Variance Estimation in Two-Step Calibration for Nonresponse Adjustment

April 12, 2016

Bernardo João Rota[1,2)]

[1)] Department of Statistics, Örebro University, 701 82 Örebro, Sweden

[2)] Statistics, Department of Mathematics and Informatics, Eduardo Mondlane University, Maputo, Mozambique.
Address: [1)] Fakultetsgatan 1, 702 81 Örebro, Sweden
[2)]Ave. Julius Nyerere/Campus Principal 3453, Maputo, Mozambique
E-mail: rotasitao@yahoo.com.br, bernardo.rota@oru.se
Tel: +46729141210

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.

Key words: 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 nonresponse-adjusted 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.

Rota and Laitila (2015) combine these calibration 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) $\mathbf{x}_k^p = (x_{1k}^p, x_{2k}^p, ..., x_{Jk}^p)^t$, a J-dimensional vector of known values for all elements k in the response set r; for each $j = 1, ..., J$, $T_{xj}^p = \sum_{k \epsilon U} x_{jk}^p$ is known. This implies that $T_x^p = (T_{x1}^p, T_{x2}^p, ..., T_{xJ}^p)^t$ is also known.
(b) $\mathbf{x}_k^s = (x_{1k}^s, x_{2k}^s, ..., x_{Lk}^s)^t$, an L-dimensional vector of known values for all elements k in the sample set, s. For each $l = 1, ..., L$, we can estimate $\hat{t}_{xl}^s = \sum_{k \epsilon s} d_k x_{lk}^s$ and compose the vector $\hat{t}_x^s = \left(\hat{t}_{x1}^s, \hat{t}_{x2}^s, ..., \hat{t}_{xL}^s\right)^t$.

Unless otherwise stated, the expected value $E_pE_q(A)$, 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 $\sum_{k\epsilon r} w_k \mathbf{x}_k = \mathbf{X}$, where $\mathbf{x}_k$ stands for $\mathbf{x}_k^p$, $\mathbf{x}_k^s$, or $\mathbf{x}_k = \left((\mathbf{x}_k^p)^t, (\mathbf{x}_k^s)^t\right)^t$ and $\mathbf{X}$ corresponds to their respective totals, i.e., T_x^p, $\hat{t}_x^s$, or $\left((T_x^p)^t, (\hat{t}_x^s)^t\right)^t$. 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 to be the 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 $w_k = d_k + d_k\mathbf{g}^t\mathbf{x}_k$, which are linear in the coefficient vector $\mathbf{g}^t = \left(\mathbf{X} - \sum_{k\epsilon r} d_k\mathbf{x}_k\right)^t \left(\sum_{k\epsilon r} d_k\mathbf{x}_k\mathbf{x}_k^t\right)^{-1}$. The resulting estimator of Y, commonly termed a linear calibration estimator, is given by

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

Other distance functions will generally produce calibrated weights that are nonlinear in their coefficients, so deriving the 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 $\sum_{k\epsilon r} d_k F(\cdot)\mathbf{x}_k = \mathbf{X}$ 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 the dimension of $\mathbf{x}_k$ is no less than that of $\mathbf{z}_k$, and 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 $\sum_{k\epsilon r} w_{1k}\mathbf{x}_k^s = \sum_{k\epsilon s} d_k \mathbf{x}_k^s$. 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 $\sum_{k\epsilon r} w_{2k}\mathbf{x}_k = \mathbf{X}$, where $\mathbf{x}_k = \mathbf{x}_k^p$ or $\mathbf{x}_k = \left((\mathbf{x}_k^p)^t, (\mathbf{x}_k^s)^t\right)^t$.

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 $q_k = q((\mathbf{x}_k^s)^t \mathbf{g})$.

In the rest of the paper we use $\hat{F}_k = F\left((\mathbf{x}_k^s)^t \hat{\mathbf{g}}\right)$, $F_k = F\left((\mathbf{x}_k^s)^t \mathbf{g}\right)$, and $F_k^\circ = F\left((\mathbf{x}_k^s)^t \mathbf{g}_\circ\right)$, where $\mathbf{g}$ is the generic parameter vector, $\mathbf{g}_\circ$ is the true value of $\mathbf{g}$, $\hat{\mathbf{g}}$ is an 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 $\sum_{k\epsilon r} d_k F_k \mathbf{x}_k^s = \hat{t}_x^s$. 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 $T_x^p = \sum_{k\epsilon r} w_{2k}\mathbf{x}_k^p$ and given by $w_{2k} = w_{1k}v_{2k}$ with $v_{2k} = 1 + \mathbf{g}_2^t \mathbf{x}_k^p$ and $\mathbf{g}_2^t = (T_x^p - \sum_{k\epsilon r} w_{1k}\mathbf{x}_k^p)^t \left(\sum_{k\epsilon r} w_{1k}\mathbf{x}_k^p (\mathbf{x}_k^p)^t\right)^{-1}$, assuming that $\sum_{k\epsilon r} w_{1k}\mathbf{x}_k^p (\mathbf{x}_k^p)^t$ 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:

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

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

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) $\mathbf{v}_k = (\mathbf{x}_k^p, \mathbf{x}_k^s, y_k)$ is nonrandom and $\|\mathbf{v}_k\| < \infty$.
(iv) $\left(\hat{\mathbf{B}}_{2Fr} - \mathbf{B}_2\right)$, $N^{-1}\left(T_x^p - \sum_{k\epsilon r} d_k F_k^\circ \mathbf{x}_k^p\right)$, and $N^{-1}\left(\hat{t}_x^s - \sum_{k\epsilon r} d_k F_k^\circ \mathbf{x}_k^s\right)$ are all $O_p(n^{-\frac{1}{2}})$, where $\mathbf{B}_2 = \left(\sum_{k\epsilon U} \mathbf{x}_k^p \left(\mathbf{x}_k^p\right)^t\right)^{-1} \sum_{k\epsilon U} \mathbf{x}_k^p y_k$ is the population analogous to $\hat{\mathbf{B}}_{2Fr}$.
(v) $N^{-1}\sum_{k\epsilon r} d_k \mathbf{x}_k^P \mathbf{F}_{1k}^\circ$ and $N^{-1}\sum_{k\epsilon r} d_k F_k^\circ \mathbf{x}_k^p \left(\mathbf{x}_k^p\right)^t$ are $O_p(1)$, where, $\mathbf{F}_1 = dF/d\mathbf{g}$.

The bias of the two-step estimator is given by $E\left(\hat{Y}_{2step}\right) - Y = E\left((T_x^p - \sum_r d_k \hat{F}_k \mathbf{x}_k^p)^t \hat{\mathbf{B}}_{2Fr}\right)$, which is of order $O\left(Nn^{-\frac{1}{2}}\right)$.
Given that $\hat{\mathbf{g}}$ is a solution to $\sum_{k\epsilon r} d_k F_k \mathbf{x}_k^s = \hat{t}_x^s$, we proceed as follows:

$\sum_{k\epsilon r} d_k F_k^\circ \mathbf{x}_k^s - \hat{t}_x^s = \sum_{k\epsilon r} d_k \hat{F}_k \mathbf{x}_k^s - \hat{t}_x^s + \sum_{k\epsilon r} d_k \mathbf{x}_k^s \tilde{\mathbf{F}}_{1k}\left(\hat{\mathbf{g}} - \mathbf{g}_\circ\right) = O_p\left(Nn^{-\frac{1}{2}}\right)$. This leads to equation (4) below:

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

where $\mathbf{\Gamma}$ is the probability limit of $N^{-1}\sum_{k\epsilon r} d_k \mathbf{x}_k^s \tilde{F}_{1k}$, assumed invertible and $\tilde{\mathbf{F}}_{1k} = \mathbf{F}_{1k}\left((\mathbf{x}_k^s)^t \tilde{\mathbf{g}}\right)$, 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:

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

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

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

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

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

The bias of $\hat{Y}^{\circ}_{2step}$ is given by $E\left(\hat{Y}^{\circ}_{2step}\right) - Y = E\left(\sum_{k\epsilon r} d_k \mathbf{F}^{\circ}_{1k} E_k \left(\hat{\mathbf{g}} - \mathbf{g}_{\circ}\right)\right) + E\left(\mathbf{R}\right)$, in which the first term on the r.h.s. is $O(Nn^{-\frac{1}{2}})$ and the second is $O(Nn^{-1})$. Thus, like the bias of $\hat{Y}_{2step}$, the bias of $\hat{Y}^{\circ}_{2step}$ is of order $O(Nn^{-\frac{1}{2}})$. Given this, in equation (6), we drop the lower-order term $\mathbf{R}$ and obtain the approximate expression for the two-step estimator of Y :

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

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

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

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:

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

where $A_{kl} = d_k d_l \left(\mathbf{x}^s_l\right)^t \left(\mathbf{F}^{\circ}_{1k} \tilde{\boldsymbol{\Gamma}}^{-1}\right)^t E_k$, 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:

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

$Var(\hat{T}^{\circ}_b) = \sum_{k\neq l\neq i\epsilon U} \frac{\pi_{kli}(F^{\circ}_l - 1)}{F^{\circ}_k F^{\circ}_i} A_{kl} A_{il} + \sum_{k\neq l\epsilon U} \frac{\pi_{kl} - \pi_k \pi_l}{F^{\circ}_k F^{\circ}_l} (F^{\circ}_l - 1)(F^{\circ}_k - 1) A_{kk} A_{ll} +$

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

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

$$\hat{V}ar\left(\hat{Y}^{\bullet}_{2step}\right) = \hat{V}ar(\hat{T}_a)+\hat{V}ar(\hat{T}_b)+2\hat{C}ov\left(\hat{T}_b, \hat{T}_a\right) \tag{11}$$

where

$\hat{V}ar(\hat{T}_a) = \sum_{k\neq l\epsilon r}(d_k d_l - d_{kl})\check{e}_k\check{e}_l + \sum_{k\epsilon r} d_k\hat{F}_k(d_k\hat{F}_k-1)e_k^2,$

$\hat{V}ar(\hat{T}_b) = \sum_{k\neq l\neq i\epsilon r} \hat{F}_l(\hat{F}_l-1)\hat{A}_{kl}\hat{A}_{il} +\sum_{k\neq l\epsilon r}(1-d_{kl}\pi_k\pi_l)(\hat{F}_k-1)(\hat{F}_l-1)\hat{A}_{kk}\hat{A}_{ll}$
$\sum_{k\neq l\epsilon r} \hat{F}_l(\hat{F}_l-1)\hat{A}_{kl}^2+\sum_{k\neq l\epsilon r}(1-\hat{F}_k)(1-\hat{F}_l)\hat{A}_{kl}\hat{A}_{lk} +\sum_{k\neq l\epsilon r}(1-\hat{F}_l)^2\hat{A}_{kl}\hat{A}_{ll}+$
$\sum_{k\epsilon r} \frac{(\hat{F}_k-1)^2(\hat{F}_k-\pi_k)}{\hat{F}_k}\hat{A}_{kk}^2,$

and

$\hat{C}ov\left(\hat{T}_b, \hat{T}_a\right) = \sum_{k\neq l\epsilon r} d_l\left((\hat{F}_l-1)\hat{A}_{kl}+(\hat{F}_k-1)\hat{A}_{kk}\right)\check{e}_l+\sum_{k\epsilon r} d_k(\hat{F}_k-1)\hat{A}_{kk}\check{e}_k-$
$\sum_{k,l\epsilon r} d_l(\hat{F}_k-1)\hat{A}_{kk}\check{e}_l,$

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

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

$$\hat{\mathbf{H}}(\mathbf{g}) = \sum_{k \epsilon r} d_k F_k \mathbf{x}_k^s - \hat{t}_x^s \tag{12}$$

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

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

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:

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

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

Now, suppose that $t_x^s = \sum_U \mathbf{x}_k^s$ is known. Then (12) is defined as:

$$\hat{\hat{\mathbf{H}}}(\mathbf{g}) = \sum_{k \epsilon r} d_k F_k \mathbf{x}_k^s - t_x^s \tag{15}$$

with the same properties as before except that

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

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

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

where $\boldsymbol{\Phi}$ and $\boldsymbol{\Psi}$ are the first and second components of $plim_{n\to\infty} n^{-1} E\left(\tilde{\hat{\mathbf{H}}}(\mathbf{g}_{\circ})\tilde{\hat{\mathbf{H}}}^{t}(\mathbf{g}_{\circ})\right)$, 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 $\tilde{\hat{\mathbf{g}}}$ be the solution to $\tilde{\hat{\mathbf{H}}}(\mathbf{g}) = \mathbf{0}$ and $\tilde{\hat{Y}}^{\bullet}_{2step} = \hat{T}^{\circ}_{a} + \hat{T}^{\circ}_{c}(\tilde{\hat{\mathbf{g}}} - \mathbf{g}_{\circ}) + \left(T^{P}_{x}\right)^{t}\mathbf{B}_{2}$ is the corresponding equation (7) when $\hat{\mathbf{g}}$ is replaced with $\tilde{\hat{\mathbf{g}}}$. 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 $\tilde{\hat{\mathbf{g}}}$, $\hat{T}^{\circ}_{c}$ 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

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

Thus, the efficiency loss of $\tilde{\hat{\mathbf{g}}}$ resulting from calibrating with population-level auxiliary information implies 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/study variable and model variable/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.

Table 1: Simulation results of study 1, first case

Sample size	Estimator of Y	Rel.bias of $\hat{Y}$ in (%)	RMSE of $\hat{Y}$	Rel.bias of $\hat{V}ar\left(\hat{Y}\right)$in (%)	RMSE of $\hat{V}ar\left(\hat{Y}\right)$	CI coverage rate
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 $\hat{Y}$ in (%)	RMSE of $\hat{Y}$	Rel.bias of $\hat{V}ar\left(\hat{Y}\right)$in (%)	RMSE of $\hat{V}ar\left(\hat{Y}\right)$	CI coverage rate
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 $\hat{Y}$ in (%)	RMSE of $\hat{Y}$	Rel.bias of $\hat{V}ar\left(\hat{Y}\right)$in (%)	RMSE of $\hat{V}ar\left(\hat{Y}\right)$	CI coverage rate
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

Table 4: Simulation results of study 2, second case

Sample size	Estimator of Y	Rel.bias of $\hat{Y}$ in (%)	RMSE of $\hat{Y}$	Rel.bias of $\hat{V}ar\left(\hat{Y}\right)$in (%)	RMSE of $\hat{V}ar\left(\hat{Y}\right)$	CI coverage rate
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

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 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 large bias for $\hat{Y}_{2step}$ than $\hat{Y}_{2LC}$ may be because $\mathbf{x}_k^s$ 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 $\mathbf{x}_k^s$ 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 of the point estimator $\hat{Y}_{2step}$ that is no less than that of the benchmark point 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 $\hat{V}_{uncor}$; for large samples this correction is less important, as we expect $\hat{\mathbf{g}}$ to be close to $\mathbf{g}_{\circ}$. 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. J Off Stat 29:329–353

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

Deville JC, Särndal CE (1992) Calibration estimators in survey sampling. J Am Stat Assoc 87:376–382

Deville JC, Särndal CE, Sautory O (1993) Generalized raking procedures in survey sampling. J Am Stat Assoc 88:1013–1020

Estevão VM, Särndal CE (2002) The Ten Cases of auxiliary Information for Calibration in Two-Phase Sampling. J Off Stat 18:233–255

Estevão VM, Särndal CE (2006) Survey Estimates by Calibration on Complex Auxiliary Information. Int Stat Rev 74:127-147

Kim JK, Park M (2010) Calibration Estimation in Survey Sampling. Int Stat Rev 78:21-39. doi:10.1111/j.1751-5823.2010.00099.x

Kott PS, Day CD (2014). Developing Calibration Weights and Standard-Error Estimates for a Survey of Drug-Related Emergency-Department Visits. J Off Stat, 30:521-532.

Kott PS, Liao D (2015) One step or two? Calibration weighting from a complete list frame with nonresponse. Surv Methodol 41:165–181

Kreuter F, Olson K (2011) Multiple auxiliary variables in nonresponse adjustment. Sociol Meth Res 40:311–332

Lundström S, Särndal C-E (1999) Calibration as a standard method for treatment of nonresponse. J Off Stat 15:305–327

Niyonsenga T (1997) Response probability estimation. J Stat Plan Infer 59:111-126

Rizzo L, Kalton G, Brick M (1996) A comparison of some weighting adjustment methods for panel nonresponse. Surv Methodol 22:43–53

Rota BJ, Laitila T (2015) Comparisons of some weighting methods for nonresponse adjustment. Lith J Stat 54:69–83

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

Särndal C-E, Lundström S (2007) Assessing auxiliary vectors for control of nonresponse bias in the calibration estimator. J Off Stat 24:167–191

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

Singh AC, Wu S, 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

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
$\hat{T}^{\circ}_{a} = \sum_{k\epsilon s} R_k d_k F^{\circ}_{k} E_k$, $\hat{T}^{\circ}_{b} = \sum_{k,l\epsilon s} R_k(1{-}R_l F^{\circ}_{l}{-}1)A_{kl}$, and $A_{kl} = d_k d_l \left(\mathbf{x}^{s}_{l}\right)^{t} \left(\mathbf{F}^{\circ}_{1k}\mathbf{\Gamma}^{-1}\right)^{t} E_k$.

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
$E_p V_q(\sum_{k,l\epsilon s} R_k(R_l F^{\circ}_{l} - 1)A_{kl}) {=} E_{pq} \sum_{k,l,i,j\epsilon U} \left(M_{kl} - E_q(M_{kl})\right)\left(M_{ij} - E_q(M_{ij})\right)$
with $M_{ab} = I_a I_b R_a(R_b F^{\circ}_{b} - 1)A_{ab}$. This leads to
$E_p V_q(\sum_{k,l\epsilon s} R_k(R_l F^{\circ}_{l} - 1)A_{kl}) = S_1 + S_2 + S_3 + 2S_4 + S_5$,
$S_1 = \sum_{k\neq l\neq i\epsilon U} \frac{\pi_{kli}}{F^{\circ}_{k}F^{\circ}_{i}}(F^{\circ}_{l} - 1)A_{kl}A_{il}$, $S_2 = \sum_{k\neq l\epsilon U} \frac{\pi_{kl}}{F^{\circ}_{k}}(F^{\circ}_{l} - 1)A^{2}_{kl}$,
$S_3 = \sum_{k\neq l\epsilon U} \frac{\pi_{kl}}{F^{\circ}_{k}F^{\circ}_{l}}(1 - F^{\circ}_{k})(1 - F^{\circ}_{l})A_{kl}A_{lk}$, $S_4 = \sum_{k\neq l\epsilon U} \frac{\pi_{kl}}{F^{\circ}_{k}F^{\circ}_{l}}(1 - F^{\circ}_{l})^2 A_{kl}A_{ll}$, 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 $C_1 = \sum_{k\neq i\epsilon U} \frac{d_i\pi_{ki}(F^{\circ}_{k}-1)}{F^{\circ}_{k}} A_{kk}E_i - \sum_{k\epsilon U} \frac{\pi_k(1-F^{\circ}_{k})}{F^{\circ}_{k}} A_{kk} \sum_{i\epsilon U} E_i$ 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 $\sum_{i\epsilon U} E_i = E_{pq}(\hat{T}^{\circ}_{a})$.