Statistical methods for combining probability and non-probability based sampling approaches

Zhao, Mingyao

Award date:
2024

Department of Mathematics and Computer Science

Statistical methods for combining probability and non-probability based sampling approaches

Master Thesis

Student Name: Mingyao Zhao
Student Number: 1678027

Supervisors:
Edwin van den Heuvel
Bart Smeulders
Maurits Kaptein

31-07-2024

Acknowledgements

I would like to express my deepest gratitude to my supervisor, Prof. Edwin Van den Heuvel, for his unwavering guidance and patience throughout this project and the numerous revisions of the thesis. Prof. Van den Heuvel introduced me to this exciting research topic, marking the start of my challenging yet rewarding journey. Under his mentorship, I gained substantial knowledge in the subject and learned the importance of working with seriousness and meticulousness, which will greatly benefit my future career.

I would also like to thank all my committee members, Prof. Maurits Kaptein and Prof. Bart Smeulders, for dedicating their time to reviewing my work.

I am profoundly grateful for the support from my parents, who have supported me both financially and emotionally during my studies in the Netherlands. I also cherish the friendships I have made along the way. I could not have reached this milestone without their help and encouragement.

Abstract

Cohort studies play a crucial role in establishing associations and causal relationships between variables, but biases in data collection methods can undermine their reliability. To better understand the health conditions in rural areas of the southern United States, the Risk Underlying Rural Areas Longitudinal Study (RURAL) is conducted. Participants are selected through probability-based sampling (PBS) using randomly collected household addresses. Additionally, individuals outside the sampling frame who voluntarily join contribute data through nonprobability-based sampling (NPBS), but nonprobability-based samples can introduce bias. This thesis explores statistical methodologies to correct selection biases in NPBS by leveraging PBS data within the context of the RURAL cohort study.

Following the approach suggested by Professor Wu and other researchers, we use two main methods in our study. First, a model-based approach, where we use "mass imputation" to predict values for the variables we're interested in within the NPBS data before calculating the average value across the population. Second, in the propensity score based methods, where we use two different methods to calculate the propensity score: pseudo maximum likelihood estimation in combination with logistic regression and pseudo-inclusion probability in combination with logistic regression or with XGBoost. After obtaining the propensity scores, we calculate the average value for the population using Hajek's average, a form of inverse probability weighing (IPW), and the double robust estimation. Overall, we compare seven different correction methods for adjusting the NPBS that will be investigated in a simulation study. The simulation study shows that using pseudo maximum likelihood estimation using the data from both NPBS and PBS, together with inverse probability weighting works best among all IPW methods.

Furthermore, the thesis also presents an analysis of the Alabama data from the RURAL study where additional complexities were addressed. These complexities refer to data specific issues that are present in RURAL, but are not addressed in the methodology to correct NPBS data. Various raking methods were used in the RURAL study to create a final set of weights for the full sample of Alabama (probability- and nonprobability-based) making use of data from the American Community Survey (ACS) as population benchmark. The findings indicate minimal differences among the raking methods tested, but the thesis justifies the choice of standardizing weights derived from propensity score data.

Keywords: Cohort study, Nonprobability sampling data, Combined Inversed probabilty weighting raking

Contents

Chapter 1

Introduction

1.1 Research Background

The Risk Underlying Rural Areas Longitudinal Study (RURAL) is a research project focusing on select rural counties in Alabama, Kentucky, Louisiana, and Mississippi. The primary goal of this cohort study is to uncover the causes behind the high burden of heart, lung, blood, and sleep (HLBS) disorders in specific rural areas in the Southeastern United States and explore potential alleviation strategies.

The data sampling procedure in this research involves two approaches. The first method employs a probability-based sampling approach, randomly selecting participants aged from 25 to 65 from all county household addresses with the assistance of a reliable vendor. Postcards and questionnaires are sent to the sampled addresses, and the research is followed up through intensive phone calls. Due to the low response rate in rural areas with the selected sampling approach, an alternative approach was implemented to increase sample sizes: gathering volunteers who wish to participate and are not included in the sampling frame (the addresses selected for the probability sample). The specific data descriptions are explained in Chapter 4.

However, data collected through the volunteer approach often introduces significant bias when estimating population parameters. Therefore, the thesis aims to explore various statistical methods that have been developed by many scientists and presented in an overview paper of sampling specialist prof. dr. Wu. These methods try to correct a nonprobability-based sample with an available probability-based sample. The main objective of this thesis is to implement some of these methods and study their performance in correcting selection bias in the nonprobability-based sample using simulation studies under different settings. Additionally, the thesis aims to evaluate how these methods perform when certain assumptions are violated. The best-performing methods were selected to be ap-

plied to the Alabama data, where additional complexities as present. Detailed discussions on these correction methods and their simulation results are presented in Chapters 2 and 3. And the case study is presented in Chapter 4. In the final chapter we provide a summary and a discussion of the results.

1.2 Sample Surveys

Sample surveys aim to acquire reliable estimates for descriptive parameters of finite populations. Traditionally, survey samplers have utilized probability-based sampling methods, often combined with census and administrative data, to facilitate the development of valid and efficient inferences regarding finite population parameters. Neyman's publication [20] lays a robust theoretical foundation for probability-based sampling approaches. Additionally, U.S. Census Bureau researchers Hansen and Hurwitz [11] contributed to the field by introducing methods for handling multiple probability-based sampling data. Their work focuses on the basic theory of stratified two-stage cluster sampling, where one cluster (or primary sampling unit) within each stratum is drawn with a probability proportional to population size (PPS).

In the early stages of survey sampling, surveys were comparatively more straightforward and simpler than today. Data collection primarily occurred through personal interviews or email questionnaires. The responsiveness of participants, known as the response rate, significantly influenced the quality of the sampling approach. The 1970s are often referred to as the "golden age of survey research" [26], characterized by generally high response rates.

Recent years have seen the development of various nonprobability-based sampling approaches [7]. Baker categorizes these approaches into three types: convenience sampling, sample matching, and network sampling. Convenience sampling, also known as volunteer sampling, is widely employed in medicine and market research. Volunteers may participate in a single study or become part of a panel recruited for different studies over time [25]. In sample matching, members of the nonprobability-based sample (NPBS) are selected to match important population characteristics [23] through estimating the propensity score, matching individuals who share similar propensity scores, and adjusting the weights to correct the selection bias. This is the main set of methods that will be implemented in the thesis. In network sampling, members of a target population are asked to identify others with whom they are somehow connected. Sirken [5] provides an early example of network or multiplicity sampling, where the network that respondents report about is clearly defined.

However, collecting data through nonprobability-based sampling can introduce bias, as specific segments of the population may be excluded. For instance, indi-

viduals without internet access or those unable to visit survey websites may not be represented when the modality of data collection is conducted through the internet. Bias is particularly common in medical research, where individuals with specific demographics may be more difficult to include in the sample due to health issues.

To mitigate selection bias, Valliant and Dever [27] proposed correcting the nonprobability-based sample using data from an independent probability-based sample available from the same target population. In this scenario, the probability-based sample has no information on the study outcome of interest but provides information on relevant independent variables that could help eliminate selection bias. In this thesis, we attempt to study this inference approach under various settings to assess its effectiveness. Detailed explanations of the settings and research approaches will be provided in Chapter 2

1.3 Complexities in Case Study

The ultimate plan for the RURAL analysis involves utilizing a hybrid approach to analyze the data, combining elements of both probability-based and nonprobability-based sampling methods to mitigate biases and improve the robustness of the analysis. Compared to the research that studies correction methods for selection bias in NPBS, this RURAL study introduces additional methodological complexities. For example, in the research on nonprobability-based sampling there is no information on the study outcomes of interest available in the probability-based sample (PBS) data. However, in the RURAL study, the PBS data includes information on all the information that is available for the NPBS (outcomes of interest, independent variables, other variables). Additionally, unlike the standard PBS data used by other researchers, the PBS data in the RURAL study may itself be biased due to non-response, which requires attention either before or after correcting the NPBS. Furthermore, missing data is a frequent issue in cohort studies, sometimes even affecting covariate selection. The methods used to impute the data or decisions on whether to impute the data can add complexity to the analysis. The research on NPBS does not discuss variable selection, it assumes we know which variables to use and assumes that they are available in the PBS, but variable selection is not straightforward when a large set of variables are present in the cohort study. Related to this issue, is the assumption that a single outcome of interest is also known in the research on NPBS, but cohort studies are typically collected to be able to study many outcomes of interest. Correcting the NPBS data for each variable in the study is nonpractical. Thus, we need a method that could adjust the NPBS data irrespective of the outcome of interest.

1.4 Outline

This thesis is divided into four parts. Chapter 2 provides a brief description of the basic setup and notations used throughout the thesis. It delivers an inferential framework based on the RURAL setting.

Chapter 3 contains a simulation study with various types of data under different settings to investigate performance differences among different inference approaches and estimators. Finally, the thesis selects the approach that yields the best results from the simulation and applies it to a real case using data from the RURAL analysis in Chapter 4. Here we also emphasize and address some of the complexities that are present in the RURAL study. Chapter 5 concludes with a discussion of the limitation of the current work and what could be studied in the future.

Chapter 2

Statistical inference methods for Nonprobability-based Samples

In this chapter, we present several statistical inference methods for nonprobability-based sampling data for which a probability-based sample with appropriate variables is available. The statistical inference methods can be divided into two categories: model-based prediction and propensity score-based methods (including a combination of the two methods). In addition to the statistical inference methods, there are several options to obtain prediction and propensity scores and there are different ways of estimating the parameter of interest. Before we introduce the statistical inference methods we will first describe the set-up and introduce some notation.

2.1 Setup and Notations

Let (x_i, y_i) be the measurement of the auxiliary variables $\mathbf{X}$ and the variables of interest $\mathbf{Y}$ associated with the unit i from the finite population $\mathrm{U} = \{1, 2, 3, \ldots, N\}$, where N is the population size. What we want to investigate is the parameter $\theta = N^{-1} \sum_{i=1}^{N} y_i$, which represents the population mean for the variable of interest. The total sampled participants can be divided into two parts: $\{(x_i, y_i), i \in S_A\}$ is the dataset for the NPBS S_A with n_A participating units, and $\{(x_i, y_i), i \in S_B\}$ is the dataset for the PBS S_B with n_B participating units. In most cases, the NPBS mean $\overline{y}_A = n_A^{-1} \sum_{i \in S_A} y_i$ is a biased estimator of θ, therefore it needs to be corrected.

The original setting described by Wu [29] and Chen [15] involves PBS data S_B with observations on auxiliary variables $\mathbf{X}$ only and NPBS data S_A with information on both auxiliary variables $\mathbf{X}$ and variables of interest $\mathbf{Y}$. Table 2.1 represent the setup for these two datasets.

Sample	Type	X	Y	Representativeness
A	Nonprobability-based sample	✓	✓	No
B	Probability-based sample	✓	x	Yes

Table 2.1: Data structure for two samples

2.1.1 Assumptions

Before embarking on the statistical correction methods and its practical implementation, it is essential to ensure the satisfaction of certain assumptions proposed by Chen [3] regarding the selection mechanism. Here we introduce R_i as the binary indicator for a participant to be part of the NPBS and the conditional probability $\pi_i^A = P(R_i = 1|(x_i, y_i))$ is then referred to as the propensity for being part of the NPBS.

- **The indicator R_i is independent of the outcome covariates y_i given the set of covariates $\mathtt{x}_i$.**
- **All estimated propensity scores are greater than zero: $\pi_i^A > 0, \quad i = 1, 2, \ldots, N$.**
- **The indicator variables R_i are independent of each other given the set of covariates $\mathtt{x}_i$.**

Under the first assumption, the selection mechanism is termed ignorable, a term that is often used in the context of missing data. Specifically, $P\left(R_i = 1 \mid x_i, y_i\right) = P\left(R_i = 1 \mid x_i\right), \quad i = 1, 2, \ldots, N$. This assumption aligns with the concept of missing at random (MAR) for missing data introduced by Rubin [24]. It implies that the set of covariates x_i encompasses all relationships between the selection mechanism and the outcome. The second assumption ensures the feasibility of subsequent computations, but it is unclear if this assumption is also realistic for studies like RURAL. The third assumption implies that participation of participants in the sample is independent of each other conditionally on the covariates.

Note that the indicator R_i is a mathematical formulation for individuals of the population to become part of the NPBS and is (in principle) unrelated to the probability of individuals being part of the PBS. The mechanisms behind such participation in NPBS may be completely different from the mechanism for PBS or there may be overlap. An indicator being equal to zero ($R_i = 0$) may in principle not say anything about participation in the PBS. However, in the RURAL study, the mechanisms are related, since volunteers are defined as participants who are not part of the framework of addresses for the PBS.

2.1.2 Choice of the covariates

In practical applications, meeting the assumptions of certain analytical approaches is often challenging. Our research shows that types of outcomes can vary greatly within a cohort study. Cohorts are designed to provide a wide range of variables, allowing researchers to focus on specific diseases, conditions, or symptoms. However, using many $\mathbf{X}$ variables for adjusting the NPBS poses a risk: these variables might later be examined as outcomes by other researchers. This conflicts potentially with the theoretical framework proposed by Wu and Chen and may undermine causality principles that have been formulated for analysis of cohort studies. Such theories suggest that we should not correct the data with variables that are being studied as outcome variables.

Therefore, it's crucial to carefully choose variables for adjustment of the NPBS. They should be relevant to the study's goals and they should be variables that are causally affecting the behavior of individuals volunteering for cohort studies. The selection process should enhance the analysis's validity without compromising the study's integrity. Balancing different types of variables is essential for robust and credible cohort research. The covariates may be divided into different subgroups, like participant attributes (e.g., sex, race), behavioral variables (e.g., smoking, alcohol), participant characteristics (e.g., age, education, BMI), and medical variables (e.g., blood pressure, diabetes). Here we decided to limit ourselves to variables in the categories attributes and characteristics to make sure we do not correct for variables that may become an outcome variable in a future research.

2.2 Model Based Prediction

One widely adopted model-based prediction inference approach is known as Mass imputation. Mass imputation was developed within the context of two-phase sampling by Breidt[19] and Rao [16]. Rivers [22] proposed the sample matching method but did not provide any theoretical proof. Chipperfield [4] discussed composite estimation when one of the surveys involves mass imputation. Subsequently, Chen [15] delivered a rigorous proof in survey sampling on mass imputation, analyzing nonprobability-based survey samples from a theoretical perspective.

In our context, the existing PBS data is assumed to lack measurements on the variables of interest $\mathbf{Y}$, treating it as if these variables are 100% missing. The NPBS data comprises observed values for both the study variable and auxiliary variables. This nonprobability-based sample serves as training data to construct a prediction model, which is then employed to impute missing values for the probability-based sample. This means that the method of mass imputation assumes that the associations in the NPBS data are representative for the population, but imbalances

in frequencies for certain variables are disturbed by the nonprobability or volunteering mechanism.

Several imputation methods can be applied, such as Nearest Neighbor Imputation suggested by Rivers [22], a non-parametric method that does not require any parametric model assumptions. However, it often faces challenges related to the dimensionality of the sampling data. In this thesis, our primary focus is on the semi-parametric model for S_A. The predicted value of the variables of interest can be formulated as:

$$\mathbb{E}(Y|X = x) = m(x, \beta) \tag{2.1}$$

The specification of $m(x, \beta)$ can be computed based on the generalized estimating equations(GEE). It can be applied to any form of regression function, where it includes all kinds of functions of $\mathbf{X}$ and interactions between $\mathbf{X}$s. For binary outcome variables of interest, the parameter β is integrated into the logistic model as $m(x, \beta) = \{1 + \exp(-x'\beta)\}^{-1}$. We assume that $\hat{\beta}$ is the unique solution to the generalized estimating equation:

$$\sum_{i \in S_A} \frac{\partial m\left(\mathbf{x}_i, \boldsymbol{\beta}\right)}{\partial \boldsymbol{\beta}} \left\{v\left(\mathbf{x}_i\right)\right\}^{-1} \left\{y_i - m\left(\mathbf{x}_i, \boldsymbol{\beta}\right)\right\} = \mathbf{0}, \tag{2.2}$$

where $v\left(\mathbf{x}_i\right)$ represents the known form of the variance function, that holds true only when we make an assumption on the distribution of $\mathbf{Y}$. Due to the fact that the GEE can work with a "working variance-covariance matrix", the exact variance term is not necessary in our case. For any general regression model $m(x, \beta)$, we have the predicted value $\hat{y}_i = m(x_i, \hat{\beta})$ for $i \in S_B$, with $\hat{\beta}$ being the estimated model parameters based on NPBS data $\{(x_i, y_i), i \in S_A\}$, The estimate for the population mean then becomes,

$$\hat{\mu}_{reg} = \frac{1}{N^B} \sum_{i \in \mathcal{S}_B} d_i^B m\left(\boldsymbol{x}_i, \hat{\boldsymbol{\beta}}_{reg}\right). \tag{2.3}$$

Here d_i^B is the sampling weight for participant i in the probability-based sample S_B, and N^B is the sum of the weights for the all the participants in S_B, which can be expressed by $N^B = \sum_{i \in S_B} d_i$. The estimator $\hat{\mu}_{reg}$ introduced by Schafer [14] is approximately unbiased under the joint framework of the prediction model and the probability sampling design for S_B .

2.3 Propensity scores: pseudo-likelihood

In the statistical analysis of observational data, the propensity score approach tries to correct imbalances in the observational data, such that the association of an

exposure with an outcome becomes unbiased and corrected for confounders[28]. In our work, we will make use of the propensity score approach to adjust or correct the NPBS using the data from the PBS. In this section we will use the pseudo-likelihood approach to estimate the propensity score $\pi_i^A = P(R_i = 1|x_i)$.

Thus, in order to correct the result of y_i for $i \in S_A$ with likelihood approaches, the propensity score must be modeled parametrically as $\pi_i^A = \pi(\boldsymbol{x}_i, \boldsymbol{\theta})$, where $R_i = 1$ if $i \in S_A$ and $R_i = 0$ if $i \notin S_A$, and with $\boldsymbol{\theta}$ representing the true value of the unknown propensity score which we assume will have a specific form. We assume that, the propensity score can be formulated using the logistic regression model as follows:

$$\pi_i^A = \frac{\exp(\boldsymbol{x}_i^T \boldsymbol{\theta})}{1 + \exp(\boldsymbol{x}_i^T \boldsymbol{\theta})} \tag{2.4}$$

We choose the logistic regression model to compute the propensity score because it is particularly well-suited for modeling binary outcomes and provides a probabilistic interpretation of the outcomes. It estimates the probability that a given outcome (here particiapting in the NPBS) will occur, which is intuitive and useful for understanding how likely individuals are part of the the NPBS using certain covariates.

2.3.1 Computing the propensity score by the reference probability sample S_B

Based on the setting described in Section 2.1.1, x_i is observed for all units in both S_A and S_B, while y_i is only observed for the S_A. The maximum likelihood estimator of $\boldsymbol{\theta}$ is computed as $\hat{\boldsymbol{\theta}}$, where $\hat{\boldsymbol{\theta}}$ maximizes the log-likelihood function over the full population[3].

$$\begin{aligned} l(\boldsymbol{\theta}) &= \sum_{i=1}^{N} \left\{ R_i \log \pi_i^A + (1 - R_i) \log\left(1 - \pi_i^A\right) \right\} \\ &= \sum_{i \in \mathcal{S}_A} \log \left\{ \frac{\pi(\boldsymbol{x}_i, \boldsymbol{\theta})}{1 - \pi(\boldsymbol{x}_i, \boldsymbol{\theta})} \right\} + \sum_{i=1}^{N} \log \left\{ 1 - \pi(\boldsymbol{x}_i, \boldsymbol{\theta}) \right\}. \end{aligned} \tag{2.5}$$

According to the log-likelihood function, we realize that the equation can only be applied if x_i is observed for all units in the population. However, for participants outside the NPBS, we only have data for the participants in the PBS. Therefore, instead of using $l(\boldsymbol{\theta})$, we replace the second term of the log-likelihood function with the reference probability sample S_B with information on x. Hence, the estimator can be obtained by the following pseudo log-likelihood function:

$$l^*(\boldsymbol{\theta}) = \sum_{i \in \mathcal{S}_A} \log \left\{ \frac{\pi(\boldsymbol{x}_i, \boldsymbol{\theta})}{1 - \pi(\boldsymbol{x}_i, \boldsymbol{\theta})} \right\} + \sum_{i \in \mathcal{S}_B} d_i^B \log \{1 - \pi(\boldsymbol{x}_i, \boldsymbol{\theta})\}. \tag{2.6}$$

with again, d_i^B represents the weight for participant i in the probability-based sample, where the weights add up to a number close or equal to the population size. The estimator can be obtained by solving the equation through the Newton-Raphson iterative procedure:

$$\frac{\partial}{\partial \boldsymbol{\theta}} l^*(\boldsymbol{\theta}) = \sum_{i \in \mathcal{S}_A} \boldsymbol{x}_i - \sum_{i \in \mathcal{S}_B} d_i^B \pi(\boldsymbol{x}_i, \boldsymbol{\theta}) \boldsymbol{x}_i = 0 \tag{2.7}$$

2.3.2 Computing the propensity score from pooled sample S_A and S_B

In addition to only using S_B as the reference sample for the population component in the log-likelihood function, Lee [17] and Isaksson [13] proposed that we can also use the data from both S_A and S_B and consider them as a whole. However, the samples must be weighted such that the sum of the weights together lead to a number close to the population size. This approach can only work when weights for the NPBS would be known. Valliant and Dever [18] considered the weighted logistic regression approach to estimate π_i^A. For each unit $i \in S_{AB} = S_A \cup S_B$ in the full sample, the pooled weight d_i is defined as follows: $d_i = 1$ if $i \in S_A$ is part of the nonprobability-based sample, and $d_i = d_i^B(1 - n_A/N^B)$ if $i \in S_B$ is part of the probability-based sample, with d_i^B the weight of participant i in the probability-based sample. The motivation behind this is to make the weights for the pooled sample S_{AB} equal to the population size that was determined for the probability-based sample S_B, which would be equal to or close to the population size. Indeed, adding up d_i over all participants in both samples lead to N^B. Subsequently, the authors incorporated this into the log-likelihood function, which then becomes:

$$l^{**}(\boldsymbol{\theta} = l^*(\boldsymbol{\theta}) + \sum_{i \in \mathcal{S}_A} \log\left(1 - \pi_i^A\right) - \frac{n_A}{N^B} \sum_{i \in \mathcal{S}_B} d_i^B \log\left(1 - \pi_i^A\right), \tag{2.8}$$

where $l^*(\boldsymbol{\theta})$ is the pseudo log-likelihood function mentioned before. We can obtain the pseudo maximum likelihood estimator by solving the following likelihood equation:

$$\frac{\partial}{\partial \boldsymbol{\theta}} l^{**}(\boldsymbol{\theta}) = \sum_{i \in \mathcal{S}_A} \{1 - \pi(\boldsymbol{x}_i, \boldsymbol{\theta})\} \boldsymbol{x}_i - \left(1 - \frac{n_A}{N^B}\right) \sum_{i \in \mathcal{S}_B} d_i^B \pi(\boldsymbol{x}_i, \boldsymbol{\theta} x_i) x_i = 0. \tag{2.9}$$

Codes for pseudo maximum likelihood estimation

The equations 2.7 and 2.9 are computed by the function 'pseudo_likelihood', which was written in Python. After defining the function, the initial values are set to zero before being solved using the 'fsolve' function. 'fsolve' is a function from the SciPy library that finds the roots of a given function. In this context, it attempts to find the parameter values that make the 'pseudo_likelihood' function equal to zero. The results are then returned as computed propensity scores, which are probabilities between 0 and 1 for each observation. The code is attached in the appendix.

2.4 Propensity scores: pseudo-inclusion probability

In 2017, Michael Elliott and Richard Valliant [8] proposed an inference approach for nonprobability-based samples called pseudo-inclusion probability to correct selection bias of the NPBS. Considering the framework described earlier, the weight they want to estimate for participants in the NPBS is $w_i = 1/P(R_i = 1|x_i)$. They made a derivation where this weight can be written in terms that can be determined from the data. This weight is approximately $w_i = \tilde{w}_i \times P(Z_i = 0|x_i)/P(Z_i = 1|x_i)$, where $\tilde{w}_i$ is the inverse of the selection probability for participant i in the nonprobability-based sample being part of the probability-based sample and Z_i is the binary indicator for NPBS and PBS (where $Z_i = 1$ means that participant is part of the nonprobability-based sample and $Z_i = 0$ means that the participant is part of the probability-based sample). The probability $P(Z_i = 1|x_i)$ can be estimated with logistic regression or machine learning methods using the data from the NPBS and PBS, and $\tilde{w}_i$ can be chosen equal to the weight of a matching participant from the probability-based sample.

2.4.1 Logistic regression

As we already discussed in previous section, logistic regression models the probability for a binary outcome as function of covariates using the logit function approach. In case logistic regression is applied to modeling the participation of NPBS with respect to PBS, i.e., modeling $P(Z_i = 1|x_i)$, the logit function is linear in the coefficients of the covariates x_i. Thus,

$$\text{logit}(P(Z_i = 1|x_i)) = \log(P(Z_i = 1|x_i)) - \log(P(Z_i = 0|x_i)) = x_i'\beta \qquad (2.10)$$

Codes for logistic regression

The parameters in the logit function can be estimated with maximum likelihood which is implemented in Python. Here we used the 'statsmodels' package to fit the logistic regression model for modeling $P(Z_i = 1|x_i)$. Codes can be found in the appendix. After the parameters are estimated, the probability $P(Z_i = 1|x_i)$ can be predicted for each participant in NPBS and used in the pseudo-inclusion probability of Elliott and Valliant [8].

2.4.2 Machine learning method : XGboost

Alternatively to logistic regression, the probability $P(Z_i = 1|x_i)$ can be estimated with machine learning approaches. The advantage is that the machine learning algorithm can more easily incorporate non-linearities and higher-order interactions between variables than logistic regression. Therefore, the probability of being part of the nonprobability-based sample compared to the probability-based sample may be more realistic with machine learning than with regression. Here we will study the machine learning technique XGBoost.

Codes for XGboost

XGBoost is a strong machine learning tool that uses decision trees to understand how different variables interact and automatically handles numerical data. It finds the best ways to split the data and predicts or classifies the variable of interest. Additionally, it deals with missing values effectively. To compute the probability $P(Z_i = 1 \mid x_i)$ using XGBoost for NPBS and PBS data, we use the `XGBClassifier` package in Python. Codes can be found in the appendix.

First, add an indicator column to distinguish between the two groups, then combine them into a single dataset. Define a set of features and the target variable (the indicator), then initialize and fit an XGBoost classifier on this combined dataset. After training, the model predicts the probability $P(Z_i = 1|x_i)$, which can then be used in the pseudo-inclusion probability approach of Elliott and Valliant [8]. Based on the setting from the Ridgeway [21], the parameters used for XGBoost are:

- Number of trees: 5000
- Interaction depth: 4
- Shrinkage: 0.01
- Stop method: es.max

2.5 Estimation of the population mean

Once the propensity score or inverse propensity score of each participant from the NPBS is obtained, they can be utilized to provide an estimator for the population mean by weighting all outcome observations of the NPBS. Additionally, they can be combined with the mass imputation approach.

2.5.1 Inverse probability weighting

Inverse Probability Weighting (IPW) assigns weights to observations based on the inverse of the propensity score for each individual, which represents the probability of being part of the NPBS. The HT estimator was originally proposed by Horvitz and Thompson [12] for a finite population with a probability survey.

$$\hat{\mu}_{HT} = \frac{1}{N} \sum_{i \in \mathcal{S}_A} \frac{y_i}{\pi_i(\hat{\boldsymbol{\theta}})} \tag{2.11}$$

Since the weights of the NPBS may not add to exactly the population size N, another version of the estimator can also be formulated based on the Hájek estimator [10]:

$$\hat{\mu}_{H} = \frac{1}{N^A} \sum_{i \in \mathcal{S}_A} \frac{y_i}{\pi_i(\hat{\boldsymbol{\theta}})} \tag{2.12}$$

where $N^A = \sum_{i \in \mathcal{S}_A} 1/\pi_i(\hat{\boldsymbol{\theta}})$. There are several examples from both the theoretical part and practical observation showing that the Hájek estimator outperforms the HT estimator [3]. Therefore, we would like to adopt the Hájek estimator to estimate the finite population mean in our research.

2.5.2 Doubly Robust

We have previously discussed model-based prediction and propensity score-based estimation in the preceding sections. However, when applied individually to estimate the population mean, both the outcome regression prediction method and the propensity score approach are unbiased only if the statistical model is correctly specified [9]. The doubly robust estimator combines these two approaches such that only one of the two models needs to be correctly specified to obtain an unbiased effect estimator.

Therefore, we aim to combine a propensity score estimator with the prediction result from mass imputation as our doubly robust estimator. The estimator for the doubly robust approach that incorporates the propensity score is:

$$\hat{\mu}_{DR}^{PS} = \frac{1}{N^A} \sum_{i \in \mathcal{S}_A} d_i^A \left\{ y_i - m_i(\hat{\boldsymbol{\beta}}) \right\} + \frac{1}{N^B} \sum_{i \in \mathcal{S}_B} d_i^B m_i(\hat{\boldsymbol{\beta}}), \tag{2.13}$$

The second term on the right-hand side of the equation represents the model-based prediction obtained through mass imputation in our study. The first term adjusts the prediction estimator using propensity scores and the errors $\varepsilon = y_i - m_i(\hat{\boldsymbol{\beta}})$ derived from the outcome regression model. Wu [29] demonstrated that the magnitude of this adjustment term is inversely related to the "goodness-of-fit" of the outcome regression model.

2.6 Summarizing the correction methods

In the previous sections we have written about approaches that could be used. Here, we summarize which methods will be used in a comparison in the simulation study:

- **Mass imputation**
  The approach is described in Section 2.2 using the formula 2.3.
- **Propensity scores: pseudo-likelihood**
  - Logistic function and pseudo-likelihood using reference S_B
  - Logistic function and pseudo-likelihood using pooled S_AB

  The approach for the propensity score is provided in formula 2.7 and 2.9 and the population mean is from the NPBS using Hajek's IPW estimator in formula 2.12.
- **Propensity scores: Pseudo-inclusion probability**
  - Logistic regression for NPBS versus PBS
  - XGBoost for NPBS versus PBS

  The approach is described in the main text in section 2.4. Here we use the logistic regression and XGBoost for predicting participation in the two samples NPBS and PBS and borrow the weights from PBS to obtain an estimate of the propensity score. Using Hajek's IPW estimator in formula 2.12, the population mean is estimated from the NPBS data.
- **Doubly robust estimator**
  - Mass imputation with IPW using S_B

– Mass imputation with IPW using S_AB

The approach is described in section 2.5.2. Here we only include the pseudo-likelihood approaches for the propensity score, since literature has only mentioned this combination (even though we could have applied it with the IPW from the pseudo-inclusion approach).

Chapter 3

Simulation Studies

In this chapter, we present several simulations to compare the performance of seven different correction methods. Our approach is to generate a full population from which we collect a probability-based sample (PBS) and nonprobability-based sample (NPBS). We will choose different populations with association structures and different choices of creating the NPBS. Here we evaluate the bias and mean squared error of the NPBS estimate to the PBS estimate and the population mean.

3.1 Simulating the full population

To generate the finite populations, we consider different populations generated by different regression functions that establish the relationship between several independent covariates x and a response y:

1. $y_{1i} = 0.5x_{1i} + 0.4x_{2i} + 0.3x_{3i} + 0.2x_{4i} + 0.1x_{5i} + \varepsilon_{1i} \quad i = 1, 2, \ldots N$
2. $y_{2i} = 0.5x_{1i} + 0.4x_{2i} + 0.3x_{3i} + 0.2x_{4i} + 0.1x_{5i} + \varepsilon_{2i} \quad i = 1, 2, \ldots N$
3. $y_{3i} = 0.5x_{1i} + 0.4x_{2i} + 0.3x_{3i} + 0.2x_{4i}^2 + 0.1x_{5i} + \varepsilon_{1i} \quad i = 1, 2, \ldots N$
4. $y_{4i} = 0.5x_{1i} + 0.4x_{2i} + 0.3x_{3i} + 0.2x_{4i}^2 + 0.1x_{5i} + \varepsilon_{2i} \quad i = 1, 2, \ldots N$

The population size was selected equal to $N = 100000$, which may represent the size of a rural county in the South of the United States. We assume that the residuals are normally distributed, with $\varepsilon_1 \sim \mathcal{N}(2, 2)$, and $\varepsilon_2 \sim \mathcal{N}(100, 2)$. The first three covariates x_1, x_2, and x_3 are considered Bernoulli distributed with probability $p = 0.5$. They may represent sex (men or women), race (black or white), and education (low or high). The fourth covariate x_4 is considered continuous and uniform distributed on the interval $[20, 60]$ and may represent age. These covariates are generated using a uniform copula with correlation matrix given in Table 3.1.

Variable	x_1	x_2	x_3	x_4
x_1	1			
x_2	0.12	1		
x_3	0.03	0.21	1	
x_4	-0.04	-0.03	0.05	1

Table 3.1: Correlation matrix for dependent $\mathbf{X}$

These correlations mimic the correlation of the variables sex (x_1), race (x_2), education (x_3), and age (x_4) in the RURAL study. A binary variable with a Bernoulli distribution ($p = 0.6$) is included, potentially reflecting the structure of the RURAL study, which involves two participating counties for Alabama. The variable x_5 is assumed to be independent of the other covariates, though in practice, the population values of these covariates may vary.

Under these conditions, four distinct types of response covariates y are generated. For population (1), the linear regression function exhibits a high R^2 value ($R^2 \approx 0.28$), indicating a strong relationship between the covariates and the response. In population (2), while the relationship between the covariates and the response is similar to that in population (1), the R^2 is notably lower ($R^2 \approx 0.09$). In populations (3) and (4), we again observe high R^2 values ($R^2 \approx 0.29$) and low R^2 values ($R^2 \approx 0.11$), respectively, but with the additional complexity of a quadratic term for the continuous variable in the association structure.

3.2 Simulating the samples

In order to obtain the sampled data, existing of PBS and NPBS. The probability-based sample is a simple random sample and the nonprobability-based sample is highly selective. The sampling approach is depicted in Figure 3.1.

The first step is to generate one large population of individuals using the procedure for the covariates x_1, x_2, x_3, x_4 and the response y. Then we simulate x_5 to split the population into two pieces (county 1 and county 2). Thus, we make the assumption that associations between variables are not changed with county. Then, each county is split into two pieces. One piece, called PBS, represents the sampling frame of addresses from which we invite individuals to participate and the other piece, called NPBS, are the remaining individuals who will not be invited to the PBS (but may become part of the study as volunteer). The split into PBS and NPBS is 40% and 60% for county 1 and 80% and 20% for county 2. Then, the two pieces PBS from the counties are combined and a random sample of 800 individuals is collected as the PBS. The two NPBS parts from the counties are combined too and 200 individuals are collected using the following structure.

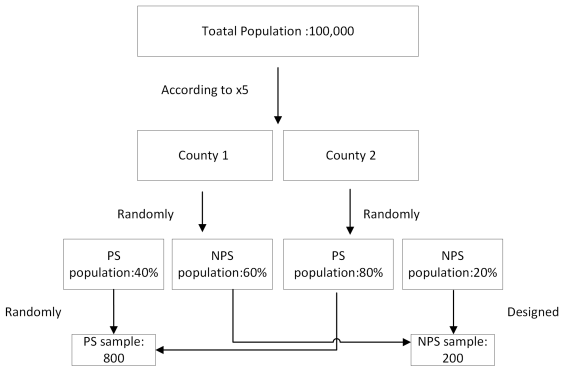


Figure 3.1: Sampling procedure

- 80 individuals from group $(x_1 = 1, x_2 = 1, x_3 = 0, x_4 > 40)$
- 60 individuals from group $(x_1 = 0, x_2 = 0, x_3 = 1, x_4 > 40)$
- 40 individuals from group $(x_1 = 1, x_2 = 0, x_3 = 0, x_4 < 40)$
- 20 individuals from group $(x_1 = 0, x_2 = 1, x_3 = 1, x_4 < 40)$

The individuals of the subgroups are collected randomly. In this scenario, we not only ensure that each x_i is selected with unequal probability but also guarantee that individuals from the PBS data are more likely to come from county 2, while individuals from the NPBS dataset are more likely to come from county 1.

This sampling approach has consequences for the methods, since the weights d_i^B for the PBS, which is used in the pseudo likelihood approach, and the weight $\tilde{w}_i$, used for the pseudo-inclusion probability, both become equal to $N/n_B = 125$.

3.2.1 Performance measures

To evaluate the performance of the correction method we make two comparisons. Firstly, we compare the difference in estimates between PBS and NPBS. We will report the mean difference in these estimates relative to the estimate of the PBS for each of the four populations and also quantify the root mean squared error relative to the PBS estimate. Secondly, we will compare the NPBS estimate with the population mean, by calculating the mean relative bias with respect to the mean population and the root mean squared error relative to the population mean. The mean relative bias is calculated as follows:

$$MRB(\%) = \frac{1}{m} \sum_{k=1}^{m} \left| \frac{\hat{\mu}_k^{\text{NPBS}} - \mu_k^{\text{B}}}{\mu_k^{\text{B}}} \right| \times 100 \tag{3.1}$$

The relative root mean squared error (RRMSE) is defined as

$$RRMSE = \sqrt{\frac{1}{m}\sum_{k=1}^{m}\left(\frac{\hat{\mu}_k^{\text{NPBS}}}{\mu_k^{\text{B}}} - 1\right)^2} \tag{3.2}$$

where the estimated value $\hat{\mu}_k^{\text{NPBS}}$ is the NPBS estimate for the population in simulation run k. The benchmark value, $\hat{\mu}_k^{B}$ is either the estimate for the population mean from the PBS in simulation run k or the population mean in simulation run k itself. To study these measures of performance, we conduct simulations to compare the 7 methods. We will use 500 runs for each population and applied all 7 correction methods for each simulation run.

3.3 Result

In this section, we begin by implementing descriptive statistics for the full population. Next, we demonstrate the differences in distributions between the PBS and NPBS datasets, taking into account both the x covariates and the outcome y under the sampling procedure. Following this, we assess whether including interactions improves the model's performance. Finally, we present the simulation results for different methods, evaluating their performance.

3.3.1 Result of data generation

Table 3.2a shows the population means of 5 covariates (x_1 to x_5) from the whole simulation (in total 500 simulation runs of 100000 participants). The values of the binary covariates are distributed evenly, with x_1, x_2, and x_3 having approximately equal proportions, as we initiated in the simulation. The indicator x_5 is equal to the setting 0.6. The value of x_4 is 40, which matches the median value of the age range of the population. Table 3.2 shows the mean values of four different population groups (y_1 to y_4) from the whole population, with values ranging from 10.5 to 506.8.

Regarding the means of x covariates in the sample, there are notable differences between the PBS dataset and the NPBS dataset, as illustrated in Table 3.3. Following the sampling structure, we observe that the mean values of x_1, x_2, x_3, and x_4 from the PBS data are close to the population means, as these 4 covariates were equally selected from the PBS. In contrast, the NPBS dataset shows a significant imbalance in the distribution of x covariates. For x_5, which indicates the two counties, the ratio of $x_5 = 1$ to $x_5 = 0$ in the PBS is approximately $3 : 4$. However, in the NPBS dataset, this ratio is about $9 : 2$, revealing that the PBS dataset

	x_1	x_2	x_3	x_4	x_5
Population	49%	50%	50%	39.7	60%

(a) Distribution of Covariates x from the whole population

	y_1	y_2	y_3	y_4
Population	10.5	108.5	408.9	506.8

(b) Mean value of y from the whole population

Table 3.2: Descriptive analysis to the whole population

	x_1	x_2	x_3	x_4	x_5
PBS	49%	50%	50%	39.7	43%
NPBS	60%	50%	40%	44	82%

Table 3.3: Means of Covariates in PBS and NPBS from the sample

predominantly includes individuals from county 2, whereas the NPBS dataset predominantly includes individuals from the first county. Note that the framework of addresses (40% vs 60% in county 1 and 80% and 20% in county 2) changes the proportion of x_5 from 0.6 to approximately 0.429.

3.3.2 Covariates Selection

For medical or epidemiological applications, it is often common practice to keep the (logistic) regression models relatively simple in terms of functions of the covariates. This means that researchers would typically include the mean effects and possibly the two-way interaction effects. Although there is no technical limitation to increase the complexity of the (logistic) function by including higher-order

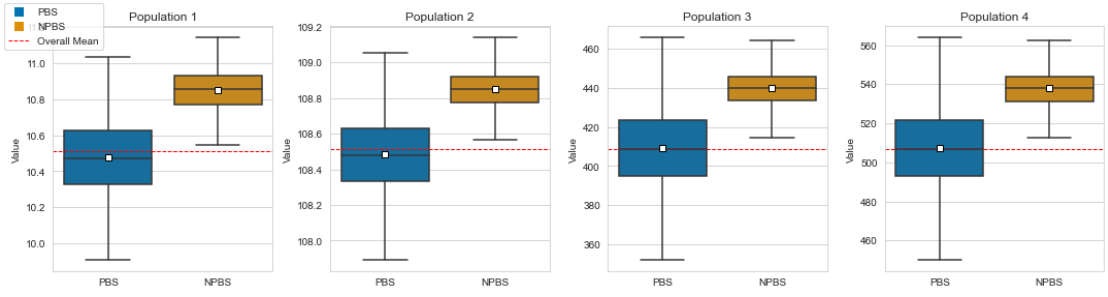


Figure 3.2: Box plots of the average values for the outcome variable over 500 simulation runs.

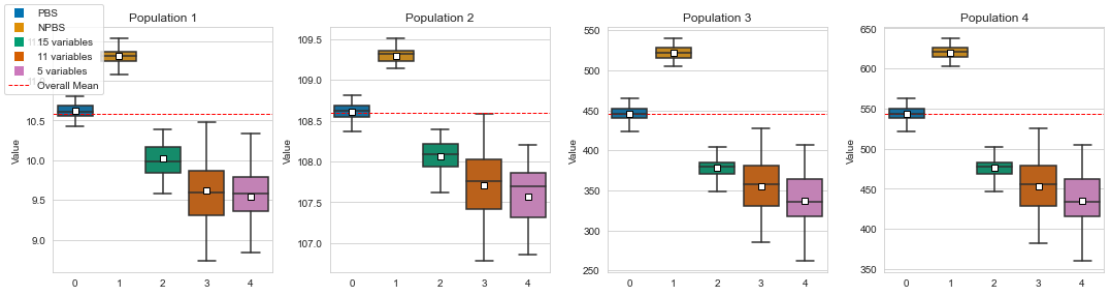


Figure 3.3: Result comparison with respect to the number of the covariates

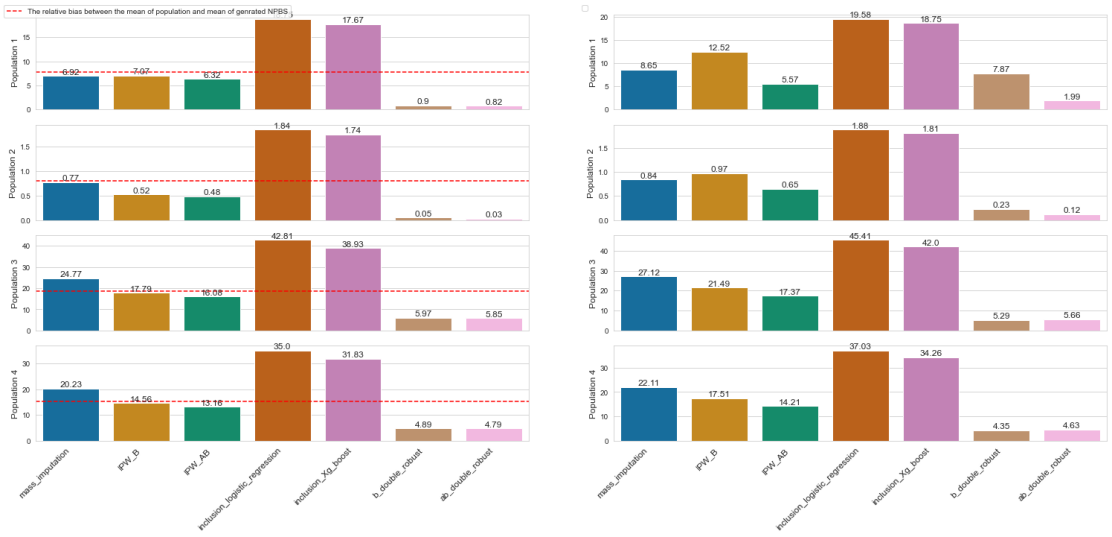
interactions, the interpretation of how strong certain variables are related (to an outcome variable for mass imputation or to the probability of being part of the NPBS) becomes more complicated. For the mass imputation we kept the prediction simple and used only the main effects, but for the propensity score we first investigate three different functional forms in the logistic regression model in combination with the pseudo-likelihood approach using S_B only (i.e., IPW_B). Here we kept relatively simple functional forms in line with medical practice, even though more complicated functional forms may be needed to describe the NPBS mechanism for participation. Note that this issue does not appear with XGBoost, since this method would be able to use higher-order interactions than just two, but it may have the disadvantage that for every simulation (or data set) a new model selection must be applied. We implemented the following three model for the logistic regression part:

- **15 variables**
  'x1','x2','x3','x4','x5'
  'x4_x1','x4_x2','x4_x3','x4_x5','x2_x1','x2_x3','x2_x5','x3_x1','x3_x5','x1_x5'

- **11 variables**
  'x1','x2','x3','x4','x5'
  'x4_x1','x4_x2','x4_x3','x4_x5','x2_x1','x2_x3','x2_x5','x3_x5'

- **5 variables**
  'x1','x2','x3','x4','x5'

The results of IPW_B are depicted in Figure 3.3. The red dotted line represents the mean value from the whole population (calculated from 5000000 (=500x100000) observations). It is evident that as the complexity of the functional form in the logistic regression analysis increases, the accuracy of the correction approach IPW_B improves. Based on this observation, we recommend that once the decision is made



(a) Relative bias comparison with respect to seven approaches

(b) RMSE comparison with respect to seven approaches

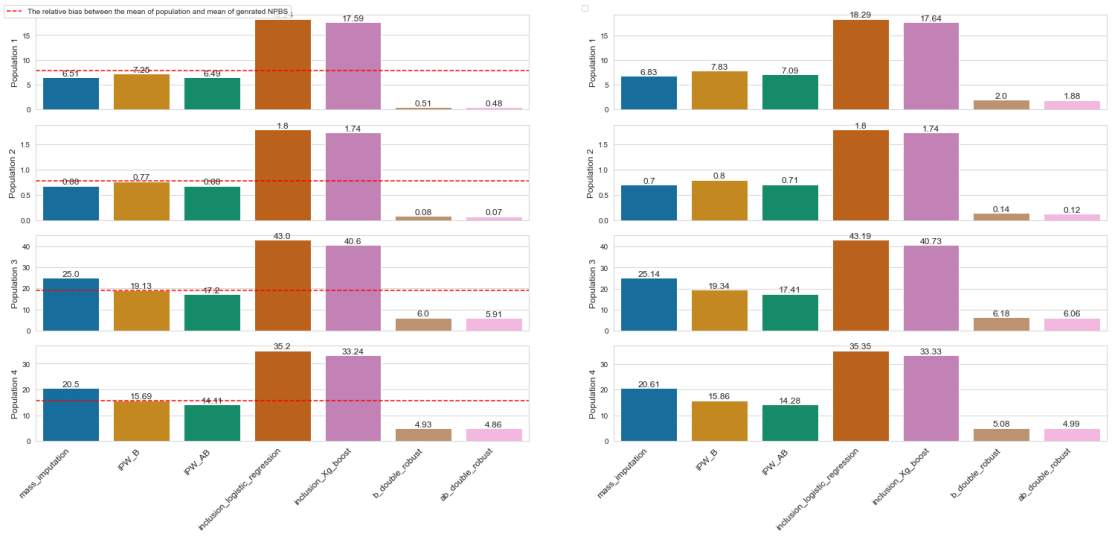
Figure 3.4: Comparison of relative bias and RMSE between the NPBS estimates and the population mean

regarding the inclusion of specific covariates, it is advisable to incorporate all possible interactions between these variables for a more accurate estimation.

3.3.3 Simulations and Method Comparison

The results of the relative bias and root mean squared error of the seven correction methods with respect to the population means and the PBS estimates for the four populations are presented in Figures 3.4 and 3.5, respectively. The red dotted line in the figures represents the mean difference between the original NPBS data (before correction) and the population, relative to the population mean. Thus, if the height of the bar in the figure is below the dotted line, the correction method has removed some or all of the bias that was present in the NPBS, but when the height of the bar is above the dotted line, the correction method made it worse..

We observe that there is no real difference between in relative biases and RMSE's between the two benachmark values (PBS estimates and population means). However, the biases and RRMSE's are either very close to each other when the two benchmark data sets are compared or the biases and RRMSE's are somewhat larger for the sample data. This implies that the imperfection of the correction methods (i.e., the existing biases) for estimating the population parameter is primarily due to the correction methods. Thus, the sampling variation for PBS itself



(a) Relative bias comparison with respect to seven approaches

(b) RMSE comparison with respect to seven approaches

Figure 3.5: Comparison of relative bias and RMSE between the NPBS estimates and the mean PBS values

does not contribute significantly to the observed biases.

We find that mass imputation has a positive effect on the bias (it makes it smaller compared to the bias of the uncorrected NPBS) for Population (1) and Population (2), which were both generated with a linear regression function containing linear functions of the covariates. However, the performance of mass imputation deteriorates for Population (3) and Population (4), which were both derived with a linear regression function containing quadratic terms of the covariates. Since the prediction model in the mass imputation has used only linear terms of the covariates, the results suggest that we should find the correct prediction model to be able to reduce bias in the NPBS with mass imputation.

Comparing the results from the propensity scores (using pseudo-likelihood estimation and pseudo-inclusion probability), we see that the pseudo-likelihood approach that uses both NPBS and PBS data simultaneously to determine the likelihood part of the population, outperforms the others. The pseudo-likelihood approach that uses only the PBS data for the likelihood part of the population is only slightly worse. The two pseudo-likelihood approaches improves the bias for all four populations, although some bias remains. The propensity score approach based on the pseudo-inclusion probability, does not perform well. The relative bias of the original sample is smaller than the relative bias of the pseudo-inclusion probability approach. We do not have a clear reason for this.

The logistic regression approach is worse in bias than the XGBoost when used in the pseudo-inclusion approach, possibly because the XGBoost would have a better estimate of the classification between the PBS and NPBS than the logistic regression. Suggesting again that it is important to find the correct predictions for the probabilities that are being estimated. However, the RRMSE of the XGBoost shows a slighty larger value than the logistic regression. This may be caused by the fact that we have to search for the best possible prediction model in each simulation run separately. Thus, the increase in complexity of predicting the participation in NPBS and PBS with XGBoost does improve the bias compared to logistic regression, which estimates the same function for each simulation, but introduce more variability.

Doubly robust methods, which combine mass imputation and IPW, produce results that provide the smallest bias and RRMSE. The bias is in most cases close to zero, even for the populations (3) and (4). The doubly robust method leverage the advantages of both the model-based approaches and the propensity score-based approaches to eliminate bias. However, given that mass imputation is a method that must be run for each outcome variable separately, it is less practical for researchers who would like to make use of the cohort study. The Doubly robust approach does not provide a set of weights that can be made available to the research community. Therefore, we still recommend the use of the pseudo-likelihood approach which implements the data from both the NPBS and PBS data simultaneously. Therefore, mass imputation and doubly robust methods are not further considered for the RURAL study.

Thus overall, the pseudo-maximum likelihood estimation combined with different propensity score estimation methods shows promise. Specifically, the IPW_AB method tends to produce better results across various populations by effectively balancing bias and variance, offering a more practical and effective solution for bias correction and variance reduction in survey data analysis. The results also shows that it is important to use the best possible model for estimating the probabilities. In case better and more complex fits to the probabilities are possible, we could potentially improve on the pseudo-likelihood approach with logistic regression, when we would allow more complex functions of the covariates, even though this may increase the variability.

3.3.4 Conclusion

Based on the results, we found that the doubly robust method with the IPW_AB approach consistently delivers the best results across all four populations. However, this choice may not be the most practical approach for the RURAL study, we therefore will implement the IPW_ab model on the real data, including all two-way interactions, for our case study.

Chapter 4

Case Study

In this chapter, we study the IPW_AB method on the RURAL study dataset. The doubly robust methods performed somewhat better than IPW_AB, but these methods are less practical, since they do not provide a set of weights that can be used for any analysis of the RURAL data. The doubly robust method requires for every outcome variable a new implementation. We will begin by providing a brief overview of the information collection procedure used in the RURAL study and describing the dataset utilized for our research. Subsequently, we will conduct descriptive analyses of the data, including assessments of missingness, mean, and median values. We will also examine the differences between the probability based sample (PBS) data and nonprobability based sample (NPBS) data concerning specific covariates. Based on the results of our descriptive statistics, we will justify the selection of the variables of interest before incorporating the data into the model.

To ensure the weights obtained for both PBS and NPBS data are directly usable, it is necessary to normalize them such that they sum up to the sample size. Therefore, we will implement several different raking approaches to align the weights with the American Community Survey Data, which serves as our benchmark. We will then compare the results of these different raking approaches in terms of the mean population value.

4.1 Background

In the RURAL study, data are collected through two different methods. Initially, the planned approach for data collection involved probability-based sampling, where the target number of participants is approximately 3.5% of the people per county. The sampled data were obtained by using the registered addresses in the county. Specialized companies or vendors have a full list of addresses that we

could use to sample from. From all these addresses a sample of addresses was provided taking into account the response rates of participants. Based on an first draw of 1000 addresses in Dallas County, Alabama, the response rate was determined at approximately 4.2% and the exhibited demographic imbalances, including a higher proportion of women and older individuals. It should be noted that this sampling was conducted during the COVID period.

To address the biases inherent in the PBS, the sampling specialists of the RURAL study designed weights for the PBS data to correct for these biases. Subsequently, to overcome the low response rate and to achieve the sample size targets, the researchers of RURAL started to support and accept volunteers into the RURAL study. Initially, the goal was to keep the volunteers rate low, but now it became a specific element of the sampling approach. This of course, required a sound statistical approach to make the two samples suitable for data analysis, such that researchers would be able to use the data without having to deal with these hybrid sampling approach.

Therefore, the ultimate plan for the RURAL study involved utilizing a hybrid approach to analyze the data, combining elements of both PBS and NPBS methods to mitigate biases and improve the robustness of the analysis.

4.2 Data Description & Pre-processing

The RURAL study contains many variables, but here we includes 26 variables in total, some of which are numerical, and others are categorical (binary) variables. The data was collected from two counties in Alabama: Dallas County and Wlicox County. To simplify further research, we converted some categorical variables. For example, for education, we set individuals who received a college degree as 1 and otherwise as 0. For people whose income is at least $30,000 per year, we set it as 1, otherwise as 0. Moreover, for some numerical variables, we implemented a log transform to ensure the normality of the data.

4.2.1 Descriptive analysis of the data

We compare the PBS with NPBS data in terms of the missingness of numerical characteristics and categorical characteristics.

Missingness

The difference between the PBS data and NPBS data in terms of missingness can be found in Table 4.1 which represents the missingness from PBS & NPBS data. Logistic regression analysis and the likelihood ratio test was implemented to

compare the difference in proportions of missingness. The results show that there are no real differences in proportion missingness for 22 variables (qualitatively and from the p-values). Moreover, it implicitly shows that there are no missing values in the variables sex, age, race, and county, since they were not part of the tables.

Numerical characteristics

We implemented descriptive statistics for each numerical variable separately, which include the median [quartile 1; quartile 3]. We also implemented likelihood ratio test with linear regression on either the original or log-transformed variable (which ever gives a better fit with normality) to investigate the difference between the two samples. The result are shown in Table 4.2. The variables that were transformed to the logarithmic scale are indicated with a an asterisk. The results show that there is a statistical difference in means for the variables FEV1 and FVC..

Categorical characteristics

For the categorical variables (incluidng the binary ones), we reported the percentages and numbers. Moreover, we used again the likelihood ratio test with logistic regression to investigate possible differences in proportions between the two samples. The result are shown in Table 4.3. The results clearly show that there is a difference in the proportion of black individuals, higher education, and a history of heart attacks between the two samples. In the volunteer group, there is a higher proportion of black participants and higher education, but a lower proportion of participants with a history of heart attacks.

4.3 Inference framework

The section on descriptives showed clearly that there are differences in variables between the two samples, thus it would make sense to correct the nonprobability-based sample with the data of the probability-based sample. However, the scenario in our case differs from the one studied in chapters 2 and 3 (as we already indicated in section 1.3). In our comparison study, the existing PBS data is assumed to have no measurements for the variables of interest$\mathbf{Y}$, effectively treating these variables as if they are 100% missing. Since all variables in the NPBS are also present in the PBS, the probability-based sample can be used to help correct the nonprobability-based sample, but also help obtain the best possible estimate for the population using both samples. Thus, we also need to come up with a procedure to combine the two samples.

Characteristic	Sampling (n = 746)	Volunteers (n = 196)	P-value
BMI	1 (0.117%)	0 (0.002%)	0.494
Weight	1 (0.101%)	0 (0.001%)	0.494
DBP	2 (0.309%)	1 (0.511%)	0.614
SBP	2 (0.316%)	1 (0.512%)	0.614
Cholesterol	43 (5.84%)	13 (6.62%)	0.651
HDL	43 (5.84%)	13 (6.63%)	0.651
LDL	47 (6.31%)	14 (7.17%)	0.673
Triglycerides	44 (5.91%)	14 (7.11%)	0.526
Fasting BGL	193 (25.9%)	50 (25.5%)	0.910
Creatinine	45 (6.01%)	16 (8.21%)	0.294
ABI	7 (0.912%)	3 (1.53%)	0.491

(a) Numerical characteristics of PBS data & NBPS data

Characteristic	Sampling (n = 746)	Volunteers (n = 196)	P-value
CAC	120 (16.1%)	33 (16.8%)	0.801
FEV1	164 (22.0%)	43 (21.9%)	0.989
FVC	164 (22.0%)	43 (21.9%)	0.989
FEV1/FVC	164 (22.0%)	43 (21.9%)	0.989
Education	2 (0.312%)	0 (0.002%)	0.334
Income	26 (3.52%)	5 (2.61%)	0.502
Smoking	32 (4.31%)	10 (5.14%)	0.629
Asthma	1 (0.102%)	1 (0.54%)	0.361
Diabetes	41 (5.51%)	16 (8.20%)	0.178
Hypertension	1 (0.187%)	1 (0.543%)	0.361
Heart attack/stroke	10 (5.51%)	5 (2.64%)	0.256

(b) Binary characteristics of PBS data & NPBS data

Table 4.1: Comparing probability sample with volunteers in terms of Missingness

Characteristic	Overall (n = 942)	Sampling (n = 746)	Volunteers (n = 196)	P-values
Age*	50.5 [40.4; 58.1]	50.5 [40.7; 58]	50.7 [39.2; 58.5]	0.696
BMI*	34.7 [29.3; 41.2]	35.3 [29.6; 41.3]	33.7 [28.3; 40.8]	0.086
Weight*	96.9 [81.2; 114.3]	97.2 [82.2; 115.6]	93.4 [78.2; 111.2]	0.019
DBP*	82.0 [74.5; 90.5]	82.5 [74.8; 91]	81.5 [73.5; 87.5]	0.198
SBP*	128 [116.5;142]	128[116.5;142.5]	127 [117;138.5]	0.238
Cholesterol*	180 [156.0;210]	180 [156;209]	183 [155;214]	0.278
HDL*	50.1 [41; 61]	49.0 [40; 61]	52.1 [42; 63]	0.119
LDL*	106 [83;132]	107 [83; 132]	105 [77.0; 133]	0.914
Triglycerides*	98.2 [72.5; 138]	97.0 [71;138]	102 [77;138]	0.103
Fasting BGL*	98.1 [90.0; 111.0]	99.0 [90; 112]	98.1 [91; 107]	0.536
Creatine*	0.812 [0.67; 0.94]	0.822 [0.66;0.95]	0.841 [0.68;0.94]	0.959
ABI*	1.01 [0.92; 1.11]	1.01 [0.92;1.11]	1.01 [0.92; 1.1]	0.688
CAC score*	0.290 [0; 22.7]	0.212 [0; 22.1]	0.127 [0; 28.8]	0.462
FEV1*	2.14 [1.76; 2.57]	2.11 [1.73; 2.54]	2.28 [1.98; 2.7]	0.003
FVC*	2.65 [2.16; 3.23]	2.87 [2.33;3.38]	2.59 [2.12; 3.16]	0.001
FEV1/FVC*	0.821 [0.77; 0.85]	0.820 [0.77;0.85]	0.803 [0.76; 0.85]	0.183

Table 4.2: Numerical characteristics

Characteristic	Overall (n = 942)	Sampling (n = 746)	Volunteers (n = 196)	P-values
Women	660 (70.1%)	523 (70.1%)	137 (69.9%)	0.955
Blacks	191 (20.3%)	125 (16.7%)	66 (33.6%)	0.001
education	309 (32.9%)	224 (30.1%)	85 (43.3%)	0.001
Income	338 (37.1%)	263 (36.5%)	75 (39.2%)	0.487
Smoking	190 (21.1%)	147 (20.5%)	43 (23.1%)	0.455
Asthma	99 (10.5%)	79 (10.6%)	20 (10.2%)	0.888
Diabetes	229 (25.9%)	183 (25.9%)	46 (25.5%)	0.912
Hypertension	734 (78.1%)	587 (78.7%)	147 (75.3%)	0.311
Heart attack	67 (7.22%)	60 (8.15%)	7 (3.66%)	0.022

Table 4.3: Categorical characteristics

4.3.1 Covariates selection

Before implementing the IPW_AB on the dataset, we will have to make a decision on which variables to use. Ideally, the determination of the propensity score model specifications should be guided by subject matter expertise, such as a thorough understanding of how a specific individual is participating in the nonprobability-based sample (i.e., what characteristics make a person a volunteer). However, in many cases, researchers lack such detailed knowledge and are instead faced with a plethora of covariates and possibly many functions of these covariates that should enter the propensity score (i.e., interaction terms, non-linearities, etc.). In this situation, the challenge for our IPW_AB approach is to decide which variables to include and which function to include in the logistic regression.

We have used the following criteria to choose the set of variables.

The first criterion is choosing variables that will not be selected as outcome variables in future research (often the clinically relevant variables). This is because it violates the principles of causal inference using potential outcome theory. Propensity score methods can only use variables that are confounders for the exposure-outcome relationship. Therefore, we would like to include as many variables that can be considered attributes and characteristics of the participants. Based on this, the variables we choose for estimation of the propensity score are: *Sex, Age, Race, Education, Income*. *Sex, age,* and *race* were complete, but for education we missed two values and 30 values for income. To accommodate the issue of missingness, we conducted a single k-nearest neighbor approach to impute the missing values for education and income (using only these variables). Since the missingness was very small we expect that this would not affect our analysis a lot. The terms involved in the logistic regression function are all main effects and all two-way interaction terms.

The second criterion is the missingness, we would like to make sure the missingness for the chosen variables is as few as possible. Multiple imputation, such as Predictive Mean Matching, would add to the uncertainty of fitting the propensity score. Imputation would work best, if we can assume that the missingness in the data is missing at random (and not missing not at random) when we can include all variables that are related to the missingness. This means we may need to add many variables to properly impute missingness and the selection of so many variables would go against the first criterion. If we would not impute, and only consider the complete case, we may introduce selection. This selection may potentially be corrected by updating the calculated weights for PBS, but may not be fully appropriate. The quality of imputation is not promising. Therefore, we would like to choose variables with few missingness.

4.3.2 Combining the NPBS and PBS

After deciding the variables used for estimating the propensity score, we would use the IPW_AB method to obtain a set of weights for the NPBS. The PBS would receive a set of weights that would consider the sampling design and the selection due to non-response. These were the weights d_i^B for the probability-based sample used in the correction methods. The RURAL sampling specialists have calculated these for the PBS. Thus, both samples (NPBS & PBS) have a set of weights that independently would add up to a number close to the population size.

To be able to use both samples, these weights must be scaled such that the total weight of both samples adds up to a number that is close to or equal to the population size. One approach that would be suitable is the method of raking [6], also referred to as proportional fitting, sample-balancing, or ratio estimation. Raking makes the marginal distribution of the covariates in our two samples equal to the marginal distribution of these covariates from a benchmark data set. It essentially scales the weights of the participants such that the marginal distributions become equal. The benchmark data set we will select is the American Community Survey [2]. Note that it is recommended to conduct raking when the distributions of the sample deviate from known distributions in the population [1]. If we are now using raking to match our two samples with an existing data set, one could argue whether any of our correction methods are necessary altogether.

Our investigation involved implementing five different raking approaches (one is used for comparison). We will use raking with the variables *Sex, Age, Race, Education, Income*. The variable age will be divided into four categories (25-34; 35-44; 45-54; and 55-64). For each combination of values for the categorical variables the participants have provided a total weight. These weights are then sequentially scaled such that the marginal distributions of the samples match the marginal distributions of the benchmarking data set.

- **Case 0:**

  This case simply compute the mean value of the PBS data with the probability-based sampled weight. No raking is performed to the data.

- **Case 1:** Here we applied our IPW_AB to generate weights for the NPBS and use the PBS weights for the NPBS. The two samples are then combined and the weights of all participants are then used in the raking method using the selected variables.

- **Case 2:** In this scenario, we did not utilize any weights, neither the probability-based sampled weight nor the nonprobability-based sampled weight. In other words, we set all weights equal to one. We then performed raking on the entire dataset to match it with the ACS data.

	Mean	Median	Standard Deviation
case 1	1.00	0.55	1.06
case 2	1.00	1.01	0.63
case 3	1.00	0.46	1.17
case 4	1.00	0.72	0.93
case 5	1.00	0.61	0.81

Table 4.4: Descriptive of weights

- **Case 3:** Here, we solely utilized the weights from the probability-based sampled data. These weights were combined with the raw nonprobability-based sampled data (i.e., we set the weights for the NPBS equal to one), followed by raking of the entire sample to align it with the ACS data.
- **Case 4:** Before we calculated the IPW_AB with the PBS sample, we applied raking to the probability-based sampled data to align the PBS with the ACS data. Then, utilizing the probability-based sample data with the raked weights, we generated weights for the nonprobability-based sample data using our IPW_AB approach. Then finally, we combined both datasets and performed raking once again to ensure alignment of the two samples together with the ACS data.
- **Case 5:** In this case, we just raked the PBS with the probability-based sampled weight to match with the ACS data.

4.4 Comparisons of estimation methods

Figure 4.1 shows the boxplots of the weights after applying the raking approaches with (or without) the IPW_AB approach for the 5 different raking approaches. Table 4.4 shows the descriptive analysis of the weights of the five raking methods. We can observe that the mean value of all the weights is one, which is expected since these weights are obtained after raking. We observe distinct patterns among the different cases analyzed. In Case 2, where weights of one are involved, the raking process concentrates the weights around 1, distributing them evenly. Introducing weights from the probability-based sampled data leads to increased variation. However, in Case 4, when the PBS data is initially raked to the standard data and then used to generate weights for the NPBS data, followed by another raking process, the variation is effectively mitigated.

Further analysis reveals that the correlation between each pair of the 4 cases of raking varies a lot. Pearson's correlation coefficients are provided in Table 4.5.

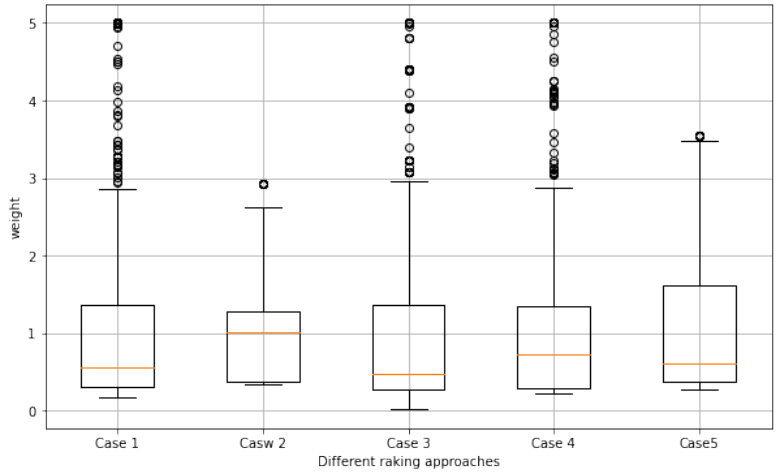


Figure 4.1: Comparison of Weight Distributions Across five raking approaches

	case 1	case 2	case 3	case 4	case 5
case 1	1.00				
case 2	0.591	1.00			
case 3	0.187	0.572	1.00		
case 4	0.932	0.671	0.129	1.00	
case 5	0.208	0.404	0.365	0.243	1.00

Table 4.5: Pearson's correlation coefficients

	case 0	case 1	case 2	case 3	case 4	case 5
Age	48.9	48.6	48.3	48.1	48.4	48.2
BMI	35.0	34.4	35.0	34.8	34.7	35.0
Weight	98.4	97.5	99.4	99.8	98.5	100
DBP	82.0	82.4	82.3	82.2	82.6	82.1
SBP	129	130	129	129	130	129
Cholesterol	183	182	185	183	184	184
HDL	53.1	52.8	52.9	52.2	53.1	52.4
LDL	107	106	108	107	108	109
Triglycerides	122	123	125	124	124	125
Fasting BGL	114	113	112	115	112	113
Creatine	0.882	0.909	0.921	0.919	0.918	0.925
ABI	0.997	1.01	1.00	1.00	1.01	0.997
CAC score	88.1	99.4	99.8	88.7	105	105
FEV1	2.29	2.40	2.36	2.37	2.38	2.34
FVC	2.88	3.03	2.98	2.91	3.01	2.96
FEV1/FVC	0.809	0.805	0.805	0.807	0.804	0.806

Table 4.6: Population mean of the numerical characteristics

It can be seen that the raking methods lead to really different weights for the participants, with the smallest correlation coefficient between case 3 and case 4 and between case 1 and case 3. Thus, the use of IPW_AB really changed the weights.

We also computed the population mean for numerical variables using the weights the different raking and correction procedures. Additionally, we calculated the frequencies and percentages for the categorical variables using the weights. The results for the numerical and categorical variables are reported in Tables 4.6 and 4.7 respectively.

Comparing the results from all variables, there isn't substantial differences in the estimated means and frequencies among the different cases. However, a concerning difference is present for the CAC score. Cases 1 and 2 exhibit relatively close values, whereas Case 3 shows a lower score. For this variable, case 4 (IPW_AB with twice raking) is closest to the case 5 (PBS with raking), possibly suggesting that case 4 is the best approach when both samples are being used.

	case 0	case 1	case 2	case 3	case 4	case 5
Women	0.654	0.548	0.548	0.548	0.548	0.548
Blacks	0.709	0.725	0.724	0.724	0.730	0.724
Education	5.65	5.64	5.53	5.63	5.55	5.54
Income	0.527	0.538	0.570	0.521	0.557	0.548
Smoking	0.199	0.226	0.226	0.216	0.237	0.223
Asthma	0.104	0.088	0.097	0.098	0.094	0.103
Diabetes	0.265	0.260	0.253	0.272	0.259	0.262
Hypertension	0.743	0.768	0.761	0.745	0.770	0.749
Heart Attack	0.082	0.059	0.075	0.080	0.066	0.088

Table 4.7: Population mean of the categorical characteristics

Chapter 5

Result & Future work

5.1 Discussion

According to the results from the case study, we did not notice a significant difference among all the raking approaches in terms of estimating the population mean, even though the weights attributed to all individuals from the PBS and NPBS were quite different between the correction methods involving raking.

We saw that when the weights for the PBS and NPBS are set equal to one (case 2), the variation in the raked weights (marginal weights) is smallest across all raking and correction methods and it provides a median value equal to the mean weight, suggesting symmetry in the weights. When we do apply the weights for PBS, and set the weight for NPBS equal to one (case 3), the raking process highly skews the weights, such that the median starts deviating from the mean and the standard deviation becomes rather large. Something similar occurs when we used the IPW_AB weights (case 1). The asymmetry and standard deviation is much less when we would first rake the PBS weights, then use the IPW_AB approach and rake all the obtained weights again (case 4). This approach shows rather symmetric weights, with a reasonable low standard deviation. Raking the PBS weights only (case 5) shows more skewed weights than case 4, but with a slightly lower standard deviation, but these results are calculated over a smaller set of participants. When the raked weights are used to compute the population mean for several numerical and binary variables, we find that the differences in means across the different raking approaches is not significant. The raking approach has a much stronger influence than the correction approach. Nevertheless, we do have to make a choice for calculating or estimating population values. Based on the results from the simulation and the descriptives of the raked weights, we believe that the double raking together with the pseudo-likelihood approach using both samples (IPW_AB) is the most suitable estimate. Once we have the standard

dataset, we aim to make the results as close as possible to the standard results. However, at the same time, we do not want to loose the characteristics of our own data. The motivation behind this is as follows: The first round of raking eliminates the bias introduced by the weights from the probability-based sample data. In this step, we use raking to standardize our data source, making it as representative as possible. We use this dataset to generate the weights for the nonprobability-based sample, using the best performing correction method (that produces weights) we found in the simulation. This approach makes the nonprobability-based sample as representative as the raked probability-based sample. Then the second round of raking ensures that the total set of weights are such that the combined data are representative to the population under study and ensures that the weights add up to the population size.

5.2 Future Work

- **Is the simulation of the logistic regression model sufficient?**
  In our study, we did not use sophisticated model selection approaches in the logistic regression function used in the pseudo-likelihood. Nevertheless, the mechanism for NPBS in the simulation study involves higher-order and may suggest that more sophisticated models for the propensity score are needed. Indeed our simulation involves 4 covariates (excluding the county), and there are $2^4 = 16$ possible subgroups in the simulation. However, we sampled from only 4 of these subgroups (using imbalances), meaning individuals from the remaining 12 subgroups could never be included in the NPBS. The simulation shows that better model fits have a positive influence on reducing bias of the NPBS sample. Thus, if we would allow all interactions (including the five-way interactions, and use a model selection step to include all the terms that are relevant, we may get a better performance of the pseudo-likelihood approach.

- **How to select the suitable covariates?**
  In terms of the covariate selection, our current strategy is trying not to include too many variables since we cannot put outcome variables in our inference step. It may violate the principles of causal inference using potential outcome theory. Indeed, propensity score methods used in the analysis of data from cohort studies can only use variables that are not outcomes and that are associated with the exposure of interest. Therefore in our case, we just choose some basic attributes. A more in depth simulation study could be conducted to investigate this aspect. The mechanism behind NPBS may then depend on clinical variables that will also be used in the correction step.

Instead of investigating the bias of the correction step we may investigate the effect of an exposure on the clinical outcome variable using weights that have been calculated with or without the clinical variable.

- **Will the results vary if different raking methods are used?**
  In terms of the raking step, we only made the marginal distribution of the single covariates equal in the sample and benchmark data. Extensions, where the bivariate marginal distributions of all pair-wise covariates are used may provide better results. It is interesting to see if the different raking method will lead to different results.

- **Whether the correction is necessary?**
  Whether correction of the NPBS is needed if a raking procedure is being used at the end. We saw that differences in averages for certain variables across different raking strategies were not very large, but we also matched the set of variables used in the IPW and in the raking approach. When studies show that many covariates should be used in the correction approach, bigger differences could occur, since benchmark data usually have limit information on participants, besides variables like age, sex, and race.

Bibliography

[1] Michael P Battaglia, David C Hoaglin, and Martin R Frankel. Practical considerations in raking survey data. *Survey practice*, 2(5), 2009.

[2] US Census Bureau. *Income, Earnings, and Poverty Data from the... American Community Survey*. US Census Bureau, 2008.

[3] Yilin Chen. Statistical analysis with non-probability survey samples. 2020.

[4] James Chipperfield, Julia Chessman, and Russell Lim. Combining household surveys using mass imputation to estimate population totals. *Australian New Zealand Journal of Statistics*, 54, 06 2012.

[5] Ellen M Daley, Robert J McDermott, Kelli R McCormack Brown, and Mark J Kittleson. Conducting web-based survey research: a lesson in internet designs. *American Journal of Health Behavior*, 27(2):116–124, 2003.

[6] Jean-Claude Deville, Carl-Erik Särndal, and Olivier Sautory. Generalized raking procedures in survey sampling. *Journal of the American statistical Association*, 88(423):1013–1020, 1993.

[7] Michael R Elliott and Richard Valliant. Inference for nonprobability samples. 2017.

[8] Michael R Elliott and Richard Valliant. Inference for nonprobability samples. 2017.

[9] Michele Jonsson Funk, Daniel Westreich, Chris Wiesen, Til Stürmer, M Alan Brookhart, and Marie Davidian. Doubly robust estimation of causal effects. *American journal of epidemiology*, 173(7):761–767, 2011.

[10] Jaroslav Hájek. Local asymptotic minimax and admissibility in estimation. In *Proceedings of the sixth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 175–194, 1972.

[11] Morris H Hansen and William N Hurwitz. On the theory of sampling from finite populations. *The Annals of Mathematical Statistics*, 14(4):333–362, 1943.

[12] Daniel G Horvitz and Donovan J Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, 47(260):663–685, 1952.

[13] Annika Isaksson and Gosta Forsman. A comparison between using the web and using the telephone to survey political opinions. In *annual meeting of the American Association for Public Opinion Research, Sheraton Music City, Nashville, TN*, 2003.

[14] Joseph D. Y. Kang and Joseph L. Schafer. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, 22(4):523–539, 2007.

[15] Jae Kim, Seho Park, Yilin Chen, and Changbao Wu. Combining non-probability and probability survey samples through mass imputation. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 184, 07 2021.

[16] JAE KWANG KIM and J. N. K. RAO. Combining data from two independent surveys: a model-assisted approach. *Biometrika*, 99(1):85–100, 2012.

[17] Sunghee Lee. Propensity score adjustment as a weighting scheme for volunteer panel web surveys. *Journal of official statistics*, 22(2):329, 2006.

[18] Sunghee Lee. Propensity score adjustment as a weighting scheme for volunteer panel web surveys. *Journal of Official Statistics*, 22:329–349, 2006.

[19] Anita McVey, F. Jay Breidt, Wayne A. Fuller, and Andrew Fuller. Two-phase estimation by imputation. 2002.

[20] Jerzy Neyman. On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. In *Breakthroughs in Statistics: Methodology and Distribution*, pages 123–150. Springer, 1992.

[21] Greg Ridgeway, Dan McCaffrey, Andrew Morral, Matthew Cefalu, Lane Burgette, Joseph Pane, and Beth Ann Griffin. Toolkit for weighting and analysis of nonequivalent groups: A guide to the twang package. *Vignette*, 2021:26, 2021.

[22] Douglas Rivers. "sampling for web surveys.". 01 2007.

[23] Richard M Royall. On finite population sampling theory under certain linear regression models. *Biometrika*, 57(2):377–387, 1970.

[24] Donald B Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

[25] Donald B Rubin. Using multivariate matched sampling and regression adjustment to control bias in observational studies. *Journal of the American Statistical Association*, 74(366a):318–328, 1979.

[26] Eleanor Singer. Reflections on surveys' past and future. *Journal of Survey Statistics and Methodology*, page smw026, 2016.

[27] Richard Valliant and Jill A Dever. Estimating propensity adjustments for volunteer web surveys. *Sociological Methods & Research*, 40(1):105–137, 2011.

[28] Ameneh Valojerdi and Leila Janani. A brief guide to propensity score analysis. *Medical Journal of the Islamic Republic of Iran*, 32:122–122, 12 2018.

[29] Changbao Wu. Statistical inference with non-probability survey samples. *Surv. Methodol*, 48:283–311, 2022.

Appendix A

Code Listings

A.1 Pseudo Likelihood Functions

Listing A.1: Pseudo Likelihood Functions and Optimization

```
import numpy as np
from scipy.optimize import fsolve

def psudo_likelihood(theta, ps_x, nps_x):
    pi = np.exp(np.dot(ps_x, theta.T)) / (1 + np.exp(np.dot(
        ps_x, theta.T)))
    res = nps_x.sum() - np.dot(weight * pi, ps_x)
    return res

def psudo_likelihood_ab(theta, ps_x, nps_x):
    pi_nps = np.exp(np.dot(nps_x, theta.T)) / (1 + np.exp(np
        .dot(nps_x, theta.T)))
    pi_ps = np.exp(np.dot(ps_x, theta.T)) / (1 + np.exp(np.
        dot(ps_x, theta.T)))
    res = np.dot(1 - pi_nps, nps_x) - (1 - len(nps_x) / sum(
        weight)) * np.dot(weight * pi_ps, ps_x)
    return res

# Example usage
initial_guess = np.zeros(15)  # Provide an initial guess for
    the value of theta
result = fsolve(psudo_likelihood, initial_guess, args=(ps_x,
    nps_x))
propensity_score_nps = np.exp(np.dot(nps_x, result.T)) / (1
   + np.exp(np.dot(nps_x, result.T)))
```

```
initial_guess = np.zeros(15)
result_combine = fsolve(psudo_likelihood_ab, initial_guess,
   args=(ps_x, nps_x))
propensity_score_combine_nps = np.exp(np.dot(nps_x,
   result_combine.T)) / (1 + np.exp(np.dot(nps_x,
   result_combine.T)))
```

A.2 XGBoost Function for Propensity Scores

Listing A.2: XGBoost Function for Computing Propensity Scores

```
import pandas as pd
from xgboost import XGBClassifier

def xgboost(ps, nps):
    ps['indicator'] = 1
    nps['indicator'] = 0
    combined_data = pd.concat([ps, nps], ignore_index=True)
    binary_columns = ['x1','x2', 'x3', 'x5']

    for col in binary_columns:
        combined_data[col] = (combined_data[col] == 1).
           astype(int)

    features = ['x1','x2','x3','x4','x5']
    target = 'indicator'

    # Initialize and fit the model
    model = XGBClassifier(n_estimators=5000, max_depth=4,
       learning_rate=0.01, verbosity=0)
    model.fit(combined_data[features], combined_data[target
       ])

    # Get propensity scores
    combined_data['propensity_score'] = model.predict_proba(
       combined_data[features])[:, 1]

    # Extract the propensity scores for the nps dataset
    nps_combined = combined_data[combined_data['indicator']
       == 0]
```

```
    return nps_combined['propensity_score']
```

A.3 Logistic Regression Function for Propensity Scores

Listing A.3: Logistic Regression Function for Computing Propensity Scores

```
import pandas as pd
import statsmodels.api as sm

def logit_regression(ps, nps):
    ps['indicator'] = 1
    nps['indicator'] = 0

    # Combine the datasets
    combined_data = pd.concat([ps, nps], ignore_index=True)

    # Extract the treatment and covariate data
    treatment = combined_data['indicator']
    covariates = combined_data[['x1','x2','x3','x4','x5','
       x4_x1','x4_x2','x4_x3','x4_x5','x2_x1','x2_x3','x2_x5
       ','x3_x1','x3_x5','x1_x5']]

    # Add a constant to the covariates for the intercept
    covariates = sm.add_constant(covariates)

    # Fit the logistic regression model
    logit_model = sm.Logit(treatment, covariates)
    result = logit_model.fit()

    # Predict the propensity scores
    propensity_scores = result.predict(covariates)

    # Add the propensity scores to the original dataset
    combined_data['propensity_score'] = propensity_scores

    nps_combined = combined_data[combined_data['indicator']
       == 0]

    return nps_combined['propensity_score']
```

List of Figures

List of Tables