Lithuanian Journal of Statistics *2015, vol. 54, No 1, pp. 69-83*
Lietuvos statistikos darbai *2015, 54 t., Nr. 1, 69-83 p.*
www.statisticsjournal.lt

COMPARISONS OF SOME WEIGHTING METHODS FOR NONRESPONSE ADJUSTMENT

Bernardo João Rota[1,3], Thomas Laitila[1,2]

[1] Department of Statistics, Örebro University
[2] Department of Research and Development, Statistics Sweden
[3] Department of Mathematics and Informatics, Eduardo Mondlane University
Address: [1] Fakultetsgatan 1, 702 81 Örebro, Sweden
[2]Klostergatan 23, 703 61 Örebro, Sweden
[3]Ave. Julius Nyerere/Campus Principal 3453, Maputo, Mozambique
E-mail: [1]bernardo.rota@oru.se, [2]thomas.laitila@oru.se

Received: August 2015 Revised: September 2015 Published: October 2015

Abstract. Sample and population auxiliary information have been demonstrated to be useful and yield approximately equal results in large samples. Several functional forms of weights are suggested in the literature. This paper studies the properties of calibration estimators when the functional form of response probability is assumed to be known. The focus is on the difference between population and sample level auxiliary information, the latter being demonstrated to be more appropriate for estimating the coefficients in the response probability model. Results also suggest a two-step procedure, using sample information for model coefficient estimation in the first step and calibration estimation of the study variable total in the second step.

Key words : calibration, auxiliary variables, response probability, maximum likelihood.

1. Introduction

Weighting is widely applied in surveys to adjust for nonresponse and correct other nonsampling errors. The literature contains many different proposals for nonresponse weighting methods. These methods usually treat the set of respondents as a second-phase sample [2], the elements of the response set being tied to a twofold weight compensating for both sampling and nonresponse. These weights, in particular those for nonresponse adjustment, are constructed with the aid of auxiliary information.

Treating the response set as a random subset of the sample set justifies associating each respondent with a probability of being included in the response set. Estimating this probability with aid the of auxiliary information and multiplying it by the sample inclusion probability gives an estimate of the probability of having a unit in the response set. The observations of target variable values are weighted by the reciprocals of these estimated probabilities and summed over the set of respondents, giving an estimated population total. This is known as direct nonresponse weighting adjustment [13]. One example of this method is the cell weighting approach described by [11].

Alternatively, the auxiliary information is incorporated into the estimation such that the second-phase weight adjustments are determined implicitly. Such estimators are known as nonresponse weighting adjustments (see [12]), and one example is the calibration method suggested by [18]. [5] combine the two approaches. They assume the response probability function to be known, and calibration serves as the means of estimating the parameters of this function. Once the parameters have been determined, the inverse of the estimated response probabilities are used as nonresponse adjustment factors.

The main feature of the calibration approach is to make the best use of available auxiliary information. When the response mechanism is assumed to be known and of the form $p(\cdot;\mathbf{g})$, parameter $\mathbf{g}$ is deemed a nuisance parameter [14]; this means that, although the information associated with its estimator $\hat{\mathbf{g}}$ is important, the primary objective is to estimate the target, say, the total $Y = \sum_U y_k$. Using calibration to estimate the unknown parameters confers a different meaning on the estimation problem, in the sense that auxiliary variables are selected to provide good auxiliary information for

estimating the parameters with good precision. This will in turn imply good precision for the estimates of response probabilities. Thus, when the response probability function is known, our principle is to view the problem of estimation in two distinct moments: estimation of parameters and estimation of targets respectively.

As noted in [4], the probabilities to respond are usually functions of the sample and survey conditions, that is, the response probability for a specific individual may change when the survey conditions also well change (see also [3]). However, the mechanism leading to response/nonresponse for a sampled individual is generally not known [14]. Thus, estimation in the presence of nonresponse requires some kind of modeling, explicitly or implicitly (see [5]). An implicit modeling for nonresponse adjustment can be found in [1], while [12] gives an example of explicit modeling. This paper considers nonresponse adjustment methods when the response probability function is assumed to be known up to a set of unknown coefficients. Under this assumption, direct weighting estimators can be used when the response probability model is estimated using, for example, the maximum likelihood estimator. An alternative here is to estimate the response probability model using calibration, as suggested by [5]. This calibration estimator requires only the values of the covariates in the response model for the sample units in the response set, while maximum likelihood needs the values of those variables for the whole sample. One issue considered is the level of information used in calibration. An option is to use either sample or population level information when calibrating for response probability coefficient estimates. This paper contributes by demonstrating that the asymptotic variance of the coefficient estimator is smaller when sample level information is used. A simulation study is performed in order to investigate the properties of the estimators for small sample sizes. We also suggest a two-step procedure in which sample level information is used for response probability model estimation in the first step, and population level information is used for estimating population characteristics in the second step. Furthermore, the importance of correlating auxiliary variables with model and study variables is addressed.

The simulation study performed is based on data from a survey on real estate, and the bias and variance properties of the estimators are considered. Several estimators are studied, including the Horvitz-Thompson (HT) estimator using true model coefficients, direct weighting using maximum likelihood (ML) estimates of coefficients, and calibration-estimated coefficients, where calibration uses sample or population information. Two-step estimators using ML-estimated and calibration-estimated coefficients, respectively, are included, as is the linear calibration (LC) estimator [21].

The estimators studied are introduced in the next section. Section 3 compares the variance of the model parameter calibration estimators when based on population and sample level information. The results of a simulation study are reported in Section 4, and a discussion of the findings is saved for the final section.

2. Estimators under nonresponse

Sample s of size n is drawn from the population $U = \{1,2,...,k,...,N\}$ of size N using a probability sampling design, $p(s)$, yielding first and second order inclusion probabilities $\pi_k = \Pr(k \in s) > 0$ and $\pi_{kl} = \Pr(k,l \in s) > 0$, respectively, for all $k,l \in U$. Let $r \subset s$ denote the response set. Units in the sample respond independently with a probability $p_k = \Pr(k \in r | k \in s) > 0$, for the known functional form $p_k = p(\mathbf{z}_k^t \mathbf{g})$ evaluated at $\mathbf{g} = \mathbf{g}_\infty$, an interior point of the parameter space $\mathbf{g} \in \mathbf{G}$, and $\mathbf{z}_k$ is a vector of model variables. Both $\mathbf{g}$ and $\mathbf{z}_k$ are column vectors of dimension K. Furthermore, we assume that conditional on the auxiliary variables, the response probability is independent of the survey variable of interest, which is known as MAR assumption (e.g. [23]). Define the indicators:

$$I_k = \begin{cases} 1 \ if \ k \in s \\ 0 \ else \end{cases} \quad \text{and} \quad R_k = \begin{cases} 1 \ if \ k \in r | I_k = 1 \\ 0 \ if \ k \notin r | I_k = 1 \end{cases}.$$

The survey variable of interest is y, and its population total, $Y = \sum_U y_k$, is to be estimated. We can then construct an estimator for Y of the form:

$$\hat{Y}_W = \sum_r w_k y_k. \tag{1}$$

The weights, w_k, can be defined in various ways but usually have the form $w_k = d_k v_k$, where $d_k = 1/\pi_k$ is the design weight and v_k is a factor adjusting for example, for nonresponse. These factors make use of auxiliary information. The auxiliary vector is $\mathbf{x}_k$, with dimension $P \times 1$, where $P \geq K$ and $\mathbf{X} = \sum_U \mathbf{x}_k$ denotes its population total.

a. Direct nonresponse weighting adjustment

One alternative of weights w_k in (1) is given by $w_k = d_k h(\mathbf{z}_k^t \hat{\mathbf{g}})$, where $h(\cdot) = p^{-1}(\cdot)$ and $\hat{\mathbf{g}}$ is an estimator of $\mathbf{g}_\infty$. Assume $p(\mathbf{z}_k^t \mathbf{g})$ to be differentiable w.r.t. $\mathbf{g}$ and define the weighted log likelihood function of the response distribution

$$l(\mathbf{g}) = \sum_s d_k \left[R_k ln\left(p(\mathbf{z}_k^t \mathbf{g})\right) + (1 - R_k) ln\left(1 - p(\mathbf{z}_k^t \mathbf{g})\right)\right]. \tag{2}$$

The first order conditions for the maximum likelihood estimator (MLE) are given by

$$\frac{\partial l(\mathbf{g})}{\partial \mathbf{g}} = \sum_s d_k \left(\frac{R_k - p(\mathbf{z}_k^t \mathbf{g})}{p(\mathbf{z}_k^t \mathbf{g})(1 - p(\mathbf{z}_k^t \mathbf{g}))} \cdot \frac{\partial p(\mathbf{z}_k^t \mathbf{g})}{\partial \mathbf{g}} \right) = \mathbf{0}. \tag{3}$$

The first order conditions in (3) are nonlinear in $\mathbf{g}$ in general, and a numerical optimization method, such as the Newton-Raphson algorithm, is required to obtain the desired $\hat{\mathbf{g}}_{ML}$. Observe that $\frac{\partial l(\mathbf{g})}{\partial \mathbf{g}}$ results in a K-dimensional column vector of partial derivatives, each with respect to one component of $\mathbf{g}$. For matrix derivations, see [19].

With a calculated $\hat{\mathbf{g}}_{ML}$, the estimator (1) takes the form

$$\hat{Y}_{DN\_ML} = \sum_r d_k h(\mathbf{z}_k^t \hat{\mathbf{g}}_{ML}) y_k \tag{4}$$

where the subscript (DN_ML) stands for direct nonresponse weighting by ML. This estimator is asymptotically unbiased for the population total Y under the assumptions established for Theorem 1 by [13].

b. The propensity score calibration estimation

[5] propose a calibration direct nonresponse adjusted estimator (1), where the weights w_k are the products of the design weight and the reciprocal of the estimated response probability $p(\mathbf{z}_k^t \hat{\mathbf{g}}_{CAL})$ for the element k in r, i.e., $w_k = d_k h(\mathbf{z}_k^t \hat{\mathbf{g}}_{CAL})$, so that the estimator (1) becomes

$$\hat{Y}_W = \sum_r d_k h(\mathbf{z}_k^t \hat{\mathbf{g}}_{CAL}) y_k. \tag{5}$$

This estimator is similar to (4) in form but makes use of calibration for the estimation of $\mathbf{g}_\infty$ instead of ML. The strategy is to estimate $\mathbf{g}_\infty$ using the solution to the calibration equation

$$\mathbf{X} = \sum_r d_k h(\mathbf{z}_k^t \mathbf{g}) \mathbf{x}_k \tag{6}$$

Assuming $h(\mathbf{z}_k^t \mathbf{g})$ to be twice differentiable, [5] suggest an estimator defined by minimizing an objective function derived from (6), assuming the difference $\mathbf{e} = \mathbf{X} - \sum_r d_k h(\mathbf{z}_k^t \mathbf{g}_\infty) \mathbf{x}_k$ to be asymptotically normal distributed. Here, we do not impose normality assumption and derive their estimator slightly differently.

Assume that $P \geq K$ and define the distance function as

$$d(\mathbf{g}) = \left(\mathbf{X} - \sum_U I_k d_k R_k h(\mathbf{z}_k^t \mathbf{g}) \mathbf{x}_k \right) \tag{7}$$

Let Σ_n be a $P \times P$ symmetric nonnegative definite matrix converging in probability to the positive definite matrix Σ, when the sample size grows arbitrarily large. Construct a weighted quadratic distance as follows:

$$D(\mathbf{g}) = 2^{-1} d^t(\mathbf{g}) \Sigma_n d(\mathbf{g}) \tag{8}$$

Then, the [5] estimator of $\mathbf{g}_\infty$ is defined as the minimizer of (8). Note that this estimator is a generalized method of moments (GMM) estimator, where minimizing (8) entails solving the estimating equations ([7], p. 378)

$$\mathbf{d}^t(\mathbf{g})\Sigma_n d(\mathbf{g}) = \mathbf{0} \tag{9}$$

that results in the equation

$$\hat{\mathbf{g}}_{c1} = \hat{\mathbf{g}}_{c0} - (\mathbf{d}^t(\hat{\mathbf{g}}_{c0})\Sigma_n \mathbf{d}(\hat{\mathbf{g}}_{c0}))^{-1}\mathbf{d}^t(\hat{\mathbf{g}}_{c0})\Sigma_n d(\hat{\mathbf{g}}_0) \tag{10}$$

after an initial guess $\hat{\mathbf{g}}_{c0}$
where,

$$\mathbf{d}(\mathbf{g}) = -\frac{\partial d(\mathbf{g})}{\partial \mathbf{g}} = \sum_U I_k d_k R_k \tilde{h}(\mathbf{z}_k^t \mathbf{g})\mathbf{x}_k \mathbf{z}_k^t \tag{11}$$

$\tilde{h}(a)$ is the first derivative of $h(a)$ and $\mathbf{d}(\mathbf{g})$ is assumed to be of full rank. Section 3 provides some details in the derivation of (10).

The [5] propensity calibration estimator is obtained upon the convergence of (10) and is given by:

$$\hat{Y}_{PS} = \sum_r d_k h(\mathbf{z}_k^t \hat{\mathbf{g}}_{c1}) y_k \tag{12}$$

c. The linear calibration estimator

The LC estimator is defined as the estimator (1) with the weights, w_k, satisfying the calibration constraint

$$\sum_r w_k \mathbf{x}_k = \mathbf{X} \tag{13}$$

where $w_k = d_k v_k$, $v_k = 1 + \lambda_r^t \mathbf{z}_k$, and $\mathbf{z}_k$ is a variable vector with the same dimension as $\mathbf{x}_k$. $\mathbf{z}_k$ is assumed known at least up to the set of respondents and is called an instrument vector if it differs from $\mathbf{x}_k$. This system yields the vector $\lambda_r^t = (\mathbf{X} - \sum_r d_k \mathbf{x}_k)^t \left(\sum_r d_k \mathbf{z}_k \mathbf{x}_k^t\right)^{-1}$. The linear calibration estimator for the total Y is then given by

$$\hat{Y}_{LC} = \sum_r d_k v_k y_k = \sum_r w_k y_k \tag{14}$$

In this setting, no explicit modeling for response or outcome is required. Instead, the method relies on the strength of the available auxiliary information. Although this is not the basic tenet, the v_k factor gives the impression of a linear approximation of the reciprocal of the response probability in the sense that a good linear approximation of $h(\mathbf{z}_k^t \mathbf{g})$ brings about a linear calibration estimator with good statistical properties (see [15]).

d. The two-step calibration estimator

[21] describe the two-step calibration approach. The first- and second-step weights are constructed according to the principle of combining population and sample levels auxiliary information. In the first step, sample level information is used to construct preliminary weights, w_{1k}, such that $\sum_r w_{1k}\mathbf{x}_k^s = \sum_s d_k \mathbf{x}_k^s$, where $\mathbf{x}_k^s$ is a J-dimensional column vector of auxiliary variables with known values for all sampled units. In the second step, weights w_{1k} replace the design weights in the derivation of the single step calibration estimator (14), and the final weights, w_k, satisfy $\sum_r w_k \mathbf{x}_k = \mathbf{X}$. Here, $\mathbf{X} = \sum_U \mathbf{x}_k^U$ if $\mathbf{x}_k = \mathbf{x}_k^U$ or $\mathbf{X} = \begin{pmatrix} \sum_U \mathbf{x}_k^U \\ \sum_s d_k \mathbf{x}_k^s \end{pmatrix}$ if $\mathbf{x}_k = \begin{pmatrix} \mathbf{x}_k^U \\ \mathbf{x}_k^s \end{pmatrix}$, with $\mathbf{x}_k^U$ being a P-dimensional column vector of auxiliary variables with known values for all respondents; moreover, their population totals are also known.

[16] also suggest a two-step calibration estimation assuming the known functional form of the response mechanism. The estimation process is conceptually different from the one suggested in [21], where the second-step weights are based on the first-step weights. The prediction approach supports the estimation setting suggested by [16].

Here, the concept of two-step estimation is implemented differently to ([21], p. 88). As in [16], we assume a specified response mechanism, $p(\mathbf{z}_k^t \mathbf{g})$, where initial weights are calculated as $w_{1k} = d_k h(\mathbf{z}_k^t \hat{\mathbf{g}})$ after calculating $\hat{\mathbf{g}}$. Depending on whether the auxiliary vector $\mathbf{z}_k$ is known up to the response set or the sample gives different options for the estimators of the true value of $\mathbf{g}$. For example, if $\mathbf{z}_k$ is known

up to the sample level, then $\hat{\mathbf{g}}$ may be the MLE. If $\mathbf{z}_k$ is known only up to the response set level, $\hat{\mathbf{g}}$ is estimated using calibration against sample level information, i.e., $\sum_s d_k \mathbf{x}_k = \sum_r d_k h(\mathbf{z}_k^t \mathbf{g}) \mathbf{x}_k$.

In the second step, the population auxiliary data are employed for estimating targets. That is, the second step weights, w_k, are given by $w_k = w_{1k} v_k$ with $v_k = 1 + \boldsymbol{\lambda}_2^t \mathbf{x}_k$ and $\boldsymbol{\lambda}_2^t = (\mathbf{X} - \sum_r w_{1k} \mathbf{x}_k)^t \left(\sum_r w_{1k} \mathbf{x}_k \mathbf{x}_k^t\right)^{-1}$.

3. Asymptotic variance of the estimated response model parameters

[12] and [13] provide analytical and empirical justification for the efficiency gain when using estimated response probabilities in place of the true response probabilities, proving what had been noted by [20], namely, the estimated probabilities outperform true probabilities. [12] and [13] demonstrate this feature in a context of direct and regression adjustments where the scores are estimated using an ML procedure. This efficiency gain by using estimated probabilities can be interpreted as resulting from the lack of the location-invariance property of the HT estimator (e.g. [9], p. 10). Using true response probabilities, observations are given weights equal to the reciprocal of the probability of having the unit in the response set. However, the size of the response set is random due to nonresponse, meaning that it is not location invariant. When using ML-estimated response probabilities, estimates satisfy moment conditions at the sample level. This can be expected to reduce variance but will not in general yield an invariance property.

Similar to the difference between true and estimated response probabilities, the difference between population and sample level information in the calibration estimator is considered. The precision of model parameters can be expected to affect the precision of target variable estimates. Here precision is auxiliary information dependent. As noted in [4] and [24], the strength of the relationships between the auxiliary variables and the response probabilities or study variables is crucial for the efficient performance of the weighting adjustment methods. Auxiliary information may be available at different levels, such as the population or sample levels [8]. Under nonresponse, this auxiliary information is used for correcting nonresponse bias and reducing the variance of the estimator. In particular, as [23] states, sample level information is suited for nonresponse adjustment rather than variance reduction, because nonresponse affects only the location of means and not their variation.

According to the quasi-randomization setup, response set generation is an experiment made conditional on the sample. On the other hand, calibrating weights against population level information means that estimation is made unconditional on the sample. Calibration based on sample level information is therefore expected to yield more efficient estimators of response probability parameters.

Reformulating the calibration equation as

$$\mathbf{X} - \sum_r w_k \mathbf{x}_k = \left(\mathbf{X} - \sum_s d_k \mathbf{x}_k\right) + \left(\sum_s d_k \mathbf{x}_k - \sum_r w_k \mathbf{x}_k\right),$$

illustrates that calibration against population level information brings a source of uncertainty that does not depend on the response probability distribution, i.e., variation due to the first phase sampling represented by the first term of the right-hand side of this equation. Calibrating against sample level information excludes this term, and the single source of randomness involved is the one defined by the conditional response distribution.

For more formal results, assume the asymptotic framework in which both the sample and population sizes are to increase to infinity (see, [10]), and assume further that the minimizer of (8) is consistent.

Using result 9.3.1 in [22], the covariance matrix of $d(\mathbf{g})$ evaluated at the true value $\mathbf{g} = \mathbf{g}_\infty$ is given by

$$E\left(d(\mathbf{g}_\infty) d^t(\mathbf{g}_\infty)\right) = \boldsymbol{\Pi}_1 + \boldsymbol{\Pi}_2 = \boldsymbol{\Pi} \tag{15}$$

where, $E(d(\mathbf{g}_\infty)) = \mathbf{0}$, $\boldsymbol{\Pi}_1 = \sum_{k \in U} \sum_{l \in U} \frac{\pi_{kl} - \pi_k \pi_l}{\pi_k \pi_l} \mathbf{x}_k \mathbf{x}_l^t$ and $\boldsymbol{\Pi}_2 = \sum_U \frac{(h(\mathbf{z}_k^t \mathbf{g}_\infty) - 1)}{\pi_k} \mathbf{x}_k \mathbf{x}_k^t$, with the expectations being taken jointly with respect to the sampling design $p(s)$ and the response distribution $p(\mathbf{z}_k^t \mathbf{g})$.

Consider equation (9) with $\mathbf{g}$ replaced by its solution $\hat{\mathbf{g}}$, and apply the mean value theorem to decompose $d(\hat{\mathbf{g}})$, obtaining the following equation:

$$d(\hat{\mathbf{g}}) = d(\mathbf{g}_\infty) + \mathbf{d}(\bar{\mathbf{g}})\,(\hat{\mathbf{g}} - \mathbf{g}_\infty)\,. \tag{16}$$

Then, we can substitute $d(\hat{\mathbf{g}})$ in (9) by the r.h.s of (16) and get:

$$\mathbf{d}^t(\hat{\mathbf{g}})\Sigma d(\mathbf{g}_\infty) + \mathbf{d}^t(\hat{\mathbf{g}})\Sigma\mathbf{d}(\bar{\mathbf{g}})(\hat{\mathbf{g}} - \mathbf{g}_\infty) = \mathbf{0} \tag{17}$$

where, $\bar{\mathbf{g}}$ lies in the segment between $\hat{\mathbf{g}}$ and $\mathbf{g}_\infty$.
We can rewrite (17) as:

$$(\hat{\mathbf{g}} - \mathbf{g}_\infty) = -\left(n^{-1}\mathbf{d}^t(\hat{\mathbf{g}})\Sigma_n n^{-1}\mathbf{d}(\bar{\mathbf{g}})\right)^{-1} n^{-1}\mathbf{d}^t(\hat{\mathbf{g}})\Sigma_n\left(n^{-1}d(\mathbf{g}_\infty)\right) \tag{18}$$

Under appropriate assumptions, we have that $\mathbf{d}(\hat{\mathbf{g}}) - \mathbf{d}(\mathbf{g}_\infty) = o_p(1)$. Let, $\mathbf{D} = plim_{n\to\infty} n^{-1}\mathbf{d}(\mathbf{a})$, where $\mathbf{a}$ stands for $\hat{\mathbf{g}}$, $\bar{\mathbf{g}}$ or $\mathbf{g}_\infty$. We replace $\mathbf{d}$ in (17) by its corresponding limit and obtain the asymptotic variance of the estimated model parameters as:

$$\begin{aligned} Avar\left(\sqrt{n}\,(\mathbf{g}_\infty - \hat{\mathbf{g}})\right) &= Avar\left(\left[\mathbf{D}^t\Sigma\mathbf{D}\right]^{-1}\mathbf{D}^t\Sigma\sqrt{n}(n^{-1}d(\mathbf{g}_\infty))\right) \\ &= \left[\mathbf{D}^t\Sigma\mathbf{D}\right]^{-1}\mathbf{D}^t\Sigma\Pi\Sigma\mathbf{D}\left[\mathbf{D}^t\Sigma\mathbf{D}\right]^{-1} \end{aligned} \tag{19}$$

where Π is the probability limit of $n^{-1}E\left(d(\mathbf{g}_\infty)d^t(\mathbf{g}_\infty)\right)$.

The choice of $\Sigma = \Pi^{-1}$ yields

$$Avar\left(\sqrt{n}\,(\hat{\mathbf{g}} - \mathbf{g}_\infty)\right) = \left[\mathbf{D}^t\Pi^{-1}\mathbf{D}\right]^{-1} \tag{20}$$

which is equivalent to expression (9.80) in [7]. Observe that equation (10) results from (17) after replacing $\mathbf{d}(\mathbf{a})$ with the computable entity $\mathbf{d}(\hat{\mathbf{g}}_{c0}) = \sum_r d_k h(\mathbf{z}_k^t\hat{\mathbf{g}}_{c0})\mathbf{x}_k$.

Now, for calibration at the sample level, rewrite equation (7) as

$$d^s(\mathbf{g}) = \left(\sum_s d_k\mathbf{x}_k - \sum_s d_k R_k h(\mathbf{z}_k^t\mathbf{g})\mathbf{x}_k\right). \tag{21}$$

The conditional expectation of $d^s(\mathbf{g}_\infty)$ with respect to the response distribution is zero. This implies that the covariance (15) in this case is $\Pi_2 = \sum_U \frac{(h(\mathbf{z}_k^t\mathbf{g}_\infty)-1)}{\pi_k}\mathbf{x}_k\mathbf{x}_k^t$, since $\Pi_1 = \mathbf{0}$. Then, with arguments similar to those that led to (20) results in asymptotic variance of the response model parameters given by

$$Avar\left(\sqrt{n}\,(\hat{\mathbf{g}}_s - \mathbf{g}_\infty)\right) = \left[\mathbf{D}^t\Pi_2^{-1}\mathbf{D}\right]^{-1}. \tag{22}$$

The additional asymptotic variance introduced by calibrating against population level instead of sample level information is expressed by the difference

$$\left[\mathbf{D}^t\left(\Pi_1 + \Pi_2\right)^{-1}\mathbf{D}\right]^{-1} - \left[\mathbf{D}^t\Pi_2^{-1}\mathbf{D}\right]^{-1} > \mathbf{0} \tag{23}$$

The positive definiteness of the difference (23) is illustrated by the positive definiteness of the difference $\left[\mathbf{D}^t\Pi_2^{-1}\mathbf{D}\right] - \left[\mathbf{D}^t\left(\Pi_1 + \Pi_2\right)^{-1}\mathbf{D}\right] > \mathbf{0}$ (see [6]). This is equivalent to demonstrating that

$$\Pi_2^{-1} - \left(\Pi_1 + \Pi_2\right)^{-1} > \mathbf{0} \tag{24}$$

because Π_1 and Π_2 are both positive definite matrices, unless $h(\mathbf{z}_k^t\mathbf{g}_\infty) = 1$ for all elements in the population, and $\mathbf{D}$ is a full rank matrix as a consequence of (11). Observe that proving (24) is in turn equivalent to demonstrating that $\left(\Pi_1 + \Pi_2\right) - \Pi_2 > \mathbf{0}$. Thus, inequality (23) follows.

4. Simulation study

Under assessment are the estimators described in points "a" to "d": the direct nonresponse weighting adjustment (a), the propensity score calibration estimator (b), the linear calibration estimator (c), and the two-step calibration estimator (d). We used data from a real case study with 4228 sampled elements, of which 1783 were nonrespondents. A two-covariate logistic regression was fitted based on this data and used as the true response probability model in the simulations. Next, we created a synthetic population based on the 2445 respondents to the survey; samples were drawn from this population, after which a response set was generated using the estimated response probability model.

Five variables were selected for the study, one categorical and the others numerical. The numerical variables were transformed into logarithmic scales to reduce variability. The categorical variable, denoted γ, was a stratum indicator in the original study having six strata, thus, $\gamma_k = (\gamma_{1k}, \gamma_{2k}, \gamma_{3k}, \gamma_{4k}, \gamma_{5k}, \gamma_{6k})$, where $\gamma_{ik} = 1_{S_i}(k)$ and S_i is the i^{th} stratum. Figure 4 presents the relationship among the original quantitative variables transformed into logarithmic form. One of them, left untransformed, was chosen to be study variable y, and estimation concerns estimating the population total $Y = 17014$, having the three auxiliary variables v_1, v_2, and v_3.

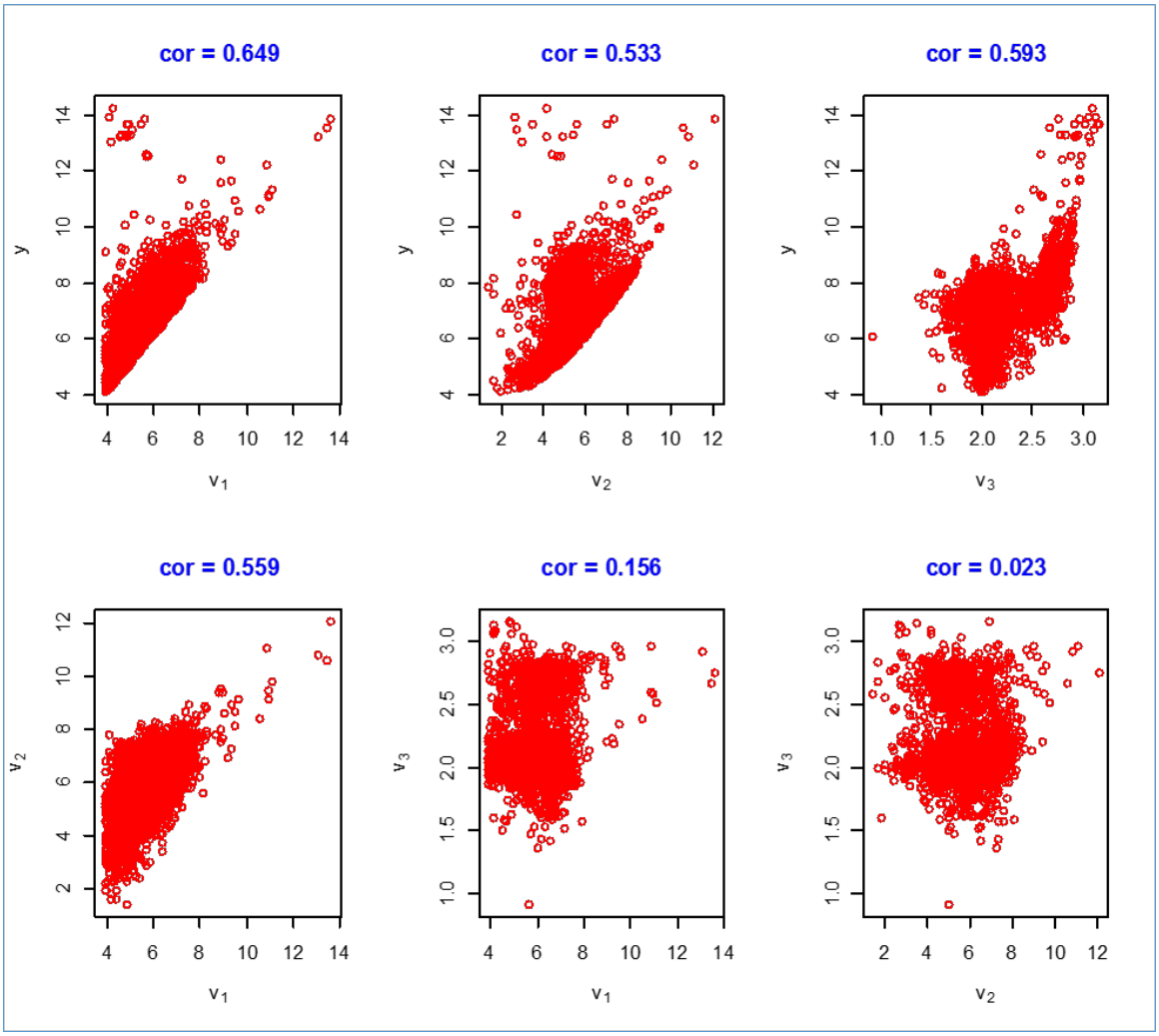

Figure 1: Pairwise correlations among the original variables used in the simulation study. Correlations calculated on the set of synthetic population.

Two additional quantitative auxiliary variables, *va* and *vb*, were created based on the equations $va = \sqrt{(v_2^2 + v_1^2)/v_3^3}$ and $vb = \sqrt[5]{v_1^6}/v_2$. The variables were created in an attempt to control for the strength of the relationship between the auxiliary variables, the study variable, and the model variables

in the response probability function. These new variables give correlation relationships not covered by the original auxiliary variables. Figure 4 shows plots of the new variables.

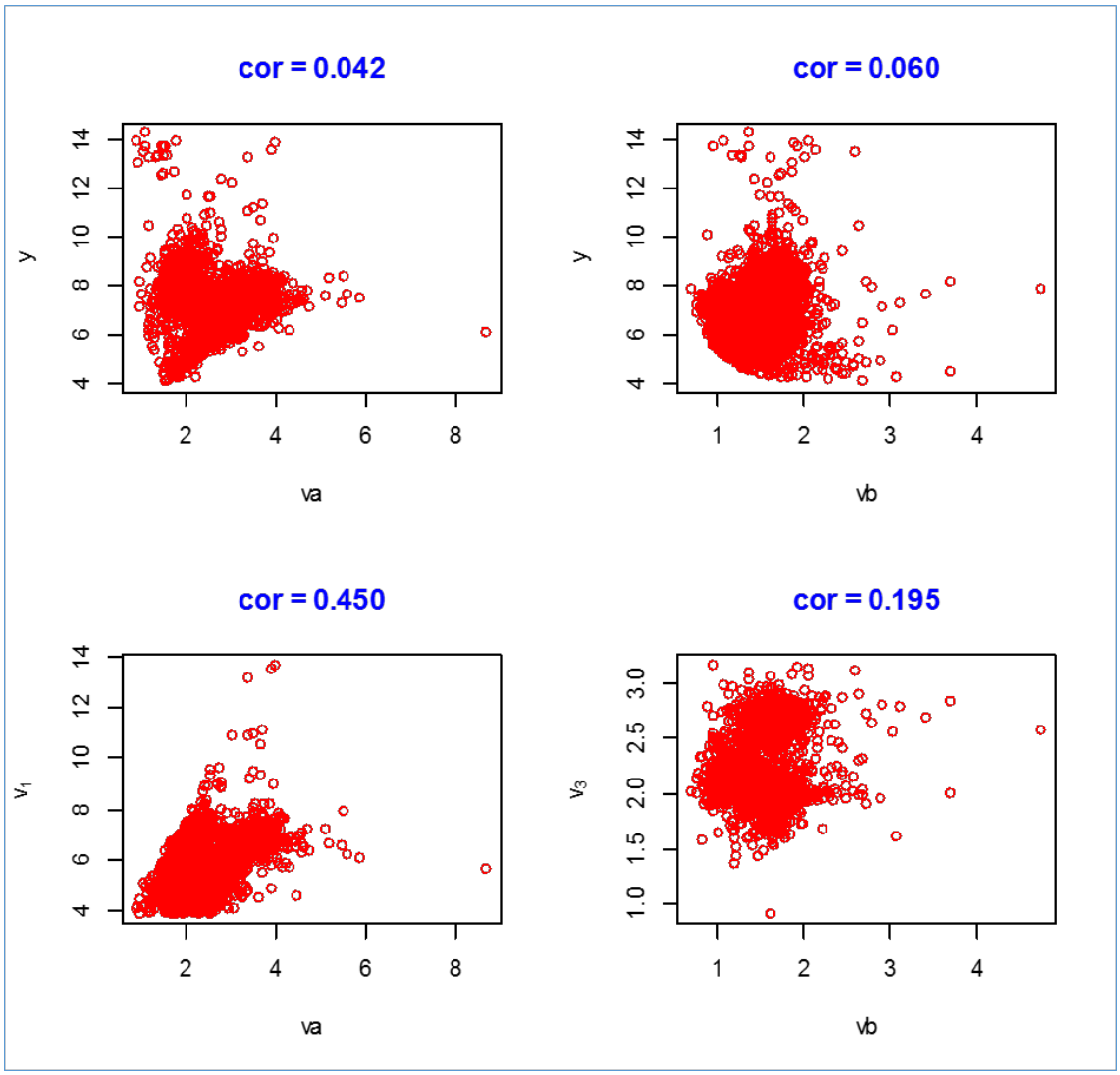

Figure 2: Pairwise correlations among artificial and original variables. Correlations calculated on the set of synthetic population.

Three simulation sets were defined using three criteria. The first criterion addresses the estimator's performance in relation to the quality of auxiliary variables, the second criterion addresses the effect of the sample size, and the last focuses on the effects of model misspecification. The response probability function is defined by the logistic regression model $p(R_k = 1|k \in s) = 1/\left(1+\exp(-\mathbf{z}_k^t\mathbf{g})\right)$, where $\mathbf{z}_k = (1, v_{1k})^t$ and the parameter values are defined by their ML fit to the original 4228 observations. The samples were selected using simple random sampling without replacement followed by Poisson sampling, in which the probability used for each Bernoulli trial was the one obtained using the response model. Each simulation result was based on 1000 replications. Initial trials with higher numbers of replications produced similar results. All estimators under study used the same samples and same response sets. The expected response rate was approximately 57%. The estimators are evaluated in terms of the relative bias (RB), standard error (SE) and mean squared error (MSE).

4.1. Simulation results

4.1.1 Correctly specified response probability model

Tables 1 – 3 present the results with the model vector defined as $\mathbf{z}_k = (1, v_{1k})^t$. In Table 1, the auxiliary vector is defined as $\mathbf{x}_k = (\gamma_{1k}, ..., \gamma_{6k}, \gamma_{1k}v_{2k}, ..., \gamma_{6k}v_{2k})^t$, a setup treated as the base case. As

seen in Figure 4, the auxiliary variable, v_2, correlates well with both the model variable and the study variable. A similar auxiliary vector was defined for the results in Table 2, with the exception that v_3 replaces v_2, that is, $\mathbf{x}_k = (\gamma_{1k}, ..., \gamma_{6k}, \gamma_{1k}v_{3k}, ..., \gamma_{6k}v_{3k})^t$. Here the auxiliary variable has a moderate level of correlation with the study variable, but carries much less information on the variation of the model variable in the response probability function. The correlations of v_3 with v_1 is approximately 0.16, with y approximately 0.59.

Again a similar auxiliary vector as in Table 1 was used for the results in Table 3, but here *va* is used in place of v_2, i.e. $\mathbf{x}_k = (\gamma_{1k}, ..., \gamma_{6k}, \gamma_{1k}va_k, ..., \gamma_{6k}va_k)^t$. The auxiliary variable *va* has approximately moderate correlation with the model variable (0.45) but low correlation with the study variable (0.04). The purpose of the simulation setup in tables 1 – 3 is partly to enable the study of the differences in the effect of having a good auxiliary variable for the model variable and the study variable respectively.

Table 1: Simulated estimates of RB, SE and MSE for the total of the survey variable in case of a true response model with $\mathbf{z}_k = (1, v_{1k})^t$, $\mathbf{x}_k = (\gamma_k, v_{2k}\gamma_k)^t$ and n=300

Estimator	Coefficients $(\hat{g}_0, \hat{g}_1)$	Coefficients variance $var(\hat{g}_0, \hat{g}_1)$	MSE $(\hat{Y})$	S.error $(\hat{Y})$	Rel.bias $(\hat{Y})$
$\hat{Y}_{DNTrue}$	(1.246,-0.157)	–	760,293	872	-0.006
$\hat{Y}_{DN\_ML}$	(1.258,-0.158)	(0.462,0.013)	45,771	214	-0.014
$\hat{Y}_{PS\_pop}$	(1.150,-0.137)	(2.640,0.077)	51,201	226	-0.042
$\hat{Y}_{PS\_samp}$	(1.168,-0.142)	(1.220,0.035)	50,708	225	-0.053
$\hat{Y}_{2stepML}$	–	–	24,495	155	-0.137
$\hat{Y}_{2stepA}$	–	–	24,727	155	-0.166
$\hat{Y}_{2stepB}$	–	–	39,566	196	-0.196
$\hat{Y}_{LC}$	–	–	191,835	438	-0.113

Table 2: Simulated estimates of RB, SE and MSE for the total of the survey variable in case of a true response model with $\mathbf{z}_k = (1, v_{1k})^t$, $\mathbf{x}_k = (\gamma_k, v_{3k}\gamma_k)^t$ and n=300

Estimator	Coefficients $(\hat{g}_0, \hat{g}_1)$	Coefficients variance $var(\hat{g}_0, \hat{g}_1)$	MSE $(\hat{Y})$	S.error $(\hat{Y})$	Rel.bias $(\hat{Y})$
$\hat{Y}_{DNTrue}$	(1.246,-0.157)	–	760,293	872	-0.006
$\hat{Y}_{DN\_ML}$	(1.258,-0.158)	(0.462,0.013)	45,771	214	-0.014
$\hat{Y}_{PS\_pop}$	(1.048,-0.106)	(8.439,0.247)	185,033	430	0.090
$\hat{Y}_{PS\_samp}$	(0.952,-0.099)	(3.665,0.106)	112,337	334	-0.160
$\hat{Y}_{2stepML}$	–	–	33,240	179	-0.192
$\hat{Y}_{2stepA}$	–	–	36,517	177	-0.429
$\hat{Y}_{2stepB}$	–	–	45,422	205	-0.342
$\hat{Y}_{LC}$	–	–	14.09×10^8	11872	-0.599

In tables 1 – 3, one can observe that the use of true probabilities ($\hat{Y}_{DNTrue}$) leads to estimated targets with larger variability than that of the estimated targets obtained using ML-estimated probabilities ($\hat{Y}_{DN\_ML}$). Observe that the standard error when using true probabilities is 872, which is four times more than the standard error when using estimated probabilities. Note that the results for these two estimators are the same over all three tables, because they are not defined by the benchmark variables used.

As predicted by the results in Section 3, the variance for the calibration estimator of the model coefficients is smaller when sample level information is used rather than population level information. This is observed in all three tables. Also, as expected, the ML estimator is associated with the smallest variance estimates, except the two-step estimators. The results also indicate that the variance decreases with increased correlation between the model and auxiliary variables; the variance estimates are the highest in Table 2. However, the comparison of tables 1 and 3 indicates that the correlation is not the only determinant of variance.

Table 3: Simulated estimates of RB, SE and MSE for the total of the survey variable in case of a true response model with $\mathbf{z}_k = (1, v_{1k})^t$, $\mathbf{x}_k = (\gamma_k, va_k\gamma_k)^t$ and n=300

Estimator	Coefficients $(\hat{g}_0, \hat{g}_1)$	Coefficients variance $var(\hat{g}_0, \hat{g}_1)$	MSE $(\hat{Y})$	S.error $(\hat{Y})$	Rel.bias $(\hat{Y})$
$\hat{Y}_{DNTrue}$	(1.246,-0.157)	–	760,293	872	-0.006
$\hat{Y}_{DN\_ML}$	(1.258,-0.158)	(0.462,0.013)	45,771	214	-0.014
$\hat{Y}_{PS\_pop}$	(1.392,-0.179)	(2.206,0.063)	78,540	277	0.231
$\hat{Y}_{PS\_samp}$	(1.318,-0.167)	(1.078,0.031)	69,218	262	0.125
$\hat{Y}_{2stepML}$	–	–	30,087	173	- 0.083
$\hat{Y}_{2stepA}$	–	–	30,139	173	-0.065
$\hat{Y}_{2stepB}$	–	–	43,848	209	-0.029
$\hat{Y}_{LC}$	–	–	2.4×10^7	4870	-1.343

A comparison between population and sample based propensity calibrations for population totals, that is, $\hat{Y}_{PS\_pop}$ and $\hat{Y}_{PS\_samp}$, indicates that under the definition of benchmark and model auxiliary information given in Table 1, these estimators perform rather similarly. The SE and RB are 226 and -0.042%, respectively, for the population-based calibration and 225 and -0.053%, respectively, for the sample-based calibration. In Table 2, however, the population-calibrated estimator displays larger variability. The SE and RB are 429 and 0.09%, respectively, for the population-calibrated estimator and 334 and -0.16%, respectively, for the sample-calibrated estimator. The same observation is made in Table 3, although the difference is smaller, i.e., 277 and 0.231% versus 262 and 0.125%.

The direct estimator based on model coefficients estimated by ML ($\hat{Y}_{DN\_ML}$) provides better results than do the single-step calibration estimators based on sample or population auxiliary information. In tables 1 – 3, the ML based estimator exhibits an SE of 214 and an RB of -0.014%.

The proposed two-step estimators provide much smaller SE and MSE estimates than do the single-step estimators. In some cases, the RB estimates are slightly larger. Estimators $\hat{Y}_{2stepML}$ (two-step with ML-estimated coefficients) and $\hat{Y}_{2stepA}$ (two-step with sample calibration-estimated coefficients), produce very similar results, with slightly smaller MSE and SE estimates for the estimator using ML-estimated model coefficients.

Finally, it is interesting to compare the effects of using different benchmark variables on the properties of the calibration-based estimators. Overall, the smallest MSE and SE estimates of the population total estimators are observed in Table 1, where the benchmark variable correlates moderately with both the study and the model variable. Table 2 contains the largest MSE and SE estimates observed among the three tables. The difference in the results of the $\hat{Y}_{2stepML}$ estimator between tables 2 and 3 is interesting. The estimator uses the same coefficient estimates but different benchmark variables, resulting in smaller MSE in Table 3.

Results are also provided for the linear calibration estimator. In all three tables, this estimator is the most penalized under the presented choices of auxiliary information. The auxiliary variables definition in Table 1 provides better results than do the definitions in Tables 2 and 3, the definition in Table 2 proving to be the worst of the three.

The results presented in tables 4 – 6 concern simulations based on the same setup as presented in Table 1, i.e., $\mathbf{z}_k = (1, v_{1k})^t$ and $\mathbf{x}_k = (\gamma_{1k}, ..., \gamma_{6k}, \gamma_{1k}v_{2k}, ..., \gamma_{6k}v_{2k})^t$, except that the sample sizes differ. In Table 4, the ordinary sample size of 300 was reduced by approximately 40%, while in tables 5 and 6 the sample was increased by approximately 100% and 400% respectively.

The results presented in the tables 4 – 6 indicate an increase and a decrease in the standard errors of the estimated targets in line with a decrease and an increase in the sample size. The standard picture in Table 1 prevails under all three sample sizes, i.e. sample calibration leads to smaller variance in model coefficient estimates than does the population calibration, while ML yields the overall smallest variance estimates. The sample-calibrated estimator, $\hat{Y}_{PS\_samp}$, yields smaller SE and MSE estimates than does the population-calibrated estimator, $\hat{Y}_{PS\_pop}$. In turn, $\hat{Y}_{DN\_ML}$ yields the smallest SE and MSE estimates of these three estimators. In addition, the two-step estimators $\hat{Y}_{2stepML}$ and $\hat{Y}_{2stepA}$ produce smaller SE and MSE estimates than do the other estimators. Interestingly, the

Table 4: Simulated estimates of RB, SE and MSE for the total of the survey variable in case of a true response model with $\mathbf{z}_k = (1, v_{1k})^t$, $\mathbf{x}_k = (\gamma_k, v_{2k}\gamma_k)^t$ and n=185

Estimator	Coefficients $(\hat{g}_0, \hat{g}_1)$	Coefficients variance $var(\hat{g}_0, \hat{g}_1)$	MSE $(\hat{Y})$	S.error $(\hat{Y})$	Rel.bias $(\hat{Y})$
$\hat{Y}_{DNTrue}$	(1.246,-0.157)	–	760,293	872	-0.006
$\hat{Y}_{DN\_ML}$	(1.271,-0.160)	(0.855,0.024)	75,694	275	-0.110
$\hat{Y}_{PS\_pop}$	(1.142,-0.132)	(4.914,0.143)	105,707	325	-0.059
$\hat{Y}_{PS\_samp}$	(1.157,-0.138)	(2.302,0.067)	91,838	303	-0.110
$\hat{Y}_{2stepML}$	–	–	45,545	207	-0.309
$\hat{Y}_{2stepA}$	–	–	46,257	207	-0.344
$\hat{Y}_{LC}$	–	–	760,742	865	-0.635

Table 5: Simulated estimates of RB, SE and MSE for the total of the survey variable in case of a true response model with $\mathbf{z}_k = (1, v_{1k})^t$, $\mathbf{x}_k = (\gamma_k, v_{2k}\gamma_k)^t$ and n=600

Estimator	Coefficients $(\hat{g}_0, \hat{g}_1)$	Coefficients variance $var(\hat{g}_0, \hat{g}_1)$	MSE $(\hat{Y})$	S.error $(\hat{Y})$	Rel.bias $(\hat{Y})$
$\hat{Y}_{DNTrue}$	(1.246,-0.157)	–	760,293	872	-0.006
$\hat{Y}_{DN\_ML}$	(1.268,-0.161)	(0.240,0.007)	22,552	150	-0.060
$\hat{Y}_{PS\_pop}$	(1.214,-0.150)	(1.188,0.034)	28,028	167	-0.060
$\hat{Y}_{PS\_samp}$	(1.206,-0.150)	(0.581,0.017)	23,659	153	-0.099
$\hat{Y}_{2stepML}$	–	–	10,889	102	-0.136
$\hat{Y}_{2stepA}$	–	–	11,043	102	-0.157
$\hat{Y}_{LC}$	–	–	21,480	146	-0.065

Table 6: Simulated estimates of RB, SE and MSE for the total of the survey variable in case of a true response model with $\mathbf{z}_k = (1, v_{1k})^t$, $\mathbf{x}_k = (\gamma_k, v_{2k}\gamma_k)^t$ and n=1200

Estimator	Coefficients $(\hat{g}_0, \hat{g}_1)$	Coefficients variance $var(\hat{g}_0, \hat{g}_1)$	MSE $(\hat{Y})$	S.error $(\hat{Y})$	Rel.bias $(\hat{Y})$
$\hat{Y}_{DNTrue}$	(1.246,-0.157)	–	760,293	872	-0.006
$\hat{Y}_{DN\_ML}$	(1.229,-0.154)	(0.125,0.004)	8246	91	-0.016
$\hat{Y}_{PS\_pop}$	(1.197,-0.148)	(0.417,0.012)	8939	94	-0.033
$\hat{Y}_{PS\_samp}$	(1.212,-0.151)	(0.273,0.008)	8638	93	-0.022
$\hat{Y}_{2stepML}$	–	–	4288	65	-0.042
$\hat{Y}_{2stepA}$	–	–	4270	65	-0.035
$\hat{Y}_{LC}$	–	–	6750	82	-0.010

linear calibration estimator displays improved properties with an increased sample size. This indicates that the estimator is competitive with the direct ML or calibration weighting.

4.1.2 Misspecified response probability model

The results in tables 7 and 8 are based on simulations with the erroneous model vector, $\mathbf{z}_k = (1, v_{3k})^t$, and the auxiliary vectors, $\mathbf{x}_k = (\gamma_{1k}, ..., \gamma_{6k}, \gamma_{1k}v_{2k}, ..., \gamma_{6k}v_{2k})^t$ and $\mathbf{x}_k = (\gamma_{1k}, ..., \gamma_{6k}, \gamma_{1k}vb_k, ..., \gamma_{6k}vb_k)^t$, respectively. The true model variable, v_1, and the specified model variable, v_3, have a correlation of approximately 0.16. Table 7 shows the results when the model variable is misspecified while the auxiliary variable is moderately correlated with the study variable and weakly correlated with the model variable. Table 8 presents the results when the auxiliary variable does not correlate well with either the study or the model variables (see 4).

The results presented in tables 7 and 8 indicate an increase in bias, compared with results in Table 1. In terms of SE, the levels are roughly the same in tables 7 and 8 as in Table 1, with the exceptions of the two-step estimators in Table 8. The MSEs for these estimators are larger in tables

Table 7: Simulated estimates of RB, SE and MSE for the total of the survey variable in case of a true response function and erroneous model variable, that is $\mathbf{z}_k = (1, v_{3k})^t$, $\mathbf{x}_k = (\gamma_k, v_{2k}\gamma_k)^t$ and n=300

Estimator	Coefficients $(\hat{g}_0, \hat{g}_1)$	Coefficients variance $var(\hat{g}_0, \hat{g}_1)$	MSE $(\hat{Y})$	S.error $(\hat{Y})$	Rel.bias $(\hat{Y})$
$\hat{Y}_{DNTrue}$	(1.246,-0.157)	–	760,293	872	-0.006
$\hat{Y}_{DN\_ML}$	(0.521,-0.085)	(0.692,0.141)	57,791	211	-0.678
$\hat{Y}_{PS\_pop}$	(0.437,-0.043)	(2.359,0.496)	55,056	204	-0.680
$\hat{Y}_{PS\_samp}$	(0.469,-0.060)	(0.958,0.196)	61,322	218	-0.692
$\hat{Y}_{2stepML}$	–	–	28,133	156	-0.367
$\hat{Y}_{2stepA}$	–	–	28,255	155	-0.376
$\hat{Y}_{LC}$	–	–	66×10^8	81,350	-3.222

Table 8: Simulated estimates of RB, SE and MSE for the total of the survey variable in case of an erroneous model and weak auxiliary variables, i.e., $\mathbf{z}_k = (1, v_{3k})^t$, $\mathbf{x}_k = (\gamma_k, vb_k\gamma_k)^t$ and n=300

Estimator	Coefficients $(\hat{g}_0, \hat{g}_1)$	Coefficients variance $var(\hat{g}_0, \hat{g}_1)$	MSE $(\hat{Y})$	S.error $(\hat{Y})$	Rel.bias $(\hat{Y})$
$\hat{Y}_{DNTrue}$	(1.246,-0.157)	–	760,293	872	-0.006
$\hat{Y}_{DN\_ML}$	(0.521,-0.085)	(0.692,0.141)	57,791	211	-0.678
$\hat{Y}_{PS\_pop}$	(0.599,-0.115)	(1.914,0.397)	57,791	206	-0.648
$\hat{Y}_{PS\_samp}$	(0.609,-0.124)	(0.934,0.190)	57,223	216	-0,602
$\hat{Y}_{2stepML}$	–	–	47,261	197	-0.537
$\hat{Y}_{2stepA}$	–	–	47,041	197	-0.528
$\hat{Y}_{LC}$	–	–	3.01×10^8	17,348	-1.169

7 and 8 than in Table 1. For the direct calibration estimators, the relationship between sample and population level information is reversed, compared with that presented in Table 1. The population level calibrated estimator yields smaller SE and MSE estimates in tables 7 and 8. Still, the two-step calibration estimators provide the smallest SE and MSE estimates. These are also associated with the smallest bias estimates.

In Table 9, estimation is carried out as in Table 1, but the true response probability model is the exponential model $p(R_k = 1|k \in s) = \left[1 - \exp(-\mathbf{z}_k^t \mathbf{g})\right]$. The coefficient vector was defined to be $\mathbf{g}^t = (0.185,\ 0.08)$. The coefficients were chosen so that the response probabilities are within the same range as in Table 1.

Table 9: Simulated estimates of RB, SE and MSE for the total of the survey variable in case of a misspecified response model with $\mathbf{z}_k = (1, v_{1k})^t$, $\mathbf{x}_k = (\gamma_k, v_{2k}\gamma_k)^t$ and n=300.

Estimator	Coefficients $(\hat{g}_0, \hat{g}_1)$	Coefficients variance $var(\hat{g}_0, \hat{g}_1)$	MSE $(\hat{Y})$	S.error $(\hat{Y})$	Rel.bias $(\hat{Y})$
$\hat{Y}_{DN\_ML}$	(-1.061,0.166)	(0.472,0.014)	52,640	229	-0.014
$\hat{Y}_{PS\_pop}$	(-1.125,0.180)	(2.374,0.072)	56,865	238	0.059
$\hat{Y}_{PS\_samp}$	(-1.105,0.175)	(1.278,0.039)	58,189	241	0.027
$\hat{Y}_{2stepML}$	–	–	29,164	170	-0.116
$\hat{Y}_{2stepA}$	–	–	29,343	170	-0.131

As a result of this setup, model coefficient estimators are inconsistent, as illustrated by the results in Table 9. For the estimators of the population total, the results in Table 9 are not very different from the ones presented in Table 1. The SE and MSE estimates for the ML-based direct weighting estimator are larger, but still the smallest estimated SE and MSE for the direct weighting estimators. In addition, as observed in tables 7 and 8, calibration using population level information yields smaller

SE and MSE estimates than does calibration using sample level information. As previously observed in all tables, the two-step calibration estimators have the smallest SE and MSE estimates.

5. Discussion

Simulation results are consistent with the principle that estimated probabilities outperform true probabilities in weighting for nonresponse, as was earlier known for ML-estimated probabilities (see [13]). The results presented here suggest that the gain in using estimated probabilities also holds for alternative model parameter estimators. This somewhat surprising principle is here interpreted as due to the random response set size whereby the HT estimator is not location invariant. The results presented also suggest that the gain in using estimated probabilities holds for alternative model parameters. In fact, even under the considered misspecifications of the response probability model, the results indicate the improved performance of the weighting estimators using estimated response probabilities.

The major concern in the paper is the use of sample or population level auxiliary information in the calibration of the response probability function. The simulation results obtained are consistent with the formal asymptotic argument presented, suggesting the use of sample auxiliary information for estimating the response probability function. Results indicate that the response function parameters are estimated with lower variance when using sample auxiliary information instead of population level information. The importance of having auxiliary information highly correlated with the model variables is observed for both levels of auxiliary information.

Using sample or population level information in the calibration estimators of population totals produces similar relative biases and standard errors. However, the sample-based calibration estimator has a smaller MSE than does the population counterpart; this is observed in all cases when the model is correctly specified. However, ML-estimated probabilities yield an estimator with the smallest SE and MSE estimates.

The auxiliary vector used in Table 2 is moderately correlated with the study variable while having virtually no relationship with the model variable; the standard errors for the single step calibration estimators are greater than when the auxiliary variable is correlated with both the study and model variables (Table 1). A much smaller difference is observed when auxiliary variables are correlated with the model variable while having virtually no correlation with the study variable (Table 3). This suggests a preference for auxiliary variables related to response propensity model variables over auxiliary variables related to the study variable.

Response probability function modelling is susceptible to misspecification. Under the erroneous choice of model variables, the major effects observed here are on the bias of the propensity based-estimators. The estimators (i.e., $\hat{Y}_{DN\_ML}$,$\hat{Y}_{PS\_pop}$ and $\hat{Y}_{PS\_samp}$) are associated with larger biases, although still at a low relative level (tables 7 and 8). The major observation is that the population-based calibrated estimator is more effective in error protection than is either the sample-calibrated or ML-based estimator. Still, good auxiliary information is important for the model variables. Although the evidence presented suggests that using sample auxiliary information is superior to using population auxiliary information in propensity calibration estimators, the population level propensity calibration is suggested to be the best alternative for reducing the MSE of the target estimates when the model is misspecified.

An erroneous functional form of the response probability model does not have a great impact on the estimator performance, according to the results in Table 9. One likely reason for this is that the two models are similar. However, the results suggest that the choice of the functional form is less important than is having the right model variables. This is partly supported by the competitive performance of the linear calibration estimator at larger sample sizes.

We suggest that estimation be performed in two steps; in the first step, the sample auxiliary data are used in the propensity calibration for estimating the response probabilities; in the second step, the products of the design weights and the reciprocals of response probabilities replace the design weights in the linear calibration estimator. The two-step estimation is to be performed using sample auxiliary information for estimating the response model through calibration, followed by the use of population auxiliary information for estimating target entities. This will generally produce more

efficient estimates.

The results presented all favor the suggested two-step calibration estimators. In some cases, these estimators are associated with larger bias, thought their relative sizes are small. In terms of MSE, the two-step estimators outperform other estimators. This is also observed when the response probability model is misspecified. A general suggestion would be to use ML to estimate model parameters in the first step, if model variables are available at the sample level. If the model variables are available only at the response set level, the [5] calibration estimator for model parameters is almost equally good. The results of the two-step estimators are of particular interest since response probability functions used in practice are models susceptible to misspecification. The effects of misspecification are usually unknown and can yield an adjusted estimator with a larger bias than the unadjusted one, depending on correlation structures among the study variable, response probability and auxiliary variables ([17]). Although small, a second calibration step reduces bias estimates in cases with a wrong auxiliary variable in the response probability model (tables 7 and 8).A question for further studies is whether a second step calibration can protect against the misspecification of the response probability function and/or if indicators of misspecification can be developed.

With large sample sizes and carefully chosen auxiliary information, the linear calibration estimator is fairly competitive with the propensity-based estimators. The linear calibration estimator is known to have good properties when good auxiliary information is available. On the other hand, poorly defined auxiliary variables may lead to negative and/or very large weights in the linear calibration ([21], remark 6.1). These problems may result in very inefficient estimates.A conclusion based on the results presented in Table 1 is that the properties of the linear calibration estimator can be improved by using efficient initial weights. These weights can be derived from a sample-based propensity calibration estimator. The combined approaches produce more efficient estimates.

Tables 1 – 3 provide results for, $\hat{Y}_{2stepB}$, an estimator not discussed here. It is a version of $\hat{Y}_{2stepA}$ in which auxiliary information exists only at the sample level, i.e. sample level information is used in both steps. This will generally provide slightly better RB, but the SE and MSE are higher than those provided by $\hat{Y}_{2stepA}$.

5.1. Limitations

In this article, we use an estimation setting in which only positive correlations among the variables in the study are considered. [17] have noted that the direction of the correlation between the variables involved in the study has an influence on the properties of the estimated entities. This suggests a further investigation whether the results here are the same when the variables are negatively correlated.

References

[1] Barranco-Chamorro, I., Jiménez-Gamero, M. D, Moreno-Rebollo, J. L. and Muñoz-Pichardo, J. M. (2012). Case-deletion type diagnostics for calibration estimators in survey sampling. *Journal of Computational Statistics and Data Analysis*, **56,** 2219–2236.

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

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

[4] Brick, J. M. (2013). Unit Nonresponse and Weighting Adjustments: A Critical Review. *Journal of Official Statistics,* **29:3***,* 329–353

[5] Chang, T. and Kott, P. S. (2008). Using calibration weighting to adjust for nonresponse under a plausible model, *Biometrika*, **95:3**, 555–571.

[6] Davidson, R. and MacKinnon, J. G. (1993). *Estimation and Inference in Econometrics*. Oxford University Press.

[7] Davidson, R. and MacKinnon, J. G. (2003). *Econometric Theory and Methods*. Oxford University Press.

[8] Estevão, V. M. and Särndal, C.-E. (2002). The Ten Cases of Auxiliary Information for Calibration in Two-Phase Sampling. *Journal of Official Statistics*, 18:2, 233–255

[9] Fuller, W. A. (2009). *Sampling Statistics*. Wiley & Sons, New Jersey.

[10] Isaki, C. T. and Fuller, W. A. (1982) Survey Design Under the Regression Superpopulation Model. *Journal of the American Statistical Association*, 77, 89–96

[11] Kalton, G. and Flores-Cervantes, I. (2003). Weighting Methods. *Journal of Official Statistics*, 19:2, 81-97

[12] Kim, J. K. (2004). Efficient nonresponse weighting adjustment using estimated response probability. ASA Section on Survey Research Methods.

[13] Kim, J. K. and Kim, J. J. (2007). Nonresponse weighting adjustment using estimated response probabilities. *The Canadian Journal of Statistics*, 35:4, 501-514.

[14] Kim, J. K. and Park , M.(2010). Calibration estimation in surveys. *International Statistical Review*, 78, 21-39.

[15] Kott, P.S. (2006). Using calibration weighting to adjust for nonresponse and coverage errors. *Survey Methodology*, 32:2, 133-142.

[16] Kott, P. S. and Liao, D. (2015). One step or two? Calibration weighting from a complete list frame with nonresponse. *Survey Methodology*, 41:1, 165–181.

[17] Kreuter, F. and Olson, K. (2011). Multiple Auxiliary Variables in Nonresponse Adjustment. *Sociological Methods & Research*, 40:2, 311–332.

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

[19] Magnus, J. R. (2010). On the concept of matrix derivative. *Journal of Multivariate Analysis*, 101, 2200-2206.

[20] Rosenbaum, P. R. (1987). Model-based direct adjustment. *Journal of the American Statistical Association*, 82:398, 387-394.

[21] Särndal, C.-E. and Lundström, S. (2005). *Estimation in Surveys with Nonresponse*. Wiley, New York.

[22] Särndal, C.-E, Swensson, B. and J. Wretman (1992). *Model Assisted Survey Sampling*. Springer, New York.

[23] Schouten, B. (2007). A selection strategy for weighting variables under a Not-Missing-at-Random assumption. *Journal of Official Statistics*, 23, 51-68.

[24] West, B. T. (2009). A Simulation Study of Alternative Weighting Class Adjustments for Nonresponse when Estimating a Population Mean from Complex Sample Survey Data. Section on Survey Research Methods-JSM2009

KAI KURIŲ PERSVĖRIMO METODŲ, SKIRTŲ ATSIŽVELGTI Į NEATSAKYMUS, PALYGINIMAS

Bernardo João Rota, Thomas Laitila

Santrauka Straipsnyje parodoma, kad imties lygio ir populiacijos lygio papildoma informacija yra naudinga ir duoda apytiksliai vienodus rezultatus didelių imčių atveju. Literatūroje siūloma keletas funkcinių svorių formų. Šiame straipsnyje nagrinėjamos kalibruotojo įvertinio savybės, laikant, kad atsakymo į apklausą tikimybės funkcinė forma yra žinoma. Dėmesys nukreipiamas į skirtumus tarp populiacijos lygio ir imties lygio papildomos informacijos, parodant, kad pastaroji yra tinkamesnė atsakymo tikimybės modelio koeficientams vertinti. Siūloma dviejų žingsnių procedūra, kurioje naudojama imties informacija modelio koeficientams vertinti pirmame žingsnyje ir tyrimo kintamojo sumos kalibruotasis įvertinys antrajame žingsnyje.

Reikšminiai žodžiai: kalibravimas, papildomi kintamieji, atsakymo tikimybė, didžiausio tikėtinumo metodas.