Catalogue no. 12-001-X
ISSN 1492-0921

Survey Methodology

Dealing with undercoverage for non-probability survey samples

by Yilin Chen, Pengfei Li and Changbao Wu

Release date: January 3, 2024

Statistics Canada Statistique Canada

Canada

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, www.statcan.gc.ca.

You can also contact us by

Email at infostats@statcan.gc.ca

Telephone, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following numbers:

- Statistical Information Service 1-800-263-1136
- National telecommunications device for the hearing impaired 1-800-363-7629
- Fax line 1-514-283-9350

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed standards of service that its employees observe. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under "Contact us" > "Standards of service to the public."

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co-operation and goodwill.

Published by authority of the Minister responsible for Statistics Canada

An HTML version is also available.

Cette publication est aussi disponible en français.

Survey Methodology, December 2023
Vol. 49, No. 2, pp. 497-515
Statistics Canada, Catalogue No. 12-001-X

Dealing with undercoverage for non-probability survey samples

Yilin Chen, Pengfei Li and Changbao Wu[1]

Abstract

Population undercoverage is one of the main hurdles faced by statistical analysis with non-probability survey samples. We discuss two typical scenarios of undercoverage, namely, stochastic undercoverage and deterministic undercoverage. We argue that existing estimation methods under the positivity assumption on the propensity scores (i.e., the participation probabilities) can be directly applied to handle the scenario of stochastic undercoverage. We explore strategies for mitigating biases in estimating the mean of the target population under deterministic undercoverage. In particular, we examine a split population approach based on a convex hull formulation, and construct estimators with reduced biases. A doubly robust estimator can be constructed if a followup subsample of the reference probability survey with measurements on the study variable becomes feasible. Performances of six competing estimators are investigated through a simulation study and issues which require further investigation are briefly discussed.

Key Words: Auxiliary information; Calibration method; Convex hull; Doubly robust estimator; Inverse probability weighting; Model-based prediction; Outcome regression; Propensity score; Split population.

1. Introduction

Probability survey samples and design-based inference have been widely used in official statistics and many other scientific fields as a standard tool for data collection and analysis. In recent years, however, "*there has been a wind of change and other data sources are being increasingly explored*" (Beaumont, 2020). One of the major reasons for looking at other data sources is the decreasing response rates for probability survey samples, and the seriousness of the undercoverage problem due to nonresponse as well as challenges in dealing with it for valid statistical inference.

Non-probability survey samples are one of the emerging data sources. Their ascent in popularity started with surveys based on web panels but the more broad definition extends to any volunteer-based and/or convenient samples or even administrative records. Statistical analysis of non-probability survey samples faces many hurdles, with the unknown sample selection and participation mechanism and the unknown coverage of the target population as the most pressing ones. Non-probability samples are biased and do not represent the target population in any tractable way, unlike probability survey samples where the survey design information is available. Valid statistical inferences with non-probability samples require additional auxiliary information at the population level and suitable inferential frameworks. A popular framework is to assume that the required population auxiliary information is available in an existing probability survey sample from the same target population. This two-sample framework has been used in several methodological developments, including the sample matching method (Rivers, 2007) and mass imputation (Kim, Park, Chen and Wu, 2021), the weighted logistic regression for propensity score estimation using the

1. Yilin Chen, Pengfei Li and Changbao Wu, Department of Statistics and Actuarial Science, University of Waterloo, Waterloo ON N2L 3G1. E-mail: cbwu@uwaterloo.ca.

pooled sample (Valliant and Dever, 2011), the pseudo maximum likelihood method (Chen, Li and Wu, 2020), the pseudo empirical likelihood approach (Chen, Li, Rao and Wu, 2022), and the Bayesian approach (Wisniowski, Sakshaug, Ruiz and Blom, 2020), among others.

Statistical inferences with non-probability samples under the two-sample framework also require another critical assumption: the propensity score or the participation probability is positive for all the units in the target population. This is the so-called positivity assumption and is the foundation for the validity of several estimation methods proposed in the literature; see Section 2 for further discussion. With probability survey samples, this is equivalent to having a complete sampling frame without nonresponses. The positivity assumption is typically violated in practice for non-probability samples, due to limited geographic coverage of the population of interest and/or failing to reach subgroups of the population that are not as easily observable through convenient sampling methods. Violations of the positivity assumption lead to undercoverage problems and invalid results based on existing estimation methods. Undercoverage is a notoriously challenge problem in finite population sampling, and there is an added layer of complications with non-probability survey samples; see Elliott and Valliant (2017) for some extended discussion on the topic.

This paper discusses two typical scenarios of undercoverage in practice for non-probability survey samples: stochastic undercoverage and deterministic undercoverage. We argue in Section 3 that methods developed under the positivity assumption can be applied directly to handle stochastic undercoverage for valid inferences. Deterministic undercoverage involves a subpopulation for which certain crucial information is missing and no rigorous and valid estimation procedures can be developed under the existing two-sample framework. In Section 4, we first discuss strategies for mitigating biases due to deterministic undercoverage using existing methods, and identify conditions under which existing methods lead to valid estimation results. We then explore estimation procedures under the split population through a convex hull formulation. We show that the correct specification of the outcome regression model is essential to several estimation procedures and a doubly robust estimator can be constructed if a followup subsample of the reference probability sample with measurements on the study variable can be obtained. Performances of six competing estimators of the finite population mean are evaluated through a simulation study and the results are reported in Section 5. Brief discussions on issues which require further investigation and some concluding remarks are given in Section 6.

2. Assumptions and existing approaches

There have been exciting methodological developments in recent years on valid statistical inference with non-probability survey samples. One of the key assumptions used by several authors is the non-zero probability of participation in the non-probability survey of all units in the target population. Let $\mathcal{U} = \{1, 2, \ldots, N\}$ be the set of N labelled units for the target population. Let y_i and $\mathbf{x}_i$ be the values of the study variable y and the vector of auxiliary variables $\mathbf{x}$ for the i^{th} unit in the population. Estimation

procedures are developed for a univariate y with the focus on the population mean $\mu_y = N^{-1}\sum_{i=1}^{N} y_i$ but extensions can be made to other inferential problems similar to the theory of the Horvitz-Thompson estimator for design-based inference with probability survey samples.

Let $\mathcal{S}_A$ be the set of n_A participating units in the non-probability survey sample and $\{(y_i, \mathbf{x}_i),\ i \in \mathcal{S}_A\}$ be the sample dataset. The most crucial feature of non-probability survey samples is the unknown sample inclusion or participation mechanism. The recent literature on the topic assumes that the mechanism is guided by an underlying stochastic process. Let $R_i = I(i \in \mathcal{S}_A)$ be the indicator variable for unit i being included in the non-probability sample $\mathcal{S}_A$. Let

$$\pi_i^A = P(i \in \mathcal{S}_A | \mathbf{x}_i, y_i) = P(R_i = 1 | \mathbf{x}_i, y_i), \quad i = 1, 2, \ldots, N.$$

The term "propensity scores" from the missing data literature (Rosenbaum and Rubin, 1983) was used for π_i^A by Chen et al. (2020), among several other authors. Some authors preferred to use the term "participation probabilities" for π_i^A; see, for instance, Beaumont (2020) and Rao (2021), among others.

2.1 Assumptions

The following assumptions have been used in the recent literature on statistical inference with non-probability survey samples; see, for instance, Wu (2022) and several key references therein.

- **A1** The sample inclusion and participation indicator R_i and the study variable y_i are independent given the set of auxiliary variables $\mathbf{x}_i$, i.e., $(R_i \perp\!\!\!\perp y_i) | \mathbf{x}_i$.
- **A2** All the units in the target population have non-zero propensity scores, i.e., $\pi_i^A > 0$, $i = 1, 2, \ldots, N$.
- **A3** The indicator variables $R_1, R_2, \ldots, R_N$ are independent given the set of auxiliary variables $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N)$.
- **A4** There exists a probability survey sample $\mathcal{S}_B$ of size n_B with information on the auxiliary variables $\mathbf{x}$ (but not on y) available in the dataset $\{(\mathbf{x}_i, d_i^B), i \in \mathcal{S}_B\}$, where d_i^B are the design weights for the probability sample $\mathcal{S}_B$.

Assumption **A1** is similar to the concept of missing-at-random (MAR) widely used for missing data analysis. Assumption **A3** is more of a convenient tool for likelihood-based estimation of propensity scores, and it is not crucial to the validity of several existing estimating procedures (Wu, 2022). Assumption **A4** is on the two-sample framework where auxiliary information on the target population is available from an existing probability survey sample. It is the basic setting for most estimation procedures proposed in the literature on non-probability survey samples.

Assumption **A2** refers to the so-called positivity condition. For probability surveys, this is equivalent to conditions that the sampling frames are complete and there are no hardcore nonrespondents. In other words, the sampled population is identical to the target population, and statistical inferences based on the survey

sample are valid for the target population. In practice, assumption **A2** is often violated for non-probability survey samples due to the voluntary and convenience nature of survey participation, resulting in undercoverage problems and invalid statistical statements on the target population.

2.2 Approaches to inference

There are three main approaches to inference using non-probability survey samples under assumptions **A1**-**A4**: (i) inverse probability weighting (IPW) based on an assumed model q for the propensity scores; (ii) model-based prediction based on an assumed outcome regression model ξ; and (iii) doubly robust (DR) procedures using both the estimated propensity scores and the outcome regression model.

Under assumption **A1**, the propensity scores $\pi_i^A = P(R_i = 1 | \mathbf{x}_i) = \pi(\mathbf{x}_i)$ are a function of the auxiliary variables $\mathbf{x}_i$ with an unknown form $\pi(\cdot)$. Let $\pi_i^A = \pi(\mathbf{x}_i, \boldsymbol{\alpha})$ be a specified parametric form with unknown model parameters $\boldsymbol{\alpha}$. Under the two-sample setting where the population auxiliary information is supplied by the reference probability sample $\mathcal{S}_B$, the pseudo log-likelihood function for $\boldsymbol{\alpha}$ proposed by Chen et al. (2020) is given by

$$\ell^*(\boldsymbol{\alpha}) = \sum_{i \in \mathcal{S}_A} \log\left(\frac{\pi_i^A}{1-\pi_i^A}\right) + \sum_{i \in \mathcal{S}_B} d_i^B \log(1-\pi_i^A). \tag{2.1}$$

The maximum pseudo-likelihood estimator $\hat{\boldsymbol{\alpha}}$ is the maximizer of $\ell^*(\boldsymbol{\alpha})$ and can be obtained as the solution to the pseudo score equations given by $\mathbf{U}(\boldsymbol{\alpha}) = \partial \ell^*(\boldsymbol{\alpha}) / \partial \boldsymbol{\alpha} = \mathbf{0}$. If the logistic regression model is assumed for the propensity scores where $\pi_i^A = 1 - \{1 + \exp(\mathbf{x}_i' \boldsymbol{\alpha})\}^{-1}$, the pseudo score functions are given by

$$\mathbf{U}(\boldsymbol{\alpha}) = \sum_{i \in \mathcal{S}_A} \mathbf{x}_i - \sum_{i \in \mathcal{S}_B} d_i^B \, \pi(\mathbf{x}_i, \boldsymbol{\alpha}) \, \mathbf{x}_i. \tag{2.2}$$

The estimated propensity scores are obtained as $\hat{\pi}_i^A = \pi(\mathbf{x}_i, \hat{\boldsymbol{\alpha}}), \, i \in \mathcal{S}_A$. The inverse probability weighted (IPW) estimator of μ_y is computed as

$$\hat{\mu}_{y\text{IPW}} = \frac{1}{\hat{N}^A} \sum_{i \in \mathcal{S}_A} \frac{y_i}{\hat{\pi}_i^A}, \tag{2.3}$$

where $\hat{N}^A = \sum_{i \in \mathcal{S}_A} (\hat{\pi}_i^A)^{-1}$ is the estimated population size. The estimator $\hat{\mu}_{y\text{IPW}}$ is consistent under the joint randomization of the propensity score model q and the probability sampling design p for the reference probability sample $\mathcal{S}_B$.

The model-based prediction approach to inference also relies heavily on the first assumption. Under assumption **A1**, the conditional distribution of y given $\mathbf{x}$ for units in the non-probability sample $\mathcal{S}_A$ (i.e., $R = 1$) is the same as the conditional distribution of y given $\mathbf{x}$ for units in the target population. It allows a valid model on y given $\mathbf{x}$ to be built using the non-probability sample dataset $\{(y_i, \mathbf{x}_i), i \in \mathcal{S}_A\}$. Under the semiparametric outcome regression model ξ as described in Wu (2022) with the first conditional moment specified as $E_\xi(y_i | \mathbf{x}_i) = m(\mathbf{x}_i, \boldsymbol{\beta})$, the model parameters $\boldsymbol{\beta}$ can be consistently estimated by $\hat{\boldsymbol{\beta}}$

using the non-probability sample. Let $y_i^* = m(\mathbf{x}_i, \hat{\boldsymbol{\beta}})$ be the fitted value of y for unit $i \in \mathcal{S}_A$ or predicted value of y for unit $i \notin \mathcal{S}_A$. A general form of the model-based prediction estimator of μ_y is computed as

$$\hat{\mu}_{y\text{MI}} = \frac{1}{\hat{N}^B} \sum_{i \in \mathcal{S}_B} d_i^B \, y_i^*, \tag{2.4}$$

where $\hat{N}^B = \sum_{i \in \mathcal{S}_B} d_i^B$. The subscript "MI" refers to "Mass Imputation", since the estimator is constructed based on the reference probability sample $\mathcal{S}_B$ with the unobserved y treated as 100% missing for the sample and imputed for all the units in the sample. The estimator $\hat{\mu}_{y\text{MI}}$ is consistent under the joint randomization of the outcome regression model ξ and the probability sampling design p for $\mathcal{S}_B$.

The doubly robust estimator of μ_y is computed by using the estimated propensity scores $\hat{\pi}_i^A = \pi(\mathbf{x}_i, \hat{\boldsymbol{\alpha}})$ and fitted or predicted values $y_i^* = m(\mathbf{x}_i, \hat{\boldsymbol{\beta}})$ and is given by (Chen et al., 2020)

$$\hat{\mu}_{y\text{DR}} = \frac{1}{\hat{N}^A} \sum_{i \in \mathcal{S}_A} \frac{y_i - y_i^*}{\hat{\pi}_i^A} + \frac{1}{\hat{N}^B} \sum_{i \in \mathcal{S}_B} d_i^B \, y_i^*. \tag{2.5}$$

The estimator $\hat{\mu}_{y\text{DR}}$ is consistent under the probability sampling design p for $\mathcal{S}_B$ and one of the correctly specified models, q or ξ.

3. Two practical scenarios with undercoverage

It is known in design-based inference that there exists an unbiased estimator of μ_y in a subclass of the so-called Godambe class of linear estimators if and only if the first order inclusion probabilities are non-zero for all the units in the finite population (Wu and Thompson, 2020). For probability survey samples, zero-inclusion probabilities are the consequences of incomplete sampling frames and nonrespondents, leading to undercoverage problems for the target population.

The positivity assumption **A2** which states that $\pi_i^A = P(R_i = 1 \mid \mathbf{x}_i, y_i) > 0$ for all i in the target population is indeed the same condition for the validity of the IPW estimator, which is adapted from the Horvitz-Thompson estimator for probability survey samples, under the propensity score model q. The positivity assumption used in missing data analysis and causal inference is not an issue since the propensity scores are only defined for units in the sample. For non-probability survey samples, assumption **A2** is often violated in practice for two major reasons: incomplete sampling frame(s) and voluntary participation. The sampling frame(s) used for selecting a non-probability survey sample is typically a convenient list such as a web panel, and it is almost surely incomplete for the target population. Participation in a non-probability survey sample is voluntary and nonresponse and refusals are an inherent part of the recruiting process.

Let $\mathcal{U}$ be the set of N units for the target population. Let $\mathcal{U}_0 = \{i \mid i \in \mathcal{U}$ and $\pi_i^A > 0\}$. It is apparent that $\mathcal{U}_0 \subset \mathcal{U}$ and $\mathcal{U}_0 \neq \mathcal{U}$ when assumption **A2** is violated. Let $\mathcal{U}_1 = \{i \mid i \in \mathcal{U}$ and $\pi_i^A = 0\}$. It follows that $\mathcal{U} = \mathcal{U}_0 \cup \mathcal{U}_1$. Let $N = N_0 + N_1$ where N_0 and N_1 are the sizes of the two subpopulations $\mathcal{U}_0$ and $\mathcal{U}_1$. Let $\mu_{y0} = N_0^{-1} \sum_{i \in \mathcal{U}_0} y_i$ and $\mu_{y1} = N_1^{-1} \sum_{i \in \mathcal{U}_1} y_i$. We have $\mu_y = W_0 \mu_{y0} + W_1 \mu_{y1}$, where $W_k = N_k / N$ for $k = 0, 1$. If

$\mathcal{S}_A$ is a sample from $\mathcal{U}_0$, and $\hat{\mu}_{yA}$ is an "unbiased estimator" based on $\mathcal{S}_A$, we usually have $E(\hat{\mu}_{yA}) = \mu_{y0}$, and the bias of using $\hat{\mu}_{yA}$ to estimate μ_y is given by

$$E(\hat{\mu}_{yA}) - \mu_y = W_1(\mu_{y0} - \mu_{y1}).$$

The two major factors for the amount of bias are (i) the size of the subpopulation (i.e., N_1) not represented by the sample $\mathcal{S}_A$; and (ii) the difference (i.e., $\mu_{y0} - \mu_{y1}$) between all potential participants of the sample and those who have no chances to be included in the sample. We discuss two practical scenarios for undercoverage problems and their implications on inference.

3.1 Stochastic undercoverage

The first scenario is termed as *stochastic undercoverage*, where the non-probability sample $\mathcal{S}_A$ is selected from a subpopulation $\mathcal{U}_0$ and the $\mathcal{U}_0$ itself can be viewed as a random sample from $\mathcal{U}$ (Chen, 2020; Wu, 2022). A typical example for this scenario is when the contact list of an existing large probability survey sample is used to recruit participants for the non-probability survey sample. Another example is when the participants for the non-probability sample are selected from a very large commercial panel, and the composition of the panel mimics the distributions of the target population in terms of key demographical variables. One can argue that it falls into the scenario of stochastic undercoverage. A less obvious example is a convenient sample of respondents recruited from shoppers at local shopping centers over certain period of time. If the target population consists of certain types of consumers in the region and there is a belief that such consumers have a non-trivial chance to visit one of the shopping centers during the time period, then it is also a case of stochastic undercoverage.

Let $D_i = 1$ if $i \in \mathcal{U}_0$ and $D_i = 0$ otherwise, $i = 1, 2, \ldots, N$. Noting that $R_i = I(i \in \mathcal{S}_A)$, we have

$$P(R_i = 1 \mid \mathbf{x}_i, y_i, D_i = 1) > 0 \quad \text{and} \quad P(R_i = 1 \mid \mathbf{x}_i, y_i, D_i = 0) = 0$$

for $i = 1, 2, \ldots, N$. If the subpopulation $\mathcal{U}_0$ is formed with an underlying stochastic mechanism such that $P(D_i = 1 \mid \mathbf{x}_i, y_i) > 0$ for all $i \in \mathcal{U}$, we have

$$\pi_i^A = P(R_i = 1 \mid \mathbf{x}_i, y_i) = P(R_i = 1 \mid \mathbf{x}_i, y_i, D_i = 1)\, P(D_i = 1 \mid \mathbf{x}_i, y_i) > 0$$

for $i = 1, 2, \ldots, N$. The positivity assumption **A2** is indeed satisfied under scenarios of stochastic undercoverage, and estimation procedures developed under the assumption can be used directly to provide valid inferences. A practical issue with stochastic undercoverage is how to specify a model for propensity scores due to the two-phase arguments involving R_i and D_i. See Section 5 for further discussion.

3.2 Deterministic undercoverage

Many non-probability samples are volunteer-based convenience samples, and the potential participants often possess certain characteristics which are unique to the group. For instance, if participation in a survey

requires the use of a computer and access to the Internet, then those who do not have Internet access or never used a computer will have no chance to be included. The severity of undercoverage in this case depends largely on the proportion of the population being excluded.

The subpopulation $\mathcal{U}_1 = \{i \mid i \in \mathcal{U} \text{ and } \pi_i^A = 0\}$ may be conceptually defined through an accessibility function. Let $\Phi(\mathbf{x}_i)$ be a function of $\mathbf{x}_i$ that measures the accessibility of unit i to the survey. An individual with a small value of $\Phi(\mathbf{x}_i)$ will have (practically) no access to the survey. More formally, we have $\pi_i^A = P(i \in \mathcal{S}_A \mid \mathbf{x}_i, y_i) = 0$ if $\Phi(\mathbf{x}_i) \le c$ for an unknown cut-off value c on accessibility. The two subpopulations can alternatively be defined as

$$\mathcal{U}_0 = \{i \mid i \in \mathcal{U} \text{ and } \Phi(\mathbf{x}_i) > c\} \text{ and } \mathcal{U}_1 = \{i \mid i \in \mathcal{U} \text{ and } \Phi(\mathbf{x}_i) \le c\}. \tag{3.1}$$

The truncation on $\Phi(\mathbf{x}_i)$ to exclude certain units from the non-probability survey can be viewed as a deterministic process, which motivates the use of the term "*deterministic undercoverage*". An overly simplified example is when x_i represents the "age" of unit i and all young individuals (i.e., $x_i \le c$ for a chosen c) are excluded from the survey.

If the two subpopulations $\mathcal{U}_0$ and $\mathcal{U}_1$ can be clearly identified, valid statistical inferences can be claimed for the subpopulation $\mathcal{U}_0$. Extending the results to the target population $\mathcal{U}$ may be possible for certain scenarios but has the risk of overstretching with unrealistic assumptions.

4. Strategies for dealing with deterministic undercoverage

Deterministic undercoverage has similarities to frame and nonresponse undercoverage for probability survey samples. There are two major difficulties with inferences on the target population: the identification of the two subpopulations $\mathcal{U}_0$ and $\mathcal{U}_1$ and the lack of information on $\mathcal{U}_1$. In this section, we discuss approaches to mitigating biases of estimation due to the undercoverage and potential issues with these methods.

4.1 Calibrated IPW approach

Under the positivity assumption **A2** and the specified parametric form $\pi_i^A = \pi(\mathbf{x}_i, \boldsymbol{\alpha})$, the pseudo score functions given by $\mathbf{U}(\boldsymbol{\alpha}) = \partial \ell^*(\boldsymbol{\alpha}) / \partial \boldsymbol{\alpha}$ from (2.1) can be replaced by a set of unbiased estimating equations (Chen et al., 2020; Wu, 2022)

$$\mathbf{G}(\boldsymbol{\alpha}) = \sum_{i \in \mathcal{S}_A} \mathbf{h}(\mathbf{x}_i, \boldsymbol{\alpha}) - \sum_{i \in \mathcal{S}_B} d_i^B \pi(\mathbf{x}_i, \boldsymbol{\alpha})\, \mathbf{h}(\mathbf{x}_i, \boldsymbol{\alpha}), \tag{4.1}$$

where $\mathbf{h}(\mathbf{x}_i, \boldsymbol{\alpha})$ is a user-specified vector of functions with the same dimension of $\boldsymbol{\alpha}$. If we let $\mathbf{h}(\mathbf{x}, \boldsymbol{\alpha}) = \mathbf{x} / \pi(\mathbf{x}, \boldsymbol{\alpha})$, the estimating functions given in (4.1) reduce to

$$\mathbf{G}(\boldsymbol{\alpha}) = \sum_{i \in \mathcal{S}_A} \frac{\mathbf{x}_i}{\pi(\mathbf{x}_i, \boldsymbol{\alpha})} - \sum_{i \in \mathcal{S}_B} d_i^B \mathbf{x}_i. \tag{4.2}$$

Note that $\mathbf{G}(\boldsymbol{\alpha}) = \mathbf{0}$ becomes the calibration equations $\sum_{i\in\mathcal{S}_A} \mathbf{x}_i/\pi(\mathbf{x}_i, \boldsymbol{\alpha}) = \sum_{i\in\mathcal{S}_B} d_i^B \mathbf{x}_i$ with the population totals $\sum_{i=1}^N \mathbf{x}_i$ estimated by the probability sample $\mathcal{S}_B$. Let $\hat{\boldsymbol{\alpha}}_C$ be the solution to $\mathbf{G}(\boldsymbol{\alpha}) = \mathbf{0}$, where the subscript "$_C$" indicates "Calibration". Let $\hat{\pi}_i^C = \pi(\mathbf{x}_i, \hat{\boldsymbol{\alpha}}_C)$. It is assumed that the first component of $\mathbf{x}$ is 1 so that

$$\hat{N}_C^A = \sum_{i\in\mathcal{S}_A} (\hat{\pi}_i^C)^{-1} = \sum_{i\in\mathcal{S}_B} d_i^B = \hat{N}^B.$$

The calibrated IPW estimator of μ_y is computed as $\hat{\mu}_{y\text{IPW}}^C = (\hat{N}_C^A)^{-1} \sum_{i\in\mathcal{S}_A} y_i/\hat{\pi}_i^C$. The term "Calibrated IPW" was first used by Chen (2020). The idea was discussed by several other authors, including Rao (2021).

Under deterministic undercoverage, the parametric form with the restriction $\pi_i^A = \pi(\mathbf{x}_i, \boldsymbol{\alpha}) > 0$ for all i is clearly misspecified. As a consequence, the conventional IPW estimator $\hat{\mu}_{y\text{IPW}}$ given by (2.3) is no longer consistent. The calibrated IPW estimator $\hat{\mu}_{y\text{IPW}}^C$ can reduce the bias if the outcome regression model is linear, i.e., $E_\xi(y_i \mid \mathbf{x}_i) = \mathbf{x}_i'\boldsymbol{\beta}$. The calibrated IPW estimator under this scenario is an approximately model-unbiased prediction estimator (with the estimated population totals from the probability sample $\mathcal{S}_B$) since

$$E_p E_\xi \left\{ \frac{1}{\hat{N}_C^A} \sum_{i\in\mathcal{S}_A} \frac{y_i}{\hat{\pi}_i^C} \right\} = E_p \left\{ \frac{1}{\hat{N}_C^A} \sum_{i\in\mathcal{S}_A} \frac{\mathbf{x}_i}{\hat{\pi}_i^C} \right\}' \boldsymbol{\beta} = E_p \left\{ \frac{1}{\hat{N}^B} \sum_{i\in\mathcal{S}_B} d_i^B \mathbf{x}_i \right\}' \boldsymbol{\beta} \doteq E_\xi(\mu_y).$$

The approximate equal sign in the last step amounts to estimating N by $\hat{N}^B$.

A question of both practical and theoretical interest is whether the solution to $\mathbf{G}(\boldsymbol{\alpha}) = \mathbf{0}$ exists, where $\mathbf{G}(\boldsymbol{\alpha})$ is given in (4.2). The answer depends on the chosen parametric form of $\pi(\mathbf{x}_i, \boldsymbol{\alpha})$. Under a generalized linear model with $\pi_i = E(R_i \mid \mathbf{x}_i) = g(\mathbf{x}_i'\boldsymbol{\alpha})$, where $g(\cdot)$ is the so-called (monotone increasing) inverse link function, we have

$$\mathbf{H}(\boldsymbol{\alpha}) = \frac{\partial}{\partial \boldsymbol{\alpha}} \mathbf{G}(\boldsymbol{\alpha}) = -\sum_{i\in\mathcal{S}_A} \frac{k(\mathbf{x}_i'\boldsymbol{\alpha})}{\{g(\mathbf{x}_i'\boldsymbol{\alpha})\}^2} \mathbf{x}_i \mathbf{x}_i',$$

where $k(t) = dg(t)/dt > 0$. The matrix $\mathbf{H}(\boldsymbol{\alpha})$ is negative definite, as long as the data matrix $\{\mathbf{x}_i, i \in \mathcal{S}_A\}$ is of full rank, and the usual Newton-Raphson iterative procedures for solving $\mathbf{G}(\boldsymbol{\alpha}) = \mathbf{0}$ is guaranteed to converge.

The calibrated IPW estimator can also be constructed when a linear regression model is not appropriate but there are sufficient grounds to use a nonlinear model in the form of $E_\xi(y_i \mid \mathbf{x}_i) = m(\mathbf{x}_i, \boldsymbol{\beta})$ with a known function $m(\cdot, \cdot)$. For instance, if y_i is a binary variable, then $m(\mathbf{x}_i, \boldsymbol{\beta})$ no longer has a linear form but may be chosen as the inverse logit function. Let $\hat{\boldsymbol{\beta}}$ be an estimator of $\boldsymbol{\beta}$ obtained by using suitable estimation method and the non-probability sample dataset $\{(y_i, \mathbf{x}_i), i \in \mathcal{S}_A\}$. Let $\hat{m}_i = m(\mathbf{x}_i, \hat{\boldsymbol{\beta}})$. The calibrated propensity scores are computed as $\hat{\pi}_i^C = (w_i^C)^{-1}$, where the calibrated weights w_i^C are obtained in two steps:

(1) Compute the initial propensity scores $\hat{\pi}_i^O = \pi(\mathbf{x}_i, \hat{\boldsymbol{\alpha}})$, $i \in \mathcal{S}_A$, where $\hat{\boldsymbol{\alpha}}$ is the solution to the pseudo score equations from (2.2). Let $w_i^O = (\hat{\pi}_i^O)^{-1}$, $i \in \mathcal{S}_A$.

(2) Obtain the model-calibrated weights w_i^C, $i \in \mathcal{S}_A$ by minimizing the distance measure $D = \sum_{i \in \mathcal{S}_A} \{w_i^C - w_i^O\}^2 / w_i^O$ subject to constraints

$$\sum_{i \in \mathcal{S}_A} w_i^C = \hat{N}^B \quad \text{and} \quad \sum_{i \in \mathcal{S}_A} w_i^C \hat{m}_i = \sum_{i \in \mathcal{S}_B} d_i^B \hat{m}_i. \tag{4.3}$$

The constraints used in (4.3) follow the general model-calibration procedures of Wu and Sitter (2001).

The calibrated IPW approach uses the outcome regression model to mitigate the potential bias of the IPW estimator with deterministic undercoverage. When a linear outcome regression model can be justified, the calibration step does not involve estimation of the model parameters $\boldsymbol{\beta}$ and hence leads to a more robust estimator than the model-based prediction estimator. The model-calibrated IPW estimator under a nonlinear outcome regression model requires the estimator $\hat{\boldsymbol{\beta}}$ which is obtained by fitting the model with the non-probability sample. There is a risk of extrapolation in computing $\hat{m}_i = m(\mathbf{x}_i, \hat{\boldsymbol{\beta}})$ for $i \in \mathcal{S}_B$. See Section 4.2 for further discussion.

4.2 Model-based prediction approach

A common scenario for deterministic undercoverage is that units lacking of certain features have no access to the non-probability survey, and the features are reflected in values of certain auxiliary variables. In practice, the first step for analyzing a non-probability survey dataset is to check the (unweighted) empirical marginal distributions of auxiliary variables which are potentially related to survey participation, and compare them to the weighted sample distributions of the variables using the reference probability sample. In particular, the observed range (or the support) of each auxiliary variable from the non-probability sample should be compared to those from the probability sample.

Model-based prediction approach through mass imputation relies on a conditional model of y given $\mathbf{x}$. While the conditional moment structure $E_\xi(y_i \mid \mathbf{x}_i) = m(\mathbf{x}_i, \boldsymbol{\beta})$, which is assumed for the target population, most likely holds for any samples, there are two problematic consequences with fitting the model using a sample which has a limited range in the observed auxiliary variables. The first is unreliable estimation of the model parameters with inflated variances for the estimators. The second is the danger of extrapolation in using the fitted model for prediction. These observations have been sufficiently documented in the existing literature on regression modelling and analysis. Tan (2007) expressed concerns on extrapolation in using a fitted outcome regression model with a biased sample in the construction of doubly robust estimators for missing data analysis and causal inference.

If the non-probability sample includes all the important auxiliary variables which are required for characterizing the participation behaviour and the outcome regression, and if the observed ranges of the auxiliary variables are similar to those from the reference probability sample, a model-based prediction estimator may be preferred over the IPW estimator in the presence of deterministic undercoverage. The calibrated IPW estimator discussed in Section 4.1 is especially attractive under a linear regression model

since estimation of the model parameters $\boldsymbol{\beta}$ is not needed and therefore the two issues with model-based prediction estimators, namely, the inflated variances for parameter estimates and the danger of extrapolation, become non-issues.

4.3 The split population approach

The conceptually defined two subpopulations $\mathcal{U}_0$ and $\mathcal{U}_1$ play a central role in the discussion of deterministic undercoverage. The non-probability sample $\mathcal{S}_A$ can be viewed as coming from $\mathcal{U}_0$ and satisfying the positivity assumption. It is tempting to develop tools to separate units in the reference probability sample $\mathcal{S}_B$ that belong to $\mathcal{U}_0$ or $\mathcal{U}_1$, and to further develop strategies for dealing with the split population.

The accessibility function $\Phi(\mathbf{x})$ introduced in Section 3.2 is a useful tool for the task. Suppose that $\Phi(\mathbf{x})$ is a convex function of $\mathbf{x}$ and $\mathcal{U}_0$ and $\mathcal{U}_1$ are defined in (3.1) with an unknown threshold c that separates units from the two subpopulations. Let $\mathcal{H}_k$ be the convex hull generated by $\{\mathbf{x}_i : i \in \mathcal{U}_k\}$ for $k = 0, 1$. It follows that $\Phi(\mathbf{x}) > c$ if $\mathbf{x} \in \mathcal{H}_0$ and $\Phi(\mathbf{x}) \le c$ if $\mathbf{x} \in \mathcal{H}_1$. There are no overlaps between $\mathcal{H}_0$ and $\mathcal{H}_1$.

Let $\mathcal{H}_A$ be the convex hull formed by $\{\mathbf{x}_i : i \in \mathcal{S}_A\}$. We have $\mathcal{H}_A \subseteq \mathcal{H}_0$ and the difference between the two becomes negligible when n_A is large. Similarly, the convex hull $\mathcal{H}_B$ formed by $\{\mathbf{x}_i : i \in \mathcal{S}_B\}$ approximates $\mathcal{H}_0 \cup \mathcal{H}_1$ when n_B is large since $\mathcal{S}_B$ represents the entire target population $\mathcal{U}$. The two subpopulations $\mathcal{U}_0$ and $\mathcal{U}_1$ can be identified through a split among units in the reference probability sample $\mathcal{S}_B = \mathcal{S}_{B,0} \cup \mathcal{S}_{B,1}$, where

$$\mathcal{S}_{B,0} = \left\{ j \,\middle|\, j \in \mathcal{S}_B \text{ and } \mathbf{x}_j \in \mathcal{H}_A \right\}$$

and $\mathcal{S}_{B,1} = \mathcal{S}_B \setminus \mathcal{S}_{B,0}$. Note that verifying $\mathbf{x}_j \in \mathcal{H}_A$ is equivalent to checking if there exists a sequence of constants $a_i \ge 0$ for $i \in \mathcal{S}_A$ such that

$$\sum_{i \in \mathcal{S}_A} a_i = 1 \quad \text{and} \quad \sum_{i \in \mathcal{S}_A} a_i \mathbf{x}_i = \mathbf{x}_j.$$

It can be done with existing computational packages. The sizes N_0 and N_1 of the two subpopulations $\mathcal{U}_0$ and $\mathcal{U}_1$ can be estimated by

$$\hat{N}_k^B = \sum_{i \in \mathcal{S}_{B,k}} d_i^B, \quad k = 0, 1,$$

which satisfy $\hat{N}_0^B + \hat{N}_1^B = \hat{N}^B$.

Kim and Rao (2018) described an idea on splitting the population using a modified nearest neighbour method. They defined $\mathcal{S}_{B,0}$ as the set of units in $\mathcal{S}_B$ which have a "close neighbour" in $\mathcal{S}_A$. More formally, they define

$$\mathcal{S}_{B,0} = \left\{ j \,\middle|\, j \in \mathcal{S}_B \text{ and } \min_{i \in \mathcal{S}_A} \left| \mathbf{x}_i - \mathbf{x}_j \right| < \epsilon \right\},$$

where $\epsilon > 0$ is a pre-specified tolerance measuring similarities in $\mathbf{x}$ among units. Choosing a value for ϵ, however, is difficult in practice and the idea has not been developed further in the literature.

4.4 Estimation for the split population

Estimation for the split population involves separate treatments for $\mathcal{U}_0$ and $\mathcal{U}_1$. Note that $\mu_y = W_0\mu_{y0} + W_1\mu_{y1}$. Estimation procedures developed under the assumptions **A1**-**A4** can be applied directly for the estimation of μ_{y0} by treating $\mathcal{S}_{B,0}$ as the reference probability sample. Let $\hat{\mu}_y = \hat{W}_0\hat{\mu}_{y0} + \hat{W}_1\hat{\mu}_{y1}$, where $\hat{W}_k = \hat{N}_k^B / \hat{N}^B$ for $k = 0, 1$. The severity of the deterministic undercoverage from using $\hat{\mu}_{y0}$ as an estimator for μ_y is partially reflected by the value of $\hat{W}_1$. When $\hat{W}_1$ is small as compared to $\hat{W}_0$, we may ignore the issue with undercoverage and proceed with estimation under the assumption that $\pi_i^A > 0$ for all i.

It is apparent that valid estimation of μ_{y1} requires additional information on y since the only relevant data in the two samples $\mathcal{S}_A$ and $\mathcal{S}_B$ on the subpopulation $\mathcal{U}_1$ are the auxiliary information $\{\mathbf{x}_i, i \in \mathcal{S}_{B,1}\}$ from the split reference probability sample. In the absence of any additional information on y for units in $\mathcal{U}_1$, we propose a hybrid estimator of μ_y as follows. We first estimate the propensity scores under the assumption that $\pi_i^A > 0$ for $i \in \mathcal{U}_0$. Let $\hat{\pi}_{i0}^A = \pi(\mathbf{x}_i, \hat{\boldsymbol{\alpha}}_0)$ under a parametric propensity score model $\pi_i^A = \pi(\mathbf{x}_i, \boldsymbol{\alpha})$, where $\hat{\boldsymbol{\alpha}}_0$ is the pseudo maximum likelihood estimator of $\boldsymbol{\alpha}$. If a logistic regression model is used, then $\hat{\boldsymbol{\alpha}}_0$ is the solution to

$$\sum_{i \in \mathcal{S}_A} \mathbf{x}_i - \sum_{i \in \mathcal{S}_{B,0}} d_i^B \pi(\mathbf{x}_i, \boldsymbol{\alpha})\, \mathbf{x}_i = \mathbf{0}.$$

A calibration-based estimator of $\boldsymbol{\alpha}$ in the form of (4.2), with $\mathcal{S}_B$ replaced by $\mathcal{S}_{B,0}$, can also be used. In this latter case we have $\hat{N}_0^A = \sum_{i \in \mathcal{S}_A} (\hat{\pi}_{i0}^A)^{-1} = \hat{N}_0^B$ if $\mathbf{x}$ contains 1 as the first component. Let $\hat{m}_i = m(\mathbf{x}_i, \hat{\boldsymbol{\beta}})$, where $\hat{\boldsymbol{\beta}}$ is an estimator of $\boldsymbol{\beta}$ obtained by using suitable estimation method and the non-probability sample dataset $\{(y_i, \mathbf{x}_i), i \in \mathcal{S}_A\}$ under the assumed outcome regression model $E_\xi(y_i \mid \mathbf{x}_i) = m(\mathbf{x}_i, \boldsymbol{\beta})$. The doubly robust estimator of μ_{y0} is computed as

$$\hat{\mu}_{y0,\mathrm{DR}} = \frac{1}{\hat{N}_0^B} \sum_{i \in \mathcal{S}_A} \frac{y_i - \hat{m}_i}{\hat{\pi}_{i0}^A} + \frac{1}{\hat{N}_0^B} \sum_{i \in \mathcal{S}_{B,0}} d_i^B \hat{m}_i. \tag{4.4}$$

Note that we used $\hat{N}_0^B$ instead of $\hat{N}_0^A$ in the first term. It leads to a simplified form of the hybrid estimator given below. Let

$$\hat{\mu}_{y1,\mathrm{REG}} = \frac{1}{\hat{N}_1^B} \sum_{i \in \mathcal{S}_{B,1}} d_i^B \hat{m}_i \tag{4.5}$$

be the model-based prediction estimator for μ_{y1}. A hybrid estimator of $\mu_y = W_0\mu_{y0} + W_1\mu_{y1}$ is constructed by using the two estimators given in (4.4) and (4.5):

$$\hat{\mu}_{y\mathrm{HYB}} = \hat{W}_0\hat{\mu}_{y0,\mathrm{DR}} + \hat{W}_1\hat{\mu}_{y1,\mathrm{REG}} = \frac{1}{\hat{N}^B} \sum_{i \in \mathcal{S}_A} \frac{y_i - \hat{m}_i}{\hat{\pi}_{i0}^A} + \frac{1}{\hat{N}^B} \sum_{i \in \mathcal{S}_B} d_i^B \hat{m}_i. \tag{4.6}$$

The form of $\hat{\mu}_{y\text{HYB}}$ is similar to the doubly robust estimator $\hat{\mu}_{y\text{DR}}$ given in (2.5), with the major difference of estimating the propensity scores through the split population.

The hybrid estimator does not have the double robustness interpretation and has the risk of extrapolation in estimating μ_{y1} using the model-based prediction estimator $\hat{\mu}_{y1,\text{REG}}$. Suppose that the split of $\mathcal{S}_B$ into $\mathcal{S}_{B,0}$ and $\mathcal{S}_{B,1}$ using the convex hull formulation on $\mathcal{U}_0$ and $\mathcal{U}_1$ is done correctly, and the propensity score model on $\pi_i^A > 0$ for $i \in \mathcal{U}_0$ is correctly specified, then the hybrid estimator $\hat{\mu}_{y\text{HYB}}$ has the asymptotic expression

$$\hat{\mu}_{y\text{HYB}} = W_0 \mu_{y0} + W_1 \, \bar{m}_1^* + o_p(1),$$

where $\bar{m}_1^* = N_1^{-1} \sum_{i \in \mathcal{U}_1} m(\mathbf{x}_i, \boldsymbol{\beta}^*)$ and $\boldsymbol{\beta}^*$ satisfies $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^* = O_p(n_A^{-1/2})$, regardless of the correctness of the outcome regression model (Chen, 2020). The potential bias of the hybrid estimator depends largely on the model-based prediction estimator $\hat{\mu}_{y1,\text{REG}}$ for estimating μ_{y1}, and the estimator has the advantage of the doubly robust estimator $\hat{\mu}_{y0,\text{DR}}$ for estimating μ_{y0}.

Under ideal situations where it is possible to take a small subsample from $\mathcal{S}_{B,1}$ and obtain measurements on y for the selected units, a rigorous development on estimation methods for μ_{y1}, and consequently for μ_y, can be carried out without much difficulties. Let $\{(y_i, d_{2i}^B), i \in \mathcal{S}_{B,1}^{(2)}\}$ be the additional dataset where $\mathcal{S}_{B,1}^{(2)}$ is a subsample from $\mathcal{S}_{B,1}$ and d_{2i}^B are the sampling weights for the subsample conditional on the given $\mathcal{S}_{B,1}$. Let $\tilde{\boldsymbol{\beta}}$ be the estimated parameters for the outcome regression model using the combined dataset $\{(y_i, \mathbf{x}_i), i \in \mathcal{S}_A \cup \mathcal{S}_{B,1}^{(2)}\}$. Let $\tilde{m}_i = m(\mathbf{x}_i, \tilde{\boldsymbol{\beta}})$. A model-assisted difference estimator (Wu and Sitter, 2001) for the subpopulation mean μ_{y1} can be constructed as

$$\hat{\mu}_{y1,\text{SS}} = \frac{1}{\hat{N}_1^B} \sum_{i \in \mathcal{S}_{B,1}^{(2)}} d_{2i}^B d_i^B \, (y_i - \tilde{m}_i) + \frac{1}{\hat{N}_1^B} \sum_{i \in \hat{\mathcal{S}}_{B,1}} d_i^B \, \tilde{m}_i, \tag{4.7}$$

where the subscript "SS" indicates "subsample". This estimator is approximately unbiased for μ_{y1} under the survey designs for $\mathcal{S}_B$ and $\mathcal{S}_{B,1}^{(2)}$ regardless the correctness of the outcome regression model. The final estimator of μ_y can then be computed as $\hat{\mu}_{y\text{SS}} = \hat{W}_0 \hat{\mu}_{y0,\text{DR}} + \hat{W}_1 \hat{\mu}_{y1,\text{SS}}$. The estimator $\hat{\mu}_{y\text{SS}}$ is doubly robust and is given by

$$\hat{\mu}_{y\text{SS}} = \frac{1}{\hat{N}^B} \left\{ \sum_{i \in \mathcal{S}_A} \frac{y_i - \tilde{m}_i}{\hat{\pi}_{i0}^A} + \sum_{i \in \mathcal{S}_{B,1}^{(2)}} d_{2i}^B d_i^B \, (y_i - \tilde{m}_i) + \sum_{i \in \mathcal{S}_B} d_i^B \, \tilde{m}_i \right\}. \tag{4.8}$$

5. Simulation studies

We evaluate the finite sample performances of several estimation strategies with deterministic undercoverage. Additional simulation results under stochastic undercoverage can be found in Chen (2020). We consider a finite population of size $N =$ 20,000, with three auxiliary variables x_1, x_2 and x_3. Independent copies of (x_{i1}, x_{i2}, x_{i3}) are generated from $x_{i1} \sim N(0,1)$, $x_{i2} \sim \text{Exp}(1)$, and $x_{i3} \sim \text{Bernoulli}(0.5)$. The response variable y_i follows the regression model,

$$y_i = 3 + x_{i1} + x_{i2} + x_{i3} - \eta x_{i1}^2 + \sigma\varepsilon_i, \quad i = 1, 2, \ldots, N, \tag{5.1}$$

where η is the coefficient for the high order term x_{i1}^2 with values representing the degree of departure from the standard linear regression model. The error terms ε_i are generated from $N(0,1)$ and the value of σ is chosen to control the correlation coefficient ρ between the response y_i and the linear predictor $3 + x_{i1} + x_{i2} + x_{i3} - \eta x_{i1}^2$. The simulation results reported in this section correspond to $\rho = 0.5$. The parameter of interest is the finite population mean μ_y.

The accessibility function $\Phi(\mathbf{x})$ is specified through an inverse logit function involving the three auxiliary variables,

$$\log\left\{\frac{\Phi(\mathbf{x}_i)}{1-\Phi(\mathbf{x}_i)}\right\} = 1 - 0.6x_{i1} + 0.5x_{i2} + 0.8x_{i3}, \quad i = 1, 2, \ldots, N.$$

Note that the inverse logit function is not a convex function. However, it can be shown that the function is convex within the subspace $\{\mathbf{x}: \Phi(\mathbf{x}) < 0.5\}$, which is sufficient for our proposed split population approach with the choices $\tau =$ 0.00, 0.20 and 0.40 used in the simulation. Let $Q(\tau)$ be the $100\tau^{\text{th}}$ sample quantile of $\{\Phi(\mathbf{x}_1), \Phi(\mathbf{x}_2), \ldots, \Phi(\mathbf{x}_N)\}$. The deterministic split of the population is decided by the measure on accessibilities. Let $\mathcal{U} = \{1, 2, \ldots, N\}$ and

$$\mathcal{U}_0 = \{i \mid i \in \mathcal{U} \text{ and } \Phi(\mathbf{x}_i) \geq Q(\tau)\}, \quad \mathcal{U}_1 = \{i \mid i \in \mathcal{U} \text{ and } \Phi(\mathbf{x}_i) < Q(\tau)\}.$$

We set $\pi_i^A = 0$ if $i \in \mathcal{U}_1$, and the size of the subpopulation $\mathcal{U}_1$ is given by $N_1 = \tau N$ (i.e., $W_1 = \tau$). The size of the subpopulation $\mathcal{U}_0$ is given by $N_0 = N - N_1$.

The true propensity scores π_i^A for $i \in \mathcal{U}_0$ are generated from a logistic regression model,

$$\log\left(\frac{\pi_i^A}{1-\pi_i^A}\right) = \theta + 0.3x_{i1} - 0.3x_{i2} + 0.5x_{i3},$$

where the intercept θ is chosen such that $\sum_{i=1}^{N_0} \pi_i^A = n_A$, where n_A is the planned size of the non-probability survey sample $\mathcal{S}_A$. We use the Poisson sampling method with inclusion probabilities π_i^A to select units for $\mathcal{S}_A$ from the subpopulation $\mathcal{U}_0$. The actual sample size of $\mathcal{S}_A$ varies from sample to sample with n_A as the expected size.

The reference probability sample $\mathcal{S}_B$ with a fixed sample size n_B is drawn from the entire finite population $\mathcal{U}$ by the randomized systematic PPS sampling method; see Section 4.4.2 of Wu and Thompson (2020) for further detail. The inclusion probabilities π_i^B are proportional to $z_i = c + x_{i2}$, where the constant c is chosen to control the variation of the survey weights such that $\max_{i \in \mathcal{U}} z_i / \min_{i \in \mathcal{U}} z_i = 50$.

We consider six estimators of the population mean μ_y discussed in Sections 2 and 4, plus the naive estimator of the sample mean of the non-probability sample, and evaluate their performances through repeated simulation samples:

(1) $\hat{\mu}_{yA}$, the naive estimator of the sample mean of the non-probability sample $\mathcal{S}_A$;

(2) $\hat{\mu}_{y\text{IPW}}$, the IPW estimator given in (2.3);

(3) $\hat{\mu}_{y\mathrm{MI}}$, the model-based prediction (mass imputation) estimator given in (2.4);

(4) $\hat{\mu}_{y\mathrm{DR}}$, the doubly robust estimator given in (2.5);

(5) $\hat{\mu}^{C}_{y\mathrm{IPW}}$, the calibrated IPW estimator described in Section 4.1;

(6) $\hat{\mu}_{y\mathrm{HYB}}$, the hybrid estimator given in (4.6) introduced in Section 4.4;

(7) $\hat{\mu}_{y\mathrm{SS}}$, the estimator specified in (4.8) using a subsample $\mathcal{S}^{(2)}_{B,1}$ from $\mathcal{S}_{B,1}$.

For each iteration of the simulation with samples $\mathcal{S}_A$ and $\mathcal{S}_B$, the working propensity score model for the estimators $\hat{\mu}_{y\mathrm{IPW}}$ and $\hat{\mu}_{y\mathrm{DR}}$ is chosen as $\log\{\pi^A_i/(1-\pi^A_i)\} = \alpha_0 + \alpha_1 x_{i1} + \alpha_2 x_{i2} + \alpha_3 x_{i3}$, the working outcome regression model for $\hat{\mu}_{y\mathrm{MI}}$ and $\hat{\mu}_{y\mathrm{DR}}$ as well as $\hat{\mu}_{y\mathrm{HYB}}$ uses $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i$. The splitting of $\mathcal{S}_B$ into $\mathcal{S}_{B,0}$ and $\mathcal{S}_{B,1}$ is done through the convex hull method. The rates of correct identification of units belonging to $\mathcal{U}_0$ and $\mathcal{U}_1$ are higher than 97% for all the settings used in the simulation. The subsample $\mathcal{S}^{(2)}_{B,1}$ is selected from $\mathcal{S}_{B,1}$ using simple random sampling without replacement and the sampling fraction is fixed at 20%. Noting that for $n_B = 500$ used in the simulation, the size of the subsample is around 20 for $\tau = W_1 = 0.2$ and 40 for $\tau = 0.4$.

The amount of misspecification of the working regression model is reflected by the value of η used in the true model (5.1) for generating the finite population $\{(y_i, x_{i1}, x_{i2}, x_{i3}),\ i = 1, 2, \ldots, N\}$. The deterministic undercoverage is represented by the value of $\tau = W_1$. We consider $3 \times 3 = 9$ different settings for the simulation, with the true values of the population and the subpopulation means $(\mu_y, \mu_{y0}, \mu_{y1})$ given in Table 5.1. The setting $(\eta = 0.0,\ \tau = 0.0)$ represents the ideal situation of no model misspecifications and no issues with undercoverage.

Table 5.1
Population and Subpopulation Means $(\mu_y, \mu_{y,1}, \mu_{y,0})$.

	$\tau = 0.0$	$\tau = 0.2$	$\tau = 0.4$
$\eta = 0.0$	(4.53, NA, NA)	(4.53, 4.49, 4.72)	(4.53, 4.53, 4.52)
$\eta = 0.5$	(4.03, NA, NA)	(4.03, 4.06, 3.91)	(4.03, 4.06, 3.97)
$\eta = 1.0$	(3.52, NA, NA)	(3.52, 3.63, 3.11)	(3.52, 3.59, 3.42)

The performance of an estimator $\hat{\mu}_y$ is measured by the simulated Relative Bias (RB%, in percentage) and the simulated Mean Squared Error (MSE), which are computed as

$$\mathrm{RB\%} = 100\left(\frac{1}{B}\sum_{b=1}^{B}\frac{\hat{\mu}_y^{(b)} - \mu_y}{\mu_y}\right),\ \ \mathrm{MSE} = \frac{1}{B}\sum_{b=1}^{B}(\hat{\mu}_y^{(b)} - \mu_y)^2,$$

where $\hat{\mu}_y^{(b)}$ is the estimator $\hat{\mu}_y$ computed from the b^{th} simulation samples. Results for $n_A = 1{,}000$ and $n_B = 500$ based on $B = 5{,}000$ simulation runs are presented in Table 5.2. The values of MSE are multiplied by 100.

Table 5.2
Simulated RB% and MSE ($\times 10^2$) of Seven Estimators of μ_y.

	Estimator	$\tau = 0$		$\tau = 0.2$		$\tau = 0.4$	
		RB%	MSE	RB%	MSE	RB%	MSE
$\eta = 0.0$	$\hat{\mu}_{yA}$	17.23	61.90	18.47	70.99	21.79	98.48
	$\hat{\mu}_{y\text{IPW}}$	-0.09	1.67	0.00	1.39	0.66	1.64
	$\hat{\mu}_{y\text{MI}}$	-0.03	1.55	-0.04	1.45	0.36	1.70
	$\hat{\mu}_{y\text{DR}}$	-0.01	1.59	0.00	1.47	0.39	1.75
	$\hat{\mu}^{C}_{y\text{IPW}}$	0.00	1.58	-0.01	1.49	0.49	2.37
	$\hat{\mu}_{y\text{HYB}}$	0.00	1.59	0.04	1.48	0.42	1.72
	$\hat{\mu}_{y\text{SS}}$	NA	NA	-0.03	4.83	0.08	7.14
$\eta = 0.5$	$\hat{\mu}_{yA}$	16.91	47.40	23.56	91.03	27.82	126.59
	$\hat{\mu}_{y\text{IPW}}$	-0.13	2.19	3.44	3.60	4.16	4.77
	$\hat{\mu}_{y\text{MI}}$	2.82	2.88	3.71	3.81	4.83	5.75
	$\hat{\mu}_{y\text{DR}}$	-0.10	2.02	3.39	3.57	4.71	5.73
	$\hat{\mu}^{C}_{y\text{IPW}}$	0.01	1.91	3.85	4.10	5.79	8.11
	$\hat{\mu}_{y\text{HYB}}$	-0.02	2.03	2.14	2.57	3.66	4.31
	$\hat{\mu}_{y\text{SS}}$	NA	NA	-1.10	6.61	-0.65	9.39
$\eta = 1.0$	$\hat{\mu}_{yA}$	16.53	35.44	30.17	114.40	35.68	159.43
	$\hat{\mu}_{y\text{IPW}}$	-0.19	3.34	7.87	10.08	8.70	12.19
	$\hat{\mu}_{y\text{MI}}$	6.48	7.58	8.53	11.14	10.63	16.81
	$\hat{\mu}_{y\text{DR}}$	-0.23	3.35	7.76	9.84	10.35	16.39
	$\hat{\mu}^{C}_{y\text{IPW}}$	0.03	2.83	8.82	11.99	12.69	23.80
	$\hat{\mu}_{y\text{HYB}}$	-0.04	3.45	4.85	5.67	7.91	10.96
	$\hat{\mu}_{y\text{SS}}$	NA	NA	-2.49	11.79	-1.60	16.17

The simulation results can be summarized as follows. The naive estimator $\hat{\mu}_{yA}$ using the sample mean from the non-probability sample is biased under all the settings and is not included in any further comparisons with other six estimators. (1) All five estimators (the sixth estimator $\hat{\mu}_{y\text{SS}}$ using a subsample is not applicable) under the setting of no model misspecification and no undercoverage (i.e., $\eta = 0.0$ and $\tau = 0.0$) perform well with no biases and similar MSEs; (2) Without issues of undercoverage (i.e., $\tau = 0.0$), the model-based prediction estimator $\hat{\mu}_{y\text{MI}}$ starts to show biases as the outcome regression model is misspecified (e.g., RB% = 6.48 for $\eta = 1.0$), while the other four estimators show no biases with similar small MSEs, including the hybrid estimator $\hat{\mu}_{y\text{HYB}}$ under the split population approach; (3) The estimators $\hat{\mu}_{y\text{MI}}$, $\hat{\mu}_{y\text{DR}}$, $\hat{\mu}_{y\text{HYB}}$ using the correctly specified outcome regression model (i.e., $\eta = 0.0$) show no biases and similar small MSEs in the presence of undercoverage (i.e., $\tau = 0.2$ or 0.4). The calibrated IPW estimator $\hat{\mu}^{C}_{y\text{IPW}}$, which requires a linear outcome regression model to justify, also shows no biases with undercoverage; (4) When the outcome regression model is misspecified (i.e., $\eta = 0.5$ or 1.0) and there is an undercoverage problem (i.e., $\tau = 0.2$ or 0.4), all five estimators (excluding the last one $\hat{\mu}_{y\text{SS}}$), which rely on the correctness of one of the two working models, demonstrate clear biases and deteriorated MSEs; (5)

The estimator $\hat{\mu}_{y\mathrm{SS}}$, which uses additional information on y through a subsample from $\mathcal{S}_{B,1}$, shows negligible biases for all scenarios but the values of MSE are larger than several other estimators. Part of the reasons is the very small size of the subsample $\mathcal{S}_{B,1}^{(2)}$ used in the simulation since the estimator $\hat{\mu}_{y\mathrm{SS}}$ involves the key component $\hat{\mu}_{y1,\mathrm{SS}}$ given in (4.7), and the latter has variance depending on the size of $\mathcal{S}_{B,1}^{(2)}$.

6. Concluding remarks

The undercoverage problem with non-probability survey samples is closely attached to issues with modelling on propensity scores, and parametric models usually fail without the positivity assumption. In practice, most non-probability samples do not represent the entire target population, rendering propensity score based weighting methods invalid under such scenarios. Model-based prediction approach is sensitive to model specification, and the quality of the model on the response variable y depends largely on the auxiliary variables which are available in both the non-probability sample and the reference probability sample. If the analyst is confident with the prediction model, the model-based prediction approach can be reliable in dealing with undercoverage problems when the support (i.e., the range) of each auxiliary variable in the non-probability sample matches the one from the probability sample. Otherwise there is a risk of extrapolation leading to biased estimation. From the theoretical view point, the deterministic undercoverage is a consequence of the violation of the positivity assumption **A2**. It leads to issues with fitting the outcome regression model using the non-probability sample data since $E(y \mid \mathbf{x}, R=1) = E(y \mid \mathbf{x})$ implicitly requires $P(R=1) \neq 0$ even if **A1** holds. This is why the calibrated IPW estimator may have some advantages under a linear outcome regression model since the estimation of the model parameters $\boldsymbol{\beta}$ is not required.

The undercoverage problem is intrinsically related to the sample selection and participation mechanism for non-probability samples, which can be further complicated by the so-called non-ignorable selection bias. Dealing with non-ignorable selection bias for non-probability samples is itself an active research topic and has been investigated in several recent publications; see, for instance, Andridge, West, Little, Boonstra and Alvarado-Leiton (2019), Boonstra, Little, West, Andridge and Alvarado-Leiton (2021), and West, Little, Andridge, Boonstra, Ware, Pandit and Alvarado-Leiton (2021), among others. Sensitive analysis and quantitative measures on selection bias developed in these papers can be valuable tools for dealing with undercoverage problems.

The split population approach has been used in survey sampling to analyze and combine data from different sources; see, for instance, Zhang (2019) for further discuss and related references. Our proposed convex hull formulation in splitting the population into two subpopulations shows some potential in dealing with undercoverage problems, but a complete removal of biases in estimation after the split requires additional information on one of the subpopulations. The modified nearest neighbour method of Kim and Rao (2018) for splitting the target population seems to be a promising idea and may deserve some conscious efforts in future research.

Another important topic which is not addressed in the paper is on variance estimation under the strategies discussed in the paper. We are currently undertaking a separate research project on variance estimation and we hope to report our progresses in the near future.

The literature on missing data and causal inference includes methodological developments in dealing with the impact of very small but positive estimated propensity scores on the estimation of the main parameters through inverse probability weighting. Some of these developments may be useful for addressing undercoverage problems with non-probability samples, such as the stable weights approach of Zubizarreta (2015). It is hoped that discussions presented in this paper will add insights to the growing field of data integration and combining data from multiple sources.

Acknowledgements

This research was supported by grants from the Natural Sciences and Engineering Research Council of Canada and a collaborative research team grant from the Canadian Statistical Sciences Institute. Part of the materials presented in this paper was taken from a chapter of the first author's doctoral dissertation completed at the University of Waterloo under the joint supervision of the other two co-authors.

References

Andridge, R.R., West, B.T., Little, R.J., Boonstra, P.S. and 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*, 68, 1465-1483.

Beaumont, J.-F. (2020). Are probability surveys bound to disappear for the production of official statistics? *Survey Methodology*, 46, 1, 1-28. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2020001/article/00001-eng.pdf.

Boonstra, P.S., Little, R.J., West, B.T., Andridge, R.R. and Alvarado-Leiton, F. (2021). A simulation study of diagnostics for selection bias. *Journal of Official Statistics*, 37, 751-769.

Chen, Y. (2020). *Statistical Analysis with Non-probability Survey Samples*, PhD Dissertation, Department of Statistics and Actuarial Science, University of Waterloo.

Chen, Y., Li, P. and Wu, C. (2020). Doubly robust inference with non-probability survey samples. *Journal of the American Statistical Association*, 115, 2011-2021.

Chen, Y., Li, P., Rao, J.N.K. and Wu, C. (2022). Pseudo empirical likelihood inference for non-probability survey samples. *The Canadian Journal of Statistics*, to appear.

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

Kim, J.K., and Rao, J.N.K. (2018). Data integration for big data analysis in finite population inference. Paper presented at the 2018 Annual Meeting of the Statistical Society of Canada.

Kim, J.K., Park, S., Chen, Y. and Wu, C. (2021). Combining non-probability and probability survey samples through mass imputation. *Journal of the Royal Statistical Society, Series A*, 184, 941-963.

Rao, J.N.K. (2021). On making valid inferences by integrating data from surveys and other sources. *Sankhyā B*, 83, 242-272.

Rivers, D. (2007). Sampling for Web surveys. *Joint Statistical Meetings, Proceedings of the Survey Research Methods Section*, American Statistical Association, Alexandria, VA, 1-26.

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

Tan, Z. (2007). Comment on "Understanding OR, PS and DR". *Statistical Science*, 22, 560-568.

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

West, B.T., Little, R.J., Andridge, R.R., Boonstra, P.S., Ware, E.B., Pandit, A. and 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, 1556-1581.

Wisniowski, A., Sakshaug, J.W., Ruiz, D.A.P. and Blom, A.G. (2020). Integrating probability and nonprobability samples for survey inference. *Journal of Survey Statistics and Methodology*, 8, 120-147.

Wu, C. (2022). Statistical inference with non-probability survey samples. *Survey Methodology*, 48, 2, 283-311. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2022002/article/00002-eng.pdf.

Wu, C., and Sitter, R.R. (2001). A model-calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association*, 96, 185-193.

Wu, C., and Thompson, M.E. (2020). *Sampling Theory and Practice*. Cham: Springer.

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

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