Treating nonresponse in the estimation of the distribution function

M. Rueda[a,], S. Martínez[b], M. Illescas[b]

[a]*Department of Statistics and Operational Research. University of Granada, Spain*
[b]*Department of Mathematics. University of Almería, Spain*

Abstract

The estimation of a finite population distribution function is considered when there are missing data. Calibration adjustment is used for dealing with nonresponse at the estimation stage. Several procedures are proposed and compared. A numerical study is carried out to evaluate the performances of estimators. Computational problems with the implementation of the proposed calibration estimators are also considered.

Keywords: Auxiliary information, missing data, calibration technique, distribution function estimates, survey sampling
2010 MSC: 62D05

1. Introduction

Surveys are a common method of data collection in economics and social sciences, but they often suffer from the problem of nonresponse which can produce biases in estimations and an increase in sampling variance if missing data follows any pattern. The standard statistical procedures developed for data with no missing values cannot be immediately and straightforwardly applied for deducing inferences in this situation.

Weighting is widely applied in surveys to adjust for nonresponse. Many different proposals for nonresponse weighting have been considered (see.eg. [1],[3],[9],[10], [11], [14], [17], [12]) in the estimation of linear parameters as total or mean, but lesser effort has been devoted in the development of efficient methods for estimating a population distribution. The distribution function is a relevant tool when the variable of interest is a measure of wages or income, since it is needed to calculate many poverty measures (the poverty line, the low income proportion, the poverty gap ...) For these reasons,

estimation of the distribution function is an important issue in sample surveys that has received much attention in the last years. On the contrary, to best of our knowledge, its estimation in the presence of missing data is a issue less investigated in the previous research. Whereas a extensive literature is available on estimation of population mean under non-response, lesser effort has been devoted in the development of efficient methods for the estimation of population distribution function.

The purpose of this paper is to estimate the distribution function in presence of missing data using the calibration method under a general sampling design. To the best of our knowledge, this is the first time that calibration techniques have been employed to remove the bias of non response in the estimation of the distribution function.

2. Estimating the distribution function when there are missing values

Given a finite population $U = \{1, \ldots, N\}$ with N different units and a sampling design d defined in U with first-order inclusion probability $\pi_i \geq 0$ and $d_i = \pi_i^{-1}$ the sampling design-basic weight for unit $i \in U$. In the presence of unit nonresponse, the character under study, say y, is observed for a subset of the original sample s. Thus, if we assume missing data on the sample s, it can be divided into the disjoint sets:

$$s_r = \{i \in s/\text{i responds}\} \text{ and } s_m = \{i \in s/\text{i does not respond}\},$$

with s_r, the respondent sample is of size r, and s_m is of size $n - r$.

Let y_i be the value of the character under study, say y, for the ith population unit. Our aim is to estimate the finite population distribution function (f.d.f.) of the study variable y, given by

$$F_y(t) = \frac{1}{N} \sum_{k \in U} \Delta(t - y_k), \text{with } \Delta(t - y_k) = \begin{cases} 0 & \text{if } t < y_k \\ 1 & \text{if } t \geq y_k \end{cases}$$

The design-unbiased Horvitz-Thompson estimator of $F_y(t)$ defined by

$$\widehat{F}_Y(t) = \frac{1}{N} \sum_{k \in s} d_k \Delta(t - y_k).$$

is impossible to compute in the presence of unit nonresponse, and a naive estimator of $F_y(t)$ is:

$$\widehat{F}_{YH}(t) = \frac{1}{N_r} \sum_{k \in s_r} d_k \Delta(t - y_k)$$

where N_r is the size of the population that would have responded if sampled, a quantity that is very rarely known in practice. When N_r is not known, one can use the Hajek estimator

$$\widehat{F}_{YHa}(t) = \frac{\sum_{k \in s_r} d_k \Delta(t - y_k)}{\sum_{k \in s_r} d_k}.$$

These estimators may lead to biased estimates because certain specific groups can be substantially under-represented. These errors can be overcome by the use of reweighting. When weighting is used, a set of weights is determined with the aid of the available auxiliary information, and estimation is carried out by applying the weights to the y-values for the responding elements. Calibration adjustment, initially conceived for correcting sampling errors ([4]), is currently one of the most appealing techniques for nonresponse adjustment ([19]).

We will use a twofold process: the sample s is first selected from the population U, then the response set s_r is realized as a subset of s. We assume that elements respond independently and the response distribution has first-order response probabilities $P(k \in s_r/s) = p_k$, positives. With the combined weights $d_k^* = \frac{1}{\pi_k p_k}$, we could reformulate the Horvitz-Thompson estimator as:

$$\widehat{F}_{Yw}(t) = \sum_{k \in s_r} d_k^* \Delta(t - y_k).$$

that is exactly unbiased albeit impossible to compute.

We can obtain a two-phase nonresponse adjusted estimator by replacing the original design weights by $d_k^o = \frac{1}{\pi_k \hat{p}_k}$. We will adapt the calibration methodology in our context.

For it, we assume the existence of auxiliary information relative to several variables related to the main variable y, $\mathbf{x}_k = (x_{k1}, x_{k2}, \ldots, x_{kJ})'$. The values $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$ are known for the entire population but y_k is known only if the kth unit is selected in the sample s_r.

3. Calibration weighting for the estimation of the distribution function with unit nonresponse.

[8] and [18] use different ways to implement the calibration approach in the estimation of the distribution function and the quantiles. The computationally simpler method proposed by [18] uses the calibration with respect to the predicted values of the variable of interest y. We use this methodology and we define a pseudo-variable $\tilde{y}_k = \widehat{\beta}^T \mathbf{x}_k$ for $k = 1, 2, ...N$, where $\widehat{\beta} = \left(\sum_{j \in s_r} d_j \mathbf{x}_j \mathbf{x}_j^T \right)^{-1} \cdot \sum_{j \in s_r} d_j \mathbf{x}_j y_j$.

Given a distance measure $G(w_k, d_k)$, the calibration process consists in finding the solution of the following minimization problem

$$\min_{w_k} \sum_{s_r} G(w_k, d_k) \tag{1}$$

while respecting the calibration equation

$$\frac{1}{N} \sum_{k \in s_r} w_k \Delta(\mathbf{t} - \tilde{y}_k) = F_{\tilde{y}}(\mathbf{t}) \tag{2}$$

where $\mathbf{t} = (t_1, t_2, \ldots, t_P')$, the term $F_{\tilde{y}}(\mathbf{t})$ denotes the f.d.f. of the pseudo-variable $\tilde{y}$ evaluated at $\mathbf{t}$ (see [15]) and t_j for $j = 1, 2, \ldots, P$ are points that we choose arbitrarily and assume that $t_1 < t_2 < \ldots < t_P$.

We can define the calibration estimator

$$\widehat{F}_{cal}(t) = \frac{1}{N} \sum_{k \in s_r} w_k \Delta(t - y_k) \tag{3}$$

What form should the weight take? It should reflect the known individual characteristics of the element $k \in r$, summarized by the vector value $\Delta(\mathbf{t} - \tilde{y}_k)$. A common way to compute calibration weights is linearly(using the chi-square distance, [14]) that produces the weights $w_k = d_k(1 + \lambda^T \Delta(\mathbf{t} - \tilde{y}_k))$. The weights w_k implicitly estimates the inverse response probability $1/p_k$, thus we acts as if $\pi_k p_k$ is the true selection probability of element k. Consequently this calibration approach has assumed a nonresponse model $p_k = 1/(1 + \lambda^T \Delta(\mathbf{t} - \beta^T \mathbf{x}_k))$. This model is difficult to deal with since the function is not differentiable on $\mathbf{x}_k$ and assume that the response mechanism depend on all auxiliary variables.

Now we consider a more flexible model and we propose a calibration method that allows the variables modeling the response mechanism to be

different from the benchmark variables in the calibration equation. Thus we define this two-step calibration method:

Step 1: Adjusting the bias of nonresponse by linear calibration.

Consider the M vector of explanatory model variables, $\mathbf{x}_k^*$ which population totals $\sum_{\mathcal{U}} \mathbf{x}_k^*$ are know. The calibration under the restrictions $\sum_{s_r} w_k^{(1)} \mathbf{x}_k^* = \sum_{\mathcal{U}} \mathbf{x}_k^*$ yields the calibrations weights $w_k^{(1)} = g_k^{(1)} * d_k, k = 1, \ldots, s_r$. Then, each unit in the sample has a weight that corrects the bias of lack of response.

Step 2: Adjusting the sample weights for the estimation of the f.d.f.

The auxiliary information of the calibration variables $\mathbf{x}$ is incorporated through the calibrated weights $w_k^{(2)} = g_k^{(2)} * w_k^{(1)}$ obtained with the restrictions $\sum_{s_r} w_k^{(2)} \Delta(\mathbf{t} - \tilde{y}_k) = F_{\tilde{y}}(\mathbf{t})$.

The two step calibration estimator proposed is

$$\hat{F}_{calTS}(t) = \frac{1}{N} \sum_{s_r} w_k^{(2)} \Delta(t - y_k) = \sum_{s_r} g_k^{(2)} g_k^{(1)} * d_k \Delta(t - y_k) \tag{4}$$

Note that the vector of model variables $\mathbf{x}_k^*$ and the vector of calibration variables $\mathbf{x}$ can be the same, can have some common component or be completely different

4. Calibration with model and calibration variables

In this section we consider a calibration approach similar to that used by [10] and [11] for the estimation of the population total. We also allow the variables modelling to be different from the calibration variables.

The probability of nonresponse can be modeled by $p_k = f(\gamma^T \mathbf{x}_k^*)$ for some vector parameter γ, where $h(\cdot) = 1/f(\cdot)$ is a known and everywhere monotonic and twice differentiable function and the vector $\mathbf{x}_k^*$ is the vector with the model variables. Some examples of usual models for the probability of response are: $p_k = \frac{1+exp(\gamma^T \mathbf{x}_k)^*/u}{l+exp(\gamma^T \mathbf{x}_k)^*}$ (the logit (l, u) method), $p_k = \frac{1}{exp(-\gamma^T \mathbf{x}_k^*)}$ (the raking model) and $p_k = \frac{exp(\gamma^T \mathbf{x}_k^*)}{1+exp(\gamma^T \mathbf{x}_k^*)}$ (the logistic-response model, a special case of logit (l, u) method, where $u = \infty$, $c = 2$ and $l = 1$ ([9])). The response probability is assumed independent of the survey variable of interest, which is known as missing at random (MAR) assumption.

We denote as $\mathbf{z}_k = \Delta(\mathbf{t} - \tilde{y}_k)$. We will use the vector $\mathbf{z}_k$ in the benchmark. Now, we generate calibrated weights by imposing the functional form $w_k =$

$\frac{d_k}{f(\hat{\gamma}^T\mathbf{x}_k^*)}\mathbf{z}_k$. The calibration equation is given by:

$$\frac{1}{N}\sum_{s_r}\frac{d_k}{f(\hat{\gamma}^T\mathbf{x}_k^*)}\mathbf{z}_k = \frac{1}{N}\sum_{s_r} d_k h(\hat{\gamma}^T\mathbf{x}_k^*)\mathbf{z}_k = F_{\tilde{y}}(\mathbf{t}) \tag{5}$$

where $\hat{\gamma}$ is a consistent estimator of vector γ and the resulting calibration estimator is given by

$$\hat{F}_{calI}(t) = \frac{1}{N}\sum_{s_r} d_k \frac{1}{\hat{p}_k}\Delta(t-y_k) = \frac{1}{N}\sum_{s_r} d_k h(\hat{\gamma}^T\mathbf{x}_k^*)\Delta(t-y_k).$$

The variance of this estimator can be obtained in a similar way as [10] but changing y_k by $\Delta(t-y_k)$, and $\mathbf{z}_k$ by $\Delta(\mathbf{t}-\tilde{y}_k)$.

Some considerations about the number of variables in the non-response model M and the number of calibration restrictions P should be taken into account to obtain the estimator $\hat{F}_{calI}(t)$. In [5], the number of model and calibration variables needed to be the same, that is $M = P$. This issue may limit the number of calibration restrictions in practice. We denote by $\hat{F}_{calID}$ the estimator $\hat{F}_{calI}$ when we use the same number of model and calibration variables.

[11] extended the Deville's weighting approach to the case where there are more calibration variables than model variables ($P > M$)through two alternatives. For it, their first extension does not look for an estimation $\widehat{\gamma}$ that satisfies the calibration equation (5), but it looks for an estimation $\widehat{\gamma}$ that minimizes $\mathbf{v}^T Q\mathbf{v}$ for some symmetric and positive definite matrix Q with dimension P, where

$$\mathbf{v} = \frac{1}{N}\left(\sum_{s_r} d_k h(\hat{\gamma}^T\mathbf{x}_k^*)\mathbf{z}_k - \sum_{s} d_k\tilde{\mathbf{z}}_k\right)$$

This minimization problem implies the reformulated calibration equation:

$$\frac{1}{N}\sum_{s_r} w_k^{(3)}\tilde{\mathbf{z}}_k = \frac{1}{N}\sum_{s_r} d_k h(\hat{\gamma}^T\mathbf{x}_k^*)\tilde{\mathbf{z}}_k = \frac{1}{N}\sum_{s} d_k\tilde{\mathbf{z}}_k \tag{6}$$

where $\tilde{\mathbf{z}}_k = A\cdot\mathbf{z}_k$ with $A = \left[\frac{1}{N}\sum_{s_r} d_i h^{'}(\hat{\gamma}^T\mathbf{x}_i^*)x_i^*\mathbf{z}_i^T Q\right]$.

Thus, the first alternative in [11] considers iterative process to find an estimation $\hat{\gamma}$ and Q that satisfied equation (6) such that $Q = H^{-1}$ with

$$H = \frac{1}{N}\sum_{s_r} d_k h^{'}(\hat{\gamma}^T\mathbf{x}_k^*)\mathbf{z}_k(\mathbf{z}_k)^T \tag{7}$$

In this alternative, each variable included in the vector $\tilde{\mathbf{z}}_k$ is a prediction for the corresponding variable in $\mathbf{x}_k^*$.

The second alternative proposed in [11] is a variant of the component reduction technique based on equation (6) with $\tilde{\mathbf{z}}_k = A_0 \cdot \mathbf{z}_k$ where

$$A_0^T = \left(\sum_{s_r} \mathbf{z}_j (\mathbf{z}_j)^T \right)^{-1} \sum_{s_r} \mathbf{z}_j (\mathbf{x}_j^*)^T$$

Consequently, this alternative does not require iteration or even rely on finding a matrix Q.

The estimator $\hat{F}_{calI}$ based on the first alternative of Kott and Liao approach ([11]) is denoted by $\hat{F}_{calIKL1}$ and the estimator $\hat{F}_{calI}$ based on second alternative, is denoted by $\hat{F}_{calIKL2}$.

5. Properties of the calibrated estimators of the distribution function.

For an estimator $\hat{F}(t)$ of $F(t)$ to be a genuine distribution function it should verify:

i. $\widehat{F}_{cal}(t)$ is continuous on the right,

ii. $\widehat{F}_{cal}(t)$ is monotone nondecreasing,

iii. a) $\lim\limits_{t \to -\infty} \widehat{F}_{cal}(t) = 0$ and b) $\lim\limits_{t \to +\infty} \widehat{F}_{cal}(t) = 1$.

It's easy to verify that conditions *i*) and *iii.a*) are satisfied for all the proposed estimators. $\hat{F}_{calTS}(t)$ meet the condition iii. b) if we take t_P such that $F_{\tilde{y}}(t_P) = 1$ (see [18]) but in general this estimator is not monotone nondecreasing. With respect to the calibration estimators based on model and calibration variables $\widehat{F}_{calID}$, $\widehat{F}_{calIKL1}$ and $\widehat{F}_{calIKL2}$, the calibration weights $\omega_k \geq 0$ for logit, raking and logistic methods and therefore, the estimators are nondecreasing. Like the previous case, for $\widehat{F}_{calID}$, if we take t_P such that $F_{\tilde{y}}(t_P) = 1$, then $\lim_{t \to +\infty} \widehat{F}_{calID}(t) = 1$. On the other hand, for the estimators $\widehat{F}_{calIKL1}$ and $\widehat{F}_{calIKL2}$, Theorem 1 establish the conditions to satisfy $\lim_{t \to +\infty} \widehat{F}_y(t) = 1$.

Theorem 1: If a component of the vector x_k^* contains all 1's and t_P should be sufficiently large, the estimators $\widehat{F}_{calIKL1}$ and $\widehat{F}_{calIKL2}$ satisfy the condition iii. b).

Proof:

We denote $\Delta(\mathbf{t}-\tilde{y}_k)^T = \Big(\Delta(t_1-\tilde{y}_k),\ldots,\Delta(t_{P-1}-\tilde{y}_k),\Delta(t_P-\tilde{y}_k)\Big) = (\Delta_{P-1}^T,1)$ and $(x_k^*)^T = (1,x_{2k},\ldots,x_{Mk}) = \Big(1,(x_k^\circ)^T\Big)$.

For the estimator $\widehat{F}_{calIKL1}(t)$, the calibration weights ω_k satisfies (6) with Q given by (7). It is clear that the matrix Q and Q^{-1} can be expressed by

$$Q = \begin{pmatrix} Q_{(P-1)\times(P-1)} & D_{(P-1)\times 1} \\ T_{1\times(P-1)} & C_{1\times 1} \end{pmatrix} \quad ; \quad Q^{-1} = \begin{pmatrix} \Gamma_{(P-1)\times(P-1)} & \Psi_{(P-1)\times 1} \\ (\Psi_{(P-1)\times 1})^T & \Upsilon_{1\times 1} \end{pmatrix}$$

with $\Psi_{(P-1)\times 1} = \frac{1}{N}\sum_{s_r} d_k h^{'}(\hat{\gamma}^T\mathbf{x}_k^*)\Delta_{P-1}$; $\Upsilon_{1\times 1} = \frac{1}{N}\sum_{s_r} d_k h^{'}(\hat{\gamma}^T\mathbf{x}_k^*)$ and $\Gamma_{(P-1)\times(P-1)}$ is given by equation (7) based on Δ_{P-1}. From $Q^{-1}\cdot Q = I_{P\times P}$, we have

$$(\Psi_{(P-1)\times 1})^T Q_{(P-1)\times(P-1)} + \Upsilon_{1\times 1} T_{1\times(P-1)} = 0_{1\times(P-1)} \tag{8}$$

$$(\Psi_{(P-1)\times 1})^T D_{(P-1)\times 1} + \Upsilon_{1\times 1} C_{1\times 1} = 1. \tag{9}$$

From (8) and (9),we have:

$$A = \left(\frac{1}{N}\sum_{s_r} d_i h^{'}(\hat{\gamma}^T\mathbf{x}_i^*)x_i^*\Delta(\mathbf{t}-\tilde{y}_k)^T Q\right) = \begin{pmatrix} 0_{1\times(P-1)} & 1 \\ A_{(M-1)\times(P-1)} & B_{(M-1)\times 1} \end{pmatrix}$$

with $B_{(M-1)\times 1} = (\Phi_{(P-1)\times(M-1)})^T D_{(P-1)\times 1} + \chi_{(M-1)\times 1} C_{1\times 1}$

$$A_{(M-1)\times(P-1)} = (\Phi_{(P-1)\times(M-1)})^T Q_{(P-1)\times(P-1)} + \chi_{(M-1)\times 1} T_{1\times(P-1)}$$

$$(\Phi_{(P-1)\times(M-1)})^T = \frac{1}{N}\sum_{s_r} d_i h^{'}(\hat{\gamma}^T\mathbf{x}_i^*)x_i^\circ\Delta_{P-1}^T \quad ; \quad \chi_{(M-1)\times 1} = \frac{1}{N}\sum_{s_r} d_i h^{'}(\hat{\gamma}^T\mathbf{x}_i^*)x_i^\circ$$

Consequently, $\tilde{\mathbf{z}}_k$ is given by $\tilde{\mathbf{z}}_k = A\Delta(\mathbf{t}-\tilde{y}_k) = \begin{pmatrix} 1 \\ Z_{(M-1)\times 1} \end{pmatrix}$ with

$$Z_{(M-1)\times 1} = A_{(M-1)\times(P-1)}\Delta_{P-1} + B_{(M-1)\times 1}$$

and the property iii. b) is fullfilled.

For the $\widehat{F}_{calIKL2}$ estimator, matrix A_0 can be expressed by $A_0 = \begin{pmatrix} A_{01} \\ A_{0(M-1)} \end{pmatrix}$ where

$$A_{01} = \left(\sum_{s_r}\Delta(\mathbf{t}-\tilde{y}_k)^T\right)\left(\sum_{s_r}\Delta(\mathbf{t}-\tilde{y}_k)\Delta(\mathbf{t}-\tilde{y}_k)^T\right)^{-1}$$

$$A_{0(M-1)} = \left(\sum_{s_r} x_k^0 \Delta(\mathbf{t} - \tilde{y}_k)^T \right) \left(\sum_{s_r} \Delta(\mathbf{t} - \tilde{y}_k) \Delta(\mathbf{t} - \tilde{y}_k)^T \right)^{-1}$$

For t_P sufficiently large, $\Delta(t_P - \tilde{y}_k) = 1$ for all $k \in U$ and we have

$$A_{01} = \left(\sum_{s_r} \Delta(t_P - \tilde{y}_k) \Delta(\mathbf{t} - \tilde{y}_k)^T \right) \left(\sum_{s_r} \Delta(\mathbf{t} - \tilde{y}_k) \Delta(\mathbf{t} - \tilde{y}_k)^T \right)^{-1} = \begin{pmatrix} 0_{1\times(P-1)} & 1 \end{pmatrix}$$

Therefore, the vector $\tilde{z}_k$ is given by $\tilde{z}_k = A_0 \cdot \Delta(\mathbf{t} - \tilde{y}_k) = \begin{pmatrix} 1 \\ A_{0(M-1)} \cdot \Delta(\mathbf{t} - \tilde{y}_k) \end{pmatrix}$ and then $\widehat{F}_{calIKL2}(t)$ satisfies property iii. b).

The property ii) could be achieved by the procedure described in [16]. This procedure, for a general estimator $\widehat{F}_y$, is defined in the following way:

$$\tilde{F}_y(y_{[1]}) = \widehat{F}_y(y_{[1]}), \quad \tilde{F}_y(y_{[i]}) = \max\{\widehat{F}_y(y_{[i]}), \tilde{F}_y(y_{[i-1]})\} \quad i = 2, \ldots, r. \quad (10)$$

On the other hand, ([12]) established sufficient conditions for consistency of estimators based on model and calibration variables. Thus, under these conditions, the estimators $\widehat{F}_{calID}$, $\widehat{F}_{calIKL1}$ and $\widehat{F}_{calIKL2}$ meet consistency. Regarding this issue, the conditions required to fulfill the properties of distribution function are not contradictory with respect to the conditions given in ([12]) and they are less restrictive than conditions from ([12]). In fact, the condition iv) of Assumption 4 from ([12]) requires $\mathbb{E}(|z_{j,k}|) < \infty$ for all components of z_k. If t_P is sufficiently large, the last component of z_k meets the condition iv). Also, under the conditions established by Theorem 1, the first component of $\tilde{z}_k$ (for $\widehat{F}_{calIKL1}$ and for $\widehat{F}_{calIKL2}$) meets the condition iv). The same occurs for the condition ii) of Assumption 5 from ([12]).

6. Simulation study

We have performed a Monte Carlo simulation study where we compare the precision of the proposed estimators with others estimators of the distribution function. All the estimators included are programmed by routines in R.

6.1. Some computational aspects

The calibration estimator $\hat{F}_{calTS}$ is programmed with routines based on the "calib" function of the package "sampling" ([20]). The "gencalib" function of the package sampling is used for obtain the calibrated weights in

$\hat{F}_{calID}$. The raking and logit (l,u) methods are available in the "gencalib" function and we have obtained two versions of this estimator based on the available methods. For the estimator $\hat{F}_{calIK1}$, we have programmed a routine that develops the Newthon-Rhapson method described in [3] and we have also obtained two versions (raking and logit (l,u)). With respect to the estimator $\hat{F}_{calIK2}$, we have also obtained two versions and for this we only needed to program a routine for the reduction of components and directly apply the original function "gencalib". For all versions of estimators based on logit (l,u) method, we used the following parameters $u = 10$; $l = 0$ and $c = 1$. Initially, a version of the estimators for the logistic-response model was considered in the simulation study but we finally decided to exclude it because of in many cases, this method did not converge.

6.2. Data

A fictitious population was simulated. The population size is $N = 5000$ and six variables were included in the study: age, nationality (native/non-native), gender, weight, access to the Internet (yes/no). These variables are generated to make its similar to the Spanish population pyramid. The study variable y is given by $y_k = 3 + 5 \cdot Internet + Age/5 + \varepsilon_k$ where ε_k are independent identically distributed random variables with distribution $\varepsilon_k \sim N(0, 0.1)$.

First we considered a raking non-response mechanism given by $p_k = \frac{1}{exp(-0.1-Internet)}$. Thus, we consider $(x_k^*)^T = (1, Internet)$ and the vector of calibration variables is $(z_k)^T = (1, Internet, Weight)$. Next, we consider a logistic non-response model based on the variable "Age": $p_k = \frac{exp(-3+0.1 \cdot Age)}{1+exp(-3+0.1 \cdot Age)}$. In this case, $(z_k)^T = (1, Age, Weight)$.

It is important to note that in both cases, the target variable is not directly related to the non-response mechanism, but this mechanism can be explained by some auxiliary variables configuring a Missing At Random (MAR) situation. We have not considered mechanism Missing Completely At Random (MCAR) since in these situations the estimators are asymptotically unbiased.

6.3. Results

The estimators considered are the Horvitz-Thompson estimator $\widehat{F}_{HT}(t)$, the Chamber-Dunstan estimator $\widehat{F}_{CD}(t)$ ([2]), the ratio estimator $\widehat{F}_r(t)$ and the Rao, Kovar and Mantel estimator $\widehat{F}_{RKM}(t)$ ([16]) based on the respondent

sample s_r. The Chamber-Dunstan estimator $\widehat{F}_{CD}(t)$ and the Rao, Kovar and Mantel estimator $\widehat{F}_{RKM}(t)$ are model-based estimator based on the following superpopulation model:

$$y_k = \kappa x_k + \mathrm{v}(x_k)u_k \quad k = 1, \ldots, N \tag{11}$$

where κ is an unknown parameter, v is a known, strictly positive function and u_k are independent and identically distributed random variables with zero mean. See ([2]) and ([16]) respectively for further details. In the simulation study, we considered in the superpopulation model (11) that $x_k = \tilde{y}_k$ and $\mathrm{v}(x_k) = 1$ for all unit k.

We drawn 10000 samples with several sizes by simple random sampling without replacement (see Table 1), both for the raking and for the logistic non-response mechanism. For each sample and for each estimator, estimates of the distribution function $F(t)$ were calculated for 11 different values of t, namely all deciles and quartiles $Q_y(0.25)$ and $Q_y(0.75)$. The performance of all the estimators is measured by means of the average relative bias (AVRB) and the average relative efficiency (AVRE), given respectively by

$$\mathrm{AVRB}(\widehat{F}) = \frac{B^{-1}}{11}\sum_{q=1}^{11}\left|\sum_{b=1}^{B}\frac{\widehat{F}(t_q)_b - F_y(t_q)}{F_y(t_q)}\right|, \quad \mathrm{AVRE}(\widehat{F}) = \frac{1}{11}\sum_{q=1}^{11}\frac{MSE[\widehat{F}(t_q)]}{MSE[\widehat{F}_{HT}(t_q)]}$$

where b indexes the bth simulation run, $\widehat{F}$ is an estimator for the distribution function, $MSE[\widehat{F}(t)] = B^{-1}\sum_{b=1}^{B}[\widehat{F}(t)_b - F_y(t)]^2$ is the empirical mean square error for $\widehat{F}(t)$.

Table 1 provides the values AVRB and AVRE for the population with two non-response mechanism considered. From results, it is observed that the usual estimators $\widehat{F}_{HT}(t)$, $\widehat{F}_{CD}(t)$, $\widehat{F}_r(t)$ and $\widehat{F}_{RKM}(t)$ have a considerable bias for all sample sizes.The proposed calibration estimators significantly reduce the bias, especially the estimators $\hat{F}_{calTS}$, and the different versions of $\hat{F}_{calIKL1}$ and $\hat{F}_{calIKL2}$. In a similar way estimators $\widehat{F}_{CD}(t)$, $\widehat{F}_r(t)$ and $\widehat{F}_{RKM}(t)$ suffer an important loss in efficiency compared to the $\widehat{F}_{HT}(t)$ estimator for the raking mechanism. All the proposed calibration estimators exhibit greater efficiency than these estimators. $\hat{F}_{calTS}$, and the different versions of $\hat{F}_{calIKL1}$ and $\hat{F}_{calIKL2}$ show the best performance. There is no significant difference between the two versions of the estimators (raking and logit methods) in terms of efficiency although the raking method produces the estimators with fewer errors in most cases. There is no estimator that is uniformly better than the rest in terms of bias and error.

Table 1: Average relative bias (AVRB) and the average relative efficiency (AVRE) of compared estimators. The lowest value is denoted in bold

	Raking non-response mechanism									
	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE
	$n = 125$		$n = 135$		$n = 145$		$n = 155$		$n = 165$	
$\widehat{F}_{HT}$	0.315	1.000	0.313	1.000	0.316	1.000	0.314	1.000	0.319	1.000
$\widehat{F}_{CD}$	0.367	2.444	0.372	2.695	0.370	2.807	0.372	3.013	0.367	2.796
$\widehat{F}_{r}$	0.359	2.318	0.363	2.502	0.360	2.623	0.360	2.725	0.358	2.525
$\widehat{F}_{RKM}$	0.339	2.242	0.344	2.431	0.340	2.535	0.344	2.717	0.340	2.515
$\widehat{F}_{calTS}$	**0.002**	0.270	**0.002**	0.256	0.002	0.247	0.007	0.243	**0.001**	0.211
$\widehat{F}_{calIDra}$	0.016	0.369	0.012	0.354	0.008	0.323	0.005	0.328	0.008	0.278
$\widehat{F}_{calIDlo}$	0.016	0.369	0.012	0.354	0.008	0.323	0.005	0.328	0.008	0.278
$\widehat{F}_{calIKL1ra}$	**0.002**	**0.264**	0.004	**0.254**	0.003	**0.241**	0.008	**0.238**	0.003	**0.210**
$\widehat{F}_{calIKL1lo}$	0.003	0.277	**0.002**	0.266	**0.001**	0.252	0.006	0.250	**0.001**	0.221
$\widehat{F}_{calIKL2ra}$	0.005	0.287	0.003	0.275	**0.001**	0.262	**0.004**	0.259	0.002	0.229
$\widehat{F}_{calIKL2lo}$	0.005	0.287	0.003	0.275	**0.001**	0.262	**0.004**	0.259	0.002	0.229
	Logistic non-response mechanism									
	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE	AVRB	AVRE
	$n = 125$		$n = 135$		$n = 145$		$n = 155$		$n = 165$	
$\widehat{F}_{HT}$	0.340	1.000	0.341	1.000	0.337	1.000	0.342	1.000	0.340	1.000
$\widehat{F}_{CD}$	0.181	0.278	0.185	0.279	0.181	0.276	0.186	0.281	0.184	0.274
$\widehat{F}_{r}$	0.191	0.401	0.197	0.404	0.195	0.399	0.199	0.396	0.199	0.395
$\widehat{F}_{RKM}$	0.181	0.318	0.184	0.317	0.180	0.314	0.185	0.316	0.184	0.310
$\widehat{F}_{calTS}$	0.040	0.101	0.035	0.091	0.032	0.088	0.036	0.083	0.036	0.079
$\widehat{F}_{calIDra}$	**0.034**	0.102	**0.029**	0.093	**0.026**	0.089	**0.030**	0.084	**0.031**	0.079
$\widehat{F}_{calIDlo}$	0.035	0.103	0.030	0.093	0.027	0.089	0.031	0.084	0.032	0.080
$\widehat{F}_{calIKL1ra}$	0.036	**0.098**	0.030	**0.089**	0.029	**0.085**	0.033	**0.080**	0.032	**0.078**
$\widehat{F}_{calIKL1lo}$	0.037	**0.098**	0.031	**0.089**	0.030	0.086	0.034	**0.080**	0.033	0.079
$\widehat{F}_{calIKL2ra}$	0.036	**0.098**	0.031	**0.089**	0.030	**0.085**	0.033	**0.080**	0.032	**0.078**
$\widehat{F}_{calIKL2lo}$	0.038	0.099	0.032	**0.089**	0.031	0.086	0.034	0.081	0.033	0.079

7. Conclusion

This paper describes how calibration weighting can be used to adjust the design weights to increase the efficiency of finite distribution function for a sample survey when there is unit nonresponse. We propose two calibration methods to reduce the non-response bias. The first method is based on two-step calibration weighting. The first calibration is designed to remove the non-response bias. The second one to decrease the sampling error in the estimation of the distribution function. This method allows different variables to be used in each phase, since the model for non-response and the predictive model can be very different. This estimator given by (4) is computationally simple. The last method is based on model and calibration variables. The calibration is done in a single stage, but different variables are also used to model the lack of response and for the calibration equation. Different model are also proposed to model the nonresponse. The problem with this methodology is the difficulty in solving the calibration equation. Various iterative methods are proposed to obtain the weights.

Our limited simulation study clearly shows the gain in reduction of bias and precision achieved when calibration is used for nonresponse that is not missing completely at random. There is no estimator that is uniformly better than the rest in terms of bias and error. The $\hat{F}_{calIKL1}$ and $\hat{F}_{calIKL2}$ estimators produce the best estimates in terms of the least error in most cases. The computational simplicity of the estimator in two stages $\hat{F}_{calTS}$ is noteworthy.

Acknowledgements

This work is partially supported by Ministerio de Economía y Competitividad of Spain (grant MTM2015-63609-R).

References

[1] Beaumont, J. F. Calibrated imputation in surveys under a quasi-model-assisted approach. Journal of the Royal Statistical Society. Series B (Statistical Methodology), 67,3 (2005) 445-458.

[2] Chambers, R.L., Dunstan, A. Estimating distribution functions from survey data. Biometrika 73 (1986) 597–604.

[3] Chang, T., and P. S. Kott Using Calibration Weighting to Adjust for Nonresponse Under a Plausible Model. Biometrika 95 (2008), 557—571.

[4] Deville,J.C., Särndal, C.E. Calibration estimators in survey sampling, J. Amer. Statist. Assoc. 87(418) (1992) 376–382.

[5] Deville, J. C. Generalized Calibration and Application to Weighting for Non-response COMPSTAT: Proceedings in Computational Statistics, 14th Symposium, Utrecht, The Netherlands, eds. J. G. Bethlehem and P. G. M. van der Heijden, New York: Springer-Verlag. 2000

[6] Estevao, V.M., and Särndal, C.E. A functional form approach to calibration. Journal of Official Statistics, 16 (2000) 379–399.

[7] Godambe, V. P., Thompson, M. E.:Parameters of Superpopulation and Survey Population: Their Relationships and Estimation. International Statistical Review. 54 (1986) 127–138.

[8] Harms, T., Duchesne, P., On calibration estimation for quantiles, Survey Methodology, 32, (2006) 37–52,

[9] Kott, P. S., Liao, D. Providing double protection for unit nonresponse with a nonlinear calibration-weighting routine. Survey Research Methods, 6(2)(2012). 105—111.

[10] Kott, P.S., Liao, D. One step or two? Calibration weihting form a complete list frame with nonresponse. Survey Methodolgy 41(1) (2015) 165–181

[11] Kott, P.S., Liao, D. Calibration weighting for nonresponse that is not missing at random: allowing more calibration than response-model variables. Journal of Survey Statistics and Methodology 5 (2017) 159-–174

[12] Lesage,E., Haziza, D., D'Haultfoeuille. A Cautionary Tale on Instrumental Calibration for the Treatment of Nonignorable Unit Nonresponse in Surveys, Journal of the American Statistical Association, 114, 526 (2019) 906–915.

[13] Little, R. J. A., Rubin, D. B. Statistical analysis with missing data. John Wiley New York, 1987

[14] Lundström, S., and Särndal, C.-E. Calibration as a standard method for the treatment of nonresponse. Journal of Official Statistics, 15 (1999) 305-327

[15] Martínez, S., Rueda, M., Arcos, A., Martínez, H. Optimum calibration points estimating distribution functions Journal of Computational and Applied Mathematics. 233 (2010) 2265–2277.

[16] Rao, J.N.K., Kovar, J.G., Mantel, H.J. On estimating distribution function and quantiles from survey data using auxiliary information, Biometrika 77(2) (1990) 365-375.

[17] Rota, B., Laitila,T. Comparisons of some weigting methods for nonresponse adjustment. Lithuanian Journal of Statistics 54,1 (2015) 69–83.

[18] Rueda, M. and Martínez, S. and Martínez, H. and Arcos, A., Estimation of the distribution function with calibration methods, J. Statist. Plann. Inference, 137(2) (2007) 435–448.

[19] Särndal, C.E., Lundström, S. Estimation in Surveys with Nonresponse. New York: John Wiley and Sons, Inc., 2005.

[20] Tille, Y., Matei, A. sampling: Survey Sampling. A software routine available online at http://cran.r-project.org/web/packages/sampling/sampling.pdf.