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

Survey Methodology

Use of nonprobability samples for official statistics, state of the art

by Danny Pfeffermann and Michael Sverchkov

Release date: June 30, 2025

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, the Agency has developed standards of service which its employees observe in serving its clients. 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, June 2025
Vol. 51, No. 1, pp. 169-196
Statistics Canada, Catalogue No. 12-001-X

Use of nonprobability samples for official statistics, state of the art

Danny Pfeffermann and Michael Sverchkov[1]

Abstract

Tightened budgets, continuing decrease of response rates in traditional probability surveys and increasing pressure by users for more timely data, has stimulated research on the use of nonprobability sample data, such as administrative records, web scraping, mobile phone data and voluntary internet surveys, for inference on finite population parameters like means and totals. These data are often easier, faster and cheaper to collect than traditional probability samples. However, a major concern with the use of this kind of data for official statistics is their nonrepresentativeness due to possible selection bias, which if not accounted for properly, could bias the inference. In this article, we review and discuss methods considered in the literature to deal with this problem and propose new methods, distinguishing between methods based on integration of the nonprobability sample with an appropriate probability sample, and methods that base the inference solely on the nonprobability sample. Empirical illustrations, based on simulated data are provided.

Key Words: Empirical likelihood; Probability and nonprobability samples; Sample integration; Selection bias.

1. Introduction

Tightened budgets, continuing decrease in response rates, due in part by increased response burden in traditional probability surveys and privacy concerns, and increasing pressure by users for more timely data, has prompted research into the use of nonprobability sample data, such as administrative records, web scraping, mobile telephone data, online panels and voluntary internet surveys for inference on finite population characteristics. These data are often easier, faster and cheaper to collect than are traditional probability samples. However, a major concern with the use of this kind of data is their possible nonrepresentativeness, due to possible selection bias, which if not accounted for properly, could bias the inference. For example, house sales advertised on the internet do not represent properly all house sales. Web scraping for job vacancies does not represent all job vacancies. Data from social media do not generally represent the general public. All these examples can be considered as “big data”, but nonprobability samples do not need to be big. Baker, Brick, Bates, Battaglia, Couper, Dever, Gile and Tourangeau (2013), Keiding and Louis (2016) and Elliott and Valliant (2017) discuss other potential problems with the use of nonprobability samples for inference on finite population parameters.

The basic definition of a probability sample is that every unit in the population has a positive probability of being included in the sample. Inference under the traditional randomization (design-based) distribution over all possible sample selections from a fixed target population requires that the first-order sample selection probabilities of the sampled units are known. The use of standard variance estimation procedures requires that the joint sample selection probabilities of the sampled units are also known, but these can be

1. Danny Pfeffermann, Hebrew University of Jerusalem, Israel & University of Southampton, UK. E-mail: msdanny@soton.ac.uk; Michael Sverchkov, Bureau of Labor Statistics, Washington DC, USA. E-mail: Sverchkov.Michael@bls.gov.

calculated to a desired approximation by repeated sampling from the sampling frame. (This is not usually available to analysts outside National Statistical Offices-NSOs.)

By definition, nonprobability samples are not selected by use of probability sampling schemes, so no selection probabilities exist. The question arising therefore is how to draw inference from such samples, regarding the population, which they are supposed to represent. In this article, we restrict our attention to inference about target population parameters such as totals or means (proportions), which are the most common target parameters in official statistics, often published in tables.

We mention in this respect that many survey statisticians claim that traditional probability samples should be replaced by external records. Citro (2014) states that "official statistical offices need to move from the probability sample survey paradigm for the past 75 years to a mixed mode data source paradigm for the future". Clearly, if the nonprobability sample data are timely, accurate, with good coverage and contain all the required information, there is no reason to select a corresponding probability sample.

However, this is seldom the case. Israel's population register covers all the population residing in Israel, but about 15% of the home addresses are wrong. Tax records of businesses are often obtained with a delay of up to 2 years. No administrative data are available on opinions, sentiments, detailed expenditures, and many other variables of interest. We also mention in this regard that government and private agencies are often reluctant to transfer data to NSOs, claiming data protection issues. Furthermore, the desired information is often contained in more than one file, requiring matching them, which is problematic if personal identifiers are unknown. (It requires probabilistic algorithms based on information in all the records.) Coverage of records might be different and may not apply to same time periods. Definitions and accuracy of information may differ between records. Finally, matching of different administrative data could magnify problems of data protection.

Methods considered in the literature to deal with possible non-representativeness of nonprobability (NP) samples can be divided into two classes:

1. Integration of the NP sample with an appropriate probability sample (PS),
2. Consideration of the NP sample on its own (no data integration).

Remark 1. The methods considered in this article for inference from NP samples alone assume known population means of some of the survey values, which are used for enhancing the inference. However, no detailed probability sample data are used.

In Section 2, we review several methods proposed in the literature for integration of a NP sample with an appropriate PS sample. We also present a new method. Section 3 reviews methods proposed for adjusting for selection bias of a NP sample without integration with a PS sample. In Section 4, we propose a new method for inference from a NP sample without integration with a PS sample. Section 5 contains simulation results illustrating the performance of our proposed method. We conclude with some summary remarks in Section 6.

2. Integration of nonprobability and probability samples

One of the earliest articles on this topic is by Lee (2006). The author proposes to create a pooled sample $S_P = S_{\mathrm{PS}} \cup S_{\mathrm{NP}}$ from the probability sample S_{PS} and the nonprobability sample S_{NP}, assuming implicitly that the two samples do not overlap, and models the selection probability to the nonprobability sample. The S_{NP} sample is treated as a "treatment sample" in observational studies, and the S_{PS} sample is treated as the "control sample". It is assumed that every unit in the population has a positive probability to be in the S_{NP} sample, estimated by use of propensity scores, $e(\mathbf{x}_j) = \Pr(j \in S_{\mathrm{NP}} \mid \mathbf{x}_j;\ j = 1, \ldots, n)$, where n is the size of the S_p sample and the $\mathbf{x}$- variables are assumed to be measured in both samples.

Next, the S_P sample is divided into C classes based on the ascending values of the estimated propensity scores. An adjustment factor f_C is computed for every class c as,

$$f_c = \frac{\sum_{k \in S_{\mathrm{PS}}^c} d_{k,\mathrm{PS}} \Big/ \sum_{k \in S_{\mathrm{PS}}} d_{k,\mathrm{PS}}}{\sum_{j \in S_{\mathrm{NP}}^c} d_{j,\mathrm{NP}} \Big/ \sum_{j \in S_{\mathrm{NP}}} d_{j,\mathrm{NP}}}, \tag{2.1}$$

where $d_{k,\mathrm{PS}}$ and $d_{j,\mathrm{NP}}$ are some base weights. An adjusted weight $d_{j,\mathrm{NP}}^A = f_c d_{j,\mathrm{NP}}$ is computed for every unit $j \in S_{\mathrm{NP}}$.

The estimator of the target population total $Y = \sum_{i \in U} Y_i$ is, $\hat{Y}_{S_{\mathrm{NP}}} = \sum_c \sum_{j \in S_{\mathrm{NP}}^c} d_{j,\mathrm{NP}}^A y_j$.

The use of this procedure for data integration requires the existence of $\mathbf{x}$- variables such that the assignment to S_{NP} and the target y- variable are independent given $\mathbf{x}$, $\Pr(j \in S_{\mathrm{NP}} \mid \mathbf{x}_j, y_j;\ j \in S_P) = \Pr(j \in S_{\mathrm{NP}} \mid \mathbf{x}_j;\ j \in S_P)$. This is a limiting assumption. An extensive empirical study revealed that the use of this approach decreases (but not eliminates) the bias of inference from the S_{NP} sample, but increases the variance. See also Beaumont (2020).

Kott and Ridenhour (2024) likewise consider the use of a pooled sample $S_P = S_{\mathrm{PS}} \cup S_{\mathrm{NP}}$ for inference from the nonprobability sample. The authors model the S_{NP} selection probabilities by a logistic model with covariates $\mathbf{z}_k$ measured in both samples and for which the true population means $\mathbf{T}_Z$ are known or estimated from the S_{PS} sample, which are used for calibration. The estimating equation is $\sum_{k \in S_{\mathrm{NP}}} [1 + \exp(\mathbf{z}_k' \mathbf{g})]\, \mathbf{z}_k = \mathbf{T}_Z\, (\hat{\mathbf{T}}_Z)$. This defines new weights $w_k = \pi_k^{-1}[1 + \exp(\mathbf{z}_k' \hat{\mathbf{g}})]$ used for inference from the S_{NP} sample, where $\pi_k = \Pr(k \in S_{\mathrm{PS}})$. When the S_{PS} sample is exposed to nonresponse, the weights $d_k = \pi_k^{-1}$ are adjusted to account for the nonresponse.

Rivers (2007) considers the case where $\mathbf{x}$ and y are measured in the S_{NP} sample but only $\mathbf{x}$ is measured in the S_{PS} sample. The author proposes to deal with the non-representativeness of the S_{NP} sample by matching to every unit $i \in S_{\mathrm{PS}}$ an element k from S_{NP}, with similar values of auxiliary (matching) variables $\mathbf{x}$.

Denote by $\mathbf{x}_i, i = 1, \ldots, n$, the $\mathbf{x}$- vectors in S_{PS} and by $\tilde{\mathbf{x}}_j$ the vectors in S_{NP}. The unit $k \in S_{\mathrm{NP}}$ satisfying $|\tilde{\mathbf{x}}_k - \mathbf{x}_i| \le |\tilde{\mathbf{x}}_j - \mathbf{x}_i|\ \forall j \in S_{\mathrm{NP}}$ is chosen as the matched element for unit $i \in S_{\mathrm{PS}}$, where $|\cdot|$ is an appropriate distance metric. Selecting a matching element for every unit $i \in S_{\mathrm{PS}}$ defines a matched sample $\tilde{S}_{\mathrm{PS}}$ of size n with y- values from the S_{NP} sample.

The proposed estimator of the population total Y is $\hat{Y}_{\text{SM}} = \sum_{k \in \tilde{S}_{\text{PS}}} w_k \tilde{y}_k$, where $w_k = (1/\pi_k)$; $\pi_k = \Pr(k \in S_{\text{PS}})$ and $\{\tilde{y}_k\}$ are the y- values measured in S_{NP}, not measured in S_{PS}. The author establishes regularity conditions under which for a scalar continuous matching variable, as $n \to \infty$, $n_{\text{NP}} \to \infty$ and $n / n_{\text{NP}} \to 0$, (n_{NP} is the size of S_{NP}), $n^{-0.5}(\hat{Y}_{\text{SM}} - Y)/N$ converges to a normal distribution with mean zero, where N is the population size.

Remark 2. Rather than matching one record, one can match k nearest records and select at random the matched record out of the k records, known as the kNN method. See, e.g., Conti, Marella and Scanu (2008). Alternatively, a weighted mean of the y- values of the nearest records can be used for matching.

Remark 3. The method requires a PS sample with similar $\mathbf{x}$ values in S_{NP} and S_{PS}. It also assumes that $f_{S_{\text{NP}}}(y_i \mid \mathbf{x}_i) = f_U(y_i \mid \mathbf{x}_i)$, implying $\Pr(i \in S_{\text{NP}} \mid \mathbf{x}_i, y_i) = \Pr(i \in S_{\text{NP}} \mid \mathbf{x}_i)$, where $f_{S_{\text{NP}}}(y \mid \mathbf{x})$ is the conditional distribution in the S_{NP} sample and $f_U(y \mid \mathbf{x})$ is the conditional distribution in the population. See Yang, Kim and Hwang (2021) for other assumptions and related theoretical properties of matching methods.

Kim and Wang (2019) propose the following procedure of integrating the data in the S_{PS} and S_{NP} samples. The authors assume that membership of the S_{PS} elements in S_{NP} is known. Let $\delta_i = 1(0)$ if $i \in S_{\text{NP}}$ ($i \notin S_{\text{NP}}$). The S_{PS} data contains therefore the values $\{(\mathbf{x}_i, \delta_i);\ i = 1, \ldots, n\}$. The procedure consists of the following step:

1. Model $p_i(\boldsymbol{\gamma}) = \Pr(\delta_i = 1 \mid \mathbf{x}_i; \boldsymbol{\gamma})$ by use of the S_{PS} data and estimate $\boldsymbol{\gamma}$ by maximizing the "pseudo likelihood" $l(\boldsymbol{\gamma}) = \sum_{i \in S_{\text{PS}}} w_i \{\delta_i \log p_i(\boldsymbol{\gamma}) + (1 - \delta_i) \log[1 - p_i(\boldsymbol{\gamma})]\}$.
2. Estimate the population total Y as,

$$\hat{Y}_{S_{\text{NP}}}(1) = \sum_{i \in S_{\text{NP}}} p_i^{-1}(\hat{\boldsymbol{\gamma}}) y_i \quad \text{or} \quad \hat{Y}_{S_{\text{NP}}}(2) = N \sum_{i \in S_{\text{NP}}} p_i^{-1}(\hat{\boldsymbol{\gamma}}) y_i \Big/ \sum_{i \in S_{\text{NP}}} p_i^{-1}(\hat{\boldsymbol{\gamma}}) \tag{2.2}$$

when N is known.

The authors consider also a doubly robust estimator under the assumption of a population regression model. Consistent variance estimators are developed.

Remark 4. This method again assumes that the sampling mechanism to S_{NP} is ignorable after controlling for the covariates, i.e. $\Pr(i \in S_{\text{NP}} \mid \mathbf{x}_i, y_i) = \Pr(i \in S_{\text{NP}} \mid \mathbf{x}_i)$, often referred to as missing at random (MAR) selection. In addition, the assumption that membership of the S_{PS} elements in S_{NP} is known, may not hold in practice.

Chen, Li and Wu (2020) likewise assume noninformative sampling after controlling for the covariates and assume a selection model $\pi_i^{S_{\text{NP}}} = \pi(\mathbf{x}_i; \boldsymbol{\gamma}) = \Pr(i \in S_{\text{NP}} \mid \mathbf{x}_i; \boldsymbol{\gamma})$, which is estimated by maximizing the pseudo loglikelihood

$$l^*(\boldsymbol{\gamma}) = \sum_{i \in S_{\text{NP}}} \log\left[\frac{\pi(\mathbf{x}_i, \boldsymbol{\gamma})}{1 - \pi(\mathbf{x}_i, \boldsymbol{\gamma})}\right] + \sum_{i \in S_{\text{PS}}} w_i \log[1 - \pi(\mathbf{x}_i, \boldsymbol{\gamma})], \tag{2.3}$$

where $w_i = 1/\pi_i$ are the sampling weights in S_{PS}. The authors consider 2 estimators of the population mean $\bar{Y} = \frac{1}{N}\sum_{i \in U} y_i$,

$$\hat{\bar{Y}}_{\text{IPW1}} = \frac{1}{N}\sum\nolimits_{i \in S_{\text{NP}}} \frac{y_i}{\pi(\mathbf{x}_i;\hat{\boldsymbol{\gamma}})} \quad \text{or} \quad \hat{\bar{Y}}_{\text{IPW2}} = \frac{1}{\hat{N}}\sum\nolimits_{i \in S_{\text{NP}}} \frac{y_i}{\pi(\mathbf{x}_i;\hat{\boldsymbol{\gamma}})};\ \hat{N} = \sum\nolimits_{i \in S_{\text{NP}}} [1/\pi(\mathbf{x}_i, \hat{\boldsymbol{\gamma}})], \tag{2.4}$$

depending on whether the population size is known or unknown.

The authors prove that for the case of a logistic selection model, both estimators have an error of order $O_P(n_{S_{NP}}^{-1/2})$. Variance estimators are also developed, correct to order $o(n_{S_{\text{NP}}}^{-1})$.

Remark 5. In a rejoinder to comments on an article by Beaumont, Bosa, Brennan, Charlebois and Chu (2024a) (see below), Beaumont, Bosa, Brennan, Charlebois and Chu (2024b) argue that the use of the likelihood (2.3) is not efficient because the second term only uses the S_{PS} data and ignores relevant S_{NP} auxiliary data. The authors propose an improved estimator of $\boldsymbol{\gamma}$ and a sample likelihood approach that properly accounts for an overlap between the two samples, when it can be identified.

Chen et al. (2020) also consider a doubly robust estimator, defined as

$$\hat{\bar{Y}}_{\text{DR}} = \frac{1}{N}\left\{\sum\nolimits_{i \in S_{\text{NP}}} [1/\pi(\mathbf{x}_i.\hat{\gamma})]\,[y_i - m_i(\mathbf{x}_i, \hat{\boldsymbol{\beta}}] + \sum\nolimits_{i \in S_{\text{PS}}} w_i m_i(\mathbf{x}_i, \hat{\boldsymbol{\beta}}]\right\}, \tag{2.5}$$

where $m_i(\mathbf{x}_i, \boldsymbol{\beta})$ is an assumed population regression model. When N is unknown, the estimator is modified by dividing the first term by $\hat{N}_{S_{\text{NP}}} = \sum_{i \in S_{\text{NP}}} [1/\pi(\mathbf{x}_i, \hat{\boldsymbol{\gamma}})]$ and the second term by $\hat{N}_{S_{\text{PS}}} = \sum_{i \in S_{\text{PS}}} (1/\pi_i)$. The estimators are shown to be consistent for $\bar{Y}$, even if the population model or the sample selection model are misspecified. Variance estimators correct to order $o(n_{S_{\text{NP}}}^{-1})$ are derived under some additional conditions.

Chen, Li, Rao and Wu (2022) consider the use of the pseudo empirical loglikelihood for inference from nonprobability samples, defined as $l_{\text{PEL}}(\mathbf{p}) = \sum_{i \in S_{\text{NP}}} d_i^{S_{\text{NP}}} \log(p_i)$, where the p_i's are the EL probabilities and $d_i^{S_{\text{NP}}} = [1/\pi(\mathbf{x}_i.\hat{\boldsymbol{\gamma}})] \big/ \hat{N}_{\text{SNP}}$. The parameters $\boldsymbol{\gamma}$ are estimated using the likelihood (2.3) and are considered fixed in the likelihood $l_{\text{PEL}}(\mathbf{p})$. Maximization of the likelihood under the constraint $\sum_{i \in S_{\text{NP}}} p_i = 1$ yields $\hat{p}_i = d_i^{S_{\text{NP}}}$.

The authors also develop a doubly robust estimator, similar to (2.5), obtained by adding the calibration constraint $\sum_{i \in S_{\text{NP}}} p_i [m_i(\mathbf{x}_i;\hat{\boldsymbol{\beta}})] = \hat{N}_{S_{\text{PS}}}^{-1} \sum_{i \in S_{\text{PS}}} w_i\, m_i(\mathbf{x}_i;\hat{\boldsymbol{\beta}})$, and corresponding pseudo empirical likelihood confidence intervals, which are shown to perform generally better than the customary normal theory intervals.

We refer the readers also to a related article by Wu (2022), which contains a critical review and some extended discussions on theoretical and practical issues with inference from non-probability samples.

Beaumont et al. (2024a) likewise consider integration of S_{NP} and S_{PS} samples, again assuming that the probability of inclusion in the S_{NP} sample only depends on $\mathbf{x}$. The authors assume a logistic model $p_i(\boldsymbol{\gamma}) = \Pr(\delta_i = 1 \mid \mathbf{x}_i;\boldsymbol{\gamma})$ for the inclusion of unit $i \in U$ in S_{NP} and estimate $\boldsymbol{\gamma}$ by solving the likelihood estimating equations $U(\boldsymbol{\gamma}) = \sum_{i \in S_{\text{NP}}} \mathbf{x}_i - \sum_{i \in S_{\text{PS}}} w_i p_i(\boldsymbol{\gamma})\,\mathbf{x}_i = 0$. The equations $U(\boldsymbol{\gamma})$ are design unbiased over all possible S_{PS} selections of the likelihood equations that would be obtained if the $\mathbf{x}$- values were known for all $i \in U$.

The authors develop a modified AIC criterion for stepwise selection of the $\mathbf{x}$- variables in the S_{NP} sample selection model $p_i(\boldsymbol{\gamma})$. However, a problem with the use of this criterion is that it ignores the relationship between y and the $\mathbf{x}$- variables. To deal with this problem, the authors extend their AIC criterion by partitioning the S_{NP} sample into homogeneous groups $S_{\text{NP}} = S_{\text{NP},1} \cup, \ldots, \cup S_{\text{NP},G}$ based on the estimated probabilities $p_i(\hat{\boldsymbol{\gamma}})$ and a ranking method, and then assigning each unit in the S_{PS} sample to one of the groups. Let $S_{\text{NP},g}$ and $S_{\text{PS},g}$ define the gth sets of units of the S_{NP} and S_{PS} samples, respectively. Assuming that the selection probabilities in each group are the same, the resulting estimated selection probabilities in group g are $\hat{p}_g = n_g^{\text{NP}} / \hat{N}_g$, where n_g^{NP} is the size of $S_{\text{NP},g}$ and $\hat{N}_g = \sum_{k \in S_{\text{PS},g}} w_k$. The estimator of Y is

$$\hat{Y}_{S_{NP}} = \sum_{k \in S_{\text{NP}}} \hat{w}_k^{\text{NP}} y_k = \sum_{g=1}^{G} \hat{N}_g \bar{y}_{S_{\text{NP},g}}; \; \hat{w}_k^{\text{NP}} = \hat{N}_g / n_g^{\text{NP}}, \, k \in S_{\text{NP},g}; \; \bar{y}_{S_{\text{NP},g}} = \sum_{i \in S_{\text{NP},g}} \frac{y_i}{n_g^{\text{NP}}}. \tag{2.6}$$

The variance of $\hat{Y}_{S_{\text{NP}}}$ is estimated by an appropriate bootstrap algorithm.

Remark 6. Rao (2021) reviews several other estimators based on data integration, distinguishing between the case where the target variable y is observed in both samples, and the case where it is only observed in the S_{NP} sample.

The common feature of all the approaches considered so far is their reliance on the assumption that the selection to the S_{NP} sample depends on known $\mathbf{x}$- variables, but not on the target y- variable. (See Remark 4 above). In practice, it is likely that the selection to S_{NP} depends also on y. For example, people participating in a voluntary web survey on political tendency, may choose not to participate in the survey, depending on their tendency. Administrative data may be missing people who do not participate in government programs, including people who do not have social security numbers, people with housing instability, or people working in the informal economy.

In addition, the S_{PS} sample used for integration with the S_{NP} sample may be subject to not missing at random (NMAR) nonresponse, in the sense that that the probability to respond depends also on the target y- variable. For example, the response of people on income may depend on their level of income. Denote by R_i the response indicator. NMAR nonresponse occurs when,

$$\Pr[R_i = 1 \mid y_i, \mathbf{x}_i, i \in s] \neq \Pr[R_i = 1 \mid \mathbf{x}_i, i \in s]. \tag{2.7}$$

Pfeffermann, Marella and Summa (2025a) consider data integration when the selection to the S_{NP} sample and the response probabilities in the S_{PS} sample depend on both y and $\mathbf{x}$, applying the empirical likelihood (EL) approach. It is assumed that $\mathbf{x}$ is observed in both samples, but y is only observed in the S_{NP} sample. Let I_i^{PS} be the sample indicator for S_{PS}, taking the value 1 if unit i is sampled and 0 otherwise. For $i \in S_{\text{PS}}$, the sample model of $\mathbf{x}_i$ is

$$p_{i,\text{PS}}^{X} = \Pr(\mathbf{x}_i \mid I_i^{\text{PS}} = 1) = \frac{\Pr(I_i^{\text{PS}} = 1 \mid \mathbf{x}_i)}{\Pr(I_i^{\text{PS}} = 1)} p_i^X, \tag{2.8}$$

where $p_i^X = \Pr_U(\mathbf{x} = \mathbf{x}_i)$ is the probability in the population. As can be seen, under informative sampling with respect to $\mathbf{x}$, the sample probability $p_{i,\mathrm{PS}}^X$ is different from p_i^X.

Additionally, it is assumed that the S_{PS} sample is exposed to NMAR nonresponse. Let R_i^{PS} be the response indicator, taking the value 1 if sample unit $i \in S_{\mathrm{PS}}$ responds and 0 otherwise. Denote by R_{PS} the set of responding units in S_{PS}. Then,

$$p_{i,R_{\mathrm{PS}}}^X = \Pr(\mathbf{x}_i \mid I_i^{\mathrm{PS}} = 1, R_i^{\mathrm{PS}} = 1) = \frac{\Pr(R_i^{\mathrm{PS}} = 1 \mid \mathbf{x}_i, I_i^{\mathrm{PS}} = 1)}{\Pr(R_i^{\mathrm{PS}} = 1 \mid I_i^{\mathrm{PS}} = 1)}\, p_{i,\mathrm{PS}}^X. \tag{2.9}$$

By (2.8) and (2.9), the respondents model is a function of the true population probability, the conditional expectations of the sampling weights, $\Pr(I_i^{\mathrm{PS}} = 1 \mid \mathbf{x}_i) = 1 / E_{\mathrm{PS}}(w_{i,\mathrm{PS}} \mid \mathbf{x}_i)$ (Pfeffermann and Sverchkov, 1999); $w_{i,\mathrm{PS}} = 1/\pi_{i,\mathrm{PS}}$ are the base sampling weights in S_{PS}, and the response probabilities $\Pr(R_i^{\mathrm{PS}} = 1 \mid \mathbf{x}_i,$ $I_i^{\mathrm{PS}} = 1)$. Assuming that the response is independent of the sample selection, $E_{\mathrm{PS}}(w_{i,\mathrm{PS}} \mid \mathbf{x}_i) = E_{R_{\mathrm{PS}}}(w_{i,\mathrm{PS}} \mid \mathbf{x}_i)$, in which case the probabilities $\Pr(I_i^{\mathrm{PS}} = 1 \mid \mathbf{x}_i)$ can be estimated by regressing $w_{i,\mathrm{PS}}$ against $\mathbf{x}_i$, using the data in R_{PS}.

The response probabilities $\Pr(R_i^{\mathrm{PS}} = 1 \mid \mathbf{x}_i, I_i^{\mathrm{PS}} = 1)$ in (2.9) are unknown and need to be estimated from the available data by postulating a parametric model,

$$\Pr(R_i^{\mathrm{PS}} = 1 \mid \mathbf{x}_i, I_i^{\mathrm{PS}} = 1, \boldsymbol{\rho}) = g(\mathbf{x}_i; \boldsymbol{\rho}) \tag{2.10}$$

for some known function g, (say, a logistic model), with $\boldsymbol{\rho}$ defining the model parameters.

Assuming independence of the sampling and the response, the *empirical respondents' likelihood* based on R_{PS} is thus,

$$\mathrm{ERL}_{R_{\mathrm{PS}}}\{p_i^X\} = \prod\nolimits_{i \in R_{\mathrm{PS}}} p_{i,R_{\mathrm{PS}}}^X = \prod_{i \in R_{\mathrm{PS}}} \frac{\Pr(R_i^{\mathrm{PS}} = 1 \mid x_i, I_i^{\mathrm{PS}} = 1)}{\Pr(R_i^{\mathrm{PS}} = 1 \mid I_i^{\mathrm{PS}} = 1)}\, \frac{\Pr(I_i^{\mathrm{PS}} = 1 \mid x_i)}{\Pr(I_i^{\mathrm{PS}} = 1)}\, p_i^X. \tag{2.11}$$

Next, consider the S_{NP} sample. Let I_i^{NP} be the sample indicator, taking the value 1 if $i \in S_{\mathrm{NP}}$ and 0 otherwise. Denote $p_i^{XY} = \Pr(\mathbf{x} = \mathbf{x}_i, y = y_i)$. For $i \in S_{\mathrm{NP}}$,

$$p_{i,\mathrm{NP}}^{XY} = \Pr(\mathbf{x}_i, y_i \mid I_i^{\mathrm{NP}} = 1) = \frac{\Pr(I_i^{\mathrm{NP}} = 1 \mid \mathbf{x}_i, y_i)}{\Pr(I_i^{\mathrm{NP}} = 1)}\, p_i^{XY}, \tag{2.12}$$

where $\Pr(I_i^{\mathrm{NP}} = 1) = \sum_{i \in S_{\mathrm{NP}}} \Pr(I_i^{\mathrm{NP}} = 1 \mid \mathbf{x}_i, y_i)\, p_i^{XY}$. Because no sampling weights for S_{NP} are available, the probabilities $P(I_i^{\mathrm{NP}} = 1 \mid \mathbf{x}_i, y_i)$ need to be modelled parametrically,

$$\Pr(I_i^{\mathrm{NP}} = 1 \mid \mathbf{x}_i, y_i; \boldsymbol{\gamma}) = h(y_i, \mathbf{x}_i; \boldsymbol{\gamma}) \tag{2.13}$$

for some known function h, with $\boldsymbol{\gamma}$ defining the model parameters. Assuming independence of the S_{NP} data, the *empirical likelihood* based on S_{NP} is

$$\mathrm{ESL}_{\mathrm{NP}}(p_i^{XY}) = \prod\nolimits_{i \in S_{\mathrm{NP}}} p_{i,\mathrm{NP}}^{XY}. \tag{2.14}$$

Assuming no overlap between the two samples, the *empirical likelihood* based on the data in S_{NP} and S_{PS} is,

$$\text{EL}_{R_{\text{PS}}\cup\text{NP}} = \text{ERL}_{\text{PS}}(p_i^X)\,\text{ESL}_{\text{NP}}(p_i^{XY}) = \prod_{i\in R_{\text{PS}}} p_{i,R_{\text{PS}}}^X \prod_{i\in S_{\text{NP}}} p_{i,\text{NP}}^{XY}. \tag{2.15}$$

The unknown parameters in (2.15) are the population probabilities p_i^X, p_i^{XY}, the sampling parameters $\boldsymbol{\gamma}$ and the response parameters $\boldsymbol{\rho}$. The likelihood is maximized subject to normalizing constraints on the unknown probabilities and calibration constraints.

Remark 7. The unknown probabilities $\{p_i^X\}$ can also be estimated from the S_{NP} sample; $\hat{p}_{i,\text{NP}}^X = \sum_{\{i;x=x_i\}} \hat{p}_{i,\text{NP}}^{XY}$. This implies two sets of estimates of the probabilities $\{p_i^X\}$, which need to be harmonized. See Marella and Pfeffermann (2023) for possible harmonization procedures. The final, integrated estimate of p_i^{XY} is $\hat{p}_i^{XY} = \hat{p}_i^X(\hat{p}_{i,\text{NP}}^{XY} / \hat{p}_{i,\text{NP}}^X)$, where $\hat{p}_i^X$ is the harmonized estimator.

The population total Y can be estimated in one of the following two ways:

$$\hat{Y}_{\text{NP}}(1) = N\sum_{i\in S_{\text{NP}}} y_i \hat{p}_i^Y;\ \hat{Y}_{\text{NP}}(2) = N\frac{\sum_{i\in\text{NP}} \hat{\Pr}^{-1}(I_i^{\text{NP}}=1\,|\,\mathbf{x}_i, y_i)\,y_i}{\sum_{i\in\text{NP}} \hat{\Pr}^{-1}(I_i^{\text{NP}}=1\,|\,\mathbf{x}_i, y_i)}, \tag{2.16}$$

where $\hat{p}_i^y = \sum_{i;y=y_i} \hat{p}_i^{XY}$. See Pfeffermann et al. (2025a) for an empirical comparison of the performance of the two estimators.

Remark 8. One of the reviewers of this article raised a concern about the model used for the selection model to the S_{NP} sample, noting that it seems difficult to obtain robustness to deviations from the model. As discussed in Section 4.3 and illustrated in Section 5, the S_{NP} model can be tested.

3. Inference from a nonprobability sample without integration

In Section 2, we considered methods of inference from a nonprobability sample, based on integration of the S_{NP} sample with an appropriate probability sample S_{PS}. In this section, we consider methods for adjusting the selection bias of the S_{NP} sample, without integration with a S_{PS} sample (see Remark 1).

We start with an approach based on calibration. The basic idea underlying this approach is to change some base weights, $d_{j,\text{NP}}$ to new weights $d_{j,\text{NP}}^{\text{cal}}$, so that when applied to a set of variables Z observed in S_{NP} and for which the true population totals are known, the S_{NP} survey estimates will equal the corresponding totals; $\sum_{j\in S_{\text{NP}}} d_{j,\text{NP}}^{\text{cal}} \mathbf{z}_j = \mathbf{T}_z$, where $\mathbf{T}_z$ are the known population totals. (In practice, the true totals can be replaced by reliable estimates from a probability sample, in which case it can be considered as "sample integration".) See AAPOR (2010) and Baker et al. (2013) for review of methods that follow this approach, and Kott and Ridenhour (2024) reviewed in Section 2.

The success of this approach depends on the availability of calibration variables, which are highly correlated with the target y- variable (good prediction power). Lee and Valliant (2009) illustrate that combining propensity scores and calibration adjustments is more effective in reducing the bias of S_{NP} estimates than using just one of the approaches. See also Elliott and Valliant (2017).

Kim and Wang (2019) propose the use of inverse sampling to obtain a representative sample from the finite population, and hence to correct for the selection bias of the S_{NP} sample. The proposed inverse sampling can be viewed as a special case of two-phase sampling, where the first phase is the S_{NP} sample and the second phase is a subsample from the first-phase sample to correct for the selection bias.

Denote, as before, by δ_i the indicator of whether unit $i \in U$ is included in the S_{NP} sample. It is assumed that $\Pr(\delta_i = 1 \mid y_i, \mathbf{x}_i) = \Pr(\delta_i = 1 \mid \mathbf{x}_i) > 0$ for all $i \in U.$ The S_{NP} sample contains the values $(y_i, \mathbf{x}_i),\ i \in S_{\text{NP}}.$ Denote by $f(\mathbf{x})$ the population distribution of the $\mathbf{x}$- variables. If $f(\mathbf{x})$ is known, an asymptotic unbiased estimator of $\theta = E(Y)$ is,

$$\hat{\theta}_{S_{\text{NP1}}} = \sum_{i \in S_{\text{NP}}} \frac{f(\mathbf{x}_i)}{f(\mathbf{x}_i \mid \delta_i = 1)} y_i \Bigg/ \sum_{i \in S_{\text{NP}}} \frac{f(\mathbf{x}_i)}{f(\mathbf{x}_i \mid \delta_i = 1)} = \sum_{i \in S_{\text{NP}}} w_{i1} y_i. \tag{3.1}$$

For the more practical case where only the mean $\bar{\mathbf{X}}_U = \sum_{i \in U} \mathbf{x}_i / N$ is known, the authors approximate $f(\mathbf{x})$ by the function $f_0(\mathbf{x}),$ which minimizes the Kullback–Leibler distance. The solution to the minimization distance is,

$$f_0(\mathbf{x}) = f(\mathbf{x} \mid \delta = 1) \frac{\exp(\mathbf{x}'\boldsymbol{\lambda})}{E[\exp(\mathbf{x}'\boldsymbol{\lambda} \mid \delta = 1)]}, \text{ with } \boldsymbol{\lambda} \text{ satisfying } \int \mathbf{x} f_0(\mathbf{x})\, d\mathbf{x} = \bar{\mathbf{X}}_U. \tag{3.2}$$

With this approximation, the estimator $\hat{\theta}_{S_{\text{NP1}}}$ in (3.1) is replaced by,

$$\hat{\theta}_{S_{\text{NP2}}} = \sum_{i \in S_{\text{NP}}} w_i^* y_i;\ w_i^* = \frac{\exp(\mathbf{x}_i'\hat{\boldsymbol{\lambda}})}{\sum_{i \in S_{\text{NP}}} \exp(\mathbf{x}_i'\hat{\boldsymbol{\lambda}})}, \text{ with } \hat{\boldsymbol{\lambda}} \text{ satisfying } \sum_{i \in S_{\text{NP}}} w_i^* \mathbf{x}_i = \bar{\mathbf{X}}_U. \tag{3.3}$$

Finally, the authors propose to select the second-phase sample from S_{NP} with probabilities $\pi_{i2|1} = n w_i^*,$ $i \in S_{\text{NP}}$ with the weights $\{w_i^*\}$ defined by (3.3) and $n \le [\max_{i \in S_{\text{NP}}} \{w_i^*\}]^{-1},$ yielding the approximately design-unbiased estimator of the $\hat{\theta}_{S_{\text{NP1}}}$ estimator defined in (3.1),

$$\hat{\theta}_{S_{\text{NP3}}} = \sum_{i \in S_{\text{NP}}} \frac{1}{\pi_{i2|1}} w_i^* y_i = \frac{1}{n} \sum_{i=1}^{n} y_i. \tag{3.4}$$

A simple estimator of the design variance of $\hat{\theta}_{S_{\text{NP3}}}$ is proposed.

The two approaches considered so far assume that the selection to the S_{NP} sample is MAR, in the sense that $\Pr(\delta_i = 1 \mid y_i, \mathbf{x}_i) = \Pr(\delta_i = 1 \mid \mathbf{x}_i) > 0$ for all $i \in U.$ However, as discussed before, this assumption may not hold and in what follows, we consider alternative approaches aimed to deal with the case of informative sample selection.

Sayag, Ben-Hur and Pfeffermann (2022) consider the following problem, underlying the computation of monthly house price indices (HPI) in many countries. A large amount of the house sales are reported several months after they occur, implying that if not accounted for, the provisional HPIs based on the on-time reported transactions are subject to large revisions, as further transactions are reported. This happens because the late-reported transactions behave differently from the transactions reported on time. This is a nice example of a nonprobability sample (the on-time reported sales), which is subject to selection bias due to late data availability of some of the sales (~40% in Israel).

To deal with this problem, the authors propose nowcasting three types of variables and adding them as input data to the hedonic regression model used for the computation of the HPI: (1)- the average characteristics of the upcoming late-reported transactions, such as the average number of rooms, the average net area size, the average age of the sold houses, etc. (2)- the average price of the late-reported transactions and (3)- the number of late-reported transactions. The three types of variables are nowcasted based on simple models fitted to data from previous months. Application of the proposed methodology shows more than 50% reduction in the magnitude of the revisions. This is a unique example of a time series of non-representative nonprobability samples for which the true population data (all the sales corresponding to a given month) become known only several months later.

Kim and Morikawa (2023) consider a non-ignorable (informative) sample selection model $\pi_i(y_i, \mathbf{x}_i; \boldsymbol{\phi}) = \Pr(\delta_i = 1 \mid y_i, \mathbf{x}_i; \boldsymbol{\phi})$, where $\delta_i = (1,0)$ is the S_{NP} sample indicator, assuming that the variables $\mathbf{x}_i$ are known for all $i \in U$ and $\pi_i(y_i, \mathbf{x}_i) > 0$ for all $i \in U$. For the case where the population model $f_U(y_i \mid \mathbf{x}_i)$ is known, the authors propose estimating $\boldsymbol{\phi}$ by maximizing the likelihood,

$$L_{\text{obs}}(\boldsymbol{\phi}) = \prod\nolimits_{i \in U} [f_U(y_i \mid \mathbf{x}_i)\, \pi(y_i, \mathbf{x}_i; \boldsymbol{\phi})]^{\delta_i} [1 - \tilde{\pi}(\mathbf{x}_i; \tilde{\boldsymbol{\phi}})]^{(1-\delta_i)}; \; \tilde{\pi}(\mathbf{x}_i; \tilde{\boldsymbol{\phi}}) \; = \; \mathrm{E}[\pi(y_i, \mathbf{x}_i; \boldsymbol{\phi}) \mid \mathbf{x}_i]. \tag{3.5}$$

However, this likelihood requires modelling the population model and the authors note that the MLE estimator obtained from (3.5) is not robust to misspecification of the model. Consequently, they develop a likelihood based on the model $f_{\text{SNP}}(y_i \mid \mathbf{x}_i) = f(y_i \mid \mathbf{x}_i, \delta_i = 1)$, which can be identified and estimated consistently.

Alternatively, the authors develop a methodology for estimating $\boldsymbol{\phi}$ and the population mean of the y-values by applying the empirical likelihood (EL) approach. For the case where the selection probabilities $\pi_i(y_i, \mathbf{x}_i)$ are known, the authors propose estimating the p_i's underlying the EL by maximizing the loglikelihood, $l(p) = \sum_{i \in S_{\text{NP}}} \log(p_i)$, subject to the constraints (1)- $\sum_{i \in S_{\text{NP}}} p_i = 1$, (2)- $\sum_{i \in S_{\text{NP}}} p_i \pi_i(y_i, \mathbf{x}_i) = n / N$, (3)- $\sum_{i \in S_{\text{NP}}} p_i \mathbf{x}_i = \bar{\mathbf{X}}_U$, where n is the size of the S_{NP} sample, N is the population size and $\bar{\mathbf{X}}_U = \sum_{i \in U} \mathbf{x}_i / N$. The constraint (2) is referred to as a bias calibration constraint, whereas the constraint (3) is added to improve the efficiency of EL estimator.

In practice, the sample selection probabilities are unknown. The authors assume a parametric model; $\pi_i(y_i, \mathbf{x}_i) = g(y_i, \mathbf{x}_i^*; \boldsymbol{\phi})$ (say, logistic, $\mathbf{x}_i^*$ is a subset of $\mathbf{x}_i$ to guarantee model identifiability, see Sections 4 and 5), and estimate $\hat{\pi}_i(y_i, \mathbf{x}_i) = g(y_i, \mathbf{x}_i^*; \hat{\boldsymbol{\phi}})$ by solving the estimating equations $\sum_{i=1}^{N} \left[\frac{\delta_i}{g(y_i, \mathbf{x}_i^*; \boldsymbol{\phi})} - 1 \right] \mathbf{x}_i = \mathbf{0}$. These equations do not require knowledge of the $\mathbf{x}$- values for every unit in the population. By considering the estimated probabilities $\hat{\pi}_i(y_i, \mathbf{x}_i) = g(y_i, \mathbf{x}_i^*; \hat{\boldsymbol{\phi}})$ as the true selection probabilities, the authors maximize the constrained EL likelihood defined above, with the bias calibration constraint (2) replaced by $\sum_{i \in S_{\text{NP}}} p_i g(y_i, \mathbf{x}_i^*; \hat{\boldsymbol{\phi}}) = N^{-1} \sum_{i=1}^{N} g(y_i, \mathbf{x}_i^*; \hat{\boldsymbol{\phi}})$, which does require knowledge of the population $\mathbf{x}^*$- values, yielding the estimates $\{\hat{p}_i\}$. The population mean of the y- values are estimated as,

$$\hat{\bar{Y}}_{\text{EL, IPW}} = \frac{1}{N} \sum_{i \in S_{\text{NP}}} \frac{y_i}{\pi_i(\mathbf{x}_i^*; \hat{\boldsymbol{\phi}})} \quad \text{or} \quad \hat{\bar{Y}}_{\text{EL}} = \sum_{i \in S_{\text{NP}}} \hat{p}_i y_i. \tag{3.6}$$

The authors derive asymptotic properties of their estimators and variance estimators.

This article proposes a novel approach for estimating finite population means from S_{NP} samples subject to nonignorable selection probabilities, but the assumption that the $\mathbf{x}$- variables are known for every unit in the population is restrictive.

Remark 9. In Section 2, we proposed a method of inference from a S_{NP} sample alone, which likewise combines a non-ignorable sample selection model with the empirical likelihood. See equations (2.12)-(2.14). This method does not require knowledge of the $\mathbf{x}$- values for every unit in the population. See also Section 4 below.

4. A new (old) approach for inference from a nonprobability sample

4.1 Relationship between the population distribution and the S_{NP} distribution

In the following, we propose an alternative approach for inference from a nonprobability sample alone. It relies in large on Pfeffermann and Sverchkov (1999).

Denote the model holding for the target variable y in U by $f_U(y_i \mid \mathbf{x}_i)$. Denote the model holding for y in the S_{NP} sample by $f_{S_{\text{NP}}}(y_i \mid \mathbf{x}_i)$, and let $\delta_i = 1(0)$ if $i \in S_{\text{NP}} \, (i \notin S_{\text{NP}})$. The target model is $f_U(y_i \mid \mathbf{x}_i)$, but observations $\{y_i, \mathbf{x}_i\}$ are only available for $f_{S_{\text{NP}}}(y_i \mid x_i)$. We assume, $\Pr(i \in S_{\text{NP}}) > 0$ for all $i \in U$ (also assumed in the other approaches considered before). The two distributions are connected via the link function $\Pr(\delta = 1 \mid y, \mathbf{x})$.

$$f_{S_{\text{NP}}}(y_i \mid \mathbf{x}_i) = f(y_i \mid \mathbf{x}_i, \delta_i = 1) \overset{\text{Bayes}}{=} \frac{\Pr(\delta_i = 1 \mid \mathbf{x}_i, y_i) f_U(y_i \mid \mathbf{x}_i)}{\Pr(\delta_i = 1 \mid \mathbf{x}_i)}. \tag{4.1}$$

As discussed below, the relationship (4.1) enables estimating the target population distribution from the observations in S_{NP} alone. Notice that $f_{S_{\text{NP}}}(y_i \mid \mathbf{x}_i) = f_U(y_i \mid \mathbf{x}_i)$ iff $\Pr(\delta_i = 1 \mid \mathbf{x}_i, y_i) = \Pr(\delta_i = 1 \mid \mathbf{x}_i) \; \forall y_i$, in which case the model fitted based on the S_{NP} sample holds for the population data and if the $\mathbf{x}$- values are known for all $i \in U$, (or in the case of a linear population model $\bar{\mathbf{X}}_U$ is known), inference based on the S_{NP} sample is valid. See Rao (2021) for discussion of this method under these conditions.

Remark 10. In the first part of their article, Kim and Morikawa (2023) also assume parametric models for the population model and the sample selection probabilities (see above), but we do not assume knowledge of the population $\mathbf{x}$- values. Additionally, the authors estimate the parameters underlying the sample selection model outside the likelihood, whereas we estimate them jointly with the population model parameters (see below). We utilize similar calibration constraints to the ones used by Kim and Morikawa (2023), see equation (4.3) below. We also test the goodness of fit of the resulting model $f_{S_{\text{NP}}}(y_i \mid \mathbf{x}_i)$, see Section 4.3.

The probabilities $\Pr(\delta_i = 1 \mid \mathbf{x}_i, y_i)$ need to be modelled. They are allowed to depend on the target y variable, thus accounting for informative sample selection. They may depend also on other variables $\mathbf{z}$, but we only need to model $\Pr(\delta_i = 1 \mid \mathbf{x}_i, y_i)$. The use of a Logistic model for δ_i has some theoretical justification. See Lemma 1 in Pfeffermann, Preminger and Sikov (2025b) for details. When $\mathbf{z}$ is observed in the S_{NP} sample, we may include it among the $\mathbf{x}$- variables.

4.2 Estimation of model parameters

Unlike the use of the empirical likelihood approach, application of this approach requires specifying the population model and the model for the sample selection probabilities, which depend on unknown parameters that need to be estimated from the observations in the S_{NP} sample. Adding parameters to (4.1), and assuming $\Pr(\delta_i = 1 \mid y_i, \mathbf{x}_i; \boldsymbol{\phi}) = \Pr(\delta_i = 1 \mid y_i, \mathbf{x}_i^*; \boldsymbol{\phi})$, with $\mathbf{x}_i^*$ denoting a subset of the vector $\mathbf{x}_i$ to guarantee the identifiability of the model (see Section 4.3), we have

$$f_{S_{\text{NP}}}(y_i \mid \mathbf{x}_i; \boldsymbol{\beta}, \boldsymbol{\phi}) = \frac{\Pr(\delta_i = 1 \mid y_i, \mathbf{x}_i^*; \boldsymbol{\phi}) f_U(y_i \mid \mathbf{x}_i; \boldsymbol{\beta})}{\Pr(\delta_i = 1 \mid \mathbf{x}_i^*; \boldsymbol{\phi}, \boldsymbol{\beta})}. \tag{4.2}$$

Assuming independence of the observations in S_{NP}, the corresponding log likelihood is $l_{S_{\text{NP}}}(\boldsymbol{\phi}, \boldsymbol{\beta}; y) = \sum_{i \in S_{\text{NP}}} \log f_{S_{\text{NP}}}(y_i \mid \mathbf{x}_i; \boldsymbol{\beta}, \boldsymbol{\phi})$, which we maximize subject to the constraints,

$$\frac{1}{N} \sum_{i \in S_{\text{NP}}} \frac{1}{\Pr(\delta_i = 1 \mid y_i, \mathbf{x}_i^*; \boldsymbol{\phi})} \mathbf{x}_i = \frac{1}{N} \sum_{j \in U} \mathbf{x}_j = \bar{\mathbf{X}}_U. \tag{4.3}$$

The constraints (4.3) are used for enhancing the estimation of the parameters $(\boldsymbol{\beta}, \boldsymbol{\phi})$. We assume throughout that the $\mathbf{x}$ and $\mathbf{x}^*$ vectors contain a "1" in the first position.

Remark 11. In the empirical study in Section 5 with continuous y, we approximated the probabilities $\Pr(\delta_i = 1 \mid \mathbf{x}_i^*; \boldsymbol{\phi}, \boldsymbol{\beta})$ by Riemann's sums over 350 sub-groups of the y- values. When y is binary,

$$\begin{aligned}\Pr(\delta_i = 1 \mid \mathbf{x}_i^*; \boldsymbol{\phi}, \boldsymbol{\beta}) = {} & \Pr(\delta_i = 1 \mid y_i = 1, \mathbf{x}_i^*; \boldsymbol{\phi}) \Pr(y_i = 1 \mid \mathbf{x}_i; \boldsymbol{\beta}) \\ & + \Pr(\delta_i = 1 \mid y_i = 0, \mathbf{x}_i^*; \boldsymbol{\phi}, \boldsymbol{\beta}) \Pr(y_i = 0 \mid \mathbf{x}_i; \boldsymbol{\beta}).\end{aligned}$$

We maximized the likelihood with the constraints by use of the SAS procedure NLIN, iterating between the maximization with respect to $\boldsymbol{\phi}$ for given $\boldsymbol{\beta}$, and the maximization of $\boldsymbol{\beta}$ for given $\boldsymbol{\phi}$, with the "given" values defined by the estimates in the previous iteration. See Section 5 for how we estimated the population mean of the y- values in our simulations.

4.3 Model testing and identifiability conditions

The application of the proposed approach assumes a model $f_U(y_i \mid \mathbf{x}_i; \boldsymbol{\beta})$ for the population values and a model $\Pr(\delta_i = 1 \mid y_i, \mathbf{x}_i^*; \boldsymbol{\phi})$ for the selection probabilities, which permits estimating the parameters $(\boldsymbol{\phi}, \boldsymbol{\beta})$ by means of (4.2) and (4.3), using the data in S_{NP}. No direct testing of the population model or the model for the selection probabilities is possible, since no data are available from the population distribution and the y- values are unknown for units $j \notin S_{\mathrm{NP}}$. However, contrary to a common perception that it is impossible to test a model fitted to the S_{NP} data, we contend this is not true. We have observations from the fitted model, so we are faced with the classical problem of testing the goodness of fit of a hypothesized model to the observed data. See Krieger and Pfeffermann (1997) and Pfeffermann and Sikov (2011) for plausible tests.

Remark 12. Rejection of the null hypothesis that the model fits the data implies that at least one of the two models is misspecified. See Section 5 for examples and the concluding remarks in Section 6.

A common argument in favor of the claim that the S_{NP} model cannot be tested is that it may be the case that there is more than one combination of a population model and a selection model, yielding the same model for the observed data, such that the model fitted to the S_{NP} data is not identifiable or "practically not identifiable". Pfeffermann and Landsman (2011) and Wang, Shao and Kim (2014) establish conditions under which the model $f_{S_{\mathrm{NP}}}(y_i \mid \mathbf{x}_i)$ is identifiable, with references to other related studies. See Section 5 for the identifiability conditions of the models considered in the simulation study.

Remark 13. In a highly cited article, Molenberghs, Beunckens and Kenward (2008) prove and illustrate that for every NMAR model fitted to a set of data, there is a MAR counterpart providing exactly the same fit to the data. The authors note that "such a construction does not lead to a member of a conventional parametric family". A simple example for this argument is where the population model $f_U(y \mid \mathbf{x})$ is assumed to be defined by the sample model $f_{S_{\mathrm{NP}}}(y_i \mid \mathbf{x}_i)$ (equation 4.2), and the sample inclusion probability satisfies $\Pr(\delta_i = 1 \mid y_i, \mathbf{x}_i^*; \boldsymbol{\phi}) = \Pr(\delta_i = 1 \mid \mathbf{x}_i^*; \boldsymbol{\phi})$. Clearly, $f_U(y_i \mid \mathbf{x}_i) = f_{S_{\mathrm{NP}}}(y_i \mid \mathbf{x}_i)$ defined by (4.2) is a very odd population distribution. Molenberghs et al. (2008) also note that "we can make progress if attention is confined to a given parametric family, in which we put sufficiently strong prior belief". This is what we do under our proposed approach. Notice that the selection model is used to obtain valid estimates of the population model, and as shown below and illustrated in Section 5, it can be tested.

Consider first the case where y is a continuous variable. In our empirical applications, we applied the following UNIF test statistic (Krieger and Pfeffermann, 1997).

Preliminaries:

1. For a continuous variable Z with cumulative distribution F, $F(z) \sim U(0,1)$.
2. Under general conditions, the set of all the moments of $F(z)$ determines the distribution.

Proposed test:

(i) Compute $T_i = F_{S_{\text{NP}}}(y_i \mid \mathbf{x}_i)$, $i = 1,\ldots,n$ based on the estimated coefficients $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\phi}})$.

(ii) Compute the sample moments $u_m = \sum_{i=1}^{n} T_i^m / n$, $m = 1,\ldots,M$.

(iii) Compute the Wald test statistic based on the estimated sample moments.

For the moments of the $U(0,1)$ distribution, $\mu_m = E(u_m) = 1/(m+1)$; $\text{Cov}(u_m, u_l) = ml / [(m+1)(l+1)(m+l+1)n]$. Assuming $\mathbf{u}' = (u_1,\ldots,u_m)$ is normal,

$$\text{UNIF} = (\mathbf{u}-\boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{u}-\boldsymbol{\mu}) \overset{H_0}{\sim} \chi_M^2, \tag{4.4}$$

where $\boldsymbol{\Sigma}$ is the Variance-Covariance matrix defined by the covariances above. The null hypothesis is that the assumed working model is “correct”.

Remark 14. In the proposed test, we replace the true moments by the estimated moments. The estimators $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\phi}})$ are obtained by MLE and under some regularity conditions, they converge almost surely (a.s.) to the true parameters $(\boldsymbol{\beta}, \boldsymbol{\phi})$, (Zacks, 1971). Then, if the true distributional function F is smooth, e.g. twice differentiable with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\phi}$, $F(y_i \mid \mathbf{x}_i, \delta_i = 1; \hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\beta}}) \overset{\text{a.s.}}{\to} F(y_i \mid \mathbf{x}_i, \delta_i = 1; \boldsymbol{\phi}, \boldsymbol{\beta})$, justifying the use of the UNIF test defined by (4.4). See Figure 5.1 in Section 5 for a simulation illustration.

Remark 15. In our simulation study we used $M = 5$ moments, which was found to perform well in Krieger and Pfeffermann (1997). Notice that $\text{Corr}^2(u_m, u_{m-l}) = 1 - (m-l)^2 / [(m+l)+1]^2$, so that higher order moments add only marginally to the power of the test.

For the case where y is binary, we apply in Section 5 the Hosmer and Lemeshow (1980, hereafter H-L) test, defined as follows:

(i) Sort the observed data in S_{NP} based on the estimated probabilities $\hat{\eta}_i = \hat{\text{P}}\text{r}(y_i = 1 \mid \mathbf{x}_i, \delta_i = 1)$, $i \in S_{\text{NP}}$.

(ii) Divide the sorted data into G groups of approximately equal size $n_g \cong (n/G)$ and compute for each group g: o_g - the number of values $y = 1$ and $\bar{\eta}_g = \frac{1}{n_g} \sum_{i \in g} \hat{\eta}_i$. The test statistic is,

$$H - L = \sum_{g=1}^{G} \frac{(o_g - n_g \bar{\eta}_g)^2}{n_g \bar{\eta}_g (1 - \bar{\eta}_g)} \overset{H_0}{\sim} \chi_{(G-2)}^2. \tag{4.5}$$

5. Simulation study

In this section, we present simulation results to illustrate the performance of our proposed approach, separately for the case where the target variable y is continuous, and for the case where y is binary.

5.1 Simulation setup with a continuous target variable- correct model

We start by repeating the same simulation study as performed by Kim and Morikawa (2023), which consists of the following steps:

S1. Generate 5,000 population values as $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$, where $x_{1i}, x_{2i} \overset{\text{Indep}}{\sim} N(2,1)$; $\varepsilon_i \sim N(0,1)$. (The values of the $\boldsymbol{\beta}$ coefficients are in Table 5.1 below.)

S2. Generate selection probabilities to the S_{NP} sample as,

$$\pi_{i,S_{\text{NP}}} = \Pr(\delta_i = 1 \mid y_i, \mathbf{x}_i^*; \boldsymbol{\phi}) = \frac{\exp(\phi_0 + \phi_1 x_{1i} + \phi_2 y_i)}{1 + \exp(\phi_0 + \phi_1 x_{1i} + \phi_2 y_i)}.$$

(The ϕ coefficients are in Table 5.1.)

S3. Repeat Steps 1 and 2 1,000 times, yielding an average selection rate of 50%.

S4. For each simulation, estimate the model parameters and the population mean $\bar{Y}_U = \sum_{i=1}^{5,000} y_i / N$.

Estimators considered:

1- $\hat{\bar{Y}}_{U,X\,\text{known}} = \frac{1}{N}\sum_{i\in U}(\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i})$. The x- variables are known for every unit $i \in U$, $\boldsymbol{\beta}$ is estimated by maximization of the likelihood $l_{S_{\text{NP}}}(\boldsymbol{\beta},\boldsymbol{\phi};\, y) = \sum_{i\in S_{\text{NP}}} \log f_{S_{\text{NP}}}(y_i \mid \mathbf{x}_i; \boldsymbol{\beta},\boldsymbol{\phi})$, under the constraints in (4.3). Note: since the population model is linear, it suffices to know the population means of the $\mathbf{x}$-variables.

2- $\hat{\bar{Y}}_{U,\text{GREG}} = \sum_{i\in S_{\text{NP}}} k_i y_i \Big/ \sum_{i\in S_{\text{NP}}} k_i + \hat{\mathbf{B}}'_{pk}\left[\bar{\mathbf{X}}_U - \sum_{i\in S_{\text{NP}}} k_i \mathbf{x}_i \Big/ \sum_{i\in S_{\text{NP}}} k_i\right]$; $k_i = (1/\hat{\pi}_{i,S_{\text{NP}}})$. The GREG estimator with the standard base sampling weights $w_i = (1/\pi_i)$ replaced by $k_i = (1/\hat{\pi}_{i,S_{\text{NP}}})$. $\hat{\mathbf{B}}_{pk}$ is the probability weighted estimator of $\boldsymbol{\beta}$, with weights k_i.

3- $\hat{\bar{Y}}_{U,\text{KM}} = \sum_{i\in S_{\text{NP}}} \hat{p}_i y_i$, the estimator of Kim and Morikawa (2023). ($\hat{\bar{Y}}_{\text{EL}}$ in equation 3.6).

4- $\hat{\bar{Y}}_{U,\text{MAR}}$- the estimator obtained by assuming that the selection probabilities only depend on the $\mathbf{x}$-variables; $\Pr(\delta_i = 1 \mid y_i, \mathbf{x}_i) = \Pr(\delta_i = 1 \mid \mathbf{x}_i)$, where $\mathbf{x}_i = (x_{1i}, x_{2i})'$. We assume a logistic model, using all the population $\mathbf{x}$- values.

The first 2 estimators are obtained by application of our approach. The estimation of the $\boldsymbol{\beta}$- coefficients in the first estimator is only based on the data in S_{NP}.

Remark 16. An important question regarding the models used in this simulation study is whether the resulting sample model $f_{S_{\text{NP}}}(y_i \mid \mathbf{x}_i; \boldsymbol{\beta},\boldsymbol{\phi}) = \Pr(\delta_i = 1 | y_i, \mathbf{x}_i^*; \boldsymbol{\phi}) f_U(y_i | \mathbf{x}_i; \boldsymbol{\beta}) \big/ \Pr(\delta_i = 1 | \mathbf{x}_i^*; \boldsymbol{\phi},\boldsymbol{\beta})$ is identifiable. By identifiability we mean that there are no different pairs $[\Pr_j(\delta_i = 1 \mid y_i, \mathbf{x}_i^*; \boldsymbol{\phi}_j), f_{Uj}(y_i \mid \mathbf{x}_i; \boldsymbol{\beta}_j)]$, $j = 1,2$ inducing the same sample model for every y and $\mathbf{x}$. Pfeffermann and Landsman (2011) consider sets of conditions guaranteeing the identifiability of the sample model. In particular, for the case of a normal population model and a logistic model for the sample selection probabilities, the sample model is identifiable if the $\mathbf{x}$- variables in the two models differ by at least one variable. Notice that in the models underlying the present simulation, the population model is a function of (x_{1i}, x_{2i}), but the selection logistic model is only a function of x_{1i}, so that the identifiability condition is satisfied.

The results in all the tables in this article are based on 1,000 simulated samples.

5.2 Results for continuous case when fitting the correct model

Table 5.1
Mean estimators and standard errors of model coefficients under the proposed method

	Population model coefficients			Selection model coefficients		
	β_0	β_1	β_2	ϕ_0	ϕ_1	ϕ_2
True coefficients	-4	1	1	-2	1	0.5
Mean estimators	-3.92	0.98	0.99	-2.15	0.80	0.43
Standard errors	0.004	0.001	0.001	0.023	0.008	0.002
Mean PWR estimators	-3.88	0.96	0.99	NA	NA	NA
Standard errors	0.006	0.002	0.001	NA	NA	NA

Note: The mean estimators are the MLE estimators. The probability weighted estimator (PWR) is computed with weights $k_i = (1/\hat{\pi}_{i,S_{\text{NP}}})$.

As can be seen, the $\boldsymbol{\beta}$ coefficients are estimated quite accurately on average. The estimators of the $\boldsymbol{\phi}$ coefficients are somewhat less accurate, but the estimators of the population mean in Table 5.2 still have a negligible bias with these estimators.

Table 5.2
Estimation of population mean. (Mean true value = -0.00)

Method	Bias	Emp. Var $\times$ 1,000	MSE $\times$ 1,000 (Bootstrap estimates)*
$\hat{\bar{Y}}_{U,X\,\text{known}}$	-0.01	2.263	2.363 (3.36)
$\hat{\bar{Y}}_{U,\text{GREG}}$	-0.02	2.423	2.823 (3.89)
$\hat{\bar{Y}}_{U,\text{KM}}$	0.01	2.030	2.080 (---)
$\hat{\bar{Y}}_{U,\text{MAR}}$	0.25	2.106	64.606 (65.11)

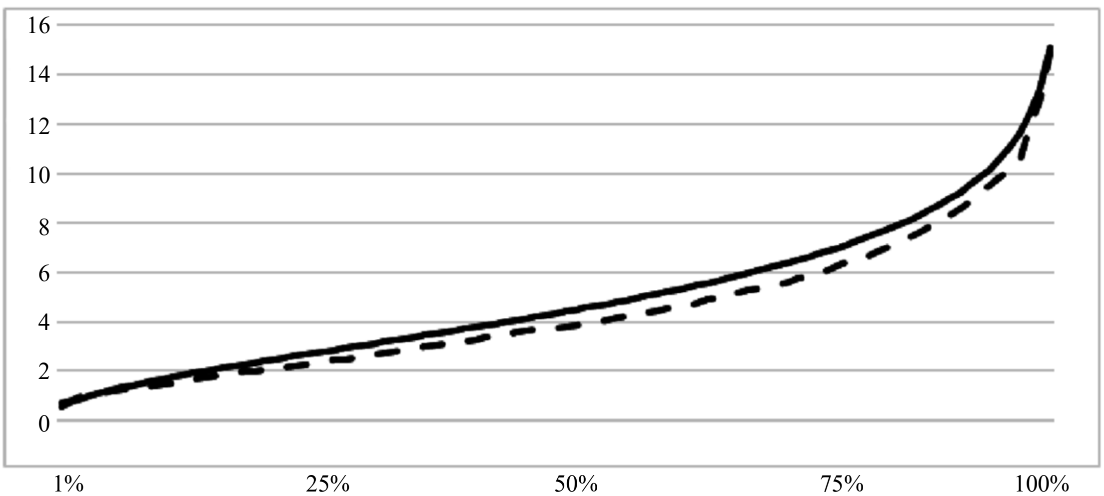
* The bootstrap MSE estimates are based on 100 simulations with 100 bootstrap samples for each simulation.

Estimation of the population mean of the y- values is the primary target of inference in the simulation study and the first three estimators are seen to be literally unbiased. The estimator $\hat{\bar{Y}}_{U,\text{KM}}$ uses all the population $\mathbf{x}$- values and performs best. The estimator $\hat{\bar{Y}}_{U,X\,\text{known}}$ likewise uses all the population $\mathbf{x}$- values (or $\mathbf{x}_i, i \in S_{\text{NP}}$ and $\bar{\mathbf{X}}_U$), but the estimation of the model coefficients is only based on the S_{NP} sample. The estimator $\hat{\bar{Y}}_{U,\text{GREG}}$ uses the S_{NP} model for estimating the $\boldsymbol{\phi}$- coefficients and likewise performs well on average, although with somewhat larger variance and MSE. The bootstrap MSE estimators are conservative with large upward bias. We selected the bootstrap samples by following the procedure proposed in Sverchkov and Pfeffermann (2004), which consists of selecting with replacement a pseudo-population from the sample with probabilities proportional to $k_i = (1/\hat{\pi}_{i,S_{\text{NP}}})$, and then selecting the bootstrap samples S_{NP}^b with the estimated probabilities $\hat{\pi}_{i,S_{\text{NP}}}$ obtained from the original sample. We only considered 100 simulations and 100 bootstrap samples for each simulation, which may explain the upward biases. As expected, the estimator $\hat{\hat{\bar{Y}}}_{U,\mathrm{MAR}}$, which assumes that the selection probabilities only depend on the $\mathbf{x}$- variables has a large positive bias and extremely large MSE. Kim and Morikawa (2023) obtained similar bias and MSE figures in this case.

Overall, the use of our proposed approach seems to perform well in this part of the simulation study.

Model testing: As discussed in Section 4.3, our proposed approach enables testing the models assumed for the population and the sample selection probabilities. Figure 5.1 compares the empirical quantiles of the UNIF statistic (equation 4.4) with the corresponding χ_M^2 quantiles under the correct model for the case of $M = 5$ moments.

Figure 5.1 Empirical quantiles of UNIF statistic (dashed curve) and χ_M^2 quantiles (solid curve) under the correct model with M = 5 moments



We applied the UNIF test for this part of the simulation study and obtained the following results for the case of $M = 5$ and $\alpha = 0.05$ significance level.

	Mean	Standard Deviation	Minimum	Maximum
UNIF statistic	4.64	2.93	0.45	22.80
P-value	0.53	0.28	~0	0.99
H0 not rejected	0.97	0.18	0	1

We conclude that the UNIF test performs well when testing the correct model, with an average non-rejection rate of 97%.

5.3 Application of the proposed procedure when the models are misspecified

In Section 5.2 we assume that the population model and the model for the selection probabilities are specified correctly. In this section, we consider the case where they are misspecified, using the same simulation setup as in Section 5.1.

Case 1. The population model is specified correctly, the sample selection model is misspecified.

In this case, we selected the S_{NP} sample with probabilities, $\pi_{i,S_{\mathrm{NP}}} = \exp(\phi_0 + \phi_1 x_{1i} + \phi_2 y_i^2) / [1 + \exp(\phi_0 + \phi_1 x_{1i} + \phi_2 y_i^2)]$, but assumed as our working model that the selection probabilities are as in Section 5.1 (with y_i in the exponent rather than y_i^2). The population model of y is specified correctly. The average selection rate over the 1,000 simulations is in this case 0.53, similar to what we had before.

Table 5.3
Estimation of model coefficients and standard errors with misspecified selection probabilities

	Population model coefficients			Selection model coefficients		
	β_0	β_1	β_2	ϕ_0	ϕ_1	ϕ_2
True coefficients	-4	1	1	-3	1	0.5
Mean estimators	-4.66	1.14	1.14	-0.52	0.33	0.02
Standard errors	0.002	0.001	0.001	0.007	0.003	0.002
Mean PWR estimators	-4.69	1.16	1.14	NA	NA	NA
Standard errors	0.006	0.002	0.002	NA	NA	NA

Estimation of the $\boldsymbol{\phi}$-coefficients is of little interest in this case because the selection model is misspecified, but notice the relative large bias in the estimation of the $\boldsymbol{\beta}$-coefficients even though the population model is specified correctly. Thus, misspecifying the selection model affects the estimation of the population model.

Table 5.4
Estimation of population mean. (Mean true value = -0.00)

Method	Bias	Emp. Var. $\times$ 1,000	MSE $\times$ 1,000
$\hat{\bar{Y}}_{U,X\text{ known}}$	0.091	1.089	9.37
$\hat{\bar{Y}}_{U,\mathrm{GREG}}$	0.096	1.369	10.585
$\hat{\bar{Y}}_{U,\mathrm{MAR}}$	0.231	0.676	54.037

As can be seen, the bias, empirical variance and MSEs are much larger in this case than under the correct model (Table 5.2). This is not surprising since we fitted a wrong selection model. Here again, we applied the UNIF test for each simulation and obtained the following results.

	Mean	Standard Deviation	Minimum	Maximum
UNIF statistic	27.24	10.75	2.23	71.18
P-value	0.01	0.04	~0	0.82
H0 not rejected	0.04	0.17	0	1

For this case, the UNIF test performs well in rejecting the model fitted, with an average rejection rate of 96%.

Case 2. The sample selection model is specified correctly, the population model is misspecified.

Here, we consider the case where the sample selection model is specified correctly (same as in Section 5.1), but the population model is misspecified. Specifically, the population values have been generated as $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}^2 + \varepsilon_i$, but the assumed working model is as in Section 5.1 (with x_{2i} instead of x_{2i}^2). All the other model specifications are as in Section 5.1.

Table 5.5
Estimation of model coefficients and standard errors with misspecified population model

	Population model coefficients			Selection model coefficients		
	β_0	β_1	β_2	ϕ_0	ϕ_1	ϕ_2
True coefficients	-4	1	0.5	-2	1	0.5
Mean estimators	-5.77	0.94	2.20	-1.39	0.600	0.390
Standard errors	0.013	0.002	0.002	0.040	0.011	0.007
Mean PWR estimators	-5.36	0.94	1.99	NA	NA	NA
Standard errors	0.01	0.005	0.007	NA	NA	NA

As expected, the estimators of the $\boldsymbol{\beta}$- coefficients are highly biased and so are the estimators of the $\boldsymbol{\phi}$- coefficients. Thus, as already noted regarding Table 5.3, misspecification of one of the models affects the estimation of both models.

Table 5.6
Estimation of population mean. (Mean true value = -0.00)

Method	Bias	Emp. Var. × 1,000	MSE × 1,000
$\hat{\bar{Y}}_{U, X \text{ known}}$	-0.024	20.16	20.74
$\hat{\bar{Y}}_{U, \text{GREG}}$	-0.010	42.03	42.11
$\hat{\bar{Y}}_{U, \text{MAR}}$	-0.209	5.85	49.53

The estimators of the population mean are less biased than for the case where the sample selection model is misspecified (Table 5.4), but with relatively large variances, particularly for the GREG estimator. Notice that the GREG estimator depends directly on the estimated sample selection probabilities, which are highly biased (Table 5.5).

Application of the UNIF test yields in this case,

	Mean	Standard Deviation	Minimum	Maximum
UNIF statistic	207.12	46.49	82.12	394.47
P-value	~0.00	~0	~0	~0
H0 not rejected	0	0	0	0

The UNIF test rejects the models fitted in each of the 1,000 simulations.

5.4 Simulation setup with binary target variable- correct model

So far, we illustrated the performance of our proposed method for the case where the target y- variable is continuous. Following, we consider the case where y is binary. We use a similar simulation setup to the setup used for the continuous case, except that the population y- values are now generated as $\Pr(y_i = 1) = \text{logit}^{-1}(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}),$ with the $\mathbf{x}$- values generated as before. We again use the logistic model $\pi_{i,S_{\text{NP}}} = \Pr(\delta_i = 1 | y_i, \mathbf{x}_i^*; \boldsymbol{\phi}) = \exp(\phi_0 + \phi_1 x_{1i} + \phi_2 y_i) / [1 + \exp(\phi_0 + \phi_1 x_{1i} + \phi_2 y_i)]$ for selecting the S_{NP} sample, maximizing the likelihood under the same constraints as before.

The question arising is whether the S_{NP} model is identifiable in this case as well. Wang et al. (2014) establish the following condition for model identifiability. The auxiliary variables $\mathbf{x}$ in the population model can be decomposed as $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2)$ with the dimension of $\mathbf{x}_2 \geq 1,$ such that $\pi_{i,S_{\text{NP}}} = \Pr(\delta_i = 1 | y_i, \mathbf{x}_i) = \Pr(\delta_i = 1 | y_i, \mathbf{x}_{1i}),$ implying that the sample selection model does not depend on $\mathbf{x}_2,$ given y and $\mathbf{x}_1.$ This condition is satisfied in our simulation setup. Recall that for a normal population model and logistic selection probabilities, the sample model is identifiable if the $\mathbf{x}$ variables in the two models differ in at least one variable, a somewhat weaker condition. See Remark 16.

The results in the following tables are based on 1,000 simulations with an average selection rate of 70%. The estimated value is again the true population mean (proportion) of the target y- variable.

Table 5.7
Mean estimators and standard errors of model coefficients

	Population model coefficients			Selection model coefficients		
	β_0	β_1	β_2	ϕ_0	ϕ_1	ϕ_2
True coefficients	-4	1	1	-2	1	5
Mean estimators	-4.40	1.18	1.01	-2.89	1.50	5.65
Standard errors	0.01	0.004	0.002	0.016	0.008	0.085
Mean PWR estimators	-0.24	0.20	0.16	NA	NA	NA
Standard errors	0.001	0.0005	0.0005	NA	NA	NA

Note: The mean estimators are the MLE estimators. The probability weighted estimator (PWR) is computed with weights $k_i = (1/\hat{\pi}_{i,S_{\text{NP}}}).$

The MLE and PWR estimators are biased, notably the PWR estimator and the MLE estimators of the $\boldsymbol{\phi}$- coefficients, but as can be seen in Table 5.8, the bias seems to have little effect on the estimation of the population mean of the target y- variable.

We consider the following estimators of the population mean:

1- $\hat{\bar{\mathbf{Y}}}_{U,H} = \sum_{i \in S_{\text{NP}}} k_i y_i \Big/ \sum_{i \in S_{\text{NP}}} k_i;\ k_i = (1/\hat{\pi}_{i,S_{\text{NP}}}).$

2- $\hat{\bar{\mathbf{Y}}}_{U,\text{EI}} = (1/N)\left\{\sum_{i \in S_{\text{NP}}} y_i + \left[(N-n) \Big/ \sum_{i \in S_{\text{NP}}} (k_i - 1)\right] \sum_{i \in S_{\text{NP}}} (k_i - 1)\, y_i\right\};$ see Sverchkov and Pfeffermann (2004) for derivation of this estimator.

3- $\hat{\bar{\mathbf{Y}}}_{U,X\text{ known}} = (1/N)\left\{\sum_{i \in S_{\text{NP}}} y_i + \sum_{j \notin S_{\text{NP}}} \hat{E}_{S_{\text{NP}}}(y_j | \mathbf{x}_j) + [(N-n)/n] \sum_{i \in S_{\text{NP}}} (k_i - 1) \Big/ (\bar{k}_{S_{\text{NP}}} - 1)\,[y_i - \hat{E}_{S_{\text{NP}}}(y_i | \mathbf{x}_i)]\right\};$ $\bar{k}_{S_{\text{NP}}} = (1/n) \sum_{i \in S_{\text{NP}}} k_i,$ $\hat{E}_{S_{\text{NP}}}$ is the estimated expectation under the model (4.2). The estimator when all the population $\mathbf{x}$'s are known. See Sverchkov and Pfeffermann (2004) for the derivation of this estimator.

4- $\hat{\hat{\mathbf{Y}}}_{U,\,\mathrm{GREG}} = \left(\sum_{i\in S_{\mathrm{NP}}} k_i y_i \Big/ \sum_{i\in S_{\mathrm{NP}}} k_i\right) + \hat{\mathbf{B}}'_{pk}\left(\overline{\mathbf{X}}_U - \sum_{i\in S_{\mathrm{NP}}} k_i \mathbf{x}_i \Big/ \sum_{i\in S_{\mathrm{NP}}} k_i\right)$; same as when y is continuous.

5- $\hat{\hat{\mathbf{Y}}}_{U,\mathrm{MAR}}$; the estimator obtained from $\hat{\hat{Y}}_{U,\mathrm{EI}}$ when replacing k_i by the MAR weight, $k_i^* = [1+\exp(\hat{\alpha}_0 + \hat{\alpha}_1 x_{1i} + \hat{\alpha}_2 x_{2i})]/\exp(\hat{\alpha}_0 + \hat{\alpha}_1 x_{1i} + \hat{\alpha}_2 x_{2i})$.

Table 5.8
Estimation of population mean. (Mean true value = 0.5)

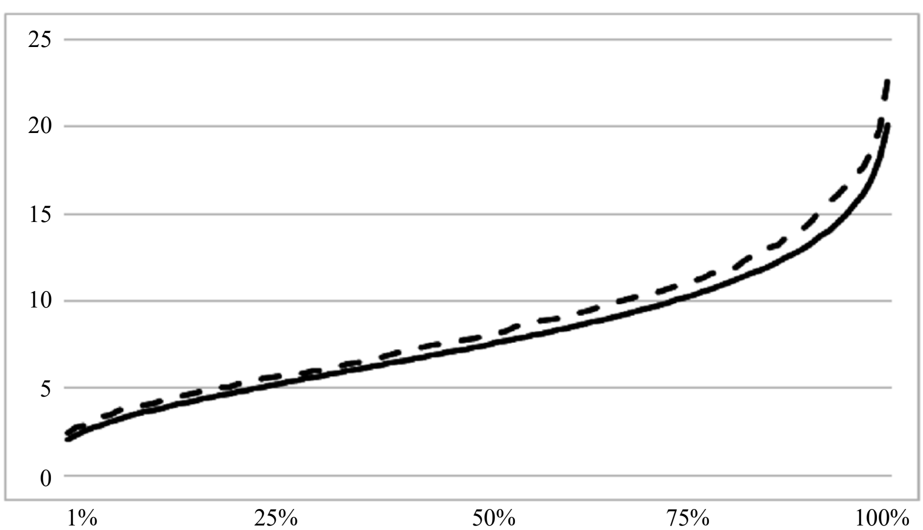
Estimator	Bias	Emp. Var × 1,000	Emp. MSE × 1,000* (Bootstrap estimate)*
$\hat{\hat{Y}}_{U,H}$	-0.051	1.600	4.201 (5.60)
$\hat{\hat{Y}}_{U,\mathrm{EI}}$	0.001	0.009	0.010 (0.026)
$\hat{\hat{Y}}_{U,X\,\mathrm{known}}$	-0.006	0.169	0.205 (0.300)
$\hat{\hat{Y}}_{U,\mathrm{GREG}}$	-0.006	0.172	0.208 (0.309)
$\hat{\hat{Y}}_{U,\mathrm{MAR}}$	0.149	0.049	22.25 (22.50)

* The bootstrap MSE estimates are based on 100 simulations with 100 bootstrap samples for each simulation.

As can be seen, all the estimators except $\hat{\hat{Y}}_{U,\mathrm{MAR}}$ have a negligible bias, despite the bias of the estimated ϕ-coefficients. Among the estimators, $\hat{\hat{Y}}_{U,\mathrm{EI}}$ is the clear winner, with surprisingly small MSE, much lower than the MSE of $\hat{\hat{Y}}_{U,X\,\mathrm{known}}$. This might be due to the fact that this estimator uses the observed y's, (~ 70% in this case), and only predicts the sum of the unobserved y's. The estimator $\hat{\hat{Y}}_{U,X\,\mathrm{known}}$ also uses the observed y's, but it uses the estimated expectation under the S_{NP} model for predicting the sum of the unobserved y's. The estimator $\hat{\hat{Y}}_{U,H}$ has a relatively large MSE due to its relatively larger bias.

Model testing: As for the continuous case, we tested the goodness of fit of our model, using in this case the Hosmer and Lemeshow (1980, H-L) test (equation 4.5). Figure 5.2 compares the empirical quantiles of the H-L statistic with the corresponding χ^2_{G-2} quantiles under the correct model with G = 10 groups.

Figure 5.2 Empirical quantiles of H-L statistic (dashed curve) and χ^2_{G-2} quantiles (solid curve) under the correct model with G = 10 groups



Application of the test in the simulations with $\alpha = 0.05$ significance level yields,

	Mean	Standard deviation	Minimum	Maximum
H-L test	8.56	4.30	1.108	30.57
p-value	0.46	0.29	~0	~1
H0 not rejected	0.934	0.248	0	1

The H-L test performs well when testing the correct model.

5.5 Application of proposed method for binary case with misspecified models

In Section 5.4, we assumed that the population model and the model for the selection probabilities are specified correctly. In this section we consider the case where they are misspecified, using the same simulation setup as before.

Case 1. The population model is specified correctly, the sample selection model is misspecified.

In this case, we selected the S_{NP} sample with probabilities, $\Pr(\delta_i = 1 \mid y_i, \mathbf{x}_i) = \exp(-2 + 5x_{1i}y_i) / [1 + \exp(-2 + 5x_{1i}y_i)]$, but assumed as our working model the same model as in Section 5.4. The population model of y is specified correctly. The average selection rate over the 1,000 simulations is in this case 54%.

Table 5.9
Estimation of model coefficients under misspecified model

	Population model coefficients			Selection model coefficients		
	β_0	β_1	β_2	ϕ_0	ϕ_1	ϕ_2
True coefficients	-4	1	1	NA	NA	NA
Mean estimators	-4.78	1.75	1.00	-3.1	0.96	3.1
Standard errors	0.01	0.003	0.003	0.02	0.003	0.14
Mean PWR estimators	-0.16	0.25	0.14	NA	NA	NA
Standard errors	0.002	0.001	0.0004	NA	NA	NA

Except for β_2, the MLE estimates of the other $\boldsymbol{\beta}$- coefficients are biased, with larger bias than when the sample selection model was specified correctly (Table 5.7).

Table 5.10
Estimation of population mean. (Mean true value = 0.5)

Estimator	Bias	Emp. Var × 1,000	Emp. MSE × 1,000
$\hat{\bar{Y}}_{U,H}$	0.04	1.60	3.2
$\hat{\bar{Y}}_{U,\text{EI}}$	0.05	0.29	2.8
$\hat{\bar{Y}}_{U,X\text{ known}}$	0.11	0.53	12.6
$\hat{\bar{Y}}_{U,\text{GREG}}$	0.11	0.53	12.6
$\hat{\bar{Y}}_{U,\text{MAR}}$	0.29	0.17	84.3

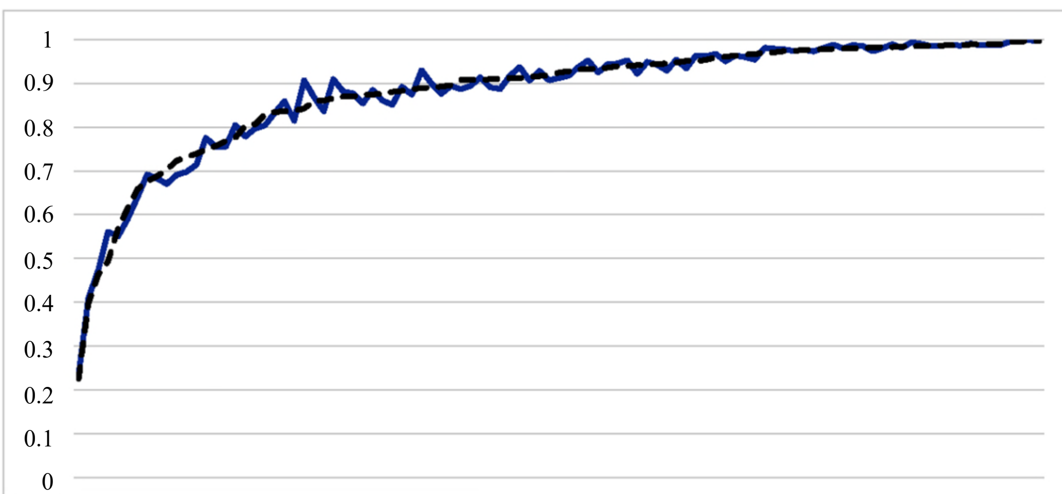
The results in Table 5.10 indicate that the first 2 estimators have small bias despite of the model misspecification, with smaller MSE of $\hat{\hat{Y}}_{U,H}$, but much larger MSEs of $\hat{\hat{Y}}_{U,\mathrm{EI}}$, $\hat{\hat{Y}}_{U,X\,\mathrm{known}}$ and $\hat{\hat{Y}}_{U,\mathrm{GREG}}$, compared to the MSEs obtained under the correct model (Table 5.8). These large MSEs are clearly explained by the misspecification of the sample selection model. As before, $\hat{\hat{Y}}_{U,\mathrm{MAR}}$ has a large bias and an extreme MSE.

We applied the H-L test with $\alpha =$ 0.05 significance level, yielding the following results:

	Mean	Standard deviation	Minimum	Maximum
H-L test	16.12	179.4	0.778	5557.8
p-value	0.43	0.30	~0	~1
H0 not rejected	0.89	0.312	0	1

Clearly, the H-L test fails to reject the misspecified model in this case. In an attempt to understand this outcome, Figure 5.3 compares the S_{NP} model $\Pr_{S_{\mathrm{NP}}}(y_i \mid \mathbf{x}_i; \boldsymbol{\beta}, \boldsymbol{\phi}) = \Pr(\delta_i = 1 \mid y_i, \mathrm{x}_{1i}; \boldsymbol{\phi}) \Pr_U(y_i \mid \mathbf{x}_i; \boldsymbol{\beta}) / \Pr(\delta_i = 1 \mid \mathrm{x}_{1i}; \boldsymbol{\phi}, \boldsymbol{\beta})$ with true coefficients used to select the sample, with the corresponding estimated model under the misspecified model, for a simple random sample of 100 observations from the S_{NP} sample. The horizontal axis is ordered based on the sampled values of $\Pr_{S_{\mathrm{NP}}}(y_i \mid \mathbf{x}_i)$ of the true model.

Figure 5.3 Comparison of correct model and estimated misspecified model*



* Dashed curve represents the correct S_{NP} model, twisted curve represents the estimated (misspecified) model.

The estimated model under wrong specification is seen to yield almost perfect estimators of the correct model producing the S_{NP} data, which explains why the H-L test does not reject the model. This is an example for what is known as "practical nonidentifiability" (Lee and Berger, 2001), meaning that even though the S_{NP} model is theoretically identifiable, another model may fit the data almost as well. Notice in Table 5.10

that the use of the misspecified working model yields two almost unbiased estimators of the true population mean.

Case 2. The population model is misspecified, the sample selection model is specified correctly.

In this case, we used the same sample selection model as in Section 5.4 (correct specification of the working model), but we generated the population values as $y_i = \text{logit}^{-1}(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}^2)$. As our working model we assumed the model of Section 5.4 (x_{2i}, instead of x_{2i}^2). The average selection rate is in this case 73%. All the other model specifications are as in Section 5.4.

Table 5.11
Estimation of model coefficients under misspecified model

	Population model coefficients			Selection model coefficients		
	β_0	β_1	β_2	ϕ_0	ϕ_1	ϕ_2
True coefficients	-4	1	1	-2	1	5
Mean estimators	-6.07	1.34	1.62	-3.50	1.55	10.66
Standard errors	0.015	0.005	0.002	0.025	0.01	0.129
Mean PWR estimators	-0.35	0.18	0.23	NA	NA	NA
Standard errors	0.001	0.001	0.001	NA	NA	NA

All the estimators are highly biased, due to misspecification of the population model.

Table 5.12
Estimation of population mean (True mean value = 0.55)

Estimator	Bias	Emp. Var $\times$ 1,000	Emp. MSE $\times$ 1,000
$\hat{\bar{Y}}_{U,H}$	-0.15	3.64	26.14
$\hat{\bar{Y}}_{U,\text{EI}}$	-0.004	0.01	0.026
$\hat{\bar{Y}}_{U,X\,\text{known}}$	-0.06	0.53	4.13
$\hat{\bar{Y}}_{U,\text{GREG}}$	-0.05	0.58	3.08
$\hat{\bar{Y}}_{U,\text{MAR}}$	0.11	0.05	12.15

All the estimators except for $\hat{\bar{Y}}_{U,H}$ and $\hat{\bar{Y}}_{\text{U,MAR}}$ have a negligible bias in this case, with $\hat{\bar{Y}}_{U,\text{EI}}$ performing really well, as in the case of correct model specification (Section 5.4). On the other hand, $\hat{\bar{Y}}_{U,X\,\text{known}}$, although having a negligible bias, has a large MSE, even larger than the MSE of $\hat{\bar{Y}}_{U,\text{GREG}}$.

Application of the H-L test with $\alpha =$ 0.05 significance level yields in this case,

	Mean	Standard deviation	Minimum	Maximum
H-L test	42.5	41.3	6.09	950.0
p-value	0.002	0.02	~0	0.637
H0 not rejected	0.006	0.08	0	1

The H-L test performs well in rejecting the misspecified model.

6. Concluding remarks

In recent years, there is growing research on the use of NP samples for inference on population parameters, as an alternative or complement to the use of probability samples. A major problem with the use of these samples is their possible nonrepresentativeness of the corresponding target population, which if not accounted for properly, may lead to large bias in the inference process. In this article, we review and discuss several approaches proposed in the literature to deal with this problem, distinguishing between methods based on integration of the NP sample with a corresponding probability sample, and methods that base the inference solely on the NP sample with added calibration constraints. Another distinction emphasized is between methods that assume that the selection to the NP sample depends on known auxiliary variables $\mathbf{x}$, but not on the target study y variable, and methods that assume that the selection depends also on y.

We also propose two additional methods for inference from a nonprobability sample, one that employs the empirical likelihood approach and one that requires specifying the population model parametrically. We discuss the conditions guaranteeing that the resulting model holding for the NP sample is identifiable, and propose simple tests for testing that the models are specified correctly. Our simulation study illustrates good performance of the proposed method and generally good performance of the test statistics.

A major problem underlying all the methods considered in this article is that they assume, at least implicitly, that every unit in the population has a positive probability to be in the NP sample. Clearly, if this is not the case, inference on the target population could be highly biased. This problem also exists with traditional probability samples when the sampling frame is not complete, known as "under-coverage". When the group of units with zero probability to be included in the NP sample is known, say certain geographical areas, industries or ethnic groups, the target population should be redefined accordingly. When this is not the case, integration of the NP sample with an appropriate PS sample and the use of known population means of the $\mathbf{x}$- variables for calibration, is a possible way to at least reduce the bias of the NP sample. This is an important topic for further research.

There are two important questions regarding the use of our proposed method that require further investigation. The first question is how to proceed when the test statistic rejects the models defining the NP model. We do not have a clear answer to this question at this stage other than a scholarly consideration of alternative models. We mention again that the use of a logistic model for the selection probabilities has some theoretical justification, and this model is in common use.

The second related question is the choice of the $\mathbf{x}$- variables in the models, when there are many of them. In practice, it may be the case that the analyst has a set of variables that he likes to include in the population model, which as explained in Section 4.1, defines also the variables included in the sample selection model. When this is not the case, one can use an appropriate stepwise algorithm. Beaumont et al. (2024a) use a forward stepwise procedure, aimed at minimizing their proposed AIC criterion.

All the methods discussed in the present article should be considered as first attempts of inference from nonprobability samples, and more theoretical research and practical applications are required before they can be used routinely for the production of official statistics.

References

AAPOR (2010). "Report on Online Survey Panels", available at http://poq.oxfordjournals.org/content/early/2010/10/19/poq.nfq048.full.html.

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

Beaumont, J.-F. (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.

Beaumont, J.-F., Bosa, K., Brennan, A., Charlebois, J. and Chu, K. (2024a). Handling non-probability samples through inverse probability weighting with an application to Statistics Canada's crowdsourcing data. *Survey Methodology*, 50(1), 77-106. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2024001/article/00004-eng.pdf.

Beaumont, J.-F., Bosa, K., Brennan, A., Charlebois, J. and Chu, K. (2024b). Authors' response to comments on "Handling non-probability samples through inverse probability weighting with an application to Statistics Canada's crowdsourcing data": Some new developments on likelihood approaches to estimation of participation probabilities for non-probability samples. *Survey Methodology*, 50(1), 123-141. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2024001/article/00001-eng.pdf.

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. *Canadian Journal of Statistics*, 50, 1166-1185.

Citro, C.F. (2014). From multiple modes for surveys to multiple data sources for estimates. *Survey Methodology*, 40(2), 137-161. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2014002/article/14128-eng.pdf.

Conti, P.L., Marella, D. and Scanu, M. (2008). Evaluation of matching noise for imputation techniques based on nonparametric local linear regression estimators. *Computational Statistics and Data Analysis*, 53, 354-365.

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

Hosmer, D.W., and Lemeshow, S. (1980). A goodness-of-fit test for the multiple logistic regression model. *Communications in Statistics*, A10, 1043-1069.

Keiding, N., and Louis, T.A. (2016). Perils and potentials of self-selected entry to epidemiological studies and surveys. *Journal of the Royal Statistical Society*, Series A, 179, 319-376.

Kim, J.K., and Morikawa, K. (2023). An empirical likelihood approach to reduce selection bias in voluntary samples. *Calcutta Statistical Association Bulletin*, 75, 8-27.

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

Kott, P., and Ridenhour, J. (2024). Calibration weighting with a blended (probability and nonprobability) sample: Mean and variance estimation when errors can come from both samples. *Methods Report, RTI Press*.

Krieger, A.M., and Pfeffermann, D. (1997). Testing of distribution functions from complex sample surveys. *Journal of Official Statistics,* 13, 123-142.

Lee, J., and Berger, J.O. (2001). Semiparametric Bayesian analysis of selection models. *Journal of the American Statistical Association,* 96, 1397-1409.

Lee, S. (2006). Propensity score adjustment as a weighting scheme for volunteer panel Web survey. *Journal of Official Statistics*, 22, 329-349.

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

Marella, D., and Pfeffermann, D. (2023). Accounting for non-ignorable sampling and non-response in statistical matching. *International Statistical Review*, 91, 269-293.

Molenberghs, G., Beunckens, C. and Kenward, M.G. (2008). Every missing not at random has a missingness at random counterpart with equal fit. *Journal of the Royal Statistical Society*, Series B, 70, 371-388.

Pfeffermann, D., and Landsman, A. (2011). Are private schools really better than public schools? Assessment by methods for observational studies. *Annals of Applied Statistics*, 5, 1726-1751.

Pfeffermann, D., Marella, D. and Summa, D. (2025a). Matching of a non-probability sample with a probability sample affected by nonignorable sampling and nonresponse. Submitted for publication.

Pfeffermann, D., Preminger, A. and Sikov, S. (2025b). Statistical inference under nonignorable sampling and nonresponse - An empirical likelihood approach. (Under revision).

Pfeffermann, D., and Sikov, A. (2011). Imputation and estimation under non ignorable nonresponse in household surveys with missing covariate information. *Journal of Official Statistics*, 27, 181-209.

Pfeffermann, D., and Sverchkov, M. (1999). Parametric and semiparametric estimation of regression models fitted to survey data. *Sankhyā*, Series B, 61, 166-186.

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

Rivers, D. (2007). Sampling from web surveys. *Proceedings of the Survey Research Methods Section*, American Statistical Association, Alexandria, VA.

Sayag, D., Ben-Hur, D. and Pfeffermann, D. (2022). Reducing revisions in hedonic house price indices by the use of nowcasts. *International Journal of Forecasting*, 38, 253-266.

Sverchkov, M., and Pfeffermann, D. (2004). Prediction of finite population totals based on the sample distribution. *Survey Methodology*, 30(1), 79-92. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2004001/article/6996-eng.pdf.

Wang, S., Shao, J. and Kim, J.K. (2014). An instrument variable approach for identification and estimation with nonignorable nonresponse. *Statistica Sinica*, 24, 1097-1116.

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.

Yang, S., Kim, J.K. and Hwang, Y. (2021). Integration of data from probability surveys and big found data for finite population inference using mass imputation. *Survey Methodology*, 47(1), 29-58. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2021001/article/00004-eng.pdf.

Zacks, S. (1971). *The Theory of Statistical Inference*. (Wiley series in probability and mathematical statistics).