29.231,31.945)	34.322 (27.864,40.779)	31.133 (28.899,33.366)	30.768 (28.385,33.152)
2	286,106	32.415 (31.372,33.458)	31.701 (29.362,34.039)	32.742 (30.989,34.494)	32.301 (30.95,33.651)
3	152,684	33.786 (32.452,35.12)	34.54 (30.838,38.242)	34.806 (32.647,36.966)	34.421 (31.549,37.293)
4	143,442	32.95 (31.524,34.376)	32.048 (29.257,34.84)	33.04 (31.09,34.99)	32.898 (30.012,35.784)
5+	77,689	32.934 (31.314,34.554)	36.522 (32.397,40.647)	36.383 (33.774,38.993)	34.731 (32.103,37.359)
Urban size					
<50k	34,987	36.147 (32.885,39.408)	34.93 (28.343,41.518)	34.077 (30.506,37.648)	32.945 (30.263,35.628)
50-200k	119,970	31.028 (29.388,32.668)	32.379 (29.47,35.288)	32.032 (29.692,34.372)	32.636 (30.058,35.215)
200-500k	44,578	36.416 (33.616,39.216)	44.143 (36.054,52.231)	35.585 (33.228,37.942)	36.461 (32.151,40.771)
500-1000k	276,629	31.973 (30.952,32.994)	32.453 (27.672,37.234)	31.005 (29.331,32.679)	32.781 (28.04,37.522)
1000k+	360,897	32.366 (31.497,33.236)	32.475 (30.577,34.373)	32.371 (31.117,33.626)	32.302 (30.134,34.471)
Vehicle make					
American	290,228	30.948 (29.995,31.9)	35.285 (31.317,39.254)	32.784 (31.234,34.335)	32.956 (30.909,35.004)
Asian	528,810	33.249 (32.476,34.022)	33.339 (31.554,35.124)	34.022 (32.623,35.42)	33.124 (31.156,35.092)
European	18,023	31.719 (28.896,34.542)	29.905 (26.439,33.37)	32.979 (30.006,35.951)	32.05 (27.154,36.946)
Vehicle type					
Car	610,245	32.126 (31.428,32.823)	33.619 (31.253,35.985)	32.916 (31.691,34.141)	32.518 (30.426,34.611)
Van	27,866	31.212 (28.225,34.199)	29.109 (24.54,33.679)	32.682 (28.052,37.312)	31.109 (24.519,37.699)
SUV	158,202	32.848 (31.558,34.137)	33.857 (30.466,37.249)	33.15 (31.374,34.926)	32.559 (29.986,35.133)
Pickup	40,748	35.958 (32.813,39.103)	35.086 (30.375,39.798)	36.557 (32.917,40.197)	36.352 (31.036,41.668)
Fuel type					
Gas/D	761,292	32.121 (31.502,32.739)	33.524 (31.522,35.526)	33.18 (32.006,34.354)	32.813 (31.021,34.605)
Other	75,769	35.409 (33.302,37.515)	43.864 (37.513,50.214)	39.259 (35.942,42.576)	37.388 (34.271,40.505)
Weekend					
Weekday	712,411	31.895 (31.307,32.482)	33.181 (31.327,35.036)	32.817 (31.666,33.968)	32.41 (30.68,34.14)
Weekend	124,650	35.41 (34.689,36.132)	37.037 (34.351,39.724)	36.64 (35.09,38.19)	35.853 (33.806,37.899)

Table 3.15: Mean daily average speed (MPH) of trips and associated 95% CIS by different covariates across DR adjustment methods

Covariate	n	Unweighted (95%CI)	GLM-AIPW-PAPP (95%CI)	GLM-AIPW-PMLE (95%CI)	BART–AIPW-PAPP (95%CI)
Total	837,061	25.03 (24.8,25.261)	25.775 (24.277,27.274)	25.562 (25.063,26.06)	25.39 (24.309,26.471)
Gender					
Male	407,312	25.474 (25.137,25.811)	26.964 (24.667,29.261)	26.199 (25.578,26.82)	25.999 (24.877,27.12)
Female	429,749	24.61 (24.297,24.923)	24.463 (23.239,25.688)	24.906 (24.32,25.492)	24.76 (23.59,25.93)
Age group					
16-24	311,106	25.239 (24.893,25.586)	25.876 (24.878,26.873)	25.921 (25.287,26.555)	25.831 (24.724,26.938)
25-34	117,758	26.951 (26.318,27.583)	27.242 (26.062,28.422)	26.571 (25.723,27.419)	27.07 (26.055,28.085)
35-44	61,908	26.065 (25.183,26.947)	25.045 (23.196,26.894)	25.696 (24.675,26.716)	25.516 (23.327,27.705)
45-54	77,903	26.527 (25.727,27.328)	27.731 (22.197,33.265)	26.412 (25.419,27.405)	25.665 (23.242,28.088)
55-64	63,891	26.22 (25.471,26.969)	26.075 (24.48,27.67)	26.275 (25.152,27.398)	25.525 (24.078,26.971)
65-74	88,762	23.956 (23.339,24.572)	22.601 (20.487,24.716)	23.618 (22.683,24.554)	23.216 (22.2,24.232)
75+	115,733	21.12 (20.559,21.681)	20.545 (19.251,21.838)	21.317 (20.539,22.094)	20.728 (19.29,22.167)
Race					
White	745,596	25.109 (24.863,25.354)	25.225 (24.255,26.196)	26.086 (25.565,26.608)	25.656 (24.95,26.363)
Black	43,109	24.227 (23.188,25.265)	27.223 (22.295,32.151)	23.198 (22.202,24.194)	23.433 (20.47,26.397)
Asian	26,265	24.35 (23.278,25.423)	25.203 (23.46,26.947)	23.038 (21.674,24.402)	24.508 (21.082,27.934)
Other	22,091	24.76 (23.473,26.046)	25.049 (23.008,27.091)	24.68 (23.128,26.232)	25.956 (22.168,29.744)
Ethnicity					
Non-Hisp	808,098	25.039 (24.805,25.274)	25.023 (24.156,25.89)	25.674 (25.197,26.152)	25.246 (24.437,26.055)
Hispanic	28,963	24.777 (23.526,26.028)	27.808 (24.042,31.574)	25.233 (23.873,26.593)	26.284 (22.881,29.688)
Education					
<High school	50,943	23.16 (22.31,24.01)	23.142 (21.364,24.919)	23.825 (22.322,25.328)	24.791 (21.638,27.943)
HS completed	78,045	25.192 (24.438,25.945)	24.851 (22.707,26.995)	26.025 (25.019,27.031)	25.165 (22.538,27.793)
College	237,206	24.506 (24.09,24.921)	25.888 (22.586,29.189)	24.988 (24.184,25.793)	24.7 (23.717,25.683)
Graduate	326,860	25.426 (25.043,25.81)	26.769 (25.764,27.775)	26.363 (25.795,26.93)	26.368 (25.077,27.659)
Post-grad	144,007	25.569 (25.027,26.11)	25.031 (23.955,26.107)	25.446 (24.775,26.117)	25.555 (24.399,26.711)
HH income					
0-49	332,586	24.333 (23.956,24.709)	23.975 (22.766,25.183)	24.659 (23.851,25.467)	24.401 (22.918,25.884)
50-99	309,387	25.25 (24.878,25.623)	27.547 (24.479,30.615)	25.971 (25.316,26.627)	25.744 (24.828,26.66)
100-149	132,757	25.963 (25.411,26.515)	25.892 (24.611,27.174)	26.281 (25.569,26.994)	25.981 (24.275,27.687)
150+	62,331	22.937 (22.564,23.31)	25.096 (21.747,28.444)	23.419 (22.632,24.206)	23.288 (22.044,24.533)
HH size					
1	177,140	23.837 (23.337,24.337)	23.986 (22.176,25.797)	24.355 (23.538,25.173)	24.024 (22.597,25.452)
2	286,106	25.155 (24.746,25.563)	25.606 (24.679,26.532)	25.778 (25.128,26.428)	25.55 (24.735,26.365)
3	152,684	25.77 (25.223,26.316)	26.042 (24.785,27.3)	25.645 (24.886,26.403)	26.035 (24.807,27.262)
4	143,442	25.423 (24.895,25.952)	25.025 (23.669,26.381)	25.766 (25.015,26.517)	25.622 (24.254,26.991)
5+	77,689	25.112 (24.48,25.745)	27.472 (22.365,32.58)	26.14 (24.981,27.3)	25.155 (23.23,27.08)
Urban size					
<50k	34,987	28.437 (27.061,29.813)	25.595 (22.943,28.247)	27.951 (26.354,29.548)	27.097 (25.536,28.659)
50-200k	119,970	24.455 (23.814,25.096)	25.081 (23.851,26.31)	24.784 (23.965,25.603)	25.031 (24.155,25.907)
200-500k	44,578	27.64 (26.634,28.645)	27.073 (23.546,30.601)	27.024 (26.049,27.999)	26.931 (24.582,29.279)
500-1000k	276,629	25.758 (25.355,26.162)	26.189 (23.701,28.678)	25.05 (24.289,25.812)	25.513 (24.121,26.904)
1000k+	360,897	20.451 (18.941,21.961)	23.557 (21.359,25.755)	21.9 (20.41,23.389)	23.488 (20.639,26.336)
Vehicle make					
American	290,228	24.799 (24.402,25.195)	27.212 (24.331,30.094)	25.766 (25.047,26.485)	25.353 (24.079,26.627)
Asian	528,810	25.174 (24.884,25.464)	24.771 (23.69,25.853)	25.509 (25.004,26.015)	25.464 (24.358,26.569)
European	18,023	24.534 (23.553,25.514)	24.974 (23.083,26.866)	24.291 (22.942,25.64)	25.307 (23.285,27.329)
Vehicle type					
Car	610,245	24.893 (24.622,25.164)	25.115 (24.327,25.904)	25.313 (24.794,25.832)	25.357 (24.467,26.247)
Van	27,866	23.562 (22.539,24.586)	23.064 (21.378,24.75)	23.484 (22.376,24.591)	23.527 (20.832,26.223)
SUV	158,202	25.398 (24.87,25.925)	26.495 (22.622,30.369)	25.635 (24.887,26.384)	25.008 (23.511,26.505)
Pickup	40,748	26.43 (25.484,27.375)	26.245 (23.43,29.059)	26.628 (25.453,27.804)	25.788 (23.842,27.733)
Fuel type					
Gas/D	761,292	24.955 (24.711,25.199)	25.727 (24.205,27.249)	25.507 (25.005,26.01)	25.361 (24.277,26.446)
Other	75,769	25.784 (25.091,26.476)	27.804 (26.114,29.493)	27.052 (25.991,28.113)	26.676 (24.691,28.66)
Weekend					
Weekday	712,411	25.077 (24.847,25.308)	25.744 (24.351,27.138)	25.598 (25.1,26.096)	25.425 (24.36,26.49)
Weekend	124,650	24.76 (24.518,25.003)	25.939 (23.843,28.034)	25.356 (24.811,25.901)	25.194 (23.987,26.401)

Table 3.16: Mean start time of the first daytrips and associated 95% CIS by different covariates across DR adjustment methods

Covariate	n	Unweighted (95%CI)	GLM-AIPW-PAPP (95%CI)	GLM-AIPW-PMLE (95%CI)	BART–AIPW-PAPP (95%CI)
Total	837,061	13.811 (13.763,13.859)	13.564 (13.391,13.737)	13.553 (13.427,13.68)	13.5 (13.364,13.636)
Gender					
Male	407,312	13.824 (13.751,13.898)	13.556 (13.304,13.807)	13.572 (13.418,13.725)	13.486 (13.304,13.667)
Female	429,749	13.799 (13.736,13.861)	13.578 (13.362,13.793)	13.533 (13.386,13.681)	13.515 (13.389,13.64)
Age group					
16-24	311,106	14.411 (14.354,14.468)	14.396 (14.254,14.537)	14.351 (14.218,14.485)	14.266 (14.13,14.402)
25-34	117,758	13.999 (13.891,14.106)	13.864 (13.562,14.165)	13.923 (13.734,14.112)	13.843 (13.65,14.037)
35-44	61,908	13.57 (13.399,13.741)	13.694 (13.178,14.211)	13.467 (13.164,13.77)	13.448 (13.117,13.78)
45-54	77,903	13.489 (13.368,13.61)	13.414 (13.028,13.8)	13.389 (13.187,13.592)	13.335 (13.043,13.626)
55-64	63,891	13.344 (13.185,13.503)	13.59 (13.248,13.933)	13.279 (13.004,13.555)	13.27 (12.902,13.637)
65-74	88,762	13.244 (13.091,13.397)	12.529 (12.082,12.975)	13.131 (12.874,13.388)	13.026 (12.663,13.388)
75+	115,733	13.047 (12.9,13.193)	13.412 (13.089,13.736)	13.051 (12.815,13.287)	13.216 (12.904,13.528)
Race					
White	745,596	13.77 (13.719,13.821)	13.522 (13.331,13.712)	13.506 (13.372,13.64)	13.46 (13.318,13.602)
Black	43,109	14.065 (13.887,14.242)	13.632 (13.387,13.876)	13.695 (13.49,13.901)	13.662 (13.21,14.114)
Asian	26,265	14.351 (14.135,14.567)	14.281 (13.405,15.157)	14.051 (13.522,14.58)	13.841 (13.51,14.172)
Other	22,091	14.055 (13.786,14.324)	13.349 (12.959,13.739)	13.585 (13.254,13.917)	13.492 (12.695,14.289)
Ethnicity					
Non-Hisp	808,098	13.803 (13.754,13.851)	13.563 (13.359,13.767)	13.547 (13.419,13.675)	13.49 (13.362,13.617)
Hispanic	28,963	14.049 (13.81,14.288)	13.602 (13.303,13.901)	13.544 (13.278,13.81)	13.558 (13.084,14.031)
Education					
<High school	50,943	14.3 (14.177,14.424)	13.601 (13.416,13.786)	13.604 (13.394,13.814)	13.513 (13.106,13.921)
HS completed	78,045	13.895 (13.73,14.06)	13.425 (12.957,13.893)	13.509 (13.236,13.781)	13.478 (13.072,13.884)
College	237,206	14.003 (13.913,14.092)	13.641 (13.36,13.922)	13.611 (13.433,13.788)	13.532 (13.339,13.725)
Graduate	326,860	13.695 (13.617,13.774)	13.68 (13.378,13.982)	13.65 (13.5,13.799)	13.558 (13.42,13.696)
Post-grad	144,007	13.539 (13.431,13.648)	13.307 (12.878,13.737)	13.369 (13.197,13.542)	13.399 (13.122,13.677)
HH income					
0-49	332,586	13.891 (13.809,13.973)	13.62 (13.357,13.882)	13.641 (13.465,13.817)	13.612 (13.339,13.886)
50-99	309,387	13.745 (13.669,13.822)	13.573 (13.341,13.805)	13.55 (13.395,13.705)	13.469 (13.323,13.615)
100-149	132,757	13.777 (13.671,13.882)	13.383 (13.062,13.704)	13.424 (13.243,13.605)	13.415 (13.247,13.584)
150+	62,331	13.531 (13.437,13.625)	13.201 (12.949,13.454)	13.457 (13.277,13.636)	13.342 (13.14,13.544)
HH size					
1	177,140	13.649 (13.533,13.765)	13.337 (12.98,13.694)	13.518 (13.349,13.688)	13.489 (13.276,13.703)
2	286,106	13.6 (13.513,13.687)	13.462 (13.164,13.761)	13.469 (13.275,13.663)	13.383 (13.215,13.551)
3	152,684	14.02 (13.918,14.122)	13.718 (13.351,14.085)	13.58 (13.395,13.765)	13.5 (13.336,13.664)
4	143,442	14.033 (13.941,14.125)	13.514 (13.189,13.838)	13.64 (13.491,13.788)	13.542 (13.362,13.723)
5+	77,689	14.138 (14.017,14.259)	13.819 (13.433,14.206)	13.581 (13.321,13.841)	13.73 (13.474,13.985)
Urban size					
<50k	34,987	13.52 (13.266,13.773)	13.383 (12.795,13.972)	13.337 (12.845,13.829)	13.328 (13.031,13.625)
50-200k	119,970	13.928 (13.794,14.062)	13.747 (13.489,14.005)	13.842 (13.672,14.011)	13.698 (13.522,13.873)
200-500k	44,578	13.918 (13.705,14.13)	13.518 (12.933,14.103)	13.817 (13.571,14.064)	13.66 (13.276,14.045)
500-1000k	276,629	13.759 (13.678,13.84)	13.395 (13.213,13.576)	13.564 (13.45,13.679)	13.503 (13.305,13.7)
1000k+	360,897	14.286 (13.859,14.713)	13.451 (12.893,14.01)	13.654 (13.317,13.992)	13.385 (12.683,14.087)
Vehicle make					
American	290,228	13.8 (13.714,13.886)	13.339 (13.067,13.61)	13.455 (13.258,13.651)	13.426 (13.221,13.631)
Asian	528,810	13.799 (13.741,13.858)	13.646 (13.485,13.808)	13.627 (13.517,13.736)	13.561 (13.43,13.692)
European	18,023	14.337 (14.094,14.58)	14.19 (13.492,14.888)	13.684 (13.334,14.034)	13.552 (13.171,13.933)
Vehicle type					
Car	610,245	13.849 (13.792,13.906)	13.649 (13.445,13.854)	13.658 (13.53,13.787)	13.611 (13.476,13.747)
Van	27,866	13.588 (13.336,13.84)	13.414 (13.064,13.764)	13.472 (13.172,13.772)	13.514 (13.168,13.861)
SUV	158,202	13.773 (13.675,13.871)	13.623 (13.279,13.966)	13.497 (13.336,13.657)	13.528 (13.264,13.793)
Pickup	40,748	13.714 (13.536,13.893)	13.725 (13.41,14.04)	13.544 (13.305,13.783)	13.502 (13.079,13.925)
Fuel type					
Gas/D	761,292	13.841 (13.791,13.891)	13.565 (13.389,13.741)	13.556 (13.428,13.685)	13.498 (13.36,13.636)
Other	75,769	13.51 (13.338,13.683)	13.525 (13.217,13.833)	13.463 (13.254,13.672)	13.56 (13.264,13.856)
Weekend					
Weekday	712,411	13.824 (13.775,13.872)	13.576 (13.408,13.744)	13.558 (13.431,13.684)	13.502 (13.364,13.64)
Weekend	124,650	13.74 (13.685,13.794)	13.496 (13.22,13.773)	13.531 (13.397,13.664)	13.486 (13.334,13.637)

Table 3.17: Mean daily maximum speed (MPH) and associated 95% CIS by different covariates across DR adjustment methods

Covariate	n	Unweighted (95%CI)	GLM-AIPW-PAPP (95%CI)	GLM-AIPW-PMLE (95%CI)	BART–AIPW-PAPP (95%CI)
Total	837,061	59.808 (59.467,60.149)	61.547 (59.717,63.377)	60.447 (59.833,61.062)	59.947 (58.623,61.27)
Gender					
Male	407,312	60.187 (59.706,60.669)	62.687 (59.483,65.89)	60.847 (60.045,61.649)	60.677 (58.953,62.402)
Female	429,749	59.448 (58.969,59.928)	60.28 (59.195,61.366)	60.023 (59.205,60.84)	59.193 (58.034,60.353)
Age group					
16-24	311,106	61.475 (60.97,61.98)	62.484 (61.235,63.733)	62.212 (61.442,62.981)	62.078 (60.788,63.368)
25-34	117,758	63.41 (62.567,64.253)	62.907 (61.082,64.733)	62.359 (61.171,63.546)	62.373 (60.598,64.148)
35-44	61,908	62.617 (61.358,63.878)	62.761 (60.791,64.731)	63.986 (62.235,65.737)	62.039 (59.911,64.166)
45-54	77,903	60.872 (59.853,61.89)	64.943 (58.295,71.591)	60.117 (58.688,61.545)	59.738 (57.435,62.041)
55-64	63,891	59.478 (58.406,60.55)	59.666 (57.225,62.107)	59.611 (57.872,61.35)	58.797 (56.332,61.262)
65-74	88,762	55.91 (55.068,56.753)	56.915 (55.026,58.805)	55.693 (54.645,56.742)	55.5 (53.256,57.744)
75+	115,733	52.613 (51.8,53.426)	52.602 (50.493,54.71)	52.88 (51.871,53.889)	52.262 (50.92,53.604)
Race					
White	745,596	59.449 (59.092,59.806)	60.312 (59.478,61.145)	60.254 (59.581,60.929)	59.586 (58.217,60.954)
Black	43,109	64.628 (62.991,66.264)	68.229 (60.656,75.802)	62.22 (60.3,64.14)	62.152 (58.85,65.453)
Asian	26,265	61.08 (59.323,62.836)	60.573 (58.414,62.733)	59.081 (56.873,61.288)	59.464 (55.818,63.11)
Other	22,091	61.008 (59.099,62.917)	60.986 (58.585,63.387)	59.968 (58.45,61.485)	60.988 (55.744,66.231)
Ethnicity					
Non-Hisp	808,098	59.718 (59.37,60.066)	60.221 (59.395,61.048)	60.232 (59.595,60.869)	59.528 (58.485,60.571)
Hispanic	28,963	62.31 (60.707,63.914)	66.308 (60.971,71.645)	61.976 (60.418,63.533)	62.437 (58.69,66.184)
Education					
<High school	50,943	58.103 (56.954,59.251)	59.199 (57.013,61.386)	58.949 (57.325,60.572)	60.162 (57.18,63.145)
HS completed	78,045	59.865 (58.83,60.901)	61.116 (59.657,62.576)	60.812 (59.457,62.166)	60.382 (57.673,63.092)
College	237,206	59.874 (59.25,60.497)	62.623 (58.482,66.764)	59.982 (58.9,61.065)	59.491 (57.762,61.219)
Graduate	326,860	60.185 (59.62,60.751)	61.474 (60.049,62.898)	61.405 (60.379,62.43)	60.718 (59.63,61.807)
Post-grad	144,007	59.414 (58.566,60.262)	59.809 (58.019,61.598)	60.113 (58.801,61.426)	59.221 (57.561,60.881)
HH income					
0-49	332,586	59.127 (58.575,59.68)	59.271 (57.864,60.679)	59.263 (58.317,60.208)	58.757 (57.339,60.175)
50-99	309,387	60.031 (59.461,60.6)	64.22 (60.289,68.151)	60.94 (60.026,61.853)	60.508 (58.903,62.113)
100-149	132,757	60.663 (59.901,61.425)	60.409 (58.784,62.035)	61.507 (60.192,62.822)	60.611 (59.169,62.052)
150+	62,331	60.513 (59.305,61.721)	61.386 (59.529,63.244)	60.484 (58.966,62.004)	60.549 (58.951,62.147)
HH size					
1	177,140	57.902 (57.123,58.682)	58.973 (57.29,60.655)	58.243 (57.246,59.24)	57.958 (56.619,59.298)
2	286,106	59.02 (58.421,59.619)	60.033 (58.585,61.48)	59.371 (58.389,60.353)	59.371 (57.722,61.019)
3	152,684	61.35 (60.569,62.132)	60.399 (58.548,62.25)	61.373 (59.998,62.748)	60.541 (58.797,62.285)
4	143,442	61.214 (60.476,61.951)	61.592 (60.031,63.152)	61.488 (60.377,62.599)	61.068 (59.616,62.52)
5+	77,689	61.428 (60.57,62.286)	66.669 (60.605,72.733)	62.759 (60.967,64.551)	60.989 (58.922,63.057)
Urban size					
<50k	34,987	60.422 (58.928,61.917)	61.118 (58.986,63.251)	59.964 (58.006,61.921)	59.665 (57.238,62.093)
50-200k	119,970	56.12 (55.22,57.021)	57.621 (55.544,59.698)	57.162 (56.071,58.254)	57.313 (55.844,58.782)
200-500k	44,578	62.847 (61.27,64.423)	64.07 (60.75,67.39)	62.507 (60.978,64.037)	62.711 (60.338,65.086)
500-1000k	276,629	60.193 (59.62,60.766)	61.073 (57.067,65.078)	59.886 (58.928,60.846)	60.296 (58.82,61.773)
1000k+	360,897	60.303 (59.8,60.807)	62.075 (59.415,64.736)	60.647 (59.977,61.317)	60.287 (59.099,61.475)
Vehicle make					
American	290,228	59.36 (58.773,59.946)	63.24 (59.63,66.85)	60.218 (59.339,61.096)	59.877 (58.058,61.695)
Asian	528,810	60.013 (59.588,60.438)	60.796 (59.891,61.701)	60.751 (60.095,61.407)	60.203 (59.036,61.371)
European	18,023	61.016 (59.074,62.958)	58.842 (56.171,61.514)	58.984 (56.944,61.025)	59.049 (55.175,62.922)
Vehicle type					
Car	610,245	59.744 (59.338,60.149)	60.92 (60.083,61.757)	60.44 (59.765,61.115)	60.119 (58.854,61.383)
Van	27,866	57.722 (56.154,59.289)	58.36 (55.812,60.907)	58.674 (56.088,61.26)	58.263 (55.586,60.94)
SUV	158,202	60.093 (59.36,60.825)	62.613 (57.431,67.795)	60.444 (59.297,61.59)	59.511 (57.678,61.345)
Pickup	40,748	61.092 (59.557,62.627)	62.97 (61.201,64.739)	61.359 (59.346,63.371)	60.707 (57.51,63.905)
Fuel type					
Gas/D	761,292	59.878 (59.516,60.239)	61.537 (59.67,63.404)	60.473 (59.842,61.105)	59.937 (58.565,61.309)
Other	75,769	59.105 (58.131,60.079)	61.685 (58.505,64.865)	61.082 (59.975,62.189)	60.645 (58.056,63.234)
Weekend					
Weekday	712,411	59.684 (59.344,60.023)	61.322 (59.663,62.982)	60.295 (59.687,60.902)	59.801 (58.483,61.119)
Weekend	124,650	60.517 (60.151,60.883)	62.809 (60.044,65.575)	61.312 (60.601,62.023)	60.768 (59.333,62.204)

Table 3.18: Mean daily frequency of brakes per driven mile and associated 95% CIS by different covariates across DR adjustment methods

Covariate	n	Unweighted (95%CI)	GLM-AIPW-PAPP (95%CI)	GLM-AIPW-PMLE (95%CI)	BART–AIPW-PAPP (95%CI)
Total	837,061	4.499 (4.387,4.611)	4.356 (3.887,4.825)	4.644 (4.345,4.942)	4.426 (3.984,4.867)
Gender					
Male	407,312	4.415 (4.247,4.583)	3.835 (3.139,4.531)	4.456 (4.129,4.784)	4.345 (3.789,4.902)
Female	429,749	4.579 (4.43,4.728)	4.957 (4.518,5.396)	4.825 (4.471,5.179)	4.508 (4.04,4.977)
Age group					
16-24	311,106	4.283 (4.114,4.451)	4.368 (3.735,5.001)	4.417 (4.068,4.766)	4.173 (3.777,4.57)
25-34	117,758	4.085 (3.819,4.351)	4.201 (3.59,4.812)	4.609 (4.084,5.133)	3.984 (3.288,4.681)
35-44	61,908	4.422 (4.052,4.792)	4.575 (3.779,5.371)	4.643 (4.05,5.236)	4.574 (3.905,5.243)
45-54	77,903	4.14 (3.846,4.435)	3.764 (2.22,5.308)	4.174 (3.719,4.628)	3.997 (3.359,4.636)
55-64	63,891	4.565 (4.136,4.995)	5.003 (4.102,5.904)	4.589 (4.042,5.136)	4.62 (3.365,5.875)
65-74	88,762	4.801 (4.41,5.193)	3.522 (1.749,5.296)	4.799 (4.195,5.402)	4.596 (3.372,5.821)
75+	115,733	5.518 (5.147,5.888)	6.902 (5.723,8.081)	5.857 (5.25,6.464)	6.44 (5.548,7.332)
Race					
White	745,596	4.521 (4.401,4.641)	4.518 (4.018,5.018)	4.583 (4.272,4.895)	4.402 (4,4.805)
Black	43,109	4.366 (3.885,4.848)	3.807 (2.117,5.496)	5.074 (4.442,5.706)	5.053 (4.117,5.988)
Asian	26,265	4.256 (3.675,4.837)	4.349 (3.393,5.304)	5.222 (4.061,6.383)	4.483 (3.405,5.561)
Other	22,091	4.319 (3.732,4.907)	4.197 (3.011,5.382)	4.438 (3.574,5.302)	3.705 (2.167,5.243)
Ethnicity					
Non-Hisp	808,098	4.488 (4.375,4.601)	4.56 (4.117,5.003)	4.597 (4.323,4.871)	4.442 (4.051,4.832)
Hispanic	28,963	4.813 (4.037,5.588)	3.842 (2.609,5.075)	4.84 (3.782,5.898)	4.3 (2.899,5.701)
Education					
<High school	50,943	4.942 (4.522,5.362)	4.779 (3.58,5.979)	5.209 (4.526,5.893)	5.204 (3.91,6.498)
HS completed	78,045	4.163 (3.8,4.526)	3.495 (1.656,5.335)	4.241 (3.662,4.819)	4.179 (3.275,5.084)
College	237,206	4.561 (4.347,4.775)	4.466 (3.609,5.323)	4.713 (4.269,5.158)	4.604 (4.062,5.145)
Graduate	326,860	4.347 (4.165,4.528)	4.174 (3.765,4.583)	4.564 (4.201,4.927)	4.17 (3.59,4.749)
Post-grad	144,007	4.769 (4.509,5.029)	5.031 (4.514,5.548)	4.788 (4.387,5.19)	4.541 (3.857,5.224)
HH income					
0-49	332,586	4.542 (4.353,4.731)	4.639 (3.669,5.608)	4.728 (4.325,5.131)	4.752 (4.157,5.347)
50-99	309,387	4.386 (4.207,4.566)	3.75 (2.847,4.653)	4.482 (4.098,4.866)	4.196 (3.682,4.71)
100-149	132,757	4.579 (4.32,4.838)	4.8 (4.301,5.3)	4.743 (4.228,5.258)	4.564 (3.933,5.194)
150+	62,331	4.66 (4.265,5.056)	4.662 (3.942,5.383)	4.721 (4.213,5.229)	4 (3.02,4.98)
HH size					
1	177,140	4.644 (4.38,4.908)	4.13 (2.528,5.733)	4.852 (4.438,5.265)	4.658 (4.058,5.258)
2	286,106	4.674 (4.46,4.888)	4.855 (4.259,5.451)	4.782 (4.302,5.262)	4.515 (3.92,5.111)
3	152,684	4.328 (4.097,4.558)	4.425 (3.937,4.913)	4.593 (4.123,5.064)	4.321 (3.785,4.857)
4	143,442	4.294 (4.064,4.524)	4.356 (3.762,4.95)	4.475 (4.051,4.9)	4.201 (3.71,4.692)
5+	77,689	4.241 (3.949,4.533)	3.851 (2.313,5.388)	4.396 (3.893,4.899)	4.459 (4.017,4.901)
Urban size					
<50k	34,987	4.051 (3.567,4.535)	4.313 (2.858,5.768)	4.293 (3.57,5.015)	4.24 (3.746,4.734)
50-200k	119,970	4.789 (4.49,5.089)	4.921 (4.454,5.389)	4.761 (4.265,5.257)	4.696 (4.02,5.372)
200-500k	44,578	4.241 (3.825,4.657)	4.177 (3.147,5.206)	4.567 (4.075,5.059)	4.231 (3.049,5.413)
500-1000k	276,629	3.969 (3.793,4.145)	3.977 (3.342,4.612)	4.21 (3.867,4.552)	4.171 (3.549,4.793)
1000k+	360,897	4.884 (4.703,5.066)	4.539 (3.985,5.093)	4.948 (4.599,5.297)	4.626 (4.025,5.227)
Vehicle make					
American	290,228	4.762 (4.548,4.975)	4.239 (3.381,5.097)	5.007 (4.549,5.466)	4.914 (4.347,5.482)
Asian	528,810	4.392 (4.261,4.524)	4.569 (4.244,4.894)	4.451 (4.181,4.721)	4.206 (3.78,4.632)
European	18,023	3.401 (2.946,3.856)	3.161 (2.359,3.963)	3.664 (3.006,4.322)	2.898 (0.958,4.837)
Vehicle type					
Car	610,245	4.504 (4.368,4.641)	4.281 (3.65,4.913)	4.564 (4.225,4.903)	4.248 (3.785,4.711)
Van	27,866	4.435 (4.064,4.806)	5.298 (4.287,6.31)	4.855 (4.41,5.3)	4.569 (3.345,5.792)
SUV	158,202	4.351 (4.148,4.555)	4.222 (3.078,5.365)	4.56 (4.215,4.904)	4.381 (3.73,5.032)
Pickup	40,748	5.043 (4.391,5.696)	4.611 (3.881,5.34)	5.022 (4.255,5.789)	5.226 (4.061,6.391)
Fuel type					
Gas/D	761,292	4.435 (4.319,4.551)	4.345 (3.865,4.825)	4.649 (4.34,4.959)	4.426 (3.992,4.859)
Other	75,769	5.145 (4.737,5.553)	4.718 (3.688,5.747)	4.934 (4.308,5.561)	4.388 (3.347,5.429)
Weekend					
Weekday	712,411	4.492 (4.379,4.605)	4.365 (3.91,4.82)	4.639 (4.339,4.938)	4.413 (3.963,4.864)
Weekend	124,650	4.54 (4.427,4.654)	4.305 (3.738,4.871)	4.675 (4.374,4.975)	4.497 (4.073,4.921)

Table 3.19: Mean daily percentage of stop time and associated 95% CIS by different covariates across DR adjustment methods

Covariate	n	Unweighted (95%CI)	GLM-AIPW-PAPP (95%CI)	GLM-AIPW-PMLE (95%CI)	BART–AIPW-PAPP (95%CI)
Total	837,061	25.518 (25.202,25.834)	25.515 (24.043,26.987)	24.949 (24.217,25.681)	0.251 (0.242,0.26)
Gender					
Male	407,312	24.618 (24.157,25.079)	24.048 (21.863,26.234)	24.06 (23.158,24.961)	0.242 (0.231,0.252)
Female	429,749	26.371 (25.945,26.797)	27.107 (25.226,28.988)	25.873 (24.968,26.779)	0.261 (0.25,0.271)
Age group					
16-24	311,106	26.713 (26.221,27.204)	26.551 (25.177,27.925)	25.913 (25.109,26.716)	0.258 (0.245,0.271)
25-34	117,758	25.199 (24.385,26.014)	24.178 (22.653,25.704)	25.013 (23.946,26.08)	0.247 (0.232,0.262)
35-44	61,908	25.575 (24.528,26.621)	27.6 (23.466,31.735)	26.828 (25.017,28.639)	0.265 (0.247,0.284)
45-54	77,903	23.406 (22.257,24.555)	24.908 (20.144,29.672)	22.926 (21.57,24.281)	0.239 (0.22,0.258)
55-64	63,891	22.879 (21.906,23.852)	22.949 (21.035,24.862)	23.408 (21.72,25.095)	0.235 (0.211,0.259)
65-74	88,762	24.425 (23.448,25.402)	27.099 (23.644,30.554)	24.739 (23.395,26.084)	0.262 (0.246,0.279)
75+	115,733	26.315 (25.367,27.264)	27.682 (25.755,29.609)	26.185 (25.039,27.332)	0.28 (0.264,0.296)
Race					
White	745,596	25.216 (24.882,25.55)	25.693 (24.172,27.213)	24.071 (23.292,24.849)	0.245 (0.237,0.253)
Black	43,109	29.711 (28.421,31.001)	27.666 (25.29,30.042)	29.697 (28.071,31.324)	0.294 (0.267,0.32)
Asian	26,265	25.989 (24.582,27.396)	23.631 (20.325,26.937)	26.098 (24.013,28.184)	0.252 (0.216,0.289)
Other	22,091	26.955 (24.952,28.958)	26.442 (23.616,29.269)	26.484 (23.923,29.045)	0.252 (0.21,0.294)
Ethnicity					
Non-Hisp	808,098	25.438 (25.118,25.758)	25.622 (24.061,27.183)	24.532 (23.79,25.274)	0.251 (0.242,0.26)
Hispanic	28,963	27.746 (25.946,29.546)	26.345 (24.033,28.657)	27.102 (25.369,28.836)	0.251 (0.224,0.277)
Education					
<High school	50,943	27.86 (26.587,29.134)	28.96 (26.302,31.618)	27.223 (25.326,29.119)	0.262 (0.233,0.292)
HS completed	78,045	26.136 (25.022,27.249)	28.155 (25.07,31.241)	25.534 (24.179,26.889)	0.263 (0.24,0.287)
College	237,206	26.881 (26.288,27.474)	27.043 (24.085,30.001)	26.128 (24.88,27.377)	0.264 (0.253,0.276)
Graduate	326,860	24.96 (24.472,25.448)	22.543 (21.102,23.983)	23.625 (22.803,24.447)	0.236 (0.218,0.254)
Post-grad	144,007	23.375 (22.656,24.094)	24.105 (22.558,25.651)	23.845 (22.697,24.992)	0.237 (0.219,0.254)
HH income					
0-49	332,586	26.578 (26.059,27.098)	27.376 (25.379,29.374)	26.485 (25.414,27.557)	0.265 (0.254,0.276)
50-99	309,387	25.205 (24.717,25.694)	24.47 (21.599,27.341)	24.537 (23.514,25.559)	0.246 (0.232,0.26)
100-149	132,757	24.218 (23.441,24.996)	24.214 (22.662,25.766)	23.707 (22.615,24.799)	0.241 (0.226,0.256)
150+	62,331	24.176 (22.962,25.39)	25.297 (20.664,29.931)	23.96 (22.384,25.536)	0.243 (0.223,0.264)
HH size					
1	177,140	26.133 (25.442,26.823)	26.166 (22.88,29.452)	25.409 (24.247,26.571)	0.257 (0.247,0.266)
2	286,106	24.412 (23.871,24.953)	24.289 (23.042,25.537)	23.693 (22.808,24.578)	0.244 (0.234,0.254)
3	152,684	25.507 (24.748,26.267)	24.228 (21.958,26.498)	25.177 (23.648,26.705)	0.243 (0.231,0.256)
4	143,442	26.236 (25.522,26.951)	26.843 (23.645,30.042)	25.755 (24.608,26.901)	0.259 (0.244,0.273)
5+	77,689	26.882 (25.884,27.88)	27.356 (22.56,32.152)	25.719 (24.458,26.98)	0.263 (0.244,0.282)
Urban size					
<50k	34,987	20.874 (19.097,22.651)	26.022 (21.85,30.194)	21.679 (19.496,23.862)	0.228 (0.21,0.247)
50-200k	119,970	23.798 (22.902,24.694)	23.87 (22.364,25.376)	24.287 (22.881,25.692)	0.239 (0.222,0.257)
200-500k	44,578	22.435 (21.355,23.515)	24.487 (18.513,30.461)	23.436 (22.035,24.837)	0.236 (0.209,0.263)
500-1000k	276,629	25.334 (24.785,25.882)	25.695 (24.531,26.86)	26.128 (25.326,26.93)	0.263 (0.249,0.278)
1000k+	360,897	27.062 (26.615,27.508)	26.883 (25.813,27.953)	27.43 (26.73,28.131)	0.271 (0.26,0.283)
Vehicle make					
American	290,228	26.28 (25.728,26.831)	24.962 (22.25,27.673)	24.957 (23.906,26.007)	0.255 (0.244,0.265)
Asian	528,810	25.075 (24.682,25.468)	26.283 (24.65,27.917)	24.94 (24.125,25.755)	0.249 (0.239,0.258)
European	18,023	26.233 (24.82,27.646)	23.294 (19.732,26.857)	25.298 (22.99,27.607)	0.243 (0.21,0.276)
Vehicle type					
Car	610,245	25.632 (25.267,25.997)	25.738 (24.14,27.335)	25.054 (24.321,25.788)	0.25 (0.24,0.261)
Van	27,866	27.205 (25.523,28.887)	28.585 (24.943,32.227)	28.5 (25.898,31.103)	0.277 (0.237,0.316)
SUV	158,202	25.274 (24.518,26.031)	25.441 (22.085,28.796)	25.053 (23.917,26.189)	0.254 (0.241,0.267)
Pickup	40,748	23.596 (22.247,24.945)	23.906 (20.704,27.107)	23.839 (21.462,26.217)	0.239 (0.211,0.267)
Fuel type					
Gas/D	761,292	25.777 (25.444,26.11)	25.638 (24.128,27.147)	25.059 (24.318,25.801)	0.252 (0.243,0.261)
Other	75,769	22.913 (21.982,23.844)	20.192 (18.427,21.956)	22.057 (20.348,23.765)	0.216 (0.195,0.237)
Weekend					
Weekday	712,411	25.395 (25.079,25.712)	25.465 (24.053,26.876)	24.83 (24.105,25.554)	0.25 (0.241,0.259)
Weekend	124,650	26.218 (25.892,26.545)	25.812 (23.822,27.803)	25.623 (24.805,26.442)	0.258 (0.248,0.268)

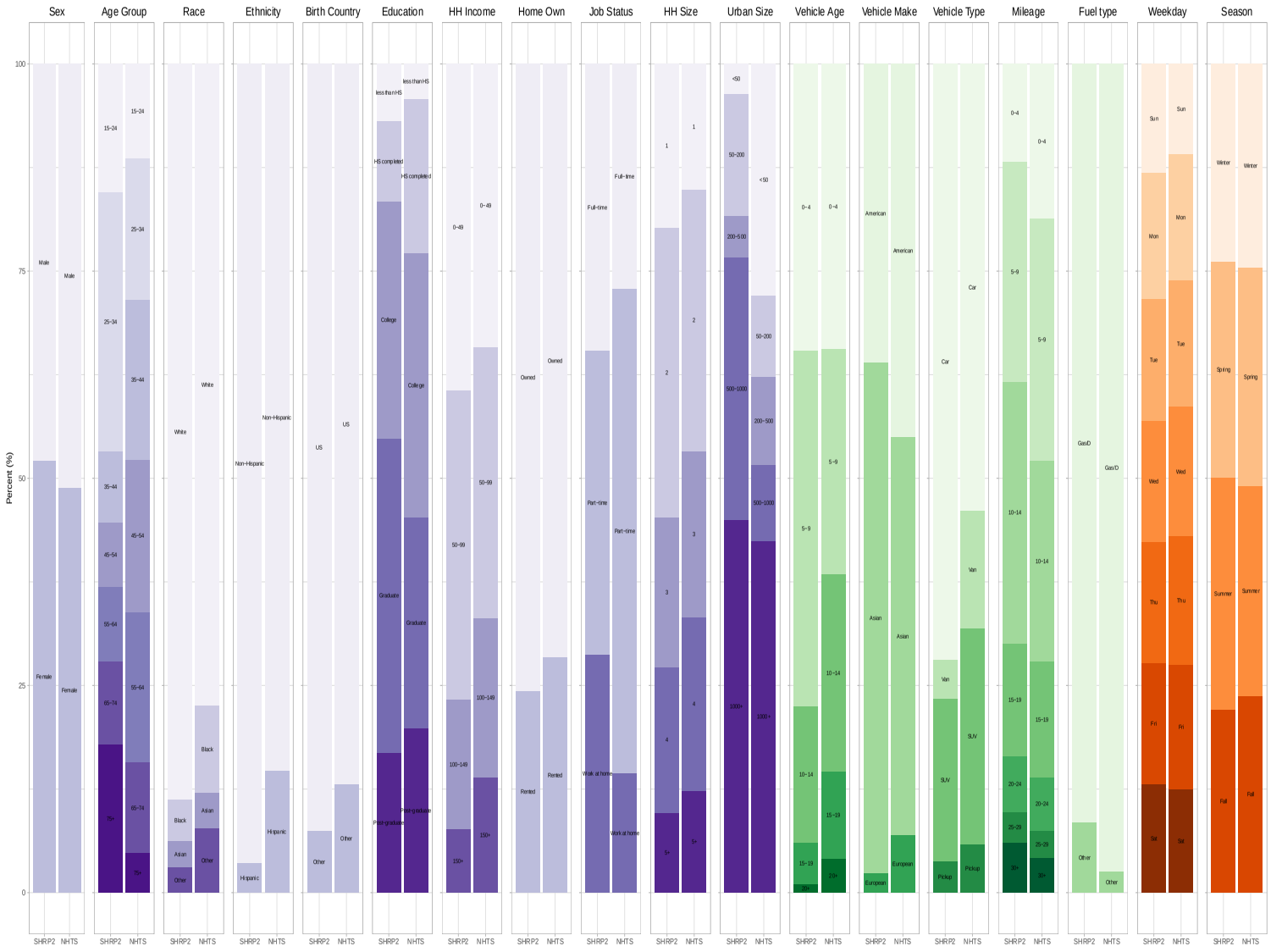

Figure 3.11: Comparing the distribution of common auxiliary variables in SHRP2 with weighted NHTS

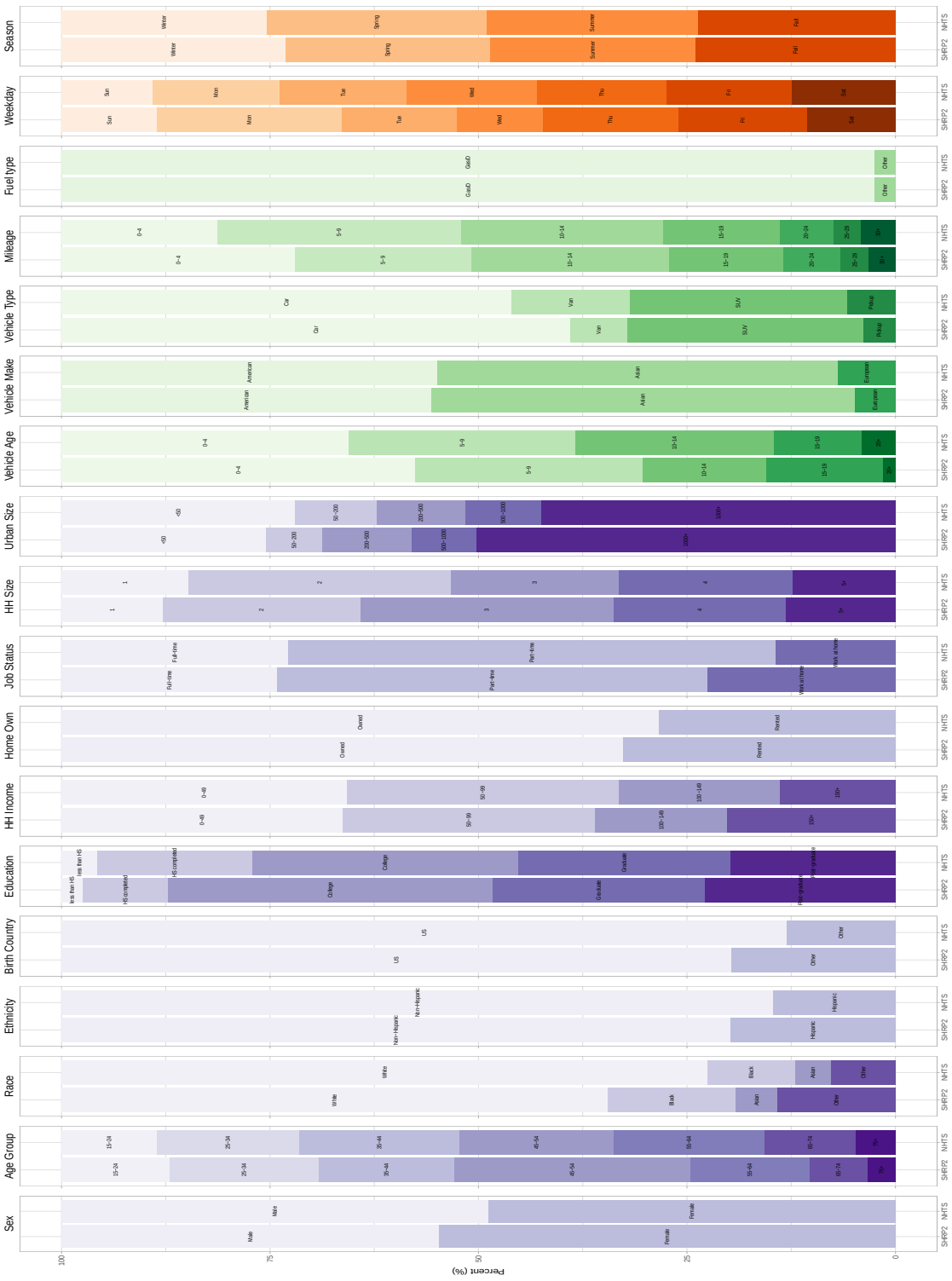

Figure 3.12: Comparing the distribution of common auxiliary variables in pseudo-weighted SHRP2 (PAPP–BART) with weighted NHTS

CHAPTER IV

Robust fully Bayesian Inference for Non-probability Samples

4.1 Introduction

The previous chapters proposed alternative approaches for robust inference in non-probability samples when there exists a reference probability survey with a common set of relevant auxiliary variables. For this setup, my literature review revealed that two distinct approaches can be chosen under a *strongly ignorable* condition: (1) *quasi-randomization* (QR)–estimating the probabilities of being included in the non-probability sample, also known as propensity scores (PS), while treating the non-probability sample as if randomly selected (Lee, 2006; Lee and Valliant, 2009; Valliant and Dever, 2011), and (2) *prediction modeling* (PM)–fitting models on the non-probability sample to predict the response variable for units in the reference survey (Rivers, 2007; Kim and Rao, 2012; Wang et al., 2015; Kim et al., 2021a). In either case, design-based approaches can then be utilized to compute point and interval estimates.

Since both ideas rely on imputation, it was demonstrated that correct specification of the underlying models is essential, especially when extrapolation is inevitable as in PM (Lenis et al., 2018). Chen et al. (2019) reconciles the QR notion with that of

PM using the idea of augmented inverse propensity weighting (AIPW) (Robins et al., 1994). This method is doubly robust (DR) in the sense that the sample estimator is consistent if either the QR or PM model holds (Scharfstein et al., 1999). Chapter III further expanded this approach to eliminate the need for a pseudo-likelihood structure when estimating the propensity scores (PS). This permitted us to exploit Bayesian modeling for prediction. However, our proposed method followed a two-step Bayesian idea which conceptually separates the design stage, i.e. estimating the PS and outcome variable(s) from the analysis stage, i.e. estimating the population quantity conditional on the imputed PS and outcome variable(s) (Kaplan and Chen, 2012; Zigler, 2016).

The proposed AIPW estimator in Chapter III, however, had a design-based structure as both QR- and PM-related terms contained a (pseudo-)weighted summation. Although an HT-type estimator is known to be design-consistent, in the presence of influential (pseudo-)weights, it becomes highly unstable with an inflated variance (Zhang and Little, 2011; Zangeneh, 2012; Chen et al., 2017a). Especially when estimating the PS, it is always likely that data suffer from a partial lack of the *positivity* assumption, which leads to extremely large estimated pseudo-weights (Stuart, 2010). Furthermore, design-based approaches lack a unified framework for quantifying the uncertainty in the point estimates (Zangeneh, 2012). Frequentist methods for variance approximation come with sophisticated theories that often hold asymptotically, especially when there are multiple sources of uncertainty (Chen et al., 2019; Kim et al., 2021a). The two-step Bayesian approach also suffers from a failure to accurately propagate the uncertainty of derived estimators (Zigler, 2016).

To minimize these limitations, Zheng and Little (2003) propose an alternative class of inferential methods for probability surveys with a probability proportional-to-size (PPS) design, which they term Penalized Spline of Propensity Prediction (PSPP). Unlike the previously discussed methods, PSPP is a fully model-based approach,

predicting the outcome variable for all non-sampled units of the population. This method borrows the idea of Linear-in-weight Prediction (LWP), in which estimated pseudo-weights are specified as a predictor in the outcome model (Zhang and Little, 2009). Scharfstein et al. (1999) and Bang and Robins (2005) demonstrate that an LWP estimator behaves equivalent to an AIPW estimator in terms of double robustness. In situations where auxiliary variables are missing for the non-sampled units, Little and Zheng (2007) recommend synthesizing the population repeatedly via finite population Bayesian bootstrapping (FPBB). An and Little (2008) extend this approach to an item-level missing data imputation context where measures of size are replaced by the estimated PS of being observed, and demonstrate its DR property in a simulation study.

PSPP is fully Bayesian allowing for direct estimation of the variance by simulating the posterior predictive distribution of the population parameters. Zangeneh and Little (2015) expand the Bayesian PSPP under a PPS design for situations where the totals of the measures of size are known from external data and where there is evidence of heteroscedasticity with respect to the estimated PS. Further extensions to probability samples with unequal selection probabilities are proposed by Chen et al. (2012). The PSPP is also suitable for situations where the design of the reference sample is complex. Zhou et al. (2016) develop a synthetic population approach based on a multi-stage cluster sample by undoing the sampling steps through a weighted Pólya posterior distribution. Recently, Tan et al. (2019) and Mercer (2018) have compared the PSPP with AIPW to make inference for incomplete data and non-probability samples, respectively, where PS are predicted using Bayesian Additive Regression Trees (BART). The authors found that the former outperforms in terms of the mean square error of the adjusted estimator.

While the use of a more flexible non-parametric function of the estimated PS may improve the efficiency of the adjusted estimator if the PM is misspecified, and

reduce the risk of model misspecification when influential pseudo-weights are present (Zhang and Little, 2011), the theoretical rationale for using a penalized spline model among a wider class of smoothers is not quite clear. Saarela et al. (2016) argue that the convergence of the posterior sampling to any well-defined joint distribution of the outcome and PS may be hard to achieve. Alternatively, one can use Gaussian Process (GP) priors to link the PS to the outcome conditional mean (Si et al., 2015). GP is a powerful non-parametric Bayesian tool for functional regression that assigns prior distributions over multidimensional non-linear functions. Because of its flexibility and generalizability, GP is gaining popularity in statistics and machine learning (Neal, 1997; Oakley and O'Hagan, 2004; Williams and Rasmussen, 2006; Kaufman et al., 2010; Yi et al., 2011; Shi and Choi, 2011; Wang and Xu, 2019).

While the correspondence between splines and GP has long been understood (Kimeldorf and Wahba, 1970; Seeger, 2000), the latter can exploit a kernel with infinite basis functions (Williams and Rasmussen, 2006). In this regard, GP may outdo the spline in terms of flexibility while depending on no arbitrary tuning parameters. In a regular spline regression, one has to determine the polynomial order as well as the frequency and location of the knots empirically. More importantly, Huang et al. (2019) demonstrate that a stationary isotropic covariance matrix in GP behaves as a non-parametric matching technique using the estimated PS as a measure of similarity. In a more *ad hoc* manner, Rivers (2007) suggests matching units of a web non-probability survey to those from a parallel reference survey. Very recently, a kernel weighting approach has been proposed by Wang et al. (2020a,b), where the weighted estimator is proved to be consistent under a weak exchangeability condition. To further weaken the modeling assumptions, Kern et al. (2020) propose to use algorithmic tree-based methods, including random forests and gradient tree boosting, for estimating the PS in kernel weighting.

Accounting for the sampling weights of the reference survey is a big hurdle in

Bayesian modeling (Gelman et al., 2007). To circumvent this issue, one possible solution is to multiply generate synthetic populations as the first step of adjustments (Dong et al., 2014; Zangeneh and Little, 2015; An and Little, 2008). However, such a method can be computationally demanding, if not impossible, when the finite population is very large (Savitsky et al., 2016; Mercer, 2018). In the present chapter, I propose an alternative robust Bayesian approach for inference in non-probability samples that models the joint distribution of the PS and outcome using a partially linear GP regression model. I call this approach a "Gaussian Process of Propensity Prediction" (GPPP). While limiting the computations to the combined non-probability and probability samples, our method links the estimated PS to the response surface non-parametrically. Therefore, the ultimate GPPP estimator can be efficient not only computationally but also with respect to variance.

As the motivating application, the present chapter aims at assessing the crash rate per distance unit driven for a subpopulation of American drivers. The current estimates are based on a ratio of the annual total police-reported crashes and annual total miles driven obtained from the General Estimates System (GES) (Administration et al., 2014) and the American Driving Survey (ADS), respectively (Kim et al., 2019; Tefft, 2017). The denominator, however, can be widely subject to measurement error as it relies on respondents' self-reported annual miles driven and often come with high item-level missing rates. In contrast, naturalistic driving studies (NDS) offer a powerful platform for capturing both of these quantities objectively by continuously monitoring traffic incidents as well as kinematic measures in their participants via a series of in-vehicle sensors and cameras (Guo et al., 2009), including miles driven. However, as discussed in Chapter III, the high administrative and technical costs of NDS force the investigators to select a volunteer sample from a limited geographical area. Therefore, inference based on such non-probabilistic samples may suffer from selection bias (Antin et al., 2015; Rafei et al., 2021). I revisit the combined

data from Strategic Highway Research Program 2 (SHPR2) and National Household Travel Survey (NHTS) used in Chapter III to address this problem.

The rest of the chapter is organized as follows: I describe the proposed method formally in Section 4.2. Section 4.3 assesses the repeated sampling properties of the proposed method and compares its performance with the LWP, AIPW, and other competing methods through two simulation studies. In Section 4.4 I describe the datasets and variables utilized in the two empirical applications of this study as well as the results after bias adjustment. Finally, Section 4.5 reviews the strengths and weaknesses of the study in more detail and suggests some future research directions. Supplemental information, including proofs, additional theory, and preliminary descriptive results, is provided in Appendix 4.6.

4.2 Methods

4.2.1 Bayesian model-based inference

I adopt the notation and conditions **C1-C4** in Section 1.2 of Chapter I. The outcome, Y, is imputed for the non-sampled units of the population with respect to S_A, i.e. $\bar{S_A} = U - S_A$. This yields a *prediction* estimator of the population mean,

$$\begin{aligned} \hat{\bar{y}}_U &= \left(\sum_{i \in S_A} y_i + \sum_{i \in \bar{S}_A} \hat{y}_i \right) /N \\ &= \left(\sum_{i \in S_A} (y_i - \hat{y}_i) + \hat{y}_U \right) /N \end{aligned} \tag{4.1}$$

where $\hat{y}_i$ is the prediction of y_i for $i \in U$, and $\hat{y}_U = \sum_{i \in U} \hat{y}_i$. Eq. 4.1 is also known as a "generalized difference estimator" (Wu and Sitter, 2001), and is more efficient than $\hat{y}_U/N$ when n_A is large.

A fully Bayesian approach specifies a model for the joint distribution of (y_i, δ_i^A)

as

$$p(y_i, \delta_i^A | x_i, d_i; \theta, \beta) = p(y_i | x_i, d_i, \delta_i^A; \theta) p(\delta_i^A | x_i; \beta), \quad i \in U \tag{4.2}$$

I denote te se of values of a variable, x, for units in U, S_A, S_R, or S_C I denote them by x_U, x_A, x_R, or x_C, respectively. Then, the likelihood of (θ, β) is

$$L(\beta, \theta | y_A, \delta_U^A, x_U, d_U) \propto p(y_A, \delta_U^A | x_U, d_U, \theta, \beta) \tag{4.3}$$

Under a Bayesian approach, the model parameters are assigned prior distributions $p(\theta, \beta | x_U, d_U)$, and analytical inference is drawn based on the posterior distribution as below:

$$p(\beta, \theta | y_A, \delta_U^A, x_U, d_U) \propto p(\theta, \beta | x_U, d_U) L(\beta, \theta | y_A, \delta_U^A, x_U, d_U) \tag{4.4}$$

Note that in a Bayesian setting, it is essential to specify independent priors, i.e. $p(\theta, \beta | x_U, d_U) = p(\theta | x_U, d_U) p(\beta | x_U, d_U)$ to preserve the ignorable assumption, i.e. **C2**, in S_A (Little and Zheng, 2007). Descriptive inference about $\bar{y}_U$ requires deriving the posterior predictive distribution conditional on the observed data, which is given by

$$p(\bar{y}_U | y_A, \delta_U^A, x_U, d_U) = \int \int p(\bar{y}_U | y_A, \delta_U^A, x_U, d_U, \theta, \beta) p(\theta, \beta | y_A, \delta_U^A, x_U, d_U) d\theta d\beta \tag{4.5}$$

For a non-conjugate model, where the posterior predictive distribution of $\bar{y}_U$ lacks a closed-form formula, one can simulate it via an appropriate MCMC algorithm.

Estimating $\hat{y}_U$ in Eq. 4.1 requires (X, D) to be observed for the entire population, but the measurement of auxiliary information is often confined to the pooled sample, S_C. One way to tackle this issue is to generate a finite set of synthetic populations, say M, non-parametrically through finite population Bayesian bootstrapping (FPBB) (Little and Zheng, 2007; Dong et al., 2014). The outcome variable is then imputed for each synthetic population non-sampled units. However, when N is large, this is

computationally expensive, if not infeasible.

The problem becomes even more serious when the joint estimation of the model parameters for QR and PM is of interest, and a custom posterior sampler, like Metropolis–Hastings algorithm, is needed (Mercer, 2018; Savitsky et al., 2016). Furthermore, the two-step algorithm proposed by Zangeneh and Little (2015) may not be fully implementable on the existing Bayesian platforms such as Stan (Carpenter et al., 2017), and therefore, Zangeneh and Little proposed to combine the estimates across synthetic populations through Rubin's combining rules (Rubin, 1976). This may not be ideal when the posterior predictive distribution of the target population quantity tends to be highly skewed, because a symmetric confidence interval will not approximate the credible intervals of the posterior predictive distribution well.

4.2.2 Proposed computationally tractable method

As stated in Section 1.2 of Chapter I, the selection probabilities in S_R can be thought as the reciprocal of the sampling weights, i.e. $\pi_i^R \propto 1/w_i^R$. Although probability surveys typically come with a set of sampling weights in their public-use dataset, all the information used for the construction of weights is not necessarily provided to the analyst. In addition, public-use survey data may lack a detailed guideline on how the sampling weights have been calculated. To simplify the problem in these situations, Si et al. (2015) assume that weights with identical values represent a unique post-stratum in the population. Therefore, one can define d as the indicator of J unique post-strata in U, and consider $w_j^R \propto N_j/n_j^R$ $(j = 1, 2, ..., J)$, where N_j and n_j^R are the j-th post-stratum size in U and S_R, respectively. For instance, in RDD telephone surveys or mail surveys, whose design involves equiprobability sampling, the inequality in weights may arise exclusively from non-response adjustment and post-stratification.

In order to directly simulate the posterior predictive distribution of $\bar{y}_U$ via a unified

algorithm that is implementable in Stan, I limit the imputation of the outcome, y_i, to units of the combined sample, i.e. $i \in S_C$. Note that it is only $\hat{y}_U$ in Eq. 4.1 that is defined across all units of U. I use the following estimator, as defined by Si et al. (2015), to multiply impute this quantity, M times, as below:

$$\begin{aligned} \hat{y}_U^{(m)} &= \sum_{j=1}^{J} \hat{N}_j^{(m)} \hat{\bar{y}}_j^{(m)} \\ &= \sum_{j=1}^{J} \frac{\hat{N}_j^{(m)}}{n_j^R} \sum_{i=1}^{n_j^R} \hat{y}_{j[i]}^{(m)} \end{aligned} \tag{4.6}$$

where $\left[\hat{N}_j^{(m)}, \hat{\bar{y}}_j^{(m)}\right]$ is the m-th draw of the joint posterior predictive distribution of the j-th post-stratum size and mean outcome. Therefore, the m-th posterior predictive draw of $\bar{y}_U$ is given by

$$\hat{\bar{y}}_U^{(m)} = \left(\sum_{i=1}^{n_A} \left(y_i - \hat{y}_i^{(m)} \right) + \hat{y}_U^{(m)} \right) / N \tag{4.7}$$

To obtain the joint posterior predictive distribution of $\left[\hat{N}_j, \hat{\bar{y}}_j\right]$ while keeping the computations confined to S_C, I propose the following model:

$$p(y_A, \delta_C^A, w^R, n^R | x_C, d_C, \theta, \beta, \xi) = p(n^R | w^R, \xi^R) p(y_A, \delta_C^A, w_R^R | x_C, d_C, \theta, \beta) \tag{4.8}$$

where $n^R = [n_1^R, n_2^R, ..., n_J^R]^T$ and $w^R = [w_1^R, w_2^R, ..., w_J^R]^T$ are the sizes of post-strata and associated weights in S_R, respectively, and ξ^R is a J-dimensional vector of parameters associated with modeling of $n_R | w_R$. Note that, while the w_j^R are fixed by design, I only observe them in the sampled data; hence I need to account for their uncertainty in the development of the full posterior distribution. While I thoroughly

discuss each component of Eq. 4.8 later, one can derive the final estimate of $\bar{y}_U$ by

$$\hat{\bar{y}}_U = \frac{1}{M} \sum_{m=1}^{M} \hat{\bar{y}}_U^{(m)} \tag{4.9}$$

and the associated $100(1-\alpha)\%$ credible interval can be constructed by sorting $(\hat{\bar{y}}_U^{(1)}, \hat{\bar{y}}_U^{(2)}, ..., \hat{\bar{y}}_U^{(M)})$ ascendingly, and finding the $\alpha/2$ and $1-\alpha/2$ percentiles of this ordered sequence that correspond to lower and upper limits of the credible interval, respectively.

4.2.2.1 Finite population Bayesian bootstrapping for modeling $p(n^R|w^R, \xi^R)$

I begin by modeling $p(n^R|w^R, \xi^R)$ non-parametrically via Bayesian bootstrapping (BB), with the aim to simulate the posterior predictive distribution of the N_j's. The idea of BB operates quite similar to the regular bootstrap approach (Efron, 1981), except for the fact that BB simulates the posterior predictive distribution of a given population parameter instead of the sample distribution of the statistic estimating that parameter (Rubin, 1981). In a finite population Bayesian bootstrap (FPBB) setting, the goal is to derive the posterior predictive distribution of the post-strata sizes for the non-sampled population units, i.e. $\bar{S}_R$. Although FPBB imposes no parametric assumptions, it is assumed that all the existing post-strata in U are limited to those observed in the collected sample (exchangeability).

Under a simple random sample, Ghosh and Meeden (1983) propose to use a Polýa Urn Scheme, in which a Dirichlet-multinomial conjugate model is considered to expand the sample to the population. Cohen (1997) generalizes this approach to a weighted sample with independent draws, and the attributed Polýa posterior distribution for the non-sampled units of U given the observed sampling weights is formulated by Dong et al. (2014). Little and Zheng (2007) propose a modified FPBB method to generate synthetic populations based on the samples with a PPS design. Further

extension based on a constrained BB is provided by Zangeneh and Little (2015) for situations where totals are known for auxiliary variables at the population level.

In the present chapter, I modify the FPBB method proposed by Little and Zheng (2007) by letting $v^R = \{v_1^R, v_2^R, ..., v_J^R\}$ represent the set of J distinct values of the sampling weights in S_R, and $\xi^R = \{\xi_1^R, \xi_2^R, ..., \xi_J^R\}$ denote the vector of conditional probabilities that $p(w^R = v_j^R|\delta^R = 1) = \xi_j^R$ for $j = 1, 2, ..., J$, where $\sum_{j=1}^J \xi_j^R = 1$. Now, suppose n_j^R and r_j^R are the frequencies of w^R taking the value v_j^R in S_R and $\bar{S}_R$, respectively, for $j = 1, 2, ..., J$. It is clear that $\sum_{j=1}^K n_j^R = n_R$, and $\sum_{j=1}^K r_j^R = N - n_R$. Considering a Dirichlet prior, i.e. $\xi^R \sim Dirichlet(\alpha_{J\times 1})$, $\alpha \in \mathbb{R}^{J>0}$, with a multinomial likelihood function of $p(n_1^R, n_2^R, ..., n_J^R|\xi) \propto \prod_{j=1}^J (\xi_j^R)^{n_j^R}$, the posterior distribution of ξ^R is given by $(\xi^R|n_1^R, n_2^R, ..., n_J^R) \sim Dirichlet(n_1^R + \alpha_1 - 1, n_2^R + \alpha_2 - 1, ..., n_J^R + \alpha_J - 1)$. Using Bayes' rule, Little and Zheng (2007) show that

$$\begin{aligned}
\xi_j^{\bar{R}} &= p(w_i^R = v_j^R|\delta_i^R = 0) \\
&= p(\delta_i^R = 0|w_i^R = v_j^R)\frac{p(w_i^R = v_j^R)}{p(\delta_i^R = 0)} \\
&= p(\delta_i^R = 0|w_i^R = v_j^R)\frac{p(w_i^R = v_j^R|\delta_i^R = 0)p(\delta_i^R = 0) + p(w_i^R = v_j^R|\delta_i^R = 1)p(\delta_i^R = 1)}{p(\delta_i^R = 0)} \\
&= p(\delta_i^R = 0|w_i^R = v_j^R)\left\{\xi^{\bar{R}} + \xi^R\frac{p(\delta_i^R = 1)}{p(\delta_i^R = 0)}\right\}
\end{aligned} \tag{4.10}$$

Since $p(\delta_i^R = 0|w_i^R = v_j^R) = 1 - \pi_j^R$, and $p(\delta_i^R = 1)/p(\delta_i^R = 0)$ can be treated as a normalizing constant,

$$\xi_j^{\bar{R}} \propto \xi_j^R\frac{1 - \pi_j^R}{\pi_j^R} \tag{4.11}$$

After normalizing $\xi^{\bar{R}}$ such that $\sum_{j=1}^J \xi_j^{\bar{R}} = 1$, the posterior predictive distribution of r^R is given by

$$p(r_1^R, r_2^R, ..., r_J^R|n_1^R, n_2^R, ..., n_J^R, \xi^R) = \binom{N - n_R}{r_1, r_2, ..., r_J}\prod_{j=1}^J \left[c\xi_j^R(1 - \pi_j^R)/\pi_j^R\right]^{r_j^R} \tag{4.12}$$

where c is the normalizing constant. The m-th posterior predictive draw of the size of post-stratum j in the population is $N_j^{(m)} = n_j^R + r_j^{R(m)}$, $(m = 1, 2, ..., M)$.

4.2.2.2 Modeling the joint distribution of (y_i, δ_i^A) given the combined sample

The goal of PM in this study is to model $p(y_i|x_i;\theta)$ in order to obtain the posterior predictive distribution of y_i for $i \in S_R$, i.e. $p(y_R|y_A, x_C) \propto \int p(y_R|y_A, x_C;\theta)p(\theta|y_A, x_C)d\theta$. Although θ is a parameter defined in U, the ignorable assumption guarantees a consistent estimate of θ by fitting $p(y|x;\theta)$ on S_A, because

$$\begin{aligned} p(y_A|x_A;\theta) &= p(y_U|\delta_U^A = 1, x_U, d_U;\theta) \\ &= \frac{p(\delta_U^A = 1|y_U, x_U;\theta)}{p(\delta_U^A = 1|x_U;\theta)} p(y_U|x_U, d_U;\theta) \\ &= p(y_U|x_U, d_U;\theta) \end{aligned} \tag{4.13}$$

If π_i^A was known for $i \in S_C$, one could augment the PM by incorporating π_i^A as a predictor into the PM, e.g. $p(y_i|x_i, f(\pi_i^A);\theta)$. A robust estimator is achieved by choosing a flexible $f(.)$, as detailed later.

While a non-probability sample is characterized by its unknown selection mechanism, given the conditions **C1-C4**, π_i^A can be estimated by modeling $p(\delta_U^A|x_U;\beta)$. Assuming that S_A is selected by a Poisson sampling, one can formulate the likelihood of β given δ_U^A as:

$$L(\beta|\delta_U^A, x_U) = \prod_{i=1}^{N} p(\delta_i^A = 1|x_i, \beta)^{\delta_i^A} \left[1 - p(\delta_i^A = 1|x_i, \beta)\right]^{1-\delta_i^A} \tag{4.14}$$

Under a logistic regression model,

$$\pi_i^A = p(\delta_i^A = 1|x_i;\beta) = \frac{exp\{x_i^T\beta\}}{1 + exp\{x_i^T\beta\}} \tag{4.15}$$

By assigning appropriate prior distributions to β, one can simulate the posterior distribution of π_i^A for $i \in U$ through an MCMC algorithm.

One major issue with Eq. 4.14 is that the observed (δ_i^A, x_i) is restricted to S_C. Although there exist several approaches restricting the estimation of β to S_C (Valliant et al., 2018; Elliott and Valliant, 2017; Chen et al., 2019; Wang et al., 2020c), the majority rely on a pseudo-maximum likelihood estimation (PMLE) idea to account for unequal w_i^R's, which necessitates solving a set of estimating equations. A corresponding method in a Bayesian setting is called pseudo-Bayesian. While such a method guarantees consistency in point estimates, the uncertainty tends to be underestimated in the posterior distribution of parameters (Savitsky et al., 2016; Gunawan et al., 2020; Williams and Savitsky, 2021). To avoid this problem, I employ a two-step pseudo-weighting approach proposed by Elliott and Valliant (2017). Assuming that $p(\delta_i^A + \delta_i^R = 2) \approx 0$, i.e. S_A and S_R have no overlap, one can show that

$$p(\delta_U^A = 1|x_U; \beta) = p(\delta_U^R = 1|x_U; \gamma)\frac{p(\delta_C^A = 1|x_C; \phi)}{1 - p(\delta_C^A = 1|x_C; \phi)} \tag{4.16}$$

where $\beta = (\gamma, \phi)^T$ is the associated model parameters. Rafei et al. (2021) call this approach propensity-adjusted probability prediction (PAPP) and prove the asymptotic properties of a pseudo-weighted estimate based on this method including consistency and variance estimation. As can be seen, this approach reduces the modeling of $p(\delta_U^A = 1|x_U)$ to the modeling of $p(\delta_C^A = 1|x_C)$ with an additional step, which is modeling $p(\delta_U^R|x_U)$. Treating π_i^R as a random variable for $i \in S_A$ conditional on x_i, one can estimate this probability by regressing the π_i^R's on the x_i's in U (Pfeffermann

and Sverchkov, 2009), because

$$\begin{aligned} p(\delta_U^R = 1|x_U;\gamma) &= \int_0^1 p(\delta_U^R = 1|\pi_U^R, x_U;\gamma)p(\pi_U^R|x_U;\gamma)d\pi_U^R \\ &= \int_0^1 \pi_U^R p(\pi_U^R|x_U;\gamma)d\pi_U^R \\ &= E(\pi_U^R|x_U;\gamma) \end{aligned} \tag{4.17}$$

Pfeffermann and Sverchkov (1999b) demonstrate that $E(\pi_U^R|x_U) = E^{-1}(w_R|x_R)$ where w_R are the sampling weights in S_R. Since π_i^R is only observed in S_R, then, the sample estimator of π_i^R is given by

$$p(\delta_C^A = 1|x_C;\gamma,\phi) = E^{-1}(w_R|x_C;\gamma)\frac{p(\delta_C^A = 1|x_C;\phi)}{1 - p(\delta_C^A = 1|x_C;\phi)} \tag{4.18}$$

$E(w_R|x)$ is modeled using a GLM with a *log* link function, as the distribution of the w_i^R's tends to be right-skewed in the actual survey data. In addition, I know that the sampling weights are usually a multiplicative factor of selection probabilities×non-response adjustment×post-stratification. Therefore, given the posterior distribution of $p(\gamma, \beta|x_C, w_R)$, one can obtain the posterior distribution of π_i^A for $i \in S_C$ by

$$p(\delta_C^A = 1|x_C;\gamma,\phi) = exp\big\{x_C^T(\phi - \gamma)\big\} \tag{4.19}$$

The joint distribution of (y_A, δ_C^A, w_R) can be written as:

$$p(y_A, \delta_C^A, w_R|x_C) = \int p(y_R, y_A|f(\pi^A[x_C, \delta_C^A, w_R;\gamma,\phi]), x_C;\theta)p(\delta_C^A|x_C;\phi)p(w_R|x_R;\gamma)dy_R \tag{4.20}$$

where $\pi^A[x_C, \delta_C^A, w_R;\gamma,\phi] = exp\big\{x_C^T(\phi-\gamma)\big\}$ according to Eq. 4.19. The correspond-

ing posterior predictive distribution of y_R is given by

$$p(y_R|y_A, \delta_C^A, w_R, x_C, \delta_C^A, \pi_R^R) = \int\int\int p(y_R|y_A, f(\pi^A[x_C, \delta_C^A, \pi_R^R; \gamma, \phi]), x_C; \theta) \times p(\phi|\delta_C^A, x_C)p(\gamma|w_R, x_R)d\theta d\phi d\gamma \tag{4.21}$$

Although Zigler (2016) argues that such a factorization of the joint distribution of (y_i, δ_i^A, w^R) does not correspond to a valid use of the Bayes' theorem, for certain reasons, it has been advocated by several studies. First, Little (2004) highlights the fact that Bayesian joint modeling can result in better repeated sampling properties. It has been well-understood that the performance of the alternative two-step Bayesian methods with respect to frequentist properties depends on the choice of priors (Kaplan and Chen, 2012). Furthermore, having both π_i^A and x_i as predictors in the PM cuts the notorious feedback between the QR and PM models, which leads to incorrect estimation of the PS posterior distribution (Zigler et al., 2013).

However, what matters most in this study is the double robustness property that the likelihood factorization in Eq. 4.20 offers. For instance, by choosing a parametric form $f(\pi_i^A) = \theta^*/\pi_i^A$, where θ^* is an unknown scalar parameter, this factorization leads to a linear-in-weight Prediction (LWP) model. Scharfstein et al. (1999) and Bang and Robins (2005) identified the correspondence between LWP and AIPW estimators. In the causal inference context, this has been termed a clever covariate by Rose and van der Laan (2008) as it characterizes the correct relationship between the propensity scores and the outcome model. In the context of item-missing data imputation, Little and An (2004) suggest that the use of a more flexible non-parametric function can improve the efficiency of the adjusted estimator, especially when there are extreme values in the estimated PS. The authors propose to use a penalized spline model, which is piecewise continuous polynomials of the estimated PS, paired with a mathematical penalization to find the best fit of PM to the data (Ruppert et al., 2003; Fahrmeir et al., 2011). Alternatively, McCandless et al. (2009) suggest categorizing

propensity scores into quantiles and using them as dummy variables to augment the PM.

In the current study, I extend the PSPP idea to a non-probability sample setting while using Gaussian process (GP) regression instead of a penalized spline model. As a flexible non-parametric Bayesian approach, GP can automatically capture non-linear associations as well as multi-way interactions (Rusmassen and Williams, 2005; Neal, 1997). Having $\pi_i^A = p(\delta_i^A = 1|x_i, w^R; \gamma, \phi)$ estimated for $i \in S_C$, for a continuous outcome variable, I fit a semiparametric model on S_A as below:

$$y_i|x_i, d_i, \hat{\pi}_i, \theta = \theta_0 + \sum_{j=1}^{p} \theta_j x_{ij} + \sum_{j=p+1}^{p+q} \theta_j d_{ij} + f\left(\hat{\pi}_i^A\right) + \epsilon_i \tag{4.22}$$

where θ denotes a $(p+q+1)$-dimensional vector of the PM parameters, and $\epsilon_i \sim N(0, \sigma^2)$ with σ^2 being unknown. Eq. 4.22 involves two parts: a linear regression parameterized by θ and a GP denoted by $f(.)$.

A GP $\{f(x) : x \in R^N\}$ is a set of random variables, any finite number of which jointly follow a multivariate Gaussian distribution. In a full-ranked GP, $f(.)$ is *a priori* defined by its mean and covariance functions as below:

$$f(x) \sim GP\left(\mu(x), K\left(x, x'\right)\right) \tag{4.23}$$

where $\mu(x)$ is the mean vector and $K(x, x')$ is the covariance matrix. The latter encompasses all our prior beliefs about the functional association between x and y, including continuity, smoothness, periodicity and scale properties (Riutort-Mayol et al., 2020). For notational simplicity, I set $\mu(x) = 0$, though it is not necessary. It is worth noting that the LWP model can be viewed as a specific type of GP with a dot product covariance matrix as $\alpha^2[1 + ((\pi_i^A)^T \pi_j^A)^{-1}]$ if the regression coefficient is specified a prior of $N(0, \alpha^2)$ (Rusmassen and Williams, 2005). While literature suggests a variety of covariance functions for GP, the most common type is the *squared*

exponential (SE) covariance matrix whose elements take the following form:

$$k(x, x') = \alpha^2 exp\Big\{ - \frac{||x - x'||^2}{2\rho} \Big\} \tag{4.24}$$

where ρ is called a *length-scale* parameter, and α is known as the *marginal standard error*. One can show that the SE covariance structure represents a kernel with an infinite number of a basis functions (Rusmassen and Williams, 2005).

From a weight-space viewpoint, Huang et al. (2019) show that with a stationary isotropic kernel, where $K(\pi_i^A, \pi_j^A) = f(||\pi_i^A - \pi_j^A||)$, GP acts as a non-parametric matching technique. Wang et al. (2020a) prove the *consistency* of a kernel-weighted estimator under certain regularity conditions. Refer to Appendix 4.6.1 to see the connection between GP and kernel weighting. In our non-probability sample setting, one can view it as matching units of S_A to units of S_R based on the estimated propensity scores, π_i^A's (Rivers, 2007). Further theoretical properties of kernel optimal matching, such as consistency, can be found in Kallus et al. (2018). Although the SE covariance has desirable properties, empirical results show that it is not a strong fit for the real-world data as it is infinitely differentiable (Rusmassen and Williams, 2005). Therefore, I propose to use a Matérn kernel added to an inhomogeneous standardized polynomial kernel of order p as below:

$$\begin{aligned} K(x_i, x_j) = \alpha^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{2\nu} \frac{||x_i - x_j||}{\rho} \right)^{\nu} K_\nu \left(\sqrt{2\nu} \frac{||x_i - x_j||}{\rho} \right) \\ + \left(\frac{\tau^2 + x_i^T x_j}{\sqrt{\tau^2 + x_i^T x_i} \sqrt{\tau^2 + x_j^T x_j}} \right)^p \end{aligned} \tag{4.25}$$

where $\Gamma(.)$ denotes the gamma function, and $K_\nu(.)$ is a modified Bessel function of the second kind. This combination of two kernels ensures capturing both local variations and long-range discrepancies in the estimated propensity scores (Vegetabile, 2018). Note that for $\nu \to \infty$, Matérn covariance will converge to the SE covariance, and the

sum of two valid kernels is still a valid kernel.

In this chapter, I set $\nu = 3/2$ and $p = 1$ throughout the simulation and empirical studies, which yields the following covariance function:

$$K(x_i, x_j) = \alpha^2 \left(1 + \frac{\sqrt{3}||x_i - x_j||}{\rho}\right) exp\left(-\frac{\sqrt{3}||x_i - x_j||}{\rho}\right) + \frac{\tau^2 + x_i x_j}{\sqrt{\tau^2 + x_i^2}\sqrt{\tau^2 + x_j^2}} \tag{4.26}$$

In addition, I propose to use a *log* transformation of the $\hat{\pi}_i^A$'s in the GP part. This is because the input of GP will become a linear combination, i.e. $x_C^T(\phi - \gamma)$, and given normal priors assigned to β, this linear combination will follow a Gaussian distribution (Si et al., 2015).

Fully Bayesian inference using GP comes with computational issues even for a moderate n_A as one has to invert the covariance matrix at each posterior sampling step that needs $O(n_A^3)$ computations. The problem becomes even more severe when the joint posterior distribution of (π_i^A, y_i) has to be simulated. I propose to use a low-ranked sparse GP based on the Laplace eigenvectors approximation (Solin and Särkkä, 2020; Riutort-Mayol et al., 2020). Such a method reduces the computational complexity up to $O(n_A l^2)$ where $l << n_A$ is the reduced rank of the covariance matrix. Further details about the partially linear GP regression and the Laplace approximation can be found in Appendix 4.6.3.

Under a standard Bayesian framework, a set of independent prior distributions are assigned to the model parameters, and conditional on the observed data through a joint likelihood function, the associated posterior distributions are obtained. To this end, I use the "black box" solver Stan (Carpenter et al., 2017), which employs a Hamiltonian Monte Carlo (HMC) technique to simulate the posterior predictive distribution of the parameters. In the following, I show the structure of our proposed method in Stan.

STEP 1: Specifying priors

$$\theta, \gamma, \phi \sim t\text{-}student(3, 0, 1)$$

$$\lambda, \alpha, \sigma \sim t\text{-}student^{+}(3, 0, 1)$$

$$\rho \sim GIG(0, 1, 2)$$

$$\xi^R \sim Dirichlet(1, 1, ..., 1)$$

STEP 2: Setting joint likelihood

$$w^R | x_R, \gamma, \lambda \sim N\left(exp\{x_R^T \gamma\}, \lambda^2\right)$$

$$\delta_C^A | x_C, \phi \sim Bernoulli\left(logit^{-1}\{x_C^T \phi\}\right)$$

$$y_A | z_A, \theta, \sigma \sim Normal\left(z_A^T \theta + f\left(x_A^T(\phi - \gamma), \alpha, \rho, \tau\right), \sigma^2\right)$$

$$n^R | \xi^R \sim Multinomial(n_R, \xi)$$

STEP 3: Obtaining posterior

$$\hat{y}_R | y_A, z_R, \theta, \sigma \sim Normal\left(z_R^T \theta + f\left(x_R^T(\phi - \gamma), \alpha, \rho, \tau\right), \sigma^2\right)$$

$$\hat{N} | \pi^R, \xi^R \sim Multinomial\left(N - n_R, c\xi^R(1 - \pi^R)/\pi^R\right)$$

$$\hat{\bar{y}}_U = \left\{ \sum_{j=1}^{J} \frac{\hat{N}_j}{n_j^R} \sum_{i=1}^{n_j^R} \hat{y}_{j[i]} + \sum_{i=1}^{n_A} \{y_i - \hat{y}_i\} \right\} / N$$

where $t\text{-}student^{+}$ denotes a half $t\text{-}student$ and GIG stands for the Generalized Inverse Gaussian distribution, which is recommended in Stan User's Guide (Stan Development Team, 2019) for the length-scale parameter of a partially linear GP regression. Also, $f(.)$ denotes a low-ranked GP approximation with $l = 10$ and a boundary condition factor of $c = 1.25$, where the covariance function is given by Eq. 4.26. I simulate the posterior predictive distribution of $\hat{\bar{y}}_U$ in Stan using $M = 500$ HMC draws after discarding the first 500 draws as the burn-in period.

4.3 Simulation study

Two simulations are presented in this section, in which I compare the performance of our proposed GPPP method with those of LWP, AIPW, and PAPP with respect to the bias magnitude, efficiency, and accuracy of the variance estimator. All of the competing methods are DR, except for the PAPP method, which is an inverse PS weighted estimate of the observed y_i for $i \in S_A$ with PS estimated from Eq. 4.16. The GPPP and LWP methods are fully implemented under a Bayesian setting, whereas AIPW and PAPP estimates are obtained under a frequentist method (Rafei et al., 2021). Therefore, for the earlier class of methods, I am able to compute 95% credible intervals (95% CIs) while for the latter, a bootstrap method with $B = 100$ replications is employed to estimate the variance and 95% confidence intervals.

Various scenarios are considered with different assumptions about the functional form of the relationship among variables. For both studies, S_A and S_R are given a random selection mechanism with unequal inclusion probabilities. Note that units of both samples are selected independently with no clustering or stratification. Once S_A and S_R are drawn from U, I assume that π_i^A for $i \in S_C$ and y_j for $j \in S_R$ are unobserved, and the aim is to adjust for the selection bias in S_A based on the combined sample, S_C. The simulation is then iterated $K = 216$ times (which is a multiple of 36, the number of cores I employed for parallel computing), where the bias-adjusted point estimates, SE and associated 95% credible/confidence interval (CI) for $\bar{y}_U$ are estimated in each iteration.

To evaluate the repeated sampling properties of the competing method, relative bias (rBias), relative root mean square error (rMSE), the nominal coverage rate of 95% CIs (crCI), relative length of 95% CIs (rlCI) and SE ratio (rSE) are calculated

as below:

$$rbias\left(\hat{\bar{y}}_U\right) = 100 \times \frac{1}{K}\sum_{k=1}^{K}\left(\hat{\bar{y}}_U^{(k)} - \bar{y}_U\right)/\bar{y}_U \tag{4.27}$$

$$rMSE\left(\hat{\bar{y}}_U\right) = 100 \times \sqrt{\frac{1}{K}\sum_{k=1}^{K}\left(\hat{\bar{y}}_U^{(k)} - \bar{y}_U\right)^2}/\bar{y}_U \tag{4.28}$$

$$crCI\left(\hat{\bar{y}}_U\right) = 100 \times \frac{1}{K}\sum_{k=1}^{K} I\left(\left|\hat{\bar{y}}_U^{(k)} - \bar{y}_U\right| < z_{0.975}\sqrt{var\left(\hat{\bar{y}}_U^{(k)}\right)}\right) \tag{4.29}$$

$$rlCI\left(\hat{\bar{y}}_U\right) = 100 \times \frac{2}{K}\sum_{k=1}^{K} z_{0.975}\sqrt{var\left(\hat{\bar{y}}_U^{(k)}\right)} \tag{4.30}$$

$$rSE\left(\hat{\bar{y}}_U\right) = \frac{1}{K}\sum_{k=1}^{K}\sqrt{var(\hat{\bar{y}}_U^{(k)})}/\sqrt{\frac{1}{K-1}\sum_{k=1}^{K}\left(\hat{\bar{y}}_U^{(k)} - \bar{\bar{y}}_U\right)^2} \tag{4.31}$$

where $\hat{\bar{y}}_U^{(k)}$ denotes the adjusted sample mean from iteration k, $\bar{\bar{y}}_U = \sum_{k=1}^{K}\hat{\bar{y}}_U^{(k)}/K$, $\bar{y}_U$ is the finite population true mean, and $var(.)$ represents the variance estimate of the adjusted mean based on the sample. Finally, to test the DR property of the proposed methods, I investigate different scenarios regarding whether models for QR and PM are correctly specified or not.

4.3.1 Simulation I

4.3.1.1 Design

The design of our first study is based on the simulation implemented in Chen et al. (2019). Consider a finite population of size $N = 10^5$ with $z = \{z_1, z_2, z_3, z_4\}$ being a set of auxiliary variables generated as follows:

$$z_1 \sim Ber(p = 0.5) \qquad z_2 \sim U(0, 2) \qquad z_3 \sim Exp(\mu = 1) \qquad z_4 \sim \chi^2_{(4)} \tag{4.32}$$

and $x = \{x_1, x_2, x_3, x_4\}$ is subsequently defined as a linear function of z as below:

$$x_1 = z_1 \qquad x_2 = z_2 + 0.3z_1 \qquad x_3 = z_3 + 0.2(x_1 + x_2) \qquad x_4 = z_4 + 0.1(x_1 + x_2 + x_3) \tag{4.33}$$

Given x, a continuous outcome variable y is constructed by

$$y_i = 2 + x_{1i} + x_{2i} + x_{3i} + x_{4i} + \sigma\epsilon_i \tag{4.34}$$

where $\epsilon_i \sim N(0, 1)$, and σ is defined such that the correlation between y_i and $\sum_{k=1}^{4} x_{ki}$ equals $\rho = 0.8$. Further, associated with the design of S_A, a set of selection probabilities are assigned to the population units through the following logistic model:

$$log\left(\frac{\pi_i^A}{1 - \pi_i^A}\right) = \gamma_0 + 0.1x_{1i} + 0.2x_{2i} + 0.1x_{3i} + 0.2x_{4i} \tag{4.35}$$

where γ_0 is determined such that $\sum_{i=1}^{N} \pi_i^A = n_A$. For the selection probabilities in S_R, I assume that $\pi_i^R \propto \gamma_1 + z_{3i}$, where γ_1 is obtained such that $max\{\pi_i^R\}/min\{\pi_i^R\} = 50$. It is important to note that in this simulation study π_i^R is assumed to be known for $i \in S_A$ as z_3 is observed in S_A.

Using these measures of size, I repeatedly draw pairs of samples corresponding to S_A and S_R from U through a Poisson sampling design. The simulation is then repeated for different pairs of expected sample sizes, i.e. $(n_A, n_R) = (500, 500)$, $(n_A, n_R) = (1,000, 500)$ and also $(n_A, n_R) = (500, 1,000)$. (Note that the actual sample size is a random variable under a Poisson sampling design.) Both Y and π^A are associated with a linear combination of X in this simulation study. Finally, in order to misspecify a model, I omit x_4 from the predictors of the working model. In Appendix 4.6.4, I provide extensions of the simulation for $\rho = \{0.3, 0.5\}$.

Table 4.1: Comparing the performance of the bias adjustment methods in the first simulation study for $\rho = 0.8$

	$n_A = 500, \quad n_R = 500$					$n_A = 1,000, \quad n_R = 500$					$n_A = 500, \quad n_R = 1,000$				
Measure	rBias	rMSE	crCI	rlCI	rSE	rBias	rMSE	crCI	rlCI	rSE	rBias	rMSE	crCI	rlCI	rSE
Probability sample (S_R)															
UW	8.866	9.093	0.926	0.724	0.982	8.866	9.093	0.926	0.724	0.982	8.819	8.942	0.000	0.513	0.953
FW	0.150	2.322	93.981	0.844	0.998	0.150	2.322	93.981	0.844	0.998	0.030	1.692	94.907	0.598	0.969
Non-probability sample (S_A)															
UW	30.675	30.794	0.000	0.940	0.950	29.958	30.006	0.000	0.657	1.063	30.675	30.794	0.000	0.940	0.950
FW	-0.038	2.354	93.519	0.811	0.944	-0.044	1.618	93.056	0.570	0.965	-0.038	2.354	93.519	0.811	0.944
Model specification: QR–True, PM–True															
GPPP	0.054	2.473	96.759	0.935	1.035	0.014	2.171	95.370	0.862	1.087	0.003	2.039	95.370	0.750	1.007
LWP	0.034	2.586	97.222	1.502	1.591	-0.027	2.193	94.907	0.860	1.074	-0.071	2.107	94.907	0.763	0.992
AIPW	-0.042	2.528	93.981	0.867	0.940	0.001	2.166	93.519	0.772	0.976	-0.094	2.086	93.981	0.709	0.932
PAPP	0.899	2.641	90.741	0.873	0.963	0.646	2.118	93.519	0.735	0.998	1.119	2.455	88.426	0.766	0.960
Model specification: QR–True, PM–False															
GPPP	0.025	2.465	96.759	0.934	1.038	0.003	2.177	93.981	0.856	1.078	-0.007	2.027	95.833	0.752	1.016
LWP	0.022	2.462	96.759	0.935	1.041	0.007	2.161	94.907	0.858	1.087	-0.028	2.036	95.833	0.749	1.007
AIPW	-0.002	2.452	92.593	0.844	0.943	0.003	2.150	93.981	0.757	0.964	-0.068	2.019	93.519	0.697	0.947
Model specification: QR–False, PM–True															
GPPP	0.945	2.855	95.370	0.983	0.999	0.798	2.453	96.759	0.898	1.061	0.990	2.461	91.204	0.792	0.963
LWP	3.989	5.740	76.852	1.413	0.937	4.233	5.511	75.463	1.278	0.992	3.959	5.245	71.759	1.136	0.904
AIPW	0.215	2.532	93.056	0.855	0.929	0.092	2.068	93.981	0.744	0.987	0.173	2.138	93.519	0.752	0.967
Model specification: QR–False, PM–False															
GPPP	27.303	27.460	0.000	1.591	1.485	26.513	26.590	0.000	1.443	1.962	27.308	27.451	0.000	1.291	1.263
LWP	27.132	27.295	0.000	1.600	1.470	26.437	26.514	0.000	1.441	1.955	27.194	27.341	0.000	1.307	1.262
AIPW	27.162	27.322	0.000	0.986	0.914	26.453	26.531	0.000	0.725	0.978	27.110	27.252	0.000	0.947	0.934
PAPP	27.946	28.097	0.000	0.976	0.918	26.996	27.070	0.000	0.715	0.979	28.166	28.303	0.000	0.954	0.941

NOTE 1: GPPP: Gaussian Process of Propensity Prediction; LWP: Linear-in-weight Prediction; AIPW: Augmented Inverse Propensity Weighting; PAPP: Propensity-adjusted Probability Prediction.
NOTE 2: The PAPP and AIPW methods have been implemented through a bootstrap method.
NOTE 3: The PAPP method is non-robust, while the rest of the methods, i.e. GPPP, LWP, and AIPW, are doubly robust.

4.3.1.2 Results

Table 4.1 summarizes the numerical results of the first simulation study across different sample size scenarios for $\rho = 0.8$. As illustrated, naive estimates of the population mean are biased in both S_R and S_A while weighting fully corrects for the bias in both samples. For the non-robust method, PAPP, estimates are unbiased as long as the QR model is correct. The DR methods produce unbiased estimates when either the QR model or PM holds, though there is evidence of residual bias for the LWP method when the QR model holds but the PM is misspecified. In terms of rMSE, all the methods perform similarly, except for the LWP method with correct and incorrect models specified the QR and PM, respectively, which shows higher degrees of rMSE compared to the alternative methods.

AIPW and PAPP have slightly narrower CIs than the Bayesian methods, GPPP and LWP. The LWP performs poorly with respect to efficiency when the PM is incorrectly specified. Generally, the values of rSE suggest that variance estimation is unbiased across different model specification scenarios with a slight overestimation and underestimation in the Bayesian and bootstrap methods, respectively. Under the situations where the working model for QR is correct while that for PM is incorrect, LWP tends to underestimate the variance. The coverage rates of 95% CIs are also close to the nominal value when at least one of the QR and PM models is correctly specified. However, I observe that 95% CIs based on the frequentist methods tend to undercover the true population mean to some degrees, and the poorest result of crCI belongs to the LWP method when the PM is wrongly specified. These findings are generalizable to all other sample size combinations, and to the other extensions of the simulation for $\rho = 0.3, 0.5$, whose tables are displayed in Appendix 4.6.4.

4.3.2 Simulation II

4.3.2.1 Design

In the previous simulation study, the ignorable assumption was violated to misspecify the working model by dropping a key auxiliary variable. Now, I focus on a situation where models misspecified with respect to the functional form of their conditional means. To this end, I consider (non-)linear associations and two-way interactions in construction of the outcome variables. In addition, to build a more realistic situation, two separate sets of auxiliary variables are generated, D associated with the design of S_R, and X associated with the design of S_A. However, I allow the two variables to be correlated through a bivariate gaussian distribution as below:

$$\begin{pmatrix} d \\ x \end{pmatrix} \sim MVN \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 4 & 2\rho \\ 2\rho & 1 \end{pmatrix} \right) \tag{4.36}$$

Note that ρ controls how strongly the sampling design of S_R is associated with that of S_A. Primarily, I set $\rho = 0.5$, but later I check other values ranging from 0 to 0.9 as well.

I then generate a continuous outcome variable (y_i^c) and the *binary* outcome variable (y_i^b) for $i \in U$ as below:

$$\begin{aligned} y_i^c &= 3 + f_k(x_i) + d_i + 0.2x_i d_i + \sigma\epsilon_i \\ p(y_i^b = 1|x_i, d_i) &= \frac{exp\{-1 + f_k(x_i) + d_i + 0.2x_i d_i\}}{1 + exp\{-1 + f_k(x_i) + d_i + 0.2x_i d_i\}} \end{aligned} \tag{4.37}$$

where $\epsilon_i \sim N(0, 1)$, and σ is determined such that the correlation between y_i^c and $f_k(x_i) + d_i + 0.2x_i d_i$ equals 0.8 for $i \in U$. The function $f_k(.)$ is assumed to take one

of the following forms:

$$\begin{aligned} LIN : f_1(x) &= x & CUB : f_2(x) &= (x/3)^3 \\ EXP : f_3(x) &= exp(x/2)/5 & SIN : f_4(x) &= 5sin(\pi x/3) \end{aligned} \tag{4.38}$$

Figure 4.1 depicts the relationships between y^c and π^A, and between y^c and $w^A = 1/\pi^A$.

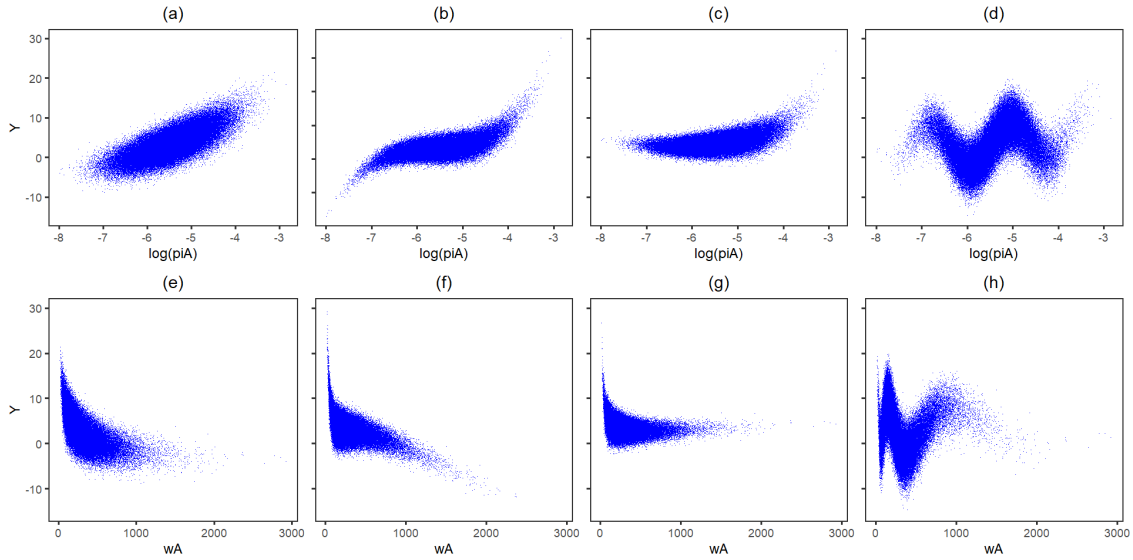

Figure 4.1: The proposed relationships between the outcome variable Y and $log(\pi^A)$ in U for (a) LIN, (b) CUB, (c) EXP and (d) SIN scenarios, and between the outcome Y and sampling weights w^A for (e) LIN, (f) CUB, (g) EXP and (h) SIN scenarios.

I then consider an informative sampling strategy with unequal probabilities of inclusion, where the selection mechanism of S_A and S_R depends on x and d, respectively. Thus, each $i \in U$ is assigned two values within $(0, 1)$ corresponding to the probabilities of selection in S_R and S_A through a *logistic* function as below:

$$\begin{aligned} \pi^R(d_i) = p(\delta_i^R = 1|d_i) &= \frac{exp\{\gamma_0 - 0.4d_i\}}{1 + exp\{\gamma_0 - 0.4d_i\}} \\ \pi^A(x_i) = p(\delta_i^A = 1|x_i) &= \frac{exp\{\gamma_1 + \gamma_2 x_i\}}{1 + exp\{\gamma_1 + \gamma_2 x_i\}} \end{aligned} \tag{4.39}$$

where δ_i^R and δ_i^A are the indicators of being selected in S_R and S_A, respectively, for $i \in U$. I initially set $\gamma_2 = 0.3$, which yields PS with a normal range. To assess how the adjustments behave in presence of influential weights, later I set $\gamma_2 = 0.6$, which yields relatively extreme weights.

Associated with S_R and S_A, independent samples of expected sizes $n_R = 1,000$ and $n_A = 500$ are selected randomly from U with a Poisson sampling design. I choose $n_A < n_R$ as is the case in the two applications of this study. The model intercepts, γ_0 and γ_1 in 4.39, are obtained such that $\sum_{i=1}^N \pi_i^R = n_R$ and $\sum_{i=1}^N \pi_i^A = n_A$, respectively. The rest of the simulation design is similar to that defined in Simulation I, except for the way I specify a working model. A QR model is misspecified by replacing x_i with x_i^2, and a PM model is misspecified by replacing $f_k(x_i)$ with x_i^2 and d_i with d_i^2, and also by dropping the interaction term $x_i d_i$.

4.3.2.2 Results

Figure 4.2 compares the relative bias (rBias) magnitude and efficiency of the competing methods for the continuous outcome variable, y^c, across different scenarios of model specification while $\gamma_2 = 0.3$. Note that the error bars reflect the relative length of 95% CIs (rlCI). As illustrated, point estimates from both S_R and S_A are biased if the sampling true weights are ignored. At the first glance, one can infer that for all f_k, $k = 1, 2, 3, 4$, the magnitude of rBias is close to *zero* as long as either QR or PM model is valid. However, in situations where π^A is non-linearly associated with y^c, i.e. plots (b), (c), and (d), the AIPW and PAPP estimators are biased when the PM is misspecified, but the QR model is valid. In contrast, the LWP method yields slightly biased estimates in all plots when the QR model is misspecified, but the PM is correct. It turns out that the GPPP is the only method that leads to unbiased estimates in all the scenarios with respect to model specification and functional form of the PM. I did not observe consistent results across the adjustment methods with

respect to efficiency. However, the GPPP method consistently shows high efficiency compared to the other methods across all the studied scenarios.

I summarize the simulation results for the binary outcome, y^b, with $\gamma_2 = 0.3$ in Figure 4.3. Again, adjusted estimates are unbiased if the working model for either QR or PM holds. Exceptions are seen for the PAPP and AIPW methods with residual bias in the plots related to (c) EXP, and (d) SIN when the PM is incorrectly specified. Unlike the simulation results for the continuous variable, the LWP consistently produces unbiased estimates for the binary outcome when the working model for QR fails. However, the magnitude of bias seems to be much larger in the LWP method when both underlying models for QR and PM are misspecified. Again, as for the continuous outcome, the proposed GPPP method consistently gives unbiased and efficient estimates. The lowest efficiency is associated with the AIPW and PAPP methods in the EXP scenario when the PM is misspecified.

Figure 4.4 displays the results of crCI and rSE for the continuous outcomes where $\gamma_2 = 0.3$. According to the rSE values, all methods perform well in variance estimation except for the LWP method which consistently underestimates the variance. A similar problem appears in the PAPP and AIPW methods for the EXP scenario when the outcome model is invalid. Generally, the Bayesian methods, i.e. GPPP and LWP, tend to slightly overestimate the variance. The values of crCI seem to be close to the nominal level for all the methods across almost all the scenarios, as long as at least one of the underlying models holds. For the non-linear associations, i.e. (b) CUB, (c) EXP and (d) SIN, the 95% CIs associated with frequentist methods, i.e. AIPW and PAPP, tend to undercover the population mean when the outcome model is false. Figure 4.5 depicts similar results for the binary outcome when $\gamma_2 = 0.3$. Overall, the results look analogous to those obtained for the continuous outcome. However, the degree of overestimation of variance by the Bayesian methods seems to be larger in the binary outcome than the continuous outcome.

Extensions of the simulation for other sample size combinations, i.e. $(n_A, n_R) = (500, 500)$ and $(n_A, n_R) = (1,000, 500)$ and also for $\gamma_2 = 0.6$, which creates extreme sampling weights in S_A, are included in Appendix 4.6.4. While I observe no major discrepancy in the simulation results for other sample size scenarios than $(n_A, n_R) = (500, 1,000)$, having influential weights presented in S_A leads to a larger magnitude of bias and lower efficiency in the estimates of PAPP, AIPW when the PM is incorrectly specified, but the QR model is valid. However, the GPPP method seems to be least affected by the presence of extreme weights.

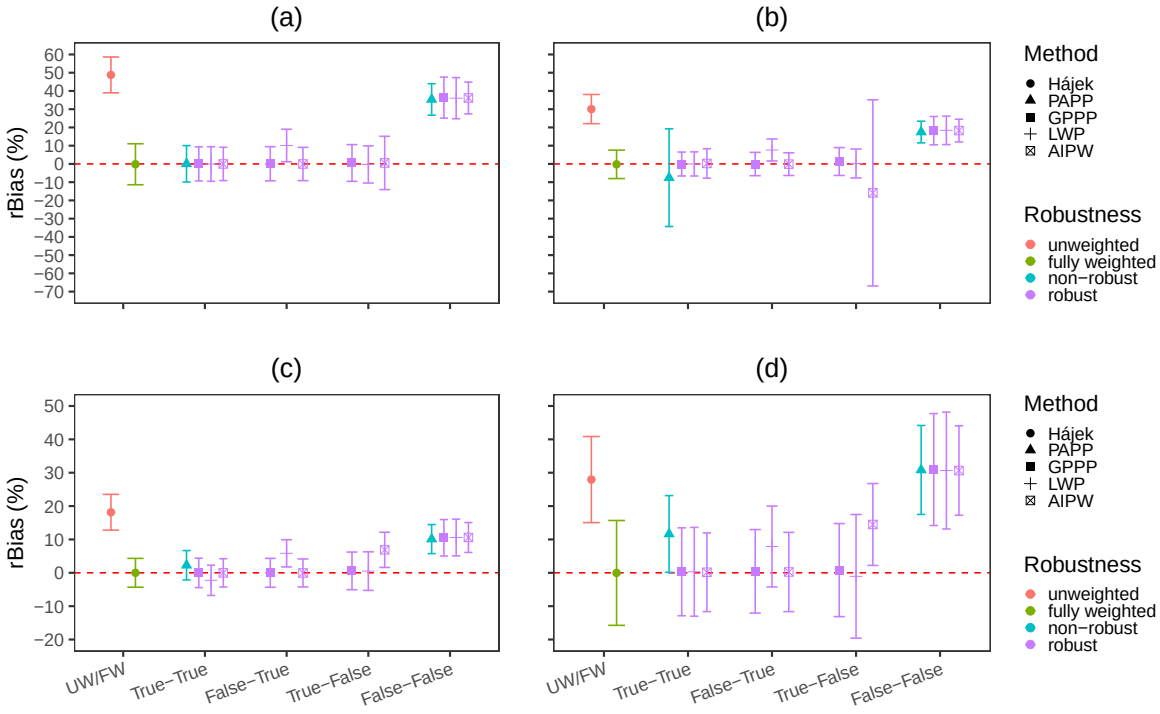

Figure 4.2: Comparing the performance of the adjusted estimators under different model-specification scenarios for the *continuous* outcome variable with $\gamma_2 = 0.3$ under (a) LIN, (b) CUB, (c) EXP, and (d) SIN scenarios. The error bars have been drawn based on the 2.5% and 97.5% percentiles of the empirical distribution of bias over the simulation iterations. UW: unweighted; FW: Fully weighted; PAPP: Propensity-adjusted Probability Prediction; GPPP: Gaussian Processes of Propensity Prediction; LWP: Linear-in-weight Prediction; AIPW: Augmented Inverse Propensity Weighting

In Figure 4.4 and 4.5, I depict the measures associated with the accuracy of the variance methods for GPPP/AIPW estimators. One can immediately infer that for

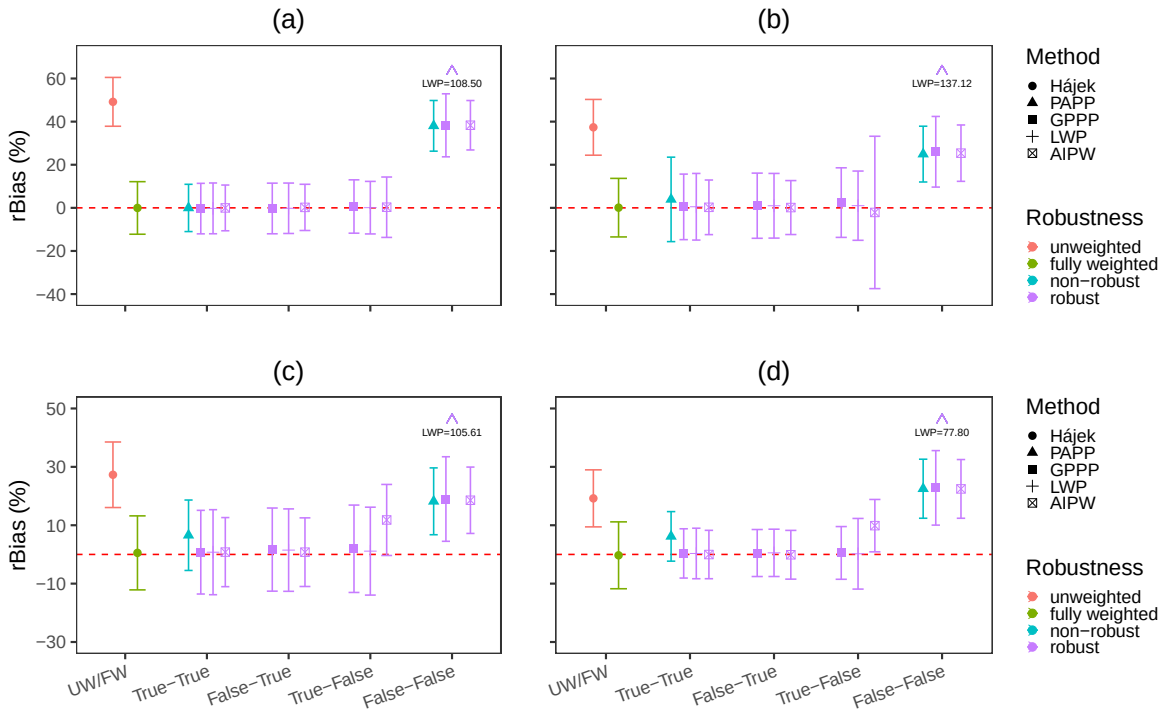

Figure 4.3: Comparing the performance of the adjusted estimators under different model-specification scenarios for the *binary* outcome variable with $\gamma_2 = 0.3$ under (a) LIN, (b) CUB, (c) EXP, and (d) SIN scenarios. The error bars have been drawn based on the 2.5% and 97.5% percentiles of the empirical distribution of bias over the simulation iterations. UW: unweighted; FW: Fully weighted; PAPP: Propensity-adjusted Probability Prediction; GPPP: Gaussian Processes of Propensity Prediction; LWP: Linear-in-weight Prediction; AIPW: Augmented Inverse Propensity Weighting

both employed methods, the variance estimator is approximately unbiased when at least one of the underlying models holds. However, in situations where both models are invalid, according to the rSE values, the AIPW estimator tends to underestimate/overestimate the variance to a significant extent, while the variance estimator under GPPP shows more robustness across the model specification scenarios as well as outcome variables. Last but not least, the proximity of the crCI values to 95% for the GPPP methods, especially when both underlying models are wrong, reflects the accuracy of both point and variance estimates under the GPPP method.

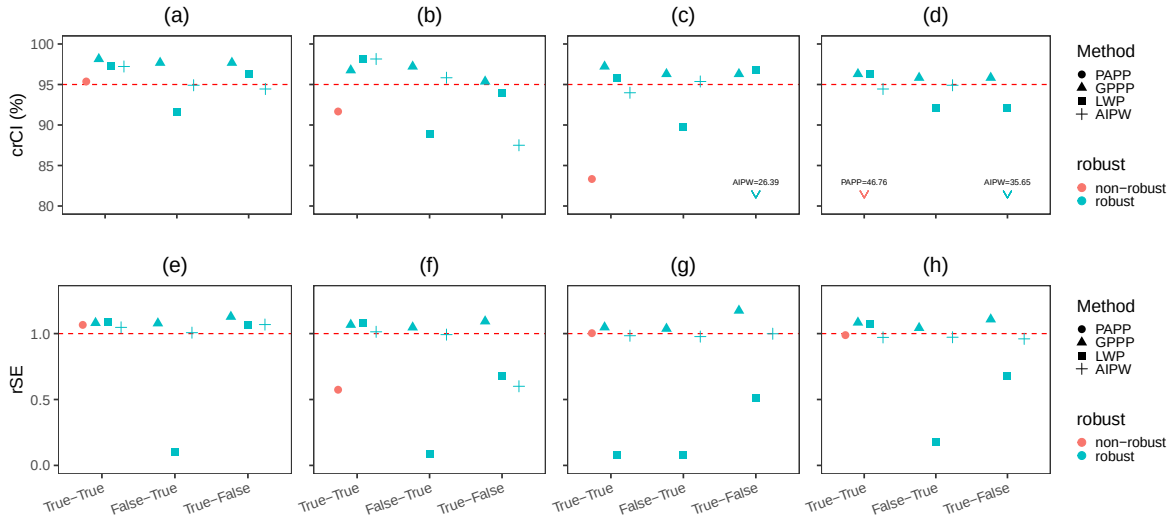

Figure 4.4: Comparing the 95% CI coverage rates (crCI) of the DR adjusted means for the *continuous* outcome variable with $\gamma_2 = 0.3$ under (a) LIN, (b) CUB, (c) EXP, and (d) SIN scenarios, and SE ratios (rSE) under (e) LIN, (f) CUB, (g) EXP, and (h) SIN scenarios, across different DR methods under different model specification scenarios. UW: unweighted; FW: Fully weighted; PAPP: Propensity-adjusted Probability Prediction; GPPP: Gaussian Processes of Propensity Prediction; LWP: Linear-in-weight Prediction; AIPW: Augmented Inverse Propensity Weighting

So far, the results I discussed were limited to a case where $\rho = 0.5$. As the final step, I replicate the simulation for different values of ρ ranging from 0 to 0.9 to show how stable the competing methods perform in terms of rbias and rMSE. Figure 4.6 depicts changes in the values of rBias and rMSE in the continuous outcome, y^c, for different adjustment methods and across different model specification scenarios as the

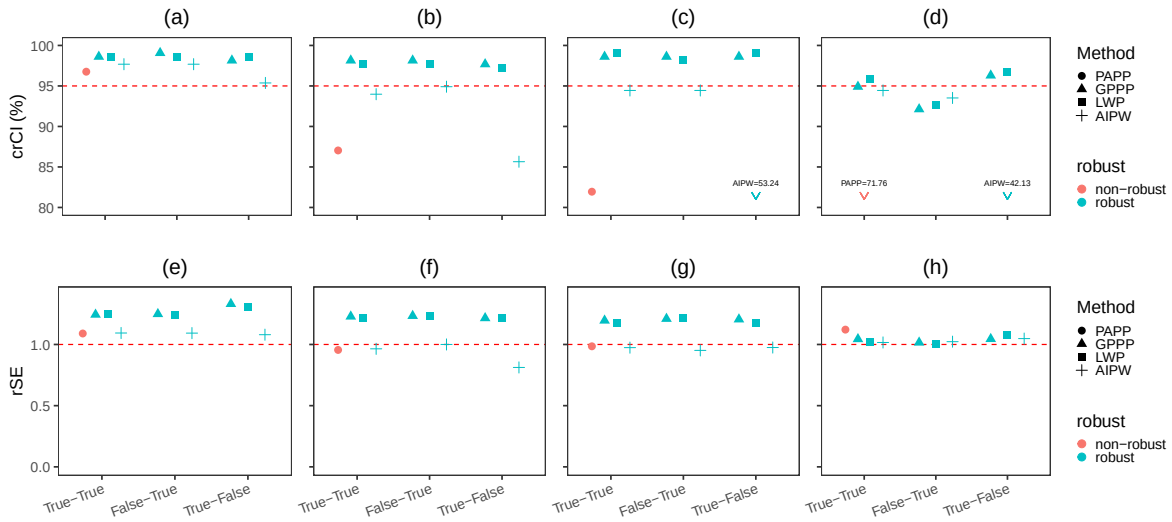

Figure 4.5: Comparing the 95% CI coverage rates (crCI) of the DR adjusted means for the *binary* outcome variable with $\gamma_2 = 0.3$ under (a) LIN, (b) CUB, (c) EXP, and (d) SIN scenarios, and SE ratios (rSE) under (e) LIN, (f) CUB, (g) EXP, and (h) SIN scenarios, across different DR methods under different model specification scenarios. UW: unweighted; FW: Fully weighted; PAPP: Propensity-adjusted Probability Prediction; GPPP: Gaussian Processes of Propensity Prediction; LWP: Linear-in-weight Prediction; AIPW: Augmented Inverse Propensity Weighting

value of ρ increases. Generally, it seems that the values of rBias and rMSE decline for all competing methods with an increase in ρ. In addition, for all values of ρ, it is evident that the GPPP method outperforms the PAPP, AIPW, LWP methods when the outcome model is wrong. This strength in GPPP is more evident when the association between the outcome and the PS is non-linear, i.e. in (b) CUB, (c) EXP, and (d) SIN. In Figure 4.7, I display corresponding comparisons for the binary outcome. The results are similar to those based on the continuous outcome, with a difference in that the values of rMSE increase with an increase in the value of ρ. Detailed numerical results of Simulation II is available in Appendix 4.6.4.

4.4 Application

I conduct an empirical study involving inference for a non-probability sample. The goal is to estimate police-reportable crash rates per 100M miles driven using the

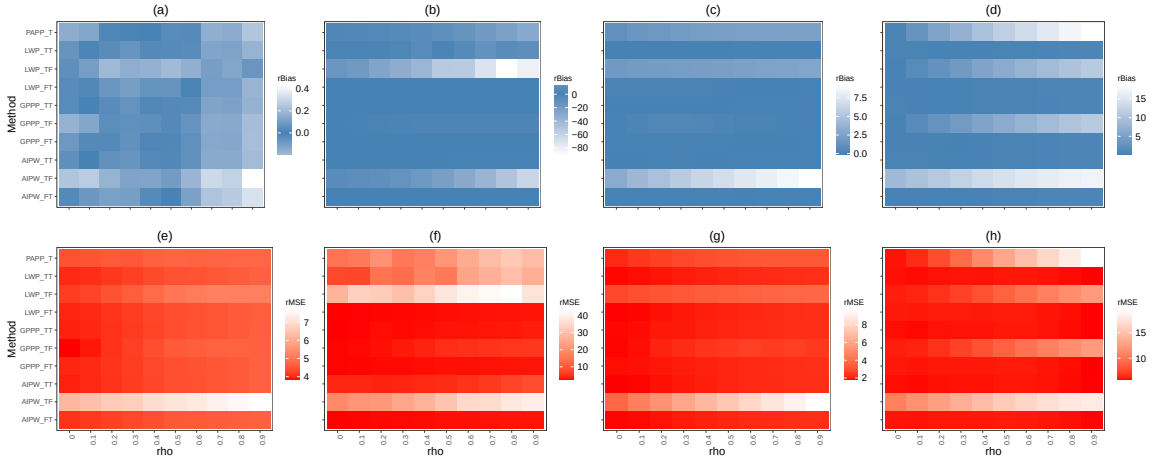

Figure 4.6: Comparing the magnitude of rBias of the DR adjusted means for the *continuous* outcome variable with $\gamma_2 = 0.3$ under (a) LIN, (b) CUB, (c) EXP, and (d) SIN, and rMSE under (e) LIN, (f) CUB, (g) EXP, and (h) SIN across different model specification scenarios and different values of ρ. UW: unweighted; FW: Fully weighted; PAPP: Propensity-adjusted Probability Prediction; GPPP: Gaussian Processes of Propensity Prediction; LWP: Linear-in-weight Prediction; AIPW: Augmented Inverse Propensity Weighting

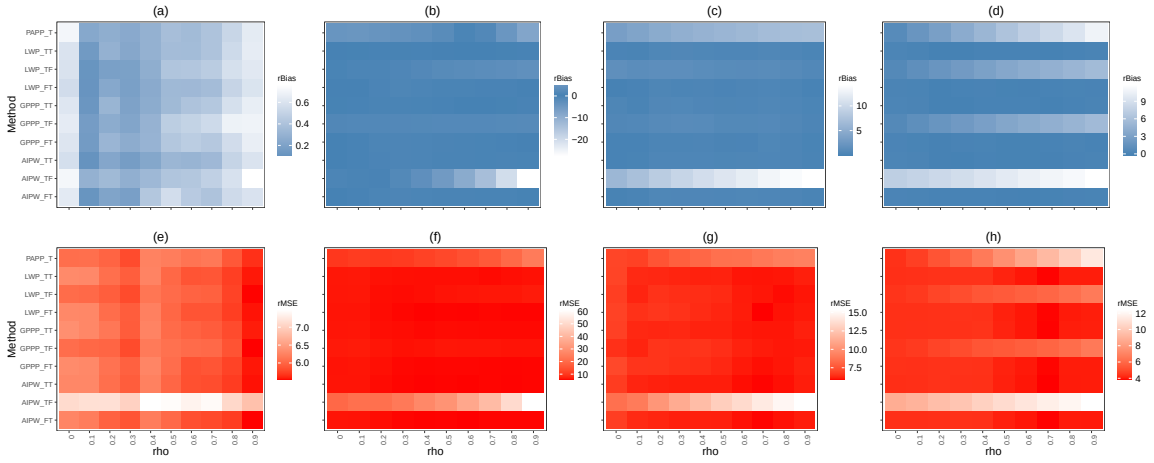

Figure 4.7: Comparing the magnitude of rBias of the DR adjusted means for the *binary* outcome variable with $\gamma_2 = 0.3$ under (a) LIN, (b) CUB, (c) EXP, and (d) SIN, and rMSE under (e) LIN, (f) CUB, (g) EXP, and (h) SIN across different model specification scenarios and different values of ρ. UW: unweighted; FW: Fully weighted; PAPP: Propensity-adjusted Probability Prediction; GPPP: Gaussian Processes of Propensity Prediction; LWP: Linear-in-weight Prediction; AIPW: Augmented Inverse Propensity Weighting

sensor-based data from the second phase of the Strategic Highway Research Program (SHRP2). To this end, I consider the National Household Travel Survey 2017 as the reference survey to adjust for the potential selection bias in crash rates. Chapter III elucidated the design of these samples in detail. In this application, however, I analyze the aggregated data at the individual level unlike Chapter III where inference was made based on the day-level data.

4.4.1 Auxiliary variables and analysis plan

To address the expressed objective of the present study, I set the outcome variable to be the frequency of police-reportable crashes by SHRP2 participants throughout their follow-up time. In addition, I utilize the total miles driven by each SHRP2 participant as the model offset to obtain the rates by a driven mile. Particular attention was paid to identify as many relevant common auxiliary variables as possible in the combined sample that are expected to govern both selection mechanism and response surface in SHRP2. Two distinct sets of variables were considered: (i) demographic and socio-economic information of the drivers including sex, age groups, race, ethnicity, birth country, education level, household size, number of owned vehicles, and state of residence, and (ii) vehicle characteristics including vehicle age, vehicle manufacturer, vehicle type and fuel type.

In order to make the two datasets more comparable, I filtered out all the subjects in NHTS who were not drivers or were younger than 16 years old or used public transportation or transportation modes other than cars, SUVs, vans, or light pickup trucks. The final sample sizes of the complete day-level datasets were $n_A = 2,862$ and $n_R = 29,572$ in SHRP2 and NHTS, respectively. I chose to use a Bayesian negative binomial (NB) regression for modeling the response surface because the outcome variable was count data and effects of overdispersion were present. I also checked and found no evidence of zero-inflation in the distribution of the outcome by comparing

the observed zeros with the expected number of zeros under the proposed NB model.

4.4.2 Results

According to Figure 4.8, one can visually infer that the largest discrepancies between the sample distribution of auxiliary variables in SHRP2 and that in the population stem from participants' age, race, and population size of the residential area as well as vehicles' age and vehicles' type. The youngest and oldest age groups are overrepresented as are Whites and non-Hispanics. In addition, I found that the proportion of urban dwellers is higher in SHRP2 than that in the NHTS. In terms of vehicle characteristics, SHRP2 participants tend to own passenger cars more than the population average, whereas individuals with other vehicle types were underrepresented in SHRP2.

Before any attempt for bias adjustment, I check the positivity assumption as well as the existence of influential pseudo-weights. To this end, I estimate the pseudo-selection probabilities for the units of the SHRP2 sample using the PAPP method as well as the PMLE method by Wang et al. (2020c). Figure 4.9 (a) compares the distribution of estimated PS in log scale between the SHRP2 and NHTS samples. As illustrated, there is a slight lack of common support in the distribution of PS, which may lead to extreme weights. The box-plot on the right side (Figure 4.9 (b)) confirms the presence of outlying pseudo-weights based on the PAPP method. However, it seems no outliers exist in the pseudo-weights based on the PMLE method. Figure 4.10 compares the distribution of auxiliary variables between the two samples after (pseudo-)weighting. As illustrated, pseudo-weighting obviates most of the previously seen discrepancies in the distribution of common covariates.

Figure 4.11 displays the adjusted estimates of police-reportable crash rates per 100M miles driven and associated 95% CIs using the LWP and GPPP methods by age groups. The plot also compares the adjusted estimates in SHRP2/NHTS data

with the naive estimate using SHRP2-only data and that based on the GES/ADS data, which is here considered as the benchmark Tefft (2017). Note that the latter represents the entire population of American drivers while our adjusted estimates represent the SHRP2 target population. As illustrated, for most of the age groups, adjustments shift the unweighted crash rates to the true population value, and the associated 95% CIs overlap, except for the last age group, i.e 80+ years old. In particular, the unweighted crash rate for the age group 50-59 years seems to be severely biased while adjusted estimates are desirably close to the true population value. While I observe no significant differences in the performance of the GPPP and LWP methods, it is evident that GPPP offers more efficient estimates than the LWP method, as the length of 95% CIs is consistently lower in GPPP than LWP. Finally, one can infer from Figure 4.11 that the risk of traffic accidents is higher among young and elder people.

In Figure 4.12, I assess the adjusted rates of police-reportable crashes across levels of auxiliary variables. The major associations I observe are as follows: Whites, more educated drivers, and those in middle-income families are at lower risk of traffic accidents. In addition, there is a positive relationship between the crash risk and household size. There is also evidence of higher crash rates among Vans, European, and gas/diesel vehicles. Numerical values associated with this plot have been provided in Table 4.15 of Appendix 4.6.5.

4.5 Discussion

The present chapter was an attempt to develop alternative Bayesian methods for inference based on non-probability samples that are robust and efficient. By robust, I mean a method that is less sensitive to misspecifying the functional form of the underlying models. In practice, the true models are almost always unknown, and double robustness does not offer a strong shield against model misspecification. By

efficiency, I mean a method that is not only smaller in variance, but also cheaper with respect to computational burden. More importantly, Bayesian approaches provide a unified framework for deriving the variance of the point estimator by simulating the posterior predictive distribution of the population's unknown parameters. A *well-calibrated* Bayesian method can appropriately capture all sources of uncertainty, meeting the desirable frequentist repeated sampling properties (Dawid, 1982). It is well-understood that joint modeling of the PS and the outcome, as was the case in our proposed method, results in good repeated sampling properties (Little, 2004).

The alternative design-based approaches, such as the AIPW estimator, are sensitive to the presence of influential pseudo-weights if the outcome model is invalid. In addition, the variance estimator proposed by Chen et al. (2019) relies on multiple asymptotic assumptions, and there is no guarantee that simultaneously solving the estimating equations leads to a unique solution. As another major limitation, such a method works only when the dimensions of the auxiliary variables are the same for the QR and PM models. According to the likelihood I factorized in Eq. 4.2, the dimension of the auxiliary variables may vary across the QR and PM methods in a non-probability sample setting ($\{X, D\}$ vs X), which can make it impossible to use Chen's AIPW method in practice. On the other hand, the existing fully model-based approaches can be extremely expensive computationally, as one needs to multiply impute the auxiliary variables for the non-sampled units of the population, and then fitting the models on each synthesized population separately (Little and Zheng, 2007; Mercer, 2018). However, the method I proposed requires fitting the model only once and on the combined sample, which makes it computationally more parsimonious, especially when dealing with Big Data.

The results of both simulation study and application reveal that our GPPP method is more efficient and less sensitive to the misspecification of the working models compared to the AIPW approach. As we observed in Simulation II, such a

method can offer extra robustness even if both underlying models are misspecified. While Bayesian joint modeling demonstrates good frequentist properties, feedback occurs between the two models (Zigler et al., 2013). This can be controversial in the sense that PS estimates should not be informed by the PM (Rubin, 2007). It is worth noting that given the currently available computational resources, Bayesian joint modeling based on the GPPP can still turn out computationally very expensive, even with the use of rank reduction techniques and for small-sized samples. This restriction made us rely on a Bayesian bootstrap approach instead of a real MCMC method to run the simulation and actual data analyses of this paper. However, increasing access and quality of high-performance computing resources may be able to overcome the computational burden of the proposed GPPP joint modeling approach.

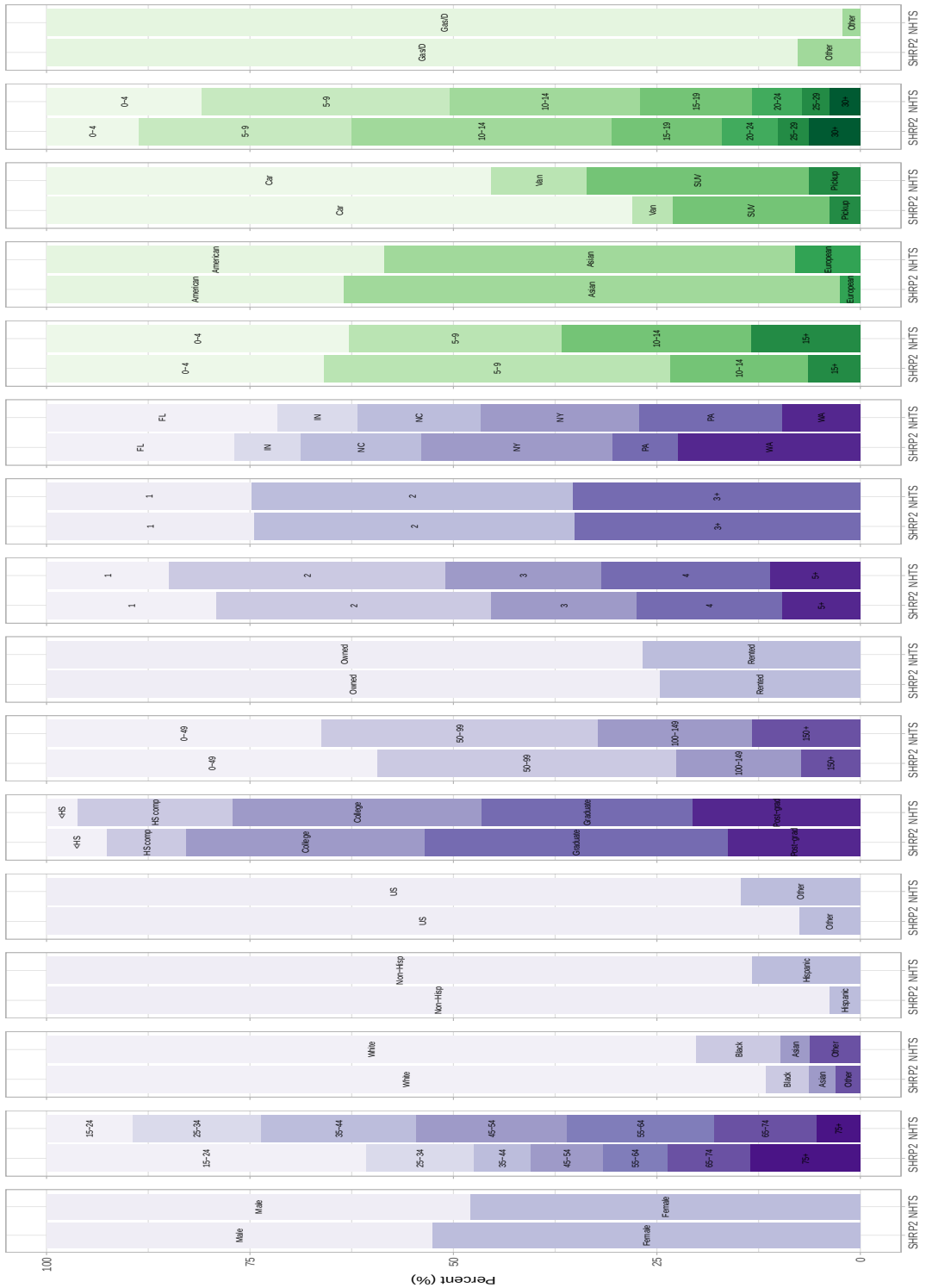

Figure 4.8: Comparing the distribution of common auxiliary variables in SHRP2 with weighted NHTS

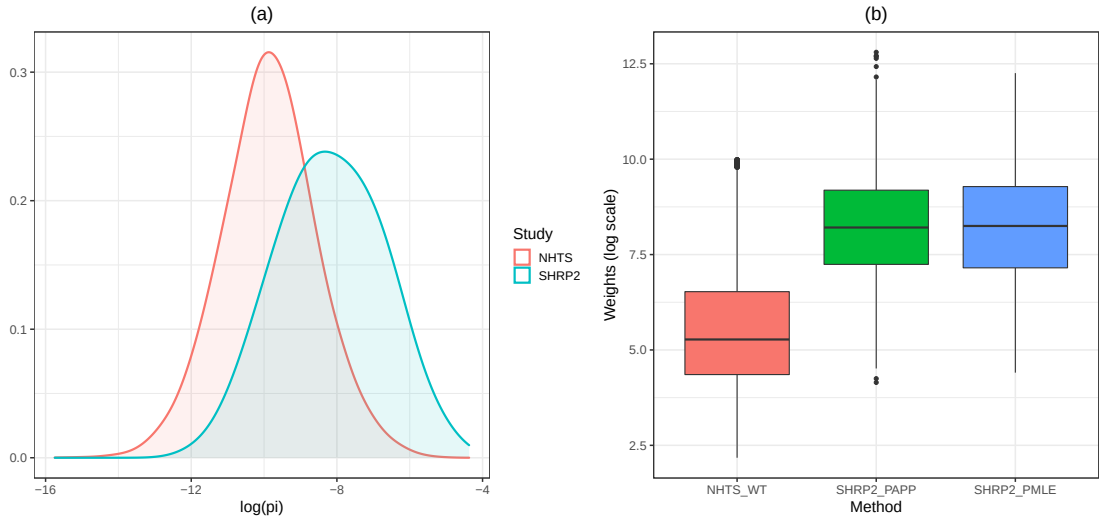

Figure 4.9: Comparing the empirical density of (a) estimated propensity scores between SHRP2 and NHTS and (b) estimated pseudo-weights in SHRP2 across the applied quasi-randomization methods

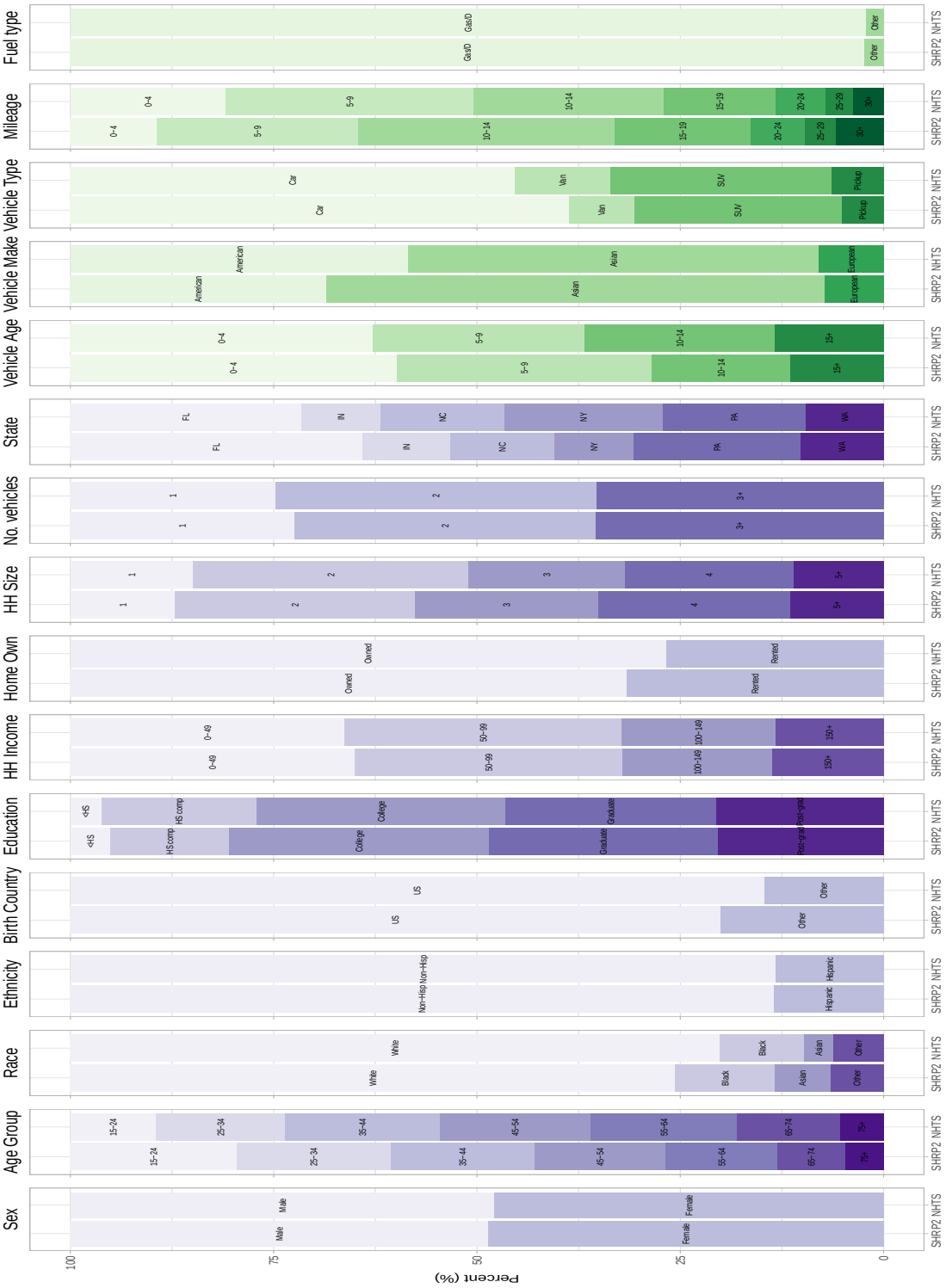

Figure 4.10: Comparing the distribution of common auxiliary variables in pseudo-weighted SHRP2 based on the PAPP method with weighted NHTS

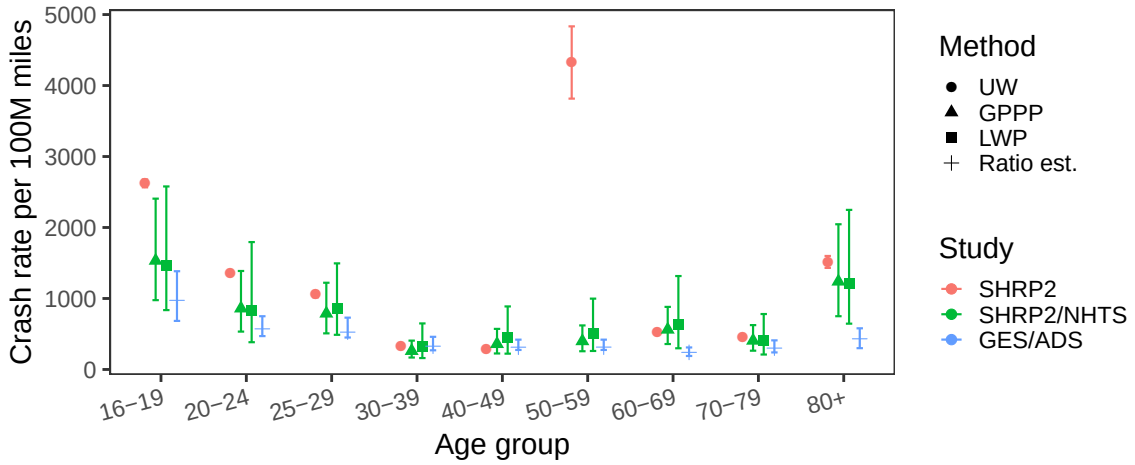

Figure 4.11: Comparing the performance of adjustment methods for estimating crash rates per 100M miles and associated 95% CIs in SHRP2/NHTS with native estimates and those based on CES/ADS as benchmark across age groups. UW: unweighted; FW: Fully weighted; PAPP: Propensity-adjusted Probability Prediction; GPPP: Gaussian Processes of Propensity Prediction; LWP: Linear-in-weight Prediction.

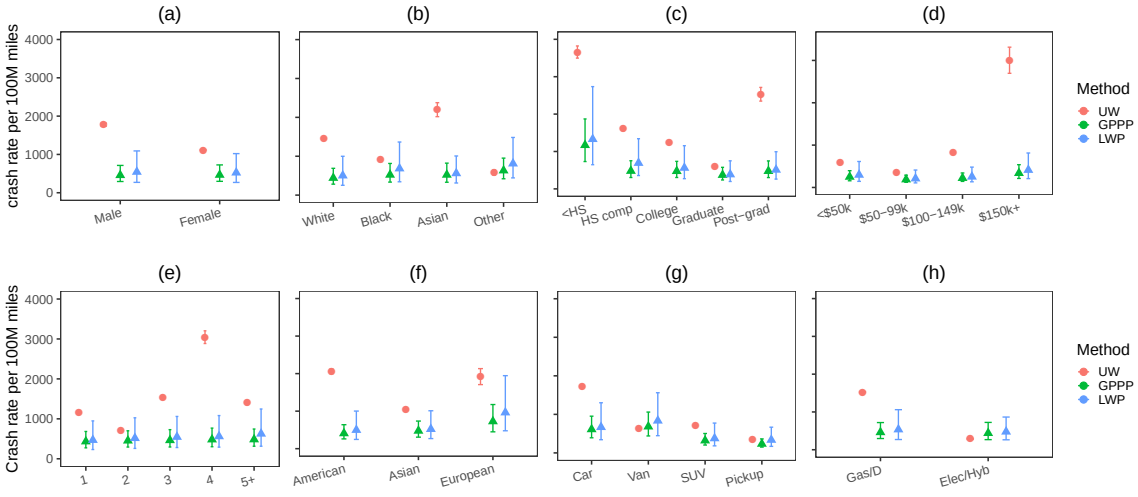

Figure 4.12: Comparing the performance of adjustment methods for estimating crash rates per 100M miles and associated 95% CIs in SHRP2/NHTS with native estimates across levels of (a) sex, (b) race, (c) education, (d) household income, (e) household size, (f) vehicle make, (g) vehicle type, and (d) fuel type. UW: unweighted; GPPP: Gaussian Process of Propensity Prediction; LWP: Linear-in-weight Prediction.

4.6 Appendix

4.6.1 Gaussian Processes and kernel weighting

Suppose $\hat{\pi}_i^A$ is the estimated PS for $i \in S_A$ based on a pseudo-weighting approach. Consider the following Gaussian Process (GP) regression model:

$$y_i = f(\hat{\pi}_i^A) + \epsilon_i \tag{4.40}$$

where $f \sim Gp(0, K)$ with $K(\pi_i^A, \pi_j^A; \alpha, \rho) = Cov\left(f(\pi_i^A), f(\pi_j^A)\right)$. From a weight-space viewpoint, one can show that the model 4.42 predicts y_i for $i \in S_R$ using a weighted sum of observed y_i in S_A as below:

$$\hat{y}_i = \sum_{j=1}^{n_A} \tilde{w}_{ij} y_j \tag{4.41}$$

where

$$\tilde{w}_{ij} = \frac{k_{ij}}{\sum_{j=1}^{n_A} k_{ij}} \quad and \quad k_{ij} = k^T(\hat{\pi}_j^A)\Sigma^{-1} \tag{4.42}$$

with $k(\hat{\pi}_j^A) = k(\hat{\pi}_j^A, \hat{\pi}_i^A)_{n_A \times 1}$. According to Huang et al. (2019), $\hat{y}_i$ can be regarded as the Nadaraya-Watson estimator of the observed outcome and selection indicator in the population.

Considering an isotropic covariance structure, which is a function of $||\hat{\pi}_j^A - \hat{\pi}_i^A||$, k_{ij} quite resembles the kernel weights Wang et al. (2020a), with the bandwidth h equivalent to the GP length-scale parameter ρ. Since the kernel weights obtained by GP is used in the PM estimator, it is clear that the final weights will be multiplied by w^R, i.e.

$$\hat{w}_j = \sum_{i=1}^{n_R} k_{ij} w_i^R \tag{4.43}$$

One can show that the major kernel-related condition determined by Wang et al. (2020a) to obtain consistency in the kernel-weighted estimates holds for a Matérn fam-

ily covariance structure, i.e. $K(x)$, $\int K(x)dx = 1$, $Sup_x|K(x)| < \infty$, and $lim_{|x|\to\infty}|x||K(x)| = 0$.

4.6.2 Partially linear Gaussian process regression

The main goal in the proposed GPPP method is to simulate the posterior predictive distribution of the outcome variable for units of S_C. Having $\pi_i^A = p(\delta_i^A = 1|x_i;\beta)$ estimated for $i \in S_C$ based on Eq. 4.18, I propose to fit a partially linear GP regression model on S_A as below:

$$y_i = \sum_{j=1}^{p+q+1} \theta_j z_{ij} + f(\hat{\pi}_i^A) + \epsilon_i \tag{4.44}$$

where θ denotes a $(p+q+1)$-dimensional vector of unknown linear regression parameters, $z_i = (1, x_i, d_i)$, f is an unknown function, and $\epsilon_i \sim N(0, \sigma^2)$. As illustrated, Eq. 4.44 consists of two parts: a linear regression parametrized by θ and a non-parametric regression denoted by $f(.)$.

In a GP regression model, I treat f a priori to follow an n_A-dimensional GP with mean 0 and an appropriately chosen covariance matrix as below:

$$f \sim Gp(0, K), \quad K(\pi_i^A, \pi_j^A; \alpha, \rho) = Cov\left(f(\pi_i^A), f(\pi_j^A)\right) \tag{4.45}$$

where K is an $n_A \times n_A$ covariance matrix taking a non-linear form with parameters (α, ρ). While there are a variety of recommended covariance structure for GP, in this section, I utilize the most popular covariance function in the GP literature, called squared exponential (SE), which is formulated as below:

$$K(\pi_i^A, \pi_j^A; \rho) = \alpha^2 exp\left\{ - \frac{||\pi_i^A - \pi_j^A||^2}{2\rho} \right\} \tag{4.46}$$

where α and ρ are often called the marginal standard error and length-scale parameters of the SE function, respectively. Note that the SE covariance function is a

special form of stationary isotropic functions as π_i^A and π_j^A depends only through their Euclidean distance, i.e. $||\pi_i^A - \pi_j^A||$ (Rusmassen and Williams, 2005).

Now, I follow Choi and Woo (2015) to fit the model in Eq. 4.44 on sample S_A. Let $f_A = [f(\pi_1^A), f(\pi_2^A), ..., f(\pi_{n_A}^A)]^T$ be a vector of covariance function values based on Eq. 4.45, evaluated at the n_A points of π_i^A. Then, the model can be re-written as:

$$Y_A|Z_A, \pi^A, f_A, \sigma^2, \theta, \alpha, \rho \sim N_{n_A}(Z_A\theta + f_A, \sigma^2 I_{n_A}) \tag{4.47}$$

where

$$f_A|\sigma^2, \theta, \alpha, \rho \sim N_{n_A}(0_{n_A}, K_{n_A}) \tag{4.48}$$

and subscript A points out that the observations are defined for $i \in S_A$ and N_{n_A} denotes an n_A-dimensional multivariate Gaussian distribution. Therefore, the posterior distribution of f_A is given by

$$\begin{aligned} p(f_A|Y_A, Z_A, \pi^A, \theta, \alpha, \rho) &\propto p(Y_A|Z_A, \pi^A, f_A, \sigma^2, \theta, \alpha, \rho)p(f_A|\sigma^2, \theta, \alpha, \rho) \\ &= N_{n_A}(Y_A - Z_A\theta, \sigma^2 I_{n_A}) \times N_{n_A}(0_{n_A}, K_{n_A}) \end{aligned} \tag{4.49}$$

Therefore, I have

$$f_A|Y_A, Z_A, \pi^A, \theta, \alpha, \rho \sim N_{n_A}(\mu_{n_A}, \Sigma_{n_A}) \tag{4.50}$$

where $\mu_{n_A} = K_{n_A}(K_{n_A} + \sigma^2 I_{n_A})^{-1}(Y_A - X_A\theta)$ and $\Sigma_{n_A} = \sigma^2 K_{n_A}(K_{n_A} + \sigma^2 I_{n_A})^{-1}$.

Now, considering normal priors for θ, i.e. $\theta \sim N_{p+q+1}(0_{p+q+1}, \Theta_0)$, the posterior distribution of θ is given by

$$\begin{aligned} p(\theta|Y_A, \sigma^2, f_A, \alpha, \rho) &\propto p(Y_A - f_A|\theta, \sigma^2, f_A)p(\theta) \\ Y_A - f_A|\theta, f_A, \sigma^2, \rho &\sim N_{n_A}(Z_A\theta, \sigma^2 I_{n_A}) \\ \theta|Y_A, f_A, \sigma^2, \rho &\sim N_{p+q+1}(\Theta_1 t, \Theta_1) \end{aligned} \tag{4.51}$$

where

$$\Theta_1^{-1} = \frac{1}{\sigma^2} Z_A^T Z_A + \Theta_0^{-1}, \quad t = \frac{1}{\sigma^2} Z_{n_A}^T (Y_A - f_A) \tag{4.52}$$

A conjugate prior for σ^2 is inverse-gamma distribution which is proportional to

$$\sigma^{-2(\gamma+1)} exp\{-\nu^{-1}\sigma^{-2}\}, \sigma^2 > 0 \tag{4.53}$$

where $\gamma > 0$ and $\nu > 0$ are two known hyperparameters. As a result, the full conditional distribution of σ^2 is given by

$$\begin{aligned} p(\sigma^2|Y_A, Z_A, \pi^A, f_A, \theta, \alpha, \rho) &\propto p(Y_A|Z_A, \pi^A, \sigma^2, \theta, \alpha, \rho) p(\sigma^2) \\ &\propto \frac{1}{\sigma^{n_A}} exp\Big\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n_A} \big[y_i - z_i^T\theta - f(\pi_i^A)\big]^2 \Big\} \\ &\times \sigma^{-2(\gamma+2)} exp\{-\nu^{-1}\sigma^{-2}\} \end{aligned} \tag{4.54}$$

which is an inverse-gamma distribution with parameters $\gamma + \frac{n_A}{2}$ and $\nu^{-1} + \frac{1}{2}\sum_{i=1}^{n_A}[y_i - z_i^T\theta - f(\pi_i^A)]^2$. The posterior distribution of the unknown parameters (f_A, θ, σ^2) can be simulated through Monte Carlo Markov Chains (MCMC).

4.6.3 Hilbert space approximation of Gaussian Processes

To reduce the GP computational burden while maintaining its accuracy, the current paper employs an approximation method proposed by Solin and Särkkä (2020), which can be implemented in Stan. Using the Laplace eigenfunctions for stationary covariance functions, this method approximates GP via a linear model by expanding the basis functions. Riutort-Mayol et al. (2020) examine the performance of this approach in several simulation and empirical studies with an attempt to identify optimal values for its tuning parameters. In the following, I briefly describe this approach for GPs with a Matérn covariance function through mathematical notations.

A stationary covariance function can be expressed uniquely with respect to a spec-

tral density function. The latter is a frequency domain representation of a stationary process, which constitutes a Fourier transform pair with the process autocovariance. One can show that the spectral density function for the Matérn covariance function is given by

$$S_\nu(x) = \alpha^2 \frac{2\pi^{1/2}\Gamma(\nu + 1/2)(2\nu)^\nu}{\Gamma(\nu)\rho^{2\nu}} \left(\frac{2\nu}{\rho^2} + 4\pi^2 x^2\right)^{\nu+1/2} \tag{4.55}$$

where $x \in \mathbb{R}$ denotes the frequency, and ρ and α are the lengthscale and marginal standard error of the kernel, respectively. For $\nu = \infty$ and $\nu = 3/2$, this function is reduced to

$$\begin{aligned} S_\infty(x) &= \alpha^2\sqrt{2\pi}\rho exp\left(-\rho^2 x^2/2\right) \\ S_{3/2}(x) &= \alpha^2 \frac{2\pi^{1/2}3^{3/2}}{\sqrt{\pi}\rho^3/2} \left(\frac{3}{\rho^2} + x^2\right)^2 \end{aligned} \tag{4.56}$$

Suppose the GP input space is given by $\Omega = [-L, L]$, where Riutort-Mayol et al. (2020) refer to $L \in \mathbb{R}^{>0}$ as the boundary condition. Within Ω, one can expand a given stationary covariance function linearly as

$$k(x, x') = \sum_{j=1}^{\infty} S(\sqrt{\lambda_j})\phi_j(x)\phi_j(x') \tag{4.57}$$

where $\{x, x'\} \in \Omega$, and $\{\lambda_j\}_{j=1}^\infty$ and $\{\phi(x)\}_{j=1}^\infty$ denote the sets of eigenvalues and eigenvectors of the Laplacian operator in the given domain, respectively. By applying the Dirichlet boundary, this implies the following eigenvalue problem in Ω :

$$\begin{aligned} -\Delta^2\phi_j(x) &= \lambda\phi_j(x), \qquad && x \in \Omega \\ \phi_j(x) &= 0, && x \notin \Omega \end{aligned} \tag{4.58}$$

Since the Laplacian is a positive definite Hermitian operator, the eigenvalues are real and positive, i.e. $\lambda_j > 0$. In addition, the eigenfunctions ϕ_j take a sinusoidal form,

and are given by

$$\begin{aligned} \lambda_j &= \left(\frac{j\pi}{2L}\right)^2 \\ \phi_j(x) &= \sqrt{\frac{1}{L}} sin\left(\sqrt{\lambda_j(x+L)}\right) \end{aligned} \tag{4.59}$$

Note that the solution to the eigenvalue problem is independent of the specific choice of covariance function. Now, one can approximate the covariance function by truncating the sum in Eq. 4.57 to the first m terms as below:

$$k(x,x') \approx \sum_{j=1}^{m} S(\sqrt{\lambda_j})\phi_j(x)\phi_j(x') = \phi(x)^T \Delta \phi(x') \tag{4.60}$$

where $\phi(x) = \{\phi_j(x)\}_{j=1}^m \in \mathbb{R}^m$ is the vector of basis functions, and $\Delta \in \mathbb{R}^{m\times m}$ denotes a diagonal matrix of the spectral density evaluated at the square root of the eigenvalues, that is, $S(\sqrt{\lambda_j})$,

$$\Delta = \begin{bmatrix} S(\sqrt{\lambda_1}) & & \\ & \ddots & \\ & & \sqrt{\lambda_m}) \end{bmatrix} \tag{4.61}$$

Thus, the Gram matrix K of the covariance function k for a set of observations $i = 1, ..., n$ and corresponding input values $\{x_i\}_{i=1}^n \in \Omega^n$ can be represented as

$$K = \Phi \Delta \Phi^T \tag{4.62}$$

where $\Phi \in \mathbb{R}^{n\times m}$ is the matrix of eigenfunctions $\phi_j(x_i)$

$$\Phi = \begin{bmatrix} \phi_1(x_1) & \dots & \phi_m(x_1) \\ \vdots & \ddots & \vdots \\ \phi_1(x_n) & \dots & \phi_m(x_n) \end{bmatrix}. \tag{4.63}$$

As a result, the model for f can be written as

$$f \sim N(\mu, \Phi\Delta\Phi^T) \tag{4.64}$$

This equivalently leads to a linear representation of f via

$$f(x) \approx \sum_{j=1}^{m} S^{\frac{1}{2}}(\sqrt{\lambda_j})\phi_j(x)\beta_j \tag{4.65}$$

where $\beta_j \sim N(0, 1)$. Therefore, Riutort-Mayol et al. (2020) approximate the function f with a finite basis function expansion, scaled by the square root of spectral density values.

As a key feature of this approximation, the eigenfunctions ϕ_j are independent of the parameters of the covariance function, i.e. (α, ρ). For a bounded covariance function, $S(.)$ goes rapidly to zero as j increases, because λ_j's are monotonically incremental with j. Note that this approximation yields a computational cost of $O(nm + m)$ for evaluating the log posterior density of a univariate GP (which is the case in this study), where n is the number of observations and m the number of basis functions (Riutort-Mayol et al., 2020). In the present study, I set $m = 10$ and $L = C \times max\{|min(x)|, |max(x)|\}$ with $C = 1.25$, which shows relatively good empirical results.

4.6.4 Further extensions of the simulation study

4.6.4.1 Simulation study I

This subsection provides additional results associated with Simulation I. Table 4.2 and Table 4.3 summarize the findings of the simulation in 4.3.1 for $\rho = 0.5$ and $\rho = 0.3$, respectively.

Table 4.2: Comparing the performance of the bias adjustment methods in the first simulation study for $\rho = 0.5$

	$n_A = 500, \quad n_R = 500$					$n_A = 1,000, \quad n_R = 500$					$n_A = 500, \quad n_R = 1,000$				
Measure	rBias	rMSE	crCI	lCI	rSE	rBias	rMSE	crCI	lCI	rSE	rBias	rMSE	crCI	lCI	rSE
Probability sample (S_R)															
UW	8.866	9.445	21.759	1.149	0.965	8.866	9.445	21.759	1.149	0.965	8.810	9.115	2.778	0.816	0.954
FW	0.260	3.770	93.056	1.363	0.991	0.260	3.770	93.056	1.363	0.991	0.080	2.736	95.833	0.967	0.967
Non-probability sample (S_A)															
UW	30.708	30.926	0.000	1.304	0.974	30.008	30.104	0.000	0.916	1.039	30.708	30.926	0.000	1.304	0.974
FW	-0.134	3.763	93.981	1.363	0.991	-0.146	2.659	95.833	0.957	0.986	-0.134	3.763	93.981	1.363	0.991
Model specification: QR–True, PM–True															
GPPP	0.129	3.876	99.537	1.912	1.350	0.000	3.055	99.537	1.724	1.544	0.092	3.567	98.148	1.576	1.209
LWP	-0.063	4.200	99.537	2.516	1.638	-0.145	3.140	99.537	1.737	1.514	-0.135	3.742	97.685	1.613	1.179
AIPW	-0.101	4.035	93.981	1.406	0.953	-0.098	3.139	94.907	1.107	0.965	-0.232	3.760	94.907	1.311	0.955
PAPP	0.832	4.012	93.056	1.406	0.980	0.547	2.998	93.981	1.078	1.000	0.984	3.844	95.370	1.339	0.985
Model specification: QR–True, PM–False															
GPPP	0.146	3.829	99.537	1.903	1.360	-0.017	3.068	99.537	1.720	1.533	0.161	3.555	98.148	1.575	1.213
LWP	-0.007	3.876	99.537	1.910	1.348	-0.124	3.084	99.537	1.730	1.535	-0.015	3.575	98.148	1.585	1.213
AIPW	-0.056	3.916	94.907	1.358	0.948	-0.087	3.061	95.833	1.070	0.957	-0.180	3.590	93.981	1.272	0.970
Model specification: QR–False, PM–True															
GPPP	1.317	4.321	99.074	1.934	1.285	1.146	3.477	99.537	1.748	1.456	1.314	4.001	96.296	1.607	1.163
LWP	4.052	6.752	89.815	2.236	1.132	4.147	6.026	92.130	2.004	1.253	4.090	6.299	87.037	1.850	1.056
AIPW	0.040	3.987	95.833	1.392	0.954	-0.010	3.025	94.444	1.074	0.971	0.073	3.716	93.519	1.338	0.984
Model specification: QR–False, PM–False															
GPPP	27.400	27.678	0.000	2.239	1.564	26.595	26.733	0.000	2.042	2.055	27.430	27.692	0.000	1.814	1.306
LWP	27.075	27.361	0.000	2.247	1.558	26.434	26.575	0.000	2.056	2.053	27.127	27.395	0.000	1.837	1.314
AIPW	27.115	27.402	0.000	1.366	0.944	26.432	26.571	0.000	0.981	0.987	27.056	27.328	0.000	1.334	0.949
PAPP	27.912	28.191	0.000	1.354	0.936	27.024	27.158	0.000	0.968	0.983	28.121	28.381	0.000	1.329	0.950

GPPP: Gaussian Process of Propensity Prediction; LWP: Linear-in-weight prediction; AIPW: Augmented Inverse Propensity Weighting; PAPP: Propensity-adjusted Probability Prediction
NOTE: The PAPP and AIPW methods have been implemented through a bootstrap method.

Table 4.3: Comparing the performance of the bias adjustment methods in the first simulation study for $\rho = 0.3$

	$n_A = 500, \quad n_R = 500$					$n_A = 1,000, \quad n_R = 500$					$n_A = 500, \quad n_R = 1,000$				
Measure	rBias	rMSE	crCI	lCI	rSE	rBias	rMSE	crCI	lCI	rSE	rBias	rMSE	crCI	lCI	rSE
Probability sample (S_R)															
UW	8.867	10.406	58.796	1.927	0.965	8.867	10.406	58.796	1.927	0.965	8.797	9.612	38.889	1.369	0.964
FW	0.424	6.361	92.593	2.303	0.989	0.424	6.361	92.593	2.303	0.989	0.154	4.580	95.370	1.633	0.973
Non-probability sample (S_A)															
UW	30.759	31.251	0.000	2.037	1.006	30.082	30.322	0.000	1.435	1.028	30.759	31.251	0.000	2.037	1.006
FW	-0.277	6.284	95.833	2.335	1.014	-0.298	4.529	93.519	1.639	0.989	-0.277	6.284	95.833	2.335	1.014
Model specification: QR–True, PM–True															
GPPP	0.438	6.297	100.000	3.413	1.482	0.061	4.754	99.537	3.094	1.775	0.489	6.052	99.074	2.821	1.275
LWP	-0.342	6.811	100.000	4.061	1.628	-0.266	4.917	99.537	3.131	1.739	-0.232	6.429	98.148	2.932	1.244
AIPW	-0.390	6.666	95.370	2.378	0.975	-0.296	4.911	94.444	1.737	0.966	-0.395	6.337	95.370	2.328	1.004
PAPP	0.574	6.537	95.370	2.366	0.991	0.360	4.696	95.370	1.709	0.995	0.873	6.291	94.907	2.334	1.022
Model specification: QR–True, PM–False															
GPPP	0.553	6.241	100.000	3.391	1.488	0.077	4.733	99.537	3.080	1.775	0.585	6.033	98.611	2.817	1.280
LWP	-0.115	6.302	100.000	3.464	1.499	-0.255	4.748	99.537	3.105	1.786	-0.020	6.114	99.074	2.879	1.284
AIPW	-0.271	6.301	95.370	2.312	1.002	-0.275	4.789	94.444	1.688	0.963	-0.287	6.042	94.907	2.233	1.009
Model specification: QR–False, PM–True															
GPPP	1.891	6.750	99.537	3.409	1.435	1.372	5.092	99.537	3.097	1.722	1.981	6.536	97.685	2.837	1.242
LWP	4.179	8.638	98.611	3.692	1.332	4.373	7.326	99.537	3.309	1.535	4.277	8.280	95.833	3.098	1.192
AIPW	-0.146	6.594	94.907	2.377	0.983	-0.187	4.794	94.444	1.708	0.973	-0.125	6.319	95.370	2.333	1.007
Model specification: QR–False, PM–False															
GPPP	27.642	28.234	5.093	3.483	1.652	26.734	27.051	2.315	3.212	2.121	27.792	28.364	0.463	2.792	1.344
LWP	26.915	27.545	11.111	3.542	1.649	26.278	26.601	2.315	3.242	2.142	27.022	27.636	1.852	2.900	1.366
AIPW	26.987	27.603	0.000	2.120	0.997	26.405	26.734	0.000	1.528	0.996	27.022	27.630	0.000	2.076	0.982
PAPP	27.881	28.473	0.000	2.122	1.002	26.994	27.300	0.000	1.501	1.004	28.146	28.716	0.000	2.104	1.008

GPPP: Gaussian Process of Propensity Prediction; LWP: Linear-in-weight prediction; AIPW: Augmented Inverse Propensity Weighting; PAPP: Propensity-adjusted Probability Prediction
NOTE: The PAPP and AIPW methods have been implemented through a bootstrap method.

4.6.4.2 Simulation study II

Table 4.4 exhibits the numerical results associated with Figure 4.2 and Figure 4.3.

Table 4.4: Comparing the performance of the bias adjustment methods in the second simulation study for the *continuous* outcome with $(n_A, n_R) = (500, 1,000)$ and $\gamma_1 = 0.3$

	LIN					*CUB*					*EXP*					*SIN*				
Measure	rBias	rMSE	crCI	lCI	rSE	rBias	rMSE	crCI	lCI	rSE	rBias	rMSE	crCI	lCI	rSE	rBias	rMSE	crCI	lCI	rSE
Probability sample (S_R)																				
UW	-23.870	24.078	0.000	12.411	1.001	-17.120	17.243	0.000	8.442	1.045	-12.097	12.179	0.000	5.427	0.976	-18.475	19.159	3.241	18.535	0.93
FW	-0.180	3.921	93.519	15.087	0.980	-0.156	2.749	94.907	10.799	1.001	-0.091	1.947	94.444	7.392	0.967	0.015	5.549	93.519	20.503	0.94
Non-probability sample (S_A)																				
UW	48.795	49.086	0.000	19.644	0.936	30.073	30.366	0.000	15.989	0.966	18.146	18.363	0.000	10.710	0.969	27.948	28.716	0.926	25.807	0.995
FW	-0.178	6.184	91.204	22.458	0.925	-0.218	4.249	92.593	15.612	0.936	0.004	2.282	93.981	8.639	0.964	-0.034	8.276	93.519	31.425	0.966
Non-robust method																				
Model specification: QR–True																				
PAPP	0.066	4.771	95.370	19.982	1.066	-7.545	24.919	91.667	53.569	0.574	2.269	3.182	83.333	8.794	1.003	11.642	13.063	46.759	23.010	0.988
Model specification: QR–False																				
PAPP	35.346	35.660	0.000	17.238	0.929	17.490	17.760	0.000	11.878	0.980	10.111	10.362	0.000	8.707	0.977	30.831	31.610	0.463	26.668	0.973
Doubly robust methods																				
Model specification: QR–True, PM–True																				
GPPP	0.025	4.439	98.148	18.703	1.081	-0.060	3.150	96.759	13.124	1.067	-0.027	2.164	97.222	8.823	1.048	0.301	6.250	96.296	26.369	1.084
LWP	-0.038	4.478	97.222	18.960	1.088	-0.005	3.133	98.148	13.227	1.084	-2.254	4.458	95.833	9.065	0.080	0.299	6.335	96.296	26.621	1.076
AIPW	0.000	4.439	97.222	18.266	1.047	0.250	4.060	98.148	16.137	1.014	-0.022	2.194	93.981	8.482	0.984	0.157	6.179	94.444	23.568	0.971
Model specification: QR–True, PM–False																				
GPPP	0.099	4.467	97.685	18.732	1.078	-0.061	3.136	97.222	12.796	1.047	0.005	2.140	96.296	8.654	1.037	0.439	6.189	95.833	25.041	1.043
LWP	10.099	4.776	91.667	17.809	0.100	7.636	3.382	88.889	12.078	0.087	5.845	17.130	89.815	8.171	0.078	7.886	14.740	92.130	24.236	0.175
AIPW	-0.033	4.608	94.907	18.233	1.007	-0.109	3.198	95.833	12.461	0.992	-0.048	2.188	95.370	8.400	0.977	0.245	6.221	94.907	23.752	0.972
Model specification: QR–False, PM–True																				
GPPP	0.523	4.628	97.685	20.143	1.128	1.310	3.829	95.370	15.280	1.093	0.575	2.531	96.296	11.294	1.175	0.819	6.523	95.833	27.890	1.108
LWP	-0.294	4.923	96.296	20.356	1.064	0.231	6.026	93.981	15.853	0.677	0.509	5.817	96.759	11.564	0.512	-1.060	14.048	92.130	37.060	0.679
AIPW	0.546	6.979	94.444	29.210	1.069	-15.872	46.090	87.500	102.060	0.600	6.885	7.391	26.389	10.549	0.999	14.490	15.884	35.648	24.548	0.96
Model specification: QR–False, PM–False																				
GPPP	36.369	36.676	0.000	22.484	1.217	18.222	18.524	0.000	15.525	1.195	10.492	10.758	0.926	10.941	1.182	30.956	31.750	3.241	33.560	1.222
LWP	36.049	36.361	0.000	22.546	1.219	18.396	18.690	0.000	15.611	1.213	10.590	10.856	0.926	11.005	1.185	30.650	31.450	2.778	35.040	1.276
AIPW	36.146	36.468	0.000	17.448	0.919	18.293	18.583	0.000	12.436	0.967	10.587	10.848	0.000	8.968	0.965	30.664	31.483	0.000	26.807	0.956

UW: Unweighted; FW: Fully weighted; GPPP: Gaussian Process of Propensity Prediction; LWP: Linear-in-weight prediction; AIPW: Augmented Inverse Propensity Weighting.
NOTE: The PAPP and AIPW methods have been implemented through a bootstrap method.

Table 4.5: Comparing the performance of the bias adjustment methods in the second simulation study for the *continuous* outcome with $(n_A, n_R) = (1,000, 500)$ and $\gamma_1 = 0.3$

	LIN					*CUB*					*EXP*					*SIN*				
Measure	rBias	rMSE	crCI	lCI	rSE	rBias	rMSE	crCI	lCI	rSE	rBias	rMSE	crCI	lCI	rSE	rBias	rMSE	crCI	lCI	rSE
Probability sample (S_R)																				
UW	-23.868	24.297	0.000	17.587	0.985	-17.158	17.440	0.000	11.959	0.974	-12.107	12.274	0.000	7.683	0.967	-18.706	19.966	22.222	26.219	0.956
FW	0.118	5.477	93.519	21.395	0.994	0.016	3.998	94.444	15.257	0.971	0.015	2.673	95.833	10.454	0.995	-0.081	7.162	96.296	29.064	1.033
Non-probability sample (S_A)																				
UW	48.581	48.727	0.000	13.878	0.935	29.971	30.129	0.000	11.296	0.933	18.042	18.159	0.000	7.569	0.935	27.647	28.037	0.000	18.285	0.999
FW	0.056	4.211	95.833	15.872	0.959	0.107	2.862	94.907	10.960	0.975	0.055	1.586	96.296	6.127	0.984	-0.297	5.417	95.833	22.208	1.045
Non-robust method																				
Model specification: QR–True																				
PAPP	0.504	4.495	96.296	18.667	1.064	-16.241	32.851	98.611	74.882	0.667	0.403	1.995	95.833	7.859	1.024	11.569	13.086	53.241	24.302	1.011
Model specification: QR–False																				
PAPP	35.317	35.521	0.000	14.654	0.982	17.616	17.838	0.000	10.819	0.982	10.136	10.324	0.000	7.535	0.977	30.431	30.802	0.000	18.895	1.008
Doubly robust methods																				
Model specification: QR–True, PM–True																				
GPPP	0.326	4.758	98.611	21.123	1.142	0.169	3.505	96.759	14.875	1.091	0.120	2.338	96.759	10.083	1.108	0.291	6.396	96.759	29.612	1.186
LWP	0.285	4.746	98.611	21.279	1.152	-1.313	3.504	96.759	16.475	0.196	-0.655	8.930	96.296	10.952	0.316	0.246	6.403	97.685	29.663	1.19
AIPW	0.287	4.808	97.222	20.161	1.069	0.767	5.589	97.222	20.610	0.947	0.140	2.361	96.296	9.838	1.062	0.292	6.293	95.370	26.295	1.065
Model specification: QR–True, PM–False																				
GPPP	0.397	4.731	97.685	21.125	1.151	0.123	3.530	97.222	14.813	1.072	0.096	2.337	98.148	10.056	1.097	0.340	6.358	97.685	29.112	1.175
LWP	21.964	5.322	84.259	19.456	0.077	16.925	4.571	81.019	13.776	0.071	12.837	31.570	78.241	9.029	0.061	17.658	14.123	85.648	26.438	0.131
AIPW	0.430	4.877	97.685	20.009	1.048	0.153	3.531	95.833	14.218	1.026	0.108	2.359	96.759	9.660	1.043	0.357	6.324	95.370	26.422	1.065
Model specification: QR–False, PM–True																				
GPPP	0.679	4.644	98.611	22.199	1.242	1.028	3.562	98.611	16.651	1.251	0.538	2.275	99.074	12.101	1.401	0.734	6.435	98.148	30.583	1.228
LWP	0.107	4.712	99.074	22.396	1.223	-1.191	3.593	97.222	17.681	0.335	0.118	2.437	99.537	12.407	1.310	-2.162	8.254	96.759	38.983	1.248
AIPW	1.102	6.334	95.833	25.788	1.052	-33.598	67.410	95.833	153.677	0.669	6.226	6.774	39.815	10.775	1.028	14.128	15.372	36.111	24.900	1.046
Model specification: QR–False, PM–False																				
GPPP	36.360	36.574	0.000	24.865	1.615	18.538	18.801	0.000	17.376	1.427	10.648	10.868	0.463	12.232	1.443	30.417	30.823	1.852	36.965	1.893
LWP	36.137	36.354	0.000	24.980	1.612	18.636	18.902	0.000	17.596	1.426	10.702	10.920	0.926	12.309	1.452	30.259	30.656	0.926	38.086	1.988
AIPW	36.248	36.471	0.000	15.439	0.976	18.546	18.800	0.000	12.121	1.003	10.730	10.951	0.000	8.629	1.005	30.357	30.749	0.000	19.215	0.999

UW: Unweighted; FW: Fully weighted; GPPP: Gaussian Process of Propensity Prediction; LWP: Linear-in-weight Prediction; AIPW: Augmented Inverse Propensity Weighting.
NOTE: The PAPP and AIPW methods have been implemented through a bootstrap method.

Table 4.6: Comparing the performance of the bias adjustment methods in the second simulation study for the *continuous* outcome with $(n_A, n_R) = (500, 500)$ and $\gamma_1 = 0.3$

	LIN					*CUB*					*EXP*					*SIN*				
Measure	rBias	rMSE	crCI	lCI	rSE	rBias	rMSE	crCI	lCI	rSE	rBias	rMSE	crCI	lCI	rSE	rBias	rMSE	crCI	lCI	rSE
Probability sample (S_R)																				
UW	-23.868	24.297	0.000	17.587	0.985	-17.158	17.440	0.000	11.959	0.974	-12.107	12.274	0.000	7.683	0.967	-18.706	19.966	22.222	26.219	0.956
FW	0.118	5.477	93.519	21.395	0.994	0.016	3.998	94.444	15.257	0.971	0.015	2.673	95.833	10.454	0.995	-0.081	7.162	96.296	29.064	1.033
Non-probability sample (S_A)																				
UW	48.795	49.086	0.000	19.644	0.936	30.073	30.366	0.000	15.989	0.966	18.146	18.363	0.000	10.710	0.969	27.948	28.716	0.926	25.807	0.995
FW	-0.178	6.184	91.204	22.458	0.925	-0.218	4.249	92.593	15.612	0.936	0.004	2.282	93.981	8.639	0.964	-0.034	8.276	93.519	31.425	0.966
Non-robust method																				
Model specification: QR–True																				
PAPP	0.487	5.489	94.907	22.317	1.039	-10.305	26.492	94.444	62.285	0.650	1.287	2.842	91.667	9.682	0.973	11.663	13.627	60.648	27.601	0.997
Model specification: QR–False																				
PAPP	35.408	35.754	0.000	18.412	0.945	17.598	17.930	0.000	13.143	0.974	10.135	10.422	0.000	9.452	0.991	30.805	31.598	0.463	26.762	0.968
Doubly robust methods																				
Model specification: QR–True, PM–True																				
GPPP	0.359	5.406	97.222	23.039	1.098	0.147	3.855	96.296	16.208	1.080	0.112	2.546	97.685	10.919	1.105	0.235	7.323	96.296	32.247	1.129
LWP	0.293	5.384	97.222	23.243	1.108	0.130	3.860	96.759	16.270	1.083	0.060	2.595	97.222	11.233	1.111	0.306	7.389	96.296	32.653	1.133
AIPW	0.292	5.410	96.296	22.264	1.049	0.570	4.891	97.222	20.056	1.051	0.100	2.610	96.759	10.631	1.037	0.311	7.290	96.296	29.106	1.017
Model specification: QR–True, PM–False																				
GPPP	0.402	5.359	96.759	23.118	1.110	0.149	3.926	95.370	15.959	1.039	0.137	2.580	96.759	10.813	1.077	0.389	7.230	97.685	31.411	1.112
LWP	10.150	5.659	91.204	22.008	0.126	7.552	4.184	88.426	14.928	0.110	5.872	23.999	88.426	10.056	0.098	7.821	17.007	91.204	29.704	0.214
AIPW	0.291	5.475	95.370	22.112	1.029	0.136	3.982	95.833	15.500	0.991	0.086	2.662	95.370	10.677	1.022	0.156	7.197	94.907	28.904	1.022
Model specification: QR–False, PM–True																				
GPPP	0.879	5.451	97.685	24.471	1.169	1.589	4.338	95.833	18.450	1.172	0.715	2.775	96.759	13.472	1.290	0.827	7.547	96.296	33.826	1.156
LWP	-1.094	5.826	97.685	24.830	0.530	-1.064	4.293	96.759	19.139	0.535	0.132	2.775	98.148	13.657	1.262	-1.868	11.807	94.444	43.580	0.96
AIPW	1.226	7.554	94.444	31.434	1.073	-20.691	52.415	92.130	121.585	0.643	6.562	7.243	43.981	12.248	1.017	14.329	16.096	50.000	28.830	1.001
Model specification: QR–False, PM–False																				
GPPP	36.537	36.882	0.000	27.271	1.392	18.359	18.731	0.000	19.077	1.320	10.586	10.905	8.333	13.442	1.317	30.765	31.605	9.259	40.674	1.443
LWP	36.191	36.544	0.000	27.447	1.391	18.536	18.894	0.000	19.235	1.350	10.681	10.998	7.870	13.431	1.314	30.488	31.305	9.259	42.309	1.529
AIPW	36.291	36.644	0.000	19.219	0.964	18.489	18.836	0.000	14.274	1.010	10.726	11.039	0.463	10.142	0.989	30.643	31.452	0.000	26.975	0.968

UW: Unweighted; FW: Fully weighted; GPPP: Gaussian Process of Propensity Prediction; LWP: Linear-in-weight Prediction; AIPW: Augmented Inverse Propensity Weighting.
NOTE: The PAPP and AIPW methods have been implemented through a bootstrap method.

Table 4.7: Comparing the performance of the bias adjustment methods in the second simulation study for the *continuous* outcome with $(n_A, n_R) = (500, 1,000)$ and $\gamma_1 = 0.6$

	LIN					*CUB*					*EXP*					*SIN*				
Measure	rBias	rMSE	crCI	lCI	rSE	rBias	rMSE	crCI	lCI	rSE	rBias	rMSE	crCI	lCI	rSE	rBias	rMSE	crCI	lCI	rSE
Probability sample (S_R)																				
UW	-23.870	24.078	0.000	12.411	1.001	-17.120	17.243	0.000	8.442	1.045	-12.097	12.179	0.000	5.427	0.976	-18.475	19.159	3.241	18.535	0.93
FW	-0.180	3.921	93.519	15.087	0.980	-0.156	2.749	94.907	10.799	1.001	-0.091	1.947	94.444	7.392	0.967	0.015	5.549	93.519	20.503	0.94
Non-probability sample (S_A)																				
UW	100.237	100.399	0.000	20.820	0.929	73.900	74.114	0.000	21.935	0.992	44.946	45.091	0.000	13.875	0.978	36.869	37.411	0.000	25.372	1.018
FW	-0.428	10.721	92.130	40.191	0.955	-0.515	8.917	91.667	27.339	0.782	-0.070	3.570	92.593	13.283	0.947	-0.451	12.399	94.907	49.982	1.027
Non-robust method																				
Model specification: QR–True																				
PAPP	0.267	8.677	94.444	36.302	1.065	1.511	27.294	48.148	46.558	0.435	5.295	5.987	49.537	10.631	0.969	27.459	28.131	0.926	22.936	0.955
Model specification: QR–False																				
PAPP	52.850	53.104	0.000	18.790	0.921	24.244	24.516	0.000	13.272	0.928	15.205	15.440	0.000	9.782	0.928	54.554	55.056	0.000	28.544	0.979
Doubly robust methods																				
Model specification: QR–True, PM–True																				
GPPP	0.023	6.062	95.833	23.490	0.995	-0.014	4.219	97.685	16.700	1.012	0.017	2.871	95.833	11.306	1.011	0.083	8.500	93.519	33.426	1.007
LWP	-0.400	6.959	96.759	28.647	1.061	-0.168	4.808	96.759	19.394	1.035	-0.076	3.301	96.296	12.962	1.008	-0.139	9.842	94.444	39.758	1.038
AIPW	-0.244	6.410	93.981	24.890	0.989	-0.214	5.745	93.056	19.238	0.853	-0.055	2.696	93.981	9.824	0.927	-0.070	6.698	95.370	25.116	0.954
Model specification: QR–True, PM–False																				
GPPP	0.578	6.177	95.370	24.644	1.027	-0.227	3.652	94.444	14.732	1.038	-0.118	2.529	95.370	9.677	0.985	0.252	7.465	93.519	28.027	0.964
LWP	102.755	6.228	61.111	36.944	0.065	77.157	4.030	58.796	26.171	0.062	52.013	28.166	60.185	21.390	0.074	68.011	7.895	61.111	27.053	0.074
AIPW	0.108	6.499	92.593	22.724	0.890	-0.148	3.903	92.593	13.978	0.912	-0.124	2.678	95.833	9.389	0.893	0.111	7.136	92.593	25.796	0.92
Model specification: QR–False, PM–True																				
GPPP	1.435	6.880	94.907	27.203	1.039	5.206	8.273	81.944	21.919	0.874	2.127	4.765	90.278	16.145	0.971	-0.452	9.473	94.444	39.581	1.071
LWP	38.447	24.018	62.037	66.505	0.166	45.933	22.233	78.704	68.192	0.145	32.100	19.000	82.407	52.561	0.202	36.851	46.251	56.481	92.284	0.236
AIPW	1.156	14.115	90.278	53.458	0.967	-5.904	64.920	48.611	92.613	0.365	11.560	11.978	3.704	11.967	0.972	34.579	35.388	0.000	28.031	0.948
Model specification: QR–False, PM–False																				
GPPP	51.343	51.589	0.000	20.925	1.069	22.943	23.242	0.000	15.244	1.053	14.179	14.437	0.000	11.227	1.060	55.161	55.704	0.000	31.475	1.04
LWP	195.557	96.444	0.000	28.177	0.255	177.674	118.013	0.463	28.728	0.250	146.864	113.140	0.000	25.490	0.260	166.531	125.351	0.000	28.693	0.288
AIPW	51.698	51.947	0.000	18.450	0.924	22.823	23.115	0.000	13.586	0.944	14.321	14.575	0.000	9.978	0.936	54.681	55.217	0.000	28.899	0.958

UW: Unweighted; FW: Fully weighted; GPPP: Gaussian Process of Propensity Prediction; LWP: Linear-in-weight Prediction; AIPW: Augmented Inverse Propensity Weighting.
NOTE: The PAPP and AIPW methods have been implemented through a bootstrap method.

Table 4.8: Comparing the performance of the bias adjustment methods in the second simulation study for the *continuous* outcome with $(n_A, n_R) = (1,000, 500)$ and $\gamma_1 = 0.6$

	LIN					*CUB*					*EXP*					*SIN*				
Measure	rBias	rMSE	crCI	lCI	rSE	rBias	rMSE	crCI	lCI	rSE	rBias	rMSE	crCI	lCI	rSE	rBias	rMSE	crCI	lCI	rSE
Probability sample (S_R)																				
UW	-23.868	24.297	0.000	17.587	0.985	-17.158	17.440	0.000	11.959	0.974	-12.107	12.274	0.000	7.683	0.967	-18.706	19.966	22.222	26.219	0.956
FW	0.118	5.477	93.519	21.395	0.994	0.016	3.998	94.444	15.257	0.971	0.015	2.673	95.833	10.454	0.995	-0.081	7.162	96.296	29.064	1.033
Non-probability sample (S_A)																				
UW	98.629	98.709	0.000	14.606	0.936	71.503	71.609	0.000	15.102	0.988	43.622	43.693	0.000	9.597	0.980	37.338	37.581	0.000	17.907	1.067
FW	-0.355	8.567	90.278	29.835	0.887	-0.483	7.679	88.889	21.564	0.716	-0.008	2.547	92.130	9.577	0.957	-0.590	9.229	93.519	36.092	0.997
Non-robust method																				
Model specification: QR–True																				
PAPP	0.837	6.971	93.981	28.496	1.048	-15.335	47.984	81.944	80.081	0.448	2.217	3.140	86.574	9.040	1.035	27.350	27.937	0.463	23.160	1.035
Model specification: QR–False																				
PAPP	51.938	52.102	0.000	16.106	0.990	23.574	23.768	0.000	11.833	0.993	14.805	14.960	0.000	8.330	0.987	55.512	55.750	0.000	20.689	1.022
Doubly robust methods																				
Model specification: QR–True, PM–True																				
GPPP	0.467	5.136	97.685	23.457	1.177	0.310	3.766	97.222	16.493	1.126	0.237	2.519	96.759	11.192	1.145	0.247	6.833	98.148	32.759	1.234
LWP	0.696	9.836	97.222	25.617	0.670	0.169	3.997	96.759	17.628	1.134	0.129	2.695	98.611	12.199	1.161	0.240	7.532	97.685	35.788	1.216
AIPW	0.328	5.304	96.296	23.712	1.140	0.323	9.295	96.759	26.822	0.735	0.236	2.504	96.296	10.389	1.061	0.292	6.326	95.833	27.065	1.09
Model specification: QR–True, PM–False																				
GPPP	0.847	5.217	96.296	23.740	1.191	-0.091	3.657	97.222	15.703	1.102	-0.035	2.398	97.222	10.482	1.121	0.301	6.461	97.685	30.374	1.209
LWP	160.667	5.868	44.444	57.995	0.103	121.455	15.991	38.889	50.917	0.120	79.393	20.870	42.130	40.288	0.139	99.141	6.863	45.370	34.496	0.092
AIPW	0.764	5.471	96.759	22.254	1.046	0.188	3.787	95.370	14.917	1.004	0.103	2.524	97.222	10.169	1.026	0.301	6.422	94.907	27.431	1.088
Model specification: QR–False, PM–True																				
GPPP	1.462	5.582	97.685	25.713	1.227	4.015	6.286	89.352	20.044	1.062	1.752	3.622	97.685	14.822	1.197	0.278	7.406	97.222	36.447	1.265
LWP	-3.714	21.454	78.704	31.849	0.335	1.510	13.601	91.667	23.485	0.307	1.474	13.319	92.593	19.695	0.383	-7.084	31.041	68.981	51.814	0.44
AIPW	1.481	12.221	88.889	44.394	0.931	-35.926	106.302	75.926	175.824	0.447	10.829	11.221	5.093	12.150	1.051	35.268	35.750	0.000	26.074	1.134
Model specification: QR–False, PM–False																				
GPPP	51.234	51.404	0.000	22.464	1.382	23.626	23.875	0.000	16.807	1.249	14.541	14.741	0.000	12.211	1.293	54.802	55.140	0.000	33.025	1.393
LWP	186.628	124.670	0.000	36.048	0.281	168.280	123.557	0.000	36.124	0.283	138.694	108.095	0.000	32.502	0.287	158.745	122.172	0.000	34.886	0.292
AIPW	51.843	52.021	0.000	16.918	1.001	23.384	23.626	0.000	13.378	1.008	14.678	14.877	0.000	9.732	1.020	55.491	55.756	0.000	21.258	0.997

UW: Unweighted; FW: Fully weighted; GPPP: Gaussian Process of Propensity Prediction; LWP: Linear-in-weight Prediction; AIPW: Augmented Inverse Propensity Weighting.
NOTE: The PAPP and AIPW methods have been implemented through a bootstrap method.

Table 4.9: Comparing the performance of the bias adjustment methods in the second simulation study for the *continuous* outcome with $(n_A, n_R) = (500, 500)$ and $\gamma_1 = 0.6$

	LIN					*CUB*					*EXP*					*SIN*				
Measure	rBias	rMSE	crCI	lCI	rSE	rBias	rMSE	crCI	lCI	rSE	rBias	rMSE	crCI	lCI	rSE	rBias	rMSE	crCI	lCI	rSE
Probability sample (S_R)																				
UW	-23.868	24.297	0.000	17.587	0.985	-17.158	17.440	0.000	11.959	0.974	-12.107	12.274	0.000	7.683	0.967	-18.706	19.966	22.222	26.219	0.956
FW	0.118	5.477	93.519	21.395	0.994	0.016	3.998	94.444	15.257	0.971	0.015	2.673	95.833	10.454	0.995	-0.081	7.162	96.296	29.064	1.033
Non-probability sample (S_A)																				
UW	100.237	100.399	0.000	20.820	0.929	73.900	74.114	0.000	21.935	0.992	44.946	45.091	0.000	13.875	0.978	36.869	37.411	0.000	25.372	1.018
FW	-0.428	10.721	92.130	40.191	0.955	-0.515	8.917	91.667	27.339	0.782	-0.070	3.570	92.593	13.283	0.947	-0.451	12.399	94.907	49.982	1.027
Non-robust method																				
Model specification: QR–True																				
PAPP	0.893	9.134	95.833	37.207	1.042	-3.022	30.594	66.667	56.091	0.469	3.650	4.758	76.852	11.529	0.961	27.447	28.332	1.852	26.590	0.963
Model specification: QR–False																				
PAPP	52.215	52.495	0.000	20.075	0.943	23.753	24.077	0.000	14.584	0.942	14.865	15.131	0.000	10.533	0.949	54.727	55.253	0.000	28.872	0.965
Doubly robust methods																				
Model specification: QR–True, PM–True																				
GPPP	0.319	6.749	94.444	27.157	1.035	0.154	4.755	95.370	19.240	1.037	0.134	3.192	96.759	12.998	1.044	0.017	9.290	96.759	38.202	1.055
LWP	0.090	7.864	95.833	32.102	1.049	0.034	5.204	95.370	21.564	1.063	0.019	3.837	94.907	14.712	0.987	-2.658	10.462	97.222	43.943	0.343
AIPW	0.180	6.969	95.370	28.099	1.027	-0.103	7.074	94.444	24.196	0.871	0.098	3.079	94.444	11.765	0.973	-0.011	7.659	94.907	30.253	1.005
Model specification: QR–True, PM–False																				
GPPP	0.876	6.819	96.759	27.893	1.060	-0.061	4.357	95.833	17.596	1.036	-0.015	2.943	95.833	11.661	1.017	0.150	8.338	95.370	33.641	1.037
LWP	103.450	6.510	62.963	44.391	0.078	77.847	4.802	59.722	36.575	0.085	51.800	18.153	58.333	26.690	0.093	67.796	8.286	62.037	33.876	0.091
AIPW	0.490	7.184	91.667	25.904	0.920	-0.024	4.629	93.981	16.579	0.912	-0.042	3.065	94.444	11.341	0.942	-0.031	7.974	94.907	30.870	0.985
Model specification: QR–False, PM–True																				
GPPP	1.843	7.386	94.444	30.532	1.097	5.574	8.660	86.111	24.260	0.939	2.304	4.919	94.907	17.944	1.057	-0.476	10.124	95.370	43.960	1.117
LWP	8.759	20.713	74.074	46.601	0.157	10.563	16.870	87.963	38.325	0.187	8.852	17.395	88.426	30.616	0.222	-4.843	41.854	68.519	74.878	0.382
AIPW	1.835	14.943	92.593	54.945	0.943	-11.650	71.384	60.185	116.080	0.420	11.277	11.832	12.500	13.911	0.988	34.536	35.418	0.926	31.247	1.012
Model specification: QR–False, PM–False																				
GPPP	51.474	51.753	0.000	25.043	1.197	23.106	23.478	0.000	18.473	1.139	14.274	14.585	0.926	13.478	1.153	54.945	55.521	0.000	37.361	1.199
LWP	191.826	108.237	0.000	35.914	0.226	173.883	112.649	0.000	35.981	0.224	143.469	104.426	0.000	32.503	0.231	163.575	116.824	0.000	36.164	0.257
AIPW	51.656	51.943	0.000	20.474	0.957	23.094	23.448	0.000	15.597	0.978	14.519	14.814	0.000	11.295	0.978	54.858	55.414	0.000	29.376	0.955

UW: Unweighted; FW: Fully weighted; GPPP: Gaussian Process of Propensity Prediction; LWP: Linear-in-weight Prediction; AIPW: Augmented Inverse Propensity Weighting.
NOTE: The PAPP and AIPW methods have been implemented through a bootstrap method.

Table 4.10: Comparing the performance of the bias adjustment methods in the second simulation study for the *binary* outcome with $(n_A, n_R) = (500, 1,000)$ and $\gamma_1 = 0.3$

	LIN					*CUB*					*EXP*					*SIN*				
Measure	rBias	rMSE	crCI	lCI	rSE	rBias	rMSE	crCI	lCI	rSE	rBias	rMSE	crCI	lCI	rSE	rBias	rMSE	crCI	lCI	rSE
Probability sample (S_R)																				
UW	-24.160	24.480	0.000	14.690	0.948	-23.859	24.179	0.000	16.035	1.040	-20.146	20.452	0.000	14.709	1.061	-11.189	11.695	10.185	13.552	1.013
FW	-0.042	4.761	94.907	17.832	0.953	-0.400	4.855	96.759	19.943	1.049	-0.706	4.262	94.907	17.407	1.054	0.169	3.866	94.444	15.066	0.993
Non-probability sample (S_A)																				
UW	49.180	49.553	0.000	22.648	0.950	37.359	38.019	0.000	25.871	0.934	27.292	27.979	0.000	22.450	0.928	19.208	19.788	1.852	19.537	1.046
FW	-0.043	6.460	93.056	24.387	0.961	0.091	6.891	94.444	27.157	1.003	0.532	6.855	94.907	25.353	0.944	-0.300	5.720	94.444	22.909	1.021
Non-robust method																				
Model specification: QR–True																				
PAPP	-0.038	5.120	96.759	21.921	1.090	3.926	11.150	87.037	39.172	0.955	6.575	9.060	81.944	24.125	0.985	6.183	7.285	71.759	16.975	1.121
Model specification: QR–False																				
PAPP	38.049	38.533	0.000	23.509	0.983	24.934	25.843	3.704	25.873	0.969	18.198	19.176	14.352	22.897	0.964	22.508	23.055	0.000	20.223	1.031
Doubly robust methods																				
Model specification: QR–True, PM–True																				
GPPP	-0.345	4.843	98.611	23.428	1.244	0.454	6.387	98.148	30.398	1.229	0.770	6.202	98.611	28.665	1.196	0.345	4.177	94.907	16.881	1.044
LWP	-0.243	4.852	98.611	23.554	1.249	0.511	6.541	97.685	30.906	1.218	0.764	6.407	99.074	29.130	1.172	0.327	4.356	95.833	17.307	1.024
AIPW	-0.028	4.938	97.685	21.216	1.094	0.248	6.688	93.981	25.328	0.964	0.792	6.236	94.444	23.674	0.974	-0.032	4.149	94.444	16.568	1.016
Model specification: QR–True, PM–False																				
GPPP	-0.286	4.831	99.074	23.492	1.249	0.996	6.370	98.148	30.231	1.234	1.659	6.255	98.611	28.479	1.210	0.485	4.095	92.130	16.104	1.016
LWP	-0.193	4.850	98.611	23.389	1.241	0.968	6.336	97.685	29.999	1.231	1.480	6.150	98.148	28.234	1.213	0.543	4.156	92.593	16.206	1.007
AIPW	0.216	4.997	97.685	21.435	1.093	0.137	6.371	94.907	25.029	1.000	0.772	6.335	94.444	23.511	0.952	-0.115	4.158	93.519	16.704	1.023
Model specification: QR–False, PM–True																				
GPPP	0.623	4.813	98.148	24.745	1.331	2.439	7.249	97.685	32.279	1.215	1.942	6.670	98.611	29.934	1.205	0.517	4.457	96.296	18.057	1.045
LWP	0.089	4.790	98.611	24.379	1.310	1.007	6.884	97.222	32.125	1.213	1.128	6.633	99.074	30.109	1.179	0.225	5.787	96.759	24.201	1.075
AIPW	0.295	6.609	95.370	28.014	1.080	-2.126	22.250	85.648	70.680	0.812	11.808	13.407	53.241	24.334	0.975	9.860	10.779	42.130	17.942	1.048
Model specification: QR–False, PM–False																				
GPPP	38.289	38.703	0.000	29.233	1.333	26.017	26.829	6.481	32.772	1.284	18.980	19.864	24.074	28.967	1.270	22.805	23.317	2.315	25.525	1.345
LWP	38.289	38.703	0.000	29.233	1.333	26.017	26.829	6.481	32.772	1.284	18.980	19.864	24.074	28.967	1.270	22.805	23.317	2.315	25.525	1.345
AIPW	38.321	38.770	0.000	22.934	0.992	25.373	26.267	4.167	26.143	0.979	18.557	19.503	11.111	22.737	0.964	22.449	23.014	0.463	20.115	1.01

UW: Unweighted; FW: Fully weighted; GPPP: Gaussian Process of Propensity Prediction; LWP: Linear-in-weight Prediction; AIPW: Augmented Inverse Propensity Weighting.
NOTE: The PAPP and AIPW methods have been implemented through a bootstrap method.

Table 4.11: Comparing the performance of the bias adjustment methods in the second simulation study for the *binary* outcome with $(n_A, n_R) = (1,000, 500)$ and $\gamma_1 = 0.3$

	LIN					*CUB*					*EXP*					*SIN*				
Measure	rBias	rMSE	crCI	lCI	rSE	rBias	rMSE	crCI	lCI	rSE	rBias	rMSE	crCI	lCI	rSE	rBias	rMSE	crCI	lCI	rSE
Probability sample (S_R)																				
UW	-24.046	24.606	0.926	20.785	1.014	-23.445	24.144	0.926	22.713	1.003	-19.875	20.549	4.167	20.818	1.015	-11.430	12.367	35.648	19.157	1.032
FW	0.329	6.211	95.833	25.283	1.038	0.153	7.020	97.222	28.280	1.026	-0.219	6.035	96.759	24.676	1.041	-0.014	4.947	96.759	21.354	1.099
Non-probability sample (S_A)																				
UW	48.655	48.836	0.000	16.025	0.970	36.988	37.324	0.000	18.292	0.932	26.828	27.149	0.000	15.876	0.970	18.929	19.164	0.000	13.818	1.176
FW	-0.155	4.474	94.907	17.199	0.979	-0.060	4.840	94.444	19.145	1.007	0.380	4.728	94.444	17.877	0.965	-0.574	3.624	96.759	16.151	1.149
Non-robust method																				
Model specification: QR–True																				
PAPP	0.236	4.935	95.370	20.326	1.049	-2.054	12.151	95.370	42.145	0.896	3.062	5.794	91.667	19.074	0.987	5.980	7.462	78.241	18.605	1.061
Model specification: QR–False																				
PAPP	37.580	37.873	0.000	18.098	0.979	24.751	25.263	0.000	19.860	0.999	17.868	18.407	1.852	17.047	0.981	22.204	22.420	0.000	14.329	1.175
Doubly robust methods																				
Model specification: QR–True, PM–True																				
GPPP	-0.350	5.172	98.148	26.705	1.330	1.049	5.831	99.074	33.701	1.509	1.324	5.279	99.074	30.631	1.539	0.701	4.796	96.296	19.836	1.072
LWP	-0.260	5.178	98.148	26.714	1.324	1.057	5.860	99.074	33.705	1.501	1.427	5.421	99.537	30.879	1.517	0.952	5.585	94.907	20.030	0.936
AIPW	0.121	5.365	96.759	22.128	1.050	0.199	5.875	93.981	23.158	1.004	0.594	5.128	95.833	20.331	1.016	-0.109	4.639	95.370	19.630	1.077
Model specification: QR–True, PM–False																				
GPPP	-0.304	5.215	98.611	26.629	1.313	1.361	5.859	99.537	33.405	1.505	1.777	5.363	99.537	30.723	1.560	0.915	4.710	94.907	19.554	1.087
LWP	-0.221	5.209	98.148	26.602	1.311	1.348	5.811	99.537	33.435	1.520	1.679	5.205	100.000	30.677	1.596	1.008	4.850	95.833	19.504	1.053
AIPW	0.251	5.271	95.833	21.900	1.059	0.163	5.703	95.370	22.900	1.022	0.565	5.074	95.370	20.202	1.020	-0.091	4.651	95.833	19.795	1.083
Model specification: QR–False, PM–True																				
GPPP	0.703	4.852	99.537	27.464	1.472	2.832	6.303	100.000	34.562	1.578	2.438	5.370	100.000	31.759	1.706	1.026	5.008	95.833	20.459	1.075
LWP	0.519	4.836	100.000	27.407	1.466	2.032	6.031	100.000	34.399	1.559	1.868	5.186	100.000	31.708	1.685	0.445	4.765	99.074	26.385	1.43
AIPW	0.619	6.290	93.519	24.931	1.014	-13.115	30.760	94.444	90.189	0.825	9.924	11.185	53.241	20.331	1.003	9.440	10.465	50.000	19.173	1.08
Model specification: QR–False, PM–False																				
GPPP	38.279	38.545	0.000	31.845	1.813	26.986	27.438	3.704	35.974	1.863	19.504	19.977	22.222	31.709	1.886	22.778	23.046	0.463	27.540	2.019
LWP	38.279	38.545	0.000	31.845	1.813	26.986	27.438	3.704	35.974	1.863	19.504	19.977	22.222	31.709	1.886	22.778	23.046	0.463	27.540	2.019
AIPW	38.013	38.302	0.000	18.348	0.995	25.475	26.004	0.000	20.612	1.005	18.409	18.940	2.315	17.500	1.001	22.053	22.284	0.000	14.343	1.141

UW: Unweighted; FW: Fully weighted; GPPP: Gaussian Process of Propensity Prediction; LWP: Linear-in-weight Prediction; AIPW: Augmented Inverse Propensity Weighting.
NOTE: The PAPP and AIPW methods have been implemented through a bootstrap method.

Table 4.12: Comparing the performance of the bias adjustment methods in the second simulation study for the *binary* outcome with $(n_A, n_R) = (500, 500)$ and $\gamma_1 = 0.3$

	LIN					*CUB*					*EXP*					*SIN*				
Measure	rBias	rMSE	crCI	lCI	rSE	rBias	rMSE	crCI	lCI	rSE	rBias	rMSE	crCI	lCI	rSE	rBias	rMSE	crCI	lCI	rSE
Probability sample (S_R)																				
UW	-24.046	24.606	0.926	20.785	1.014	-23.445	24.144	0.926	22.713	1.003	-19.875	20.549	4.167	20.818	1.015	-11.430	12.367	35.648	19.157	1.032
FW	0.329	6.211	95.833	25.283	1.038	0.153	7.020	97.222	28.280	1.026	-0.219	6.035	96.759	24.676	1.041	-0.014	4.947	96.759	21.354	1.099
Non-probability sample (S_A)																				
UW	49.180	49.553	0.000	22.648	0.950	37.359	38.019	0.000	25.871	0.934	27.292	27.979	0.000	22.450	0.928	19.208	19.788	1.852	19.537	1.046
FW	-0.043	6.460	93.056	24.387	0.961	0.091	6.891	94.444	27.157	1.003	0.532	6.855	94.907	25.353	0.944	-0.300	5.720	94.444	22.909	1.021
Non-robust method																				
Model specification: QR–True																				
PAPP	0.448	6.024	93.981	24.516	1.039	1.301	12.124	93.981	44.579	0.941	4.807	8.095	88.426	25.182	0.984	6.097	7.752	78.241	20.696	1.1
Model specification: QR–False																				
PAPP	38.110	38.636	0.000	24.026	0.963	24.997	25.965	6.944	26.728	0.968	18.257	19.247	15.741	23.253	0.971	22.514	23.073	0.000	20.041	1.01
Doubly robust methods																				
Model specification: QR–True, PM–True																				
GPPP	0.041	5.778	98.148	29.040	1.290	0.907	7.182	98.611	37.212	1.342	1.092	6.455	99.537	34.268	1.382	0.340	5.187	95.833	21.357	1.06
LWP	0.119	5.848	97.685	29.079	1.279	0.923	7.407	98.148	37.189	1.304	1.167	6.632	100.000	34.385	1.356	0.372	5.302	97.685	21.684	1.054
AIPW	0.329	5.941	94.444	25.256	1.084	0.566	7.494	93.056	28.540	0.972	0.795	6.444	94.444	25.518	1.016	-0.091	5.004	95.833	21.217	1.079
Model specification: QR–True, PM–False																				
GPPP	0.133	5.778	98.611	29.108	1.294	1.387	7.130	99.537	36.942	1.358	1.909	6.569	99.537	34.123	1.395	0.740	6.331	95.370	20.815	0.85
LWP	0.255	5.782	98.148	28.810	1.285	1.365	7.102	98.611	36.637	1.356	1.755	6.472	100.000	33.880	1.396	0.512	4.994	94.907	20.754	1.075
AIPW	0.509	5.932	96.296	25.084	1.080	0.439	7.218	93.056	27.731	0.980	0.910	6.627	95.370	25.364	0.983	-0.118	5.034	95.370	21.062	1.065
Model specification: QR–False, PM–True																				
GPPP	1.066	5.654	99.537	30.395	1.408	2.959	7.837	99.537	38.926	1.378	2.388	6.928	100.000	35.583	1.406	0.495	5.498	95.370	22.425	1.057
LWP	0.551	5.573	99.537	30.016	1.392	1.474	7.532	99.537	38.708	1.346	1.462	6.800	100.000	35.656	1.377	0.175	6.160	98.611	29.437	1.226
AIPW	0.939	7.448	97.222	30.643	1.056	-6.206	26.540	92.593	84.822	0.837	10.963	12.774	60.648	26.141	1.015	9.697	10.937	57.407	21.505	1.082
Model specification: QR–False, PM–False																				
GPPP	38.601	39.056	0.463	35.225	1.524	26.372	27.263	16.667	39.480	1.465	19.248	20.128	36.574	35.137	1.535	22.752	23.322	8.333	30.490	1.528
LWP	38.601	39.056	0.463	35.225	1.524	26.372	27.263	16.667	39.480	1.465	19.248	20.128	36.574	35.137	1.535	22.752	23.322	8.333	30.490	1.528
AIPW	38.532	39.009	0.000	24.171	1.011	25.585	26.540	4.630	26.995	0.974	18.697	19.645	11.111	23.391	0.987	22.342	22.938	0.926	20.299	0.995

UW: Unweighted; FW: Fully weighted; GPPP: Gaussian Process of Propensity Prediction; LWP: Linear-in-weight Prediction; AIPW: Augmented Inverse Propensity Weighting.
NOTE: The PAPP and AIPW methods have been implemented through a bootstrap method.

Table 4.13: Comparing the performance of the bias adjustment methods in the second simulation study for the *continuous* outcome with $(n_A, n_R) = (500, 1,000)$ and $\gamma_1 = 0.6$

	LIN					*CUB*					*EXP*					*SIN*				
Measure	rBias	rMSE	crCI	lCI	rSE	rBias	rMSE	crCI	lCI	rSE	rBias	rMSE	crCI	lCI	rSE	rBias	rMSE	crCI	lCI	rSE
Probability sample (S_R)																				
UW	-24.160	24.480	0.000	14.690	0.948	-23.859	24.179	0.000	16.035	1.040	-20.146	20.452	0.000	14.709	1.061	-11.189	11.695	10.185	13.552	1.013
FW	-0.042	4.761	94.907	17.832	0.953	-0.400	4.855	96.759	19.943	1.049	-0.706	4.262	94.907	17.407	1.054	0.169	3.866	94.444	15.066	0.993
Non-probability sample (S_A)																				
UW	93.235	93.380	0.000	20.067	0.982	79.514	79.811	0.000	25.341	0.937	58.010	58.307	0.000	21.829	0.945	22.488	22.962	0.000	19.491	1.069
FW	-0.792	10.053	91.667	37.625	0.956	-1.068	10.904	93.981	41.622	0.976	-0.125	11.375	89.352	40.607	0.909	-0.444	9.445	94.444	37.557	1.013
Non-robust method																				
Model specification: QR–True																				
PAPP	-0.227	8.039	95.833	33.477	1.060	13.742	22.143	56.019	50.263	0.737	12.275	14.469	66.667	29.468	0.979	14.626	15.234	6.944	17.300	1.033
Model specification: QR–False																				
PAPP	59.708	60.092	0.000	26.975	1.012	37.229	38.028	0.000	29.931	0.982	27.691	28.525	1.852	26.075	0.969	41.013	41.367	0.000	21.652	1.02
Doubly robust methods																				
Model specification: QR–True, PM–True																				
GPPP	-2.388	6.355	96.296	27.586	1.201	-2.760	8.776	96.296	37.554	1.155	-1.181	8.926	93.056	37.261	1.087	-0.179	5.066	95.370	21.096	1.069
LWP	-1.908	8.425	94.907	30.108	0.941	-2.168	10.318	93.981	39.769	1.011	0.601	12.671	89.815	39.277	0.803	-0.367	7.087	91.667	23.998	0.879
AIPW	-0.429	6.143	95.833	25.367	1.054	-0.311	7.571	95.370	29.165	0.981	0.596	7.700	92.130	29.438	0.976	0.211	4.192	94.444	17.162	1.043
Model specification: QR–True, PM–False																				
GPPP	-1.367	9.372	96.759	28.818	0.802	0.141	7.570	98.611	36.271	1.230	2.018	7.965	96.759	35.445	1.183	0.338	4.147	93.981	17.033	1.061
LWP	-12.060	133.689	95.833	47.857	0.098	0.028	14.769	98.148	33.546	0.584	—	—	96.759	—	0.024	0.526	4.355	93.519	17.391	1.033
AIPW	0.315	6.036	95.370	25.426	1.074	0.007	7.330	94.444	29.316	1.018	0.783	7.440	94.444	28.708	0.988	0.051	4.317	93.981	17.622	1.039
Model specification: QR–False, PM–True																				
GPPP	-0.218	6.315	97.685	31.009	1.257	3.623	10.786	96.296	43.539	1.105	2.203	10.529	93.056	39.155	0.986	-0.932	6.565	92.593	23.803	0.941
LWP	-2.276	6.534	96.759	27.687	1.160	-0.705	10.418	93.519	41.871	1.032	0.000	11.326	90.278	38.376	0.871	-0.156	10.385	83.796	29.606	0.731
AIPW	0.857	12.269	94.907	50.214	1.044	8.524	37.241	53.241	81.350	0.571	19.411	20.900	26.389	28.560	0.938	26.569	27.089	0.000	21.309	1.027
Model specification: QR–False, PM–False																				
GPPP	53.302	53.677	0.000	31.071	1.263	32.374	33.276	2.778	34.849	1.163	24.967	25.879	7.407	31.078	1.171	40.756	41.104	0.000	25.637	1.24
LWP	53.302	53.677	0.000	31.071	1.263	32.374	33.276	2.778	34.849	1.163	24.967	25.879	7.407	31.078	1.171	40.756	41.104	0.000	25.637	1.24
AIPW	55.682	56.049	0.000	26.339	1.047	33.889	34.796	0.463	29.653	0.956	25.863	26.799	3.704	26.079	0.945	41.401	41.769	0.000	22.047	1.015

UW: Unweighted; FW: Fully weighted; GPPP: Gaussian Process of Propensity Prediction; LWP: Linear-in-weight Prediction; AIPW: Augmented Inverse Propensity Weighting.
NOTE: The PAPP and AIPW methods have been implemented through a bootstrap method.

Table 4.14: Comparing the performance of the bias adjustment methods in the second simulation study for the *binary* outcome with $(n_A, n_R) = (1,000, 500)$ and $\gamma_1 = 0.6$

	LIN					*CUB*					*EXP*					*SIN*				
Measure	rBias	rMSE	crCI	lCI	rSE	rBias	rMSE	crCI	lCI	rSE	rBias	rMSE	crCI	lCI	rSE	rBias	rMSE	crCI	lCI	rSE
Probability sample (S_R)																				
UW	-24.046	24.606	0.926	20.785	1.014	-23.445	24.144	0.926	22.713	1.003	-19.875	20.549	4.167	20.818	1.015	-11.430	12.367	35.648	19.157	1.032
FW	0.329	6.211	95.833	25.283	1.038	0.153	7.020	97.222	28.280	1.026	-0.219	6.035	96.759	24.676	1.041	-0.014	4.947	96.759	21.354	1.099
Non-probability sample (S_A)																				
UW	92.334	92.413	0.000	14.252	0.948	78.359	78.506	0.000	17.952	0.951	56.892	57.036	0.000	15.470	0.973	22.907	23.094	0.000	13.777	1.195
FW	-0.489	7.243	92.593	26.827	0.945	0.076	7.212	93.519	29.956	1.057	0.340	7.973	93.519	29.311	0.936	-1.071	6.592	96.296	26.831	1.05
Non-robust method																				
Model specification: QR–True																				
PAPP	0.544	6.512	94.907	26.660	1.046	-0.222	21.896	81.019	62.841	0.731	6.897	9.340	76.852	23.256	0.940	14.493	15.104	11.574	18.440	1.104
Model specification: QR–False																				
PAPP	58.629	58.883	0.000	20.647	0.962	36.297	36.745	0.000	23.009	1.025	26.921	27.371	0.000	19.501	1.005	41.494	41.660	0.000	15.831	1.084
Doubly robust methods																				
Model specification: QR–True, PM–True																				
GPPP	-3.903	6.864	95.833	28.565	1.295	-2.421	7.354	98.611	37.021	1.372	-0.993	7.089	99.537	35.230	1.293	-0.285	5.107	96.759	21.765	1.097
LWP	-3.647	6.913	96.296	29.888	1.304	-2.105	7.801	97.222	38.341	1.312	-0.166	8.573	97.685	36.597	1.097	0.242	8.279	93.056	23.532	0.734
AIPW	-0.017	5.864	94.444	24.464	1.062	0.362	6.715	95.370	26.837	1.019	0.738	6.379	96.759	24.283	0.975	0.158	4.599	96.759	19.998	1.107
Model specification: QR–True, PM–False																				
GPPP	-3.729	6.809	96.759	28.835	1.297	-0.979	6.280	99.074	36.689	1.521	0.846	6.314	99.537	34.586	1.425	0.275	4.658	96.296	19.899	1.1
LWP	-4.364	7.228	94.444	28.283	1.261	-1.685	6.469	99.074	35.185	1.447	-0.398	5.941	99.074	32.773	1.415	-128.960	1104.353	92.593	307.225	0.071
AIPW	0.392	5.823	94.907	24.088	1.055	0.428	6.352	95.370	25.181	1.011	0.772	6.207	96.296	23.403	0.967	0.005	4.710	95.833	20.270	1.095
Model specification: QR–False, PM–True																				
GPPP	-1.790	5.718	100.000	30.256	1.428	3.429	8.509	98.611	40.538	1.339	1.870	7.991	99.074	36.386	1.207	-0.683	5.880	94.444	23.425	1.029
LWP	-2.976	6.166	99.074	29.028	1.380	0.070	7.947	96.296	39.401	1.271	—	—	98.148	—	0.074	0.056	9.039	89.352	28.784	0.816
AIPW	1.373	9.986	94.444	39.167	1.008	-13.701	54.515	73.611	128.401	0.619	17.160	18.219	19.444	23.805	0.989	26.793	27.156	0.000	19.895	1.145
Model specification: QR–False, PM–False																				
GPPP	51.913	52.162	0.000	32.167	1.621	32.802	33.285	0.926	36.718	1.668	25.110	25.585	3.241	32.362	1.697	40.375	40.592	0.000	26.482	1.623
LWP	51.913	52.162	0.000	32.167	1.621	32.802	33.285	0.926	36.718	1.668	25.110	25.585	3.241	32.362	1.697	40.375	40.592	0.000	26.482	1.623
AIPW	55.890	56.136	0.000	20.741	1.006	34.495	34.980	0.000	23.541	1.033	25.807	26.291	0.000	20.174	1.023	41.565	41.743	0.000	16.317	1.077

UW: Unweighted; FW: Fully weighted; GPPP: Gaussian Process of Propensity Prediction; LWP: Linear-in-weight Prediction; AIPW: Augmented Inverse Propensity Weighting.
NOTE: The PAPP and AIPW methods have been implemented through a bootstrap method.

Table 4.15: Comparing the performance of the bias adjustment methods in the second simulation study for the *binary* outcome with $(n_A, n_R) = (500, 500)$ and $\gamma_1 = 0.6$

	LIN					*CUB*					*EXP*					*SIN*				
Measure	rBias	rMSE	crCI	lCI	rSE	rBias	rMSE	crCI	lCI	rSE	rBias	rMSE	crCI	lCI	rSE	rBias	rMSE	crCI	lCI	rSE
Probability sample (S_R)																				
UW	-24.046	24.606	0.926	20.785	1.014	-23.445	24.144	0.926	22.713	1.003	-19.875	20.549	4.167	20.818	1.015	-11.430	12.367	35.648	19.157	1.032
FW	0.329	6.211	95.833	25.283	1.038	0.153	7.020	97.222	28.280	1.026	-0.219	6.035	96.759	24.676	1.041	-0.014	4.947	96.759	21.354	1.099
Non-probability sample (S_A)																				
UW	93.235	93.380	0.000	20.067	0.982	79.514	79.811	0.000	25.341	0.937	58.010	58.307	0.000	21.829	0.945	22.488	22.962	0.000	19.491	1.069
FW	-0.792	10.053	91.667	37.625	0.956	-1.068	10.904	93.981	41.622	0.976	-0.125	11.375	89.352	40.607	0.909	-0.444	9.445	94.444	37.557	1.013
Non-robust method																				
Model specification: QR–True																				
PAPP	0.291	8.528	98.148	34.586	1.033	8.356	21.288	76.389	57.366	0.746	9.612	12.558	78.704	31.136	0.981	14.580	15.439	22.222	20.750	1.04
Model specification: QR–False																				
PAPP	59.044	59.464	0.000	27.586	0.995	36.477	37.351	0.000	30.541	0.968	27.229	28.064	2.315	26.766	1.002	41.224	41.593	0.000	21.907	1.008
Doubly robust methods																				
Model specification: QR–True, PM–True																				
GPPP	-1.904	6.915	97.222	32.515	1.253	-2.381	9.423	98.148	42.778	1.203	-0.828	9.124	94.907	41.719	1.180	-0.203	5.952	95.833	24.740	1.067
LWP	-1.841	7.877	96.759	33.629	1.126	-1.526	10.156	96.759	45.327	1.157	0.746	12.656	92.593	44.049	0.898	-0.293	7.747	91.204	27.358	0.91
AIPW	-0.147	6.802	95.833	28.899	1.082	-0.122	8.356	93.056	32.175	0.980	0.891	8.193	93.056	31.509	0.985	0.128	5.051	96.296	21.700	1.094
Model specification: QR–True, PM–False																				
GPPP	-1.465	6.874	98.148	33.459	1.279	0.499	8.277	97.685	42.119	1.307	2.102	7.892	99.537	39.760	1.346	0.354	5.216	96.759	21.606	1.069
LWP	-2.607	7.164	96.296	31.969	1.229	-0.384	8.034	98.611	39.763	1.274	0.531	7.418	98.611	37.732	1.308	-11.862	162.050	95.370	46.573	0.076
AIPW	0.646	6.852	97.222	28.536	1.065	0.208	8.127	94.444	31.529	0.988	0.865	7.761	93.056	30.408	1.004	-0.006	5.236	96.296	21.870	1.063
Model specification: QR–False, PM–True																				
GPPP	0.154	6.819	99.537	35.644	1.338	4.130	11.390	96.759	48.346	1.172	2.647	10.628	97.222	43.869	1.096	-0.911	7.270	94.444	27.402	0.977
LWP	-1.833	6.836	97.685	32.806	1.276	-0.516	10.511	98.148	46.784	1.143	0.254	11.133	92.593	42.761	0.988	1.071	13.095	85.648	33.602	0.661
AIPW	1.505	12.999	95.370	51.268	1.011	2.173	41.495	65.741	101.955	0.626	18.273	19.986	36.574	30.913	0.972	26.480	27.090	0.463	23.873	1.063
Model specification: QR–False, PM–False																				
GPPP	53.595	54.003	0.000	36.316	1.413	32.656	33.628	4.630	41.353	1.322	25.278	26.197	13.426	36.607	1.366	40.628	41.023	0.000	30.304	1.374
LWP	53.595	54.003	0.000	36.316	1.413	32.656	33.628	4.630	41.353	1.322	25.278	26.197	13.426	36.607	1.366	40.628	41.023	0.000	30.304	1.374
AIPW	55.803	56.209	0.000	27.460	1.037	34.159	35.144	1.389	30.859	0.951	25.913	26.866	5.556	27.008	0.969	41.485	41.875	0.000	22.470	1.003

UW: Unweighted; FW: Fully weighted; GPPP: Gaussian Process of Propensity Prediction; LWP: Linear-in-weight Prediction; AIPW: Augmented Inverse Propensity Weighting.
NOTE: The PAPP and AIPW methods have been implemented through a bootstrap method.

4.6.5 Supplemental results on SHRP2/HNTS data

Table 4.16: Police-reportable crash rates per 100M driven miles and associated 95% CIs by different covariates across DR adjustment methods

Covariate	n	Unweighted (95% CI)	GPPP (95% CI)	LWP (95% CI)	AIPW (95% CI)
Total	2,862	1430.59 (1417.66,1443.52)	461.29 (296,718.88)	534.2 (270.47,1055.11)	464.58 (294.06,734)
Gender					
Male	1,357	1778.61 (1740.38,1816.84)	457.56 (293.54,713.22)	543.75 (270.88,1091.48)	464.59 (293.9,734.42)
Female	1,505	1116.79 (1107.96,1125.62)	465.4 (298.14,726.49)	524.11 (268.95,1021.35)	464.45 (293.01,736.18)
Age group					
16-19	453	2621.22 (2565.47,2676.97)	1532.82 (976.35,2406.45)	1468.28 (836.11,2578.44)	1535.73 (970.36,2430.52)
20-24	671	1357.97 (1334.13,1381.82)	860.25 (533.43,1387.32)	829.97 (383.85,1794.61)	879.91 (534.14,1449.52)
25-29	254	1058.64 (1017.15,1100.13)	788.27 (508.57,1221.8)	854.11 (488.08,1494.64)	794.45 (499.52,1263.49)
30-39	237	331.76 (313.61,349.9)	261.38 (168.1,406.41)	322.84 (160.73,648.47)	265 (168.61,416.48)
40-49	214	290.26 (273.98,306.54)	358.91 (225.32,571.7)	444.7 (222.7,887.98)	361.39 (220.82,591.43)
50-59	235	4324.69 (3815.74,4833.64)	399.07 (256.67,620.47)	509.73 (260.39,997.82)	402.83 (253.23,640.81)
60-69	276	529.89 (509.18,550.6)	561.92 (358.13,881.66)	626.86 (298.56,1316.14)	544.86 (343.45,864.39)
70-79	345	450.48 (433.74,467.23)	406.44 (264.44,624.69)	405.47 (210.49,781.06)	417.89 (272.45,640.96)
80+	177	1514.88 (1430.84,1598.91)	1238.85 (750.24,2045.68)	1204.41 (645.22,2248.24)	1248.12 (736.97,2113.81)
Race					
White	2,530	1461.22 (1445.75,1476.7)	440.54 (281.63,689.11)	502.06 (252.37,998.76)	446.8 (282.12,707.59)
Black	150	910.16 (860.84,959.49)	521.84 (334.41,814.31)	683.96 (342.54,1365.71)	511.11 (323.81,806.74)
Asian	96	2197.74 (2017.8,2377.68)	521.55 (330.77,822.36)	560.17 (311.8,1006.42)	513.48 (313.78,840.28)
Other	86	580.72 (517.11,644.34)	632.26 (420.17,951.43)	810.56 (443.05,1482.92)	634.01 (403.55,996.07)
Gender					
Non-Hisp	2,754	1442.75 (1429.07,1456.44)	434.32 (277.78,679.07)	490.06 (247.28,971.23)	434.68 (274.25,688.97)
Hispanic	108	1120.45 (1053.82,1187.08)	684.24 (472.5,990.85)	1023.74 (563.45,1860.07)	716.38 (471.67,1088.05)
Ethnicity					
0<HS	213	3659.76 (3497.14,3822.38)	1169.05 (730.81,1870.09)	1329.65 (646.48,2734.8)	1158.09 (703.44,1906.6)
HS comp	279	1606.27 (1554.61,1657.93)	478.5 (304.26,752.52)	692.02 (356.97,1341.54)	472.07 (295.39,754.42)
College	837	1248.89 (1231.23,1266.54)	473.56 (304.61,736.21)	561.22 (272.75,1154.81)	483.62 (303.5,770.65)
Graduate	1,068	603.63 (597.39,609.87)	370.9 (238.69,576.34)	385.55 (198.1,750.37)	374.99 (238.43,589.77)
Post-grad	465	2530.49 (2347.36,2713.63)	475.47 (302.42,747.56)	509.52 (261.3,993.53)	473.87 (299.47,749.84)
HH income					
0-49	1,164	1179.27 (1167.12,1191.42)	499.59 (315.31,791.56)	594.45 (288.18,1226.2)	497.9 (308.84,802.71)
150-99	1,049	709.89 (702.91,716.88)	375.32 (243.18,579.25)	421.01 (215.49,822.54)	376.41 (242.88,583.35)
100-149	442	1658.96 (1605.49,1712.43)	442.6 (286.32,684.19)	506.46 (266.98,960.73)	455.88 (289.48,717.9)
150+	207	6008.42 (5391.28,6625.55)	676.53 (425.78,1074.96)	824.7 (417.97,1627.23)	685.64 (410.65,1144.77)
HH size					
1	598	1155.7 (1128.21,1183.18)	432.75 (273.84,683.87)	468.57 (231.47,948.5)	439.9 (272.26,710.74)
2	967	698.78 (690.22,707.34)	453.73 (293.66,701.06)	516.25 (259.7,1026.23)	442.13 (283.57,689.34)
3	510	1536.75 (1493.83,1579.67)	463.86 (295.02,729.33)	546.31 (281.53,1060.12)	470.79 (293.94,754.04)
4	512	3045.14 (2885.96,3204.33)	481.49 (301.8,768.17)	561.53 (291.39,1082.1)	486.84 (300.07,789.85)
5+	275	1398.78 (1353.14,1444.42)	483.9 (314.29,745.04)	626.53 (314.87,1246.67)	512.19 (323.83,810.11)
Vehicle make					
American	1,045	2058.42 (2003.06,2113.78)	407.92 (260.59,638.55)	496.88 (246.37,1002.13)	414.98 (260.96,659.91)
Asian	1,745	1034.41 (1025.7,1043.12)	475.22 (306.94,735.75)	521.23 (268.96,1010.09)	478.85 (305.11,751.51)
European	72	1920.11 (1707.15,2133.08)	726.7 (448.99,1176.18)	963.13 (476.84,1945.38)	690.15 (409.89,1162.04)
Vehicle type					
Car	2,061	1736.09 (1715.31,1756.86)	611.19 (392.09,952.73)	667.49 (342.48,1300.92)	607.53 (381.97,966.28)
Van	109	629.14 (581.73,676.55)	682.49 (439.71,1059.32)	835.22 (446.81,1561.3)	754.96 (480.49,1186.19)
SUV	551	724.36 (696.5,752.21)	320.33 (203.8,503.48)	376.21 (182.91,773.77)	325.25 (208.23,508.02)
Pickup	141	344.53 (311.67,377.4)	233.9 (151.03,362.22)	336.55 (169.77,667.17)	238.51 (148.96,381.9)
Vehicle age					
0-4	320	2821.77 (2738.31,2905.24)	511.45 (319.47,818.8)	631.43 (316.66,1259.1)	536.52 (324.72,886.47)
5-9	742	838.31 (826.17,850.46)	483.2 (313.51,744.74)	555.03 (293.54,1049.46)	478.57 (305.12,750.62)
10-14	905	977.22 (968.18,986.25)	438.34 (281.64,682.23)	489.77 (238.5,1005.77)	442.87 (279.4,701.98)
15-19	382	607.65 (592.55,622.76)	412.86 (264.47,644.5)	480.04 (248.96,925.61)	404.66 (256.76,637.75)
20-24	197	5119.19 (4543.74,5694.65)	433.61 (279.84,671.86)	478.1 (219.52,1041.29)	436.13 (281.22,676.37)
25-29	108	545.61 (515.12,576.09)	418.01 (257.83,677.71)	469.24 (214.37,1027.12)	429.71 (265.65,695.09)
30+	178	2030.45 (1897.07,2163.84)	466.84 (277.69,784.84)	580.56 (283.85,1187.42)	486.75 (288.21,822.08)
Fuel type					
Gas/D	2,641	1526.32 (1511.74,1540.9)	461.79 (296.55,719.13)	535.61 (270.59,1060.18)	465.34 (294.5,735.28)
Other	221	286.58 (273.29,299.87)	439.6 (267.26,723.09)	476.92 (262.85,865.33)	432.82 (262.46,713.76)

CHAPTER V

Conclusion and Future Research Directions

5.1 Summary

With recent advances in automated measurement technologies, such as interactive web portals, public cameras, Global Positioning System (GPS), pedometers, and other types of tracking sensors, novel unconventional sources of data are becoming increasingly accessible in various research fields. Since collecting design-based data is often time-consuming and expensive, more and more researchers approach these pre-existing sources of data to conduct their projects. Although many social and clinical studies typically focus on the internal validity of the results for fair assessments across different experimental groups, nowadays, growing attention is paid to the generalizability of their findings to a larger population (Stuart et al., 2018, 2015, 2011; Susukida et al., 2017). When it comes to external validity for finite population inference, one has to either rely on the randomization distribution and design-based sampling under a total survey error (TSE) framework or trust merely on models and assumptions. It is undoubtedly safest to choose the earlier route, but the use of the latter is becoming increasingly inevitable.

Probability sampling suffers from declining response rates, which incur excessive costs with reduced validity. Besides, there are many situations where this long-standing touchstone for finite population inference may not be practical. Examples

include but are not limited to rare population studies, small area estimation and those studies requiring expensive and limited measurement equipment. The nature of the data-generating process in alternative sources of data is non-probabilistic and often appears as self-selection. Therefore, concepts like response and completion rates may no longer be meaningful, but valid inference for such samples becomes highly challenging as the selection mechanism seems like a "black box" to the analyst.

The extent to which one can correct for the potential selection bias in such data depends on how strong the fundamental assumptions are met in reality, and how accurately the external data represent the target population. That is perhaps the main reason why empirical studies show relatively contradictory results concerning the quality of non-probability samples (Rivers and Bailey, 2009; Gittelman et al., 2015; Wang et al., 2015; Dutwin and Buskirk, 2017; Mercer et al., 2018; Cornesse et al., 2020). According to Mercer (2018), the most critical conditions involve *exchangeability*, *positivity* and *composition*. I referred to the first two as *strong ignorability* collectively, which implies that all the auxiliary variables governing the selection mechanism of the non-probability sample or the response surface structure in the population are observed, and there are adequate sample units within each level of the auxiliary variables. By *composition*, the author means correctly specifying the underlying models so that the target composition on the auxiliary variables can be properly replicated. Mercer (2018) also proposes a unified framework for the evaluation of these assumptions, but this framework requires the key outcome variables to be observed for both samples.

Assuming that a perfect benchmark survey is present with these auxiliary variables fully measured, this dissertation attempted to develop alternative Bayesian approaches that weaken some of the other necessary assumptions for valid inference. First, it is unknown to the analyst how the observed auxiliary variables are linked to the selection propensity and the response surface in reality. Misspecifying the models

explaining these relationships can result in biased inference. Second, a partial lack of common support in the joint distribution of the auxiliary variables (partial lack of positivity) may result in the prediction of extremely low propensity scores (PS). This not only leads to biased point estimates but inflates the uncertainty of the estimates.

This dissertation addresses the first problem in two ways. The first is to employ flexible non-parametric Bayesian tools for modeling, which possess an embedded variable selection procedure and detect non-linear associations and multi-way interactions automatically. As the second approach, I reconcile the idea of propensity modeling, also known as quasi-randomization (QR), with that of prediction modeling (PM) to construct a doubly robust (DR) estimator, which maintains its consistency even if one of the underlying models for QR or PM is incorrectly specified. Chapter II proposes a two-step QR method using Bayesian Additive Regression Trees (BART). The strong flexibility of BART as a predictive tool is believed to be protecting the QR estimator against model misspecification. In addition, the posterior predictive distribution simulated by BART permits me to directly quantify the uncertainty of the adjusted estimator. Despite these advantages, it is well-understood that BART performs poorly when there is evidence of a partial lack of common support in the joint distribution of the auxiliary variables between the two samples. In addition, there is no guarantee that the observed set of common auxiliary variables fully meet the ignorable assumption.

To further protect against model misspecification, Chapter III combines the QR estimator in Chapter II with a PM estimator through a modified augmented inverse propensity weighting (AIPW) method, which is DR. Since I propose to use BART for multiply imputing both propensity scores and the outcome, the ultimate estimator is expected to be "robust squared" (Tan et al., 2019). This means that even if the true functional form of both PS and the response surface is unknown to the analyst, the proposed estimator may still remain consistent. There are, however, two other major

concerns in addition to BART's weakness against partial lack of positivity. First, the proposed methods in both chapters II and III are two-step Bayesian approaches, in the sense that imputed PS are treated as known quantities at the outcome stage. A well-known problem with such two-step methods is that the uncertainty of the ultimate estimator is misstated (Zigler et al., 2013). The simulation results in both chapters indicated that variance estimation based on BART's posterior predictive distribution consistently overestimates the variance. The second problem is that the ultimate form of the AIPW estimator is design-based. As a major drawback, design-based estimates are sensitive to the presence of influential pseudo-weights, which can potentially lead to biased estimates with inflated variance.

The second strategy of this thesis for weakening the modeling assumptions involved fully model-based inference. The basic idea is to impute the outcome for all non-sampled units of the population. Having the outcome known for the entire population units eliminates the need for design-based estimators, such as inverse PS weighting (IPSW). A DR estimator can be achieved under this setting by including the estimated PS as a predictor in the PM. The main advantage of this alternative class of DR methods is that it can be fully implemented in a Bayesian framework. By jointly estimating the PS and response surface, one can integrate out the estimated PS from the PM. Therefore, a fully model-based approach can propagate the uncertainty of the final estimator accurately, which reduces the concern of overestimated variance in the two-step Bayesian methods proposed in chapters II and III. Furthermore, Zhang and Little (2011) recognized that non-parametrically linking the estimated PS to the outcome mean reduces the sensitivity to outlying pseudo-weights. While the authors suggest fitting a penalized spline model, in Chapter IV, I proposed to use a Gaussian process (GP) regression model as the PM and showed that GP behaves as an optimal matching technique based on the estimated PS.

As another major advantage of the method I proposed in Chapter IV, one can

directly simulate the posterior predictive distribution of the finite population quantity, which allows for drawing credible intervals. This is unlike the two-step methods proposed by Zangeneh and Little (2015); Little and Zheng (2007) where a synthetic population is initially generated, and then models are fitted on the generated populations. Therefore, their method eventually requires Rubin's combining rules to derive the final point and interval estimates. My proposed method is especially advantageous when the posterior predictive distribution is not symmetric, which is usually the case for non-normal outcomes. Although the proposed method uses an embedded finite population bootstrapping (FPBB) technique to undo the sampling mechanism of the reference survey, it assumes that units of the reference survey are selected independently albeit with unequal selection probabilities. While extensions of FPBB can be suggested that handle more complex sampling designs, such a method would require knowing selection probabilities at different stages of sampling. Such information is not unusually included in the public-use datasets of probability surveys.

To recap, I list the estimators proposed across different chapters in the following:

$$Chapter\ II: \qquad \hat{\bar{y}}_U = \frac{1}{N}\sum_{i=1}^{n_A}\frac{y_i}{\hat{\pi}_i^A} \tag{5.1}$$

$$Chapter\ III: \qquad \hat{\bar{y}}_U = \frac{1}{N}\sum_{i=1}^{n_A}\frac{(y_i-\hat{y}_i)}{\hat{\pi}_i^A} + \frac{1}{N}\sum_{j=1}^{n_R}\frac{\hat{y}_i}{\pi_i^R} \tag{5.2}$$

$$Chapter\ IV: \qquad \hat{\bar{y}}_U = \frac{1}{N}\sum_{i=1}^{n_A}(y_i-\hat{y}_i) + \frac{1}{N}\sum_{j=1}^{n_R}\frac{\hat{y}_i}{\pi_i^R} \tag{5.3}$$

As illustrated, the proposed estimator in Chapter III deals with two design-based terms, whereas Chapter IV eliminates the need for the first design-based term. Note that the omitted term $\hat{\pi}_i^A$ in the last formula has appeared in the PM model as a predictor. Modifying the second term, however, demands the generation of synthetic populations based on S_R. Since models have to be fitted on the full synthetic population, for a large N, there would be a trade- off between enhanced efficiency and

computational intensity of the adjustments. In Chapter IV, because the goal was developing a unified Bayesian framework with the joint estimation of (π_i^A, y_i), the second term was kept as a HT-estimator to be able to directly simulate the posterior predictive distribution of the population mean.

5.2 Weaknesses and limitations

There were several limitations identified in this dissertation. First and foremost, the methods I proposed throughout the thesis were built upon the ignorable condition where it is assumed that all the auxiliary variables governing the selection mechanism of the non-probability sample or response surface in the population are observed and available for the analyst. In reality, this assumption may hold for neither the selection mechanism nor the response surface. In such situations, the use of more flexible modeling tools, such as BART, would no longer help remove the potential selection bias in the estimates. However, for a given non-probability sample with a known outcome variable and a fixed set of auxiliary variables, one can assess the extent of departure from this assumption before any attempt for bias adjustment using the measure proposed by Little et al. (2020). As a drawback, this measure depends on an inestimable parameter, whose valid range can be identified through sensitivity analysis.

The second most critical concern is about the construction of the PM. According to the decomposition of the joint likelihood in Eq. 1.1, we observed that the outcome may depend not only on the auxiliary variables associated with selection variables of the non-probability sample but also on the design features of the reference survey such that both samples are informative in design. Since the PM has to be fitted on the non-probability sample, it is unlikely that all the design variables of the reference survey, including sampling weights, strata, and clusters, are available for units of the non-probability sample. This problem was the case in all actual data applications

in this dissertation. All I could do was to assume that missing design variables play no significant role in explaining the variation of the outcome variable. For the DR methods, this was of less concern as the propensity model does not necessarily depend on the design variables of the reference survey. Given the ignorable assumption, and if the QR model is correctly specified, estimates are expected to be unbiased even if none of the design features of the reference survey are known for units of the non-probability sample.

As the third weakness, I assumed that auxiliary variables are error-free in measurement. Although it is well-understood that the presence of classical measurement error attenuates the estimate of model coefficients, this may not affect systematic bias in prediction and therefore the population-adjusted estimates. The challenge arises when the auxiliary variables have different measurement error structures across the non-probability sample and reference survey. The presence of differential measurement error can deteriorate the performance of the bias adjustment methods.

Last but not least, the computational intensity was a major obstacle throughout the entire data analysis of this thesis, especially when dealing with large-scale datasets. Despite the strong flexibility of BART as a predictive tool, fitting it on an even moderately sized sample can be very demanding computationally. Although different resources of high-performance computing were used for parallel processing throughout the analysis, I had to keep the number of MCMC draws as low as possible for simulating the posterior predictive distribution. This may have endangered the convergence of the MCMC sequence to the true posterior predictive distribution. I also came across computational obstacles when implementing a fully Bayesian approach for the joint estimation of PS and outcome as I had to use a custom HMC algorithm for simulating the posterior predictive distributions. To reduce the computational costs, however, I proposed a method limiting the computations to the combined sample and employed approximation methods to train the GP on the data.

For a fully model-based approach, the computational problem becomes more conspicuous because models have to be fitted repeatedly on synthetic populations. Note that the standard Bayesian software precludes one from generating the synthetic population and fitting the models under a unified framework because it is not possible to use the posterior predictive draws simulated in one step as the input for the following step.

5.3 Future research directions

This subsection strives to suggest a couple of distinct directions to enthusiastic researchers for future developments. First and foremost, one may be interested in relaxing the strongly ignorable condition, which was the main fixed assumption throughout this dissertation. One elegant solution is to use proxy pattern-mixture analysis, which has been well-developed in both causal inference and incomplete data analysis domains (Andridge and Little, 2011). However, such a method relies on an unknown parameter controlling the degree of non-ignorability, whose true value can only be assessed through sensitivity analysis. Yang and Little (2021) propose to use a penalized spline extension to pattern mixture models to reduce the risk of model misspecification. Alternatively, one may use GP instead.

Second, all the methods I proposed throughout this dissertation dealt with estimating the finite population mean and associated 95% CI, which lies in the descriptive inference domain. One may be interested in taking one further step and modifying these methods to estimate some non-smoothed population quantities such as quantiles or mode. It might also be of interest to estimate the coefficient of a regression model for analytical inference. In such situations, the use of a fully model-based approach is expected to perform best, as one would no longer need to deal with sampling weights and the stratification/clustering effects of the reference survey. Furthermore, one may intend to use the proposed approaches for improving the external validity

of multiple treatment comparisons in observational studies while the internal validity of estimates, e.g the average treatment effect (ATE), is simultaneously taken into account.

Third, future work could be to further expand the DR methods under a situation where a subset of common auxiliary variables is subject to measurement error in either the non-probability sample or reference survey. To address this issue, one has to build a measurement error model, and training such a model may demand an external validation dataset where both mismeasured and error-free covariates are observed for each sample unit. In a recent study by Hong et al. (2017), a DR method with a Bayesian framework was proposed for situations in which differential measurement errors between treated and untreated groups are present in covariates. The authors examined several scenarios including systematic, heteroscedastic, and mixed measurement errors. In the absence of a validation sample, their method relies on sensitivity analysis with respect to the parameters of the measurement error model. Antonelli et al. (2017) proposed a guided Bayesian imputation to adjust for confounders where a large portion of covariates suffer from missingness. Their approach combines the idea of Bayesian model averaging, confounding selection, and missing data imputation into a single framework.

When the non-probability sample is extremely large in size, implementing the proposed adjustment methods is computationally demanding. In the Bayesian setting, we saw that generating synthetic populations is inevitable, whose size should be, at best, several times larger than the non-probability sample. In addition, there are situations where Big Data are stored in distributed clusters of computers owing to either the large volume or confidentiality protection of the data. There is a surge of research exploring novel methods to reduce the computational burden of fitting statistical models on Big Data. A state-of-the-art solution involves parallel processing through the *divide-and-recombine* techniques, in which Big Data are initially partitioned into

independent batches, models are fitted separately on each batch, and eventually, parameter estimates are recombined such that the pooled estimator remains consistent. This method has been well-developed for generalized linear models, mixed effect models LASSO (Tang, 2018) and ridge regression (Zhang and Yang, 2017), splines (Xu and Wang, 2018) and GP smoothing (Guhaniyogi et al., 2017).

Therefore, as the final suggestion, I propose to address this important gap in the existing literature by extending the *divide-and-recombine* technique based on the idea of confidence distribution to both classes of DR estimators, AIPW and GPPP, for finite population quantities. The idea of confidence distribution provides a unified framework for combining the estimators obtained from each subsample (Xie et al., 2011). One can limit the study to a situation where the true underlying models lie within the family of generalized additive models. Further extensions can be given to a situation where data are naturally correlated, and under the high-dimensional setting, where a LASSO regularization technique is used for variable selection.

Appendix: R/Stan codes

The following GitHub link provides annotated R/Stan codes developed for generating the results of simulation and empirical studies across the chapters II-IV: https://github.com/arafei/Mythesis.

BIBLIOGRAPHY

BIBLIOGRAPHY

Abu-Nimeh, S., Nappa, D., Wang, X., and Nair, S. (2008). Bayesian additive regression trees-based spam detection for enhanced email privacy. In *Availability, Reliability and Security, 2008. ARES 08. Third International Conference on*, pages 1044–1051. IEEE.

Administration, N. H. T. S. et al. (2014). National automotive sampling system (nass) general estimates system (ges) analytical user's manual 1988-2012 (dot publication no. dot hs 811 853).

An, H. and Little, R. J. (2008). Robust model-based inference for incomplete data via penalized spline propensity prediction. *Communications in Statistics–Simulation and Computation*, 37(9):1718–1731.

An, W. (2010). Bayesian propensity score estimators: Incorporating uncertainties in propensity scores into causal inference. *Sociological Methodology*, 40(1):151–189.

Andridge, R. R. and Little, R. J. (2011). Proxy pattern-mixture analysis for survey nonresponse. *Journal of Official Statistics*, 27(2):153.

Antin, J., Stulce, K., Eichelberger, L., and Hankey, J. (2015). Naturalistic driving study: descriptive comparison of the study sample with national data. Technical report.

Antin, J. F., Lee, S., Perez, M. A., Dingus, T. A., Hankey, J. M., and Brach, A. (2019). Second strategic highway research program naturalistic driving study methods. *Safety Science*, 119:2–10.

Antonelli, J., Zigler, C., and Dominici, F. (2017). Guided bayesian imputation to adjust for confounding when combining heterogeneous data sources in comparative effectiveness research. *Biostatistics*, 18(3):553–568.

Baker, R., Blumberg, S. J., Brick, J. M., Couper, M. P., Courtright, M., Dennis, J. M., Dillman, D., Frankel, M. R., Garland, P., et al. (2010). Research synthesis: Aapor report on online panels. *Public Opinion Quarterly*, 74(4):711–781.

Baker, R., Brick, J. M., Bates, N. A., Battaglia, M., Couper, M. P., Dever, J., Gile, K. J., and Tourangeau, R. (2013). Summary report of the aapor task force on non-probability sampling. *Journal of Survey Statistics and Methodology*, 1:90–143.

Bang, H. and Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973.

Beręsewicz, M., Lehtonen, R., Reis, F., Di Consiglio, L., and Karlberg, M. (2018). An overview of methods for treating selectivity in big data sources. Technical report, Eurostat Statistical Working Paper. Doi: https://doi. org/10.2785/312232.

Brick, J. M. (2015). Compositional model inference. *JSM Proceedings (Survey Research Methods Section)*, pages 299–307.

Brick, J. M. and Kalton, G. (1996). Handling missing data in survey research. *Statistical Methods in Medical Research*, 5(3):215–238.

Brick, J. M. and Tourangeau, R. (2017). Responsive survey designs for reducing nonresponse bias. *Journal of Official Statistics*, 33(3):735–752.

Brick, J. M. and Williams, D. (2013). Explaining rising nonresponse rates in cross-sectional surveys. *The Annals of the American Academy of Political and Social Science*, 645(1):36–59.

Buelens, B., Daas, P., Burger, J., Puts, M., and van den Brakel, J. (2014). *Selectivity of Big data*. Statistics Netherlands.

Campbell, K. L. (2012). The shrp 2 naturalistic driving study: Addressing driver performance and behavior in traffic safety. *Tr News*, (282).

Cao, W., Tsiatis, A. A., and Davidian, M. (2009). Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data. *Biometrika*, 96(3):723–734.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M. A., Guo, J., Li, P., and Riddell, A. (2017). Stan: a probabilistic programming language. *Grantee Submission*, 76(1):1–32.

Cassel, C. M., Särndal, C. E., and Wretman, J. H. (1976). Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika*, 63(3):615–620.

Castanedo, F. (2013). A review of data fusion techniques. *The Scientific World Journal*, pages 1–19.

Chen, Q., Elliot, M., Haziza, D., Yang, Y., Gosh, M., Little, R., Sedransk, J., and Thompson, M. (2017a). Weights and estimation of a survey population mean: A review. *Statistical Science*, 32:227–248.

Chen, Q., Elliott, M. R., Haziza, D., Yang, Y., Ghosh, M., Little, R. J., Sedransk, J., Thompson, M., et al. (2017b). Approaches to improving survey-weighted estimates. *Statistical Science*, 32(2):227–248.

Chen, Q., Elliott, M. R., and Little, R. J. (2012). Bayesian inference for finite population quantiles from unequal probability samples. *Survey Methodology*, 38(2):203.

Chen, Y., Li, P., and Wu, C. (2019). Doubly robust inference with nonprobability survey samples. *Journal of the American Statistical Association*, pages 1–11.

Chipman, H. A., George, E. I., and McCulloch, R. E. (2007). Bayesian ensemble learning. In *Advances in Neural Information Processing Systems*, pages 265–272.

Chipman, H. A., George, E. I., McCulloch, R. E., et al. (2010). Bart: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1):266–298.

Choi, T. and Woo, Y. (2015). A partially linear model using a gaussian process prior. *Communications in Statistics-Simulation and Computation*, 44(7):1770–1786.

Cochran, W. G. (1977). *Sampling Techniques: 3d Ed.* Wiley.

Cohen, M. P. (1997). The bayesian bootstrap and multiple imputation for unequal probability sample designs. In *Proceedings of the Survey Research Methods Section, American Statistical Association*, pages 635–638.

Cornesse, C., Blom, A. G., Dutwin, D., Krosnick, J. A., De Leeuw, E. D., Legleye, S., Pasek, J., Pennay, D., Phillips, B., Sakshaug, J. W., et al. (2020). A review of conceptual approaches and empirical evidence on probability and nonprobability sample survey research. *Journal of Survey Statistics and Methodology*, 8(1):4–36.

Couper, M. P. (2013). Is the sky falling? new technology, changing media, and the future of surveys. *Survey Research Methods*, 7(3).

Daas, P. J., Puts, M. J., Buelens, B., and van den Hurk, P. A. (2015). Big data as a source for official statistics. *Journal of Official Statistics*, 31(2):249–262.

Dawid, A. P. (1982). The well-calibrated bayesian. *Journal of the American Statistical Association*, 77(379):605–610.

Deville, J.-C. and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87(418):376–382.

Deville, J.-C. and Tillé, Y. (2004). Efficient balanced sampling: the cube method. *Biometrika*, 91(4):893–912.

Dong, Q., Elliott, M. R., and Raghunathan, T. E. (2014). A nonparametric method to generate synthetic populations to adjust for complex sampling design features. *Survey Methodology*, 40(1):29.

Dutwin, D. and Buskirk, T. D. (2017). Apples to oranges or gala versus golden delicious? comparing data quality of nonprobability internet samples to low response rate probability samples. *Public Opinion Quarterly*, 81(S1):213–239.

Dutwin, D. and Lavrakas, P. (2016). Trends in telephone outcomes, 2008-2015. *Survey Practice*, 9(2):1–9.

Efron, B. (1981). Nonparametric estimates of standard error: the jackknife, the bootstrap and other methods. *Biometrika*, 68(3):589–599.

Elliott, M. R. (2009). Combining data from probability and non-probability samples using pseudo-weights. *Survey Practice*, 2(6):1–7.

Elliott, M. R. (2016). Comments on keiding & louis's perils and potentials of self-selected entry to epidemiological studies and surveys. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 179(2):319–376.

Elliott, M. R. and Little, R. J. A. (2000). Model-based alternatives to trimming survey weights. *Journal of Official Statistics*, 16:191–209.

Elliott, M. R., Resler, A., Flannagan, C. A., and Rupp, J. D. (2010). Appropriate analysis of ciren data: using nass-cds to reduce bias in estimation of injury risk factors in passenger vehicle crashes. *Accident Analysis & Prevention*, 42(2):530–539.

Elliott, M. R. and Valliant, R. (2017). Inference for nonprobability samples. *Statistical Science*, 32(2):249–264.

Fahrmeir, L., Kneib, T., et al. (2011). Bayesian smoothing and regression for longitudinal, spatial and event history data. *OUP Catalogue*.

Ferrari, S. and Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of Applied Statistics*, 31(7):799–815.

Forsberg, O. J. (2020). Polls and the us presidential election: real or fake? *Significance*, 17(5):6–7.

Fuller, W. A. (2011). *Sampling statistics*, volume 560. John Wiley & Sons.

Gelman, A. et al. (2007). Struggles with survey weighting and regression modeling. *Statistical Science*, 22(2):153–164.

Ghosh, M. and Meeden, G. (1983). Estimation of the variance in finite population sampling. *Sankhyā: The Indian Journal of Statistics, Series B*, pages 362–375.

Gittelman, S. H., Thomas, R. K., Lavrakas, P. J., and Lange, V. (2015). Quota controls in survey research: a test of accuracy and intersource reliability in online samples. *Journal of Advertising Research*, 55(4):368–379.

Green, D. P. and Kern, H. L. (2012). Modeling heterogeneous treatment effects in survey experiments with bayesian additive regression trees. *Public Opinion Quarterly*, 76(3):491–511.

Griffin, B. A., McCaffrey, D. F., Almirall, D., Burgette, L. F., and Setodji, C. M. (2017). Chasing balance and other recommendations for improving nonparametric propensity score models. *Journal of Causal Inference*, 5(2).

Groves, R. M. (2011). Three eras of survey research. *Public Opinion Quarterly*, 75(5):861–871.

Groves, R. M. and Heeringa, S. G. (2006). Responsive design for household surveys: tools for actively controlling survey errors and costs. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 169(3):439–457.

Guhaniyogi, R., Li, C., Savitsky, T. D., and Srivastava, S. (2017). A divide-and-conquer bayesian approach to large-scale kriging. *arXiv preprint arXiv:1712.09767*.

Gunawan, D., Panagiotelis, A., Griffiths, W., and Chotikapanich, D. (2020). Bayesian weighted inference from surveys. *Australian & New Zealand Journal of Statistics*, 62(1):71–94.

Guo, F., Hankey, J. M., et al. (2009). Modeling 100-car safety events: A case-based approach for analyzing naturalistic driving data. Technical report, Virginia Tech. Virginia Tech Transportation Institute.

Hahn, P. R., Murray, J. S., and Carvalho, C. M. (2020). Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects (with discussion). *Bayesian Analysis*, 15(3):965–1056.

Hájek, J. (1971). Comment on a paper by d. basu. *Foundations of statistical inference*, 236.

Han, P. and Wang, L. (2013). Estimation with missing data: beyond double robustness. *Biometrika*, 100(2):417–430.

Hargittai, E. (2015). Is bigger always better? potential biases of big data derived from social network sites. *The ANNALS of the American Academy of Political and Social Science*, 659(1):63–76.

Haziza, D. and Rao, J. N. (2006). A nonresponse model approach to inference under imputation for missing survey data. *Survey Methodology*, 32(1):53.

Hill, J. and Su, Y.-S. (2013). Assessing lack of common support in causal inference using bayesian nonparametrics: Implications for evaluating the effect of breastfeeding on children's cognitive outcomes. *The Annals of Applied Statistics*, pages 1386–1420.

Hill, J., Weiss, C., and Zhai, F. (2011). Challenges with propensity score strategies in a high-dimensional setting and a potential alternative. *Multivariate Behavioral Research*, 46(3):477–513.

Holt, D. and Smith, T. F. (1979). Post stratification. *Journal of the Royal Statistical Society: Series A (General)*, 142(1):33–46.

Hong, H., Rudolph, K. E., and Stuart, E. A. (2017). Bayesian approach for addressing differential covariate measurement error in propensity score methods. *Psychometrika*, 82(4):1078–1096.

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

Huang, B., Chen, C., and Liu, J. (2019). Gpmatch: A bayesian doubly robust approach to causal inference with gaussian process covariance function as a matching tool. *arXiv preprint arXiv:1901.10359*.

Huisingh, C., Owsley, C., Levitan, E. B., Irvin, M. R., Maclennan, P., and McGwin, G. (2018). Distracted driving and risk of crash or near-crash involvement among older drivers using naturalistic driving data with a case-crossover study design. *risk*, 6:12.

Hunsberger, S., Graubard, B. I., and Korn, E. L. (2008). Testing logistic regression coefficients with clustered data and few positive outcomes. *Statistics in Medicine*, 27(8):1305–1324.

Japec, L., Kreuter, F., Berg, M., Biemer, P., Decker, P., Lampe, C., Lane, J., O'Neil, C., and Usher, A. (2015). Big data in survey research: Aapor task force report. *Public Opinion Quarterly*, 79(4):839–880.

Johnson, T. P. and Smith, T. W. (2017). Big data and survey research: Supplement or substitute? In *Seeing Cities Through Big Data*, pages 113–125. Springer.

Kallus, N., Pennicooke, B., and Santacatterina, M. (2018). More robust estimation of sample average treatment effects using kernel optimal matching in an observational study of spine surgical interventions. *arXiv preprint arXiv:1811.04274*.

Kalton, G. and Flores-Cervantes, I. (2003). Weighting methods. *Journal of Official Statistics*, 19(2):81.

Kang, J. D., Schafer, J. L., et al. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, 22(4):523–539.

Kaplan, D. and Chen, J. (2012). A two-step bayesian approach for propensity score analysis: Simulations and case study. *Psychometrika*, 77(3):581–609.

Kaufman, C. G., Sain, S. R., et al. (2010). Bayesian functional {ANOVA} modeling using gaussian process prior distributions. *Bayesian Analysis*, 5(1):123–149.

Keeter, S., Hatley, N., Kennedy, C., and Lau, A. (2017). What low response rates mean for telephone surveys. *Pew Research Center*, 15:1–39.

Kern, C., Li, Y., and Wang, L. (2020). Boosted kernel weighting–using statistical learning to improve inference from nonprobability samples. *Journal of Survey Statistics and Methodology*, 00:1–26.

Kern, H. L., Stuart, E. A., Hill, J., and Green, D. P. (2016). Assessing methods for generalizing experimental impact estimates to target populations. *Journal of Research on Educational Effectiveness*, 9(1):103–127.

Kiaer, A. (1897). The representative method of statistical surveys. *Norwegian Academy of Science and Letters. The Historical, Philosophical Section*, 4:37–56.

Kim, J. K., Brick, J. M., Fuller, W. A., and Kalton, G. (2006). On the bias of the multiple-imputation variance estimator in survey sampling. *Journal of the Royal Statistical Society, Series B*, 68:509–521.

Kim, J. K. and Haziza, D. (2014). Doubly robust inference with missing data in survey sampling. *Statistica Sinica*, 24(1):375–394.

Kim, J. K. and Park, H. (2006). Imputation using response probability. *Canadian Journal of Statistics*, 34(1):171–182.

Kim, J. K., Park, S., Chen, Y., and Wu, C. (2021a). Combining non-probability and probability survey samples through mass imputation. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 184(3):941–963.

Kim, J. K. and Rao, J. N. (2012). Combining data from two independent surveys: a model-assisted approach. *Biometrika*, 99(1):85–100.

Kim, J.-K., Tam, S.-M., et al. (2021b). Data integration by combining big data and survey sample data for finite population inference. *International Statistical Review*, 89(2):382–401.

Kim, J. K., Wang, Z., Zhu, Z., and Cruze, N. B. (2018). Combining survey and non-survey data for improved sub-area prediction using a multi-level model. *Journal of Agricultural, Biological and Environmental Statistics*, 23(2):175–189.

Kim, W., Anorve, V., and Tefft, B. (2019). American driving survey, 2014–2017. *AAA Foundation for Traffic Safety*, pages 1–8.

Kimeldorf, G. S. and Wahba, G. (1970). A correspondence between bayesian estimation on stochastic processes and smoothing by splines. *The Annals of Mathematical Statistics*, 41(2):495–502.

Kish, L. (1965). *Survey sampling*. John Wiley and Sons.

Kitchin, R. (2015). The opportunities, challenges and risks of big data for official statistics. *Statistical Journal of the IAOS*, 31(3):471–481.

Kohler, U., Kreuter, F., and Stuart, E. A. (2019). Nonprobability sampling and causal analysis. *Annual Review of Statistics and Its Application*, 6:149–172.

Korn, E. and Graubard, B. (1999). Sample weights and imputation. analysis of health surveys. *Hoboken: Wiley*.

Kott, P. S. (1994). A note on handling nonresponse in sample surveys. *Journal of the American Statistical Association*, 89(426):693–696.

Kott, P. S. (2006). Using calibration weighting to adjust for nonresponse and coverage errors. *Survey Methodology*, 32(2):133.

Kott, P. S. and Chang, T. (2010). Using calibration weighting to adjust for non-ignorable unit nonresponse. *Journal of the American Statistical Association*, 105(491):1265–1275.

Kreuter, F. and Peng, R. D. (2014). 12 extracting information from big data: Issues of measurement, inference and linkage. *Privacy, Big Data, and the Public Good: Frameworks for Engagement*, page 257.

Lane, J. (2016). Big data for public policy: The quadruple helix. *Journal of Policy Analysis and Management*, 35(3):708–715.

Lee, S. (2006). Propensity score adjustment as a weighting scheme for volunteer panel web surveys. *Journal of Official Statistics*, 22(2):329.

Lee, S. and Valliant, R. (2009). Estimation for volunteer panel web surveys using propensity score adjustment and calibration adjustment. *Sociological Methods & Research*, 37(3):319–343.

Lenis, D., Ackerman, B., and Stuart, E. A. (2018). Measuring model misspecification: Application to propensity score methods with complex survey data. *Computational Statistics & Data Analysis*, 128:48–57.

Little, R. and An, H. (2004). Robust likelihood-based analysis of multivariate data with missing values. *Statistica Sinica*, pages 949–968.

Little, R. and Vartivarian, S. L. (2005). Does weighting for nonresponse increase the variance of survey means? *Survey Methodology*, 31(2):161.

Little, R. J. (2004). To model or not to model? competing modes of inference for finite population sampling. *Journal of the American Statistical Association*, 99(466):546–556.

Little, R. J., West, B. T., Boonstra, P. S., and Hu, J. (2020). Measures of the degree of departure from ignorable sample selection. *Journal of Survey Statistics and Methodology*, 8(5):932–964.

Little, R. J. and Zheng, H. (2007). The bayesian approach to the analysis of finite population surveys. *Bayesian Statistics*, 8(1):1–20.

Lohr, S. L., Raghunathan, T. E., et al. (2017). Combining survey data with other data sources. *Statistical Science*, 32(2):293–312.

Luiten, A., Hox, J., and de Leeuw, E. (2020). Survey nonresponse trends and fieldwork effort in the 21st century: Results of an international study across countries and surveys. *Journal of Official Statistics*, 36(3):469–487.

Mayer, I., Sverdrup, E., Gauss, T., Moyer, J.-D., Wager, S., Josse, J., et al. (2020). Doubly robust treatment effect estimation with missing attributes. *Annals of Applied Statistics*, 14(3):1409–1431.

McCandless, L. C., Gustafson, P., and Austin, P. C. (2009). Bayesian propensity score analysis for observational data. *Statistics in Medicine*, 28(1):94–112.

McConnell, K. J. and Lindner, S. (2019). Estimating treatment effects with machine learning. *Health Services Research*, 54(6):1273–1282.

McGuckin, N. and Fucci, A. (2018). Summary of travel trends: 2017 national household travel survey (report fhwa-pl-18-019). *Washington, DC: Federal Highway Administration, US Department of Transportation*.

Meng, X.-L. (2016). Statistical paradises and paradoxes in big data. *https://statistics.fas.harvard.edu/event/colloq-xiao-li-meng*.

Meng, X.-L. et al. (2018). Statistical paradises and paradoxes in big data (i): Law of large populations, big data paradox, and the 2016 us presidential election. *The Annals of Applied Statistics*, 12(2):685–726.

Mercer, A., Lau, A., and Kennedy, C. (2018). For weighting online opt-in samples, what matters most? *Pew Research Center, URL: https://www.pewresearch.org/methods/2018/01/26/for-weighting-online-opt-in-samples-what-matters-most/*.

Mercer, A. W. (2018). *Selection Bias in Nonprobability Surveys: A Causal Inference Approach*. PhD thesis.

Mercer, A. W., Kreuter, F., Keeter, S., and Stuart, E. A. (2017). Theory and practice in nonprobability surveys: Parallels between causal inference and survey inference. *Public Opinion Quarterly*, 81(S1):250–271.

Meyer, B. D., Mok, W. K., and Sullivan, J. X. (2015). Household surveys in crisis. *Journal of Economic Perspectives*, 29(4):199–226.

Miller, P. V. (2017). Is there a future for surveys? *Public Opinion Quarterly*, 81(S1):205–212.

Murdoch, T. B. and Detsky, A. S. (2013). The inevitable application of big data to health care. *JAMA*, 309(13):1351–1352.

Narain, R. (1951). On sampling without replacement with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, 3(2):169–175.

Narla, S. R. (2013). The evolution of connected vehicle technology: From smart drivers to smart cars to... self-driving cars. *Institute of Transportation Engineers. ITE Journal*, 83(7):22.

Neal, R. M. (1997). Monte carlo implementation of gaussian process models for bayesian regression and classification. *arXiv preprint physics/9701026*.

Neyman, J. (1934). On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97(4):558–625.

Oakley, J. E. and O'Hagan, A. (2004). Probabilistic sensitivity analysis of complex models: a bayesian approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(3):751–769.

Oman, S. D. and Zucker, D. M. (2001). Modelling and generating correlated binary variables. *Biometrika*, 88(1):287–290.

Pfeffermann, D. and Sverchkov, M. (1999a). Parametric and semi-parametric estimation of regression models fitted to survey data. *Sankhyā: The Indian Journal of Statistics, Series B*, 61:166–186.

Pfeffermann, D. and Sverchkov, M. (1999b). Parametric and semi-parametric estimation of regression models fitted to survey data. *Sankhyā: The Indian Journal of Statistics, Series B*, pages 166–186.

Pfeffermann, D. and Sverchkov, M. (2009). Inference under informative sampling. In *Handbook of statistics*, volume 29, pages 455–487. Elsevier.

Potter, F. and Zheng, Y. (2015). Methods and issues in trimming extreme weights in sample surveys. In *Proceedings of the American Statistical Association, Section on Survey Research Methods*, pages 2707–2719.

Presser, S. and McCulloch, S. (2011). The growth of survey research in the united states: Government-sponsored surveys, 1984–2004. *Social Science Research*, 40(4):1019–1024.

Rafei, A., Flannagan, C. A., and Elliott, M. R. (2020). Big data for finite population inference: applying quasi-random approaches to naturalistic driving data using bayesian additive regression trees. *Journal of Survey Statistics and Methodology*, 8(1):148–180.

Rafei, A., Flannagan, C. A., West, B. T., and Elliott, M. R. (2021). Robust bayesian inference for big data: Combining sensor-based records with traditional survey data. *arXiv preprint arXiv:2101.07456*.

Raghunathan, T. (2015). Statistical challenges in combining information from big and small data sources.

Rao, J. N. (2015). *Small-Area Estimation*. Wiley Online Library.

Rao, J. N. and Wu, C. (1988). Resampling inference with complex survey data. *Journal of the American Statistical Association*, 83(401):231–241.

Rao, J. N. K. and Fuller, W. A. (2017). Sample survey theory and methods: Past, present, and future directions. *Survey Methodology*, 43(2):145–160.

Riutort-Mayol, G., Bürkner, P.-C., Andersen, M. R., Solin, A., and Vehtari, A. (2020). Practical hilbert space approximate bayesian gaussian processes for probabilistic programming. *arXiv preprint arXiv:2004.11408*.

Rivers, D. (2007). Sampling for web surveys. In *Joint Statistical Meetings*.

Rivers, D. and Bailey, D. (2009). Inference from matched samples in the 2008 us national elections. In *Proceedings of the Joint Statistical Meetings*, volume 1, pages 627–639. YouGov/Polimetrix Palo Alto, CA.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89(427):846–866.

Rose, S. and van der Laan, M. J. (2008). Simple optimal weighting of cases and controls in case-control studies. *The International Journal of Biostatistics*, 4(1).

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3):581–592.

Rubin, D. B. (1981). The bayesian bootstrap. *The Annals of Statistics*, pages 130–134.

Rubin, D. B. (2004). *Multiple imputation for nonresponse in surveys*, volume 81. John Wiley & Sons.

Rubin, D. B. (2007). The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. *Statistics in Medicine*, 26(1):20–36.

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003). *Semiparametric regression*. Number 12. Cambridge university press.

Rusmassen, C. and Williams, C. (2005). *Gaussian process for machine learning*. MIT Press, Cambridge, MA, USA.

Saarela, O., Belzile, L. R., and Stephens, D. A. (2016). A bayesian view of doubly robust causal inference. *Biometrika*, 103(3):667–681.

Santos, A., McGuckin, N., Nakamoto, H. Y., Gray, D., and Liss, S. (2011). Summary of travel trends: 2009 national household travel survey. Technical report.

Särndal, C.-E., Swensson, B., and Wretman, J. (2003). *Model assisted survey sampling*. Springer Science & Business Media.

Savitsky, T. D., Toth, D., et al. (2016). Bayesian estimation under informative sampling. *Electronic Journal of Statistics*, 10(1):1677–1708.

Scharfstein, D. O., Rotnitzky, A., and Robins, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association*, 94(448):1096–1120.

Seeger, M. (2000). Relationships between gaussian processes, support vector machines and smoothing splines. *Machine Learning*.

Senthilkumar, S., Rai, B. K., Meshram, A. A., Gunasekaran, A., and Chandrakumarmangalam, S. (2018). Big data in healthcare management: A review of literature. *American Journal of Theoretical and Applied Business*, 4(2):57–69.

Shi, J. Q. and Choi, T. (2011). *Gaussian process regression analysis for functional data*. CRC Press.

SHRP2 (2013). *The 2nd Strategic Highway Research Program Naturalistic Driving Study Dataset*. Transportation Research Board National Academy of Sciences.

Si, Y., Pillai, N. S., Gelman, A., et al. (2015). Bayesian nonparametric weighted sampling inference. *Bayesian Analysis*, 10(3):605–625.

Smith, T. (1983). On the validity of inferences from non-random sample. *Journal of the Royal Statistical Society. Series A (General)*, pages 394–403.

Solin, A. and Särkkä, S. (2020). Hilbert space methods for reduced-rank gaussian process regression. *Statistics and Computing*, 30(2):419–446.

Spertus, J. V. and Normand, S.-L. T. (2018). Bayesian propensity scores for high-dimensional causal inference: A comparison of drug-eluting to bare-metal coronary stents. *Biometrical Journal*, 60:721–733.

Struijs, P., Braaksma, B., and Daas, P. J. (2014). Official statistics and big data. *Big Data & Society*, 1(1):1–6.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 25(1):1.

Stuart, E. A., Ackerman, B., and Westreich, D. (2018). Generalizability of randomized trial results to target populations: design and analysis possibilities. *Research on Social Work Practice*, 28(5):532–537.

Stuart, E. A., Bradshaw, C. P., and Leaf, P. J. (2015). Assessing the generalizability of randomized trial results to target populations. *Prevention Science*, 16(3):475–485.

Stuart, E. A., Cole, S. R., Bradshaw, C. P., and Leaf, P. J. (2011). The use of propensity scores to assess the generalizability of results from randomized trials. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 174(2):369–386.

Susukida, R., Crum, R. M., Ebnesajjad, C., Stuart, E. A., and Mojtabai, R. (2017). Generalizability of findings from randomized controlled trials: application to the national institute of drug abuse clinical trials network. *Addiction*, 112(7):1210–1219.

Tam, S.-M. and Clarke, F. (2015). Big data, official statistics and some initiatives by the australian bureau of statistics. *International Statistical Review*, 83(3):436–448.

Tan, Y. V., Elliott, M. R., and Flannagan, C. A. (2017). Development of a real-time prediction model of driver behavior at intersections using kinematic time series data. *Accident Analysis & Prevention*, 106:428–436.

Tan, Y. V., Flannagan, C. A., and Elliott, M. R. (2016). Predicting human-driving behavior to help driverless vehicles drive: random intercept bayesian additive regression trees. *arXiv preprint arXiv:1609.07464*.

Tan, Y. V., Flannagan, C. A., and Elliott, M. R. (2019). Robust-squared imputation models using bart. *Journal of Survey Statistics and Methodology*, 7(4):465–497.

Tan, Z. (2006). A distributional approach for causal inference using propensity scores. *Journal of the American Statistical Association*, 101(476):1619–1637.

Tang, L. (2018). *Statistical Methods of Data Integration, Model Fusion, and Heterogeneity Detection in Big Biomedical Data Analysis*. PhD thesis.

Tefft, B. (2017). Rates of motor vehicle crashes, injuries and deaths in relation to driver age, united states, 2014-2015. *AAA Foundation for Traffic Safety*, pages 1–5.

Terhanian, G., Bremer, J., Smith, R., and Thomas, R. (2000). Correcting data from online surveys for the effects of nonrandom selection and nonrandom assignment. *Harris Interactive White Paper*, pages 1–13.

Thompson, M. E. (2019). Combining data from new and traditional sources in population surveys. *International Statistical Review*, 87:S79–S89.

Tourangeau, R., Brick, J. M., Lohr, S., and Li, J. (2017). Adaptive and responsive survey designs: A review and assessment. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 1(180):203–223.

Valliant, R. (2020). Comparing alternatives for estimation from nonprobability samples. *Journal of Survey Statistics and Methodology*, 8(2):231–263.

Valliant, R. and Dever, J. A. (2011). Estimating propensity adjustments for volunteer web surveys. *Sociological Methods & Research*, 40(1):105–137.

Valliant, R., Dever, J. A., and Kreuter, F. (2018). Nonprobability sampling. In *Practical Tools for Designing and Weighting Survey Samples*, pages 565–603. Springer International Publishing.

Vegetabile, B. G. (2018). *Methods for Optimal Covariate Balance in Observational Studies for Causal Inference*. PhD thesis, UC Irvine.

Vittert, L., Enos, R. D., and Ansolabehere, S. (2020). Predicting the 2020 presidential election. *Harvard Data Science Review*, 2(4).

Wang, B. and Xu, A. (2019). Gaussian process methods for nonparametric functional regression with mixed predictors. *Computational Statistics & Data Analysis*, 131:80–90.

Wang, L., Graubard, B. I., Katki, H. A., Li, and Yan (2020a). Improving external validity of epidemiologic cohort analyses: a kernel weighting approach. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 183(3):1293–1311.

Wang, L., Graubard, B. I., Katki, H. A., and Li, Y. (2020b). Efficient and robust propensity-score-based methods for population inference using epidemiologic cohorts. *arXiv preprint arXiv:2011.14850*.

Wang, L., Valliant, R., and Li, Y. (2020c). Adjusted logistic propensity weighting methods for population inference using nonprobability volunteer-based epidemiologic cohorts. *arXiv preprint arXiv:2007.02476*.

Wang, W., Rothschild, D., Goel, S., and Gelman, A. (2015). Forecasting elections with non-representative polls. *International Journal of Forecasting*, 31(3):980–991.

Wendling, T., Jung, K., Callahan, A., Schuler, A., Shah, N., and Gallego, B. (2018). Comparing methods for estimation of heterogeneous treatment effects using observational data from health care databases. *Statistics in Medicine*, 37:3309–3324.

West, B. T. and Little, R. J. (2013). Non-response adjustment of survey estimates based on auxiliary variables subject to error. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 62(2):213–231.

Westreich, D., Cole, S. R., Funk, M. J., Brookhart, M. A., and Stürmer, T. (2011). The role of the c-statistic in variable selection for propensity score models. *Pharmacoepidemiology and Drug Safety*, 20:317–320.

Williams, C. K. and Rasmussen, C. E. (2006). *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA.

Williams, D. and Brick, J. M. (2018). Trends in us face-to-face household survey nonresponse and level of effort. *Journal of Survey Statistics and Methodology*, 6(2):186–211.

Williams, M. R. and Savitsky, T. D. (2021). Uncertainty estimation for pseudo-bayesian inference under complex sampling. *International Statistical Review*, 89(1):72–107.

Wu, C. and Sitter, R. R. (2001). A model-calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association*, 96(453):185–193.

Xie, M., Singh, K., and Strawderman, W. E. (2011). Confidence distributions and a unifying framework for meta-analysis. *Journal of the American Statistical Association*, 106(493):320–333.

Xu, D. and Wang, Y. (2018). Divide and recombine approaches for fitting smoothing spline models with large datasets. *Journal of Computational and Graphical Statistics*, 27(3):677–683.

Yang, S. and Kim, J. K. (2018). Integration of survey data and big observational data for finite population inference using mass imputation. *arXiv preprint arXiv:1807.02817*.

Yang, S., Kim, J. K., and Song, R. (2019). Doubly robust inference when combining probability and non-probability samples with high dimensional data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*.

Yang, Y. and Little, R. J. (2021). Spline pattern-mixture models for missing data. *Journal of Data Science*, 19(1):75–95.

Yi, G., Shi, J., and Choi, T. (2011). Penalized gaussian process regression and classification for high-dimensional nonlinear data. *Biometrics*, 67(4):1285–1294.

Zangeneh, S. Z. (2012). *Model-based Methods for Robust Finite Population Inference in the Presence of External Information.* PhD thesis.

Zangeneh, S. Z. and Little, R. J. (2015). Bayesian inference for the finite population total from a heteroscedastic probability proportional to size sample. *Journal of Survey Statistics and Methodology*, 3(2):162–192.

Zhang, G. and Little, R. (2009). Extensions of the penalized spline of propensity prediction method of imputation. *Biometrics*, 65(3):911–918.

Zhang, G. and Little, R. (2011). A comparative study of doubly robust estimators of the mean with missing data. *Journal of Statistical Computation and Simulation*, 81(12):2039–2058.

Zhang, L. C. (2019). On valid descriptive inference from non-probability sample. *Statistical Theory and Related Fields*, 3(2):103–113.

Zhang, T. and Yang, B. (2017). An exact approach to ridge regression for big data. *Computational Statistics*, 32(3):909–928.

Zheng, H. and Little, R. J. (2003). Penalized spline model-based estimation of the finite populations total from probability-proportional-to-size samples. *Journal of Official Statistics*, 19(2):99–118.

Zhou, H., Elliott, M. R., and Raghunathan, T. E. (2016). Synthetic multiple-imputation procedure for multistage complex samples. *Journal of Official Statistics*, 32(1):231.

Zhou, Q., McNeal, C., Copeland, L. A., Zachariah, J. P., and Song, J. J. (2020). Bayesian propensity score analysis for clustered observational data. *Statistical Methods & Applications*, 29(2):335–355.

Zigler, C. M. (2016). The central role of bayes' theorem for joint estimation of causal effects and propensity scores. *The American Statistician*, 70(1):47–54.

Zigler, C. M., Watts, K., Yeh, R. W., Wang, Y., Coull, B. A., and Dominici, F. (2013). Model feedback in bayesian propensity score estimation. *Biometrics*, 69(1):263–273.