VARIANCE ESTIMATION FOR THE TWO-PHASE REGRESSION ESTIMATOR - A CALIBRATION APPROACH

Martin Axelson, Statistics Sweden
Research and Development, Statistics Sweden, SE-701 89 Örebro, Sweden

Key Words: Two-phase sampling, Regression estimator, Variance estimation, Calibration

1 Introduction

In situations where weak or no auxiliary information is available at the population level, two-phase regression estimation constitutes a possible tool for cost-efficient estimation. In this respect, important references are Särndal and Swensson (1987), Dupont (1995), and Hidiroglou and Särndal (1998). Särndal and Swensson presented general results regarding generalized regression estimation under two-phase samling, while both Dupont and Hidiroglou and Särndal discussed two-phase calibration estimation and its possible relation to generalized regression estimation. Hidiroglou and Särndal showed that under their suggested calibration approach, the generalized regression estimator (*GREG*) under two-phase sampling alternatively may derived using a two-step calibration approach, a result in line with the findings regarding the relation between regression and calibration estimation under single-phase sampling in Deville and Särndal (1992).

In this paper we focus on variance estimation for the *GREG* under two-phase sampling, and the question we seek to answer is whether or not the available auxiliary information may be used more extensively than is generally the case, in order to obtain more efficient variance estimators. This is by no means a new question (e.g. Rao and Sitter, 1995; Sitter, 1997; Axelson, Breidt, and Carriquiry, 1996), but to our knowledge, the approach presented in this paper is a new development. The main goal of this paper is to extend the calibration technique to allow not only for point estimation purposes under two-phase sampling, but for variance estimation purposes as well. A general framework for calibration estimation of the variance of the two-phase regression estimator will be presented in detail and the method will be evaluated empirically through a small-scale simulation study. Related work under single-phase sampling includes Singh, Horn, and Yu (1998) and Théberge (1999)

The paper is organized as follows. In Section 2 some basic notation is introduced. In Section 3 the *GREG* is defined and a large-sample approximation to its variance is given. Section 4 deals with variance estimation. The standard approach is presented in Section 4.1, while a detailed account of the proposed calibration approach is given in Section 4.2. In Section 5, finally, the results of a small-scale Monte Carlo study are presented.

2 Preliminaries

Let $U = \{1, \ldots, k, \ldots, N\}$ denote the finite population of interest. Associated with each population element $k \in U$ there are fixed values of of the auxiliary column vector $\mathbf{x}_1$ and the variable of interest y, respectively. The value of $\mathbf{x}_1$, $\mathbf{x}_{1k}$, is known for all $k \in U$, and the objective of the survey is to estimate $t_y = \sum_U y_k$, the finite population total of y, which is unknown at the outset of the study. (If S_1 is any set of elements such that $S_1 \subseteq U$ and a_k is a quantity associated with element k, $\sum_{S_1} a_k$ is our shorthand for $\sum_{k \in S_1} a_k$.) To estimate t_y, two-phase regression estimation will be used.

Let s_a be a sample drawn from U according to a sampling design $p_a(\cdot)$. For all n_a elements included in the first-phase sample s_a, information on the auxiliary vector $\mathbf{x}$ is recorded. (For a discussion about the possible relationships between the auxiliary variables $\mathbf{x}_1$ and $\mathbf{x}$, see Hidiroglou and Särndal, 1998.) While $\mathbf{x}_1$ is often primarily of an administrative nature, $\mathbf{x}$ is typically chosen because it is assumed to be a powerful, yet relatively inexpensive, predictor for y. However, since both of the auxiliary variables $\mathbf{x}_1$ and $\mathbf{x}$ eventually will serve as predictors for the study variable y in the *GREG*, we assume that each variable in itself has a predictive ability strong enough to motivate the use of regression estimation. Next, a sample s is drawn from s_a according to a sampling design $p(\cdot|s_a)$, and the value of the study variable, y_k, is recorded for the n elements included in the second-phase sample. Throughout the paper, it is assumed that s_a and s are organized in such a way that the elements are sorted after increasing size of k. When necessary, the ν_ath element in s_a will be referred to as k_{ν_a} ($\nu_a = 1, \ldots, n_a$), while the νth element in s will be referred to as k^*_ν ($\nu = 1, \ldots, n$).

The first-phase first- and second-order inclusion probabilities induced by $p_a(\cdot)$ are denoted π_{ak} and π_{akl}, respectively, while the conditional first-

and second-order inclusion probabilities induced by $p(\cdot|s_a)$ are denoted $\pi_{k|s_a}$ and $\pi_{kl|s_a}$, respectively. In this paper, we only consider two-phase designs such that (i) $\pi_{akl} > 0$ for all $k\&l \in U$ and (ii) $\pi_{kl|s_a} > 0$ for all $k\&l \in s_a$ and every s_a.

For any quantity associated with element k, we let $\check{}$ symbolize division by π_{ak} and let $\check{\check{}}$ symbolize division by $\pi_{ak}\pi_{k|s_a}$. Thus, for example, $\check{y}_k = y_k/\pi_{ak}$, which is defined for all $k \in U$, and $\check{\check{y}}_k = \check{y}_k/\pi_{k|s_a} = y_k/(\pi_{ak}\pi_{k|s_a})$, which is defined for all $k \in s_a$.

3 Generalized regression estimation under two-phase sampling

Let $\hat{t}_{ys} = \sum_s \check{\check{y}}_k$ and let $\hat{\mathbf{t}}_{xs}$ be analogous to $\hat{t}_{ys}$. Moreover, let $\hat{\mathbf{t}}_{x_1 s_a} = \sum_{s_a} \check{\mathbf{x}}_{1k}$, let $\hat{\mathbf{t}}_{xs_a}$ be analogous to $\hat{\mathbf{t}}_{x_1 s_a}$, and let $\mathbf{t}_{x_1} = \sum_U \mathbf{x}_{1k}$. In line with Särndal and Swensson (1987) and Särndal et al. (1992, section 9.7), we define the *GREG* under two-phase sampling as

$$\begin{aligned}\hat{t}_{yr} = \hat{t}_{ys} &+ (\hat{\mathbf{t}}_{xs_a} - \hat{\mathbf{t}}_{xs})'\widehat{\mathbf{B}}_s \\ &+ (\mathbf{t}_{x_1} - \hat{\mathbf{t}}_{x_1 s_a})'\widehat{\mathbf{B}}_{1s},\end{aligned} \tag{1}$$

where

$$\widehat{\mathbf{B}}_s = (\textstyle\sum_s \mathbf{x}_k \check{\check{\mathbf{x}}}_k'/c_k)^{-1} \sum_s \mathbf{x}_k \check{\check{y}}_k/c_k \tag{2}$$

and

$$\widehat{\mathbf{B}}_{1s} = (\textstyle\sum_{s_a} \mathbf{x}_{1k} \check{\mathbf{x}}_{1k}'/c_{1k})^{-1} \sum_{s_a} \mathbf{x}_{1k} \check{y}_k^{\circ}/c_{1k}, \tag{3}$$

with

$$\check{y}_k^{\circ} = \begin{cases} \check{\mathbf{x}}_k'\widehat{\mathbf{B}}_s & \text{if } k \in s_a - s. \\ \check{\check{y}}_k + (\check{\mathbf{x}}_k - \check{\check{\mathbf{x}}}_k)'\widehat{\mathbf{B}}_s & \text{if } k \in s. \end{cases}$$

In (2), c is a weight which is assigned after the first phase of sampling but prior to the second phase of sampling. The weight serves to reflect the relative importance the statistician is willing to assign to element k on the basis of the auxiliary information available for $k \in s_a$. Similarly, in (3), c_1 is a weight which serves to reflect the relative importance the statistician is willing to assign to element k on the basis of the auxiliary information available for $k \in U$.

Define

$$\begin{aligned} g_{1ks_a} = 1 &+ (\mathbf{t}_{x_1} - \hat{\mathbf{t}}_{x_1 s_a})' \\ &\times (\textstyle\sum_{s_a} \mathbf{x}_{1k} \check{\mathbf{x}}_{1k}'/c_{1k})^{-1} \mathbf{x}_{1k}/c_{1k} \end{aligned}$$

for $k \in s_a$, and

$$\begin{aligned} g_{2ks} = 1 &+ (\textstyle\sum_{s_a} g_{1ks_a} \check{\mathbf{x}}_k - \sum_s g_{1ks_a} \check{\check{\mathbf{x}}}_k)' \\ &\times (\textstyle\sum_s \mathbf{x}_k \check{\check{\mathbf{x}}}_k'/c_k)^{-1} \mathbf{x}_k/c_k \end{aligned}$$

for $k \in s$, and let $g_{ks} = g_{1ks_a} + g_{2ks} - 1$, which thus is defined $k \in s$. It is matter of algebra to show that an alternative expression for (1) is given by

$$\hat{t}_{yr} = \textstyle\sum_s g_{ks} \check{\check{y}}_k.$$

This expression is due to Hidiroglou and Särndal (1998), who derived $\hat{t}_{yr}$ using a two-phase calibration approach.

The variance of $\hat{t}_{yr}$ may be written as $V(\hat{t}_{yr}) = V_1 + V_2$, where $V_1 = V_{p_a}[E(\hat{t}_{yr}|s_a)]$ and $V_2 = E_{p_a}[V(\hat{t}_{yr}|s_a)]$ sometimes are referred to as the first- and second-phase variance component, respectively. Let

$$\mathbf{B}_{1U} = (\textstyle\sum_U \mathbf{x}_{1k}\mathbf{x}_{1k}'/c_{1k})^{-1} \sum_U \mathbf{x}_{1k} y_k/c_{1k},$$

let

$$\widehat{\mathbf{B}}_{s_a} = (\textstyle\sum_{s_a} \mathbf{x}_k \check{\mathbf{x}}_k'/c_k)^{-1} \sum_{s_a} \mathbf{x}_k \check{y}_k/c_k,$$

and define the prediction errors $E_{1kU} = y_k - \mathbf{x}_{1k}'\mathbf{B}_{1U}$ and $E_{ks_a} = y_k - \mathbf{x}_k'\widehat{\mathbf{B}}_{s_a}$, which are defined for $k \in U$ and $k \in s_a$, respectively. Moreover, let

$$\check{\mathbf{E}}_{1U} = (\check{E}_{11U}, \ldots, \check{E}_{1kU}, \ldots, \check{E}_{1NU})'$$

and

$$\check{\check{\mathbf{E}}}_{s_a} = (\check{\check{E}}_{k_1 s_a}, \ldots, \check{\check{E}}_{k_{\nu_a} s_a}, \ldots, \check{\check{E}}_{k_{n_a} s_a})',$$

and define the matrices $\boldsymbol{\Delta}_{1U} = [\Delta_{akl}]$ and $\boldsymbol{\Delta}_{s_a} = [\Delta_{kl|s_a}]$, where $\Delta_{akl} = \pi_{akl} - \pi_{ak}\pi_{al}$ $(k\&l \in U)$ and $\Delta_{kl|s_a} = \pi_{kl|s_a} - \pi_{k|s_a}\pi_{l|s_a}$ $(k\&l \in s_a)$. Large sample approximations for V_1 and V_2 are given by

$$V_1 \doteq AV_1 = \check{\mathbf{E}}_{1U}'\boldsymbol{\Delta}_{1U}\check{\mathbf{E}}_{1U} \tag{4}$$

and

$$V_2 \doteq AV_2 = E_{p_a}[AV(\hat{t}_{yr}|s_a)], \tag{5}$$

where $AV(\hat{t}_{yr}|s_a) = \check{\check{\mathbf{E}}}_{s_a}'\boldsymbol{\Delta}_{s_a}\check{\check{\mathbf{E}}}_{s_a}$, respectively.

4 Variance estimation

4.1 The standard approach

For $k \in s$, let $e_{1ks} = y_k - \mathbf{x}_{1k}'\widehat{\mathbf{B}}_{1s}$ and $e_{ks} = y_k - \mathbf{x}_k'\widehat{\mathbf{B}}_s$, and define the vectors

$$\check{\mathbf{e}}_{1s} = (\check{e}_{1k_1^* s}, \ldots, \check{e}_{1k_\nu^* s}, \ldots, \check{e}_{1k_n^* s})'$$

and

$$\check{\check{\mathbf{e}}}_s = (\check{\check{e}}_{k_1^* s}, \ldots, \check{\check{e}}_{k_\nu^* s}, \ldots, \check{\check{e}}_{k_n^* s})'.$$

Moreover, define the matrices $\check{\check{\boldsymbol{\Delta}}}_{1s} = [\check{\check{\Delta}}_{akl}]$ and $\check{\boldsymbol{\Delta}}_s = [\check{\Delta}_{kl|s_a}]$, where $\check{\check{\Delta}}_{akl} = \Delta_{akl}/(\pi_{akl}\pi_{kl|s_a})$

$(k\&l \in s)$ and $\check{\Delta}_{kl|s_a} = \Delta_{kl|s_a}/\pi_{kl|s_a}$ $(k\&l \in s)$. From (4) and (5) it follows that a possible estimator for $V(\hat{t}_{yr})$ is given by

$$\hat{V}_{REF} = \hat{V}_{REF,1} + \hat{V}_{REF,2}, \tag{6}$$

where

$$\hat{V}_{REF,1} = \check{\mathbf{e}}'_{1s}\check{\check{\mathbf{\Delta}}}_{1s}\check{\mathbf{e}}_{1s}$$

and

$$\hat{V}_{REF,2} = \check{\mathbf{e}}'_{s}\check{\mathbf{\Delta}}_{s}\check{\mathbf{e}}_{s}$$

When the sample size is large in each of the two phases, (6) is approximately unbiased for $V(\hat{t}_{yr})$ (e.g., Särndal and Swensson, 1987).

Remark 1. Let

$$\mathbf{G}_{1s} = \mathrm{diag}(g_{1k_1^* s_a}, \ldots, g_{1k_\nu^* s_a}, \ldots, g_{1k_n^* s_a})$$

and

$$\mathbf{G}_{s} = \mathrm{diag}(g_{k_1^* s}, \ldots, g_{k_\nu^* s}, \ldots, g_{k_n^* s}.$$

As an alternative to $\hat{V}_{REF}$, Axelson (2000a) proposed

$$\hat{V}^{g_1 g}_{REF} = \hat{V}^{g_1}_{REF,1} + \hat{V}^{g}_{REF,2},$$

where

$$\hat{V}^{g_1}_{REF,1} = \check{\mathbf{e}}'_{1s}\mathbf{G}_{1s}\check{\check{\mathbf{\Delta}}}_{1s}\mathbf{G}_{1s}\check{\mathbf{e}}_{1s}$$

and

$$\hat{V}^{g}_{REF,2} = \check{\mathbf{e}}'_{s}\mathbf{G}_{s}\check{\mathbf{\Delta}}_{s}\mathbf{G}_{s}\check{\mathbf{e}}_{s}.$$

Other alternatives to $\hat{V}_{REF}$ based on the so-called g-weighted residual technique have been proposed by Särndal et al. (1992, section 9.7) and Hidiroglou and Särndal (1998).

4.2 The calibration approach

In this section we show how the results in Théberge (1999) regarding calibration estimation of bilinear estimators may be extended to allow for potentially efficient estimation of $V(\hat{t}_{yr})$. To this end, we need to introduce some new notation. Let $\mathbf{z}_1$ $(J_1 \times 1)$ and $\mathbf{z}$ $(J \times 1)$ denote vectors such that $\mathbf{z}_{1k}$ is known for (i) $k \in U$ or (ii) $k \in s_a$, and $\mathbf{z}_k$ is known for $k \in s_a$, and define the matrices

$$\mathbf{Z}_{1U} = (\mathbf{z}_{11}, \ldots, \mathbf{z}_{1k}, \ldots, \mathbf{z}_{1N})',$$

$$\mathbf{Z}_{1s_a} = (\mathbf{z}_{1k_1}, \ldots, \mathbf{z}_{1k_{\nu_a}}, \ldots, \mathbf{z}_{1k_{n_a}})',$$

$$\mathbf{Z}_{1s} = (\mathbf{z}_{1k_1^*}, \ldots, \mathbf{z}_{1k_\nu^*}, \ldots, \mathbf{z}_{1k_n^*})',$$

$$\mathbf{Z}_{s_a} = (\mathbf{z}_{k_1}, \ldots, \mathbf{z}_{k_{\nu_a}}, \ldots, \mathbf{z}_{k_{n_a}})',$$

and

$$\mathbf{Z}_{s} = (\mathbf{z}_{k_1^*}, \ldots, \mathbf{z}_{k_\nu^*}, \ldots, \mathbf{z}_{k_n^*})'.$$

In deriving a calibration estimator for $V(\hat{t}_{yr})$, it is assumed that $\mathbf{z}_1$ and $\mathbf{z}$ may be regarded as auxiliary vectors for the residuals $\check{e}_{1U}$ and $\check{\check{e}}_{s_a}$, respectively.

Let $\mathbf{F}_1$ $(A \times B)$, let $\mathrm{vec}(\mathbf{F}_1)$ denote the vector obtained by stacking the successive columns of $\mathbf{F}_1$ with the first column on top, and let $\mathbf{F}_2$ $(AB \times AB)$ be a positive diagonal matrix. Following Théberge, we define the distance measure $\|\mathbf{F}_1\|^2_{\mathbf{F}_2} = \|\mathrm{vec}(\mathbf{F}_1)\|^2_{\mathbf{F}_2} = \mathrm{vec}(\mathbf{F}_1)'\mathbf{F}_2\mathrm{vec}(\mathbf{F}_1)$. Now, a calibration estimator for V_1 based on $\mathbf{Z}_{1U}$ is given by

$$\hat{V}_{CAL,1} = \check{\mathbf{e}}'_{1s}\mathbf{W}_{1s}\check{\mathbf{e}}_{1s}, \tag{7}$$

where $\mathbf{W}_{1s}$ is the matrix which minimizes

$$\|\mathbf{W}^*_{1s} - \check{\check{\mathbf{\Delta}}}_{1s}\|^2_{\mathbf{Q}_1}, \tag{8}$$

subject to the condition that it minimizes

$$\|\mathbf{Z}'_{1s}\mathbf{W}^*_{1s}\mathbf{Z}_{1s} - \mathbf{Z}'_{1U}\mathbf{\Delta}_{1U}\mathbf{Z}_{1U}\|^2_{\mathbf{T}_1}. \tag{9}$$

The matrices $\mathbf{Q}_1$ and $\mathbf{T}_1$ in (8) and (9), respectively, are positive diagonal matrices, assigned by the statistician, which serve to reflect the relative importance of the units in the distance measures. When $\mathbf{z}_{1k}$ is known only for $k \in s_a$, we simply minimize (8) with respect to $\mathbf{W}^*_{1s}$, subject to the condition that

$$\|\mathbf{Z}'_{1s}\mathbf{W}^*_{1s}\mathbf{Z}_{1s} - \mathbf{Z}'_{1s_a}\check{\mathbf{\Delta}}_{1s_a}\mathbf{Z}_{1s_a}\|^2_{\mathbf{T}_1}, \tag{10}$$

where $\check{\mathbf{\Delta}}_{1s_a} = [\check{\Delta}_{akl}]$ with $\check{\Delta}_{akl} = \Delta_{akl}/\pi_{akl}$ $(k\&l \in s_a)$, is minimized. A calibration estimator for V_2 based on $\mathbf{Z}_{s_a}$ is given by

$$\hat{V}_{CAL,2} = \check{\check{\mathbf{e}}}'_{s}\mathbf{W}_{s}\check{\check{\mathbf{e}}}_{s}, \tag{11}$$

where $\mathbf{W}_s$ is the matrix which minimizes

$$\|\mathbf{W}^*_{s} - \check{\mathbf{\Delta}}_{s}\|^2_{\mathbf{Q}}, \tag{12}$$

subject to the condition that it minimizes

$$\|\mathbf{Z}'_{s}\mathbf{W}^*_{s}\mathbf{Z}_{s} - \mathbf{Z}'_{s_a}\mathbf{\Delta}_{s_a}\mathbf{Z}_{s_a}\|^2_{\mathbf{T}}. \tag{13}$$

The matrices $\mathbf{Q}$ and $\mathbf{T}$ in (12) and (13) are positive diagonal matrices, analogous to $\mathbf{Q}_1$ and $\mathbf{T}_1$, respectively. Combining (7) and (11), we thus have

$$\hat{V}_{CAL} = \hat{V}_{CAL,1} + \hat{V}_{CAL,2}, \tag{14}$$

which may be viewed as a calibration estimator for $V(\hat{t}_{yr})$.

To find closed-form expressions for $\mathbf{W}_{1s}$ and $\mathbf{W}_s$, respectively, let $\mathbf{V}_{1U} = \mathbf{Z}_{1U} \otimes \mathbf{Z}_{1U}$, $\mathbf{V}_{1s_a} = \mathbf{Z}_{1s_a} \otimes \mathbf{Z}_{1s_a}$, $\mathbf{V}_{1s} = \mathbf{Z}_{1s} \otimes \mathbf{Z}_{1s}$, $\mathbf{V}_{s_a} = \mathbf{Z}_{s_a} \otimes \mathbf{Z}_{s_a}$, and $\mathbf{V}_s = \mathbf{Z}_s \otimes \mathbf{Z}_s$, where $\otimes$ denotes the Kronecker product (e.g., Rao, 1973, p. 29). Moreover, for a matrix $\mathbf{F}_1$, let $\mathbf{F}_1^{\dagger}$ denote the Moore-Penrose generalized inverse (e.g., Rao, 1973, p. 26) of $\mathbf{F}_1$. Using the results in Théberge (1999, section 5), it may be shown that the minimization of (8) with respect to $\mathbf{W}_{1s}^*$, subject to (9), yields

$$\begin{aligned}\text{vec}(\mathbf{W}_{1s}) = \text{vec}(\breve{\breve{\mathbf{\Delta}}}_{1s}) &+ \mathbf{Q}_1^{-1}\mathbf{V}_{1s}\mathbf{T}_1^{1/2} \\ &\times [\mathbf{T}_1^{1/2}\mathbf{V}_{1s}'\mathbf{Q}_1^{-1}\mathbf{V}_{1s}\mathbf{T}_1^{1/2}]^{\dagger} \\ &\times \mathbf{T}_1^{1/2}[\mathbf{V}_{1U}'\text{vec}(\mathbf{\Delta}_{1U}) \\ &- \mathbf{V}_{1s}'\text{vec}(\breve{\breve{\mathbf{\Delta}}}_{1s})].\end{aligned}$$

When (10) is used as the calibration condition, we get

$$\begin{aligned}\text{vec}(\mathbf{W}_{1s}) = \text{vec}(\breve{\breve{\mathbf{\Delta}}}_{1s}) &+ \mathbf{Q}_1^{-1}\mathbf{V}_{1s}\mathbf{T}_1^{1/2} \\ &\times [\mathbf{T}_1^{1/2}\mathbf{V}_{1s}'\mathbf{Q}_1^{-1}\mathbf{V}_{1s}\mathbf{T}_1^{1/2}]^{\dagger} \\ &\times \mathbf{T}_1^{1/2}[\mathbf{V}_{1s_a}'\text{vec}(\breve{\mathbf{\Delta}}_{1s_a}) \\ &- \mathbf{V}_{1s}'\text{vec}(\breve{\breve{\mathbf{\Delta}}}_{1s})].\end{aligned}$$

Furthermore, the minimization of (12) with respect to $\mathbf{W}_s^*$, subject to (13), yields

$$\begin{aligned}\text{vec}(\mathbf{W}_s) = \text{vec}(\breve{\mathbf{\Delta}}_s) &+ \mathbf{Q}^{-1}\mathbf{V}_s\mathbf{T}^{1/2} \\ &\times [\mathbf{T}^{1/2}\mathbf{V}_s'\mathbf{Q}^{-1}\mathbf{V}_s\mathbf{T}^{1/2}]^{\dagger} \\ &\times \mathbf{T}_1^{1/2}[\mathbf{V}_{s_a}'\text{vec}(\mathbf{\Delta}_{s_a}) \\ &- \mathbf{V}_s'\text{vec}(\breve{\mathbf{\Delta}}_s)].\end{aligned}$$

Remark 2. An alternative to (14), based on g-weighted residuals, is given by

$$\hat{V}_{CAL}^{g_1;g} = \hat{V}_{CAL,1}^{g_1} + \hat{V}_{CAL,2}^{g},$$

where

$$\hat{V}_{CAL,1}^{g_1} = \breve{\mathbf{e}}_{1s}'\mathbf{G}_{1s}\mathbf{W}_{1s}\mathbf{G}_{1s}\breve{\mathbf{e}}_{1s}$$

and

$$\hat{V}_{CAL,2}^{g} = \breve{\breve{\mathbf{e}}}_s'\mathbf{G}_s\mathbf{W}_s\mathbf{G}_s\breve{\breve{\mathbf{e}}}_s.$$

Remark 3. Under stratified sampling designs, alternative calibration estimators are obtained if the minimization is carried out separately within each stratum.

5 A Simulation Study

To study the design-based properties of $\hat{V}_{CAL}$ relative to $\hat{V}_{REF}$, a small-scale Monte Carlo study was performed. For the study, the population generated by Axelson (2000b) was used. The population U was generated according to a slightly modified version of the method suggested by Vale and Maurelli (1983), which allows for the generation a multivariate non-normal distribution with specified correlation structure and given marginal means, variances, and coefficients of skewness and kurtosis. In generating U, Axelson used the correlation structure and the univariate moments for the variables in the real-world population $MU281$ (Särndal et al., 1992, Appendix B, pp. 652–659) as input. Hence, although artificial, U is similar to $MU281$ in terms of the univariate moments as well as the correlation structure.

As an initial choice, the study variable and the auxiliary information was chosen in accordance to the choices made by Särndal et al. (1992, p. 278). That is, y corresponded to $RMT85$, x_1 corresponded to $CS82$, and x_2 corresponded to $SS82$. The population correlation matrix of $(y, x_1, x_2)'$ is given by

$$\mathbf{r} = (r_{ij}) = \begin{bmatrix} 1.00 & 0.65 & 0.66 \\ 0.65 & 1.00 & 0.14 \\ 0.66 & 0.14 & 1.00 \end{bmatrix}$$

and $R^2 = 0.75$, where R^2 is the multiple coefficient of determination associated with regression of y on $(1, x_1, x_2)$. To get an indication of the extent to which the behavior of $\hat{V}_{CAL}$ relative to $\hat{V}_{REF}$ depends on the predictive ability of the vector $(1, x_1, x_2)$, the choice of x_2 corresponding to $REV84$ was also considered. For this choice, the population correlation matrix is given by

$$\mathbf{r} = (r_{ij}) = \begin{bmatrix} 1.00 & 0.65 & 0.91 \\ 0.65 & 1.00 & 0.59 \\ 0.91 & 0.59 & 1.00 \end{bmatrix}$$

and $R^2 = 0.85$. These two choices of auxiliary information will henceforth be referred to as *X1* and *X2*.

While the choice of the first-phase design often is governed by administrative arguments, the choice of the second-phase design is typically more directly related to efficiency arguments. It is not uncommon that the auxiliary information collected for the elements included in the first-phase sample is used in order to obtain an efficient stratified design for the second phase. To study the extent to which the behavior of $\hat{V}_{CAL}$ relative to $\hat{V}_{REF}$ depends on the choice of the second-phase design, the following two-phase sampling designs were considered:

D1 $p_a(\cdot)$: simple random sampling $(n_a/N = 0.1)$
$p(\cdot|s_a)$: simple random sampling $(n/n_a = 0.1)$

D2 $p_a(\cdot)$: simple random sampling $(n_a/N = 0.1)$
$p(\cdot|s_a)$: s_a divided into $H = 2$ equally sized strata by increasing size of x_2, stratified simple random sampling $(n_h/n_{ah} = 0.1)$

D3 $p_a(\cdot)$: simple random sampling, $n_a/N = 0.1$
$p(\cdot|s_a)$: s_a divided into $H = 4$ equally sized strata by increasing size of x_2, stratified simple random sampling $(n_h/n_{ah} = 0.1)$

Now, let $\delta_{hk} = 1$ if k belongs to second-phase strata h and define $\boldsymbol{\delta}_k = (\delta_{1k}, \ldots, \delta_{hk}, \ldots, \delta_{Hk})'$. Throughout the simulation study, x_1 was assumed available for all $k \in U$ while x_2 was known only for all $k \in s_a$, and the *GREG* was defined according to (1) with $\mathbf{x}_{1k} = (1, x_{1k})'$, $\mathbf{x}_k = \boldsymbol{\delta}_k \otimes (1, x_{1k}, x_{2k})'$ and $c_{1k} = c_k = 1$.

It is likely that the performance of $\hat{V}_{CAL}$ to a large extent is governed by the choice of the auxiliary vectors $\mathbf{z}_1$ and $\mathbf{z}$. In the Monte Carlo study, the following two choices of $\mathbf{z}_1$ and $\mathbf{z}$ was considered:

Z1 $\mathbf{z}_{1k} = \check{\mathbf{x}}_k'\widehat{\mathbf{B}}_s - \check{\mathbf{x}}_{1k}'\widehat{\mathbf{B}}_{1s} = \check{d}_{ks}$ $(k \in s_a)$ and $\mathbf{z}_k = -\check{\check{d}}_{ks}$ $(k \in s_a)$

Z2 $\mathbf{z}_{1k} = \check{\mathbf{x}}_{1k}'\widehat{\mathbf{B}}_{1s}$ $(k \in U)$ and $\mathbf{z}_k = \check{\check{\mathbf{x}}}_k'\widehat{\mathbf{B}}_s$ $(k \in s_a)$

The motivation for *Z1* is that $\check{d}_{ks} = \check{e}_{1ks} - \check{e}_{ks}$ and $-\check{\check{d}}_{ks} = \check{\check{e}}_{ks} - \check{\check{e}}_{1ks}$ may be regarded as proxies for $\check{E}_{1ks_a}$ and $\check{\check{E}}_{ks_a}$, respectively. *Z2* may be motivated through an extension of the framework presented by Singh et al. (1998), to allow for generalized regression estimation under two-phase sampling based on multivariate auxiliary information.

Let

$$\boldsymbol{\Pi}_{as} = \mathrm{diag}(\pi_{ak_1^*}, \ldots, \pi_{ak_\nu^*}, \ldots, \pi_{ak_n^*})$$

and

$$\boldsymbol{\Pi}_{s|s_a} = \mathrm{diag}(\pi_{k_1^*|s_a}, \ldots, \pi_{k_\nu^*|s_a}, \ldots, \pi_{k_n^*|s_a}).$$

The following two choices of $\mathbf{Q}_1$ and $\mathbf{Q}$ were considered:

Q1 $\mathbf{Q}_1 = (\boldsymbol{\Pi}_{s|s_a}\boldsymbol{\Pi}_{as}^{-1}) \otimes (\boldsymbol{\Pi}_{as}^{-1}\boldsymbol{\Pi}_{s|s_a})$ and $\mathbf{Q} = (\boldsymbol{\Pi}_{s|s_a}\boldsymbol{\Pi}_{as})^{-1} \otimes (\boldsymbol{\Pi}_{as}\boldsymbol{\Pi}_{s|s_a})^{-1}$

Q2 $\mathbf{Q}_1 = \mathrm{diag}(\mathbf{V}_{1s})\mathrm{diag}[\mathrm{vec}(\check{\check{\boldsymbol{\Delta}}}_{1s})]^{-1}$ and $\mathbf{Q} = \mathrm{diag}(\mathbf{V}_s)\mathrm{diag}[\mathrm{vec}(\check{\boldsymbol{\Delta}}_s)]^{-1}$

Throughout the study, the matrices $\mathbf{T}_1 = \mathbf{T} = 1$ were used. Hence, in addition to $\hat{V}_{REF}$, the following four choices of $\hat{V}_{CAL}$ were included in the study:

$\hat{V}_{C_1}$ $\hat{V}_{CAL}$ according to (14) using *Z1* and *Q1*

$\hat{V}_{C_2}$ $\hat{V}_{CAL}$ according to (14) using *Z1* and *Q2*

$\hat{V}_{C_3}$ $\hat{V}_{CAL}$ according to (14) using *Z2* and *Q1*

$\hat{V}_{C_4}$ $\hat{V}_{CAL}$ according to (14) using *Z2* and *Q2*

For each of combination of sampling design and auxiliary information, $M = 10000$ two-phase samples were realized, and for each simulation run, the variance estimates corresponding to $\hat{V}_{REF}$, $\hat{V}_{C_1}$, $\hat{V}_{C_2}$, $\hat{V}_{C_3}$, and $\hat{V}_{C_4}$ were computed. Let $\hat{V}_m$ and $\hat{t}_{yr,m}$ denote the estimates of $V(\hat{t}_{yr})$ and t_y, respectively, corresponding to the mth simulation run. The Monte Carlo mean squared error of $\hat{V}$ was calculated as

$$MSE_{MC}(\hat{V}) = \sum_{m=1}^{M} (\hat{V}_m - S^2_{\hat{t}_{yr}})^2/(M-1),$$

where

$$S^2_{\hat{t}_{yr}} = \sum_{m=1}^{M} (\hat{t}_{yr,m} - \bar{\hat{t}}_{yr})^2/(M-1),$$

with $\bar{\hat{t}}_{yr} = \sum_{m=1}^{M} \hat{t}_{yr,m}/M$. Also computed for each variance estimator were 95% confidence intervals of the form

$$CI(\hat{t}_{yr,m}, \hat{V}_m) = \hat{t}_{yr,m} \pm z_{0.975}\sqrt{\hat{V}_m},$$

where $z_{0.975}$ is the 97.5 percentile of the standard normal distribution, and the empirical coverage rate was calculated as

$$EC_{MC}(\hat{t}_{yr}, \hat{V}) = 100 \sum_{m=1}^{M} I_{[CI(\hat{t}_{yr,m}, \hat{V}_m) \ni t_y]}/M,$$

where $I_{[\cdot]}$ is the indicator function.

In Table 1, some of the results from the Monte Carlo study are presented .

Design & auxiliary information	$MSE_{MC}(\hat{V})/MSE_{MC}(\hat{V}_{REF})$ $(EC_{MC}(\hat{t}_{yr},\hat{V}))$				
	$\hat{V}_{REF}$	$\hat{V}_{C_1}$	$\hat{V}_{C_2}$	$\hat{V}_{C_3}$	$\hat{V}_{C_4}$
D1, X1	1.00 (93.7)	0.95 (93.7)	1.04 (94.0)	1.00 (93.7)	0.79 (94.0)
D2, X1	1.00 (93.8)	0.90 (93.8)	1.32 (94.1)	1.00 (93.8)	1.00 (94.1)
D3, X1	1.00 (93.1)	0.91 (93.2)	1.45 (93.8)	1.00 (93.2)	1.13 (93.6)
D1, X2	1.00 (94.0)	0.88 (94.1)	1.02 (94.2)	1.00 (94.0)	0.83 (94.4)
D2, X2	1.00 (93.6)	0.87 (93.7)	1.17 (93.8)	1.00 (93.6)	1.20 (94.2)
D3, X2	1.00 (93.3)	0.90 (93.4)	1.16 (94.0)	1.00 (93.3)	1.56 (94.0)

Table 1. Relative Monte Carlo MSEs and empirical coverage rates of 95% confidence intervals

Clearly, all of the studied variance estimators display acceptable properties in terms of the empirical coverage rates. Comparing the estimators on the basis of the difference between the the empirical and the nominal coverage rate, the data implies that either $\hat{V}_{C_2}$ or $\hat{V}_{C_4}$ is to be preferred. However, since the differences between the studied estimators are quite small, we refrain from drawing any far-reaching conclusions solely on the basis of the empirical coverage rates.

If the relative MSEs are included in the comparison a slightly different picture emerges, in which $\hat{V}_{C_1}$ appears as the best choice. The major reasons for this are:

i Apart for the combinations *D1, X1* and *D1, X2*, $\hat{V}_{C_1}$ is the most efficient estimator.

ii Even under *D1, X1* and *D1, X2*, $\hat{V}_{C_1}$ is more efficient than $\hat{V}_{REF}$.

iii Throughout the simulation study, $\hat{V}_{C_2}$ is less efficient than $\hat{V}_{REF}$.

iv $\hat{V}_{C_3}$ displays the same efficiency as $\hat{V}_{REF}$ throughout the simulation study. This may seem surprising, but it is a matter of algebra to show that $\hat{V}_{C_3}$ is almost identical to $\hat{V}_{REF}$ under the used definition of the *GREG*.

v Although $\hat{V}_{C_4}$, which is most efficient under *D1, X1* and *D1, X2*, performs reasonably well under *D2, X1* and *D3, X1*, it is the most inefficient choice under *D2, X2* and *D3, X2*.

vi The degree to which the performance depends on the choice of the auxiliary vector $\mathbf{x}$ seems to be smaller for $\hat{V}_{C_1}$ than for any of the other three calibration-type variance estimators.

In conclusion, the results from the Monte Carlo study imply that more efficient variance estimators may be obtained through the use of the suggested calibration approach, but they also indicate that the performance depends on the choice of auxiliary information to be used. However, before any general recommendations can be made, both the theoretical and practical properties of the suggested approach to variance estimation need to be further investigated.

References

Axelson, M. (2000a). A modified approach to variance estimation for the generalized regression estimator under two-phase sampling. In *On variance estimation for the two-phase regression estimator.* Ph. D. thesis, Department of Statistics, Uppsala University.

Axelson, M. (2000b). A note on variance estimation for the regression estimator under two-phase sampling. In *On variance estimation for the two-phase regression estimator.* Ph. D. thesis, Department of Statistics, Uppsala University.

Axelson, M., Breidt, F. J., and Carriquiry, A. L. (1996). Two-phase regression estimation for policy analysis using computer simulation experiments. In *Proceedings of the Section on Survey Research Methods*, pp. 320–325. American Statistical Association.

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

Dupont, F. (1995). Alternative adjustments when there are several levels of auxiliary information. *Survey Methodology* **21**, 125–135.

Hidiroglou, M. A. and Särndal, C. E. (1998). Use of auxiliary information for two-phase sampling. *Survey Methodology* **24**, 11–20.

Rao, C. R. (1973). *Linear statistical inference and its applications* (2 ed.). New York: Wiley.

Rao, J. N. K. and Sitter, R. R. (1995). Variance estimation under two-phase sampling with application to imputation for missing data. *Biometrika* **82**, 453–460.

Särndal, C. E. and Swensson, B. (1987). A general view of estimation for two phases of selection with applications to two-phase sampling and non-response. *International Statistical Review* **55**, 279–294.

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

Singh, S., Horn, S., and Yu, F. (1998). Estimation of variance of general regression estimator: Higher level calibration approach. *Survey Methodology* **24**, 41–50.

Sitter, R. R. (1997). Variance estimation for the regression estimator in two-phase sampling. *Journal of the American Statistical Association* **92**, 780–787.

Théberge, A. (1999). Extensions of calibration estimators in survey sampling. *Journal of the American Statistical Association* **94**, 635–644.

Vale, C. D. and Maurelli, V. A. (1983). Simulating multivariate nonnormal distributions. *Psychometrika* **48**, 465–471.