Statistical Analysis with Non-probability Survey Samples

by

Yilin Chen

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Statistics

Waterloo, Ontario, Canada, 2020

Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

External Examiner:	Louis-Paul Rivest Professor (Université Laval)
Supervisors:	Pengfei Li Professor
	Changbao Wu Professor
Internal Member:	Audrey Béliveau Assistant Professor
	Mary Thompson Distinguished Professor Emerita
Internal-External Member:	Liping Fu Professor (Department of Civil and Environmental Engineering)

Author's Declaration

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

A version of Chapter 2 of this thesis has been accepted for publication in Journal of the American Statistical Association. A version of Chapter 3 of this thesis has been published in the proceedings of the Survey Methods Section, Statistical Society of Canada. Both papers are co-authored with my supervisors, and my contributions include deriving main theorems, implementing the proposed methods through R programming, performing the simulations and real data analysis, and writing the initial draft. A version of Section 2.7 in Chapter 2 serves as the real data application in the preprint paper "Combining non-probability and probability survey samples through mass imputation", which is co-authored with Dr. Jae Kwang Kim, Dr. Seho Park and Dr. Changbao Wu.

Abstract

The goal of this thesis is to develop inferential procedures with non-probability survey samples. In recent years, the use of non-probability survey samples has become one of the most important topics in the area. Contrast to the burdensome process of obtaining probability samples, non-probability survey samples, empowered by the information technology, can be acquired through the internet and other convenient measures timely and efficiently. These prompt and affordable data have facilitated online researches for both academic and industrial uses.

Nevertheless, non-probability survey samples are biased samples, from which no valid inferences about the target population can be obtained immediately. A popular tool for bias correction is the propensity score associated with each unit in the population, which is defined as the probability of selection conditional on observed auxiliary variables. Propensity scores need to be estimated in practice, but existing estimation methods are mainly derived on an ad hoc basis. This thesis establishes a general framework for statistical inferences with non-probability survey samples when relevant auxiliary information is available from a reference probability survey sample. Under this framework, we develop a rigorous procedure of estimating propensity scores. The main idea of the procedure is to approximate the required but unknown population-level information by its estimate based on the reference sample. Given the estimated propensity scores, we further present two parallel approaches to estimate the finite population mean: the quasi-randomization (QR) approach and the pseudo-empirical likelihood (PEL) approach.

In Chapter 1, we provide an overview of non-probability survey samples, especially the recent evolution driven by the information technology. We also briefly discuss commonly used statistical methods such as QR approach and mass imputation approach, and inferential barriers for non-probability samples. A separate section is dedicated to the PEL approach, where important concepts such as empirical likelihood, calibration weighting and model-calibration technique, are briefly discussed.

In Chapter 2, we introduce a two-sample setup, where a probability reference sample

is adopted in addition to the non-probability sample. Under this two-sample setup, we propose consistent estimators for propensity scores, which lead to two types of quasi-randomization based estimators for the finite population mean: the inverse probability weighted estimator and the doubly robust estimator. Lastly, a comprehensive analysis is conducted on a real non-probability survey dataset by using our proposed methods.

The pseudo-empirical likelihood approach is considered in Chapter 3. The PEL approach is comparable to the QR approach, but is motived from an entirely different non-parametric perspective. We show that the PEL approach often leads to more desirable results when the sample size is small and/or the population distribution is skewed. The approach also has flexibility to include additional constraints for double robust or multiple robust inferences.

Statistical inference with survey samples, no matter probability based or non-probability based, often relies on a complete sampling frame where every unit has a non-zero probability to be selected. This assumption, however, is not always met in practice. Chapter 4 discusses issues with incomplete sampling frames where units have zero propensity scores and illustrates the danger of applying regular procedures such as the QR approach when the sampling frame is incomplete. In particular, we describe two generating mechanisms for incomplete sampling frames, and explore inferential consequences for regular procedures under the incomplete frames. We also present a split-population approach to estimate the finite population mean, which provides more sensible and robust inferential results in the presence of zero propensity scores.

Chapter 5 discusses a few extensions and potential research directions which follow the current work.

Acknowledgements

I would like to express my deepest gratitude to my principal supervisor Prof. Changbao Wu for his guidance and encouragement during the course of my Ph.D. study. Prof. Wu introduced this exciting research topic to me, which started my challenging but enjoyable journey of academic life. It is also attributed to his generous support and constant encouragement, I had the chance to participate in numerous academic events and competitions, which have earned me valuable prizes and precious experience.

My deepest gratitude also goes to my co-supervisor Prof. Pengfei Li. He offered his expertise and gave me the sincerest suggestions whenever there were challenges on my path of research. His ingenuousness and modesty sets an example for me and constantly inspires me.

I would like to thank all my committee members, Prof. Mary Thompson, Prof. Audrey Béliveau, Prof. Liping Fu and Prof. Louis-Paul Rivest for dedicating their time to reviewing my work.

I am grateful for unconditional love and support from my parents, aunt Julia and aunt Grace, who have financially and emotionally supported me for my nine-year study in Canada. I cherish the friendship with my dearest roommate Liuyan, officemates Junhan, Bingfeng, friends Qi, Qimiao, Meng, Menglu and many others. The support from my family and friends gives me every strength to bravely face any hardship. In addition, many thanks to Mary Lou for all the help I received from her, to Prof. Song Cai at Carleton University and Dr. Fei Xu who both encouraged me to pursue Ph.D. study.

Last but not least, I want to give special thanks to my husband Anderson for being there for me in the past four years. Because of him and my parents-in-law, I have a home in Canada. This thesis can never be finished without his unwavering love and tolerance.

Dedication

In memory of Guoqing.

Table of Contents

List of Tables

Chapter 1

Introduction

1.1 Non-probability Survey Samples

Probability sampling methods have become a universally accepted approach in survey sampling since the seminal paper of Neyman (1934). Design-based inferences for finite populations using probability survey samples are widely adopted by official statistics and researchers in areas such as social studies and health sciences where surveys are one of the primary tools for data collection (Hansen, 1987; Rao, 2005). There exists an extensive literature with continued research activities on probability sampling and design-based inferences for complex surveys.

The use of non-probability survey samples has a very long history. Quota sampling, for instance, serves as a cost-effective alternative method to select a survey sample when one is limited by resources and/or the availability of reliable sampling frames. However, non-probability sampling methods have never gained true momentum in survey practice of the 20th century due to the lack of theoretical foundation for statistical inferences under the conventional design-based framework.

The success of probability sampling has led to more frequent surveys and more ambitious research projects that involve long and sophisticated questionnaires and measure-

ments. Response burden and privacy concerns, along with many other factors, have led to a dramatic decrease in response rates for almost all surveys. The challenge of low participation rates and the ever-increasing costs for conducting surveys using probability sampling methods, coupled with technology advances, has resulted in a shift of paradigm in recent years for government agencies, research institutions and industrial organizations to seek other cheaper and quicker alternatives for data collection (Citro, 2014). In particular, a great deal of attention has been given to non-probability survey samples.

The rise of the web based surveys has reshaped our views on non-probability sampling in terms of cost-and-time efficiency (Couper, 2000). The most popular type of web surveys is based on the so-called *opt-in panels*. These panels consist of volunteers who agreed to participate and are recruited through various convenient but non-probability methods. Online research through opt-in panel surveys has become popular in recent years due to its efficient recruitment process, quick responses, and low maintenance expenses. Tourangeau et al. (2013) contains many examples for web based surveys.

As much as the excitement brought by these changes, there are serious issues and major challenges for the use of web surveys and other non-probability survey samples. Unlike probability survey samples which are often selected by carefully designed sampling strategies with selection probabilities being fully available, non-probability survey samples are not governed by any clearly specified sampling schemes. This non-probability based sampling mechanism leads to selection bias, which makes the direct inferences for the population impossible. To remove the selection bias and achieve valid inferences with non-probability survey samples, the existing literature, e.g., Valliant and Dever (2011) commonly adopts the assumption that an independent probability sample is also available from the same target population (*two-sample setup*). This probability sample does not contain any measurements on study variables, but has some other variables which provide representative auxiliary information of the population.

Under this two-sample setup, two types of inferential procedures are especially popular in empirical studies. The first type is the quasi-randomization (QR) approach (see Lee,

2006; Terhanian and Bremer, 2000; Brick, 2015), where the non-probability survey sample is viewed as the primary sample while the probability survey sample is regarded as the *reference sample*. The most crucial step of this approach is to estimate the propensity scores of individuals in the population, which are defined as the conditional selection probabilities to the non-probability survey sample given some covariates. Then the non-probability survey sample can be analyzed in a similar manner to a probability survey sample, with the inverse of the propensity scores being treated as weights. Similarly to randomization approach for probability samples, this QR approach is practically appealing since a single set of weights can be used to estimate a wide range of population parameters. Popular estimating methods like inverse probability weighting (IPW) and doubly robust (DR) estimation both belong to QR approach. The second type of approach is the mass imputation (MI), under which the probability survey sample is viewed as the primary sample with responses of study variables being missing for all units. The essential idea of mass imputation is extracting information from the non-probability survey sample to fill in missing values of the probability sample. For example, Rivers (2007) and Vavreck and Rivers (2008) conducted imputation through the nearest neighbours (NN) method, i.e., for each unit in the probability sample, its missing response is imputed with the response of its closest neighbour in the non-probability sample. Kim et al. (2018) and Elliott and Valliant (2017) considered model-based methods by predicting the missing response through regression modelling. One of the main attractions of the MI approach is that the design-based nature of the probability survey sample remains intact. These two approaches are further introduced in Chapter 2.

A variety of issues need to be taken into account when employing either QR or MI approach. One of the major steps of conducting QR approach is to estimate propensity scores, but a rigorous estimating procedure is not available from the existing literature. The failure of estimating propensity scores does not only cause biased inferences, but fundamentally hinders the theoretical development of the QR approach. Erratic inferential results is the other danger of using QR approach. QR approach relies on weighting, but if small propensity scores exist, weights of non-probability samples can get excessively

large by inverting small propensity scores. As a result of large weights, IPW estimators and other weight-adjusted estimators under the QR approach can undesirably acquire massive variances and large finite sample biases. The MI approach also has drawbacks. NN imputation can be computational expensive when handling data with large dimensions. Prediction based imputation depends on modelling, and different study variables require different prediction models. So it is a less unified and productive method compared to the QR approach.

Besides the above complications, the rudimentary hurdle of applying QR and MI approach is stemming from the statistical assumptions they rely on. Under the two-sample setup, several strong assumptions are often adopted for the generating mechanism of the non-probability sample to achieve inferential validity. For example, it is common to assume that every unit in the population has a positive propensity score. However, this assumption is hard to meet given the uncontrolled generating process of non-probability survey samples, and also difficult to check due to the unavailability of the sampling frame. We discuss these assumptions and related issues in Chapter 2 and Chapter 5.

More high-level discussions and comments of using non-probability survey samples are available from "*Summary Report of the AAPOR Task Force on Non-probability Sampling*" by Baker et al. (2013), which was commissioned by the American Association of Public Opinion Research (AAPOR) Executive Council. The task force's conclusions include: (i) unlike probability sampling, there is no single framework that adequately encompasses all of non-probability sampling; (ii) making inferences for any probability or non-probability survey requires some reliance on modeling assumptions; and (iii) if non-probability samples are to gain wider acceptance among survey researchers there must be a more coherent framework and accompanying set of measures for evaluating their quality.

1.2 Pseudo-empirical Likelihood Method for Survey Data

Analysis of non-probability survey samples and probability survey samples both belong to inferential problems for finite populations so it is appealing to adapt existing inferential procedures from the probability survey sample context to solve problems in non-probability survey samples. In this thesis, we apply pseudo-empirical likelihood (PEL) approach, which was developed for probability survey samples originally, to our current research problem.

In survey data analysis, how to take advantage of auxiliary information is a frequently visited problem. In particular, a variety of approaches have been derived to utilize auxiliary information such that efficiency of estimating population means and totals can be improved for study variables. One of the earliest solutions is generalized regression (GREG) estimators (Cassel et al., 1976; Särndal, 1980), which incorporates auxiliary information by adding an augmentation component to a Horvitz-Thompson (HT) or Hájek estimator. GREG estimators are broadly used since they are easy to compute and the augmentation component does not effect the design-based consistency of HT or Hájek estimator. The other influential approach is Deville and Särndal (1992)'s calibration weighting. The idea of this approach is to obtain a set of weights which satisfies pre-specified calibration constraints, and also has the minimum discrepancy to design weights with respect to a pre-specified distance function. Specifically, calibration constraints usually force the weighted auxiliary variables and some benchmark values to be equal. GREG estimators can also be derived through calibration weighting with properly specified distance function and calibration constraints. One of the drawbacks of calibration weighting is the possible disagreement on the selection of distance functions among researchers. Moreover, some commonly used distance functions could yield negative weights which are troublesome for a series of inferential tasks. For example, when estimating the distribution function, negative weights could lead to counter-intuitive negative estimates.

Another popular approach, PEL approach, is motivated by the empirical likelihood

method. Empirical likelihood (EL) (Owen, 1988) is a non-parametric analogue of likelihood method, which has many successful applications in the area such as econometrics (Kitamura, 2007), survival analysis (Zhou, 2015), etc. EL approach is known for a few attractive features. (1) Compared to parametric based approaches, EL method is data-driven and less dependent on the assumption of the underlying distribution, so its empirical performance is usually more robust than its competitors. (2) The general goal of EL approach is to maximize EL function under a set of user-specified constraints, which means auxiliary information can be utilized through constraints in a similar way to calibration weighting. (3) The resulting weights under EL approach are strictly positive, which is a rather appreciable quality for the current research. (4) EL ratio statistics asymptotically follow chi-squared distributions under mild conditions, which offer additional solutions to construct confidence intervals (CI). These EL-ratio-based CIs are range-preserving and transformation-respecting, which are appealing properties in many practical scenarios. Extensive study of EL method is available in Owen (2001). Chen and Qin (1993) was one of the pioneers who extended EL approach to the finite population context, and they mainly considered the scenario of simple random samples. Later on, Chen and Sitter (1999) formally proposed PEL approach for probability samples. PEL approach inherits many appealing characteristics from EL method, and more importantly, it can be applied to general complex survey data. There are some equivalencies among GREG estimation, calibration weighting and PEL approach; see Wu and Lu (2016) for more discussions.

In general, efficiency gain can be achieved by these three approaches if there is a linear relation between the study variable and auxiliary variables. But efficiency may not be improved if other complex relations, besides linear relation exist between them. Wu and Sitter (2001) proposed to use *model-calibrated* constraints under the calibration weighting and PEL approach. The basic idea is to build a prediction model between the response variable and auxiliary variables first, and then constructing constraints which calibrate over the fitted values of the obtained model, as opposed to calibration on auxiliary variables directly. The advantage of this technique is its compatibility with more general relations between the response variable and its predictors. In Chapter 3, we demonstrate

the inferential procedure with non-probability survey samples by using PEL approach and model-calibration technique.

1.3 Outline of the Thesis

In the current society where information is becoming the most valuable resource, we can foresee that the use of web based survey samples and other non-probability samples would only grow wider and faster. It is urgent for us to advance in the theoretical development of non-probability survey samples to keep up with this unstoppable trend. Under this primary goal, our thesis mainly focuses on three aspects of non-probability survey samples: the establishment of a coherent inferential framework, the development of inferential procedures, and the investigation of practical issues.

In particular, the thesis is built towards a very specific task, the estimation of finite population means. As simple as it seems, this task is a thought-provoking starting point which motives us to: (1) establish a general framework for statistical inferences with non-probability survey samples, (2) develop inferential procedures which can be extended to other finite population parameters, and (3) identify general inferential barriers with non-probability survey samples. The following four chapters are composed to achieve this task from different angles.

In Chapter 2, we first define the two-sample setup which is considered throughout the thesis, and postulate a few assumptions which are critical for the inference with non-probability samples. We then discuss the estimation of propensity scores in great detail, which is considered as a key step of the QR approach. While the existing solutions are mainly ad hoc, we propose a rigorous procedure of estimating propensity scores in Section 2.3. Not only does our procedure lead to valid IPW and DR estimators given in Section 2.4, but also makes it possible adopting inferential procedures from other contexts. For example, achieving double robustness in variance estimation is not straightforward, but we are able to construct a DR variance estimator (see Section 2.5) based on Kim and Haziza

(2014)'s technique, which was originally derived for the general missing data problem. This adoption is unavailable without our proposed procedure. In Section 2.7, we apply the proposed procedure to a non-probability survey sample collected by the Pew Research Center, with auxiliary information from the Behavioral Risk Factor Surveillance System survey and the prestigious Current Population Survey. This real data application shows our method is easy to use and capable of removing selection bias.

In Chapter 3, we propose a different strategy, the PEL approach, to estimate the finite population mean. We show that consistent point estimators are obtainable under the PEL approach; and by utilizing auxiliary information through model-calibrated constraint, the obtained point estimators could also acquire double robustness property (see Section 3.2.1). In addition, doubly robust inference under PEL approach can be easily extended to multiple robustness by simple modifications (see Section 3.4). We illustrate two types of PEL-ratio-based CIs in Section 3.2.2. One method is based on the limiting distribution of the adjusted PEL ratio statistics (Wu and Rao, 2006), and the other is derived from the bootstrap-calibrated PEL ratio statistics (Wu and Rao, 2010). These PEL-ratio-based CIs are generally comparable with typical Wald-type CIs, and have more attractive data-driven features when the response variable is binary and the parameter of interest is the finite population proportion. Specifically, they outperform Wald-type CIs with respect to coverage rates and balance of tail errors under simulated scenarios where the sample size is small and the population proportion is close to 0 or 1.

To achieve valid inferences, both QR and PEL approach require positive propensity scores for every unit in the population (*positivity assumption*). However, positivity, as a strong assumption, often fails to hold in practice (*positivity violation*), and the failure of the positivity assumption would cause the *incomplete sampling frame*. Incomplete sampling frames usually refer to the scenarios where the sample is drawn from the partial population according to some randomization mechanism, with the rest of the population being excluded from the process. It is a severe issue for both probability and non-probability survey samples since the uncovered population by the sampling frame may not be represented in the sample. To the best of our knowledge, most of existing inferential procedures in

the current context are derived under the positivity assumption and a complete sampling frame, but the violation of these assumptions is merely investigated. In Chapter 4, we are filling in this research gap by examining two generating mechanisms for incomplete frames (see Section 4.1) and investigating several methods which can potentially deal with the positivity violation (see Section 4.2). Moreover, to mitigate the danger of ignoring zero propensity scores, we recommend a split-population approach in Section 4.3. Under this approach, the target population is viewed as the union of two subpopulations, one consisting of units with zero propensity scores and the other containing the rest. Then based on the ratio and characteristics of two subpopulations, proper inferential procedures for the target population can be chosen sensibly and flexibly.

In Chapter 5, we first summarize the work we have done in the thesis, then briefly explore some potential extensions based on our current work, and lastly provide the general outlook of non-probability survey samples.

Chapter 2

Inference for Non-probability Samples

2.1 Setup and Notation

Let $\mathcal{U} = \{1,2,\cdots,N\}$ represent the set of N units for the finite population, with N being the population size. Associated with unit i are values of the k-dimensional vector of auxiliary variables $\boldsymbol{x}_i$, and the value y_i for the response variable y, $i = 1, 2, \cdots, N$. Under the design-based framework, the set of finite population values $\mathcal{F}_N = \{(\boldsymbol{x}_i,y_i), i \in \mathcal{U}\}$ is viewed as fixed. Let $\mu_y = N^{-1}\sum_{i=1}^{N} y_i$ be the finite population mean for the response variable, and our goal is to estimate μ_y.

Consider a non-probability sample $\mathcal{S}_A$ consisting of n_A units from the finite population. Let $\{(\boldsymbol{x}_i,y_i), i \in \mathcal{S}_A\}$ be the dataset from the non-probability sample. Let $R_i = I(i \in \mathcal{S}_A)$ be the indicator variable for unit i being included in the sample $\mathcal{S}_A$, i.e., $R_i = 1$ if $i \in \mathcal{S}_A$ and $R_i = 0$ if $i \notin \mathcal{S}_A$, $i = 1,2,\cdots,N$. Then the conditional selection probability for unit i given $\boldsymbol{x}_i$ and y_i is computed as

$$\pi_i^A = E_q(R_i \mid \boldsymbol{x}_i, y_i) = P(R_i = 1 \mid \boldsymbol{x}_i, y_i)\,, \quad i = 1,2,\cdots,N\,,$$

where subscript q indicates that the operator is taken under the selection mechanism for sample $\mathcal{S}_A$. The value of π_i^A is the so-called propensity score (Rubin, 1976), which is usually unknown in practice and need to be estimated.

The selection mechanism is called ignorable if $\pi_i^A = P(R_i = 1 \mid \boldsymbol{x}_i, y_i) = P(R_i = 1 \mid \boldsymbol{x}_i), i = 1,2,\cdots,N$. This is referred to as *ignorability condition*, which corresponds to *missing at random* (MAR) as defined by Rubin (1976). Ignorability condition simplifies the estimation of propensity scores by requiring no measurements of missing values of the response variable. If a selection mechanism is non-ignorable, then estimating propensity scores is very challenging or can be even impossible (see Little and Rubin, 2002; Kim and Shao, 2013). More formally, we assume that the selection mechanism for sample $\mathcal{S}_A$ satisfies the following assumptions.

A1 The selection indicator R_i and the response variable y_i are independent given the set of covariates $\boldsymbol{x}_i$.

A2 All units have a non-zero propensity score, i.e., $\pi_i^A > 0$ for all i.

A3 The indicator variables R_i and R_j are independent given $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ for $i \neq j$.

As pointed out by Rivers (2007), the term "ignorable" is an unfortunate choice of terminology for the missing data and causal inference literature, since it certainly cannot be ignored by the analyst. Similarly, the term "missing at random" should not be confused with "randomly missing". Assumption **A1** means that covariates $\boldsymbol{x}_i$ is a set of confounding variables which fully captures the relation between the selection mechanism and the response variable, and there is no other unmeasured confounding variables. Assumption **A2** is often referred to as positivity assumption, and it cannot be satisfied by scenarios where certain units will for sure not be included in the sample. This is a rather complicated issue which we further explore in Chapter 4. Assumptions **A1** and **A2** together is the strong ignorability condition as discussed by Rosenbaum and Rubin (1983), which directly implies igonorability condition. From now on, we assume that Assumptions **A1**–**A3** always hold until Chapter 4 where some assumptions are relaxed for the further investigation.

The propensity scores π_i^A cannot be estimated from the sample $\mathcal{S}_A$ itself, and information on the rest of the finite population is required. So we adopt the assumption that a reference probability sample, denoted as $\mathcal{S}_B$, is also available from the target population. Let $\{(\boldsymbol{x}_i, d_i^B), i \in \mathcal{S}_B\}$ be the data from the reference probability sample, where $d_i^B = 1/\pi_i^B$ are the survey weights and $\pi_i^B = P(i \in \mathcal{S}_B)$ are the inclusion probabilities under the probability sampling design for the sample $\mathcal{S}_B$. Note that the response variable y is not part of the dataset for the reference sample.

It is usually assumed that reference sample $\mathcal{S}_B$ is relatively inexpensive and easy to obtain from existing data sources, otherwise the purpose of using non-probability survey samples is defeated. If covariates $\boldsymbol{x}$ are some typical items in surveys, such as demographic variables, then it is not hard to find some existing census or probability survey samples which contain measurements of $\boldsymbol{x}$. This type of sample $\mathcal{S}_B$ is the most ideal reference sample since it is subject to no extra cost. If no existing sample is suitable to use, one can consider drawing a small-size probability based sample as sample $\mathcal{S}_B$. In contrast with conventional probability samples, a reference probability sample $\mathcal{S}_B$ does not require information of the response variable, which means it can be collected more easily and timely. This is a major advantage especially when y is some sensitive item, such as income, health condition, etc. Due to these merits, this two-sample setup is considered in a growing volume of literature recently, e.g., Zhang (2019), Rafei et al. (2020), etc.

2.2 Mass Imputation and Quasi-randomization Approach

In the existing literature, a variety of statistical methods have been developed to estimate finite population means with non-probability survey samples. We mainly focus on the mass imputation and quasi-randomization approach, which together encompass a wide range of popular methods. For other relevant statistical methods, one may refer to Yang and Kim (2020) for a detailed and up-to-date review.

2.2.1 Mass imputation approach

Mass imputation is a conventional approach to deal with missing values in survey data (see Chen and Shao, 2000; Kim and Rao, 2012; Yang and Kim, 2018). In this section, we discuss its extensions to the current two-sample setup.

Compared to non-probability survey samples, probability survey samples better represent the target population. This motivates analysts to treat sample $\mathcal{S}_B$ as the primary sample with missingness on response y. Sample $\mathcal{S}_A$ here serves as auxiliary dataset to help fill in missing y values of sample $\mathcal{S}_B$. Let $\hat{y}_i$ denote the imputed value for unit $i \in \mathcal{S}_B$, then based on the imputed sample $\mathcal{S}_B$ and the survey weights d_i^B, we can construct the following estimator of parameter μ_y,

$$\hat{\mu}_{MI} = \frac{1}{\hat{N}^B} \sum_{i \in \mathcal{S}_B} d_i^B \hat{y}_i,$$

where $\hat{N}^B = \sum_{i \in \mathcal{S}_B} d_i^B$. Estimator $\hat{\mu}_{MI}$ with the estimated population count $\hat{N}^B$ is a well-known Hájek estimator and is preferred to use in practice even if N is known. If $\hat{N}^B$ is replaced by the true N, then the estimator becomes a HT-type estimator. Särndal et al. (1992) discussed several scenarios where Hájek estimators are likely to outperform the counterpart HT estimators.

The imputed value $\hat{y}_i$ can be obtained through different procedures. For instance, Rivers (2007) investigated a non-parametric procedure, the nearest neighbour (NN) imputation. Specifically, the imputation value is given by $\hat{y}_i = y_j$ for unit $i \in \mathcal{S}_B$, where $j \in \mathcal{S}_A$ and its associated value $\boldsymbol{x}_j$ minimizes the distance $\|\boldsymbol{x}_k - \boldsymbol{x}_i\|$ for all $k \in \mathcal{S}_A$. In other words, each missing y_i for $i \in \mathcal{S}_B$ is imputed with an observed y_j for some $j \in \mathcal{S}_A$. We denote the estimator of μ_y under the NN imputation by $\hat{\mu}_{NN}$. The NN imputation is a specific form of the so-called donor imputation. In addition to being non-parametric, it can be used to impute a vector of survey variables, which would preserve the distributions of the variables or the relationships among the variables.

Model-based prediction approach has also been explored for inferences with non-probability samples. Suppose that the finite population $\{(\boldsymbol{x}_i, y_i), i \in \mathcal{U}\}$ can be viewed as

a random sample from the model

$$y_i = m(\boldsymbol{x}_i) + \varepsilon_i\,, \quad i = 1,2,\cdots\,, \tag{2.2.1}$$

where $m(\boldsymbol{x}_i) = E_\xi(y_i \mid \boldsymbol{x}_i)$, and the subscript ξ indicates that the operator is taken under the conditional distribution of y. Model $m(\boldsymbol{x}_i)$ is called prediction model or outcome regression model, which can take a parametric form such as $m(\boldsymbol{x}_i) = \boldsymbol{x}_i^\intercal\boldsymbol{\beta}$ or an unspecified non-parametric form. The error terms ε_i are independent with $E_\xi(\varepsilon_i) = 0$ and $Var_\xi(\varepsilon_i) = v(\boldsymbol{x}_i)\sigma^2$. The variance function $v(\boldsymbol{x}_i)$ has a known form, and the homogeneous variance structure with $v(\boldsymbol{x}_i) = 1$ might be used for certain applications.

Under Assumptions **A1** and **A2**, we have $E_\xi(y_i \mid \boldsymbol{x}_i, R_i = 1) = E_\xi(y_i \mid \boldsymbol{x}_i) = m(\boldsymbol{x}_i)$ for every $\boldsymbol{x}_i$, which means the dataset $\{(\boldsymbol{x}_i,y_i), i \in \mathcal{S}_A\}$ from the non-probability sample can be used to build the model (2.2.1). For example, for the linear regression model where $m(\boldsymbol{x}_i) = \boldsymbol{x}_i^\intercal\boldsymbol{\beta}$ and $v(\boldsymbol{x}_i) = 1$, the least square estimator of $\boldsymbol{\beta}$ is given by

$$\hat{\boldsymbol{\beta}}_{ls} = \Big(\sum_{i=1}^{N} R_i\boldsymbol{x}_i\boldsymbol{x}_i^\intercal\Big)^{-1}\Big(\sum_{i=1}^{N} R_i\boldsymbol{x}_i y_i\Big) = \Big(\sum_{i\in\mathcal{S}_A} \boldsymbol{x}_i\boldsymbol{x}_i^\intercal\Big)^{-1}\Big(\sum_{i\in\mathcal{S}_A} \boldsymbol{x}_i y_i\Big)\,,$$

and the predicted value for y_i with associated $\boldsymbol{x}_i$ is given by $\hat{y}_i = \boldsymbol{x}_i^\intercal\hat{\boldsymbol{\beta}}_{ls}$. Similarly, for any general regression model $m(\boldsymbol{x},\boldsymbol{\beta})$, we have the predicted value $\hat{y}_i = m(\boldsymbol{x}_i,\hat{\boldsymbol{\beta}})$ for $i \in \mathcal{S}_B$, with $\hat{\boldsymbol{\beta}}$ being the estimated model parameters based on $\{(\boldsymbol{x}_i,y_i), i \in \mathcal{S}_A\}$. If imputing missing y_i for $i \in \mathcal{S}_B$ with their corresponding predicted value $\hat{y}_i$, then we obtain the following regression-type estimator,

$$\hat{\mu}_{REG} = \frac{1}{\hat{N}^B}\sum_{i\in\mathcal{S}_B} d_i^B m(\boldsymbol{x}_i,\hat{\boldsymbol{\beta}}).$$

Estimator $\hat{\mu}_{REG}$ is approximately unbiased under the joint framework of the prediction model and the probability sampling design for $\mathcal{S}_B$. The regression prediction estimator $\hat{\mu}_{REG}$ tends to perform well if the model $y_i = m(\boldsymbol{x}_i) + \varepsilon_i$ has strong prediction power (Kang and Schafer, 2007). More discussions about this approach are available in Kim et al. (2018).

2.2.2 Quasi-randomization approach

The barrier of utilizing sample $\mathcal{S}_A$ is the intrinsic selection bias inherited from non-probability sampling. This bias nevertheless can be potentially removed if sample $\mathcal{S}_A$ is adjusted by the propensity scores. Specifically, under Assumptions **A1**–**A3**, a representative dataset is given by $\{(\boldsymbol{x}_i, y_i, 1/\pi_i^A), i \in \mathcal{S}_A\}$, with $1/\pi_i^A$ being treated as the survey weights. Weighted sample $\mathcal{S}_A$ can be viewed as an analogue of weighted probability survey data, and sequentially, many design-based statistical methods can be naturally extended to $\mathcal{S}_A$. But unlike probability samples where survey weights are accessible as part of the design, weights of sample $\mathcal{S}_A$ are always estimated in practice due to the unavailability of the true propensity scores π_i^A. This major difference is reflected on the name "quasi"-randomization approach (Kott, 1994), or "pseudo" design-based approach, which contrasts with conventional randomization theory and design-based inferences under the probability sampling design.

Estimating propensity scores is obviously a crucial step in the QR approach. Under the two-sample setup defined in Section 2.1, a few estimating procedures are suggested in the existing literature. However, we notice that these procedures are either ad hoc based, or dependent on stronger assumptions than what we postulate here.

For instance, the approach described in Lee (2006), Isaksson and Forsman (2003) and Lee and Valliant (2009) attempted to estimate $\tilde{\pi}_i^A = P(\tilde{R}_i = 1 \mid \boldsymbol{x}_i)$, where $\tilde{R}_i = 1$ if $i \in \mathcal{S}_A$ and $\tilde{R}_i = 0$ if $i \in \mathcal{S}_B$. The estimation of $\tilde{\pi}_i^A$ is carried out through binary regression model like logistic regression model based on the pooled sample $\mathcal{S}_{AB} = \mathcal{S}_A \cup \mathcal{S}_B$. It is obvious that $\tilde{\pi}_i^A$ is not the same as the true propensity score π_i^A, and pooling the two samples $\mathcal{S}_A$ and $\mathcal{S}_B$ together in such a way does not provide required information for the estimation. Valliant and Dever (2011) considered a weighted logistic regression procedure to estimate $\tilde{\pi}_i^A$. For each unit $i \in \mathcal{S}_{AB}$, the pooled weight d_i is defined as $d_i = 1$ if $i \in \mathcal{S}_A$, and $d_i = d_i^B(1 - n_A/\hat{N}_B)$ if $i \in \mathcal{S}_B$. Compared to the first unweighted approach, pooled sample $\mathcal{S}_{AB}$ now are weighted up to the $\hat{N}_B$, which is the approximated total count of the population. Including weights in the estimation is obviously a substantial improvement,

but we show in Section 2.3 that this procedure is only valid under very limited scenarios.

There are other estimating procedures for propensity scores which are derived under extra assumptions. For instance, Elliott and Valliant (2017) proposed a "pseudo-weights" method, which is valid if sampling fractions for both $\mathcal{S}_A$ and $\mathcal{S}_B$ are small. Kim and Wang (2019) assumed that the selection indicator R_i of sample $\mathcal{S}_A$ can be observed in sample $\mathcal{S}_B$, and π_i^A can be estimated based on the data $\{(R_i,\boldsymbol{x}_i, d_i^B), i \in \mathcal{S}_B\}$.

Once π_i^A are estimated, $\mathcal{S}_A$ can be viewed as a sample obtained by the Poisson sampling under Assumption **A3**, with the probabilities of selection being specified by the estimated propensity scores. As a result, various statistical procedures derived under the design-based framework can be used to make inferences with sample $\mathcal{S}_A$. But one may need to be cautious about the additional variation resulted from estimating π_i^A, especially when doing variance estimation. The estimation of means and totals under the QR approach will be discussed in great depth in the following section.

2.3 Estimation of Propensity Scores

Consider the hypothetical situation where $\boldsymbol{x}_i$ is observed for all units in the finite population $\mathcal{U}$ while y_i is only observed for the non-probability sample $\mathcal{S}_A$. Estimation of the propensity scores under this scenario becomes the standard MAR problem with observations $\{(R_i, R_i y_i, \boldsymbol{x}_i), i = 1,2,\cdots,N\}$. Suppose that the propensity scores can be modelled parametrically as $\pi_i^A = P(R_i = 1 \mid \boldsymbol{x}_i) = \pi(\boldsymbol{x}_i,\boldsymbol{\theta}_0)$, where $\boldsymbol{\theta}_0$ is the true value of the unknown model parameters. The maximum likelihood estimator of π_i^A is computed as $\pi(\boldsymbol{x}_i,\hat{\boldsymbol{\theta}})$, where $\hat{\boldsymbol{\theta}}$ maximizes the log-likelihood function

$$\begin{aligned} l(\boldsymbol{\theta}) &= \sum_{i=1}^{N} \left\{ R_i \log \pi_i^A + (1 - R_i) \log (1 - \pi_i^A) \right\} \\ &= \sum_{i \in \mathcal{S}_A} \log \left\{ \frac{\pi(\boldsymbol{x}_i,\boldsymbol{\theta})}{1 - \pi(\boldsymbol{x}_i,\boldsymbol{\theta})} \right\} + \sum_{i=1}^{N} \log \left\{ 1 - \pi(\boldsymbol{x}_i,\boldsymbol{\theta}) \right\}. \end{aligned} \tag{2.3.2}$$

However, the log-likelihood function specified in (2.3.2) cannot be used in practice since we do not observe $\boldsymbol{x}_i$ for all units in the finite population. This is where we need the reference probability sample $\mathcal{S}_B$ with information on $\boldsymbol{x}$. Instead of using $l(\boldsymbol{\theta})$, we compute the estimator by maximizing the following pseudo log-likelihood function

$$l^*(\boldsymbol{\theta}) = \sum_{i \in \mathcal{S}_A} \log \left\{ \frac{\pi(\boldsymbol{x}_i, \boldsymbol{\theta})}{1 - \pi(\boldsymbol{x}_i, \boldsymbol{\theta})} \right\} + \sum_{i \in \mathcal{S}_B} d_i^B \log \left\{ 1 - \pi(\boldsymbol{x}_i, \boldsymbol{\theta}) \right\}, \tag{2.3.3}$$

where the population total $\sum_{i=1}^{N} \log\{1 - \pi(\boldsymbol{x}_i, \boldsymbol{\theta})\}$ in $l(\boldsymbol{\theta})$ is replaced by the HT estimator $\sum_{i \in \mathcal{S}_B} d_i^B \log\{1 - \pi(\boldsymbol{x}_i, \boldsymbol{\theta})\}$ using the reference sample $\mathcal{S}_B$.

Under a logistic regression model for the propensity scores where $\pi_i^A = \pi(\boldsymbol{x}_i, \boldsymbol{\theta}_0) = \exp(\boldsymbol{x}_i^\intercal \boldsymbol{\theta}_0)/\{1 + \exp(\boldsymbol{x}_i^\intercal \boldsymbol{\theta}_0)\}$, the pseudo log-likelihood function (2.3.3) becomes

$$l^*(\boldsymbol{\theta}) = \sum_{i \in \mathcal{S}_A} \boldsymbol{x}_i^\intercal \boldsymbol{\theta} - \sum_{i \in \mathcal{S}_B} d_i^B \log \left\{ 1 + \exp(\boldsymbol{x}_i^\intercal \boldsymbol{\theta}) \right\}.$$

The maximum pseudo likelihood estimator, denoted by $\hat{\boldsymbol{\theta}}_{ml}$, can be obtained by solving the score equations $U(\boldsymbol{\theta}) = \mathbf{0}$ where

$$U(\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} l^*(\boldsymbol{\theta}) = \sum_{i \in \mathcal{S}_A} \boldsymbol{x}_i - \sum_{i \in \mathcal{S}_B} d_i^B \pi(\boldsymbol{x}_i, \boldsymbol{\theta}) \boldsymbol{x}_i \,. \tag{2.3.4}$$

Note that the intercept term is suppressed in the estimating equation for notational simplicity. The solution can be found by using the following Newton-Raphson iterative procedure

$$\boldsymbol{\theta}^{(m+1)} = \boldsymbol{\theta}^{(m)} + \left\{ H(\boldsymbol{\theta}^{(m)}) \right\}^{-1} U(\boldsymbol{\theta}^{(m)}),$$

where $H(\boldsymbol{\theta}) = \sum_{i \in \mathcal{S}_B} d_i^B \pi(\boldsymbol{x}_i, \boldsymbol{\theta})\{1 - \pi(\boldsymbol{x}_i, \boldsymbol{\theta})\} \boldsymbol{x}_i \boldsymbol{x}_i^\intercal$ and the initial value for the iteration can be chosen as $\boldsymbol{\theta}^{(0)} = \mathbf{0}$.

We notice that Valliant and Dever (2011)'s weighted logistic regression procedure leads

to the following log-likelihood function,

$$l^*(\boldsymbol{\theta}) + \sum_{i \in \mathcal{S}_A} \log(1 - \pi_i^A) - \frac{n_A}{\hat{N}_B} \sum_{i \in \mathcal{S}_B} d_i^B \log(1 - \pi_i^A)\,, \tag{2.3.5}$$

where the first term $l^*(\boldsymbol{\theta})$ is the pseudo log-likelihood function given in (2.3.3). Obviously, the resulting estimating equations are not approximately unbiased unless $\mathcal{S}_A$ is a simple random sample from the population or its sampling fraction is negligibly small. The procedure in general does not lead to a consistent estimator for $\boldsymbol{\theta}_0$.

Besides from likelihood methods, alternative estimators for propensity scores can be derived with the method of estimating equations. Consider the following class of estimating equations,

$$\sum_{i=1}^{N} R_i \boldsymbol{h}(\boldsymbol{x}_i, \boldsymbol{\theta}) - \sum_{i=1}^{N} \pi(\boldsymbol{x}_i, \boldsymbol{\theta}) \boldsymbol{h}(\boldsymbol{x}_i, \boldsymbol{\theta}) = 0, \tag{2.3.6}$$

where $\boldsymbol{h}(\boldsymbol{x}_i, \boldsymbol{\theta})$ is a pre-specified smooth function of $\boldsymbol{\theta}$ which ensures that the equation system (2.3.6) has a unique solution. The equation system (2.3.6) was previously considered by Beaumont (2005) and Kim and Kim (2007). Under the current setting, we replace $\sum_{i=1}^{N} \pi(\boldsymbol{x}_i, \boldsymbol{\theta}) \boldsymbol{h}(\boldsymbol{x}_i, \boldsymbol{\theta})$ in (2.3.6) by $\sum_{i \in \mathcal{S}_B} d_i^B \pi(\boldsymbol{x}_i, \boldsymbol{\theta}) \boldsymbol{h}(\boldsymbol{x}_i, \boldsymbol{\theta})$, which leads to the class of solvable estimating equations,

$$\sum_{i \in \mathcal{S}_A} \boldsymbol{h}(\boldsymbol{x}_i, \boldsymbol{\theta}) - \sum_{i \in \mathcal{S}_B} d_i^B \pi(\boldsymbol{x}_i, \boldsymbol{\theta}) \boldsymbol{h}(\boldsymbol{x}_i, \boldsymbol{\theta}) = 0\,. \tag{2.3.7}$$

The maximum pseudo likelihood estimator $\hat{\boldsymbol{\theta}}_{ml}$ can be obtained from (2.3.7) by taking $\boldsymbol{h}(\boldsymbol{x}_i, \boldsymbol{\theta}) = \boldsymbol{x}_i$. Another natural choice is $\boldsymbol{h}(\boldsymbol{x}_i, \boldsymbol{\theta}) = \pi(\boldsymbol{x}_i, \boldsymbol{\theta})^{-1} \boldsymbol{x}_i$, where the system (2.3.6) becomes the conventional calibration type equations

$$\sum_{i \in \mathcal{S}_A} \frac{\boldsymbol{x}_i}{\pi(\boldsymbol{x}_i, \boldsymbol{\theta})} = \sum_{i=1}^{N} \boldsymbol{x}_i\,. \tag{2.3.8}$$

If the population totals $\sum_{i=1}^{N} \boldsymbol{x}_i$ are available from external sources, an estimator for $\boldsymbol{\theta}_0$

can be obtained immediately and does not require a probability sample. Kim and Riddles (2012) showed that, among the class of estimating equations (2.3.6), the choice $\boldsymbol{h}(\boldsymbol{x}_i, \boldsymbol{\theta}) = \pi(\boldsymbol{x}_i, \boldsymbol{\theta})^{-1}\boldsymbol{x}_i$ leads to the optimal estimation when a linear regression model holds for y given $\boldsymbol{x}$. Wu and Sitter (2001) and Wu (2003) contain detailed discussions on the efficiency of conventional calibration estimation and its relation to linear regression models.

Our proposed estimation procedure with two samples can be viewed as a data combination problem. Semiparametric models are one of the most widely used approaches to handle this type of problems; see, for instance, Chen et al. (2008), Graham et al. (2016) and Shu and Tan (2020). Standard semiparametric methods, however, are not applicable to the setting of the thesis, since the non-probability sample and the probability sample cannot be linked directly. It requires a joint randomization framework, which combines semiparametric models for the propensity scores and the outcome regression for the non-probability sample, and the design-based inference for finite populations from the probability sample. The similarities and differences between the current setting and other combining data problems are further highlighted in Section 2.5 on variance estimation, where two distinct variance components are involved, one from the semiparametric models and the other from the probability sample involving the variance of the survey weighted HT estimator.

2.4 Estimation of Finite Population Means

2.4.1 Inverse probability weighted estimators

The inverse probability weighted estimator is the most successful adoption of the HT estimator for missing data problems and causal inferences. The HT estimator was originally proposed by Horvitz and Thompson (1952) for a finite population with probability survey samples where the weights are determined by the sampling design. The IPW estimator, however, requires modelling on the propensity scores and its use in the survey context is referred to as the quasi-randomization approach or pseudo design-based approach.

For brevity, let $\pi_i(\boldsymbol{\theta}) = \pi(\boldsymbol{x}_i, \boldsymbol{\theta})$ for some $\boldsymbol{\theta}$. Then the estimated propensity scores based on the maximum pseudo likelihood estimator $\hat{\boldsymbol{\theta}}_{ml}$ is computed as $\pi_i(\hat{\boldsymbol{\theta}}_{ml})$ for $i \in \mathcal{S}_A$, which gives following two types of IPW estimators for the population mean μ_y, depending on whether the population size N is known or not:

$$\hat{\mu}_{IPW1} = \frac{1}{N} \sum_{i \in \mathcal{S}_A} \frac{y_i}{\pi_i(\hat{\boldsymbol{\theta}}_{ml})} \quad \text{and} \quad \hat{\mu}_{IPW2} = \frac{1}{\hat{N}^A} \sum_{i \in \mathcal{S}_A} \frac{y_i}{\pi_i(\hat{\boldsymbol{\theta}}_{ml})}, \tag{2.4.9}$$

where $\hat{N}^A = \sum_{i \in \mathcal{S}_A} 1/\pi_i(\hat{\boldsymbol{\theta}}_{ml})$. The estimator $\hat{\mu}_{IPW1}$ can be viewed as a HT-type estimator while $\hat{\mu}_{IPW2}$ can be viewed as a Hájek-type estimator. For the further theoretical development of the proposed estimators, we consider the asymptotic framework described in Section 2.9. Moreover, the properties of the IPW estimators, summarized in Theorem 2.1, are developed under both the model for the propensity scores and the survey design for the probability sample $\mathcal{S}_B$. Proof of the theorem is given in Section 2.9.

Theorem 2.1. *Under regularity conditions* **C1**–**C4** *specified in Section 2.9 and assuming the logistic regression model for the propensity scores, we have* $\hat{\mu}_{IPW1} - \mu_y = O_p(n_A^{-1/2})$, $\hat{\mu}_{IPW2} - \mu_y = O_p(n_A^{-1/2})$, $Var(\hat{\mu}_{IPW1}) = V_{IPW1} + o(n_A^{-1})$, $Var(\hat{\mu}_{IPW2}) = V_{IPW2} + o(n_A^{-1})$, *with*

$$V_{IPW1} = \frac{1}{N^2} \sum_{i=1}^{N} (1 - \pi_i^A) \pi_i^A \left(\frac{y_i}{\pi_i^A} - \mathbf{a}_1^\intercal \boldsymbol{x}_i \right)^2 + \mathbf{a}_1^\intercal \mathbf{D} \mathbf{a}_1 , \tag{2.4.10}$$

$$V_{IPW2} = \frac{1}{N^2} \sum_{i=1}^{N} (1 - \pi_i^A) \pi_i^A \left(\frac{y_i - \mu_y}{\pi_i^A} - \mathbf{a}_2^\intercal \boldsymbol{x}_i \right)^2 + \mathbf{a}_2^\intercal \mathbf{D} \mathbf{a}_2 , \tag{2.4.11}$$

where $\mathbf{a}_1^\intercal = \big\{ \sum_{i=1}^N (1 - \pi_i^A) y_i \boldsymbol{x}_i^\intercal \big\} \big\{ \sum_{i=1}^N \pi_i^A (1 - \pi_i^A) \boldsymbol{x}_i \boldsymbol{x}_i^\intercal \big\}^{-1}$, $\mathbf{a}_2^\intercal = \big\{ \sum_{i=1}^N (1 - \pi_i^A)(y_i - \mu_y) \boldsymbol{x}_i^\intercal \big\} \big\{ \sum_{i=1}^N \pi_i^A (1 - \pi_i^A) \boldsymbol{x}_i \boldsymbol{x}_i^\intercal \big\}^{-1}$, *and* $\mathbf{D} = N^{-2} V_p\big(\sum_{i \in \mathcal{S}_B} d_i^B \pi_i^A \boldsymbol{x}_i \big)$, *where* $V_p(\cdot)$ *denotes the design-based variance under the probability sampling design for* $\mathcal{S}_B$.

Under slightly tightened conditions for the propensity score model and the survey design on the sample $\mathcal{S}_B$ where both $N^{-1} \sum_{i=1}^N R_i \boldsymbol{x}_i$ and $N^{-1} \sum_{i \in \mathcal{S}_B} d_i^B \pi_i^A \boldsymbol{x}_i$ are asymptotically normally distributed, we have that both $(V_{IPW1})^{-1/2}(\hat{\mu}_{IPW1} - \mu_y)$ and $(V_{IPW2})^{-1/2}(\hat{\mu}_{IPW2} - \mu_y)$

converge to the standard normal distribution $N(0,1)$.

2.4.2 Doubly robust estimators

The IPW estimators are sensitive to model misspecifications, especially when certain units have very small values in the estimated propensity scores. See, for instance, Tan (2007) for further discussions. The efficiency and the robustness of IPW estimators can be improved by incorporating a prediction model for the response variable. Robins et al. (1994) identified a class of augmented inverse probability weighted (AIPW) estimators under the two-model framework, and showed the improved efficiency of AIPW estimators over the IPW estimators when both models are correct. Scharfstein et al. (1999) further noticed that this class of AIPW estimators remains consistent as long as one of the two models is correctly specified. This is the so-called double robustness property that is widely studied in the recent literature on missing data problems.

Consider parametric model $E_\xi(y \mid \boldsymbol{x}) = m(\boldsymbol{x},\boldsymbol{\beta}_0)$ for the response y given the $\boldsymbol{x}$, where $\boldsymbol{\beta}_0$ is the true model parameter. For notational convenience, we let $m_i(\boldsymbol{\beta}) = m(\boldsymbol{x}_i,\boldsymbol{\beta})$ for some $\boldsymbol{\beta}$. Then the typical form of doubly robust estimators for μ_y is given by

$$\hat{\mu}_{DR} = \frac{1}{N}\sum_{i=1}^{N} \frac{R_i\{y_i - m_i(\hat{\boldsymbol{\beta}})\}}{\pi_i(\hat{\boldsymbol{\theta}})} + \frac{1}{N}\sum_{i=1}^{N} m_i(\hat{\boldsymbol{\beta}})\,, \tag{2.4.12}$$

where $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\beta}}$ are some estimators of parameters $\boldsymbol{\theta}_0$ and $\boldsymbol{\beta}_0$ under each of the two models. In particular, due to ignorability condition, $\hat{\boldsymbol{\beta}}$ can be easily obtained by the method such as least square and maximum likelihood estimation with data of sample $\mathcal{S}_A$ only (see Section 2.2.1). The estimator $\hat{\mu}_{DR}$ given by (2.4.12) is identical to the model-assisted "generalized difference estimator" discussed in Wu and Sitter (2001) under scenarios where the complete auxiliary information $\{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N\}$ is available. Our proposed DR estimator for μ_y under

the current setting is given by

$$\hat{\mu}_{DR1} = \frac{1}{N}\sum_{i\in\mathcal{S}_A} d_i^A\{y_i - m_i(\hat{\boldsymbol{\beta}})\} + \frac{1}{N}\sum_{i\in\mathcal{S}_B} d_i^B m_i(\hat{\boldsymbol{\beta}}), \tag{2.4.13}$$

where $d_i^A = 1/\pi_i(\hat{\boldsymbol{\theta}}_{ml})$. An alternative estimator using the estimated population size is given by

$$\hat{\mu}_{DR2} = \frac{1}{\hat{N}^A}\sum_{i\in\mathcal{S}_A} d_i^A\{y_i - m_i(\hat{\boldsymbol{\beta}})\} + \frac{1}{\hat{N}^B}\sum_{i\in\mathcal{S}_B} d_i^B m_i(\hat{\boldsymbol{\beta}}), \tag{2.4.14}$$

and estimator $\hat{\mu}_{DR2}$ can be viewed as the sum of two Hájek estimators.

The development of theoretical properties of $\hat{\mu}_{DR1}$ and $\hat{\mu}_{DR2}$ requires a joint randomization framework involving the propensity score model for $\mathcal{S}_A$, the outcome regression model ξ, and the probability sampling design for $\mathcal{S}_B$. We allow model misspecification and assume that $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}^* + O_p(n_A^{-1/2})$ for some fixed $\boldsymbol{\beta}^*$. The value of $\boldsymbol{\beta}^*$ is the same as the true parameter $\boldsymbol{\beta}_0$ when the regression model is correctly specified but has no practical meanings otherwise. An important feature of estimators $\hat{\mu}_{DR1}$ and $\hat{\mu}_{DR2}$ is that variation induced by estimating $\boldsymbol{\beta}_0$ does not have any impact on asymptotic variances. Formally, we define $\tilde{\mu}_{DR1}$ and $\tilde{\mu}_{DR2}$ by replacing $\hat{\boldsymbol{\beta}}$ with $\boldsymbol{\beta}^*$ in $\hat{\mu}_{DR1}$ and $\hat{\mu}_{DR2}$ respectively. Under correctly specified propensity score model, we have $\tilde{\mu}_{DR1} = \hat{\mu}_{DR1} + o_p(n_A^{-\frac{1}{2}})$ and $\tilde{\mu}_{DR2} = \hat{\mu}_{DR2} + o_p(n_A^{-\frac{1}{2}})$, which further means that $Var(\hat{\mu}_{DR1})$ is asymptotically equivalent to $Var(\tilde{\mu}_{DR1})$, and $Var(\hat{\mu}_{DR2})$ is asymptotically equivalent to $Var(\tilde{\mu}_{DR2})$. Justification can be found in the proof of Theorem 2.2 in Section 2.9. Similarly, we assume that $\hat{\boldsymbol{\theta}}_{ml} = \boldsymbol{\theta}^* + O_p(n_A^{-1/2})$ for some fixed $\boldsymbol{\theta}^*$, which is not necessarily equal to $\boldsymbol{\theta}_0$ when the propensity score model is misspecified.

We consider the logistic regression model for the propensity scores and focus on the practically useful estimator $\hat{\mu}_{DR2}$ in the following theorem.

Theorem 2.2. *The estimator $\hat{\mu}_{DR2}$ is doubly robust in the sense that it is a consistent estimator of μ_y if either the propensity score model or the outcome regression model is correctly specified. Furthermore, under the regularity conditions* **C1**–**C6** *specified in Section 2.9 and the correctly specified logistic regression model for the propensity scores, we have*

$Var\big(\hat{\mu}_{DR2}\big) = V_{DR2} + o(n_A^{-1})$ *where*

$$V_{DR2} = \frac{1}{N^2}\sum_{i=1}^{N}(1-\pi_i^A)\pi_i^A\big[\{y_i - m_i(\boldsymbol{\beta}^*) - h_N\}/\pi_i^A - \mathbf{a}_3^\intercal \boldsymbol{x}_i\big]^2 + W_1\,, \tag{2.4.15}$$

where $\mathbf{a}_3^\intercal = \big[\sum_{i=1}^N(1-\pi_i^A)\big\{y_i - m_i(\boldsymbol{\beta}^*) - h_N\big\}\boldsymbol{x}_i^\intercal\big]\big\{\sum_{i=1}^N \pi_i^A(1-\pi_i^A)\boldsymbol{x}_i\boldsymbol{x}_i^\intercal\big\}^{-1}$, $h_N = N^{-1}\sum_{i=1}^N\big\{y_i - m_i(\boldsymbol{\beta}^*)\big\}$, *and* $W_1 = N^{-2}V_p\big(\sum_{i\in\mathcal{S}_B} d_i^B t_i\big)$ *is the design-based variance with* $t_i = \pi_i^A\boldsymbol{x}_i^\intercal\mathbf{a}_3 + m_i(\boldsymbol{\beta}^*) - N^{-1}\sum_{i=1}^N m_i(\boldsymbol{\beta}^*)$.

The comparison of efficiency between IPW estimators and DR estimators is not a straightforward topic and has been studied extensively in the missing data literature. See, for instance, Robins et al. (1994), Tan (2007), Cao et al. (2009), among others. The DR estimators are constructed through the residual variable $e_i = y_i - m(\boldsymbol{x}_i,\boldsymbol{\beta})$ and usually has smaller variance if the regression model provides a good fit to the non-probability survey data.

2.5 Variance Estimation

The asymptotic variance formulas presented in Section 2.4 provide a simple plug-in method for variance estimation. However, the asymptotic variance formulas for the DR estimators are derived under the assumption that the model for propensity scores is correctly specified. The plug-in variance estimator becomes inconsistent when the propensity score model is misspecified. The doubly robust variance estimation technique proposed by Kim and Haziza (2014) is a preferred approach and can be implemented under the current context.

2.5.1 Plug-in variance estimators

We show the details of the plug-in variance estimator for the IPW estimator $\hat{\mu}_{IPW2}$. Using the asymptotic variance formula (2.4.11) presented in Theorem 2.1, the first component

$N^{-2}\sum_{i=1}^{N}(1-\pi_i^A)\pi_i^A\{(y_i-\mu_y)/\pi_i^A-\mathbf{a}_2^\intercal\boldsymbol{x}_i\}^2$ can be consistently estimated by

$$\frac{1}{N^2}\sum_{i\in\mathcal{S}_A}\{1-\pi_i(\hat{\boldsymbol{\theta}}_{ml})\}\left\{\frac{y_i-\hat{\mu}_{IPW2}}{\pi_i(\hat{\boldsymbol{\theta}}_{ml})}-\hat{\mathbf{a}}_2^\intercal\boldsymbol{x}_i\right\}^2,$$

where N might be replaced by $\hat{N}^A$ if necessary, and

$$\hat{\mathbf{a}}_2^\intercal=\Big[\sum_{i\in\mathcal{S}_A}\{1/\pi_i(\hat{\boldsymbol{\theta}}_{ml})-1\}(y_i-\hat{\mu}_{IPW2})\boldsymbol{x}_i^\intercal\Big]\Big[\sum_{i\in\mathcal{S}_B}d_i^B\pi_i(\hat{\boldsymbol{\theta}}_{ml})\{1-\pi_i(\hat{\boldsymbol{\theta}}_{ml})\}\boldsymbol{x}_i\boldsymbol{x}_i^\intercal\Big]^{-1}.$$

The second piece $\mathbf{a}_2^\intercal\mathbf{D}\mathbf{a}_2$ can be estimated by $\hat{\mathbf{a}}_2^\intercal\hat{\mathbf{D}}\hat{\mathbf{a}}_2$, where $\hat{\mathbf{D}}$ is the design-based variance estimator and is given by

$$\hat{\mathbf{D}}=\frac{1}{N^2}\sum_{i\in\mathcal{S}_B}\sum_{j\in\mathcal{S}_B}\frac{\pi_{ij}^B-\pi_i^B\pi_j^B}{\pi_{ij}^B}\frac{\pi_i(\hat{\boldsymbol{\theta}}_{ml})}{\pi_i^B}\frac{\pi_j(\hat{\boldsymbol{\theta}}_{ml})}{\pi_j^B}\boldsymbol{x}_i\boldsymbol{x}_j^\intercal,$$

where π_i^B and π_{ij}^B are the first and second order inclusion probabilities for the probability sample $\mathcal{S}_B$. For certain sampling designs, determining the second order inclusion probabilities π_{ij}^B can incur theoretical or computational complexity. If so, approximate estimators for the design-based variance $\mathbf{D}$ are available from the survey sampling literature such as Berger (2004) and Brewer and Donadio (2003).

When the propensity score model is valid, a plug-in variance estimator for the DR estimator $\hat{\mu}_{DR2}$ can be similarly constructed based on the asymptotic variance formula V_{DR2} presented in Theorem 2.2.

2.5.2 Doubly robust variance estimators

Let E_q, E_ξ, E_p, V_q, V_ξ and V_p denote the expectation and variance under the propensity score model q, the outcome regression model ξ, and the probability sampling design p for $\mathcal{S}_B$, respectively. We have $E_q(R_i\mid\boldsymbol{x}_i)=\pi(\boldsymbol{x}_i,\boldsymbol{\theta}_0)$ and $V_q(R_i\mid\boldsymbol{x}_i)=\pi(\boldsymbol{x}_i,\boldsymbol{\theta}_0)\{1-\pi(\boldsymbol{x}_i,\boldsymbol{\theta}_0)\}$. We also have $E_\xi(y_i\mid\boldsymbol{x}_i)=m(\boldsymbol{x}_i,\boldsymbol{\beta}_0)$ and $V_\xi(y_i\mid\boldsymbol{x}_i)=v(\boldsymbol{x}_i)\sigma^2$.

The concept of DR variance estimation is appealing and has been discussed by several authors, including Haziza and Rao (2006) and Kim and Park (2006). The variance estimator is doubly robust if it is approximately unbiased for the variance of the DR point estimator when one of the models q or ξ is correctly specified. The uncertainty of not knowing which of the two models is valid for DR estimators presents a real challenge for variance estimation. In this section, we illustrate how to implement the method proposed by Kim and Haziza (2014) under the current setting of non-probability survey samples. Firstly, compute the following DR point estimator which is different from $\hat{\mu}_{DR1}$ and $\hat{\mu}_{DR2}$,

$$\hat{\mu}_{KH} = \frac{1}{N}\sum_{i=1}^{N} \frac{R_i\{y_i - m_i(\hat{\boldsymbol{\beta}}_{kh})\}}{\pi_i(\hat{\boldsymbol{\theta}}_{kh})} + \frac{1}{N}\sum_{i\in\mathcal{S}_B} d_i^B m_i(\hat{\boldsymbol{\beta}}_{kh})\,, \tag{2.5.16}$$

where the subscript "KH" indicates Kim and Haziza (2014)'s method. The form of this estimator is identical to $\hat{\mu}_{DR1}$ given in (2.4.13). However, instead of estimating $\boldsymbol{\theta}_0$ and $\boldsymbol{\beta}_0$ separately using the propensity score model and the regression model, estimated model parameters $(\hat{\boldsymbol{\theta}}_{kh}, \hat{\boldsymbol{\beta}}_{kh})$ are obtained by solving the following system of estimating equations:

$$\frac{1}{N}\sum_{i=1}^{N} \frac{R_i\dot{\pi}_i(\boldsymbol{\theta})}{\{\pi_i(\boldsymbol{\theta})\}^2}\{y_i - m_i(\boldsymbol{\beta})\} = \mathbf{0}\,, \tag{2.5.17}$$

$$\frac{1}{N}\sum_{i=1}^{N} \frac{R_i}{\pi_i(\boldsymbol{\theta})}\dot{m}_i(\boldsymbol{\beta}) - \frac{1}{N}\sum_{i\in\mathcal{S}_B} d_i^B \dot{m}_i(\boldsymbol{\beta}) = \mathbf{0}\,, \tag{2.5.18}$$

where $\dot{\pi}_i(\boldsymbol{\theta}) = \partial\pi(\boldsymbol{x}_i,\boldsymbol{\theta})/\partial\boldsymbol{\theta}$ and $\dot{m}_i(\boldsymbol{\beta}) = \partial m(\boldsymbol{x}_i,\boldsymbol{\beta})/\partial\boldsymbol{\beta}$. And to keep simplicity, we assume that $\boldsymbol{\beta}^*$ and $\boldsymbol{\theta}^*$ are also the limits of $\hat{\boldsymbol{\beta}}_{kh}$ and $\hat{\boldsymbol{\theta}}_{kh}$ respectively.

There are two major consequences from this approach. The first is the asymptotic expansion of $\hat{\mu}_{KH}$ given by

$$\hat{\mu}_{KH} - \mu_y = \frac{1}{N}\sum_{i=1}^{N}\left\{\frac{R_i}{\pi_i(\boldsymbol{\theta}^*)} - 1\right\}\left\{y_i - m_i(\boldsymbol{\beta}^*)\right\} + \frac{1}{N}\sum_{i\in\mathcal{S}_B} d_i^B m_i(\boldsymbol{\beta}^*)$$

$$-\frac{1}{N}\sum_{i=1}^{N} m_i(\boldsymbol{\beta}^*) + o_p(n_A^{-1/2})\,. \tag{2.5.19}$$

The second consequence is the construction of a variance estimator which is approximately unbiased under the joint randomization involving either q or ξ (but not both), and the sampling design p, as shown below.

We first derive a variance estimator for $\hat{\mu}_{KH}$ under the joint randomization of q and p. It follows from (2.5.19) that $V_{qp}(\hat{\mu}_{KH}) = V_1 + V_2 + o(n_A^{-1})$ where

$$V_1 = V_q\left[\frac{1}{N}\sum_{i=1}^{N}\Big\{\frac{R_i}{\pi_i(\boldsymbol{\theta}^*)} - 1\Big\}\Big\{y_i - m_i(\boldsymbol{\beta}^*)\Big\}\right] = \frac{1}{N^2}\sum_{i=1}^{N}\frac{(1-\pi_i^A)\pi_i^A}{\{\pi_i(\boldsymbol{\theta}^*)\}^2}\Big\{y_i - m_i(\boldsymbol{\beta}^*)\Big\}^2\,,$$

and $V_2 = V_p\big\{N^{-1}\sum_{i\in\mathcal{S}_B} d_i^B m_i(\boldsymbol{\beta}^*)\big\}$. The second term V_2 is the design-based variance and can be estimated using standard methods for the sample $\mathcal{S}_B$. Let $\hat{V}_2$ be the estimator for V_2. We can estimate the first term V_1 by

$$\hat{V}_1 = \frac{1}{N^2}\sum_{i=1}^{N}\frac{R_i\{1-\pi_i(\hat{\boldsymbol{\theta}}_{kh})\}}{\{\pi_i(\hat{\boldsymbol{\theta}}_{kh})\}^2}\big\{y_i - m_i(\hat{\boldsymbol{\beta}}_{kh})\big\}^2\,. \tag{2.5.20}$$

The asymptotic variance of $\hat{\mu}_{KH}$ under the joint randomization ξ and p is given by $V_{\xi p}(\hat{\mu}_{KH} - \mu_y) = K_1 + V_2 + o(n_A^{-1})$, where V_2 is the same design-based variance as defined in $V_{qp}(\hat{\mu}_{KH})$ and

$$K_1 = V_\xi\left[\frac{1}{N}\sum_{i=1}^{N}\Big\{\frac{R_i}{\pi_i(\boldsymbol{\theta}^*)} - 1\Big\}\Big\{y_i - m_i(\boldsymbol{\beta}^*)\Big\}\right] = \frac{1}{N^2}\sum_{i=1}^{N}\Big\{\frac{R_i}{\pi_i(\boldsymbol{\theta}^*)} - 1\Big\}^2\sigma_i^2\,,$$

where $\sigma_i^2 = V_\xi(y_i \mid \boldsymbol{x}_i) = v(\boldsymbol{x}_i)\sigma^2$. It is apparent that $\hat{V}_1$ is not a valid estimator for K_1 under the model ξ, and the bias is given by

$$E_\xi(\hat{V}_1) - K_1 = \frac{1}{N^2}\sum_{i=1}^{N}\Big\{\frac{R_i}{\pi_i(\boldsymbol{\theta}^*)} - 1\Big\}\sigma_i^2 + o(n_A^{-1})\,.$$

An important observation is that the bias is non-negligible under the outcome regression model ξ but the expectation of the leading term in the bias under the propensity score model q is approximately zero. This leads to the following DR variance estimator for $\hat{\mu}_{KH}$,

$$v_{KH} = \hat{V}_1 + \hat{V}_2 - \frac{1}{N^2}\left\{ \sum_{i\in\mathcal{S}_A} \frac{\hat{\sigma}_i^2}{\pi_i(\hat{\boldsymbol{\theta}}_{kh})} - \sum_{i\in\mathcal{S}_B} d_i^B \hat{\sigma}_i^2 \right\},$$

where $\hat{\sigma}_i^2$ is the estimator of σ_i^2 for $i \in \mathcal{S}_A$. It should be noted that the variance estimator v_{KH} has several limitations. First of all, it is derived for the point estimator with known N. Secondly, it is constructed under the logistic regression model $\pi(\boldsymbol{x}_i,\boldsymbol{\theta}) = \exp(\boldsymbol{x}_i^\intercal\boldsymbol{\theta})/\{1 + \exp(\boldsymbol{x}_i^\intercal\boldsymbol{\theta})\}$ and is not valid if the propensity score model q is specified differently. Lastly, solutions to the equation system (2.5.17) and (2.5.18) may not exist under certain scenarios. For example, if the two working models are specified as $m(\boldsymbol{x}_i,\boldsymbol{\beta}) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}$ and $\pi(\boldsymbol{x}_i,\boldsymbol{\theta}) = \exp\{\theta_0 + \theta_1(x_{1i} + x_{2i})\}/[1 + \exp\{\theta_0 + \theta_1(x_{1i} + x_{2i})\}]$, then the equations in (2.5.18) becomes

$$\frac{1}{N}\sum_{i=1}^{N} \frac{R_i}{\pi_i(\boldsymbol{\theta})}\boldsymbol{x}_i = \frac{1}{N}\sum_{i\in\mathcal{S}_B} d_i^B \boldsymbol{x}_i\,,$$

where $\boldsymbol{x}_i = (1, x_{i1}, x_{i2})^\intercal$, which is an over-identified system with three equations for $\boldsymbol{\theta} = (\theta_0,\theta_1)$ and usually has no solutions. Bootstrap method can be used as an alternative approach to DR variance estimation. This is explored in the simulation studies reported in the next section.

2.6 Simulation Studies

Simulation studies consist of three parts. The first part is to compare the proposed procedure of estimating the propensity scores with other existing procedures discussed in Section 2.2.2. The second part is to examine different methods of estimating μ_y, including naive sample mean, IPW, regression prediction, and DR estimation. Simulations are specially designed such that double robustness property of DR estimators can be verified. The last

part is to evaluate variance estimators developed in the previous section.

To generate finite populations, we consider the following candidate models for the response variable y,

$$\xi 1: \; y_i = 1 + 2x_{1i} + 2x_{2i} + 2x_{3i} + \sigma_a \varepsilon_i \,, \quad i = 1,2,\cdots,N \,;$$

$$\xi 2: \; y_i = 1 + 2x_{1i} + 2x_{2i} + 2x_{3i} + 0.2x_{3i}^4 + \sigma_b \varepsilon_i \,, \quad i = 1,2,\cdots,N \,;$$

$$\xi 3: \; y_i = 1 + 2x_{1i} + 2x_{2i} + 2x_{3i} + 0.5x_{3i}^4 + \sigma_c \varepsilon_i \,, \quad i = 1,2,\cdots,N \,;$$

where N = 20,000, $x_{1i} = z_{1i}$, $x_{2i} = z_{2i} + 0.3x_{1i}$, $x_{3i} = z_{3i} + 0.3(x_{1i} + x_{2i})$, with $z_{1i} \sim Bernoulli(0.5)$, $z_{2i} \sim Uniform(0,2)$, $z_{3i} \sim N(0,1)$. The error term ε_i's are independently generated from $N(0,1)$, and values of σ_a, σ_b and σ_c are chosen such that the correlation coefficient ρ, between y and the linear predictor is controlled at some desirable level for model $\xi 1$, $\xi 2$ and $\xi 3$ respectively. Higher ρ means higher predicting power of the model.

Three candidate logistic regression models are considered for generating true propensity scores π_i^A for the non-probability sample $\mathcal{S}_A$,

$$q1: \; \log\left\{\pi_i^A/(1-\pi_i^A)\right\} = \theta_a + 0.3x_{1i} + 0.3x_{2i} + 0.3x_{3i} \,, \quad i = 1,2,\cdots,N \,;$$

$$q2: \; \log\left\{\pi_i^A/(1-\pi_i^A)\right\} = \theta_b + 0.3x_{1i} + 0.3x_{2i} + 0.3x_{3i} + 0.1x_{3i}^2 \,, \quad i = 1,2,\cdots,N \,;$$

$$q3: \; \log\left\{\pi_i^A/(1-\pi_i^A)\right\} = \theta_c + 0.3x_{1i} + 0.3x_{2i} + 0.3x_{3i} + 0.2x_{3i}^2 \,, \quad i = 1,2,\cdots,N \,;$$

where θ_a, θ_b and θ_c are set such that $\sum_{i=1}^N \pi_i^A = n_A$ for model $q1$, $q2$ and $q3$ respectively, and n_A is the target sample size.

We consider seven finite populations based on seven combinations of above candidate models, i.e., (ξ1,q1), (ξ1,q2), (ξ1,q3), (ξ2,q1), (ξ3,q1), (ξ2,q2) and (ξ3,q3). For example, when the combination (ξ1,q1) is adopted, then we generate y from model $\xi 1$, and generate π_i^A from model $q1$. In the meanwhile, no matter which population, among the seven, is used in the analysis, we always consider a simple linear regression $m(\boldsymbol{x}_i, \boldsymbol{\beta}) = \beta_0 + \beta_1 x_{1i} +$

$\beta_2 x_{2i} + \beta_3 x_{3i}$ as the working model for the outcome prediction, and logistic regression $\log\left[\pi(\boldsymbol{x}_i, \boldsymbol{\theta})/\{1 - \pi(\boldsymbol{x}_i, \boldsymbol{\theta})\}\right] = \theta_0 + \theta_1 x_{1i} + \theta_2 x_{2i} + \theta_3 x_{3i}$ as the working model for the propensity scores.

Based on the discrepancy between the true models for finite populations and specified working models $m(\boldsymbol{x}_i, \boldsymbol{\beta})$ and $\pi(\boldsymbol{x}_i, \boldsymbol{\theta})$, we can further categorize seven model combinations into four scenarios. (i) Both prediction and propensity score model are correctly specified, denoted by "TT"; and combination ($\xi 1$,$q1$) belongs to "TT"; (ii) Prediction model is correctly specified and propensity score model is misspecified, denoted by "TF"; and combination ($\xi 1$,$q2$) and ($\xi 1$,$q3$) belong to "TF". (iii) Prediction model is misspecified and propensity score model is correctly specified, denoted by "FT"; and combination ($\xi 2$,$q1$) and ($\xi 3$,$q1$) belong to "FT". (iv) Both models are misspecified, denoted by "FF"; and combination ($\xi 2$,$q2$) and ($\xi 3$,$q3$) belong to "FF". Moreover, according to the coefficient of covariate x_{3i}^4 in the model $\xi 2$ and $\xi 3$, we can view $m(\boldsymbol{x}_i, \boldsymbol{\beta})$ as a heavier misspecification for model $\xi 3$ than for model $\xi 2$. Similarly, $\pi(\boldsymbol{x}_i, \boldsymbol{\theta})$ is a heavier misspecification for model $q3$ than for model $q2$.

Once a finite population is generated, we repeatedly draw sample $\mathcal{S}_A$ and $\mathcal{S}_B$ from the population B times. The non-probability sample $\mathcal{S}_A$ with the target size n_A is selected by the Poisson sampling method with inclusion probabilities specified by π_i^A. The probability sample $\mathcal{S}_B$ with the target size n_B is taken by the randomized systematic PPS sampling method (Goodman and Kish, 1950; Hartley and Rao, 1962) with the inclusion probabilities π_i^B proportional to $z_i = c - x_{2i}$. The value of c is chosen to control the variation of the survey weights such that $\max z_i / \min z_i = 30$.

Our first task is to assess the proposed procedure of estimating propensity scores. A valid procedure will lead to consistent IPW estimators, so we directly check the performance of estimators $\hat{\mu}_{IPW1}$ and $\hat{\mu}_{IPW2}$. To show that the proposed procedure is an improvement over pre-existing methods, we also include IPW estimators $\hat{\mu}_{AB1}$ and $\hat{\mu}_{AB2}$, which have the same forms as $\hat{\mu}_{IPW1}$ and $\hat{\mu}_{IPW2}$ respectively, but with "propensity score" being defined by $\tilde{\pi}_i$ and estimated through unweighted logistic regression model (Lee,

2006). The subscript "AB" indicates estimation is based on the pooled sample $\mathcal{S}_{AB} = \mathcal{S}_A \cup \mathcal{S}_B$. Under the current setup, we compute $\hat{\mu}_{AB1} = N^{-1} \sum_{i \in \mathcal{S}_A} y_i / \pi_i(\hat{\boldsymbol{\theta}}_{AB})$ and $\hat{\mu}_{AB2} = \{ \sum_{i \in \mathcal{S}_A} 1/\pi_i(\hat{\boldsymbol{\theta}}_{AB}) \}^{-1} \sum_{i \in \mathcal{S}_A} y_i / \pi_i(\hat{\boldsymbol{\theta}}_{AB})$, where estimator $\hat{\boldsymbol{\theta}}_{AB}$ is the solution to the following equation system,

$$\sum_{i \in \mathcal{S}_A} \{1 - \pi(\boldsymbol{x}_i, \boldsymbol{\theta})\} \boldsymbol{x}_i - \sum_{i \in \mathcal{S}_B} \pi(\boldsymbol{x}_i, \boldsymbol{\theta}) \boldsymbol{x}_i = 0.$$

Similarly, we compute and include the weighted version of $\hat{\mu}_{AB1}$ and $\hat{\mu}_{AB2}$ given by Valliant and Dever (2011), i.e., $\hat{\mu}_{AB1,w} = N^{-1} \sum_{i \in \mathcal{S}_A} y_i / \pi_i(\hat{\boldsymbol{\theta}}_{AB,w})$ and $\hat{\mu}_{AB2,w} = \{ \sum_{i \in \mathcal{S}_A} 1/\pi_i(\hat{\boldsymbol{\theta}}_{AB,w}) \}^{-1} \sum_{i \in \mathcal{S}_A} y_i / \pi_i(\hat{\boldsymbol{\theta}}_{AB,w})$, where estimator $\hat{\boldsymbol{\theta}}_{AB,w}$ is the solution to estimating equations

$$\sum_{i \in \mathcal{S}_A} \{1 - \pi(\boldsymbol{x}_i, \boldsymbol{\theta})\} \boldsymbol{x}_i - (1 - \frac{n_A}{\hat{N}_B}) \sum_{i \in \mathcal{S}_B} d_i^B \pi(\boldsymbol{x}_i, \boldsymbol{\theta}) \boldsymbol{x}_i = 0.$$

The performance of a given estimator $\hat{\mu}$ is evaluated through the relative bias (in percentage, $\%RB$) and the mean squared error (MSE), which are computed as

$$\%RB = \frac{1}{B} \sum_{b=1}^{B} \frac{\hat{\mu}^{(b)} - \mu_y}{\mu_y} \times 100\,, \qquad MSE = \frac{1}{B} \sum_{b=1}^{B} (\hat{\mu}^{(b)} - \mu_y)^2\,,$$

where $\hat{\mu}^{(b)}$ is the estimator computed from the bth simulated sample, and B is set to 2,000.

For this part of the investigation, we only use model $(\xi 1, q1)$ to generate the finite population, which means that both propensity score model and prediction model are correctly specified ("TT"). We choose different sampling fractions, and see how sample sizes effect the performance of procedures. Specifically, we consider size combination $(500, 1000)$, $(2000, 1000)$, $(2000, 2000)$, $(5000, 1000)$ and $(5000, 2000)$, where values in the brackets indicate n_A and n_B respectively. Moreover, different predicting power of outcome regression models are considered by setting $\rho = 0.3$, 0.6 and 0.9. Simulated results are reported in Table 2.1, with some major observations being summarized as follows.

Table 2.1: Simulated $\%RB$ and MSE of IPW Estimators under Model $(\xi 1, q1)$

		$\rho = 0.3$		$\rho = 0.6$		$\rho = 0.9$	
Sample Sizes	Estimators	$\%RB$	MSE	$\%RB$	MSE	$\%RB$	MSE
$(500, 1000)$	$\hat{\mu}_A$	24.28	1.86	24.67	1.76	24.84	1.75
	$\hat{\mu}_{AB1}$	-92.72	24.16	-92.73	24.14	-92.74	24.13
	$\hat{\mu}_{AB2}$	-4.83	0.38	-5.02	0.14	-5.11	0.10
	$\hat{\mu}_{AB1,w}$	0.63	0.32	0.59	0.11	0.56	0.08
	$\hat{\mu}_{AB2,w}$	0.37	0.29	0.32	0.09	0.30	0.05
	$\hat{\mu}_{IPW1}$	-0.04	0.32	-0.10	0.12	-0.13	0.08
	$\hat{\mu}_{IPW2}$	-0.41	0.30	-0.47	0.09	-0.50	0.06
$(2000, 1000)$	$\hat{\mu}_A$	21.98	1.40	22.42	1.42	22.62	1.44
	$\hat{\mu}_{AB1}$	-83.64	19.66	-83.61	19.62	-83.60	19.61
	$\hat{\mu}_{AB2}$	8.60	0.26	8.80	0.23	8.90	0.23
	$\hat{\mu}_{AB1,w}$	2.34	0.12	2.43	0.08	2.47	0.08
	$\hat{\mu}_{AB2,w}$	2.17	0.09	2.26	0.05	2.30	0.04
	$\hat{\mu}_{IPW1}$	-0.08	0.12	-0.03	0.08	-0.01	0.07
	$\hat{\mu}_{IPW2}$	-0.52	0.10	-0.48	0.05	-0.45	0.05
$(2000, 2000)$	$\hat{\mu}_A$	21.98	1.40	22.42	1.42	22.62	1.44
	$\hat{\mu}_{AB1}$	-79.48	17.76	-79.47	17.73	-79.46	17.72
	$\hat{\mu}_{AB2}$	1.60	0.06	1.67	0.02	1.71	0.01
	$\hat{\mu}_{AB1,w}$	2.22	0.10	2.31	0.06	2.35	0.05
	$\hat{\mu}_{AB2,w}$	2.23	0.08	2.32	0.04	2.36	0.03
	$\hat{\mu}_{IPW1}$	-0.19	0.09	-0.15	0.05	-0.13	0.04
	$\hat{\mu}_{IPW2}$	-0.38	0.07	-0.34	0.03	-0.31	0.02
$(5000, 1000)$	$\hat{\mu}_A$	18.15	0.94	18.47	0.96	18.62	0.97
	$\hat{\mu}_{AB1}$	-66.3	12.35	-66.23	12.31	-66.20	12.30
	$\hat{\mu}_{AB2}$	12.27	0.44	12.48	0.44	12.57	0.45
	$\hat{\mu}_{AB1,w}$	4.83	0.13	4.91	0.12	4.95	0.12
	$\hat{\mu}_{AB2,w}$	4.77	0.09	4.86	0.08	4.90	0.08
	$\hat{\mu}_{IPW1}$	0.02	0.08	0.02	0.07	0.02	0.07
	$\hat{\mu}_{IPW2}$	-0.40	0.05	-0.40	0.04	-0.40	0.04
$(5000, 2000)$	$\hat{\mu}_A$	18.15	0.94	18.47	0.96	18.62	0.97
	$\hat{\mu}_{AB1}$	-62.13	10.85	-62.09	10.82	-62.07	10.81
	$\hat{\mu}_{AB2}$	7.95	0.19	8.08	0.19	8.13	0.19
	$\hat{\mu}_{AB1,w}$	4.71	0.10	4.79	0.09	4.83	0.09
	$\hat{\mu}_{AB2,w}$	4.79	0.09	4.87	0.08	4.91	0.08
	$\hat{\mu}_{IPW1}$	-0.11	0.05	-0.10	0.04	-0.10	0.04
	$\hat{\mu}_{IPW2}$	-0.28	0.04	-0.27	0.02	-0.27	0.02

(i) Estimators $\hat{\mu}_{AB1}$ and $\hat{\mu}_{AB2}$, which are obtained by the unweighed logistic regression model method, fail almost for each case; estimator $\hat{\mu}_{AB1}$, which has true N as population count, yields especially unreliable results. (ii) Estimators $\hat{\mu}_{AB1,w}$ and $\hat{\mu}_{AB2,w}$, which are obtained by the weighed logistic regression model method, have close performance to $\hat{\mu}_{IPW1}$ and $\hat{\mu}_{IPW2}$ when $n_A = 500$ and $n_B = 1{,}000$. However, the biases of $\hat{\mu}_{AB1,w}$ and $\hat{\mu}_{AB2,w}$ increase with size n_A, which agrees with the observation stated in (2.3.5). (iii) Proposed estimators $\hat{\mu}_{IPW1}$ and $\hat{\mu}_{IPW2}$ always have excellent performance regarding $\%RB$ and MSE in all the situations considered, and their efficiencies increase with n_A and n_B. (iv) Estimators with estimated N generally have better performance in terms of MSE compared to their counterpart having the true N.

In summary, the weighted logistic regression model based on $\mathcal{S}_{AB}$ is a conditionally valid procedure for estimating propensity scores. The bias of its resulting estimators for μ_y is negligibly small when the sample size n_A is relatively small, but there is a clear trend of growing bias as the sampling fraction gets larger. This is a counter-intuitive phenomena that augmenting sample would unexpectedly aggravate estimating results, so it is important to check the sampling fraction before applying this procedure. On the contrary, our proposed approach yields consistent estimators, whose accuracy improves with the sample size.

The second part is to compare point estimators obtained by different methods. We generate finite populations according to seven model combinations, with ρ being set to 0.3, 0.6 and 0.9. The results for sample size (500,1000) are reported in Table 2.2, with some key observations being highlighted below.

Table 2.2: Simulated $\%RB$ and MSE of Estimators of μ_y ($n_A = 500$, $n_B = 1{,}000$)

			$\rho = 0.3$		$\rho = 0.6$		$\rho = 0.9$	
Scenarios	Models	Estimators	$\%RB$	MSE	$\%RB$	MSE	$\%RB$	MSE
	(ξ1,q1)	$\hat{\mu}_A$	24.28	1.86	24.67	1.76	24.84	1.75
		$\hat{\mu}_{IPW1}$	-0.04	0.32	-0.10	0.12	-0.13	0.08
		$\hat{\mu}_{IPW2}$	-0.41	0.30	-0.47	0.09	-0.50	0.06

Continued on next page

Table 2.2 – *Continued from previous page*

			$\rho = 0.3$		$\rho = 0.6$		$\rho = 0.9$	
Scenarios	Models	Estimators	$\%RB$	MSE	$\%RB$	MSE	$\%RB$	MSE
TT		$\hat{\mu}_{REG}$	0.07	0.25	0.00	0.06	-0.04	0.02
		$\hat{\mu}_{DR1}$	0.03	0.30	-0.01	0.10	-0.03	0.06
		$\hat{\mu}_{DR2}$	0.03	0.26	-0.02	0.06	-0.04	0.02
		$\hat{\mu}_{KH}$	0.04	0.30	-0.01	0.10	-0.03	0.06
TF	(ξ1,q2)	$\hat{\mu}_{A}$	33.65	3.38	34.17	3.33	34.41	3.35
		$\hat{\mu}_{IPW1}$	-3.09	0.46	-3.02	0.17	-2.98	0.12
		$\hat{\mu}_{IPW2}$	-8.01	0.61	-7.94	0.35	-7.91	0.30
		$\hat{\mu}_{REG}$	-0.18	0.26	-0.14	0.06	-0.11	0.02
		$\hat{\mu}_{DR1}$	-0.30	0.34	-0.18	0.11	-0.13	0.06
		$\hat{\mu}_{DR2}$	-0.30	0.29	-0.19	0.07	-0.13	0.02
		$\hat{\mu}_{KH}$	-0.25	0.32	-0.16	0.10	-0.12	0.06
	(ξ1,q3)	$\hat{\mu}_{A}$	46.25	6.21	46.75	6.19	46.98	6.22
		$\hat{\mu}_{IPW1}$	-6.84	1.72	-6.19	0.56	-5.88	0.33
		$\hat{\mu}_{IPW2}$	-26.33	3.13	-25.94	2.46	-25.76	2.31
		$\hat{\mu}_{REG}$	0.30	0.32	0.09	0.07	0.00	0.02
		$\hat{\mu}_{DR1}$	-1.84	0.88	-0.78	0.20	-0.30	0.08
		$\hat{\mu}_{DR2}$	-1.21	0.56	-0.54	0.11	-0.23	0.03
		$\hat{\mu}_{KH}$	-0.06	0.39	-0.04	0.12	-0.03	0.07
FT	(ξ2,q1)	$\hat{\mu}_{A}$	32.25	4.80	32.81	4.48	33.07	4.46
		$\hat{\mu}_{IPW1}$	0.01	0.84	-0.07	0.25	-0.11	0.14
		$\hat{\mu}_{IPW2}$	-0.36	0.80	-0.44	0.21	-0.48	0.11
		$\hat{\mu}_{REG}$	-4.44	0.84	-4.54	0.28	-4.59	0.18
		$\hat{\mu}_{DR1}$	0.25	0.81	0.19	0.23	0.16	0.13
		$\hat{\mu}_{DR2}$	0.22	0.76	0.15	0.18	0.12	0.07
		$\hat{\mu}_{KH}$	0.04	0.80	-0.03	0.23	-0.06	0.12
	(ξ3,q1)	$\hat{\mu}_{A}$	40.13	12.18	40.97	11.03	41.36	10.93
		$\hat{\mu}_{IPW1}$	0.09	2.71	-0.03	0.68	-0.09	0.32
		$\hat{\mu}_{IPW2}$	-0.27	2.64	-0.40	0.63	-0.46	0.26
		$\hat{\mu}_{REG}$	-8.96	3.13	-9.11	1.22	-9.19	0.87
		$\hat{\mu}_{DR1}$	0.50	2.71	0.40	0.73	0.35	0.37
		$\hat{\mu}_{DR2}$	0.43	2.63	0.33	0.65	0.28	0.28
		$\hat{\mu}_{KH}$	0.06	2.63	-0.04	0.66	-0.09	0.31
	(ξ2,q2)	$\hat{\mu}_{A}$	61.05	15.59	61.81	15.52	62.17	15.60
		$\hat{\mu}_{IPW1}$	2.33	1.18	2.45	0.35	2.50	0.20

Continued on next page

Table 2.2 – *Continued from previous page*

			$\rho = 0.3$		$\rho = 0.6$		$\rho = 0.9$	
Scenarios	Models	Estimators	$\%RB$	MSE	$\%RB$	MSE	$\%RB$	MSE
FF		$\hat{\mu}_{IPW2}$	-2.89	1.12	-2.78	0.37	-2.73	0.24
		$\hat{\mu}_{REG}$	-5.79	0.97	-5.73	0.39	-5.69	0.28
		$\hat{\mu}_{DR1}$	15.63	2.13	15.81	1.49	15.9	1.38
		$\hat{\mu}_{DR2}$	14.25	1.82	14.42	1.19	14.51	1.08
		$\hat{\mu}_{KH}$	6.74	1.06	6.88	0.46	6.94	0.35
	($\xi 3$,$q3$)	$\hat{\mu}_{A}$	161.15	163.27	162.34	163.8	162.89	164.47
		$\hat{\mu}_{IPW1}$	33.80	25.74	35.24	17.39	35.91	16.51
		$\hat{\mu}_{IPW2}$	3.68	8.94	4.54	3.52	4.94	2.58
		$\hat{\mu}_{REG}$	-20.04	6.90	-20.50	4.53	-20.72	4.12
		$\hat{\mu}_{DR1}$	163.78	232.32	166.20	235.75	167.31	239.16
		$\hat{\mu}_{DR2}$	116.13	103.00	117.68	101.86	118.39	102.53
		$\hat{\mu}_{KH}$	34.69	11.55	34.78	8.86	34.82	8.37

(i) IPW estimators $\hat{\mu}_{IPW1}$ and $\hat{\mu}_{IPW2}$ perform well under the correctly specified propensity score model ("TT" and "FT"). But both estimators show noticeable bias when the true propensity score model is $q2$ or $q3$; and compared to using model $q2$, generating propensity scores from $q3$ leads to much more bias. Estimator $\hat{\mu}_{IPW2}$ has smaller MSE than $\hat{\mu}_{IPW1}$ under the correctly specified propensity score model, but this pattern does not hold under the misspecified propensity score model ("TF" and "FF"). (ii) The regression based estimator $\hat{\mu}_{REG}$ performs very well under the correctly specified prediction model ("TT" and "TF"), but the bias emerges when $\xi 2$ or $\xi 3$ is the true model to generate y. (iii) DR estimator $\hat{\mu}_{DR1}$, $\hat{\mu}_{DR2}$ and $\hat{\mu}_{KH}$ have excellent performance as long as at least one of the two models is correctly specified ("TT" , "FT" and "TF"). Estimator $\hat{\mu}_{DR2}$ has smaller MSE than $\hat{\mu}_{DR1}$ for all the cases we consider. (iv) As expected, none of estimators remain consistent when both models are misspecified ("FF"), and their performance under model combination $(\xi 3, q3)$ is much worse than that under combination $(\xi 2, q2)$.

Besides sample size (500,1000), we have also examined the size combination (500,500),

(1000,500), (1000,1000) and (2000,2000). The performance of considered methods under different size combinations basically demonstrate the same pattern as that of case (500,1000). The results therefore are not presented here to avoid repetition. The failure of DR estimation under "FF" scenario reminds us that DR estimators are not foolproof. It can be observed from the simulation results that DR estimators could even have worse performance than the unadjusted naive estimator. Moreover, the performance of DR estimators under "FF" scenario is highly unpredictable, and heavily dependent on the level of the model misspecification.

The last task of this section is to examine the behaviour of variance estimators. We consider variance estimators v_{IPW1} and v_{IPW2} associated with $\hat{\mu}_{IPW1}$ and $\hat{\mu}_{IPW2}$ based on Theorem 2.1 and the plug-in method described in Section 2.5.1. We also consider the variance estimator $v_{DR2,plug}$ for $\hat{\mu}_{DR2}$ using the plug-in method, and the DR variance estimator v_{KH} along with $\hat{\mu}_{KH}$ described in Section 2.5.2. Lastly, a bootstrap variance estimator for $\hat{\mu}_{DR2}$, denoted by $v_{DR2,bst}$ is obtained by the following procedure. Let $\mathcal{S}_A^{(j)}$ and $\mathcal{S}_B^{(j)}$ be the bootstrap samples of sizes n_A and n_B, taken respectively from $\mathcal{S}_A$ and $\mathcal{S}_B$ using simple random sampling (SRS) with replacement. For taking $\mathcal{S}_B^{(j)}$, weights $d_i^B = 1/\pi_i^B$ are treated as an intrinsic part of the dataset $\mathcal{S}_B$, i.e., both d_i^B and $\boldsymbol{x}_i$ are attached to unit i, and $\{(d_i^B,\boldsymbol{x}_i), i \in \mathcal{S}_B^{(j)}\}$ is the bootstrap dataset from $\mathcal{S}_B$. Note that the bootstrap procedures may select certain units more than once and all duplicated units are kept in the bootstrap samples. The bootstrap DR estimator $\hat{\mu}_{DR2}^{(j)}$ is computed by using the same procedure as computing $\hat{\mu}_{DR2}$, but based on $\mathcal{S}_A^{(j)}$ and $\mathcal{S}_B^{(j)}$. The final estimator $v_{DR2,bst}$ is the variance of the sequence $\{\hat{\mu}_{DR2}^{(1)}, \cdots \hat{\mu}_{DR2}^{(j)}, \cdots \hat{\mu}_{DR2}^{(J)}\}$, where $J = 1{,}000$. Note that the with-replacement bootstrap procedure provides valid variance estimator for single-stage PPS sampling with negligible sampling fractions; see, for instance, Rao et al. (1992).

The performance of a given variance estimator v along with the point estimator $\hat{\mu}$ is assessed by the percentage relative bias ($\%RB$) and the coverage probability ($\%CP$)

computed as

$$\%RB = \frac{1}{B}\sum_{b=1}^{B}\frac{v^{(b)} - V}{V} \times 100\,, \qquad \%CP = \frac{1}{B}\sum_{b=1}^{B} I\big(\mu_y \in CI^{(b)}\big) \times 100\,,$$

where $v^{(b)}$ is the variance estimator computed from the bth simulated sample, V is the Monte-carlo simulated variance of $\hat{\mu}$ obtained through a separate set of B simulation runs, $I(\cdot)$ is the indicator function, and $CI^{(b)} = \big[\hat{\mu}^{(b)} - 1.96(v^{(b)})^{1/2}, \hat{\mu}^{(b)} + 1.96(v^{(b)})^{1/2}\big]$ is the 95% confidence interval for μ_y based on the normal approximation.

Simulation results for $n_A = 500$ and $n_B = 1{,}000$ are reported in Table 2.3. The most important observation is that all the variance estimators and associated CIs have excellent performance when the propensity score model is correctly specified ("TT" and "FT"); the biases of the variance estimators are all small and the coverage probabilities of 95% CIs are close to the nominal value. When the propensity score model is misspecified, and the prediction model is correctly specified ("TF"), the IPW point estimators are invalid and the related CIs cannot be used; the plug-in variance estimator $v_{DR2,plug}$ for $\hat{\mu}_{DR2}$ has enormous positive bias, which incurs serious over-coverage issue for the resulting CIs; the DR variance estimator v_{KH} coupled with $\hat{\mu}_{KH}$ has excellent performance, while bootstrap estimator $v_{DR2,bst}$ also has noticeable improvement over $v_{DR2,plug}$. In general, none of variance estimators have adequate performance under "FF" scenarios, especially under model $(\xi 3, q3)$.

2.7 Real Data Application

In this section, we apply our proposed methods to a dataset collected by the Pew Research Centre (http://www.pewresearch.org) in 2015. The dataset consists of nine non-probability samples with a total of 9,301 individuals and a wide range of measurements over 56 variables. The nine non-probability samples are supplied by eight vendors, which have different but unknown strategies in panel recruitment, sampling, incentives for participation, etc.

Table 2.3: Simulated $\%RB$ and $\%CP$ of Variance Estimators ($n_A = 500$, $n_B = 1{,}000$)

			$\rho = 0.3$		$\rho = 0.6$		$\rho = 0.9$	
Scenarios	Models	Estimators	$\%RB$	$\%CP$	$\%RB$	$\%CP$	$\%RB$	$\%CP$
TT	(ξ1,q1)	v_{IPW1}	-5.56	94.50	-3.30	94.55	-1.56	93.90
		v_{IPW2}	-5.68	94.55	-4.99	94.60	-4.39	94.05
		$v_{DR2,plug}$	-4.49	94.35	-2.95	94.80	-0.48	94.75
		$v_{DR2,bst}$	-2.39	94.65	-1.56	94.85	-0.58	94.70
		v_{KH}	-5.24	94.70	-2.28	94.95	0.10	94.25
TF	(ξ1,q2)	v_{IPW1}	0.49	95.00	-1.03	92.55	-0.99	88.85
		v_{IPW2}	10.04	93.35	5.20	87.95	2.94	83.65
		$v_{DR2,plug}$	25.18	97.15	18.50	97.05	5.94	95.45
		$v_{DR2,bst}$	5.76	95.60	3.34	95.20	0.06	94.45
		v_{KH}	-1.99	94.70	-2.33	93.85	-1.07	93.70
	(ξ1,q3)	v_{IPW1}	9.41	97.00	7.05	95.10	3.56	93.50
		v_{IPW2}	79.12	93.90	66.65	80.25	60.35	65.20
		$v_{DR2,plug}$	141.03	99.10	120.61	99.00	57.61	98.05
		$v_{DR2,bst}$	12.89	96.35	10.24	95.95	3.55	95.40
		v_{KH}	-5.65	94.40	-3.93	94.65	-1.11	93.85
FT	(ξ2,q1)	v_{IPW1}	-5.79	94.20	-4.02	94.5	-2.33	94.15
		v_{IPW2}	-5.48	94.45	-4.78	94.20	-4.34	94.50
		$v_{DR2,plug}$	-4.33	94.55	-2.46	94.05	0.34	95.30
		$v_{DR2,bst}$	-1.92	94.85	0.40	94.70	4.19	95.75
		v_{KH}	-5.82	94.50	-3.98	94.60	-2.22	94.60
	(ξ3,q1)	v_{IPW1}	-5.90	94.40	-4.43	94.25	-2.76	94.60
		v_{IPW2}	-5.36	94.40	-4.45	94.00	-3.68	94.40
		$v_{DR2,plug}$	-4.33	94.45	-2.52	94.40	-0.08	94.85
		$v_{DR2,bst}$	-1.60	94.60	1.69	94.70	6.77	95.60
		v_{KH}	-6.17	94.45	-5.19	93.90	-4.34	94.10
FF	(ξ2,q2)	v_{IPW1}	1.44	95.35	0.59	95.00	1.05	94.25
		v_{IPW2}	11.28	96.30	5.85	95.35	1.56	93.40
		$v_{DR2,plug}$	25.71	91.45	21.29	73.95	17.87	53.65
		$v_{DR2,bst}$	9.54	88.40	14.21	71.90	18.11	54.75
		v_{KH}	-1.77	92.50	-2.92	87.40	-2.84	82.30
	(ξ3,q3)	v_{IPW1}	10.68	93.45	9.15	86.60	7.91	75.85
		v_{IPW2}	86.19	97.95	90.46	98.05	98.62	97.75
		$v_{DR2,plug}$	112.99	67.20	104.58	34.85	103.06	29.55
		$v_{DR2,bst}$	7.20	34.75	4.55	14.05	3.78	11.10
		v_{KH}	-8.86	69.85	-10.20	30.05	-9.19	7.80

In this analysis we treat the dataset as a single non-probability sample with $n_A = 9{,}301$. The dataset is referred to as PRC from now. Four study variables of the PRC dataset are of particular interest, but no valid inferences are immediately available from the PRC sample given its non-probability based nature.

We let two probability samples, which are taken in the same period as the PRC sample, as sources of auxiliary information. The first sample is the volunteer supplement survey data from the Current Population Survey (CPS), which is one of the most reliable sources of official statistics in the United States. The CPS dataset contains 80,075 cases with measurements on volunteerism, which is highly relevant to the study variables considered in the PRC dataset. The second probability sample is the Behavioral Risk Factor Surveillance System (BRFSS) survey data (https://www.cdc.gov/brfss/index.html). It is designed to measure behavioral risk factors for US residents and has a large sample size of 441,456. Neither of probability samples contain measurements of the study variables, but both share a rich set of common auxiliary variables with the PRC dataset as shown in Table 2.4.

We first examine marginal distributions of the variables from three datasets. Table 2.4 contains the estimated population means using each of the three datasets. For the PRC dataset, the sampling strategy is unknown and no survey weights are available, so estimates presented are unadjusted simple sample means. For the BRFSS and the CPS dataset where survey weights are available as part of the datasets, survey weighted estimates are used. "NA" in the table indicates that the variable is not available from the dataset. While the two reference probability samples provide similar results over most of the variables, there is a clear discrepancy between the non-probability PRC sample and the two reference samples on age, race, origin and socioeconomic status. For instance, the PRC sample has 9.3% participants with Hispanic/Latino origin and close to 42% with a bachelor's degree or above, the corresponding numbers from the CPS sample are 15.6% and 30.9%. It is a strong evidence that the PRC dataset is not a representative sample for the population.

Four study variables of PRC dataset are listed at the end of the Table 2.4 along with their simple sample means. There are three binary variables: *Talk with neighbors frequently*

Table 2.4: Estimated Population Means of Survey Items from the Three Samples

Categories	Items	PRC	CPS	BRFSS
Age	<30	0.183	0.212	0.209
	>=30, <50	0.326	0.336	0.333
	>=50, <70	0.387	0.326	0.327
	>=70	0.104	0.126	0.131
Gender	Female	0.544	0.518	0.513
Race	White only	0.823	0.786	0.750
	Black only	0.088	0.125	0.126
Origin	Hispanic/Latino	0.093	0.156	0.165
Region	Northeast	0.200	0.180	0.177
	South	0.275	0.373	0.383
	West	0.299	0.235	0.232
Marital status	Married	0.503	0.528	0.508
Employment	Working	0.521	0.589	0.566
	Retired	0.243	0.143	0.179
Education	High school or less	0.216	0.407	0.427
	Bachelor's degree and above	0.416	0.309	0.263
Household	Presence of child in household	0.289	NA	0.368
	Home ownership	0.654	NA	0.672
Health	Smoke everyday	0.157	NA	0.115
	Smoke never	0.798	NA	0.833
Financial status	No money to see doctors	0.207	NA	0.133
	Having medical insurance	0.891	NA	0.878
	Household income < 20K	0.161	0.153	NA
	Household income >100K	0.199	0.233	NA
Volunteerism	Volunteered	0.510	0.248	NA
Study variables	Talk with neighbors frequently (y_1)	0.461	NA	NA
	Participated in school groups (y_2)	0.210	NA	NA
	Participated in service organizations (y_3)	0.141	NA	NA
	Days had at least one drink last month (y_4)	5.301	NA	NA

(y_1), *Participated in school groups* (y_2), *Participated in service organizations* (y_3); and one count variable: *Days had at least one drink last month* (y_4), which is treated as a continuous variable in the analysis. While keeping the estimation of the population means as the primary goal, we approach the problem from three specific aspects: (i) the impact of sample size on the DR estimation, (ii) the covariate availability and selection, (iii) comparison of different estimation methods.

2.7.1 Impact of relative sample size

Note sample size n_B is much larger than n_A no matter CPS or BRFSS dataset is used as the reference sample. To investigate other possible scenarios where n_A and n_B has different ratios, we draw three subsamples from original BRFSS dataset by SRS with replacement method. The resulting subsamples, denoted by BRFSS$^{(\mathrm{L})}$, BRFSS$^{(\mathrm{M})}$ and BRFSS$^{(\mathrm{S})}$, have sample size $n_B^* = 80{,}000$, 8,000 and 800, respectively. Survey weights for each of the subsamples are computed as $d_i^B n_B / n_B^*$, where d_i^B is the weight of unit i in the original BRFSS sample with $n_B = 441{,}456$.

The choice of covariates for modelling is constrained by the availability of covariates. In other words, only the common covariates between the chosen reference sample and the non-probability sample can be used to obtain models. We consider variables listed in Table 2.4, except four study variables, as candidate covariates for modelling. Note the set of common covariates between BRFSS and PRC differs from the set between CPS and PRC. For the current investigation, we only include covariates which are available from all three datasets, such that the same set of covariates is used for modelling under either probability samples. Throughout the entire real data analysis, logistic regression model is adopted for the propensity scores; and for the prediction model, logistic regression model and linear regression model are considered under binary and continuous response respectively.

We first compute proposed estimator $\hat{\mu}_{DR2}$ for four response variables using five different probability samples, and the results are presented in Table 2.5. The first row specifies which probability sample is used as sample $\mathcal{S}_B$ except for the column under "PRC" which provides

simple sample means based on PRC dataset. It can be seen that the three larger probability samples BRFSS, BRFSS$^{(L)}$ and BRFSS$^{(M)}$ with $n_B = 441{,}456$ and $n_B^* = 80{,}000$ and 8,000 produce almost identical results. The smallest probability sample BRFSS$^{(S)}$ with $n_B^* = 800$ leads to noticeably different results for responses y_2 and y_4, indicating potential inconsistent estimates when the size of the probability sample is too small. Nevertheless, compared to simple sample means, DR estimators generally agree with each others regardless the choice of $\mathcal{S}_B$.

Table 2.5: Estimator $\hat{\mu}_{DR2}$ by Different Reference Samples

Response	PRC	CPS	BRFSS	BRFSS$^{(L)}$	BRFSS$^{(M)}$	BRFSS$^{(S)}$
y_1	0.461	0.458	0.457	0.458	0.457	0.456
y_2	0.210	0.207	0.202	0.202	0.204	0.208
y_3	0.141	0.136	0.133	0.133	0.133	0.135
y_4	5.301	5.013	4.971	4.980	4.951	4.910

We further look at variance estimators with different probability samples, using the plug-in method described in Section 2.5.1 and bootstrap method described in Section 2.6. The plug-in variance estimator $v_{DR2,plug}$ can be further decomposed into two components according to the formula given by Theorem 2.2. One component is attributed to the selection mechanism for $\mathcal{S}_A$, and the other component is the design-based variation of obtaining $\mathcal{S}_B$. Let two components be denoted by v_A and v_B respectively, then the plug-in variance estimator is computed as $v_{DR2,plug} = v_A + v_B$. Unfortunately, detailed design information other than the survey weights is not available for either the BRFSS or the CPS sample. We use an approximate variance formula for v_B by assuming that the survey design is single-stage PPS sampling with replacement, a strategy often used by survey data analyst for the purpose of variance estimation. The bootstrap variance estimator $v_{DR2,bst}$ is computed based on $J = 5{,}000$ bootstrap samples, however variance decomposition cannot be done for the bootstrap method.

Results for variance estimators with decomposition (v_A, v_B) for $v_{DR2,plug}$ are reported in Table 2.6. The variance estimators have been multiplied by 10^5 for binary response variables and by 10^2 for the continuous variable to facilitate reading. We have the fol-

lowing major observations from Table 2.6: (i) the two original probability samples CPS and BRFSS produce similar total variances for all cases, with the design-based variance component v_B making a negligible contribution; (ii) when the size of the probability sample becomes smaller, from BRFSS, BRFSS$^{(L)}$, BRFSS$^{(M)}$ to BRFSS$^{(S)}$, the total variance becomes larger and the variance component v_B from the probability sampling design becomes dominant; (iii) two types of variance estimators $v_{DR2,plug}$ and $v_{DR2,bst}$ show the similar pattern of behaviour in terms of different sample sizes, (iv) estimator $v_{DR2,bst}$ is always larger than $v_{DR2,plug}$ for each case, and the reason could be the discrepancy between the assumed sampling strategy for $\mathcal{S}_B$ and the true sampling design for $\mathcal{S}_B$.

Table 2.6: Variance and Variance Components of Estimator $\hat{\mu}_{DR2}$ by Different Reference Samples

Response	Estimators	CPS	BRFSS	BRFSS$^{(L)}$	BRFSS$^{(M)}$	BRFSS$^{(S)}$
y_1	$v_{DR2,plug}$	3.998 (3.992,0.006)	3.784 (3.780, 0.004)	3.790 (3.768, 0.021)	4.098 (3.796, 0.302)	5.988 (4.173, 1.815)
	$v_{DR2,bst}$	4.212	4.141	4.280	4.665	6.636
y_2	$v_{DR2,plug}$	2.519 (2.502, 0.017)	2.329 (2.320, 0.008)	2.368 (2.322, 0.046)	2.783 (2.367, 0.416)	6.611 (2.659, 3.952)
	$v_{DR2,bst}$	2.635	2.592	2.693	3.134	7.470
y_3	$v_{DR2,plug}$	1.790 (1.783, 0.007)	1.656 (1.652, 0.004)	1.680 (1.661, 0.019)	1.915 (1.681, 0.234)	3.668 (1.876, 1.792)
	$v_{DR2,bst}$	1.973	1.906	2.017	2.228	4.264
y_4	$v_{DR2,plug}$	0.911 (0.908, 0.003)	0.900 (0.898, 0.002)	0.910 (0.899, 0.011)	1.046 (0.890, 0.156)	1.707 (0.922, 0.785)
	$v_{DR2,bst}$	0.965	1.046	1.021	1.155	1.966

2.7.2 Covariate selection

This part is to explore the covariate selection. The computation of the DR estimator require both the prediction model and the propensity score model. According to Assumption **A1**, the definition of the propensity scores and the model-based prediction approach discussed in Section 2.2.1, every covariate which is simultaneously related to the response variable and selection mechanism for sample $\mathcal{S}_A$ should be included in both models to fully remove

the selection bias in sample $\mathcal{S}_A$. In practice, this requirement is hard to achieve. The largest barrier is the potential unavailability of certain key covariates from two samples. This type of issue is usually hard to resolve since obtaining measurements of variables which were originally not in the samples could be unrealistic or extremely challenging. So our first task in this part is to show that the availability of significant covariates is extremely important to proposed methods, and if key covariates are missing from models, the proposed methods will be less effective in removing the selection bias. The second issue we want to investigate is that under the scenario where the available covariates are already very limited, whether we should include all the available covariates in the model no matter they are significant or not.

More specifically, let us investigate the following three covariate selection strategies. (i) Only use the covariates which are available in all three datasets. This is also the strategy which has been considered in Section 2.7.1. The resulting set of covariates are denoted by $\boldsymbol{x}.partial$, since only partial common covariates are used for modelling under a given reference sample. (ii) Use the set of common covariates between PRC and a given probability sample, i.e., all the available common covariates between PRC and CPS or all the available common covariates between PRC and BRFSS. This strategy leads to two larger but different set of covariates for the two probability samples. The resulting set is denoted as $\boldsymbol{x}.large$. (iii) Use set $\boldsymbol{x}.large$ as candidate, but covariates are further selected through the backward variable selection algorithm for the prediction model, i.e., only covariates which are significant for the prediction model are kept to construct both models. The resulting set is denoted as $\boldsymbol{x}.select$. P-values for covariates in the $\boldsymbol{x}.large$ and $\boldsymbol{x}.select$ based prediction models are listed in Table 2.7 and Table 2.8 respectively.

Estimator $\hat{\mu}_{DR2}$ for the four response variables using different sets of covariates and probability samples are presented in Table 2.9, with the plug-in variance estimators displayed in the parentheses (multiplied by 10^5 for binary variables and by 10^2 for the continuous variable). Major observations from Table 2.9 can be summarized as follows. (i) With a chosen sample $\mathcal{S}_B$, the point estimators using covariates $\boldsymbol{x}.large$ are very similar to the estimators using covariates $\boldsymbol{x}.select$ which drops some non-significant covariates. (ii) With a

Table 2.7: P-values by Using Covariate Set $\boldsymbol{x}.large$

	CPS				BRFSS			
Covariates	y_1	y_2	y_3	y_4	y_1	y_2	y_3	y_4
Intercept	$*$	$*$	$*$	$*$	$*$	$*$	$*$	$*$
Age	$*$	$*$	0.188	$*$	$*$	0.044	0.305	0.071
Female	$*$	0.990	$*$	$*$	$*$	0.175	$*$	$*$
White only	0.010	0.163	0.665	0.066	0.064	0.074	0.879	0.086
Black only	$*$	$*$	$*$	0.589	0.002	$*$	$*$	0.805
Hispanic/Latino	0.122	$*$	$*$	0.335	0.164	0.001	$*$	0.108
Northeast	$*$	0.514	0.068	0.987	$*$	0.617	0.125	0.688
South	0.007	$*$	0.252	0.985	0.011	$*$	0.461	0.933
West	0.087	0.001	0.668	0.060	0.032	$*$	0.908	0.013
Married	0.021	$*$	$*$	0.769	0.142	$*$	$*$	0.030
Working	0.864	0.062	0.010	$*$	0.406	0.001	$*$	$*$
Retired	0.007	0.912	0.002	$*$	$*$	0.006	$*$	$*$
High school or less	0.364	0.134	0.304	0.016	0.314	$*$	$*$	$*$
Bachelor's degree and above	0.023	0.039	0.189	0.005	0.038	$*$	$*$	$*$
Presence of child in household	NA	NA	NA	NA	$*$	$*$	$*$	$*$
Home ownership	NA	NA	NA	NA	0.036	$*$	$*$	0.001
Smoke everyday	NA	NA	NA	NA	0.847	0.893	0.067	0.311
Smoke never	NA	NA	NA	NA	$*$	0.031	$*$	$*$
No money to see doctors	NA	NA	NA	NA	$*$	$*$	$*$	0.018
Having medical insurance	NA	NA	NA	NA	0.006	$*$	0.009	0.338
Household income $<$ 20K	0.003	0.025	0.038	0.003	NA	NA	NA	NA
Household income $>$ 100K	0.075	0.057	0.213	$*$	NA	NA	NA	NA
Volunteered	$*$	$*$	$*$	0.438	NA	NA	NA	NA

"$*$" indicates that p-value < 0.001.

Table 2.8: P-values by Using Covariate Set $\boldsymbol{x}.select$

	CPS				BRFSS			
Covariates	y_1	y_2	y_3	y_4	y_1	y_2	y_3	y_4
Intercept	$*$	$*$	$*$	$*$	$*$	$*$	$*$	$*$
Age	$*$	$*$	✗	$*$	$*$	0.044	✗	0.072
Female	$*$	✗	$*$	$*$	$*$	✗	$*$	$*$
White only	0.009	✗	✗	0.003	0.068	0.074	✗	0.011
Black only	$*$	$*$	$*$	✗	0.002	$*$	$*$	✗
Hispanic/Latino	0.118	$*$	$*$	✗	✗	0.001	$*$	0.094
Northeast	$*$	✗	0.013	✗	$*$	✗	0.095	✗
South	0.007	$*$	0.095	✗	0.009	$*$	✗	✗
West	0.097	$*$	✗	0.009	0.016	$*$	✗	0.003
Married	0.020	$*$	$*$	✗	0.147	$*$	$*$	0.022
Working	✗	0.027	0.005	$*$	✗	0.002	$*$	$*$
Retired	0.002	✗	$*$	$*$	$*$	0.008	$*$	$*$
High school or less	✗	0.128	✗	0.013	✗	$*$	0.001	$*$
Bachelor's degree and above	0.005	0.031	0.029	0.003	0.004	$*$	$*$	$*$
Presence of child in household	NA	NA	NA	NA	$*$	$*$	$*$	$*$
Home ownership	NA	NA	NA	NA	0.026	$*$	$*$	0.001
Smoke everyday	NA	NA	NA	NA	✗	✗	0.081	✗
Smoke never	NA	NA	NA	NA	$*$	$*$	$*$	$*$
No money to see doctors	NA	NA	NA	NA	$*$	$*$	$*$	0.010
Having medical insurance	NA	NA	NA	NA	0.004	$*$	0.008	✗
Household income $<$ 20K	0.002	0.027	0.021	0.001	NA	NA	NA	NA
Household income $>$ 100K	0.076	0.048	✗	$*$	NA	NA	NA	NA
Volunteered	$*$	$*$	$*$	✗	NA	NA	NA	NA

"$*$" indicates that p-value < 0.001. "✗" indicates that the covariate is not selected by the backward variable selection algorithm.

chosen sample $\mathcal{S}_B$, the point estimators using covariates $\boldsymbol{x}.large$ and $\boldsymbol{x}.select$ behave quite differently from the estimators using $\boldsymbol{x}.partial$ which excludes some covariates based on the availability from a second probability sample. (iii) The plug-in variance estimators under covariates $\boldsymbol{x}.select$ are almost always smaller than the corresponding variance estimators under covariates $\boldsymbol{x}.large$, showing some efficiency gain by eliminating non-significant factors from the model. So one can similarly conduct the variable selection procedure to the propensity score model to further increase estimation efficiency. (iv) The point estimators do not differ much by using CPS and BRFSS when the same set of covariates $\boldsymbol{x}.partial$ is used. (v) The point estimators using CPS are quite different from the estimators using BRFSS when covariates $\boldsymbol{x}.large$ is considered. Note that covariate sets $\boldsymbol{x}.large$ are different under BRFSS and CPS, and some highly relevant covariates are only available from one particular probability sample. For instance, the covariate *Volunteered*, which has a relatively high correlation to the response variables y_1, y_2 and y_3, is available in CPS but not in BRFSS. Similarly, health related covariates which have strong association with response y_4, are available in BRFSS but not in CPS. The discrepancy between two sets of $\boldsymbol{x}.large$ could possibly explain the different behavior of estimator $\hat{\mu}_{DR2}$ between two probability samples.

Table 2.9: Estimator $\hat{\mu}_{DR2}$ by Different Covariates

Response	$\mathcal{S}_B$	$\hat{\mu}_A$	$\hat{\mu}_{DR2}$		
			$\boldsymbol{x}.partial$	$\boldsymbol{x}.large$	$\boldsymbol{x}.select$
y_1	CPS	0.461	$0.458_{(3.998)}$	$0.401_{(4.265)}$	$0.399_{(3.987)}$
	BRFSS		$0.457_{(3.784)}$	$0.446_{(4.285)}$	$0.446_{(3.918)}$
y_2	CPS	0.210	$0.207_{(2.519)}$	$0.132_{(1.271)}$	$0.132_{(1.227)}$
	BRFSS		$0.202_{(2.329)}$	$0.198_{(2.540)}$	$0.198_{(2.497)}$
y_3	CPS	0.141	$0.136_{(1.790)}$	$0.086_{(0.900)}$	$0.086_{(0.805)}$
	BRFSS		$0.133_{(1.656)}$	$0.120_{(1.635)}$	$0.119_{(1.641)}$
y_4	CPS	5.301	$5.013_{(0.911)}$	$5.114_{(1.156)}$	$5.113_{(0.960)}$
	BRFSS		$4.971_{(0.900)}$	$4.819_{(0.918)}$	$4.820_{(0.883)}$

2.7.3 Comparisons of estimation methods

In the final part of the analysis, we compare estimators of population means obtained by different methods. The covariate set $\boldsymbol{x}.select$ given in the previous section is adopted for both the propensity score model and the prediction model. The point estimators $\hat{\mu}_A$, $\hat{\mu}_{IPW1}$, $\hat{\mu}_{IPW2}$, $\hat{\mu}_{REG}$, $\hat{\mu}_{DR1}$ and $\hat{\mu}_{DR2}$, along with their associated plug-in variance estimators (multiplied by 10^5 for binary variables and by 10^2 for the continuous variable) for the four response variables are reported in Table 2.10. Estimator $\hat{\mu}_{KH}$ is also computed, with its associated variance estimator v_{KH} being displayed in the parentheses. Kim and Haziza (2014)'s method requires the knowledge of the true N. This requirement is met here since the estimated population count $\hat{N}_B$ based on the survey weights was already calibrated to the true N in both CPS and BRFSS datasets. We notice that proposed estimators differ from naive simple sample means for every response variable; and with a given probability sample, proposed estimators for the same response variable are very close to each other, which indicates reasonable fit of models and the relevance of the auxiliary variables. For the estimating efficiency, it can be observed that Hájek-type of estimators are generally more efficient than HT-type estimators. In particular, estimator $\hat{\mu}_{DR2}$ has smaller variance than $\hat{\mu}_{DR1}$, and estimator $\hat{\mu}_{IPW2}$ has smaller variance than $\hat{\mu}_{IPW1}$ for the most of cases considered.

Table 2.10: Estimators of Population Means by Using Different Methods

Response	$\mathcal{S}_B$	$\hat{\mu}_A$	$\hat{\mu}_{IPW1}$	$\hat{\mu}_{IPW2}$	$\hat{\mu}_{REG}$	$\hat{\mu}_{DR1}$	$\hat{\mu}_{DR2}$	$\hat{\mu}_{KH}$
y_1	CPS	0.461	$0.396_{(4.170)}$	$0.400_{(4.078)}$	$0.402_{(4.149)}$	$0.399_{(4.041)}$	$0.399_{(3.987)}$	$0.397_{(4.315)}$
	BRFSS		$0.443_{(4.170)}$	$0.443_{(3.964)}$	$0.447_{(3.952)}$	$0.446_{(4.054)}$	$0.446_{(3.918)}$	$0.446_{(4.073)}$
y_2	CPS	0.210	$0.136_{(1.309)}$	$0.141_{(1.352)}$	$0.134_{(1.388)}$	$0.132_{(1.235)}$	$0.132_{(1.227)}$	$0.132_{(1.457)}$
	BRFSS		$0.193_{(2.686)}$	$0.196_{(2.616)}$	$0.198_{(2.786)}$	$0.198_{(2.532)}$	$0.198_{(2.497)}$	$0.198_{(2.990)}$
y_3	CPS	0.141	$0.088_{(0.829)}$	$0.090_{(0.845)}$	$0.087_{(0.893)}$	$0.086_{(0.808)}$	$0.086_{(0.805)}$	$0.086_{(0.913)}$
	BRFSS		$0.120_{(1.735)}$	$0.121_{(1.718)}$	$0.120_{(1.701)}$	$0.119_{(1.652)}$	$0.119_{(1.641)}$	$0.119_{(1.780)}$
y_4	CPS	5.301	$5.059_{(0.989)}$	$5.050_{(0.965)}$	$5.086_{(0.959)}$	$5.113_{(0.970)}$	$5.113_{(0.960)}$	$5.113_{(0.996)}$
	BRFSS		$4.717_{(0.916)}$	$4.777_{(0.901)}$	$4.807_{(0.984)}$	$4.820_{(0.897)}$	$4.820_{(0.883)}$	$4.821_{(1.008)}$

We have also analyzed other response variables from the PRC non-probability survey

samples, such as *Tended to trust neighbors*, *Expressed opinions at a government level*, *Voted local elections*, *Participated in sports organizations*, and *No money to buy food*. Their results, not reported here to save space, convey the same messages as the results from Tables 2.5-2.10.

2.8 Discussion

The inferential procedures developed in the current chapter focus on the estimation of the finite population mean. Extensions to other finite population parameters such as the distribution function and quantiles are straightforward, since the estimated propensity scores play the same role as the sample inclusion probabilities. This is in line with the classic survey sampling theory where the basic estimation procedures are typically developed for the population mean. The proposed procedures can also be extended to cover parameters defined through estimating functions, similar to the survey weighted estimating equation methods (Godambe and Thompson, 2009) for analytic use of survey data, with weights defined as the inverse of the estimated propensity scores.

Assumptions **A1**–**A3** listed in Section 2.1 are part of the foundation for the estimation procedures presented in the chapter. In particular, Assumption **A1** requires the availability of a complete set of confounding variables. In practice, however, it is often difficult to decide whether the auxiliary variables $\boldsymbol{x}$ contain all the components for characterizing the selection mechanism. One of the general principles for collecting data using any non-probability method is to include essential auxiliary variables such as gender, age and measurements on socioeconomic status, as well as other variables which not only provide tendencies for participation in the non-probability sample but also have the potential to be useful predictors for the response variables. The other extreme scenario, contrary to having limited number of covariates is having a large set of covariates. As we learned from the real data analysis, including too many covariates would add computational complexity and also decrease estimating efficiency. Yang et al. (2020) and Chen et al. (2019) discuss

the covariate selection when auxiliary variables are high dimensional, and their proposed treatments make our research ready to use on modern data types such as big data.

Assumption **A2**, i.e., positivity assumption, is also too important to ignore. The scenario of having zero propensity scores for certain units in the target population requires a careful evaluation of the population represented by the non-probability sample. This is the same issue of the under-coverage problem in probability sampling and the severity of the problem depends on the proportion of the uncovered population units and the discrepancies between the two parts of the population in terms of the response variables. Corrections for biases due to under-coverage problems require additional source of information on the uncovered units. We extend the current general framework and methodology to the scenario with zero propensity scores in Chapter 4.

The procedures developed in this chapter calls for the availability of high quality probability survey samples with relevant auxiliary information. Census data and large scale probability samples collected by statistical agencies can serve as a rich source of information for statistical analysis of non-probability samples. As more and more data are collected by non-probability methods, the traditional survey-centric approach by many statistical agencies needs to evolve to stay relevant and effective for the new data era.

2.9 Technical Details

Asymptotic Framework.

Consider the following asymptotic framework for theoretical development. Suppose that there is a sequence of finite populations $\mathcal{U}_\nu$ of size N_ν, indexed by ν. Associated with each $\mathcal{U}_\nu$ are a non-probability sample $\mathcal{S}_{A,\nu}$ of size $n_{A,\nu}$ and a probability sample $\mathcal{S}_{B,\nu}$ of size $n_{B,\nu}$. The population size $N_\nu \to \infty$ and the sample sizes $n_{A,\nu} \to \infty$ and $n_{B,\nu} \to \infty$ as $\nu \to \infty$. For notational simplicity the index ν is suppressed for the rest of the thesis and the limiting process is represented by $N \to \infty$.

Regularity Conditions.

In regularity conditions **C3**–**C7** below, value $\boldsymbol{\beta}^*$ is defined as the limit of $\hat{\boldsymbol{\beta}}$ under the asymptotic framework, where $\hat{\boldsymbol{\beta}}$ is the estimated model parameter of prediction model $m(\boldsymbol{x}_i,\boldsymbol{\beta})$. The value of $\boldsymbol{\beta}^*$ is the same as the true parameter $\boldsymbol{\beta}_0$ when the regression model is correctly specified but has no practical meanings otherwise.

C1 The population size N and the sample sizes n_A and n_B satisfy $\lim_{N\to\infty} n_A/N = f_A \in (0,1)$ and $\lim_{N\to\infty} n_B/N = f_B \in (0,1)$.

C2 There exist c_1 and c_2 such that $0 < c_1 \leq N\pi_i^A/n_A \leq c_2$ and $0 < c_1 \leq N\pi_i^B/n_B \leq c_2$ for all units i.

C3 The finite population and the sampling design for $\mathcal{S}_B$ satisfy $N^{-1}\sum_{i\in\mathcal{S}_B} d_i^B \boldsymbol{v}_i - N^{-1}\sum_{i=1}^N \boldsymbol{v}_i = O_p(n_B^{-1/2})$ for $\boldsymbol{v}_i = 1, \boldsymbol{x}_i$, y_i and $m(\boldsymbol{x}_i,\boldsymbol{\beta}^*)$.

C4 The finite population and the propensity scores satisfy $N^{-1}\sum_{i=1}^N \|\boldsymbol{x}_i\|^3 = O(1)$, $N^{-1}\sum_{i=1}^N y_i^2 = O(1)$, $N^{-1}\sum_{i=1}^N \{m(\boldsymbol{x}_i,\boldsymbol{\beta}^*)\}^2 = O(1)$, and $N^{-1}\sum_{i=1}^N \pi_i^A(1-\pi_i^A)\boldsymbol{x}_i\boldsymbol{x}_i^\intercal$ is a positive definite matrix.

C5 For each $\boldsymbol{x}$, $\partial m(\boldsymbol{x},\boldsymbol{\beta})/\partial\boldsymbol{\beta}$ is continuous in $\boldsymbol{\beta}$ and $|\partial m(\boldsymbol{x},\boldsymbol{\beta})/\partial\boldsymbol{\beta}| \leq h(\boldsymbol{x},\boldsymbol{\beta})$ for $\boldsymbol{\beta}$ in the neighborhood of $\boldsymbol{\beta}^*$, and $N^{-1}\sum_{i=1}^N h(\boldsymbol{x}_i,\boldsymbol{\beta}^*) = O(1)$.

C6 For each $\boldsymbol{x}$, $\partial^2 m(\boldsymbol{x},\boldsymbol{\beta})/\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^\intercal$ is continuous in $\boldsymbol{\beta}$ and $\max_{j,l} |\partial^2 m(\boldsymbol{x},\boldsymbol{\beta})/\partial\beta_j\partial\beta_l| \leq k(\boldsymbol{x},\boldsymbol{\beta})$ for $\boldsymbol{\beta}$ in the neighborhood of $\boldsymbol{\beta}^*$, and $N^{-1}\sum_{i=1}^N k(\boldsymbol{x}_i,\boldsymbol{\beta}^*) = O(1)$.

C7 Both $N^{-1}\sum_{i=1}^N (R_i/\pi_i^A)\dot{\boldsymbol{v}}_i$ and $N^{-1}\sum_{i\in\mathcal{S}_B} d_i^B \dot{\boldsymbol{v}}_i$ are asymptotically normally distributed for $\dot{\boldsymbol{v}}_i = 1$, y_i, $\pi_i^A\boldsymbol{x}_i$, $m(\boldsymbol{x}_i,\boldsymbol{\beta}^*)$.

Conditions **C1** and **C3** are commonly used for survey samples. Under regularity condition **C1**, we do not need to distinguish among $O_p(n_A^{-1/2})$, $O_p(n_B^{-1/2})$ and $O_p(N^{-1/2})$. Condition **C2** states that the inclusion probabilities for the samples $\mathcal{S}_A$ and $\mathcal{S}_B$ do not differ in terms of order of magnitude from simple random sampling. Condition **C4** is the

typical finite moment conditions and is used for making valid Taylor series expansions. Conditions **C5** and **C6** are the usual smoothness and boundedness conditions (Wu and Sitter, 2001). Condition **C7** is the normality assumption, and is critical for theoretical development in Chapter 3.

Proof of Theorem 2.1.

Let $\boldsymbol{\eta}^\intercal = (\mu, \boldsymbol{\theta}^\intercal)$. The IPW estimator $\hat{\mu} = \hat{\mu}_{IPW1}$ or $\hat{\mu}_{IPW2}$ along with the estimated parameter $\hat{\boldsymbol{\theta}}_{ml}$ for the propensity score model, can be combined as $\hat{\boldsymbol{\eta}}^\intercal = (\hat{\mu}, \hat{\boldsymbol{\theta}}_{ml}^\intercal)$ which is the solution to the combined estimating equation system given by

$$\boldsymbol{\Phi}_n(\boldsymbol{\eta}) = \begin{bmatrix} \frac{1}{N}\sum_{i=1}^{N}\left\{\frac{R_i(y_i-\mu)}{\pi_i(\boldsymbol{\theta})} + \Delta\frac{R_i-\pi_i(\boldsymbol{\theta})}{\pi_i(\boldsymbol{\theta})}\right\} \\ \frac{1}{N}\sum_{i=1}^{N} R_i\boldsymbol{x}_i - \frac{1}{N}\sum_{i\in\mathcal{S}_B} d_i^B\pi_i(\boldsymbol{\theta})\boldsymbol{x}_i \end{bmatrix} = \mathbf{0}. \tag{2.9.21}$$

where $\Delta = \mu$ if $\hat{\mu} = \hat{\mu}_{IPW1}$ and $\Delta = 0$ if $\hat{\mu} = \hat{\mu}_{IPW2}$. This formation is similar to the one used by Lunceford and Davidian (2004). Under the joint randomization of the propensity score model and the sampling design for $\mathcal{S}_B$, we have $E\{\boldsymbol{\Phi}_n(\boldsymbol{\eta})\} = \mathbf{0}$ when $\boldsymbol{\eta}^\intercal = \boldsymbol{\eta}_0^\intercal = (\mu_y, \boldsymbol{\theta}_0^\intercal)$. Consistency of the estimator $\hat{\boldsymbol{\eta}}$ follows similar arguments in Section 3.2 of Tsiatis (2006).

Under regularity conditions **C1**–**C4**, we have $\boldsymbol{\Phi}_n(\hat{\boldsymbol{\eta}}) = \mathbf{0}$ and $\boldsymbol{\Phi}_n(\boldsymbol{\eta}_0) = O_p(n_A^{-1/2})$. By applying the first order Taylor expansion to $\boldsymbol{\Phi}_n(\hat{\boldsymbol{\eta}})$ around $\boldsymbol{\eta}_0$, we further have

$$\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0 = \left\{\boldsymbol{\phi}_n(\boldsymbol{\eta}_0)\right\}^{-1}\boldsymbol{\Phi}_n(\boldsymbol{\eta}_0) + o_p(n_A^{-1/2}) = \left[E\left\{\boldsymbol{\phi}_n(\boldsymbol{\eta}_0)\right\}\right]^{-1}\boldsymbol{\Phi}_n(\boldsymbol{\eta}_0) + o_p(n_A^{-1/2}), \tag{2.9.22}$$

where $\boldsymbol{\phi}_n(\boldsymbol{\eta}_0) = \partial\boldsymbol{\Phi}_n(\boldsymbol{\eta})/\partial\boldsymbol{\eta}|_{\boldsymbol{\eta}=\boldsymbol{\eta}_0}$ and is given by

$$\boldsymbol{\phi}_n(\boldsymbol{\eta}_0) = -\frac{1}{N}\begin{bmatrix} \sum_{i=1}^{N} R_i(1-\Delta_0/\mu_y)/\pi_i^A + \Delta_0/\mu_y & \sum_{i=1}^{N} R_i(1-\pi_i^A)/\pi_i^A(y_i-\mu_y+\Delta_0)\boldsymbol{x}_i^\intercal \\ \mathbf{0} & \sum_{i\in\mathcal{S}_B} d_i^B\pi_i^A(1-\pi_i^A)\boldsymbol{x}_i\boldsymbol{x}_i^\intercal \end{bmatrix},$$

and $\Delta_0 = \mu_y$ if $\hat{\mu} = \hat{\mu}_{IPW1}$ and $\Delta_0 = 0$ if $\hat{\mu} = \hat{\mu}_{IPW2}$. It follows that $\hat{\mu} = \mu_y + O_p(n_A^{-1/2})$ and

$$Var(\hat{\boldsymbol{\eta}}) = \big[E\big\{\boldsymbol{\phi}_n(\boldsymbol{\eta}_0)\big\}\big]^{-1} Var\big\{\boldsymbol{\Phi}_n(\boldsymbol{\eta}_0)\big\}\big[E\big\{\boldsymbol{\phi}_n(\boldsymbol{\eta}_0)\big\}^{\intercal}\big]^{-1} + o(n_A^{-1})\,.$$

It can be shown that

$$\big[E\big\{\boldsymbol{\phi}_n(\boldsymbol{\eta}_0)\big\}\big]^{-1} = \begin{bmatrix} -1 & \Delta_0/\mu_y \mathbf{a}_1^{\intercal} + (1-\Delta_0/\mu_y)\mathbf{a}_2^{\intercal} \\ \mathbf{0} & -\{\frac{1}{N}\sum_{i=1}^{N} \pi_i^A(1-\pi_i^A)\boldsymbol{x}_i\boldsymbol{x}_i^{\intercal}\}^{-1} \end{bmatrix},$$

where the expressions for $\mathbf{a}_1$ and $\mathbf{a}_2$ are given in Theorem 2.1. The other major piece $Var\big\{\boldsymbol{\Phi}_n(\boldsymbol{\eta}_0)\big\}$ can be found by using the decomposition $\boldsymbol{\Phi}_n(\boldsymbol{\eta}_0) = \mathbf{A}_1 + \mathbf{A}_2$ where

$$\mathbf{A}_1 = \frac{1}{N}\sum_{i=1}^{N} \begin{bmatrix} R_i(y_i - \mu_y)/\pi_i^A + \Delta_0(R_i - \pi_i^A)/\pi_i^A \\ R_i\boldsymbol{x}_i - \pi_i^A\boldsymbol{x}_i \end{bmatrix}, \quad \mathbf{A}_2 = \frac{1}{N}\begin{bmatrix} 0 \\ \sum_{i=1}^{N}\pi_i^A\boldsymbol{x}_i - \sum_{i\in S_B} d_i^B\pi_i^A\boldsymbol{x}_i \end{bmatrix}.$$

It follows that $Var\big\{\boldsymbol{\Phi}_n(\boldsymbol{\eta}_0)\big\} = \mathbf{V}_1 + \mathbf{V}_2$, where $\mathbf{V}_1 = Var(\mathbf{A}_1)$ which only involves the propensity score model and $\mathbf{V}_2 = Var(\mathbf{A}_2)$ which only involves the sampling design for $\mathcal{S}_B$. We have

$$\mathbf{V}_1 = \frac{1}{N^2}\sum_{i=1}^{N} \begin{bmatrix} \{(1-\pi_i^A)/\pi_i^A\}(y_i - \mu_y + \Delta_0)^2 & (1-\pi_i^A)(y_i - \mu_y + \Delta_0)\boldsymbol{x}_i^{\intercal} \\ (1-\pi_i^A)(y_i - \mu_y + \Delta_0)\boldsymbol{x}_i & \pi_i^A(1-\pi_i^A)\boldsymbol{x}_i\boldsymbol{x}_i^{\intercal} \end{bmatrix}$$

and

$$\mathbf{V}_2 = \begin{bmatrix} 0 & \mathbf{0}^{\intercal} \\ \mathbf{0} & \mathbf{D} \end{bmatrix},$$

where $\mathbf{D} = N^{-2}V_p\big(\sum_{i\in\mathcal{S}_B} d_i^B\pi_i^A\boldsymbol{x}_i\big)$ is the design-based variance-covariance matrix under the probability sampling design for $\mathcal{S}_B$. The asymptotic variance for the IPW estimator $\hat{\mu}$ is obtained as the first diagonal element of the matrix $\big[E\big\{\boldsymbol{\phi}_n(\boldsymbol{\eta}_0)\big\}\big]^{-1}\big\{\mathbf{V}_1 + \mathbf{V}_2\big\}\big[E\big\{\boldsymbol{\phi}_n(\boldsymbol{\eta}_0)\big\}^{\intercal}\big]^{-1}$ due to (2.9.22).

Proof of Theorem 2.2.

The double robustness property is straightforward from the construction of the estimator. We first show that the estimation of the outcome regression model parameters $\boldsymbol{\beta}$ has no impact on the asymptotic variance of $\hat{\mu}_{DR2}$. We assume that $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^* = O_p(n_A^{-1/2})$ for some fixed $\boldsymbol{\beta}^*$ regardless of the true regression model. Treating $\hat{\mu}_{DR2}$ as a function of $\hat{\boldsymbol{\beta}}$ and making a Taylor expansion around $\boldsymbol{\beta}^*$, then we have the following expression under regularity conditions **C1**–**C6**.

$$\begin{aligned}\hat{\mu}_{DR2} &= \frac{1}{\hat{N}^A}\sum_{i=1}^{N}\frac{R_i\{y_i - m_i(\boldsymbol{\beta}^*)\}}{\pi_i(\hat{\boldsymbol{\theta}}_{ml})} + \frac{1}{\hat{N}^B}\sum_{i\in\mathcal{S}_B} d_i^B m_i(\boldsymbol{\beta}^*) \\ &+ \left\{\frac{1}{\hat{N}^B}\sum_{i\in\mathcal{S}_B} d_i^B \dot{m}_i(\boldsymbol{\beta}^*) - \frac{1}{\hat{N}^A}\sum_{i=1}^{N}\frac{R_i\dot{m}_i(\boldsymbol{\beta}^*)}{\pi_i(\hat{\boldsymbol{\theta}}_{ml})}\right\}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) \\ &+ O_p(n_A^{-1}),\end{aligned} \tag{2.9.23}$$

where $\dot{m}(\boldsymbol{\beta}) = \partial m(\boldsymbol{x},\boldsymbol{\beta})/\partial\boldsymbol{\beta}$. Under regularity conditions **C3** and **C5**, we have

$$\begin{aligned}\frac{1}{\hat{N}^B}\sum_{i\in\mathcal{S}_B} d_i^B \dot{m}_i(\boldsymbol{\beta}^*) - \frac{1}{N}\sum_{i=1}^{N}\dot{m}_i(\boldsymbol{\beta}^*) &= o_p(1), \\ \frac{1}{\hat{N}^A}\sum_{i=1}^{N}\frac{R_i\dot{m}_i(\boldsymbol{\beta}^*)}{\pi_i(\hat{\boldsymbol{\theta}}_{ml})} - \frac{1}{N}\sum_{i=1}^{N}\dot{m}_i(\boldsymbol{\beta}^*) &= o_p(1),\end{aligned}$$

which implies that $(\hat{N}^B)^{-1}\sum_{i\in\mathcal{S}_B} d_i^B\dot{m}_i(\boldsymbol{\beta}^*) - (\hat{N}^A)^{-1}\sum_{i=1}^{N} R_i\dot{m}_i(\boldsymbol{\beta}^*)/\pi_i(\hat{\boldsymbol{\theta}}_{ml}) = o_p(1)$ and

$$\hat{\mu}_{DR2} = \frac{1}{\hat{N}^A}\sum_{i=1}^{N}\frac{R_i\{y_i - m_i(\boldsymbol{\beta}^*)\}}{\pi_i(\hat{\boldsymbol{\theta}}_{ml})} + \frac{1}{\hat{N}^B}\sum_{i\in\mathcal{S}_B} d_i^B m_i(\boldsymbol{\beta}^*) + o_p\big(n_A^{-1/2}\big). \tag{2.9.24}$$

We now derive the asymptotic variance of $\hat{\mu}_{DR2}$ under the propensity score model and the sampling design for $\mathcal{S}_B$. The first part of $\hat{\mu}_{DR2}$ given in (2.9.24) is the IPW estimator $\hat{\mu}_{IPW2}$ given in (2.4.9) with y_i replaced by $y_i - m_i(\boldsymbol{\beta}^*)$. Using the asymptotic expansion

developed in (2.9.22) on $\hat{\mu}_{IPW2}$, we have

$$\frac{1}{\hat{N}^A}\sum_{i=1}^{N}\frac{R_i\{y_i - m_i(\boldsymbol{\beta}^*)\}}{\pi_i(\hat{\boldsymbol{\theta}}_{ml})} = h_N + \frac{1}{N}\sum_{i=1}^{N} R_i\left\{\frac{y_i - m_i(\boldsymbol{\beta}^*) - h_N}{\pi_i^A} - \mathbf{a}_3^\intercal \boldsymbol{x}_i\right\} + \mathbf{a}_3^\intercal \frac{1}{N}\sum_{i\in\mathcal{S}_B} d_i^B \pi_i^A \boldsymbol{x}_i + o_p(n_A^{-1/2}),$$

where $h_N = N^{-1}\sum_{i=1}^{N}\{y_i - m_i(\boldsymbol{\beta}^*)\}$ and

$$\mathbf{a}_3^\intercal = \Big[\sum_{i=1}^{N}(1-\pi_i^A)\{y_i - m_i(\boldsymbol{\beta}^*) - h_N\}\boldsymbol{x}_i^\intercal\Big]\Big\{\sum_{i=1}^{N}\pi_i^A(1-\pi_i^A)\boldsymbol{x}_i\boldsymbol{x}_i^\intercal\Big\}^{-1}.$$

The second part of $\hat{\mu}_{DR2}$ given in (2.9.24) is the Hájek estimator under the probability sampling design for $\mathcal{S}_B$, which has the following expansion

$$\frac{1}{\hat{N}^B}\sum_{i\in\mathcal{S}_B} d_i^B m_i(\boldsymbol{\beta}^*) = \frac{1}{N}\sum_{i=1}^{N} m_i(\boldsymbol{\beta}^*) + \frac{1}{N}\sum_{i\in\mathcal{S}_B} d_i^B\Big\{m_i(\boldsymbol{\beta}^*) - \frac{1}{N}\sum_{i=1}^{N} m_i(\boldsymbol{\beta}^*)\Big\} + O_p(n_B^{-1}).$$

Putting the two parts together leads to

$$\hat{\mu}_{DR2} - \mu_y = \frac{1}{N}\sum_{i=1}^{N} R_i\left\{\frac{y_i - m_i(\boldsymbol{\beta}^*) - h_N}{\pi_i^A} - \mathbf{a}_3^\intercal\boldsymbol{x}_i\right\} + \frac{1}{N}\sum_{i\in\mathcal{S}_B} d_i^B t_i + o_p(n_A^{-1/2}),$$

where $t_i = \pi_i^A \boldsymbol{x}_i^\intercal \mathbf{a}_3 + m_i(\boldsymbol{\beta}^*) - N^{-1}\sum_{i=1}^{N} m(\boldsymbol{\beta}^*)$. It follows that the asymptotic variance of $\hat{\mu}_{DR2}$ is given by V_{DR2} as specified in Theorem 2.2.

Chapter 3

Pseudo-empirical Likelihood Approach to Non-probability Samples

In this chapter, we propose to use the PEL approach (Chen and Sitter, 1999) to estimate the finite population mean. We show in the following sections that PEL approach is comparable with QR approach given in Chapter 2 in several aspects. For example, when the prediction model is not considered, the resulting estimator under the PEL approach is identical to normalized IPW estimator $\hat{\mu}_{IPW2}$; and under the scenario where the prediction model is incorporated into the PEL approach through model-calibration technique, equivalency can be found between the PEL based estimator and DR estimator $\hat{\mu}_{DR2}$. In spite of the similarities between two approaches, we are motivated to use the PEL approach in many scenarios given the following reasons.

First of all, PEL is a unified approach for both point and interval estimation. Not only does the PEL approach result in point estimators which are comparable with $\hat{\mu}_{IPW2}$ and $\hat{\mu}_{DR2}$, but it also provides an alternative way to construct CIs. Specifically, we construct two types of CIs through PEL ratio functions.

Secondly, its non-parametric nature leads to more robust inferential results compared to parametric or semi-parametric approaches. One of our research interest is the finite population proportion, which is a special type of mean when the response variable is binary. However, based on some preliminary studies, we notice that Wald-type CIs given by either IPW or DR method fail to give satisfactory coverage rates when the sample size is small and the true proportion is close to 0 or 1. The data-driven features of the PEL approach can be utilized to accommodate this issue. Our proposed PEL-ratio-based CIs show improved performance over Wald-type CIs based on QR approach under various sample sizes and different values of the true proportion.

Lastly, calibration constraints under PEL approach are powerful tools to utilize auxiliary information. While the QR approach developed in Chapter 2 is more suitable to combine two datasets, adding constraints under PEL approach could flexibly integrate information in broader forms, regardless of the data sources. This property ideally serves the rising trend that relevant data are often available from multiple sources. Moreover, model-calibration technique under PEL approach conveniently allows for multiple working models for the outcome regression, which is an improvement in robustness over previous single specification. This immediately extends the current doubly robust inference to an emerging area, multiply robust inference; see Han (2014) and Zhang et al. (2019) for instance.

3.1 PEL with Non-probability Samples

PEL approach was proposed by Chen and Sitter (1999) to make inferences about a finite population with probability survey samples. As an extension of the empirical likelihood approach (Owen, 1988), PEL approach is known for its superiority in empirical performance to its competitors when the sample size is relatively small.

Recall that $\mathcal{F}_N = \{(\boldsymbol{x}_i, y_i), i \in \mathcal{U}\}$ is the data of the finite population $\mathcal{U}$, where $\boldsymbol{x}_i$ is the associated value of auxiliary variables, and y_i is the associated value of the response

variable. The parameter of interest is the population mean μ_y. Under PEL approach, data $\mathcal{F}_N$ is treated as a random sample from some super population $\mathcal{F}$, and the log-empirical likelihood function based on $\mathcal{F}_N$ is given by,

$$l(\boldsymbol{p}) = \sum_{i=1}^{N} \log p_i,$$

where $\boldsymbol{p} = (p_1, \cdots, p_N)$, and p_i is the point mass at $(\boldsymbol{x}_i, y_i)$, for $i = 1, \cdots, N$.

Instead of working with $l(\boldsymbol{p})$ directly, we consider a probability sample $\mathcal{S}$, which is drawn from population $\mathcal{U}$. Let $\{(\boldsymbol{x}_i, y_i, d_i), i \in \mathcal{S}\}$ be the data of sample $\mathcal{S}$, where d_i are the survey design weights, then the PEL in the sense of Chen and Sitter (1999) is given by

$$\hat{l}(\boldsymbol{p}) = \sum_{i \in \mathcal{S}} d_i \log p_i.$$

Notice the weights d_i ensure $\hat{l}(\boldsymbol{p})$ to be a valid approximation of the population level information $l(\boldsymbol{p})$ given the relation $E\{\hat{l}(\boldsymbol{p})\} = E\{\sum_{i \in \mathcal{S}} d_i \log p_i\} = l(\boldsymbol{p})$, where expectation is taken under the probability sampling for $\mathcal{S}$. So likewise, we need obtain a set of weights for $\mathcal{S}_A$ to construct a non-probability sample based PEL function. We found that the inverse of estimated propensity scores are natural weights for $\mathcal{S}_A$ according to the development in Chapter 2.

Following Wu and Rao (2006), with the given estimated propensity score $\pi_i(\hat{\boldsymbol{\theta}}_{ml})$, the non-probability sample based PEL function is constructed as,

$$\hat{l}^A(\boldsymbol{p}) = n_A \sum_{i \in \mathcal{S}_A} \hat{d}_i^A \log p_i,$$

where $\hat{d}_i^A = (\hat{N}^A)^{-1}\{\pi_i(\hat{\boldsymbol{\theta}}_{ml})\}^{-1}$. Note the normalized weights $\hat{d}_i^A$ rather than naive inverse $1/\pi_i(\hat{\boldsymbol{\theta}}_{ml})$ are considered in the PEL function. This modification simplifies the derivation of the theorems we are about to propose, but does not effect the estimation of p_i. PEL $\hat{l}^A(\boldsymbol{p})$ also contains a scaling term n_A, which let $\hat{l}^A(\boldsymbol{p})$ reduce to the regular log-empirical

likelihood function $\sum_{i\in\mathcal{S}_A}\log p_i$ when weights $\hat{d}_i^A$ are equal for every $i\in\mathcal{S}_A$.

To maximize $\hat{l}^A(\boldsymbol{p})$, we start with the simplest case where the normalization constraint $\sum_{i\in\mathcal{S}_A} p_i = 1$ is the only constraint. Trivially, we observe that the resulting pseudo-empirical maximum likelihood estimator (PEMLE) of p_i is equal to $\hat{d}_i^A$ for $i\in\mathcal{S}_A$, and the PEMLE of μ_y is equivalent to the Hájek-type IPW estimator $\hat{\mu}_{IPW2}=\sum_{i\in\mathcal{S}_A}\hat{d}_i^A y_i$.

3.2 Doubly Robust Inference through PEL

We show in this section that the estimator under PEL approach can also achieve double robustness through the model-calibration technique. Moreover, two methods of constructing CIs are illustrated in Section 3.2.2. One method is based on limiting distributions of the adjusted PEL ratio statistics; and the other method is based on the bootstrap-calibrated PEL ratio statistics.

3.2.1 Model calibration and point estimation

Another piece of information we have not considered yet is the prediction model $E_\xi(y\mid\boldsymbol{x})=m(\boldsymbol{x},\boldsymbol{\beta})$. We further simplify notations by letting $\hat{m}_i=m_i(\hat{\boldsymbol{\beta}})=m(\boldsymbol{x}_i,\hat{\boldsymbol{\beta}})$, where $\hat{\boldsymbol{\beta}}$ is the estimated model parameter (see Section 2.2.1). To utilize fitted value $\hat{m}_i$ of the prediction model under the PEL approach, we propose to use the model-calibration technique which is carried out by maximizing $\hat{l}^A(\boldsymbol{p})$ under the following constraints,

$$\sum_{i\in\mathcal{S}_A} p_i = 1\,,\ \sum_{i\in\mathcal{S}_A} p_i\hat{m}_i=\hat{\bar{m}}^B, \tag{3.2.1}$$

where $\hat{\bar{m}}^B=(\hat{N}^B)^{-1}\sum_{i\in\mathcal{S}_B} d_i^B\hat{m}_i$. Note $\sum_{i\in\mathcal{S}_A} p_i\hat{m}_i=\hat{\bar{m}}^B$ is the so-called model-calibrated constraint, which was proposed by Wu and Sitter (2001) in the probability sample data context. Maximizing $\hat{l}^A(\boldsymbol{p})$ under (3.2.1) leads to model-calibrated PEMLE $\hat{\mu}_{PEL}=\sum_{i\in\mathcal{S}_A}\hat{p}_i y_i$.

By Wu and Sitter (2001), we have $\hat{p}_i = \hat{d}_i^A / \{1 + \hat{\lambda}(\hat{m}_i - \hat{\bar{m}}^B)\}$, where $\hat{\lambda}$ is the solution to

$$\sum_{i \in \mathcal{S}_A} \frac{\hat{d}_i^A (\hat{m}_i - \hat{\bar{m}}^B)}{1 + \lambda(\hat{m}_i - \hat{\bar{m}}^B)} = 0.$$

The use of the model calibration method under the current setup is motivated by two aspects. On the one hand, we learned from the last chapter, that IPW estimators can gain efficiency and double robustness property if the prediction model is properly incorporated in the estimation. On the other hand, model-calibrated constraint is a natural way to incorporate the fitted value of the prediction model; and Wu and Sitter (2001) already showed that under the probability sample context, model-calibrated constraint leads to model-assisted estimators, which share similar properties with DR estimators.

To verify the DR property of the proposed estimator $\hat{\mu}_{PEL}$, we postulate the same assumptions as in Section 2.4.2 for the model parameters, i.e., $\hat{\boldsymbol{\theta}}_{ml} = \boldsymbol{\theta}^* + O_p(n_A^{-1/2})$, $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}^* + O_p(n_A^{-1/2})$ for some fixed $\boldsymbol{\theta}^*$ and $\boldsymbol{\beta}^*$, regardless of the model specification. Moreover, let $m_i^* = m_i(\boldsymbol{\beta}^*) = m(\boldsymbol{x}_i, \boldsymbol{\beta}^*)$, $\bar{m}^* = N^{-1} \sum_{i=1}^N m_i^*$ for simplicity. Related discussions about parameter estimation of misspecified models can be found in White (1982). Asymptotic properties of estimator $\hat{\mu}_{PEL}$ are given in the following theorem.

Theorem 3.1. *Estimator $\hat{\mu}_{PEL}$ is doubly robust in the sense that it is a consistent estimator of μ_y if either the propensity score model or the outcome regression model is correctly specified. Furthermore, under regularity conditions* **C1**–**C6** *specified in Section 2.9, and the correctly specified model for the propensity scores, we have*

$$\hat{\mu}_{PEL} = \hat{\mu}_{IPW2} + (\hat{\bar{m}}^B - \hat{\bar{m}}_{IPW2}) \hat{B}_m + o_p(n_A^{-\frac{1}{2}}), \tag{3.2.2}$$

with $\hat{B}_m = \sum_{i \in \mathcal{S}_A} \hat{d}_i^A (\hat{m}_i - \hat{\bar{m}}^B) y_i / \sum_{i \in \mathcal{S}_A} \hat{d}_i^A (\hat{m}_i - \hat{\bar{m}}^B)^2$, and $\hat{\bar{m}}_{IPW2} = \sum_{i \in \mathcal{S}_A} \hat{d}_i^A \hat{m}_i$.

In addition, the asymptotic variance of $\hat{\mu}_{PEL}$ is given by $Var(\hat{\mu}_{PEL}) = V_{PEL} + o(n_A^{-1})$

under the correctly specified model for the propensity scores, with

$$V_{PEL} = \frac{1}{N^2}\sum_{i=1}^{N}\frac{1-\pi_i^A}{\pi_i^A}(y_i - m_i^* B_m^* - k_N - \pi_i^A \boldsymbol{x}_i^\intercal \mathbf{b}_1)^2 + W_2,$$

where

$$\mathbf{b}_1 = \Big\{\sum_{i=1}^{N}(1-\pi_i^A)(y_i - m_i^* B_m^* - k_N)\boldsymbol{x}_i^\intercal\Big\}\Big\{\sum_{i=1}^{N}\pi_i^A(1-\pi_i^A)\boldsymbol{x}_i\boldsymbol{x}_i^\intercal\Big\}^{-1},$$

$B_m^* = \big\{\sum_{i=1}^N (m_i^* - \bar{m}^*)^2\big\}^{-1}\{\sum_{i=1}^N (m_i^* - \bar{m}^*)y_i\}$, $k_N = N^{-1}\sum_{i=1}^N \big(y_i - m_i^* B_m^*\big)$, *and* $W_2 = N^{-2}V_p(\sum_{i\in\mathcal{S}_B} d_i^B q_i)$ *is a design-based variance with* $q_i = m_i^* B_m^* + \pi_i^A \boldsymbol{x}_i^\intercal \mathbf{b}_1 - \bar{m}^* B_m^*$.

Note that the expansion in (3.2.2) is a bridge to reach many important observations which directly follow Theorem 3.1. (1) Estimator $\hat{\mu}_{PEL}$ has no close form expression, but the non-negligible terms of expansion (3.2.2), which can be viewed as a computable estimator, clearly reveals the double robustness nature of $\hat{\mu}_{PEL}$. (2) It can be easily derived from the expansion in (3.2.2) that variation induced by estimating $\boldsymbol{\beta}_0$ does not have any impact on the asymptotic variance of $\hat{\mu}_{PEL}$. More specifically, the expansion in (3.2.2) can be equivalently written as $\hat{\mu}_{IPW2} + (\bar{m}^{*B} - \bar{m}^*_{IPW2})\tilde{B}_m + o_p(n_A^{-\frac{1}{2}})$, which is a $\hat{\boldsymbol{\beta}}$-free expression with $\bar{m}^{*B} = (\hat{N}^B)^{-1}\sum_{i\in\mathcal{S}_B} d_i^B m_i^*$, $\bar{m}^*_{IPW2} = \sum_{i\in\mathcal{S}_A}\hat{d}_i^A m_i^*$, and $\tilde{B}_m = \sum_{i\in\mathcal{S}_A}\hat{d}_i^A(m_i^* - \bar{m}^{*B})y_i / \sum_{i\in\mathcal{S}_A}\hat{d}_i^A(m_i^* - \bar{m}^{*B})^2$. This is similar to the observation we made on $\hat{\mu}_{DR1}$ and $\tilde{\mu}_{DR1}$, and $\hat{\mu}_{DR2}$ and $\tilde{\mu}_{DR2}$ in Section 2.4.2. (3) The expansion in (3.2.2) also shows the equivalency among PEL approach, calibration weighting approach, and DR estimation under certain conditions. We consider a calibration weighting based estimator $\hat{\mu}_{MC}$, which is computed as $\hat{\mu}_{MC} = \sum_{i\in\mathcal{S}_A} w_i y_i$, where w_i for $i \in \mathcal{S}_A$ is a set of weights such that the distance function $\sum_{i\in\mathcal{S}_A}(w_i - \hat{d}_i^A)^2/\hat{d}_i^A$ achieves the minimum under the normalization constraint and the model-calibrated constraint $\sum_{i\in\mathcal{S}_A} w_i\hat{m}_i = \hat{\bar{m}}^B$. Estimator $\hat{\mu}_{MC}$ has the following explicit expression which is in a similar form to (3.2.2),

$$\hat{\mu}_{MC} = \hat{\mu}_{IPW2} + (\hat{\bar{m}}^B - \hat{\bar{m}}_{IPW2})\hat{B}_{m,MC}$$

where $\hat{B}_{m,MC} = \sum_{i\in\mathcal{S}_A}\hat{d}_i^A(\hat{m}_i - \hat{\bar{m}}_{IPW2})y_i / \sum_{i\in\mathcal{S}_A}\hat{d}_i^A(\hat{m}_i - \hat{\bar{m}}_{IPW2})^2$. We observe that under

the correctly specified propensity score model and prediction model, both $\hat{B}_{m,MC}$ and $\hat{B}_m$ converge to one, and $\hat{\mu}_{PEL} - \hat{\mu}_{MC} = \hat{\mu}_{PEL} - \hat{\mu}_{DR2} = o_p(n_A^{-\frac{1}{2}})$, i.e., all three estimators are asymptotically equivalent.

Based on the variance formula V_{PEL}, a variance estimator for $\hat{\mu}_{PEL}$ is computed as,

$$v_{PEL} = \frac{1}{N^2} \sum_{i \in \mathcal{S}_A} \{1 - \pi_i(\hat{\boldsymbol{\theta}}_{ml})\} \Big\{ \frac{y_i - \hat{m}_i \hat{B}_m - \hat{k}}{\pi_i(\hat{\boldsymbol{\theta}}_{ml})} - \hat{\mathbf{b}}_1^\intercal \boldsymbol{x}_i \Big\}^2 + \hat{W}_2 \,, \tag{3.2.3}$$

where

$$\hat{\mathbf{b}}_1^\intercal = \Big[\sum_{i \in \mathcal{S}_A} \Big\{ 1/\pi_i(\hat{\boldsymbol{\theta}}_{ml}) - 1 \Big\} (y_i - \hat{m}_i \hat{B}_m - \hat{k}) \boldsymbol{x}_i^\intercal \Big] \Big[\sum_{i \in \mathcal{S}_B} d_i^B \pi_i(\hat{\boldsymbol{\theta}}_{ml}) \{1 - \pi_i(\hat{\boldsymbol{\theta}}_{ml})\} \boldsymbol{x}_i \boldsymbol{x}_i^\intercal \Big]^{-1},$$

$\hat{k} = \sum_{i \in \mathcal{S}_A} \hat{d}_i^A (y_i - \hat{m}_i \hat{B}_m)$ is a consistent estimator of k_N, and $\hat{W}_2$ is a design-based variance estimator based on W_2, whose exact form depends on the sampling scheme for $\mathcal{S}_B$. Estimator v_{PEL} is consistent as long as the propensity score model is correctly specified. Note the population size N in the formula can be replaced by its estimator $\hat{N}^A$ or $\hat{N}^B$ if N is unavailable.

3.2.2 PEL-ratio-based confidence intervals

Confidence interval is an important type of statistic which normally presents a range of values where the parameter of interest is likely to lie in. For a given parameter, one can usually construct multiple CIs which enjoy different properties and performance. For example, Chen and Kim (2014), Rao and Wu (2010) and Berger and Torres (2016) provide different methods of constructing CIs for finite population parameters using probability based complex survey data. According to Theorem 3.1, we can construct a $100(1-a)\%$ Wald-type CI for μ_y based on estimator $\hat{\mu}_{PEL}$ and its associated variance estimator v_{PEL},

$$NA_{PEL} : \big[\hat{\mu}_{PEL} - z_{a/2} v_{PEL}^{1/2}, \quad \hat{\mu}_{PEL} + z_{a/2} v_{PEL}^{1/2} \big],$$

where $z_{a/2}$ is the $(1-a/2)$th quantile of the standard normal distribution. Wald-type CIs rely on normal approximation, but the approximation is not accurate when the sample size is small. So it is foreseeable that NA_{PEL} would not be a substantial improvement over Wald-type CIs based on $\hat{\mu}_{IPW2}$ and $\hat{\mu}_{DR2}$ in our interested scenarios. A more natural approach under the current framework is the PEL-ratio-based CIs, and we notice that the adjusted PEL ratio method given by Wu and Rao (2006) can be directly applied.

We first consider a simple scenario where the prediction model is not considered. Let $\tilde{\boldsymbol{p}} = (\tilde{p}_1, \ldots, \tilde{p}_{n^A})$ be the maximizer of $\hat{l}^A(\boldsymbol{p})$ under constraint $\sum_{i\in\mathcal{S}} p_i = 1$, and let $\tilde{\boldsymbol{p}}(\mu) = \big(\tilde{p}_1(\mu), \ldots \tilde{p}_{n^A}(\mu)\big)$ be the maximizer of $\hat{l}^A(\boldsymbol{p})$ under constraints

$$\sum_{i\in\mathcal{S}_A} p_i = 1, \text{and } \sum_{i\in\mathcal{S}_A} p_i y_i = \mu,$$

where μ is some constant. Based on $\tilde{\boldsymbol{p}}$ and $\tilde{\boldsymbol{p}}(\mu)$, we construct the following PEL ratio function about μ,

$$\Lambda_1(\mu) = -2\Big\{\hat{l}^A\big(\tilde{\boldsymbol{p}}(\mu)\big) - \hat{l}^A(\tilde{\boldsymbol{p}})\Big\}.$$

We also consider an adjustment factor

$$s_1 = n_A^{-1}\sum_{i\in\mathcal{S}_A} \hat{d}_i^A (y_i - \hat{\mu}_{IPW2})^2 / v_{IPW2},$$

where v_{IPW2} is the variance estimator for estimator $\hat{\mu}_{IPW2}$, and its expression is available in Section 2.5.1. Note that s_1 is a computable quantity based on the observed data. We have the following result for PEL ratio function $\Lambda_1(\mu)$.

Theorem 3.2. *Under regularity conditions* **C1**–**C4** *and* **C7** *specified in Section 2.9 and assuming the correctly specified model for the propensity scores, we have* $s_1\Lambda_1(\mu_y) \xrightarrow{d} \chi_1^2$, *where* $\xrightarrow{d}$ *denotes asymptotic convergence in distribution, and* χ_1^2 *denotes chi-squared distribution with one degree of freedom.*

Under Theorem 3.2, an approximate $100(1-a)\%$ CI for μ_y is given by $PEL_{1,adj} = \big\{\mu \mid s_1\Lambda_1(\mu) \le \chi_1^2(a)\big\}$, where $\chi_1^2(a)$ is the $(1-a)$th quantile of χ_1^2 distribution.

The model-calibrated constraint can also be incorporated in the process of constructing CIs. Consider two sets of constraint,

$$\sum_{i\in\mathcal{S}_A} p_i = 1 \ , \sum_{i\in\mathcal{S}_A} p_i\hat{m}_i = \hat{\bar{m}}^B, \tag{3.2.4}$$

and

$$\sum_{i\in\mathcal{S}_A} p_i = 1 \ , \sum_{i\in\mathcal{S}_A} p_i\hat{m}_i = \hat{\bar{m}}^B, \ \sum_{i\in\mathcal{S}_A} p_i y_i = \mu, \tag{3.2.5}$$

which lead to PEL ratio function $\Lambda_2(\mu) = -2\Big\{\hat{l}^A\big(\hat{\boldsymbol{p}}(\mu)\big) - \hat{l}^A(\hat{\boldsymbol{p}})\Big\}$, where $\hat{\boldsymbol{p}} = (\hat{p}_1, \ldots \hat{p}_{n^A})$ is the maximizer of $\hat{l}^A(\boldsymbol{p})$ subject to the constraints in (3.2.4), and $\hat{\boldsymbol{p}}(\mu) = \big(\hat{p}_1(\mu), \ldots \hat{p}_{n^A}(\mu)\big)$ is the maximizer of $\hat{l}^A(\boldsymbol{p})$ subject to the constraints in (3.2.5). Wu and Rao (2006) observed that the constraints in (3.2.5) can be rewritten as,

$$\sum_{i\in\mathcal{S}_A} p_i = 1\,, \ \sum_{i\in\mathcal{S}_A} p_i\hat{m}_i = \hat{\bar{m}}^B\,, \ \sum_{i\in\mathcal{S}_A} p_i r_i = 0, \tag{3.2.6}$$

where $r_i = y_i - \mu - (\hat{m}_i - \hat{\bar{m}}^B)\hat{B}_m$. We show in Section 3.6 that this reformatting largely simplifies the proof for Theorem 3.3 below.

Again, we define a computable adjustment factor

$$s_2 = n_A^{-1}\sum_{i\in\mathcal{S}_A} \hat{d}_i^A \hat{r}_i^2 / v_{PEL},$$

where $\hat{r}_i = y_i - \hat{\mu}_{PEL} - (\hat{m}_i - \hat{\bar{m}}^B)\hat{B}_m$, and the expression of v_{PEL} is already given in (3.2.3). Then we obtain the following result for PEL ratio function $\Lambda_2(\mu)$.

Theorem 3.3. *Under regularity conditions* **C1**–**C7** *specified in Section 2.9 and assuming the correctly specified model for the propensity scores, we have* $s_2\Lambda_2(\mu_y) \xrightarrow{d} \chi_1^2$.

Under Theorem 3.3, an approximate $100(1-a)\%$ CI for μ_y is given by $PEL_{2,adj} = \big\{\mu \mid s_2\Lambda_2(\mu) \le \chi_1^2(a)\big\}$.

One of the major hurdles of obtaining adjusted PEL ratio CIs is constructing adjustment factors. Values of adjustment factors require case-by-case computation since they vary with sampling designs, model assumptions, target parameters, etc. However, the derivation of adjustment factors for other parameters may not be so straightforward as for our current interest μ_y. To bypass this complication, we consider a bootstrap-calibrated PEL procedure, which was investigated by Wu and Rao (2010).

We take unadjusted ratio function $\Lambda_1(\mu)$ as an example for illustration. Let $d(a)$ be the $(1-a)$th quantile of the distribution of $\Lambda_1(\mu_y)$. If $d(a)$ is known, then a $100(1-a)\%$ CI for μ_y is given by $\{\mu \mid \Lambda_1(\mu) \leq d(a)\}$. However, the value of $d(a)$ is unavailable and we therefore apply the following bootstrap procedure to approximate $d(a)$.

Step 1 Draw bootstrap sample $\mathcal{S}_A^{(j)}$ from $\{(\boldsymbol{x}_i, y_i), i \in \mathcal{S}_A\}$ and bootstrap sample $\mathcal{S}_B^{(j)}$ from $\{(\boldsymbol{x}_i, d_i^B), i \in \mathcal{S}_B\}$.

Step 2 Replace sample $\mathcal{S}_A$ and $\mathcal{S}_B$ by $\mathcal{S}_A^{(j)}$ and $\mathcal{S}_B^{(j)}$ respectively, and then obtain quantity $\Lambda_1^{(j)}(\hat{\mu}_{IPW2})$ by the same procedure as that for obtaining $\Lambda_1(\mu)$ with $\mu = \hat{\mu}_{IPW2}$.

Step 3 Repeat Step 1 and Step 2 for $j = 1, \cdots, J$ times to obtain $\{\Lambda_1^{(1)}(\hat{\mu}_{IPW2}), \cdots, \Lambda_1^{(J)}(\hat{\mu}_{IPW2})\}$. Then $d(a)$ can be approximated by $\tilde{d}(a)$, which is the $(1-a)$th quantile of $\{\Lambda_1^{(1)}(\hat{\mu}_{IPW2}), \cdots, \Lambda_1^{(J)}(\hat{\mu}_{IPW2})\}$. Finally, a bootstrap-calibrated interval is given by $PEL_{1,bts} = \{\mu \mid \Lambda_1(\mu) \leq \tilde{d}(a)\}$.

Through a similar procedure, we can also construct a bootstrap-calibrated CI based on the unadjusted ratio $\Lambda_2(\mu)$. We denote the resulting interval by $PEL_{2,bts}$.

Remarks. Bootstrap sample $\mathcal{S}_A^{(j)}$ can be taken by with replacement SRS method with sample size n_A. How to draw bootstrap sample $\mathcal{S}_B^{(j)}$ and choose the bootstrap sample size depends on the original sampling design of $\mathcal{S}_B$. For certain designs, such as SRS and single-stage PPS sampling with small sampling fraction, we can apply the same strategy as that for computing bootstrap variance estimator $v_{DR2,bst}$ in Section 2.6. When $\mathcal{S}_B$ comes from more sophisticated sampling designs, one can refer to bootstrap procedures from Antal and Tillé (2011), and Rao and Wu (1988).

3.3 Extension to Other Parameters

The resulting PEMLE $\hat{p}_i$ of p_i for $i \in \mathcal{S}_A$ can be used to construct estimators for other population parameters. In this section, we in particular discuss the estimation of population proportions, distribution function and quantiles.

3.3.1 Estimation of proportions

In survey questionnaires, binary responses such as, yes/no, agree/disagree, satisfied/not satisfied are one of the most commonly used formats to collect information, and collected binary data are used to estimate the proportion of the population who has certain characteristics. Formally, let $y_i = 1$ if individual i has the characteristic of interest, and let $y_i = 0$ otherwise, then the finite population proportion is given by $P = N^{-1} \sum_{i=1}^{N} y_i$.

Our development in Section 3.2 is for general mean estimation, which can be used to estimate proportion P without any modifications. For the specification of outcome regression model, we can consider binary regression models such as a logistic regression model and a probit model. There are several advantages of using PEL approach for the proportion estimation in comparison with QR approach. Point estimators under PEL approach are range-preserving, which means their values will not fall outside of the interval [0,1] when P is the parameter of interest. QR based estimators such as $\hat{\mu}_{DR1}$ and $\hat{\mu}_{DR2}$, however do not always have this property. Moreover, when Wald-type CIs have unsatisfactory performance, PEL-ratio-based CIs have the potential to provide better results. The shape and orientation of PEL-ratio-based CIs are totally determined by observed data, which would increase the robustness of resulting CIs against small sample sizes. Our simulation studies in Section 3.5 mainly focus on the estimation of proportion, where a range of values for P are considered.

3.3.2 Distribution functions and quantiles

Estimating distribution functions and quantiles are important tasks in many survey data analysis. A wide range of powerful tools and indices are defined through distribution functions and quantiles, such as, Lorenz curve, Gini coefficient and Suits index (a measure of tax progressiveness).

There are many similarities between estimating population means and distribution functions. The true finite population distribution function of response variable at some value y is defined by $F_{Y,N}(y) = N^{-1}\sum_{i=1}^{N} I(y_i \leq y)$, which is essentially the mean of the indicator function $I(y_i \leq y)$. So our proposed techniques for estimating population means are still applicable here, but the variable of interest becomes $I(y_i \leq y)$. For example, PEL approach with normalization constraint gives normalized IPW estimator $\hat{F}_{Y,IPW}(y) = \sum_{i\in\mathcal{S}_A} \hat{d}_i^A I(y_i \leq y)$ for $F_{Y,N}(y)$. It is notable that this estimator is a genuine distribution function which can be inverted to obtain quantile estimators immediately.

Some extra work need to be done to achieve doubly robust inferences. First we need postulate a prediction model for $I(y_i \leq y)$ given $\boldsymbol{x}_i$, at some y. Let $G_i(y) = E_\delta\big\{I(y_i \leq y) \mid \boldsymbol{x}_i\big\} = P(y_i \leq y \mid \boldsymbol{x}_i)$, where E_δ indicates the expectation under the prediction model for $I(y_i \leq y)$. We briefly discuss two parametric model approaches, given by Chen and Wu (2002), to obtain $G_i(y)$.

For the first approach, assume a super population model

$$y_i = m(\boldsymbol{x}_i,\boldsymbol{\beta}_0) + \varepsilon_i,$$

where $\boldsymbol{\beta}_0$ is the true model parameter, and ε_i's are independent with $E_\xi(\varepsilon_i) = 0$ and $Var_\xi(\varepsilon_i) = v(\boldsymbol{x}_i)\sigma_0^2$, and $v(\boldsymbol{x}_i)$ is a function of $\boldsymbol{x}_i$ with some known form. Under the assumption that ε_i's are normally distributed, we have

$$G_i(y) = \Phi\Big[\big\{y - m(\boldsymbol{x}_i,\boldsymbol{\beta}_0)\big\}/\big\{v(\boldsymbol{x}_i)^{\frac{1}{2}}\sigma_0\big\}\Big],$$

where $\Phi(\cdot)$ is the cumulative distribution function (cdf) of the standard normal distribution. The fitted value of $G_i(y)$ is given by $\hat{G}_i(y) = \Phi\Big[\big\{y - m(\boldsymbol{x}_i,\hat{\boldsymbol{\beta}})\big\}/\big\{v(\boldsymbol{x}_i)^{\frac{1}{2}}\hat{\sigma}\big\}\Big]$, where $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}$ are estimators of $\boldsymbol{\beta}_0$ and σ_0 respectively.

For the second approach, consider logistic regression model

$$G_i(y) = \frac{\exp(\boldsymbol{x}_i^\intercal \boldsymbol{\kappa}_0)}{1 + \exp(\boldsymbol{x}_i^\intercal \boldsymbol{\kappa}_0)},$$

where $\boldsymbol{\kappa}_0$ is the true model parameter. Compared to the first model, logistic regression model does not rely on the normality, and its parameters are easier to obtain. But notice that different values of y lead to model $G_i(y)$ with different $\boldsymbol{\kappa}_0$; for example, if the interest is two distinct points, say y_1 and y_2, then two estimation procedures are required to obtain model $G_i(y_1)$ and $G_i(y_2)$ independently, and two resulting models have different estimators for $\boldsymbol{\kappa}_0$. Let S_y^* be the set which contains distinct values of y from the sample $\mathcal{S}_A$. If S_y^* has n_y^* elements, then we need obtain a total of n_y^* prediction models to cover every observed y values.

Besides aforementioned parametric models, non-parametric approach is also natural to consider. Cheng and Chu (1996) and Wang and Qin (2010) promoted Nadaraya-Watson kernel estimators for $G_i(y)$. These estimation methods are more robust compared to parametric approach, but require bandwidth selection.

Assume we are interested in the distribution function at $y = y_0$, i.e., $F_{Y,N}(y_0)$. Let $\hat{G}_i(y)$ denote an estimator of $G_i(y)$ which is obtained by one of Chen and Wu (2002)'s parametric approach, then a straightforward DR estimator of $F_{Y,N}(y_0)$ under PEL approach is given by $\hat{F}_{y_0,PEL}(y_0) = \sum_{i\in\mathcal{S}_A} \hat{p}_i I(y_i \le y_0)$, where subscript "$y_0, PEL$" indicates that y_0 is specified and fixed for the prediction model, and $\hat{p}_i$ is the PEMLE under constraints

$$\sum_{i\in\mathcal{S}_A} p_i = 1\,,\ \sum_{i\in\mathcal{S}_A} p_i\hat{G}_i(y_0) = 1/\hat{N}^B \sum_{i\in\mathcal{S}_B} d_i^B \hat{G}_i(y_0).$$

Notice that estimator $\hat{F}_{y_0,PEL}(y)$ is not doubly robust besides at $y = y_0$, e.g., $\hat{F}_{y_0,PEL}(y_1)$

is not doubly robust when $y_1 \neq y_0$. If the interest is the entire range of y, then $\hat{F}_{y,PEL}(y)$ need to be calculated for each $y \in S_y^*$ with different prediction models. But the resulting estimators under this procedure do not lead to a genuine distribution function. In other words, $y_1 \leq y_2$ does not ensure $\hat{F}_{y_1,PEL}(y_1) \leq \hat{F}_{y_2,PEL}(y_2)$ for arbitrary y_1 and y_2.

DR estimator of $F_{Y,N}(y_0)$ in the classic DR form give by (2.4.12) is compute as

$$\hat{F}_{y_0,DR}(y_0) = \sum_{i \in \mathcal{S}_A} \hat{d}_i^A \big\{ I(y_i \leq y_0) - \hat{G}_i(y_0) \big\} + \frac{1}{\hat{N}^B} \sum_{i \in \mathcal{S}_B} d_i^B \hat{G}_i(y_0).$$

Similarly to estimator $\hat{F}_{y_0,PEL}(y)$, estimator $\hat{F}_{y_0,DR}(y)$ is not doubly robust for the entire range of y.

This issue can be partially solved by using multiple model-calibrated constraints under the PEL approach. For example, when y_1 and y_2 are our interested points, we can obtain a single set of $\hat{p}_i$ by using constraints $\sum_{i \in \mathcal{S}_A} p_i \hat{G}_i(y_1) = 1/\hat{N}^B \sum_{i \in \mathcal{S}_B} d_i^B \hat{G}_i(y_1)$ and $\sum_{i \in \mathcal{S}_A} p_i \hat{G}_i(y_2) = 1/\hat{N}^B \sum_{i \in \mathcal{S}_B} d_i^B \hat{G}_i(y_2)$ simultaneously. The resulting estimator, denoted by $\hat{F}_{y_1,y_2,PEL}(y)$, is doubly robust at both points y_1 and y_2. If the entire range of y is the interest, we can pick serval points such as 0.2th, 0.4th, 0.6th and 0.8th quantiles of set S_y^*, to construct multiple constraints. This multiple constraints technique was also discussed in Rueda and Muñoz (2009). Compared to DR estimation above, this approach still enjoys some robustness, and more importantly, generates genuine distribution functions which are desirable for quantile estimation.

Quantiles are obtained by inverting the distribution function. Let ζ be some value which satisfies $0 < \zeta < 1$, and $F(t)$ be an arbitrary cdf. Then the ζth quantile of $F(t)$ is defined as $q_\zeta = F^{-1}(\zeta) = \inf \big\{ t : F(t) \geq \zeta \big\}$. In the current setting of finite population, the target cdf is $F_{Y,N}(y)$, and the parameter of interest is $q_{\zeta,N} = F_{Y,N}^{-1}(\zeta)$.

Since estimators $\hat{F}_{Y,IPW}(y)$, $\hat{F}_{y_0,PEL}(y)$ and $\hat{F}_{y_1,y_2,PEL}(y)$ are genuine distribution functions, they can be inverted directly to obtain quantile estimators. Let $\hat{F}_{Y,N}(y)$ be any of $\hat{F}_{Y,IPW}(y)$, $\hat{F}_{y_0,PEL}(y)$ or $\hat{F}_{y_1,y_2,PEL}(y)$, and let $\hat{q}_{\zeta,N} = \hat{F}_{Y,N}(\zeta)^{-1}$. We assume there is a twice differentiable distribution function $F_Y(y)$, such that $F_{Y,N}(y) \to F_Y(y)$ in distribution

as $N \to \infty$. Then under the correctly specified propensity model and regularity conditions similarly to Chen and Wu (2002), we have the following weak version of Bahadur representation for $\hat{q}_{\zeta,N}$,

$$\hat{q}_{\zeta,N} = q_\zeta + \frac{\zeta - \hat{F}_{Y,N}(q_\zeta)}{f_Y(q_\zeta)} + o_p(n^{-\frac{1}{2}}),$$

where $q_\zeta = F_Y^{-1}(\zeta)$, and $f_Y(y)$ is the density function of $F_Y(y)$. This representation can be justified by similar arguments from Serfling (1980), Chen and Chen (2000) and Chen and Wu (2002). From the expression, the asymptotic normality of $\hat{q}_{\zeta,N}$ can be immediately established through the asymptotic normality of $\hat{F}_{Y,N}(q_\zeta)$. It also reveals that the efficiency of the quantile estimator is determined by the choice of estimators for the distribution function. Usually, estimators $\hat{F}_{y_0,PEL}(y)$ and $\hat{F}_{y_1,y_2,PEL}(y)$ are more efficient than $\hat{F}_{Y,IPW}(y)$, especially when y_0, and one of y_1 or y_2 are chosen close to q_ζ.

3.4 Multiply Robust Inference

Under the PEL framework, doubly robust inference developed in Section 3.2 can be extended to multiply robust inference through model-calibration technique. The notion of multiple robustness was introduced by Han and Wang (2013), which allows for multiple working models for propensity scores and outcome regression, and the resulting estimator is consistent if one of the working models is correctly specified. Relevant work can be found in Han (2014), Chen and Haziza (2017), Zhang et al. (2019), etc.

Let $\mathcal{P}_\xi = \{m^{(j)}(\boldsymbol{x}_i, \boldsymbol{\beta}^{(j)}), j = 1, \cdots J_1\}$ be a set of working models for the outcome regression, where $\boldsymbol{\beta}^{(j)}$ is the corresponding model parameter for jth working model, and J_1 is the total number of working models. To construct model-calibrated constraints, let $\hat{m}_i^{(j)} = m^{(j)}(\boldsymbol{x}_i, \hat{\boldsymbol{\beta}}^{(j)})$, and $\hat{\bar{m}}^{B,(j)} = 1/\hat{N}^B \sum_{i \in \mathcal{S}_B} d_i^B \hat{m}_i^{(j)}$, where $\hat{\boldsymbol{\beta}}^{(j)}$ is the estimator of parameter $\boldsymbol{\beta}^{(j)}$. According to the relation $E_\xi(y_i \mid \boldsymbol{x}_i, R_i = 1) = E_\xi(y_i \mid \boldsymbol{x}_i)$, estimator $\hat{\boldsymbol{\beta}}^{(j)}$ for $j = 1, \cdots J_1$, can still be obtained with data of sample $\mathcal{S}_A$ alone as we obtain estimator

$\hat{\boldsymbol{\beta}}$ for model $m(\boldsymbol{x}_i, \boldsymbol{\beta})$ in Section 2.4.2. Then under regularity conditions, the PEL based estimator, subject to the following $J_1 + 1$ constraints,

$$\sum_{i \in \mathcal{S}_A} p_i = 1 \,, \ \sum_{i \in \mathcal{S}_A} p_i \hat{m}_i^{(j)} = \hat{\bar{m}}^{B,(j)} \ , \text{for } j = 1, \cdots J_1,$$

is consistent if the propensity score model or any model in set $\mathcal{P}_\xi$ is correctly specified,

To adopt multiple working models for propensity scores, it is more natural to consider the regular EL function $l_s(\boldsymbol{p}) = \sum_{i \in \mathcal{S}_A} \log p_i$. See Zhang et al. (2019) for further discussions. Let $\mathcal{P}_q = \{\pi^{(j)}(\boldsymbol{x}_i, \boldsymbol{\theta}^{(j)}), j = 1, \cdots J_2\}$ be a set of working models for the propensity scores, where $\boldsymbol{\theta}^{(j)}$ is the corresponding model parameter for jth working model, and J_2 is the total number of working models. Let $\hat{\pi}_i^{(j)} = \pi^{(j)}(\boldsymbol{x}_i, \hat{\boldsymbol{\theta}}^{(j)})$, and $\hat{\bar{\pi}}^{B,(j)} = 1/\hat{N}^B \sum_{i \in \mathcal{S}_B} d_i^B \hat{\pi}_i^{(j)}$, where $\hat{\boldsymbol{\theta}}^{(j)}$ is the estimator of parameter $\boldsymbol{\theta}^{(j)}$. Note that set $\mathcal{P}_q$ is a class of propensity score models, so the proposed method of estimating model parameter $\boldsymbol{\theta}$ for $\pi(\boldsymbol{x}_i, \boldsymbol{\theta})$ in Section 2.3 can still be used for obtaining $\hat{\boldsymbol{\theta}}^{(j)}$ for $j = 1, \cdots J_2$, but under different model specifications. Then a multiply robust estimator of μ_y is given by $\hat{\mu}_{MR} = \sum_{i \in \mathcal{S}_A} \hat{p}_i y_i$, where $\hat{p}_i$'s maximize $l_s(\boldsymbol{p})$ under the following $J_1 + J_2 + 1$ constraints,

$$\begin{aligned}
&\sum_{i \in \mathcal{S}_A} p_i = 1 \ , \\
&\sum_{i \in \mathcal{S}_A} p_i \hat{m}_i^{(j)} = \hat{\bar{m}}^{B,(j)} \ , \text{for } j = 1, \cdots J_1 \ , \\
&\sum_{i \in \mathcal{S}_A} p_i \hat{\pi}_i^{(j)} = \hat{\bar{\pi}}^{B,(j)} \ , \text{for } j = 1, \cdots J_2.
\end{aligned}$$

Under regularity conditions, the estimator $\hat{\mu}_{MR}$ is consistent if one of the working models in set $\mathcal{P}_\xi$ or $\mathcal{P}_q$ is correctly specified. The multiple robustness property can be proved by using techniques in Han and Wang (2013).

3.5 Simulation Studies

We consider a finite population of size $N = 10{,}000$, with binary response y and auxiliary variable x_1, x_2, and x_3. Each y_i is generated from a Bernoulli distribution with mean u_i, which follows logistic regression model (ξ),

$$\log\left(\frac{u_i}{1-u_i}\right) = \beta_0 + 0.5x_{1i} + 0.5x_{2i} + 0.5x_{3i}\,,$$

where $x_{1i} = z_{1i}$, $x_{2i} = z_{2i} + 0.1x_{1i}$, $x_{3i} = z_{3i} + 0.1x_{2i}$, with $z_{1i} \sim Bernoulli(0.5)$, $z_{2i} \sim Uniform(0{,}1)$, and $z_{3i} \sim Exponential(mean = 0.5)$. The value of parameter β_0 is set such that $N^{-1}\sum_{i=1}^{N} u_i$ equals to the proportion P we are interested in.

The propensity scores π_i^A follow logistic regression model (q),

$$\log\left(\frac{\pi_i^A}{1-\pi_i^A}\right) = \theta_0 + x_{1i} + x_{2i} + x_{3i}\,,$$

where θ_0 is chosen such that $\sum_{i=1}^{N}\pi_i^A = n_A$, with n_A being the target sample size. The non-probability sample $\mathcal{S}_A$ is selected by the Poisson sampling method with the inclusion probabilities specified by π_i^A.

The probability sample $\mathcal{S}_B$, with the target size n_B, is taken by the randomized systematic PPS sampling method with the inclusion probabilities π_i^B proportional to $z_i = c + x_{3i}$. The value of c is chosen to control the variation of the survey weights such that $\max z_i / \min z_i = 20$.

We consider three scenarios of model specification. (i) Both models are correctly specified, denoted by "TT". (ii) The prediction model is misspecified, and the propensity score model is correctly specified, denoted by "FT"; the working model for ξ is chosen as $\log\{u_i/(1-u_i)\} = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}$, with x_{3i} being omitted. (iii) The prediction model is correctly specified, and the propensity score model is misspecified, denoted by "TF"; the working model for q is chosen as $\log\{\pi_i^A/(1-\pi_i^A\} = \theta_0 + \theta_1 x_{1i} + \theta_2 x_{2i}$, with x_{3i} being omitted.

In the first part of simulation studies, we examine performance of point estimators $\hat{\mu}_{IPW2}$ and $\hat{\mu}_{PEL}$, which are obtained through the proposed PEL approach. We also include simulation results of naive estimator $\hat{\mu}_A = n_A^{-1} \sum_{i \in \mathcal{S}_A} y_i$, prediction based estimator $\hat{\mu}_{REG} = \hat{\bar{m}}^B$, and doubly robust estimator $\hat{\mu}_{DR2}$ for the purpose of comparisons. For a given estimator, its performance is evaluated through $\%RB$ and MSE based on $B = 5{,}000$ simulated samples. Rather than focusing on a single specification of parameters, we take different sizes for $\mathcal{S}_A$ and consider a range of values for P. Results for setting $n_A = 100$, $n_B = 100$, and $P = 0.1,\ 0.2,\ 0.5, 0.7$ are reported in Table 3.1; and results for setting $n_A = 500$, $n_B = 100$, and $P = 0.02,\ 0.03, 0.05,\ 0.95$ are reported in Table 3.2. To facilitate reading, MSE have been multiplied by 10^3 and 10^5 in Table 3.1 and Table 3.2 respectively. We have following interesting observations based on the two tables.

(1) Estimators $\hat{\mu}_{IPW2}$ and $\hat{\mu}_{PEL}$ both have small $\%RB$ under the correctly specified q model ("TT" and "FT"), while $\hat{\mu}_{PEL}$ has smaller MSE than $\hat{\mu}_{IPW2}$. (2) Estimators $\hat{\mu}_{IPW2}$ and $\hat{\mu}_{REG}$ are not robust against model misspecification. Estimator $\hat{\mu}_{IPW2}$ fails under scenario "TF", while $\hat{\mu}_{REG}$ collapses under scenario "FT". (3) Estimators $\hat{\mu}_{PEL}$ and $\hat{\mu}_{DR2}$ are robust against model misspecification, and they have comparable performance in terms of the $\%RB$ and MSE under all the scenarios considered. Moreover, the close performance of $\hat{\mu}_{PEL}$ and $\hat{\mu}_{DR2}$ under scenario "TT" further verifies their asymptotic equivalency. (4) When the prediction model ξ is correctly specified ("TT" and "TF"), estimator $\hat{\mu}_{REG}$ generally has the smallest MSE among all the estimators considered. (5) Two tables with different n_A demonstrate similar patterns.

The second part of the simulation study focuses on the CIs. We are mainly interested in the PEL-ratio-based CIs, including $PEL_{1,adj}$, $PEL_{2,adj}$, $PEL_{1,bts}$ and $PEL_{2,bts}$. Their performance is compared with the following normal approximation based CIs,

$$NA_{IPW2} : [\hat{\mu}_{IPW2} - z_{a/2} v_{IPW2}^{1/2},\ \hat{\mu}_{IPW2} + z_{a/2} v_{IPW2}^{1/2}],$$

$$NA_{DR2,plug} : [\hat{\mu}_{DR2} - z_{a/2} v_{DR2,plug}^{1/2},\ \hat{\mu}_{DR2} + z_{a/2} v_{DR2,plug}^{1/2}],$$

Table 3.1: Simulated $\%RB$ and $MSE_{\times 10^3}$ of Estimators of P ($n_A = 100$)

		TT		FT		TF	
P	Estimators	$\%RB$	MSE	$\%RB$	MSE	$\%RB$	MSE
0.1	$\hat{\mu}_A$	55.73	4.42	55.73	4.42	55.73	4.42
	$\hat{\mu}_{IPW2}$	-0.49	1.48	-0.49	1.48	36.43	2.77
	$\hat{\mu}_{PEL}$	1.64	1.46	0.61	1.45	-2.23	1.39
	$\hat{\mu}_{REG}$	0.76	1.22	40.65	3.13	0.76	1.22
	$\hat{\mu}_{DR2}$	1.27	1.45	0.94	1.45	-1.99	1.22
0.2	$\hat{\mu}_A$	39.56	8.05	39.56	8.05	39.56	8.05
	$\hat{\mu}_{IPW2}$	-0.48	2.86	-0.48	2.86	22.87	4.52
	$\hat{\mu}_{PEL}$	0.79	2.75	0.19	2.75	-0.18	2.66
	$\hat{\mu}_{REG}$	-1.57	2.23	23.02	4.38	-1.57	2.23
	$\hat{\mu}_{DR2}$	0.70	2.75	0.40	2.77	-1.31	2.31
0.5	$\hat{\mu}_A$	19.51	11.78	19.51	11.78	19.51	11.78
	$\hat{\mu}_{IPW2}$	-0.34	5.52	-0.34	5.52	11.47	7.06
	$\hat{\mu}_{PEL}$	0.24	5.07	-0.02	5.17	0.34	5.20
	$\hat{\mu}_{REG}$	0.36	4.37	11.32	6.84	0.36	4.37
	$\hat{\mu}_{DR2}$	0.24	5.12	0.02	5.23	0.50	4.45
0.7	$\hat{\mu}_A$	9.70	6.35	9.70	6.35	9.70	6.35
	$\hat{\mu}_{IPW2}$	-0.52	5.16	-0.52	5.16	5.40	4.59
	$\hat{\mu}_{PEL}$	-0.19	4.83	-0.29	4.82	-0.22	5.06
	$\hat{\mu}_{REG}$	-0.50	4.17	5.39	4.44	-0.50	4.17
	$\hat{\mu}_{DR2}$	-0.18	4.87	-0.30	4.89	-0.46	4.29

Table 3.2: Simulated $\%RB$ and $MSE_{\times 10^5}$ of Estimators of P ($n_A = 500$)

		TT		FT		TF	
P	Estimators	$\%RB$	MSE	$\%RB$	MSE	$\%RB$	MSE
0.02	$\hat{\mu}_A$	44.39	1.30	44.39	1.30	44.39	1.30
	$\hat{\mu}_{IPW2}$	-1.07	0.55	-1.07	0.55	31.35	1.00
	$\hat{\mu}_{PEL}$	-0.93	0.55	-0.70	0.54	0.87	0.51
	$\hat{\mu}_{REG}$	3.80	0.48	33.56	1.06	3.80	0.48
	$\hat{\mu}_{DR2}$	-0.50	0.54	-0.74	0.55	2.26	0.49
0.03	$\hat{\mu}_A$	38.20	2.34	38.20	2.34	38.20	2.34
	$\hat{\mu}_{IPW2}$	-1.24	0.89	-1.24	0.89	20.72	1.34
	$\hat{\mu}_{PEL}$	-1.01	0.88	-0.72	0.87	0.30	0.83
	$\hat{\mu}_{REG}$	2.26	0.77	22.32	1.41	2.26	0.77
	$\hat{\mu}_{DR2}$	-0.57	0.87	-0.71	0.87	1.19	0.78
0.05	$\hat{\mu}_A$	41.57	5.63	41.57	5.63	41.57	5.63
	$\hat{\mu}_{IPW2}$	-1.31	1.33	-1.31	1.33	26.17	3.20
	$\hat{\mu}_{PEL}$	-1.30	1.32	-0.78	1.29	-1.27	1.28
	$\hat{\mu}_{REG}$	4.69	1.24	29.11	3.64	4.69	1.24
	$\hat{\mu}_{DR2}$	-0.65	1.29	-0.86	1.30	2.29	1.20
0.95	$\hat{\mu}_A$	1.31	2.16	1.31	2.16	1.31	2.16
	$\hat{\mu}_{IPW2}$	0.00	2.23	0.00	2.23	0.68	1.75
	$\hat{\mu}_{PEL}$	0.03	2.15	0.03	2.15	-0.02	2.29
	$\hat{\mu}_{REG}$	-0.11	2.00	0.65	1.68	-0.11	2.00
	$\hat{\mu}_{DR2}$	0.04	2.15	0.03	2.17	-0.07	2.00

$$NA_{DR2,bst} : [\hat{\mu}_{DR2} - z_{a/2} v_{DR2,bst}^{1/2},\ \ \hat{\mu}_{DR2} + z_{a/2} v_{DR2,bst}^{1/2}]\,,$$

$$NA_{PEL} : [\hat{\mu}_{PEL} - z_{a/2} v_{PEL}^{1/2},\ \ \hat{\mu}_{PEL} + z_{a/2} v_{PEL}^{1/2}]\,;$$

as well as the *bootstrap hybrid* confidence interval (Shao and Tu, 1996) based on $\hat{\mu}_{DR2}$,

$$Bst_{DR2} : [\hat{\mu}_{DR2} - H_{boot}^{-1}(a/2),\ \ \hat{\mu}_{DR2} - H_{boot}^{-1}(1-a/2)]\,,$$

where $H_{boot}^{-1}(a/2)$ and $H_{boot}^{-1}(1-a/2)$ are the $(1-a/2)$th and the $(a/2)$th quantile of the bootstrap distribution of $\hat{\mu}_{DR2}$ based on $J = 1{,}000$ bootstrap samples.

Performance of CIs is evaluated through simulated coverage probability (%CP), lower tail error rate (%L), upper tail error rate (%U), and average length (AL) based on $B =$

5,000 simulated samples. Variance estimation is crucial in computing adjustment factors and constructing normal approximation based CIs, so we also assess variance estimators v_{IPW2}, v_{PEL} and $v_{DR2,plug}$ through their $\%RB$ in comparison with Monte-carlo simulated variances.

Performance of CIs with $n_A = 100$ and $n_A = 500$ are presented in Table 3.3 and Table 3.4, respectively; and for each case, two CIs whose $\%CP$ are closest to the nominal value 95% are underlined. Variance estimators with $n_A = 100$ and $n_A = 500$ are reported in Table 3.5. We have following key observations based on these three tables.

(1) When the q model is correctly specified ("TT" and "FT"), bootstrap-calibrated PEL ratio CIs, including $PEL_{1,bts}$ and $PEL_{2,bts}$, generally have coverage rates closer to 95% than other CIs reported. This advantage is especially notable when the true proportion P is relatively closer to zero or one. (2) When the q model is correctly specified ("TT" and "FT"), incorporating information of the prediction model incurs shorter AL in general. Specifically, $PEL_{2,adj}$ has shorter AL than $PEL_{1,adj}$, $PEL_{2,bst}$ has shorter AL than $PEL_{1,bst}$, and $NA_{DR2,plug}$ has shorter AL than NA_{IPW2} for most cases. (3)When the q model is correctly specified ("TT" and "FT"), the performance of CIs deteriorates when the true proportion moves closer to boundary values no matter which approaches are taken. (4) We observe that $PEL_{2,bts}$, $NA_{DR2,bst}$ and Bst_{DR2} have some robustness against misspecification of q model ("TF") while other CIs do not. This result indeed can be predicted since other CIs either involve biased point estimators or biased variance estimators. In the meanwhile, $PEL_{2,bts}$ generally outperforms $NA_{2,bts}$ in coverage rates; and Bst_{DR2}, even its performance is relatively insensitive to model misspecification, suffers the most severe under-coverage issue for every case. (5) When $n_A = 500$ and $P = 0.02, 0.03, 0.05$, we observe that NA_{IPW2}, $NA_{DR2,plug}$ and NA_{PEL} have better performance under scenario "TF" than under scenarios "FT" and "TT". This is very counter-intuitive since NA_{IPW2}, $NA_{DR2,plug}$ and NA_{PEL} are supposed to fail in theory when the q model is misspecified. This mystery can be easily unrevealed from Table 3.5. When $P = 0.02, 0.03, 0.05$, variance estimators for $\hat{\mu}_{IPW2}$, $\hat{\mu}_{DR2}$ and $\hat{\mu}_{PEL}$ have large positive $\%RB$. Large variance estimators lead to wide CIs which coincidentally compensate the under-coverage issue. Once biases of variance

Table 3.3: 95% CIs for P Obtained by Different Approaches ($n_A = 100$)

P	Scenarios		$PEL_{1,adj}$	$PEL_{2,adj}$	$PEL_{1,bts}$	$PEL_{2,bts}$	NA_{IPW2}	$NA_{DR2,plug}$	$NA_{DR2,bts}$	NA_{PEL}	Bst_{DR2}
0.1	TT	$\%CP$	91.83	91.15	93.20	93.40	88.55	88.38	90.18	88.48	86.78
		$\%L$	1.52	1.65	0.98	0.80	0.50	0.57	0.30	0.62	0.22
		$\%U$	6.65	7.20	5.83	5.80	10.95	11.05	9.53	10.90	13.00
		AL	0.1407	0.1388	0.1490	0.1487	0.1414	0.1380	0.1475	0.1375	0.1454
	FT	$\%CP$	91.83	91.57	93.10	92.83	88.55	88.92	91.17	88.62	88.33
		$\%L$	1.52	1.50	0.98	0.98	0.50	0.57	0.27	0.57	0.25
		$\%U$	6.65	6.93	5.92	6.20	10.95	10.50	8.55	10.80	11.43
		AL	0.1407	0.1402	0.1489	0.1467	0.1414	0.1397	0.1513	0.1382	0.1497
	TF	$\%CP$	83.83	87.78	86.22	93.47	90.83	92.12	90.88	90.28	86.58
		$\%L$	15.82	2.25	13.45	1.03	8.33	0.80	0.48	1.12	0.32
		$\%U$	0.35	9.98	0.32	5.50	0.85	7.07	8.65	8.60	13.10
		AL	0.1488	0.1226	0.1554	0.1533	0.1497	0.1378	0.1376	0.1360	0.1363
0.2	TT	$\%CP$	93.25	93.08	94.85	94.60	91.75	91.40	92.50	91.22	90.48
		$\%L$	1.82	1.77	1.12	1.00	0.90	0.92	0.70	1.03	0.52
		$\%U$	4.92	5.15	4.03	4.40	7.35	7.67	6.80	7.75	9.00
		AL	0.2011	0.1976	0.2160	0.2112	0.2041	0.1984	0.2100	0.1971	0.2086
	FT	$\%CP$	93.25	92.85	94.85	93.85	91.75	91.40	92.85	91.15	90.83
		$\%L$	1.82	1.70	1.12	1.18	0.90	0.95	0.68	1.07	0.65
		$\%U$	4.92	5.45	4.03	4.98	7.35	7.65	6.48	7.78	8.53
		AL	0.2011	0.1984	0.2159	0.2089	0.2041	0.1996	0.2126	0.1980	0.2113
	TF	$\%CP$	85.05	90.50	87.67	94.85	88.95	92.75	93.08	91.20	91.30
		$\%L$	14.42	2.83	11.85	1.10	10.17	1.40	0.62	2.35	0.45
		$\%U$	0.52	6.68	0.48	4.05	0.88	5.85	6.30	6.45	8.25
		AL	0.1934	0.1774	0.2030	0.2143	0.1956	0.1859	0.1933	0.1843	0.1923
0.5	TT	$\%CP$	94.60	94.05	96.08	95.35	93.75	93.30	94.50	93.23	93.60
		$\%L$	2.85	2.57	2.05	1.92	3.23	2.83	2.27	3.02	2.75
		$\%U$	2.55	3.38	1.88	2.73	3.02	3.88	3.23	3.75	3.65
		AL	0.2783	0.2707	0.3018	0.2929	0.2845	0.2751	0.2911	0.2716	0.2912
	FT	$\%CP$	94.60	93.83	96.08	95.10	93.75	93.20	94.58	93.05	93.35
		$\%L$	2.85	2.70	2.05	2.10	3.23	3.15	2.43	3.23	2.95
		$\%U$	2.55	3.48	1.88	2.80	3.02	3.65	3.00	3.72	3.70
		AL	0.2783	0.2715	0.3018	0.2898	0.2845	0.2768	0.2934	0.2734	0.2938
	TF	$\%CP$	83.90	91.38	86.50	96.10	82.95	91.00	94.70	88.20	93.83
		$\%L$	15.95	3.92	13.43	1.92	16.85	4.85	2.57	5.90	2.97
		$\%U$	0.15	4.70	0.08	1.98	0.20	4.15	2.73	5.90	3.20
		AL	0.2370	0.2456	0.2495	0.3004	0.2406	0.2317	0.2689	0.2295	0.2686
0.7	TT	$\%CP$	93.97	92.92	95.45	95.08	92.95	92.25	93.60	91.88	92.03
		$\%L$	3.55	3.88	2.75	2.75	5.12	4.92	4.32	5.20	5.70
		$\%U$	2.48	3.20	1.80	2.17	1.92	2.83	2.08	2.93	2.27
		AL	0.2646	0.2586	0.2879	0.2852	0.2707	0.2648	0.2815	0.2605	0.2817
	FT	$\%CP$	93.97	93.33	95.30	94.73	92.95	92.45	93.65	92.00	92.33
		$\%L$	3.55	3.85	2.85	3.00	5.12	5.15	4.45	5.38	5.67
		$\%U$	2.48	2.83	1.85	2.27	1.92	2.40	1.90	2.62	2.00
		AL	0.2646	0.2587	0.2860	0.2780	0.2707	0.2649	0.2807	0.2611	0.2811
	TF	$\%CP$	88.02	90.62	89.88	95.97	84.50	87.95	94.08	84.45	92.27
		$\%L$	11.58	4.67	9.78	2.17	15.15	6.95	3.65	8.70	5.30
		$\%U$	0.40	4.70	0.35	1.85	0.35	5.10	2.27	6.85	2.43
		AL	0.2135	0.2364	0.2263	0.2925	0.2165	0.2119	0.2614	0.2097	0.2613

Table 3.4: 95% CIs for P Obtained by Different Approaches ($n_A = 500$)

P	Scenarios		$PEL_{1,adj}$	$PEL_{2,adj}$	$PEL_{1,bts}$	$PEL_{2,bts}$	NA_{IPW2}	$NA_{DR2,plug}$	$NA_{DR2,bts}$	NA_{PEL}	Bst_{DR2}
0.02	TT	$\%CP$	91.47	91.17	92.77	93.17	88.78	88.78	89.85	88.55	87.10
		$\%L$	1.43	1.52	1.07	1.10	0.35	0.35	0.22	0.38	0.12
		$\%U$	7.10	7.30	6.15	5.73	10.88	10.88	9.93	11.07	12.78
		AL	0.0282	0.0280	0.0296	0.0293	0.0278	0.0275	0.0290	0.0275	0.0286
	FT	$\%CP$	91.47	91.83	92.65	92.80	88.78	88.90	90.33	88.83	87.50
		$\%L$	1.43	1.38	1.07	1.12	0.35	0.35	0.25	0.38	0.15
		$\%U$	7.10	6.80	6.28	6.08	10.88	10.75	9.43	10.80	12.35
		AL	0.0282	0.0282	0.0295	0.0290	0.0278	0.0277	0.0294	0.0276	0.0290
	TF	$\%CP$	87.75	91.05	89.68	94.00	94.40	94.80	92.42	93.80	90.50
		$\%L$	11.55	2.25	9.93	1.23	4.30	0.52	0.35	0.68	0.18
		$\%U$	0.70	6.70	0.40	4.78	1.30	4.67	7.22	5.53	9.32
		AL	0.0309	0.0264	0.0323	0.0302	0.0307	0.0304	0.0282	0.0303	0.0279
0.03	TT	$\%CP$	93.33	93.00	94.53	94.45	91.17	91.03	92.30	90.55	89.72
		$\%L$	1.65	1.62	1.18	1.20	0.52	0.52	0.38	0.52	0.22
		$\%U$	5.03	5.38	4.30	4.35	8.30	8.45	7.32	8.92	10.05
		AL	0.0364	0.0361	0.0384	0.0379	0.0362	0.0357	0.0375	0.0356	0.0372
	FT	$\%CP$	93.33	93.27	94.50	94.17	91.17	90.83	92.42	90.80	90.18
		$\%L$	1.65	1.70	1.18	1.27	0.52	0.52	0.32	0.57	0.15
		$\%U$	5.03	5.03	4.32	4.55	8.30	8.65	7.25	8.62	9.68
		AL	0.0364	0.0364	0.0384	0.0374	0.0362	0.0359	0.0378	0.0358	0.0375
	TF	$\%CP$	89.98	91.65	91.67	94.85	93.58	94.45	93.27	93.47	91.88
		$\%L$	9.35	2.43	7.80	1.47	5.08	0.80	0.62	0.98	0.25
		$\%U$	0.68	5.92	0.52	3.67	1.35	4.75	6.10	5.55	7.88
		AL	0.0372	0.0333	0.0390	0.0378	0.0371	0.0366	0.0358	0.0365	0.0356
0.05	TT	$\%CP$	94.03	93.65	95.50	94.90	92.42	92.45	93.40	91.70	92.35
		$\%L$	1.80	1.65	1.27	1.25	0.80	0.82	0.65	0.90	0.32
		$\%U$	4.17	4.70	3.23	3.85	6.78	6.73	5.95	7.40	7.32
		AL	0.0454	0.0447	0.0482	0.0471	0.0453	0.0444	0.0465	0.0443	0.0463
	FT	$\%CP$	94.03	94.05	95.50	94.75	92.42	92.53	93.53	92.53	92.50
		$\%L$	1.80	1.75	1.27	1.27	0.80	0.75	0.60	0.80	0.32
		$\%U$	4.17	4.20	3.23	3.98	6.78	6.73	5.88	6.68	7.17
		AL	0.0454	0.0452	0.0482	0.0464	0.0453	0.0447	0.0471	0.0446	0.0469
	TF	$\%CP$	79.97	92.10	82.27	95.17	86.17	96.47	95.08	94.53	94.33
		$\%L$	19.88	2.23	17.57	1.30	13.55	1.12	1.00	1.12	0.80
		$\%U$	0.15	5.67	0.15	3.52	0.27	2.40	3.92	4.35	4.88
		AL	0.0476	0.0412	0.0498	0.0471	0.0475	0.0464	0.0444	0.0461	0.0443
0.95	TT	$\%CP$	93.03	92.38	94.67	94.23	90.70	90.53	91.50	90.53	89.53
		$\%L$	5.53	5.73	4.28	4.47	8.88	8.70	8.12	8.67	10.35
		$\%U$	1.45	1.90	1.05	1.30	0.43	0.78	0.38	0.80	0.12
		AL	0.0568	0.0564	0.0643	0.0626	0.0565	0.0565	0.0593	0.0561	0.0591
	FT	$\%CP$	93.03	92.77	94.20	93.58	90.70	90.75	91.85	90.65	89.88
		$\%L$	5.53	5.55	4.75	5.00	8.88	8.72	7.83	8.70	10.03
		$\%U$	1.45	1.68	1.05	1.43	0.43	0.52	0.32	0.65	0.10
		AL	0.0568	0.0563	0.0601	0.0580	0.0565	0.0562	0.0590	0.0560	0.0588
	TF	$\%CP$	88.92	92.62	91.65	95.40	84.30	87.42	93.40	84.88	91.55
		$\%L$	10.78	5.17	8.12	3.65	15.60	9.28	6.12	10.70	8.28
		$\%U$	0.30	2.20	0.22	0.95	0.10	3.30	0.48	4.42	0.18
		AL	0.0448	0.0549	0.0508	0.0657	0.0446	0.0452	0.0566	0.0450	0.0565

Table 3.5: Simulated $\%RB$ of Variance Estimators

$n_A = 100$					$n_A = 500$				
P	Estimators	TT	FT	TF	P	Estimators	TT	FT	TF
0.1	v_{IPW2}	-0.03	-0.03	7.07	0.02	v_{IPW2}	4.62	4.62	12.47
	v_{PEL}	-3.17	-2.15	-7.47		v_{PEL}	2.93	4.92	20.54
	$v_{DR2,plug}$	-1.25	0.02	7.54		$v_{DR2,plug}$	4.70	5.60	34.10
0.2	v_{IPW2}	1.35	1.35	3.78	0.03	v_{IPW2}	5.62	5.62	7.15
	v_{PEL}	-1.09	-0.37	-13.97		v_{PEL}	4.06	5.93	9.54
	$v_{DR2,plug}$	1.07	1.28	0.83		$v_{DR2,plug}$	5.95	6.51	18.16
0.5	v_{IPW2}	-2.05	-2.05	1.08	0.05	v_{IPW2}	4.66	4.66	9.68
	v_{PEL}	-2.91	-3.54	-33.06		v_{PEL}	3.03	5.37	10.94
	$v_{DR2,plug}$	-1.21	-2.10	-20.14		$v_{DR2,plug}$	5.02	5.24	22.90
0.7	v_{IPW2}	-2.70	-2.70	-0.22	0.95	v_{IPW2}	1.76	1.76	2.40
	v_{PEL}	-4.83	-4.08	-42.02		v_{PEL}	1.35	1.90	-40.76
	$v_{DR2,plug}$	-2.68	-2.62	-29.92		$v_{DR2,plug}$	1.73	1.93	-30.37

estimators decrease at $P = 0.95$, then confidence intervals NA_{IPW2}, $NA_{DR2,plug}$ and NA_{PEL} no longer hold valid under scenario "TF". (6) PEL approach in general tends to provide more balanced tail error rates for resulting CIs, compared to the normal approximation based approaches.

3.6 Technical Details

Proof of Theorem 3.1.

(1) Justification of Double Robustness.

Define $\hat{u}_i = \hat{m}_i - \hat{\bar{m}}^B$ for notational simplicity. We rewrite $\hat{\mu}_{PEL}$ as $\hat{\mu}_{PEL} = \sum_{i\in\mathcal{S}_A} \hat{p}_i(y_i - \hat{m}_i) + \sum_{i\in\mathcal{S}_A} \hat{p}_i\hat{m}_i$. Consider the case where the prediction model is correctly specified. By using the first order Taylor expansion, we have

$$\begin{aligned}\sum_{i\in\mathcal{S}_A} \hat{p}_i(y_i - \hat{m}_i) &= \sum_{i=1}^{N} \frac{R_i\hat{d}_i^A}{1+\hat{\lambda}\hat{u}_i}(y_i - \hat{m}_i) \\ &= \frac{1}{N}\sum_{i=1}^{N} \frac{R_i(y_i - m_i^*)}{\pi_i(\boldsymbol{\theta}^*)\{1+\lambda^*(m_i^* - \bar{m}^*)\}} \Big/ \frac{1}{N}\sum_{i=1}^{N} \frac{R_i}{\pi_i(\boldsymbol{\theta}^*)} + o_p(1),\end{aligned}$$

which immediately leads to

$$\sum_{i\in\mathcal{S}_A} \hat{p}_i(y_i - \hat{m}_i) \xrightarrow{p} \frac{1}{N}\sum_{i=1}^{N} \frac{E_q(R_i)E_\xi(y_i - m_i^*)}{\pi_i(\boldsymbol{\theta}^*)\{1+\lambda^*(m_i^* - \bar{m}^*)\}} \Big/ \frac{1}{N}\sum_{i=1}^{N} E_q\left\{\frac{R_i}{\pi_i(\boldsymbol{\theta}^*)}\right\},$$

where λ^* is the limiting point of $\hat{\lambda}$. Since $\boldsymbol{\beta}^* = \boldsymbol{\beta}_0$ under the true prediction model, then we have $E_\xi(y_i - m_i^*) = 0$, which gives $\sum_{i\in\mathcal{S}_A} \hat{p}_i(y_i - \hat{m}_i) = o_p(1)$. In addition, we have $\sum_{i\in\mathcal{S}_A} \hat{p}_i\hat{m}_i = \hat{\bar{m}}^B$ by the model-calibrated constraint, and $\hat{\bar{m}}^B - \mu_y = o_p(1)$ under regularity conditions **C1**–**C3**. Hence $\hat{\mu}_{PEL} = \mu_y + o_p(1)$.

When the propensity score model is correctly specified, we have $\hat{\lambda}$ converges to zero. Using the first order Taylor expansion, we get

$$\begin{aligned}\hat{\mu}_{PEL} &= \sum_{i=1}^{N} \frac{R_i\hat{d}_i^A}{1+\hat{\lambda}\hat{u}_i} y_i \\ &= \frac{1}{N}\sum_{i=1}^{N} \frac{R_i}{\pi_i^A} y_i \Big/ \frac{1}{N}\sum_{i=1}^{N} \frac{R_i}{\pi_i^A} + o_p(1).\end{aligned}$$

Note that

$$\frac{1}{N}\sum_{i=1}^{N}\frac{R_i}{\pi_i^A}y_i\Big/\frac{1}{N}\sum_{i=1}^{N}\frac{R_i}{\pi_i^A}\xrightarrow{p}\frac{1}{N}\sum_{i=1}^{N}E_q\left(\frac{R_i}{\pi_i^A}\right)y_i\Big/\frac{1}{N}\sum_{i=1}^{N}E_q\left(\frac{R_i}{\pi_i^A}\right)=\mu_y,$$

where $\xrightarrow{p}$ indicates convergence in probability. Hence $\hat{\mu}_{PEL}\xrightarrow{p}\mu_y$.

(2) Asymptotic Expansion of $\hat{\mu}_{PEL}$.

Define $\hat{u}_i=\hat{m}_i-\hat{\bar{m}}^B$. First of all, we show following two statements,

(i) $\max_{i\in\mathcal{S}_A}\left\{|\hat{u}_i|\right\}=o_p(n_A^{\frac{1}{2}})$,

(ii) $\sum_{i\in\mathcal{S}_A}\hat{d}_i^A\hat{u}_i/\sum_{i\in\mathcal{S}_A}\hat{d}_i^A\hat{u}_i^2=O_p(n_A^{-\frac{1}{2}})$,

which together implies $\hat{\lambda}=\sum_{i\in\mathcal{S}_A}\hat{d}_i^A\hat{u}_i/\sum_{i\in\mathcal{S}_A}\hat{d}_i^A\hat{u}_i^2+o_p(n_A^{-\frac{1}{2}})$ by Wu and Sitter (2001).

To prove Statement (i), we first obtain the second-order Taylor expansion of $\hat{u}_i$ around $\boldsymbol{\beta}^*$,

$$\begin{aligned}\hat{u}_i &= \hat{m}_i-\hat{\bar{m}}^B\\ &= \hat{m}_i-N^{-1}\sum_{i=1}^{N}\hat{m}_i+o_p(1)\\ &= (m_i^*-\bar{m}^*)+\left\{\partial(m_i-\bar{m})/\partial\boldsymbol{\beta}^{\intercal}|_{\boldsymbol{\beta}=\boldsymbol{\beta}_n}\right\}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)+o_p(1),\end{aligned}$$

where $\boldsymbol{\beta}_n$ is between $\boldsymbol{\beta}^*$ and $\hat{\boldsymbol{\beta}}$, $m_i=m(\boldsymbol{x}_i,\boldsymbol{\beta})$, and $\bar{m}=N^{-1}\sum_{i=1}^{N}m_i$. Obviously, Statement (i) follows directly if $\max_{i\in\mathcal{S}_A}\{|m_i^*-\bar{m}^*|\}=o_p(n_A^{\frac{1}{2}})$ and $\max_{i\in\mathcal{S}_A}\left\{|\partial(m_i-\bar{m})/\partial\boldsymbol{\beta}^{\intercal}|_{\boldsymbol{\beta}=\boldsymbol{\beta}_n}|\right\}=o_p(n_A)$. Observe that

$$\max_{i\in\mathcal{S}_A}\{|m_i^*-\bar{m}^*|\}\leq\max_{i\in\mathcal{S}_A}\{|m_i^*|\}+|\bar{m}^*|,$$

where $|\bar{m}^*|=O(1)$, and the order of $\max_{i\in\mathcal{S}_A}\{|m_i^*|\}$ can not be larger than $o(n_A^{\frac{1}{2}})$ under assumption $N^{-1}\sum_{i=1}^{N}m_i^{*2}=O(1)$, which is stated in regularity condition **C4**. Similarly, it

can be shown that $\max_{i\in\mathcal{S}_A}\left\{|\partial(m_i-\bar{m})/\partial\boldsymbol{\beta}^{\intercal}|_{\boldsymbol{\beta}=\boldsymbol{\beta}_n}|\right\}=o_p(n_A)$ under regularity condition **C5**. We thus have Statement (i) verified.

To find the order of $\sum_{i\in\mathcal{S}_A}\hat{d}_i^A\hat{u}_i/\sum_{i\in\mathcal{S}_A}\hat{d}_i^A\hat{u}_i^2$, we first obtain the first order Taylor expansion of $\sum_{i\in\mathcal{S}_A}\hat{d}_i^A\hat{u}_i$ around $\boldsymbol{\beta}^*$,

$$\sum_{i\in\mathcal{S}_A}\hat{d}_i^A\hat{u}_i=(\sum_{i\in\mathcal{S}_A}\hat{d}_i^Am_i^*-\bar{m}^{*B})+(\sum_{i\in\mathcal{S}_A}\hat{d}_i^A\dot{m}_i^*-\dot{\bar{m}}^{*B})(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)+o_p(n_A^{-\frac{1}{2}}),$$

where $\bar{m}^{*B}=(\hat{N}^B)^{-1}\sum_{i\in\mathcal{S}_B}d_i^Bm_i^*$, $\dot{m}_i^*=\partial m(\boldsymbol{x}_i,\boldsymbol{\beta}^*)/\partial\boldsymbol{\beta}^{\intercal}$ and $\dot{\bar{m}}^{*B}=(\hat{N}^B)^{-1}\sum_{i\in\mathcal{S}_B}d_i^B\dot{m}_i^*$. The first component of this expansion can be rewritten as $(\sum_{i\in\mathcal{S}_A}\hat{d}_i^Am_i^*-\bar{m}^*)+(\bar{m}^*-\bar{m}^{*B})$, where $(\sum_{i\in\mathcal{S}_A}\hat{d}_i^Am_i^*-\bar{m}^*)$ has order of $O_p(n_A^{-\frac{1}{2}})$ by the similar augment for Theorem 2.1, and $(\bar{m}^*-\bar{m}^{*B})$ also has order of $O_p(n_A^{-\frac{1}{2}})$ under regularity condition **C3**. Hence, the first component of this expansion has order of $O_p(n_A^{-\frac{1}{2}})$. The second component of the expansion has order of $o_p(n_A^{-\frac{1}{2}})$. Then $\sum_{i\in\mathcal{S}_A}\hat{d}_i^A\hat{u}_i=O_p(n_A^{-\frac{1}{2}})$ follows directly.

It can be easily shown that the denominator $\sum_{i\in\mathcal{S}_A}\hat{d}_i^A\hat{u}_i^2$ has order of $O_p(1)$, and does not converge to zero. Thus Statement (ii) is verified.

Under Statements (i) and (ii), the asymptotic expression for $\hat{\mu}_{PEL}$ can be derived. First we get $\hat{\mu}_{PEL}=\sum_{i\in\mathcal{S}_A}\hat{p}_iy_i=\hat{d}_i^A(1-\hat{u}_i\hat{\lambda})y_i+o_p(n_A^{-\frac{1}{2}})$ through linearization technique. Then by substituting $\hat{\lambda}$ with $\sum_{i\in\mathcal{S}_A}\hat{d}_i^A\hat{u}_i/\sum_{i\in\mathcal{S}_A}\hat{d}_i^A\hat{u}_i^2$, estimator $\hat{\mu}_{PEL}$ can be further written as,

$$\hat{\mu}_{PEL}=\hat{\mu}_{IPW2}+(\hat{\bar{m}}^B-\hat{\bar{m}}_{IPW2})\hat{B}_m+o_p(n_A^{-\frac{1}{2}}).$$

Finally, under regularity conditions **C1**–**C6**, it can be shown that

$$\hat{\mu}_{PEL}-\mu_y=\frac{1}{N}\sum_{i=1}^N\frac{R_i}{\pi_i^A}(y_i-m_i^*B_m^*-k_N-\pi_i^A\boldsymbol{x}_i^{\intercal}\mathbf{b}_1)+\frac{1}{N}\sum_{i\in\mathcal{S}_B}d_i^Bq_i+o_p(n_A^{-\frac{1}{2}}),$$

where B_m^*, k_N, $\mathbf{b}_1$ and q_i are defined in Theorem 3.1. This expression naturally leads to the variance formula for $\hat{\mu}_{PEL}$.

Proof of Theorem 3.2.

Applying the second order Taylor expansion to $\sum_{i\in\mathcal{S}_A}\hat{d}_i^A\log\tilde{p}_i(\mu)$ around $\lambda=0$, we obtain

$$\begin{aligned}
&\sum_{i\in\mathcal{S}_A}\hat{d}_i^A\log\tilde{p}_i(\mu)-\sum_{i\in\mathcal{S}_A}\hat{d}_i^A\log\tilde{p}_i\\
=&\sum_{i\in\mathcal{S}_A}\hat{d}_i^A\log\frac{\hat{d}_i^A}{1+\hat{\lambda}(y_i-\mu)}-\sum_{i\in\mathcal{S}_A}\hat{d}_i^A\log(\hat{d}_i^A)\\
=&-\Big\{\sum_{i\in\mathcal{S}_A}\hat{d}_i^A(y_i-\mu)\hat{\lambda}-\frac{1}{2}\sum_{i\in\mathcal{S}_A}\hat{d}_i^A(y_i-\mu)^2\hat{\lambda}^2\Big\}+o_p(n_A^{-1})\\
=&-\frac{1}{2}(\sum_{i\in\mathcal{S}_A}\hat{d}_i^Ay_i-\mu)^2\Big\{\sum_{i\in\mathcal{S}_A}\hat{d}_i^A(y_i-\mu)^2\Big\}^{-1}+o_p(n_A^{-1}).
\end{aligned}$$

Since $\sum_{i\in\mathcal{S}_A}\hat{d}_i^Ay_i-\mu_y$ is asymptotically normally distributed with zero mean under regularity conditions **C1**–**C4** and **C7**, then

$$\Lambda_1(\mu_y)\frac{n_A^{-1}\sum_{i\in\mathcal{S}_A}\hat{d}_i^A(y_i-\mu_y)^2}{Var(\sum_{i\in\mathcal{S}_A}\hat{d}_i^Ay_i-\mu_y)}\xrightarrow{d}\chi_1^2.$$

Moreover, $Var(\sum_{i\in\mathcal{S}_A}\hat{d}_i^Ay_i-\mu_y)$ can be consistently estimated by v_{IPW2} and $\sum_{i\in\mathcal{S}_A}\hat{d}_i^A(y_i-\mu_y)^2-\sum_{i\in\mathcal{S}_A}\hat{d}_i^A(y_i-\hat{\mu}_{IPW2})^2=o_p(1)$. Then we have $s_1\Lambda_1(\mu_y)\xrightarrow{d}\chi_1^2$ by Slutsky's theorem, where $s_1=n_A^{-1}\sum_{i\in\mathcal{S}_A}\hat{d}_i^A(y_i-\hat{\mu}_{IPW2})^2/v_{IPW2}$.

Proof of Theorem 3.3.

We first write constraints in (3.2.6) in a more compact form, i.e.,

$$\sum_{i\in\mathcal{S}_A}p_i=1\,,\ \sum_{i\in\mathcal{S}_A}p_i\boldsymbol{a}_i=\bar{\boldsymbol{a}},$$

where $\boldsymbol{a}_i=(\hat{m}_i,r_i)^\intercal$, and $\bar{\boldsymbol{a}}=(\hat{\bar{m}}^B,0)^\intercal$.

By using similar techniques as those used in Theorem 3.2, we can show that

$$\begin{aligned}
&\sum_{i\in\mathcal{S}_A} \hat{d}_i^A \log \hat{p}_i(\mu) \\
= \; & -\frac{1}{2}(\sum_{i\in\mathcal{S}_A} \hat{d}_i^A \boldsymbol{a}_i - \bar{\boldsymbol{a}})^\intercal \Big\{ \sum_{i\in\mathcal{S}_A} \hat{d}_i^A (\boldsymbol{a}_i - \bar{\boldsymbol{a}})(\boldsymbol{a}_i - \bar{\boldsymbol{a}})^\intercal \Big\}^{-1} (\sum_{i\in\mathcal{S}_A} \hat{d}_i^A \boldsymbol{a}_i - \bar{\boldsymbol{a}}) + o_p(n_A^{-1}) \\
= \; & -\frac{1}{2}\frac{(\sum_{i\in\mathcal{S}_A} \hat{d}_i^A \hat{m}_i - \hat{\bar{m}}^B)^2}{\sum_{i\in\mathcal{S}_A} \hat{d}_i^A (\hat{m}_i - \hat{\bar{m}}^B)^2} - \frac{1}{2}\frac{(\sum_{i\in\mathcal{S}_A} \hat{d}_i^A r_i)^2}{\sum_{i\in\mathcal{S}_A} \hat{d}_i^A r_i^2} + o_p(n_A^{-1}),
\end{aligned}$$

and

$$\sum_{i\in\mathcal{S}_A} \hat{d}_i^A \log \hat{p}_i = -\frac{1}{2}\frac{(\sum_{i\in\mathcal{S}_A} \hat{d}_i^A \hat{m}_i - \hat{\bar{m}}^B)^2}{\sum_{i\in\mathcal{S}_A} \hat{d}_i^A (\hat{m}_i - \hat{\bar{m}}^B)^2} + o_p(n_A^{-1}).$$

The above two asymptotic expansions lead to the result

$$\sum_{i\in\mathcal{S}_A} \hat{d}_i^A \log \hat{p}_i(\mu) - \sum_{i\in\mathcal{S}_A} \hat{d}_i^A \log \hat{p}_i = -\frac{1}{2}(\sum_{i\in\mathcal{S}_A} \hat{d}_i^A r_i)^2 (\sum_{i\in\mathcal{S}_A} \hat{d}_i^A r_i^2)^{-1} + o_p(n_A^{-1}).$$

Thus, when μ is evaluated at μ_y and under the asymptotic normality of $\sum_{i\in\mathcal{S}_A} \hat{d}_i^A r_i$, we have

$$\Lambda_2(\mu)\frac{n_A^{-1}\sum_{i\in\mathcal{S}_A} \hat{d}_i^A r_i^2}{Var(\sum_{i\in\mathcal{S}_A} \hat{d}_i^A r_i)} \xrightarrow{d} \chi^2(1),$$

To estimate $Var(\sum_{i\in\mathcal{S}_A} \hat{d}_i^A r_i)$, we notice $Var(\sum_{i\in\mathcal{S}_A} \hat{d}_i^A r_i) = V_{PEL} + o(n_A^{-1})$ when $\mu = \mu_y$; and v_{PEL} in (3.2.3) is a consistent estimator of V_{PEL}. Then by Slutsky's theorem, we get

$$s_2 \Lambda_2(\mu_y) \xrightarrow{d} \chi_1^2,$$

where $s_2 = n_A^{-1}\sum_{i\in\mathcal{S}_A} \hat{d}_i^A \{y_i - \hat{\mu}_{PEL} - (\hat{m}_i - \hat{\bar{m}}^B)\hat{B}_m\}^2 / v_{PEL}$ is a computable adjustment factor.

Chapter 4

Statistical Inference with Incomplete Frames

Assumptions **A1**–**A3** stated in Chapter 2 are extremely important when making inferences with non-probability survey samples. In fact, all the major results through Chapter 2 to Chapter 3 are developed under these three assumptions. In this Chapter, we investigate the consequences when Assumption **A2** is not satisfied and the broader issue of incomplete sampling frames.

Recall that Assumption **A2** is referred to as the positivity assumption, which requires every unit in the population to have a positive propensity score. The failure of the positivity assumption is often due to the incomplete sampling frame for $\mathcal{S}_A$ where units with zero propensity score can never be selected into the sample. Zero propensity scores can be viewed as a parallel phenomena to the under-coverage issue in probability survey samples, where the sampling frame only covers the population partially. However, the issue of zero propensity scores can be far more complicated. Unlike probability survey samples, most of non-probability survey samples are not governed by any sampling scheme, which means zero propensity scores can come from a variety of unforeseen sources. Moreover, units with zero propensity score can be difficult to identify in practice since the sample generating

mechanism is always unknown for non-probability survey samples.

The violation to the positivity assumption raises huge inferential obstacles for both the QR approach and the MI approach. Due to the absence of design information such as sampling frame and sampling strategy, positivity assumption is often implicitly used in practice without verification. This chapter focuses on two generating mechanisms for incomplete frames: stochastic mechanism and deterministic mechanism. Stochastic mechanism leads to an incomplete frame but does not violate the positivity assumption. We show that our developed methodologies in Chapters 2 and 3 can be directly extended to this scenario. The positivity assumption does not hold under the deterministic mechanism, which means regular QR and MI approaches are no longer feasible. Under the deterministic mechanism, we review some novel approaches which have potentials to mitigate the impact of zero propensity scores, and evaluate these approaches through simulation studies. We also propose a convex hull method to identify units with zero propensity score, and based on the identified results, a split-population type estimator is constructed to estimate the finite population mean.

4.1 Mechanisms for Incomplete Frames

It is not hard to demonstrate the complications that the QR approach and the MI approach would encounter when zero propensity scores exist. We divide the finite population $\mathcal{U}$ into two subpopulations $\mathcal{U}_1$ and $\mathcal{U}_0$, where $\mathcal{U}_1 = \{i \mid i \in \mathcal{U}, \pi_i^A > 0\}$ with size of N_1, and $\mathcal{U}_0 = \{i \mid i \in \mathcal{U}, \pi_i^A = 0\}$ with size of N_0. The subpopulations $\mathcal{U}_1$ and $\mathcal{U}_0$ have their own corresponding finite population means $\mu_{y,1} = N_1^{-1} \sum_{i \in \mathcal{U}_1} y_i$ and $\mu_{y,0} = N_0^{-1} \sum_{i \in \mathcal{U}_0} y_i$ respectively. Let $\tau = N_0/N$ be the proportion of zero propensity scores.

When $\tau \neq 0$, applying the QR approach is problematic both in practice and theory. Practically, commonly used propensity score models such as logistic model are not compatible with zero selection probability, and using these conventional models by ignoring positivity violation could lead to biased inferences. Theoretically, two IPW estimators do

not converge to μ_y. In particular, the estimator $\hat{\mu}_{IPW1}$ converges to $(1-\tau)\mu_{y,1}$, while the estimator $\hat{\mu}_{IPW2}$ converges to $\mu_{y,1}$. One may notice that propensity scores are not explicitly involved in the MI approach. However, the positivity assumption is still inevitable. One direct consequence of the violation is that the prediction model, which is built upon relation $E_\xi(y_i \mid \boldsymbol{x}_i) = E_\xi(y_i \mid \boldsymbol{x}_i, R_i = 1)$, might not give valid predictions for those $\boldsymbol{x}_i$ with $P(R_i = 1 \mid \boldsymbol{x}_i) = 0$. We use the following two examples to illustrate this phenomenon.

For the first example, we assume that the target population of a study is all the students in a school, but the sample $\mathcal{S}_A$ contains no data on female students. Suppose that gender, related to the variable of interest, is a component of covariates $\boldsymbol{x}$, then the propensity score of female students can be treated as zero. If using NN imputation in this case, then no female student in sample $\mathcal{S}_B$ has a close match in sample $\mathcal{S}_A$ in terms of gender. If adopting regression prediction method described in Section 2.2.1, then the estimation of parameter $\boldsymbol{\beta}_0$ is unattainable since the design matrix based on data $\{(\boldsymbol{x}_i, y_i), i \in \mathcal{S}_A\}$ is not of full rank.

For the second example, the research interest is the average vacation spending of the target population. We assume that personal income belongs to the set of covariates $\boldsymbol{x}$, and the selection process for sample $\mathcal{S}_A$ accidentally excludes the part of the population which have lower income, i.e., units with lower income have zero propensity score. Under this scenario, no close match can be found for those units with lower income when applying the NN imputation. Regression prediction method can also be challenging. Prediction model is obtained based on sample $\mathcal{S}_A$ which only contains units with higher income, but predictions are also required for units with lower income for computing regression based estimates. This is actually the issue of extrapolation, which is further illustrated in Section 4.1.2.

In summary, statistical inferences in the presence of zero propensity scores require suitable adaptions of the methodologies developed in Chapters 2 and 3. In this section, we discuss two generating mechanisms for propensity scores and incomplete frames.

4.1.1 Stochastic mechanism

We consider a two-stage stochastic mechanism to generate sample $\mathcal{S}_A$. This mechanism is frequently visited in the existing literature, and the example of online survey panel is often used for illustrations. Online panels typically recruit panel members through some non-probability based method such as sending out invitations and posting advertisements online. Once being recruited, panel members will receive notifications whenever some survey need to be filled. The survey participation is mostly voluntary, and cash or non-cash prize is usually used as incentives for participation. Under this example, the first stage of obtaining $\mathcal{S}_A$ is the panel recruitment, and the second stage is panel members choosing to fill out surveys.

Formally, let z be the indicator variable of being in the sampling frame or not, then we accordingly have $z_i = 1$ if unit i belongs to the online panel, and $z_i = 0$ otherwise. Furthermore, we assume $P(z_i = 1 \mid \boldsymbol{x}_i) > 0$ for $i \in \mathcal{U}$, i.e., every unit in the population has a positive chance to be part of the sampling frame. This assumption is realistic in many scenarios; for the survey panel example, the assumption is met if the panel registration link is posted on the website where all the target population would browse. Due to the randomization of variable z, this process is referred to as "stochastic" mechanism. The population $\mathcal{U}$ is then divided by indicator z into subpopulations $\breve{\mathcal{U}}_1 = \{i \mid i \in \mathcal{U}, z_i = 1\}$ and $\breve{\mathcal{U}}_0 = \{i \mid i \in \mathcal{U}, z_i = 0\}$, where $\breve{\mathcal{U}}_1$ consists of units which are used as the sampling frame for the final sample $\mathcal{S}_A$, and $\breve{\mathcal{U}}_0$ consists of the rest. Based on the survey panel example, it is reasonable for us to assume that every unit in the set $\breve{\mathcal{U}}_1$ has a positive chance to enter sample $\mathcal{S}_A$, i.e., $P(R_i = 1 \mid \boldsymbol{x}_i = 1, z_i = 1) > 0$ for every i. Note the positivity assumption still holds under this mechanism since

$$\pi_i^A = P(R_i = 1 \mid \boldsymbol{x}_i) = P(R_i = 1 \mid \boldsymbol{x}_i, z_i = 1)P(z_i = 1 \mid \boldsymbol{x}_i), \tag{4.1.1}$$

where both $P(R_i = 1 \mid \boldsymbol{x}_i, z_i = 1)$ and $P(z_i = 1 \mid \boldsymbol{x}_i)$ are positive values based on the assumptions. The validity of the positivity assumption means that propensity scores

can be estimated by the parametric model $\pi(\boldsymbol{x}_i, \boldsymbol{\theta})$ we propose in Section 2.3, and the methodologies developed in Chapters 2 and 3 can be used here.

Instead of adopting a single model like logistic regression model for the propensity scores, we can also choose to model $P(R_i = 1 \mid \boldsymbol{x}_i, z_i = 1)$ and $P(z_i = 1 \mid \boldsymbol{x}_i)$ separately. Note that measurements of $\boldsymbol{x}$ are required for each $z_i = 1$ to identify both models for $P(R_i = 1 \mid \boldsymbol{x}_i, z_i = 1)$ and $P(z_i = 1 \mid \boldsymbol{x}_i)$.

The stochastic mechanism is obviously a simplification of real case scenarios, and it is possible that multiple layers of panel sign-up indicators exist. For example, besides panel sign-up variable z, assume there is a variable z_1 which indicates the status of individual's internet access. Notice that the set $\{i \mid i \in \mathcal{U}, z_i = 1\}$ is a subset of $\{i \mid i \in \mathcal{U}, z_{1i} = 1\}$ since the sign-up requires internet access, and an individual can only enter into sample $\mathcal{S}_A$ if $z = 1$ and $z_1 = 1$. Under this mechanism, the propensity scores are given by

$$\pi_i^A = P(R_i = 1 \mid \boldsymbol{x}_i, z_i = 1, z_{1i} = 1) P(z_i = 1 \mid \boldsymbol{x}_i, z_{1i} = 1) P(z_{1i} = 1 \mid \boldsymbol{x}_i). \tag{4.1.2}$$

To estimate π_i^A by modelling each part of the decomposition (4.1.2), we need measurements of indicator z_i, z_{1i} and $\boldsymbol{x}$ for $z_{1i} = 1$, which however are often hard to obtain in practice. Elliott and Valliant (2017) gives a more sophisticated example about how complicated the mechanism can get by taking more relevant variables into account. So even if multiple model approach are theoretically more suitable for the mechanism in (4.1.1) and (4.1.2), approximating propensity scores with a carefully chosen single model is often more feasible in practice.

Besides modelling issue, analysts should be aware that this stochastic mechanism relies on the assumption that indicator variable z is not a confounding variable for R and y. This assumption may not always hold in reality. For example, if variable y is the amount of hours spent on online activities, then it is very likely that individuals on the panel have higher values of y than those who are not. When both z_i and $\boldsymbol{x}_i$ are confounding variables for R and y, the propensity scores are computed as $\pi_i^A = P(R_i = 1 \mid \boldsymbol{x}_i, z_i = 0) = 0$ and $\pi_i^A = P(R_i = 1 \mid \boldsymbol{x}_i, z_i = 1) > 0$. It is obvious that the positivity assumption fails

under this scenario, and neither the QR approach or the MI approach is applicable. In the simulation study reported in Section 4.4, we will evaluate the performance of QR and MI approach with $\boldsymbol{x}_i$ being treated as the confounding variables, while $(z_i, \boldsymbol{x}_i)$ is actually used as confounding variables to generate sample $\mathcal{S}_A$.

4.1.2 Deterministic mechanism

Assume that sample $\mathcal{S}_A$ is obtained by the following strategy. A researcher has a complete list of individuals in the target population. And for each individual i, there is an associated measurement of accessibility $\Phi(\boldsymbol{x}_i)$, where $\Phi(\cdot)$ is some function about confounding variables $\boldsymbol{x}_i$. The value of $\Phi(\boldsymbol{x}_i)$ for each i is also available to the researcher, but measurement of $\boldsymbol{x}_i$ can not be observed before sampling. To save costs and time, the researcher chooses to contact individuals who are more accessible. So the researcher orders individuals by their accessibility $\Phi(\boldsymbol{x})$ from the highest to the lowest, and contact ordered individuals one by one. Individual i, once being contacted, has probability $\pi(\boldsymbol{x}_i, \boldsymbol{\theta})$ of taking the survey. The procedure stops when the total of $(1-\tau)N$ individuals are contacted. The sample selection mechanism of this example can be written as,

$$\pi_i^A = \begin{cases} \pi(\boldsymbol{x}_i, \boldsymbol{\theta}) & \text{if } \Phi(\boldsymbol{x}_i) > Q(\tau), \\ 0 & \text{otherwise,} \end{cases} \tag{4.1.3}$$

where $Q(\tau)$ is the τth percentile of $\{\Phi(\boldsymbol{x}_1), \cdots, \Phi(\boldsymbol{x}_N)\}$. In other words, the original selection mechanism in Section 2.1 is truncated so that units with small value of $\Phi(\boldsymbol{x})$ have no chance of entering into sample $\mathcal{S}_A$. The frame indicator z under this mechanism is computed as $z_i = 1$ if $\Phi(\boldsymbol{x}_i) > Q(\tau)$, and $z_i = 0$ otherwise. This means that $\breve{\mathcal{U}}_1 = \mathcal{U}_1$ and $\breve{\mathcal{U}}_0 = \mathcal{U}_0$ under this mechanism. We refer to this mechanism as the deterministic mechanism, since the sampling frame is determined by fixed values $\Phi(\boldsymbol{x}_i)$. Similarly to the two-stage stochastic mechanism in the last section, sample $\mathcal{S}_A$ here can also be viewed as the result of a two-stage process. The first stage is to determine the sampling frame by a

known function $\Phi(\boldsymbol{x}_i)$, and the second stage is to obtain the final sample from the frame by certain randomized process. Note that the first stage in both mechanisms is dependent on some function of covariates, but the frame indicator z follows a randomized process in the stochastic mechanism, while being deterministically generated in the deterministic mechanism.

When the propensity scores follow the non-truncated model $\pi(\boldsymbol{x}_i,\boldsymbol{\theta})$, every unit in $\mathcal{U}$ has a positive probability of being selected in sample $\mathcal{S}_A$, which implies that $\mathcal{S}_A$ and the population $\mathcal{U}$ share the same support with respect to covariates $\boldsymbol{x}$. However, the supports of $\boldsymbol{x}$ for $\mathcal{U}_1$ and $\mathcal{U}_0$ form two sets that do not overlap with each other based on (4.1.3). This means that units in sample $\mathcal{S}_A$ are not representative for the subpopulation $\mathcal{U}_0$, and thus it is not sensible to use the IPW method to infer the characteristics of the entire population.

Regression prediction approach is also questionable due to the possible failure of the relation $E(y_i \mid \boldsymbol{x}_i) = E_\xi(y_i \mid \boldsymbol{x}_i, R_i = 1)$ for $i \in \mathcal{U}_0$ under the deterministic mechanism. Furthermore, it has been noticed by several researchers, for example, Tan (2007) that regression approach suffers the issue of extrapolation. Recall from Section 2.2.1 that the model $m(\boldsymbol{x}_i,\boldsymbol{\beta})$ is posited and estimated based on $\mathcal{S}_A$ exclusively, and $\mathcal{S}_A$ does not share a common support with $\mathcal{U}_0$. It is difficult for analysts to check how closely the obtained model fits the data from $\mathcal{U}_0$, since no response y is available from $\mathcal{U}_0$. The NN imputation can also be problematic under this mechanism. For units which belong to $\mathcal{U}_0$, no exact match can be found in $\mathcal{S}_A$ in theory due to the non-overlapped supports for $\mathcal{U}_1$ and $\mathcal{U}_0$.

4.2 Existing Approaches

We argue in Section 4.1 that the QR, MI and PEL approaches can be immediately extended to the stochastic mechanism, but are not valid under the deterministic mechanism due to the positivity violation. From now on, we mainly focus on the deterministic mechanism and explore inferential procedures which are robust against zero propensity scores. The following three procedures are examined first, which do not aim at solving the positivity

issue directly but require no positivity assumption explicitly for inferences.

4.2.1 Calibrated IPW approach

When the propensity scores follow the non-truncated model $\pi(\boldsymbol{x}_i,\boldsymbol{\theta})$, we used the score equations (2.3.7) to obtain the maximum pseudo likelihood estimator $\hat{\boldsymbol{\theta}}_{ml}$. In addition, we discussed an alternative estimator of $\boldsymbol{\theta}$, that is, the solution of the following calibration type estimating equations,

$$\sum_{i\in\mathcal{S}_A} \boldsymbol{x}_i/\pi_i(\boldsymbol{\theta}) - \sum_{i\in\mathcal{S}_B} d_i^B \boldsymbol{x}_i = 0. \tag{4.2.4}$$

Assume that a unique solution exists to (4.2.4), and let $\hat{\boldsymbol{\theta}}_{cal}$ denote the solution. This equation forces the weighted estimator $\sum_{i\in\mathcal{S}_A} \boldsymbol{x}_i/\pi_i(\hat{\boldsymbol{\theta}}_{cal})$ to be equal to the estimated population totals $\sum_{i\in\mathcal{S}_B} d_i^B \boldsymbol{x}_i$. Assume that the intercept is included in the model, the estimated population sizes based on two samples are also calibrated to each other in the sense that $\hat{N}_{cal}^A = \hat{N}^B$, where $\hat{N}_{cal}^A = \sum_{i\in\mathcal{S}_A} 1/\pi_i(\hat{\boldsymbol{\theta}}_{cal})$. Given $\hat{\boldsymbol{\theta}}_{cal}$, the normalized IPW estimator for μ_y is computed as $\hat{\mu}_{IPW,cal} = (\hat{N}_{cal}^A)^{-1} \sum_{i\in\mathcal{S}_A} y_i/\pi_i(\hat{\boldsymbol{\theta}}_{cal})$.

The estimator $\hat{\mu}_{IPW,cal}$ is robust against zero propensity scores if there is a linear relation between y and $\boldsymbol{x}$. Specifically, if y_i and $\boldsymbol{x}_i$ satisfies $E_\xi(y_i \mid \boldsymbol{x}_i) = \boldsymbol{x}_i^\intercal\boldsymbol{\beta}$, for some $\boldsymbol{\beta}$, for $i \in \mathcal{U}$, then we have

$$E_\xi\left\{\frac{1}{\hat{N}_{cal}^A}\sum_{i\in\mathcal{S}_A}\frac{y_i}{\pi_i(\hat{\boldsymbol{\theta}}_{cal})}\right\} = \frac{1}{\hat{N}_{cal}^A}\sum_{i\in\mathcal{S}_A}\frac{\boldsymbol{x}_i^\intercal}{\pi_i(\hat{\boldsymbol{\theta}}_{cal})}\boldsymbol{\beta} = \frac{1}{\hat{N}^B}\sum_{i\in\mathcal{S}_B} d_i^B \boldsymbol{x}_i^\intercal\boldsymbol{\beta}. \tag{4.2.5}$$

The relation (4.2.5) shows that under the linearity assumption, the estimator $\hat{\mu}_{IPW,cal}$ is approximately unbiased even if $\mathcal{S}_A$ is not generated from the non-truncated model $\pi(\boldsymbol{x}_i,\boldsymbol{\theta}_0)$.

However, one may easily argue that the regression prediction approach can be used directly if there is a known linear relation between y and $\boldsymbol{x}$. Besides, relation (4.2.5) only holds if (4.2.4) has a solution, which is not guaranteed in practice.

4.2.2 Modified nearest neighbour approach

Recall from Section 2.2.1 that the NN procedure assigns a match to every unit in sample $\mathcal{S}_B$. However, close matches may not exist for units in sample $\mathcal{S}_B$ under the deterministic mechanism. Kim and Rao (2018) modifies this NN idea by only conducting matching on units which potentially belong to $\mathcal{U}_1$. We refer to their approach as KR-NN.

More specifically, the KR-NN method is achieved by two steps. The first step is to classify units in sample $\mathcal{S}_B$ into two subsamples $\mathcal{S}_{B,1}$ and $\mathcal{S}_{B,0}$, where $\mathcal{S}_{B,1} = \mathcal{S}_B \cap \mathcal{U}_1$ and $\mathcal{S}_{B,0} = \mathcal{S}_B \cap \mathcal{U}_0$. Since the true propensity scores are not always available to partition $\mathcal{S}_B$, the following set $\tilde{\mathcal{S}}_{B,1}$ is proposed to approximate $\mathcal{S}_{B,1}$,

$$\tilde{\mathcal{S}}_{B,1} = \{i \mid min_{j \in \mathcal{S}_A} ||\boldsymbol{x}_i - \boldsymbol{x}_j|| < \epsilon, i \in \mathcal{S}_B\},$$

where ϵ is a pre-specified constant. In other words, unit i belongs to $\tilde{\mathcal{S}}_{B,1}$ if its distance to its closest neighbour in sample $\mathcal{S}_A$ is less than the value of ϵ. Next, similarly to the original NN method, the missing response for units in $\tilde{\mathcal{S}}_{B,1}$ is imputed with the response value of its closest match in $\mathcal{S}_A$.

The second step is the calibration weighting, i.e., finding some weights w_i which satisfy $\sum_{i \in \tilde{\mathcal{S}}_{B,1}} w_i x_i = \sum_{i \in \mathcal{S}_B} d_i^B x_i$. Finally, the resulting estimator for μ_y under KR-NN is given by $\hat{\mu}_{KR} = 1/\hat{N}^B \sum_{i \in \tilde{\mathcal{S}}_{B,1}} w_i \hat{y}_i$. Note if $\tilde{\mathcal{S}}_{B,1} = \mathcal{S}_B$, then estimator $\hat{\mu}_{KR}$ coincides with the original NN based estimator $\hat{\mu}_{NN}$. The original NN method can be viewed as a special case of of the KH-NN method, where ϵ is set large enough to guarantee a match for each $i \in \mathcal{S}_B$. If the size of $\tilde{\mathcal{S}}_{B,1}$ is much smaller than the size of $\mathcal{S}_B$, then the adequacy of the estimator $\hat{\mu}_{KR}$ depends heavily on the calibration adjustment.

4.2.3 Stable weights approach

It is a broadly noticed phenomenon that small propensity scores, even being positive, could result in large variances for IPW estimators. To find a set of stable weights, Zubizarreta

(2015) takes a design-based approach by treating weights of $\mathcal{S}_A$ as unknown parameters. This approach does not require the positivity assumption explicitly, since the propensity score model is not used for the estimation. Let w_i denote the weight of unit i, for $i \in \mathcal{S}_A$. The goal of the approach is to find a set of weights $\{w_1, w_2, \cdots, w_{n_A}\}$ which has the minimum variance under the specified calibration constraints. Specifically, the problem is formulated as minimizing $\sum_{i\in\mathcal{S}_A} w_i^2$ subject to $\sum_{i\in\mathcal{S}_A} w_i = 1$, $w_i \geq 0$ for $i \in \mathcal{S}_A$ and

$$\Big| \sum_{i\in\mathcal{S}_A} w_i \boldsymbol{x}_i - 1/\hat{N}^B \sum_{i\in\mathcal{S}_B} d_i^B \boldsymbol{x}_i \Big| < \boldsymbol{\delta}, \tag{4.2.6}$$

where (4.2.6) are calibration constraints, and the constant vector $\boldsymbol{\delta}$ is the user-specified calibration tolerance. The resulting estimator is given by $\hat{\mu}_{SW} = \sum_{i\in\mathcal{S}_A} \hat{w}_i y_i$, where $\hat{w}_i$ is obtained from the optimization problem described above. As the goal of this procedure is to obtain stable weights, calibration constraints are exclusively responsible for reducing selection bias. However, similarly to our previous argument, unless there is an approximate linear relation between y and $\boldsymbol{x}$, the calibration constraints cannot be accounted to remove all the selection bias.

In summary, even though the positivity assumption is not postulated explicitly in these three approaches, the linearity between $\boldsymbol{x}$ and y is still implicitly required to obtain valid inferences under the deterministic mechanism. Like the positivity assumption, linearity is also a strong assumption which tends to oversimplify the reality for most of cases.

4.3 Split-population Approach

We propose to use a split-population approach under the deterministic mechanism. Split-population approach has been used in the survey sampling setting to analyze and combine data from different sources; e.g., Zhang (2019) applied this approach to non-probability samples. For the current setup, the "split-population" refers to subpopulations $\mathcal{U}_1$ and $\mathcal{U}_0$. While the sample $\mathcal{S}_A$ belongs to $\mathcal{U}_1$ automatically, units in the sample $\mathcal{S}_B$ need to be

classified into $\mathcal{U}_1$ and $\mathcal{U}_0$. In fact, the first step of KR-NN imputation is an example of splitting sample $\mathcal{S}_B$. Under the split-population approach, estimators for μ_y take the form of

$$\hat{\mu}_y = (1-\hat{\tau})\hat{\mu}_{y,1} + \hat{\tau}\hat{\mu}_{y,0},$$

where $\hat{\mu}_{y,1}$ and $\hat{\mu}_{y,0}$ are estimators of $\mu_{y,0}$ and $\mu_{y,1}$ respectively, and $\hat{\tau}$ is the estimated proportion of $\mathcal{U}_0$. Once sample $\mathcal{S}_B$ is split, then computing $\hat{\tau}$ is trivial, and estimators $\hat{\mu}_{y,1}$ and $\hat{\mu}_{y,0}$ can be obtained separately by suitable methods. The splitting step can be computationally expensive, but it provides analysts with more insights into the data structure. Based on the classification result, estimation methods can be chosen in a more sensible manner. For example, if the proportion of zero propensity scores is small, using classic methods in Section 2.2 may still be reasonable; but if the proportion of $\mathcal{U}_0$ is large, then more sophisticated procedures should be considered.

4.3.1 Splitting method

Splitting sample $\mathcal{S}_B$ is essentially to identify subsamples $\mathcal{S}_{B,1}$ and $\mathcal{S}_{B,0}$, which are defined in Section 4.2.2. We propose the following convex hull method for classifications under the deterministic mechanism.

For the purpose of illustration, we assume that $\boldsymbol{x}$ are continuous, and the underlying function $\Phi(\boldsymbol{x})$ used for truncation follows a logistic regression model, with $\boldsymbol{x}$ being the covariates. Let $\boldsymbol{R}_1$ and $\boldsymbol{R}_0$ denote the support of $\boldsymbol{x}$ for subpopulations $\mathcal{U}_1$ and $\mathcal{U}_0$ respectively. Since the supports $\boldsymbol{R}_1$ and $\boldsymbol{R}_0$ do not overlap according to (4.1.3), the set $\mathcal{S}_{B,1}$, as a sample of $\mathcal{U}_1$, can be distinguished from $\mathcal{S}_{B,0}$ once $\boldsymbol{R}_1$ is identified. The support $\boldsymbol{R}_1$ can not be found directly, but can be approximated through sample $\mathcal{S}_A$, which shares a common support with $\mathcal{U}_1$. Let C_{n_A} be the convex hull generated by the set $\{\boldsymbol{x}_i \mid i \in \mathcal{S}_A\}$, then the following condition holds under the assumptions made above,

C8 For any $\boldsymbol{x}_j \in \boldsymbol{R}_1 \cup \boldsymbol{R}_0$, we have $I(\boldsymbol{x}_j \in C_{n_A}) \to I(\boldsymbol{x}_j \in \boldsymbol{R}_1)$.

When Condition **C8** holds, the convex hull C_{n_A} can be seen as a substitute of $\boldsymbol{R}_1$, it follows that the sets $\mathcal{S}_{B,1}$ and $\mathcal{S}_{B,0}$ can be respectively approximated by $\hat{\mathcal{S}}_{B,1} = \mathcal{S}_B \cap C_{n_A}$ and $\hat{\mathcal{S}}_{B,0} = \mathcal{S}_B / \hat{\mathcal{S}}_{B,1}$. The subpopulation counts N_1, N_0 and the proportion τ can be estimated by $\hat{N}_1^B = \sum_{i \in \hat{\mathcal{S}}_{B,1}} d_i^B$, $\hat{N}_0^B = \sum_{i \in \hat{\mathcal{S}}_{B,0}} d_i^B$ and $\hat{\tau} = \hat{N}_0^B / \hat{N}^B$.

To see if $\boldsymbol{x}_j$ belongs to C_{n_A} for a given j, it suffices to check if there exist some a_i for $i \in \mathcal{S}_A$ which satisfy the following constraints,

$$\sum_{i \in \mathcal{S}_A} a_i \boldsymbol{x}_i = \boldsymbol{x}_j \,, \ \sum_{i \in \mathcal{S}_A} a_i = 1, \quad \text{and} \quad a_i \geq 0, \ \ \forall \, i \in \mathcal{S}_A.$$

Condition **C8** holds for a broad class of $\Phi(\boldsymbol{x})$ functions besides the logistic function. Let $C_{A,1} = \lim_{n_A \to \infty} C_{n_A}$, then it is trivial that the relations $\boldsymbol{R}_1 \subseteq C_{A,1}$ and $C_{A,1} \cap \boldsymbol{R}_0 = \emptyset$ are sufficient conditions for **C8** being valid. A variety of forms for $\Phi(\boldsymbol{x})$ satisfy these two relations, and a special case $\Phi(\boldsymbol{x}) = \pi(\boldsymbol{x}_i, \boldsymbol{\theta})$ also meets sufficient conditions. Moreover, Condition **C8** can also be extended for non-continuous covariates $\boldsymbol{x}$.

Even if Condition **C8** is not satisfied, this classification procedure can also be used as a diagnostic tool to check assumptions on propensity scores. When the non-truncated model $\pi(\boldsymbol{x}_i, \boldsymbol{\theta})$ is adopted, then the assumption that $\mathcal{S}_A$ and $\mathcal{S}_B$ have the same support in terms of $\boldsymbol{x}$ is also made implicitly. Checking the proportion of $\mathcal{S}_B$ falling into C_{n_A} can be treated as checking the overlap between $\mathcal{S}_A$ and $\mathcal{S}_B$. If the non-truncated model $\pi(\boldsymbol{x}, \boldsymbol{\theta})$ holds, then $I\{\boldsymbol{x}_j \in C_{n_A}\} \xrightarrow{p} 1$ for any $\boldsymbol{x}_j \in \mathcal{S}_B$, and $\hat{\tau} \xrightarrow{p} 0$ under mild conditions. If the overlap turns out scarce, then the non-truncated model assumption should be further investigated.

4.3.2 Estimation under the split-population approach

To construct estimators for μ_y based on the split-population, the most important step is to estimate $\mu_{y,1}$ and $\mu_{y,0}$. Since the positivity assumption holds for subpopulation $\mathcal{U}_1$ under the deterministic mechanism, the estimation of $\mu_{y,1}$ can still be achieved through IPW approach. To estimate the propensity scores for $i \in \mathcal{U}_1$, the following unbiased estimating

equations are considered,

$$\sum_{i\in\mathcal{S}_A} \boldsymbol{x}_i - \sum_{i\in\mathcal{S}_{B,1}} d_i^B \pi_i(\boldsymbol{\theta})\boldsymbol{x}_i = 0. \tag{4.3.7}$$

Let $\tilde{\boldsymbol{\theta}}$ be the solution to (4.3.7), then $\pi_i(\tilde{\boldsymbol{\theta}})$ is a consistent estimator of π_i^A for $i \in \mathcal{U}_1$. However, the exact components of $\mathcal{S}_{B,1}$ are unknown under the deterministic mechanism, so we replace set $\mathcal{S}_{B,1}$ by its approximation $\hat{\mathcal{S}}_{B,1}$ to obtain a different solution, denoted by $\hat{\boldsymbol{\theta}}_{cv}$. In fact, we show in Section 4.6 that $\hat{\boldsymbol{\theta}}_{cv}$ and $\tilde{\boldsymbol{\theta}}$ are asymptotically equivalent in the sense that $\hat{\boldsymbol{\theta}}_{cv} = \tilde{\boldsymbol{\theta}} + o_p(n_A^{-\frac{1}{2}})$ if condition **C8** holds. Given $\hat{\boldsymbol{\theta}}_{cv}$, the resulting IPW estimator of $\mu_{y,1}$ is computed as $\hat{\mu}_{1,IPW} = (\hat{N}_1^B)^{-1}\sum_{i\in\mathcal{S}_A} y_i/\pi_i(\hat{\boldsymbol{\theta}}_{cv})$.

The regression prediction and doubly robust methods can also be used to estimate $\mu_{y,1}$, with the resulting estimators being given by $\hat{\mu}_{1,REG} = (\hat{N}_1^B)^{-1}\sum_{i\in\hat{\mathcal{S}}_{B,1}} d_i^B m_i(\hat{\boldsymbol{\beta}})$ and $\hat{\mu}_{1,DR} = (\hat{N}_1^B)^{-1}\sum_{i\in\mathcal{S}_A}\{y_i - m_i(\hat{\boldsymbol{\beta}})\}/\pi_i(\hat{\boldsymbol{\theta}}_{cv}) + \hat{\mu}_{1,REG}$ respectively. The NN imputation is straightforward to conduct by assigning matching unit for $i \in \hat{\mathcal{S}}_{B,1}$ from sample $\mathcal{S}_A$.

On the other hand, estimating $\mu_{y,0}$ is a challenging task. The IPW approach is not applicable since units in $\mathcal{U}_0$ have zero propensity scores. If adopting NN method, close matches for $i \in \mathcal{S}_{B,0}$ may not exist. Regression prediction is a relatively reasonable choice among methods we considered in the thesis, which gives the estimator $\hat{\mu}_{0,REG} = (\hat{N}_0^B)^{-1}\sum_{i\in\hat{\mathcal{S}}_{B,0}} d_i^B m_i(\hat{\boldsymbol{\beta}})$. If we let $\hat{\mu}_{y,1} = \hat{\mu}_{1,REG}$ and $\hat{\mu}_{y,0} = \hat{\mu}_{0,REG}$, then the resulting estimator for μ_y becomes the classic regression prediction estimator $\hat{\mu}_{REG}$. A more sensible option regarding robustness is to take $\hat{\mu}_{y,1} = \hat{\mu}_{1,DR}$ and $\hat{\mu}_{y,0} = \hat{\mu}_{0,REG}$, which leads to the following estimator for μ_y,

$$\begin{aligned}
\hat{\mu}_{HYB} &= (1-\hat{\tau})\hat{\mu}_{1,DR} + \hat{\tau}\hat{\mu}_{0,REG} \\
&= \frac{\hat{N}_1^B}{\hat{N}^B}\Big\{\frac{1}{\hat{N}_1^B}\sum_{i\in\mathcal{S}_A}\frac{y_i - m_i(\hat{\boldsymbol{\beta}})}{\pi_i(\hat{\boldsymbol{\theta}}_{cv})} + \frac{1}{\hat{N}_1^B}\sum_{i\in\hat{\mathcal{S}}_{B,1}} d_i^B m_i(\hat{\boldsymbol{\beta}})\Big\} + \frac{\hat{N}_0^B}{\hat{N}^B}\Big\{\frac{1}{\hat{N}_0^B}\sum_{i\in\hat{\mathcal{S}}_{B,0}} d_i^B m_i(\hat{\boldsymbol{\beta}})\Big\} \\
&= \frac{1}{\hat{N}^B}\sum_{i\in\mathcal{S}_A}\frac{y_i - m_i(\hat{\boldsymbol{\beta}})}{\pi_i(\hat{\boldsymbol{\theta}}_{cv})} + \frac{1}{\hat{N}^B}\sum_{i\in\mathcal{S}_B} d_i^B m_i(\hat{\boldsymbol{\beta}}).
\end{aligned}$$

We refer to $\hat{\mu}_{HYB}$ as the hybrid estimator, since different estimating methods are adopted

for the two subpopulations. We notice that the estimator $\hat{\mu}_{HYB}$ has a similar form to a classic doubly robust estimator, and parameter τ and its estimator do not explicitly appear in the formula. However, estimator $\hat{\mu}_{HYB}$ does not have the DR property if τ is non-negligible.

To derive asymptotic properties of estimator $\hat{\mu}_{HYB}$, we assume that parameter $\boldsymbol{\beta}$ for the prediction model is obtained from the following estimating equations,

$$\sum_{i \in \mathcal{S}_A} \{y_i - m(\boldsymbol{x}_i, \boldsymbol{\beta})\}\boldsymbol{x}_i = 0. \tag{4.3.8}$$

Moreover, similarly to assumptions made on the DR estimators in Chapter 2, we assume that there exists a constant vector $\boldsymbol{\beta}^*$ such that $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}^* + O_p(n_A^{-\frac{1}{2}})$ regardless of the model specification.

Theorem 4.1. *Under Assumptions* **A1** *and* **A3**, *as well as regularity conditions* **C1**–**C6** *and* **C8**, *the estimator $\hat{\mu}_{HYB}$ has the following asymptotic properties when the incomplete sampling frame is generated by the deterministic mechanism.*

(i) The estimator $\hat{\mu}_{HYB}$ can be expressed as,

$$\hat{\mu}_{HYB} = (1-\tau)\mu_{y,1} + \tau\bar{m}_0^* + o_p(1),$$

where $\bar{m}_0^* = N_0^{-1}\sum_{i \in \mathcal{U}_0} m_i(\boldsymbol{\beta}^*)$.

(ii) The asymptotic variance formula of $\hat{\mu}_{HYB}$ is given by $Var(\hat{\mu}_{HYB}) = V_{HYB} + o_p(n_A^{-1})$, where

$$V_{HYB} = \frac{1}{N^2}\sum_{i=1}^{N} \pi_i^A\{1-\pi_i^A\}s_i^2 + \frac{1}{N^2}V_p\Big(\sum_{i \in \mathcal{S}_B} d_i^B l_i\Big),$$

$s_i = \{y_i - m_i(\boldsymbol{\beta}^*)\}/\pi_i^A - \mathbf{c}_2^\intercal \boldsymbol{x}_i + \mathbf{c}_1^\intercal\{y_i - m_i(\boldsymbol{\beta}^*)\}\boldsymbol{x}_i$, $l_i = f_i - f_N - N^{-1}\sum_{i \in \mathcal{S}_A} s_i$, $f_i = m_i(\boldsymbol{\beta}^*) + \pi_i^A \mathbf{c}_2^\intercal \boldsymbol{x}_i I(\boldsymbol{x}_i \in \boldsymbol{R}_1)$, $f_N = N^{-1}\sum_{i=1}^N f_i$, *and*

$$\mathbf{c}_1^\intercal = \Big\{\sum_{i \in \mathcal{U}_0} \dot{m}_i(\boldsymbol{\beta}^*)^\intercal\Big\}\Big\{\sum_{i \in \mathcal{U}_1} \pi_i^A \dot{m}_i(\boldsymbol{\beta}^*)^\intercal \boldsymbol{x}_i\Big\}^{-1},$$

$$\mathbf{c}_2^{\intercal} = \Big[\sum_{i\in\mathcal{U}_1}(1-\pi_i^A)\{y_i - m_i(\boldsymbol{\beta}^*)\}\boldsymbol{x}_i^{\intercal}\Big]\Big\{\sum_{i\in\mathcal{U}_1}\pi_i^A(1-\pi_i^A)\boldsymbol{x}_i\boldsymbol{x}_i^{\intercal}\Big\}^{-1}.$$

Based on the asymptotic variance formula V_{HYB}, a consistent variance estimator for $\hat{\mu}_{HYB}$ is computed as,

$$v_{HYB} = \frac{1}{N^2}\sum_{i\in\mathcal{S}_A}\{1-\pi_i(\hat{\boldsymbol{\theta}}_{cv})\}\hat{s}_i^2 + \frac{1}{N^2}\sum_{i\in\mathcal{S}_B}\sum_{j\in\mathcal{S}_B}\frac{\pi_{ij}^B - \pi_i^B\pi_j^B}{\pi_{ij}^B\pi_j^B\pi_i^B}\hat{l}_i\hat{l}_j + o_p(n_A^{-\frac{1}{2}}),$$

with

$$\hat{s}_i = \frac{y_i - m_i(\hat{\boldsymbol{\beta}})}{\pi_i(\hat{\boldsymbol{\theta}}_{cv})} - \hat{\mathbf{c}}_2^{\intercal}\boldsymbol{x}_i + \hat{\mathbf{c}}_1^{\intercal}\big\{y_i - m_i(\hat{\boldsymbol{\beta}})\big\}\boldsymbol{x}_i,$$

and

$$\hat{l}_i = \hat{f}_i - \frac{1}{N}\sum_{i\in\mathcal{S}_B}d_i^B\hat{f}_i - \frac{1}{N}\sum_{i\in\mathcal{S}_A}\hat{s}_i,$$

where $\hat{f}_i = m_i(\hat{\boldsymbol{\beta}}) + \pi_i(\hat{\boldsymbol{\theta}}_{cv})\hat{\mathbf{c}}_2^{\intercal}\boldsymbol{x}_i I(\boldsymbol{x}_i \in C_{n_A})$, $\hat{\mathbf{c}}_1^{\intercal} = \big\{\sum_{i\in\hat{\mathcal{S}}_{B,0}}d_i^B\dot{m}_i(\hat{\boldsymbol{\beta}})^{\intercal}\big\}\big\{\sum_{i\in\mathcal{S}_A}\dot{m}_i(\hat{\boldsymbol{\beta}})^{\intercal}\boldsymbol{x}_i\big\}^{-1}$, $\hat{\mathbf{c}}_2^{\intercal} = \big[\sum_{i\in\mathcal{S}_A}\{1-\pi_i(\hat{\boldsymbol{\theta}}_{cv})\}/\pi_i(\hat{\boldsymbol{\theta}}_{cv})\{y_i - m_i(\hat{\boldsymbol{\beta}})\}\boldsymbol{x}_i^{\intercal}\big]\big[\sum_{i\in\hat{\mathcal{S}}_{B,1}}d_i^B\pi_i(\hat{\boldsymbol{\theta}}_{cv})\{1-\pi_i(\hat{\boldsymbol{\theta}}_{cv})\}\boldsymbol{x}_i\boldsymbol{x}_i^{\intercal}\big]^{-1}$.

Result (i) shows that the hybrid estimator $\hat{\mu}_{HYB}$ is approximately unbiased if the regression model is correctly specified or $\tau = 0$. The regression model assumption however is hard to check, and the issue of extrapolation is of concern for the subpopulation $\mathcal{U}_0$. So we suggest getting more information from $\mathcal{U}_0$ when τ is relatively large.

If budget permits, the conundrum of the positivity violation can be solved by using a second-phase sample. Assume that a second-phase sample $\mathcal{S}_{B,0}^{(2)}$, which contains measurements on y, is drawn from the sample $\hat{\mathcal{S}}_{B,0}$. Then we can consider the following model-assisted estimator of $\mu_{y,0}$,

$$\hat{\mu}_{0,SP} = \frac{1}{\hat{N}_0^B}\sum_{i\in\mathcal{S}_{B,0}^{(2)}}d_{2i}^B d_i^B\{y_i - m_i(\hat{\boldsymbol{\beta}}_{pl})\} + \frac{1}{\hat{N}_0^B}\sum_{i\in\hat{\mathcal{S}}_{B,0}}d_i^B m_i(\hat{\boldsymbol{\beta}}_{pl}),$$

where $\hat{\boldsymbol{\beta}}_{pl}$ is obtained from the pooled sample $\mathcal{S}_A \cup \mathcal{S}_{B,0}^{(2)}$, and d_{2i}^B is the design weights for sample $\mathcal{S}_{B,0}^{(2)}$ conditional on $\mathcal{S}_B$. The estimator $\hat{\mu}_{0,SP}$ is approximately unbiased irrespective

of the specification of the regression model, and has efficiency gain when the regression model is correctly specified. Finally, if we let $\hat{\mu}_{y,1} = \hat{\mu}_{1,DR}$ and $\hat{\mu}_{y,0} = \hat{\mu}_{0,SP}$, then the resulting estimator of μ_y under the deterministic mechanism is given by

$$\hat{\mu}_{SP} = \frac{1}{\hat{N}^B}\left[\sum_{i\in\mathcal{S}_A}\frac{y_i - m_i(\hat{\boldsymbol{\beta}}_{pl})}{\pi_i(\hat{\boldsymbol{\theta}}_{cv})} + \sum_{i\in\mathcal{S}^{(2)}_{B,0}} d^B_{2i}d^B_i\{y_i - m_i(\hat{\boldsymbol{\beta}}_{pl})\} + \sum_{i\in\mathcal{S}_B} d^B_i m_i(\hat{\boldsymbol{\beta}}_{pl})\right].$$

The estimator $\hat{\mu}_{SP}$ has the DR property, and it is also free of the extrapolation issue since sample $\mathcal{S}_A$ and sample $\mathcal{S}^{(2)}_{B,0}$ are used collectively to derive the regression model.

4.3.3 Extension to practical scenarios

The split-population approach is also a useful technique in analyzing practical violations of the positive assumption. There are two types of positivity violations in general, theoretical violation and *practical violation* (Petersen et al., 2012). Theoretical violation occurs when there exists some unit $i \in \mathcal{U}$ with $\pi^A_i = 0$, which is also the interest of this chapter. The practical violation generally refers to scenarios where some units have extremely small (near zero) propensity scores. Small propensity scores are as problematic as zero propensity scores. For instance, if units with small propensity score are drawn into sample $\mathcal{S}_A$, then IPW estimators could get highly inflated due to inverting these small propensity scores. The erratic behaviour of IPW estimators would be reflected on the large variance and the large finite sample bias.

Various procedures have been developed under practical violations to obtain stable IPW estimators. One type of strategy is to avoid extreme weights for $\mathcal{S}_A$ through specific weighting procedures, for example, Molina et al. (2019), Li et al. (2018), Zubizarreta (2015). Trimming is also a popular strategy, but usually at the cost of increased bias. For example, Crump et al. (2009) suggested to discard units with extreme propensity scores, so certain efficiency optimization can be achieved. Ma and Wang (2019) investigated asymptotic properties of trimmed IPW estimators, and suggested a trimming threshold

which achieves small mean squared errors.

However, it is still rather challenging to apply the above methods if sample $\mathcal{S}_A$ only contains a limited amount of units with small propensity score. We divide population $\mathcal{U}$ into subpopulations $\dot{\mathcal{U}}_1$ and $\dot{\mathcal{U}}_0$, where $\dot{\mathcal{U}}_0$ contains units with near zero propensity score, and $\dot{\mathcal{U}}_1$ consists of the rest of units. If $\dot{\mathcal{U}}_0$ takes a large portion of $\mathcal{U}$, but with few units being selected into sample $\mathcal{S}_A$, then the behaviour of sample $\mathcal{S}_A$ can hardly represent the behaviour of $\dot{\mathcal{U}}_0$ no matter which type of adjustment is applied. Data following this structure are very similar to those generated from the deterministic mechanism with $\Phi(\boldsymbol{x}_i) = \pi(\boldsymbol{x}_i,\boldsymbol{\theta})$. One may still consider the split-population approach and second-phase sampling as potential treatments for practical violations.

4.4 Simulation Studies

In this section, we conduct simulation studies to investigate performances of aforementioned inferential procedures when the sampling frame is not complete. The stochastic mechanism and the deterministic mechanism are adopted in Section 4.4.1 and Section 4.4.2 respectively.

4.4.1 Performance under the stochastic mechanism

We consider a finite population of size $N = 20{,}000$, with frame indicator variable z_i and two auxiliary variables $x_{1i} \sim N(0,1)$ and $x_{2i} \sim Exponential(mean = 1)$. Denote the proportion of the population uncovered by the frame by γ, which is calculated as $\gamma = N^{-1}\sum_{i=1}^{N}(1-z_i)$. We generate z_i from the Bernoulli distribution with mean value ψ_i, which follows logistic regression model,

$$\log\left(\frac{\psi_i}{1-\psi_i}\right) = \omega + 0.5x_{1i} + 0.5x_{2i}\,, \quad i = 1,2,\cdots,N\,,$$

where intercept ω is chosen such that $N^{-1}\sum_{i=1}^{N}\psi_i = 1-\gamma$.

For the response variable y, we consider a linear regression model (ξ),

$$y_i = 3 + x_{1i} - x_{2i} + \alpha z_i + \sigma\varepsilon_i\,, \quad i = 1,2,\cdots,N\,,$$

where α is the coefficient for covariate z which we will specify later.

The error terms ε_i's are generated independently from $N(0,1)$, and the value of σ is chosen such that $\rho = 0.5$, where ρ is the correlation coefficient between y and the linear predictor.

Sample $\mathcal{S}_A$ is selected by the Poisson sampling method from set $\breve{\mathcal{U}}_1 = \{i \in \mathcal{U} : z_i = 1\}$ with probabilities $\breve{\pi}_i^A$, which follow

$$\log\left(\frac{\breve{\pi}_i^A}{1-\breve{\pi}_i^A}\right) = \phi + 0.6x_{1i} - 0.3x_{2i}\,,$$

where intercept ϕ is set such that $\sum_{i\in\breve{\mathcal{U}}_1}\breve{\pi}_i^A = n_A$, and n_A is the target sample size of $\mathcal{S}_A$. Note $\breve{\pi}_i^A$ is not the same as the propensity score $\pi_i^A = P(R_i = 1 \mid x_{1i}, x_{2i})$.

The probability sample $\mathcal{S}_B$ with the target size n_B is drawn by the randomized systematic PPS sampling method. The inclusion probability π_i^B is proportional to $v_i = c + x_{2i}$, where the constant c is chosen to control the variation of the survey weights such that $\max v_i / \min v_i = 50$.

We fix the finite population and indicator variable z once generated, and repeatedly draw sample $\mathcal{S}_A$ with $n_A = 1{,}000$ and sample $\mathcal{S}_B$ with $n_B = 500$, for $B = 5{,}000$ times. For each simulation run, we compute the following estimators based on $\mathcal{S}_A$ and $\mathcal{S}_B$.

- The naive estimator based on the sample mean of $\mathcal{S}_A$, i.e., $\hat{\mu}_A = 1/n_A \sum_{i\in\mathcal{S}_A} y_i$.
- The regression type estimator $\hat{\mu}_{REG}$. The working model used for computation is linear regression model $m(\boldsymbol{x}_i,\boldsymbol{\beta}) = \beta_0+\beta_1 x_{1i}+\beta_2 x_{2i}$, and the parameter $\boldsymbol{\beta}$ is obtained by the least square method.

- The two IPW estimators $\hat{\mu}_{IPW1}$ and $\hat{\mu}_{IPW2}$. The working model we use to estimate propensity scores π_i^A is a single logistic regression model $\pi(\boldsymbol{x}_i,\boldsymbol{\theta}) = \left\{1+\exp(-\theta - \theta_1 x_{1i} - \theta_2 x_{2i})\right\}^{-1}$, and parameter $\boldsymbol{\theta}$ is estimated by the maximum pseudo likelihood method.

- The calibration type IPW estimator $\hat{\mu}_{IPW,cal}$. The working model considered for propensity scores is the same as that in the IPW approach, but the parameter is obtained from the calibration type estimating equations (4.2.4).

- The NN imputation based estimator $\hat{\mu}_{NN}$.

- The two estimators obtained by the KR-NN imputation. One has $\epsilon = 0.5$, and denoted by $\hat{\mu}_{KR,0.5}$. The other has $\epsilon = 0.1$, and denoted by $\hat{\mu}_{KR,0.1}$. The calibration step follows the typical Deville and Särndal (1992)'s calibration weighting method with the distance function being chosen as $\sum_{i\in\tilde{\mathcal{S}}_{B,1}}(w_i - d_i^B)^2/d_i^B$.

- The two estimators obtained by the stable weights method: one with $\boldsymbol{\delta} = (0.05, 0.05)^\intercal$, denoted by $\hat{\mu}_{SW,0.05}$; and the other with $\boldsymbol{\delta} = (0.01, 0.01)^\intercal$, denoted by $\hat{\mu}_{SW,0.01}$.

- A doubly robust estimator in the form of (2.4.14), denoted by $\hat{\mu}_{DR2}$, with two working models being specified by $\pi(\boldsymbol{x}_i,\boldsymbol{\theta})$ and $m(\boldsymbol{x}_i,\boldsymbol{\beta})$.

For a given estimator $\hat{\mu}$, its performance is evaluated through the $\%RB$ and MSE. We consider scenarios with different proportions of sampling frame by setting $\gamma = 0, 0.2$, and 0.4; and for each value of γ, we let parameter α in the model ξ take the value of 0, 0.5 and 1 respectively. When $\alpha = 0$, the covariates (x_1,x_2) are confounding variables for R and y, which is in line with the defined stochastic mechanism. When $\alpha \neq 0$, the covariates (x_1,x_2,z) are confounding variables for R and y, which is a more complicated process than the interested stochastic mechanism. In Table 4.1, the values of $(\mu_y,\mu_{y,1},\mu_{y,0})$ for each combination of γ and α are listed, which roughly indicate the discrepancy between

two subpopulations. The simulated performance of 11 point estimators are presented in Table 4.2, with some key observations being summarized as follows.

Table 4.1: Population and Subpopulation Means (μ_y,$\mu_{y,1}$,$\mu_{y,0}$) under the Stochastic Mechanism

	$\gamma = 0$	$\gamma = 0.2$	$\gamma = 0.4$
$\alpha = 0$	(2.01, NA, NA)	(2.01, 2.05, 1.86)	(2.01, 2.04, 1.97)
$\alpha = 0.5$	(2.51, NA, NA)	(2.41, 2.55, 1.86)	(2.31, 2.54, 1.97)
$\alpha = 1$	(3.01, NA, NA)	(2.82, 3.05, 1.85)	(2.61, 3.04, 1.98)

(1) The unadjusted sample mean $\hat{\mu}_A$ always has the largest bias among reported estimators. The rest of estimators, except for $\hat{\mu}_{SW,0.05}$, have comparable performance for each case, and demonstrate very similar trend under changes in α and γ. (2) Under the defined stochastic mechanism, i.e., $\alpha = 0$, all the approaches considered, expect for the stable weights method with $\boldsymbol{\delta} = (0.05, 0.05)^\intercal$, have acceptable performance in terms of both $\%RB$ and MSE. The IPW estimators $\hat{\mu}_{IPW1}$ and $\hat{\mu}_{IPW2}$ show negligible biases when $\alpha = 0$, which means the single logistic regression model is a reasonable fit to the two-stage sampling process. (3) When $\gamma = 0$, i.e., the sampling frame is complete, no discernible bias is observed for any of the adjusted estimators (except for $\hat{\mu}_{SW,0.05}$) considered. (4) When $\alpha \neq 0$ and $\gamma \neq 0$, bias emerges for every adjusted estimators; and there is an obvious pattern that bias grows with α and γ. But compared to $\hat{\mu}_A$, adjusted estimators still have dramatic improvement regarding $\%RB$ and MSE. (5) The performance of stable weights estimators and NN based estimators heavily depends on the specification of $\boldsymbol{\delta}$ and ϵ respectively. Stable weights estimator $\hat{\mu}_{SW,0.01}$ has much better performance than estimator $\hat{\mu}_{SW,0.05}$ in terms of $\%RB$ and MSE. For the NN based approach, we compare estimators $\hat{\mu}_{NN}$, $\hat{\mu}_{KR,0.5}$ and $\hat{\mu}_{KR,0.1}$, and found biases decrease with the value of ϵ.

4.4.2 Performance under the deterministic mechanism

We consider a finite population of size N = 20,000, with three auxiliary variables x_1, x_2 and x_3. Variables x_1 and x_2 are generated by the same way as in the previous section,

Table 4.2: Simulated $\%RB$ and $MSE_{\times 10^2}$ of Estimators of μ_y under the Stochastic Mechanism

		$\gamma = 0$		$\gamma = 0.2$		$\gamma = 0.4$	
	Estimator	$\%RB$	MSE	$\%RB$	MSE	$\%RB$	MSE
$\alpha = 0$	$\hat{\mu}_A$	38.81	61.64	41.27	69.63	40.83	68.13
	$\hat{\mu}_{REG}$	-0.50	1.42	0.19	1.49	-0.84	1.45
	$\hat{\mu}_{IPW1}$	-0.65	4.61	0.38	4.42	-0.67	4.01
	$\hat{\mu}_{IPW2}$	-1.80	4.44	-0.42	4.25	-0.85	4.09
	$\hat{\mu}_{IPW,cal}$	-0.12	1.49	0.78	1.59	-0.06	1.55
	$\hat{\mu}_{NN}$	1.21	3.79	2.17	4.08	1.47	3.88
	$\hat{\mu}_{KR,0.5}$	0.58	3.67	1.33	4.38	0.97	5.54
	$\hat{\mu}_{KR,0.1}$	0.05	3.76	0.72	4.65	0.00	6.09
	$\hat{\mu}_{SW,0.05}$	4.65	2.00	5.47	2.40	4.46	1.96
	$\hat{\mu}_{SW,0.01}$	0.69	1.19	1.47	1.32	0.56	1.20
	$\hat{\mu}_{DR2}$	-0.10	1.54	0.79	1.61	-0.09	1.57
$\alpha = 0.5$	$\hat{\mu}_A$	31.08	61.64	38.49	87.02	44.31	105.45
	$\hat{\mu}_{REG}$	-0.40	1.42	4.24	2.56	8.02	4.88
	$\hat{\mu}_{IPW1}$	-0.25	4.84	4.60	5.98	8.25	7.96
	$\hat{\mu}_{IPW2}$	-1.45	4.44	3.75	5.09	8.02	7.53
	$\hat{\mu}_{IPW,cal}$	-0.10	1.49	4.74	2.90	8.70	5.63
	$\hat{\mu}_{NN}$	0.97	3.79	5.90	6.01	10.04	9.29
	$\hat{\mu}_{KR,0.5}$	0.46	3.67	5.20	5.99	9.61	10.60
	$\hat{\mu}_{KR,0.1}$	0.04	3.76	4.69	6.03	8.76	10.38
	$\hat{\mu}_{SW,0.05}$	3.73	2.00	8.64	5.57	12.64	9.70
	$\hat{\mu}_{SW,0.01}$	0.55	1.19	5.32	2.90	9.24	5.77
	$\hat{\mu}_{DR2}$	-0.08	1.54	4.75	2.93	8.68	5.63
$\alpha = 1$	$\hat{\mu}_A$	25.92	61.64	36.51	106.44	46.99	150.95
	$\hat{\mu}_{REG}$	-0.33	1.42	7.14	5.61	14.83	16.49
	$\hat{\mu}_{IPW1}$	0.01	5.21	7.63	9.87	15.14	20.45
	$\hat{\mu}_{IPW2}$	-1.21	4.44	6.73	7.93	14.86	19.21
	$\hat{\mu}_{IPW,cal}$	-0.08	1.49	7.58	6.22	15.46	17.95
	$\hat{\mu}_{NN}$	0.81	3.79	8.58	10.04	16.65	23.09
	$\hat{\mu}_{KR,0.5}$	0.39	3.67	7.97	9.70	16.28	24.18
	$\hat{\mu}_{KR,0.1}$	0.04	3.76	7.54	9.52	15.51	23.19
	$\hat{\mu}_{SW,0.05}$	3.11	2.00	10.92	10.72	18.93	25.63
	$\hat{\mu}_{SW,0.01}$	0.46	1.19	8.07	6.47	15.92	18.54
	$\hat{\mu}_{DR2}$	-0.07	1.54	7.59	6.25	15.44	17.92

and variable x_3 follows Bernoulli distribution with mean of 0.5. The response variable y_i follows the regression model,

$$y_i = 3 + x_{1i} + x_{2i} + x_{3i} - \eta x_{1i}^2 + \sigma\varepsilon_i\,, \quad i = 1,2,\cdots,N\,,$$

where η is the coefficient for covariate x_1^2. Error terms ε_i's and the value of σ are generated by the same way as in previous section such that $\rho = 0.5$ for the response model.

Propensity scores follow the deterministic mechanism, and the underlying values used for truncation are given by,

$$\log\left(\frac{\Phi_i}{1-\Phi_i}\right) = 1 - 0.6x_{1i} + 0.5x_{2i} + 0.8x_{3i}\,, \quad i = 1,2,\cdots,N\,.$$

Let $Q(\tau)$ denote the τth percentile of $\{\Phi_1,\cdots,\Phi_N\}$. If $\Phi_i \leq Q(\tau)$, then set $\pi_i^A = 0$; if $\Phi_i > Q(\tau)$, then generate π_i^A from model

$$\log\left(\frac{\pi_i^A}{1-\pi_i^A}\right) = \theta + 0.3x_{1i} - 0.3x_{2i} + 0.5x_{3i}\,,$$

where the intercept θ is chosen such that $\sum_{i=1}^N \pi_i^A = n_A$. Sample $\mathcal{S}_A$ is selected by the Poisson sampling method with inclusion probabilities specified by π_i^A, and $\mathcal{S}_B$ is selected by the same strategies as described in the previous section.

The 11 point estimators described in the last section, plus the estimators $\hat{\mu}_{HYB}$ and $\hat{\mu}_{SP}$ are investigated. For estimators $\hat{\mu}_{IPW1}$, $\hat{\mu}_{IPW2}$, $\hat{\mu}_{IPW,cal}$, $\hat{\mu}_{DR2}$, we adopt the logistic regression model $\pi(\boldsymbol{x}_i,\boldsymbol{\theta}) = \left\{1+exp(-\theta_0-\theta_1x_{1i}-\theta_2x_{2i}-\theta_3x_{3i})\right\}^{-1}$ for the propensity scores, but do not take the truncation step into consideration. For estimators $\hat{\mu}_{REG}$, $\hat{\mu}_{DR2}$ and $\hat{\mu}_{HYB}$, the model $m(\boldsymbol{x}_i,\boldsymbol{\beta}) = \beta_0 + \beta_1x_{1i} + \beta_2x_{2i} + \beta_3x_{3i}$ is adopted for the outcome regression, i.e., covariate x_{1i}^2 is missing in the working model. The misspecified model is considered here since the specification of the outcome regression is especially challenging under the deterministic mechanism. For the proposed hybrid estimator $\hat{\mu}_{HYB}$, model $\pi(\boldsymbol{x}_i,\boldsymbol{\theta})$ and $m(\boldsymbol{x}_i,\boldsymbol{\beta})$ are considered as working models, and sample $\mathcal{S}_B$ is split by the convex hull method. To

calculate estimator $\hat{\mu}_{SP}$, a second-phase sample $\mathcal{S}_{B,0}^{(2)}$ is obtained by SRS without replacement, and the sample size is set to the 20% of the size of $\hat{\mathcal{S}}_{B,0}$. Both the propensity score model and the prediction model are correctly specified under the second-phase sampling approach.

We compare these approaches under several scenarios, and in particular consider $\tau = 0, 0.2, 0.4$, and $\eta = 0, 0.5, 1$. Values of $(\mu_y,\mu_{y,1},\mu_{y,0})$ are listed in Table 4.3 for different combinations of τ and η.

Table 4.3: Population and Subpopulation Means $(\mu_y,\mu_{y,1},\mu_{y,0})$ under the Deterministic Mechanism

	$\tau = 0$	$\tau = 0.2$	$\tau = 0.4$
$\eta = 0$	(4.53, NA, NA)	(4.53, 4.49, 4.72)	(4.53, 4.53, 4.52)
$\eta = 0.5$	(4.03, NA, NA)	(4.03, 4.06, 3.91)	(4.03, 4.06, 3.97)
$\eta = 1$	(3.52, NA, NA)	(3.52, 3.63, 3.11)	(3.52, 3.59, 3.42)

Simulation results of point estimators with n_A = 1,000, n_B = 500 and B = 5,000, are reported in Table 4.4, and based on which we have following observations. (1) The naive estimator $\hat{\mu}_A$ has relatively large biases in all the scenarios. (2) When $\eta = 0$, estimators dependent on the regression prediction show small $\%RB$. The reason is that the working model $m(\boldsymbol{x}_i,\boldsymbol{\beta})$ is a correctly specified model for the outcome regression when $\eta = 0$. Some noticeable bias can be observed for the estimator $\hat{\mu}_{IPW1}$, especially when $\tau = 0.4$; while $\hat{\mu}_{IPW2}$ has surprisingly good performance. (3) When $\tau = 0$ and $\eta \neq 0$, estimators which rely on the propensity score model have promising performance, while $\hat{\mu}_{REG}$, $\hat{\mu}_{SW,0.05}$, $\hat{\mu}_{SW,0.01}$, $\hat{\mu}_{KR,0.5}$ and $\hat{\mu}_{KR,0.1}$ have biases which increase with the value of η. (4) In general, performance of all the estimators deteriorate as τ and η increase. Estimator $\hat{\mu}_{SP}$ is an exception since it gains the DR property by using additional information from $\mathcal{S}_{B,0}^{(2)}$. (5) The proposed hybrid estimator has relatively small $\%RB$ across all the scenarios, and its performance is especially robust when $\tau \neq 0$. (6) Similarly to results in Table 4.2, we still found that the stable weights estimator $\hat{\mu}_{SW,0.01}$ has better performance than $\hat{\mu}_{SW,0.05}$. For the NN based approach, estimator $\hat{\mu}_{KR,0.1}$ which has the most precise matching criteria

however has the worst performance when $\eta \neq 0$. This is an opposite pattern to what we observe from the stochastic mechanism.

Moreover, we evaluate the convex hull method by its classification accuracy (AC), which is defined as,

$$AC = \frac{\sum_{\hat{\mathcal{S}}_{B,0}} I(i \in \mathcal{U}_0) + \sum_{\hat{\mathcal{S}}_{B,1}} I(i \in \mathcal{U}_1)}{n_B} \times 100.$$

Given $\tau = 0, 0.2$ and 0.4, simulated AC of the convex hull method is 97.54, 97.88 and 98.29 respectively. To examine its performance under smaller sizes for $\mathcal{S}_A$, we also conduct a simulation with $n_A = 500$, and the new setting gives corresponding results of 95.44, 96.1 and 96.9. So the convex hull method generally works better with a larger sample size, but still has acceptable performance when the sample size is fairly small.

Lastly, we briefly check the performance of the variance estimator v_{HYB}. Its $\%RB$ by comparing with the Monte-carlo simulated variance, and $\%CP$ of its associated Wald-type 95% CIs are reported in Table 4.5. It can be observed that the $\%RB$ of v_{HYB} is relatively small for all the cases considered. The coverage rates of associated Wald-type CIs are close to the nominal value when either $\eta = 0$ or $\tau = 0$, but the under-coverage issue rises when $\eta \neq 0$ and $\tau \neq 0$ due to the biased point estimator $\hat{\mu}_{HYB}$ (see Table 4.4).

4.5 Discussion

In the current chapter, we have investigated two mechanisms for incomplete frames. It can be summarized from the limited simulation studies that making inferences with non-probability samples is possible under the incomplete frame, but it is particularly challenging if the positivity assumption is violated. In fact, the mechanism of the positivity violation in practice can be far more complicated than the simple process we consider. Therefore, we can expect issues with conventional procedures with more severe departures from scenarios we discussed.

The issue of zero propensity scores should be tackled in a systematic manner. The

Table 4.4: Simulated $\%RB$ and $MSE_{\times 10^2}$ of Estimators of μ_y under the Deterministic Mechanism

	Estimator	$\tau = 0$		$\tau = 0.2$		$\tau = 0.4$	
		$\%RB$	MSE	$\%RB$	MSE	$\%RB$	MSE
$\eta = 0$	$\hat{\mu}_A$	17.23	61.90	18.47	70.99	21.79	98.48
	$\hat{\mu}_{REG}$	-0.03	1.55	-0.04	1.45	0.36	1.70
	$\hat{\mu}_{IPW1}$	0.22	3.01	-1.43	2.68	-9.68	20.88
	$\hat{\mu}_{IPW2}$	-0.09	1.67	0.00	1.39	0.66	1.64
	$\hat{\mu}_{IPW,cal}$	0.00	1.58	-0.01	1.49	0.49	2.37
	$\hat{\mu}_{NN}$	0.12	4.15	0.22	5.55	2.72	10.32
	$\hat{\mu}_{KR,0.5}$	0.02	4.14	-0.24	5.12	1.04	9.17
	$\hat{\mu}_{KR,0.1}$	-0.06	4.84	-0.10	5.33	0.57	9.07
	$\hat{\mu}_{SW,0.05}$	1.28	1.54	2.34	2.10	1.31	1.86
	$\hat{\mu}_{SW,0.01}$	0.24	1.29	0.35	1.19	0.54	1.86
	$\hat{\mu}_{DR2}$	-0.01	1.59	0.00	1.47	0.39	1.75
	$\hat{\mu}_{HYB}$	0.00	1.59	0.04	1.48	0.42	1.72
	$\hat{\mu}_{SP}$	NA	NA	-0.03	4.83	0.08	7.14
$\eta = 0.5$	$\hat{\mu}_A$	16.91	47.40	23.56	91.03	27.82	126.59
	$\hat{\mu}_{REG}$	2.82	2.88	3.71	3.81	4.83	5.75
	$\hat{\mu}_{IPW1}$	0.16	2.90	1.96	2.93	-6.54	8.67
	$\hat{\mu}_{IPW2}$	-0.13	2.19	3.44	3.60	4.16	4.77
	$\hat{\mu}_{IPW,cal}$	0.01	1.91	3.85	4.10	5.79	8.11
	$\hat{\mu}_{NN}$	0.69	5.13	2.26	7.63	6.02	16.67
	$\hat{\mu}_{KR,0.5}$	1.03	5.20	3.49	8.25	6.86	18.65
	$\hat{\mu}_{KR,0.1}$	4.78	9.52	5.82	11.94	7.80	20.99
	$\hat{\mu}_{SW,0.05}$	3.48	3.22	6.32	7.64	6.38	8.43
	$\hat{\mu}_{SW,0.01}$	2.59	2.44	4.24	4.25	5.78	7.56
	$\hat{\mu}_{DR2}$	-0.10	2.02	3.39	3.57	4.71	5.73
	$\hat{\mu}_{HYB}$	-0.02	2.03	2.14	2.57	3.66	4.31
	$\hat{\mu}_{SP}$	NA	NA	-1.10	6.61	-0.65	9.39
$\eta = 1$	$\hat{\mu}_A$	16.53	35.44	30.17	114.40	35.68	159.43
	$\hat{\mu}_{REG}$	6.48	7.58	8.53	11.14	10.63	16.81
	$\hat{\mu}_{IPW1}$	0.09	3.56	6.33	7.76	-2.48	2.99
	$\hat{\mu}_{IPW2}$	-0.19	3.34	7.87	10.08	8.70	12.19
	$\hat{\mu}_{IPW,cal}$	0.03	2.83	8.82	11.99	12.69	23.80
	$\hat{\mu}_{NN}$	1.41	8.02	4.81	13.33	10.43	30.02
	$\hat{\mu}_{KR,0.5}$	2.30	8.40	8.26	18.08	14.59	43.35
	$\hat{\mu}_{KR,0.1}$	10.99	23.84	13.41	32.07	17.21	53.89
	$\hat{\mu}_{SW,0.05}$	6.32	6.86	11.43	17.93	12.95	23.46
	$\hat{\mu}_{SW,0.01}$	5.61	5.95	9.24	12.51	12.57	22.74
	$\hat{\mu}_{DR2}$	-0.23	3.35	7.76	9.84	10.35	16.39
	$\hat{\mu}_{HYB}$	-0.04	3.45	4.85	5.67	7.91	10.96
	$\hat{\mu}_{SP}$	NA	NA	-2.49	11.79	-1.60	16.17

Table 4.5: Simulated $\%RB$ of the Variance Estimator v_{HYB}.

	$\tau = 0$		$\tau = 0.2$		$\tau = 0.4$	
	$\%RB$	$\%CP$	$\%RB$	$\%CP$	$\%RB$	$\%CP$
$\eta = 0$	6.57	95.60	5.01	95.60	5.44	95.34
$\eta = 0.5$	5.05	95.60	3.66	90.04	5.77	83.58
$\eta = 1$	3.88	95.06	2.39	81.14	5.71	66.78

primary step is to check if the positivity assumption is met. The proposed split-population method is a useful technique to help understand the data structure and check the potential violation of the positivity assumption. In particular, the convex hull classification method, although developed for the deterministic mechanism, can be viewed as a diagnostic tool to examine if sample $\mathcal{S}_A$ and sample $\mathcal{S}_B$ have enough overlap. One can also refer to external studies to understand zero propensity scores. For example, as internet users grow rapidly, there is an increasing amount of national surveys aiming at the online behaviour of the population. The following list contains a series of questions found among these surveys.

- The frequency of using internet to express opinions about political or community issues within the last 12 months (Current Population Survey, Civic Engagement Supplement, 2013);
- Whether or not the adult uses the internet or email and the frequency at which they are used (National Health Interview Survey, 2017);
- Internet use in the past 30 days (Behavioural Risk Factor Surveillance System, 2017);
- Internet access and frequency of use (National Household Education Survey, 2016).

These questions are extremely useful in indicating the characteristics of web-survey participants, and shed light on the differences between internet users and non-users.

After preliminary investigations, specific estimation strategies can be derived based on the newly acquired knowledge about samples and the population. The proposed hybrid estimator is an example of choosing estimation procedures according to the re-categorized

sample, and the resulting estimator shows robustness in the simulation studies considered. When the issue of positivity violation is severe, the second-phase sampling method is a possible remedy. Even though taking a second-phase sample is subject to extra costs, both theory and simulation studies show that it can improve inferential results by getting crucial data of the response variable from the subpopulation $\mathcal{U}_0$.

4.6 Technical Details

Proof of Theorem 4.1.

If the asymptotic expansion of $\hat{\mu}_{HYB}$ is available, then results (i) and (ii) of Theorem 4.1 can be easily proved. We expand $\hat{\mu}_{HYB}$ and derive V_{HYB} through following steps. Step 1 proves that $\hat{\boldsymbol{\theta}}_{cv} = \tilde{\boldsymbol{\theta}} + o_p(n_A^{-\frac{1}{2}})$; Step 2 finds asymptotic expressions of estimated model parameters $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\theta}}_{cv}$; Step 3 uses the linearization method to deal with variations from multiple sources; Step 4 obtains variance formula through the final asymptotic expression.

Step 1:

For the proposed hybrid estimator $\hat{\mu}_{HYB}$, its estimated model parameter $\hat{\boldsymbol{\theta}}_{cv}$ is obtained by solving equation

$$\sum_{i\in\mathcal{S}_A} \boldsymbol{x}_i - \sum_{i\in\hat{\mathcal{S}}_{B,1}} d_i^B \pi_i(\boldsymbol{\theta})\boldsymbol{x}_i = 0. \tag{4.6.9}$$

By comparing (4.6.9) to (4.3.7), it is easy to observe that the extra variation is induced by identifying set $\hat{\mathcal{S}}_{B,1}$. Our first goal is to show that this additional variation is negligible, i.e., $\hat{\boldsymbol{\theta}}_{cv} = \tilde{\boldsymbol{\theta}} + o_p(n^{-\frac{1}{2}})$, where $\tilde{\boldsymbol{\theta}}$ is the solution to (4.3.7). To prove this equivalency, it suffices to show

$$\frac{1}{N_1}\left\{ \sum_{i\in\hat{\mathcal{S}}_{B,1}} d_i^B \pi_i(\boldsymbol{\theta}_0)\boldsymbol{x}_i - \sum_{i\in\mathcal{S}_{B,1}} d_i^B \pi_i(\boldsymbol{\theta}_0)\boldsymbol{x}_i \right\} = o_p(n_A^{-\frac{1}{2}}).$$

This result is immediately implied by the following derivations if Condition **C8** holds uniformly over all i:

$$
\begin{aligned}
&\frac{1}{N_1}\sum_{i\in\mathcal{S}_A}\boldsymbol{x}_i-\frac{1}{N_1}\sum_{i\in\hat{\mathcal{S}}_{B,1}}d_i^B\pi_i(\boldsymbol{\theta}_0)\boldsymbol{x}_i\\
=&\frac{1}{N_1}\sum_{i\in\mathcal{S}_A}\boldsymbol{x}_iI(\boldsymbol{x}_i\in C_{n_A})-\frac{1}{N_1}\sum_{i\in\mathcal{S}_{B,1}}d_i^B\pi_i(\boldsymbol{\theta}_0)\boldsymbol{x}_iI(\boldsymbol{x}_i\in C_{n_A})\\
=&\frac{1}{N_1}\sum_{i\in\mathcal{S}_A}\boldsymbol{x}_i\big\{I(\boldsymbol{x}_i\in\boldsymbol{R}_1)+o_p(1)\big\}-\frac{1}{N_1}\sum_{i\in\mathcal{S}_{B,1}}d_i^B\pi_i(\boldsymbol{\theta}_0)\boldsymbol{x}_i\big\{I(\boldsymbol{x}_i\in\boldsymbol{R}_1)+o_p(1)\big\}\\
=&\frac{1}{N_1}\Big\{\sum_{i\in\mathcal{S}_A}\boldsymbol{x}_iI(\boldsymbol{x}_i\in\boldsymbol{R}_1)-\sum_{i\in\mathcal{S}_{B,1}}d_i^B\pi_i(\boldsymbol{\theta}_0)\boldsymbol{x}_iI(\boldsymbol{x}_i\in\boldsymbol{R}_1)\Big\}+\frac{1}{N_1}\Big\{\sum_{i\in\mathcal{S}_A}\boldsymbol{x}_i-\sum_{i\in\mathcal{S}_{B,1}}d_i^B\pi_i(\boldsymbol{\theta}_0)\boldsymbol{x}_i\Big\}o_p(1)\\
=&\frac{1}{N_1}\Big\{\sum_{i\in\mathcal{S}_A}\boldsymbol{x}_i-\sum_{i\in\mathcal{S}_{B,1}}d_i^B\pi_i(\boldsymbol{\theta}_0)\boldsymbol{x}_i\Big\}+o_p(n_A^{-\frac{1}{2}}).
\end{aligned}
$$

Step 2:

According to estimating equations in (4.3.8) and by the linearization technique, we have

$$\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*=M_1\Big[\sum_{i\in\mathcal{S}_A}\big\{y_i-m_i(\boldsymbol{\beta}^*)\big\}\boldsymbol{x}_i\Big]+o_p(n_A^{-\frac{1}{2}}),$$

where $M_1=\Big\{\sum_{i\in\mathcal{U}_1}\pi_i^A\dot{m}_i(\boldsymbol{\beta}^*)^\intercal\boldsymbol{x}_i\Big\}^{-1}$. Similarly, based on the result from Step 1, the estimator $\hat{\boldsymbol{\theta}}_{cv}$ can be written as

$$\hat{\boldsymbol{\theta}}_{cv}-\boldsymbol{\theta}_0=M_2\Big(\sum_{i\in\mathcal{S}_A}\boldsymbol{x}_i-\sum_{i\in\mathcal{S}_{B,1}}d_i^B\pi_i^A\boldsymbol{x}_i\Big)+o_p(n_A^{-\frac{1}{2}}),$$

where $M_2=\Big\{\sum_{i\in\mathcal{U}_1}\pi_i^A(1-\pi_i^A)\boldsymbol{x}_i\boldsymbol{x}_i^\intercal\Big\}^{-1}$.

Step 3:

We investigate the variation of $\hat{\mu}_{HYB}$ induced by different sources. We first consider the

estimated parameter $\hat{\boldsymbol{\beta}}$, and rewrite $\hat{\mu}_{HYB}$ as,

$$\hat{\mu}_{HYB} = \frac{1}{\hat{N}^B}\left\{ \sum_{i\in\mathcal{S}_A} \frac{y_i - m_i(\hat{\boldsymbol{\beta}})}{\pi_i(\hat{\boldsymbol{\theta}}_{cv})} + \sum_{i\in\mathcal{S}_{B,1}} d_i^B m_i(\hat{\boldsymbol{\beta}}) + \sum_{i\in\mathcal{S}_{B,0}} d_i^B m_i(\hat{\boldsymbol{\beta}}) \right\},$$

By the similar argument for Theorem 2.2, we can show that the variation of estimating $\boldsymbol{\beta}$ is negligible for the first two components of above expression. The estimator $\hat{\mu}_{HYB}$ can be further simplified to

$$\hat{\mu}_{HYB} = \frac{1}{\hat{N}^B}\left\{ \sum_{i\in\mathcal{S}_A} \frac{y_i - m_i(\boldsymbol{\beta}^*)}{\pi_i(\hat{\boldsymbol{\theta}}_{cv})} + \sum_{i\in\mathcal{S}_{B,1}} d_i^B m_i(\boldsymbol{\beta}^*) + \sum_{i\in\mathcal{S}_{B,0}} d_i^B m_i(\hat{\boldsymbol{\beta}}) \right\} + o_p(n_A^{-\frac{1}{2}}). \tag{4.6.10}$$

The third component of (4.6.10) is the only one which contains $\hat{\boldsymbol{\beta}}$, and the expansion of this component is given by

$$\frac{1}{\hat{N}^B}\sum_{i\in\mathcal{S}_{B,0}} d_i^B m_i(\hat{\boldsymbol{\beta}}) = \frac{1}{\hat{N}^B}\sum_{i\in\mathcal{S}_{B,0}} d_i^B m_i(\boldsymbol{\beta}^*) + \mathbf{c}_1^\intercal\left[\frac{1}{\hat{N}^B}\sum_{i\in\mathcal{S}_A}\left\{y_i - m_i(\boldsymbol{\beta}^*)\boldsymbol{x}_i\right\}\right] + o_p(n_A^{-\frac{1}{2}}),$$

where $\mathbf{c}_1^\intercal = \left\{\sum_{i\in\mathcal{U}_0} \dot{m}_i(\boldsymbol{\beta}^*)^\intercal\right\} M_1$.

Next, we examine the first component of (4.6.10), which contains the variation coming from computing $\hat{\boldsymbol{\theta}}_{cv}$. By Taylor expansion, we have

$$\begin{aligned}\frac{1}{\hat{N}^B}\left\{ \sum_{i\in\mathcal{S}_A} \frac{y_i - m_i(\boldsymbol{\beta}^*)}{\pi_i^A} \right\} =& \frac{1}{\hat{N}^B}\left\{ \sum_{i\in\mathcal{S}_A} \frac{y_i - m_i(\boldsymbol{\beta}^*)}{\pi_i^A} \right\} \\ & - \mathbf{c}_2^\intercal\left(\frac{1}{\hat{N}^B}\sum_{i\in\mathcal{S}_A}\boldsymbol{x}_i - \frac{1}{\hat{N}^B}\sum_{i\in\mathcal{S}_{B,1}} d_i^B\pi_i^A\boldsymbol{x}_i\right) \\ & + o_p(n_A^{-\frac{1}{2}}),\end{aligned} \tag{4.6.11}$$

where $\mathbf{c}_2^\intercal = \left[\sum_{i\in\mathcal{U}_1}(1-\pi_i^A)\{y_i - m_i(\boldsymbol{\beta}^*)\}\boldsymbol{x}_i^\intercal\right] M_2$.

Finally, we apply the linearization technique again to deal with the variation contributed

by the probability sampling, and obtain the following asymptotic expansion,

$$\hat{\mu}_{HYB} = (2 - \frac{\hat{N}^B}{N})\frac{1}{N}\sum_{i\in\mathcal{S}_A} s_i + f_N + \frac{1}{N}\sum_{i\in\mathcal{S}_B} d_i^B(f_i - f_N) + o_p(n_A^{-\frac{1}{2}}),$$

where $s_i = \{y_i - m_i(\boldsymbol{\beta}^*)\}/\pi_i^A - \mathbf{c}_2^\intercal \boldsymbol{x}_i + \mathbf{c}_1^\intercal\big\{y_i - m_i(\boldsymbol{\beta}^*)\big\}\boldsymbol{x}_i$, $f_i = m_i(\boldsymbol{\beta}^*) + \pi_i^A \mathbf{c}_2^\intercal \boldsymbol{x}_i I(\boldsymbol{x}_i \in \boldsymbol{R}_1)$, and $f_N = N^{-1}\sum_{i=1}^N f_i$.

Step 4:

Based on the above expansion, we can easily obtain asymptotic variance formula V_{HYB} such that $Var(\hat{\mu}_{HYB}) = V_{HYB} + o_p(n_A^{-1})$. Specifically, V_{HYB} has the form of

$$V_{HYB} = \frac{1}{N^2}\sum_{i=1}^N \pi_i^A(1-\pi_i^A)s_i^2 + Var_p\bigg\{\frac{1}{N}\sum_{i\in\mathcal{S}_B} d_i^B\Big(f_i - f_N - \frac{1}{N}\sum_{i\in\mathcal{S}_A} s_i\Big)\bigg\},$$

where the first component is the variance attributed to the selection mechanism for sample $\mathcal{S}_A$, and the second component is resulted from the probability sampling for $\mathcal{S}_B$ and its exact formula depends on the specific strategy of taking $\mathcal{S}_B$.

Chapter 5

Discussion and Future Work

5.1 Summary

Researchers frequently encounter the dilemma that pertinent data is not available while obtaining new data is inefficient and expensive. Facing this challenge, more and more researchers choose to incorporate data from unconventional sources into their project. Among various novel data sources, web-based non-probability surveys have received the most attention and have rapidly become one of the most exciting topics in the area. However, inferences made from web-based survey samples and other non-probability samples are frequently questioned due to the absence of a mature theoretical framework.

My thesis establishes a general framework for statistical inferences with non-probability survey samples when relevant auxiliary information is available from a probability survey sample. Under this setup, discussions are made upon: model assumptions, data integration, inferential methods and applications on different data types. Moreover, the potential issue of zero propensity scores is highlighted and further investigated.

The starting point of the thesis is to adjust the intrinsic selection bias and make valid inferences with web-based and other non-probability survey samples. In particular, a rigorous procedure for estimating propensity scores is proposed in Chapter 2. While existing

methods of estimating propensity scores are largely ad hoc, our method gives a class of estimators which are consistent under commonly used assumptions. This is an important contribution to the area since estimating propensity scores is the most fundamental step of the QR approach. Doubly robust estimation is another major topic. Several DR point estimators and associated variance formulas are given in Chapter 2, which can be immediately applied to real data analyses.

In Chapter 3, we consider the use of PEL approach with non-probability survey samples. While the PEL approach and the QR approach are comparable to some extent, we found that PEL approach has more advantages in certain scenarios. For example, it has more flexible structures to utilize auxiliary information, which leads to multiple robustness naturally. Coupled with the model-calibration technique, PEL approach is also a robust way to estimate distribution functions and quantiles. Moreover, PEL-ratio-based CIs show relatively stable performance in the simulation studies considered, and outperform Wald-type CIs under the scenario of our interest.

In Chapter 4, two mechanisms for incomplete frames, namely, stochastic mechanism and deterministic mechanism, are investigated. We show that the positivity assumption holds under the stochastic mechanism, which means the discussed QR, MI and PEL approach can be directly applied. Meanwhile, zero propensity scores occur under the deterministic mechanism, which raises inferential challenges for aforementioned approaches. To construct more robust estimators under the deterministic mechanism, we suggest a two-step split-population approach, which is carried out by (1) dividing the population into subpopulations by zero and non-zero propensity scores, and (2) choosing suitable estimating method based on the features of the two subpopulations. At the end of the chapter, we also conduct a series of simulation studies to evaluate some popular procedures under incomplete frames. The results further reveal issues of ignoring zero propensity scores when using non-probability samples for inferences.

5.2 Extensions and Future Directions

In the thesis, we mainly focus on the estimation of finite population means under the context of non-probability survey samples. As a matter of fact, our work can be potentially extended for broader uses in a wider range of settings. A few interesting directions are listed below.

5.2.1 Applications to multiple data sources problem

Online activities leave digital traces, and the nature of these traces is data. As more and more activities have being moved online, researchers in either the traditional area like survey sampling or the modern field such as machine learning, all face the same challenge of extracting useful information from different data sources.

The idea of constructing pseudo likelihood functions given in (2.3.3) and estimating equations given in (2.3.7) by using multiple data sources can be easily extended beyond the survey data context. In particular, our idea enjoys two features which could be meaningful to many research topics.

Firstly, data linkage is not required. Data linkage is a popular strategy to combine multiple datasets, which is often performed before statistical analysis. The gist of data linkage is to link the records from different datasets if they belongs to the same entity. The link among datasets is usually the ID of the entity, for example, sample $\mathcal{S}_A$ and sample $\mathcal{S}_B$ can be merged by entity if both datasets contain entities' phone number. But different datasets often use different items as ID, which means the extra information is required to link IDs first. Moreover, IDs which are critical for the data linkage, such as name, account number, IP address are often removed for confidentiality. Our method bypasses the data linkage step, and can effectively cope with independently generated and unpaired datasets.

Secondly, by replacing unknown population quantities with sample based estimators, less of data needs to be gathered for our proposed method than a typical classification

method. This technique is appealing to many practical situations. Consider the following hypothetical scenario. A medical researcher aims to build a model to predict if a person has the disease or not. Assume the database of positive cases is fairly complete, but no data have been collected for negative cases. By convention, researchers need obtain the data of all the negative cases in the population to build a logistic model. However, our method could largely save time and costs by only requiring a sample of the population. This technique is also a potential treatment to the computational difficulty attributed to the large data size. Specifically, if the computational complexity of some quantity grows with the data size substantially, then we can draw a sample from the original data and compute the sample based estimator corresponding to the quantity of interest.

5.2.2 Model and variable selections

In non-probability survey samples, there are not many researches available for selecting the model of propensity scores. A popular method of model selection is comparing the adjusted auxiliary information with benchmark auxiliary information. For example, a small discrepancy between $\sum_{i\in\mathcal{S}_A} \boldsymbol{x}_i/\hat{\pi}_i$ and $\sum_{i\in\mathcal{S}_B} d_i^B \boldsymbol{x}_i$ is interpreted as the indication of an adequate candidate propensity score model, where $\hat{\pi}_i$ is the estimated propensity score based on the candidate model. This method is not applicable if model parameters are obtained from calibration type estimating equation (4.2.4), where the discrepancy is always forced to be zero. Moreover, since the gap between adjusted values and benchmark values can often be decreased by adding more relevant covariates in the model, this method is inclined to arrive at over-fitted and less efficient models. Results given in Table 2.9 is a clear manifestation of this issue. There are other suggestions on the covariate selection for the propensity score model, for example, including all the covariates related to either the outcome variable or the selection mechanism, only including covariates related to the outcome variable and the selection mechanism simultaneously, including covariates only related to the outcome variable, including covariates only related to the selection mechanism, etc.

The pseudo likelihood function we build in (2.3.3) provides a different angle of the

model selection. Since the pseudo likelihood function is available, we can compute AIC and BIC and other criteria for the model selection. Moreover, the error matrix (also known as confusion matrix in the field of machine learning), a tool often used to assess the performance of a classification model, can also be approximated. We use the inclusion in sample $\mathcal{S}_A$ as an example. Let $\hat{R}_i$ be the predicted status of inclusion for unit i, based on the candidate model. If the estimated propensity score for unit i is larger than 0.5, then $\hat{R}_i = 1$; otherwise $\hat{R}_i = 0$. The approximated error matrix is given below.

	Predicted 1	Predicted 0
Actual 1	$\sum_{i\in\mathcal{S}_A} I(\hat{R}_i = 1)$	$\sum_{i\in\mathcal{S}_A} I(\hat{R}_i = 0)$
Actual 0	$\sum_{i\in\mathcal{S}_B} d_i^B I(\hat{R}_i = 1) - \sum_{i\in\mathcal{S}_A} I(\hat{R}_i = 1)$	$\sum_{i\in\mathcal{S}_B} d_i^B I(\hat{R}_i = 0) - \sum_{i\in\mathcal{S}_A} I(\hat{R}_i = 0)$

Given these evaluation metrics, many other model selection methods become applicable. For example, k-fold cross validation can be conducted by dividing each datasets into k subsamples; LASSO regression, also being used to prevent overfitting, can be built upon pseudo likelihood functions.

When computing DR estimators, we also need to postulate a prediction model. Since prediction regression model is built by a complete case analysis, many existing model selection methods are already available. According to Assumption **A1**, only covariates which are related to both outcome variable and selection mechanism need to be included in the analysis. So we can conduct variable selection for two models individually, and then only use the common covariates which are selected in the both procedures to obtain final models.

5.2.3 Non-ignorable selection mechanism

Our previous work focuses on the ignorable selection mechanism, i.e., given covariates, selection mechanism does not depend on the response variable. However, this ignorability condition may not hold in practice. Consider the following hypothetical scenario.

There is a self-selection web survey posted at a website, which aims to collect some income information of site viewers. For each survey participant, a cash incentive will be offered once survey is completed. However, we can make an educated guess that site viewers with high income would be less motivated by the cash reward, and be more cautious about their confidentiality. In this situation where the study variable highly relates to the response model and no confounding variables available to entirely explain their relation, our previous approaches are no longer effective for bias adjustment.

Contrary to ignorable mechanism, the scenario describes a non-ignorable missing data problem. Formally, under the non-ignorable mechanism, the propensity scores are given by

$$\pi_i^A = P(R_i = 1 \mid \boldsymbol{x}_i, y_i), \text{ for } i = 1, \cdots N,$$

which is a function of y, and possibly of $\boldsymbol{x}$, and relation $P(R_i = 1 \mid \boldsymbol{x}_i, y_i) = P(R_i = 1 \mid \boldsymbol{x}_i)$ for $i = 1, \cdots N$, does not hold.

In general missing data context, a few approaches have been developed to tackle non-ignorable selection mechanism. See Liu et al. (2020) for a comprehensive review about existing works. But related researches are still very limited for non-probability samples. Nevertheless, we found that several treatments which were derived for the general missing data are promising for the adaptation to the current setting. For instance, methods in Wang et al. (2014) and Ai et al. (2018) depend on the specially designed estimating equations, and corresponding estimating equations can be constructed for non-probability samples by using techniques of obtaining equation (2.3.7), i.e., replacing unknown information with reference sample based estimators.

5.3 Outlook on Future Development

In spite of the rising applications of non-probability survey samples, there are still many unexplored aspects about this topic. In the process of our development, we found at least three broad areas which call for more investigations.

The first area is the relation between non-probability survey samples and probability survey samples. While the goal of our thesis is to analyze non-probability survey samples, the journey of the research however, reveals and confirms the irreplaceable role of representative datasets. As internet has wider and wider coverage, many datasets are generated by online activities. These datasets are mostly non-probability based, so adjustments are often required to infer the larger population. According to our work in previous chapters, the existing adjustments all rely on the benchmark information provided by some representative datasets such as probability sample and census. How to conduct traditional probability samples and what information to be gathered are important topics to study in order to take advantage of modern data sources.

The second area is the relation between non-probability survey samples and the general missing data problem. Non-probability survey samples contain the data of participants only, which obviously belongs to missing data problem. While there are already many techniques developed to deal with missingness for general missing data, it is sensible to borrow and adapt these techniques to the current context. However, possibly due to the difference between the design-based framework and the independent and identically distributed random variables assumption, as well as the unique two-sample setup in non-probability survey samples, this kind of extensions have been rarely made. We believe further explorations on this subject are especially meaningful for data analysts in the current field since the extensions directly lead to a richer set of analytic tools for non-probability samples.

The third area is the relation between non-probability survey samples and the modern data sources. As data can be collected more and more easily through internet, analysts are facing an increasing amount of large datasets or the so-called big data. Under the belief that the more the better, these large datasets are often treated as quality data which contain the unbiased information of the larger population. However, if data itself is generated by a non-randomized mechanism, more data does not make unrepresentative datasets more representative (Meng, 2018). It is a practically important task to identify the hidden applications of non-probability survey samples in these novel fields, and advocate

the adjustment methods we derive in the current context.

References

Ai, C., Linton, O., and Zhang, Z. (2018). A simple and efficient estimation method for models with nonignorable missing data. *Statistica Sinica*. In press.

Antal, E. and Tillé, Y. (2011). A direct bootstrap method for complex sampling designs from a finite population. *Journal of the American Statistical Association*, 106:534–543.

Baker, R., Brick, J. M., Bates, N. A., Battaglia, M., Couper, M. P., Dever, J. A., Gile, K. J., and Tourangeau, R. (2013). Report of the aapor task force on non-probability sampling. *Journal of Survey Statistics and Methodology*, 1:90–143.

Beaumont, J. F. (2005). Calibrated imputation in surveys under a quasi-model-assisted approach. *Journal of the Royal Statistical Society: Series B*, 67:445–458.

Berger, Y. G. (2004). A simple variance estimator for unequal probability sampling without replacement. *Journal of Applied Statistics*, 31:305–315.

Berger, Y. G. and Torres, O. D. L. R. (2016). Empirical likelihood confidence intervals for complex sampling designs. *Journal of the Royal Statistical Society: Series B*, 78:319–341.

Brewer, K. R. W. and Donadio, M. E. (2003). The high entropy variance of the Horvitz-Thompson estimator. *Survey Methodology*, 29:189–196.

Brick, J. M. (2015). Compositional model inference. *Proceedings of the Survey Research Methods Section, Joint Statistical Meetings, American Statistical Association*, pages 299–307.

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*, 96:723–734.

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*, 63:615–620.

Chen, H. and Chen, J. (2000). Bahadur representations of the empirical likelihood quantile processes. *Journal of Nonparametric Statistics*, 12:645–660.

Chen, J. and Qin, J. (1993). Empirical likelihood estimation for finite populations and the effective usage of auxiliary information. *Biometrika*, 80:107–116.

Chen, J. and Shao, J. (2000). Nearest neighbor imputation for survey data. *Journal of Official Statistics*, 16:113–131.

Chen, J. and Sitter, R. R. (1999). A pseudo empirical likelihood approach to the effective use of auxiliary information in complex surveys. *Statistica Sinica*, 9:385–406.

Chen, J. and Wu, C. (2002). Estimation of distribution function and quantiles using the model-calibrated pseudo empirical likelihood method. *Statistica Sinica*, 12:1223–1239.

Chen, J. K. T., Valliant, R., and Elliott, M. R. (2019). Calibrating non-probability surveys to estimated control totals using lasso, with an application to political polling. *Journal of the Royal Statistical Society: Series C*, 68:657–681.

Chen, S. and Haziza, D. (2017). Multiply robust imputation procedures for the treatment of item nonresponse in surveys. *Biometrika*, 104:439–453.

Chen, S. and Kim, J. K. (2014). Population empirical likelihood for nonparametric inference in survey sampling. *Statistica Sinica*, 24:335–355.

Chen, X., Hong, H., and Tarozzi, A. (2008). Semiparametric efficiency in GMM models with auxiliary data. *The Annals of Statistics*, 36:808–843.

Cheng, P. E. and Chu, C. K. (1996). Kernel estimation of distribution functions and quantiles with missing data. *Statistica Sinica*, 6:63–78.

Citro, C. F. (2014). From multiple modes for surveys to multiple data sources for estimates. *Survey Methodology*, 40:137–161.

Couper, M. P. (2000). Review: web surveys: a review of issues and approaches. *The Public Opinion Quarterly*, 64:464–494.

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

Deville, J. C. and Särndal, C. E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87:376–382.

Elliott, M. R. and Valliant, R. (2017). Inference for nonprobability samples. *Statistical Science*, 32:249–264.

Godambe, V. P. and Thompson, M. E. (2009). Estimating functions and survey sampling. *Handbook of Statistics: Sample Surveys: Inference and Analysis*, 29B:83–101.

Goodman, R. and Kish, L. (1950). Controlled selection -a technique in probability sampling. *Journal of the American Statistical Association*, 45:350–372.

Graham, B., Pinto, C., and Egel, D. (2016). Efficient estimation of data combination models by the method of auxiliary-to-study tilting (AST). *Journal of Business & Economic Statistics*, 34:288–301.

Han, P. (2014). Multiply robust estimation in regression analysis with missing data. *Journal of the American Statistical Association*, 109:1159–1173.

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

Hansen, M. H. (1987). Some history and reminiscences on survey sampling. *Statistical Science*, 2:180–190.

Hartley, H. O. and Rao, J. N. K. (1962). Sampling with unequal probabilities and without replacement. *Annals of Mathematical Statistics*, 33:350–374.

Haziza, D. and Rao, J. N. K. (2006). A nonresponse model approach to inference under imputation for missing survey data. *Survey Methodology*, 32:53–64.

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

Isaksson, A. and Forsman, G. (2003). A comparison between using the web and using telephone to survey political opinions. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pages 100–106.

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

Kim, J. K. and Haziza, D. (2014). Doubly robust inference with missing data in survey sampling. *Statistica Sinica*, 24:375–394.

Kim, J. K. and Kim, J. J. (2007). Nonresponse weighting adjustment using estimated probability. *The Canadian Journal of Statistics*, 35:501–514.

Kim, J. K. and Park, H. A. (2006). Imputation using response probability. *The Canadian Journal of Statistics*, 34,:171–182.

Kim, J. K., Park, S., Chen, Y., and Wu, C. (2018). Combining non-probability and probability survey samples through mass imputation. *arXiv preprint arXiv: 1812.10694*.

Kim, J. K. and Rao, J. N. K. (2012). Combining data from two independent surveys: a model-assisted approach. *Biometrika*, 99:85–100.

Kim, J. K. and Rao, J. N. K. (2018). Data integration for big data analysis in finite population inference. *Talk at SSC2018, Montreal*.

Kim, J. K. and Riddles, M. K. (2012). Some theory for propensity-score-adjustment estimators in survey sampling. *Survey Methodology*, 38:157–165.

Kim, J. K. and Shao, J. (2013). *Statistical Methods for Handling Incomplete Data*. Chapman & Hall/CRC, New York.

Kim, J. K. and Wang, Z. (2019). Sampling techniques for big data analysis. *International Statistical Review*, 87:177–191.

Kitamura, Y. (2007). *Empirical Likelihood Methods in Econometrics: Theory and Practice*. Cambridge University Press, Cambridge.

Kott, P. S. (1994). A Note on handling nonresponse in sample surveys. *Journal of the American Statistical Association*, 89:693–696.

Lee, S. (2006). Propensity score adjustment as a weighting scheme for volunteer panel web surveys. *Journal of Official Statistics*, 22:329–349.

Lee, S. and Valliant, R. (2009). Estimation for volunteer panel web surveys using propensity score adjustment and calibration adjustment. *Sociological Methods & Research*, 37:319–343.

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

Little, R. J. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. Wiley, New York, second edition.

Liu, Y., Li, P., and Qin, J. (2020). Full-semiparametric-likelihood-based inference for non-ignorable missing data. *Statistica Sinica*. In press.

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

Ma, X. and Wang, J. (2019). Robust inference using inverse probability weighting. *Journal of the American Statistical Association*. In press.

Meng, X.-L. (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*, 12:685–726.

Molina, J., Sued, M., Valdora, M., and Yohai, V. (2019). Robust doubly protected estimators for quantiles with missing data. *TEST*. In press.

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*, 97:558–625.

Owen, A. B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, 75:237–249.

Owen, A. B. (2001). *Empirical Likelihood*. CRC Press, London.

Petersen, M. L., Porter, K. E., Wang, Y., and van der Laan, M. J. (2012). Diagnosing and responding to violations in the positivity assumption. *Statistical Methods in Medical Research*, 21:31–54.

Rafei, A., Flannagan, C. A. C., 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*, 8:148–180.

Rao, J. N. K. (2005). Interplay between sample survey theory and practice: an appraisal. *Survey Methodology*, 31:117–138.

Rao, J. N. K. and Wu, C. (2010). Bayesian pseudo-empirical-likelihood intervals for complex surveys. *Journal of the Royal Statistical Society: Series B*, 72:533–544.

Rao, J. N. K. and Wu, C. F. J. (1988). Resampling inference with complex survey data. *Journal of the American Statistical Association*, 83:231–241.

Rao, J. N. K., Wu, C. F. J., and Yue, K. (1992). Some recent work on resampling methods for complex surveys. *Survey Methodology*, 18:209–217.

Rivers, D. (2007). Sampling for web surveys. *Proceedings of the Survey Research Methods Section, Joint Statistical Meetings, American Statistical Association*, pages 1–26.

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

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

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63:581–592.

Rueda, M. and Muñoz, J. F. (2009). New model-assisted estimators for the distribution function using the pseudo empirical likelihood method. *Statistica Neerlandica*, 63:227–244.

Särndal, C. E. (1980). On π-inverse weighting versus best linear unbiased weighting in probability sampling. *Biometrika*, 67:639–650.

Särndal, C. E., Swensson, B., and Wretman, J. H. (1992). *Model Assisted Survey Sampling*. Springer-Verlag, New York.

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*, 94:1096–1120.

Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. Wiley Series in Probability and Statistics, New York.

Shao, J. and Tu, D. (1996). *The Jackknife and Bootstrap*. Springer, New York.

Shu, H. and Tan, Z. (2020). Improved methods for moment restriction models with data combination and an application to two-sample instrumental variable estimation. *The Canadian Journal of Statistics*, 48:259–284.

Tan, Z. (2007). Comment: understanding OR, PS and DR. *Statistical Science*, 22:560–568.

Terhanian, G. and Bremer, J. (2000). Confronting the selection-bias and learning effects problems associated with Internet research. *Research paper, Harris Interactive.*

Tourangeau, R., Conrad, F. G., and Couper, M. P. (2013). *The Science of Web Surveys*. Oxford University Press, New York, first edition.

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

Valliant, R. and Dever, J. A. (2011). Estimating propensity adjustments for volunteer web surveys. *Sociological Methods & Research*, 40:105–137.

Vavreck, L. and Rivers, D. (2008). The 2006 cooperative congressional election study. *Journal of Elections, Public Opinion and Parties*, 18:355–366.

Wang, Q. and Qin, Y. (2010). Empirical likelihood confidence bands for distribution functions with missing responses. *Journal of Statistical Planning and Inference*, 140:2778–2789.

Wang, S., Shao, J., and Kim, J. K. (2014). An instrumental variable approach for identification and estimation with nonignorable nonresponse. *Statistics Sinica*, 24:1097–1116.

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

Wu, C. (2003). Optimal calibration estimators in survey sampling. *Biometrika*, 90:935–951.

Wu, C. and Lu, W. W. (2016). Calibration weighting methods for complex surveys. *International Statistical Review*, 84:79–98.

Wu, C. and Rao, J. N. K. (2006). Pseudo-empirical likelihood ratio confidence intervals for complex surveys. *The Canadian Journal of Statistics*, 34:359–375.

Wu, C. and Rao, J. N. K. (2010). Bootstrap procedures for the pseudo empirical likelihood method in sample surveys. *Statistics and Probability Letters*, 80:1472–1478.

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*, 96:185–193.

Yang, S. and Kim, J. K. (2018). Predictive mean matching imputation in survey sampling. *arXiv preprint arXiv: 1703.10256*.

Yang, S. and Kim, J. K. (2020). Statistical data integration in survey sampling: a review. *arXiv preprint arXiv: 2001.03259*.

Yang, S., Kim, J. K., and Song, R. (2020). Doubly robust inference when combining probability and non-probability samples with high dimensional data. *Journal of the Royal Statistical Society: Series B*, 82:445–465.

Zhang, L.-C. (2019). On valid descriptive inference from non-probability sample. *Statistical Theory and Related Fields*, 3:103–113.

Zhang, S., Han, P., and Wu, C. (2019). Empirical likelihood inference for non-randomized pretest-posttest studies with missing data. *Electronic Journal of Statistics*, 13:2012–2042.

Zhou, M. (2015). *Empirical Likelihood Method in Survival Analysis*. CRC Press, New York.

Zubizarreta, J. R. (2015). Stable weights that balance covariates for estimation with incomplete outcome data. *Journal of the American Statistical Association*, 110:910–922.