Correcting Selection Bias in Nonprobability Samples by Pseudo Weighting

Proefschrift ter verkrijging van de graad van doctor aan Tilburg University op gezag van de rector magnificus, prof. dr. W.B.H.J. van de Donk, in het openbaar te verdedigen ten overstaan van een door het college voor promoties aangewezen commissie in de Aula van de Universiteit op maandag 16 juni 2025 om 13.30 uur

door

An-Chiao Liu,

geboren te Taipei, Taiwan

Promotores:	prof. dr. A. G. de Waal	(Tilburg University)
	prof. dr. K. Van Deun	(Tilburg University)
Leden promotiecommissie:	prof. dr. J.K. Vermunt	(Tilburg University)
	prof. dr. J.A. van den Brakel	(Maastricht University)
	dr. B. Struminskaya	(Utrecht University)
	dr. G. Vink	(Utrecht University)

Contents

Chapter 1

Introduction

When we try to answer a question about a population, for example, how happy the Dutch PhD students are, we often do not have information about all the units in the population, but only for a sample of it. The representativeness of the sample is then crucial to make a reliable inference of the population. To ensure representativeness, the gold standard of drawing a sample is in a probabilistic manner. That is, we first define the scope of the population we are interested in such as PhD students enrolled in a Dutch university in 2024. The probability of each unit being included in the sample is then decided based on the research question and perhaps also the amount of resources the researcher has. The inclusion probability can be the same for every unit, as in simple random sampling, or the probability can be specified given some unit characteristics, as in probability proportional to size. The problem of drawing a probability sample based on different needs or limitations has been well-studied in the past decades. See, for example, Cochran (1977); Kish (1965). The acquired sample is then called a *probability sample*.

To estimate the target estimand from the probability sample, such as the population total or mean of the target variable, a series of estimators can be used. For instance, for simple random sampling, the sample mean is a well-known and often-used estimator for the population mean.

Here we introduce some notation for the following discussion. Suppose we are interested in a finite population U of size N. The target variable, or study variable, Y is expressed for unit $i \in \{1, 2, 3, ..., N\}$ as y_i and we are interested in the population mean of the target variable $\mu = N^{-1} \sum_{i \in U} y_i$. If the probability sample (S) of size n is obtained by simple random sampling, μ is often estimated by $\bar{y}_{SRS} = n^{-1} \sum_{i \in S} y_i$. If the probability sample is from an unequal probability design, that is, $\mathbb{P}(i \in S)$ is not the same for every i, μ may be estimated by design-based estimators such as the Horvitz-Thompson estimator $\bar{y}_{HT} = N^{-1} \sum_{i \in S} w_i y_i$ or Hájek estimator $\bar{y}_w = \sum_{i \in S} w_i y_i / \sum_{i \in S} w_i$, where the design weight $w_i = 1/\mathbb{P}(i \in S)$ (Hájek, 1971; Horvitz & Thompson, 1952).

Although probability sampling generally guarantees the unbiasedness and/or consistency of statistical inference, it has limitations as well. The collection of a probability sample is usually time and resource-consuming, and response rates to surveys have been decreasing over the last few decades. At the same time, a large amount of *nonprobability samples*, samples for which the inclusion mechanism is unknown to the researcher, are available or can be easily collected. Think of data scraped from social media, sensor data, or administrative data. If the nonprobability sample is treated as a simple random sample, selection bias/error may occur. Specifically, if μ is estimated by the nonprobability sample (NP) with $\bar{y}_{NP} = n_{NP}^{-1} \sum_{i \in NP} y_i$ where n_{NP} is the size of the nonprobability sample, the error can be expressed as

$$\text{Error}(\bar{y}_{NP}) = \bar{y}_{NP} - \mu.$$

If the nonprobability sample is from an underlying sampling mechanism (b), the bias is then

$$\text{Bias}(\bar{y}_{NP}) = E_b(\bar{y}_{NP}) - \mu$$

where $E(\cdot)$ is an expectation function. Also, since nonprobability samples are usually easy to collect, we often have a large n_{NP}. If we again treat the nonprobability sample as a simple random sample, the resulting estimated variance is small and the confidence interval can barely cover the true value (Meng, 2018).

To avoid the threat of selection bias, a correction is needed before inferring from a nonprobability sample. Correcting selection bias in a nonprobability sample is certainly not a new topic, but it draws lots of attention from statisticians nowadays given the emerging data sources and new analysis techniques. Many influential statistical journals have published special issues on nonprobability samples or closely related topics, such as the Journal of Survey Methodology and Statistics, Survey Methodology, and Journal of Official Statistics. In this dissertation, we propose a selection bias correction method for large nonprobability samples in terms of inclusion fraction $f_{NP} = n_{NP}/N$. Some practical analysis issues such as model selection, imbalanced samples, and small area estimations are also discussed.

1.1 Correcting Selection Bias by Pseudo Weights

One of the often-used methods for selection bias correction is pseudo-weighing. It assigns a set of unit weights to the nonprobability sample so that the target estimand can be estimated by design-based estimators. The weights may be constructed by assuming the nonprobability sample comes from probability sampling (with unknown inclusion probabilities) and trying to estimate the inclusion probabilities $\mathbb{P}(i \in NP)$ by utilizing some auxiliary information. Ideally, if the weights are properly constructed, the resulting estimators are unbiased or consistent no matter which target variable is of interest. Most pseudo-weighting methods are based on the following assumptions:

1. The nonprobability sample is from the target population U.

2. The inclusion probability for every unit of the population is positive. That is, $\mathbb{P}(i \in NP) > 0$ for every $i \in U$.

3. The auxiliary information (x) governs the inclusion mechanism of the nonprobability sample so that $\mathbb{P}(i \in NP|x, y) = \mathbb{P}(i \in NP|x)$.

The first assumption may be violated by overcoverage. For example, if the nonprobability sample is collected from social media, people who specified themselves from the Netherlands may not be from the Netherlands. The second assumption is necessary for applying design-based estimators after the pseudo-weights are constructed. This assumption may not always be valid. For example, people who do not have an account on the social medium used to collect the data for the nonprobability sample will have a zero chance of being included. The third assumption is similar to the Missing At Random assumption in missing data literature (Rubin, 1978). In practice, researchers may apply auxiliary variables that are believed to be related to the inclusion mechanism of the nonprobability sample or related to the target variables, but it is hard to verify whether the assumption is met or not.

Many methods for constructing the pseudo-weights have been proposed in the literature. Elliott and Valliant (2017) give a review of selection bias correction and discuss one of the pseudo-weighting methods originally proposed in Elliott and Davis (2005). The method assumes the inclusion fraction of the nonprobability sample is small so that the nonprobability sample and the reference probability sample do not have any overlapping parts. This may be valid in large populations such as the United States. In the Netherlands, many registers cover a large proportion of the population, and this assumption may be easily violated. Therefore, in Chapter 2 we propose a pseudo-weighting framework for large inclusion fractions. The proposed method requires a reference probability sample that shares some auxiliary variables with the nonprobability sample but does not necessarily have the target variable.

This is referred to as a *two-sample setup*. A propensity model is then fitted given the non-overlapping part of the two samples. We also discuss different relations between the nonprobability sample and the reference probability sample. That is, the inclusion mechanism of the nonprobability sample may be affected by the reference probability sample because of, for instance, the response burden. A nonparametric variance estimation framework is also proposed to estimate the uncertainty of the estimated population mean under various kinds of propensity models.

1.2 Model Selection for Weight Construction

The proposed framework in Chapter 2 and many proposed methods in the literature allow researchers to fit all kinds of models to estimate the inclusion probability of the nonprobability sample. Many machine learning techniques for probability estimation can then be used. However, machine learning techniques often rely on a proper model selection process, for example cross-validation, to guarantee predictive performance. Those model selection processes may not be immediately useful for selection bias correction since they mainly focus on prediction problems. Here we borrow the terminology from Efron (2020). That is, many model selection techniques try to find a model with a minimum Mean Squared Error on the unit level

$$\text{MSE}(\hat{y}) = E_b[n^{-1}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2].$$

The estimation problem, on the other hand, focuses on having the smallest MSE of the estimated population parameter (e.g., the estimated population mean $\bar{y}$) and it tries to find a model with minimum

$$\text{MSE}(\bar{y}) = E_b[(\bar{y} - \mu)^2].$$

Of course in a perfect scenario, we may find a model that returns the smallest MSE($\hat{y}$) and MSE($\bar{y}$) at the same time, but this is not always the case in practice.

Given that the fundamental aims of prediction and estimation are different, in Chapter 3 we try to find a suitable performance measure that can indicate the performance of the constructed weights. Since constructing weights may involve a propensity estimation step, we examined the model performance measures which are often used for probability estimation to see whether they are useful for selection bias correction. Some possible measures developed from related literature such as calibration are discussed as well.

1.3 Imbalanced Samples

Under the two-sample setup, since the nonprobability sample is often much larger than the probability sample, their union results in an imbalanced dataset when fitting a propensity model. Imbalanced samples often cause modeling difficulties, and many approaches have been proposed to deal with it. For example, undersampling the majority group (i.e., nonprobability sample) or oversampling the minority group (i.e., probability sample). Instead of merely under- or oversampling, some other methods such as the Synthetic Minority Oversampling Technique (SMOTE) are widely used in machine learning literature. SMOTE tries to create a sample that is similar to the minority group and is shown to be effective in machine learning literature. However, these approaches cannot immediately be used for selection bias correction since the estimation steps in Chapter 2 need to be adjusted accordingly. Therefore, in Chapter 4, we discuss the adjustment of these approaches and try to find the ideal way to deal with the imbalance in the two-sample setup.

1.4 Small Area Estimation Based on a Nonprobability Sample

Besides estimating the population mean or total, often the estimate of a certain subgroup is of interest as well. As in the example at the beginning, we may be interested in not only how happy the Dutch PhD students are, but also how happy the Dutch PhD students in each department are. Since the sample size of each department may not be large enough to draw a reliable inference, some adjustments are needed. This type of problem is referred to as *Small Area Estimation* (SAE). The massive literature on SAE is based on probability samples, see, for example, reviews in Parker, Janicki, and Holan (2023); Pfeffermann (2013), while the discussion of doing SAE on nonprobability samples is limited. When doing SAE on a nonprobability sample, some differences should be taken into account. For example, the design weights of the probability sample are replaced by the constructed pseudo-weights of the nonprobability sample. The weights are therefore subject to the variation of the propensity model. Also, as in the difference between $\text{MSE}(\hat{y})$ and $\text{MSE}(\bar{y})$, a model that returns the smallest MSE in small areas may not necessarily return the smallest $\text{MSE}(\bar{y})$. If both SAE and overall estimates are important to the researcher, some decisions should be made during the modeling process.

1.5 Outline

In the following chapters, the proposed selection bias correction framework for large inclusion fractions is introduced in Chapter 2. A discussion of suitable performance measures for selection bias correction is given in Chapter 3. The proposed remedy for unbalanced samples is in Chapter 4. The approaches for small area estimation from a nonprobability sample are in Chapter 5. All of these chapters are supported

by simulation studies and applications of real-life data. Finally, the dissertation is concluded with some discussion and suggestions for practitioners.

Chapter 2

Correcting Selection Bias in Big Data by Pseudo Weighting

Abstract

Nonprobability samples, for example observational studies, online opt-in surveys, or register data, do not come from a sampling design and therefore may suffer from selection bias. To correct for selection bias, Elliott and Valliant (2017) (EV) proposed a pseudo-weight estimation method which applies a two-sample setup. That is, a set-up where other than the target nonprobability sample, a probability sample that shares some common auxiliary variables with the nonprobability sample is used. By estimating the propensities of inclusion in the nonprobability sample given the two samples, we may correct the selection bias by (pseudo) design-based approaches.

However, the EV method is not suitable for large inclusion fractions of the population or for units having high inclusion probabilities for either sample, which is often seen in administrative data sets and more and more common for Big Data. In this research, we extend the EV method to be suitable for all ranges of inclusion probabilities, while retaining the attractive properties of the original study. Any

model that is suitable for propensity estimation can be easily applied, for instance, a machine learning model.

Furthermore, the possible dependency between the selection of the nonprobability sample and the probability sample is discussed, to deal with the scenario where inclusion in the nonprobability sample is affected by being included in the probability sample. For variance estimation, two finite population bootstrap algorithms are proposed that account for the two-sample setup. These algorithms can be applied to a wide range of estimators and estimation methods. We show in a simulation study based on a real data set that the proposed method outperforms other comparing methods, and that the pseudo population bootstrap algorithms give reasonable variance estimates.

Keywords: Selection Bias, Big Data, Nonprobability Sample, Propensity Score, Pseudo Population Bootstrap

Liu, A.-C., Scholtus, S., & De Waal, T. (2023). Correcting selection bias in big data by pseudo-weighting. *Journal of Survey Statistics and Methodology*, 11(5), 1181-1203. https://doi.org/10.1093/jssam/smad042

2.1 Introduction

2.1.1 Nonprobability and Probability Samples

Probability sampling according to a well-designed sampling design, enables one to obtain valid estimates for population parameters of interest. However, the collection of probability samples is time-consuming and expensive. On the contrary, a wide diversity of new data sources, for example, Big Data, register data, and opt-in online surveys, can provide a massive amount of information at a low cost within a short time (Baker et al., 2013; Beaumont, 2020; Cornesse et al., 2020). These nonprobability samples do not come from a known sampling design, and therefore, it is uncertain whether we can make an unbiased inference from the nonprobability sample to the population of interest.

2.1.2 Selection Bias

Since the inclusion mechanism of a nonprobability sample is unknown, estimates based on a nonprobability sample may suffer from selection bias because under- or over-representation of certain units in the sample is unknown. Treating the nonprobability sample as if collected by simple random sampling may result in biased estimates even when the sample size is large. In fact, as noted in Bethlehem (2010); Meng (2018); J. Rao (1966) and Kim and Wang (2019), when the nonprobability sample is treated as a simple random sample from the population, selection bias occurs when the inclusion in the sample is correlated to the target variable we are interested in.

2.1.3 Existing Approaches

To solve the issue of selection bias, some approaches focus on estimating the target variables or parameters at the population level, while other approaches focus on estimating the inclusion probabilities of the units in the nonprobability sample. The two approaches can be combined to achieve doubly robust estimation (Y. Chen, Li, & Wu, 2020). For reviews of existing methods, we refer to Elliott and Valliant (2017), J. Rao (2020), and Cornesse et al. (2020). No matter which kind of approach is chosen, auxiliary information of the population is necessary, such as the full auxiliary information for the entire population (Heckman, 1976), population totals of the auxiliary variables (Little, West, Boonstra, & Hu, 2020), or a probability sample that shares some common auxiliary variables with the nonprobability sample. The last situation is termed as the two-sample setup. In this setup it is assumed that the two samples come from the same population we are interested in and the probability sample is a (relatively) good representation of the population we are interested in as long as the design weights are considered. By combining the information from the two samples, we may be able to correct for the selection bias in the nonprobability sample. This two-sample setup may allow researchers to have more data sources to choose from compared to methods requiring auxiliary information for the full population. For example, variables from an attitude questionnaire can be used as auxiliary information, which are usually not available on the population level. Many researchers also apply the two-sample setup, for example, Y. Chen et al. (2020); Elliott and Valliant (2017); Kim and Wang (2019); Valliant and Dever (2011); and Valliant (2020).

One approach that can be applied to the two-sample setup is a pseudo-weighting approach proposed by Elliott and Valliant (2017) which we will refer to as the EV method. The EV method assumes there exists an underlying (unknown) stochastic

mechanism for units to be included into the nonprobability sample, and the selection bias can be corrected by estimating the propensity of the units to be included in the nonprobability sample. The nonprobability sample is then treated as having been obtained from an unequal probability sampling design as known from the sampling literature. For example, one may use the Horvitz-Thompson estimator to obtain estimates for the population mean, or apply a weighted regression (Horvitz & Thompson, 1952).

The EV method is appealing for four reasons. First, when more than one target variable or the relation of multiple variables is of interest, which is often the case in practice (Haziza & Beaumont, 2017), only one propensity model needs to be fitted. Even though we may trade off with efficiency, estimating only one propensity model is less cumbersome than the approaches that model every target variable individually. The second reason is that the EV method handles the design weights of the probability sample *after* the propensity model is fitted, which tends to make the propensity estimation more stable compared to methods that consider the design weights during the propensity estimation (Gelman, 2007; Little, 2004). The post-processing of the design weights leads to the third appealing point. Any method focusing on probability estimation of binary categories, such as machine learning methods, can easily be plugged into the EV method in order to estimate propensity scores without considering the design weights during the model fitting process. For example Rafei, Flannagan, and Elliott (2020) apply Bayesian additive regression trees as the estimation model to mitigate the effect of model misspecification. Fourth, the EV method can also serve as the inclusion propensity estimation part to achieve doubly robust estimation (Y. Chen et al., 2020).

2.1.4 The EV Method in Big Data

Unfortunately, the EV method is not suitable for large inclusion probabilities, while large inclusion probabilities are often found in Big Data, or simply in a small population of interest, e.g., all the listed enterprises in a certain industry. When an administrative data set suffers from incompleteness due to delayed reporting, it can also be seen as a nonprobability sample with a large inclusion fraction. For example, in the simulation study in Section 2.4, we use the records of odometers of all privately-owned cars in the Netherlands from the Dutch Online Kilometer Registration. The record is entered when a car visits a garage for maintenance, repair, or testing. When the records are used to produce road traffic statistics, some cars may have not visited a garage yet, which results in a nonprobability sample (Buelens, Burger, & van den Brakel, 2018). Also, when some units have large propensities of being included in both samples, even when the overall inclusion fraction of the nonprobability sample is relatively small, the original EV method is not suitable. The latter situation can occur for business surveys, where large inclusion fractions are often used in strata containing larger enterprises (Ouwehand & Schouten, 2014).

In this research, we extend the EV method so that it not only enjoys the merits of the original method, but is also suitable for large inclusion fractions. Our extension assumes a unique linkage key for units in both samples to identify the overlapping part between the nonprobability sample and the probability sample. The unique key can be license plate numbers in the above example, personal IDs, or company registration numbers. If a unique key is not available, the overlapping part can also be estimated by auxiliary variables as in probabilistic record linkage, or by the method in Kim and Tam (2020).

We also address the issue of the possible dependency between the inclusion in the nonprobability sample and the inclusion in the probability sample. When the

probability sample is drawn before the nonprobability sample, the inclusion in the probability sample may affect the inclusion in the nonprobability sample. For example, businesses in the Netherlands are obliged to fill in a short-term business statistics panel survey when asked (probability sample) and are also always obliged to report data to the value-added tax (VAT) register. Since some businesses are late reporters, the available VAT data for producing early estimates of short-term statistics are incomplete and can be considered a nonprobability sample that is likely to be selective (Ouwehand & Schouten, 2014). When businesses are drawn into the panel survey, some of them might also be more likely to become an early VAT reporter, since they already have the data at hand. This would create an unintended dependency between the two samples. Dependency of the two samples may also be introduced intentionally in order to reach groups that are underrepresented in a probability survey. For example, one may try to increase the precision of estimates for minority groups by snowball sampling, and a screening procedure could be used to give the units already in the probability sample a lower (or zero) probability to be included in the nonprobability sample. We will show in the simulation study that this dependency of inclusion is important to be considered during the estimation process.

Variance estimation by resampling is discussed given the two-sample setup. Since we are focusing on large inclusion fractions of the samples from a finite population, the assumption of independent and identically distributed observations in the standard bootstrap method is not met (Efron, 1979). We apply algorithms extended from pseudo population bootstrapping, which can be applied for any estimator which is a smooth function of population means, for example, a total, a ratio, or a coefficient of linear regression. Also, the algorithms enjoy the generality that many kind of sampling design of the probability sample and any kind of propensity estimation process, e.g., machine learning methods, can be applied in this framework.

In the next section, we will first introduce the original approach by Elliott and Valliant (2017) and then propose our extended approach for the cases where the two samples are independent, respectively dependent. Section 2.3 discusses variance estimation algorithms. A simulation study comparing the proposed method and other methods is given in Section 2.4, and Section 2.5 concludes the paper with a discussion.

2.2 Methods

2.2.1 The EV Method

Following Elliott and Valliant (2017), we consider a finite population U of size N. Let $S_i^* \in \{0, 1\}$ denote an inclusion indicator for a nonprobability sample (NP), and let $S_i \in \{0, 1\}$ denote an inclusion indicator for a probability sample (P) for unit i. Three common assumptions are made:

A1 For all the units i in U, $\mathbb{P}(i \in NP)$ and $\mathbb{P}(i \in P)$ are non-zero.

A2 A common set of auxiliary variables $\mathbf{x}_i$ which govern the inclusion mechanism of the nonprobability sample is available in both the nonprobability sample and the probability sample, while the target variable y_i is only available in the nonprobability sample.

A3 The design weights of the probability sample d_i are available for both the probability sample and the nonprobability sample, or it is possible to estimate $d_i = \mathbb{P}^{-1}(S_i = 1|\mathbf{q}_i)$ with a set of variables $\mathbf{q}_i$ which govern the inclusion mechanism of the probability sample for all units in the two samples, as explained in Elliott and Valliant (2017) and Rafei et al. (2020).

For notational simplicity, below we presume $\mathbf{q}_i \subset \mathbf{x}_i$, while the inclusion mechanism of the nonprobability sample is not necessarily governed by $\mathbf{q}_i$. Elliott and Davis (2005) derive

$$\begin{aligned}\mathbb{P}(S_i^* = 1|\mathbf{x}_i = \mathbf{x}_o) &= \frac{\mathbb{P}(\mathbf{x}_i = \mathbf{x}_o|S_i^* = 1)\mathbb{P}(S_i^* = 1)}{\mathbb{P}(\mathbf{x}_i = \mathbf{x}_o)} \\ &= \frac{\mathbb{P}(\mathbf{x}_i = \mathbf{x}_o|S_i^* = 1)\mathbb{P}(S_i^* = 1)\mathbb{P}(S_i = 1|\mathbf{x}_i = \mathbf{x}_o)}{\mathbb{P}(S_i = 1)\mathbb{P}(\mathbf{x}_i = \mathbf{x}_o|S_i = 1)} \\ &\propto \frac{\mathbb{P}(\mathbf{x}_i = \mathbf{x}_o|S_i^* = 1)\mathbb{P}(S_i = 1|\mathbf{x}_i = \mathbf{x}_o)}{\mathbb{P}(\mathbf{x}_i = \mathbf{x}_o|S_i = 1)}.\end{aligned} \tag{2.1}$$

using Bayes' rule; this derivation is equation (2.4) in Elliott and Valliant (2017). Avoiding direct estimation of $\mathbb{P}(\mathbf{x}_i = \mathbf{x}_o|S_i^* = 1)$ and $\mathbb{P}(\mathbf{x}_i = \mathbf{x}_o|S_i = 1)$, Elliott and Valliant (2017) use discriminant analysis on the combined nonprobability and probability samples, setting $Z_i^{EV} = 1$ for units from the nonprobability sample and $Z_i^{EV} = 0$ for units from the probability sample, thus obtaining

$$\begin{aligned}\frac{\mathbb{P}(\mathbf{x}_i = \mathbf{x}_o|Z_i^{EV} = 1)}{\mathbb{P}(\mathbf{x}_i = \mathbf{x}_o|Z_i^{EV} = 0)} &= \frac{\mathbb{P}(Z_i^{EV} = 1|\mathbf{x}_i = \mathbf{x}_o)\mathbb{P}(\mathbf{x}_i = \mathbf{x}_o)/\mathbb{P}(Z_i^{EV} = 1)}{\mathbb{P}(Z_i^{EV} = 0|\mathbf{x}_i = \mathbf{x}_o)\mathbb{P}(\mathbf{x}_i = \mathbf{x}_o)/\mathbb{P}(Z_i^{EV} = 0)} \\ &\propto \frac{\mathbb{P}(Z_i^{EV} = 1|\mathbf{x}_i = \mathbf{x}_o)}{\mathbb{P}(Z_i^{EV} = 0|\mathbf{x}_i = \mathbf{x}_o)},\end{aligned} \tag{2.2}$$

which yields

$$\mathbb{P}(S_i^* = 1|\mathbf{x}_i = \mathbf{x}_o) \propto \mathbb{P}(S_i = 1|\mathbf{x}_i = \mathbf{x}_o)\frac{\mathbb{P}(Z_i^{EV} = 1|\mathbf{x}_i = \mathbf{x}_o)}{\mathbb{P}(Z_i^{EV} = 0|\mathbf{x}_i = \mathbf{x}_o)}.$$

Note that Elliott and Valliant (2017) make a fourth additional assumption:

A4 The sampling fractions of the probability sample and the nonprobability sample are small so that there are no common units in both samples. Specifically requiring that

$$\begin{aligned}\mathbb{P}(\mathbf{x}_i|Z_i^{EV} = 1) = \mathbb{P}(\mathbf{x}_i|S_i^* = 1, S_i = 0) \approx \mathbb{P}(\mathbf{x}_i|S_i^* = 1) \\ \mathbb{P}(\mathbf{x}_i|Z_i^{EV} = 0) = \mathbb{P}(\mathbf{x}_i|S_i^* = 0, S_i = 1) \approx \mathbb{P}(\mathbf{x}_i|S_i = 1).\end{aligned} \tag{2.3}$$

Our proposed method (see Section 2.2.2), does not require this assumption.

The estimated pseudo weight is then the inverse of the estimated propensity (Czajka, Hirabayashi, Little, & Rubin, 1992)

$$w_{i,EV} \propto d_i \frac{\hat{\mathbb{P}}(Z_i^{EV} = 0|\mathbf{x}_i = \mathbf{x}_o)}{\hat{\mathbb{P}}(Z_i^{EV} = 1|\mathbf{x}_i = \mathbf{x}_o)}. \tag{2.4}$$

The estimated $w_{i,EV}$ can then be plugged into, for example, a Hájek estimator

$$\sum_{i \in NP} w_i y_i / \sum_{i \in NP} w_i$$

for estimating the population mean of a target variable y.

It is worth to note that some literature suggests to use the inverse of the estimated odds as the weights without considering the design weights of the probability sample, that is, $w_i = \mathbb{P}(Z_i^{EV} = 0|\mathbf{x}_i = \mathbf{x}_o)/\mathbb{P}(Z_i^{EV} = 1|\mathbf{x}_i = \mathbf{x}_o)$ (Schonlau & Couper, 2017). However, this is only valid when the designed inclusion probability of the probability sample $\mathbb{P}(S_i = 1|\mathbf{x}_i = \mathbf{x}_o)$ is the same for all the units in the two samples, such as simple random sampling.

2.2.2 The Proposed Method

In practice, assumption **A4** does not always hold. Under a Big data scenario, the nonprobability sample may cover a large proportion of the population so that the sampling fraction is not ignorable. Even when the inclusion fraction of the nonprobability sample is small, it is possible that some units in the population have large propensities to be included in both nonprobability and probability samples, and therefore assumption **A4** is again invalid. In the proposed method, we inherit the assumptions **A1** to **A3** but relax assumption **A4**. Therefore, the merits of the original paper are also valid in our extension.

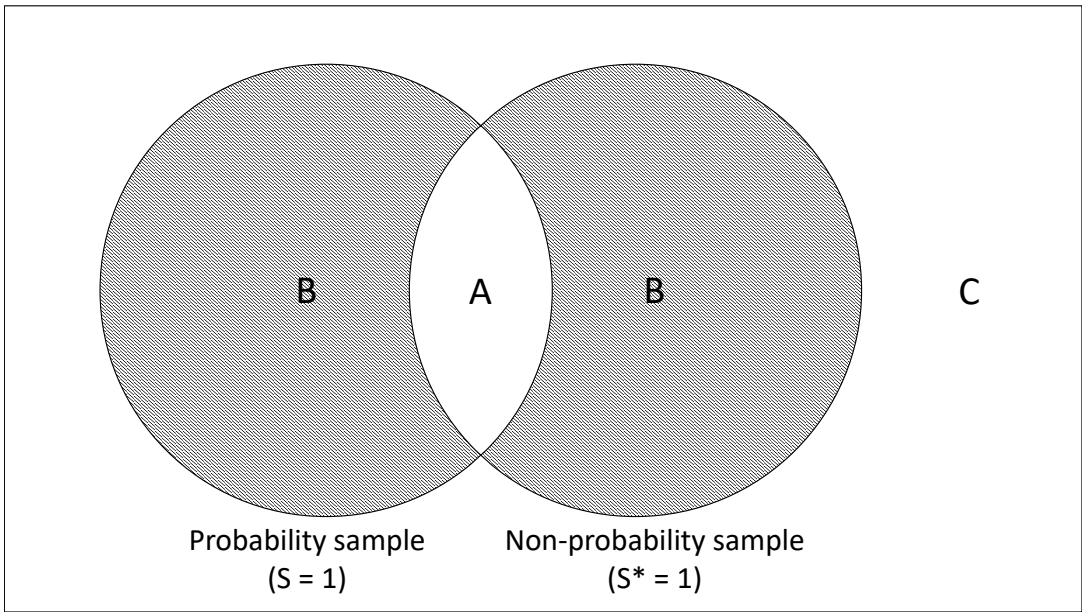

FIGURE 2.1: Venn diagram of the samples used. The A set contains the overlapping units of the two samples, which is removed and only the units in the B set are used for fitting the propensity model.

The target population is divided into three nonoverlapping subpopulations: $U = A \cup B \cup C$, with $A = \{i : S_i = S_i^* = 1\}, B = \{i : S_i + S_i^* = 1\}$, and $C = \{i : S_i + S_i^* = 0\}$. See Figure 2.1 for the Venn diagram of the three subpopulations. That is, we assume that

A5 It is possible to identify (in any case with a high probability) the overlapping units between the two samples, i.e. subpopulation A.

Within subpopulation B, define $Z_i = 1$ if $(S_i^*, S_i) = (1, 0)$ and $Z_i = 0$ if $(S_i^*, S_i) = (0, 1)$. In the subpopulation B, it can be derived analogously to Equation (2.1) that

$$\begin{aligned}
&\mathbb{P}(S_i^* = 1, i \in B | \mathbf{x}_i = \mathbf{x}_o) \\
&= \frac{\mathbb{P}(\mathbf{x}_i = \mathbf{x}_o | S_i^* = 1, i \in B)\mathbb{P}(S_i^* = 1, i \in B)}{\mathbb{P}(\mathbf{x}_i = \mathbf{x}_o)} \\
&= \frac{\mathbb{P}(\mathbf{x}_i = \mathbf{x}_o | S_i^* = 1, i \in B)\mathbb{P}(S_i^* = 1, i \in B)\mathbb{P}(S_i = 1, i \in B | \mathbf{x}_i = \mathbf{x}_o)}{\mathbb{P}(S_i = 1, i \in B)\mathbb{P}(\mathbf{x}_i = \mathbf{x}_o | S_i = 1, i \in B)} \\
&= g_1 \times \frac{\mathbb{P}(\mathbf{x}_i = \mathbf{x}_o | S_i^* = 1, i \in B)\mathbb{P}(S_i = 1, i \in B | \mathbf{x}_i = \mathbf{x}_o)}{\mathbb{P}(\mathbf{x}_i = \mathbf{x}_o | S_i = 1, i \in B)},
\end{aligned} \tag{2.5}$$

where $g_1 = \mathbb{P}(S_i^* = 1, i \in B)/\mathbb{P}(S_i = 1, i \in B)$, and Equation (2.2) becomes

$$\begin{aligned}
&\frac{\mathbb{P}(\mathbf{x}_i = \mathbf{x}_o | Z_i = 1, i \in B)}{\mathbb{P}(\mathbf{x}_i = \mathbf{x}_o | Z_i = 0, i \in B)} \\
&= \frac{\mathbb{P}(Z_i = 1 | \mathbf{x}_i = \mathbf{x}_o, i \in B)\mathbb{P}(\mathbf{x}_i = \mathbf{x}_o | i \in B)/\mathbb{P}(Z_i = 1 | i \in B)}{\mathbb{P}(Z_i = 0 | \mathbf{x}_i = \mathbf{x}_o, i \in B)\mathbb{P}(\mathbf{x}_i = \mathbf{x}_o | i \in B)/\mathbb{P}(Z_i = 0 | i \in B)} \\
&= g_2 \times \frac{\mathbb{P}(Z_i = 1 | \mathbf{x}_i = \mathbf{x}_o, i \in B)}{\mathbb{P}(Z_i = 0 | \mathbf{x}_i = \mathbf{x}_o, i \in B)},
\end{aligned} \tag{2.6}$$

where $g_2 = \mathbb{P}(Z_i = 0 | i \in B)/\mathbb{P}(Z_i = 1 | i \in B)$. Since within subpopulation B, it holds that $S_i^* = 1$ if and only if $Z_i = 1$ and that $S_i^* = 0$ if and only if $Z_i = 0$, we now obtain from Equation (2.5) and Equation (2.6) for all possible sampling fractions:

$$\begin{aligned}
&\mathbb{P}(S_i^* = 1, i \in B | \mathbf{x}_i = \mathbf{x}_o) \\
&= g \times \mathbb{P}(S_i = 1, i \in B | \mathbf{x}_i = \mathbf{x}_o) \frac{\mathbb{P}(Z_i = 1 | \mathbf{x}_i = \mathbf{x}_o, i \in B)}{\mathbb{P}(Z_i = 0 | \mathbf{x}_i = \mathbf{x}_o, i \in B)},
\end{aligned} \tag{2.7}$$

with $g = g_1 g_2$, where

$$\begin{aligned}
g &= \frac{\mathbb{P}(S_i^* = 1, i \in B)\mathbb{P}(Z_i = 0 | i \in B)}{\mathbb{P}(S_i = 1, i \in B)\mathbb{P}(Z_i = 1 | i \in B)} \\
&= \frac{\mathbb{P}(S_i^* = 1, i \in B)\mathbb{P}(S_i = 1 | i \in B)}{\mathbb{P}(S_i = 1, i \in B)\mathbb{P}(S_i^* = 1 | i \in B)} \\
&= \frac{\mathbb{P}(i \in B)}{\mathbb{P}(i \in B)} = 1.
\end{aligned}$$

Since $\mathbb{P}(S_i = 1 | \mathbf{x}_i = \mathbf{x}_o)$ is assumed to be known for all the units in the two samples, we can write:

$$\begin{aligned}
\mathbb{P}(S_i^* = 1, i \in B | \mathbf{x}_i = \mathbf{x}_o) &= \mathbb{P}(S_i^* = 1 | \mathbf{x}_i = \mathbf{x}_o)\mathbb{P}(i \in B | S_i^* = 1, \mathbf{x}_i = \mathbf{x}_o) \\
&= \mathbb{P}(S_i^* = 1 | \mathbf{x}_i = \mathbf{x}_o)\mathbb{P}(S_i = 0 | S_i^* = 1, \mathbf{x}_i = \mathbf{x}_o) \\
&= \mathbb{P}(S_i^* = 1 | \mathbf{x}_i = \mathbf{x}_o)\mathbb{P}(S_i = 0 | \mathbf{x}_i = \mathbf{x}_o).
\end{aligned} \tag{2.8}$$

By design, the probability of unit i to be included in the probability sample does not depend on S_i^* after conditioning on $\mathbf{x}_i$. Similarly,

$$\mathbb{P}(S_i = 1, i \in B|\mathbf{x}_i = \mathbf{x}_o) = \mathbb{P}(S_i = 1|\mathbf{x}_i = \mathbf{x}_o)\mathbb{P}(S_i^* = 0|S_i = 1, \mathbf{x}_i = \mathbf{x}_o). \quad (2.9)$$

Often it is assumed that $\mathbb{P}(S_i^* = 0|S_i = 1, \mathbf{x}_i = \mathbf{x}_o) = \mathbb{P}(S_i^* = 0|\mathbf{x}_i = \mathbf{x}_o)$, that is, the inclusion of the nonprobability sample is independent of the inclusion of the probability sample; see, e.g., Robbins, Ghosh-Dastidar, and Ramchand (2021). Therefore,

$$\mathbb{P}(S_i = 1, i \in B|\mathbf{x}_i = \mathbf{x}_o) = \mathbb{P}(S_i = 1|\mathbf{x}_i = \mathbf{x}_o)\mathbb{P}(S_i^* = 0|\mathbf{x}_i = \mathbf{x}_o). \quad (2.10)$$

As noted in the introduction, there may also exist applications where this assumption is not reasonable. In the next subsection we will first discuss the scenario where the two samples are independent, and later in Subsection 2.2.2 the scenario where the two samples are dependent is discussed.

Independent Samples

Combining Equations (2.8) and (2.10) with (2.7), we obtain:

$$\frac{\mathbb{P}(S_i^* = 1|\mathbf{x}_i = \mathbf{x}_o)}{\mathbb{P}(S_i^* = 0|\mathbf{x}_i = \mathbf{x}_o)} = \frac{\mathbb{P}(S_i = 1|\mathbf{x}_i = \mathbf{x}_o)}{\mathbb{P}(S_i = 0|\mathbf{x}_i = \mathbf{x}_o)} \frac{\mathbb{P}(Z_i = 1|\mathbf{x}_i = \mathbf{x}_o, i \in B)}{\mathbb{P}(Z_i = 0|\mathbf{x}_i = \mathbf{x}_o, i \in B)}. \quad (2.11)$$

Since $\mathbb{P}^{-1}(S_i = 1|\mathbf{x}_i = \mathbf{x}_o) = d_i$,

$$\frac{\mathbb{P}(S_i = 1|\mathbf{x}_i = \mathbf{x}_o)}{\mathbb{P}(S_i = 0|\mathbf{x}_i = \mathbf{x}_o)} = \frac{1/d_i}{1 - 1/d_i} = \frac{1}{d_i - 1}.$$

Also we denote

$$O_i = \frac{\mathbb{P}(Z_i = 1|\mathbf{x}_i = \mathbf{x}_o, i \in B)}{\mathbb{P}(Z_i = 0|\mathbf{x}_i = \mathbf{x}_o, i \in B)}.$$

O_i can be estimated by any model that is suitable for binary class probability estimation, for example, logistic regression or some machine learning method. Combining this notation together with Equation (2.11) we obtain

$$\hat{\mathbb{P}}(S_i^* = 1|\mathbf{x}_i = \mathbf{x}_o) = \frac{\hat{O}_i/(d_i - 1)}{1 + \hat{O}_i/(d_i - 1)} = \frac{\hat{O}_i}{\hat{O}_i + d_i - 1}. \tag{2.12}$$

Note that $\hat{\mathbb{P}}(S_i^* = 1|\mathbf{x}_i = \mathbf{x}_o)$ is computed for all units in the nonprobability sample, including those in subpopulation A. The weight is again the inverse of the propensity

$$w_{i,ind} = \frac{1}{\hat{\mathbb{P}}(S_i^* = 1|\mathbf{x}_i = \mathbf{x}_o)} = 1 + \frac{d_i - 1}{\hat{O}_i} \tag{2.13}$$

Dependent Samples

In some cases it is not appropriate to assume that the inclusion in the nonprobability sample is independent of the inclusion in the probability sample. In such a case, the following formulation should be used. Again combining Equations (2.8) and (2.9) with (2.7) we now have,

$$\begin{aligned}&\mathbb{P}(S_i^* = 1|\mathbf{x}_i = \mathbf{x}_o)\\ &= \frac{\mathbb{P}(S_i = 1|\mathbf{x}_i = \mathbf{x}_o)}{\mathbb{P}(S_i = 0|\mathbf{x}_i = \mathbf{x}_o)} \frac{\mathbb{P}(Z_i = 1|\mathbf{x}_i = \mathbf{x}_o, i \in B)}{\mathbb{P}(Z_i = 0|\mathbf{x}_i = \mathbf{x}_o, i \in B)} \mathbb{P}(S_i^* = 0|S_i = 1, \mathbf{x}_i = \mathbf{x}_o).\end{aligned} \tag{2.14}$$

The new term $\mathbb{P}(S_i^* = 0|S_i = 1, \mathbf{x}_i = \mathbf{x}_o)$ can be modelled on units in the probability sample using, e.g., logistic regression:

$$\log \frac{\mathbb{P}(S_i^* = 1|S_i = 1, \mathbf{x}_i = \mathbf{x}_o)}{\mathbb{P}(S_i^* = 0|S_i = 1, \mathbf{x}_i = \mathbf{x}_o)} = \boldsymbol{\beta}^T \mathbf{x}_i,$$

which yields the following estimate

$$\hat{\mathbb{P}}(S_i^* = 0|S_i = 1, \mathbf{x}_i = \mathbf{x}_o) = 1/(1 + \exp(\boldsymbol{\beta}^T \mathbf{x}_i)) \equiv L_i.$$

The rest of Equation (2.14) is estimated as before, and we find

$$\hat{\mathbb{P}}(S_i^* = 1|\mathbf{x}_i = \mathbf{x}_o) = \hat{O}_i\hat{L}_i/(d_i - 1). \tag{2.15}$$

Hence, we have obtained an estimated inclusion probability that can be computed for all units in the nonprobability sample, including those in subpopulation A. The resulting pseudo-weight is

$$w_{i,dep} = \frac{1}{\hat{\mathbb{P}}(S_i^* = 1|\mathbf{x}_i = \mathbf{x}_o)} = \frac{d_i - 1}{\hat{O}_i\hat{L}_i}. \tag{2.16}$$

For some applications it may be reasonable to assume that the inclusion in the nonprobability sample is independent of the inclusion in the probability sample. In particular, this assumption seems reasonable if it is known that the probability sample was drawn *after* the nonprobability sample became available. By definition, all units in the population with a given combination of values in $\mathbf{x}$ have the same (known) probability of being included in the probability sample and this sample was selected randomly using these probabilities. Therefore, in theory, if the nonprobability sample was 'selected' before the probability sample was drawn, then conditional on $\mathbf{x}_i$ the fact that $S_i = 1$ cannot provide any additional information about the likelihood that $S_i^* = 1$ or $S_i^* = 0$.

By contrast, when the probability sample is selected before the nonprobability sample, it is possible to construct examples where $\mathbb{P}(S_i^* = 0|S_i = 1, \mathbf{x}_i = \mathbf{x}_o) \neq \mathbb{P}(S_i^* = 0|\mathbf{x}_i = \mathbf{x}_o)$ because for some or all units in the population the event of being included in the probability sample somehow affects the likelihood of being included (later) in the nonprobability sample. However, in general it is not possible to test this using only the observed data. At best, one could test the stronger assumption that $\mathbb{P}(S_i^* = 0|S_i = 1, \mathbf{x}_{1i} = \mathbf{x}_{1o}) = \mathbb{P}(S_i^* = 0|\mathbf{x}_{1i} = \mathbf{x}_{1o})$ for the subset of auxiliary

variables that are available for all units in the target population (in particular, the design variables of the probability sample). This could be done by estimating a logistic regression model of the form

$$\log \frac{\mathbb{P}(S_i^* = 1 | S_i = s_o, \mathbf{x}_{1i} = \mathbf{x}_{1o})}{\mathbb{P}(S_i^* = 0 | S_i = s_o, \mathbf{x}_{1i} = \mathbf{x}_{1o})} = \boldsymbol{\beta}_1^T \mathbf{x}_{1i} + \gamma S_i$$

on the entire target population. If the null hypothesis that $\gamma = 0$ is not rejected, then one would conclude that $\mathbb{P}(S_i^* = 0 | S_i = 1, \mathbf{x}_{1i} = \mathbf{x}_{1o}) = \mathbb{P}(S_i^* = 0 | \mathbf{x}_{1i} = \mathbf{x}_{1o})$ and therefore also that the weaker assumption $\mathbb{P}(S_i^* = 0 | S_i = 1, \mathbf{x}_i = \mathbf{x}_o) = \mathbb{P}(S_i^* = 0 | \mathbf{x}_i = \mathbf{x}_o)$ holds. However, in practice for large populations this test may be too sensitive against small deviations from $\gamma = 0$.

2.3 Variance Estimation

Since the proposed propensity estimation method allows any type of propensity estimation model to be applied, our goal for the variance estimation is to construct a general framework that is suitable for a wide range of target parameters and propensity estimation models. Given the two-sample setup, the variance of the target parameter estimates can be decomposed into the variation from the 'sampling mechanism' of the nonprobability sample, the variation from the design of the probability sample, and the propensity estimation process.

In the nonprobability sample part, so far we have assumed that the nonprobability sample came from some unequal probability sampling mechanism and the first order inclusion probability $\pi_i = \mathbb{P}(S_i^* = 1 | \mathbf{x}_i = \mathbf{x}_o)$ is estimated by Equation (2.12) or (2.15). The design-based estimators for the target parameter can then be applied, for example, the Horvitz-Thompson estimator $\hat{Y}_{HT} = \sum_{i=1}^{n} y_i / \pi_i$ for the population

total estimation. Usually to complete the variance estimation for a certain estimator, the second order inclusion probability $\pi_{ij} = \mathbb{P}(S_i^* = 1, S_j^* = 1)$ for every $i \neq j$ should be considered. For instance, the variance of the Horvitz-Thompson estimator is (Cochran, 1977)

$$Var(\hat{Y}_{HT}) = \sum_{i=1}^{N} \frac{(1-\pi_i)}{\pi_i} y_i^2 + 2 \sum_{i=1}^{N} \sum_{j>i}^{N} \frac{(\pi_{ij} - \pi_i \pi_j)}{\pi_i \pi_j} y_i y_j.$$

However, the "true" sampling mechanism is unknown for the nonprobability sample, and therefore the second order inclusion probability is, in general, unknown and hard to estimate. A choice may be to assume that the nonprobability sample is from a high entropy sampling design so that $\pi_{ij} \approx \pi_i \pi_j$ (Yang, Kim, & Song, 2020), although it is not obvious how to test whether this assumption is valid or not.

Rather than assuming the nonprobability sample is from a high entropy sampling design, we offer a framework where researchers' understanding of the nonprobability sample can be considered and assist to obtain a reasonable variance estimate. For example, if the goal of a nonprobability survey is to collect one thousand participants, then we may see it as a fixed size design. Or if some units in the nonprobability sample are observed more than once, we may assume the sample is drawn with replacement. When the understanding of the inclusion mechanism of the nonprobability sample is limited, it may still be possible to have a reasonable estimate of the variance. Since the sample size of the probability sample is often smaller than the size of the nonprobability sample, the uncertainty of the point estimate may mainly come from the probability sample. As long as the sampling design of the probability sample is considered, the resulting variance can give an idea of the order of magnitude of the uncertainty.

Many resampling methods have been proposed to capture the variation in a probability sample, see for example J. Rao, Wu, and Yue (1992). In order to achieve the

flexibility for researchers to include their knowledge of the inclusion mechanism of the nonprobability sample, and allow the possibility of considering the dependency between two samples, we implement a pseudo population bootstrap. The pseudo population bootstrap was proposed by Booth, Butler, and Hall (1994), and an extension suitable for Poisson sampling is considered by Chauvet (2007) with fixed sample size. See Mashreghi, Haziza, and Léger (2016) for an excellent overview paper on finite population bootstrapping. The pseudo population bootstrap can be applied for estimators that are smooth functions of population means, for example, totals, ratios, or coefficients of linear regression. In the simulation of Antal and Tillé (2014), they found that the pseudo population approach also works well for estimating the variance of an estimated median. Pseudo population bootstrap has also been applied for variance estimation for imputed survey data (S. Chen, Haziza, Léger, & Mashreghi, 2019).

The traditional pseudo population bootstrap for a probability sample creates a pseudo population by copying the units in the sample (approximately) d_i times, and then repeatedly drawing samples according to the sampling design to capture the uncertainty of the sampling process. Since pseudo population bootstrap mimics the drawing process in the population, it is more flexible for a large variety of sampling designs compared to other resampling methods for finite populations. The tradeoff may be when the population size is large, it is computationally demanding. We extend the traditional pseudo population bootstrap to our situation with a probability sample and a nonprobability sample by assuming $E[\hat{\mathbb{P}}(S_i^* = 1|\mathbf{x}_i = \mathbf{x}_o)] = \mathbb{P}(S_i^* = 1|\mathbf{x}_i = \mathbf{x}_o)$. The nonprobability sample is then expanded into a pseudo population from which both samples can be drawn and the overlapping units can be identified.

2.3.1 Variance Estimate for Independent Samples

Algorithm 1 Pseudo population bootstrap for independent samples

1: Estimate $\hat{O}_i$, and the weights of the nonprobability sample ($w_i \equiv w_{i,ind}$).
2: Normalize the weights by $w_i N / \sum_{i \in NP} w_i$ to obtain $\sum_{i \in NP} w_i = N$.
3: Randomly round w_i to its ceiling with probability $w_i - \lfloor w_i \rfloor$ and to its floor otherwise to obtain $[w_i]$, with the property that $\sum_{i \in NP} [w_i] = N$.
4: Create a pseudo population by copying unit i $[w_i]$ times.
5: Draw a bootstrap probability sample (S_P) from the pseudo population according to the design of the probability sample, with inclusion probabilities $1/d_i$.
6: Draw a bootstrap nonprobability sample (S_{NP}) from the pseudo population with inclusion probabilities $1/w_i$.
7: Remove the overlapping units in S_P and S_{NP}, and then estimate $\hat{O}_i$ to complete the estimation of the weights and the target parameter.
8: Repeat Step 5 to 7 for R times to acquire R estimates.
9: Compute the bootstrap variance of the R estimates.

The variance of the proposed estimator in independent samples is estimated by Algorithm 1. A pseudo population is created by replicating units in the nonprobability sample, and both the nonprobability bootstrap sample (S_{NP}) and the probability bootstrap sample (S_P) are drawn from it. We create only one pseudo population instead of two separate pseudo populations (one for the nonprobability sample and one for the probability sample) since in Step 7 the overlapping part of the two samples should be recognized and accounted for. Therefore, our estimation approach for the variance is based on an underlying assumption that the nonprobability sample and the probability sample have similar quality in terms of measurement and selectivity, so that we can draw a bootstrap probability sample similar to the original probability sample from a pseudo population which is created from the nonprobability sample only. If the probability sample size is much larger than the nonprobability sample, which would be unusual in practice, this method may not be suitable since too many copies of the units in the nonprobability sample are used to represent the probability sample.

To create the pseudo population, the estimated weight for each unit in the nonprobability sample is treated as the 'design weight.' The weight is rounded to its ceiling with probability $w_i - \lfloor w_i \rfloor$ and to its floor otherwise with setting $\sum_{i \in NP}[w_i] = N$, to obtain $[w_i]$ (Fellegi, 1975). Units in the original nonprobability sample are replicated $[w_i]$ times to become the pseudo population. Note that the $[w_i]$ copies are treated as different units in the population. Therefore, an overlap of the P and NP samples in a bootstrap sample will occur when exactly the same copy of a unit is selected for both samples. Bootstrap samples S_P are drawn from the pseudo population based on the original design. S_{NP} are drawn with probability $1/w_i$ considering the knowledge of the (possible) sampling mechanism of the nonprobability sample, for example, whether the sample size is fixed or not. A target parameter θ is estimated by S_{NP} and by S_P. That is, first discard the overlapping units of S_{NP} and S_P, and then estimate w_{ind} as in Equation (2.13). If, for example, the aim is to estimate the population mean, then θ can be estimated by $\hat{\theta} = \sum_{i \in NP} w_i y_i / \sum_{i \in NP} w_i$, where y_i is the variable of interest. A large number R of bootstrap (non-)probability samples are drawn, and the estimated variance of the estimators is $\widehat{Var}(\hat{\theta}) = (R-1)^{-1} \sum_{r=1}^{R} (\hat{\theta}_r - \bar{\theta})^2$ with $\bar{\theta} = R^{-1} \sum_{r=1}^{R} \hat{\theta}_r$.

It is worth noting that some pseudo population bootstrap literature suggests creating D pseudo populations to deal with the effect of random rounding in Step 3. That is, Steps 3 to 9 are repeated D times, and the variance is estimated by taking the mean of the D variance estimates. However, it is sufficient to have only one pseudo population, see Mashreghi et al. (2016) and especially Chauvet (2007). Later in the Simulation study, we will also give results for $D = 10$ which do not show obvious differences with $D = 1$.

2.3.2 Variance Estimate for Dependent Samples

Algorithm 2 Pseudo population bootstrap for dependent samples

1: Estimate $\hat{L}_i$, $\hat{O}_i$, and the weights of the nonprobability sample ($w_i \equiv w_{i,dep}$).
2: Normalize the weights by $w_i N / \sum_{i \in NP} w_i$ to obtain $\sum_{i \in NP} w_i = N$.
3: Randomly round w_i to its ceiling with probability $w_i - \lfloor w_i \rfloor$ and to its floor otherwise to obtain $[w_i]$, with the property that $\sum_{i \in NP} [w_i] = N$.
4: Create a pseudo population by copying unit i $[w_i]$ times.
5: Calculate $\mathbb{P}(S_i^* = 1 | S_i = 0, \mathbf{x}_i = \mathbf{x}_o)$ and $\mathbb{P}(S_i^* = 1 | S_i = 1, \mathbf{x}_i = \mathbf{x}_o)$ for all the units in the pseudo population.
6: Draw a bootstrap probability sample (S_P) from the pseudo population according to the design of the probability sample, with inclusion probabilities $1/d_i$.
7: Draw a bootstrap nonprobability sample (S_{NP}) from the pseudo population with probability $\hat{\mathbb{P}}(S_i^* = 1 | S_i = 0, \mathbf{x}_i = \mathbf{x}_o)$ or $\hat{\mathbb{P}}(S_i^* = 1 | S_i = 1, \mathbf{x}_i = \mathbf{x}_o)$ given the unit is in S_P or not.
8: Estimate $\hat{L}_i$, $\hat{O}_i$ using S_P and S_{NP} to complete the estimation of the weights and the target parameter.
9: Repeat Step 6 to 8 for R times to acquire R estimates.
10: Compute the bootstrap variance of the R estimates.

Algorithm 2 is for the variance estimation for dependent samples, which requires two extra steps. The probability bootstrap sample is drawn first from the pseudo population, then the nonprobability bootstrap sample is drawn given the units drawn in the probability bootstrap sample. That is, the probability of drawing a unit i into S_{NP} is $\hat{\mathbb{P}}(S_i^* = 1 | S_i = 1, \mathbf{x}_i = \mathbf{x}_o) = 1 - \hat{L}_i$ for units that are drawn in S_P, and $\hat{\mathbb{P}}(S_i^* = 1 | S_i = 0, \mathbf{x}_i = \mathbf{x}_o)$ for units that are not drawn in S_P. We can calculate

$\mathbb{P}(S_i^* = 1|S_i = 0, \mathbf{x}_i = \mathbf{x}_o)$ by the law of total probability,

$$\begin{aligned}\mathbb{P}(S_i^* = 1|\mathbf{x}_i = \mathbf{x}_o) = \mathbb{P}(S_i^* = 1|S_i = 1, \mathbf{x}_i = \mathbf{x}_o)\mathbb{P}(S_i = 1|\mathbf{x}_i = \mathbf{x}_o) \\ + \mathbb{P}(S_i^* = 1|S_i = 0, \mathbf{x}_i = \mathbf{x}_o)\mathbb{P}(S_i = 0|\mathbf{x}_i = \mathbf{x}_o),\end{aligned}$$

which can be re-arranged and plugged into the estimates before to obtain

$$\hat{\mathbb{P}}(S_i^* = 1|S_i = 0, \mathbf{x}_i = \mathbf{x}_o) = 1 - d_i + d_i\hat{O}_i\hat{L}_i + \hat{L}_i(d_i - 1).$$

The rest of the process is the same as for the independent samples.

2.4 Simulation Study

2.4.1 Design

The proposed methods are compared with other methods that apply the same two-sample setup for selection bias correction. The compared methods are:

- Naive estimator: treat the nonprobability sample as a simple random sample.
- Proposed methods: w_{ind} from Equation (2.13) and w_{dep} from Equation (2.16). To illustrate the effect of removing the overlapping part, a second choice, $w_{ind,2}$, is estimated as if the overlapping part is unknown so that all the data in both samples are used in the estimation procedure.
- EV: the original pseudo-weight estimation from Equation (2.4) proposed by Elliott and Valliant (2017).
- CLW: a weighted logistic regression with solving the score equation:

$$l(\beta) = \sum_{i \in NP} \mathbf{x}_i - \sum_{i \in P} d_i \pi(\mathbf{x}_i, \beta)\mathbf{x}_i = 0,$$

where NP and P stand for the nonprobability sample and the probability sample, and $\pi(\mathbf{x}_i, \beta)$ stands for the estimated propensity given the auxiliary variables $\mathbf{x}_i$ and the parameters for the logistic regression β (Y. Chen et al., 2020). The final weight is the inverse of the estimated propensity $w_i = \pi^{-1}(\mathbf{x}_i, \beta)$.

- VD: a weighted logistic regression with solving the score equation (Valliant, 2020; Valliant & Dever, 2011):

$$l(\beta) = \sum_{i \in NP} \mathbf{x}_i[1 - \pi(\mathbf{x}_i, \beta)] - \sum_{i \in P} d_i \pi(\mathbf{x}_i, \beta)\mathbf{x}_i = 0,$$

  where the final weight is the inverse of the estimated propensity $w_i = \pi^{-1}(\mathbf{x}_i, \beta)$.

- WVL: an adjusted version of VD. The modeling process is the same as VD, but the weights are calculated differently. Rather than using the inverse of $p_i \equiv \pi(\mathbf{x}_i, \beta)$ as weights, WVL use the inverted odds ratio $(1 - p_i)/p_i$ as weights, which is constrained by $p_i \leq 0.5$ (Wang, Valliant, & Li, 2021).

- KW: a weighted logistic regression where the propensity is estimated only by the probability sample. That is, assign $z_i = 1$ for the overlapping units and $z_i = 0$ for the non-overlapping units in the probability sample (Kim & Wang, 2019). The fitted model is then applied for calculating the propensity of being included in NP for each unit i of the nonprobability sample given $\mathbf{x}_i$.

It is worth to note that other than the methods proposed in this paper and EV, the rest of the methods consider the design weights of the probability sample during the logistic regression fitting procedure.

A real data set is used as the population for the simulation study. The data set is the registered data from the Dutch Online Kilometer Registration, which contains around 6.7 million records of the privately-owned cars in the Netherlands in 2012.

The variables include x_1 = first registration year of the car, x_2 = engine type of the car, x_3 = age of car owner, and the target variable y = mileage of the car. See Buelens et al. (2018) for more details of the data set. Following these authors, we treat all variables, including x_2 (using the codes 1, 2, 3, and 4 for different engine types), as numerical ones. We draw a simple random sample without replacement of size $N = 100,000$ to serve as the population of the simulation study.

The inclusion fractions of the nonprobability sample $f_{NP} = n_{NP}/N$ are 0.05, 0.3, 0.5, and the inclusion fractions of the probability sample $f_P = n_P/N$ are 0.01, 0.1, which reflect the common setting of a large inclusion fraction of the nonprobability sample and a smaller inclusion fraction of the probability sample. The two samples are either drawn independently or dependently, see the next paragraph. In total, 12 ($= 3 \times 2 \times 2$) combined scenarios are run.

Both samples are drawn by fixed-size unequal probability sampling without replacement with random systematic sampling where

$$\mathbb{P}(S^* = 1) = \exp(c_1 + x)/(1 + \exp(c_1 + x))$$

where $x = x_2 - x_3/20$ for the nonprobability sample, and

$$\mathbb{P}(S = 1) = \exp(c_2 + x_1/30)/(1 + \exp(c_2 + x_1/30))$$

for the probability sample. Constants c_1 and c_2 are used for controlling the sample size, and x_2 is treated as a numerical data to allow c_1 to be easier to compute. When drawing two dependent samples, the units already drawn in the probability sample are given the auxiliary value $x - 2$ instead of x, resulting in a smaller chance to be included in the nonprobability sample.

The target parameter is the population mean $\bar{y} = N^{-1} \sum_{i=1}^{N} y_i = 11741.79$,

which is estimated by $\hat{\bar{y}} = \sum_{i \in NP} w_i y_i / \sum_{i \in NP} w_i$ with the corresponding estimated weights w_i. The relative bias

$$M^{-1} \sum_{m=1}^{M} (\hat{\bar{y}}_m - \bar{y}) / \bar{y} \times 100\%$$

and the root mean square error

$$\text{RMSE} = (M^{-1} \sum_{m=1}^{M} (\hat{\bar{y}}_m - \bar{y})^2)^{1/2}$$

are shown to reflect the performance of the methods, where $M = 1,000$ is the number of replications.

2.4.2 Results

Table 2.1 and 2.2 show the relative bias and RMSE of the compared methods. The proposed methods w_{ind} and w_{dep} perform better when the samples are drawn with the corresponding dependency. Theoretically, w_{dep} is unbiased for both independent and dependent two samples. However, the estimate is less efficient when using w_{dep} for the independent two samples. Bias increases when the overlapping part is not removed. However, $w_{ind,2}$ still enjoys the efficiency that outperforms some of the other methods in terms of RMSE.

Relative to other comparing methods, CLW is more robust to the dependency of the two samples in terms of relative bias. The bias of KW is large when the probability sample size is small and when the two samples are dependent, since the estimation process only uses units in the probability sample. It is not shown here but we also observed that when the probability sample has a large inclusion fraction, KW also works well.

2.4.3 Variance Estimate

The simulation design for the validation of the variance estimates is the same as before. When the number of pseudo populations D is set equal to 1, the pseudo population bootstrap is run $R = 500$ times. That is, 500 (non-)probability samples are drawn from the same one pseudo population for each scenario. We also show the results for $D = 10$ and $R = 50$ for independent samples, to illustrate the effect of multiple pseudo populations. The (non-)probability bootstrap samples are drawn by fixed-size random systematic sampling given the known inclusion probability for the probability sample or the estimated propensities for the nonprobability sample. The true variances are calculated by means of Monte Carlo Simulation. Table 2.3 shows the relative bias of the estimated variances and the coverage rate of the 95% confidence intervals $CI_m = (\hat{\bar{y}}_m - 1.96se_m,\ \hat{\bar{y}}_m + 1.96se_m)$ where se_m is the estimated standard error for replication m. The coverage rate is calculated as $CR = M^{-1}\sum_{m=1}^{M} \mathbf{1}(\bar{y} \in CI_m) \times 100\%$ where $\mathbf{1}(\cdot)$ is the indicator function and $\bar{y}$ is the true mean of the constructed pseudo-population. The variance estimates for the proposed methods are generally close to the true variances as long as the inclusion fraction of the nonprobability sample is large, as noted in Section 2.3. The coverage rate of the confidence intervals are in general close to 95%. Only in the dependent samples when $f_{NP} = 0.3, f_P = 0.1$, the coverage rate is notably lower. This can be explained because the point estimate (see Table 2.1) has a relatively large bias compared to the other conditions. Although this bias is not particularly large (0.58%), its magnitude is similar to the relative standard error of the point estimate (0.60%, as can be computed from Table 2.1 and 2.2) and this is enough to have a significant impact on the coverage rate of the confidence interval.

2.5 Conclusion and Discussion

To correct the possible selection error in a nonprobability sample, we have extended the pseudo-weight approach from Elliott and Valliant to more general formulations. The proposed methods can be applied for all kinds of inclusion fractions of nonprobability and probability samples, and retain the merits of the original method. In a simulation study we showed that the proposed methods outperform other related methods and the dependency between the two samples is crucial to be considered to obtain a reliable estimate. Algorithms for estimating the variance by a pseudo population bootstrap approach are demonstrated, and we showed that the algorithms guarantee reasonable variance estimates in the simulation study.

Some limitations may be found in practice when applying the proposed method. Like with many other related works, we applied the two-sample setup for propensity estimation. It is assumed that the inclusion probabilities of both the probability sample and nonprobability sample are all non-zero and the two samples are from exactly the same population. This may not be the case in practice. Also, the two-sample setup is highly relying on the quality of the probability sample, while in practice often probability samples suffer from non-response or other quality issues, which may affect the estimated propensities. The measurement of the variables in the two samples may not always be the same and result in estimation difficulties. The design variable(s) of the probability sample are needed in the nonprobability sample or it should be possible to estimate probabilities to be included into the probability sample for the units in the nonprobability sample as noted in Rafei et al. (2020), which may not always be the case.

We use the non-overlapping parts of the probability sample and the nonprobability sample for the estimation. Therefore, if one sample is a subset of the other one, or the non-overlapping part of the two samples is trivial, it is not suitable to

apply the proposed method. However, if the nonprobability sample is a subset of the probability sample, the situation reduces to the well-known problem of missing data inside a probability sample. If, on the other hand, the probability sample is a subset of the nonprobability sample, it is a data integration problem since the target variable is known for both samples. See, e.g., J. Rao (2020) for a review of methods that can be used in this situation. Identifying the overlapping part between the two samples may not be easy. Although we offer the result of including the overlapping part in the sample, attempting to remove the overlap by probabilistic linkage may be an alternative option in practice. How probabilistic linkage will affect the proposed method remains unknown for now. If the design variable(s) or unique ID are not available, the CLW method may be a good option since the simulation shows that the CLW method is relatively robust to the possible dependency between the inclusion of the two samples. If the required information is available, efficiency is gained by applying our proposed method.

In this research we only applied logistic regression for the odds estimate, while any model suitable for binary class probability estimation can be easily plugged into the proposed methods. If researchers are interested in applying the doubly robust framework for selection bias correction (Y. Chen et al., 2020), the method discussed here can also be applied for the propensity estimation step in the doubly robust approach.

Although we did not look into it in this research, some post procedures of pseudo weights may also benefit the parameter estimation. For example, after the weights are estimated by the proposed method, the estimation may be improved by calibrating the estimated weights with other auxiliary variables that are related to the target variable, as traditional calibration in probability samples (Deville & Särndal, 1992; Wu & Sitter, 2001). At that stage, one might use auxiliary variables for which (estimated) population totals are available but not values for individual units outside

the nonprobability sample. Other techniques such as trimming or post-stratification of the pseudo weights may also be a remedy for the large variance of the pseudo weights (Cochran, 1977; Rafei et al., 2020).

To estimate the uncertainty of the estimated target parameter, we propose pseudo population bootstrap algorithms in order to be suitable for many possible sampling designs, propensity estimation models, and target parameters. In principle, the proposed algorithms can be applied for not only the proposed pseudo weights but also other selection bias correction methods which focus on propensity estimation. It is important to note that the validity of the estimated confidence interval depends on the unbiasedness of the point estimate as shown in the simulation study. In the simulation, we illustrate the algorithm with unequal probability sampling. Other sampling methods, for example multi-stage sampling, can also apply the proposed algorithms with some adjustments. See Mashreghi et al. (2016) and Chauvet (2007) for detail. Other than the pseudo population bootstrap, the Bayesian bootstrap may also be a good alternative to understand the uncertainty (Dong, Elliott, & Raghunathan, 2014; Little & Zheng, 2007), while research is needed in order to be suitable for the two-sample setup.

	f_{NP}	f_P	Naive	w_{ind}	w_{dep}	$w_{ind,2}$	EV	CLW	VD	WVL	KW
Ind	0.05	0.01	20.22	0.12	0.10	1.38	0.45	**-0.05**	2.07	0.48	20.22
		0.10	20.21	**-0.00**	-0.00	1.46	0.76	-0.02	2.11	0.77	-0.20
	0.30	0.01	15.36	**-0.01**	0.02	3.94	0.50	-0.23	12.16	0.52	-1.48
		0.10	15.35	0.04	0.04	4.55	1.25	0.03	5.69	1.28	**-0.01**
	0.50	0.01	11.43	-0.04	**0.03**	4.06	0.37	-0.33	-0.91	0.38	-0.48
		0.10	11.45	0.01	0.01	4.43	0.85	**-0.00**	5.30	0.87	-0.01
Dep	0.05	0.01	20.23	1.15	0.30	1.34	0.41	**-0.12**	2.00	0.44	0.47
		0.10	19.69	1.24	0.37	1.45	0.77	**0.05**	2.11	0.77	-2.36
	0.30	0.01	15.31	2.88	**0.16**	3.95	0.53	-0.17	12.13	0.55	15.30
		0.10	14.72	3.19	0.58	4.38	1.23	**0.25**	5.48	1.24	-6.06
	0.50	0.01	11.37	2.13	**-0.21**	4.04	0.37	-0.30	-0.46	0.39	7.81
		0.10	10.78	2.17	**0.18**	4.16	0.82	0.19	4.97	0.83	-9.23

TABLE 2.1: The Relative Bias (%) of compared methods. *Ind* stands for independent samples, and *Dep* stands for dependent samples. Different inclusion fractions are shown. The smallest absolute relative bias of each scenario is bolded.

	f_{NP}	f_P	Naive	w_{ind}	w_{dep}	$w_{ind,2}$	EV	CLW	VD	WVL	KW
Ind	0.05	0.01	2379	**246**	248	285	249	341	380	253	2379
		0.10	2377	**237**	238	278	243	245	332	243	321
	0.30	0.01	1805	**122**	146	475	150	240	1851	157	845
		0.10	1803	**73**	77	538	163	90	671	166	106
	0.50	0.01	1343	**98**	158	486	141	229	2581	149	294
		0.10	1345	**46**	61	521	112	73	623	115	65
Dep	0.05	0.01	2379	282	**258**	292	259	335	373	264	3446
		0.10	2317	262	**228**	275	240	241	330	241	664
	0.30	0.01	1799	356	**138**	476	149	236	1837	156	1798
		0.10	1729	380	98	517	160	**95**	646	162	735
	0.50	0.01	1335	266	**138**	483	139	225	2415	148	1326
		0.10	1266	259	**57**	491	108	73	585	110	1092

TABLE 2.2: The RMSE of compared methods. *Ind* stands for independent samples, and *Dep* stands for dependent samples. Different inclusion fractions are shown. The smallest RMSE of each scenario is bolded.

		Ind	$D = 1$	Ind	$D = 10$	Dep	$D = 1$
f_{NP}	f_P	RB	CR	RB	CR	RB	CR
0.05	0.01	-7.81	94.10	10.20	94.02	398.00	96.80
	0.10	11.61	96.40	-1.75	94.76	-2.98	95.10
0.30	0.01	2.58	94.70	1.61	94.51	-2.00	93.60
	0.10	0.86	95.00	1.54	94.53	4.16	87.60
0.50	0.01	3.06	95.40	-2.44	93.83	-5.24	93.60
	0.10	-6.91	93.80	-4.16	94.02	-2.05	93.60

TABLE 2.3: The Relative Bias (%) of the estimated variances and the Coverage Rate of the confidence intervals (%) of the variance estimates. *Ind* refers to independent samples where the variances of w_{ind} are estimated by Algorithm 1. *Dep* refers to dependent samples where the variances of w_{dep} are estimated by Algorithm 2. D is the number of the pseudo populations.

Chapter 3

Performance Measures for Sample Selection Bias Correction by Weighting

Abstract

When estimating a population parameter by a nonprobability sample, i.e., a sample without a known sampling mechanism, the estimate may suffer from sample selection bias. To correct selection bias, one of the often-used methods is assigning a set of unit weights to the nonprobability sample, and estimating the target parameter by a weighted sum. Such weights are often obtained with classification methods. However, a tailor-made framework to evaluate the quality of the assigned weights is missing in the literature, and the evaluation framework for prediction may not be suitable for population parameter estimation by weighting. We try to fill in the gap by discussing several promising performance measures, which are inspired by classical calibration and measures of selection bias. In this paper, we assume that the population parameter of interest is the population mean of a target variable. A simulation study and real data examples show that some performance measures

have a strong positive relationship with the mean squared error and/or error of the estimated population mean. These performance measures may be helpful for model selection when constructing weights by logistic regression or machine learning algorithms.

Keywords: Model Evaluation, Nonprobability Sample, Population Parameter Estimation, Data Integration

Liu, A.-C., Scholtus, S., Van Deun, K. & De Waal, T. Performance Measures for Sample Selection Bias Correction by Weighting. *Journal of Official Statistics* (Accepted)

3.1 Introduction

Probability samples have long been the gold standard for drawing reliable conclusions from the target population. However, probability samples often require much time and resources to collect. On the other hand, more and more naturally occurring data are available nowadays. For example, social media data, administrative data, or sensor data. These data sources are easier to collect in terms of time and cost or are already available due to digitalization (Cornesse et al., 2020). However, the inclusion mechanisms of these data sources are often unknown. Such data sets that have not been obtained through a known sampling mechanism are termed nonprobability samples. Without a known sampling mechanism, nonprobability samples are often treated as a simple random sample when estimating a population parameter (e.g., a population mean), and therefore may suffer from sample selection bias. When the inclusion mechanism of the nonprobability sample depends on the target variable we are interested in, selection bias is critical even with a large sample size (Meng, 2018). For example, during the COVID-19 pandemic, an intensive survey was conducted by Facebook to investigate COVID-related features. Over 250,000 participants in the U.S. filled in the survey which was invited by the Facebook pop-on ad. The participants had a self-selection process after they saw the ad, and this process may have been affected by the COVID-related features. Although the number of participants was massive, the vaccination uptake rate was overestimated by 17% compared to the official figure. The large sample size also resulted in a narrow estimated variance so that the confidence interval could hardly cover the true value (Bradley et al., 2021).

Intensive research on selection bias correction methods has appeared. In general, selection bias correction methods can be categorized into y-modeling (modeling the target variables), weighting, and combining y-modeling and weighting (e.g., doubly

robust estimation). See Elliott and Valliant (2017); Meng (2022); J. Rao (2020); Wu (2022) for reviews. Here we focus on the weighting methods. In a weighting method, a set of unit weights is derived from, for example, inverse inclusion propensities (i.e. inclusion probabilities) or calibration given some estimated or known population values of auxiliary variables. The population parameter of interest is then estimated by a design-based estimator, e.g., the Horvitz-Thompson estimator or Hájek estimator (Hájek, 1971; Horvitz & Thompson, 1952).

Such weights can be constructed in many different ways. The main aim of this paper is to select the best approach for constructing weights for a nonprobability sample out of a set of candidate approaches. However, a tailor-made model evaluation framework for the constructed weights is missing in the literature. Given the nature of selection bias correction frameworks, model evaluation methods for common statistical analyses, such as scoring rules in prediction (Gneiting & Raftery, 2007), may not be suitable, since the interest is in having an unbiased estimated population parameter instead of a perfect unit-level prediction of the inclusion propensity. The relation between the constructed weights and the target variable will affect the performance of the weights (Meng, 2018). For example, as an extreme case, if the target variable is a constant for every unit in the population, the performance of any constructed set of weights should be the same (assuming the weights sum up to the number of units in the population), while many often-used model evaluation indexes such as AIC fail to reflect this.

Besides, finding the correct propensity model for the nonprobability sample is not necessarily the goal when correcting for selection bias, similar to that it is not necessary to find the correct imputation model when imputing missing data (Vidotto, Kaptein, & Vermunt, 2015). The inclusion mechanism of the nonprobability sample at hand may be unique. Without any strong reason, it is hard to believe that the acquired model can be applied to any other nonprobability sample. Having

a correct model can assist in deriving many nice properties, as noted in Wu and Thompson (2020). However, it is hard to know even if the correct model exists or is considered in the candidate set of models (Zhang, 2019). Instead of trying to find the correct model, we try to find the best model out of the candidate set of models by a performance measure. That is, a performance measure that reflects the underlying Mean Squared Error (MSE) and/or error of the estimated parameter. Ideally, that measure has a strong correlation with the MSE or error and we will be able to choose the best model based on the measure. We may then conduct variable selection or model selection, which is especially critical for the weighting method to prevent the variance of the correction method from outweighing the corrected bias. Literature suggests that only auxiliary variables that have strong relations with the target variable should be considered, while how to perform variable selection is not clear (Brick, 2013; Mercer, Kreuter, Keeter, & Stuart, 2017).

In the following sections, we will start by discussing the background in Section 3.2, and some possible performance measures for selection bias correction are described in Section 3.3. A simulation study and examples of real data sets will follow in Sections 3.4 and 3.5. Section 3.6 ends this paper with a discussion and by drawing some conclusions.

3.2 Background

Before discussing the performance measures, it is useful to discuss the source of the selection bias and the mechanism of weighting methods. The discussion will focus on taking the population mean as the parameter of the target variable, while it can also be extended to more than one target variable or other parameters that are a linearizable function of the population mean.

3.2.1 Selection Bias

We assume that we are interested in a finite population (U) with index $i \in \{1, 2, ..., N\}$. The population mean of the target variable y, $\mu = N^{-1} \sum_{i \in U} y_i$, is the parameter of interest. We also assume that we have observed a nonprobability sample (NP) of size n_{NP} where $NP \subset U$. If NP is treated as a simple random sample without replacement from the population and used to estimate the population mean, the error in the nonprobability sample can be expressed as (Meng, 2018)

$$\text{Error}(\bar{y}_{NP}) = \bar{y}_{NP} - \mu = \frac{1}{n_{NP}} \sum_{i \in NP} y_i - \mu = \frac{\text{Cov}(s, y)}{\bar{s}}, \tag{3.1}$$

where s is the inclusion indicator of the nonprobability sample, that is $s_i = 1$ if $i \in NP$ and 0 otherwise, and the population mean of the inclusion indicator $\bar{s} = N^{-1} \sum_{i \in U} s_i$. The population covariance of s and y is

$$\text{Cov}(s, y) = \frac{1}{N} \sum_{i \in U} (s_i - \bar{s})(y_i - \mu).$$

Assuming that the nonprobability sample is drawn from the population by means of some sampling mechanism (b) with unknown inclusion propensities $\mathbb{P}(s_i = 1)$ ($i \in U$), we can find the bias of ($\bar{y}_{NP}$) by taking the expectation of (3.1) over repeated sampling (E_b). We then get

$$\text{Bias}(\bar{y}_{NP}) = \text{E}_\text{b} \left(\frac{\text{Cov}(s, y)}{\bar{s}} \right). \tag{3.2}$$

Take the vaccination rate case in the Introduction as an example, the target variable y_i is whether a person is vaccinated and s_i is whether a person responds to the Facebook survey. If a vaccinated person has a higher or lower tendency to

respond to the Facebook survey, selection error will occur in the estimated vaccination rate. Neither Error($\bar{y}_{NP}$) nor Bias($\bar{y}_{NP}$) can be estimated by the nonprobability sample only. Even if $\mathbb{P}(s_i = 1)$ is known for units in NP, we still need to assume that the relationship between s and y in the nonprobability sample is the same as the relationship between s and y in the population (Nishimura, Wagner, & Elliott, 2016).

3.2.2 Correcting Selection Bias by Weighting

When correcting selection bias by weighting, often the target parameter μ is estimated by a design-based estimator $\bar{y}_w = \sum_{i \in NP} w_i y_i / \sum_{i \in NP} w_i$ where w_i is a weight assigned to unit i. The usage of a design-based estimator implies that we assume all units in the population have a non-zero probability of being included in the nonprobability sample. Our goal in this paper is to obtain a set of weights that minimizes MSE($\bar{y}_w$) = $\mathrm{E_b}[\bar{y}_w - \mu]^2$. Besides minimizing the MSE, it is often also of interest to minimize the error and the bias incurred in estimating the population mean. The error can be expressed as (Meng, 2018, 2022)

$$\mathrm{Error}(\bar{y}_w) = \bar{y}_w - \mu = \frac{\mathrm{Cov}(sw, y)}{\overline{sw}} \tag{3.3}$$

where $\overline{sw} = N^{-1} \sum_{i \in U} s_i w_i$ and the bias as

$$\mathrm{Bias}(\bar{y}_w) = \mathrm{E_b}\left(\frac{\mathrm{Cov}(sw, y)}{\overline{sw}}\right) \tag{3.4}$$

Construct Weights by Inverse Propensity Estimation

From (3.4) we can see that the bias can be corrected if the constructed weights satisfy $w_i \propto \mathbb{P}(s_i = 1)^{-1}$, since then $\mathrm{E_b}(s_i w_i)$ becomes a constant and therefore $\mathrm{E_b}\left(\mathrm{Cov}(sw, y)\right)$ becomes zero no matter what the values of y are. That is, if the

true inclusion propensities are used, the bias vanishes for any target variable, similar to what normally happens in a design-based estimator. However, as mentioned in the Introduction, it is hard to know whether true inclusion propensities are obtained.

Construct Weights by Calibration

Besides propensity weighting, another often used correction method is calibration, see for example, Y. Chen et al. (2020); Kim and Wang (2019); Yang et al. (2020), and also a review in Wu (2022). Unlike inverse propensity weighting, it does not consider the underlying inclusion mechanism of the nonprobability sample but merely tries to obtain a set of weights that allows the weighted sum of the values of the target variable y observed in the nonprobability sample to be equal or close to the population total of y. The constructed weights are only valid for the target variable y under consideration but are not necessarily valid for other target variables. To construct the weights, a set of auxiliary variables $\mathbf{x}$ with known or estimated population totals is needed. Ideally, the relations between the auxiliary variables and the target variable are strong so that if a set of weights can (approximately) obtain the known totals of the auxiliary variables, it can also assist in obtaining the population total of the target variable. The relation between y and $\mathbf{x}$ can usually only be observed in the nonprobability sample. An assumption is needed that the relation between y and $\mathbf{x}$ is the same when $s = 1$ and $s = 0$ so that $f(y|\mathbf{x}, s) = f(y|\mathbf{x})$ given the density function $f(\cdot)$ for y or $\mathbb{P}(s|\mathbf{x}, y) = \mathbb{P}(s|\mathbf{x})$ (Little et al., 2020). Instead of directly constructing the weights by $\mathbf{x}$, an alternative may be to fit the model $\mathbb{P}(s|\mathbf{x}, y)$ with $\mathbb{P}(s|\mathbf{x}, \hat{y})$ or $\mathbb{P}(s|\hat{y})$. That is, construct the weights with the assistance of a model for y. This y-model should ideally be correctly specified (Marella, 2023). Of course, the y-model may not always be correctly specified. Later in the simulation we also explore the scenario when the y-model is incorrect.

3.3 Performance Measures for Selection Bias Correction

In this section, we discuss some possible measures to evaluate a set of weights for selection bias correction. A performance measure is a nonnegative function of the weights, which has positive linear dependence with the mean squared error of the estimator of the finite population parameter of interest, constructed using these weights. That is, a measure that can give a good indication of the underlying unknown MSE($\bar{y}_w$) or Error($\bar{y}_w$) defined in Section 3.2.2. All the measures we present here are expected to have a positive relation with MSE($\bar{y}_w$) and/or absolute Error($\bar{y}_w$).

The performance measures are presented under the two-sample setup, which has often been used in the selection bias correction literature, e.g., Y. Chen et al. (2020); Elliott and Valliant (2017). In the two-sample setup, along with the nonprobability sample, a probability sample (P) from the same population of size n_P is available. For both P and NP, the design weights $d_i = 1/\mathbb{P}(i \in P)$ and a common set of auxiliary variables $\mathbf{x}$ are available. Here we do not limit the possibility of whether the two samples are overlapping or not. The sample resulting by merging the nonprobability sample and the probability sample is denoted as S_c so that the size of S_c is $n = n_{NP} + n_P$, i.e. overlapping units (if any) are counted twice.

3.3.1 Measures without y-Model

Here we discuss some measures that are often used for probability estimation or model evaluation in general. Since propensity estimation may be applied to construct weights, one may wonder whether performance measures for probability estimation will be helpful for evaluating the propensities.

In the following, we first discuss the mean cross entropy (MXE) and Brier score under the pseudo-weight method from Elliott and Valliant (2017). That pseudo-weight method first estimates $\hat{\pi}_i = \mathbb{P}(s_i = 1|i \in S_c)$ and constructs the final weights with $w_i = d_i(1 - \hat{\pi}_i)/\hat{\pi}_i$, for details see Elliott and Valliant (2017). That is, the design weights of the probability sample are considered after modeling $\hat{\pi}_i$, which allows the probability estimation methods to be applied in a standard way for estimating π_i (i.e., without weighting the units by the design weights) and therefore many nonparametric or machine learning methods can be applied, for example, Bayesian Additive Regression Trees (BART) as proposed in Rafei et al. (2020); Rafei, Flannagan, West, and Elliott (2022) (these articles also offer a broader discussion of estimation for nonprobability samples).

MXE

For MXE and Brier score the performance of the model is evaluated on $\hat{\pi}_i$ instead of the estimated propensity to be included in the nonprobability sample $\hat{p}_i = 1/w_i$, since the goal is to maximize the impurity of the estimated $\hat{\pi}_i$ but not of the underlying propensities. MXE under the two-sample setup is (Caruana & Niculescu-Mizil, 2004; Kullback, 1997),

$$\text{MXE}(\hat{\pi}) = -\frac{1}{n}[\sum_{i \in NP} \log(\hat{\pi}_i) + \sum_{i \in P} \log(1 - \hat{\pi}_i)]. \tag{3.5}$$

A smaller value for MXE indicates better performance according to this measure. So, $\hat{\pi}_i$ closer to 0 or 1 will be preferable by MXE.

Brier's Score

A similar measure is Brier's score which is a distance-based measure. A smaller value of the Brier score reflects a smaller distance between the $\hat{\pi}_i$ and the s_i and

therefore $\hat{\pi}_i$ close to 0 or 1 is also preferred. The formula is (Brier et al., 1950)

$$\text{Brier}(\hat{\pi}, s) = \frac{1}{n}\sum_{i \in S_c}(\hat{\pi}_i - s_i)^2. \tag{3.6}$$

AIC

For model selection, the Akaike information criterion (AIC) is one of the often-used measures (Akaike, 1974; Schwarz, 1978). AIC is based on the value of the likelihood function $\hat{L}$ of the estimated model with a penalty on the used number of parameters (k) of the model,

$$\text{AIC}(k, \hat{L}) = 2k - 2\ln(\hat{L}). \tag{3.7}$$

An AIC for complex design survey data has also been proposed by Lumley and Scott (2015). For many machine learning methods, it is difficult or even impossible to calculate AIC since the likelihood function is unknown, and sometimes even the number of parameters is unknown as well (e.g., a tree model).

Cal1

As noted in Section 3.2.2, calibration is an often-used method for selection bias correction. The calibration property may be suitable for not only constructing the weights but also serving as a performance measure. Based on the calibration property, we may examine the performance of the weights by auxiliary variables which are strongly correlated to the target variable (Deville & Särndal, 1992). If the differences between weighted totals of the auxiliary variables and the corresponding known totals are small, we may conclude that we have a good set of weights. Under the two-sample setup, the population means of the auxiliary variables can be estimated from the probability sample by $\sum_{i \in P} d_i \mathbf{x}_i / \sum_{i \in P} d_i$. Therefore we can

calculate

$$\text{Cal1}(w,d) = \sum_{j=1}^{J} |\frac{\sum_{i\in NP} w_i x_{ij}}{\sum_{i\in NP} w_i} - \frac{\sum_{i\in P} d_i x_{ij}}{\sum_{i\in P} d_i}|, \tag{3.8}$$

which is the sum of the absolute differences between the weighted and design-based estimates of the population means over J auxiliary variables from the two samples. A set of weights with the smallest value of Cal1 will then be chosen. This approach has been applied by Yang et al. (2020) as a loss function for weight construction and as a performance measure for tuning. Since Cal1 is sensitive to the scale of the auxiliary variables, it may be standardized by

$$\sum_{j=1}^{J} \frac{1}{\sigma_j} |\frac{\sum_{i\in NP} w_i x_{ij}}{\sum_{i\in NP} w_i} - \frac{\sum_{i\in P} d_i x_{ij}}{\sum_{i\in P} d_i}|, \tag{3.9}$$

where σ_j is the standard deviation of x_j and may be estimated by $((\sigma^2_{j,NP}+\sigma^2_{j,P})/2)^{0.5}$. $\sigma^2_{j,NP}$ and $\sigma^2_{j,P}$ are the estimated variances of x_j from the nonprobability sample and the probability sample, where $\sigma^2_{j,NP}$ is estimated by assuming that NP was obtained by a simple random sample, which may not be a realistic assumption in practice. (3.9) has been applied in Austin (2009); Kern, Li, and Wang (2021); McCaffrey, Ridgeway, and Morral (2004).

3.3.2 Measures with y-Model

Cal2 and Cal3

As noted in Section 3.2, it may be useful to also consider a y-model when using the weighting approach. The first two measures with y-models that we consider are transformations of (3.8). An underlying assumption of (3.8) and (3.9) is that the auxiliary variables and the target variable have a linear relation, which may not be met in practice (Deville & Särndal, 1992). As an alternative, Wu and Sitter (2001) proposed a model-calibration method that allows all types of relationships

between the auxiliary variables and the target variable. Rather than using $\mathbf{x}$ as in (3.8), a model $\hat{y} = \hat{m}(\mathbf{x})$ is fitted on the nonprobability sample, where $\hat{m}(\cdot)$ can be any function. The weights are then evaluated by the weighted total of $\hat{y}$ in the two samples, that is, the performance measure will be

$$\text{Cal2}(w, d) = |\frac{\sum_{i \in NP} w_i \hat{y}_i}{\sum_{i \in NP} w_i} - \frac{\sum_{i \in P} d_i \hat{y}_i}{\sum_{i \in P} d_i}|. \tag{3.10}$$

We also look at the difference between the weighted sum of the observed y and the weighted sum of $\hat{y}$,

$$\text{Cal3}(w, d) = |\frac{\sum_{i \in NP} w_i y_i}{\sum_{i \in NP} w_i} - \frac{\sum_{i \in P} d_i \hat{y}_i}{\sum_{i \in P} d_i}|. \tag{3.11}$$

With the usage of y in the nonprobability sample, (3.11) may be less subject to model misspecification compared to (3.10).

MSB

One way to estimate the selection bias is the Measure of Unadjusted Bias (MUB) proposed by Little et al. (2020), which may also be useful for evaluating the performance of the acquired weights. In Boonstra, Little, West, Andridge, and Alvarado-Leiton (2021), it is shown that MUB outperforms other measures such as the R-indicator, Coefficient of Variation (CV), or Area Under the receiver-operating characteristic Curve (AUC) (for details on these measures, see Boonstra et al. (2021)) in reflecting the amount of selection bias. MUB aims to estimate $\bar{y}_{NP} - \mu$ by assuming the inclusion mechanism is from a function of $\mathbb{P}(s = 1|\mathbf{x}, y) = g[(1 - \phi)\hat{y} + \phi y]$, where $\phi \in [0, 1]$ is an unknown model parameter that allows different degrees of ignorability to be considered, and $g[\cdot]$ is some function. A model $\hat{y} = \hat{m}(\mathbf{x})$ is fitted on the nonprobability sample, and $\hat{m}(\mathbf{x})$ is used in the probability sample to calculate

$\hat{y}$. When $\phi = 1$, MUB(ϕ) completely depends on the observed y in the nonprobability sample, and when $\phi = 0$, MUB(ϕ) completely depends on $\hat{y}$, which is aligned with Cal3, see Little et al. (2020) for details. The definition of MUB(ϕ) under the two-sample setup is

$$\text{MUB}(\phi) = \frac{\phi + (1-\phi) r_{\hat{y}y}}{\phi r_{\hat{y}y} + (1-\phi)} \sqrt{\frac{\sigma_y^2}{\sigma_{\hat{y}}^2}} (\bar{\hat{y}}_{NP} - \bar{\hat{y}}_U), \tag{3.12}$$

where $\bar{\hat{y}}_{NP} = n_{NP}^{-1} \sum_{i \in NP} \hat{y}_i$, $\bar{\hat{y}}_U = \sum_{i \in P} d_i \hat{y}_i / \sum_{i \in P} d_i$, $r_{\hat{y}y}$ is the correlation coefficient of $\hat{y}$ and y, i.e.

$$r_{\hat{y}y} = \frac{\sum_{i \in NP} (y_i - \bar{y}_{NP})(\hat{y}_i - \bar{\hat{y}}_{NP})}{n_{NP} \sigma_{\hat{y}} \sigma_y},$$

and σ_y^2, $\sigma_{\hat{y}}^2$ are estimated by $\hat{\sigma}_y^2 = n_{NP}^{-1} \sum_{i \in NP} (y_i - \bar{y}_{NP})^2$ and $\hat{\sigma}_{\hat{y}}^2 = n_{NP}^{-1} \sum_{i \in NP} (\hat{y}_i - \bar{\hat{y}}_{NP})^2$. Note that in this setup we assume that the auxiliary variables used to obtain $\hat{y}$ are not available outside of the two samples. If the population totals of the auxiliary variables are available, MUB may give a more accurate error estimation since using the estimated population value is naturally losing efficiency compared to a known population parameter (Zhang, 2019). A performance measure that borrows the strength of MUB(ϕ) may be:

$$\text{MSB}(\phi) \equiv |\text{MUB}(\phi) - \bar{y}_{NP} + \bar{y}_w|. \tag{3.13}$$

That is, if the difference between the naive estimate $\bar{y}_{NP}$ for y and the weighted mean $\bar{y}_w$ is close to MUB(ϕ), we may conclude that the acquired set of weights can correct the underlying selection bias.

It is worth noting that, since μ is fixed, $\bar{y}_w$ naturally has a perfect positive relationship with Error($\bar{y}_w$), see (3.3). We can have an idea of the direction of the

error by merely looking at whether the estimated $\bar{y}_w$ is moving away from or towards $\bar{y}_{NP}$. However, merely looking at $\bar{y}_w$ does not prevent us from over-correction, i.e., when $\bar{y}_{NP} - \mu$ has an opposite sign to Error($\bar{y}_w$). We hope to understand whether the selection error is over-corrected by considering the MUB(ϕ) in the measure. If Error($\bar{y}_w$) is zero, ideally the value of MSB should also be zero.

KS

Another model evaluation index in the nonresponse literature is the Kolmogorov-Smirnov (KS) distance (Chambers, 2001). KS distance is a non-parametric index that calculates the maximum difference between two empirical cumulative distribution functions. Unlike AIC, KS can be applied to the result from any model. Under the two-sample setup, we calculate the maximum difference by

$$\text{KS}(w, d) = \max_t \left| \frac{\sum_{i \in NP} w_i \text{I}(y_i \leq t)}{\sum_{i \in NP} w_i} - \frac{\sum_{i \in P} d_i \text{I}(\hat{y}_i \leq t)}{\sum_{i \in P} d_i} \right| \tag{3.14}$$

for $t \in (-\infty, \infty)$, where $\text{I}(\cdot)$ is an indicator function.

3.4 Simulation

3.4.1 Simulated Data

We evaluate the performance measures by examining the relation of the performance measures with MSE($\bar{y}_w$) and absolute Error($\bar{y}_w$). A population of size 10,000 with auxiliary variables $x_1, x_2, x_3, x_4 \sim N(1, 1)$ is created. The target variable $y = 3x_1 + x_2 - 5x_3 + 0.1x_4 + e$, where $e \sim N(0, 1)$. The finite population mean of y is $\mu \approx -0.9$. The auxiliary variables are available both for the probability and nonprobability samples while the target variable is only available in the nonprobability sample. The probability sample is repeatedly drawn by means of simple random sampling without

replacement with inclusion probability 0.05, that is, $d = 20$ is the design weight for all units in the probability sample, and results in $n_P = 500$. The nonprobability sample is repeatedly drawn by means of fixed-size unequal probability sampling without replacement. We do this by randomized systematic sampling with inclusion probability $p = \exp(c + x_1 + x_2 - 0.5x_3)/(1 + \exp(c + x_1 + x_2 - 0.5x_3))$, where c is a constant so that the inclusion fraction of the nonprobability sample is fixed at 0.1, which results in $n_{NP} = 1,000$ (Madow, 1949). The result of a nonlinear propensity model $p = \exp(c + x_1 + 0.5x_2x_4)/(1 + \exp(c + x_1 + 0.5x_2x_4))$ is shown in Section 3.7, which shows a similar conclusion as the linear one.

3.4.2 Estimation and Evaluation

The weights are constructed by Elliott and Valliant's (2017) pseudo-weight method as discussed in Section 3.3.1, since Elliott and Valliant's method offers a relatively stable estimation compared to methods considering design weights during the propensity model estimation (Liu, Scholtus, & De Waal, 2023). To reflect different possible model choices, the propensity model $\hat{\pi}_i = \mathbb{P}(s_i = 1 | i \in S_c)$ is fitted by a machine learning algorithm, `XGBoost`, and logistic regression (T. Chen & Guestrin, 2016).

`XGBoost` is a flexible and powerful algorithm in prediction problems, and it has been applied in Castro-Martín, Rueda, and Ferri-García (2020) and Klingwort and Burger (2023) for selection bias correction. As for many machine learning algorithms, hyperparameters should be chosen before fitting the `XGBoost` model (see, e.g., T. Chen and Guestrin (2016) for these hyperparameters). In the simulation, we use the default hyperparameters in `XGBoost`. A more detailed tuning scheme will be performed later in the real data examples. Note that the AIC cannot be calculated for `XGBoost` since the number of parameters and the likelihood function

are unknown.

For logistic regression, the correct model and 33 incorrectly specified or over-specified models are fitted. These incorrectly specified or over-specified models, for example, miss some auxiliary variables, have some extra interactions between variables, or have higher-order terms of the variables. The incorrectly specified models may reflect the effect of a Not Missing At Random mechanism. See the online supplement for details on the models used or Table 3.1 for a few examples.

In total, 35 models/methods are used to estimate the propensity scores to reflect the relation between the measures and different degrees of the estimated MSE of the estimated population mean, which is $\widehat{\text{MSE}}(\bar{y}_w) = R^{-1}\sum_{r=1}^{R}[\bar{y}_{w,r} - \mu]^2$, where $\bar{y}_w = \sum_{i\in NP} w_i y_i / \sum_{i\in NP} w_i$, and $R = 1,000$ is the number of replicates in drawing a probability and nonprobability sample. The averages of the performance measures under each model are recorded.

For measures considering a y-model as in Section 3.3.2, we apply linear regression with the correct model (using x_1, x_2, x_3, x_4 as the auxiliary variables) and an incorrect model (using x_2 and x_4 only) to show the effect of different model use. The adjusted R^2 of the correct model in the nonprobability sample is around 0.967, and the adjusted R^2 of the incorrect model is around 0.011.

3.4.3 Results

Figures 3.1 and 3.2 show the relations between each measure and $\widehat{\text{MSE}}(\bar{y}_w)$ for the 35 specified models. The number after MSB is the value of ϕ, e.g, MSB05 indicates that $\phi = 0.5$ is used. Figure 3.1 shows that in general all measures are positively related to $\widehat{\text{MSE}}(\bar{y}_w)$ under the correct y-model. Figure 3.2 shows the effect of an incorrect y-model, therefore only measures with a y-model are shown. For Cal2, the relation is only clear when the correct y-model is used, while under the incorrect y-model,

a low correlation between Cal2 and $\widehat{\text{MSE}}(\bar{y}_w)$ is observed. Cal3 and MSB0 show the opposite relation between the measures and $\widehat{\text{MSE}}(\bar{y}_w)$. However, if the unknown parameter ϕ is well chosen, between 0.5 and 0.75 in this case, zero estimated MSB is then corresponding to zero $\widehat{\text{MSE}}(\bar{y}_w)$. KS has a negative relation with $\widehat{\text{MSE}}(\bar{y}_w)$ when the wrong y-model is applied.

MXE and Brier have a similar tendency since $\hat{\pi}$ is mostly around 0.7 and the difference between these performance measures will only be obvious when the estimated probability is close to 0 or 1. `XGBoost` indeed gives a good estimation in terms of impurity (low MXE and Brier), however, low impurity does not necessarily guarantee a good population parameter estimate. In fact, when the estimated propensity is close to 0, although this results in a low impurity, this also causes a large weight and a large variation of the parameter estimates.

In Table 3.1 we list the best 10 models in terms of Bias $= R^{-1}\sum_{r=1}^{R}\bar{y}_{w,r} - \mu$. It is interesting to see that the correct propensity model does not necessarily perform the best in terms of both Bias and MSE. Some overfitting models may capture the underlying variation and allow a better parameter estimation. A similar discussion can also be found in the imputation literature, see, e.g., Vermunt, Van Ginkel, Van der Ark, and Sijtsma (2008); Vidotto et al. (2015).

3.4.4 Selecting Smallest Error

We also examine whether the performance measures are able to pick out the best model in terms of absolute Error$(\bar{y}_w)$. In every set of the drawn samples, 35 models are fitted as before. Kendall rank correlation coefficient (τ) between absolute Error$(\bar{y}_w)$ of the 35 models and each measure is calculated to reflect whether the measures are able to rank the models correctly (Kendall, 1948). Kendall τ is calculated by the probability of the same order of pairs of two units of a variable,

FIGURE 3.1: Relations between $\widehat{\text{MSE}}(\bar{y}_w)$ and the mean performance measures under different propensity model specifications. The correct y-model is used. The dots are the estimates from logistic regression and the cross is the estimate from `XGBoost`. Measures include different variants of calibration (Cal1, Cal2, Cal3), measure of selection bias (MSB) with different ϕ, Mean cross entropy (MXE), Brier score, Akaike information criterion (AIC), and Kolmogorov-Smirnov distance (KS).

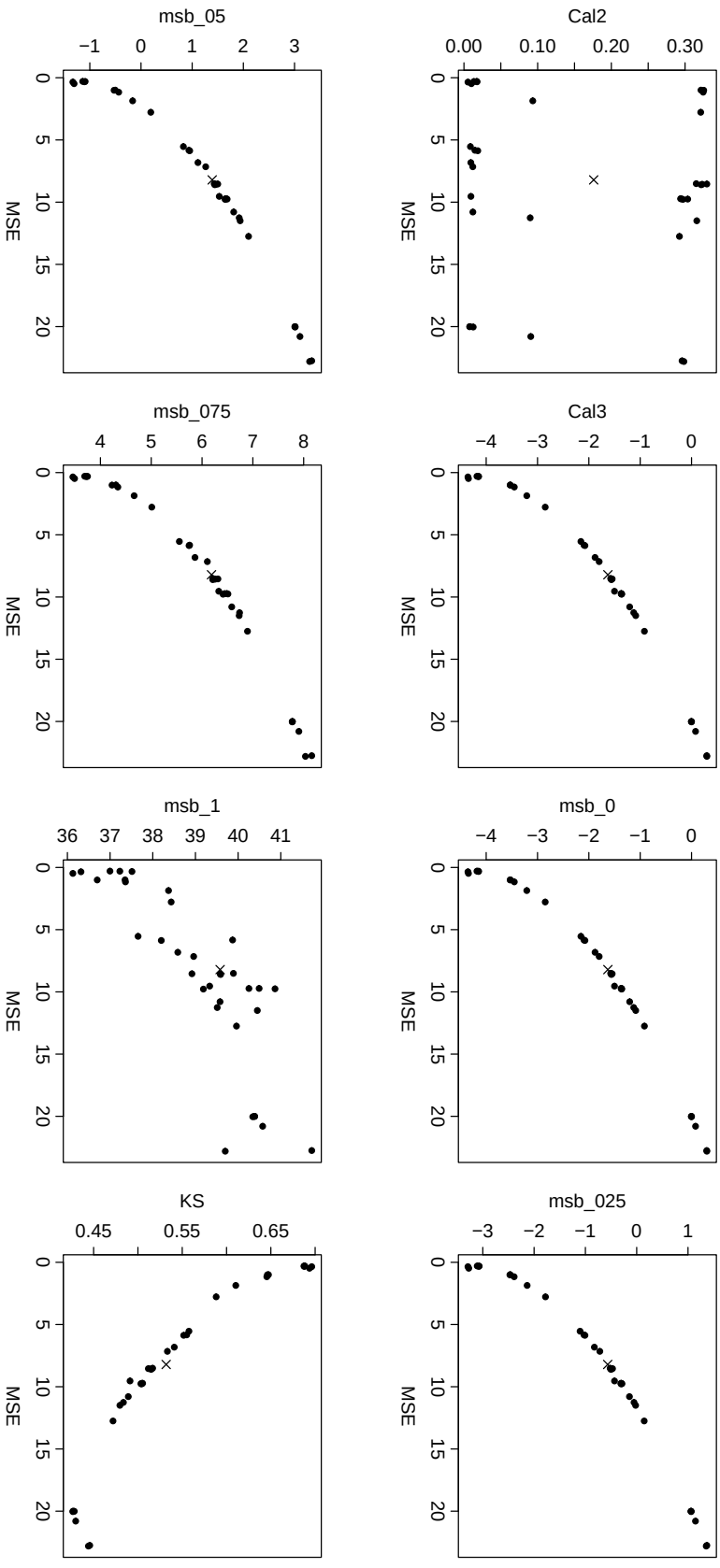

FIGURE 3.2: Relations between $\widehat{\text{MSE}}(\bar{y}_w)$ and mean performance measures when the incorrect y-model is used. The dots are the estimates from logistic regression and the cross is the estimate from `XGBoost`. Measures with a y-model include two variants of calibration (Cal2, Cal3), measure of selection bias (MSB) with different ϕ, and Kolmogorov-Smirnov distance (KS).

TABLE 3.1: Best 10 propensity model ranked by Bias. The bold model is the correct model for drawing the nonprobability sample, and the bold Bias and MSE are the smallest values.

Propensity Model	Bias	MSE
$(x_1 + x_2 + x_3 + x_4)^2$	**0.12**	0.48
$(x_1 + x_2 + x_3)^2$	0.12	0.35
$x_2 + x_4 + x_1 * x_3$	0.29	**0.30**
$\mathbf{x_1 + x_2 + x_3}$	0.32	0.30
$x_1 + x_2 + x_3 + x_4$	0.32	0.33
$x_1 * x_3$	0.94	1.00
$(x_1 + x_3)^2$	0.95	1.01
$x_1 + x_3$	1.03	1.16
$x_1 + x_2^2 + x_3^2$	1.27	1.86
$x_1 + x_3^2$	1.62	2.78

subtracting the probability of different order of pairs of two units, that is, how consistent the orders between the two variables are (here the performance measure and absolute Error). Kendall $\tau = 1$ if two variables share the same rank, and $\tau = -1$ if two variables have totally opposite rankings. That is, if $\tau = 1$ for a performance measure and absolute Error($\bar{y}_w$), it means that in every possible subset of the 35 models, we will be able to pick out the best model based on the value of the performance measure. The averages of τ for all measures over a thousand runs are reported.

Tables 3.2 and 3.3 show the average Kendall rank correlation coefficient between absolute Error($\bar{y}_w$) and each measure. In general, measures considering a y-model are strongly correlated with the actual underlying error, that is, we are able to pick out the smallest error model based on Cal2, Cal3, MSB, and KS under the correct y-model. Under the incorrect y-model, only MSB with $\phi \in [0.75, 1]$ may give a good indication.

TABLE 3.2: Kendall τ between absolute Error($\bar{y}_w$) and each measure under the correct y-model. Measures include different variants of calibration (Cal1, Cal2, Cal3), measure of selection bias (MSB) with different ϕ, Mean cross entropy (MXE), Brier score, Kolmogorov-Smirnov distance (KS), and Akaike information criterion (AIC)

	Cal1	Cal2	Cal3	MSB0	MSB0.5	MSB1	MXE	Brier	KS	AIC
τ	0.66	0.98	0.99	0.99	0.99	0.99	0.59	0.59	0.91	0.66

TABLE 3.3: Kendall τ between absolute Error($\bar{y}_w$) and each measure under the incorrect y-model. Measures with a y-model include two variants of calibration (Cal2, Cal3), measure of selection bias (MSB) with different ϕ, and Kolmogorov-Smirnov distance (KS).

	Cal2	Cal3	MSB0	MSB0.25	MSB0.5	MSB0.75	MSB1	KS
τ	0.02	-0.96	-0.96	-0.54	0.61	0.97	0.99	-0.89

3.5 Experiments on Real Data Sets

The experiment looks at the performance measures under various real data sets. Since in practice the true models for target variable y and the inclusion mechanism of the nonprobability sample are usually unknown to the investigator, we try to mimic this situation in the experiment. Three data sets in R packages are used, i.e., Iris data (Anderson, 1935), Election data from the Survey package (Lumley, 2020), and MU284 data from the Sampling package (Tillé & Matei, 2021). See the references therein for the details of the data sets. These data sets are treated as populations where one of the variables is treated as the indicator of inclusion in the nonprobability sample, one continuous variable is treated as the target variable of which the population mean is of interest, and the rest of the variables are the auxiliary variables. The variable specification is shown in Table 3.4.

Since the propensity model of the nonprobability sample is unknown in this case, we may only look at absolute Error($\bar{y}_w$) but not MSE($\bar{y}_w$). A simple random sample with an inclusion fraction of 0.1 is drawn from the population and treated as the available probability sample. The construction of the weights is again the

TABLE 3.4: Data sets and the used variables in the experiment

Data	**Inclusion Indicator**	**Y**	**X**
Iris	$s_i = 1$ if *Species* is *setosa*, else $s_i = 0$	*Sepal.Length*	*Sepal.Width*, *Petal.Length*, *Petal.Width*
Election	$s_i = 1$ if $p < 0.005$, else $s_i = 0$	*TotPrecincts*	*PrecinctsReporting*, *Bush*, *Kerry*, *Nader*, *votes*
MU284	$s_i = 1$ if $CL < 15$, else $s_i = 0$	*P85*	*P75*, *RMT85*, *CS82*, *SS82*, *S82*, *ME84*, *REV84*, *REG*

pseudo-weight method from Elliott and Valliant (2017) where `XGBoost` is used for the propensity estimation. We grid search a range of tuning parameters of `XGBoost` under reasonable ranges including

- Learning rate (η): Learning rate is between 0 and 1, and the default value is 0.3. A higher learning rate means a larger contribution of each tree. A set of $\{0, 0.3, 0.5, 0.7\}$ is used.

- Minimum loss (γ): The minimum loss that needs to be reduced when partitioning. It ranges from 0 to ∞ and the default value is 0. Here we use $\{0, 1, 2\}$.

- Minimum child weight: The minimum of the sum of weights in a child node. It ranges from 0 to ∞ and the default value is 1. $\{1, 3, 5, 7\}$ is used.

In total, there are $4 \times 3 \times 4 = 48$ combinations of the tuning parameters. Error($\bar{y}_w$) of each combination is calculated, and the relations between Error($\bar{y}_w$) and the measures are shown. Note that, unlike a prediction task, we do not use cross-validation for tuning since the goal is not to find a model for future prediction but to estimate population parameters by means of the probability sample and the nonprobability sample. The y-model is fitted by linear regression, using all the auxiliary variables as predictors.

Figures 3.3, 3.4, and 3.5 show the relationships between the performance measures and absolute Error($\bar{y}_w$). In general, the patterns are similar to those in the simulation. The measures with a y-model for target variable y better reflect the underlying absolute Error($\bar{y}_w$). It also can be seen that the three data sets offer sufficient auxiliary information so that Cal2, Cal3, and MSB reflect clear indications of model performance, while Cal1 may not be useful when some auxiliary variables have a negative correlation with the target parameter. MXE, brier, and KS show a positive relationship with absolute Error($\bar{y}_w$) in Iris data but that is not the case in Election and MU284 data.

3.6 Conclusion and Discussion

Weighting is one of the popular methods for selection bias correction. In order to evaluate the constructed weights, we discussed several performance measures that can be considered in practice. However, unfortunately we are not able to identify the best performance measure for a given situation. One reason for this is that, given the nature of the weighting, many often-used performance measures are not suitable. What we may conclude is that, based on the results of the simulation and the examples, measures considering a y-model have the potential to perform well. Among all the discussed measures, MSB is especially a reliable measure of performance given that it is less sensitive to model misspecification of y. However, it may still be challenging to reveal the actual error left in the data set after weighting because of the uncertainty with respect to the parameter ϕ. In Little et al. (2020), it is suggested that $\phi = 0.5$ may be used and also checking $\phi = 0$ and $\phi = 1$ as a sensitivity analysis.

An interesting result from our simulation study is that the best-performing inclusion propensity model in terms of Bias and MSE of the estimated population

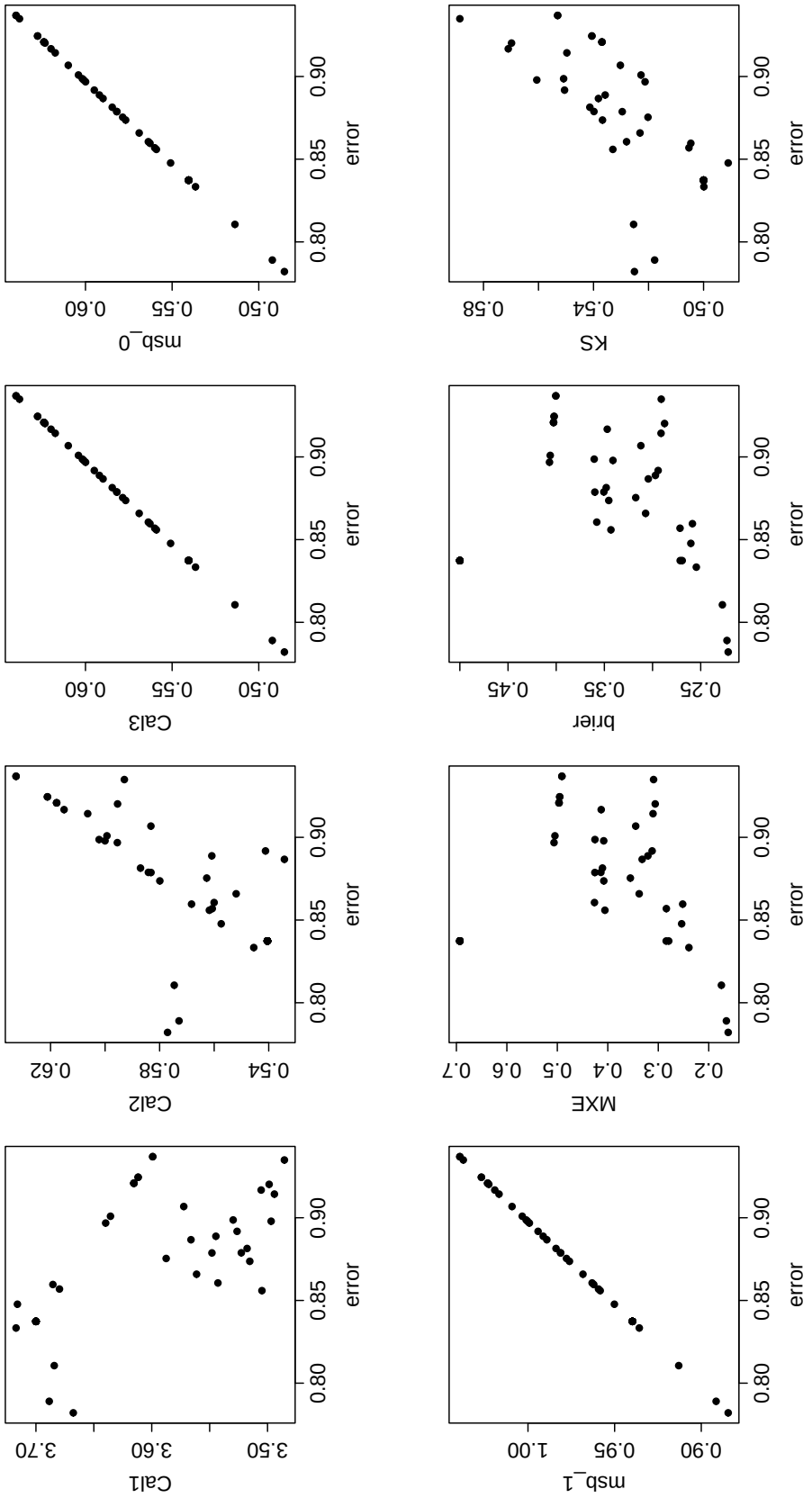

FIGURE 3.3: Relations between the absolute error and the performance measures in Iris data. Measures include different variants of calibration (Cal1, Cal2, Cal3), measure of selection bias (MSB) with different ϕ, Mean cross entropy (MXE), Brier score, and Kolmogorov-Smirnov distance (KS).

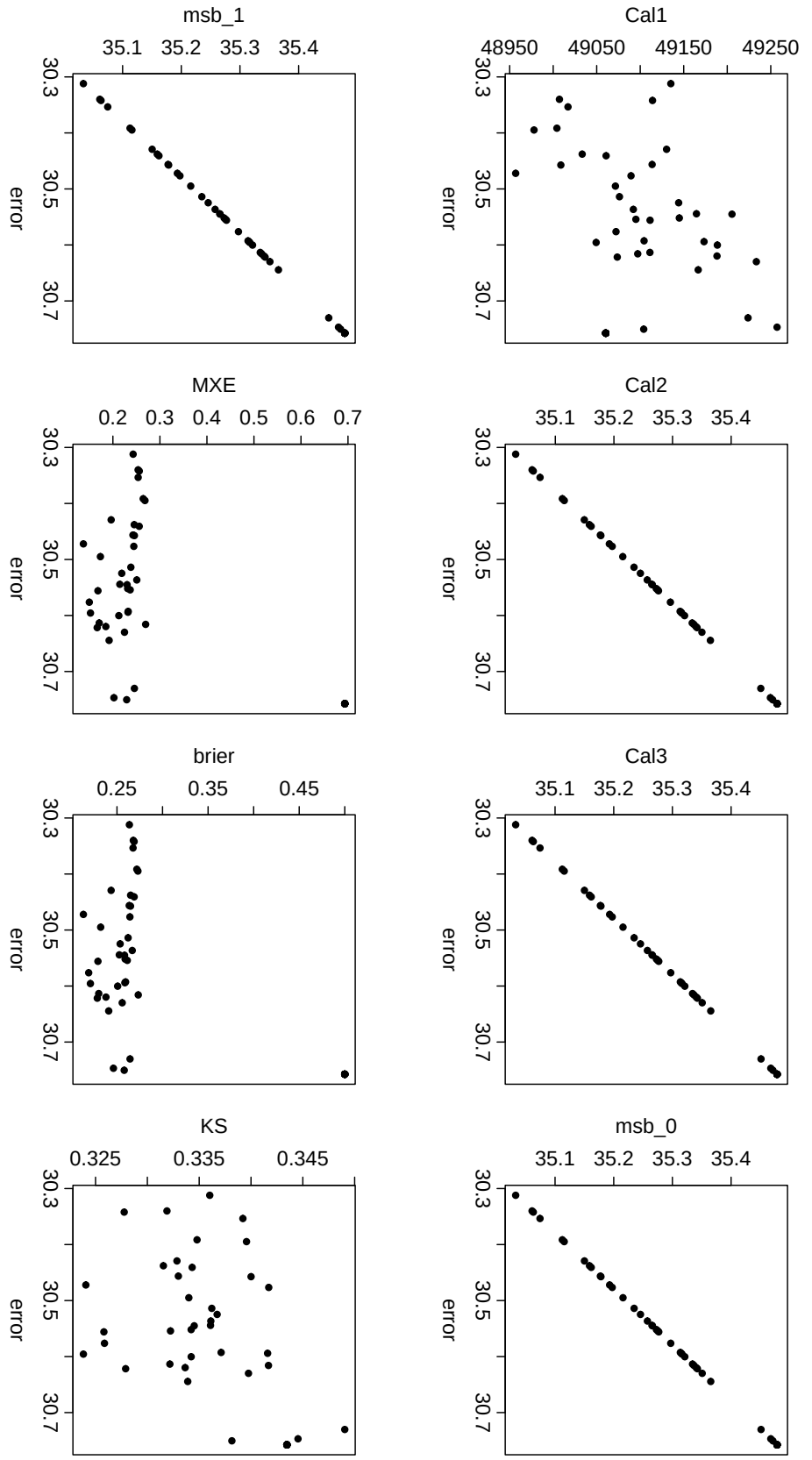

FIGURE 3.4: Relations between the absolute error and the performance measures in Election data. Measures include different variants of calibration (Cal1, Cal2, Cal3), measure of selection bias (MSB) with different ϕ, Mean cross entropy (MXE), Brier score, and Kolmogorov-Smirnov distance (KS).

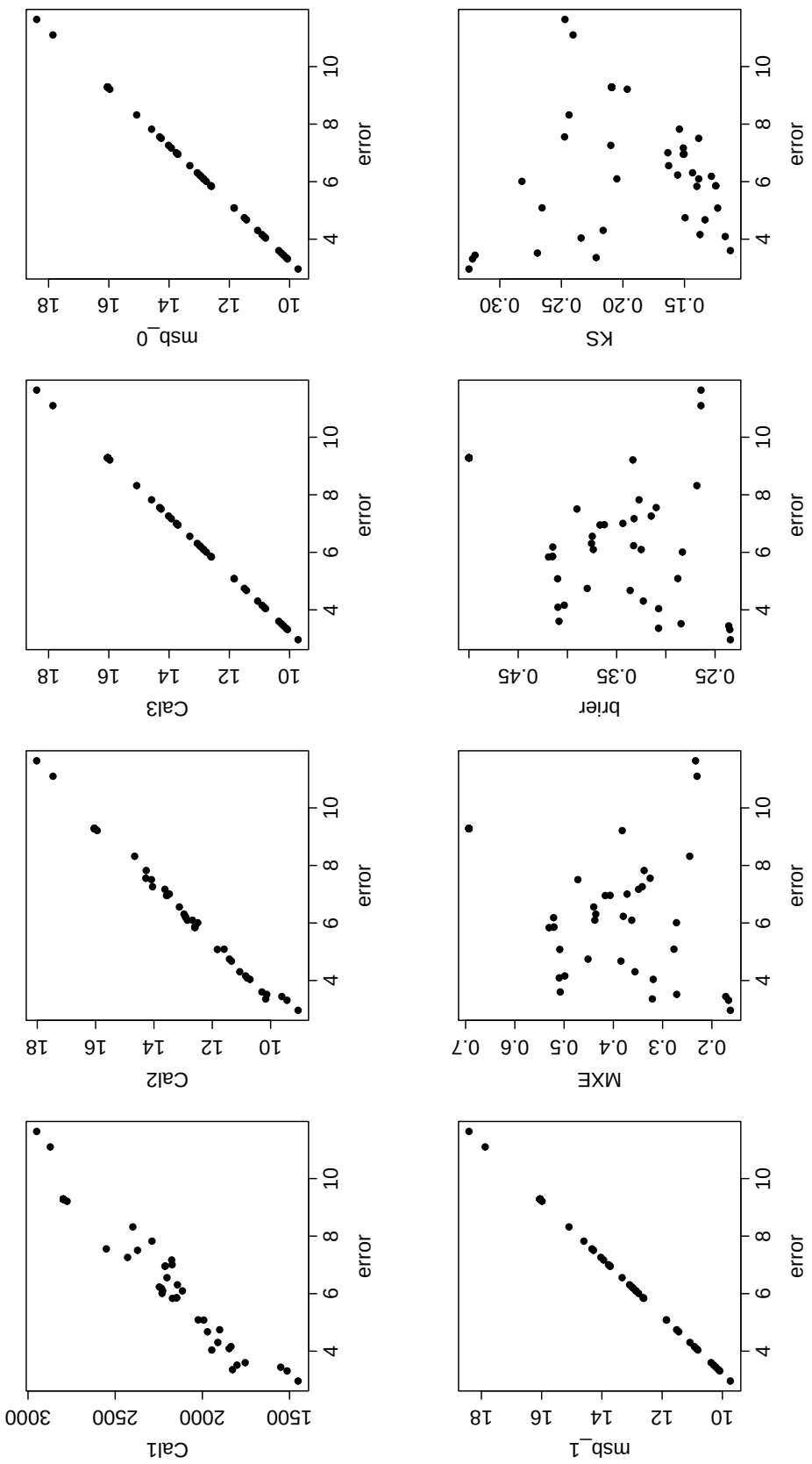

FIGURE 3.5: Relations between the absolute error and the performance measures in MU284 data. Measures include different variants of calibration (Cal1, Cal2, Cal3), measure of selection bias (MSB) with different ϕ, Mean cross entropy (MXE), Brier score, and Kolmogorov-Smirnov distance (KS).

parameter is not necessarily the correct model for these inclusion propensities.

The performance measures considered in this paper aim to assess the usefulness of weights constructed to correct for selection bias when estimating population means. Measures of bias other than those discussed in the paper have been proposed that may serve as a performance measure for other population parameters, see for example, Andridge, West, Little, Boonstra, and Alvarado-Leiton (2019) for proportion estimation or West et al. (2021) for regression coefficient estimation. Future research is needed to examine the usage of these measures for other population parameters than population means.

We illustrated the performance evaluation framework by Elliott and Valliant's method since it is flexible to many kinds of models/algorithms and gives stable estimates. The weights may also come from other approaches, for example, approaches by Y. Chen et al. (2020); Valliant (2020), and still fit in the framework we discussed here. Also, if more than one target variable is of interest, performance measures can be calculated with regard to different variables at the same time and one can choose a set of weights that fits well for most of the target variables.

3.7 Appendix: Simulation Results for Non-linear Propensity Model

The simulation results when the propensity model is $p = \exp(c + x_1 + 0.5x_2x_4)/(1 + \exp(c + x_1 + 0.5x_2x_4))$. See Section 3.4 for the rest of the simulation setup.

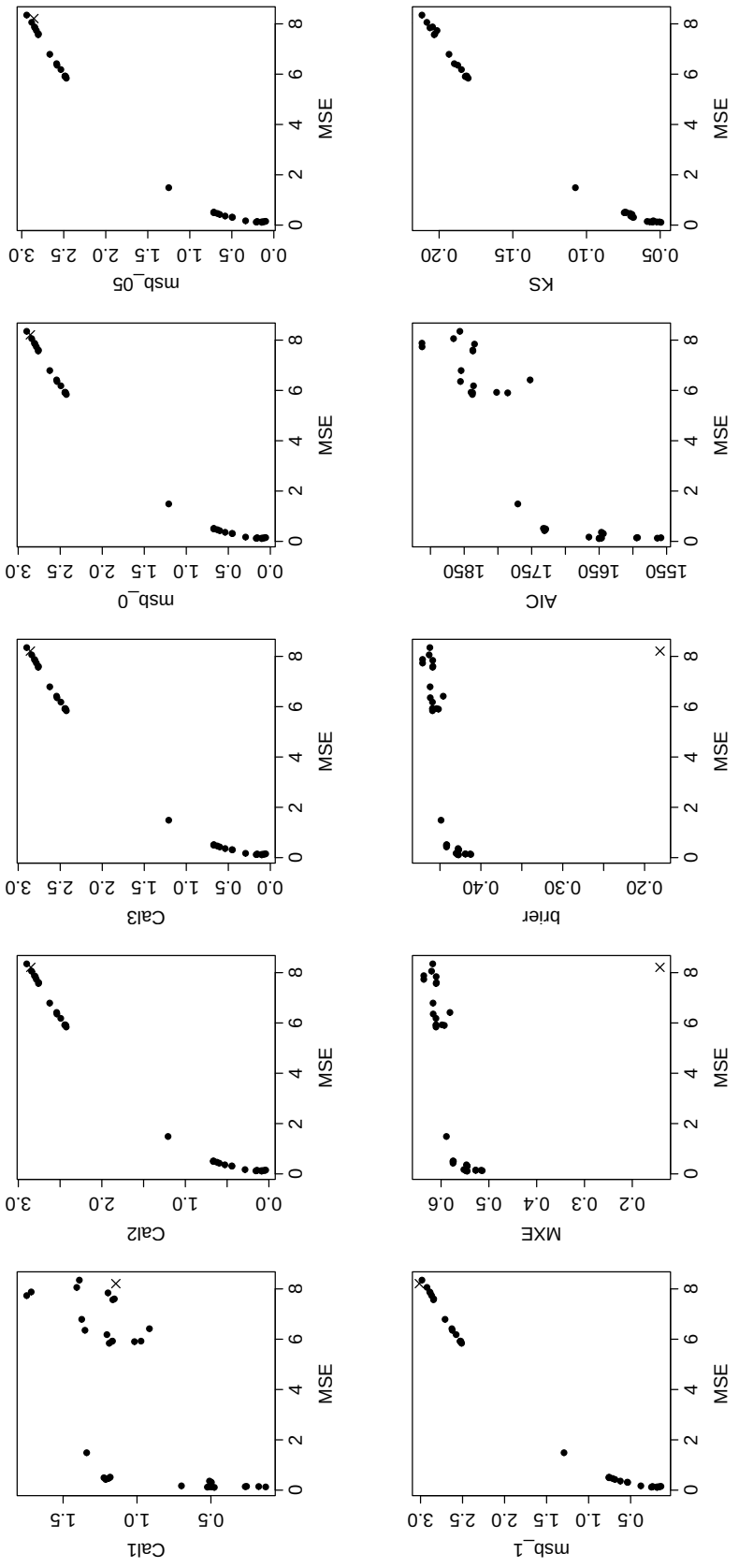

FIGURE 3.6: Relations between the MSE and the mean performance measures under different propensity model specifications. The correct y-model is used. The dots are the estimates from logistic regression and the cross is the estimate from `XGBoost`. Measures include different variants of calibration (Cal1, Cal2, Cal3), measure of selection bias (MSB) with different ϕ, Mean cross entropy (MXE), brier score, Akaike information criterion (AIC), and Kolmogorov-Smirnov distance (KS).

FIGURE 3.7: Relations between the MSE and the mean performance measures under different propensity model specifications. The correct y-model is used. The dots are the estimates from logistic regression and the cross is the estimate from `XGBoost`. Measures with a y-model include two variants of calibration (Cal2, Cal3), measure of selection bias (MSB) with different ϕ, and Kolmogorov-Smirnov distance (KS)

Chapter 4

Selection Bias Correction for Imbalanced Samples

Abstract

Selection bias correction is generally crucial to have reliable inference for a nonprobability sample. Often the correction is applied with the two-sample setup. That is, along with the nonprobability sample we are interested in, a probability sample sharing some common auxiliary variables is used for constructing correction weights for the nonprobability sample. The two-sample setup allows one to calculate weighted estimates for population parameters of interest based on the nonprobability sample. Since the nonprobability sample is usually easy to collect, we often end up with a large nonprobability sample and a small probability sample. The imbalance of the two samples may cause difficulties in modeling the propensity of units to be included in the nonprobability sample, which is an essential step for constructing correction weights for the nonprobability sample. This paper discusses some often-seen solutions for imbalanced samples in machine learning literature, i.e., undersampling, Synthetic Minority Oversampling Technique (SMOTE), and a mixture of both. A

selection bias correction framework is adjusted to incorporate the imbalance solutions. A simulation study and a real data application are shown. The results indicate that SMOTE has the potential to deal with imbalances in selection bias correction.

Keywords: Descriptive Inference, Imbalanced Sample, Nonprobability Sample, SMOTE, Undersampling

Liu, A.-C., Scholtus, S., Ang, L., Van Deun, K. & De Waal, T. Selection Bias Correction for Imbalanced Samples. (Submitted)

4.1 Introduction

Because of digitalization, many new data types emerge, for example, social media data, sensor data, or administrative data. These data are usually not based on a known sampling design and are termed nonprobability samples. If a nonprobability sample is treated as a simple random sample, selection bias may occur for population parameter estimation. To correct for selection bias, many correction methods adopt the two-sample setup. That is, along with the nonprobability sample, a probability sample sharing some common auxiliary variables from the same population is used, see for example, Y. Chen et al. (2020); Elliott and Valliant (2017); Liu et al. (2023). In the two-sample setup, by combining the nonprobability sample and the probability sample, the inclusion probabilities (or propensities) of the units in the nonprobability sample can be estimated, and the reciprocals of these inclusion propensities can be used as weights for the units in the nonprobability sample. A design-based estimator (e.g., Hájek estimator) can then be applied for population parameter estimation. Other correction methods can also be found in recent reviews of selection bias correction methods, see for example, J. Rao (2020); Wu (2022).

Given that a nonprobability sample is often time and cost-efficient to collect, researchers often have a massive nonprobability sample and a relatively small probability sample. This means that researchers frequently end up with an imbalanced combined sample, which may cause difficulties when estimating the inclusion propensities of the nonprobability sample. The problem of imbalanced samples has been well-studied in prediction; see, for example, discussions in Batista, Prati, and Monard (2004); Estabrooks, Jo, and Japkowicz (2004). One straightforward approach to deal with imbalanced samples is to downsize the majority group. In our case, this would mean drawing a subsample of the nonprobability sample so that it has a similar size as the probability sample. A large proportion of the nonprobability sample may

then be discarded from the propensity modeling process. Another popular way to deal with an imbalanced sample is the Synthetic Minority Oversampling Technique (SMOTE) proposed by Chawla, Bowyer, Hall, and Kegelmeyer (2002). SMOTE creates synthetic samples for the minority group in a nonparametric manner. That is, weighted combinations of a unit and its nearest neighbors are used to create a synthetic "new sample". It has been shown that SMOTE improves massively the prediction performance. Many variants of SMOTE have also been proposed for different types of data or different goals of modeling. For example, SMOTE-ENC is for a data set containing both continuous and categorical data (Mukherjee & Khushi, 2021). A review of the developments around SMOTE can be found in Fernández, Garcia, Herrera, and Chawla (2018).

However, neither under- nor oversampling can immediately be applied for selection bias correction since the probability sample does not always come from simple random sampling. Also, after under- or oversampling, the sample size is changed and the calculation of the correction weights should also be adjusted accordingly. In Section 4.2, we will start by introducing a selection bias correction framework. Since we aim for a scenario where the nonprobability sample is larger than the probability sample, the method proposed by Liu et al. (2023) is considered. Algorithms for under- and oversampling and a mixture of both are discussed in Section 4.2.2. A weighted version of SMOTE is also discussed for a probability sample resulting from random sampling schemes different from simple random sampling. A simulation study and an application to a real data set are presented respectively in Section 4.3 and 4.4 to evaluate the algorithms under different scenarios. Section 4.5 ends this paper with a short discussion and some conclusions.

4.2 Methods

4.2.1 LSW Method for Selection Bias Correction

We start by presenting the method for correcting for selection in a nonprobability sample proposed by Liu et al. (2023), which we will refer to as the LSW method (for details of this method, see Liu et al. (2023)).

We consider a nonprobability sample (NP) with size n_{NP} from a finite population (U) indexed by $i = 1, 2, ..., N$ with N the population size. In NP, y and $\mathbf{x}$ are available where y is a target variable and $\mathbf{x}$ is a set of auxiliary variables. For simplicity, here we consider only a single target variable y, but this can easily be generalized to more target variables. The target parameter of interest is the population mean of y, i.e. $\mu = N^{-1} \sum_{i \in U} y_i$.

Along with the nonprobability sample, a reference probability sample (P) with size n_P from the same population is available with the same set of auxiliary variables $\mathbf{x}$ and design weights $d_i = 1/\mathbb{P}(i \in P)$, but not necessarily with y. Some assumptions are considered:

1. For all the units in the population $0 < \mathbb{P}(i \in NP) < 1$ and $0 < \mathbb{P}(i \in P) < 1$;
2. $\mathbb{P}(i \in NP)$ is fully governed by $\mathbf{x}$;
3. The design weights of the probability sample are available (or can be calculated) for both NP and P;
4. The overlapping units between NP and P can be identified.

Assumptions 1 and 2 are common assumptions in the selection bias correction literature to allow for valid inference. Assumptions 3 and 4 are especially for the LSW method. If Assumption 4 does not hold, the pseudo-weighting method from

Elliott and Valliant (2017) may be applied, although the efficiency of the resulting estimator is then less than the LSW estimator (Liu et al., 2023).

In order to estimate propensities for units to be included in NP, we fit a propensity model $f(\mathbf{x})$. This model is fitted with the non-overlapping part of the two samples (B). Within B, we assign $z_i = 1$ if $i \in NP \setminus P$, and $z_i = 0$ if $i \in P \setminus NP$ as the dependent variable. The model $f(\mathbf{x})$ can be any probability estimation method, for example, logistic regression or a classification tree. After fitting $f(\mathbf{x})$, the odds

$$O_i = \frac{\mathbb{P}(z_i = 1 \mid i \in B, \mathbf{x}_i)}{\mathbb{P}(z_i = 0 \mid i \in B, \mathbf{x}_i)} \tag{4.1}$$

are calculated given $f(\mathbf{x})$. The final weights, i.e. the reciprocals of the inclusion propensities, are calculated as

$$w_i = \frac{1}{\mathbb{P}(i \in NP \mid \mathbf{x}_i)} = 1 + \frac{d_i - 1}{O_i}. \tag{4.2}$$

The population mean is estimated by

$$\hat{y} = \sum_{i \in NP} w_i y_i (\sum_{i \in NP} w_i)^{-1}.$$

Ideally, it satisfies

$$E(\hat{y}) = \mu.$$

Here we focus on the scenario where the inclusion in the nonprobability sample is independent of the inclusion in the probability sample, which might not always be the case as discussed in Liu et al. (2023).

4.2.2 Balancing the Two Samples

When fitting a propensity model $f(\mathbf{x})$ on a nonprobability sample and a probability sample, the model may be sensitive to the imbalances between the two samples since we often have a large nonprobability sample and a small probability sample. To deal with imbalances, either under- or oversampling can be used. After fitting $f(\mathbf{x})$ with under- or oversampling samples, the weights are then constructed for the whole NP for inference.

Unweighted under- and oversampling

In undersampling, a subset is selected from the majority group by simple random sampling without replacement. Therefore, one will end up with a smaller sample size than the original data. Alternatively, one can oversample the minority group by either simply adding copies of the original data of the minority group, or creating data points that are similar to the units in the minority group.

After under- or oversampling, the calculation in (4.2) needs to be adjusted. Here we introduce further notation. We denote the set of units $i \in B$ where $z_i = 0$ by B_0 and the set of units $i \in B$ with $z_i = 1$ by B_1. The size of B_0 is $n_{B0} = \sum_{i \in B}(1 - z_i)$, the size of B_1 is $n_{B1} = \sum_{i \in B} z_i$, and $n_{B0} + n_{B1} = n_B$. We limit our discussion to the scenario where $0 < n_{B0} < n_{B1}$. Suppose we aim to have a sample where the (original) majority group has a proportion of F. One choice of F, for example, may be 0.5 to have a fully balanced sample. We can achieve $F = 0.5$ by undersampling the B_1 set to obtain a subset B_1^- of size $n_{B1}^- = n_{B0}$, or oversampling the B_0 set to acquire a larger set B_0^+ of size $n_{B0}^+ = n_{B1}$. Assuming all the units have the same probability of being undersampled, respectively oversampled, the calculation of weights in (4.2) is completed by $O_i = O_i^F/p$, where O_i^F is the odds calculated

from the balanced sample and a fixed term

$$p = \frac{F}{(1-F)}\frac{n_{B0}}{n_{B1}}. \tag{4.3}$$

Finally,

$$w_i = 1 + \frac{(d_i - 1)p}{O_i^F}. \tag{4.4}$$

The detailed derivation of (4.4) can be found in the Appendix 4.6.

Algorithm 3 Undersampling

Require: d_i, B_0, B_1, NP, F

1: Calculate p with (4.3)
2: Draw a simple random sample without replacement from B_1 with sampling fraction p, and denote the drawn subsample by B_1^-
3: Fit propensity model $f(\mathbf{x})$ on $B_0 \cup B_1^-$, where $z_i = 1$ if $i \in B_1^-$ and $z_i = 0$ otherwise.
4: Estimate the odds O_i^F of the units in NP given $f(\mathbf{x})$.
5: Calculate the weights with $w_i = 1 + (d_i - 1)p/O_i^F$.

For undersampling, p is aligned with the sampling fraction of the units in B_1. Algorithm 3 presents the process of undersampling. In oversampling, studies have shown that simply adding copies of units in the minority group contributes little to model performance (Blagus & Lusa, 2013). Therefore, here we introduce SMOTE which creates a set of synthetic data points from the minority group. The general idea is that SMOTE creates synthetic data points by combining data points with their neighbors. The corresponding process is shown in Algorithm 4. The ratio of the majority group (F) is used to calculate the number of synthetic units (R) that need to be created. For every unit in B_0, K nearest neighbors from other units in B_0 are found using a distance function based on $\mathbf{x}_i$. If the design weights of the probability sample are not included as one of the auxiliary variables in $\mathbf{x}_i$, the distance should be calculated not only by using the $\mathbf{x}_i$ but also by using the design weights. Considering weights in the distance calculation can leverage the

auxiliary information to improve the selection of the neighbors. A similar concept can be found in the nonresponse literature where including the response propensity in the imputation model can improve the performance of the imputation (Andridge & Little, 2009).

From the literature on nearest neighbor imputation, we also learn that the efficiency of the imputation can be increased if $K > 1$ (Fuller, 2011). If K is small, the imputation bias is smaller while the imputation variance is larger. If K is large, the smoothness and symmetry of the density around the unit and its neighbors are crucial to guarantee reasonable synthetic data points. In SMOTE, after K neighbors ($r_{i1}, r_{i2}, ..., r_{iK}$ for unit i) are found, a random draw of the neighbors is performed and the drawn neighbor is assigned as r_q. The drawing process can be conducted by giving all the K neighbors an equal chance to be drawn or letting $\mathbb{P}(r_q = r_{ik}) \propto \mathbb{D}(x_i, r_{ik})^{-1}$, where $\mathbb{D}$ denotes the distance function. That is, the closer the neighbor, the higher the chance to be drawn. Another random value α is drawn from $U(a, b)$, where $U(\cdot)$ is the uniform distribution on the interval (a, b). We use $\alpha \sim U(0, 1)$ as in the original paper by Chawla et al. (2002). The synthetic data point is created by $s_q = x_i + \alpha(r_q - x_i)$. The process is repeated until $[R]$ synthetic units are created, where $[\cdot]$ denotes rounding into the closest integer. Here we presume the records in the sample are placed in a random order, or else the order should be shuffled before creating the synthetic sample, since the units with a smaller i have a higher chance of being replicated.

When the data is extremely imbalanced, $n_{B0} \ll n_{B1}$, the choice between under- or oversampling may be difficult. Undersampling implies that a large proportion of B_1 is discarded, and the model is fitted with a small sample size. On the other hand, in oversampling, a large amount of synthetic data is created given limited true data. The distribution of the synthetic data may differ a lot from the true distribution of the minority group. As a trade-off, instead of merely under- or oversampling, another

Algorithm 4 Oversampling by SMOTE

Require: d_i, B_0, B_1, NP, F, K

1: Calculate the size of synthetic data $R = n_{B1}/F - n_B$ and round it to $[R]$
2: Find K nearest neighbors (r_{ik}) from B_0 for every $i \in B_0$
3: Assign $i \leftarrow 1$
4: **for** $q = 1$ to $[R]$ **do**
5: Draw a neighbor of i from $\{r_{i1}, r_{i2}, ..., r_{ik}\}$ and assign as r_q
6: Draw $\alpha \sim U(0, 1)$
7: Create a synthetic data point by $s_q = x_i + \alpha(r_q - x_i)$
8: Assign $i \leftarrow i + 1$. If $i = n_{B0} + 1$, then $i \leftarrow 1$
9: **end for**
10: Row combine all the s_q with B_0 to form B_0^+
11: Fit propensity model $f(\mathbf{x})$ on $B_0^+ \cup B_1$, where $z_i = 1$ if $i \in B_1$ and $z_i = 0$ otherwise.
12: Estimate the odds O_i^F of the units in NP given $f(\mathbf{x})$.
13: Calculate p with (4.3) and the weights with $w_i = 1 + (d_i - 1)p/O_i^F$.

option is to perform both under- and oversampling at the same time. The process is shown in Algorithm 5 where the aim of the algorithm is to create a balanced sample with a sample size equal to n_B.

Oversampling by weighted SMOTE

When the probability sample was not obtained by simple random sampling, besides the original SMOTE, a weighted type of SMOTE may be used to create a synthetic data set that ideally has a similar distribution as a simple random sample from the population. It is achieved by deciding the number of synthetic units created from a unit by the design weight of the unit. This weighted type of SMOTE is shown in Algorithm 6. First, the sum of the design weights in B_0 is regularized to be equal to $n_{B1}(1 - F)/F$ by $d_i^* = d_i C_{B0}$, where

$$C_{B0} = n_{B1} \frac{1 - F}{F} \left(\sum_{i \in B_0} d_i \right)^{-1}. \tag{4.5}$$

Algorithm 5 A mixture of SMOTE and undersampling

Require: d_i, B_0, B_1, NP, F, K

1: Draw a simple random sample without replacement from B_1 with size $F \times n_B$, and denote this sample by B_1^-
2: Calculate the size of synthetic data $R = n_B(1-F) - n_{B0}$ and round it to $[R]$
3: Find K nearest neighbors (r_{ik}) from B_0 for every $i \in B_0$
4: Assign $i \leftarrow 1$
5: **for** $q = 1$ to $[R]$ **do**
6: Draw a neighbor of i from $\{r_{i1}, r_{i2}, ..., r_{ik}\}$ and assign as r_q
7: Draw $\alpha \sim U(0,1)$
8: Create a synthetic data point by $s_q = x_i + \alpha(r_q - x_i)$
9: Assign $i \leftarrow i+1$. If $i = n_{B0}+1$, then $i \leftarrow 1$.
10: **end for**
11: Row combine all the s_q with B_0 to form B_0^+
12: Fit propensity model $f(\mathbf{x})$ on $B_0^+ \cup B_1^-$, where $z_i = 1$ if $i \in B_1^-$ and $z_i = 0$ otherwise.
13: Estimate the odds O_i^F of the units in NP given $f(\mathbf{x})$.
14: Calculate p with (4.3) and the weights with $w_i = 1 + (d_i - 1)p/O_i^F$.

In the second step, a fully synthetic data set can be created by assigning $\tilde{d}_i = d_i^*$, while units with small weights may be removed due to rounding in the next step. Another option, and also the one we apply in the simulation, is to assign $\tilde{d}_i = d_i^* - 1$, and the original probability sample is combined with the synthetic data. In the third step, stochastic controlled rounding is applied to prevent systematic rounding bias (Fellegi, 1975). That is, $\tilde{d}_i$ is rounded up to the smallest integer that is not less than $\tilde{d}_i$ with probability $\tilde{d}_i - \lfloor \tilde{d}_i \rfloor$ and rounded down otherwise, where $\lfloor \tilde{d}_i \rfloor$ is the rounded-down value of $\tilde{d}_i$. After the process, we have $E([\tilde{d}_i]) = d_i^*$ when $\tilde{d}_i = d_i^*$. The process is repeated until for each i, $[\tilde{d}_i]$ corresponding synthetic data points are created. The calculation of the final weights is now

$$w_i = 1 + \frac{d_i - 1}{O_i^F d_i C_{B0}}. \tag{4.6}$$

See Appendix 4.6 for a detailed derivation of (4.6).

Algorithm 6 Oversampling by weighted SMOTE

Require: d_i, B_0, B_1, NP, F, K

1: Calculate C_{B0} by (4.5), and regularize the weights to obtain $\sum_{i \in B_0} d_i^* = n_{B1}(1-F)/F$
2: Assign $\tilde{d}_i \leftarrow d_i^* - 1$
3: Randomly round $\tilde{d}_i$ to its ceiling with probability $\tilde{d}_i - \lfloor \tilde{d}_i \rfloor$ and to its floor otherwise to obtain $[\tilde{d}_i]$
4: Find K nearest neighbors (r_{ik}) from B_0 for every $i \in B_0$
5: **for** every $i \in B_0$ **do**
6: **for** $q = 1$ to $[\tilde{d}_i]$ **do**
7: Draw a neighbor of i from $\{r_{i1}, r_{i2}, ..., r_{ik}\}$ and assign as r_q
8: Draw $\alpha \sim U(0, 1)$
9: Create a synthetic data point by $s_{iq} = x_i + \alpha(r_q - x_i)$
10: **end for**
11: **end for**
12: Combine all the s_{iq} with B_0 to form B_0^+
13: Fit propensity model $f(\mathbf{x})$ on $B_0^+ \cup B_1$, where $z_i = 1$ if $i \in B_1$ and $z_i = 0$ otherwise.
14: Estimate the odds O_i of units in NP given $f(\mathbf{x})$.
15: Calculate the weights with $w_i = 1 + (d_i - 1)/(O_i^F d_i C_{B0})$.

4.3 Simulation

4.3.1 Setup

A finite population U of size $N = 30,000$ is created with $y, x_1, x_2, x_3 \sim N(\mu = 1, \sigma = 1)$, i.e., the normal distribution with mean $\mu = 1$ and standard deviation $\sigma = 1$. We are interested in estimating the population mean of the target variable y. The true population mean $\mu = N^{-1} \sum_U y_i \approx 1$. To reflect different amounts of auxiliary information researchers have, the correlation between y and every auxiliary variable x is 0.7 as more informative auxiliary information, and the correlation between y and each x is 0.3 as less informative auxiliary information. The correlation between each x is 0.3. Selective nonprobability samples are drawn with inclusion probabilities $\pi_i = \exp(\beta_0 - 2y_i)/[1 + \exp(\beta_0 - 2y_i)]$, where β_0 is a constant allowing control of the sample size of the nonprobability sample. The inclusion fractions of

the nonprobability sample (f_{NP}) are 10, 30, 50, and 70 percent, and the inclusion fractions of the probability sample (f_P) are 1, 5, 10, 30, and 50 percent to reflect different levels of imbalances. We only look at the combinations of inclusion fractions where $f_P < f_{NP}$, which results in 14 combinations. Two sampling designs of the probability sample are considered, namely Simple Random Sampling (SRS) and Probability Proportional to Size (PPS). When the probability sample is from a PPS design, the inclusion probability is proportional to x_1. Several approaches are conducted for comparison:

- Naive: Treat the nonprobability sample as a simple random sample.
- LSW: Apply the LSW method without any adjustment for imbalance.
- Undersampling (US): Draw a subsample of the nonprobability sample by Algorithm 3.
- Oversampling with SMOTE (OS): Create synthetic samples by Algorithm 4.
- Mix: A combination of undersampling and SMOTE by Algorithm 5.
- Oversampling with Weighted SMOTE (OS2): Create synthetic samples by Algorithm 6. This is only applied to the PPS probability samples.

The under- or oversampling is done in such a way that the B sample is fully balanced, that is, $F = 0.5$ so that around half of the units in the final sample are units with $z_i = 1$ and the other half with $z_i = 0$. The number of nearest neighbors is $K = 5$ for all the algorithms applying SMOTE and the R package *smotefamily* is used. The propensity model is a logistic regression fitted with all the auxiliary variables. The performance of each approach is evaluated with estimated Relative Bias $\text{RB}(\hat{y}) = M^{-1}\sum_{m=1}^{M}(\hat{y}_m - \mu)/\mu$ and mean squared error $\text{MSE}(\hat{y}) = M^{-1}\sum_{m=1}^{M}(\hat{y}_m - \mu)^2$ where m is the index of simulation runs with $m = 1, 2, ..., M$ and $M = 1000$.

4.3.2 Results

The results from SRS samples are in Table 4.1, and the results of PPS samples are in Table 4.2. In general, the original SMOTE performs very well across different scenarios. Undersampling does contribute to decreasing bias in some cases, however, it does not compensate for the large variation. Although the mixture of under- and oversampling does not perform the best, it is still better than doing nothing to balance the samples (as in LSW) most of the time. When the sample is from a PPS sampling, performing the original SMOTE or the weighted SMOTE may both be good options.

4.4 Application

In the simulation study, the data are created with a symmetric distribution and the number of variables is small, which is an ideal scenario for SMOTE (Blagus & Lusa, 2013). In this section, we apply the algorithms on a (semi) real-world data set to evaluate the performance of each algorithm in a practical scenario.

The data set we use represents a synthetic business population consisting of 900,000 records; see Ang, Clark, Loong, and Holmberg (2024) for the location of the data set, along with information on how it was created. The data set aims to resemble the real-world population of employing businesses in Australia in terms of the distribution of businesses across industry groupings, geographic locations and business size. Data for the population was generated by using a combination of published business outputs from the Australian Bureau of Statistics (ABS), as well as business survey information and employee tax data sourced from the Business Longitudinal Analysis Data Environment (BLADE) in the ABS DataLab. A goal of the data generation process was to ensure the characteristics of the variables

TABLE 4.1: MSE and Relative Bias under SRS samples. The upper block corresponds to the condition with informative auxiliary variables, and the lower block with less informative auxiliary variables. In each row, the smallest (absolute) value is printed in bold.

%		100*MSE					RB (%)				
f_P	f_{NP}	Naive	LSW	OS	US	Mix	Naive	LSW	OS	US	Mix
1	10	163.80	3.78	**3.44**	5.02	3.66	-127.88	-15.49	**-10.01**	-14.49	-10.49
5	10	163.86	4.07	**3.85**	4.29	4.18	-127.91	-17.02	**-15.89**	-17.10	-17.09
1	30	76.59	1.65	**1.29**	2.55	1.31	-87.44	-11.79	**-6.00**	-9.52	-6.11
5	30	76.49	1.70	**1.43**	1.98	1.51	-87.39	-12.31	**-10.62**	-12.62	-11.02
10	30	76.53	1.76	**1.62**	1.94	1.69	-87.41	-12.69	**-11.95**	-13.04	-12.15
1	50	36.93	0.78	**0.64**	1.64	0.64	-60.72	-8.06	**-2.41**	-5.92	-2.44
5	50	36.94	0.76	**0.55**	0.92	0.58	-60.73	-8.38	**-6.45**	-8.05	-6.60
10	50	36.92	0.78	**0.62**	0.87	0.65	-60.72	-8.57	**-7.34**	-8.58	-7.51
30	50	36.94	0.79	**0.73**	0.80	0.79	-60.73	-8.63	**-8.22**	-8.62	-8.62
1	70	14.22	**0.31**	0.64	1.63	0.65	-37.68	-4.67	1.87	**0.15**	1.87
5	70	14.21	0.26	**0.15**	0.35	0.15	-37.67	-4.84	**-2.71**	-3.99	-2.75
10	70	14.21	0.25	**0.15**	0.28	0.16	-37.67	-4.86	**-3.33**	-4.28	-3.41
30	70	14.20	0.25	**0.19**	0.25	0.20	-37.66	-4.87	**-4.14**	-4.66	-4.23
50	70	14.21	0.25	**0.22**	0.24	0.25	-37.66	-4.88	**-4.54**	-4.73	-4.85
1	10	163.05	125.48	**124.90**	126.11	125.07	-127.36	-111.70	**-111.45**	-111.97	-111.52
5	10	163.09	126.02	**126.02**	126.27	126.10	-127.37	-111.96	**-111.95**	-112.07	-111.99
1	30	75.76	59.29	**58.69**	59.07	58.69	-86.81	-76.78	**-76.39**	-76.62	-76.39
5	30	75.83	59.49	**59.40**	59.63	59.44	-86.85	-76.92	**-76.87**	-77.01	-76.89
10	30	75.80	59.55	**59.53**	59.68	59.55	-86.84	-76.96	**-76.95**	-77.05	-76.97
1	50	36.34	28.27	**27.83**	28.15	27.83	-60.12	-53.01	**-52.60**	-52.89	-52.60
5	50	36.31	28.27	**28.11**	28.25	28.12	-60.10	-53.03	**-52.87**	-53.01	-52.88
10	50	36.32	28.29	**28.21**	28.31	28.22	-60.11	-53.05	**-52.97**	-53.07	-52.98
30	50	36.31	28.28	**28.26**	28.29	28.29	-60.10	-53.04	**-53.02**	-53.05	-53.05
1	70	13.91	10.67	**10.37**	10.42	10.37	-37.19	-32.56	**-32.09**	-32.15	-32.09
5	70	13.91	10.61	**10.48**	10.58	10.48	-37.20	-32.48	**-32.28**	-32.43	-32.29
10	70	13.91	10.61	**10.53**	10.58	10.53	-37.20	-32.49	**-32.36**	-32.44	-32.37
30	70	13.90	10.60	**10.56**	10.59	10.57	-37.19	-32.47	**-32.42**	-32.45	-32.42
50	70	13.91	10.60	**10.58**	10.59	10.60	-37.20	-32.47	**-32.44**	-32.45	-32.47

TABLE 4.2: MSE and Relative Bias under PPS samples. The upper block is the result of informative auxiliary variables, and the lower block is the result of less informative auxiliary variables. The smallest (absolute) value of each row is bold.

%		100*MSE						RB (%)					
f_P	f_{NP}	Naive	LSW	OS	OS2	US	Mix	Naive	LSW	OS	OS2	US	Mix
1	10	163.86	3.85	**3.54**	3.61	5.20	3.68	-127.90	-15.70	-8.60	**-8.43**	-13.71	-9.38
5	10	163.80	4.16	**3.92**	3.93	4.37	4.19	-127.88	-15.55	**-13.93**	-14.54	-15.04	-15.38
1	30	76.52	1.59	**1.30**	1.41	2.63	1.33	-87.41	-11.48	-4.80	**-4.65**	-9.23	-4.96
5	30	76.56	1.60	**1.22**	1.28	1.68	1.28	-87.43	-11.89	**-9.61**	-9.89	-11.25	-9.84
10	30	76.58	1.60	**1.37**	1.44	1.70	1.43	-87.44	-11.90	**-10.69**	-11.01	-11.84	-10.89
1	50	36.94	0.78	**0.64**	0.71	1.61	0.66	-60.73	-8.14	-2.05	**-1.68**	-4.92	-2.09
5	50	36.95	0.75	**0.49**	0.51	0.85	0.51	-60.74	-8.34	**-5.94**	-6.04	-7.60	-6.09
10	50	36.92	0.74	**0.54**	0.57	0.74	0.58	-60.71	-8.29	**-6.74**	-6.97	-7.80	-7.00
30	50	36.95	0.75	**0.67**	0.72	0.76	0.75	-60.74	-8.42	**-7.85**	-8.16	-8.29	-8.37
1	70	14.21	**0.31**	0.55	0.67	1.37	0.55	-37.66	-4.81	1.96	2.62	**-0.05**	1.89
5	70	14.21	0.25	**0.12**	0.13	0.31	0.13	-37.67	-4.76	-2.33	**-2.27**	-3.60	-2.41
10	70	14.22	0.24	**0.14**	0.15	0.25	0.15	-37.67	-4.78	**-3.08**	-3.12	-4.02	-3.19
30	70	14.21	0.24	**0.18**	0.18	0.22	0.20	-37.67	-4.76	**-3.95**	-4.01	-4.32	-4.19
50	70	14.22	0.28	**0.25**	0.27	0.28	0.28	-37.68	-5.18	**-4.80**	-5.03	-5.05	-5.12
1	10	163.22	125.60	**124.55**	124.57	125.74	124.65	-127.42	-111.76	**-111.28**	-111.29	-111.80	-111.33
5	10	163.05	125.86	**125.68**	125.83	126.04	125.96	-127.35	-111.88	**-111.80**	-111.87	-111.96	-111.93
1	30	75.82	59.61	58.67	**58.57**	59.50	58.68	-86.84	-76.99	-76.38	**-76.31**	-76.90	-76.39
5	30	75.79	59.48	59.19	**59.18**	59.54	59.23	-86.83	-76.92	-76.73	**-76.72**	-76.96	-76.76
10	30	75.79	59.52	59.38	**59.37**	59.58	59.42	-86.83	-76.94	-76.85	**-76.85**	-76.98	-76.88
1	50	36.31	28.43	27.73	**27.64**	28.13	27.74	-60.10	-53.17	-52.51	**-52.41**	-52.86	-52.51
5	50	36.32	28.36	28.06	**28.03**	28.28	28.07	-60.11	-53.11	-52.83	**-52.80**	-53.03	-52.84
10	50	36.31	28.35	28.15	**28.12**	28.28	28.17	-60.10	-53.10	-52.91	**-52.89**	-53.03	-52.93
30	50	36.34	28.33	**28.27**	28.27	28.31	28.33	-60.12	-53.09	**-53.03**	-53.03	-53.06	-53.08
1	70	13.91	10.69	10.27	**10.20**	10.40	10.27	-37.19	-32.60	-31.94	**-31.84**	-32.10	-31.94
5	70	13.91	10.67	10.47	**10.45**	10.59	10.48	-37.19	-32.58	-32.27	**-32.23**	-32.44	-32.28
10	70	13.91	10.66	10.51	**10.49**	10.58	10.51	-37.20	-32.57	-32.33	**-32.30**	-32.44	-32.34
30	70	13.91	10.63	10.55	**10.55**	10.58	10.56	-37.19	-32.51	-32.39	**-32.39**	-32.44	-32.40
50	70	13.90	10.63	**10.61**	10.63	10.63	10.62	-37.19	-32.51	**-32.48**	-32.52	-32.51	-32.51

created would mimic their real-world counterparts. Variables available in the data set include:

- *earnings* - Total weekly wages and salaries reported by the business
- *rep.emp* - Total number of employees reported by the business
- *ovt* - Total overtime paid to employees by the business
- *frame.emp* - An out-of-date value for the total number of employees in the business
- *indgrp* - Industry class for the business
- *size2* - Business size grouping based on the *frame.emp* variable
- *state* - Broad geographic identifier for the business

The target variables are *ovt* and *earnings*, which descriptive statistics are shown in Table 4.3. Both target variables are extremely skewed. The mean of *ovt* is 519 and 76% of the data is zero. For *earnings*, the mean is $20,090$ and 9.3% of the data is zero.

TABLE 4.3: Descriptive statistics of target variables

	ovt	*earnings*
Minimum	0	0
1st quartile	0	3819
Median	0	9459
Mean	519	20090
3rd quartile	0	16619
Maximum	1259194	8262828

The nonprobability sample is drawn with inclusion probabilities proportional to *pi.mar*, where *pi.mar* was generated using a two-stage process. At the first stage, an initial probability $pi.mar_1$ was produced, where

$$pi.mar_1 = \frac{\exp(\phi_0 + \phi_1 x)}{1 + \exp(\phi_0 + \phi_1 x)}$$

with $(\phi_0, \phi_1) = (0.09, 0.009)$, and x is the *frame.emp* variable. At the second stage, the $pi.mar_1$ values were adjusted downwards by a pre-specified factor in some *indgrp* classes, to simulate a reduced likelihood of being present on the big data set for those industries. The resulting probabilities were the final *pi.mar* values. The modelled inclusion probabilities result in units with smaller frame employment having a lower chance of being included in the nonprobability sample.

The same approaches are compared as in the simulation study. Since the data set is a combination of continuous data and categorical data, the package *themis* is used for the classic SMOTE where the categorical variable is created by the majority category of the neighbors. For weighted SMOTE for categorical data (Algorithm 6), the values of the drawn nearest neighbors are used.

4.4.1 SRS sample

Again we look at the scenario where the probability sample is a simple random sample. The inclusion fractions of the probability sample are 0.1% and 0.5%, and the inclusion fractions of the nonprobability sample are 1% and 3%. Both samples are drawn 300 times, The propensity model is fitted with auxiliary variables *size2* and *indgrp* with logistic regression and the results are shown in Table 4.4. The presented MSE of *ovt* is 10^{-2} of the original MSE, and the presented MSE of *earnings* is 10^{-5} times the original MSE. Both oversampling with SMOTE and the mixture of under- and oversampling improve the estimates regarding MSE and relative bias. Undersampling, however, is worse than no adjustment in most cases. The weighted SMOTE introduces a large MSE and bias, which may result from the skewness of the target variables and how the categorical variables were created.

TABLE 4.4: MSE and Relative Bias for SRS samples. The smallest (absolute) value of each row is bold.

	%		MSE				
	f_P	f_{NP}	Naive	LSW	OS	US	Mix
ovt	0.1	1	231.328	51.221	28.869	3679.292	**27.036**
	0.5	1	247.183	38.716	**29.056**	40.843	32.122
	0.1	3	211.640	36.530	**17.763**	1499.057	17.893
	0.5	3	215.571	27.035	**13.617**	31.120	15.737
earnings	0.1	1	308.846	45.490	21.734	538.798	**20.599**
	0.5	1	307.609	33.824	**21.470**	35.251	26.204
	0.1	3	296.198	46.346	21.396	657.055	**20.916**
	0.5	3	291.531	30.083	**12.718**	35.645	15.240
			RB (%)				
ovt	0.1	1	26.434	7.704	-2.642	19.832	**-1.677**
	0.5	1	27.039	7.876	**5.784**	7.866	6.365
	0.1	3	27.024	8.081	-3.174	16.814	**-2.612**
	0.5	3	27.152	8.067	**4.539**	8.362	5.116
earnings	0.1	1	27.218	7.838	-2.964	14.207	**-1.908**
	0.5	1	27.115	8.055	**5.908**	7.936	6.583
	0.1	3	26.947	8.191	-3.392	14.083	**-2.887**
	0.5	3	26.701	7.836	**4.339**	8.047	4.913

4.4.2 Stratified SRS sample

A stratified SRS sample is considered to reflect a practical scenario. A realistic business survey sample design was conducted by first assigning the population data set into strata based on State (*state*), Industry class (*indgrp*), and Size (*size2*). The size categories were: 0-4 employees, 5-19 employees, 20-299 employees, and 300+ employees. The Bethel-Chromy algorithm (Bethel (1989), Chromy (1987)) was used to produce an optimal allocation of sample to strata, to meet the following accuracy constraints for the *earnings* variable:

- Relative Standard Error (RSE) of 1.5% at the National level
- RSE of 5% for each Industry class
- RSE of 5% for each State

A minimum sample size of 6 was applied for each sampled stratum, while the large units (that is, those in the 300+ size strata) were fully sampled with a sampling fraction of 1. In total, the sample size of the stratified SRS sample is 7715, with 3259 units having $\mathbb{P}(i \in P) = 1$.

Being fully sampled, the large units in the population have $\mathbb{P}(i \in P) = 1$, which is often seen in the business survey scenario but it violates the first assumption as discussed in Section 4.2. Therefore, we separate the population into two sets. Set CE is units with $\mathbb{P}(i \in P) = 1$, and set $U \setminus CE$ is units with $\mathbb{P}(i \in P) < 1$. In the set $U \setminus CE$, all the estimation steps are the same as before and the weight is denoted as $W_{U \setminus CE}$. The propensity model is fitted with *size2*, *indgrp*, and *state*. In CE, since only the units in the nonprobability sample have a value of y, another set of weights is constructed. A propensity model is fitted in CE with $z_i = 1$ if $i \in NP$ and $z_i = 0$ otherwise. The weights are then

$$W_{CE} = \frac{1}{\mathbb{P}(i \in NP \mid i \in CE, \mathbf{x}_i)}.$$

The final estimate of μ is

$$\hat{y} = f_{CE} \frac{\sum_{i \in CE} y_i \, W_{CE,i}}{\sum_{i \in CE} W_{CE,i}} + (1 - f_{CE}) \frac{\sum_{i \in U \setminus CE} y_i \, W_{U \setminus CE,i}}{\sum_{i \in U \setminus CE} W_{U \setminus CE,i}},$$

where f_{CE} is the proportion of the CE set in the population. The nonprobability samples are drawn as in the SRS sample cases, with inclusion fractions $f_{NP} = \{3, 5, 7, 10, 30\}$ percent. The results are shown in Table 4.5. The presented MSE is again the original MSE times 10^{-2} for *ovt* and times 10^{-5} for *earnings*. Balancing two samples with either under- or oversampling does not improve the performance in many cases when the probability sample is from a stratified SRS.

TABLE 4.5: MSE and Relative Bias for stratified SRS samples. The smallest (absolute) value of each row is bold.

	%	MSE					
	f_{NP}	Naive	LSW	OS	OS2	US	Mix
ovt	3	217.845	**4.459**	4.773	83.233	4.903	4.844
	5	206.626	**3.305**	3.809	50.437	4.209	3.880
	7	203.416	**1.968**	2.462	39.838	2.615	2.481
	10	195.689	**1.145**	1.596	30.212	1.708	1.619
	30	198.902	**0.457**	2.585	26.257	2.567	2.589
earnings	3	299.790	1.019	0.973	78.840	0.994	**0.939**
	5	293.386	**0.740**	0.942	44.595	0.977	0.978
	7	293.815	**0.539**	1.004	37.236	0.848	1.007
	10	291.984	**0.371**	1.041	30.370	0.809	1.033
	30	291.497	**0.592**	3.319	29.066	2.468	3.314
		RB (%)					
ovt	3	27.471	0.615	1.014	16.259	**0.223**	0.950
	5	27.014	**0.486**	1.173	12.465	0.573	1.138
	7	27.082	**0.314**	1.364	11.409	0.550	1.333
	10	26.694	**-0.043**	1.259	10.093	0.480	1.249
	30	27.106	**0.550**	2.793	9.585	2.277	2.792
earnings	3	27.091	1.073	1.045	13.831	0.227	**0.94**0
	5	26.883	0.991	1.199	10.386	**0.450**	1.162
	7	26.927	0.910	1.409	9.499	**0.546**	1.378
	10	26.861	0.728	1.453	8.556	**0.693**	1.434
	30	26.868	**1.028**	2.742	8.313	2.096	2.738

The only improvement we found is after undersampling, *earning* has a smaller relative bias.

4.5 Conclusion and Discussion

When correcting selection bias with a two-sample setup, an imbalanced combined sample is often seen and may hinder the propensity estimation. In this research, we evaluate algorithms that aim to deal with imbalance in the prediction problem in a selection bias correction framework. In general, as long as the target variable is not skewed, oversampling by SMOTE is the most promising approach for decreasing the

MSE and bias. The mixture of under- and oversampling also improves the parameter estimation. On the other hand, undersampling the nonprobability sample introduces additional variance and sometimes bias. In the application on (semi-) real data, we have two extremely skewed target variables with many zeroes. It can be seen that when the probability sample is a simple random sample, under- or oversampling still improves the performance, while this is not the case when the probability sample is from a stratified SRS. Variants of SMOTE may be examined for future studies. For example, different distance functions for choosing the nearest neighbors, different ways to draw the nearest neighbors, and different ways to combine the neighbors and the sample.

Data availability

The data set of the application can be found at https://doi.org/10.5281/zenodo.11095755, and the code can be found at https://github.com/cheerup731/Imbalanced.

4.6 Appendix

Here we present the derivation of (4.4) and (4.6).

Undersampling

For the proposed undersampling approach (Algorithm 3), the propensity model is estimated on $B^- = B_0 \cup B_1^-$. That is to say, instead of $O_i(B)$ from (4.1) we model

$$O_i(B^-) = \frac{\mathbb{P}(z_i = 1 \mid i \in B^-, \mathbf{x}_i)}{\mathbb{P}(z_i = 0 \mid i \in B^-, \mathbf{x}_i)}.$$

To find the correct pseudo-weights based on (4.2), we can work out the relationship between $O_i(B^-)$ and $O_i(B)$.

Define the subsampling fraction $p^- = n_{B0}F/n_{B1}(1-F)$. We obtain:

$$\begin{aligned}
\frac{\mathbb{P}(z_i = 1 \mid i \in B^-, \mathbf{x}_i)}{\mathbb{P}(z_i = 0 \mid i \in B^-, \mathbf{x}_i)} &= \frac{\mathbb{P}(z_i = 1, i \in B^- \mid \mathbf{x}_i)}{\mathbb{P}(z_i = 0, i \in B^- \mid \mathbf{x}_i)} \\
&= \frac{\mathbb{P}(z_i = 1, i \in B^-, i \in B \mid \mathbf{x}_i)}{\mathbb{P}(z_i = 0, i \in B^-, i \in B \mid \mathbf{x}_i)} \\
&= \frac{\mathbb{P}(z_i = 1, i \in B \mid \mathbf{x}_i) \cdot \mathbb{P}(i \in B^- \mid z_i = 1, i \in B, \mathbf{x}_i)}{\mathbb{P}(z_i = 0, i \in B \mid \mathbf{x}_i) \cdot \mathbb{P}(i \in B^- \mid z_i = 0, i \in B, \mathbf{x}_i)} \\
&= \frac{\mathbb{P}(z_i = 1, i \in B \mid \mathbf{x}_i) \cdot p^-}{\mathbb{P}(z_i = 0, i \in B \mid \mathbf{x}_i)} \\
&= \frac{\mathbb{P}(z_i = 1 \mid i \in B, \mathbf{x}_i)}{\mathbb{P}(z_i = 0 \mid i \in B, \mathbf{x}_i)} p^-.
\end{aligned}$$

In the second line, it was used that $B^- \subseteq B$. In the fourth line, it was used that

$$\mathbb{P}(i \in B^- \mid z_i = 1, i \in B, \mathbf{x}_i) = \mathbb{P}(i \in B^- \mid z_i = 1, i \in B) = p^-,$$

since the subsample is obtained by simple random sampling without replacement from B_1. In the same line, it was also used that $\mathbb{P}(i \in B^- \mid z_i = 0, i \in B, \mathbf{x}_i) = 1$.

We conclude that $O_i(B^-) = O_i(B)p^-$ or, equivalently, $O_i(B) = O_i(B^-)/p^-$. Hence, with this undersampling approach we can compute the pseudo-weights according to (4.2) as follows:

$$w_i = \frac{1}{\mathbb{P}(S_i^* = 1 \mid \mathbf{x}_i)} = 1 + \frac{d_i - 1}{O_i(B^-)} p^-. \tag{4.7}$$

Oversampling by SMOTE

For the proposed oversampling approach based on the original SMOTE method (Algorithm 4), the propensity model is estimated on $B^+ = B_0^+ \cup B_1$ to predict

$$O_i(B^+) = \frac{\mathbb{P}(z_i = 1 \mid i \in B^+, \mathbf{x}_i)}{\mathbb{P}(z_i = 0 \mid i \in B^+, \mathbf{x}_i)}.$$

This time, we work out the relationship between $O_i(B^+)$ and $O_i(B)$.

If we ignore the effect of rounding in Algorithm 4 and assume that the records in B_0 are placed in a random order before running the algorithm, it holds that

$$\mathbb{P}(i \in B \mid i \in B_0^+, \mathbf{x}_i) = \mathbb{P}(i \in B \mid i \in B_0^+) = \frac{|B \cap B_0^+|}{|B_0^+|} = \frac{|B_0|}{|B_0^+|} \equiv p^+.$$

Furthermore,

$$|B_0^+| = |B_0| + R = n_{B0} + \frac{n_{B1}}{F} - (n_{B0} + n_{B1}) = n_{B1}\frac{1 - F}{F}.$$

Hence, the probability p^+ can be expressed as $p^+ = n_{B0}F/n_{B1}(1 - F)$ $(= p^-)$.

We can now apply a similar derivation as in the case of undersampling, but here it is convenient to start with $O_i(B)$ rather than $O_i(B^+)$. (Note that this time

$B \subseteq B^+$ whereas previously we had $B^- \subseteq B$.) We obtain:

$$\begin{aligned}
\frac{\mathbb{P}(z_i = 1 \mid i \in B, \mathbf{x}_i)}{\mathbb{P}(z_i = 0 \mid i \in B, \mathbf{x}_i)} &= \frac{\mathbb{P}(z_i = 1, i \in B \mid \mathbf{x}_i)}{\mathbb{P}(z_i = 0, i \in B \mid \mathbf{x}_i)} \\
&= \frac{\mathbb{P}(z_i = 1, i \in B, i \in B^+ \mid \mathbf{x}_i)}{\mathbb{P}(z_i = 0, i \in B, i \in B^+ \mid \mathbf{x}_i)} \\
&= \frac{\mathbb{P}(z_i = 1, i \in B^+ \mid \mathbf{x}_i) \cdot \mathbb{P}(i \in B \mid z_i = 1, i \in B^+, \mathbf{x}_i)}{\mathbb{P}(z_i = 0, i \in B^+ \mid \mathbf{x}_i) \cdot \mathbb{P}(i \in B \mid z_i = 0, i \in B^+, \mathbf{x}_i)} \\
&= \frac{\mathbb{P}(z_i = 1, i \in B^+ \mid \mathbf{x}_i)}{\mathbb{P}(z_i = 0, i \in B^+ \mid \mathbf{x}_i) \cdot p^+} \\
&= \frac{\mathbb{P}(z_i = 1 \mid i \in B^+, \mathbf{x}_i)}{\mathbb{P}(z_i = 0 \mid i \in B^+, \mathbf{x}_i)} \frac{1}{p^+}.
\end{aligned}$$

In the fourth line, in addition to the above-defined probability p^+ it was also used that $\mathbb{P}(i \in B \mid z_i = 1, i \in B^+, \mathbf{x}_i) = 1$.

Hence, it follows that $O_i(B) = O_i(B^+)/p^+$. Within Algorithm 4 we can therefore compute the pseudo-weights according to (4.2) by:

$$w_i = \frac{1}{\mathbb{P}(S_i^* = 1 \mid \mathbf{x}_i)} = 1 + \frac{d_i - 1}{O_i(B^+)} p^+. \tag{4.8}$$

Note that, as $p^+ = p^-$, this expression actually has the same form as (4.7).

The expression for the pseudo-weight used in Algorithm 5 can be obtained by a combination of the two above derivations for undersampling and oversampling by SMOTE.

Oversampling by weighted SMOTE

For the proposed oversampling approach based on weighted SMOTE (Algorithm 6), the propensity model is again estimated on $B^+ = B_0^+ \cup B_1$ to predict $O_i(B^+)$. An

important difference compared to Algorithm 4 is that here the probability

$$\mathbb{P}(i \in B \mid i \in B_0^+, \mathbf{x}_i) = \frac{\mathbb{P}(i \in B, i \in B_0^+ \mid \mathbf{x}_i)}{\mathbb{P}(i \in B_0^+ \mid \mathbf{x}_i)} = \frac{\mathbb{P}(i \in B_0 \mid \mathbf{x}_i)}{\mathbb{P}(i \in B_0^+ \mid \mathbf{x}_i)}$$

cannot be reduced to a single value that does not depend on i. Given that each record $i \in B_0$ with regularized weight d_i^* is associated in B_0^+ with one copy of the original record and $d_i^* - 1$ synthetic records, it might be reasonable to assume that

$$\mathbb{P}(i \in B \mid i \in B_0^+, \mathbf{x}_i) = \frac{1}{1 + (d_i^* - 1)} = \frac{1}{d_i^*}.$$

Here, the regularized weight $d_i^* = d_i C_{B0}$, with C_{B0} as defined in Section 4.2.2 so that $\sum_{i \in B_0} d_i^* = n_{B1}(1 - F)/F$.

With the above assumption in place, we obtain analogously to the above derivation for Algorithm 4:

$$\frac{\mathbb{P}(z_i = 1 \mid i \in B, \mathbf{x}_i)}{\mathbb{P}(z_i = 0 \mid i \in B, \mathbf{x}_i)} = \frac{\mathbb{P}(z_i = 1 \mid i \in B^+, \mathbf{x}_i)}{\mathbb{P}(z_i = 0 \mid i \in B^+, \mathbf{x}_i)} d_i^*.$$

That is to say, $O_i(B) = O_i(B^+)d_i^* = O_i(B^+)d_i C_{B0}$. Within Algorithm 6 we can therefore compute the pseudo-weights according to (4.2) by:

$$w_i = \frac{1}{\mathbb{P}(S_i^* = 1 \mid \mathbf{x}_i)} = 1 + \frac{d_i - 1}{O_i(B^+)d_i C_{B0}}. \tag{4.9}$$

Chapter 5

Correcting Selection Bias in Contingency Tables

Abstract

When inferring population characteristics from a nonprobability sample, i.e., a sample that does not come from a known sampling scheme, it is crucial to correct the possible selection bias therein. Often selection bias correction methods focus on estimating the means or totals of target variables on the population level. However, researchers are often also interested in estimates within some subgroups of the population. In this paper, we apply two small area estimation methods on nonprobability samples. One is a design-based method using iterative proportional fitting, and the other one is a model-based method based on a hierarchical Bayesian model. These methods are combined with an often-used method for selection bias correction, namely pseudo-weighting. A simulation study and an application on a real dataset are presented. Although we do not find one method that is suitable in all scenarios, there are some patterns we observed that may assist the choice of method in the future.

Keywords: Nonprobability Sample, Selection Bias, Small Area Estimation, Pseudo Weighting

Liu, A.-C., Scholtus, S., Van Deun, K. & De Waal, T. Correcting Selection Bias in Contingency Tables. (Preparing for submission)

5.1 Introduction

Nonprobability samples, for example, online opt-in surveys, and administrative data, can offer a massive amount of data within a short period. Since the inclusion mechanism is unknown to researchers, the inference from a nonprobability sample may suffer from selection bias. A massive amount of literature has been discussing how to correct selection bias, see for example, Elliott and Valliant (2017); J. Rao (2020). Much of the existing literature focuses on estimating the population mean or total of a target variable. However, we are often not only interested in estimates at the highest level of aggregation but also in estimates for domains. For example, if the vaccination rate of a country was estimated from an online survey, researchers may be interested in not only the vaccination rate of the whole country but also the vaccination rate in different age groups.

In this paper, we discuss possible methods to estimate a contingency table of population proportions from a nonprobability sample. This is closely related to Small Area Estimation (SAE). Although small area estimation often refers to estimating population parameters of geographical areas, it is not necessarily limited to geographical areas but can also refer to domains of the population such as different age groups or education levels (J. N. Rao & Molina, 2015). A nonprobability sample is often large and it may not suffer that much from a common problem in small area estimation, such as observing only a few or even zero sample units in an area. However, area estimation from nonprobability samples faces the challenge of an unknown inclusion mechanism of the sample. Estimating the inclusion mechanism of the nonprobability sample may be done by for example the approaches by Elliott and Valliant (2017) or Liu et al. (2023) for large inclusion fractions. These approaches calculate pseudo-weights for the units in the nonprobability sample. The

pseudo-weights may be used in the same way as survey weights for a probability sample to obtain estimates for population parameters. However, these pseudo-weights generally result in a larger variability compared to survey weights in a probability sample.

We examine two promising methods in small area estimation to see how these methods perform in a nonprobability sample context. In small area estimation, the domain estimates may be improved by a design-based method or a model-based method. See Parker et al. (2023); Pfeffermann (2013) for recent reviews. In the current paper, we first of all apply a design-based method, namely Iterative Proportional Fitting (IPF), or raking (Ireland & Kullback, 1968). IPF can resolve possible measurement errors or sampling errors in tables (Little & Wu, 1991). IPF presumes the size N_k of each domain k $(k = 1, ..., K)$ is known, and the cell values of a contingency table of the target variable by domain are calibrated proportionally given the known domain size. IPF has been applied in Villalobos-Alíste, Scholtus, and de Waal (2024) and has shown promising results when integrating a probability sample and a nonprobability sample. A variety of estimation choices can be made when applying IPF to a nonprobability sample. For example, after a set of pseudo-weights is constructed for selection bias correction in the nonprobability sample, researchers can choose between either applying IPF on the weighted contingency table or using the estimated population parameters as known margins for IPF.

Besides IPF, we also look at a model-based method with hierarchical Bayesian models proposed by Vandendijck, Faes, Kirby, Lawson, and Hens (2016). This method was meant for small area estimation from a probability sample, where the inclusion probabilities or design weights of the probability sample are considered when modeling the target variable. For the nonprobability sample context, we replace the design weights with the constructed pseudo-weights so that the prediction model can also consider the information of the estimated selection mechanism of the

nonprobability sample.

In the following, we will start by introducing the estimation goal and the methods in Section 5.2. A simulation study with an artificial data set is presented in Section 5.3 to reflect different relations between the target variable, domains, and auxiliary variables. In Section 5.4, the methods are examined on a real data set, and the paper is concluded with a discussion in Section 5.5.

5.2 Methods

5.2.1 Setup

	Z_1	Z_2	Z_3	$\sum$
Y_0	p_{01}	p_{02}	p_{03}	$p_{0\cdot}$
Y_1	p_{11}	p_{12}	p_{13}	$p_{1\cdot}$
$\sum$	$p_{\cdot 1} = \mathrm{d}_1$	$p_{\cdot 2} = \mathrm{d}_2$	$p_{\cdot 3} = \mathrm{d}_3$	$p_{\cdot\cdot} = 1$

FIGURE 5.1: Illustration of the target contingency table with $K = 3$

Consider a finite population U of size N with index $i \in \{1, 2, ..., N\}$. The first goal is to estimate the population contingency table $Y \times Z$ of a target variable Y and a domain variable Z. We limit $Y \in \{0, 1\}$ to simplify the discussion. The domain variable Z may be a combination of more than one variable, for example, *Gender* × *Age group*. The proportion of each cell in $Y \times Z$ is denoted as p_{jk} where $j \in \{0, 1\}$ is the indicator of the corresponding categories of Y, and $k \in \{1, 2, ..., K\}$ is the index of the categories of Z. We use $\cdot$ to denote the sum over categories, for example, $p_{\cdot k} = \sum_j p_{jk}$ and $p_{\cdot\cdot} = 1$. Along with the contingency table, often the population

proportions of the target variable $p_{j\cdot}$, i.e., the marginal proportions of Y, are also of interest. An illustration of the target contingency table is in Figure 5.1.

Assume the population proportion of each domain k of Z is known and denoted as $d_k = N_k/N$, with N_k the population size of units in domain k. We have a nonprobability sample NP of size n_{NP} with Y, Z, and perhaps auxiliary variables X that govern the inclusion mechanism of the nonprobability sample or are related to Y to assist the selection bias correction process.

5.2.2 Correcting Selection Bias by Pseudo-Weighting

One often-used method for selection bias correction is pseudo-weighting. It assumes that the nonprobability sample comes from an unknown inclusion mechanism, and tries to capture the inclusion mechanism given the auxiliary information. The pseudo-weights for the units in the nonprobability sample (w_i) may be constructed from known population totals of the auxiliary variables, or from a probability sample drawn from the same finite population sharing some common auxiliary variables. Here we do not limit our discussion to one specific pseudo-weight construction method. Readers interested in various options of pseudo-weight construction may consult Elliott and Valliant (2017); J. Rao (2020); Wu (2022).

Many pseudo-weighting methods focus on estimating the population total or mean. A design-based estimator can then be used after the pseudo-weights are constructed. For example, the estimated $p_{1\cdot}$ may be

$$y_w = \frac{\sum_{i \in NP} w_i y_i}{\sum_{i \in NP} w_i}. \tag{5.1}$$

To estimate proportions for the cells of $Y \times Z$, one choice is simply summing up the pseudo-weights in each cell and dividing the sum by N or by the sum of pseudo-weights in the sample. However, similar to using design weights of a probability

sample in small area estimation, directly using pseudo-weights for SAE may not return a reliable estimate because of the large variation of the (pseudo-)weights and small sample sizes in the domains. In the following, we discuss two methods in SAE and combine them with pseudo-weights. The first one is a designed-based method, namely IPF, and the second one is a method based on a hierarchical Bayesian model.

5.2.3 Design-Based SAE Model

Here we introduce the IPF process from Ireland and Kullback (1968). IPF adjusts a contingency table given some known marginal distributions, and it is not limited to binary variables but is applicable to all types of categorical variables. Assume we have a seed table with cell values $p_{jk}^{(0)}$. Given d_k is known, the cell values of the target table may be estimated by $p_{jk}^{(1)} = p_{jk}^{(0)} d_k / p_{\cdot k}^{(0)}$. If more than one margin is known, the cell values will be updated iteratively. That is, if the marginal distribution of Y is also known, then in iteration $t+1$ the cell values are updated in two steps: $p_{jk}^{*(t)} = p_{jk}^{(t)} d_k / p_{\cdot k}^{(t)}$ and $p_{jk}^{(t+1)} = p_{jk}^{*(t)} p_{j\cdot} / p_{j\cdot}^{*(t)}$, where the superscript (t) denotes the iteration index. If the maximum of $p_{jk}^{(t+1)} - p_{jk}^{(t)}$ within the table is larger than a given tolerance, the process is repeated by iterating over all known margins until the estimates converge or until a specified maximum number of iterations T is reached. Since IPF multiplies each cell value with some number, it is important that if the target population table does not contain any zeroes, the seed table should also not contain any zeroes to guarantee a reliable estimate. To combine pseudo-weights with IPF, we have several options:

1. Apply IPF to the seed table from the unweighted NP with d_k and $p_{1\cdot} = y_w$, $p_{0\cdot} = 1 - y_w$ as the known margins.

2. Apply IPF to the seed table from the weighted NP with only d_k as the known margin.

3. Apply IPF to the seed table from the weighted NP with d_k and $p_{1\cdot} = y_w$, $p_{0\cdot} = 1 - y_w$ as the known margins.

In the first option, the constructed pseudo-weights are not directly used for SAE, only the estimated margin y_w is used. This may prevent large variations of pseudo-weights in each cell, while the cell values may still be subject to selection bias. The second and the third options both apply pseudo-weights for SAE. In the third one, the estimated margin y_w is again used as the known margin, which results in the information of the pseudo-weights being used twice in this case. Therefore, the quality of the constructed pseudo-weights is especially important if option three is used.

5.2.4 Model-Based SAE Model

For the model-based SAE, we apply a hierarchical Bayesian model proposed by Vandendijck et al. (2016). The model is mainly for binary target variables. The proportion of $Y = 1$ for domain k $(k = 1, \cdots, K)$ is

$$\tilde{p}_k = \frac{p_{1k}}{p_{\cdot k}}.$$

We may estimate $\tilde{p}_k$ by a combination of the observed values y_{ik} in NP and estimated values $\hat{y}_{ik}$ in $U \setminus NP$

$$\text{est}(\tilde{p}_k) = \frac{1}{N_k}\{\sum_{i \in Z_k} R_i y_{ik} + \sum_{i \in Z_k}(1 - R_i)\hat{y}_{ik}\}, \tag{5.2}$$

where R_i is the inclusion indicator of NP. To acquire $\hat{y}_{ik}$, a model for y_{ik} is fitted in NP, which is assumed to have a form of

$$y_{ik}|\tilde{p}_{ik} \sim \text{Bernoulli}(\tilde{p}_{ik})$$

$$\text{logit}(\tilde{p}_{ik}) = \beta_0 + f(\tilde{w}_{ik}) + v_k + u_i, \tag{5.3}$$

where β_0 is a constant, and $f(\tilde{w}_{ik})$ is a function of regularized (pseudo-)weights. As suggested by Vandendijck et al. (2016), we regularize the sum of (pseudo-)weights by the sample size n_k in domain k, i.e.,

$$\tilde{w}_{ik} = \frac{w_{ik} n_k}{\sum_{i \in Z_k} w_{ik}}.$$

For the function $f(\tilde{w}_{ik})$, we apply a B-spline model which is shown to be promising (Vandendijck et al., 2016). It has a form of

$$f(\tilde{w}_{ik}) = \sum_{b=1}^{B} \theta_b B_b(\tilde{w}_{ik}),$$

where $B(\cdot)$ is a B-spline basis function and θ_b is a corresponding regression coefficient. We follow the original paper by Vandendijck et al. (2016) where the degree of B-spline functions equals 2 and $B = 20$ quantile knots. See Vandendijck et al. (2016) for more details. A domain-dependent random effect v_k in the original paper is the effect of the geographical area which is estimated by the mean of neighbors' areas. Instead of geographical areas, here we look at a scenario where the domains consist of some demographic traits of the units. Therefore, v_k is the random effect of the domain variable. In the simulation and application, we apply two domain variables, and the effect of each domain variable is then modeled separately. A unit random effect u_i with $E(u_i) = 0$ is also used.

To calculate $\hat{y}_{ik}$, since N_k is known, the only missing element in $U \setminus NP$ is $\tilde{w}_{ik}$. To estimate $\tilde{w}_{ik}$, Si, Pillai, and Gelman (2015); Vandendijck et al. (2016) assume that the (pseudo-)weights in the population consist of several strata $h \in \{1, 2, ..., H\}$, and units in the same stratum share the same (pseudo-)weight. The number of strata (H) is aligned with the number of unique values of the (pseudo-)weights in

each domain in the sample. As long as the population size of each stratum N_{hk} is known, we can then calculate $\hat{y}_{ik}$. To estimate the population size of each stratum, the sample size n_{hk} is assumed to follow a multinomial distribution.

$$(n_{1k}, n_{2k}, ..., n_{Hk}) \sim \text{Multinom}(n_k; \frac{N_{1k}/w_{1k}}{\sum_{h=1}^{H} N_{hk}/w_{hk}}, \frac{N_{2k}/w_{2k}}{\sum_{h=1}^{H} N_{hk}/w_{hk}}, ..., \frac{N_{Hk}/w_{Hk}}{\sum_{h=1}^{H} N_{hk}/w_{hk}})$$

The estimated population size of stratum h is regularized by the known population size of each domain N_k

$$\tilde{N}_{hk} = \frac{\hat{N}_{hk} N_k}{\sum_{h=1}^{H} \hat{N}_{hk}}$$

Finally, the second term in (5.2) is estimated by

$$\sum_{i \in Z_k} (1 - R_i)\hat{y}_{ik} = \sum_{i \in Z_k} \sum_{h=1}^{H} (\tilde{N}_{hk} - n_{hk})\hat{p}_{hk}.$$

5.3 Simulation

5.3.1 Simulation Setup

In the simulation, we examine the performance of the methods under different relations between the target variable, domain variables, and auxiliary variables. A finite population of size $N = 5,000$ is created. Two domain variables z_1 and z_2 are drawn from a Bernoulli distribution, where z_1 has four categories $\{1, 2, 3, 4\}$ with probabilities $\{0.1, 0.2, 0.3, 0.4\}$, and z_2 has three categories $\{1, 2, 3\}$ with probabilities $\{0.3, 0.3, 0.4\}$. Two auxiliary variables x_1 and x_2 are drawn from a normal distribution with mean equal to 5 and variance equal to one. The target binary variable y is created by the values of the domain variables and/or the auxiliary variables. That is, $p_i = \mathbb{P}(y_i = 1)$ is calculated by three y-models:

- Model 1: $\text{logit}\, p = -0.1x_1 - 0.1x_2 + 0.1z_1 + 0.1z_2$

- Model 2: $\text{logit}\, p = -0.1x_1 - 0.1x_2$
- Model 3: $\text{logit}\, p = -0.1x_2 + 0.1z_1 + e$, where $e \sim N(\mu = 0.2, \sigma = 1)$

Domain variables z_1 and z_2 are treated as numeric variables when specifying p. Nonprobability samples are drawn repeatedly 500 times by fixed-size unequal probability sampling without replacement with random systematic sampling. The inclusion probability of the nonprobability sample depends on the auxiliary variables with $\mathbb{P}(R = 1) = \exp(c + x_1 + x_2)/(1 + \exp(c + x_1 + x_2))$, where c governs the sizes of nonprobability samples with $\{500, 1000, 1500, 2000\}$. In Model 1, p depends on both domain and auxiliary variables, and therefore p_{jk} differs in each domain and $\hat{p}_{jk}$ from the nonprobability sample is subject to selection bias. In Model 2, p only depends on the auxiliary variables. That is, $p_{jk}/p_{\cdot k}$ is the same across domains. In Model 3, a noise variable e is added to reflect the scenario where not all the information of the target variable can be explained by domain and auxiliary variables, and therefore the selection bias is relatively small.

To construct the pseudo-weights, we apply the method proposed by Liu et al. (2023) since the inclusion fractions of the nonprobability samples are large. Probability samples with x_1 and x_2 are repeatedly drawn 500 times from the population by simple random sampling without replacement of size 500. The non-overlapping units of the nonprobability sample and the probability sample, referred as set V, are used to fit a logistic propensity model. The propensity model is fitted with the dependent variable $\tilde{R}_i = 1$ if unit i ($i \in V$) is from the nonprobability sample and $\tilde{R}_i = 0$ otherwise. Both auxiliary variables x_1 and x_2 are used as the independent variables in the propensity model. The weights for the units in the nonprobability sample are then

$$w_i = 1 + \frac{D_P - 1}{O_i}$$

where $D_P = 10$ is the design weight of the probability sample and

$$O_i = \frac{\mathbb{P}(\tilde{R}_i = 1|\mathbf{x}_i, i \in V)}{\mathbb{P}(\tilde{R}_i = 0|\mathbf{x}_i, i \in V)}$$

is calculated from the propensity model. The following methods are compared:

- naive: Treat NP as if it has been obtained by simple random sampling and use the corresponding direct estimator.
- LSW: Sum up w_i in each cell and divide by the sum of w_i in the whole non-probability sample.
- IPF: Apply IPF to NP with d_k as the known margin. This method reflects the ability of selection bias correction merely by IPF, and the information of w_i is not used. IPF is conducted by R package *mipfp* (Barthélemy & Suesse, 2018).
- IPF+y_w: Apply IPF to NP with both d_k and $p_{1\cdot} = y_w$, $p_{0\cdot} = 1 - y_w$ as the known margins.
- IPF.w: Apply IPF to the weighted NP with d_k as the known margin.
- IPF.w+y_w: Apply IPF to the weighted NP with d_k and $p_{1\cdot} = y_w$, $p_{0\cdot} = 1 - y_w$ as the known margins.
- Bayes.w: Hierarchical Bayesian Model as described in Section 5.2.4, where $v_k \in \{z_1, z_2\}$. R package *INLA* is used for Bayesian estimation and *splines* is used for B-splines (R Core Team, 2024; Rue, Martino, & Chopin, 2009).
- Bayes: Hierarchical Bayesian Model without modeling weights. That is, (5.3) is replaced by

$$\text{logit}(\tilde{p}_{ik}) = \beta_0 + v_k + u_i, \tag{5.4}$$

with $v_k \in \{z_1, z_2\}$.

The target table $y \times z_1 \times z_2$ has $2 \times 4 \times 3 = 24$ cells. The performance of the methods in SAE is evaluated by the Relative Mean Squared Error over all cells

$$\text{RMSE}_\text{A} = \frac{1}{M} \sum_{m=1}^{M} \sum_{j=0}^{1} \sum_{k=1}^{K} \frac{(\hat{p}_{jk}^m - p_{jk})^2}{p_{jk}},$$

and Mean Squared Error over all cells

$$\text{MSE}_\text{A} = \frac{1}{M} \sum_{m=1}^{M} \sum_{j=0}^{1} \sum_{k=1}^{K} (\hat{p}_{jk}^m - p_{jk})^2$$

where $\hat{p}_{jk}^m$ is the estimated value of cell (j, k) in replication $m \in \{1, 2, ..., M = 500\}$, and p_{jk} is the fixed value of cell (j, k) in the true table of the finite population. Note that RMSE_A equals the average value across M simulations of a normalized version of Pearson's chi-square test statistic for the estimated table with the true table as its reference distribution. The performance of the marginal distribution of y is estimated by bias of $p_{1\cdot}$

$$\text{Bias}_\text{Y} = \frac{1}{M} \sum_{m=1}^{M} \hat{p}_{1\cdot}^m - p_{1\cdot},$$

and Mean Squared Error of $p_{1\cdot}$

$$\text{MSE}_\text{Y} = \frac{1}{M} \sum_{m=1}^{M} (\hat{p}_{1\cdot}^m - p_{1\cdot})^2.$$

$\hat{p}_{1\cdot}^m$ is the estimated value of $p_{1\cdot}$ of replication m.

5.3.2 Results

Tables 5.1, 5.2, and 5.3 show the results of Models 1, 2, and 3 respectively. To make these tables easier to read, the presented values are 10^3 times the actual values. As expected, we can see that the pseudo-weighting method (LSW) worsens the small area estimation compared to naive estimation due to the additional variance from the estimation of the pseudo-weights. IPF-related methods improve the performance when the correction weights are not considered in the table (IPF, IPF+y_w). On the other hand, the hierarchical Bayesian models (Bayes.w and Bayes) perform better than IPF-related models. The Bayesian models considering the correction weights (Bayes.w) give the smallest RMSE_A and MSE_A in Models 1 and 2. When the quality of weights is not ideal, as in Model 3, the Bayesian model without considering the correction weights (Bayes) is better.

When estimating $p_{1\cdot}$, the pseudo-weighting method (LSW) in general improves the performance of the naive estimates. While in Model 3, Bias_Y of the naive estimate is small so that the MSE_Y of LSW is mainly subject to large variation. As long as y_w is used as the known marginal value, the performance of IPF for marginal estimation is always the same as LSW as expected. The weighted IPF (IPF.w) performs well in some cases in terms of bias. The Bayesian models return large Bias_Y but much smaller variance, and therefore result in smaller MSE_Y as well.

5.4 Application

5.4.1 Dataset

We use a real data set SD2011 from the R package *synthpop* as the target population to reflect a scenario where the relations between the target variable, auxiliary variables, and domain variables are unknown. The population size is 4,867, with the

target variable *sport* equal to one representing "Active engagement in some form of sport or exercise" and otherwise zero. The domain variables are *age group* and *education level*. *Age group* consists of six categories namely 16-24, 25-34, 35-44, 45-59, 60-64, and above 65. *Education level* consists of four categories namely "primary school or no education", "vocational or grammar school", "secondary school", and "post-secondary or higher". For each domain variable, one of the categories is treated as a reference category when estimating v_k in Bayesian related model. In total, we have $K = 6 \times 4 = 24$ domains and 48 cells in the table. The target population table is shown in Table 5.4. The inclusion probability of the nonprobability sample depends on two continuous variables: *depress* and *no friend* where

$$\mathbb{P}(R_i = 1) = \frac{n_{NP}(30 - \ depress_i + \ no\ friend_i)}{\sum_{i \in U}(30 - \ depress_i + \ no\ friend_i)}$$

and n_{NP} are $\{1500, 2000, 2500\}$. Simple random probability samples with *depress* and *no friend* are drawn with size 500. Both nonprobability samples and probability samples are repeatedly drawn 500 times. The estimation process is the same as that in the simulation. In practice, the inclusion information of the nonprobability sample may not always be known or available. To reflect the scenario of limited information, the propensity model is fitted by the correct model (*depress* and *no friend*) and an incorrect model (*depress* only).

5.4.2 Results

Table 5.5 and 5.6 show the results of the correct propensity model and incorrect propensity model. The presented RMSE_A are 10 times the actual values, and MSE_A, Bias_Y, and MSE_Y are 10^3 times the actual values. In terms of small area estimation, IPF, IPF+y_w, and Bayes.w all improve the performance compared to the naive approach. IPF+y_w is especially the best out of all the scenarios. For marginal

estimation, Bias_Y and MSE_Y of LSW are larger when the incorrect model is used, while the performance of LSW is still the best so that the estimators with constrained y_w also perform the best. It can also be seen that the Bayesian model is sensitive to the quality of the correction weights especially for marginal estimation.

5.5 Conclusion and Discussion

We examined two SAE methods and combined them with pseudo-weights to improve the estimation of contingency tables from a nonprobability sample. The simulation and application results show that no single method performs best for all the scenarios and estimands. A possible explanation for the differences between the results of the simulation study and the application, especially those for RMSE_A and MSE_A, maybe that in the application the naive approach already estimates the internal cell values of the contingency table quite accurately. We leave it to future research to examine if this is the cause for the differences between the results. Besides, the best model for SAE is not necessarily the best model for marginal estimation, although we may hope they both have good quality in practice. Some patterns are observed and may be useful for developing new methods in the future. For example, if the quality of the pseudo-weights is good, using the estimated margins as constraints in IPF improves the estimation compared to the naive approach. The weighted table, however, may return a worse SAE compared to naive estimates. On the other hand, the Bayesian model in general improves the SAE performance but does not always improve the performance of the marginal distribution. It seems interesting to investigate the possibility of constraining the marginal distribution in a Bayesian model so that both small area and marginal estimation can perform well. Also, although it is not shown in this paper, another benefit of the Bayesian model is that it can easily estimate the uncertainty of the estimates (Vandendijck et al., 2016).

In this paper, we only looked at the scenario where the target variable is only available in a nonprobability sample. Future studies may look into a scenario where the target variable is available in nonprobability and probability samples. Also, in practice, the target variable is not necessarily binary but may have more than two categories. Especially the Bayesian model needs to be extended for future studies involving target variables with more than two categories.

TABLE 5.1: Results of Model 1. The smallest (absolute) value of each row is bold.

	n_{NP}	naive	LSW	IPF	IPF+y_w	IPF.w	IPF.w+y_w	Bayes.w	Bayes
$RMSE_A$	500	49.95	196.96	30.68	30.22	77.06	79.64	**16.90**	23.02
($\times 10^3$)	1000	24.54	82.18	16.11	12.88	35.57	36.18	**7.95**	12.99
	1500	14.85	40.81	10.14	6.97	18.23	18.53	**4.69**	8.74
	2000	10.05	22.88	6.96	4.38	11.00	11.10	**3.15**	6.27
MSE_A	500	2.19	8.14	1.26	1.25	3.16	3.28	**0.72**	1.01
($\times 10^3$)	1000	1.06	3.36	0.67	0.54	1.44	1.47	**0.33**	0.57
	1500	0.63	1.68	0.42	0.29	0.74	0.76	**0.20**	0.38
	2000	0.42	0.94	0.29	0.18	0.44	0.45	**0.13**	0.27
$Bias_Y$	500	49.55	**-1.22**	49.29	**-1.22**	3.89	**-1.22**	28.02	51.46
($\times 10^3$)	1000	41.31	**1.40**	41.00	**1.40**	3.29	**1.40**	20.11	41.91
	1500	34.83	-1.23	34.53	-1.23	**-0.22**	-1.23	12.55	35.01
	2000	29.55	-0.39	29.26	-0.39	**0.16**	-0.39	7.41	29.53
MSE_Y	500	2.78	1.96	2.76	1.96	1.46	1.96	**1.32**	2.98
($\times 10^3$)	1000	1.85	0.83	1.82	0.83	0.71	0.83	**0.65**	1.90
	1500	1.29	0.42	1.27	0.42	0.36	0.42	**0.31**	1.31
	2000	0.93	0.25	0.91	0.25	0.22	0.25	**0.16**	0.92

TABLE 5.2: Results of Model 2. The smallest (absolute) value of each row is bold.

	n_{NP}	naive	LSW	IPF	IPF+y_w	IPF.w	IPF.w+y_w	Bayes.w	Bayes
$RMSE_A$	500	45.39	200.17	26.26	30.59	80.09	82.86	**15.69**	18.60
($\times 10^3$)	1000	21.92	84.99	13.56	12.93	37.53	38.48	**7.80**	10.66
	1500	13.22	41.79	8.51	6.85	19.34	19.64	**4.78**	7.26
	2000	9.08	23.41	5.97	4.45	11.59	11.73	**3.24**	5.29
MSE_A	500	1.96	8.12	0.90	1.08	2.78	2.89	**0.56**	0.68
($\times 10^3$)	1000	0.92	3.38	0.47	0.46	1.31	1.34	**0.28**	0.39
	1500	0.54	1.68	0.30	0.24	0.66	0.67	**0.17**	0.27
	2000	0.36	0.93	0.21	0.15	0.39	0.40	**0.11**	0.19
$Bias_Y$	500	36.77	**-1.37**	36.71	**-1.37**	2.32	**-1.37**	24.30	39.99
($\times 10^3$)	1000	31.84	**0.14**	31.81	**0.14**	1.99	**0.14**	18.64	33.22
	1500	27.22	-1.34	27.15	-1.34	**-0.46**	-1.34	12.29	27.92
	2000	23.40	-0.94	23.33	-0.94	**-0.41**	-0.94	7.77	23.78
MSE_Y	500	1.62	1.69	1.63	1.69	1.35	1.69	**1.07**	1.88
($\times 10^3$)	1000	1.13	0.72	1.13	0.72	0.60	0.72	**0.56**	1.22
	1500	0.81	0.35	0.81	0.35	0.31	0.35	**0.29**	0.85
	2000	0.59	0.21	0.58	0.21	0.19	0.21	**0.14**	0.60

TABLE 5.3: Results of Model 3. The smallest (absolute) value of each row is bold.

	n_{NP}	naive	LSW	IPF	IPF+y_w	IPF.w	IPF.w+y_w	Bayes.w	Bayes
$RMSE_A$	500	39.25	192.96	20.73	27.45	73.74	75.47	11.79	**10.36**
($\times 10^3$)	1000	16.88	79.42	8.65	11.28	33.41	33.91	5.60	**4.93**
	1500	9.59	40.40	4.88	6.23	17.75	17.89	3.48	**3.05**
	2000	6.27	22.24	3.16	3.86	10.36	10.46	2.46	**2.14**
MSE_A	500	1.60	8.25	0.82	1.18	3.28	3.37	0.54	**0.46**
($\times 10^3$)	1000	0.69	3.36	0.35	0.48	1.47	1.50	0.26	**0.22**
	1500	0.40	1.71	0.21	0.27	0.79	0.79	0.16	**0.14**
	2000	0.26	0.95	0.13	0.17	0.45	0.46	0.11	**0.10**
$Bias_Y$	500	1.47	**-0.62**	0.96	**-0.62**	1.78	**-0.62**	6.17	1.06
($\times 10^3$)	1000	4.09	**0.29**	3.95	**0.29**	1.57	**0.29**	6.43	3.94
	1500	4.72	-1.21	4.61	-1.21	**-0.65**	-1.21	6.31	4.58
	2000	5.42	**0.15**	5.38	**0.15**	0.36	**0.15**	6.90	5.31
MSE_Y	500	**0.33**	2.07	0.34	2.07	1.51	2.07	0.61	0.34
($\times 10^3$)	1000	**0.17**	0.84	0.17	0.84	0.69	0.84	0.32	0.17
	1500	0.11	0.45	0.11	0.45	0.41	0.45	0.21	**0.11**
	2000	0.09	0.27	0.09	0.27	0.24	0.27	0.16	**0.09**

TABLE 5.4: Target population table of the application (%)

Y	Age	Education	Freq	Y	Age	Education	Freq
Yes	16-24	primary/no edu.	0.23	No	16-24	primary/no edu.	0.62
Yes	25-34	primary/no edu.	0.43	No	25-34	primary/no edu.	0.23
Yes	35-44	primary/no edu.	1.17	No	35-44	primary/no edu.	0.25
Yes	45-59	primary/no edu.	3.35	No	45-59	primary/no edu.	0.47
Yes	60-64	primary/no edu.	2.32	No	60-64	primary/no edu.	0.43
Yes	65+	primary/no edu.	9.08	No	65+	primary/no edu.	0.64
Yes	16-24	vocational/grammar	2.14	No	16-24	vocational/grammar	4.73
Yes	25-34	vocational/grammar	2.26	No	25-34	vocational/grammar	0.76
Yes	35-44	vocational/grammar	3.99	No	35-44	vocational/grammar	1.44
Yes	45-59	vocational/grammar	8.92	No	45-59	vocational/grammar	2.38
Yes	60-64	vocational/grammar	2.34	No	60-64	vocational/grammar	0.76
Yes	65+	vocational/grammar	2.12	No	65+	vocational/grammar	0.39
Yes	16-24	secondary	1.81	No	16-24	secondary	3.45
Yes	25-34	secondary	2.77	No	25-34	secondary	2.22
Yes	35-44	secondary	2.65	No	35-44	secondary	1.58
Yes	45-59	secondary	5.36	No	45-59	secondary	2.45
Yes	60-64	secondary	1.97	No	60-64	secondary	1.01
Yes	65+	secondary	3.60	No	65+	secondary	0.86
Yes	16-24	post-secondary/higher	0.41	No	16-24	post-secondary/higher	0.49
Yes	25-34	post-secondary/higher	2.18	No	25-34	post-secondary/higher	3.80
Yes	35-44	post-secondary/higher	1.99	No	35-44	post-secondary/higher	2.01
Yes	45-59	post-secondary/higher	1.93	No	45-59	post-secondary/higher	2.38
Yes	60-64	post-secondary/higher	0.95	No	60-64	post-secondary/higher	0.58
Yes	65+	post-secondary/higher	1.34	No	65+	post-secondary/higher	0.78
$p_{1\cdot}$			**65.3**	$p_{0\cdot}$			**34.7**

TABLE 5.5: Application results with the correct propensity model. The smallest (absolute) value of each row is bold.

	n_{NP}	naive	LSW	IPF	IPF+y_w	IPF.w	IPF.w+y_w	Bayes.w	Bayes
$RMSE_A$	1500	0.33	38.75	0.13	**0.12**	14.97	5.64	0.13	0.81
($\times 10$)	2000	0.25	39.43	0.08	**0.07**	15.17	5.64	0.09	0.59
	2500	0.20	39.44	0.06	**0.05**	15.29	5.60	0.07	0.42
MSE_A	1500	1.10	40.78	0.19	**0.18**	27.97	12.52	0.21	1.66
($\times 10^3$)	2000	0.93	41.43	0.13	**0.11**	28.41	12.62	0.15	1.22
	2500	0.84	41.51	0.09	**0.08**	28.51	12.48	0.11	0.84
$Bias_Y$	1500	-28.77	**-0.05**	-12.87	**-0.05**	-308.33	**-0.05**	10.51	-16.16
($\times 10^3$)	2000	-28.97	**0.03**	-13.48	**0.03**	-314.05	**0.03**	8.37	-16.49
	2500	-28.30	**1.06**	-12.64	**1.06**	-316.15	**1.06**	8.16	-15.89
MSE_Y	1500	0.93	**0.11**	0.25	**0.11**	97.28	**0.11**	0.22	0.35
($\times 10^3$)	2000	0.90	**0.08**	0.23	**0.08**	99.16	**0.08**	0.14	0.33
	2500	0.84	**0.05**	0.19	**0.05**	100.04	**0.05**	0.12	0.29

TABLE 5.6: Application results with the incorrect propensity model. The smallest (absolute) value of each row is bold.

	n_{NP}	naive	LSW	IPF	IPF+y_w	IPF.w	IPF.w+y_w	Bayes.w	Bayes
RMSE_A	1500	0.33	37.52	0.13	**0.12**	14.76	5.72	0.23	0.95
($\times 10$)	2000	0.25	38.22	0.08	**0.08**	14.90	5.72	0.19	0.70
	2500	0.20	38.18	0.06	**0.05**	14.98	5.68	0.19	0.49
MSE_A	1500	1.10	40.06	0.19	**0.18**	27.43	12.51	0.50	2.06
($\times 10^3$)	2000	0.93	40.67	0.13	**0.12**	27.79	12.60	0.44	1.51
	2500	0.84	40.69	0.09	**0.08**	27.83	12.47	0.47	1.03
Bias_Y	1500	-28.77	**-7.90**	-12.87	**-7.90**	-300.73	**-7.90**	29.28	-20.62
($\times 10^3$)	2000	-28.97	**-8.22**	-13.48	**-8.22**	-305.62	**-8.22**	30.56	-20.84
	2500	-28.30	**-7.73**	-12.64	**-7.73**	-307.20	**-7.73**	35.91	-19.73
MSE_Y	1500	0.93	**0.17**	0.25	**0.17**	92.45	**0.17**	1.22	0.52
($\times 10^3$)	2000	0.90	**0.15**	0.23	**0.15**	93.88	**0.15**	1.20	0.49
	2500	0.84	**0.11**	0.19	**0.11**	94.45	**0.11**	1.49	0.43

Chapter 6

Epilogue

6.1 Summary

Statistics are often estimated from a sample rather than from the entire population. If the inclusion probability of the sample is unknown to the researcher, that is, a nonprobability sample, naively treating the sample as a simple random sample may result in selection bias. Attention to correcting selection bias is increasing due to the availability of new data sources. These data are often easy to collect and may be so-called "Big Data" considering the large inclusion fraction of the population. This dissertation proposes a novel framework for correcting selection bias in nonprobability samples. The general idea is to construct a set of unit weights for the nonprobability sample by borrowing the strength of a reference probability sample. If a proper set of weights is constructed, design-based estimators can be used for population parameter estimation given the weights. To evaluate the uncertainty of the estimated population parameter, a pseudo population bootstrap procedure is proposed given different relations between the nonprobability sample and the probability sample.

Three practical challenges for pseudo-weighting are also discussed. The proposed framework is flexible and many kinds of probability estimation models can be used.

The question is raised about how to select a proper model given the population parameter in question. A series of performance measures are tested, and we found that modeling the target variable when evaluating the performance of weights may be useful. The second challenge comes from the large size of the nonprobability sample. Since we often have a large nonprobability sample assisted with a small probability sample, we end up with an imbalanced combined sample which can cause problems when estimating model parameters. Several remedies for imbalanced samples are discussed and the proposed framework is also adjusted accordingly. The results show that SMOTE is a promising technique for dealing with imbalanced samples. Finally, we look at the scenario where not only the population level estimates are of interest but also subpopulation estimates. Several approaches to combine pseudo-weights with small area estimation are discussed. Of all approaches, we found that combining a hierarchical Bayesian model with weights is a relatively stable estimation approach. If both population-level and area-level estimates are of interest, aligning the weighted estimates with estimated marginal totals may be a better option.

6.2 Discussion

The dissertation mainly focuses on correcting selection bias for descriptive inference such as a finite population mean or total, since often this is the goal for official statistics. Other population quantities may also be of interest, such as the relation between more than one target variable or the causal relation between variables. The attention may then be on not only a certain finite population but also a population across different times and space.

Also, in our method and many other pseudo-weighting methods, we assume the inclusion mechanism of the nonprobability sample exists but is unknown to the researcher. Some may find assuming the existence of an inclusion mechanism a strong

assumption. However, if the relation between the auxiliary variables and the target variables are strong enough, the discrepancy between the nonprobability sample and the finite population can be reduced by weighting even if the nonprobability sample is not from a probability sampling process.

It is necessary to again stress the importance of the auxiliary variables since many correction methods largely rely on the common auxiliary variables between datasets as the bridge between the nonprobability sample and the population in mind. Practitioners may already think of the selection bias correction step before collecting a nonprobability sample. That is, if the goal is to collect an online opt-in survey, the auxiliary variables that are already collected in some probability sample are worth being included in the opt-in survey as well.

Furthermore, since we look at the context of large inclusion fractions, an overlap between the nonprobability sample and the reference probability sample is inevitable. Some scenarios, such as the probability sample being a subset of the nonprobability sample, are not discussed in this dissertation. It may also be interesting to investigate a scenario where the researcher has more than one probability sample from which to choose. Given probability samples, deciding which sample is better for selection bias correction is still an open question. One direction may be considering the design of the probability sample to calculate the variance as shown in Scholtus, Liu, and De Waal (2024). It may be also interesting to combine more than one probability sample during the correction process.

Finally, it is important to note that, in practice, we often have multiple problems at the same time. We may need to deal with the possible measurement error in the common auxiliary variables before correcting the selection bias, or missingness may occur in the datasets. More research is needed to deal with multiple quality issues in one data set.

Chapter 7

Samenvatting

Statistieken worden vaak gebaseerd op een steekproef in plaats van de gehele populatie. Als de insluitkansen van de steekproef onbekend zijn bij de onderzoeker, kan het naïef behandelen van de steekproef als een enkelvoudig aselecte steekproef leiden tot vertekening (selectiebias). De aandacht voor het corrigeren van selectiebias neemt toe vanwege de beschikbaarheid van nieuwe gegevensbronnen. Deze gegevens zijn vaak eenvoudig te verzamelen en kunnen zogenaamde "Big Data" worden genoemd vanwege de grote inclusiefractie van de populatie. Dit proefschrift stelt een nieuw raamwerk voor om selectiebias in niet-kanssteekproeven te corrigeren. Het algemene idee is om een set gewichten voor eenheden van de niet-kanssteekproef te construeren door informatie van een referentiekanssteekproef te lenen. Als een juiste set gewichten wordt geconstrueerd, kunnen op deze gewichten gebaseerde schatters worden gebruikt voor het schatten van populatieparameters. Om de onzekerheid van de geschatte populatieparameter te evalueren, wordt een pseudo-populatiebootstrap voorgesteld, gegeven verschillende relaties tussen de niet-kanssteekproef en de kanssteekproef.

Drie praktische uitdagingen voor pseudo-weging worden ook besproken. Het

voorgestelde raamwerk is flexibel en er kunnen veel soorten schattingsmodellen worden gebruikt. De vraag hoe een geschikt model te selecteren gegeven de populatieparameter waarin we geïnteresseerd zijn, werd gesteld. Een reeks prestatiematen wordt getest en dit laat zien dat het modelleren van de doelvariabele bij het evalueren van de prestatie van gewichten nuttig kan zijn. De tweede uitdaging komt door de grote omvang van de niet-kanssteekproef. Omdat we vaak een grote niet-kanssteekproef hebben met een kleine kanssteekproef, eindigen we met een onevenwichtige gecombineerde steekproef en dit kan leiden tot schattingsproblemen. Verschillende oplossingen voor onevenwichtige steekproeven worden besproken en het voorgestelde raamwerk wordt ook dienovereenkomstig aangepast. De resultaten laten zien dat SMOTE veelbelovend is voor het omgaan met onevenwichtige steekproeven. Tot slot kijken we naar het scenario waarin niet alleen de schattingen op populatieniveau van belang zijn, maar ook schattingen van subpopulaties. Verschillende manieren om pseudo-gewichten te combineren met schattingen van kleine domeinen worden besproken. Van alle manieren vonden we dat het combineren van een hiërarchisch Bayesiaans model met gewichten een relatief stabiele schattingsmethode is. Als zowel schattingen op populatieniveau als op domeinniveau van belang zijn, kan het benchmarken van de gewogen schattingen op de geschatte marginale totalen een betere optie zijn.

Acknowledgements

This dissertation would not be possible without the help of many people. First of all, I would like to express my appreciation to Ton, who started this super interesting project. Ton has guided me since my master's studies. He read every draft of mine in detail and patiently gave me all the suggestions. He introduced me to the big academic world, showing me how things work, and never forgot to embarrass me by telling people, "This is my brilliant PhD student." I learned a lot from him, not only professional skills but also life lessons. It is a great pleasure to work with Ton.

A great appreciation also to Katrijn. I started my PhD during the pandemic. Although the time was hard, Katrijn did her best to make sure I felt at home, even though we could not see each other in person for quite a while. Somehow, having an online meeting in a toilet became something I was looking forward to. She is always willing to share how she dealt with all the difficulties at work and is the best role model for being a warm and welcoming person. She is definitely the coolest one among the cool people.

I would also like to give great appreciation to Sander. Having discussions with Sander is the best highlight in my research life. Many ideas are generated during the discussions with him, and he always manages to drag me into more interesting questions in a magical way. Every now and then, we receive a long, detailed document from him that answers certain unsolved problems we had before. Not to mention how he managed to spot all the errors, logical mistakes, and inconsistencies in the papers, while still being the most supportive person and patiently listening

to all my silly thoughts. I still do not understand how a human being can achieve all these things altogether at the same time.

A special thanks to our co-author, Lyndon. I met Lyndon during his visit to CBS and found that we share the same research interests. This common interest resulted in some chestnut pictures and our imbalanced paper. It is a pleasure to collaborate with someone from the other side of the earth.

During my studies, I was very lucky to be involved in many study groups. From the cool people to the latter data science group, I had an opportunity to see many kinds of research fields, and most importantly, I got wonderful friendships from them. In CBS, we have reading groups for nonprobability samples and data integration. It is such a privilege to have a reading group with lots of experienced researchers discussing topics that are close to mine.

I would like to thank my dear family. Without their support, I would not be in this country far away in the first place. Whenever I encounter a difficulty, I think of how they dealt with all those challenges in even trickier scenarios in the past. It gives me lots of strength and power. Finally, Yeh, thank you for being by my side all the time. I would not be able to go this far without you.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, *19*(6), 716–723. doi: doi:10.1109/TAC.1974.1100705

Anderson, E. (1935). The irises of the gaspe peninsula. *Bull. Am. Iris Soc.*, *59*, 2–5.

Andridge, R. R., & Little, R. J. (2009). The use of sample weights in hot deck imputation. *Journal of Official Statistics*, *25*(1), 21.

Andridge, R. R., West, B. T., Little, R. J., Boonstra, P. S., & Alvarado-Leiton, F. (2019). Indices of non-ignorable selection bias for proportions estimated from non-probability samples. *Journal of the Royal Statistical Society. Series C, Applied Statistics*, *68*(5), 1465. doi: doi:10.1111/rssc.12371

Ang, L., Clark, R., Loong, B., & Holmberg, A. (2024). Synthetic business population containing simulated business variables. Zenodo. doi: doi:10.5281/zenodo.11095755

Antal, E., & Tillé, Y. (2014). A new resampling method for sampling designs without replacement: the doubled half bootstrap. *Computational Statistics*, *29*(5), 1345–1363.

Austin, P. C. (2009). Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Statistics in Medicine*, *28*(25), 3083–3107. doi: doi:10.1002/sim.3697

Baker, R., Brick, J. M., Bates, N. A., Battaglia, M., Couper, M. P., Dever, J. A.,

... Tourangeau, R. (2013). Summary report of the aapor task force on non-probability sampling. *Journal of Survey Statistics and Methodology*, *1*(2), 90–143.

Barthélemy, J., & Suesse, T. (2018). mipfp: An R package for multidimensional array fitting and simulating multivariate Bernoulli distributions. *Journal of Statistical Software, Code Snippets*, *86*(2), 1–20. doi: doi:10.18637/jss.v086.c02

Batista, G. E., Prati, R. C., & Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD explorations newsletter*, *6*(1), 20–29.

Beaumont, J.-F. (2020). Are probability surveys bound to disappear for the production of official statistics? *Survey Methodology*, *46*(1), 1–29.

Bethel, J. (1989). Sample Allocation in Multivariate Surveys. *Survey Methodology*, *15*(1), 47–57. Retrieved 2023-03-29, from `https://www150.statcan.gc.ca/n1/en/catalogue/12-001-X198900114578`

Bethlehem, J. (2010). Selection bias in web surveys. *International Statistical Review*, *78*(2), 161–188. doi: doi:10.1111/j.1751-5823.2010.00112.x

Blagus, R., & Lusa, L. (2013). Smote for high-dimensional class-imbalanced data. *BMC bioinformatics*, *14*, 1–16.

Boonstra, P. S., Little, R. J., West, B. T., Andridge, R. R., & Alvarado-Leiton, F. (2021). A simulation study of diagnostics for selection bias. *Journal of official statistics*, *37*(3), 751–769. doi: doi:10.2478/jos-2021-0033

Booth, J. G., Butler, R. W., & Hall, P. (1994). Bootstrap methods for finite populations. *Journal of the American Statistical Association*, *89*(428), 1282–1289.

Bradley, V. C., Kuriwaki, S., Isakov, M., Sejdinovic, D., Meng, X.-L., & Flaxman, S. (2021). Unrepresentative big surveys significantly overestimated us vaccine uptake. *Nature*, *600*(7890), 695–700. doi: doi:10.1038/s41586-021-04198-4

Brick, J. M. (2013). Unit nonresponse and weighting adjustments: A critical review. *Journal of Official Statistics*, *29*(3), 329.

Brier, G. W., et al. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, *78*(1), 1–3. doi: doi:10.1175/1520-0493(1950)078%3C0001:VOFEIT%3E2.0.CO;2

Buelens, B., Burger, J., & van den Brakel, J. A. (2018). Comparing inference methods for non-probability samples. *International Statistical Review*, *86*(2), 322–343.

Caruana, R., & Niculescu-Mizil, A. (2004). Data mining in metric space: an empirical analysis of supervised learning performance criteria. In *Proceedings of the tenth acm sigkdd international conference on knowledge discovery and data mining* (pp. 69–78).

Castro-Martín, L., Rueda, M. d. M., & Ferri-García, R. (2020). Inference from non-probability surveys with statistical matching and propensity score adjustment using modern prediction techniques. *Mathematics*, *8*(6), 879. doi: doi:10.3390/math8060879

Chambers, R. L. (2001). *Evaluation criteria for statistical editing and imputation* (No. 28). Office for National Statistics.

Chauvet, G. (2007). *Méthodes de bootstrap en population finie* (Unpublished doctoral dissertation). ENSAI.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, *16*, 321–357.

Chen, S., Haziza, D., Léger, C., & Mashreghi, Z. (2019). Pseudo-population bootstrap methods for imputed survey data. *Biometrika*, *106*(2), 369–384.

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In

Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785–794). doi: doi:10.1145/2939672.2939785

Chen, Y., Li, P., & Wu, C. (2020). Doubly robust inference with nonprobability survey samples. *Journal of the American Statistical Association*, *115*(532), 2011–2021. doi: doi:10.1080/01621459.2019.1677241

Chromy, J. (1987). Design Optimization With Multiple Objectives. In *Proceedings of the Section on Survey Research Methods* (pp. 194–199). American Statistical Association.

Cochran, W. G. (1977). *Sampling techniques*. John Wiley & Sons Location New York.

Cornesse, C., Blom, A. G., Dutwin, D., Krosnick, J. A., De Leeuw, E. D., Legleye, S., ... others (2020). A review of conceptual approaches and empirical evidence on probability and nonprobability sample survey research. *Journal of Survey Statistics and Methodology*, *8*(1), 4–36.

Czajka, J. L., Hirabayashi, S. M., Little, R. J., & Rubin, D. B. (1992). Projecting from advance data using propensity modeling: An application to income and tax statistics. *Journal of Business & Economic Statistics*, *10*(2), 117–131.

Deville, J.-C., & Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, *87*(418), 376–382. doi: doi:10.1080/01621459.1992.10475217

Dong, Q., Elliott, M. R., & Raghunathan, T. E. (2014). A nonparametric method to generate synthetic populations to adjust for complex sampling design features. *Survey Methodology*, *40*(1), 29.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, *7*(1), 1–26.

Efron, B. (2020). Prediction, estimation, and attribution. *International Statistical Review*, *88*, S28–S59. doi: doi:10.1111/insr.12414

Elliott, M. R., & Davis, W. W. (2005). Obtaining cancer risk factor prevalence estimates in small areas: combining data from two surveys. *Journal of the Royal Statistical Society Series C: Applied Statistics*, *54*(3), 595–609.

Elliott, M. R., & Valliant, R. (2017). Inference for nonprobability samples. *Statistical Science*, 249–264. doi: doi:10.1214/16-STS598

Estabrooks, A., Jo, T., & Japkowicz, N. (2004). A multiple resampling method for learning from imbalanced data sets. *Computational intelligence*, *20*(1), 18–36.

Fellegi, I. P. (1975). Controlled random rounding. *Survey Methodology*, *1*, 123–135.

Fernández, A., Garcia, S., Herrera, F., & Chawla, N. V. (2018). Smote for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. *Journal of artificial intelligence research*, *61*, 863–905.

Fuller, W. A. (2011). *Sampling statistics* (Vol. 560). John Wiley & Sons.

Gelman, A. (2007). Struggles with survey weighting and regression modeling. *Statistical Science*, *22*(2), 153–164.

Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, *102*(477), 359–378.

Hájek, J. (1971). Comment on "an essay on the logical foundations of survey sampling, part one". *The foundations of survey sampling*, *236*.

Haziza, D., & Beaumont, J.-F. (2017). Construction of weights in surveys: A review. *Statistical Science*, *32*(2), 206–226.

Heckman, J. J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. In *Annals of economic and social measurement, volume 5, number 4* (pp. 475–492). NBER.

Horvitz, D. G., & Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical*

Association, *47*(260), 663–685. doi: doi:10.1080/01621459.1952.10483446

Ireland, C. T., & Kullback, S. (1968). Contingency tables with given marginals. *Biometrika*, *55*(1), 179–188.

Kendall, M. G. (1948). Rank correlation methods.

Kern, C., Li, Y., & Wang, L. (2021). Boosted kernel weighting–using statistical learning to improve inference from nonprobability samples. *Journal of Survey Statistics and Methodology*, *9*(5), 1088–1113. doi: doi:10.1093/jssam/smaa028

Kim, J.-K., & Tam, S.-M. (2020). Data integration by combining big data and survey sample data for finite population inference. *International Statistical Review*.

Kim, J. K., & Wang, Z. (2019). Sampling techniques for big data analysis. *International Statistical Review*, *87*, S177–S191. doi: doi:10.1111/insr.12290

Kish, L. (1965). *Survey sampling.* Wiley.

Klingwort, J., & Burger, J. (2023). A framework for population inference: Combining machine learning, network analysis, and non-probability road sensor data. *Computers, Environment and Urban Systems*, *103*, 101976. doi: doi:10.1016/j.compenvurbsys.2023.101976

Kullback, S. (1997). *Information theory and statistics.* Courier Corporation.

Little, R. J. (2004). To model or not to model? competing modes of inference for finite population sampling. *Journal of the American Statistical Association*, *99*(466), 546–556.

Little, R. J., West, B. T., Boonstra, P. S., & Hu, J. (2020). Measures of the degree of departure from ignorable sample selection. *Journal of Survey Statistics and Methodology*, *8*(5), 932–964. doi: doi:10.1093/jssam/smz023

Little, R. J., & Wu, M.-M. (1991). Models for contingency tables with known margins when target and sampled populations differ. *Journal of the American*

Statistical Association, *86*(413), 87–95.

Little, R. J., & Zheng, H. (2007). The Bayesian approach to the analysis of finite population surveys. *Bayesian Statistics*, *8*(1), 1–20.

Liu, A.-C., Scholtus, S., & De Waal, T. (2023). Correcting selection bias in big data by pseudo weighting. *Journal of Survey Statistics and Methodology*. doi: doi:http://doi.org/10.1093/jssam/smac029

Lumley, T. (2020). *Survey: analysis of complex survey samples.* (R package version 4.0)

Lumley, T., & Scott, A. (2015). Aic and bic for modeling with complex survey data. *Journal of Survey Statistics and Methodology*, *3*(1), 1–18. doi: doi:10.1093/jssam/smu021

Madow, W. G. (1949). On the theory of systematic sampling, ii. *The Annals of Mathematical Statistics*, *20*(3), 333–354. doi: doi:10.1214/aoms/1177729988

Marella, D. (2023). Adjusting for selection bias in nonprobability samples by empirical likelihood approach. *Journal of Official Statistics*, *39*(2), 151–172. doi: doi:10.2478/jos-2023-0008

Mashreghi, Z., Haziza, D., & Léger, C. (2016). A survey of bootstrap methods in finite population sampling. *Statistics Surveys*, *10*, 1–52.

McCaffrey, D. F., Ridgeway, G., & Morral, A. R. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological Methods*, *9*(4), 403. doi: doi:10.1037/1082-989X.9.4.403

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. *Annals of Applied Statistics*, *12*(2), 685–726. doi: doi:10.1214/18-AOAS1161SF

Meng, X.-L. (2022). Comments on "statistical inference with non-probability

survey samples" – miniaturizing data defect correlation: A versatile strategy for handling non-probability samples. *Survey Methodology*, *48*(2), 339–360. Retrieved from `http://www.statcan.gc.ca/pub/12-001-x/2022002/article/00006-eng.htm`

Mercer, A. W., Kreuter, F., Keeter, S., & Stuart, E. A. (2017). Theory and practice in nonprobability surveys: parallels between causal inference and survey inference. *Public Opinion Quarterly*, *81*(S1), 250–271.

Mukherjee, M., & Khushi, M. (2021). Smote-enc: A novel smote-based method to generate synthetic data for nominal and continuous features. *Applied System Innovation*, *4*(1), 18.

Nishimura, R., Wagner, J., & Elliott, M. (2016). Alternative indicators for the risk of non-response bias: a simulation study. *International Statistical Review*, *84*(1), 43–62. doi: doi:10.1111/insr.12100

Ouwehand, P., & Schouten, B. (2014). Measuring representativeness of short-term business statistics. *Journal of Official Statistics*, *30*(4), 623–649.

Parker, P. A., Janicki, R., & Holan, S. H. (2023). A comprehensive overview of unit-level modeling of survey data for small area estimation under informative sampling. *Journal of Survey Statistics and Methodology*, *11*(4), 829–857.

Pfeffermann, D. (2013). New important developments in small area estimation. *Statistical Science*, 40–68.

R Core Team. (2024). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org/`

Rafei, A., Flannagan, C. A., & Elliott, M. R. (2020). Big data for finite population inference: Applying quasi-random approaches to naturalistic driving data using bayesian additive regression trees. *Journal of Survey Statistics and Methodology*, *8*(1), 148–180.

Rafei, A., Flannagan, C. A., West, B. T., & Elliott, M. R. (2022). Robust bayesian inference for big data: Combining sensor-based records with traditional survey data. *The Annals of Applied Statistics*, *16*(2), 1038–1070.

Rao, J. (1966). Alternative estimators in pps sampling for multiple characteristics. *Sankhyā: The Indian Journal of Statistics, Series A*, 47–60.

Rao, J. (2020). On making valid inferences by integrating data from surveys and other sources. *Sankhya B*, 1–31. doi: doi:10.1007/s13571-020-00227-w

Rao, J., Wu, C., & Yue, K. (1992). Some recent work on resampling methods for complex surveys. *Survey Methodology*, *18*(2), 209–217.

Rao, J. N., & Molina, I. (2015). *Small area estimation*. John Wiley & Sons.

Robbins, M. W., Ghosh-Dastidar, B., & Ramchand, R. (2021). Blending probability and nonprobability samples with applications to a survey of military caregivers. *Journal of Survey Statistics and Methodology*, *9*(5), 1114–1145.

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

Rue, H., Martino, S., & Chopin, N. (2009). Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *71*(2), 319–392.

Scholtus, S., Liu, A.-C., & De Waal, T. (2024). Notes on the variance of a pseudo-weighted estimator for selection bias correction. *METRON*. doi: doi:10.1007/s40300-024-00284-5

Schonlau, M., & Couper, M. P. (2017). Options for conducting web surveys. *Statistical Science*, *32*(2), 279–292.

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, 461–464. doi: doi:10.1214/aos/1176344136

Si, Y., Pillai, N. S., & Gelman, A. (2015). Bayesian nonparametric weighted sampling inference. *Bayesian Analysis*, *10*(3), 605–625.

Tillé, Y., & Matei, A. (2021). sampling: Survey sampling [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=sampling` (R package version 2.9)

Valliant, R. (2020). Comparing alternatives for estimation from nonprobability samples. *Journal of Survey Statistics and Methodology*, *8*(2), 231–263. doi: doi:10.1093/jssam/smz003

Valliant, R., & Dever, J. A. (2011). Estimating propensity adjustments for volunteer web surveys. *Sociological Methods & Research*, *40*(1), 105–137. doi: doi:10.1177/0049124110392533

Vandendijck, Y., Faes, C., Kirby, R. S., Lawson, A., & Hens, N. (2016). Model-based inference for small area estimation with sampling weights. *Spatial Statistics*, *18*, 455–473.

Vermunt, J. K., Van Ginkel, J. R., Van der Ark, L. A., & Sijtsma, K. (2008). Multiple imputation of incomplete categorical data using latent class analysis. *Sociological Methodology*, *38*(1), 369–397.

Vidotto, D., Kaptein, M., & Vermunt, J. (2015). Multiple imputation of missing categorical data using latent class models: State of art. *Psychological test and assessment modeling*, *57*(4), 542–576.

Villalobos-Alíste, S., Scholtus, S., & de Waal, T. (2024). Combining probability and non-probability samples on an aggregated level. *Accepted for publication by Journal of Official Statistics*.

Wang, L., Valliant, R., & Li, Y. (2021). Adjusted logistic propensity weighting methods for population inference using nonprobability volunteer-based epidemiologic cohorts. *Statistics in Medicine*, *40*(24), 5237–5250.

West, B. T., Little, R. J., Andridge, R. R., Boonstra, P. S., Ware, E. B., Pandit,

A., & Alvarado-Leiton, F. (2021). Assessing selection bias in regression coefficients estimated from nonprobability samples with applications to genetics and demographic surveys. *The annals of applied statistics*, *15*(3), 1556–1581. doi: doi:10.1214/21-AOAS1453

Wu, C. (2022). Statistical inference with non-probability survey samples. *Survey Methodology*, *48*(2), 283–311. Retrieved from `http://www.statcan.gc.ca/pub/12-001-x/2022002/article/00002-eng.htm`

Wu, C., & Sitter, R. R. (2001). A model-calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association*, *96*(453), 185–193. doi: doi:10.1198/016214501750333054

Wu, C., & Thompson, M. E. (2020). *Sampling theory and practice*. Springer.

Yang, S., Kim, J. K., & 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 (Statistical Methodology)*, *82*(2), 445–465. doi: doi:10.1111/rssb.12354

Zhang, L.-C. (2019). On valid descriptive inference from non-probability sample. *Statistical Theory and Related Fields*, *3*(2), 103–113. doi: doi:10.1080/24754269.2019.1666241