ACTA ET COMMENTATIONES UNIVERSITATIS TARTUENSIS DE MATHEMATICA
Volume 23, Number 2, December 2019
Available online at http://acutm.math.ut.ee

Some competing nonresponse adjustment estimators

Imbi Traat

Abstract. The nonresponse adjustment estimator, derived in this paper by standard regression tools, is surprising by its form. The weights of the new estimator, called the f-estimator, are (general) inverses of the respective weights in the classical linear calibration estimator and propensity adjusted estimator. In a simulation experiment on real data, the new estimator is the best for several study variables.

1. Introduction

Nowadays, response rates are low in many sample surveys. It is difficult to estimate population parameters from the response set since the response mechanism is unknown. In the full-sample design-based theory, there are universial results holding for any study variable (y-variable) and any sampling design. Examples of these important properties are unbiasedness of the Horvitz–Thompson estimator [5], or nearly unbiasedness of the calibration estimator [3]. Such universal results are not available for the estimators under nonresponse. Estimators are usually biased, and the situation is even more complicated – the bias is different for different variables in the same sample survey.

There is a long history and large literature on the estimation under nonresponse [10]. The extensive development of new methods took place after auxiliary information (from various registers) became available. Auxiliary variables ($\mathbf{x}$-variables) are used to construct nonresponse adjusted estimators. A review on nonresponse weighting adjustments is given in [2]. Classical adjustment estimators are the calibration and the response propensity (called also double expansion) estimators. Their task is to reduce bias. The

Received September 27, 2019.

2010 *Mathematics Subject Classification.* 62D05.

Key words and phrases. Nonresponse; calibration; f-estimator; mixture estimator; overall bias.

https://doi.org/10.12697/ACUTM.2019.23.25

arguments here are: (1) calibration estimator is exact for auxiliary variables, therefore, perhaps, is less biased for study variables, (2) response propensities estimate response probabilities, the latter, if known, produce unbiased estimator by double expansion.

Additional requirements are postulated for auxiliary variables. For estimating response probabilities they have to be related to both, the response indicator and the y-variable ([6], [1]). In calibration, the $\mathbf{x}$-variables have to be related to the y-variable.

Theory concentrates on one fixed y-variable and seeks the best possible, least biased estimator for this variable. The received estimator involves weights computed with auxiliary variables. In practice, these weights are uniformly applied for all study variables in the survey at hand. The problem occurs when the response is poorly explained by available auxiliary variables; usually response also depends on study variables. More important, study variables (there are many in a survey) have different relationships with auxiliary variables, and therefore the weights computed with the same auxiliary variables cannot work well for all study variables. There are even cases where adjustment increases bias compared with simple unadjusted estimator [4].

It is important to evaluate goodness of the estimator not just for one study variable, but rather over all study variables of the survey. Here we use the average absolute relative bias. We call the estimator best if its absolute relative bias is smallest on the average (average over all study variables in the current survey).

We derive a new estimator by regressing the study variable y on the auxiliary vector $\mathbf{x}$. The estimation of the regression slope can go in two different ways. One of these gives well-known classical linear calibration estimator, another gives the new f-estimator. Basically, the f-estimator is the mean of fitted values in the response set r. The received mean uses auxiliary information outside r, which may decrease bias for some study variables.

The f-estimator is theoretically compared with well-known estimators, such as the simple unweighted, propensity adjusted and calibration estimators. The weights of the f-estimator and of the calibration estimator are each other's inverses in a general sense. Two modifications are defined: the scaled f-estimator and the mixture estimator. The scaled f-estimator uses rescaled weights that have mean 1 in the response set, similarly to the weights of the calibration estimator. The mixture estimator chooses for each y-variable either calibration or f-estimator. The decision is made using upper bounds for the differences from the target sample mean of both estimators.

2. Preliminaries

The sample s is drawn from the population $U = \{1, \ldots, k, \ldots, N\}$ so that unit k has the known inclusion probability $\pi_k > 0$ and the sampling weight $d_k = 1/\pi_k$. The response r is the set of units k having delivered their values of study variables. The mechanism that generates r from s is unknown, $r \subset s \subset U$. The (sample-weighted) survey response rate is $P = \sum_{k\in r} d_k / \sum_{k\in s} d_k$, where $0 < P < 1$ is assumed.

In the nonresponse context, three types of variables play a role. The study variable (continuous or categorical) y has values y_k observed for $k \in r$ only, and is used to estimate the population total $Y = \sum_{k\in U} y_k$, or the mean $\bar{Y} = \sum_{k\in U} y_k/N$. The response indicator I has value $I_k = 1$ for $k \in r$ and $I_k = 0$ for $k \in s - r$. The auxiliary vector $\mathbf{x}$ with value $\mathbf{x}_k$ is available at least for $k \in s$, possibly for $k \in U$. The $J \geq 1$ variables in the vector $\mathbf{x}$ can be continuous or categorical. They are recorded from registers or available as paradata from the data collection process. More particularly, $\mathbf{x}$ can be a group vector, that is, of the form $\mathbf{x}_k = (0, \ldots, 1, \ldots, 0)'$ with a single entry 1 to indicate the group membership of k.

We assume that all $\mathbf{x}$-vectors used here have the following feature: there exists a constant vector $\boldsymbol{\mu}$ (not depending on k) such that

$$\boldsymbol{\mu}'\mathbf{x}_k = 1 \text{ for all } k. \tag{1}$$

Most vectors of interest satisfy this requirement. When the vector $\mathbf{x}_k$ has one constant element, e.g., 1 in the first position for all k, then $\boldsymbol{\mu} = (1, 0, \ldots, 0)'$ satisfies (1). When $\mathbf{x}$ is a group vector, the vector $\boldsymbol{\mu} = (1, 1, \ldots, 1)'$ satisfies (1). The reason for the requirement is convenience in many derivations and simple forms of the results.

In the following, the design-weighted means and the second moments of the $\mathbf{x}$-vector are needed, both in r and in s:

$$\bar{\mathbf{x}}_r = \frac{\sum_{k\in r} d_k \mathbf{x}_k}{\sum_{k\in r} d_k}, \quad \bar{\mathbf{x}}_s = \frac{\sum_{k\in s} d_k \mathbf{x}_k}{\sum_{k\in s} d_k}, \tag{2}$$

$$\boldsymbol{\Sigma}_r = \frac{\sum_{k\in r} d_k \mathbf{x}_k \mathbf{x}_k'}{\sum_{k\in r} d_k}, \quad \boldsymbol{\Sigma}_s = \frac{\sum_{k\in s} d_k \mathbf{x}_k \mathbf{x}_k'}{\sum_{k\in s} d_k}, \tag{3}$$

where the matrices $\boldsymbol{\Sigma}_r$ and $\boldsymbol{\Sigma}_s$ are assumed to be nonsingular. Also, the following quadratic forms are used:

$$Q_s = (\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s)' \boldsymbol{\Sigma}_s^{-1} (\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s), \quad Q_r = (\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s)' \boldsymbol{\Sigma}_r^{-1} (\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s). \tag{4}$$

Both Q_s and Q_r express the balance of auxiliary variables in the response set with respect to the full sample s. Särndal [8] has defined the imbalance measure as $IMB = P^2 Q_s$. There are limits for Q_s, $0 \leq Q_s \leq (1-P)/P$.

3. Estimators based on regression

Our target is the sample mean of the study variable,

$$\bar{y}_s = \frac{\sum_{k\in s} d_k y_k}{\sum_{k\in s} d_k}.$$

The sample mean $\bar{y}_s$ is unbiased for the population mean but, regrettably, not computable under nonresponse. Estimators can be constructed from the response set r, with the aim to be unbiased or close to $\bar{y}_s$. Note that speaking about the bias with respect to $\bar{y}_s$, we mean the conditional bias over the response mechanism (over realizations of the response r, given a sample s). Below we consider estimators derived by the linear regression – two well-known estimators and one new estimator.

Under nonresponse, the classical two-phase approach [10] leads to the estimator

$$\bar{y}_{\text{2ph}} = \frac{\sum_{k\in r} d_k y_k/\theta_k}{\sum_{k\in s} d_k}, \tag{5}$$

where $\theta_k = P(I_k = 1|k \in s)$ is the response probability. For fixed s, the resulting $\bar{y}_{\text{2ph}}$ is unbiased for $\bar{y}_s$. Since θ_k is rarely known, the natural move is to estimate it. Under available auxiliary information, it is possible to estimate the response propensity $P(I_k = 1|\mathbf{x}_k, k \in s)$ and use it as an estimator for θ_k. The propensity adjusted or double expansion estimator is received. For estimating response propensities, and also θ_k, we regress I_k on $\mathbf{x}_k$ in s, and find coefficient-vector $\mathbf{b}$ in $\mathbf{b}'\mathbf{x}_k$ by weighted least square method (WLS). We get $\mathbf{b}' = P\,\bar{\mathbf{x}}_r'\mathbf{\Sigma}_s^{-1}$, and use the respective linear combination, the fitted value, as an estimator for θ_k,

$$\hat{\theta}_k = \mathbf{b}'\mathbf{x}_k = P\,\bar{\mathbf{x}}_r'\mathbf{\Sigma}_s^{-1}\mathbf{x}_k = Pf_k, \tag{6}$$

where

$$f_k = \bar{\mathbf{x}}_r'\mathbf{\Sigma}_s^{-1}\mathbf{x}_k. \tag{7}$$

The propensity adjusted estimator follows from (5) to (7) and from the expression for P,

$$\bar{y}_{\text{pro}} = \frac{\sum_{k\in r} d_k y_k/\hat{\theta}_k}{\sum_{k\in s} d_k} = \frac{\sum_{k\in r} d_k y_k/f_k}{\sum_{k\in r} d_k}. \tag{8}$$

For the linear calibration estimator and for the new f-estimator we regress y_k on $\mathbf{x}_k$ in s. The WLS method gives the coefficient-vector

$$\mathbf{b}_s' = \frac{\sum_{k\in s} d_k y_k \mathbf{x}_k'}{\sum_{k\in s} d_k}\mathbf{\Sigma}_s^{-1}. \tag{9}$$

The problem with (9) is missing y_k in s, they are only known in r, $r \subset s$. We consider two ways for estimating $\mathbf{b}_s'$. First, estimating both factors in (9) by the respective means in r, we get the coefficient-vector

$$\mathbf{b}_r' = \frac{\sum_{k\in r} d_k y_k \mathbf{x}_k'}{\sum_{k\in r} d_k}\mathbf{\Sigma}_r^{-1}. \tag{10}$$

The respective fitted values are $\mathbf{b}_r'\mathbf{x}_k$. It appears that the mean of the fitted values in s is the classical linear calibration estimator,

$$\frac{\sum_{k\in s} d_k \mathbf{b}_r'\mathbf{x}_k}{\sum_{k\in s} d_k} = \mathbf{b}_r'\bar{\mathbf{x}}_s = \frac{\sum_{k\in r} d_k y_k \mathbf{x}_k'}{\sum_{k\in r} d_k}\mathbf{\Sigma}_r^{-1}\bar{\mathbf{x}}_s,$$

from which we get

$$\bar{y}_{\text{cal}} = \frac{\sum_{k\in r} d_k y_k g_k}{\sum_{k\in r} d_k}, \tag{11}$$

where g_k is the calibration weight,

$$g_k = \mathbf{x}_k'\mathbf{\Sigma}_r^{-1}\bar{\mathbf{x}}_s. \tag{12}$$

It is easy to check that calibration property holds, i.e., applying g_k on the auxiliary vector $\mathbf{x}_k$ results with the exact sample mean in s,

$$\frac{\sum_{k\in r} d_k \mathbf{x}_k g_k}{\sum_{k\in r} d_k} = \frac{\sum_{k\in r} d_k \mathbf{x}_k \mathbf{x}_k'}{\sum_{k\in r} d_k}\mathbf{\Sigma}_r^{-1}\bar{\mathbf{x}}_s = \mathbf{\Sigma}_r\mathbf{\Sigma}_r^{-1}\bar{\mathbf{x}}_s = \bar{\mathbf{x}}_s.$$

The second way for estimating (9) avoids estimating Σ_s. It can be computed if auxiliary variables are known for each $k \in s$ (standard situation). Thus, estimating only the first factor in (9) gives

$$\mathbf{b}_{sr}' = \frac{\sum_{k\in r} d_k y_k \mathbf{x}_k'}{\sum_{k\in r} d_k}\mathbf{\Sigma}_s^{-1}. \tag{13}$$

Taking the mean of the respective fitted values $\mathbf{b}_{sr}'\mathbf{x}_k$ in r gives the new estimator $\bar{y}_{\text{f}}$ which we call f-estimator,

$$\frac{\sum_{k\in r} d_k \mathbf{b}_{sr}'\mathbf{x}_k}{\sum_{k\in r} d_k} = \mathbf{b}_{sr}'\bar{\mathbf{x}}_r = \frac{\sum_{k\in r} d_k y_k \mathbf{x}_k'}{\sum_{k\in r} d_k}\mathbf{\Sigma}_s^{-1}\bar{\mathbf{x}}_r,$$

from which

$$\bar{y}_{\text{f}} = \frac{\sum_{k\in r} d_k y_k f_k}{\sum_{k\in r} d_k}, \tag{14}$$

where f_k is given in (7). Comparing $\bar{y}_{\text{f}}$ and the well-known propensity adjusted estimator $\bar{y}_{\text{pro}}$ in (8), we notice the surprising opposite role of the factor f_k: in one case it divides y_k, in the other case it multiplies it. The estimators $\bar{y}_{\text{f}}$ and $\bar{y}_{\text{cal}}$ are also opposite in a certain sense. They both have their weights, respectively f_k and g_k, in the numerator, but these weights are inverses of each other in a general sense, namely the means of their product are equal to 1 both in s and in r,

$$\frac{\sum_{k\in r} d_k g_k f_k}{\sum_{k\in r} d_k} = \frac{\sum_{k\in s} d_k g_k f_k}{\sum_{k\in s} d_k} = 1.$$

The above property and many others of the factors f_k and g_k are proved in [11]. The relationship (1) for the auxiliary vector is important in derivations. Inserting $\boldsymbol{\mu}'\mathbf{x}_k = \mathbf{x}_k'\boldsymbol{\mu} = 1$ in suitable places helps to simplify matrix expressions. If the auxiliary vector is a group vector, then $g_k = 1/f_k$, and $\bar{y}_{\text{cal}} = \bar{y}_{\text{pro}}$, i.e., $\bar{y}_{\text{cal}}$ and $\bar{y}_{\text{f}}$ have opposite use of the weight f_k.

In [11] the means of weights f_k and g_k in r are given as

$$\bar{g}_r = \frac{\sum_{k\in r} d_k g_k}{\sum_{k\in r} d_k} = 1, \quad \bar{f}_r = \frac{\sum_{k\in r} d_k f_k}{\sum_{k\in r} d_k} = 1 + Q_s,$$

where Q_s is in (4). Here comes the motivation to rescale the weights f_k to have average 1. One more comparable estimator, the scaled f-estimator is given by

$$\bar{y}_{\text{scf}} = \frac{\sum_{k\in r} d_k y_k f_k}{(1+Q_s)\sum_{k\in r} d_k}. \tag{15}$$

When deriving $\bar{y}_{\text{f}}$, we have taken the mean of the fitted values over r. However, it uses auxiliary information outside r; through Σ_s. On the contrary, taking the mean over s and using property (1) results in the simple unweighted mean,

$$\frac{\sum_{k\in s} d_k \mathbf{b}'_{sr}\mathbf{x}_k}{\sum_{k\in s} d_k} = \mathbf{b}'_{sr}\bar{\mathbf{x}}_s = \frac{\sum_{k\in r} d_k y_k \mathbf{x}'_k}{\sum_{k\in r} d_k}\mathbf{\Sigma}_s^{-1}\bar{\mathbf{x}}_s = \frac{\sum_{k\in r} d_k y_k}{\sum_{k\in r} d_k} = \bar{y}_{\text{unw}}.$$

The estimators $\bar{y}_{\text{unw}}$, $\bar{y}_{\text{f}}$, $\bar{y}_{\text{pro}}$, $\bar{y}_{\text{cal}}$ are related, as studied below.

4. Relationships between estimators

The Cauchy–Schwarz inequality gives for nonnegative y_k and f_k

$$\bar{y}_{\text{f}} \cdot \bar{y}_{\text{pro}} = (\sum_{k\in r} w_k y_k f_k) \cdot (\sum_{k\in r} w_k y_k / f_k) \geq (\sum_{k\in r} w_k y_k)^2 = \bar{y}^2_{\text{unw}},$$

where $w_k = d_k / \sum_{k\in r} d_k$. In another way,

$$\frac{\bar{y}_{\text{f}}}{\bar{y}_{\text{unw}}} \cdot \frac{\bar{y}_{\text{pro}}}{\bar{y}_{\text{unw}}} \geq 1. \tag{16}$$

We see from (16) that if $\bar{y}_{\text{f}} \leq \bar{y}_{\text{unw}}$, then $\bar{y}_{\text{pro}} \geq \bar{y}_{\text{unw}}$, i.e., in each response set r the propensity weighted and the f-estimator are on either sides of the unweighted $\bar{y}_{\text{unw}}$. Suppose that we know that $\bar{y}_{\text{unw}}$ is biased and, e.g., underestimates $\bar{y}_s$. Most probably, for a given r, one has $\bar{y}_{\text{unw}} < \bar{y}_s$. We want to make nonresponse adjustment with estimators using auxiliary information. Suppose that $\bar{y}_{\text{pro}} < \bar{y}_{\text{unw}}$. Clearly, it is not wise to use $\bar{y}_{\text{pro}}$, but rather stay with $\bar{y}_{\text{unw}}$, or perhaps choose instead $\bar{y}_{\text{f}}$. Similar arguments hold for the calibration estimator, at least for the group-vector $\mathbf{x}$ case, then $\bar{y}_{\text{pro}} = \bar{y}_{\text{cal}}$.

The important question is, which is closer to $\bar{y}_s$? Is it $\bar{y}_{\text{pro}}$, $\bar{y}_{\text{cal}}$, or is it $\bar{y}_{\text{f}}$ or $\bar{y}_{\text{scf}}$? A drawback of $\bar{y}_{\text{pro}}$ is that f_k placed in the denominator may be near zero for some k. Below we compare $\bar{y}_{\text{cal}}$ and $\bar{y}_{\text{f}}$ to $\bar{y}_s$. We get the upper bounds,

$$\begin{aligned} |\bar{y}_s - \bar{y}_{\text{f}}| &\leq |\bar{y}_s - \bar{y}_{\text{unw}}| + |\bar{y}_{\text{unw}} - \bar{y}_{\text{f}}|, \\ |\bar{y}_s - \bar{y}_{\text{cal}}| &\leq |\bar{y}_s - \bar{y}_{\text{unw}}| + |\bar{y}_{\text{unw}} - \bar{y}_{\text{cal}}|. \end{aligned}$$

The above inequalities suggest to choose $\bar{y}_{\text{f}}$ for estimating $\bar{y}_s$ if it is closer to $\bar{y}_{\text{unw}}$ than $\bar{y}_{\text{cal}}$, and otherwise to choose $\bar{y}_{\text{cal}}$. We call the respective estimator mixture estimator, and denote it by $\bar{y}_{\text{mix}}$.

Different y-variables of the same survey may require different estimators. Correlations between y and the response indicator, y and $\mathbf{x}$-variables affect behavior of the estimators. Since $\mathbf{x}$ is a vector, we rather observe one-dimensional summaries of it, the f- and g-factors. The covariance in s of two variables a and b is defined by the formula

$$\text{cov}_s(a, b) = \frac{\sum_{k \in s} d_k a_k b_k}{\sum_{k \in s} d_k} - \bar{a}_s \, \bar{b}_s, \tag{17}$$

where $\bar{a}_s$, $\bar{b}_s$ are the weighted means in s, similarly to (2). The covariance in r is defined analogously. We get:

$$\begin{aligned} \text{cov}_s(I, y) &= P\,(\bar{y}_{\text{unw}} - \bar{y}_s), & (18) \\ \text{cov}_r(f, y) &= \bar{y}_{\text{f}} - (1 + Q_s)\,\bar{y}_{\text{unw}}, & (19) \\ \text{cov}_r(g, y) &= \bar{y}_{\text{cal}} - \bar{y}_{\text{unw}}, & (20) \\ \text{cov}_r(f, g) &= -Q_s, \ \text{cov}_s(f, g) = -Q_r. & (21) \end{aligned}$$

We see from (18) that if response is negatively correlated with the study variable y, the simple unweighted estimator underestimates $\bar{y}_s$. We know that the factors f_k and g_k are (general) inverses of each others. Negative covariances in (21) confirm their opposite behavior. In fact, their correlation is nearly -1, $\text{cor}_s(f, g) \approx -1$, [11]. Inverse relationship between f and g causes opposite signs to their covariances with y in (19)–(20). Consequently, if $\bar{y}_{\text{cal}}$ is smaller than $\bar{y}_{\text{unw}}$, then $\bar{y}_{\text{f}}$ is bigger than $\bar{y}_{\text{unw}}$, more bigger for unbalance response sets where $Q_s > 0$.

The response r is called perfectly balanced with respect to the vector $\mathbf{x}$ if $\bar{\mathbf{x}}_r = \bar{\mathbf{x}}_s$, [8]. For a balanced response set, $f_k = 1$ and $g_k = 1$ for all k. This follows from their expressions and the property (1). For example, $f_k = \bar{\mathbf{x}}'_r \mathbf{\Sigma}_s^{-1} \mathbf{x}_k = \bar{\mathbf{x}}'_s \mathbf{\Sigma}_s^{-1} \mathbf{x}_k = 1$. All the estimators $\bar{y}_{\text{pro}}$, $\bar{y}_{\text{cal}}$, $\bar{y}_{\text{f}}$, and $\bar{y}_{\text{scf}}$ reduce to $\bar{y}_{\text{unw}}$ for a balanced response set. Nevertheless, since balancing means conditions on $\mathbf{x}$-variables, the resulting $\bar{y}_{\text{unw}}$ may still be biased for $\bar{y}_s$. Särndal and Lundquist [9] have confirmed that deviation of $\bar{y}_{\text{cal}}$ from $\bar{y}_s$ decreases, but not to zero, if balance of the response set increases.

Strong relationship between y and $\mathbf{x}$-variables is expected to reduce bias. Suppose, we have the exact linear relationship $y_k = \mathbf{b}'\mathbf{x}_k$. In this extreme case, the calibration estimator is exact,

$$\bar{y}_{\mathrm{cal}} = \frac{\sum_{k\in r} d_k y_k g_k}{\sum_{k\in r} d_k} = \frac{\sum_{k\in r} d_k \mathbf{b}'\mathbf{x}_k\mathbf{x}_k'\boldsymbol{\Sigma}_r^{-1}\bar{\mathbf{x}}_s}{\sum_{k\in r} d_k} = \mathbf{b}'\bar{\mathbf{x}}_s = \bar{y}_s.$$

The estimator $\bar{y}_{\mathrm{f}}$ does not have this property. It needs additional condition – balanced response, or even a stronger condition, $\boldsymbol{\Sigma}_r = \boldsymbol{\Sigma}_s$. Exact linear relationship does not exist in real life for variables under interest. Even if, for some y-variable the relationship with the $\mathbf{x}$-variables is strong, and calibration reduces bias, then for other y-variables the same calibration estimator may increase bias. As simulation shows, in this case one of $\bar{y}_{\mathrm{unw}}$, $\bar{y}_{\mathrm{f}}$, $\bar{y}_{\mathrm{scf}}$, or $\bar{y}_{\mathrm{mix}}$ may be better.

5. Simulation set-up

Real data from the Estonian household surveys in 2004–2007 were used. The data set for our use consisted of 1000 households (considered as a simple random sample s), and of the following variables, where H denotes Household and HD the Household Head:

HD_sex (binary; 1 for woman),
HD_active (binary; 1 for employed),
HD_educ (discrete; 1, 2, 3 coding education levels "low", "medium", "high"),
H_size (discrete; 1,2, ...,12),
H_No_of_Children (discrete; 0, 1, ...,9),
HD_educ1 (binary; 1 for education level "low"),
HD_educ2 (binary; 1 for education level "medium"),
HD_educ3 (binary; 1 for education level "high"),
H_With_Children (binary; 1 if yes),
H_big (binary; 1 for H_size bigger than 1),
H_income (min=0, median=7233, mean=9392, max=57163),
H_transfer (social benefits; min=0, median=2442, mean=2827, max=47284),
H_expenditure (min=577, median=6020, mean=7812, max=54723).

The values of variables H_income, H_transfer, H_expenditure are monthly values in Estonian Kroons (EEK).

Response probabilities θ_k (where k designates a household) were computed for $k \in s$ by the model

$$\mathrm{logit}(\theta) = 5 - 4 \times \mathrm{HD\_sex} + 2 \times \mathrm{HD\_active} - 0.0003 \times \mathrm{H\_income}. \quad (22)$$

The model deliberately assigns lower response probability to high income households where the head is an unemployed female.

The auxiliary vector used in calibration and f-estimator is a four-dimensional group vector formed by crossing 2 binary variables,

$$\mathbf{x} = (\mathrm{HD\_sex} \times \mathrm{HD\_active}). \quad (23)$$

We see that the $\mathbf{x}$-vector is related to the response probabilities by its 2 variables, but $\mathbf{x}$ does not involve information about income. This choice aspires to mimic the real life situation, where auxiliary variables are various demographic and other variables obtainable from registers, but response, however, depends on study variables, such as income. We also consider another choice of the $\mathbf{x}$-vector, the 5-dimensional

$$\mathbf{x} = (\text{HD\_sex} \times \text{HD\_active}, \text{H\_expenditure}). \tag{24}$$

The $\mathbf{x}$-vector in (24) is unrealistic in practice since it is very hard to find out H_expenditure for nonrespondents. Here, it helps to see the behavior of estimators under stronger auxiliary information.

The response r of $m = 600$ households was generated according to the response probabilities θ_k. The order-sampling scheme was used. Accordingly, the value $u_k = U(0,1)/\theta_k$ was generated for each household $k \in s$, where $U(0,1)$ denotes a random outcome from the uniform distribution. Then the 600 households with the smallest values of u_k were selected as respondents r. As a result, the household k responds with probability θ_k^0, very close to θ_k. As shown in [7], $\lim_{m\to\infty} \theta_k/\theta_k^0 = 1$.

The response set was generated 1000 times. From each response set the estimators $\bar{y}_{\text{unw}}$, $\bar{y}_{\text{cal}}$, $\bar{y}_{\text{pro}}$, $\bar{y}_{\text{f}}$, $\bar{y}_{\text{scf}}$, $\bar{y}_{\text{mix}}$ were computed. They are also referred to as UNW, CAL, PRO, f, SCf, and MIX. One by one, all 13 variables of our data set were taken in the role of the y-variable. The arrangement where $\mathbf{x}$-variables can also serve as study variables may seem unusual. The reason to do so was to observe the behavior of estimators under perfect linear relationship.

The absolute relative bias was computed for each estimator and for each y-variable according to the formula

$$\text{ARB} = \frac{|E_{rep}\hat{\bar{y}} - \bar{y}_s|}{\bar{y}_s}, \tag{25}$$

where $\hat{\bar{y}}$ denotes an estimator and E_{rep} refers to its mean over all 1000 repetitions.

The overall absolute relative bias of an estimator is just the average of ARB in (25) over all y-variables. This is a measure characterizing the behavior of an estimator for the entire survey.

6. Simulation results

Response probabilities (Table 1) were varying from very small to very high. Their mean is equal to the response rate 0.6.

Table 1. Characteristics of the response probabilities.

Min	1st quartile	Median	Mean	3rd quartile	Max
0.6×10^{-6}	0.415	0.685	0.600	0.846	0.871

Table 2 displays correlations in s. The $\mathbf{x}$-variable HD_active is correlated with the income related variables. Those households with employed heads get less transfers, they have higher income and higher expenditures. Income and expenditure have very high correlation (0.716). Correlations with response probabilities follow the model (22). The strongest correlation is with the $\mathbf{x}$-variable HD_sex (-0.655).

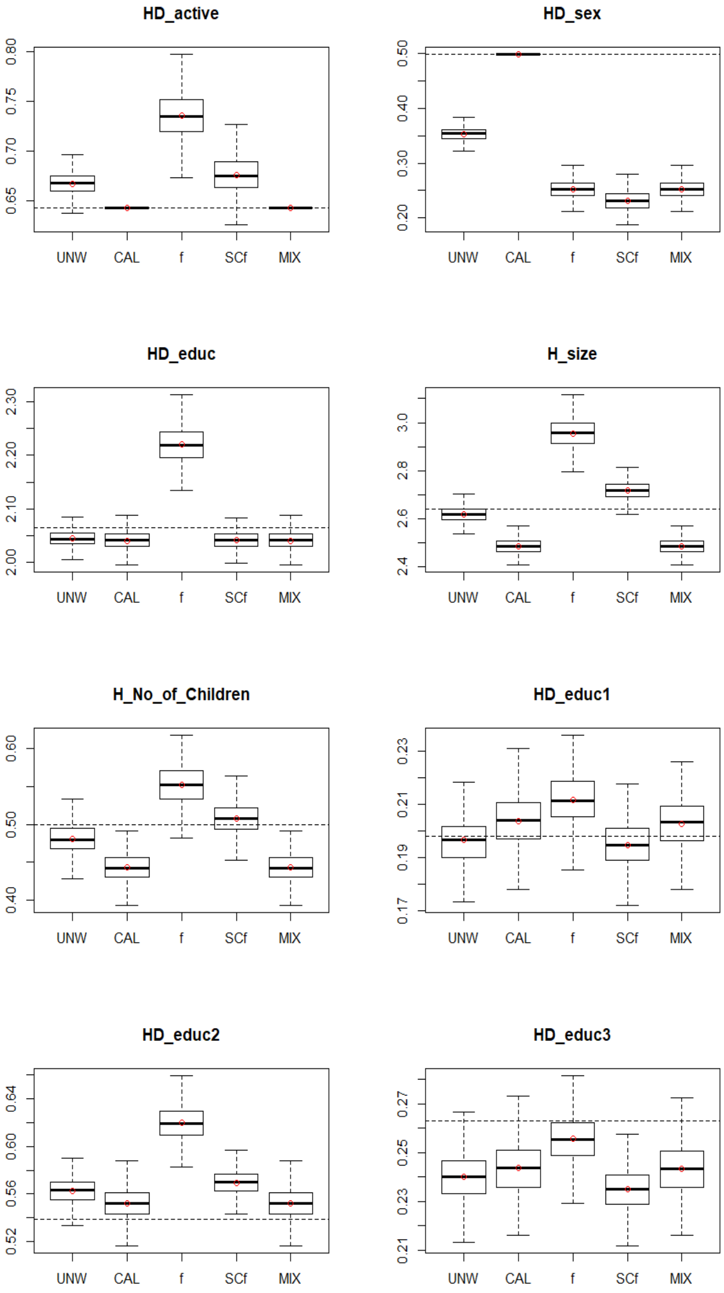
Table 2. Correlations between $\mathbf{x}$-variables, income related variables and response probabilities θ.

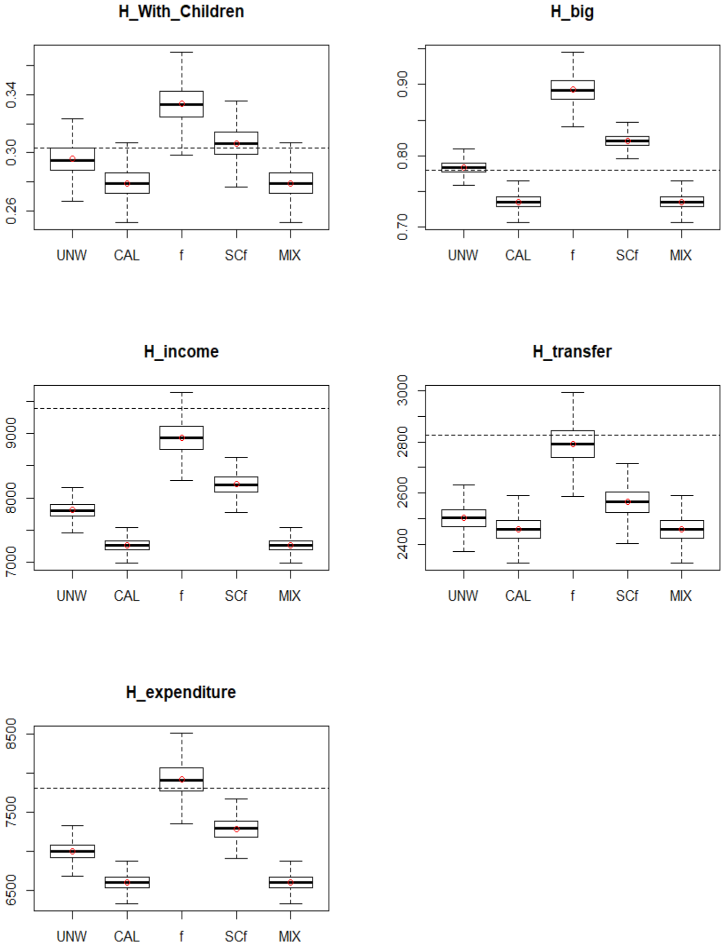
	HD_sex	HD_active	H_inc.	H_trans.	H_exp.	θ
HD_sex	1.000	-0.176	-0.190	0.006	-0.174	-0.655
HD_active	-0.176	1.000	0.422	-0.329	0.412	0.120
H_inc.	-0.190	0.422	1.000	0.189	0.716	-0.453
H_trans.	0.006	-0.329	0.189	1.000	0.097	-0.254
H_exp.	-0.174	0.412	0.716	0.097	1.000	-0.288
θ	-0.655	0.120	-0.453	-0.254	-0.288	1.000

On Figure 1 we see distributions of our estimators around the true value $\bar{y}_s$ over 1000 repeated response sets. The $\mathbf{x}$-vector is a group vector (23). In this particular case, $\bar{y}_{\text{pro}} = \bar{y}_{\text{cal}}$. Therefore, $\bar{y}_{\text{pro}}$ was dropped from Figure 1 and Tables 3–4. Calibration estimator is exact for HD_sex and HD_active, these variables are used as auxiliary variables in the calibration estimator. Basically, here is the case where y-variable depends linearly on $\mathbf{x}$-variables. We see that f- and SCf-estimators still have variability and are biased in this case, even more than the simple UNW-estimator.

The behavior of the CAL-estimator is surprising. Usually it is expected to decrease bias compared to the UNW-estimator. But here, it increases bias in many cases. Especially large bias appears for income related variables – income, transfer, and expenditure. This happens despite the correlations, not very strong but of order 0.4, that these y-variables have with the calibration variable HD_active. Quite strong correlation between another calibration variable HD_sex and response probabilities (≈ -0.7) has not been able to reduce bias either. Increased bias in calibration with some calibration functions has been also experienced by [4]; this has been in spite of the strong relationship between the study and auxiliary variables.

FIGURE 1. Distributions of estimators over 1000 response sets for 13 variables, the true $\bar{y}_s$ is shown by dashed line, the mean of each distribution by circle, the median by bold line. The $\mathbf{x}$-vector of estimators is in (23).





One notes in Table 5 that CAL is considerably improved when the calibration is done instead with the **x**-vector (24) that contains the income related variable H_expenditure. Nevertheless, in our case, for the income related variables (excluding due to natural reasons H_expenditure), the f-estimator is the best.

Compared with unweighted estimator UNW, the f- and the CAL-estimators tend to be on opposite side of it. Very often, they are even on opposite side of the true $\bar{y}_s$. The mixture estimator MIX behaves similarly to the CAL-estimator for almost all study variables.

Absolute relative biases in Table 3 show that for different variables different estimators are the best. The overall relative bias is smallest for the UNW-estimator (Table 4). In Table 4, the average is taken over 11 variables, i.e., HD_sex and HD_active were dropped out. These variables are used in the auxiliary **x**-vector, they are known in s, and there is no need to estimate them. They are used here to demonstrate the case where the calibration estimator is exact.

We have also studied the effect of the auxiliary vector (24) on the estimators CAL, PRO, f, SCf, MIX. The vector (24) involves additional auxiliary variable H_expenditure. Stronger auxiliary vector reduces bias of CAL-estimator for almost all variables. Note that with auxiliary vector (23), PRO- and CAL-estimators are equal, but with auxiliary vector (24) they are different, CAL appeared to be better than PRO. Comparing Tables 4 and 6, we see that the overall absolute relative bias has strongly decreased for the CAL-estimator, but increased for f- and SCf-estimators. Comparing absolute relative biases by each of the 13 variables (Tables 3 and 5), we see the same tendency: ARB of CAL-estimator has decreased, but of f-estimator increased. Remarkable is the position of the UNW-estimator. In spite of the stronger calibration, the CAL-estimator has larger ARB than the UNW-estimator for several study variables. In the overall, these two estimators have almost equal ARB (Table 6).

Table 3. Absolute relative bias, **x**-vector in (23).

	UNW	CAL	f	SCf	MIX
HD_active	0.0385	0.0000	0.1438	0.0518	0.0000
HD_sex	0.2910	0.0000	0.4936	0.5339	0.4936
HD_educ	0.0100	0.0117	0.0752	0.0113	0.0117
H_size	0.0092	0.0592	0.1191	0.0284	0.0592
H_No_of_Children	0.0357	0.1111	0.1066	0.0174	0.1111
HD_educ1	0.0072	0.0280	0.0692	0.0170	0.0237
HD_educ2	0.0436	0.0241	0.1493	0.0567	0.0241
HD_educ3	0.0870	0.0736	0.0278	0.1060	0.0740
H_With_Children	0.0237	0.0788	0.1008	0.0117	0.0788
H_big	0.0059	0.0566	0.1461	0.0541	0.0566
H_income	0.1683	0.2259	0.0485	0.1259	0.2259
H_transfer	0.1144	0.1306	0.0127	0.0926	0.1306
H_expenditure	0.1041	0.1550	0.0140	0.0673	0.1550

Table 4. Overall absolute relative bias, **x**-vector in (23).

UNW	CAL	f	SCf	MIX
0.0505	0.0800	0.0855	0.0521	0.0796

Table 5. Absolute relative bias, $\mathbf{x}$-vector in (24).

	UNW	CAL	PRO	f	SCf	MIX
HD_active	0.0385	0.0000	0.0443	0.1987	0.0643	0.0000
HD_sex	0.2910	0.0000	0.0282	0.4552	0.5160	0.4552
HD_educ	0.0100	0.0085	0.0227	0.1042	0.0198	0.0085
H_size	0.0092	0.0287	0.0668	0.1408	0.0124	0.0287
H_No_of_Children	0.0357	0.0527	0.0988	0.1138	0.0109	0.0527
HD_educ1	0.0072	0.0431	0.0260	0.1346	0.0072	0.0431
HD_educ2	0.0436	0.0006	0.0008	0.2067	0.0713	0.0006
HD_educ3	0.0870	0.0308	0.1831	0.0475	0.1542	0.0475
H_With_Children	0.0237	0.0392	0.0809	0.1223	0.0041	0.0392
H_big	0.0059	0.0329	0.0382	0.1721	0.0410	0.0329
H_income	0.1683	0.1653	0.2162	0.0510	0.1578	0.1653
H_transfer	0.1144	0.0903	0.1554	0.0100	0.1213	0.0903
H_expenditure	0.1041	0.0000	0.0320	0.0414	0.1487	0.0421

Table 6. Overall absolute relative bias, $\mathbf{x}$-vector in (24).

UNW	CAL	PRO	f	SCf	MIX
0.0505	0.0492	0.0889	0.1103	0.0600	0.0509

7. Conclusion

This paper studied estimation under nonresponse. We have constructed a new estimator, the f-estimator, using regression of the study variable y on the vector of auxiliary variables $\mathbf{x}$. Uncomputable regression slope in sample s is estimated in two different ways, one of them results in the well-known linear calibration estimator, the other one in the new f-estimator. The estimators are compared theoretically and experimentally. The estimators have opposite behavior: if one of them is smaller than the simple unweighted estimator, then the other tends to be bigger. Weights of the calibration estimator and of the f-estimator are general inverses (sometimes exact inverses) of each other. Based on the opposite nature of these two estimators, we have defined a mixture estimator. It has smaller upper bound for the absolute difference from target sample mean.

In simulation experiment on real data, the f-estimator is better than the calibration estimator for some study variables. In fact, for different variables different estimators are the best. Sometimes calibration may even increase bias compared to the simple unweighted estimator. One can say that in the overall sense, the UNW-estimator is superior. Its average absolute relative bias (average over all study variables) is the smallest (Table 4) or nearly the smallest (Table 6). In the latter case the winner is the CAL-estimator. The reader should keep in mind that ranking of estimators is based on simulation results that depend on choices made in this paper.

In survey practice the same auxiliary vector $\mathbf{x}$ and, consequently, the same set of weights is used for the nonresponse adjustment in the entire survey. The vector $\mathbf{x}$ cannot be strongly related to each survey variable, and therefore cannot work well for each variable in widely used calibration estimator. Even if strongly related, it does not remove all the bias in that estimator. Based on the same vector $\mathbf{x}$, the constructed f-estimator is not perfect either. But it works well for some variables where calibration does not work. We proposed here a mixture of the CAL- and the f-estimator, but it appeared to behave similarly to the CAL-estimator.

The calibration and the f-estimator are both constructed by using regression tools. As noticed by many authors, the dilemma with nonresponse is inconsistent regression. A regression model in s does not hold in r due to selective response mechanism, the unequal response probabilities. Bias of the calibration estimator is largely defined by the difference in slopes $\mathbf{b}_r - \mathbf{b}_s$, [9]. Bias of the f-estimator is respectively defined by the difference $\mathbf{b}_{sr} - \mathbf{b}_s$. But for different y-variables one of these two biases is smaller. Consequently, it is profitable to choose between these estimators.

Acknowledgements

The author thanks the referee for constructive comments and suggestions. This work was partly supported by the Institutional Research Funding IUT34-5 of Estonia.

References

[1] J. F. Beaumont, *On the use of data collection process information for the treatment of unit nonresponse through weight adjustment*, Survey Methodol. **31** (2005), 227–231.

[2] J. M. Brick, *Unit nonresponse and weighting adjustments: A critical review*, J. Offic. Statist. **29** (2013), 329–353.

[3] J. C. Deville and C.-E. Särndal, *Calibration estimation in survey sampling*, J. Amer. Statist. Assoc. **87** (1992), 375-382.

[4] D. Haziza and É. Lesage, *A discussion of weighting procedures for unit nonresponse*, J. Offic. Statist. *3*2 (2016), 129–145.

[5] D. G. Horvitz and D. J. Thompson,*A generalization of sampling without replacement from a finite universe*, J. Amer. Statist. Assoc. **47** (1952), 663–685.

[6] R. J. A. Little and S. Vartivarian, *Does weighting for nonresponse increase the variance of survey means?*, Survey Methodol. **31** (2005), 161–168.

[7] B. Rosen, *On inclusion probabilities for order πps sampling*, J. Statist. Plann. Inference **90** (2000), 117–143.

[8] C.-E. Särndal, *The 2010 Morris Hansen lecture: Dealing with survey nonresponse in data collection, in estimation*, J. Offic. Statist. **27** (2011), 1–21.

[9] C.-E. Särndal and P. Lundquist, *Inconsistent regression and nonresponse bias: Exploring their relationship as a function of response imbalance*, J. Offic. Statist. **33** (2017), 709–733.

[10] C.-E. Särndal and S. Lundström, *Estimation in Surveys with Nonresponse*, John Wiley & Sons, Ltd., Chichester, 2005.

[11] C.-E. Särndal, I. Traat, and K. Lumiste, *Interaction between data collection and estimation phases in surveys with nonresponse*, Statistics in Transition **19** (2018), 183–200.

Institute of Mathematics and Statistics, University of Tartu, 50090 Tartu, Estonia

E-mail address: `imbi.traat@ut.ee`