sciendo

Journal of Official Statistics, Vol. 39, No. 2, 2023, pp. 151–172, http://dx.doi.org/10.2478/JOS-2023-0008

Adjusting for Selection Bias in Nonprobability Samples by Empirical Likelihood Approach

Daniela Marella[1]

Large amount of data are today available, that are easier and faster to collect than survey data, bringing new challenges. One of them is the nonprobability nature of these big data that may not represent the target population properly and hence result in highly biased estimators. In this article two approaches for dealing with selection bias when the selection process is nonignorable are discussed. The first one, based on the empirical likelihood, does not require parametric specification of the population model but the probability of being in the nonprobability sample needed to be modeled. Auxiliary information known for the population or estimable from a probability sample can be incorporated as calibration constraints, thus enhancing the precision of the estimators. The second one is a mixed approach based on mass imputation and propensity score adjustment requiring that the big data membership is known throughout a probability sample. Finally, two simulation experiments and an application to income data are performed to evaluate the performance of the proposed estimators in terms of robustness and efficiency.

Key words: Big data; informative sample; mass imputation.

1. Introduction

The main characteristic of big data sources is that they provide us with detailed information often in real time, since they are generated in an automated way using information technology systems or sensors. This results in massive datasets of very large volume and in a huge variety of forms of data. For an overview on big data, see Beresewicz et al. (2018). Large amount of data are therefore available, that are easier and faster to collect than the standard data sources as census and survey data, bringing new opportunities and challenges, see Pfeffermann (2015).

However, if on one hand big data represent potentially new data sources, on the other we need to know how much they can help the inferential process and under which assumptions. From the statistical inference point of view, what really matters is the way these data are generated. Big data sources are nonprobability samples, which often fail to represent the target population properly then the analysis results may be subject to selection biases, see Elliott and Valliant (2017) and Meng (2018). In this article, this concern is addressed.

Let A and B be two data sources, where B is a nonprobability sample while A is an independent probability sample. We assume that $(\mathbf{x}, y)$ are available from B while $\mathbf{x}$ is available from survey data A, where $\mathbf{x}$ is a vector of p auxiliary variables and y is the

[1]University of Rome La Sapienza, Piazzale Aldo Moro 5, Rome, 00185, Italy. Email: daniela.marella@uniroma1.it

Acknowledgment: The author is grateful to the referees for very careful reading of the manuscript and thoughtful comments.

variable of interest. Generally speaking, there are three possible methods to draw reliable statistical inference from nonprobability samples. The first method is the so called propensity score adjustment, see Rosenbaum and Rubin (1983). In this approach the unknown probability of selection for the units in B is estimated from sample A (propensity or sampling score) by the covariates $\mathbf{x}$. The second approach is based on calibration. That is, information on the auxiliary variables in sample B is calibrated with that in the population or at least can be estimated from the probability sample A, see Kott (2006) and DiSogra et al. (2011). The third approach is mass imputation where imputed values of y are created for all units in the probability sample A. Then, an estimator of the parameter of interest based on imputed data is computed. Survey data integration for combining a probability sample with a nonprobability sample is also discussed in Yang et al. (2021a) where a formal framework for mass imputation is developed and asymptotic results for the k nearest neighbor estimator are established. The nearest neighbor imputation estimator of Rivers (2007) is also covered as a special case. Finally, Yang et al. (2021b) propose a doubly robust estimator of the finite population mean using the estimated propensity scores as well as an outcome linear regression model. The double robustness entails that the final estimator is consistent for the true value if either the probability of selection into the nonprobability sample or the outcome model is correctly specified, not necessarily both.

All the aforementioned methods assume that the selection mechanism for sample B is ignorable after controlling on $\mathbf{x}$. Since selection mechanism and nonresponse are closely related, it is essentially the missing at random (MAR) assumption of Rubin (1976). However, the MAR assumption is not always realistic because survey participation may be related to the survey topic of interest. For instance, we might expect that the selection process (self-selection) to be nonignorable on Twitter data, since the propensity to tweet (sample inclusion probability) might depend on the particular subject, which will often be related to the target variable. When the inclusion probabilities are related to the value of the target outcome variable even after conditioning on the model covariates, the observed outcomes are no longer representative of the population outcomes and the model holding for the sample data is then different from the model holding in the population. This allows the possibility that being in the sample or analogously being a respondent depends in some stochastic way on the variable of interest y. It is essentially the not missing at random (NMAR) assumption of Rubin (1976).

If MAR assumption does not hold, then we can build a NMAR model for the selection mechanism and estimate the model parameters, see Chang and Kott (2008) and Riddles et al. (2016). Existing approaches for parameter estimation for a propensity score model under nonignorable nonresponse can be classified as fully parametric approaches or method of moments approaches. A fully parametric approach, which makes parametric assumptions about the population distribution of the study variable, is considered in Beaumont (2000). Also, the Heckman (1979) selection model approach is a fully parametric approach in the sense that the outcome regression model and the response model are linked by a joint normal distribution on the error terms of the two models. In Galimard et al. (2018) an imputation model for missing binary data with NMAR mechanism from Heckman's model using a onestep maximum likelihood estimator is derived. These fully parametric approaches can be used to estimate the parameters in the

response model, but the estimates can be very sensitive to failure of the assumed model. The method of moments approach does not directly use the outcome model while the response model is assumed to be specified. In Chang and Kott (2008) and Kott and Chang (2010) propensity score weighting for nonignorable missing mechanism is introduced together with instrumental variable calibration. The authors extended the notion of calibration weighting by allowing the number of explanatory variables in the assumed response model to be less than the number of calibrations variables. Instead of the fully parametric or the calibration approach, Riddles et al. (2016) consider an alternative modeling approach that uses parametric model assumptions about the study variable among the respondents only. Such a modeling approach has been considered in Pfeffermann and Sikov (2011).

Evidently, accounting for nonignorable selection mechanism is a major undertaking and the present article attempts to address this challenge. In this article two approaches for dealing with selection bias when the selection process is nonignorable are discussed. The first one, based on the empirical likelihood, does not require parametric specification of the population model but the probability of being in the nonprobability sample needed to be modeled. An important advantage of this approach is that it facilitates the use of calibration constraints that can help to correct for selection bias in nonprobability samples. That is, auxiliary information known for the population or estimable from the probability sample A can be incorporated as calibration constraints, thus enhancing the precision of the estimators. The success of the proposed approach depends on proper modeling of the unknown selection probabilities. However, the resulting sample model can be tested based on the observations in nonprobability sample by standard test statistics. Then, model diagnostics are more feasible and the method is less sensitive to failure of the assumed selection model. The approach relies on work by Feder and Pfeffermann (2019) for dealing with problems such as observational studies, informative sampling and nonignorable nonresponse. Such an approach has also been proposed to deal with the statistical matching problem under nonignorable sampling and nonresponse in Marella and Pfeffermann (2021).

The second one is a mixed approach based on mass imputation and propensity score adjustment. We consider the case when additionally the membership to the nonprobability sample B can be determined throughout the probability sample A, as in Yang et al. (2021a). First of all, imputed values $\tilde{y}$ are created for all units in A and the selection probabilities for units in B are estimated from A by $(\mathbf{x}, \tilde{y})$. Next, the sample empirical likelihood is maximized by a two steps estimation procedure.

The article is organized as follows. In Section 2 the basic setup in a fully parametric context is briefly introduced. In Section 3 a semiparametric approach based on the empirical likelihood (EL) is discussed and its performance is evaluated by a simulation study in Sections 5 and 6. The failure in proper modeling the unknown selection probabilities is investigated in Subsection 6.2.1. The robustness of the approach to violations of the population normality assumption is evaluated in Subsection 6.2.2, where skewed and binary data are considered. Furthermore, an application to income data is presented in Section 7. In Section 4 the mixed approach is described. Its performance is assessed by a simulation study in Section 8. Section 9 draws final conclusions.

2. Adjusting for Selection Bias: Basic Setup

The main challenges in using nonprobability samples are under-coverage and self selection. In the sequel we assume that the target population is fully covered, then the inclusion probabilities are nonzero for all the population units. Suppose that we have two independent samples A and B selected from a finite population of size N generated from a joint probability distribution function (pdf) $f(\mathbf{x}, y; \boldsymbol{\theta})$, governed by a vector parameter $\boldsymbol{\theta}$. Let y be the study variable and $\mathbf{x} = (x_1, x_2, ..., x_p)'$ the vector of p auxiliary variables. Let B be a nonprobability sample of size n_B, such as a voluntary sample or a self-selected sample, and A an independent probability sample. We assume that $(\mathbf{x}, y)$ are available from B while $\mathbf{x}$ is available from survey data A. Then, B contains rich information on $(\mathbf{x}, y)$ but the sampling mechanism is unknown while the sample A, representing the finite population, does not observe the study variable of interest. Let δ_i be the sample inclusion indicator, that is, a Bernoulli random variable taking value $\delta_i = 1$ if population unit $i \in B$, $\delta_i = 0$ otherwise. The sampling mechanism for the nonprobability sample B is ignorable (noninformative) after controlling on $\mathbf{x}$ if,

$$P(\delta_i = 1|\mathbf{x}_i, y_i) = P(\delta_i = 1|\mathbf{x}_i), \tag{1}$$

for each $\mathbf{x}_i$. Unfortunately, the ignorability condition is a strong assumption and it is not verifiable based on the observed data. If the sampling mechanism for sample B is not ignorable, the inclusion probabilities are related to the value of the target outcome variable y even after conditioning on the model covariates $\mathbf{x}$, then the observed outcomes are no longer representative of the population outcomes and the model holding for the sample data is then different from the model holding in the population, see Pfeffermann and Sverchkov (2009) and Pfeffermann (2011) for discussion of the notion of informative sampling. This is equivalent to assume that the sample B is subject to not missing at random (NMAR) nonresponse, by which the response probabilities depend in some stochastic way on the study variable of interest.

In this section an approach of reducing the selection bias associated with the nonprobability sample B is briefly illustrated in a parametric context. In Section 3 the use of the EL is proposed. From Pfeffermann et al. (1998), the marginal sample pdf of $(\mathbf{x}_i, y_i)$ for $i \in B$ is defined as,

$$f_B\left(\mathbf{x}_i, y_i; \boldsymbol{\theta}, \boldsymbol{\gamma}_B\right) = \frac{P\left(\delta_i = 1|\mathbf{x}_i, y_i; \boldsymbol{\gamma}_B\right)}{P\left(\delta_i = 1; \boldsymbol{\theta}, \boldsymbol{\gamma}_B\right)} f_p(\mathbf{x}_i, y_i; \boldsymbol{\theta}), \tag{2}$$

where $f_p(\mathbf{x}_i, y_i; \boldsymbol{\theta})$ is the population pdf governed by $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}_B$ represents any additional parameters defining the sample distribution, resulting from the sampling process. Under independence between observations corresponding to different sampling units, the sample likelihood can be approximated by the product of the sample pdfs over the corresponding sample observations. Hence, the sample likelihood is,

$$L_B(\boldsymbol{\theta}, \boldsymbol{\gamma}_B) = \prod_{i=1}^{n_B} f_B(\mathbf{x}_i, y_i; \boldsymbol{\theta}, \boldsymbol{\gamma}_B). \tag{3}$$

The probabilities $P(\delta_i = 1|\mathbf{x}_i, y_i; \boldsymbol{\gamma}_B)$ appearing in the sample pdf (2) needed to be modeled. To this aim, a parametric model indexed by the unknown parameter $\boldsymbol{\gamma}_B = (\boldsymbol{\gamma}_x, \gamma_y)'$

of length $p + 1$ can be assumed, which is allowed to depend on the observed data (the outcome and auxiliary variables). Formally,

$$P(\delta_i = 1|\mathbf{x}_i, y_i; \boldsymbol{\gamma}_B) = g(\boldsymbol{\gamma}_x'\mathbf{x}_i + \gamma_y y_i) \tag{4}$$

for some known function g, taking values in the range [0, 1].

Remark 1 The sample inclusion probabilities in B may depend on several unobserved variables and yet, by definition of the sample pdf, one only needs to model the probabilities $P(\delta_i = 1|\mathbf{x}_i, y_i)$. As discussed and illustrated in subsequent sections, the resulting sample model can be tested based on the observations in B.

Modeling the probabilities by the logistic or probit functions is common, but notice that in our case the probabilities depend also on the study variable y. Then, the two models, the population model $f_p(\mathbf{x}_i, y_i; \boldsymbol{\theta})$ and the parametric model (4), define the model holding for the observed units in B. Notice that, the sample likelihood in Equation (3) only depends on the observed data in sample B. Furthermore, it needs to be maximized with respect to the population and selection model parameters ($\boldsymbol{\theta}$, $\boldsymbol{\gamma}_B$). Thus, the unknown sampling parameters $\boldsymbol{\gamma}_B$ are estimated jointly from the likelihood.

Remark 2 If the main target of inference is the mean of y (μ_y), after having estimated $\boldsymbol{\theta}$, the following estimators can be computed,

$$\hat{\mu}_y = E_p(y_i; \hat{\boldsymbol{\theta}}), \quad \hat{\mu}_{y,H} = \frac{\sum_{i \in B} y_i / \hat{P}\left(\delta_i = 1|\mathbf{x}_i, y_i\right)}{\sum_{i \in B} 1 / \hat{P}\left(\delta_i = 1|\mathbf{x}_i, y_i\right)}. \tag{5}$$

where $\hat{\mu}_{y,H}$ is the Hájek estimator, see Hájek (1964). Large differences between the two estimators may indicate misspecification of either the population model or the parametric model (4). Notice that both the estimators take into account the informative sampling design in **B** since $\hat{P}\left(\delta_i = 1|\mathbf{x}_i, y_i\right)$ instead of the propensity scores $\hat{P}\left(\delta_i = 1|\mathbf{x}_i\right)$ are used.

However, the maximization of sample likelihood in Equation (3) with respect to ($\boldsymbol{\theta}$, $\boldsymbol{\gamma}_B$) can be complicated numerically and result in unstable estimates, depending on the population model and the model assumed for the selection probabilities. One may also face identifiability or practical identifiability problems, see Pfeffermann and Landsman (2011) and Lee and Berger (2001). For this reason, we propose in the next section the use of the empirical likelihood approach.

3. Adjusting for Selection Bias: A Semiparametric Approach

The approach described in Section 2 is fully parametric, since it makes parametric assumptions about both the population distribution of the study variable and the selection mechanism. In this section we propose a semiparametric approach based on the use of the EL which enables estimating the parameter $\boldsymbol{\gamma}_B$, governing the sampling process, without specifying the population model. The EL combines the robustness of nonparametric methods with the efficiency of the likelihood approach, see Owen (2001, 2013) and references therein. The EL is essentially the likelihood of the multinomial distribution, where the parameters are the point masses assigned to the distinct sample values. An important advantage of the empirical likelihood approach is that it facilitates the use of

calibration constraints. That is, auxiliary information on known population means for some auxiliary variables can be incorporated by placing additional constraints on the maximization process. See Chaudhuri et al. (2010) for details of the constrained estimation procedure and the asymptotic properties of the resulting empirical likelihood estimators. Last, but not least, not requiring to specify the population model the approach is more robust and often easier to implement.

The basic idea of the empirical likelihood approach is to approximate the population distribution by a multinomial model with probabilities $p_i^{xy} = Pr(\mathbf{x}_i, y_i)$, which support is given by the empirical observations $\{(\mathbf{x}_i, y_i), i = 1,...,n_B\}$. This means that a multinomial probability is assigned just to the observed values in sample B. Notice that, the statement regarding the support is a basic assumption underlying the EL approach which can be justified by having sufficiently large sample. Then, the sample distribution in B is,

$$p_{i,B}^{xy} = \frac{P(\delta_i = 1|\mathbf{x}_i, y_i; \boldsymbol{\gamma}_B)}{P(\delta_i = 1; \{p_i^{xy}\}, \boldsymbol{\gamma}_B)} p_i^{xy}, \tag{6}$$

where $P(\delta_i = 1; \{p_i^{xy}\}, \boldsymbol{\gamma}_B) = \sum_{i \in B} P(\delta_i = 1|\mathbf{x}_i, y_i, \boldsymbol{\gamma}_B) p_i^{xy}$. The sample EL based on B is thus,

$$EL_B(\{p_i^{xy}\}, \boldsymbol{\gamma}_B) = \prod_{i \in B} p_{i,B}^{xy} = \prod_{i \in B} \frac{P(\delta_i = 1|\mathbf{x}_i, y_i; \boldsymbol{\gamma}_B)}{\Sigma_{i \in B} P(\delta_i = 1|\mathbf{x}_i, y_i; \boldsymbol{\gamma}_B) p_i^{xy}} p_i^{xy}. \tag{7}$$

Then, the semiparametric approach defines the sample EL and combines it with a parametric model for the probabilities $P(\delta_i = 1|\mathbf{x}_i, y_i; \boldsymbol{\gamma}_B)$, as specified in Equation (4).

As previously stressed, an important advantage of the EL approach is that it facilitates the use of calibration constraints that can help to correct for selection bias in nonprobability samples. Specifically, known population means of auxiliary variables related to the study variable and measured for the nonprobability sample B can be incorporated by placing additional constraints (calibration constraints) on the maximization process. For instance, in the simulation study of Section 5 the constraint,

$$\sum_{i \in B} \mathbf{x}_i \sum_{\{i \in B: \mathbf{X} = \mathbf{X}_i\}} p_i^{xy} = \boldsymbol{\mu}_x \tag{8}$$

is considered, where the population mean $\boldsymbol{\mu}_x$ of $\mathbf{x}$ is assumed known. Then, the likelihood (7) must be maximized with respect to $(\{p_i^{xy}\}, \boldsymbol{\gamma}_B)$ under the constraints,

$$p_i^{xy} \geq 0, \sum_{i \in B} p_i^{xy} = 1, \tag{9}$$

for all i, and the calibration constraint (8). One only needs the estimates of the multinomial population model parameters $\{p_i^{xy}\}$ and thus, we may consider $\boldsymbol{\gamma}_B$, as nuisance parameter. In order to write the likelihood in Equation (7) as only a function of the unknown probabilities $\{p_i^{xy}\}$, we adopt the profile likelihood approach. We use some initial estimates for the set of probabilities $\{p_i^{xy}\}$ and we solve the constrained maximization problem by first computing the profile likelihood of $\boldsymbol{\gamma}_B$ and then maximizing the profile likelihood over $\boldsymbol{\gamma}_B$. For a given $\boldsymbol{\gamma}_B$, we then maximize the resulting likelihood under the constraints in Equastions (8) and (9) with respect to the unknown probabilities $\{\hat{p}_i^{xy}\}$,

yielding $\{p_i^{xy}\}$. This completes the first iteration in the estimation process. In the second iteration, we consider the estimates $\{\hat{p}_i^{xy}\}$ as known, re-estimate the parameters $\boldsymbol{\gamma}_B$, and then the unknown probabilities $\{p_i^{xy}\}$. The iterations continue until convergence. See Feder and Pfeffermann (2019) for conditions guaranteeing the convergence of the maximization process. There are cases where a solution does not exist. An example is where all the observed values of a constraining variable are greater (or smaller) than its known population mean. Furthermore, a combination of multivariate constraints can also preclude a solution. For instance, when the sum of two variables used in the constraints is greater for all the observed units than the sum of the corresponding population means.

Remark 3 The simulation study has been carried out by using the software R, R Core Team (2021). The maximization with respect to $\boldsymbol{\gamma}_B$ can be performed by using the R numerical optimization function optim. For a given $\boldsymbol{\gamma}_B$, the maximization with respect to $\{p_i^{xy}\}$ can be carried out by using the function emplik in the R package mev, see Belzile et al. (2022). See Owen (2013) for related theory and further details.

Notice that, inference on the unknown model parameters is based on the sample EL which requires that the corresponding sample model is identifiable. The sample model is not identifiable if there is more than one combination of a population model and a sampling mechanism yielding the same sample model. See Pfeffermann and Landsman (2011) and references therein for conditions guaranteeing the identifiability of the sample model. Notice that, for a given parameter $\boldsymbol{\gamma}_B$ and without any constraints the EL is not identifiable. In the proposed approach the empirical likelihood is maximized under a set of calibration constraints. Then, the main question is how the survey variables defining the constraints should be chosen. As in Chang and Kott (2008), such variables should be correlated as highly as possible with y and $\mathbf{x}$ because otherwise they provide little or no information on the probabilities $P(\delta_i = 1|\mathbf{x}_i, y)$.

Remark 4 If $\boldsymbol{\mu}_x$ is unknown but a probability sample A is available then the auxiliary infomation in sample B can be calibrated with that in sample A. Then, in Equastion (8) the mean vector $\boldsymbol{\mu}_x$ can be replaced by its Horvitz-Thompson estimator computed from sample A. Formally,

$$\sum_{i\in B}\mathbf{x}_i \sum_{\{i\in B:\mathbf{X}=\mathbf{X}_i\}} p_i^{xy} = \frac{1}{N}\sum_{j\in A} d_j\mathbf{x}_j, \tag{10}$$

where d_j is the sampling weight associated to the jth unit in sample A. Constraint (10) is used in the application to income data of Section 7.

The success of the proposed approach depends on proper modeling of the unknown selection probabilities for sample B, that is the estimates can be sensitive to failure of the assumed model. However, once the parameters $(\{p_i^{xy}\}, \boldsymbol{\gamma}_B)$ have been estimated, the null hypothesis that the sample model fits the sample data can be tested successfully by classical test statistics, because the sample model refers to the observed data. An overview of the plausible test statistics that can be used for assessing the goodness of fit of the sample pdf is in Pfeffermann (2011). For instance, in the simulation study of Section 5 the Kolmogorov-Smirnov test has been used to compare the theoretical and the empirical sample pdfs of y. The asymptotic distribution of test statistic and correct critical values have been obtained by use of parametric bootstrap, as established theoretically by Babu and Rao (2004).

Finally, the bias and the standard deviation of the population parameters estimates can be obtained by resampling method. Formally, once the estimated model has been validated M bootstrap samples can be selected from it and for each bootstrap sample the unknown parameters can be estimated according the proposed approach. Then, bootstrap estimates of bias and standard deviation can be computed.

4. A Mixed Approach Based on Mass Imputation and Propensity Score Adjustment

In this section a data integration approach for combining the nonprobability sample B with an independent probability sample A is described. It is a mixed approach based on mass imputation and propensity score adjustment requiring that we can observe δ_i, the B sample inclusion indicator, from the probability sample A. That is, among the elements in the sample A, it is possible to obtain the membership information from the nonprobability sample B, as in Kim and Wang (2019). As stressed in Yang et al. (2021a), the key insight is that the subsample of units in probability sample A with the membership information ($\delta_i = 1$) constitutes a second phase sample from B, which acts as a new population. Clearly, this condition is more plausible in the big data context where the nonprobability sample B is so large that any probability sample A is bound to overlap with it.

As previously stressed, unlike the usual imputation for missing data analysis, in mass imputation imputed values for all units in the probability sample A are created. The mass imputation methods and their statistical properties are discussed in Yang et al. (2021a). The nearest neighbor imputation estimator of Rivers (2007) is also covered as a special case. The parameter of interest is μ_y. The proposed approach can be described by the following steps:

Step 1. Create imputed values $\tilde{y}_i$ for all units $i \in A$ by nearest neighbor method. The basic idea is to find the nearest neighbor in sample B to create an imputed value of y for each unit in sample A. Formally, the unit $k \in B$ closest to unit $i \in A$ is determined by the Euclidean distance based on the auxiliary variables $\mathbf{x}$ and the corresponding y value from this unit is used as the imputed value.

Step 2. Regress the membership indicator δ against $(\mathbf{x}, \tilde{y})$ in sample A, estimate the selection probabilities for all units in B and compute their inverse. Let $\tilde{w}_i$ be the estimated sample weight (pseudo-weight) for the ith unit in B, for $i = 1,...,n_B$.

Step 3. Estimate the parameter μ_y by

(3.1) the Horvitz-Thompson estimator

$$\mu_W = \frac{1}{N}\sum_{i\in B}\tilde{w}_i y_i. \tag{11}$$

(3.2) the maximum sample EL estimator (μ_{EL}). Formally, once the sampling weights $\tilde{w}_i$ in B are computed a two steps estimation procedure can be applied in the maximization of the EL (7). More specifically, since

$$P\left(\delta_i = 1|\mathbf{x}_i, y_i; \boldsymbol{\gamma}_B\right) \approx \frac{1}{E_B(\tilde{w}_i|\mathbf{x}_i, y_i; \boldsymbol{\gamma}_B)}, \tag{12}$$

from Equation (6) the EL (7) becomes,

$$EL_B\left(\{p_i^{xy}\}, \boldsymbol{\gamma}_B\right) = \prod_{i=1}^{n_B} p_{i,B}^{xy} \approx \prod_{i=1}^{n_B} \frac{E_B\left(\tilde{w}_i; \left\{p_{i,}^{xy}\right\}, \boldsymbol{\gamma}_B\right)}{E_B\left(\tilde{w}_i | \mathbf{x}_i, y_i; \boldsymbol{\gamma}_B\right)} p_i^{xy}. \quad (13)$$

Then, in the first step the expectations displayed in Equation (13) are estimated from the observed data, using classical model fitting procedures. Specifically, the expectation $E_B(\tilde{w}_i|\mathbf{x}_i, y_i; \boldsymbol{\gamma}_B)$ could be estimated by regressing the sampling weights $\tilde{w}_i$ against $(\mathbf{x}_i, y_i)$. See, for example, Pfeffermann and Sverchkov (2009) and Pfeffermann (2011) for examples of regression models that can be used for this purpose, depending on the problem at hand. In the second step, fixing the unknown parameters $\boldsymbol{\gamma}_B$ featuring in these expectations at their estimated values allows to maximize the EL in Equation (13) only with respect to the parameter $\{p_i^{xy}\}$ indexing the population pdf, thus simplifying and stabilizing the maximization process.

The basic idea of the proposed method is to create predicted values for y in the probability sample A. In order to accomplish this in Step 1 the covariates $\mathbf{x}$ are used, then the predictions are based on the ignorability assumption of the selection mechanism acting in the nonprobability sample B. A class of nonparametric imputation procedures based on k-nearest neighbors methods (kNN), including 1NN, is discussed in Marella et al. (2008), where both theoretical and simulation results are obtained. Furthermore, a nonparametric technique based on local linear regression is discussed in Conti et al. (2008). In Step 2 the inclusion probabilities in B are computed by applying the estimated regression of δ on $(\mathbf{x},\tilde{y})$ to the observed values $(\mathbf{x}, y)$ in B.

When auxiliary information is available it can be incorporated into the method to avoid the ignorability assumption in Step 1 and to improve the quality of the imputed values $\tilde{y}$ in sample A. For instance, auxiliary information may refer to a set of proxy variables $\mathbf{z} \subset \mathbf{x}$ expected to behave similarly to the variable of interest. Under this circumstance better predicted values can be obtained in Step 1. Furthermore, the proxy variables $\mathbf{z}$, if sufficiently associated with y, can help studying the relationship between y and δ and in particular, help verifying or refuting the ignorability assumption.

In Section 8 a simulation study is employed to investigate the performance of the proposed method when the selection process is nonignorable, comparing it with other existing methods. As discussed in Section 9, new developments of the present work include the use of proxy variables.

5. Simulation Study 1

In order to evaluate the performance of the approach discussed in Sections 2 and 3 in its parametric and semiparametric form, a simulation experiment is performed. Suppose that the primary target of inference is to estimate μ_y. The simulation study consists of the following steps:

Step 1 Generate a population of $N = 1,000,000$ observations $(\mathbf{x}_i, y_i)$, where (x, y) has a bivariate normal distribution with mean $\boldsymbol{\mu} = (\mu_x, \mu_y)'$ and variance covariance matrix (V-C matrix) $\boldsymbol{\Sigma}$ ($\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, for short). Specifically, the marginal distribution of x is $\mathcal{N}(\mu_x, \sigma_x^2)$ with $\mu_x = 5$ and $\sigma_x^2 = 1$. The conditional distribution of y given x

is $\mathcal{N}(\mu_{y|x}, \sigma^2_{y|x})$, with $\mu_{y|x} = \beta_0 + \beta_1 x$, $\beta_0 = \mu_y - \beta_1 \mu_x$, $\beta_1 = \sigma_{xy}/\sigma^2_x$, $\sigma^2_{y|x} = \sigma^2_y - \beta^2_1 \sigma^2_x$. We assume that $\beta_0 = 2$, $\beta_1 = 1$, $\sigma_{y|x} = 2$.

Step 2 Draw a sample B from the population generated in Step 1 by a Poisson sampling design with expected sample size $E(n_B) = 0.2N$ and sample inclusion probabilities given by,

$$E_p(\pi_i | x_i, y_i; \boldsymbol{\gamma}_B) = \kappa \exp\{\gamma_x x_i + \gamma_y y_i\}, \tag{14}$$

where $\boldsymbol{\gamma}_B = (\gamma_x, \gamma_y)'$ is the sampling model parameter and κ guarantees that the expectation is less or equal to one. We use different sampling model parameters $\boldsymbol{\gamma}_B$, so as to distinguish between informative and noninformative samples. From Marella and Pfeffermann (2019) the joint sample pdf $f_B(x_i, y_i)$ is $\mathcal{N}(\boldsymbol{\mu}_B, \boldsymbol{\Sigma}_B)$, with mean vector $\boldsymbol{\mu_B} = (\mu_x + (\gamma_x + \beta_1 \gamma_y)\sigma^2_x, \mu_y + \beta_1 \gamma_x \sigma^2_x + \gamma_y \sigma^2_y)'$ and V-C matrix $\boldsymbol{\Sigma}_B = \boldsymbol{\Sigma}$, that is the sample V-C matrix is the same as for the population distribution. Then, the sample model and the population model are in the same family and only differ in the mean ($\boldsymbol{\mu}_B \neq \boldsymbol{\mu}$).

Step 3 For computational reasons, as in Kim and Wang (2019), we generate 500 samples S of size $n_S = 2{,}000$ from sample B drawn in Step 2 by a simple random sampling (*srs*). In *srs* the selection probabilities are equal for all units in sample B and the sample S can also be regarded as a set of independent and identically distributed observations from the sample model $f_B(x_i, y_i)$.

The population model parameters are estimated by parametric and semiparametric approach.

Parametric approach. For each sample S drawn in Step 3, the population model parameters $(\mu_x, \sigma_x, \beta_0, \beta_1, \sigma_{y|x}, \mu_y)$ are estimated under the following scenarios:

Scenario 1 The sample B and then each sample S are simply treated as simple random samples (*srs*). The estimates of the population parameters are denoted by $\{\hat{\mu}_{x,I}, \hat{\sigma}_{x,I}, \hat{\beta}_{0,I}, \hat{\beta}_{1,I}, \hat{\sigma}_{y|x,I}, \hat{\mu}_{x,I}\}$, where I means that the selection mechanism acting in B is ignored.

Scenario 2 The sample likelihood in Equation (3) is maximized with respect to the population parameters and the sampling parameters $\boldsymbol{\gamma}_B$. The estimates of the population parameters are denoted by $\{\hat{\mu}_{x,P}, \hat{\sigma}_{x,P}, \hat{\beta}_{0,P}, \hat{\beta}_{1,P}, \hat{\sigma}_{y|x,P}, \hat{\mu}_{x,P}\}$, where P stands for parametric approach.

Semiparametric approach. In what follows we assume knowledge of the population mean μ_x (Equation (8)). Hereafter the calibration constraint. For each sample S drawn in Step 3, the population model parameters $\{p_i^{xy}\}$ are estimated under the following scenarios:

Scenario 3 The sample B and then each sample S are simply treated as simple random samples but the knowledge of μ_x is assumed, so as to enhance the precision of the estimator for the mean μ_y. Formally, the EL under the independent and identically distributed assumption,

$$EL\left(\{p_i^{xy}\}\right) = \prod_{i \in B} p_i^{xy}, \tag{15}$$

is maximized under the constraints in Equation (9) and the calibration constraint in Equation (8). Denote by $\hat{\mu}_{y,ISP}$ the estimate of μ_y, where *ISP*

means that such an estimate is obtained under the semiparametric approach by ignoring the selection mechanism acting in B.

Scenario 4 The sample empirical likelihood in Equation (7) is maximized with respect to $(\{p_i^{xy}\}, \boldsymbol{\gamma}_B)$ under the constraints in Equation (9) and the calibration constraint in Equation (8). Denote by $\hat{\mu}_{y,ISP}$ the estimate of μ_y, where SP stands for semiparametric approach.

In order to evaluate the performance of the proposed approach as the informativeness of sampling design acting in B changes, in scenarios 2 and 4 we assume that the model (14) for the inclusion probabilities $P(\delta_i = 1|\mathbf{x}_i, y_i)$ is known and the sample EL is maximized with respect to the sampling and the population parameters $(\{p_i^{xy}\}, \boldsymbol{\gamma}_B)$. The robustness of the semiparametric approach to misspecification of the selection model $P(\delta_i = 1|\mathbf{x}_i, y_i)$ is assessed by a sensitivity analysis in Subsection 6.2.1. Finally, the robustness to violations of the population normality assumption is evaluated in Subsection 6.2.2. Notice that we generated the population values only once, so as to assess the design-based properties of the various estimation procedures.

6. Results of Simulation Study 1

In this section the simulation results obtained by the parametric approach (Subsection 6.1) and by the semiparametric approach (Subsection 6.2) are reported.

6.1. Simulation Results for the Parametric Approach

We start by studying the effect of ignoring the sampling mechanism used for drawing the sample B in the parametric approach. This is done by comparing the estimates of the population parameters under the scenarios 1 and 2, described in Section 5. In Table 1 the bias (B), the standard deviation (Sd) and the root mean square error (RMSE) of the estimates $\hat{\mu}_{y,I}, \hat{\mu}_{y,P}$ over the 500 samples are reported, for different $\boldsymbol{\gamma}_B$ coefficients so to distinguish between informative and noninformative samples. In Table 2 and 3 the mean and the standard deviation of the remaining parameters $(\mu_x, \sigma_x, \beta_0, \beta_1, \sigma_{y|x})$ over the 500 samples are presented. As stated previously, since $\Sigma_B = \Sigma$ it follows that $\hat{\sigma}_{x,I} = \hat{\sigma}_{x,P}$, $\hat{\beta}_{1,I} = \hat{\beta}_{1,P}$, and $\hat{\sigma}_{y|x,I} = \hat{\sigma}_{y|x,P}$, for details see Marella and Pfeffermann (2019). Then, in Table 2 and 3 just the means of $(\hat{\sigma}_{x,I}, \hat{\beta}_{1,I}, \hat{\sigma}_{y|x,I})$ and the corresponding standard deviations are reported.

As results in Table 1 show, for $\boldsymbol{\gamma}_B = (0,0)'$ the estimate $\hat{\mu}_{y,I}$ coincides with $\hat{\mu}_{y,P}$ since the sampling process acting in B is ignorable, $B(\hat{\mu}_{y,I}) = B(\hat{\mu}_{y,P}) = 0$. When $\boldsymbol{\gamma}_B \neq (0,0)'$ the sampling design is informative and the bias in $\hat{\mu}_{y,I}$ (last two rows in Table 1), coming from

Table 1. Bias (B), standard deviation (Sd) and RMSE of $\hat{\mu}_{y,I}$ and $\hat{\mu}_{y,P}$ over the 500 samples for different $\boldsymbol{\gamma}_B$ coefficients. True parameter is $\mu_y = 7$.

$\boldsymbol{\gamma}_B$	$B(\hat{\mu}_{y,I})$	$B(\hat{\mu}_{y,P})$	$Sd(\hat{\mu}_{y,I})$	$Sd(\hat{\mu}_{y,P})$	$RMSE(\hat{\mu}_{y,I})$	$RMSE(\hat{\mu}_{y,P})$
(0,0)	0.00	0.00	0.05	1.43	0.05	1.43
(0, 0.5)	2.16	-0.11	0.04	2.02	2.16	2.02
(0.25, 0.5)	2.26	-0.28	0.04	2.06	2.26	2.08

Table 2. Mean of the estimates of (μ_x, σ_x, β_0, β_1, $\sigma_{y|x}$) under scenarios 1 and 2, over the 500 samples for different $\boldsymbol{\gamma}_B$ coefficients. True parameters are $\mu_x = 5$, $\sigma_x = 1$, $\beta_0 = 2$, $\beta_1 = 1$, $\sigma_{y|x} = 2$.

$\boldsymbol{\gamma}_B$	$\bar{\hat{\mu}}_{x,I}$	$\bar{\hat{\mu}}_{x,P}$	$\bar{\hat{\sigma}}_{x,I}$	$\bar{\hat{\beta}}_{0,I}$	$\bar{\hat{\beta}}_{0,P}$	$\bar{\hat{\beta}}_{1,I}$	$\bar{\hat{\sigma}}_{y\|x,I}$
(0,0)	5.00	5.00	1.00	2.00	2.00	1.00	2.00
(0,0.5)	5.43	4.89	0.98	4.65	2.83	0.83	1.82
(0.25, 0.5)	5.62	4.85	0.95	5.21	3.23	0.72	1.80

Table 3. Standard deviation of the estimates of (μ_x, σ_x, β_0, β_1, $\sigma_{y|x}$) under scenarios 1 and 2, over the 500 samples with different $\boldsymbol{\gamma}_B$ coefficients.

$\boldsymbol{\gamma}_B$	$Sd(\hat{\mu}_{x,I})$	$Sd(\hat{\mu}_{x,P})$	$Sd(\hat{\sigma}_{x,I})$	$Sd(\hat{\beta}_{0,I})$	$Sd(\hat{\beta}_{0,P})$	$Sd(\hat{\beta}_{1,I})$	$Sd(\hat{\sigma}_{y\|x,I})$
(0,0)	0.02	0.62	0.02	0.22	1.10	0.04	0.03
(0, 0.5)	0.02	0.63	0.02	0.23	1.68	0.04	0.03
(0.25, 0.5)	0.02	0.65	0.01	0.25	1.86	0.04	0.03

the bias affecting $\hat{\beta}_{0,I}$ and $\hat{\mu}_{x,I}$ (last two rows in Table 2), increases considerably. Then, ignoring the sample selection process in sample B affects negatively the quality of the estimates of μ_y. The estimator $\hat{\mu}_{y,I}$ works poorly, even though $\hat{\mu}_{y,I}$ has the smallest standard deviation as shown in Table 1, a well known phenomenon from other studies, see Marella and Pfeffermann (2019). Furthermore, the larger is the informativeness of the sampling process the larger will be the bias in $\hat{\mu}_{y,I}$. Finally, the bias of the estimates $\hat{\mu}_{y,P}$ for $\boldsymbol{\gamma}_B \neq (0,0)'$ reduces since scenario 2 takes into account the selection mechanism acting in B. Same consideration holds for the estimates of the other population parameters, see Table 2 and 3.

Figure 1 (left) exhibits the population pdf, the sample *pdf* and the estimated sample pdf of y for one of the 500 samples, for the case $\boldsymbol{\gamma}_B = (0.25, 0.5)'$. As can be seen, the sample pdf is very different from the population pdf, but the distribution of the estimated pdf is close to the true population distribution. Finally, with regard to the variable of interest y we test the model fitted for the sample units by Kolmogorov-Smirnov (KS) test statistic given by

$$KS_Y = max_{y \in B} |\hat{F}_{emp}(y) - F_B(y; \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\gamma}}_B)|, \tag{16}$$

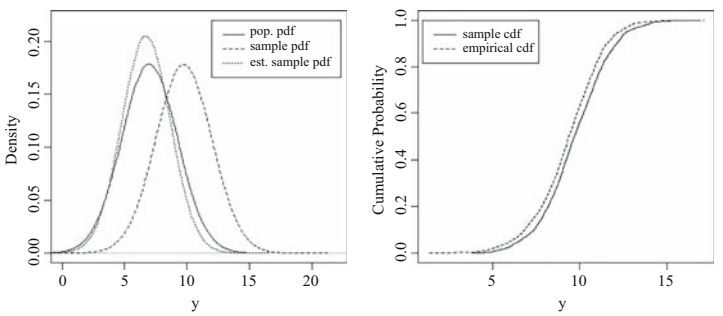

Fig. 1. Population pdf, sample pdf and estimated sample pdf of y (left); sample cdf and empirical cdf of y (right), for $\boldsymbol{\gamma}_B = (0.25, 0.5)'$.

where $\hat{F}_{emp}(y) = \frac{1}{n_B}\sum_{i\in B} I(y_i \leq y)$ is the empirical cumulative distribution (cdf), $I(y_i \leq y)$ is the indicator function taking the value 1 if $y_i \leq y$ and 0 otherwise, and

$$F_B(y; \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\gamma}}_B) = \int_{-\infty}^{y} f_B(y_i; \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\gamma}}_B)dyi, \tag{17}$$

is the sample cdf. The asymptotic distribution of test statistic in Equation (16) and correct critical values can be obtained by use of parametric bootstrap, as established theoretically by Babu and Rao (2004) and applied in Pfeffermann (2011). Specifically, first of all $M = 1{,}000$ samples are generated from the estimated sample model. Next, for each bootstrap sample the unknown parameters and the corresponding test statistic are computed. The empirical distribution of test statistic provides approximate critical values for the null distribution. In Figure 1 (right), the estimated sample cdf and the empirical cdf of y are reported. The KS statistic is 0.069 and the critical value corresponding to a significance level $\alpha = 0.05$ is 0.197. Then, the null hypothesis that the estimated model fits the sample data in B is not rejected.

6.2. *Simulation Results for the Semiparametric Approach*

In this section we proceed to estimate μ_y by the semiparametric approach described in Section 3. The robustness with respect to misspecification of the parametric model for $P(\delta_i = 1|\mathbf{x}_i, y_i)$ and to violations of the population normality assumption is evaluated in Subsections 6.2.1 and 6.2.2, respectively. Table 4 shows the bias (B), the standard deviation (Sd) and the RMSE of the estimates $\hat{\mu}_{y,ISP}$ and $\hat{\mu}_{y,SP}$ obtained under scenarios 3 and 4, respectively.

Notice that, if sample B is treated as a simple random sample the estimates of μ_y obtained maximizing the EL in Equation (15) under the constraints in Equation (9) match the estimates $\hat{\mu}_{y,I}$ obtained under scenario 1. The conclusions of Table 4 are similar to those obtained from Table 1. When $\boldsymbol{\gamma}_B = (0, 0)'$ the estimates $\hat{\mu}_{y,ISP}$ and $\hat{\mu}_{y,SP}$ are equal since the sampling design is not informative, $B(\hat{\mu}_{y,ISP}) = B(\hat{\mu}_{y,SP}) = 0$. When $\boldsymbol{\gamma}_B \neq (0, 0)'$ the estimates $\hat{\mu}_{y,ISP}$ ignoring the sampling process show as light reduction in the bias compared to the estimates $\hat{\mu}_{y,I}$ (scenario 1) because of the introduction of the calibration constraint in Equation (8) in the EL maximization (scenario 3). As results in the Table 4 show, the estimates $\hat{\mu}_{y,SP}$ obtained by maximizing the sample EL (7) under the constraints in Equation (9) and the calibration constraint in Equation (8) (scenario 4) are characterized by lower selection bias and standard deviation illustrating the good performance of our proposed methodology. Finally, for the sample in Figure 1 the goodness of fit of the estimated model to the observed data is tested by the KS statistic in Equation (16). Its value is 0.098 and the critical value corresponding to a significance level $\alpha = 0.05$ is 0.332. Then, the null

Table 4. Bias (B), standard deviation (Sd) and RMSE of $\hat{\mu}_{y,ISP}$ and $\hat{\mu}_{y,SP}$, over the 500 samples with different $\boldsymbol{\gamma}_B$ coefficients. True parameter is $\mu_y = 7$.

$\boldsymbol{\gamma}_B$	$B(\hat{\mu}_{y,ISP})$	$B(\hat{\mu}_{y,SP})$	$Sd(\hat{\mu}_{y,ISP})$	$Sd(\hat{\mu}_{y,SP})$	$RMSE(\hat{\mu}_{y,ISP})$	$RMSE(\hat{\mu}_{y,SP})$
(0, 0)	0.00	0.00	0.04	0.05	0.04	0.05
(0,0.5)	1.78	-0.02	0.07	0.09	1.78	0.09
(0.25,0.5)	1.74	-0.08	0.17	0.11	1.75	0.14

hypothesis that the estimated model fits the sample data is not rejected. The results in the Table 4 suggest that μ_y can be estimated almost unbiasedly and with acceptable standard error estimates when external auxiliary information is incorporated in the EL maximization, as the estimates $\hat{\mu}_{y,SP}$ and their standard deviations show.

6.2.1. Misspecification of the Selection Model

As previously stated, the EL approach does not require to specify the population pdf while the relationship between the probabilities $P(\delta_i = 1|\mathbf{x}_i, y_i)$ and the variables $(\mathbf{x}, y)$ is parametrically specified, see Equation (4). Hence, its performance depends on how well the assumed parametric model describes the unknown selection mechanism acting in B. In this section a sensitivity analysis is performed to assess the impact on μ_y estimate due to misspecification of the selection model. First of all, suppose that the sample B is selected by a Poisson sampling design with expected sample size $E(n_B) = 0.2N$ and unknown selection probabilities given by Equation (14) with $\boldsymbol{\gamma}_B = (0.25, 0.5)'$. Next, 500 samples S of size $n_S = 2000$ are drawn from B by a *srs* (Step 1–3, Section 5). Let us assume that the probabilities $P(\delta_i = 1|x_i, y_i)$ are modeled by:

Model A: a linear logistic model

$$P(\delta_i = 1|x_i, y_i) = logit^{-1}(\gamma_x x_i + \gamma_y y_i). \tag{18}$$

Model B: a quadratic logistic model. In Equation (18) x is squared and y is linear;

Model C: a quadratic logistic model. In Equation (18) both x and y are squared.

For each sample S, the estimates of μ_y under scenario 4 and models A-C are computed. Table 5 reports the bias (B) and the standard deviation (Sd) of such estimates over the 500 samples. Finally, in Table 6 the corresponding RMSEs are computed.

As results in Table 5 show, $B(\hat{\mu}_{y,SP})$ increases from 0.89 (model A) to 1.53 (model B) with a reduction in the standard deviation from 0.36 to 0.18. An additional increase is obtained under model C where $B(\hat{\mu}_{y,SP}) = 1.78$. Recall that under a correct specification of the selection model $B(\hat{\mu}_{y,SP}) = -0.08$ (see Table 4). Hence, as results in Table 5 show, the reduction of the bias in estimating μ_y depends on proper modeling the probabilities $P(\delta_i = 1|x_i, y_i; \boldsymbol{\gamma}_B)$. The larger is the distance between the true selection model and the assumed selection model the lower will be the performance of the semiparametric

Table 5. Bias (B) and standard deviation (Sd) of $\hat{\mu}_{y,SP}$ over the 500 samples under models A-C, for $\boldsymbol{\gamma}_B = (0.25,0.5)'$. True parameter is $\mu_y = 7$.

Model A		Model B		Model C	
$B(\hat{\mu}_{y,SP})$	$Sd(\hat{\mu}_{y,SP})$	$B(\hat{\mu}_{y,SP})$	$Sd(\hat{\mu}_{y,SP})$	$B(\hat{\mu}_{y,SP})$	$Sd(\hat{\mu}_{y,SP})$
0.89	0.36	1.53	0.18	1.78	0.16

Table 6. RMSE of $\hat{\mu}_{y,SP}$ over the 500 samples under models A-C, for $\boldsymbol{\gamma}_B = (0.25,0.5)'$

$RMSE$ $(\hat{\mu}_{y,SP})$		
Model A	Model B	Model C
0.96	1.54	1.79

approach in removing the bias in μ_y estimator. Same consideration holds for the RMSEs in Table 6. The RMSE value of $\hat{\mu}_{y,SP}$ under model A is lower than that of comparison models, implying that the model A is better.

As previously discussed, the combined model can be tested based on the observations in sample B by standard test statistics because the sample model refers to the observed data. For instance, with regard to the sample used in Figure 1, after having modeled the probabilities $P(\delta_i = 1|\mathbf{x}_i, y_i)$ by the logistic model in Equation (18) (model A) the goodness of fit of the estimated model is tested by the KS statistic. The KS statistic is 0.166 and the critical value corresponding to a significance level $\alpha = 0.05$ is 0.244, then the null hypothesis that the estimated model fits the sample data is not rejected. Recall that the KS statistic when the model for the selection probabilities is assumed known is 0.098 (critical value 0.332) much smaller than 0.166 (critical value 0.244) when the model (18) is assumed. The same consideration holds for model B. Finally, under model C the null hypothesis is rejected. Specifically, the KS statistic is 0.129 and the critical value is 0.112. Finally, setting the significance level $\alpha = 0.01$ both models B and C are rejected. Notice that, the relative bias under model A is 13%. A further reduction in the bias can be obtained introducing additional calibration constraints in the empirical likelihood maximization.

6.2.2. Violations of the Population Normality Assumption

In this section we employ a simulation study to assess the impact associated to violations of the normality assumption on the proposed EL approach. With this regard, two population pdfs are considered:

1. Generate a population of $N = 1,000,000$ observations $(\mathbf{x}_i, y_i)$, where x has a Gamma distribution with shape 3 and scale 1 and $\log(y|x)$ is normal with parameters $\boldsymbol{\theta}_{y|x} = \beta_0 + \beta_1 \mathbf{x}_i$ with $\beta_0 = 0.1$, $\beta_1 = 0.2$, and $\sigma^2_{y|x} = 0.3$. A sample B is selected by a Poisson sampling design with expected sample size $E(n_B) = 0.2N$ and sample inclusion probabilities given by Equation (14) where $\boldsymbol{\gamma}_B = (0.25, 0.5)'$.
2. As in Feder and Pfeffermann (2019), generate a population of $N = 1,000,000$ observations $(\mathbf{x}_i, y_i)$, where x has a Gamma distribution with parameters (2, 2). For each $\mathbf{x}_i$ a binary outcome y_i is generated with $P(y_i = 1|x_i; \boldsymbol{\beta}) = logit^{-1}(-0.8 + 0.8\mathbf{x}_i)$ where $\boldsymbol{\beta} = (-0.8, 0.8)$. Next, a value of a design variable z is generated as $z_i = max[(x_i + 1.1)(2y_i + 1) + v_i; 0.01]$ where v_i follows a uniform distribution $(-0.2, 0.2)$.

   The sample B is drawn by a Poisson sampling with inclusion probability,

$$\pi_i = min(200000 z_i^{-1} / \sum_{j=1}^{N} z_j^{-1}, 0.9999). \tag{19}$$

Finally, 500 samples S of size 2,000 are drawn from B by a *srs* and the logistic model in Equation (18) is used to model the selection probability $P(\delta_i = 1|x_i, y_i)$. Results are shown in Table 7 where the bias (B), the standard deviation (Sd) and the RMSE of the estimates $\hat{\mu}_{y,ISP}$ and $\hat{\mu}_{y,SP}$, obtained under scenarios 3 and 4, are reported.

As Table 7 shows, if the nonprobability sample B is treated as a simple random sample the estimates $\hat{\mu}_{y,ISP}$ obtained maximizing the EL in Equation (15) under the constraints in Equations (8) and (9) are biased. The bias is 1.21 (relative bias 48.4%) and -0.26 (relative bias -0.32%) for the lognormal and the binary case, respectively. A reduction in

Table 7. Bias (B), standard deviation (Sd) and RMSE of $\hat{\mu}_{y,ISP}$ and $\hat{\mu}_{y,SP}$, over the 500 samples for $\gamma_B = (0.25,0.5)'$. True parameter is $\mu_y = 2.5$ for the lognormal variable and $\mu_y = 0.8$ for the binary variable.

Population	$B(\hat{\mu}_{y,ISP})$	$B(\hat{\mu}_{y,SP})$	$Sd(\hat{\mu}_{y,ISP})$	$Sd(\hat{\mu}_{y,SP})$	$RMSE(\hat{\mu}_{y,ISP})$	$RMSE(\hat{\mu}_{y,SP})$
Lognormal	1.21	0.58	0.09	0.05	1.21	0.58
Binary	-0.26	-0.14	0.02	0.01	0.26	0.14

the bias is obtained when the selection process in B is taken into account as the estimates $\hat{\mu}_{y,SP}$ show. More specifically, in the lognormal case the bias decreases to 0.58 (relative bias 0.23), in the binary case to -0.14 (relative bias -0.18). A further reduction in the bias can be obtained introducing additional calibration constraints in the empirical likelihood maximization.

7. An Application to Income Data

In this section the approach based on the EL is applied to real sample data. In Italy, reliable information on households income (y) is provided by the Survey on Household Income and Wealth (SHIW) conducted by the Bank of Italy (Banca d'Italia) every two years. Its main goal is to study the economic status of Italian households, focusing on income and wealth. The sample for the SHIW survey is drawn in two stages, with municipalities and households as, respectively, the primary and secondary sampling units. The primary units are stratified by region and population size. Bigger municipalities (with more than 40,000 inhabitants) are all included in the sample, while the smaller towns are selected using a probability proportional to size sampling (PPS). The individual households to be interviewed are then selected by simple random sampling. In the present article we use the 2010 wave, whose sample consists of 7,951 households and 387 municipalities. The variable of interest is the household income, defined as the combined disposable annual income of all the people living in the household. The average annual household income in 2010 is $\mu_y =$ EUR 32,714, as published by Bank of Italy (Banca d'Italia 2012). To reproduce the situation where a nonprobability sample B and a probability sample A are available the following procedure has been implemented:

1. A sample B is selected from SHIW according to a Poisson sampling design with expected sample size $E(n_B) = 2{,}000$ and unknown inclusion probabilities proportional to $(y_i - min_i y_i + 10)$.
2. Suppose that the Household Budget Survey (HBS) run by Italian National Institute of Statistics, (ISTAT) in 2010 (sample A) which consists of 22,227 households is available. The HBS uses a sampling design similar to SHIW and collects detailed information on sociodemographic characteristics and expenditures on a disaggregated set of commodities (durable and nondurable). Let $\mathbf{x} = (x_1, x_2)$ be the available auxiliary variables, where x_1 is the household size and x_2 is the monthly expenditure on food. Furthermore, let 2.4 and 507.46 be the estimates of the average size of households and the monthly mean expenditure on food in 2010, respectively, as obtained from HBS. Then, we can add the calibration constraint in Equation (10) where,

$$\frac{1}{N}\sum_{i\in A} d_i x_{1i} = 2.4,\ \frac{1}{N}\sum_{i\in A} d_i x_{2i} = 507.46. \tag{20}$$

Step 3 The probabilities $P(\delta_i = 1|\mathbf{x}_i, y_i)$ are modeled by the logistic model $logit^{-1}(\boldsymbol{\gamma}_x{}'\mathbf{x}_i + \gamma_y y_i)$ with $\boldsymbol{\gamma}_x = (\gamma_{x1}, \gamma_{x2})'$ and the average annual household income μ_y is estimated from B by the EL under scenarios 1,3,4 described in Section 5. We recall that under scenario 1 the selection process in B is not taken into account. Scenario 3 is as scenario 1 but we add the calibration constraint in Equation (10). Under scenario 4 we maximize the EL in Equation (7) under the constraints in Equation (9) and the calibration constraint in Equation (10), respectively.

Step 4 Steps 1–3 are repeated 500 times.

For one of the 500 samples B, Figure 2 shows the weighted kernel density of y estimated from SHIW for the purpose of benchmark comparison and the kernel density estimate of y estimated from B. The bandwidth selection rule is as proposed in Sheather and Jones (1991). As clearly seen, both the distributions are right-skewed but the B sample pdf is very different from the SHIW pdf. In Table 8 the bias (B) and the standard deviation (Sd) of the estimates $\hat{\mu}_{y,I}, \hat{\mu}_{y,ISP}$ and $\hat{\mu}_{y,SP}$ obtained under scenarios 1,3,4, over the 500 samples, are reported. Furthermore, in Table 9 the corresponding RMSEs are computed.

As results in Table 8 show, the estimate $\hat{\mu}_{y,ISP}$ ignoring the sampling process in B shows a slight reduction in the bias compared to the estimates $\mu_{y,I}$ because of the introduction of the calibration constraint in Equation (10). A larger reduction in the bias is obtained when the selection process in B is taken into account as the estimates $\hat{\mu}_{y,SP}$ show. Finally, the RMSE of $\hat{\mu}_{y,SP}$ is lower than that of comparison estimators as shown in Table 9.

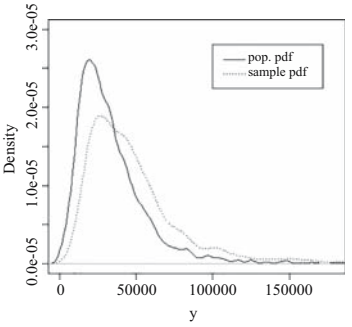

Fig. 2. Income pdf from SHIW data set, sample pdf from B.

Table 8. Bias (B) and standard deviation (Sd) of $\hat{\mu}_{y,I}, \hat{\mu}_{y,ISP}$ and $\hat{\mu}_{y,SP}$, over the 500 samples. True parameter is $\mu_y = 32{,}714$.

$B(\hat{\mu}_{y,I})$	$B(\hat{\mu}_{y,ISP})$	$B(\hat{\mu}_{y,SP})$	$Sd(\hat{\mu}_{y,I})$	$Sd(\hat{\mu}_{y,ISP})$	$Sd(\hat{\mu}_{y,SP})$
15803.01	13331.07	-2527.36	420.49	328.01	261.54

Table 9. RMSE of $\hat{\mu}_{y,I}, \hat{\mu}_{y,ISP}$ and $\hat{\mu}_{y,SP}$, over the 500 samples

$RMSE(\hat{\mu}_{y,I})$	$RMSE(\hat{\mu}_{y,ISP})$	$RMSE(\hat{\mu}_{y,SP})$
15808.60	13335.10	2540.86

8. Simulation Study 2

In this section we evaluate the performance of the mixed approach described in Section 4. We generate a finite population $\{\mathbf{x} = (x_{1i},\ x_{21}),\ \mathbf{y} = (y_{1i},\ y_{2i}):\ i = 1,..,\ N\}$ with size $N = 1{,}000{,}000$ where y_1 is a continuous outcome while y_2 is a binary outcome. From the finite population we select a sample B where the inclusion indicator $\delta_i \sim Ber(p_i)$ with p_i the inclusion probability for unit i. We obtain a representative sample $A = \{x_{1i}, x_{2i}, d_i\}$ of size $n = 1{,}000$ using *srs*. As in Yang et al. (2021a), for generating the finite population we consider the following models,

$$y_{1i} = 1 + x_{1i} + x_{2i} + \alpha_i + \varepsilon_i, \tag{21}$$

$$P(y_{2i} = 1|x_{1i}, x_{2i}; \alpha_i) = logit^{-1}(1 + x_{1i} + x_{2i} + \alpha_i), \tag{22}$$

where $x_1 \sim N(1,\ 1)$, $x_2 \sim Exp(1)$, $\alpha \sim N(0,\ 1)$, $\varepsilon \sim N(0,\ 1)$, and x_1, x_2, α and ε are mutually independent. The variable α induces the dependence of y_1 and y_2 even adjusting for x_1 and x_2. The point biserial correlation coefficient between y_1 and y_2 is 0.32. For the inclusion probability in B, we consider the following logistic linear model,

$$p_i = logit^{-1}(y_{1i}). \tag{23}$$

The expected size of the subsample of units in probability sample A with the membership information ($\delta_i = 1$) is 873. Notice that, in the logistic regression there should be an adequate number of outcomes per predictor variable to avoid an overfit model. Agresti (2007) suggests that there should be ten outcomes for each independent variable. However, the issue has not been definitively settled. We compare the following estimators:

1. $\hat{\mu}_{HT}$, the Horvitz-Thompson estimator assuming that (y_{1i}, y_{2i}) are observed in sample A for the purpose of benchmark comparison.
2. $\hat{\mu}_{NN}$, the nearest neighbor imputation estimator where the imputed values (y_{1i}, y_{2i}) are obtained by nearest neighbor method, as described in section 4.
3. $\hat{\mu}_{RC}$, the regression calibration estimator based on $\hat{\mu}_{NN}$ with calibration variables $H(\delta, \mathbf{x}, \mathbf{y}) = (\delta,\ 1 - \delta,\ \delta\mathbf{x},\ \delta\,\mathbf{y})'$, as described in Yang et al. (2021a).
4. $\hat{\mu}_W$, the Horvitz-Thompson estimator with weights $\tilde{w}_i$ obtained by regressing the membership indicator δ against $(x_{1i}, x_{2i}, \tilde{y}_{1i}, \tilde{y}_{2i})$ in sample A, as described in section 4 (Step 3, point 3.1),
5. $\hat{\mu}_{EL}$, the estimator based on maximization of empirical likelihood, as described in section 4 (Step 3, point 3.2).

The simulation is based on 1,000 Monte Carlo runs. Table 10 summarizes the simulation results with biases, standard deviations and coverage rates of 95% confidence intervals using asymptotic normality of the aforementioned estimators. All the results are multiplied by 100. The population means of y_1 and y_2 are 3 and 0.89, respectively.

First of all, if sample B is simply treated as a simple random sample the bias is 33.1% and 2.3% for the mean of y_1 and y_2, respectively. In Table 10 the estimators $\hat{\mu}_{NN}$ and $\hat{\mu}_{RC}$ have the larger bias, but $\hat{\mu}_{RC}$ has a smaller standard error than $\hat{\mu}_{NN}$. Recall that both estimators implicitly assume that the selection mechanism for sample B is ignorable. With regard to the mean of y_1, $\hat{\mu}_W$ has the smaller bias (13.3%) followed by $\hat{\mu}_{EL}$

Table 10. Bias (B_h), standard deviation (Sd_h), RMSE and coverage rate of 95% confidence interval (CR_h) for the population mean of y_h, $h = 1, 2$, based on 1,000 Monte Carlo samples. The population means of y_1 and y_2 are 3 and 0.89, respectively. All the results are multiplied by 100.

Estimator	B_1	Sd_1	$RMSE_1$	CR_1	B_2	Sd_2	$RMSE_2$	CR_2
$\hat{\mu}_{HT}$	0.2	6.2	6.2	95.4	0.0	1.0	1.0	96.6
$\hat{\mu}_{NN}$	20.2	5.6	21.0	93.8	1.2	0.9	1.5	96.6
$\hat{\mu}_{RC}$	20.1	1.8	20.2	96.0	1.3	0.4	1.4	95.4
$\hat{\mu}_{W}$	13.3	1.9	13.4	95.8	-0.4	0.3	0.5	93.4
$\hat{\mu}_{EL}$	16.12	0.1	16.1	99.4	-0.02	0.3	0.3	94.4

(16.12%) even if $\hat{\mu}_{EL}$ has a smaller standard error and a larger coverage rate ($CR_1 =$ 99.4%). The opposite occurs for the mean of y_2, the bias is $B_2 =$ -0.4% for $\hat{\mu}_W$ against $B_2 =$ -0.02% for $\hat{\mu}_{EL}$. The coverage rates are all close to the nominal level. In conclusion, both μ_W and μ_{EL} reduce the bias in estimating μ_y respect to $\hat{\mu}_{NN}$ and $\hat{\mu}_{RC}$. The Horvitz-Thompson estimator ($\hat{\mu}_W$) seems to perform better for continuous variables while the estimator based on maximization of empirical likelihood ($\hat{\mu}_{EL}$) according the two steps procedure described in Section 4 seems to show a better performance for binary variables.

9. Concluding Remarks

In this article two approaches for reducing selection bias when the selection process is non-ignorable are proposed. The first one based on EL requires to model parametrically the unknown selection probabilities and to maximize the sample likelihood with respect to the sampling and the population parameters. Auxiliary information known for the population or estimable from a probability sample can be incorporated in the maximization process, thus enhancing the precision of the estimators. As previously stressed, the success of the proposed approach depends on proper modeling of the unknown selection probabilities. However, the resulting sample model can be tested from the data by standard test statistics, see Subsection 6.2.1. A broad simulation study illustrates the good performance of the EL approach also when skewed and binary data are considered, see Subsection 6.2.2. Finally, the proposed approach can be extended to the multivariate case when several variables of interest are considered. For variables selection in modeling $P(\delta_i = 1|\mathbf{x}_i, \mathbf{y}_i)$ see Variyath et al. (2010) and Chen et al. (2022). We obviously hope that other researchers will apply our proposed approach with appropriate modifications required for their data.

The second one is a mixed approach based on mass imputation and propensity score adjustment. It requires that the membership to nonprobability sample can be determined throughout the probability sample A. As indicated by the results in Section 8, the method seems to show a good performance in terms of bias, standard error and confidence interval coverage probabilities. Empirical studies with alternative population and selection models are needed to further ascertain the results of the mixed approach obtained in the present article. Finally, new theoretical developments of the present work include the use of proxy variables that can help studying the relationship between y and δ and in particular, help verifying or refuting the ignorability assumption.

10. References

Agresti, A. 2007. *An Introduction to Categorical Data Analysis* (second edition). John Wiley & Sons, Inc., Hoboken: New Jersey.

Babu, G.J., and C.R. Rao. 2004. "Goodness-of-Fit Tests when Parameters are Estimated." *Sankhyā*. Series A 66(1): 63–74. DOI: https://doi.org/10.2307/25053332.

Banca d'Italia. 2012. *Supplement to the Statistical Bulletin, Sample Surveys, Household income and wealth in 2010*: 12(6). Available at: https://www.bancaditalia.it/pubblicazioni/indagine-famiglie/bil-fam2010.

Beaumont, J.F. 2000. "An Estimation Method for Nonignorable Nonresponse." *Survey Methodology* 26(2): 131–136. Available at: https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2000002/article/5532-eng.pdf?st=WJWdN_3l.

Belzile, L., J.L. Wadsworth, P.J. Northrop, S.D. Grimshaw, J. Zhang, M.A. Stephens, A.B. Owen, and R. Huser. 2022. *mev: Modelling Extreme Values*. R package version 1.14 Available at: https://CRAN.R-project.org/package = mev (accessed June 2022).

Beresewicz, M., R. Lehtonen, F. Reis, L. Di Consiglio and M. Karlberg. 2018. *An overview of methods for treating selectivity in big data sources*. Statistical Working Papers, Eurostat. Available at: https://ec.europa.eu/eurostat/web/products-statistical-working-papers/-/ks-tc-18-004 (accessed July 2022).

Chang, T., and P.S. Kott. 2008. "Using Calibration Weighting to Adjust for Nonresponse under a Plausible Model." *Biometrika* 95(3): 555–571. DOI: https://doi.org/10.1093/-biomet/asn022.

Chaudhuri S., M.S. Handcock, and M.S. Rendall. 2010. *A conditional empirical likelihood approach to combine sampling design and population level information*. Technical report No. 3/2010, National University of Singapore, Singapore. Available at: https://cpb-us-w2.wpmucdn.com/blog.nus.edu.sg/dist/0/14452/files/2020/10/tr03 2010.pdf (accessed July 2022).

Chen, C., M. Wang, R. Wu, and R. Li. 2022. "A Robust Consistent Information Criterion for Model Selection Based on Empirical Likelihood." *Statistica Sinica* 32: 1205–1223. DOI: https://doi.org/10.5705/ss.202020.0254.

Conti P.L., D. Marella, and M. Scanu. 2008. "Evaluation of matching noise for imputation techniques based on nonparametric local linear regression estimators." *Computational Statistics & Data Analysis* 53(2): 354–365. DOI: https://doi.org/10.1016/j.csda.2008.07.041.

DiSogra, C., C. Cobb, E. Chan, and J. M. Dennis. 2011. "Calibrating non-probability internet samples with probability samples using early adopter characteristics." In Proceedings of the Section on Survey Research Methods, Joint Statistical Meetings. Miami Beach, Florida, July 30-August 4, 2011: 4501–4515. Alexandria, VA: American Statistical Association. Available at: http://www.asasrms.org/Proceedings/y2011/Files/30270468925.pdf (accessed June 2022).

Elliott, M., and R. Valliant. 2017. "Inference for non-probability samples." *Statistical Science* 32(2): 249–264. DOI: https://doi.org/10.1214/16-STS598.

Feder, M., and D. Pfeffermann. 2019. *Statistical Inference Under Non-ignorable Sampling and Non-response. An Empirical Likelihood Approach*. Working paper. University of Southampton. Available at: https://eprints.soton.ac.uk/378245/ (accessed July 2022).

Galimard J.E., S. Chevret, E. Curis, and M. Resche-Rigon. 2018. "Heckman imputation models for binary or continuous MNAR outcomes and MAR predictors." *BMC Medical Research Methodology* 18(90). DOI: https://doi.org/10.1186/s12874-018-0547-1.

Hájek, J. 1964. "Asymptotic theory of rejective sampling with varying probabilities from a finite population." *The Annals of Mathematical Statistics* 35(4): 1491–1523. DOI: 10.1214/aoms/1177700375.

Heckman, J.J. 1979. "Sample Selection Bias as a Specification Error." *Econometrica* 47(1): 153–161. DOI: http://dx.doi.org/10.2307/1912352.

Kim, J.K., and Z. Wang. 2019. "Sampling techniques for big data analysis in finite population inference." *International Statistical Review* 87(S1): S177–S191. DOI: https://doi.org/10.1111/insr.12290.

Kott, P.S. 2006. "Using calibration weighting to adjust for nonresponse and coverage errors." *Survey Methodology* 32(2): 133–142. Available at: https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2006002/article/9547-eng.pdf?st=B2aZNvo0.

Kott, P.S., and T. Chang. 2010. "Using Calibration Weighting to Adjust for Nonignorable Unit Nonresponse." *Journal of the American Statistical Association* 105(491): 1265–1275. DOI: https://doi.org/10.1198/jasa.2010.tm09016.

Lee, J., and J.O. Berger. 2001. "Semiparametric Bayesian analysis of selection models." *Journal of the American Statistical Association* 96(456): 1397–1409. DOI: https://doi.org/10.1198/016214501753382318.

Marella D., M. Scanu, and P.L. Conti. 2008. "On the matching noise of some nonparametric imputation procedures." *Statistics & Probability Letters* 78(12): 1593–1600. DOI: https://doi.org/10.1016/j.spl.2008.01.020.

Marella, D., and D. Pfeffermann. 2019 "Matching Information from two independent informative samples." *Journal of Statistical Planning and Inference* 203: 70–81. https://doi.org/10.1016/j.jspi.2019.03.001.

Marella, D., and D. Pfeffermann. 2021 "Accounting for nonignorable sampling and nonresponse in statistical matching." *International Statistical Review*. *Accepted for pubblication*. DOI: https://doi.org/10.1111/insr.12524.

Meng, X-L., 2018. "Statistical paradises and paradoxes in big data (I): law of large populations, big data paradox and the 2016 US presidential election." *The Annals of Applied Statistics* 12(2): 685–726. DOI: https://doi.org/10.1214/18-AOAS1161SF.

Owen, A.B. 2001. *Empirical Likelihood*. Chapman & Hall/CRC: New York.

Owen, A.B. 2013. "Self-concordance for empirical likelihood." *Canadian Journal of Statistics* 41(3): 387–397. DOI: https://doi.org/10.1002/cjs.11183.

Pfeffermann, D., A.M. Krieger, and Y. Rinott. 1998. "Parametric distribution of complex survey data under informative probability sampling." *Statistica Sinica* 8(4): 1087–1114.

Pfeffermann, D., and M. Sverchkov. 2009. "Inference under Informative Sampling." In *Handbook of Statistics 29B: Sample Surveys: Inference and Analysis*, edited by D. Pfeffermann and C.R. Rao.: 455–487. North Holland.

Pfeffermann, D. 2011. "Modelling of complex survey data: Why model? Why is it a problem? How can we approach it?" *Survey Methodology* 37(2): 115–136. Available at: https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2011002/article/11602-eng.pdf?st=_vXrCcPb.

Pfeffermann, D., and V. Landsman. 2011. "Are private schools really better than public schools? Assessment by methods for observational studies." *Annals of Applied Statistics* 5(3): 1726–1751. DOI: https://doi.org/10.1214/11-AOAS456.

Pfeffermann, D., and A. Sikov. 2011. "Imputation and Estimation under Nonignorable Non-response in Household Surveys with Missing Covariate Information." *Journal of Official Statistics* 27(2): 181–209. Available at: https://www.scb.se/contentassets/ff271eeeca694f47ae99b942de61df83/imputation-and-estimation-under-nonignorable-nonresponse-in-household-surveys-with-missing-covariate-information.pdf.

Pfeffermann, D. 2015. "Methodological issues and challenges in the production of official statistics: 24th Annual Morris Hansen Lecture." *Journal of Survey Statistics and Methodology* 3(4): 425–483. DOI: https://doi.org/10.1093/jssam/smv035.

R Core Team 2021. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Available at: http://www.R-project.org/.

Riddles, M.K., J. K. Kim, and J. Im. 2016. "A propensity-score-adjustment method for non-ignorable nonresponse." *Journal of Survey Statistics and Methodology"* 4(2): 215–245. DOI: https://doi.org/10.1093/jssam/smv047.

Rivers, D. 2007. "Sampling for web surveys." In Proceedings of the Section on Survey Research Methods, Joint Statistical Meetings. Salt Lake City, Utah, July 29-August 2, 2007: 4127–4134. Alexandria, VA: American Statistical Association. Available at: http://www.websm.org/uploadi/editor/1368187629Rivers 2007 Sampling for web surveys.pdf (accessed June 2022).

Rosenbaum, P.R., and D.B. Rubin. 1983. "The central role of the propensity score in observational studies for causal effects." *Biometrika* 70(1): 41–55. DOI: https://doi.org/10.1093/biomet/70.1.41.

Rubin, D.B. 1976. "Inference and missing data." *Biometrika* 63(3): 581–592. DOI: https://doi.org/10.1093/biomet/63.3.581.

Sheather, S.J., and M.C. Jones. 1991. "A reliable data-based bandwidth selection method for Kernel density estimation." *Journal of the Royal Statistical Society. Series B-Statistical Methodology* 53(3): 683–690. DOI: https://doi.org/10.2307/2345597.

Variyath, A. M., J. Chen, and B. Abraham. 2010. "Empirical likelihood based variable selection." *Journal of Statistical Planning and Inference* 140(4): 971–981. DOI: https://doi.org/10.1016/j.jspi.2009.09.025.

Yang, S., J.K. Kim, and Y. Hwang. 2021a. "Integration of data from probability surveys and big found data for finite population inference using mass imputation." *Survey Methodology* 47(1): 29–58. Available at: https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2021001/article/00004-eng.pdf?st=-WLDQdr7.

Yang, S., J.K. Kim, and R. Song. 2021b. "Doubly Robust Inference when Combining Probability and Nonprobability Samples with High-dimensional Data." *Journal of the Royal Statistical Society. Series B-Statistical Methodology* 82(2): 445–465. DOI: https://doi.org/10.1111/rssb.12354.

Received July 2021
Revised April 2022
Accepted July 2022