mathematics

MDPI

Article

Estimating General Parameters from Non-Probability Surveys Using Propensity Score Adjustment

Luis Castro-Martín, María del Mar Rueda * and Ramón Ferri-García

Department of Statistics and O.R., University of Granada, 18071 Granada, Spain; luiscastro193@ugr.es (L.C.-M.); rferri@ugr.es (R.F.-G.)

* Correspondence: mrueda@ugr.es

Received: 20 October 2020; Accepted: 21 November 2020; Published: 23 November 2020

Abstract: This study introduces a general framework on inference for a general parameter using nonprobability survey data when a probability sample with auxiliary variables, common to both samples, is available. The proposed framework covers parameters from inequality measures and distribution function estimates but the scope of the paper is broader. We develop a rigorous framework for general parameter estimation by solving survey weighted estimating equations which involve propensity score estimation for units in the non-probability sample. This development includes the expression of the variance estimator, as well as some alternatives which are discussed under the proposed framework. We carried a simulation study using data from a real-world survey, on which the application of the estimation methods showed the effectiveness of the proposed design-based inference on several general parameters.

Keywords: nonprobability surveys; propensity score adjustment; survey sampling

MSC: 62D05

1. Introduction

Nonprobability samples are increasingly common in empirical sciences. The rise of online and smartphone surveys, along with the decrease of response rates in traditional survey modes, have contributed to the popularization of volunteer surveys where sampling is non-probabilistic. Moreover, the development of Big Data involves the analysis of large scale datasets whose obtention is conditioned by data availability and not by a probabilistic selection, and therefore they can be considered large nonprobability samples of a population [1].

The lack of a probability sampling scheme can be responsible for selection bias. Following the description from [1,2], we can distinguish the target population, U_T, the subpopulation that a given selection method can potentially cover, U_{pc}, and the fraction of the subpopulation that is finally covered, U_{fc}, and whose individuals might participate in the survey. Selection bias occurs when the characteristics of the individuals in U_{fc} differ significantly from those in U_T in a way that could affect final estimates. Typically, differences between individuals in U_T and individuals in U_{pc} are caused by a lack of coverage induced by the survey administration mode (for example, an online questionnaire cannot be administered to the population without internet access), while differences between U_{pc} and U_{fc} are caused by the variability in the propensities to participate between social-demographic groups (for example, an online questionnaire accesible in a thematic website might only be fulfilled by visitors of the website who have a specific interests that could influence the results).

Following the rise of nonprobability samples, a class of methods for reducing selection bias have been proposed in the last decades. These methods were developed from different perspectives according to the availability of auxiliary information. We can mention calibration, Propensity Score

Adjustment (PSA), Statistical Matching and superpopulation modelling as the most relevant techniques to mitigate selection bias produced by coverage and self-selection errors.

Calibration weighting was originally developed by [3] as a method to correct representation issues in samples with coverage or non-response errors. It only requires a vector of auxiliary variables available for each individual of the sample and the population totals of those variables. Calibration is able to remove selection bias in nonprobability samples if the selection mechanism is ignorable [4], and despite being originally developed for parametric estimation, further work [5–7] has extended calibration to distribution function, quantile and poverty measures estimation.

Propensity Score Adjustment (PSA) and Statistical Matching require, apart from the nonprobability sample, a probability sample to do the adjustments. PSA was originally developed for balancing groups in non-randomized clinical trials [8] and it was adapted for non-response adjustments shortly after [9,10]. The application of PSA for removing bias in nonprobability surveys was theoretically developed in [11,12]. Statistical Matching was firstly proposed in [13] and extended in [14] for non-response adjustments. The difference between both methods is the sample used in the estimators: PSA estimates the propensity of each individual of the nonprobability sample to participate in the survey and then this propensity is used to construct the weights of the estimators, while Statistical Matching adjusts a prediction model using data from the nonprobability sample, applies it in the probability sample to predict their values for the target variable y and uses them in the parametric estimators. To the best of our knowledge, PSA and Statistical Matching has not been developed for nonparametric estimation.

Superpopulation modelling requires data from the complete census of the target population for the covariates used in the adjustment, which is assumed to be a realization (sample) of a superpopulation where the (unknown) target values follow a model. It is based on the works by [15,16], where the main idea is to fit a regression model on the target variable with data from the nonprobability sample, and use the model to predict the values of the target variable for each individual in the population. The prediction can be used for estimation using a model-based approach or some alternative versions such as model-assisted and model-calibrated. LASSO models [17] and Machine Learning predictors [18,19] have been studied as alternatives to ordinary least squares regression in superpopulation modelling.

The interest of society on poverty and inequality has increased in the last decades given the successive economic cycles and crisis. In such a context, official poverty rates and the percentage of people in poverty (or under a poverty threshold) are some important measures of a country's wealth. The common characteristic of many poverty measures is their complexity. The literature on survey sampling is usually focused on the goal of estimating linear parameters. However, it is usual that the variable of interest in poverty studies is a measure of wages or income, where the distribution function becomes a relevant tool because it is required to calculate the proportion of people with low income, the poverty gap and other measures. Estimators for the cumulative distribution function, quantiles [20,21] and poverty measures [22] can be found in literature regarding probability samples, but there is hardly any work on the estimation of these parameters when the samples are obtained from volunteers.

In this paper, we aim to develop a framework for statistical inference on a general parameter with non probability survey samples when a reference probability sample is available. After introducing the problem of the mean estimation for volunteer samples in Section 2, in Section 3, we consider the problem of the estimation for a general parameter through general estimating equations. Section 4 presents a new estimator for a general parameter through the use of PSA to estimate the propensity score of each individual in the survey weighted estimating equation and major theoretical results are presented. Results from simulation studies are reported in Sections 5 and 6 presents the concluding remarks.

2. Approaches to Estimation of a Mean for Volunteer Online Samples

Let U_T be the target population with N elements and s_v a nonprobability sample drawn from a subset of U_T, U_v, with a size of $n_v \leq N$. Let y be the target variable of the survey, whose mean in the

population U_T is denoted as $\overline{Y}$. The sample estimation of $\overline{Y}$, $\hat{\overline{Y}}$, is done using the Horvitz-Thompson estimator:

$$\hat{\overline{Y}}_{HT} = \frac{\sum_{i \in s_v} w_i y_i}{\sum_{i \in s_v} w_i} \tag{1}$$

where w is a vector of weights that accounts for the lack of representativity of s_v caused by selection bias. If no auxiliary information is given, the weight would be the same for every unit, $w_i = N/n_v$, which requires to assume that the sample was drawn under a simple random sampling scheme. This is a naïve assumption given that s_v is not probabilistic, this is, the probability of being in the sample is unknown and/or null for any of the units in U_T.

Let $\mathbf{x}$ be a matrix of covariates measured in s_v along with y. If the population totals of the covariates, $\mathbf{X}$, are available, it is possible to estimate the mean using a vector of weights obtained with calibration, w^{CAL}. The calibration weights aim to minimize the distance between the original and the new weights

$$\min_{w_i^{CAL}} E\left[\sum_{i \in s_v} G(w_i, w_i^{CAL})\right] \tag{2}$$

while respecting the calibration equations

$$\sum_{i \in s_v} w_i^{CAL} \mathbf{x}_i = \mathbf{X}. \tag{3}$$

Some choices for the distance $G(.,.)$ were listed in [3], along with the resulting estimators. Calibration weighting for selection bias treatment was studied in [4], where post-stratification, which is a special case of calibration [23], was used to mitigate the bias caused by different selection mechanisms, showing its efficacy when the selection of the units of s_v is Missing At Random (MAR).

If a reference sample, s_r, drawn from the population U_T is available and a number of covariates $\mathbf{x}$ have been measured both in s_v and s_r, two procedures can be done to reduce selection bias present in s_v. Let I_v be an indicator variable of an element being in s_v, this is

$$I_{vi} = \begin{cases} 1 & i \in s_v \\ 0 & i \notin s_v \end{cases} \tag{4}$$

Propensity Score Adjustment (PSA) assumes that each element of U_T has a probability (propensity) of being selected for s_v which can be formulated as

$$\pi_i^v = Pr(I_{vi} = 1 | \mathbf{x}_i, y_i) \tag{5}$$

where π_i^v is the propensity of the i-th individual to participate in s_v. The random mechanism behind this probability is the selection mechanism that governs the nonprobability sample. If the selection is Missing Completely At Random (MCAR), then $\pi_i^v = Pr(I_{vi} = 1)$ and the selection bias is null, while if the selection is MAR then $\pi_i^v = Pr(I_{vi} = 1 | \mathbf{x_i})$ and the selection mechanism is considered ignorable. This does not mean that the selection bias should be ignored but rather it can be treated with the right techniques.

In PSA, we consider the situation where true propensities are not known and therefore have to be estimated; we do it by combining s_v and s_r into a sample. The probability that $I_v = 1$ is then estimated using a prediction model, traditionally a logistic regression one:

$$\hat{\pi}_i^v = \frac{1}{1 + exp\{-\beta \mathbf{x_i}\}} \tag{6}$$

Alternative models, such as non-linear regression and Machine Learning classification algorithms, have been studied in literature as a substitute of logistic regression (see [24] for a review). The resulting propensities can be used to adjust new weights, w^{PSA}, with different alternatives:

- A simple inverse probability weighting is proposed by [25]

$$w_i^{PSA1} = \frac{w_i}{\hat{\pi}_i^v} \tag{7}$$

  which is a similar approach to the formula used in [26]

$$w_i^{PSA2} = \frac{1 - \hat{\pi}_i^v}{\hat{\pi}_i^v} \tag{8}$$

- Alternatively, individuals of the combined sample ($s_v \cup s_r$) can be grouped in g equally-sized strata of similar propensity scores from which an average propensity is calculated for each group. Let $\overline{\pi}_g$ be the mean propensities of the g-th strata. [2] use the means as in (7) to calculate the new weights:

$$w_i^{PSA3} = \frac{w_i}{\overline{\pi}_{g_i}} \tag{9}$$

  where g_i refers to the strata to which the i-th individual of s_v belongs.
- A similar approach can be found in [12], but instead of using the means, a factor is calculated for each strata:

$$f_g = \frac{\sum_{k \in s_{r_g}} \tilde{w}_k / \sum_{k \in s_r} \tilde{w}_k}{\sum_{i \in s_{v_g}} w_i / \sum_{i \in s_v} w_i} \tag{10}$$

  where s_{r_g} and s_{v_g} are respectively the individuals from the probability and nonprobability sample that belong to the g-th strata, and $\tilde{w}$ is the vector of design weigths of the reference sample. The final weights are obtained by multiplying the original weights and the correction factor:

$$w_i^{PSA4} = w_i \cdot f_{g_i} \tag{11}$$

PSA has been proven to successfully remove selection bias when prognostic covariates are chosen [11] and further adjustments, such as calibration, are applied in the estimations [2,12,27]. A recent paper [28] shows a real application of PSA in web panel surveys where the reductions in bias, although present, were not large enough to consider the estimates as unbiased.

As an alternative to PSA, Statistical Matching is another method to mitigate selection bias when a reference sample is available. For the matter, a prediction model for y using $\mathbf{x}$ as the dependent variables is built using data from s_v. The model is subsequently applied on the reference sample to obtain the estimates from the predicted values of y in s_r, $\hat{y}$:

$$\hat{\bar{Y}} = \sum_{k \in s_r} w_k \hat{y}_k \tag{12}$$

The choice of prediction models has been studied in literature; the usual method is linear regression but other approaches such as donor imputation [13] or Machine Learning algorithms [19,29] have been listed as alternatives. Under certain conditions, Statistical Matching can reduce bias and mean square error to a greater extent than PSA [29].

When a complete census of the entire target population is available, with information on the covariates present in s_v, superpopulation modelling can be applied to remove selection bias [19]. In this paper we consider the case when auxiliary information is available only from a reference probability survey.

3. Estimation of a General Parameter by Using PSA

Let y be the variable of interest in a survey and y_i be the value of the i-th unit in that variable, $i = 1, ..., N$. Suppose we want to estimate a finite population parameter θ_N of dimension $p \geq 1$ defined as the solution of the census estimating equations:

$$U(\theta_N) = \frac{1}{N} \sum_U u_i(y_i, \theta_N) = \mathbf{0} \tag{13}$$

where $u_i(y_i, \theta_N)$ is be a function of θ_N. Some unidimensional parameters of interest can be:

- the population total T_y for $u_i = (y_i - \theta_N/N)$,
- the population mean $\bar{Y}$ for $u_i = (y_i - \theta_N)$,
- the population distribution function $F_y(t)$ for $u_i = (1(y_i \leq t) - \theta_N)$ with $1(\cdot)$ being the indicator function,
- the finite population quantile of order j, Q_j for $u_i = (1(y_i \leq \theta_N) - j$, where $0 < j < 1$,

We denote by $\hat{\theta}$ the solution of the equation:

$$\hat{U}(\theta_N) = \sum_U I_{vi} u_i(y_i, \theta_N)/\pi_i^v = \sum_{s_v} u_i(y_i, \theta_N)/\pi_i^v = \mathbf{0}. \tag{14}$$

It is clear the $E_r(\hat{U}(\theta_N)) = U(\theta_N)$ where r stands for the model of the selection mechanism for the sample s_v, this is, the true model that fits propensity scores. If π_i^v are known we can get the consistent estimator of θ_N by solving the equation above. For the study of the properties of this estimator we consider a quasi-probability approach or pseudo-design-based inference ([19]) and we treat the volunteer sample as a realization of a Poisson sampling with probabilities π_i^v.

For any sample design that verifies certain regularity conditions, the solution to $\widehat{U}(\theta) = \mathbf{0}$ provides a consistent estimator for the parameter θ_N (see [30]). Poisson sampling verifies these conditions, so that the consistency of the estimator is obtained immediately from the result of [30]. The normality of the estimator is demonstrated by [31], who also obtains the asymptotic variance of the estimator. From said expression and taking into account that in Poisson sampling the extractions are independent and therefore the probability of second order is given by $\pi_{ij}^v = \pi_i^v \pi_j^v$ we can obtain the variance of $\hat{\theta}$:

$$V(\hat{\theta}) = J(\hat{\theta})^{-1} var(\hat{U}(\theta)) J'(\hat{\theta})^{-1} \tag{15}$$

being $J(\theta) = \frac{1}{N} \sum_U \partial u_i/\partial\theta$ and $var(\hat{U}(\theta)) = \sum_U (1 - \pi_i^v) u_i^2/(\pi_i^v)^2$

4. Estimation of a General Parameter with Estimated Propensities

The propensity scores π_i^v are not known are impossible to estimate using the nonprobability sample s_v alone, so additional information must be included. Let s_r be a reference probability sample, of size n_r, selected from U_T under a sampling design (s_d, p_d) where the first order inclusion probabilities, $\pi_i^p = \sum_{s_r \ni i} p_d(s_r), i = 1, ..., n_r$, are known and non-null.

The covariates of the propensity model $\mathbf{x}$ have been measured both in s_v and s_r, while the variable of interest y is only available for those individuals in s_v.

Suppose that the propensity scores can be modelled parametrically as

$$\pi_i^v = P(I_{vi} = 1/\mathbf{x}_i) = m(\lambda_o, \mathbf{x}_i) \;\; i = 1, ..., N \tag{16}$$

for some known function $m(\cdot)$ with second continuous derivatives with respect to an unknown parameter λ_o.

We estimate the propensity scores by using data of both the volunteer and the probability sample. The maximum likelihood estimator (MLE) of π_i^v is $m(\hat{\lambda}, \mathbf{x}_i)$ where $\hat{\lambda}$ corresponds to the value of lambda that maximizes the log-likelihood function:

$$l(\lambda) = \sum_{U}(I_{vi}log(m(\lambda, \mathbf{x}_i)) + (1 - I_{vi})log(1 - m(\lambda, \mathbf{x}_i)) =$$

$$\sum_{s_v} log\frac{m(\lambda, \mathbf{x}_i)}{1 - m(\lambda, \mathbf{x}_i)} + \sum_{U} log(1 - m(\lambda, \mathbf{x}_i)). \tag{17}$$

As it is usual in survey sampling, we consider the pseudo-likelihood given that some units of the population have not been sampled:

$$\tilde{l}(\lambda) = \sum_{s_v} log\frac{m(\lambda, \mathbf{x}_i)}{1 - m(\lambda, \mathbf{x}_i)} + \sum_{s_p} \frac{1}{\pi_i^p} log(1 - m(\lambda, \mathbf{x}_i)). \tag{18}$$

We propose thus a two phase procedure in this manner:

Step 1: Calculate $\hat{\lambda}_{pl}$ by solving the score equations:

$$\partial \tilde{l}(\mathbf{x}_i, \lambda)/\partial\lambda = 0$$

Step 2: Calculate $\hat{\theta}_v$ as the solution of the estimating function:

$$\hat{U}_V(\theta) = \sum_{U} I_{vi} u_i(y_i, \theta) \frac{1}{m(\hat{\lambda}_{pl}, \mathbf{x}_i)} = 0 \tag{19}$$

We consider the following asymptotic framework for theoretical development, which is equivalent to the framework in [32]. Let $U_{T\nu}$ be a sequence of finite populations of size N_ν. Each $U_{T\nu}$ has an associated non-probability sample $s_{v\nu}$ of size $n_{v\nu}$ and an associated probability sample $s_{p\nu}$ of size $n_{p\nu}$. We consider that the population size $N_\nu \to \infty$, the nonprobability sample size $n_{v\nu} \to \infty$ and the probability sample size $n_{p\nu} \to \infty$ as $\nu \to \infty$. For notational simplicity the index ν is suppressed for the rest of the paper. The properties of the estimator $\hat{\theta}_v$ are developed under both the model for the propensity scores and the survey design for the probability sample.

We make the following assumptions:

- A.1. The estimating function $u_i(y_i, \theta, \lambda)$ is twice differentiable with respect to θ and λ.
- A.2. The propensities and the sampling design ensure that $\hat{U}_V(\theta) - U(\theta) = O_p(n^{-1/2})$ for any $\theta \in \Theta$.
- A.3. The propensities and the sampling design ensure that $\hat{U}_V(\theta)$ is asymptotically Normal with mean $\mathbf{U}(\theta)$ and entries of the variance at the order $O(n^{-1})$ for any fixed $\theta \in \Theta$.

Theorem 1. *Under the conditions A.1, A.2 and A.3, $\hat{\theta}_v$ is a consistent and asymptotically normal estimator for θ.*

Proof. Under assumed conditions,

$\hat{U}_V(\theta) = U(\theta) + O_p(n^{-1/2})$, thus by using the mean value theorem, $\hat{\theta}_v$ has the same asymptotic behaviour that $\hat{\theta}$ which is consistent for θ and asymptotically normal distributed (see Section 3). □

Variance estimation for $\hat{\theta}_v$ can be handled by combining the two estimating equations, $\tilde{l}$ and $\hat{U}_v$, into a single system as it is done in [33].

The MLE of λ, $\hat{\lambda}_{pl}$ is the solution to the equations:

$$U_2(\lambda) = \sum_{s_v} \partial log\frac{m(\lambda, \mathbf{x}_i)}{1 - m(\lambda, \mathbf{x}_i)}/\partial\lambda + \sum_{s_p} \partial \frac{1}{\pi_i^p} log(1 - m(\lambda, \mathbf{x}_i))/\partial\lambda = 0$$

and the PSA estimator of θ_N is the solution to the estimating equations

$$U_1(\theta,\lambda) = \sum_{s_v} u_i(y_i,\theta)\frac{1}{m(\lambda_{pl},\mathbf{x}_i)} = \sum_{s_v} g_1(y_i,\mathbf{x}_i,\theta,\lambda) = 0.$$

Let $\mathbf{U}(\theta,\lambda) = (U_1'(\theta,\lambda), U_2'(\lambda))'$. Let $\boldsymbol{\psi} = (\theta_N', \lambda_o)'$ be the true parameter values defined through the census estimating equations and $\hat{\psi} = (\hat{\theta}_N', \hat{\lambda_o}')'$ the solutions to $\mathbf{U}(\theta,\lambda) = 0$.

We need an additional assumption:

- A.4. The propensities, the sampling design and the estimating function satisfy $\partial\widehat{\mathbf{U}}/\partial\boldsymbol{\psi} = O_p(1)$ and $\partial^2\widehat{\mathbf{U}}/\partial\boldsymbol{\psi}\partial\boldsymbol{\psi}' = O_p(1)$.

Theorem 2. *Under the conditions A.1, A.2, A.3 and A.4, the asymptotic variance-covariance matrix of* $\hat{\boldsymbol{\psi}}$ *is given by the expression:*

$$\mathbf{V}(\hat{\boldsymbol{\psi}}) = \mathbf{H}^{-1}V(\hat{\mathbf{U}}(\theta,\lambda))\mathbf{H}'^{-1} \tag{20}$$

with $\mathbf{H} = \begin{pmatrix} H_{11} & H_{12} \\ 0 & H_{22} \end{pmatrix}$

$$H_{11} = E\{\frac{\partial}{\partial\theta}U_1(\theta_N,\lambda)\}$$

$$H_{21} = E\{\frac{\partial}{\partial\lambda}U_1(\theta_N,\lambda)\}$$

$$H_{22} = E\{\frac{\partial}{\partial\lambda}U_2(\lambda)\}$$

Proof. Since $\hat{\theta}_v$ and $\hat{\lambda}$ are consistent estimator of respective parameters, we can write $\hat{\boldsymbol{\psi}} = \boldsymbol{\psi} + O_p(1)$ and the Taylor series expansion gives:

$$\hat{\boldsymbol{\psi}} = \boldsymbol{\psi} - \mathbf{H}^{-1}\widehat{\mathbf{U}}(\theta,\lambda) + O_p(\|\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}\|^2),$$

Thus the asymptotic variance of $\hat{\psi}$ is given by:

$$\mathbf{V}(\hat{\boldsymbol{\psi}}) = \mathbf{H}^{-1}V(\widehat{\mathbf{U}}(\theta,\lambda))\mathbf{H}'^{-1}.$$

Taking into account the two random mechanisms, and the probabilities of the conditional expectation, we have $V(\widehat{\mathbf{U}}(\theta,\lambda)) = V_pE_r(\widehat{\mathbf{U}}(\theta,\lambda)) + E_pV_r(\widehat{\mathbf{U}}(\theta,\lambda))$ where r stands for the model of the selection mechanism for the sample s_v and p refers to the probability sampling design for s_p. □

The asymptotic variance of $\hat{\boldsymbol{\psi}}$ depends on the probability of selecting the sample s_p under the given sampling design and the selection mechanism described by the propensity model. Plug-in estimators can be used to construct variance estimators for all the required components but it is not a simple issue.

In practice, and as described in [7], the use of jackknife [34] and bootstrap techniques [35] in the variance estimation for nonlinear parameters should be more advantageous because of their wide applicability for different cases and conditions. Direct applications of bootstrap methods for estimating the variance-covariance matrix of $\hat{\boldsymbol{\psi}}$ involve solving the equation $\mathbf{U}(\theta,\lambda) = 0$ repeatedly for each bootstrap sample. Multiplier Bootstrap with Estimating Functions was proposed by [36].

5. Simulation Study

5.1. Data

Data for the simulation study come from a wave of the Spanish Living Conditions Survey collected between 2011 and 2012 [37], which contains an annual thematic module that, in 2012, was dedicated to

household conditions. The survey sampling follows a two-phase cluster sampling, where the primary units are the households and the secondary units are their members. In 2012, the final sample included 33,573 individuals. For this study, the dataset was filtered to rule out individuals and variables with high quantities of missing data. After this procedure, the dataset employed as pseudopopulation of the study had a size of N = 28,210 individuals and p = 60 available variables.

From this pseudopopulation, two probability samples of size n_r were drawn according to the following sampling strategies:

- The first sample, s_{r1}, was drawn with a stratified cluster sampling, where the strata were defined by the Autonomous Communities (NUTS2 regions) and the clusters were the households, which were drawn with probabilities proportional to the household size. The number of households to be selected, m, was estimated dividing n_r by the medium household size in order to reach the aforementioned size of $n_r = 2000$, resulting in $m = 902$ households. The final sample size of s_{r1} was $n_{r1} = 2003$.
- The second sample, s_{r2}, was drawn with an unequal probability sampling, where probabilities were proportional to the minimum income of the individual's household to make ends meet (variable HS130 in [37]).

The extraction of the nonprobability sample, s_v, was done with unequal probability sampling from the full pseudopopulation, where the probability of selection for the i-th individual, p_i, was given by the formula:

$$p_i = \frac{1}{1 + exp(-2x_i^1 + 0.2x_i^2 + 0.01x_i^3 + 0.2x_i^{41} + 0.4x_i^{42})} \tag{21}$$

where

- $x_i^1 = 1$ when the i-th sampled individual has a computer at home, and $x_i^1 = 0$ otherwise.
- $x_i^2 = 1$ when the i-th sampled individual is a man, and $x_i^2 = 0$ otherwise.
- x_i^3 is the age (in years) of the i-th sampled individual.
- $x_i^{41} = 1$ when the i-th sampled individual lives in a medium population density area, and $x_i^{41} = 0$ otherwise.
- $x_i^{42} = 1$ when the i-th sampled individual lives in a low population density area, and $x_i^{42} = 0$ otherwise.

The reasoning behind this sampling procedure is to take into account more similar mechanisms to self-selection procedures that take place in real nonprobability surveys.

We have considered three different sample sizes, n_v = 2000, 4000, 6000. 1000 simulation runs were performed for each procedure and sample size, drawing a sample in each run.

5.2. Simulation

In each simulation, the parameters to be estimated were the following:

- The Gini coefficient [38], which measures the income inequality, estimated as

$$\widehat{G}_y = \frac{\sum_{k \in s_v} \frac{1}{\pi_k} (2\widehat{F}_y(y_k) - 1) y_k}{\sum_{k \in s_v} y_k / \pi_k}$$

- The proportion of individuals with a disposable income below the at-risk-at-poverty threshold. This measure can be referred to as poverty incidence, poverty proportion, poverty risk or HCI ([39] and is estimated as

$$\widehat{HCI} = \frac{1}{N} \sum_{k \in s_v} \frac{1}{\pi_k} I(y < 0.6Q_{0.5})$$

- The interquartile range, estimated as

$$\widehat{IQR} = \frac{\widehat{Q}_{0.75}}{\widehat{Q}_{0.25}}$$

- The interdecile range, estimated as

$$\widehat{IDR} = \frac{\widehat{Q}_{0.9}}{\widehat{Q}_{0.1}}.$$

Every parameter was estimated with and without applying PSA so we could evaluate its performance. In order to estimate the propensities, a logistic regression model was chosen:

$$m(\hat{\lambda}, \mathbf{x}_i) = \frac{exp(\hat{\lambda}^\top \mathbf{x}_i)}{1 + exp(\hat{\lambda}^\top \mathbf{x}_i)}$$

1000 simulations were executed for each context. The resulting mean bias, standard deviation and Root Mean Square Error were measured in relative numbers to make them comparable across different scenarios. The formulas used for their calculation can be found below:

$$RBias\ (\%) = \left| \frac{\sum_{i=1}^{1000} \hat{\theta}^{(i)}}{1000} - \theta_N \right| \cdot \frac{100}{\theta_N} \tag{22}$$

$$RStandard\ deviation\ (\%) = \sqrt{\frac{\sum_{i=1}^{1000} (\hat{\theta}^{(i)} - \hat{\bar{\theta}})^2}{999}} \cdot \frac{100}{\theta_N} \tag{23}$$

$$RMSE\ (\%) = \sqrt{RBias^2 + RSD^2} \tag{24}$$

with $\hat{\theta}^{(i)}$ the estimation in the i-th simulation and $\hat{\bar{\theta}}$ the mean of the 1000 estimations.

5.3. Results

The relative mean bias of the estimations can be observed in Tables 1–3. We can observe that PSA reduces the bias in all situations, specially in the estimation of HCI. PSA using the reference sample drawn with probabilities proportional to the income, s_{r2}, provided much less biased estimates overall.

Table 1. Relative mean bias (%) of each parameter without applying PSA.

Size	Gini	HCI	IQR	IDR
6000	6.7	80.4	7.9	9.4
2000	3.2	93	3.8	3.1
4000	3.1	86	3.7	3
6000	3	79	3.5	3

Table 2. Relative mean bias (%) of each parameter applying PSA with the stratified reference sample.

Size	Gini	HCI	IQR	IDR
2000	1.7	3	2.5	1
4000	2.1	3.3	2.7	1
6000	2.2	3.1	2.7	0.9

Table 3. Relative mean bias (%) of each parameter applying PSA with the proportional reference sample.

Size	Gini	HCI	IQR	IDR
2000	0.3	1.1	0	0.2
4000	0.1	1.3	0.1	0.3
6000	0	1.1	0.1	0.5

The relative standard deviation of the estimations can be observed in Tables 4–6. The standard deviation remained stable across estimates of Gini coefficient, IQR and IDR, even with small gains for

the latter when using the reference sample with probabilities proportional to the minimum income to make ends meet, s_{r2}, but increased after applying PSA in the estimation of HCI.

Table 4. Relative standard deviation (%) of each parameter without applying PSA.

Size	Gini	HCI	IQR	IDR
2000	1.6	0.2	2.2	4.2
4000	1.1	0.3	1.5	2.9
6000	0.8	0.4	1.2	2.2

Table 5. Relative standard deviation (%) of each parameter applying PSA with the stratified reference sample.

Size	Gini	HCI	IQR	IDR
2000	1.7	4.1	2.7	4
4000	1.1	2.8	1.8	2.6
6000	0.9	2.2	1.4	2

Table 6. Relative standard deviation (%) of each parameter applying PSA with the proportional reference sample.

Size	Gini	HCI	IQR	IDR
2000	1.3	3.9	2.1	3.2
4000	0.9	2.8	1.5	2.2
6000	0.8	2.3	1.2	2.3

The relative Root Mean Square Error of the estimations can be observed in Tables 7–9. As a result of the stability of standard deviation and the reduction in bias, the RMSE of the estimates of the four parameters has a similar pattern than the observed for bias. Although RMSE is reduced after applying PSA in all cases, PSA was more efficient when the reference sample was drawn with probabilities proportional to the minimum income to make ends meet, s_{r2}.

Table 7. Relative RMSE (%) of each parameter without applying PSA.

Size	Gini	HCI	IQR	IDR
2000	3.6	93	4.4	5.2
4000	3.3	86	4	4.2
6000	3.1	79	3.7	3.7

Table 8. Relative RMSE (%) of each parameter applying PSA with the stratified reference sample.

Size	Gini	HCI	IQR	IDR
2000	2.4	5.1	3.7	4.2
4000	2.4	4.3	3.2	2.8
6000	2.4	3.8	3	2.2

Table 9. Relative RMSE (%) of each parameter applying PSA with the proportional reference sample.

Size	Gini	HCI	IQR	IDR
2000	1.4	4.1	2.1	3.2
4000	0.9	3.1	1.5	2.2
6000	0.8	2.5	1.2	2.4

PSA performance could be deeply affected by the selection mechanisms, which could lead to model misspecifications in propensity estimations. To test limitation and robustness of the proposed approach we have repeated the simulation with different patterns of non-response. The selection procedures can be described as follows:

NP1 Simple Random Sampling Without Replacement (SRSWOR) from the population fraction of individuals with a computer at home, U_v.

NP2 The probability of selection for the i-th individual, is given by

$$p_i = \frac{1}{1 + exp(-2x_i^1 + 0.2x_i^2 + 0.01x_i^3 + 0.2x_i^{41} + 0.4x_i^{42})} \tag{25}$$

NP3 The probability of selection for the i-th individual, is given by

$$p_i = (x_i^3 - 1925)^3/(1995 - 1925)^3 \tag{26}$$

NP4 The probability of selection for the i-th individual, is given by

$$p_i = 0.35 + 0.1 * x_i^1 - cos((2012 - x_i^3)/5)/3 \tag{27}$$

The procedure 1 is a typical case of coverage error (which is a type of selection bias itself [1]). The third scheme represents a cubic relationship between age and the probability of selection, with young people being the individuals with the highest probabilities and decreasing as age increases. The last scheme has two components: one dichotomous and the other cosine-shaped.

Tables 10 and 11 show the results of bias and relative ecm for the HCI parameter, where the selection bias of the unweighted estimator is large.

Table 10. Relative mean bias (%) for estimating HCI without and with applying PSA.

	Unadjusted	**PSA with Stratified Sample**	**PSA with Proportional Sample**
NP1 2000	93.5	1.7	4.5
NP1 4000	86.9	1.8	4.5
NP1 6000	80.4	1.9	4.5
NP2 2000	93	3	1.1
NP2 4000	86	3.3	1.3
NP2 6000	79	3.1	1.1
NP3 2000	92.9	1.3	1.3
NP3 4000	85.8	0.1	0.2
NP3 6000	78.7	0.5	0.5
NP4 2000	92.8	3	1.4
NP4 4000	85.5	3.2	1.5
NP4 6000	78.3	3.2	1.4

Table 11. Relative RMSE(%) for estimating HCI without and with applying PSA.

	Unadjusted	PSA with Stratified Sample	PSA with Proportional Sample
NP1 2000	93.5	3.6	5.3
NP1 4000	86.9	2.8	4.9
NP1 6000	80.4	2.5	4.7
NP2 2000	93	5.1	4.1
NP2 4000	86	4.3	3.1
NP2 6000	79	3.8	2.5
NP3 2000	92.9	9.6	8.6
NP3 4000	85.8	6.4	5.6
NP3 6000	78.7	5	4.3
NP4 2000	92.8	4.7	4
NP4 4000	85.5	4.1	3.1
NP4 6000	78.3	3.6	2.4

The results show a large decrease in bias and MSE for all response patterns for both PSA methods, which shows the robustness of the adjustment method. The reduction in bias and MSE is different across them. Using PSA with the reference sample drawn under a stratified design, s_{r1}, provided less RMSE when the convenience sample was drawn using NP1. On the other hand, PSA using the reference sample drawn with probabilities proportional to the income, s_{r2} provided much less biased estimates overall when the selection mechanism depended on NP2, NP3 or NP4.

6. Conclusions

Technological development has made large amounts of inexpensive data (commonly known as Big Data) available for researchers to be used for inference. New survey administration methods have also favoured the rise of data from nonprobability samples. Inferences from Big Data and nonprobability surveys have important sources of error ([4,24,28], ...). Given the characteristics of these data collection procedures, selection bias is particularly relevant.

Despite the growing interest raised by nonprobability data (both coming from Big Data or nonprobability surveys), there is still a lack of rigorous theory to make statistical inferences for general parameters through estimating equations. The current paper aims to fill this gap by establishing a theoretical framework for estimation of general parameters with nonprobability samples.

Results observed in our simulation study provide strong evidence on the efficiency of methods based in estimating equations with estimated propensities. However, it must be noted that the efficiency depends on the selection mechanisms of nonprobability samples and the availability of covariates for propensity estimation. In our simulations, results showed that Propensity Score Adjustment is more efficient when the propensity of being in the nonprobability sample is less related to the variable of interest. This behavior has been observed in literature regarding PSA for parametric estimation [11,24].

We used parametric methods to obtain the estimated propensities but we could use machine learning techniques as regression trees, spline regression, random forests etc. Recently [24,29] presented simulation studies where decision trees, k-nearest neighbors, Naive Bayes, Random Forest, Gradient Boosting Machine and Model Averaged Neural Networks are used for propensity score estimation. These studies compare the empirical efficiency of the use of linear models and Machine Learning prediction algorithms in estimation of linear parameters, but the theory is more complex and has not yet been developed. Other way to reduce the bias of the PSA estimates is to combine the PSA technique with other techniques as Statistical Matching or calibration. [27] apply a combination of propensity score adjustment and calibration on auxiliary variables in a real volunteer survey aimed to a population for which a complete census was available. [32] propose a doubly robust estimator for population mean estimation by incorporating the model-based estimator framework to PSA methods, improving their efficiency and making it robust to model misspecifications. Further research should focus on extensions of those methods for general parameter estimation.

Author Contributions: The authors contributed equally to this work in conceptualization, methodology, software and original draft preparation. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Ministerio de Educación y Ciencia (grant MTM2015-63609-R) and by Ministerio de Ciencia e Innovación (grant PID2019-106861RB-I00-AEI- 10.13039/501100011033).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Elliott, M.R.; Valliant, R. Inference for Nonprobability Samples. *Stat. Sci.* **2017**, *32*, 249–264.
2. Valliant, R.; Dever, J.A. Estimating Propensity Adjustments for Volunteer Web Surveys. *Sociol. Method Res.* **2011**, *40*, 105–137.
3. Deville, J.C.; Särndal, C.E. Calibration estimators in survey sampling. *J. Am. Stat. Assoc.* **1992**, *87*, 376–382.
4. Bethlehem, J. Selection Bias in Web Surveys. *Int. Stat. Rev.* **2010**, *78*, 161–188.
5. Martínez, S.; Rueda, M.; Arcos, A.; Martínez, H. Optimum calibration points estimating distribution functions. *J. Comput. Appl. Math.* **2010**, *233*, 2265–2277.
6. Martínez, S.; Rueda, M.; Martínez, H.; Arcos, A. Optimal dimension and optimal auxiliary vector to construct calibration estimators of the distribution function. *J. Comput. Appl. Math.* **2017**, *318*, 444–459.
7. Martínez, S.; Rueda, M.; Illescas, M. The optimization problem of quantile and poverty measures estimation based on calibration. *J. Comput. Appl. Math.* **2020**, https://doi.org/10.1016/j.cam.2020.113054
8. Rosenbaum, P.R.; Rubin, D.B. The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika* **1983**, *70*, 41–55.
9. David, M.; Little, R.J.A.; Samuhel, M.E.; Triest, R.K. Nonrandom nonresponse models based on the propensity to respond. In Proceedings of the Business and Economic Statistics Section, American Statistical Association, Toronto, Canada, August 15-18, 1983; 168–173.
10. Little, R.J. Survey nonresponse adjustments for estimates of means. *Int. Stat. Rev. Int. Stat.* **1986**, *54*, 139–157.
11. Lee, S. Propensity Score Adjustment as a Weighting Scheme for Volunteer Panel Web Surveys. *J. Off. Stat.* **2006**, *22*, 329–349.
12. Lee, S.; Valliant, R. Estimation for Volunteer Panel Web Surveys Using Propensity Score Adjustment and Calibration Adjustment. *Sociol. Method. Res.* **2009**, *37*, 319–343.
13. Rivers, D. Sampling for Web Surveys. In *Presented in Joint Statistical Meetings*; Stanford University and Polimetrix, Inc.: Salt Lake City, UT, USA, 2007.
14. Beaumont, J.F.; Bissonnette, J. Variance estimation under composite imputation: The methodology behind SEVANI. *Surv. Methodol.* **2011**, *37*, 171–179. Available online: https://es.overleaf.com/project/5eb2de68d45d5000014608e2 (accessed on 19 November 2020).
15. Hartley, H.O.; Sielken, R.L., Jr. A "super-population viewpoint" for finite population sampling. *Biometrics* **1975**, *31*, 411–422.
16. Royall, R.M.; Herson, J. Robust estimation in finite populations I. *J. Am. Stat. Assoc.* **1973**, *68*, 880–889.
17. Chen, J.K.T.; Valliant, R.L.; Elliott, M.R. Calibrating non-probability surveys to estimated control totals using LASSO, with an application to political polling. *J. R. Stat. Soc. Ser. C (Appl. Stat.)* **2019**, *68*, 657–681.
18. Breidt, J.; Opsomer, J. Model-assisted survey estimation with modern prediction. *Stat. Sci.* **2017**, *32*, 190–205.
19. Buelens, B.; Burger, J.; van den Brakel, J.A. Comparing Inference Methods for Non-probability Samples. *Int. Stat. Rev.* **2018**, *86*, 322–343.
20. Buskirk, T.D.; Lohr, S.L. Asymptotic properties of kernel density estimation with complex survey data. *J. Stat. Plan. Inference* **2005**, *128*, 165–190.
21. Francisco, C.A.; Fuller, W.A. Quantile estimation with a complex survey design. *Ann. Stat.* **1991**, *19*, 454–469.
22. Conti, P.L.; Di Iorio, A.; Guandalini, A.; Marella, D.; Vicard, P.; Vitale, V. On the estimation of the Lorenz curve under complex sampling designs. *Stat. Methods Appl.* **2019**, *29 (1)*, 1–24.
23. Deville, J.C.; Särndal, C.E.; Sautory, O. Generalized raking procedures in survey sampling. *J. Am. Stat. Assoc.* **1993**, *88*, 1013–1020.
24. Ferri-García, R.; Rueda, M.D.M. Propensity score adjustment using machine learning classification algorithms to control selection bias in online surveys. *PLoS ONE* **2020**, *15*, e0231500.
25. Valliant, R. Comparing alternatives for estimation from nonprobability samples. *J. Surv. Stat. Methodol.* **2020**, *8*, 231–263.

26. Schonlau, M.; Couper, M. Options for Conducting Web Surveys. *Stat. Sci.* **2017**, *32*, 279–292.
27. Ferri-García, R.; Rueda, M.M. Efficiency of Propensity Score Adjustment and calibration on the estimation from non-probabilistic online surveys. *SORT-Stat. Oper. Res. T.* **2018**, *42*, in press.
28. Copas, A.; Burkill, S.; Conrad, F.; Couper, M.P.; Erens, B. An evaluation of whether propensity score adjustment can remove the self-selection bias inherent to web panel surveys addressing sensitive health behaviours. *BMC Med. Res. Methodol.* **2020**, *20*, 251, doi:10.1186/s12874-020-01134-4.
29. Castro-Martín, L.; Rueda, M.D.M.; Ferri-García, R. Inference from Non-Probability Surveys with Statistical Matching and Propensity Score Adjustment Using Modern Prediction Techniques. *Mathematics* **2020**, *8*, 879.
30. Godambe, V.P.; Thompson, M.E. Estimating equations in presence of a nuisance parameter. *Ann. Stat.* **1974**, *2*, 568–571.
31. Binder, D.A. On the Variances of Asymptotically Normal Estimators from Complex Surveys. *Int. Stat. Rev. Rev. Int. Stat.* **1983**, *51*, 279–292.
32. Chen, Y.; Li, P.; Wu, C. Doubly Robust Inference With Nonprobability Survey Samples. *J. Am. Stat. Assoc.* **2020**.
33. Wu, C.; Thompson, M.E. *Sampling Theory and Practice*; Springer Nature: Berlin, Germany, 2020.
34. Wolter, K.M. *Introduction to Variance Estimation*, 2nd ed.; Springer, Inc.: New York, NY, USA, 2007.
35. Efron, B. Bootstrap methods: Another look at the jackknife. *Ann. Stat.* **1979**, *7*, 1–26.
36. Zhao, P.; Haziza, D.; Wu, C. Survey weighted estimating equation inference with nuisance functionals. *J. Econom.* **2020**, *216*, 516–536.
37. National Institute of Statistics. Living Conditions Survey. Microdata. 2012. https://www.ine.es/en/prodyser/microdatos_en.htm
38. Handcock, M.S.; Morris, M. *Relative Distribution Methods in the Social Sciences*; Springer Science & Business Media: Berlin, Germany, 2006.
39. Martínez, S.; Illescas, M.; Martínez, H.; Arcos, A. Calibration estimator for Head Count Index. *Int. J. Comput. Math.* **2020**, *97*, 51–62.