To my parents Bernardo and Jacinta

Örebro Studies in Statistics 8

BERNARDO JOÃO ROTA

Calibration Adjustment for Nonresponse in Sample Surveys

Title: Calibration Adjustment for Nonresponse in Sample Surveys

Publisher: Örebro University 2016
www.oru.se/publikationer-avhandlingar

Print: Örebro University, Repro 09/2016

ISSN 1651-8608
ISBN 978-91-7529-160-4

Abstract

Bernardo João Rota (2016): Calibration Adjustment for Nonresponse in Sample Surveys. Örebro Studies in Statistics 8.

In this thesis, we discuss calibration estimation in the presence of nonresponse with a focus on the linear calibration estimator and the propensity calibration estimator, along with the use of different levels of auxiliary information, that is, sample and population levels. This is a four-papers-based thesis, two of which discuss estimation in two steps. The two-step-type estimator here suggested is an improved compromise of both the linear calibration and the propensity calibration estimators mentioned above. Assuming that the functional form of the response model is known, it is estimated in the first step using calibration approach. In the second step the linear calibration estimator is constructed replacing the design weights by products of these with the inverse of the estimated response probabilities in the first step. The first step of estimation uses sample level of auxiliary information and we demonstrate that this results in more efficient estimated response probabilities than using population-level as earlier suggested. The variance expression for the two-step estimator is derived and an estimator of this is suggested. Two other papers address the use of auxiliary variables in estimation. One of which introduces the use of principal components theory in the calibration for nonresponse adjustment and suggests a selection of components using a theory of canonical correlation. Principal components are used as a mean to accounting the problem of estimation in presence of large sets of candidate auxiliary variables. In addition to the use of auxiliary variables, the last paper also discusses the use of explicit models representing the true response behavior. Usually simple models such as logistic, probit, linear or log-linear are used for this purpose. However, given a possible complexity on the structure of the true response probability, it may raise a question whether these simple models are effective. We use an example of telephone-based survey data collection process and demonstrate that the logistic model is generally not appropriate.

Keywords: Auxiliary variables, Calibration, Nonresponse, principal com-ponents, regression estimator, response probability, survey sampling, two-step estimator, variance estimator, weighting.

Bernardo João Rota, School of Business
Örebro University, SE-701 82 Örebro, Sweden, bernardo.rota@oru.se

List of Papers

This thesis consists of four papers:

- Rota, B. J. and Laitila, T. (2015) Comparisons of some weighting methods for nonresponse adjustment. *Lithuanian Journal of Statistics*, **54:1,** 69–83.
- Rota, B. J. (2016). Variance Estimation in Two-Step Calibration for Nonresponse Adjustment. Manuscript
- Rota, B. J. and Laitila, T. (2016) Calibrating on Principal Components in the Presence of Multiple Auxiliary Variables for Nonresponse Adjustment. This paper is accepted in *South African Statistical Journal*
- Rota, B. J. and Laitila, T. (2016). On the Use of Auxiliary Variables and Models in Estimation in Surveys with Nonresponse. Manuscript

Acknowledgements

The path I have chosen does not end with achievement of a PhD degree; rather, it simply goes another way around to start another path. However, it is an amazing feeling to realize that you are capable of such achievement after a long journey on thorny ground. I would never be able to walk this thorny ground and succeed without support.

I thank Professor Thomas Laitila, my Supervisor since I was a master student and mentor of my success in this endeavor. Thank you for being patient in your guidance, particularly in those moments when I wrote "senseless stuff".

I cannot forget Professor Sune Karlsson; thank you for your support which has been extended since I was a master student.

My parents, aged as they are, were subjected to living years without seeing their son but had a strong belief in my success. I thank my brother Victor and my sister Bernardete and their respective families, my brothers Solano and Flaviano for everything. My nieces and nephews, you are always the reason for my happiness. Thank you for your tireless support.

Mónica Mucocana thanks for everything.

My gratitude extends to the Örebro School of Business administrative personnel, the list is extensive. Thank you all of you for being friendly and helpful every moment that I needed your support. My colleagues from department of Mathematics and Informatics at Eduardo Mondlane University. Thank you all. I also express my gratitude to Professor João Munembe, co-supervisor for the mozambican part, Professor Manuel Alves I still remember your support.

My friends and fellow PhD colleagues, particularly Jose Nhavoto, with whom I started this journey and who witnessed my struggle day after day and Göran Bergstrand and Pari Bergstrand the best friendship I have made in Sweden. To all of you thank you very much.

I would like to express my gratitude to the Swedish SIDA Foundation - International Science Program for the cooperation with Eduardo Mondlane University in Maputo and especially for funding and supporting my studies. For all who have been involved to tight this cooperation both from Swedish and Mozambican side, my deepest gratitude.

Contents

Part I: Introduction

Background

Sample surveys have long been used as an effective means of obtaining information about populations of interest. According to Särndal, et al. (1992) and Rao (2003), the use of sample surveys gained emphasis from 1930 as time- and cost-effective, although providing reliable information about the population characteristics of interest through statistical inferences.

The reliability of the survey-based information depends upon how the total survey error is approached. Total survey error is the joint survey error resulting from sampling and nonsampling errors, that is, survey errors due to the use of a sample instead of the whole population and survey errors that relate to how the data are collected and processed, respectively.

In the absence of nonsampling errors, basic statistical estimation methods such as the Horvitz-Thompson estimators can yield reliable statistics that can be used to make inferences. With nonsampling errors such as nonresponse, these basic methods are no longer effective in producing reliable information. These problems boosted the discovery of more sophisticated methods for production of survey-based statistics. Among these methods is the use of auxiliary information through weighting the observed values of the variables of interest. Here, nonsampling errors are restricted to nonresponse errors.

When refering to weghting methods, it comes along with the calibration weighting approach which is one of the fastest emerging weighting adjustment methods. Calibration started as a procedure for improving the accuracy of survey estimates in a full-response setting (see, Deville and Särndal, 1992 and Deville, et al., 1993). In later advances, calibration also became a tool for estimation in small domains or small areas (Chambers, 2005; Lehtonen and Veijanen, 2012, 2015) and estimation in surveys with incomplete data due to nonresponse (e.g. Lundström and Särndal, 1999). Observe that in complete-data surveys, bias is not a concern; simple methods can yield unbiased estimation. Thus, in this context, variance is a concern. Under nonresponse, accuracy is measured in terms of both bias and variance, with particular emphasis on the former.

Nonresponse, which is the failure to obtain data from a sampled unit, will generally bias estimates, whatever the estimation method. In weighting for nonresponse adjustment, the general setting is to view the response set as a random subsample of the selected sample. The observed values of the respondents are weighted, attempting to make the response set representative

of the original sample, in which case survey estimates can be used to make reasonable inferences.

Notation

Consider a finite population U consisting of N units labelled 1, ..., N. A sample s of size n is drawn from U with a given probability sampling design $p(s)$ yielding first- and second-order inclusion probabilities $\pi_k > 0$ and $\pi_{kl} > 0$, respectively, where $\pi_{kk} = \pi_k$ for all $k \epsilon U$. The survey variable of interest is y, and we are interested in estimating its total $Y = \sum_U y_k$, where $\sum_A = \sum_{k \epsilon A}$. Data are assumed to be observed for a subset $r \subset s$; each y_k, $k \epsilon r$ is observed with probability $Pr(R_k = 1 | I_k = 1) > 0$ where $R_k = 1$ if $k \epsilon r$ and $R_k = 0$ otherwise and I_k is defined analogously whether $k \epsilon s$ or not. Here, we assume that R_k and R_l are independent for all $k \neq l$. Let $\mathbf{x}_k$ be an L-dimensional column vector of auxiliary variables known for all $k \epsilon U$ and $\mathbf{z}_k$ is a J-dimensional vector of model variables known for all k in r. Assume that $Pr(R_k = 1 | I_k = 1) = q(\mathbf{z}_k^t \mathbf{g})$ evaluated at $\mathbf{g} = \mathbf{g}_\circ$, which is an interior point of parameter space $\mathbf{G}$.

Calibration estimators

Calibration estimators use weights w_k that satisfy the calibration constraint $\sum_r w_k \mathbf{x}_k = \mathbf{X}$, where $\mathbf{X} = \sum_U \mathbf{x}_k$ or $\mathbf{X} = \sum_s d_k \mathbf{x}_k$, that is, a population or an estimated population total of $\mathbf{x}_k$, respectively. The weights w_k in the linear calibration estimators minimize a Chi-Square distance function (see, Kim and Park, 2010). The resulting estimators have the following form:

$$\hat{Y}_{LC} = \left(\sum_U \mathbf{x}_k - \sum_r d_k \mathbf{x}_k \right)^t \left(\sum_r d_k \mathbf{x}_k \mathbf{x}_k^t \right)^{-1} \sum_r d_k \mathbf{x}_k y_k + \sum_r d_k y_k \quad (1)$$

where $d_k = \pi_k^{-1}$.

The propensity calibration (Chang and Kott, 2008) is an estimator of the following form:

$$\hat{Y}_{PSC} = \sum_r d_k q^{-1}(\mathbf{z}_k^t \hat{\mathbf{g}}) y_k \quad (2)$$

where $\hat{\mathbf{g}}$ is a solution to the calibration constraint $\sum_r d_k q^{-1}(\mathbf{z}_k^t \mathbf{g}) \mathbf{x}_k = \sum_U \mathbf{x}_k$.

Contribution

In this thesis, we discuss calibration estimation in the presence of nonresponse with a focus on the linear calibration estimator (Särndal and Lundström, 2005) and the propensity calibration estimator (Chang and Kott, 2008), along with the use of different levels of auxiliary information, that is, sample and population levels. This is a four-papers-based thesis, two of which discuss estimation in two steps. The two-step-type estimator here suggested is an improved compromise of both the linear calibration and the propensity calibration estimators mentioned above. Assuming that the functional form of the response model is known, it is estimated in the first step following the principle suggested by Chang and Kott (2008). In the second step the linear calibration estimator is constructed replacing the design weights by products of these with the inverse of the estimated response probabilities in the first step. The first step of estimation uses sample level of auxiliary information and we demonstrate that this results in more efficient estimated response probabilities than using population-level as suggested by Chang and Kott (2008). The resulting two-step estimator is given by

$$\hat{Y}_{2step} = \left(\sum_{U} \mathbf{x}_k - \sum_{r} g_k \mathbf{x}_k\right)^t \left(\sum_{r} g_k \mathbf{x}_k \mathbf{x}_k^t\right)^{-1} \sum_{r} g_k \mathbf{x}_k y_k + \sum_{r} g_k y_k \quad (3)$$

where $g_k = d_k q^{-1}(\mathbf{z}_k^t \hat{\boldsymbol{\gamma}})$.

The variance expression for (3) is derived and an estimator of this is suggested. Two other papers address the use of auxiliary variables in estimation. One of which introduces the use of principal components theory in the calibration for nonresponse adjustment. Principal components are used as a mean to accounting the problem of estimation in presence of large sets of candidate auxiliary variables. In addition to the use of auxiliary variables, the last paper also discusses the use of explicit models representing the true response behavior. Usually simple models such as logistic, probit, linear or log-linear are used for this purpose. However, given a possible complexity on the structure of the true response probability (see Kaminska, 2013), it may raise a question whether these simple models are effective. We use an example of telephone-based survey data collection process and demonstrate that the logistic model can be effective under very restrictive assumptions.

Part II: Summary of the papers

Paper I: Comparisons of Some Weighting Methods for Nonresponse Adjustment

This paper proposes combining the linear calibration estimator (1) and the propensity calibration estimator (2) in two steps of estimation. That is, we suggest improving the linear calibration estimator by a preliminary adjusting of design weights through multiplication of these with reciprocals of calibration-estimated response propensities. The resulting two-step-based calibration estimator given in (3) is compared with some single step nonresponse adjusted estimators and a two-step estimator with maximum likelihood-estima-ted response probabilities in the first step.

Asymptotic variance expressions for the model parameter estimator are derived for both the sample and population levels of auxiliary information. These expressions illustrate that the model parameter estimates have smaller variance when sample level auxiliary information is used rather than population level. This paper also addresses issues related to the choice of auxiliary variables by assessing the effect of different correlation relationships between auxiliary, model and study variables.

Numeric illustrations were based on real survey data. Three simulation sets were defined using three criteria. The first criterion addressed the es-timator's performance in relation to the quality of auxiliary variables, the second criterion addressed the effect of the sample size, and the last focuses on the effects of model misspecification. We did not find any strongly corre-lated pair of variables, the maximum correlation between pairs of the chosen variables was 0.649. Nevertheless, we believe that the results obtained are illustrative of the simulation objectives.

Among the results obtained are that two-step estimators are more efficient than any single step estimator, with maximum likelihood-based two step being fairly competitive with the calibration-based two-step estimator. Still good auxiliary variables are necessary especially for the linear calibration, an estimator that tend to be more penalized with the choice of poor auxiliary variables. The population level of auxiliary information provides more protection under model misspecification than does the sample level. The linear calibration estimator tend to be competitive with increasing sample size and use of good auxiliary variables.

Remark 1: The results displayed in some tables, in particular tables 2,3,7 and 8 for the linear calibration estimator are abnormal, which is a result of some extreme weights in the simulation. This result confirms one of the features of this estimator, which is documented on page 59, *remark 6.1* in Särndal and Lundström (2005). The estimator would have performed better if it were assigned a weight restriction.

Remark 2: This paper uses notations $\hat{Y}_{2stepA}$ and $\hat{Y}_{2stepB}$. These notations are only used to distinguish that the former uses sample level auxiliary information in the first step of estimation and population level in the second step, whereas the latter uses the sample level auxiliary information in both steps. Thus, these notations should not be confused with estimators defined by Särndal and Lundström (2005), who use the same notation.

Note: for further clarification of text in paper 1 see the appendix section below.

Paper II: Variance Estimation in Two-Step Calibration for Nonresponse Adjustment

Paper 1 combines linear calibration and propensity calibration estimators and constructs an alternative estimator of the total Y of a survey variable y by means of two-step estimation in the presence of sample- and population-level auxiliary information under the assumption of a known functional form of the response mechanism.

In this paper, a variance expression for the two-step estimator is derived and an estimator of this is suggested. The variance expression has an extra component that accounts for model parameter estimation in the first step. We show that the reduced variability due to the use of sample-level auxiliary information in the estimation of model parameters in the first step, which has been demonstrated in paper 1, implies reduced variance in the estimation of population characteristics.

The numerical illustration for the properties of the suggested estimator is based on two simulation setups, one of which is on real survey data whereas another is on simulated data. Simulation results suggest that the estimator performs well when good auxiliary variables are used. For large sample sizes and good auxiliary variables, the extra component in the variance expression has negligible contribution to the variance of population characteristics.

Remark: The variance and variance estimator developed in this paper is relative to the two-step calibration estimator suggested in paper 1. However,

it is not clearly stated which estimator is being referenced in paper 1. We use here the notation $\hat{Y}_{2step}$, which refers to $\hat{Y}_{2stepA}$ in the notation of paper 1.

Paper III: Calibrating on Principal Components in the Presence of Multiple Auxiliary Variables for Nonresponse Adjustment

When adjusting for nonresponse in sample surveys, auxiliary information has important role in successful estimation. This has been noted by Rizzo, Kalton and Brick (1996), who claim that the choice of auxiliary variables may be of greater significance than the choice of the weighting method.

This implies that the lack of auxiliary variables to assist in estimation is undesired. Conversely, large sets of auxiliary variables being available, can also bring problems such as strong correlation or multicollinearity among the variables which might result in an increased standard error of the estimated statistics. Another problem is the difficulty in selecting auxiliary variables related to a number of study variables simultaneously.

Thus, in accounting for these problems, we suggest reducing the dimensionality of the auxiliary data using principal components. The standard data variation is nearly maintained but in lower dimensional data.

We implement a rejection of principal components based on their canonical correlation with the model variables. The rejection based on canonical correlation is advantageous when samples are of small sizes whilst in large samples the results are similar to the obtained using the eigen-value-one stopping criterion of the principal components theory.

Simulation results confirmed that the use of principal components is effective both in the linear calibration and in the propensity calibration estimators.

Because the use of principal components auxiliary data is effective in estimation, the variance expression and the variance estimator derived in paper 2 can be adapted to use these dimension-reduced auxiliary data. However, this is left as a topic for future research.

This paper has been accepted for publication in South African Statistical Journal (Rota and Laitila, 2017)

Paper IV: On the Use of Auxiliary Variables and Models in Estimation in Surveys with Nonresponse

In weighting for nonresponse adjustment, the general framework is to characterize the response set as a random realization from a selected sample. This approach resembles estimation in two-phase sampling (e.g. Keen, 2005). In one version, the estimation is performed with explicit modelling of the response propensity whereas another version provides implicit modelling.

In papers 1 to 3, we use the linear calibration estimator, which is a case of implicit modelling of the response probability and the propensity calibration estimator illustrating explicit modelling.

Both weighting alternatives rely on the use of powerful auxiliary variables. A question rarely raised in the literature can be formulated as follows: how does weighting affect estimates if the response set mean is unbiased? One potential reason for this problem not being addressed is the adaptation of concepts on the relationship between the study variable and the generation of the response set from the model-based inference literature, e.g., MAR (miss-ing at random) and MCAR (missing completely at random).

Conditional on these auxiliary variables, the data are assumed to be missing at random. However, as with any other such missingness mechanism, this one cannot be tested statistically (Thoemmes and Rose, 2014). This problem leads to selection of auxiliary variables based on the correlation relationships they share with the variables of interest and the response behavior. We show here that such a guiding rule for selection of auxiliary variables can lead in a wrong direction, that is, we can increase rather than reduce the bias.

Furthermore, response mechanisms can be of complex structure, and applications tend to use simple models such as logit, probit or exponential in representing the true response mechanism (see e.g. Chang and Kott, 2008; Kim and Riddles, 2012; Haziza and Lesage, 2016). One might question whether it is appropriate to use such simple models. With an example of telephone-based survey data collection, we show that a logit model conditional on restrictive assumptions can be a valid choice. However, these models are not realistic in general, and better models reflecting the data collection process are needed. In addition to this, there is a need to develop tools to judge when the use auxiliary variables give valid estimates.

References

Chambers, R. L. (2005) *Calibrated Weighting for Small Area Estimation.* Southampton, UK, Southampton Statistical Sciences Research Institute, 26pp. (S3RI Methodology Working Papers, M05/04).

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

Deville, J. C. and Särndal, C.E. (1992). Calibration estimators in survey sampling, *Journal of the American Statistical Association*, **87**, 376–382.

Deville, J. C., Särndal, C. E. and Sautory, O. (1993). Generalized raking procedures in survey sampling, *Journal of the American Statistical Association*, **88:423**, 1013–1020.

Haziza, D. and Lesage, É. (2016) Journal of Official Statistics, **32:1**, 129–145.

Keen, K. J. (2005). Two-Phase Sampling. Wiley Online Library. DOI: 10.1002/0470011815.b2a05094 2005.

Kaminska, O. (2013). Discussion.*Journal of Official Statistics*, **29:3**, 355–358

Kim, J. K. and Park, M. (2010). Calibration Estimation in Survey Sampling, *International Statistical Review*, **78:1**, 21–39.

Lehtonen, R. and Veijanen, A. (2012). Small area poverty estimation by model calibration. *Journal of the Indian Society of Agricultural Statistics*, **66**, 125–133.

Lehtonen, R. and Veijanen, A. (2015). Estimation of poverty rate for small areas by model calibration and hybrid calibration methods. Retrieved from http://dx.doi.org/10.2901/EUROSTAT.C2015.001.

Lundström, S. and Särndal, C.-E. (1999). Calibration as a Standard Method for Treatment of Nonresponse. *Journal of Official Statistics.***15:2**,305–327.

Rao, J. N. K. (2003). *Small Area Estimation.* Wiley, New Jersey.

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

Särndal, C.-E. and Lundström, S. (2007). Assessing auxiliary vectors for control of nonresponse bias in the calibration estimator. *Journal of Official Statistics*, **24:2**, 167–191.

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

Thoemmes, F. and Rose, N. (2014). A Cautious Note on Auxiliary Variables That Can Increase Bias in Missing Data Problems, *Multivariate Behavioral Research*, **49**, 443–459.

Appendix
Clarification of text in paper 1

1. On page 71 and thereafter, d_k is used as design weight for element k, whereas $d(\cdot)$, e.g., in equation 7, is a function.

2. On page 71, the D used in equation 8 cannot be confounded with $\mathbf{D}$ (boldfaced) on page 74.

3. On page 72, *c. The linear calibration estimator.*

   The linear calibration estimator is defined with weights $w_k = d_k v_k$, where $v_k = 1 + \boldsymbol{\lambda}_r^t \mathbf{z}_k$ and $\boldsymbol{\lambda}_r^t = (\mathbf{X} - \sum_r d_k \mathbf{x}_k)^t \left(\sum_r d_k \mathbf{z}_k \mathbf{x}_k^t\right)^{-1}$. However, the simulations are held using the standard definition (Särndal and Lundström, 2005, p. 62), that is, the vector $\mathbf{z}_k$ is replaced by $\mathbf{x}_k$ leading to $v_k = 1 + \boldsymbol{\lambda}_r^t \mathbf{x}_k$ and $\boldsymbol{\lambda}_r^t = (\mathbf{X} - \sum_r d_k \mathbf{x}_k)^t \left(\sum_r d_k \mathbf{x}_k \mathbf{x}_k^t\right)^{-1}$.

4. On page 74 line 3 after equation 20, replace the word "rewrite" with "redefine".

5. On page 74 line 1 after equation 23, replace the words "is illustrated by" with "follows from".

6. On Tables 1–9, we use $\hat{Y}_{2stepA}$ and $\hat{Y}_{2stepB}$ without a clear distinction between them. The former uses sample-level auxiliary information in the first step and population-level in the second step whereas the latter uses sample level in both steps.

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

PAPER I

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

ISSN 2029-7262 online

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 $p_k = p(\mathbf{z}_k'\mathbf{g})$ 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 $w_k = d_k h(\mathbf{z}_k'\hat{\mathbf{g}})$, where $h(\cdot) = p^{-1}(\cdot)$ and $\hat{\mathbf{g}}$ is an estimator of $\mathbf{g}_\infty$. Assume $p(\mathbf{z}_k'\mathbf{g})$ to be differentiable w.r.t. $\mathbf{g}$ and define the weighted log likelihood function of the response distribution

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

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

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

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

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

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 $p(\mathbf{z}_k'\hat{\mathbf{g}}_{CAL})$ for the element k in r, i.e., $w_k = d_k h(\mathbf{z}_k'\hat{\mathbf{g}}_{CAL})$, so that the estimator (1) becomes

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

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

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

Assuming $h(\mathbf{z}_k'\mathbf{g})$ to be twice differentiable, [5] suggest an estimator defined by minimizing an objective function derived from (6), assuming the difference $\mathbf{e} = \mathbf{X} - \sum_r d_k h(\mathbf{z}_k'\mathbf{g}_\infty)\mathbf{x}_k$ 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

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

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:

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

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 $w_k = d_k v_k$, $v_k = 1 + \boldsymbol{\lambda}_r^t \mathbf{z}_k$, 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 $\boldsymbol{\lambda}_r^t = (\mathbf{X} - \sum_r d_k \mathbf{x}_k)^t \left(\sum_r d_k \mathbf{z}_k \mathbf{x}_k^t\right)^{-1}$. 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

$$E\left(d(\mathbf{g}_\infty) d^t(\mathbf{g}_\infty)\right) = \Pi_1 + \Pi_2 = \Pi \tag{15}$$

where, $E(d(\mathbf{g}_\infty)) = \mathbf{0}$, $\Pi_1 = \sum_{k \in U} \sum_{l \in U} \frac{\pi_{kl} - \pi_k \pi_l}{\pi_k \pi_l} \mathbf{x}_k \mathbf{x}_l^t$ and $\Pi_2 = \sum_U \frac{(h(\mathbf{z}_k^t \mathbf{g}_\infty) - 1)}{\pi_k} \mathbf{x}_k \mathbf{x}_k^t$, 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.

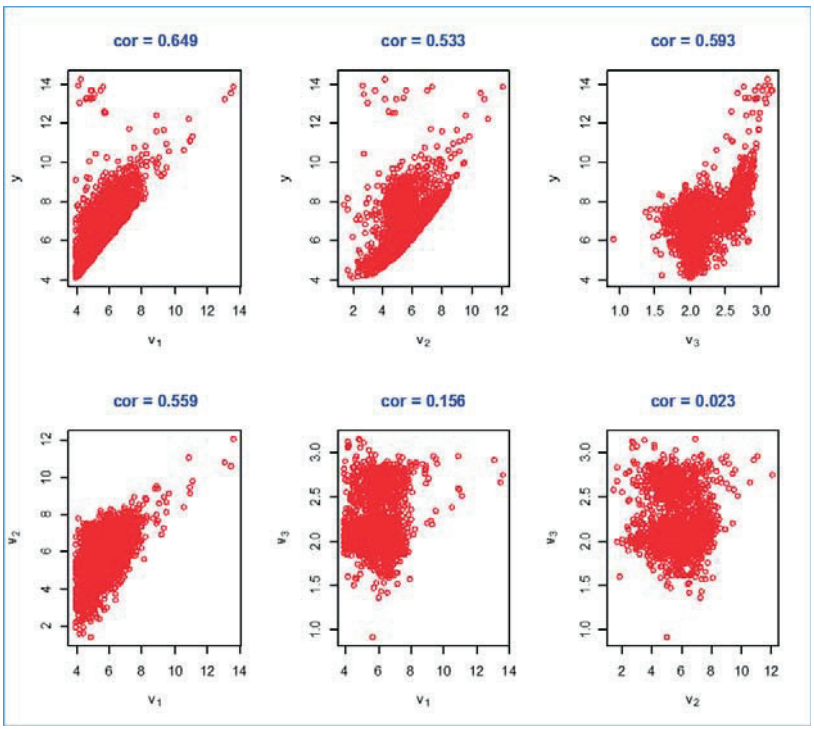


Figure 1: Pairwise correlations among the original variables used in the simulation study. Correlations calculated on the set of synthetic population.

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.

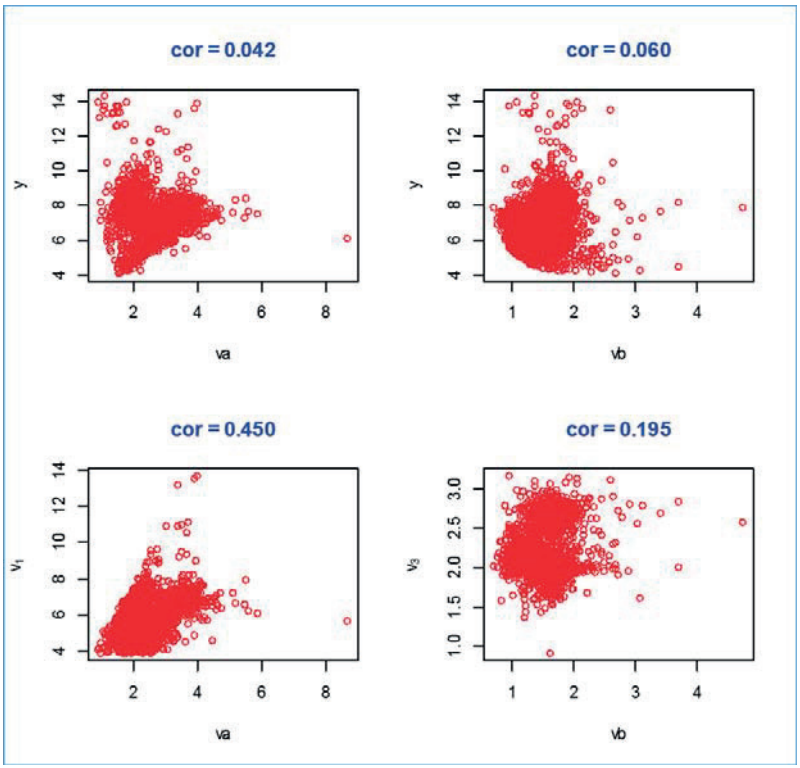


Figure 2: Pairwise correlations among artificial and original variables. Correlations calculated on the set of synthetic population.

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.

PAPER II

Variance Estimation in Two-Step Calibration for Nonresponse Adjustment

Bernardo João Rota

Abstract

Rota and Laitila (2015) suggest an alternative two-step calibration estimation resulting from combining two calibration estimation approaches, i.e., linear calibration (Särndal and Lundström 2005) and propensity score calibration (Chang and Kott 2008), when the functional form of the response probability is assumed to be known. The first step focuses on estimating this function and the second step on estimating the total of a survey variable. This paper extends these previous findings by deriving an approximate variance expression and suggesting a variance estimator for the two-step estimator. The paper also justifies the use of sample-level auxiliary information in the first step of estimation, deferring the use of population-level auxiliary information to the second step of estimation.

Keywords: Two-step, Variance estimator, Calibration, Nonresponse, Auxiliary information, Response probability.

1 Introduction

Efficient estimation in surveys affected by nonresponse requires the appropriate use of auxiliary information. This theme is emphasized by, for example, Rizzo et al. (1996), Särndal and Lundström (2007), and Brick (2013). Various approaches to accounting for the negative effects of nonresponse are proposed in the literature, with weighting the units in the response being one alternative. Auxiliary information can be available at different levels, such as the sample-level, population-level, or both. When both these levels of auxiliary information are available, they offer alternative ways of constructing the auxiliary vectors (see Estevão and Särndal 2002). Moreover, the combined use

of population and sample level auxiliary information gives further alternatives when estimating population characteristics. One such alternative is the estimation in two steps.

A two-step estimation by calibration approach is suggested by, for example, Särndal and Lundström (2005), with linear calibration acting in both steps. Kott and Liao (2015) also suggest a two-step calibration estimation approach assuming a known functional form of the response mechanism.

In two-step estimation, sample-level auxiliary information can be used in the initial adjustment to correct for nonresponse bias and population-level auxiliary information in the final adjustment intended to reduce the sampling variance. One reason for employing sample auxiliary data for preliminary adjustment is that these data may well capture important respondent characteristics. For example, if the sample auxiliary data are process data, they will generally embody information about the nonresponse pattern, which may be important in correcting for nonresponse bias (e.g., Brick 2013).

Calibration adjustment, initially conceived for correcting sampling errors (Deville and Särndal 1992; Deville et al. 1993), is currently one of the most appealing techniques for nonresponse adjustment. The rationale of calibration is to construct adjustment weights that replicate known quantities. Several calibration schemes have been proposed in the literature, including:

1. Linear calibration (LC) (e.g., Lundström and Särndal 1999) is derived from a Chi-square type function that minimizes the distance between the sampling weights and the calibrated weights. In the absence of nonresponse, this calibration estimator takes the form of a generalized regression (GREG) estimator (Särndal et al. 1992). An important feature of this version of calibration is that it simply relies on the strength of the auxiliary variables in explaining either variables of interest, the response pattern, or both, without an explicit need for modeling.

2. Propensity calibration (PC) (e.g., Chang and Kott 2008; Kim and Park 2010; Kott and Day 2014; Kott and Liao 2015) relies on explicit modeling of the response pattern, that is, the functional form of the response model is assumed to be known and its parameters are estimated by means of the calibration principle.

3. Model calibration (MC) (e.g. Wu and Sitter, 2001; Särndal, 2007; Lehtonen et al., 2008; Rueda et al., 2010). Here, the data is assumed to be generated by an underlying process described by specific model that links the survey variable of interest to some covariates, and calibration is used in construction of weights that are consistent with population totals of the predicted targets obtained using that model.

4. Hybrid calibration (HC) (Lehtonen and Veijanen, 2015). This calibration scheme combines MC and model-free calibration estimator, attempting thus to exploit their favourable properties. The auxiliary vector encompasses both the auxiliary data and predictions of the study variable values.

The last two calibration schemes illustrate the recent advances in the calibration approach. They have been used for domain or small area estimation, e.g., (Lehtonen and Veijanen, 2012, 2015).

Rota and Laitila (2015) combine LC and PC schemes and construct an alternative estimator of the total Y of a survey variable y by means of two-step estimation in the presence of sample- and population-level auxiliary information under the assumption of a known functional form of the response mechanism. In line with this setup, this paper contributes by deriving an approximate variance expression and suggesting a variance estimator for this alternative two-step estimator. Moreover, we demonstrate that the use of sample-level auxiliary information generally yields more efficient two-step estimator than does the use of population-level auxiliary information. Simulation studies are carried out to illustrate the properties of the two-step estimator and its variance.

The rest of the paper is organized as follows: section 2 introduces calibration theory; the two-step estimator is presented in section 3 and the variance and variance estimator in section 4; in section 5, we provide arguments justifying the use of sample auxiliary information in the first step of estimation; the simulation study is presented in section 6 and the results are discussed in the final section.

2 Introduction of calibration estimation

2.1 Notations

Sample s of n elements is drawn from population $U = \{1, 2, ..., k, ..., N\}$ of size N using a probability sampling design, $p(s)$, that yields the first- and second-order inclusion probabilities $\pi_k = \Pr(k \in s) > 0$ and $\pi_{kl} = \Pr(k, l \in s) > 0$, respectively, and $\pi_{kk} = \pi_k$ for all $k, l \in U$. Let $r \subset s$ denote the response set. Units in the sample respond independently of each other with probability $q_k = \Pr(k \in r \,|k \in s) > 0$. Assume y to be the survey variable of which we are interested in estimating its total $Y = \sum_{k\epsilon U} y_k$ using auxiliary information defined as:

(a) $\mathbf{x}_k = (x_{1k}, x_{2k}, ..., x_{Jk})^t$, a J-dimensional vector of known values for all elements k in the response set r; for each $j = 1, ..., J$, $T_{xj} = \sum_{k\epsilon U} x_{jk}$ is

known. This implies that $T_x = (T_{x1}, T_{x2}, ..., T_{xJ})^t$ is also known.

(b) $\mathbf{z}_k = (z_{1k}, z_{2k}, ..., z_{Lk})^t$, an L-dimensional vector of known values for all elements k in the sample set, s. For each $l = 1, ..., L$, we can estimate $\hat{t}_{zl} = \sum_{k\epsilon s} d_k z_{lk}$ and compose the vector $\hat{t}_z = \left(\hat{t}_{z1}, \hat{t}_{z2}, ..., \hat{t}_{zL}\right)^t$.

Unless otherwise stated, the expected value $E_p E_q(A)$, is written simply as $E(A)$.

2.2 Calibration estimators

Calibration estimators are a class of weighted estimators of the form $\hat{Y}_{cal} = \sum_{k\epsilon r} w_k y_k$, with weights w_k satisfying the calibration constraint $\sum_{k\epsilon r} w_k \check{\mathbf{x}}_k = \check{\mathbf{X}}$, where $\check{\mathbf{x}}_k$ stands for $\mathbf{x}_k$, $\mathbf{z}_k$, or $\check{\mathbf{x}}_k = (\mathbf{x}_k^t, \mathbf{z}_k^t)^t$ and $\check{\mathbf{X}}$ corresponds to their respective totals, i.e., T_x, $\hat{t}_z$, or $\left(T_x^t, \hat{t}_z^t\right)^t$. Papers by Deville and Särndal (1992) and Deville et al. (1993), benchmarks in calibration estimation theory, approach calibration in the context of full-sample responses and their main purpose was the reduction of sampling errors. The approach was then extended to cases of samples with nonresponse in order to reduce nonresponse bias (e.g., Singh et al. 1995; Niyonsenga 1997; Lundström and Särndal 1999; Kreuter and Olson 2011).

The minimum-distance approach to deriving calibration weights aims to determine calibrated weights as close as possible to the design weights by means of a distance function, $D(w, d)$. Deville and Särndal (1992) required the distance D to be positive and a convex function of its arguments, with $D(0) = dD(0) = 1$, where d stands for the first derivative. Minimizing D, subject to the above calibration constraint and using a Lagrange function, leads to calibrated weights of the form $w_k = d_k F(\cdot)$, where $F^{-1}(a) = dD(a)$ and $d_k = 1/\pi_k$. When D is chosen to be

$$D(w, d) = \sum_{k\epsilon r} \left[d_k^{-2}(w_k - d_k)\right]^2 /2, \tag{1}$$

the calibrated weights are given by $w_k = d_k + d_k \mathbf{g}^t \check{\mathbf{x}}_k$, which are linear in the coefficient vector $\mathbf{g}^t = \left(\check{\mathbf{X}} - \sum_{k\epsilon r} d_k \check{\mathbf{x}}_k\right)^t \left(\sum_{k\epsilon r} d_k \check{\mathbf{x}}_k \check{\mathbf{x}}_k^t\right)^{-1}$. The resulting estimator of Y, commonly termed a linear calibration estimator, is given by

$$\hat{Y}_{LC} = \sum_{k\epsilon r} d_k y_k + \mathbf{g}^t \sum_{k\epsilon r} d_k \check{\mathbf{x}}_k y_k. \tag{2}$$

Other distance functions will generally produce calibrated weights that are nonlinear in their coefficients and deriving these weights may require some

iterative procedures. Deville et al. (1993) provide a set of common distance functions that can be used in generating calibrated weights.

A direct approach when adjusting for nonresponse is to assume that $F(\cdot)$ is the nonresponse adjustment weight and to choose it suitably. The principle is known as response propensity, in which $F^{-1}(\cdot)$ is a probability function. The calibration equation $\sum_{k\epsilon r} d_k F(\cdot)\check{\mathbf{x}}_k = \check{\mathbf{X}}$ is employed in estimating the function $F(\cdot)$. Chang and Kott (2008) use this principle in constructing the estimator $\hat{Y}_{cal}$, with $F(\cdot) = F(\mathbf{z}_k^t\mathbf{g})$, where $\mathbf{z}_k$ with dimension less or equal to that of $\mathbf{x}_k$, is known only for $k\epsilon r$. They also suggest an iterative algorithm for estimating $\mathbf{g}$.

3 Calibrating in two steps

Särndal and Lundström (2005) suggest a two-step calibration estimator, here denoted by $\hat{Y}_{2LC}$. The first- and second-step weights are constructed according to the principle of combining population- and sample-level auxiliary information. In the first step, sample-level information is used to construct intermediate weights, w_{1k}, such that $\sum_{k\epsilon r} w_{1k}\mathbf{z}_k = \sum_{k\epsilon s} d_k\mathbf{z}_k$. In the second step, weights w_{1k} replace the design weights in the optimization problem that led to calibration estimator (2), and the final weights, w_{2k}, satisfy $\sum_{k\epsilon r} w_{2k}\check{\mathbf{x}}_k = \check{\mathbf{X}}$, where $\check{\mathbf{x}}_k = \mathbf{x}_k$ with $\check{\mathbf{X}} = T_x$ or $\check{\mathbf{x}}_k = (\mathbf{x}_k^t, \mathbf{z}_k^t)^t$ with $\check{\mathbf{X}} = \left(T_x^t, \hat{t}_z^t\right)^t$.

The two-step estimator suggested by Rota and Laitila (2015) assumes that the functional form of the response probability is known and is given by $q_k = q(\mathbf{z}_k^t\mathbf{g})$.

In the rest of the paper we use $\hat{F}_k = F\left(\mathbf{z}_k^t\hat{\mathbf{g}}\right)$, $F_k = F\left(\mathbf{z}_k^t\mathbf{g}\right)$, and $F_k^\circ = F\left(\mathbf{z}_k^t\mathbf{g}_\circ\right)$, where $\mathbf{g}$ is a generic parameter vector, $\mathbf{g}_\circ$ is the true value of $\mathbf{g}$, $\hat{\mathbf{g}}$ is a consistent estimator of $\mathbf{g}_\circ$, and $F_k = 1/q_k$.

Rota and Laitila (2015) define intermediate weights as $w_{1k} = d_k\hat{F}_k$, after calculating $\hat{\mathbf{g}}$ in the first step from the calibration equation $\sum_{k\epsilon r} d_k F_k\mathbf{z}_k = \hat{t}_z$. The second-step weights, w_{2k}, are derived from the problem $\underset{\{w_{2k}\}}{min} \sum_{k\epsilon r} \frac{(w_{2k}-w_{1k})^2}{2w_{1k}}$ subject to $T_x = \sum_{k\epsilon r} w_{2k}\mathbf{x}_k$ and given by $w_{2k} = w_{1k}v_{2k}$ with $v_{2k} = 1 + \mathbf{g}_2^t\mathbf{x}_k$ and $\mathbf{g}_2^t = \left(T_x - \sum_{k\epsilon r} w_{1k}\mathbf{x}_k\right)^t \left(\sum_{k\epsilon r} w_{1k}\mathbf{x}_k\mathbf{x}_k^t\right)^{-1}$, assuming that $\sum_{k\epsilon r} w_{1k}\mathbf{x}_k\mathbf{x}_k^t$ is invertible. Then, the two-step estimator for the total Y is given by $\hat{Y}_{2step} = \sum_{k\epsilon r} w_{2k}y_k$. This estimator can be equivalently written as:

$$\hat{Y}_{2step} = \sum_{k\epsilon r} d_k\hat{F}_k y_k + \left(T_x - \sum_{k\epsilon r} d_k\hat{F}_k\mathbf{x}_k\right)^t \hat{\mathbf{B}}_{2Fr} \qquad (3)$$

where $\hat{\mathbf{B}}_{2Fr} = \left(\sum_{k\epsilon r} d_k \hat{F}_k \mathbf{x}_k \mathbf{x}_k^t\right)^{-1} \sum_{k\epsilon r} d_k \hat{F}_k \mathbf{x}_k y_k$.

4 The variance and variance estimator

The following assumptions are used in deriving the variance of the two-step estimator:

(i) The sequence of populations and samples increases to infinity, as in Isaki and Fuller (1982).

(ii) Function $F(\cdot\mathbf{g})$ is monotonic and continuous for all $\mathbf{g}$ in $\mathbf{G}$, with finite first derivatives.

(iii) $\mathbf{v}_k = (\mathbf{x}_k, \mathbf{z}_k, y_k)$ is nonrandom and $\|\mathbf{v}_k\| < \infty$.

(iv) $\left(\hat{\mathbf{B}}_{2Fr} - \mathbf{B}_2\right)$, $N^{-1}\left(T_x - \sum_{k\epsilon r} d_k F_k^{\circ} \mathbf{x}_k\right)$, and $N^{-1}\left(\hat{t}_z - \sum_{k\epsilon r} d_k F_k^{\circ} \mathbf{z}_k\right)$ are all $O_p(n^{-\frac{1}{2}})$, where $\mathbf{B}_2 = \left(\sum_{k\epsilon U} \mathbf{x}_k \mathbf{x}_k^t\right)^{-1} \sum_{k\epsilon U} \mathbf{x}_k y_k$ is the population analogous to $\hat{\mathbf{B}}_{2Fr}$.

(v) $N^{-1}\sum_{k\epsilon r} d_k \mathbf{x}_k \mathbf{F}_{1k}^{\circ}$, $N^{-1}\sum_{k\epsilon r} d_k \mathbf{F}_{1k}^{\circ} y_k$, and $N^{-1}\sum_{k\epsilon r} d_k F_k^{\circ} \mathbf{x}_k \left(\mathbf{x}_k\right)^t$ are $O_p(1)$, where, $\mathbf{F}_1 = dF/d\mathbf{g}$.

Given that $\hat{\mathbf{g}}$ is a solution to $\sum_{k\epsilon r} d_k F_k \mathbf{z}_k = \hat{t}_z$, we proceed as follows: $\sum_{k\epsilon r} d_k F_k^{\circ} \mathbf{z}_k - \hat{t}_z = \sum_{k\epsilon r} d_k \hat{F}_k \mathbf{z}_k - \hat{t}_z + \sum_{k\epsilon r} d_k \mathbf{z}_k \tilde{\mathbf{F}}_{1k}\left(\hat{\mathbf{g}} - \mathbf{g}_{\circ}\right) = O_p\left(N n^{-\frac{1}{2}}\right)$. This leads to equation (4) below:

$$\left(\hat{\mathbf{g}} - \mathbf{g}_{\circ}\right) = \mathbf{\Gamma}^{-1} N^{-1} \left(\sum_{k\epsilon r} d_k F_k^{\circ} \mathbf{z}_k - \hat{t}_z\right) + o_p\left(n^{-\frac{1}{2}}\right) = O_p\left(n^{-\frac{1}{2}}\right) \qquad (4)$$

where $\mathbf{\Gamma}$ is the probability limit of $N^{-1}\sum_{k\epsilon r} d_k \mathbf{z}_k \tilde{F}_{1k}$, assumed invertible and $\tilde{\mathbf{F}}_{1k} = \mathbf{F}_{1k}\left(\mathbf{z}_k^t \tilde{\mathbf{g}}\right)$, with $\tilde{\mathbf{g}}$ being a convex combination of $\hat{\mathbf{g}}$ and $\mathbf{g}_{\circ}$.

A first-order Taylor approximation of $\hat{Y}_{2step}$ at $\mathbf{g}_{\circ}$ gives:

$$\begin{aligned}\hat{Y}_{2step} \approx & \sum_{k\epsilon r} d_k F_k^{\circ} y_k + \left(T_x - \sum_{k\epsilon r} d_k F_k^{\circ} \mathbf{x}_k\right)^t \hat{\mathbf{B}}_{2Fr}^{\circ} \\ & + \sum_{k\epsilon r} d_k \mathbf{F}_{1k}^{\circ}\left(\hat{\mathbf{g}} - \mathbf{g}_{\circ}\right)\left(y_k - \mathbf{x}_k^t \hat{\mathbf{B}}_{2Fr}^{\circ}\right) + \boldsymbol{\lambda}_{\circ}^t \sum_{k\epsilon r} d_k \mathbf{x}_k \mathbf{F}_{1k}^{\circ}\left(\hat{\mathbf{g}} - \mathbf{g}_{\circ}\right)\left(y_k - \mathbf{x}_k^t \hat{\mathbf{B}}_{2Fr}^{\circ}\right)\end{aligned} \qquad (5)$$

where $\boldsymbol{\lambda}_{\circ}^t = N^{-1}\left(T_x - \sum_{k\epsilon r} d_k F_k^{\circ} \mathbf{x}_k\right)^t \left(N^{-1}\sum_{k\epsilon r} d_k F_k^{\circ} \mathbf{x}_k \mathbf{x}_k^t\right)^{-1}$ is $O_p(n^{-\frac{1}{2}})$.

Now, as in Estevão and Särndal (2006), we can replace $\hat{\mathbf{B}}^{\circ}_{2Fr}$ in (5) with $\left(\mathbf{B}_2 + \hat{\mathbf{B}}^{\circ}_{2Fr} - \mathbf{B}_2\right)$ and obtain:

$$\hat{Y}^{\circ}_{2step} = \sum_{k\epsilon r} d_k F^{\circ}_k E_k + T^t_x \mathbf{B}_2 + \sum_{k\epsilon r} d_k \mathbf{F}^{\circ}_{1k} \left(\hat{\mathbf{g}} - \mathbf{g}_{\circ}\right) E_k + \mathbf{R} \tag{6}$$

where $\mathbf{R} = \sum_{k\epsilon r} d_k \mathbf{F}^{\circ}_{1k} \left(\hat{\mathbf{g}} - \mathbf{g}_{\circ}\right) \boldsymbol{\lambda}^t_{\circ} \mathbf{x}_k E_k +$
$\left[\left(T_x - \sum_{k\epsilon r} d_k F^{\circ}_k \mathbf{x}_k\right)^t - \sum_{k\epsilon r} d_k v^{\circ}_k \mathbf{F}^{\circ}_{1k} \left(\hat{\mathbf{g}} - \mathbf{g}_{\circ}\right) \mathbf{x}_k\right]^t \left(\hat{\mathbf{B}}^{\circ}_{2Fr} - \mathbf{B}_2\right)$, $v^{\circ}_k = 1 + \boldsymbol{\lambda}^t_{\circ} \mathbf{x}_k$, and $E_k = y_k - \mathbf{x}^t_k \mathbf{B}_2$.

In (6), $\sum_{k\epsilon r} d_k F^{\circ}_k E_k$ and $\sum_{k\epsilon r} d_k \mathbf{F}^{\circ}_{1k} \left(\hat{\mathbf{g}} - \mathbf{g}_{\circ}\right) E_k$ are $O_p(Nn^{-\frac{1}{2}})$, whereas $\mathbf{R}$ is $O_p(Nn^{-1})$, thus, of lower order. This lower-order term is then dropped to obtain the approximate expression for the two-step estimator of Y :

$$\hat{Y}^{\bullet}_{2step} = \sum_{k\epsilon r} d_k F^{\circ}_k E_k + \sum_{k\epsilon r} d_k \mathbf{F}^{\circ}_{1k} \left(\hat{\mathbf{g}} - \mathbf{g}_{\circ}\right) E_k + T^t_x \mathbf{B}_2. \tag{7}$$

If we replace $(\hat{\mathbf{g}} - \mathbf{g}_{\circ})$ in (7) with the corresponding expression in (4), we get

$$\hat{Y}^{\bullet}_{2step} = \sum_{k\epsilon r} d_k F^{\circ}_k E_k + \sum_{k\epsilon r} d_k \mathbf{F}^{\circ}_{1k} \tilde{\boldsymbol{\Gamma}}^{-1} \left(\sum_{k\epsilon r} d_k F^{\circ}_k \mathbf{z}_k - \hat{t}_z\right) E_k + T^t_x \mathbf{B}_2 + o_p(Nn^{-\frac{1}{2}}) \tag{8}$$

where $\tilde{\boldsymbol{\Gamma}}^{-1} = \boldsymbol{\Gamma}^{-1} N^{-1}$. Let $\sum_{k,l\epsilon A} = \sum_{k\epsilon A} \sum_{l\epsilon A}$ and write (8) as:

$$\hat{Y}^{\bullet}_{2step} = \sum_{k\epsilon s} R_k d_k F^{\circ}_k E_k + \sum_{k,l\epsilon s} R_k (R_l F^{\circ}_l - 1) A_{kl} + T^t_x \mathbf{B}_2 \tag{9}$$

where $A_{kl} = d_k d_l \mathbf{z}^t_l \left(\mathbf{F}^{\circ}_{1k} \tilde{\boldsymbol{\Gamma}}^{-1}\right)^t E_k$, and $R_k = 1$ if k is a respondent; $R_k = 0$, otherwise.

The variance of (3) is approximated by the variance of (9) given by:

$$Var\left(\hat{Y}^{\bullet}_{2step}\right) = Var(\hat{T}^{\circ}_a) + Var(\hat{T}^{\circ}_b) + 2Cov\left(\hat{T}^{\circ}_b, \hat{T}^{\circ}_a\right). \tag{10}$$

where $\hat{T}^{\circ}_a = \sum_{k\epsilon s} R_k d_k F^{\circ}_k E_k$ and $\hat{T}^{\circ}_b = \sum_{k,l\epsilon s} R_k (R_l F^{\circ}_l - 1) A_{kl}$.

The variances on the r.h.s. of (10) are obtained using result 9.3.1 in Särndal et al. (1992, p. 348) and given by:

$Var(\hat{T}_a^\circ) = \sum_{k\neq l\epsilon U}(\pi_{kl}d_k d_l - 1)E_k E_l + \sum_{k\epsilon U}(d_k F_k^\circ - 1)E_k^2,$

$Var(\hat{T}_b^\circ) = \sum_{k\neq l\neq i\epsilon U}\frac{\pi_{kli}(F_l^\circ-1)}{F_k^\circ F_i^\circ}A_{kl}A_{il} + \sum_{k\neq l\epsilon U}\frac{\pi_{kl}-\pi_k\pi_l}{F_k^\circ F_l^\circ}(F_l^\circ - 1)(F_k^\circ - 1)A_{kk}A_{ll} + \sum_{k\neq l\epsilon U}\frac{\pi_{kl}(F_l^\circ-1)}{F_k^\circ}A_{kl}^2 + \sum_{k\neq l\epsilon U}\frac{\pi_{kl}}{F_k^\circ F_l^\circ}(1-F_k^\circ)(1-F_l^\circ)A_{kl}A_{lk} + \sum_{k\neq l\epsilon U}\frac{2\pi_{kl}}{F_k^\circ F_l^\circ}(1-F_l^\circ)^2 A_{kl}A_{ll} + \sum_{k\epsilon U}\frac{\pi_k(F_k^\circ-\pi_k)(F_k^\circ-1)^2}{(F_k^\circ)^2}A_{kk}^2,$
and
$Cov\left(\hat{T}_a^\circ, \hat{T}_b^\circ\right) = \sum_{k\neq l\epsilon U}\frac{d_l\pi_{kl}}{F_k^\circ}\left((F_l^\circ-1)A_{kl} + (F_k^\circ-1)A_{kk}\right)E_l + \sum_{k\epsilon U}(F_k^\circ - 1)A_{kk}E_k - \sum_{k,l\epsilon U}\frac{\pi_k(F_k^\circ-1)}{F_k^\circ}A_{kk}E_l.$

Some details of the derivation of these formulae are given in Appendix. The corresponding variance estimator is given by:

$$\hat{Var}\left(\hat{Y}_{2step}^\bullet\right) = \hat{Var}(\hat{T}_a) + \hat{Var}(\hat{T}_b) + 2\hat{Cov}\left(\hat{T}_b, \hat{T}_a\right) \tag{11}$$

where, $\hat{Var}(\hat{T}_a) = \sum_{k\neq l\epsilon r}(d_k d_l - d_{kl})\check{e}_k\check{e}_l + \sum_{k\epsilon r} d_k\hat{F}_k(d_k\hat{F}_k - 1)e_k^2,$
$\hat{Var}(\hat{T}_b) = \sum_{k\neq l\neq i\epsilon r}\hat{F}_l(\hat{F}_l - 1)\hat{A}_{kl}\hat{A}_{il} + \sum_{k\neq l\epsilon r}(1 - d_{kl}\pi_k\pi_l)(\hat{F}_k - 1)(\hat{F}_l - 1)\hat{A}_{kk}\hat{A}_{ll}$
$\sum_{k\neq l\epsilon r}\hat{F}_l(\hat{F}_l-1)\hat{A}_{kl}^2 + \sum_{k\neq l\epsilon r}(1-\hat{F}_k)(1-\hat{F}_l)\hat{A}_{kl}\hat{A}_{lk} + \sum_{k\neq l\epsilon r}2(1-\hat{F}_l)^2\hat{A}_{kl}\hat{A}_{ll}$
$+\sum_{k\epsilon r}\frac{(\hat{F}_k-1)^2(\hat{F}_k-\pi_k)}{\hat{F}_k}\hat{A}_{kk}^2,$
and
$\hat{Cov}\left(\hat{T}_b, \hat{T}_a\right) = \sum_{k\neq l\epsilon r} d_l\left((\hat{F}_l-1)\hat{A}_{kl} + (\hat{F}_k-1)\hat{A}_{kk}\right)\check{e}_l + \sum_{k\epsilon r} d_k(\hat{F}_k-1)\hat{A}_{kk}\check{e}_k - \sum_{k,l\epsilon r} d_l(\hat{F}_k-1)\hat{A}_{kk}\check{e}_l,$

with
$\hat{T}_a = \sum_{k\epsilon s} R_k d_k\hat{F}_k e_k$, $\hat{T}_b = \sum_{k,l\epsilon s} R_k(R_l\hat{F}_l - 1)\hat{A}_{kl}$, $\hat{A}_{kl} = d_k d_l \mathbf{z}_l^t\left(\hat{\mathbf{F}}_{1k}\hat{\tilde{\mathbf{\Gamma}}}^{-1}\right)^t e_k$,
$\hat{\tilde{\mathbf{\Gamma}}} = \sum_{k\epsilon r} d_k\mathbf{z}_k\hat{\mathbf{F}}_{1k}$, $d_{kl} = 1/\pi_{kl}$, $\check{e}_k = \hat{F}_k e_k$, and $e_k = y_k - \mathbf{x}_k^t\hat{\mathbf{B}}_{2Fr}$.

Note: As the third-order inclusion probability in variance estimator (11) vanishes, the triple sum involved is easily factorized into a product of double and single sums, making the computation easier. Below we provide the factorization of this sum:

$\sum_{k\neq l\neq i\epsilon r}\hat{F}_l(\hat{F}_l-1)\hat{A}_{kl}\hat{A}_{il} = \sum_{k\neq l\epsilon r} d_l\hat{F}_l(\hat{F}_l-1)\hat{A}_{kl}(\hat{\tilde{\mathbf{\Gamma}}}^{-1}\mathbf{x}_l^s)^t\sum_{i\epsilon r} d_i(\hat{\mathbf{F}}_{1i})^t e_i$
$-\sum_{k\neq l\epsilon r}\hat{F}_l(\hat{F}_l - 1)\left(\hat{A}_{kl}^2 + \hat{A}_{kl}\hat{A}_{ll}\right).$

Remark: The last two terms on the r.h.s. of equation (10) represent the contribution of the variance of the model parameter estimates to the variance of the two-step estimator. A question may therefore be raised: Is it worthwhile correcting for the uncertainty in model parameter estimates when estimating the variance of the two-step estimator?

5 Efficiency gain with calibration at sample level

5.1 Efficiency in estimating the model parameters

The principal goal of the first step is the appropriate estimation of the response model. This is of particular importance in protecting the target estimates against nonresponse bias. We can formally illustrate this in the following:

Let

$$\hat{\mathbf{H}}(\mathbf{g}) = \sum_{k\epsilon r} d_k F_k \mathbf{z}_k - \hat{t}_z \tag{12}$$

with $E\left(\hat{\mathbf{H}}(\mathbf{g}_\circ)\right) = \mathbf{0}$.

From Särndal et al. (1992) result 9.3.1, the covariance of $\hat{\mathbf{H}}(\mathbf{g}_\circ)$ is given by

$$E\left(\hat{\mathbf{H}}(\mathbf{g}_\circ)\hat{\mathbf{H}}^t(\mathbf{g}_\circ)\right) = \sum_{k\epsilon U} d_k(F_k^\circ - 1)\mathbf{z}_k\mathbf{z}_k^t. \tag{13}$$

We assume that the vector of estimating equations, $\hat{\mathbf{H}}(\mathbf{g}) = \mathbf{0}$, is uniquely solved for $\mathbf{g} = \hat{\mathbf{g}}$ and consider assumptions (i) and (ii) in section 4. From (4) we observe that the asymptotic variance of the response model coefficients is given by:

$$Avar\left(\sqrt{n}\left(\hat{\mathbf{g}} - \mathbf{g}_\circ\right)\right) = \left[\left(\mathbf{M}\left(\mathbf{g}_\circ\right)\right)^{-1}\right] \boldsymbol{\Psi} \left[\left(\mathbf{M}\left(\mathbf{g}_\circ\right)\right)^{-1}\right] \tag{14}$$

where $\mathbf{M}\left(\mathbf{g}_\circ\right) = plim_{n\to\infty}\frac{1}{n}\frac{d\hat{\mathbf{H}}(\mathbf{g}_\circ)}{d\mathbf{g}}$ and $\boldsymbol{\Psi} = plim_{n\to\infty} n^{-1} E\left(\hat{\mathbf{H}}(\mathbf{g}_\circ)\hat{\mathbf{H}}^t(\mathbf{g}_\circ)\right)$.

Now, suppose that $t_z = \sum_U \mathbf{z}_k$ is known. Then (12) is defined as:

$$\hat{\tilde{\mathbf{H}}}(\mathbf{g}) = \sum_{k\epsilon r} d_k F_k \mathbf{z}_k - t_z \tag{15}$$

with the same properties as before except that

$$E\left(\hat{\tilde{\mathbf{H}}}(\mathbf{g}_\circ)\hat{\tilde{\mathbf{H}}}^{\,t}(\mathbf{g}_\circ)\right) = \sum_{k,l\epsilon U} d_k d_l(\pi_{kl} - \pi_k\pi_l)\mathbf{z}_k\mathbf{z}_k^t + \sum_{k\epsilon U} d_k(F_k^\circ - 1)\mathbf{z}_k\mathbf{z}_k^t. \tag{16}$$

Using similar arguments as those that led to (14), we have that

$$Avar\left(\sqrt{n}\left(\hat{\mathbf{g}}-\mathbf{g}_{\circ}\right)\right)=\left[\left(\mathbf{M}\left(\mathbf{g}_{\circ}\right)\right)^{-1}\right]\boldsymbol{\Phi}\left[\left(\mathbf{M}\left(\mathbf{g}_{\circ}\right)\right)^{-1}\right] + \left[\left(\mathbf{M}\left(\mathbf{g}_{\circ}\right)\right)^{-1}\right]\boldsymbol{\Psi}\left[\left(\mathbf{M}\left(\mathbf{g}_{\circ}\right)\right)^{-1}\right] \tag{17}$$

where $\boldsymbol{\Phi}$ and $\boldsymbol{\Psi}$ are the first and second components of $plim_{n\to\infty}n^{-1}E\left(\hat{\tilde{\mathbf{H}}}(\mathbf{g}_{\circ})\hat{\tilde{\mathbf{H}}}^{t}\left(\mathbf{g}_{\circ}\right)\right)$, respectively.

The difference between equations (17) and (14) is $\widetilde{\mathbf{M}}\left(\mathbf{g}_{\circ}\right)=\left[\left(\mathbf{M}\left(\mathbf{g}_{\circ}\right)\right)^{-1}\right]\boldsymbol{\Phi}\left[\left(\mathbf{M}\left(\mathbf{g}_{\circ}\right)\right)^{-1}\right]$, which is a positive definite matrix, unless it is a case of census. This illustrates that (12) is more appropriate than (15) in the first step of estimation.

5.2 Efficiency in estimating the total Y

Let $\tilde{\tilde{\mathbf{g}}}$ be the solution to $\hat{\tilde{\mathbf{H}}}(\mathbf{g})=\mathbf{0}$ and $\tilde{\hat{Y}}^{\bullet}_{2step}=\hat{T}^{\circ}_{a}+\hat{T}^{\circ}_{c}(\tilde{\tilde{\mathbf{g}}}-\mathbf{g}_{\circ})+T^{t}_{x}\mathbf{B}_{2}$ is the corresponding equation (7) when $\hat{\mathbf{g}}$ is replaced with $\tilde{\tilde{\mathbf{g}}}$. Furthermore, if $\hat{\mathbf{g}}_{a}$ is uncorrelated with either $\hat{T}^{\circ}_{a}=\sum_{k\epsilon r}d_{k}F^{\circ}_{k}E_{k}$ or $\hat{T}^{\circ}_{c}=\sum_{k\epsilon r}d_{k}\mathbf{F}^{\circ}_{1k}E_{k}$, where $\hat{\mathbf{g}}_{a}$ stands for $\hat{\mathbf{g}}$ or $\tilde{\tilde{\mathbf{g}}}$, $\hat{T}^{\circ}_{c}$ is a non-zero vector, and given that $E(\hat{\mathbf{g}}_{a}-\mathbf{g}_{\circ})\to\mathbf{0}$ (see equation 4), we have that

$$\begin{aligned}
&Var\left(\tilde{\hat{Y}}^{\bullet}_{2step}\right)-Var\left(\hat{Y}^{\bullet}_{2step}\right)=\\
&\quad Var\left(\hat{T}^{\circ}_{c}(\tilde{\tilde{\mathbf{g}}}-\mathbf{g}_{\circ})\right)-Var\left(\hat{T}^{\circ}_{c}(\hat{\mathbf{g}}-\mathbf{g}_{\circ})\right)\\
&\quad +2Cov\left(\hat{T}^{\circ}_{a},\hat{T}^{\circ}_{c}\right)\left(E(\tilde{\tilde{\mathbf{g}}}-\mathbf{g}_{\circ})-E(\hat{\mathbf{g}}-\mathbf{g}_{\circ})\right)=\\
&\quad E\left(\hat{T}^{\circ}_{c}(\tilde{\tilde{\mathbf{g}}}-\mathbf{g}_{\circ})(\tilde{\tilde{\mathbf{g}}}-\mathbf{g}_{\circ})^{t}\hat{T}^{\circ t}_{c}\right)-E\left(\hat{T}^{\circ}_{c}(\hat{\mathbf{g}}-\mathbf{g}_{\circ})(\hat{\mathbf{g}}-\mathbf{g}_{\circ})^{t}\hat{T}^{\circ t}_{c}\right)\\
&\quad =E\left(\hat{T}^{\circ}_{c}\widetilde{\mathbf{M}}\left(\mathbf{g}_{\circ}\right)\hat{T}^{\circ t}_{c}\right)>0.
\end{aligned}$$

Thus, the efficiency loss of $\tilde{\tilde{\mathbf{g}}}$ resulting from calibrating with population-level auxiliary information is indicated to yield an efficiency loss of the two-step estimator (3).

6 Simulations

Two simulation studies were performed to illustrate the properties of the two-step estimator and its variance. In the following, we describe the setup of each

simulation study.

6.1 The setup

6.1.1 Study 1

We used data from a real estate survey with 4228 sampled elements of which 1783 were nonrespondents. We selected five variables from the study. A categorical variable that was a stratum indicator in the original six-strata study is denoted by $\boldsymbol{\gamma}_k = (\gamma_{1k}, \gamma_{2k}, \gamma_{3k}, \gamma_{4k}, \gamma_{5k}, \gamma_{6k})$, where $\gamma_{ik} = 1(k \epsilon S_i)$ and S_i is the i^{th} stratum. Three numerical variables denoted x_1, x_2, and z were transformed into logarithmic scales to reduce the variability, with the first two being used as benchmarks and the last as a model variable. Another numerical variable, y, was left untransformed and is the study variable. Here, the estimation concerned estimating the population total, Y.

We performed a logistic regression fit of R to a constant and z, and the resulting model was used as the true response probability function. Here, R is a dichotomous variable of 1/0, i.e., respondent/nonrespondent. The true response probabilities obtained using the model were then attached to the respective elements and used for Bernoulli trials to generate the response sets.

The population consists of the 2445 respondents to the survey and samples of sizes 200, 400, and 600 were selected using simple random sampling without replacement. We assume that the chosen response model is correct, that is, the response probabilities are estimated according to the equation $\hat{q}_k = 1/\left(1 + \exp(-\mathbf{z}_k^t \hat{\mathbf{g}})\right)$, where $\mathbf{z}_k = (1, z_k)^t$ and $\hat{\mathbf{g}}$ is obtained from the first step of estimation. The benchmark vector was a combination of $\boldsymbol{\gamma}$ and x given by $\mathbf{x}_k = (\boldsymbol{\gamma}_k^t, x_k \boldsymbol{\gamma}_k^t)^t$, while x stands for x_1 or x_2. The choices of x_1,x_2, z, and y were based on their relationships in satisfying the following two cases:

In the first case, the estimator's performance is analysed when the correlation between benchmark and model variable is $cor(x_1, z) = 0.16$, while the correlations between the benchmark and the study variable and the model variable and the study variable are $cor(x_1, y) = 0.59$ and $cor(z, y) = 0.65$, respectively. This may be the case when the model and benchmark variables are obtained from different sources, for example, when model variables are process data while the benchmark variables are obtained from administrative registers. The benchmark variables are selected based on their relationship with the survey variable, and the model variables are selected with the intention of capturing the response behavior. This means that, in general, we do not expect a good relationship between the model and benchmark variables, although such a relationship is possible. In the second case, we consider the

possibility of having model variables at least moderately correlated with the benchmark variable and want to observe the impact of this possibility on the variance of the two-step estimator in relation to the first case. The correlations between the variables are the following: $cor(x_2, z) = 0.56$, $cor(x_2, y) = 0.53$, and $cor(z, y) = 0.65$. Each simulation result was based on 1000 replications. The expected response rate was approximately 55%. The estimators are evaluated in terms of relative bias (Rel.bias) and root mean squared error (RMSE).

6.1.2 Study 2

The previous study was based on real survey data, which are important in empirical studies because theoretical findings need to be evaluated in real environments. Although use of real data is important, sometimes freedom to control the environment is desired, for which simulated data are usually appropriate. Accordingly, this study is based on simulated population data of size 2445. The estimation setup is as in the former study except that the variables are generated as follows: $x \sim U(0, 1)$, $z = \rho x + \xi$, where ρ is the required correlation between x and z, $\xi \sim U(0, a)$, and $a = \sqrt{1 - \rho^2}$. The study variable is given by $y = c_1 U(0, x) + c_2 U(0, z)$, where $c_1 = c_2 = 1$ and U is the uniform distribution. The coefficients c_1 and c_2 can be varied to change the mean of y and/or balance or unbalance the correlations ρ_{xy} and ρ_{zy} between x and y and between z and y, respectively. The response model is the same as in study 1 except that the coefficient vector is given by $\mathbf{g}_\circ = (-1.5, 2.0)^t$. We also created a categorical variable, $\boldsymbol{\gamma}_k = (\gamma_{1k}, \gamma_{2k}, \gamma_{3k}, \gamma_{4k})$, where $\gamma_{ik} = 1(k\epsilon S_i)$ and S_i is the i^{th} quartile of x, so that the benchmark vector is given by $\mathbf{x}_k = (\gamma_{1k}, \gamma_{2k}, \gamma_{3k}, \gamma_{4k}, x_k\gamma_{1k}, x_k\gamma_{2k}, x_k\gamma_{3k}, x_k\gamma_{4k})^t$. In the first case, we have a correlation between x and z of 0.2, between x and y of 0.49, and between z and y of 0.53, while in the second case these correlations are 0.7, 0.62, and 0.65, respectively.

6.2 Simulation results

Below we present the simulation results of each of the above studies. The simulations illustrate the ability of the suggested two-step variance estimator to estimate the variance of the two-step calibration estimator. The variance estimator of the two-step estimator (Särndal and Lundström 2005) is used as a benchmark in assessing the performance of our suggested method. The results also enable us to respond to the question raised in Remark, that is, whether it is important to correct for the variance in model parameter estimation when estimating the variance of the two-step estimator. In Tables 1–4

below, $\hat{Y}$ stands for $\hat{Y}_{2LC}$ or $\hat{Y}_{2step}$. In each table, $\hat{Y}_{2step}$ is followed by two results in the column "Rel.bias of $\hat{V}ar(\hat{Y})$", the first of which is the relative bias of the corrected variance estimator, $\hat{V}_{cor} = \hat{V}ar\left(\hat{Y}_{2step}\right)$, and the second, within parentheses, is the relative bias of the uncorrected variance estimator, $\hat{V}_{uncor} = \hat{V}ar\left(\hat{T}_a\right)$.

6.2.1 Results of study 1

Table 1 presents the results of the first simulation study when the correlation between model and benchmark variables is 0.16, while in Table 2 their correlation is 0.56. In all tables CICR stands for confidence interval coverage rate.

Table 1: Simulation results of study 1, first case

Sample size	Estimator of Y	Rel.bias of $\hat{Y}$	RMSE of $\hat{Y}$	Rel.bias of $\hat{V}ar\left(\hat{Y}\right)$	RMSE of $\hat{V}ar\left(\hat{Y}\right)$)	CI CR
200	$\hat{Y}_{2LC}$	–0.23%	194	–25%	10627	86
	$\hat{Y}_{2step}$	–0.35%	203	03(01)%	29915	94
400	$\hat{Y}_{2LC}$	–0.16%	130	–35 %	5999	68
	$\hat{Y}_{2step}$	–0.17 %	131	09 (19) %	4950	82
600	$\hat{Y}_{2LC}$	–0.09%	103	–17 %	2082	85
	$\hat{Y}_{2step}$	–0.11%	106	06 (15)%	4325	84

Table 2: Simulation results of study 1, second case

Sample size	Estimator of Y	Rel.bias of $\hat{Y}$	RMSE of $\hat{Y}$	Rel.bias of $\hat{V}ar\left(\hat{Y}\right)$	RMSE of $\hat{V}ar\left(\hat{Y}\right)$	CI CR
200	$\hat{Y}_{2LC}$	–0.09%	188	–32	15184	61
	$\hat{Y}_{2step}$	–0.19%	188	–17(–21)%	13978	64
400	$\hat{Y}_{2LC}$	–0.14 %	123	–36%	6011	84
	$\hat{Y}_{2step}$	–0.14 %	124	–06(–07) %	4469	90
600	$\hat{Y}_{2LC}$	–0.12%	99	–19%	2550	84
	$\hat{Y}_{2step}$	–0.14%	99	–03(–03)%	2487	90

Tables 1–2 present the results of the first simulation study, which is based

on real survey data. The results suggest that the two-step estimator $\hat{Y}_{2step}$ is almost unbiased, having generally slightly larger Rel.bias and RMSE than the benchmark. With regard to variance estimators, the results indicate that the Rel.bias of the corrected $\hat{V}_{cor}$ and uncorrected $\hat{V}_{uncor}$ variance estimators are low compared with the benchmark. In Table 1, the biases of these variance estimators are positive while those of the benchmark variance estimator are negative. In Table 2, all variance estimators have negative biases. In Table 1, the RMSE of $\hat{V}_{cor}$ is larger than that of the benchmark, except when the sample size (n) is 400, while in Table 2, $\hat{V}_{cor}$ has smaller RMSE values for all sample sizes. The tables also show that $\hat{V}_{cor}$ has a smaller absolute relative bias than does $\hat{V}_{uncor}$, except in Table 1 for $n = 200$ and in Table 2 for $n = 600$. In Table 2, the Rel.bias values of $\hat{V}_{cor}$ and $\hat{V}_{uncor}$ are decreasing in absolute values and converging to the same level. These properties are not observed in Table 1, however. The estimated confidence interval coverage rates (CICR) are generally larger for $\hat{Y}_{2step}$ than the benchmark, increasing for both estimators with increasing sample size, but are less than 95%.

6.2.2 Results of study 2

The results of the second simulation study are shown in Tables 3–4.

Table 3: Simulation results of study 2, first case

Sample size	Estimator of Y	Rel.bias of $\hat{Y}$	RMSE of $\hat{Y}$	Rel.bias of $\hat{Var}\left(\hat{Y}\right)$	RMSE of $\hat{Var}\left(\hat{Y}\right)$	CI CR
200	$\hat{Y}_{2LC}$	–0.22%	70	–17 %	1090	52
	$\hat{Y}_{2step}$	–0.67%	71	–09(–04)%	1307	67
400	$\hat{Y}_{2LC}$	–0.15%	50	–18%	529	87
	$\hat{Y}_{2step}$	–0.30%	50	–14 (–01)%	474	85
600	$\hat{Y}_{2LC}$	–0.08%	39	–15%	261	88
	$\hat{Y}_{2step}$	–0.15%	40	–19 (–24)%	342	88

Tables 3–4 present the results of the second simulation study based on simulated data. As in the former study, the two-step estimator $\hat{Y}_{2step}$ is almost unbiased but presenting slightly larger Rel.bias (except in Table 4 when $n = 400$) than the benchmark estimator. Regarding the variance estimators, Table 4 also shows that the Rel.bias of the corrected $\hat{V}_{cor}$ and uncorrected $\hat{V}_{uncor}$ variance estimators are low compared with the benchmark and tend

Table 4: Simulation results of study 2, second case

Sample size	Estimator of Y	Rel.bias of $\hat{Y}$	RMSE of $\hat{Y}$	Rel.bias of $\hat{Var}\left(\hat{Y}\right)$	RMSE of $\hat{Var}\left(\hat{Y}\right)$	CI CR
200	$\hat{Y}_{2LC}$	–0.04%	83	–25%	2045	80
	$\hat{Y}_{2step}$	–0.33%	84	–09 (–17)%	5317	81
400	$\hat{Y}_{2LC}$	–0.13%	63	–33%	1338	82
	$\hat{Y}_{2step}$	–0.27%	59	–07 (–14)%	909	88
600	$\hat{Y}_{2LC}$	0.13%	46	–19%	442	91
	$\hat{Y}_{2step}$	0.07%	47	–06(–09)%	450	91

to decrease in absolute value with increasing sample size. Furthermore, the relative biases of these variance estimators tend to converge to the same level. Table 4 also shows that the RMSE is larger for $\hat{V}_{cor}$ than for $\hat{V}ar(\hat{Y}_{2LC})$, except when $n = 400$, which is the same behavior in Table 3. The estimated coverage rates for $\hat{Y}_{2step}$ are generally not less than the benchmark and, for both estimators, tend to increase with increasing sample size, but remain less than 95%.

7 Discussion

Above we present the illustrative results of the two-step calibration estimator $\hat{Y}_{2step}$. The results are based on two simulation setups, one based on data from a real estate survey, the other based on simulated data. The results given in Tables 1–4 indicate that $\hat{Y}_{2step}$ have very low bias levels, however, tends to have a slightly larger bias than $\hat{Y}_{2LC}$, except when $n = 600$ in Table 4, in which case the sign of the bias is positive. The slightly larger bias for $\hat{Y}_{2step}$ than $\hat{Y}_{2LC}$ may be because $\mathbf{z}_k$ is reused in the second step of the $\hat{Y}_{2LC}$ estimator, while the estimator $\hat{Y}_{2step}$, uses it only in the first step. One alternative is to reuse $\mathbf{z}_k$ in the second step of estimation, which we expect to further reduce the bias of $\hat{Y}_{2step}$. The RMSE values for $\hat{Y}_{2LC}$ and $\hat{Y}_{2step}$ are generally comparable. To assess the role of the auxiliary information used here, we have also calculated the expansion estimator, $\hat{Y}_{Exp}$ (Särndal and Lundström 2005, p. 68), obtaining relative biases of –7% and –8% for the first and second studies, respectively. These relative biases are much larger than those obtained with the two-step estimators under consideration.

In virtually all tables, the Rel.bias of $\hat{V}_{cor}$ is smaller in absolute value

than the benchmark, except in Table 3 when $n = 600$. The Rel.bias of $\hat{V}_{cor}$ is positive in Table 1 and negative in others, this inconsistency is associated with some very small probability estimates producing very large weights that influence the estimated entities. When the benchmark is at least moderately correlated with the model variable, the Rel.bias of $\hat{V}_{cor}$ tends to decrease in absolute value with increasing sample size. The properties mentioned above are no longer observed when the correlation between benchmark and model variables is low. Another indicator of the performance of the suggested variance is the estimated confidence interval coverage rate, which suggests that our proposed variance estimator works well, as it generally leads to a coverage rate that is no less than that of the benchmark estimator. In Tables 2 and 4, the coverage rates increase with decreasing Rel.bias of $\hat{V}_{cor}$.

Regarding the question in the Remark, the results indicate that, with correlated model and benchmark variables, it is worth correcting for the uncertainty in model parameter estimation for small sample sizes in which $\hat{V}_{cor}$ tends to have a smaller bias than does $\hat{V}_{uncor}$. In large samples, the differences between $\hat{V}_{uncor}$ and $\hat{V}_{cor}$ are small. With low correlation between model and benchmark variables, it is not clear whether or not this correction is important, as we can see in Tables 1 and 3 that some situations favour $\hat{V}_{cor}$ while others favour $\hat{V}_{uncor}$.

The overal conclusion is that inferences will be reasonably valid when good benchmarks are available and not too small samples are considered.

References

Brick, M. (2013) Unit Nonresponse and Weighting Adjustments: A Critical Review. *Journal of Official Statatistics*, **29**, 329– 353.

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

Deville, J. C. and Särndal, C. E. (1992) Calibration estimators in survey sampling. *Journal of the American Statatistical Association*,**87**, 376–382

Deville, J. C., Särndal, C.-E. and Sautory, O.(1993) Generalized raking procedures in survey sampling.*Journal of the American Statatistical Association*,**88**, 1013–1020

Estevão, V. M. and Särndal, C.-E. (2002) The Ten Cases of auxiliary Information for Calibration in Two-Phase Sampling. *Journal of Official Statatistics*,**18**, 233–255

Estevão, V. M. and Särndal, C.-E. (2006) Survey Estimates by Calibration on Complex Auxiliary Information. *International Statistical Review*,**74**, 127–147

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

Kim, J. K. and Park, M. (2010) Calibration Estimation in Survey Sampling. *International Statistical Review*,**78**, 21–39. doi: 10.1111/j.1751-5823.2010.00099.x

Kott, P. S. and Day, C. D. (2014). Developing Calibration Weights and Standard Error Estimates for a Survey of Drug-Related Emergency-Departm-ent Visits. *Journal of Official Statatistics*, **30**, 521–532.

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

Kreuter, F. and Olson, K. (2011) Multiple auxiliary variables innonresponse adjustment.*Sociological Methods & Research*,**40**, 311–332

Lehtonen, R., Särndal C.-E., and Veijanen, A. (2008) Generalized regression and model-calibration estimation for domains. Invited paper, NORDSTAT 2008 Conference, Vilnius, June 2008.

Lehtonen, R. and Veijanen, A. (2012). Small area poverty estimation by model calibration. *Journal of the Indian Society of Agricultural Statistics*, **66**, 125133.

Lehtonen, R. and Veijanen, A. (2015). Estimation of poverty rate for small areas by model calibration and "hybrid" calibration methods. Retrieved from http://dx.doi.org/10.2901/EUROSTAT.C2015.001.

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

Niyonsenga, T. (1997) Response probability estimation. *Journal of Statistical Planning and Inference*, **59**, 111-126

Rizzo, L., Kalton, G. and Brick, M. (1996) A comparison of some weighting adjustment ment methods for panel nonresponse. *Survey Methodolology*, **22**, 43–53

Rota, B. J. and Laitila, T. (2015) Comparisons of some weighting methods for nonresponse adjustment. *Lithuanian Journal of Statistics*, **54**, 69–83

Rueda, M., Snchez-Borrego, I., Arcos, A., and Martnez, S. (2010). Model-calibration estimation of the distribution function using nonparametric regression, *Metrika* **71**, 33–44. DOI 10.1007/s00184-008-0199-y

Särndal, C.-E. and Lundström, S. (2005) *Estimation in surveys with nonresponse*. Wiley, New York.

Särndal, C.-E. and Lundström, S. (2007) Assessing auxiliary vectors for control of nonresponse bias in the calibration estimator. *Journal of Official Statistics*, **24**, 167–191

Särndal, C.-E., Swensson, B. and Wretman, J. (1992) *Model assisted survey sampling*. Springer, New York.

Singh, A. C., Wu, S. and Boyer, R. (1995) Longitudinal survey nonresponse adjustment by weight calibration for estimation of gross flows. *In JSM Proceedings, Survey research methods section. American Statistical Association, Alexandria, VA*, pp. 390–396. Retrieved from http://www.amstat.org/sections/srms/proceedings

Wu, C. and Sitter, R.R. (2001) A model-calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association* **96**, 185–193

Appendix

From (9)we have that: $Var\left(\hat{Y}^{\bullet}_{2step}\right) = Var(\hat{T}^{\circ}_a) + Var(\hat{T}^{\circ}_b) + 2Cov(\hat{T}^{\circ}_b, \hat{T}^{\circ}_a)$, where
$\hat{T}^{\circ}_a = \sum_{k\epsilon s} R_k d_k F^{\circ}_k E_k$, $\hat{T}^{\circ}_b = \sum_{k,l\epsilon s} R_k (R_l F^{\circ}_l - 1) A_{kl}$
and

$A_{kl} = d_k d_l \mathbf{z}^t_l \left(\mathbf{F}^{\circ}_{1k} \mathbf{\Gamma}^{-1}\right)^t E_k$.

From Särndal et al. (1999), $Var(\hat{T}_w) = E_p V_q(\hat{T}_w) + V_p E_q(\hat{T}_w)$ with $\hat{T}_w$ standing for $\hat{T}^{\circ}_a$ or $\hat{T}^{\circ}_b$.
Then,
$Var(\hat{T}^{\circ}_b) = V_p E_q(\sum_{k,l\epsilon s} R_k (R_l F^{\circ}_l - 1) A_{kl}) + E_p V_q(\sum_{k,l\epsilon s} R_k (R_l F^{\circ}_l - 1) A_{kl})$
where
$V_p E_q(\sum_{k,l\epsilon s} R_k (R_l F^{\circ}_l - 1) A_{kl}) = V_p(\sum_{k\epsilon s} \frac{1-F^{\circ}_k}{F^{\circ}_k} A_{kk}) = \sum_{k\neq l\epsilon U} \frac{\pi_{kl}-\pi_k\pi_l}{F^{\circ}_k F^{\circ}_l}(F^{\circ}_k - 1)(F^{\circ}_l - 1) A_{kk} A_{ll} + \sum_{k\epsilon U} \frac{\pi_k(1-\pi_k)(F^{\circ}_k-1)^2}{(F^{\circ}_k)^2} A^2_{kk}$
and
$E_p V_q(\sum_{k,l\epsilon s} R_k (R_l F^{\circ}_l - 1) A_{kl}) = E_{pq} \sum_{k,l,i,j\epsilon U} (M_{kl} - E_q(M_{kl}))(M_{ij} - E_q(M_{ij}))$
with $M_{ab} = I_a I_b R_a (R_b F^{\circ}_b - 1) A_{ab}$. This leads to
$E_p V_q(\sum_{k,l\epsilon s} R_k (R_l F^{\circ}_l - 1) A_{kl}) = S_1 + S_2 + S_3 + 2S_4 + S_5$, with
$S_1 = \sum_{k\neq l\neq i\epsilon U} \frac{\pi_{kli}}{F^{\circ}_k F^{\circ}_i}(F^{\circ}_l - 1) A_{kl} A_{il}$, $S_2 = \sum_{k\neq l\epsilon U} \frac{\pi_{kl}}{F^{\circ}_k}(F^{\circ}_l - 1) A^2_{kl}$,
$S_3 = \sum_{k\neq l\epsilon U} \frac{\pi_{kl}}{F^{\circ}_k F^{\circ}_l}(1-F^{\circ}_k)(1-F^{\circ}_l) A_{kl} A_{lk}$, $S_4 = \sum_{k\neq l\epsilon U} \frac{\pi_{kl}}{F^{\circ}_k F^{\circ}_l}(1-F^{\circ}_l)^2 A_{kl} A_{ll}$,
and $S_5 = \sum_{k\epsilon U} \frac{\pi_k}{(F^{\circ}_k)^2}(F^{\circ}_k - 1)^3 A^2_{kk}$.

Where S_1 is for $l = j$, S_2 for $l = j$ and $k = i$, S_3 for $k = j$ and $l = i$, S_4 for $l = i = j$ and $k = l = j$, S_5 for $k = l = i = j$, and zero for other index combinations.

$Cov(\hat{T}^{\circ}_b, \hat{T}^{\circ}_a) =$
$E_{pq}(\sum_{k,l\epsilon s} R_k (R_l F^{\circ}_l - 1) A_{kl} \sum_{i\epsilon s} R_i d_i F^{\circ}_i E_i) - E_{pq}(\sum_{k,l\epsilon s} R_k (R_l F^{\circ}_l - 1) A_{kl}) \sum_{i\epsilon U} E_i =$
$E_{pq}(\sum_{k,l,i\epsilon U} d_i I_k I_l I_i R_k R_l R_i F^{\circ}_l F^{\circ}_i A_{kl} E_i - d_i I_k I_l I_i R_k R_i F^{\circ}_i A_{kl} E_i) -$
$E_{pq}(\sum_{k,l\epsilon U} I_k I_l R_k A_{kl}(R_l F^{\circ}_l - 1) \sum_{i\epsilon U} E_i) = C_1 + C_2 + C_3$.
where $C_1 = \sum_{k\neq i\epsilon U} \frac{d_i \pi_{ki}(F^{\circ}_k - 1)}{F^{\circ}_k} A_{kk} E_i - \sum_{k\epsilon U} \frac{\pi_k (F^{\circ}_k - 1)}{F^{\circ}_k} A_{kk} \sum_{i\epsilon U} E_i$ when $k = l$;
$C_2 = \sum_{k\neq l\epsilon U} \frac{d_l \pi_{kl}(F^{\circ}_l - 1)}{F^{\circ}_k} A_{kl} E_l$, when $l = i$; and $C_3 = \sum_{k\epsilon U}(F^{\circ}_k - 1) A_{kk} E_k$ when $k = l = i$. Note that $\sum_{i\epsilon U} E_i = E_{pq}(\hat{T}^{\circ}_a)$.

CALIBRATING ON PRINCIPAL COMPONENTS IN THE PRESENCE OF MULTIPLE AUXILIARY VARIABLES FOR NONRESPONSE ADJUSTMENT

Bernardo João Rota [1]
Örebro University, Eduardo Mondlane University
e-mail: *bernardo.rota@oru.se, bernardo.rota@uem.mz*

Thomas Laitila
Örebro University, Statistics Sweden

PAPER III

Key words: Weighting, Nonresponse, Calibration, Principal components.

Abstract: Nonresponse is a major impediment to valid inference in sample surveys. In the nonresponse scenario, the driver of successful estimation is the efficient use of available auxiliary information. As electronic devices provide considerable data storage capacities, at the estimation stage it is natural for survey statisticians to face large datasets of auxiliary variables. It is unwise to use all available data as doing so may lead to poor estimators, especially if some variables are strongly correlated. Furthermore, selecting a subset of available auxiliary variables may not be the best alternative given the issues related to selection criteria. In this paper, we propose reducing the dimensions of the original set of auxiliary variables by using principal components. The use of principal components in place of the original auxiliary variables is evaluated via two calibration approaches, linear calibration using no explicit response model and propensity calibration of a known response model. For the latter, we propose selecting components based on their canonical correlation with the model variables. The results of two simulation studies suggest that using principal components is appropriate, as it offers the great advantage of reducing the computational burden.

1. Introduction

When adjusting for nonresponse in sample surveys, auxiliary information plays a prominent role in successful estimation. Rizzo, Kalton and Brick (1996) note that, providing it is carefully chosen, the particular adjustment scheme used at the estimation stage is not that important. The relation with the study variable or response pattern is usually taken as a benchmark in the choice of auxiliary variables (see Kreuter and Olson, 2011; Särndal and Lundström, 2005, p. 110).

Calibration estimation (Deville and Särndal, 1992), initially designed to reduce sampling error in surveys with complete response, was eventually extended to surveys affected by nonresponse, (see

[1]Corresponding author.

AMS: 62GXX, 62FXX, 62DXX

e.g. Lundström and Särndal, 1999; Kott, 2006). The method relies on an efficient choice of auxiliary variables.

When many auxiliary variables are available, calibrating on all of them may lead to 'over-calibration', the term used by Guggemos and Tillé (2010). According to Särndal and Lundström (2005), a problem may arise when the candidate auxiliary vector contains variables likely to cause multicollinearity or variables with highly skewed distributions. These problems may result in a very inefficient estimator almost less efficient than, for example, the Horvitz-Thompson estimator (Cardot, Goga and Shehzad, 2015).

Large sets of auxiliary variables have also been considered by many authors in various estimation settings, as in the following examples, Bardsley and Chambers (1984) propose a ridge-type estimator in the context of model-based estimation, an approach that relaxes the principle that the calibration weights 'exactly' reproduce the totals of known characteristics by holding only 'approximately'. Guggemos and Tillé (2010) introduce a penalized calibration estimator. Bilen, Khan and Yadav (2004) suggest a principal component approach for reducing the multicollinearity and dimensions of the auxiliary variables in a regression context. Cardot et al. (2015) propose calibration on reduced data via principal components (PCs) in surveys with complete response.

Variable selection criteria are also suggested in the literature as an alternative way to deal with large sets of auxiliary variables and related problems. McHenry (1978) suggests an algorithm to select the best subset of auxiliary variables in the context of multiple regression or multivariate analysis. Silva and Skinner (1997) suggest a selection criterion based on the variability of the regression estimator. Särndal and Lundström (2007) propose a selection device based on the variability of estimated inverse propensities determined under the assumption that the auxiliary variables satisfy some pre-specified condition. The variable selection is conditioned on an increase in the variability of the inverse propensities. A potential auxiliary variable must predict the key survey variables and the propensities to respond. Geuzinge, Rooijen and Bakker (2000) propose a selection indicator based on the product of (a) the correlation between the auxiliary vector and the study variables and (b) the correlation between the auxiliary vector and the response propensity. When adjusting for nonresponse through regression estimation, Bethlehem and Schouten (2004) and Schouten (2007) propose a selection based on minimizing the maximal absolute bias of the estimator; the method relies on computing an interval for the maximal absolute bias and selecting those variables that minimize its width.

The common practice of using a subset of the full set of potential auxiliary variables and discarding others may result in the loss of important information. For example, in a regression context, it is known that the R^2 tends to decrease with the removal of regressors from the regression equation. This phenomenon can be interpreted in many ways, but in some cases is due to the loss of valuable information. Furthermore, most of the suggested selection algorithms are computationally intensive and, impractical for large sets of candidate auxiliary variables.

In this paper, we calibrate on reduced data via principal components. Thus, we account for the exponential growth in computing time due to dimensionality in the auxiliary data and most importantly, the problem of large weights due to outliers is also accounted turning the estimator more efficient. The idea was initially suggested by Cardot et al. (2015) in surveys with complete response, and we extend it to estimation in surveys affected by nonresponse. Furthermore, the ideas in Cardot et al. (2015) are centered on the Greg-type-calibration (the complete response linear

calibration), while we study this and the propensity score calibration estimators in the nonresponse context. Note that the use of principal components in weighting does not stand for data interpretation, but is a tool for alleviating the problem of managing high-dimensional auxiliary data. Specifically, the PCs approach assists in the construction of new auxiliary variables from the original variables by taking into account all available candidate variables through linear combinations. Furthermore, we implement a rejection of PCs based on their canonical correlation (Hotelling, 1939) with the model variables.

Two calibration estimators are considered in the paper:

1. Linear calibration (LC) using no explicit form of response model Särndal and Lundström (2005).

2. Instrumental variable or propensity score calibration (PSC) with an explicit form of response model (Chang and Kott, 2008).

This suggests two sources of auxiliary information for estimation: an $\mathbf{X}_{(N\times P)}$ data matrix carrying information on the N population elements of a P-dimensional vector of auxiliary variables and an $\mathbf{H}_{(m\times L)}$ data matrix carrying information on the m respondent elements of an L-dimensional vector of instrumental variables. The LC estimator uses only the first source of auxiliary information, while the PSC combines the two sources.

The rest of the article is organized as follows: section 2 provides background information on calibration estimators for nonresponse adjustment; section 3 provides a summary theoretical framework on principal components; section 4 provides a theoretical combination of calibration estimators and principal components; section 5 provides numerical support for section 4; and the final section discusses the results.

2. Calibration Estimators

Define a finite population, U, of distinguishable units indexed by integers $1,2,...,k,...,N$. A probability sample, s, of distinguishable elements indexed by integers $1,2,...,k,...,n$ is drawn from U according to a probability sampling design, $p(s)$, yielding the first- and second-order inclusion probabilities, $\pi_k = P(k \in s) > 0$ and $\pi_{kl} = P(k\&l \in s) > 0$, respectively for all $k,l \in \{1,2,...,N\}$, where $\pi_{kk} = \pi_k$. Suppose that data are observed for subset $r \subset s$ with $|r| = m$. The elements of r are assumed to be generated by a random process, $q(r)$, on s. Thus, each element $k \in r$ is associated with probability $\theta_k = P(k \in r|k \in s) > 0$. The random process $q(r)$ on a given s is usually termed a response mechanism, while θ_k is the response probability for the individual k. Here, it is assumed that events $k \in r$ and $l \in r$ for a given s are independent of one another given that $k \neq l$.

Calibration estimators were introduced by Deville and Särndal (1992) in the context of surveys with complete response; the approach was then extended to surveys affected by nonresponse. In this context, Särndal and Lundström (2005) define the calibration estimator for total $t_y = \sum_U y_k$ as,

$$\hat{t}_{ycal} = \mathbf{w}_{(r)}^t \mathbf{y}_{(r)} \tag{1}$$

where $\mathbf{w}_{(r)} = vec\{w_k\}^m$ and $\mathbf{y}_{(r)} = vec\{y_k\}^m$ are m-dimensional column vectors of calibrated weights w_k and study variable values y_k respectively. The term 'calibrated weights' means that the weights

satisfy the calibration property $\mathbf{X}'_{(r)}\mathbf{w}_{(r)} = \mathbf{T}_x$, where $\mathbf{T}_x = \sum_U \mathbf{X}_k$ and $\mathbf{X}_k$ being the transpose of the k^{th} line of $\mathbf{X}_{(N\times P)}$. Calibrated weights, w_k, are constructed to be as close as possible to the reciprocals of the sample inclusion probabilities, $d_k = 1/\pi_k$, according to a distance metric $\Omega\left(\mathbf{w}_{(r)};\mathbf{d}_{(r)}\right)$, while satisfying the above calibration property. Using Lagrange reasoning, calibrated weights can be derived by minimizing $\Omega(\mathbf{w}_{(r)};\mathbf{d}_{(r)}) + \gamma'\left(\mathbf{T}_x - \mathbf{X}'_r\mathbf{w}_{(r)}\right)$, where γ is a column vector of Lagrange multipliers, $\mathbf{d}_{(r)} = vec\{d_k\}^m$. The resulting calibrated weights take the form

$$w_k = d_k h(\gamma' \mathbf{X}_k) \tag{2}$$

where $d_k h_k = \psi^{-1}(\cdot, d_k)$, $\psi = \partial\Omega/\partial w$, given the assumptions in Deville and Särndal (1992).

A different choice of Ω leads to a different weight system (2). Deville and Särndal (1992) establish conditions under which any choice of distance function leads to estimators that are asymptotically equivalent to the regression estimator obtained through a Chi-square-type distance measure. Thus, the choice of distance measure may be influenced by the computational aspects or other properties of w_k, such as its non-negativity or degree of stability.

Using the Chi-square distance, i.e., $\Omega(\mathbf{w}_{(r)};\mathbf{d}_{(r)}) = \left(\mathbf{w}_{(r)} - \mathbf{d}_{(r)}\right)^t (2\mathbf{D})^{-1}\left(\mathbf{w}_{(r)} - \mathbf{d}_{(r)}\right)$, with $\mathbf{D} = diag\{d_1, d_2, ..., d_k, ..., d_m\}$, leads to the linear calibrated weights of the form

$$w_k = d_k + d_k \gamma' \mathbf{X}_k \tag{3}$$

where $\gamma = \left(\mathbf{X}'_{(r)}\mathbf{D}\mathbf{X}_{(r)}\right)^{-1}\left(\mathbf{T}_x - \mathbf{X}'_{(r)}\mathbf{d}_{(r)}\right)$.
The linear calibration estimator for t_y is:

$$\hat{t}_{ycal} = \mathbf{w}'_{(r)}\mathbf{y}_{(r)} = \mathbf{d}'_{(r)}\mathbf{e}_{(r)} + \mathbf{T}'_x\left(\mathbf{X}'_{(r)}\mathbf{D}\mathbf{X}_{(r)}\right)^{-1}\mathbf{X}'_{(r)}\mathbf{D}\mathbf{y}_{(r)} \tag{4}$$

where, $\mathbf{e}_{(r)} = vec\{e_k\}^m$ and $\mathbf{y}_{(r)} = vec\{y_k\}^m$ are m-dimensional column vectors of residuals $e_k = y_k - \hat{y}_k$ and study variable values y_k respectively, and $\hat{y}_k = \mathbf{X}'_k\left(\mathbf{X}'_{(r)}\mathbf{D}\mathbf{X}_{(r)}\right)^{-1}\mathbf{X}'_{(r)}\mathbf{D}\mathbf{y}_{(r)}$.

In the complete response context, estimator (4) is equivalent to the GREG estimator (Särndal, Swensson and Wretman, 1992) derived under superpopulation model ξ, which assumes a linear relationship between the survey variable, y_k, and the auxiliary vector, $\mathbf{X}_k$, given by $\xi : y_k = \beta'\mathbf{X}_k + \varepsilon_k$. Since, $\mathbf{X}'_{(s)}\mathbf{d}_{(s)}$ is unbiased for $\mathbf{T}_x$, the weights (3) are in average equal to d_k which leads to zero average differences $y_k - \hat{y}_k$.

3. A brief summary of principal components

Suppose that $\mathbf{X}$ is defined as in Section 1 except that each $\mathbf{X}_j, j = 1, ..., P$ is rescaled to zero mean and unit variance, then, $\mathbf{X}'\mathbf{X}$ is the covariance matrix of $\mathbf{X}$. Let $(\lambda_j, \mathbf{b}_j; j = 1, ..., P)$ be eigenvalue-eigenvector pairs of $\mathbf{X}'\mathbf{X}$. The j^{th} principal component is given by $\mathbf{Z}_j = \mathbf{b}'_j\mathbf{X} = \sum_{l=1}^P \mathbf{b}_{lj}\mathbf{X}_l$ with the properties $cov(\mathbf{Z}_j, \mathbf{Z}_i) = \begin{cases} 0, & j \neq i \\ \lambda_i, & j = i \end{cases}$, $\mathbf{b}_j$ is a P-dimensional column vector and the $\lambda_i, i = 1, ..., P$ satisfy $\lambda_1 \geq \lambda_2 \geq, ..., \geq \lambda_P \geq 0$. The proportion of total variance accounted for by the first $R < P$ principal components is given by $\left(\sum_{i=1}^R \lambda_i / \sum_{i=1}^P \lambda_i\right) \times 100\%$.

Suppose now that $\mathbf{X} = \mathbf{X}_{(\mathbf{s})}$, that is, auxiliary data observed only at sample level. The covariance matrix of $\mathbf{X}_{(\mathbf{s})}$ is estimated without bias by $\mathbf{X}^t\mathbf{D}\mathbf{X}$, where $\mathbf{D} = diag\{d_1, d_2, ..., d_k, ..., d_n\}$. The estimated principal components are given by $\hat{\mathbf{Z}}_j = \hat{\mathbf{b}}_j^t \mathbf{X}_{(\mathbf{s})} = \sum_{l=1}^{P} \hat{\mathbf{b}}_{lj} \mathbf{X}_{l(s)}$. The pair $(\hat{\lambda}_j, \hat{\mathbf{b}}_j; j = 1, ..., P)$ comprise the eigenvalue and eigenvector of $\mathbf{X}^t\mathbf{D}\mathbf{X}$.

4. Calibrating on principal components

The calibration estimator in the principal components setting can be derived by solving the following problem:

$$\min \Omega(\mathbf{w}_{(r)}^{pc}; \mathbf{d}_{(r)}) sub: \ \mathbf{Z}_{(r)}^t \mathbf{w}_{(r)}^{pc} = \mathbf{T}_z \ , \tag{5}$$

4.1. The linear calibration estimator based on principal components

If we follow the same reasoning that led to weights (3), we will then arrive at principal components calibrated weights given by

$$\mathbf{w}_k^{pc} = d_k + d_k \boldsymbol{\gamma}_{(pc)}^t \mathbf{Z}_k \tag{6}$$

where, $\boldsymbol{\gamma}_{(pc)} = \left(\mathbf{Z}_{(r)}^t \mathbf{D} \mathbf{Z}_{(r)}\right)^{-1} \left(\mathbf{T}_z^t - \mathbf{Z}_{(r)}^t \mathbf{d}_{(r)}\right)$ and $\mathbf{Z}_k = \{Z_{k1}, Z_{k2}, ..., Z_{kR} | R < P\}$ is the vector whose elements are the retained components. The nonresponse principal-components-based calibration estimator for t_y is given by

$$\hat{t}_{ycal(pc)} = \mathbf{d}_{(r)}^t \mathbf{e}_{(r)}^{pc} + \mathbf{T}_z^t \left(\mathbf{Z}_{(r)}^t \mathbf{D} \mathbf{Z}_{(r)}\right)^{-1} \mathbf{Z}_{(r)}^t \mathbf{D} \mathbf{y}_{(r)} \tag{7}$$

where $\mathbf{e}_{(r)}^{pc} = vec\left\{ y_k - \mathbf{Z}_k^t \left(\mathbf{Z}_{(r)}^t \mathbf{D} \mathbf{Z}_{(r)}\right)^{-1} \mathbf{Z}_{(r)}^t \mathbf{D} \mathbf{y}_{(r)} \right\}^r$.

4.2. The propensity score calibration based on principal components

Consider a framework of unit response resulting according to a known parametric model, $\phi^{-1}(\cdot; \mathbf{H}_k)$. Observe that this model is known only up to an unknown L-dimensional vector of parameters, $\delta = \delta^*$, where $\delta \in \Upsilon$, $\dim(\mathbf{H}_k) = L \leqslant R$ and R is the number of selected PCs. Then, the model parameters can be estimated from the calibration constraint below (see Kott, 2012).

$$\mathbf{Z}_{(r)}^t \Phi(\delta) \mathbf{d}_{(r)} - \mathbf{T}_z = \mathbf{0} \tag{8}$$

where $\dim(\mathbf{Z}_{(r)}) = m \times R$ and $\Phi(\delta) = diag\{\phi(\delta; \mathbf{H}_1), \phi(\delta; \mathbf{H}_2), ..., \phi(\delta; \mathbf{H}_k), ..., \phi(\delta; \mathbf{H}_m)\}$. This is a principle suggested by Chang and Kott (2008). The solution to (8) is the minimizer of the objective function:

$$\left(\mathbf{Z}_{(r)}^t \Phi(\delta) \mathbf{d}_{(r)} - \mathbf{T}_z\right)^t \mathbf{W}_n \left(\mathbf{Z}_{(r)}^t \Phi(\delta) \mathbf{d}_{(r)} - \mathbf{T}_z\right). \tag{9}$$

When $L = R$, the form of weighting matrix $\mathbf{W}_n$ of dimension $R \times R$ is irrelevant, as system (8) is just identified, otherwise $\mathbf{W}_n$ is a suitably chosen nonnegative definite matrix. Note that $\mathbf{Z}_k$ is an R-dimensional column vector of retained principal components of P originals. Under this setting, to make the system of equations (8) feasible, the minimal requirement is that the number of PCs in $\mathbf{Z}_k$ be at least L retained components.

Having estimated the response model parameter, δ^*, the calibration estimator for t_y (the propensity score calibration) is

$$\hat{t}_{PSC(pc)} = \sum_r d_k \phi(\hat{\delta}^t_{(pc)} Z_k) y_k \tag{10}$$

where $\hat{\delta}_{(pc)}$ is the estimated value of δ. To obtain $\hat{\delta}_{(pc)}$, we can follow the ideas by Beaumont (2006), who propose an iterative procedure based on the Taylor approximation of (8). This is similar to the procedure suggested by Binder (1983). We apply a slightly different perspective in the estimation of δ in (8).

Assume the following conditions to hold:

1. Function $\phi(\delta)$ is continuous and twice differentiable with respect to δ.
2. $E_{pq}\left(\mathbf{Z}^t_{(r)}\Phi(\delta)\mathbf{d}_{(r)} - \mathbf{T}_z\right) = \mathbf{0}$ if and only if $\delta = \delta^*$ for all $\delta \in \Upsilon$
3. Set Υ is a compact set .
4. $E_{pq}\left[\left(\mathbf{Z}^t_{(r)}\Phi(\delta)\mathbf{d}_{(r)} - \mathbf{T}_z\right)\left(\mathbf{Z}^t_{(r)}\Phi(\delta)\mathbf{d}_{(r)} - \mathbf{T}_z\right)^t\right]$ is finite
5. $\mathbf{Z}^t_{(r)}\Psi(\delta)\mathbf{H} = \frac{\partial}{\partial\delta}\left(\mathbf{Z}^t_{(r)}\Phi(\delta)\mathbf{d}_{(r)} - \mathbf{T}_z\right) = \sum_r d_k \phi_1(\mathbf{H}_k;\delta)\mathbf{Z}_k\mathbf{H}^t_k$ exists and is continuous in Υ, where $\phi_1(\mathbf{H}_k;\delta) = \partial\phi(\mathbf{H}_k;\delta)/\partial\delta$ and the $m \times m$ diagonal matrix $\Psi(\delta)$ has its k^{th} diagonal element given by $d_k\phi_1(\mathbf{H}_k;\delta)$
6. $\mathbf{Z}^t_{(r)}\Psi(\delta)\mathbf{H}$ is a full-column rank matrix.

Define the quadratic distance as follows:

$$\left(\mathbf{Z}^t_{(r)}\Phi(\delta)\mathbf{d}_{(r)} - \mathbf{T}_z\right)^t \frac{\mathbf{W}_n}{2}\left(\mathbf{Z}^t_{(r)}\Phi(\delta)\mathbf{d}_{(r)} - \mathbf{T}_z\right). \tag{11}$$

The solution to (8) is defined as the minimizer of objective function (11). In the generalized method of moments setting, minimizing (11) is equivalent to solving the set of estimating equations defined by

$$\left(\mathbf{Z}^t_{(r)}\Psi(\delta)\mathbf{H}\right)^t \mathbf{W}_n\left(\mathbf{Z}^t_{(r)}\Phi(\delta)\mathbf{d}_{(r)} - \mathbf{T}_z\right) = \mathbf{0}. \tag{12}$$

We use the following approximation:

$$\left(\mathbf{Z}^t_{(r)}\Phi(\delta^*)\mathbf{d}_{(r)} - \mathbf{T}_z\right) \approx \left(\mathbf{Z}^t_{(r)}\Phi(\hat{\delta}_{(pc)})\mathbf{d}_{(r)} - \mathbf{T}_z\right) + \left(\mathbf{Z}^t_{(r)}\Psi(\hat{\delta}_{(pc)})\mathbf{H}\right)(\delta^* - \hat{\delta}_{(pc)}). \tag{13}$$

Introducing equation (13) into (12) yields the following updating equation:

$$\hat{\delta}^1_{(pc)} \approx \hat{\delta}^0_{(pc)} + \left[\left(\mathbf{Z}^t_{(r)}\Psi^0\mathbf{H}\right)^t \mathbf{W}_n\left(\mathbf{Z}^t_{(r)}\Psi^0\mathbf{H}\right)\right]^{-1}\left(\mathbf{Z}^t_{(r)}\Psi^0\mathbf{H}\right)^t \mathbf{W}_n\left(\mathbf{Z}^t_{(r)}\Phi(\delta^0)\mathbf{d}_{(r)} - \mathbf{T}_z\right) \tag{14}$$

where $\Psi^0 = \Psi(\hat{\delta}^0_{(pc)})$. In (9), $\hat{\delta}_{(pc)}$ is the value of $\hat{\delta}^1_{(pc)}$ obtained upon convergence of (14).

In the appendix section we provide the derivation of the asymptotic variances of the estimated coefficients of the propensity functions when population- or sample-level auxiliary information is used. A comparison of these variances shows that sample-level auxiliary information provides more accurate estimated coefficients than population-level does.

4.3. Suggested retention criterion (a canonical correlation-based criterion)

Many authors have discussed PCs retention criteria, for example, Jollifé (1972), Cadima and Jollifé (1995), Jollifé, Trendafilov and Uddin (2003), and McCabe (1984), though there is no unified recommendation on this matter (Johnson and Wichern, 2007). Common practice is based on one or combinations of the following three criteria: the eigenvalue-one, scree plot, and proportion of total variance explained criteria. Mansfield, Webster and Gunst (1977) noted that it is common in PCs analysis for significant data variation to be accounted for by the first few components. According to these criteria, the components with small variability are excluded. Note, however that we are not concerned with interpreting PCs, instead using them as a tool for constructing new auxiliary variables that take into account all original candidate auxiliary variables.

In a canonical correlation setting, the goal is to determine sets of linearly independent vectors for two groups of variables that result in the maximum correlation between the projections of these variables onto the space spanned by these linearly independent vectors. According to Borga (2001), the correlation between two sets of multidimensional variables, if it exists, may be blurred if an inappropriate coordinate system is used to represent the variables. However, in canonical correlation, each of the two sets is linearly transformed, so that the corresponding pairs of coordinates of these transformed variables have the maximum correlation.

Recall that $\mathbf{H}$ is an $m \times L$ data matrix where $H_1, H_2, ..., H_L$ are the model variables and let, $\tilde{Z}$ be an $m \times D$ data matrix, where $1 \leq D \leq P$ is the number of principal component variables in $\tilde{Z}$. Let $\mathbf{P_H}$ be the projection of $\mathbf{H}$ onto the space spanned by linear combinations of its elements and suppose that $\mathbf{P}_{\tilde{\mathbf{Z}}}$ is the analogous projection of elements in $\tilde{Z}$. We want to approximate the correlation $(\tilde{\rho}_{\mathbf{H},\tilde{\mathbf{Z}}})$ of sets $\mathbf{H}$ and $\tilde{Z}$ by the canonical correlation defined by $\max_{\mathbf{P_H},\mathbf{P}_{\tilde{\mathbf{Z}}}} \Gamma\left(\mathbf{P_H}\mathbf{H}^t, \mathbf{P}_{\tilde{\mathbf{Z}}}\tilde{\mathbf{Z}}^t\right)$.

$$\tilde{\rho}_{\mathbf{H},\tilde{\mathbf{Z}}} \equiv \max_{\mathbf{P_H},\mathbf{P}_{\tilde{\mathbf{Z}}}} \Gamma\left(\mathbf{P_H}\mathbf{H}^t, \mathbf{P}_{\tilde{\mathbf{Z}}}\tilde{\mathbf{Z}}^t\right) = \max_{\mathbf{P_H},\mathbf{P}_{\tilde{\mathbf{Z}}}} \frac{\left[\mathbf{P_H}\left(\mathbf{H}^t\tilde{\mathbf{Z}}\right)\mathbf{P}^t_{\tilde{\mathbf{Z}}}\right]}{\left[\mathbf{P_H}\left(\mathbf{H}^t\mathbf{H}\right)\mathbf{P}^t_{\mathbf{H}}\right]^{1/2}\left[\mathbf{P}_{\tilde{\mathbf{Z}}}\left(\tilde{\mathbf{Z}}^t\tilde{\mathbf{Z}}\right)\mathbf{P}^t_{\tilde{\mathbf{Z}}}\right]^{1/2}} \tag{15}$$

We can equivalently reformulate (15) as

$$\begin{cases} \max\left[\mathbf{P_H}\left(\mathbf{H}^t\tilde{\mathbf{Z}}\right)\mathbf{P}^t_{\tilde{\mathbf{Z}}}\right] \\ sub \begin{cases} \left[\mathbf{P_H}\left(\mathbf{H}^t\mathbf{H}\right)\mathbf{P}^t_{\mathbf{H}}\right]^{1/2} = 1 \\ \left[\mathbf{P}_{\tilde{\mathbf{Z}}}\left(\tilde{\mathbf{Z}}^t\tilde{\mathbf{Z}}\right)\mathbf{P}^t_{\tilde{\mathbf{Z}}}\right]^{1/2} = 1 \end{cases} \end{cases} \tag{16}$$

Using Lagrange multiplier principle, (16) is solved by maximizing the objective function

$$\mathbf{L}(\mu_*, \mathbf{P}_*) = \left[\mathbf{P_H}\left(\mathbf{H}^t\tilde{\mathbf{Z}}\right)\mathbf{P}^t_{\tilde{\mathbf{Z}}}\right] - 2^{-1}\left(\mu_1\left[\mathbf{P_H}\left(\mathbf{H}^t\mathbf{H}\right)\mathbf{P}^t_{\mathbf{H}} - 1\right] - \mu_2\left[\mathbf{P}_{\tilde{\mathbf{Z}}}\left(\tilde{\mathbf{Z}}^t\tilde{\mathbf{Z}}\right)\mathbf{P}^t_{\tilde{\mathbf{Z}}} - 1\right]\right)$$

yielding the system of equations

$$\begin{cases} \frac{\partial L}{\partial \mathbf{P_H}} = \left(\mathbf{H}^t\tilde{\mathbf{Z}}\right)\mathbf{P}^t_{\tilde{\mathbf{Z}}} - \mu_1\left(\mathbf{H}^t\mathbf{H}\right)\mathbf{P}^t_{\mathbf{H}} = 0 \\ \frac{\partial L}{\partial \mathbf{P}_{\tilde{\mathbf{Z}}}} = \left(\mathbf{H}^t\tilde{\mathbf{Z}}\right)^t\mathbf{P}^t_{\mathbf{H}} - \mu_2\left(\tilde{\mathbf{Z}}^t\tilde{\mathbf{Z}}\right)\mathbf{P}^t_{\tilde{\mathbf{Z}}} = 0. \end{cases} \tag{17}$$

Premultiplying the first equation in (17) by $\mathbf{P_H}$ and subtracting $\mathbf{P}_{\tilde{\mathbf{Z}}}$ times the second equation from the first, results in $\mu_1\mathbf{P_H}\left(\mathbf{H}^t\mathbf{H}\right)\mathbf{P}^t_{\mathbf{H}} = \mu_2\mathbf{P}_{\tilde{\mathbf{Z}}}\left(\tilde{\mathbf{Z}}^t\tilde{\mathbf{Z}}\right)\mathbf{P}^t_{\tilde{\mathbf{Z}}}$, because $\mathbf{P_H}\left(\mathbf{H}^t\tilde{\mathbf{Z}}\right)\mathbf{P}^t_{\tilde{\mathbf{Z}}} = \left(\mathbf{P_H}\left(\mathbf{H}^t\tilde{\mathbf{Z}}\right)\mathbf{P}^t_{\tilde{\mathbf{Z}}}\right)^t$, where by $\mu_1 = \mu_2 = \mu$.

Assuming that $\mathbf{H}^t\mathbf{H}$ is invertible, the first equation gives

$$\mu \mathbf{P}^t_{\mathbf{H}} = \left(\mathbf{H}^t\mathbf{H}\right)^{-1}\left(\mathbf{H}^t\tilde{\mathbf{Z}}\right)\mathbf{P}^t_{\tilde{\mathbf{Z}}}. \tag{18}$$

After appropriate replacements in the second, we get $\left(\mathbf{H}^t\tilde{\mathbf{Z}}\right)^t\left(\mathbf{H}^t\mathbf{H}\right)^{-1}\left(\mathbf{H}^t\tilde{\mathbf{Z}}\right)\mathbf{P}^t_{\tilde{\mathbf{Z}}} - \mu^2\left(\tilde{\mathbf{Z}}^t\tilde{\mathbf{Z}}\right)\mathbf{P}^t_{\tilde{\mathbf{Z}}} = 0$, which is equivalent to writing this last equation as

$$Qx = \lambda Rx \tag{19}$$

where $Q = \left(\mathbf{H}^t\tilde{\mathbf{Z}}\right)^t\left(\mathbf{H}^t\mathbf{H}\right)^{-1}\left(\mathbf{H}^t\tilde{\mathbf{Z}}\right)$, $x = \mathbf{P}^t_{\tilde{\mathbf{Z}}}$, $\lambda = \mu^2$ and $R = \left(\tilde{\mathbf{Z}}^t\tilde{\mathbf{Z}}\right)$.

Equation (19) is in the form of a generalized eigenvalue equation (Parra and P. Sajda, 2003). Let, $R = MM^t$ be a Cholesky decomposition of R; then (19) becomes

$$\left(M^{-1}QM^{-1t}\right)M^t x = \lambda M^t x \Leftrightarrow \tilde{Q}\tilde{x} = \lambda\tilde{x},$$

which is the standard eigenvalue equation. Solving this, we obtain a solution for $\mathbf{P}_{\tilde{\mathbf{Z}}}$, which naturally leads to a solution for $\mathbf{P}_{\mathbf{H}}$ in (18). These solutions represent the optimal projections of the variables in $\{\tilde{\mathbf{Z}}\}$ and $\{\mathbf{H}\}$ onto spaces spanned by their respective linear combinations. The coordinate systems resulting from $\mathbf{P}_{\tilde{\mathbf{Z}}}$, and $\mathbf{P}_{\mathbf{H}}$ are mutually maximally correlated. See, for example, Borga (2001) and Hardoon, Szedmak and Shawe-Taylor (2004), for more insight on canonical correlation analysis.

Our PCs selection criterion is based on the value of the canonical correlation between the PCs and the instrumental variables. The PCs are selected in order of their appearance and the canonical correlations are used to measure the representativeness of the selected components. The canonical correlations are calculated in a forward stepwise manner: the first canonical correlation is the correlation between the instrumental vector and a vector comprinsing the first PC; the second canonical correlation is the maximal correlation between the instrument vector and the vector comprising the first two PCs, and so on. The values of these canonical correlations are obtained in an increasing order. The stopping rule is based on the amount by which this correlation increases from a previous step to the actual step. If the addition of a further component to the vector of PCs does not significantly change the correlation among these two groups, then that component and the remaining components are discarded from the final auxiliary vector.

Remark 1 Unlike $\mathbf{Z}^t\mathbf{Z}$, which is a diagonal matrix with eigenvalues of $\mathbf{X}^t\mathbf{X}$ being its diagonal elements, matrix $\tilde{\mathbf{Z}}^t\tilde{\mathbf{Z}}$ is no longer a diagonal since $\tilde{\mathbf{Z}}$ is made of elements of $\mathbf{Z}$ falling into response set r.

Remark 2 We maximize the relation $(\mathbf{H}, \tilde{\mathbf{Z}})$ rather than $(\mathbf{H}, \mathbf{Z})$ as the latter is impossible because information on $\mathbf{H}$ is assumed to be known at response level. The variables' distributions are generally distorted by nonresponse and the resulting correlation is expected to deviate from the true correlation. This is not of concern here, as the main goal is to guarantee at the response level selected auxiliary variables closely linked to the instruments.

5. Simulation Study

This section provides empirical illustrations of the points discussed in the previous sections. It is known, that the principal components data reduction approach is effective when the relations among

the variables involved are strong. In this article, we present two simulation studies: in the first study, the structure of correlation among the variables is very strong, the first principal component alone explaining more than 90% of the total data variation, as can be observed in Figure 1; in the second study, the structure of correlation among the variables is weak, and several components are needed to meaningfully explain the total variation of the data, as illustrated by, the scree plot shown in Figure 2. The data source for the first study is 'Unemployment and median household income for the U.S., States, and counties, 2006–2014' from the Unemployment – Bureau of Labor Statistics – LAUS data. The data are freely and publicly accessible for use at http://www.bls.gov/lau/. According to the source, 'the concepts and definitions underlying LAUS data come from the Current Population Survey (CPS), the household survey that is the official measure of the labor force for the nation. State monthly model estimates are controlled in real time to sum to national monthly labor force estimates from the CPS. These models combine current and historical data from the CPS, the Current Employment Statistics (CES) program, and State unemployment insurance (UI) systems'.

The data source for the second study is 'Small Area Income and Poverty Estimates (SAIPE)', which is a 1989, 1993, and 1995–2013 dataset, also freely and publicly accessible at https://www.census.gov/did/www/saipe/. According to the source, 'Small Area Income and Poverty Estimates (SAIPE) are produced for school districts, counties, and states. The main objective of this program is to provide updated estimates of income and poverty statistics for the administration of federal programs and the allocation of federal funds to local jurisdictions'.

5.1. Simulation setup

5.1.1. Study 1

From the data of the first study we selected 27 quantitative variables. We applied data transformation to induce the correlation among them to a desired pattern. The transformed variables are named v_1 to v_{27}. For example, from uncorrelated variables x_1 and x_2 we can generate new corresponding correlated variables $v_1 = x_1$ and $v_2 = sqrt(x_1 * x_2)$, respectively. From these 27 variables, two (v_1, v_5), were chosen to be the model variables, that is, the variables governing the response behaviour and one (v_{27}), was chosen to be the study variable y. The correlations between each model variable with the study variable are approximately 0.5. The remaining 24 were assumed to be auxiliary variables. These data correspond to our population of 3260 observations. The simulation process was the following:

1. From this population, we draw a sample of size 300 according to a simple random sampling without raplacement.
2. A response set was generated using a logistic regression model $p_k = 1/(1 + exp(-\delta^t \mathbf{H}_k))$, where $\mathbf{H} = \{1, v_1, v_5\}^t$ is a vector of model variables whereas δ is a vector of model parameters. The elements $k\varepsilon s$ for which a Bernoulli trial was 1 with probability p_k, were selected to the reponse set.
3. Estimates of interest were calculated using the data in the respose set.
4. The process was repeated 1000 times. Higher replication numbers basically produced similar results.

5. Indicators of the properties of the estimators were calculated. These indicators are the relative bias ($Rel.bias = \frac{bias(\hat{\theta})}{\theta} * 100\%$), the standard error ($S.E. = sqrt(var(\hat{\theta}))$), and the root mean squared error ($RMSE = sqrt(bias(\hat{\theta})^2 + var(\hat{\theta}))$, where, $bias(\hat{\theta}) = mean(\hat{\theta}) - \theta$, $mean(\hat{\theta}) = \frac{\sum_{i=1}^{1000} \hat{\theta}_i}{1000}$, and $var(\hat{\theta}) = \frac{1}{999} \sum_{i=1}^{1000} (\hat{\theta}_i - mean(\hat{\theta}))^2$.

The points 1 to 5 were repeated for samples of sizes 400, 500, and 600. We chose $\delta = \{1.311, -0.199, -0.083\}^t$, which led to an average response rate of 57% for each sample size.

Recall that we base this article on two calibration approaches, the linear calibration (LC) estimator of Särndal and Lundström (2005) and the propensity score calibration (PSC) of Chang and Kott (2008). For the former estimator, the auxiliary vector was given by $\mathbf{X}_k = \{1, v_{1k}, ..., v_{26k}\}^t$, whereas the latter used $\mathbf{X}_k = \{1, v_{2k}, ..., v_{4k}, v_{6k}, ..., v_{26k}\}^t$ and $\mathbf{H}_k = \{1, v_{1k}, v_{5k}\}^t$.

The principal components auxiliary variables for both the LC and PSC estimators were generated from their corresponding values of $\mathbf{X}$. The retention criterion for the LC estimator was the proportion of total variance explained by the set of selected components. This led to the selection of three principal components in population LC, while for the PSC estimator, the retention criterion is the one suggested in subsection 4.3. The scree plot given in Figure 1 below illustrates the population correlation structure of the variables.

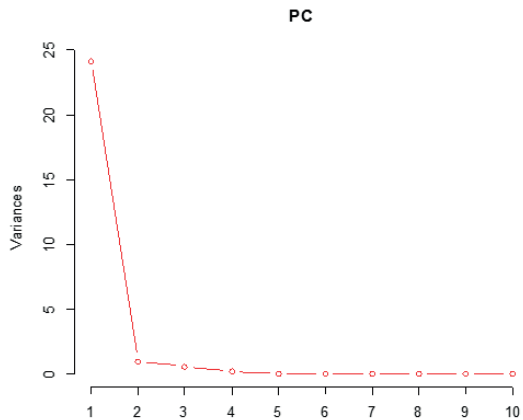


Figure 1: Scree plot of the auxiliary data in thefirst study.

5.1.2. Study 2

In this study, we perform the simulations following the same 5 points of the study 1. However, here we use the 'Small Area Income and Poverty Estimates' dataset (size 3173). We selected some variables from the 2006 and others from the 2013 data, for a total of 19 variables. The original data

were square root transformed and the new variables were named, v_1 to v_{19}. Again we chose three variables, two of which were used as model variables, say, v_1 and v_5, and the other was the study variable, say, v_{19}. The correlations of v_1 and v_5 with the study variable v_{19} were $cor(v_1, v_{19}) = 0.52$ and $cor(v_5, v_{19}) = 0.45$, respectively. This resembles the correlation structure of the corresponding variables in the study 1. As in the study 1, we used $\delta = \{1.311, -0.199, -0.083\}^t$. The LC estimator uses $\mathbf{X}_k = \{1, v_{1k}, ..., v_{18k}\}^t$ whereas the PSC estimator uses $\mathbf{X}_k = \{1, v_{2k}, ..., v_{4k}, v_{6k}, ..., v_{18k}\}^t$ and $\mathbf{H}_k = \{1, v_{1k}, v_{5k}\}^t$. The proportion of total variance explained by the selected PCs is again the retention criterion used for the LC estimator based on PCs. This criterion led to a selection of eight components. The retention criterion for the PSC estimator based on PCs is again the one described in subsection 4.3. The following is the scree plot of the principal components of the population auxiliary data used in the second simulation study.

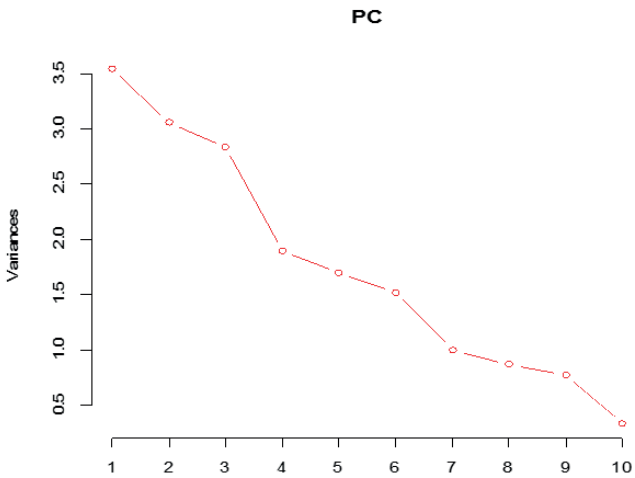


Figure 2: Scree plot of the auxiliary data in the second study

5.2. Simulation results

5.2.1. Results of study 1

The results are presented in two versions, a tabular version in Tables 1–4 and a graphic version in Figures 3–6 (the figures are in the appendix). These representations show the behaviour of each considered estimator when the sample size increases. For each table or graph, the performance of the estimator is evaluated from two perspectives: when the estimator is based on the complete original auxiliary variables (X) and when it is based on the PCs of the auxiliary variables.

Tables 1 and 2 show the LC estimator results when auxiliary information is observed at the population and sample levels, respectively. Both tables show that, the relative bias, standard error, and the root mean squared error values of the PCs-based linear calibration are smaller than their counterparts computed based on the original auxiliary variables.

Table 1: LC on original population auxiliary variables vs. LC on population PCs – Study 1

Sample size	Properties	Estimators	
		L. Calibration on X	L. Calibration on PCs
300	Rel.bias(%)	5.474	1.296
	S.E.	3519	935
	RMSE	8661	2094
400	Rel.bias(%)	4.771	1.224
	S.E.	3231	872
	RMSE	7616	1973
500	Rel.bias(%)	4.462	1.222
	S.E.	3083	804
	RMSE	7150	1941
600	Rel.Bias(%)	3.974	1.149
	S.E.	3135	846
	RMSE	6544	1864

Table 2: LC on original sample auxiliary variables vs. LC on sample PCs – Study 1.

Sample size	Properties	Estimator	
		L.Calibration on X	L. Calibration on PCs
300	Rel.bias(%)	3.930	0.192
	S.E.	21,936	11,202
	RMSE	22,660	11,206
400	Rel.bias(%)	3.341	0.037
	S.E.	17,089	9621
	RMSE	17,758	9621
500	Rel.bias(%)	3.551	0.328
	S.E.	14,332	8250
	RMSE	15,224	8263
600	Rel.bias(%)	2.951	0.369
	S.E.	12,422	7608
	RMSE	13,134	7626

Tables 3 and 4 show results obtained under conditions similar to those used to obtain the results in Tables 1 and 2, except that PSC replaces LC. The results obtained by using PCs of the auxiliary variables are comparable to the obtained using original auxiliary variables, this is true in both levels of auxiliary information.

Table 3: PSC on original population auxiliary variables vs. PSC on population PCs – Study 1.

Sample size	Properties	Estimator			
		PS on X	Time (in hr)	PS on PCs	Time (in hr)
300	Rel.bias(%)	0.280	7	0.153	0.35
	S.E.	16,182		15,912	
	RMSE	16,188		15,914	
400	Rel.bias(%)	0.105	13	0.209	0.57
	S.E.	13,815		13,660	
	RMSE	13,816		13,663	
500	Rel.bias(%)	0.338	22	0.434	0.83
	S.E.	11,953		11,837	
	RMSE	11,963		11,854	
600	Rel.bias(%)	0.169	36	0.264	1.30
	S.E.	10,899		10,757	
	RMSE	10,902		10,764	

Table 4: PSC on original sample auxiliary variables vs. PSC on sample PCs – Study 1.

Sample size	Properties	Estimator			
		PS on X	Time (in hr)	PS on PCs	Time (in hr)
300	Rel.bias(%)	0.255	0.25	0.125	0.18
	S.E.	16,162		16,010	
	RMSE	16,166		16,011	
400	Rel.bias(%)	0.120	0.32	0.189	0.22
	S.E.	13,820		13,711	
	RMSE	13,821		13,713	
500	Rel.bias(%)	0.353	0.45	0.421	0.23
	S.E.	11,952		11,834	
	RMSE	11,963		11,850	
600	Rel.bias(%)	0.191	0.50	0.263	0.25
	S.E.	10,880		10,795	
	RMSE	10,884		10,801	

5.2.2. Results of study 2

Tables 5–10 below present the results of this study. The process of evaluating the estimators is similar to that used in study 1. The results of the LC, presented in Tables 5 and 6, display consistency when comparing X-and PCs- based estimators and when comparing population- and sample-based estimators.

Table 5: LC on original population auxiliary variables vs LC on population PCs – Study2.

Sample size	Properties	Estimator	
		L. Calibration on X	L. Calibration on PCs
300	Rel.bias(%)	0.735	0.899
	S.E.	2262	2136
	MSE	2282	2168
400	Rel.bias(%)	0.810	1.077
	S.E.	1871	1798
	MSE	1901	1852
500	Rel.bias(%)	0.829	1.029
	S.E.	1558	1529
	MSE	1596	1588
600	Rel.bias(%)	0.672	0.836
	S.E.	1402	1382
	MSE	1429	1425

Table 6: LC on original sample auxiliary variables vs. LC on sample PCs – Study 2.

Sample size	Properties	Estimator	
		L. Calibration on X	L. Calibration on PCs
300	Rel.bias(%)	0.725	0.949
	S.E.	2489	2494
	MSE	2507	2525
400	Rel.bias(%)	0.882	1.205
	S.E.	2068	2068
	MSE	2100	2129
500	Rel.bias(%)	0.841	1.104
	S.E.	1792	1814
	MSE	1825	1871
600	Rel.bias(%)	0.711	0.937
	S.E.	1639	1658
	MSE	1666	1703

The results of the PSC estimators for the second study are displayed in Tables 7 and 8. As the LC estimator, the PSC results are also consistent in terms of the type (X or PCs) and level (population or sample) of the auxiliary information used.

Table 7: PSC on original population auxiliary variables vs. PSC on population PCs – Study 2.

Sample size	Properties	Estimator	
		PSC on X	PSC on PCs
300	Rel.bias(%)	1.345	1.566
	S.E.	3748	3791
	MSE	3789	3846
400	Rel.bias(%)	1.627	1.925
	S.E.	3293	3385
	MSE	3362	3478
500	Rel.bias(%)	1.487	1.848
	S.E.	2921	2994
	MSE	2985	3091
600	Rel.bias(%)	1.757	2.072
	S.E.	2708	2846
	MSE	2804	2973

Table 8: PSC on original sample auxiliary variables vs. PSC on sample PCs – Study 2.

Sample size	Properties	Estimator	
		PSC on X	PSC on PCs
300	Rel.bias(%)	1.347	1.480
	S.E.	3634	3643
	MSE	3676	3695
400	Rel.bias(%)	1.482	1.701
	S.E.	3166	3219
	MSE	3226	3296
500	Rel.bias(%)	1.559	1.648
	S.E.	2775	2815
	MSE	2849	2897
600	Rel.bias(%)	1.778	1.908
	S.E.	2567	2651
	MSE	2672	2767

The results shown in Tables 9 and 10 comprise estimated model parameters (with associated standard errors in parentheses) in the PSC estimation using the data of the second study.

Table 9: Estimated model coefficients (population auxiliary information – Study 2.)

Coefficient estimates						
True coefficients	$(\delta_0, \delta_1, \delta_2)$ (1.311, -0.199, -0.083)					
Sample size	PSC on X			PSC on PCs		
	$\hat{\delta}_0$	$\hat{\delta}_1$	$\hat{\delta}_2$	$\hat{\delta}_0$	$\hat{\delta}_1$	$\hat{\delta}_2$
300	1.129	-0.182	-0.044	1.097	-0.188	-0.026
	(0.197)	(0.008)	(0.011)	(0.247)	(0.012)	(0.014)
400	1.125	-0.174	-0.052	1.079	-0.176	-0.035
	0.149	0.006	0.008	(0.213)	(0.010)	(0.010)
500	1.147	-0.178	-0.056	1.096	-0.179	-0.038
	(0.133)	(0.005)	(0.006)	(0.182)	(0.008)	(0.009)
600	1.140	-0.178	-0.054	1.092	-0.179	-0.037
	(0.117)	(0.005)	(0.005)	(0.190)	(0.007)	(0.008)

Table 10: Estimated model coefficients (sample auxiliary information – Study 2.)

Coefficient estimates						
True coefficients	$(\delta_0, \delta_1, \delta_2)$ (1.311, -0.199, -0.083)					
Sample size	PSC on X			PSC on PCs		
	$\hat{\delta}_0$	$\hat{\delta}_1$	$\hat{\delta}_2$	$\hat{\delta}_0$	$\hat{\delta}_1$	$\hat{\delta}_2$
300	1.139	-0.177	-0.054	1.119	-0.179	-0.046
	(0.092)	(0.003)	(0.005)	(0.122)	(0.005)	(0.005)
400	1.149	-0.175	-0.061	1.118	-0.174	-0.052
	(0.068)	(0.003)	(0.003)	(0.108)	(0.004)	(0.005)
500	1.148	-0.175	-0.061	1.132	-0.175	-0.054
	(0.063)	(0.0002	(0.003)	(0.003)	(0.094)	(0.004)
600	1.145	-0.175	-0.060	1.123	-0.174	-0.054
	(0.06)	(0.002)	(0.002)	(0.099)	(0.003)	(0.004)

6. Discussion

The results of two simulation studies are presented in the last section, and for each study we assess two calibration approaches, the LC estimator using no explicit form of response function and the PSC estimator with explicit functional form. Both estimators are evaluated using the original large set of auxiliary variables (X) and using the principal components (PCs) of the original auxiliary variables. The results of the first study are given in two versions, tabular and graphic, while the results of the second study are given in tabular form only. The graphic form enables the convenient visual inspection of the estimator behaviour, while the tabular form gives a quantitative illustration. Study 1 demonstrates that the LC estimator based on principal components auxiliary variables is always superior in terms of relative bias, standard error, and root mean squared error (RMSE), to

its counterpart using the original auxiliary information. This is true regardless of the level of the auxiliary information, that is, the population or sample levels, as demonstrated in Tables 1 and 2, respectively.

There is a large discrepancy of the standard errors and RMSEs between population- and sample-based LC estimators. The RMSE of the population-based LC estimator based on the original auxiliary information ranges from 6544 to 8661 while the range for its counterpart based on sample-level auxiliary information is 13,134 to 22,660. The RMSE of the LC based on PCs auxiliary information ranges from 1864 to 2094 and from 7626 to 11,206 for population- and sample-based auxiliary information, respectively. Thus, the results differ greatly when comparing estimators of population- and sample-based auxiliary information. Considerable differences are also observed in the standard errors and RMSEs when comparing the estimators in terms of the type of auxiliary information used, that is, original X- auxiliaries and PCs- auxiliaries. This is not a surprising behaviour of the LC calibration estimator as this is a regression-type estimator.

In the response propensity calibration approach, auxiliary information is used in estimating response propensities; the estimation of population characteristics then proceeds by adjusting the design weights through multiplication by the corresponding reciprocals of the estimated propensities, which is usually called 'double weighting'. Here, it is observed that these results are more consistent. As Tables 3 and 4 illustrate, the principal-components-based estimator provides results similar to those obtained using the original auxiliary information. This is true regardless of the level of information on which the estimator is based. Furthermore, the results display consistency when comparing the properties of the corresponding estimators when population- and sample-based auxiliary data are used. The corresponding interval ranges of the RMSEs when using population-level auxiliary information are close to those when sample-level auxiliary information is used. As the sample size increases, the RMSEs tend to converge to the same level, irrespective of the type (X or PCs) or level (population or sample) of the auxiliary information. One of the major advantages of using PCs in place of the original auxiliary variables is the computational effort measured in terms of computational time; as reported in Tables 3–4, due to dimensionality reduction, the principal-components-based estimates are computed much more quickly than are the estimates based on the original auxiliary information.

Tables 5–10 report the results of the second simulation study. In contrast to the previous study, here, the LC calibration (Tables 5–6) results are consistent regardless of the type of auxiliary information used for estimation as well as when comparing the properties of the estimator across levels of information. The RMSEs of the estimators lie in virtually the same interval, regardless of the level of auxiliary information (population or sample levels) or type (original X or PCs) . A similar observation can be made with respect to the PSC estimator in Tables 7–8. We can still compare the performances of the LC and PSC estimators as we are using the same set of auxiliary variables, however, the estimators are conceptually different in terms of how auxiliary information is used.

The levels of bias are approximately the same: they are less than 0.1% in study 2 while in study 1 some differences are observed, especially in the LC estimator where the bias level attains 5.5%, as Tables 1 and 2 demonstrate. An interesting property of the auxiliary information in the PSC scheme, is the ability to appropriately estimate the response model. Tables 9–10 provide the population- and sample-based model-estimated coefficients, and the results suggest equally good model coefficients estimates when PCs are used compared with estimates resulting from the use of the original X

variables. As the results of model estimates are good, we can further improve the target estimates by performing a two-step estimation in which the products of design weigths and the reciprocal of the estimated response probabilities are used as initial weights in the linear calibration estimator.

Both study 1 and study 2 illustrate how the use of principal components in place of original auxiliary data when adjusting for nonresponse does not lead to distorted results and has the great advantage of reducing the computational effort.

The reported PSC results based on principal components are very similar to those obtained using a fixed number of components via the eigenvalue-one rule. However, the eigenvalue-one results are worse than those of our approach based on canonical correlation for very small samples. When the sample size increases, the number of selected components converges to the number of components based on the eigenvalue-one rule. The Figure 7 in the appendix illustrates the behaviour of our components selection method using the data of the study 1.

References

BARDSLEY, P. AND CHAMBERS, R. L. G. (1984). Multipurpose estimation from unbalanced samples. *Journal of the Royal Statistical Society. Series C(Applied Statistics)*, **33** (3), 290–299.

BEAUMONT, J. F. (2006). On the use of data collection process information for the treatment of unit nonresponse through weight adjustment. *Survey Methodology*, **31**, 227–231.

BETHLEHEM, J. AND SCHOUTEN, B. (2004). Nonresponse adjustment in household surveys. discussion paper 04007. *Statistics Netherlands*.

BILEN, C., KHAN, A., AND YADAV, O. P. (2004). Principal components regression control for multivariate autocorrelated cascade process. *International Journal of Quality Engeneering and Technology*, **3** (1), 301–316.

BINDER, D. A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, **51**, 279–292.

BORGA, M. (2001). Canonical correlation a tutorial.
URL: https://www.cs.cmu.edu/ +tom/10701 sp11/slides/CCA tutorial.pdf

CADIMA, J. AND JOLLIFÉ, I. T. (1995). Loading and correlations in the interpretation of principle components. *Journal of Applied Statistics*, **22** (2), 203–214.

CARDOT, H., GOGA, C., AND SHEHZAD, M. A. (2015). Calibration and partial calibration on principal components when the number of auxiliary variables is large. *arXiv:1406.7686 [stat.ME]*.

CHANG, T. AND KOTT, P. S. (2008). Using calibration weighting to adjust for nonresponse under a plausible model. *Biometrika*, **95** (3), 555–571.

DEVILLE, J. C. AND SÄRNDAL, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, **87**, 376–382.

GEUZINGE, L., ROOIJEN, J. V., AND BAKKER, B. F. M. (2000). The use of administrative registers to reduce non-response bias in household surveys. *Netherlands Official Statistics*, **2000** (2), 32–39.

GUGGEMOS, F. AND TILLÉ, Y. (2010). Penalized calibration in survey sampling: Design based estimation assisted by mixed models. *Journal of Statistical Planning and Inference*, **140**, 3199–3212.

HARDOON, D., SZEDMAK, S., AND SHAWE-TAYLOR, J. (2004). Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, **16**, 2639–2664.

HOTELLING, A. (1939). Relation between two sets of variables. *Biometrika*, **28** (4), 321–377.

JOHNSON, R. A. AND WICHERN, D. W. (2007). *Applied Multivariate Statistical Analysis*. Pearson, Prentice Hall.

JOLLIFÉ, I. T. (1972). Discarding variables in a principal component analysis. i: Articial data. *Journal of the Royal Statistical Society Series C (Applied Statistics)*, **21** (1), 160–173.

JOLLIFÉ, I. T., TRENDAFILOV, N. T., AND UDDIN, M. (2003). A modified principal component technique based on the lasso. *Journal of Computational and Graphical Statistics*, **12** (3), 531–547.

KOTT, P. S. (2006). Using calibration weighting to adjust for nonresponse and coverage errors. *Survey Methodology*, **32** (2), 133–142.

KOTT, P. S. (2012). Exploring some uses for instrumental-variable calibration weighting. *In Section on Survey Research Methods-JSM2012*.

KREUTER, F. AND OLSON, K. (2011). Multiple auxiliary variables in nonresponse adjustment. *Sociological Methods and Research*, **40** (2), 311–332.

LUNDSTRÖM, S. AND SÄRNDAL, C.-E. (1999). Calibration as a standard method for treatment of nonresponse. *Journal of Official Statistics*, **15** (2), 305–327.

MANSFIELD, E. R., WEBSTER, J. T., AND GUNST, R. F. (1977). An analytic variable selection technique for principal component regression. *Applied Statistics*, **6**, 34–40.

MCCABE, G. P. (1984). Principal variables. *Technometrics*, **26** (2), 137–144.

MCHENRY, C. E. (1978). Computation of a best subset in multivariate analysis. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, **27** (3), 291–296.

PARRA, L. AND P. SAJDA, P. (2003). Blind source separation via generalized eigenvalue decomposition. *Journal of Machine Learning Research*, **4**, 1261–1269.

RIZZO, L., KALTON, G., AND BRICK, M. (1996). A comparison of some weighting adjustment methods for panel nonresponse. *Survey Methodology*, **22**, 43–53.

SÄRNDAL, C.-E. AND LUNDSTRÖM, S. (2005). *Estimation in Surveys with Nonresponse*. Wiley: New York.

SÄRNDAL, C.-E. AND LUNDSTRÖM, S. (2007). Assessing auxiliary vectors for control of nonresponse bias in the calibration estimator. *Journal Of Official Statistics*, **24** (2), 167–191.

SÄRNDAL, C.-E., SWENSSON, B., AND WRETMAN, J. F. (1992). *Model Assisted Survey Sampling*. Springer: New York.

SCHOUTEN, B. (2007). A selection strategy for weighting variables under a not-missingat-random assumption. *Journal Official Statistics*, **23**, 51–68.

SILVA, N. AND SKINNER, C. J. (1997). Variable selection for regression estimation in finite populations. *Survey Methodology*, **23** (1), 23–32.

PAPER III

Appendix A: Asymptotic variance of the estimated coefficients of the propensity functions

Let

$$E_{pq}\left[\left(\mathbf{Z}_{(r)}^{t}\Phi(\delta^{*})\mathbf{d}_{(r)}-\mathbf{T}_{z}\right)\left(\mathbf{Z}_{(r)}^{t}\Phi(\delta^{*})\mathbf{d}_{(r)}-\mathbf{T}_{z}\right)^{t}\right]=\Pi_{1}+\Pi_{2} \tag{20}$$

where $\Pi_1 = \sum_{k\in U}\sum_{l\in U}(\pi_{kl}-\pi_k\pi_l)d_k d_l \mathbf{Z}_k\mathbf{Z}_l^t$ and $\Pi_2 = \sum_U d_k(h(\mathbf{H}_k^t\delta^*)-1)\mathbf{Z}_k\mathbf{Z}_k^t$ (see Chang and Kott, 2008).
Then

$$Avar\sqrt{n}\left(\hat{\delta}_{(pc)}-\delta^*\right) = \left[\mathbf{F}^t\mathbf{W}\mathbf{F}\right]^{-1}\mathbf{F}^t\mathbf{W}\Theta\mathbf{W}\mathbf{F}\left[\mathbf{F}^t\mathbf{W}\mathbf{F}\right]^{-1}$$

with $\mathbf{F} = \mathbf{Z}^t\Psi\mathbf{H}$ and $\Theta = Avar\left[n^{-1/2}\left(\mathbf{Z}_{(r)}^t\Phi(\delta^*)\mathbf{d}_{(r)}-\mathbf{T}_z\right)\right]$
We choose

$$\mathbf{W}^{-1} = \Theta$$

and obtain,

$$Avar\sqrt{n}\left(\hat{\delta}_{(pc)}-\delta^*\right) = \left[\left(\mathbf{Z}^t\Psi\mathbf{H}\right)^t\Theta^{-1}\left(\mathbf{Z}^t\Psi\mathbf{H}\right)\right]^{-1} \tag{21}$$

Where, $\mathbf{W} = p\lim_{n\to\infty}\mathbf{W}_n$, is a positive definite matrix,
$(\mathbf{Z}^t\Psi\mathbf{H}) = p\lim_{n\to\infty}\frac{1}{n}E_{pq}\left(\mathbf{Z}_{(r)}^t\Psi^0\mathbf{H}\right)$ and

$$\Theta = p\lim_{n\to\infty}\frac{1}{n}E_{pq}\left[\left(\mathbf{Z}_{(r)}^t\Phi(\delta^*)\mathbf{d}_{(r)}-\mathbf{T}_z\right)\left(\mathbf{Z}_{(r)}^t\Phi(\delta^*)\mathbf{d}_{(r)}-\mathbf{T}_z\right)^t\right]$$

Alternatively, the calibration (8) is on estimated principal components, that is,

$$\hat{\mathbf{Z}}_{(r)}^t\Phi(\delta)\mathbf{d}_{(r)}-\hat{\mathbf{T}}_z = \mathbf{0} \tag{22}$$

where $\hat{\mathbf{T}} = \sum_s d_k\hat{\mathbf{Z}}_k$.
Observe that,

$$var_{pq}(\hat{\mathbf{Z}}_{(r)}^t\Phi(\delta^*)\mathbf{d}_{(r)}-\hat{\mathbf{T}}_z) = V_1+V_2$$

where, $V_1 = var_pE_q\left(\hat{\mathbf{Z}}_{(r)}^t\Phi(\delta^*)\mathbf{d}_{(r)}-\hat{\mathbf{T}}_z|s\right)$ and $V_2 = E_pvar_q\left(\hat{\mathbf{Z}}_{(r)}^t\Phi(\delta^*)\mathbf{d}_{(r)}-\hat{\mathbf{T}}_z|s\right)$.
The first variance component is zero, implying that

$$E_{pq}\left[\left(\hat{\mathbf{Z}}_{(r)}^t\Phi(\delta^*)\mathbf{d}_{(r)}-\hat{\mathbf{T}}_z\right)\left(\hat{\mathbf{Z}}_{(r)}^t\Phi(\delta^*)\mathbf{d}_{(r)}-\hat{\mathbf{T}}_z\right)^t\right] = \sum_U d_k(h(\mathbf{H}_k^t\delta^*)-1)\mathbf{Z}_k\mathbf{Z}_k^t,$$

therefore, the sample version analogous to (**W**) in (21) is $\tilde{\mathbf{W}} = p\lim_{n\to\infty}\frac{1}{n}\sum_U d_k(h(\mathbf{H}_k^t\delta^*)-1)\mathbf{Z}_k\mathbf{Z}_k^t$.

Appendix B: Figures

Manuscript received, 20XX-01-01, revised, 20XX-01-02, accepted, 20XX-01-03.

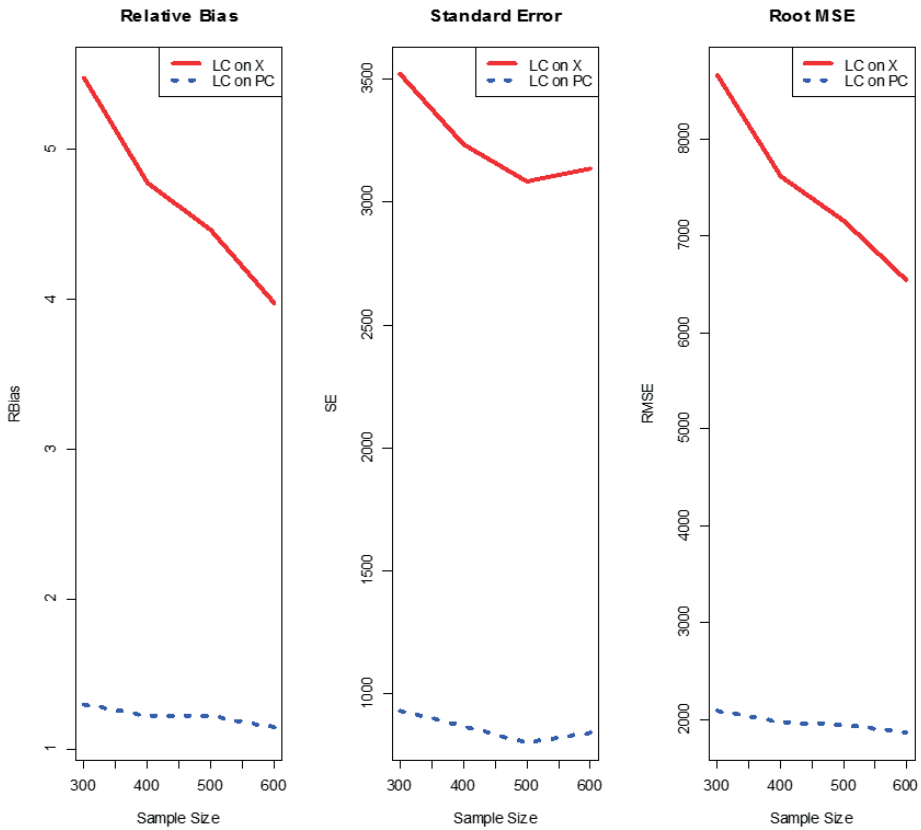


Figure 3: LC on original population auxiliary variables vs. LC on population PCs – Study 1.

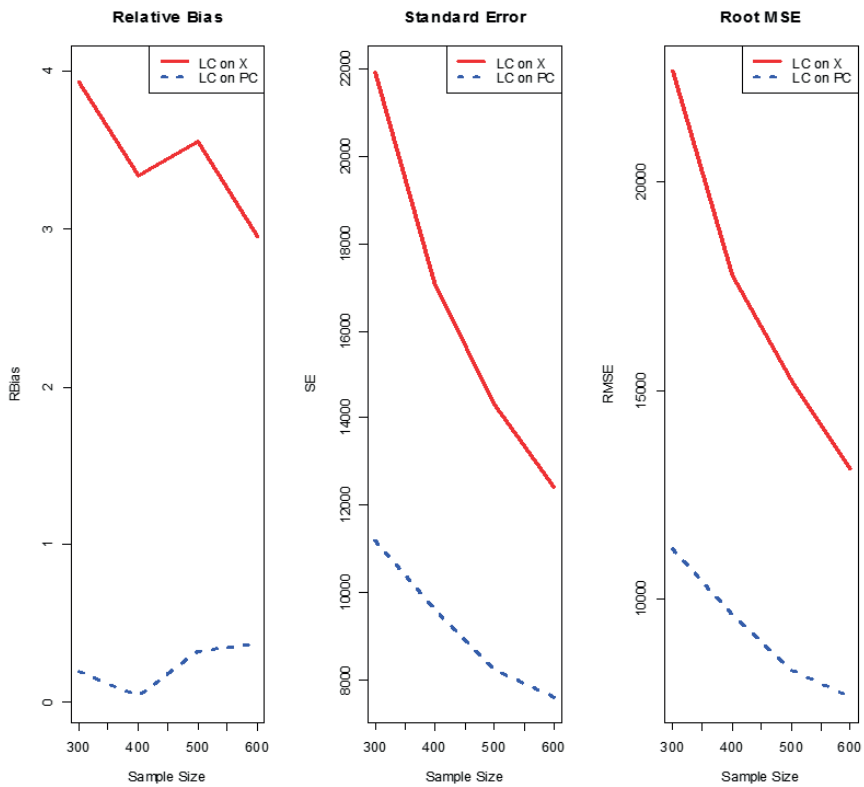


Figure 4: LC on original sample auxiliary variables vs. LC on sample PCs – Study 1.

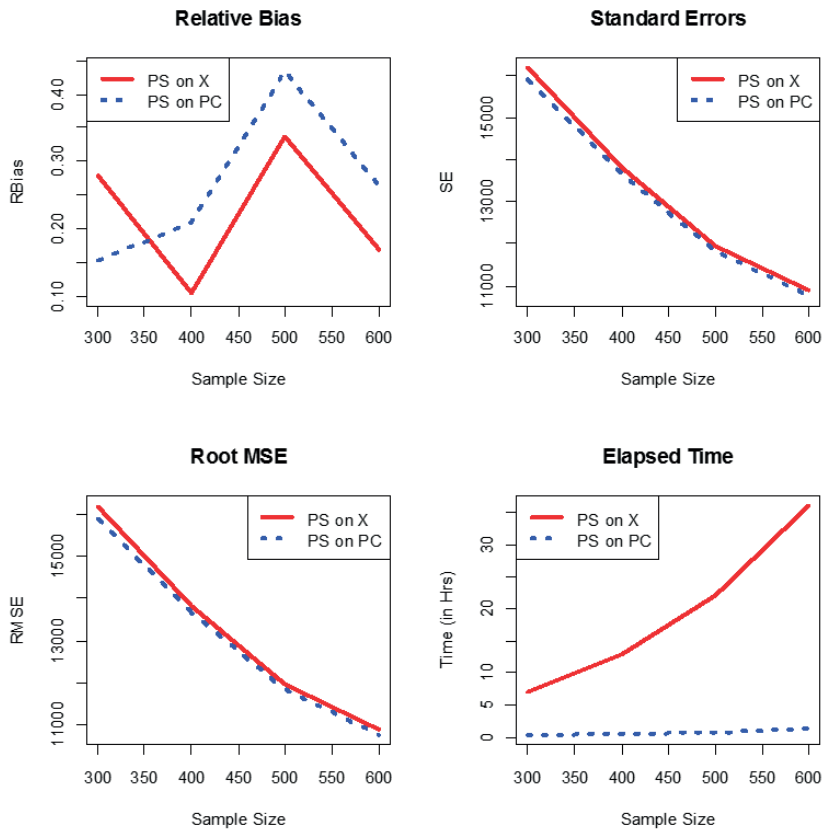


Figure 5: PSC on original population auxiliary variables vs. PSC on population PCs – Study 1.

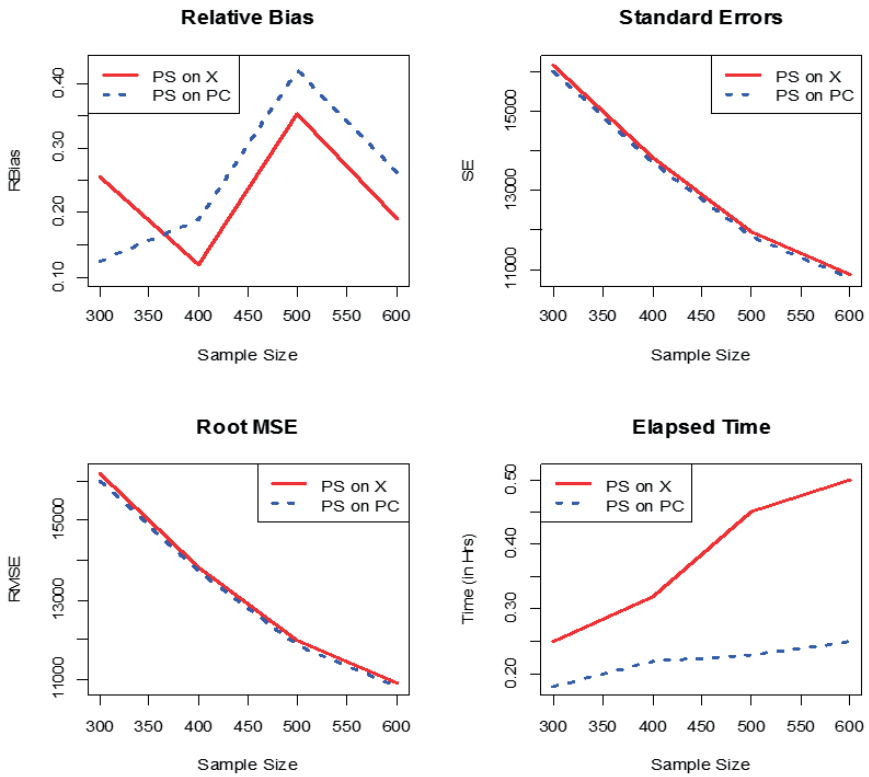


Figure 6: PSC on original sample auxiliary variables vs. PSC on sample PCs – Study 1.

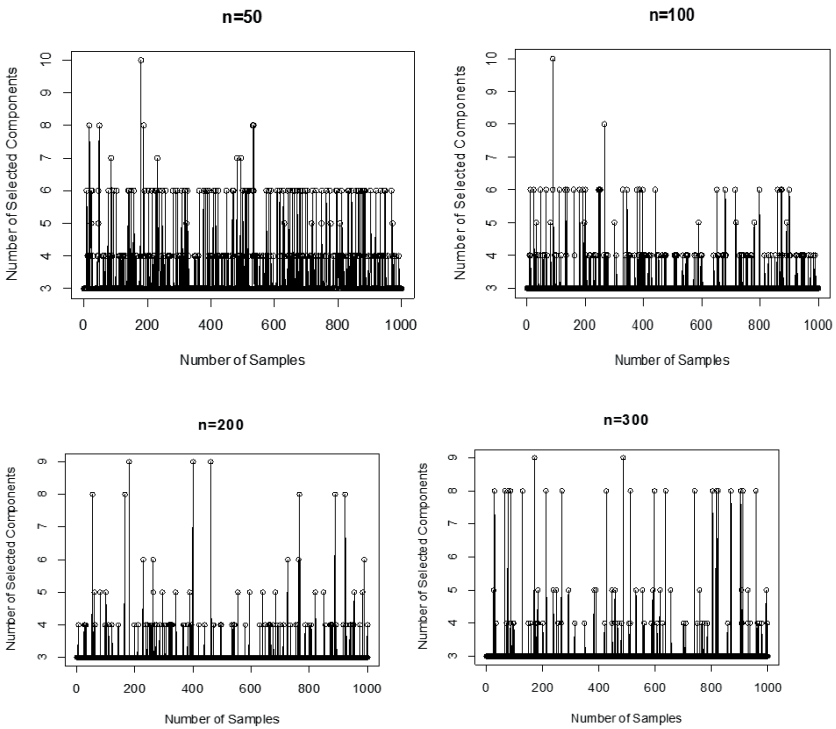


Figure 7: Behaviour of the number of selected components when sample sizes increase – Study 1.

On the Use of Auxiliary Variables and Models in Estimation in Surveys with Nonresponse*

Bernardo João Rota
and
Thomas Laitila

PAPER IV

Abstract

This paper contains a discussion on two alternative weighting procedures, that is, weighting with and without explicitly modeling of the response mechanism, known as the *direct weighting* and the *weighting* approaches respectively. The generalized regression estimator benchmarks the weighting methods, whereas a general double-weighted Horvitz-Thompson estimator represents the direct-weighting approach. A general reliance on the strength of the correlation between the auxiliary variables, the response behavior and the study variables prevailing mostly in weighting approaches is shown to be inappropriate in some cases, that is, such reliace increases the bias of the resulting estimator. Conversely, the traditional use of simple models in representing the true response behavior is addressed through an example in which it is shown to be adequate only under very restrictive assumptions. Results presented for both weighting procedures reveal the need to develop new tools and methodologies when performing estimation in surveys with nonresponse.

Keywords:Auxiliary Variables, Nonresponse, Regression Estimator, Weighting.

*This is a joint work with Professor Thomas Laitila, Department of Statistics, Örebro University.

1 Introduction

In adjusting for nonresponse, weighting is a commonly used approach by survey methodologists. Weighting relies on auxiliary variables, which can be defined as variables on which information is available for respondents and nonrespondents. In weighting for nonresponse adjustment, the role of auxiliary variables is crucial in reducing nonresponse errors. Rizzo, Kalton, and Brick (1996) noted that the selection of auxiliary variables could be more important than the weighting scheme. Furthermore, Särndal (2011) claims that in the case of bias inflation by nonresponse, access to powerful auxiliary variables becomes key in minimizing the problem. These auxiliary variables are required to predict (a) the propensities to respond and (b) the key survey variables to adjust effectively for nonresponse (West and Little, 2012).

Use of auxiliary variables in estimation can be found in for example, Bethlehem (1988), Estevão and Särndal (2000), Kalton and Flores-Cervantes (2003), Särndal and Lundström (2005), and Särndal (2007). In practice, there can be a wide choice of variables (Särndal and Lundström, 2008), and one must decide on their selection for effective adjustment. The literature provides suggestions to guide the selection of auxiliary variables. Särndal and Lundström (2008) propose a selection device based on the variability of the reciprocals of estimated propensities. The propensities are determined under the assumption that the auxiliary variables satisfy some pre-specified condition. Geuzinge, Rooijen and Bakker (2000) propose a selection indicator based on a product of correlations arising from (a) and (b). In adjusting for nonresponse using the regression estimator, Bethlehem and Schouten (2004) and Schouten (2007) propose a selection based on minimizing a maximal absolute bias of the estimator. The method relies on computing an interval for the maximal absolute bias and selecting those variables that minimize its width.

Searching for auxiliary variables satisfying requirements (a) and (b) simultaneously can be a difficult task. Survey practices involve many variables of interest; as Kott (2013) comments on Brick's (2013) discussion paper, one can seldom encounter auxiliary variables fulfilling (b) for every variable of interest in a multipurpose survey. Kreuter and Olson (2011) also noted the same difficulty. Furthermore, as we illustrate with an example, fulfilling requirements (a) and (b) simultaneously does not generally guaranty effectiveness in bias protection for target estimates. Doing so can even introduce a larger bias. It might not be appropriate to rely entirely on correlation relationships between the variables involved in the study.

For adjustment methods in which the response behavior is explicitly mod-

eled, the primary goal is in observing those variables that are linked to a response pattern; thus, the estimation of targets is viewed as a second objective after the estimation of the response model. However, the approach is also challenging in the sense that it is difficult or even impossible to guess the appropriate response behavior when some can have simple forms, whereas others are complex. Simple models such as the logit and probit models are usually used to represent the true response model. We use a telephone survey case to show that such simple models are adequate under very specific assumptions.

To produce good adjustment weights, weighting methods rely on the proper use of available auxiliary information (e.g. Falk, 2012; Brick, 2013). We emphasize here weighting in two directions, that is, with and without an explicit modelling of response function. Kim and Kim (2007) note that the weighting procedures for nonresponse adjustment are mainly made by applying one of the two approaches: *weighting adjustment* or *direct weighting adjustment*. They are treated in the next two sections followed by a discussion of results in the final section.

2 Weighting adjustment

In *weighting adjustment,* the auxiliary information is embedded into the estimation of targets, improving the efficiency of the resulting estimators. The generalized regression (GREG) estimator is an example of this type of adjustment.

Suppose a sample $s = \{1, 2, ..., k, ..., n\}$ of size n is drawn from a population $U = \{1, 2, ..., k, ..., N\}$ of size N with a probability sampling design $p(s)$, yielding sample inclusion probabilities $\pi_k = \Pr(k \epsilon s) > 0$ and corresponding design weights $d_k = 1/\pi_k$ for all $k \in U$. Let y and $\mathbf{x}$ be the survey variable of interest and an L-dimensional column vector of auxiliary variables, respectively. We want to estimate $Y = \sum_U y_k$.

Assume the following relationship between y and $\mathbf{x}$, described through the model:

$$\zeta : y_k = \boldsymbol{\beta}^t \mathbf{x}_k + \varepsilon_k, k = 1, .., N \tag{1}$$

where $\boldsymbol{\beta}$ is an L-dimensional column vector of model parameters, and ε_k is a zero-mean random variable with $V_\zeta(\varepsilon_k) = \sigma_k^2$.

The generalized regression (GREG) estimators for Y based on the relationship between y and $\mathbf{x}$ given by equation (1) are a class of estimators of the form

$$\hat{Y}_{reg} = \left(\sum_U \mathbf{x}_k - \sum_s d_k \mathbf{x}_k \right)^t \hat{B}_s + \sum_s d_k y_k \tag{2}$$

where $\hat{B}_s = \left(\sum_s d_k c_k \mathbf{x}_k \mathbf{x}_k^t\right)^{-1} \sum_s d_k c_k \mathbf{x}_k y_k$ and we assume $c_k = 1$.

According to Cobben (2009), the GREG estimator was introduced by Särndal (1980) and Bethlehem and Keller (1987). The GREG estimator (2) is extensively studied in Särndal et al. (1992); its properties rely on the sampling design and a close linear fit between y and $\mathbf{x}$ without explicitly depending upon whether (1) is true. In this setting, the regression estimator (2) is deemed model assisted rather than dependent (Särndal, 2007). The model-dependent regression estimator is extensively reviewed in Fuller (2002). The regression estimator can be written in a simpler form as a weighted sum of the values of the survey variable by writing $\left(\sum_U \mathbf{x}_k - \sum_s d_k \mathbf{x}_k\right)^t \left(\sum_s d_k \mathbf{x}_k \mathbf{x}_k^t\right)^{-1} \sum_s d_k \mathbf{x}_k y_k$ as $\sum_s d_k M_k y_k$, where $M_k = \left(\sum_U \mathbf{x}_k - \sum_s d_k \mathbf{x}_k\right)^t \left(\sum_s d_k \mathbf{x}_k \mathbf{x}_k^t\right)^{-1} \mathbf{x}_k$. Thus, equation (2) becomes

$$\hat{Y}_{reg} = \sum_s w_k y_k \tag{3}$$

with $w_k = d_k(1 + M_k)$.

This particular form of the regression estimator is advantageous because the weights w_k can be applied to any survey variable and have the following property:

$$\sum_s w_k \mathbf{x}_k = \sum_U \mathbf{x}_k. \tag{4}$$

Furthermore, when $\sum_U \mathbf{x}_k$ can be constructed by summing $\mathbf{x}_k$ in the sampling frame, a number of regression estimators can be constructed. However, observing $\mathbf{x}_k, k = 1, .., N$ is not a requirement for the regression estimator based on (1); it suffices to know only $\sum_U \mathbf{x}_k$, which can be information obtained from other sources.

Letting $M_k^* = \left(\sum_U \mathbf{x}_k - \sum_s d_k \mathbf{x}_k\right)^t \left(\sum_s d_k \mathbf{z}_k \mathbf{x}_k^t\right)^{-1} \mathbf{z}_k$, where $\mathbf{z}_k$ is a vector of auxiliary variables conceptually different, but of the same dimension as $\mathbf{x}_k$, we obtain the more general regression estimator given in (5). In this case, the regression estimator resembles the instrumental variable regression estimator learned from econometric theory.

$$\hat{Y}_{IVreg} = \left(\sum_U \mathbf{x}_k - \sum_s d_k \mathbf{x}_k \right)^t \hat{B}_{IVs} + \sum_s d_k y_k. \tag{5}$$

where $\hat{B}_{IVs} = \left(\sum_s d_k \mathbf{z}_k \mathbf{x}_k^t\right)^{-1} \sum_s d_k \mathbf{z}_k y_k$.

In our case, where sampling is followed by nonresponse (e.g. An, 1996; Fuller and An, 1998; Singh and Kumar, 2011), the GREG estimator (5) is given by

$$\hat{Y}_{IVreg^*} = \left(\sum_U \mathbf{x}_k - \sum_r d_k \mathbf{x}_k\right)^t \hat{B}_{IVr} + \sum_r d_k y_k \tag{6}$$

where $\hat{B}_{IVr} = \left(\sum_r d_k \mathbf{z}_k \mathbf{x}_k^t\right)^{-1} \sum_r d_k \mathbf{z}_k y_k$, and r is the set of respondents.

Let us assume that the condition $\boldsymbol{\lambda}^t \mathbf{z}_k = 1$ holds for all k, where $\boldsymbol{\lambda}$ is a constant vector. Equation (6) becomes

$$\hat{Y}_{IVreg^*} = \sum_U \mathbf{x}_k^t \hat{B}_{IVr} \tag{7}$$

The expected value of $\hat{Y}_{IVreg^*}$ is approximately

$$E\left(\hat{Y}_{IVreg^*}\right) \approx \sum_U \mathbf{x}_k^t B_{IV\theta} \tag{8}$$

where $B_{IV\theta} = \left(\sum_U \theta_k \mathbf{x}_k \mathbf{z}_k^t\right)^{-1} \sum_U \theta_k \mathbf{z}_k y_k$ and $\theta_k = \Pr(k\epsilon r | k\epsilon s)$.

Equation (8) says that the bias of the regression estimator almost entirely depends upon the properties of the response-based regression coefficients $\hat{B}_{IVr}$. If all $\theta_k = 1$, then the regression estimator is approximately unbiased for $\sum_U y_k$. The following interesting statement is given by Cobben (2009): "*Practical experience (at least in the Netherlands) shows that nonresponse often seriously affects estimators like means and totals, but less often causes estimates of relationships to be biased. Particularly if relationships are strong, i.e., the regression line fits the data well, the risk of finding wrong relationships is small*". Furthermore, Särndal and Lundström (2005, p. 100) shows the existence of the following relationship between $B_{IV\theta}$ and $B = \left(\sum_U \mathbf{z}_k \mathbf{x}_k^t\right)^{-1} \sum_U \mathbf{z}_k y_k$:

$$B_{IV\theta} - B = \left(\sum_U \theta_k \mathbf{z}_k \mathbf{x}_k^t\right)^{-1} \sum_U \theta_k \mathbf{z}_k e_k \tag{9}$$

where $e_k = y_k - \mathbf{x}_k^t B$.

Equation (9) says that estimator (7) is approximately unbiased for Y if the linear fit between y_k and $\mathbf{x}_k$ is strong or the regression errors are uncorrelated with the response probabilities. Thus, a need exists for strong relationships

between the explanatory variables, the response probabilities and the variables of interest.

- **The auxiliary variable can introduce bias**

Fuller and An (1998) emphasize that the level of bias reduction depends upon the relationships between the auxiliary variable, the variable of interest, and the response probability.

A question rarely raised in the literature can be formulated as follows: how does weighting affect estimates if the response set mean is unbiased? One potential reason for this problem not being addressed is the adaptation of concepts on the relationship between the study variable and the generation of the response set from the model-based inference literature, e.g., MAR (missing at random) and MCAR (missing completely at random) and ignorable and nonignorable nonresponse.

For estimation of population means or totals in the finite population framework, such concepts can be misleading. MCAR is a stronger concept than MAR, usually meaning that if MCAR holds, so does MAR. Methods derived to handle MAR cases then also encompass MCAR cases. However, this meaning might not hold in the finite population context for similar concepts; MCAR might hold but not MAR.

If MCAR is defined as $\sum_U \theta_k y_k = \bar{\theta} \sum_U y_k$ and $U_x \subset U$, a subset e.g. defined by values on auxiliary variables, then MCAR does not imply $\sum_{U_x} \theta_k y_k = \bar{\theta} \sum_{U_x} y_k$. The same argument can be derived by considering a random draw from the population and observing y and R, where R is a response indicator variable . Then MCAR, defined as $F(y|R=1) = F(y|R=0)$, does not imply $F(y|R=1,x) = F(y|R=0,x)$, where $F(\cdot)$ denotes the cdf.

Consider the following relationships:

$$\begin{aligned} y_k &= \beta_0 + \beta_1 x_k + e_k \\ \theta_k &= \Pr(k \epsilon r | k \epsilon s) \end{aligned} \tag{10}$$

where

1. $\sum_U e_k = 0$ and $\sum_U x_k e_k = 0$
2. $\sum_U \theta_k x_k \neq \bar{\theta} \sum_U x_k$

Here the auxiliary variable x_k correlates with both the study variable ($\beta_1 \neq 0$) and the response probability θ_k.

The approximate bias of the expanded Horvitz-Thompson estimator for the total of y obtained from (6) by setting $\mathbf{x}_k = \mathbf{z}_k = 1$,

$$\hat{Y}_{\text{exp}} = \frac{N}{\sum_r d_k} \sum_r d_k y_k$$

is given by

$$NearBias(\hat{Y}_{exp}) = N \frac{E\left(\sum_r d_k y_k\right)}{E\left(\sum_r d_k\right)} - \sum_U y_k = \frac{N \cdot cov(\theta, y)}{\bar{\theta}}.$$

where $cov(\theta, y)$ denotes the covariance between θ and y in the population and $\bar{\theta}$ is the population mean of θ. Thus, the expansion estimator is approximately unbiased if $cov(\theta, y)$ is zero.

One special case of the regression estimator is the ratio estimator, which is obtained from (6) by setting $\mathbf{x}_k = x_k$ and $\mathbf{z}_k = 1$. In the literature, the ratio estimator is suggested to have smaller bias due to nonresponse than does the expansion estimator. The approximate bias for the GREG estimator in this case is given by

$$NearBias(\hat{Y}_{RA}) = \frac{N^2 \bar{x}}{\sum_U \theta_k x_k} \cdot \sigma_{\theta e} \tag{11}$$

where $\bar{x} = N^{-1} \sum_U x_k$, $\sigma_{\theta e} = cov(\theta, e)$, and in model (10), we assume $\beta_0 = 0$.

Now, in this case, if $cov(\theta, y) = 0$ then $cov(\theta, e) \neq 0$ and the expansion estimator is approximately unbiased whilst the ratio estimator is not. Hence, using auxiliary information in estimation introduces errors in estimates because of estimator bias.

Defining $\mathbf{x}_k = \mathbf{z}_k = \begin{pmatrix} 1 & x_k \end{pmatrix}^t$, equation (7) yields another well-known estimator, the simple linear regression, which is also suggested to be more efficient than the expansion estimator. In this case, the bias of the regression estimator is given by:

$$NearBias(\hat{Y}_{reg}) = N \frac{\sigma_{xh}\sigma_{\theta e} - \sigma_{x\theta}\sigma_{he}}{\bar{\theta}\bar{hx} - \bar{h}^2} \tag{12}$$

where $h_k = \theta_k x_k$, $\sigma_{xh} = cov(x, h)$, $\sigma_{he} = cov(h, e)$, $\sigma_{x\theta} = cov(x, \theta)$, $\bar{h}^2 = (N^{-1} \sum_U h_k)^2$, and $\bar{hx} = N^{-1} \sum_U h_k x_k$. Additionally, we no longer assume $\beta_0 = 0$ in (10). Again, if $cov(\theta, y) = 0$ assumptions made imply $cov(\theta, e) \neq 0$ and the numerator in this bias expression cannot be claimed zero in general. Thus, using auxiliary information via the regression estimator also yields a biased estimator although the expansion estimator is approximately unbiased.

Thus, the ratio and the linear regression estimators have nonzero approximate bias resulting from the choice of auxiliary variable whereas the expansion estimator is approximately unbiased. The bias of the ratio estimator is proportional to $\sigma_{\theta e}$ whereas the approximate bias of the regression estimator is a weighted sum of $\sigma_{\theta e}$ and $\sigma_{\theta x}$. This might both reduce or increase bias compared with the ratio estimator.

The NearBias expressions presented above shows that the recommendation of selecting powerful auxiliary variables in the sense of being correlated with variables of interest and response probability can introduce, instead of reducing bias due to nonresponse. The bias expression for the regression estimator is more complex and a numerical example is used for illustration.

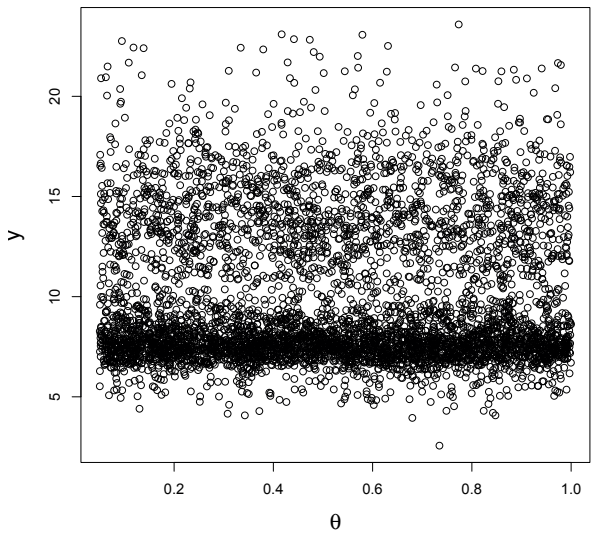


Figure 1. Plot of study variable y against response probability θ.
(y obtained from a real data set ($n = 5214$) and $\theta = 0.05 + 0.95 \cdot u$
where u are independent $U(0, 1)$ numbers.)

Figure 1 depicts 5214 values of a study variable y, obtained from real data. To obtain a case where the expansion estimator is approximately unbiased re-

sponse probabilities are independently generated from a uniform distribution as $\theta = 0.05 + 0.95 \cdot u$ where $u \sim U(0,1)$. The figure depicts a plot of y against the generated response probabilities (θ). For the data in Figure 1, the total of y is 51925.49 and the approximate relative bias of the expansion estimator is 0.06%. The correlation between y and θ is $cor(\theta, y)$=0.0034.

As an auxiliary variable, the variable $x = \theta + 0.1 \cdot y + a \cdot e$ is generated, where a is a fixed number and e is generated from a standard normal distribution. The correlations between x and the study variable and the response probability, respectively, are controlled by the value of a.

Calculated approximate biases for the regression estimator are shown in Table 1. In terms of correlations, the auxiliary variable x is strongest when $a = 0$, where $cor(x, \theta) = 0.6082$ and $cor(x, y) = 0.7958$. This case also shows the largest bias, -10.05%. By increasing a the correlations $cor(x, \theta)$ and $cor(x, y)$ decrease and so does also the relative bias in absolute numbers.

The average response probability is 0.5223 for the data underlying the results in Table 1. In two additional calculations, the average response probabilities are increased to 0.698. In one of them $cor(x, \theta)$ and $cor(x, y)$ are similar to the ones in Table 1. In the other, $cor(x, y)$ is larger and $cor(x, \theta)$ is smaller compared with those in Table 1. However, both additional calculations give much smaller biases. In the case with $a = 0$, bias reduces to -3.76% and -3.51%. Thus, the response rate is indicated an important determinant of the potential bias introduced by using the regression estimator.

Table 1: Approximate relative bias of the regression estimator of the total of the variable y depicted in Figure 1. Auxiliary variable generated as $x = \theta + 0.1 \cdot y + a \cdot e$ where e is standard normal distributed.

a	$cor(x, \theta)$	$cor(x, y)$	Rel.Bias (%)
0	0.6082	0 .7958	-10.05
0.5	0.4167	0.5339	-4.35
1	0.2618	0.3286	-1.62
1.5	0.1871	0.2303	-0.78
2	0.1458	0.1759	-0.45

Note: y has total 51925.49 and standard deviation 3.58. Approximate relative bias of the expansion estimator is 0.06%

3 Direct weighting adjustment

In *direct weighting adjustment,* it is assumed that the functional form of the response probability is known and given by $\theta_k = p(\cdot \mathbf{z}_k)$, where $\mathbf{z}_k$ is a vector of model variables. The primary goal is to estimate this function so that the observed values of the target variable are double weighted, that is, each y_k is multiplied by $d_k \hat{\theta}_k^{-1}$, where $\hat{\theta}$ estimates θ. The target population Y can then be estimated by

$$\hat{Y}_{nr} = \sum_r d_k \hat{\theta}_k^{-1} y_k. \tag{13}$$

The estimator (13) is widely suggested in the literature of nonresponse adjustment (see e.g. Chang and Kott, 2008; Kim and Park, 2010; Kim and Riddles, 2012). The properties of $\hat{Y}_{nr}$ are conditioned on the properties of $\hat{\theta}$. For example, the consistency of $\hat{Y}_{nr}$ depends in general on a correct specification of the function θ. Thus, a wrongly specified θ leads to an inconsistent $\hat{Y}_{nr}$ estimator, in general. Given the limitation on knowledge of the response mechanism (Särndal and Lundström, 2005), it is difficult to determine whether a proposed response mechanism is the appropriate one. Simple models such as the logit and probit models are often suggested and used in applications (e.g. Chang and Kott, 2008). An immediate question is when such simple models are appropriate?

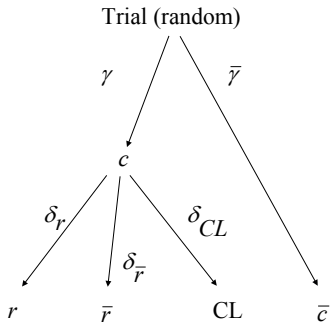
Here we take on a process view with repeated contact attempts in a telephone survey. Figure 2 below illustrates the possible outcomes in one attempt. We also consider an attempt to make contact with a unit in the sample and collect a response as a random trial. From this, calling can result in a contact (c) with probability γ or in a failure to make contact ($\bar{c}$) with probability $\bar{\gamma} = (1 - \gamma)$.

Given a contact is made with a unit in the sample, it may result in a response (r), a refusal to participate ($\bar{r}$), or an agreement to call back later (CL). Conditionally on c, let the probabilities of these outcomes be denoted δ_r, $\delta_{\bar{r}}$, and δ_{CL}, respectively.

Factors affecting the probability of a contact include an incorrect telephone number, the unit cannot at the time respond to a telephone call and, the respondent is not willing to respond to a call from an unknown displayed telephone number. If a contact is made, other factors are involved in a decision to respond, refuse or agree to be contacted later. The presentation of the survey, the topic of the survey and the time required to respond come into play. Thus, Figure 2 indicate the probability of a response at one attempt to be a function of two different probabilities which are determined by different

factors and variables. The summarizing of terms for e.g. the probability of a response into a simple function as the logistic cdf seems less appropriate.

The probabilities of the outcomes of the trial are $Pr(r) = \theta_1 = \gamma\delta_r$, $Pr(\bar{r}) = \theta_2 = \gamma\delta_{\bar{r}}$, $Pr(CL) = \theta_3 = \gamma\delta_{CL}$, and $Pr(\bar{c}) = (1-\gamma) = \bar{\gamma}$. If the outcome of the trial is either a failure to make contact or an agreement to call back, a second trial to obtain a response from the unit can be made. The same potential outcomes are possible. For simplicity, it is here assumed that given an agreement to call back later, contact is made in the next trial and the outcome is either a response (r) or a nonresponse ($\bar{r}$).



r = response, $\bar{r}$ = refusal, CL = call later,
$\bar{c}$ = no contact, γ = probability of contact (c),
δ_a = probability of outcome a given c

Figure 2: Tree diagram of potential outcome of a contact trial in a telephone survey.

Now, consider a sequence of contact trials, and let $\mathbf{P}_t$ denote a column vector of probabilities of the outcomes (r,$\bar{r}$, CL, $\bar{c}$) after t trials. The sequence of

trials can be modelled as a stochastic process with a transition matrix $\mathbf{\Gamma}_t$ which includes probabilities of the outcomes at trial t given the outcome on trial $t-1$. Then we can write $\mathbf{P}_t = \mathbf{\Gamma}_t \mathbf{P}_{t-1} (t \geq 2)$ with $\boldsymbol{P}_1 = \begin{pmatrix} \theta_1 & \theta_2 & \theta_3 & \bar{\gamma} \end{pmatrix}^t$.

There is some information on the transition matrix restricting the values on its entries. First, the response and nonresponse outcomes are both absorbing. That is, if a response is obtained on trial $t-1$ we still have a response on trial t, and similar for a nonresponse. Also, above the outcome CL is assumed to yield either a response or a nonresponse in the following trial.

Suppose $\mathbf{\Gamma}_t = \mathbf{\Gamma}_2$ for all $t \geq 2$, and consider

$$\mathbf{\Gamma}_2 = \begin{pmatrix} 1 & 0 & \theta_{31} & \theta_1 \\ 0 & 1 & \theta_{32} & \theta_2 \\ 0 & 0 & 0 & \theta_3 \\ 0 & 0 & 0 & \bar{\gamma} \end{pmatrix}$$

The fourth column in the matrix equals $\boldsymbol{P}_1$, meaning that the conditional probabilities in a trial following a series of no contacts are the same; they do not change with the number of trials made earlier. The third column contains the conditional probabilities of a response (θ_{31}) or nonresponse (θ_{32}) given the outcome CL in the previous trial. The first and second columns are obtained for the absorbing outcomes response and nonresponse, respectively.

With this model, the probabilities of the different outcomes can be expressed as $\boldsymbol{P}_t = \mathbf{\Gamma}_2^{t-1} \mathbf{P}_1$, and letting the number of trials converge to infinity yields the probability vector

$$\boldsymbol{P}_\infty = \begin{pmatrix} \delta_r + \theta_{31}\delta_{CL} \\ \delta_{\bar{r}} + \theta_{32}\delta_{CL} \\ 0 \\ 0 \end{pmatrix}$$

if $\bar{\gamma} < 1$.

This model gives a probability of a response from a unit being a function of three unknown probabilities, i.e., $Pr(r) = \delta_r + \theta_{31}\delta_{CL}$. Again modeling of response with e.g. the logit model is less appropriate with respect to the expression obtained.

Adding the assumption $\delta_{CL} = 0$ yields the traditional dichotomy of response/nonresponse, suggesting modelling δ_r with e.g., a normal or a logistic distribution function. The same modelling approach can also be motivated if $\theta_{31} = 0$.

A different case is obtained by setting $\theta_{31} = 1$ whereby $Pr(r) = \delta_r + \delta_{CL}$, and the modeling of $Pr(r)$ using probit or logit models seem less appropriate.

Rather, these models imply the modeling of nonresponse $Pr(\bar{r})$ due to their symmetry. With a distribution function F having a symmetric density and $Pr(\bar{r}) = F(-\mathbf{x}^t\alpha)$, then $Pr(r) = 1 - Pr(\bar{r}) = F(\mathbf{x}^t\alpha)$. If the distribution is asymmetric, modeling of nonresponse instead of response is more appropriate.

A final special case of interest is obtained with $\theta_{31} = \delta_r/(\delta_r + \delta_{\bar{r}})$, which corresponds to the independence of irrelevant alternative (IIA) assumption underlying the multinomial logit model (e.g. McFadden, 1978). With this specification, $Pr(r) = \delta_r/(\delta_r + \delta_{\bar{r}})$. Now suppose $\delta_a = e^{V_a}/(e^{V_r} + e^{V_{\bar{r}}} + e^{V_{CL}})$ $(a \in \{r, \bar{r}\})$, where $V_\cdot$ are nonrandom scalars. Then, $Pr(r) = e^{V_r}/(e^{V_r} + e^{V_{\bar{r}}}) = e^{V_D}/(1 + e^{V_D})$, where $V_D = V_r - V_{\bar{r}}$, and the logit model is obtained.

In the discrete choice literature (e.g. McFadden, 1978), V_r, $V_{\bar{r}}$ and V_{CL} represent systematic parts of the utilities of choosing alternatives r, $\bar{r}$ and CL, respectively. The utilities for the units are obtained by adding individual specific components ϵ_a $(a \in \{r, \bar{r}, CL\})$ yielding $U_a = V_a + \epsilon_a$. Under the maximum utility paradigm, the unit selects the alternative yielding maximum utility, that is, a unit responds if $U_r > max(U_{\bar{r}}, U_{CL})$.

Let $\mathbf{x}$ denote a vector characterizing the respondent and $V_a = \mathbf{x}^t\alpha_a$ such that $U_a = \mathbf{x}^t\alpha_a + \epsilon_a$ $(a \in \{r, \bar{r}, CL\})$. Suppose ϵ_a $(a \in \{r, \bar{r}, CL\})$ are independent and identically Gumbel distributed; then, $Pr(r) = e^{\mathbf{x}^t\alpha_D}/(1 + e^{\mathbf{x}^t\alpha_D})$, where $\alpha_D = \alpha_r - \alpha_{\bar{r}}$ (e.g. McFadden, 1978). Again, the logit model is obtained.

4 Discussion

Sampling theory shows how to utilize randomization to achieve valid, objective inferences from empirical observations. Its application in the social sciences and for official statistics production, however, is hampered by nonresponse because the theory assumes observations are obtained for all units in the sample.

There are early suggestions on how to correct for nonresponse in which the theory is applied in two or more steps. One example is the Hansen and Hurwitz (1946) method, in which a subset of the set of nonrespondents is sampled and measured. A similar idea is advanced by Bartholomew (1961). Again, however, for these theories to work in practice, full response is required when sampling from the subset of nonrespondents.

Later, the view of response as an outcome of a random trial was adopted. Oh and Scheuren (1983) consider this interpretation a quasi-randomization approach, treating the response set generated as a second sampling phase with an unknown second-phase sampling design. The idea makes standard

theory on estimation applicable by using estimated response probabilities.

In one direction, these response probabilities are estimated implicitly. In this case, reliance is attributed to the ability of the auxiliary variables in capturing response pattern given that these are also related to variables of interest. There are many results in the literature showing this approach to be successful in reducing bias due to nonresponse.

As pointed out earlier, auxiliary variables are generally used in attempting to achieve a MAR situation, a property that cannot be tested statistically. Also, the example illustrated in Section 2, the dependence on the relationships between the variables involved in the estimation can lead to undesired effects. The use of auxiliary variables in an attempt to reduce bias due to nonresponse can introduce a bias more severe than the one of the expansion estimator. This is an issue usually not addressed in the literature and has to be considered in future research.

Although our results point to the risk of introducing bias, the recommendation on using powerful auxiliary information still hold. If an auxiliary variable is e.g. strongly and positively correlated with both the study variable and the response probability, respectively, it implies a strong positive correlation between the study variable and the response probability as well. This follows by the results of e.g. Olkin (1981); if two correlations in a trivariate distribution are fixed, they bound the value on the third correlation. This result is also used by Schouten (2007) for selection of auxiliary variables.

What is needed are new tools and methods which can be used for judging when it is appropriate to use the auxiliary information available. This is an issue for future research and there are some ideas worthwhile studying. One idea is to further consider restrictions on correlations in multivariate distributions. These can be used to bound the correlation between the study variable and the response probability given estimates of the other correlations involved. It is also possible to construct simple tests of a zero correlation, that is, potential unbiasedness of the expansion estimator.

Another idea is to evaluate properties of estimators where the study variable itself is assumed to directly affect the response probability. Examples are the calibration estimators suggested by Chang and Kott (2008) and Särndal and Lundström (2005). Another example is Heckman's (1979) estimator for the sample selection model.

In another direction, explicit models are used in representation of the true response mechanism. Using a model by definition means use of approximations. A model cannot be assumed correct, and valid inference cannot be guaranteed from its application. An essential part for valid inference is how

well the model approximates the true response probabilities.

A popular response probability model is the binary logit model. Taking on a process view on the sequence of contact trials in a telephone survey, the example in Section 3 shows a response probability of a more complicated structure than what is implied by the logit model. The response probability obtained is defined by a function of three different probabilites. Thus, simple models of response probabilities do not capture the complex process of attempts to reach contacts and the choices of the units to respond or not. This discrepancy is a source of bias in estimation, and new models capturing the specific characteristics of the data collection process are of interest.

Graph models in combination with models for discrete choice data can here provide new tools for modeling response probabilities. It is interesting to note how contributions in the discrete choice literature can be adapted in modeling response probabilities. In particular, theories describing individual choice behavior is a source for improving model specifications.

References

An, A.B. (1996) Regression estimation for finite population means in the presence of nonresponse. Retrospective Theses and Dissertations. Paper 11357.

Bartholomew, B. J. (1961). A method for allowing for 'not-at-home' bias in sample surveys, *Journal of the Royal Statistical Society, Series C*, 10:1, 52-59.

Bethlehem, J. G. (1988). Reduction of nonresponse bias through regression estimation, *Journal of Official Statistics*, **4:3**, 251-260.

Bethlehem, J. G. and Keller, W. J. (1987). Linear Weighting of Sample Survey Data, *Journal of Official Statistics*, **3:2**, 141–153.

Bethlehem, J. and Schouten, B. (2004). Nonresponse adjustment in household surveys, *Discussion paper 04007. Statistics Netherlands.*

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

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

Cobben, F. (2009). Nonresponse in sample surveys : methods for analysis and adjustment, *Statistics Netherlands. http://hdl.handle.net/11245/2.69219.*

Estevão, V.M. and Särndal, C.-E. (2000). A functional form approach to calibration, *Journal of Official Statistics*, **16:4**, 379–399.

Falk, G. (2012). Calibration adjustment for nonresponse in cross-classified data.*Section on Survey Research-JSM 201.*

Fuller, W. A. (2002). Regression Estimation for Survey Samples, *Survey Methodology*, **28:1**, 5–23.

Fuller, W. A. and An, A.B. (1998). Regression Adjustment for Nonresponse, *Jour. Ind. Soc. Ag. Statistics,* **51**, 331–342.

Geuzinge, L., Rooijen, J. Van and Bakker, B.F.M. (2000), The use of administrative registers to reduce non-response bias in household surveys, *Netherlands Official Statistics* **2000:2**, 32–39.

Hansen, M. H. and Hurwitz, W. N. (1946) The Problem of Non-Response in Sample Surveys. *Journal of the American Statistical Association*, **41:236**, 517–529.

Heckman, J.J. (1979). Sample selection as a specification error, *Econometrica* **47:1**, 153-161.

Kalton, G. and Flores-Cervantes, I. (2003). Weighting methods, *Journal of Official Statistics,* **19:2**, 81–97.

Kim, J. K. and Kim, J. J. (2007). Nonresponse weighting adjustment us-

ing estimated response probabilities,*The Canadian Journal of Statistics,* **35:4**, 501–514.
Kim, J. K. and Park, M. (2010). Calibration Estimation in Survey Sampling. *International Statistical Review,* **78**, 21-39.
Kim, J. K. and Riddles, M. K. (2012). Some theory for propensity-score-adjustment estimators in survey sampling, *Survey Methodology*, **38:2**, 157–165.
Kott, P. S. (2013). Discussion, *Journal of Official Statistics*, **29:3**, 359–362.
Kreuter, F. and Olson, K. (2011). Multiple Auxiliary Variables in Nonresponse Adjustment. *Sociological Methods & Research,* **40:2**, 311–332.
Luce, R. D. (1959). *Individual Choice Behavior: A Theoretical Analysis.* Wiley, New York.
McFadden, D. (1978). Modelling the Choice of Residential Location. In A. Karlqvist, L. Lundqvist, F. Snickars and J. Weibull (eds), *Spatial Interaction Theory and Planning Models*, North-Holland, Amsterdam, pp. 75 -96.
Oh, H.L. and Scheuren, F. J. (1983). Weighting adjustment for unit nonresponse. In: Madow, W.G, Olkin, I. and Rubin, D.B. (Eds.), Incomplete Data in Sample Surveys:Vol 2, Academic Press, New York, pp. 143-184.
Rizzo, L., Kalton, G., and Brick, M. (1996). A comparison of some weighting adjustment methods for panel. *Survey Methodology,***22**, 43-53.
Särndal, C.-E. (1980). A Two-Way Classification of Regression Estimation Strategies in Probability Sampling, *The Canadian Journal of Statistics/ La Revue Canadienne de Statistique*, **8:2**, 165–177.
Särndal, C.-E. (2007). The calibration approach in survey theory and practice, *Survey Methodology,***33:2**, 99-119.
Särndal, C.-E. (2011). Three Factors to Signal Non-Response Bias With Applications to Categorical Auxiliary Variables, *International Statistical Review,* **79:2**, 233–254.
Särndal, C.-E. and Lundström, S. (2005). *Estimation in Surveys with Nonresponse.* Wiley, New York.
Särndal, C.-E. and Lundström, S. (2008). Assessing Auxiliary Vectors for Control of Nonresponse Bias in the Calibration Estimator. *Journal of Official Statistics*, **24:2**, 167–191.
Särndal, C.-E, Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling.* Springer, New York.
Schouten, B. (2007). A selection strategy for weighting variables under a Not-Missing-at-Random assumption, *Journal of Official Statistics,* **23**, 51–68.
Singh, H.P. and Kumar, S. (2011). Combination of regression and ratio estimate in presence of nonresponse. *Brazilian Journal of Probability and*

Statistics, **25:2**, 205–217 DOI: 10.1214/10-BJPS117

West, B.T. and Little, R.J.A. (2012). Non-response adjustment of survey estimates based on auxiliary variables subject to error, *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, **176**, 211–225.

Publications *in the series*
Örebro Studies in Statistics

1. Werner, Peter (2003). *On the Cost-Efficiency of Mixed Mode Surveys Using the Web*.
2. Wahlström, Helen (2004). *Nonparametric Tests for Comparing Two Treatments by Using Ordinal Data*.
3. Westling, Sara (2008). *Cost efficiency of nonresponse rate reduction efforts – an evaluation approach*.
4. Högberg, Hans (2010). *Some properties of measures of disagreement and disorder in paired ordinal data*.
5. Alam, Moudud Md. (2010). *Feasible computation of the generalized linear mixed models with application to credit risk modelling*.
6. Li, Dao (2013): *Common Features in Vector Nonlinear Time Series Models*.
7. Ding, Shutong (2014): *Model Choice in Bayesian VAR Models*.
8. Rota, Bernardo João (2016): *Calibration Adjustment for Non-response in Sample Surveys*.