A NEW PERSPECTIVE ON CALIBRATION ESTIMATORS

Victor M. Estevao and Carl-Erik Särndal
Statistics Canada, Ottawa, Ontario, Canada, K1T 3L2

KEY WORDS: Auxiliary Information, Calibration Weights, Automated Linearization, Population Residuals.

1 Introduction

Complex sampling design and *complex parameter* are familiar concepts in survey sampling. Another survey feature that often extends beyond the simple ordinary formulation is the use of auxiliary information at the estimation stage. As the recent literature has shown, auxiliary information can be more or less complex, depending on the survey design. Estimation for complex cases is not well-covered by standard textbook techniques. A broader framework for estimation is needed with such auxiliary information. We use the term *complex auxiliary information* for the several non-standard cases examined in this paper.

The most basic use of auxiliary information occurs for a single stage, single phase sampling design, where a known set of auxiliary variables and their corresponding totals are used to compute calibrated weights for the estimate of a population total. This procedure is reviewed in section 2. More complex cases considered in later sections arise when auxiliary information is available at different stages and phases of sampling. Then it is not always evident how to make efficient use of the available auxiliary information. In this paper, we look at different ways of using complex auxiliary information to produce efficient calibration estimators in two-stage and two-phase sampling. The derivation of the variance of these estimators requires a simple procedure to linearize the expression for a nonlinear calibration estimator. This simple procedure is introduced as the method of automated linearization.

The paper is arranged as follows. In section 2 we explain automated linearization for calibration estimators in a one-stage, one-phase sampling design. Section 3 examines estimation for two-phase sampling designs, and section 4 looks at estimation in two-stage sampling with and without integrated weighting. A brief summary and comments are given in the concluding section 5.

2 Automated linearization

We first look at the simple case of auxiliary information for a one-stage, one-phase unit sampling design. Consider a finite population $U = \{1, 2, ..., k, ..., N\}$ from which a probability sample s is drawn. We denote by π_k the inclusion probability of unit k and by $a_k = 1/\pi_k$ the sampling weight of k. Let y be the variable of interest. Its value for unit k, y_k, is observed for $k \in s$. The unknown total to be estimated is $Y = \sum_U y_k$.

We denote by $\mathbf{x}$ an *auxiliary vector* of dimension $J \geq 1$, and by $\mathbf{x}_k$ its value for unit k. We assume that we have the following *auxiliary information*:

(i) The population vector total $\mathbf{X} = \sum_U \mathbf{x}_k$ is known.

(ii) The vector value $\mathbf{x}_k$ is known for every $k \in s$.

Here, $\mathbf{X}$ is assumed known from an outside source such as a census. If we know the value $\mathbf{x}_k$ for every $k \in U$, as when $\mathbf{x}_k$ is on the population frame U for every k, then both (i) and (ii) are met. We can compute the simple unbiased estimator of the known $\mathbf{X}$ as $\hat{\mathbf{X}} = \sum_s a_k \mathbf{x}_k$. Under general conditions $N^{-1}(\hat{\mathbf{X}} - \mathbf{X})$ is $\mathrm{O}_p(n^{-1/2})$.

Our objective is to estimate $Y = \sum_U y_k$. One possibility is the simple unbiased Horvitz-Thompson (HT) estimator $\hat{Y} = \sum_s a_k y_k$. However, a more efficient weighting of the observed y_k is one that takes the auxiliary information into account. Let us consider instead $\hat{Y}_{CAL} = \sum_s w_k y_k$, where the weights $\{w_k ; k \in s\}$ satisfy the calibration equation $\sum_s w_k \mathbf{x}_k = \mathbf{X}$. We say that the weights $\{w_k ; k \in s\}$ are *calibrated* to $\mathbf{X} = \sum_U \mathbf{x}_k$.

Alternative sets of calibrated weights can be derived by the distance measure approach, as for example in Huang and Fuller (1978) and Deville and Särndal (1992). The minimization of each distance measure produces a different set of calibrated weights. However, the proposed distance measures are fairly similar so they tend to produce estimators with similar properties. Instead we use the instrument vector approach, also called generalized calibration, as in Deville (2002) and Le Guennec and Sautory (2002). This method allows a more general parameterization of

the calibration weights. We specify a vector $\mathbf{z}_k$ of the same dimension as $\mathbf{x}_k$ and compute the weights

$$w_k = a_k(1+\lambda_s^T\mathbf{z}_k), \quad k \in s \tag{2.1}$$

where $\lambda_s^T = (\mathbf{X}-\hat{\mathbf{X}})^T(\sum_s a_k\mathbf{z}_k\mathbf{x}_k^T)^{-1}$. The mapping from $\mathbf{z}_k$ to w_k is not one-to-one. Different choices for $\mathbf{z}_k$ produce the same weights w_k. We are free to choice the form of $\mathbf{z}_k$ as long as the $J \times J$ matrix $(\sum_s a_k\mathbf{z}_k\mathbf{x}_k^T)$ has an inverse for every possible sample *s*. We refer to such a $\mathbf{z}_k$ as a *valid instrument vector*. The standard choice of $\mathbf{z}_k = \mathbf{x}_k$ produces the generalized regression estimator, although as explained later, this choice is not necessarily optimal for any given design. For any valid instrument vector $\mathbf{z}_k$, the weights satisfy $\sum_s w_k\mathbf{x}_k = \mathbf{X}$, and the estimator can be written as $\hat{Y}_{CAL} = \sum_s w_k y_k = \hat{Y} + (\mathbf{X}-\hat{\mathbf{X}})^T\hat{\mathbf{B}}$, where $\hat{\mathbf{B}} = (\sum_s a_k\mathbf{z}_k\mathbf{x}_k^T)^{-1}\sum_s a_k\mathbf{z}_k y_k$.

Here $\hat{\mathbf{B}}$ is a nonlinear design-weighted statistic, thus, it is not what we call a HT statistic. Although $\hat{Y}$ is a linear statistic, the term $(\mathbf{X}-\hat{\mathbf{X}})^T\hat{\mathbf{B}}$ makes $\hat{Y}_{CAL}$ a nonlinear estimator. This causes no problem for point estimation since $\hat{Y}_{CAL}$ can be readily computed. But the nonlinear form of the estimator creates a problem for obtaining a simple exact expression for the variance of $\hat{Y}_{CAL}$ and for finding a corresponding sample-based estimate of this variance. Linearization is the usual technique for circumventing this difficulty with nonlinear statistics. Woodruff (1971) is a basic reference. Since then, many papers have appeared on the linearization of complex statistics of interest in survey sampling, for example, Binder (1996), Binder and Kovačević (1995), and Deville (1999). The emphasis in these references is on linearization of statistics for estimating complex parameters, a purpose somewhat different from ours, which is the study of calibration estimators of a total. Théberge (1999) presents a linearization approach similar to the one given here. His development is based on the use of distance functions rather than an instrument vector.

Linearization of the nonlinear $\hat{Y}_{CAL}$ involves a power series expansion, including an evaluation of partial derivatives. The rather lengthy derivation is given for example in Särndal, Swensson and Wretman (1992). This method isolates a main term, $\hat{Y}_{CAL,lin}$, which is a linear statistic. The remainder term is of lower order in probability and assumed negligible compared to the main term. The expression for the remainder term is usually not made explicit in Woodruff linearization. This is not a serious drawback, because standard practice is to discard this term and simply take $\hat{Y}_{CAL,lin}$ to be a "sufficiently good" linear approximation to $\hat{Y}_{CAL}$. Under general conditions, $N^{-1}(\hat{Y}_{CAL}-\hat{Y}_{CAL,lin})$ is $O_p(n^{-1})$, not just $O_p(n^{-1/2})$, permitting the easily derived variance of $\hat{Y}_{CAL,lin}$ to be used as an accurate approximation of the variance of $\hat{Y}_{CAL}$, even for modest sample sizes.

Instead of the standard linearization approach, we introduce the method of automated linearization. This simple two-step procedure "automatically" makes explicit both the linearized statistic and the lower order term. In contrast to Woodruff linearization, automated linearization requires no evaluation of partial derivatives. For the case of simple auxiliary information in this section, we confirm the well-known expression for the variance of $\hat{Y}_{CAL,lin}$. Automated linearization has two steps:

Step 1. In the expression $\hat{Y}_{CAL} = \hat{Y} + (\mathbf{X}-\hat{\mathbf{X}})^T\hat{\mathbf{B}}$, create a term of lower order in probability by centering $\hat{\mathbf{B}}$ on the population vector $\mathbf{B} = (\sum_U \mathbf{z}_k\mathbf{x}_k^T)^{-1}\sum_U \mathbf{z}_k y_k$ to which $\hat{\mathbf{B}}$ converges in probability. Then $\hat{\mathbf{B}}-\mathbf{B}$ is $O_p(n^{-1/2})$, and we have

$$\hat{Y}_{CAL} = \hat{Y} - (\hat{\mathbf{X}}-\mathbf{X})^T\mathbf{B} - (\hat{\mathbf{X}}-\mathbf{X})^T(\hat{\mathbf{B}}-\mathbf{B}) \tag{2.2}$$

where $N^{-1}(\hat{\mathbf{X}}-\mathbf{X})^T(\hat{\mathbf{B}}-\mathbf{B})$ is $O_p(n^{-1})$, a lower order compared to $N^{-1}(\hat{\mathbf{X}}-\mathbf{X})^T\mathbf{B}$ which is $O_p(n^{-1/2})$.

Step 2. Rewrite (2.2) as

$$\hat{Y}_{CAL} = (\hat{Y}-\hat{\mathbf{X}}^T\mathbf{B}) + \mathbf{X}^T\mathbf{B} - (\hat{\mathbf{X}}-\mathbf{X})^T(\hat{\mathbf{B}}-\mathbf{B}). \tag{2.3}$$

The calibration estimator is the sum of three terms: the constant term $\mathbf{X}^T\mathbf{B}$, the design-based linear term $\hat{Y}-\hat{\mathbf{X}}^T\mathbf{B}$, and the design-based nonlinear term $-(\hat{\mathbf{X}}-\mathbf{X})^T(\hat{\mathbf{B}}-\mathbf{B})$ of lower order. The first two terms on the right hand side of (2.3) define the linearized statistic

$$\hat{Y}_{CAL,lin} = (\hat{Y}-\hat{\mathbf{X}}^T\mathbf{B}) + \mathbf{X}^T\mathbf{B} = \sum_s a_k e_k + \mathbf{X}^T\mathbf{B} \tag{2.4}$$

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

Our point estimator of Y is $\hat{Y}_{CAL}$. It has a small bias, $\mathrm{E}(\hat{Y}_{CAL}) - Y = -\mathrm{E}\{(\hat{\mathbf{X}} - \mathbf{X})^T(\hat{\mathbf{B}} - \mathbf{B})\}$, since $-N^{-1}\mathrm{E}\{(\hat{\mathbf{X}} - \mathbf{X})^T(\hat{\mathbf{B}} - \mathbf{B})\}$ is of order $\mathrm{O}(n^{-1})$. Therefore, the variance of $\hat{Y}_{CAL}$ is approximately the variance of the linearized statistic $\hat{Y}_{CAL,lin}$. Since $\mathbf{X}^T\mathbf{B}$ is a constant, the use of auxiliary information reduces the variance of the estimator from $\mathrm{Var}(\sum_s a_k y_k)$ to approximately $\mathrm{Var}(\sum_s a_k e_k)$. It is important to note that the e_k are fixed but unknown values and that $\sum_s a_k e_k$ is a HT statistic in the e_k. Although the e_k resemble regression residuals, they arise *automatically* from steps 1 and 2, without any explicit regression model or fit. Because $\sum_s a_k e_k$ is a HT statistic, we obtain immediately,

$$\mathrm{Var}(\hat{Y}_{CAL}) \cong \mathrm{Var}(\hat{Y}_{CAL,lin}) = \sum\sum_U F_{k\ell} e_k e_\ell \qquad (2.5)$$

where $F_{k\ell} = \dfrac{a_k a_\ell}{a_{k\ell}} - 1$ for $\ell \neq k$, $F_{kl} = F_{kk} = a_k - 1$ for $\ell = k$, with $a_{kl} = 1/\pi_{k\ell}$ where $\pi_{k\ell}$ is the joint inclusion probability of k and ℓ. We use $\sum\sum_U$ as shorthand for the double sum $\sum_{k\in U}\sum_{\ell\in U}$. To estimate the variance of $\hat{Y}_{CAL}$, we use the sample-based analogue of (2.5),

$$\hat{\mathrm{V}}(\hat{Y}_{CAL}) = \sum\sum_s (a_k a_\ell - a_{k\ell})\hat{e}_k \hat{e}_\ell \qquad (2.6)$$

where $\hat{e}_k = y_k - \mathbf{x}_k^T\hat{\mathbf{B}}$ and $\sum\sum_s$ stands for $\sum_{k\in s}\sum_{\ell\in s}$.

The weights w_k in $\hat{Y}_{CAL} = \sum_s w_k y_k$ depend on the instrument vector $\mathbf{z}_k$. For every choice of $\mathbf{z}_k$ for $k \in U$, there corresponds a vector $\mathbf{B}$ satisfying the equation $(\sum_U \mathbf{z}_k \mathbf{x}_k^T)\,\mathbf{B} = \sum_U \mathbf{z}_k y_k$. We can find an optimal $\mathbf{B}$, and a corresponding $\mathbf{z}_k$, by minimizing $\mathrm{Var}(\hat{Y}_{CAL,lin})$ given by (2.5). This $\mathbf{z}_k$ is asymptotically optimal for $\hat{Y}_{CAL}$ in that it minimizes $\mathrm{Var}(\hat{Y}_{CAL,lin}) \cong \mathrm{Var}(\hat{Y}_{CAL})$. The optimal $\mathbf{B}$ is $\mathbf{B}^0$, defined as the solution of the normal equation

$$(\sum\sum_U F_{k\ell}\mathbf{x}_\ell \mathbf{x}_k^T)\,\mathbf{B}^0 = \sum\sum_U F_{k\ell}\mathbf{x}_\ell y_k . \qquad (2.7)$$

A comparison with the general form $(\sum_U \mathbf{z}_k \mathbf{x}_k^T)\,\mathbf{B} = \sum_U \mathbf{z}_k y_k$ defining $\mathbf{B}$, suggests that an optimal instrument vector is $\mathbf{z}_k = \mathbf{z}_k^0$, where $\mathbf{z}_k^0 = \sum_{\ell\in U} F_{k\ell}\mathbf{x}_\ell$. The result agrees with Montanari's (1987) determination of $\mathbf{B}$ so as to minimize the variance of the unbiased difference estimator $\hat{Y} + (\mathbf{X} - \hat{\mathbf{X}})^T\mathbf{B}$.

To see the features of the weights, let us write them as $w_k = a_k\{1 + (\sum_U \mathbf{x}_k - \sum_s a_k \mathbf{x}_k)^T(\sum_s a_k \mathbf{z}_k \mathbf{x}_k^T)^{-1}\mathbf{z}_k\}$. We note the following:

(i) The computation of the weights w_k for $k \in s$ requires the *design weights* a_k, the *auxiliary vector values* $\mathbf{x}_k$, the *instrument vector values* $\mathbf{z}_k$, and the known *auxiliary vector of totals* $\sum_U \mathbf{x}_k$.

(ii) The a_k are fixed by the design.

(iii) We are free to choose the $\mathbf{z}_k$ as long as the matrix $\sum_s a_k \mathbf{z}_k \mathbf{x}_k^T$ is invertible.

(iv) The weights w_k calibrate to the known totals $\sum_U \mathbf{x}_k$ for any valid instrument vector $\mathbf{z}_k$.

(v) The weights w_k are not dependent on y or on any presumed relationship between y and $\mathbf{x}$, as in a model dependent approach.

Some choices of $\mathbf{z}_k$ are "better" than others. The optimal choice, as noted above, is $\mathbf{z}_k = \mathbf{z}_k^0 = \sum_{\ell\in U} F_{k\ell}\mathbf{x}_\ell$. It makes sense that the optimal choice depends on the sampling design. The sample-based choice corresponding to $\mathbf{z}_k^0$ is $\mathbf{z}_k = \mathbf{z}_k^* = a_k^{-1}\sum_{\ell\in s} a_{k\ell} F_{k\ell}\mathbf{x}_\ell$. The weights w_k do not depend on the values y_k of the variable of interest y and thus the optimal weights do not depend on y_k. Once the $\mathbf{z}_k$ are specified, the same weights can be used for all y-variables in the survey. The estimator $\hat{Y}_{CAL}$ is free of any unverifiable assumptions about a possible regression of y on $\mathbf{x}$. In the application of this approach it does not matter whether there exists a linear relationship between y and $\mathbf{x}$. Furthermore, no assumptions are required on the properties of the residuals e_k. These are treated as fixed but unknown values over the population U rather than random variables from a hypothetical superpopulation model.

As a simple illustration, consider Simple Random Sampling (SRS) from U with the sampling fraction $f = n/N$ and consider $\mathbf{x}_k = (1, x_k)^T$, where x_k is a scalar variable value. The required population information is $\sum_U \mathbf{x}_k = (N, \sum_U x_k)^T$. Then the optimal instrument is found to be $\mathbf{z}_k = \mathbf{z}_k^0 = \sum_{\ell\in U} F_{k\ell}\mathbf{x}_\ell =$

$\frac{N}{N-1}(\frac{1}{f}-1)(\mathbf{x}_k-\bar{\mathbf{x}}_U)$ where $\bar{\mathbf{x}}_U=\sum_U \mathbf{x}_k/N$. The corresponding sample based choice is $\mathbf{z}_k=\mathbf{z}_k^*=\frac{n}{n-1}(\frac{1}{f}-1)(\mathbf{x}_k-\bar{\mathbf{x}}_s)$ where $\bar{\mathbf{x}}_s=\sum_s \mathbf{x}_k/n$. However both $\mathbf{z}_k^0$ and $\mathbf{z}_k^*$ are invalid because the first component of these vectors is always zero, leading to a singular matrix $\sum_s a_k\mathbf{z}_k\mathbf{x}_k^T$. We drop the first auxiliary variable with the known total *N* and work instead with the vector $\mathbf{x}_k=x_k$. This gives $z_k=z_k^*=\frac{n}{n-1}(\frac{1}{f}-1)(x_k-\bar{x}_s)$. The result is the familiar $\hat{Y}_{CAL}=N\{\bar{y}_s+(\bar{x}_U-\bar{x}_s)b\}$ with $b=\{\sum_s(x_k-\bar{x}_s)(y_k-\bar{y}_s)/\sum_s(x_k-\bar{x}_s)^2\}$. As is easily verified, this estimator gives $\hat{Y}_{CAL}=N$ when $y_k=1$ for all *k*, and $\hat{Y}_{CAL}=N\bar{x}_U=\sum_U x_k$ when $y_k=x_k$. That is, even though we must reduce the auxiliary vector from $\mathbf{x}_k=(1,x_k)^T$ to $\mathbf{x}_k=x_k$, the resulting set of weights still reproduce the two known quantities *N* and $\sum_U x_k$. No loss of information is incurred from the non-invertibility of $\sum_s a_k\mathbf{z}_k\mathbf{x}_k^T$ with $\mathbf{x}_k=(1,x_k)^T$.

We end this section by listing the steps of the preceding argument. These important steps are applied in each of the subsequent sections, where the auxiliary information is more complex.

Step 1 The auxiliary information: Specify an $\mathbf{x}_k$-vector with known totals.

Step 2 Point estimation: Specify a valid $\mathbf{z}_k$, compute the calibrated weights and the resulting point estimate.

Step 3 Variance and variance estimation: Use automated linearization to (a) identify the linearized statistic, (b) obtain the residuals that determine the variance, and (c) transform that variance into an estimated variance.

3 Calibration estimation in two-phase sampling

We now consider the setup for sampling in two phases. From the population $U=\{1, 2, ..., k, ..., N\}$, a large probability sample, s_1, is drawn with known first-phase inclusion probabilities π_{1k}. The first-phase sampling weights are $a_{1k}=1/\pi_{1k}$ for $k\in s_1$. One or more variables are observed for $k\in s_1$. Then, from s_1, a sub-sample, *s*, is drawn with known conditional probabilities π_{2k}. The second-phase sampling weights are $a_{2k}=1/\pi_{2k}$, conditionally on the realized s_1. We denote by $a_k=a_{1k}a_{2k}$ the overall sampling weight for unit *k*. The value y_k of the variable of interest is observed for all $k\in s$. The objective is to find a more efficient alternative for estimating $Y=\sum_U y_k$ than the two-phase double expansion estimator $\hat{Y}=\sum_s a_k y_k$.

We need to consider two auxiliary vectors for each unit *k*. We denote these by $\mathbf{x}_1$ and $\mathbf{x}_2$, with $\mathbf{x}_{1k}$ and $\mathbf{x}_{2k}$ representing their respective values for unit *k*. Their dimensions are $J_1\geq 1$ and $J_2\geq 1$ respectively. The auxiliary information for $\mathbf{x}_1$ and $\mathbf{x}_2$ is as follows:

(i) The population vector total $\mathbf{X}_1=\sum_U \mathbf{x}_{1k}$ is known while the population vector total $\mathbf{X}_2=\sum_U \mathbf{x}_{2k}$ is not known.

(ii) For every $k\in s_1$, the vector values $\mathbf{x}_{1k}$ and $\mathbf{x}_{2k}$ are known.

(iii) For every $k\in s$, the vector values $\mathbf{x}_{1k}$ and $\mathbf{x}_{2k}$ are known.

The information given by (i), (ii) and (iii) is used to compute the weights for the calibration estimator $\hat{Y}_{CAL}=\sum_s w_k y_k$ in an effort to improve on $\hat{Y}=\sum_s a_k y_k$. There are different ways to produce these weights w_k, depending on how we use (i) to (iii). For example, we can carry out a single calibration step, or arrive at the w_k in two steps by first producing a set of first-phase weights w_{1k}. Each step requires starting weights, an auxiliary vector and a valid instrument vector. We consider the following alternatives:

(a) One step calibration. Starting from $a_k=a_{1k}a_{2k}$, compute directly final weights w_k for $k\in s$, calibrated to satisfy $\sum_s w_k\mathbf{x}_k=\begin{pmatrix}\sum_U \mathbf{x}_{1k}\\ \sum_{s_1} a_{1k}\mathbf{x}_{2k}\end{pmatrix}$, with $\mathbf{x}_k=\begin{pmatrix}\mathbf{x}_{1k}\\ \mathbf{x}_{2k}\end{pmatrix}$ of dimension J_1+J_2. This is case B1 in Estevao and Särndal (2002).

(b) Two step calibration. In step one, starting from a_{1k}, compute first-phase weights w_{1k} for $k\in s_1$, such that $\sum_{s_1} w_{1k}\mathbf{x}_{1k}=\sum_U \mathbf{x}_{1k}$. In step two, starting from $a_k=a_{1k}a_{2k}$, and using the w_{1k} from step (i), compute final weights w_k for $k\in s$, such that

$\sum_s w_k \mathbf{x}_k = \sum_{s_1} w_{1k} \mathbf{x}_k$. The final weights satisfy $\sum_s w_k \mathbf{x}_k = \begin{pmatrix} \sum_U \mathbf{x}_{1k} \\ \sum_{s_1} w_{1k} \mathbf{x}_{2k} \end{pmatrix}$. This is case A1 in Estevao and Särndal (2002).

The two procedures make slightly different use of the auxiliary information and in general, they produce different weights w_k for $k \in s$. The use of information is somewhat more extensive in (b) than (a), in that it requires the information about the individual values $\mathbf{x}_{1k}$ for $k \in s_1$. This may or may not lead to an increase in efficiency, depending on the relation between $\mathbf{x}_{1k}$, $\mathbf{x}_{2k}$ and y_k. These questions are discussed in Estevao and Särndal (2002). We can use automated linearization to obtain the form of the residuals and the variance of each estimator in (a) and (b). We examine case (b) below.

The first-phase calibrated weights for case (b) are computed as

$w_{1k} = a_{1k}(1 + \lambda_{s_1}^T \mathbf{z}_{1k})$ with
$\lambda_{s_1}^T = (\sum_U \mathbf{x}_{1k} - \sum_{s_1} a_{1k} \mathbf{x}_{1k})^T (\sum_{s_1} a_{1k} \mathbf{z}_{1k} \mathbf{x}_{1k}^T)^{-1}$

for some valid instrument vector $\mathbf{z}_{1k}$. The w_{1k} are then used as input to compute the final calibrated weights as

$w_k = a_k(1 + \lambda_s^T \mathbf{z}_k)$ where
$\lambda_s^T = (\sum_{s_1} w_{1k} \mathbf{x}_k - \sum_s a_k \mathbf{x}_k)^T (\sum_s a_k \mathbf{z}_k \mathbf{x}_k^T)^{-1}$

where $\mathbf{x}_k = \begin{pmatrix} \mathbf{x}_{1k} \\ \mathbf{x}_{2k} \end{pmatrix}$ and $\mathbf{z}_k$ is another valid instrument vector. We derive the variance by using automated linearization. First, we insert into $\hat{Y}_{CAL} = \sum_s w_k y_k$ the expression for w_k. Then, we define $\hat{\mathbf{B}} = \begin{pmatrix} \hat{\mathbf{B}}_1 \\ \hat{\mathbf{B}}_2 \end{pmatrix} = \left(\sum_s a_k \mathbf{z}_k \mathbf{x}_k^T\right)^{-1} (\sum_s a_k \mathbf{z}_k y_k)$ and center it on its population counterpart, the non-random vector $\mathbf{B} = \begin{pmatrix} \mathbf{B}_1 \\ \mathbf{B}_2 \end{pmatrix} = \left(\sum_U \mathbf{z}_k \mathbf{x}_k^T\right)^{-1} (\sum_U \mathbf{z}_k y_k)$. After some algebra, we then define the statistic $\hat{\mathbf{B}}_1^* = (\sum_{s_1} a_{1k} \mathbf{z}_{1k} \mathbf{x}_{1k}^T)^{-1} \left(\sum_{s_1} a_{1k} \mathbf{z}_{1k} (\mathbf{x}_k^T \hat{\mathbf{B}})\right)$, which we center on its population counterpart $\mathbf{B}_1^* = (\sum_U \mathbf{z}_{1k} \mathbf{x}_{1k}^T)^{-1} \left(\sum_U \mathbf{z}_{1k} (\mathbf{x}_k^T \mathbf{B})\right)$. The result is

$$\begin{aligned} \hat{Y}_{CAL} &= \sum_s a_k (y_k - \mathbf{x}^T \mathbf{B}) \\ &\quad + \sum_{s_1} a_{1k} (\mathbf{x}_k^T \mathbf{B} - \mathbf{x}_{1k}^T \mathbf{B}_1^*) + \mathbf{X}_1^T \mathbf{B}_1^* + R \end{aligned} \tag{3.1}$$

where R is the lower order term given by

$$R = -\begin{pmatrix} \hat{\hat{\mathbf{X}}}_1 - \hat{\mathbf{X}}_1 \\ \hat{\hat{\mathbf{X}}}_2 - \hat{\mathbf{X}}_2 \end{pmatrix}^T \begin{pmatrix} \hat{\mathbf{B}}_1 - \mathbf{B}_1 \\ \hat{\mathbf{B}}_2 - \mathbf{B}_2 \end{pmatrix} - (\hat{\mathbf{X}}_1 - \mathbf{X}_1)^T (\hat{\mathbf{B}}_1^* - \mathbf{B}_1^*)$$

with $\hat{\mathbf{X}}_j = \sum_{s_1} a_{1k} \mathbf{x}_{jk}$ and $\hat{\hat{\mathbf{X}}}_j = \sum_s a_k \mathbf{x}_{jk}$, $j = 1, 2$. The term of main interest is the linear statistic

$$\hat{Y}_{CAL,lin} = \sum_s a_k e_k + \sum_{s_1} a_{1k} e_k^* + \mathbf{X}_1^T \mathbf{B}_1^* \tag{3.2}$$

where $\mathbf{X}_1^T \mathbf{B}_1^* = (\sum_U \mathbf{x}_{1k})^T \mathbf{B}_1^*$ is a constant, and the residuals in the two random terms are

$e_k = y_k - \mathbf{x}_k^T \mathbf{B} = y_k - \mathbf{x}_{1k}^T \mathbf{B}_1 - \mathbf{x}_{2k}^T \mathbf{B}_2$ and
$e_k^* = \mathbf{x}_k^T \mathbf{B} - \mathbf{x}_{1k}^T \mathbf{B}_1^*$.

By ignoring the lower-order term in (3.1), we can use the linear design-weighted statistic (3.2) to obtain the approximate variance of $\hat{Y}_{CAL}$. Then, conditioning on s_1, we obtain

$$\mathrm{Var}(\hat{Y}_{CAL}) \cong \mathrm{Var}(\hat{Y}_{CAL,lin}) = \mathrm{V}_1(\mathrm{E}_c) + \mathrm{E}_1(\mathrm{V}_c)$$

where $\mathrm{E}_c = \sum_{s_1} a_{1k} (y_k - \mathbf{x}_{1k}^T \mathbf{B}_1^*)$ and V_c is the conditional variance of $\sum_s a_k e_k$, given s_1. The expressions for $\mathrm{V}_1(\mathrm{E}_c)$ and $\mathrm{E}_1(\mathrm{V}_c)$, and for their respective estimates are not detailed here. They follow well-known patterns for two-phase sampling as shown for example in Estevao and Särndal (2002). For the estimated variance, we use $\hat{\mathbf{B}}$ and $\hat{\mathbf{B}}_1^*$ instead of $\mathbf{B}$ and $\mathbf{B}_1^*$. Note that $\mathrm{E}_1(\mathrm{V}_c) = 0$ if sample selection stops after the first phase.

The first term, $\mathrm{V}_1(\mathrm{E}_c)$, is reduced by the presence in e_k^* of the regressor $\mathbf{x}_{1k}$ only, whereas the second term, $\mathrm{E}_1(\mathrm{V}_c)$, gets reduced by both regressors, $\mathbf{x}_{1k}$ and $\mathbf{x}_{2k}$. These features seem logical under the survey conditions. An interesting question, which we leave unresolved here, is the jointly optimal choice for the two instruments, $\mathbf{z}_{1k}$ and $\mathbf{z}_k$. The simple standard choices are $\mathbf{z}_{1k} = \mathbf{x}_{1k}$ and $\mathbf{z}_k = \mathbf{x}_k$.

We comment briefly on the automated linearization in case (a). The outcome is also a variance of the form $\mathrm{V}_1(\mathrm{E}_c) + \mathrm{E}_1(\mathrm{V}_c)$, with one residual for the first

component, $V_1(E_c)$, and another residual for the second, $E_1(V_c)$. But these residuals are somewhat different in (a) and (b). For (b), we stated earlier in this section the first component residuals as $e_k = y_k - \mathbf{x}_{1k}^T\mathbf{B}_1 - \mathbf{x}_{2k}^T\mathbf{B}_2$, showing a removal of the influence of both $\mathbf{x}_{1k}$ and $\mathbf{x}_{2k}$, and those of the second component as $e_k^* = \mathbf{x}_k^T\mathbf{B} - \mathbf{x}_{1k}^T\mathbf{B}_1^*$, showing a removal of the influence on $\mathbf{x}_k^T\mathbf{B}$ (rather than on y_k) of $\mathbf{x}_{1k}$ alone. The same pattern holds for (a), in that both $\mathbf{x}_{1k}$ and $\mathbf{x}_{2k}$ are removed in the first residual, and $\mathbf{x}_{1k}$ alone in the second. Cases (a) and (b) differ in the **B**-coefficients of the two kinds of residuals. The automated linearization of (a) readily reveals the form of these **B** -coefficients. We do not show them here. The important point is that the influence of the **x**-vectors is removed according to a common pattern, although the values of e_k and e_k^* are not the same. Thus we can expect that (a) and (b) will usually generate rather small differences in the variance of the corresponding $\hat{Y}_{CAL}$ estimators. This is confirmed by the simulations in Estevao and Särndal (2002). For unusual relationships between y_k, $\mathbf{x}_{1k}$ and $\mathbf{x}_{2k}$, the differences can be more significant. Further studies are needed to examine this.

4 Calibration estimation in two-stage sampling

We start from the usual formulation of sampling in two stages. A sample of units is realized by two-stage selection from a population $U = \{1, 2, ..., k, ..., N\}$ grouped into clusters. This design involves sampling from two distinct populations of interest: (i) the population of first stage clusters, $U_I = \{1, 2, ..., i, ..., N_I\}$, and (ii) the population of second stage units $U = \{1, 2, ..., k, ..., N\}$. For simplicity, we refer to them as the population of clusters and the population of units respectively. The population $U = \{1, 2, ..., k, ..., N\}$ is the union of all the units U_i, $i \in U_I$ in the N_I clusters.

First, we draw a sample of clusters s_I from U_I, with known first-stage inclusion probabilities π_{1i}. The first-stage sampling weights are $a_{1i} = 1/\pi_{1i}$ for $i \in s_I$. At the second stage, we sample units within each of the selected clusters. From U_i, we draw a sample s_i of units, with known second-stage probabilities $\pi_{k|i}$ conditional on s_i. The conditional sampling weights are $a_{k|i} = 1/\pi_{k|i}$ for $k \in s_i$. Thus, $a_k = a_{1i}a_{k|i}$ is the overall sampling weight for unit *k*, and $s = \bigcup_{i \in s_I} s_i$ is the sample of units. The value y_k of the variable of interest is observed for all units $k \in s$. We want to estimate the total $Y = \sum_U y_k$, but more efficiently than with the simple unbiased $\hat{Y} = \sum_s a_k y_k = \sum_{s_I} a_{1i}(\sum_{s_i} a_{k|i} y_k)$.

In general, auxiliary information exists for both the units and the clusters. We denote by $\mathbf{x}_{1i}$ an auxiliary vector value associated with cluster *i*, and by $\mathbf{x}_k$ an auxiliary vector value associated with unit *k*. We consider the following information to be available:

(i) The cluster population vector total $\sum_{U_I} \mathbf{x}_{1i}$ is known.
(ii) For every $i \in s_I$, the cluster vector $\mathbf{x}_{1i}$ is known.
(iii) The unit population vector total $\sum_U \mathbf{x}_k$ is known.
(iv) For every $k \in s$, the unit vector $\mathbf{x}_k$ is known.

If $\mathbf{x}_{1i}$ is known for every $i \in U_I$, then (i) and (ii) are met. This occurs, for example, in area sampling where each cluster is a geographical entity for which we have a useful auxiliary measurement vector, for example, the surface area and/or the number of inhabitants. On the other hand, it is unlikely that we would have information $\mathbf{x}_k$ about every unit $k \in U$ in a survey where the absence of a list frame of units precludes single stage sampling and forces us to use two-stage sampling. But conditions (iii) and (iv) are met if $\mathbf{x}_k$ is recorded for all sampled units and the total $\sum_U \mathbf{x}_k$ can be "imported" from an accurate outside source, a census or a census projection, as it is, for example, in the Canadian Labour Force survey. This section examines calibration estimators derived from some or all of the information (i) to (iv).

The information is somewhat different when there is a known value $\mathbf{x}_k$ for every unit $k \in U_i$, where U_i is a selected cluster, $i \in s_I$. This case is covered by (i) to (iv) and we need not consider it, because the known cluster total $\mathbf{t}_{\mathbf{x}i} = \sum_{U_i} \mathbf{x}_k$ for $i \in s_I$ can then be entered into $\mathbf{x}_{1i}$ in (ii), assuming $\sum_{U_I} \mathbf{t}_{\mathbf{x}i} = \sum_U \mathbf{x}_k$ is also known.

Surveys involving sampling of clusters often have the double objective of computing estimates of totals for both the population of units *U* (referred here as *unit statistics*) and the population of clusters U_I (*cluster statistics*). Then, we observe both the value of a cluster

variable of interest, $y_{\mathrm{I}i}$ for $i \in s_{\mathrm{I}}$, and the value of a unit variable of interest, y_k for $k \in s$. For example, if households are clusters, $y_{\mathrm{I}i}$ may be the value for household *i* of the variable y_{I} = household income; and if units are persons in the households, y_k may be the value of the variable of interest *y* = employment status (0 if employed, 1 if unemployed).

The totals to be estimated are then $Y_{\mathrm{I}} = \sum_{U_{\mathrm{I}}} y_{\mathrm{I}i}$ for statistics on household income, and $Y = \sum_U y_k$ for statistics on individuals' employment. We thus examine the calibration estimators $\hat{Y}_{\mathrm{I},CAL} = \sum_{s_{\mathrm{I}}} w_{\mathrm{I}i} y_{\mathrm{I}i}$ and $\hat{Y}_{CAL} = \sum_s w_k y_k$ with cluster weights $w_{\mathrm{I}i}$ satisfying

$$\sum_{s_{\mathrm{I}}} w_{\mathrm{I}i} \mathbf{x}_{\mathrm{I}i} = \sum_{U_{\mathrm{I}}} \mathbf{x}_{\mathrm{I}i} \qquad (4.1)$$

and unit weights w_k satisfying

$$\sum_s w_k \mathbf{x}_k = \sum_U \mathbf{x}_k \ . \qquad (4.2)$$

We also allow for the fact that many two-stage designs call for some form of *integrated weighting*. Its objective is to impose a simple relation between a cluster weight $w_{\mathrm{I}i}$ and the weights w_k for the selected units k of that cluster. The interest in integrated weighting is promoted by Eurostat in its efforts to harmonize the estimation methods used by the member states of the European Union. Also, integrated weighting schemes are of interest for the further development of generalized estimation systems such as Bascula, CLAN and GES. We examine two options for integrated weighting:

(1) $\sum_{s_i} w_k = N_i w_{\mathrm{I}i}$ for every $i \in s_{\mathrm{I}}$, where N_i is the known size of cluster *i*.

(2) $w_k = a_{k|i} w_{\mathrm{I}i}$ for the selected units k in cluster $i \in s_{\mathrm{I}}$.

Each option imposes a simple relationship between the w_k and the $w_{\mathrm{I}i}$. Depending on the option selected, we can write (4.1) and (4.2) in terms of either w_k or $w_{\mathrm{I}i}$. We assume that the resulting set of equations is consistent. Depending on the choice, there is some effect on the precision of the resulting calibration estimates, as discussed in this section. Option (1) is based on the requirement that the estimated number of units within any group of clusters must be the same whether the cluster weights or the unit weights are used to create that estimate. Option (2) preserves the conditional design weights. One can argue that option (2) is slightly simpler than (1) but it actually imposes more severe restrictions on the unit weights. As we see later, this has implications on the variance of the estimators.

A special case of (2) that has drawn considerable attention occurs for single stage cluster sampling, see for example Lemaître and Dufour (1987), Andersson (1997), and Nieuwenbroek (1993). Since all *k* in cluster *i* are observed, $a_{k|i} = 1$ and (2) implies $w_k = w_{\mathrm{I}i}$. It is practical to assign the same weight to all units in a cluster, for the calculation of unit statistics, and this common weight is the cluster weight for cluster statistics. The approach of Lemaître and Dufour (1987) differs from ours. They find w_k to satisfy (4.2) but in such a way that the known auxiliary vector value $\mathbf{x}_k$ is replaced by one and the same constructed value, $\sum_{U_i} \mathbf{x}_k / N_i$, for every k in cluster i. By contrast, we keep the individual $\mathbf{x}_k$ and use one of the integrated weighting options to set up the calibration problem, leading to integrated w_k and $w_{\mathrm{I}i}$. The calculation of the weights of the calibration estimators $\hat{Y}_{CAL} = \sum_s w_k y_k$ and $\hat{Y}_{\mathrm{I},CAL} = \sum_{s_{\mathrm{I}}} w_{\mathrm{I}i} y_{\mathrm{I}i}$ is described below.

(a) Non-integrated calibration: Starting from $a_{\mathrm{I}i}$, compute cluster weights $w_{\mathrm{I}i}$ for $i \in s_{\mathrm{I}}$, calibrated to the cluster information in the manner of (4.1); in an independent second calibration, starting from $a_k = a_{\mathrm{I}i} a_{k|i}$, compute unit weights w_k for $k \in s$ calibrated to the unit information as stated in (4.2).

(b) Calibration with integration option (1): In (4.1), replace $w_{\mathrm{I}i}$ by $\sum_{s_i} w_k / N_i$, making that equation a function of the w_k. Assign the "equal shares" value $\mathbf{x}_{ik} = \mathbf{x}_{\mathrm{I}i} / N_i$ to every selected unit k in cluster *i*. Then starting from $a_k = a_{\mathrm{I}i} a_{k|i}$, compute unit weights w_k for $k \in s$, calibrated to satisfy $\sum_s w_k \begin{pmatrix} \mathbf{x}_{ik} \\ \mathbf{x}_k \end{pmatrix} = \begin{pmatrix} \sum_{U_{\mathrm{I}}} \mathbf{x}_{\mathrm{I}i} \\ \sum_U \mathbf{x}_k \end{pmatrix}$. Then compute the cluster weights as $w_{\mathrm{I}i} = \sum_{s_i} w_k / N_i$.

(c) Calibration with integration option (2): In (4.2), replace w_k by $a_{k|i} w_{\mathrm{I}i}$, making that equation a function of the $w_{\mathrm{I}i}$. Starting from $a_{\mathrm{I}i}$, compute cluster weights $w_{\mathrm{I}i}$ for $i \in s_{\mathrm{I}}$, calibrated to satisfy

Case	Integrated Weighting Option	Method	Calibration Equation(s)
(a)	None	Using a_{1i} as starting weights, compute w_{1i} to satisfy (4.1). Independently, using a_k as starting weights, compute w_k to satisfy (4.2).	$\sum_{s_1} w_{1i}\mathbf{x}_{1i} = \sum_{U_1}\mathbf{x}_{1i}$ $\sum_s w_k\mathbf{x}_k = \sum_U \mathbf{x}_k$
(b)	$\sum_{s_i} w_k = N_i w_{1i}$	In (4.1), replace w_{1i} by $\sum_{s_i} w_k / N_i$. With a_k as starting weights, compute w_k to satisfy both (4.1) and (4.2). Then compute the w_{1i}.	$\sum_s w_k \begin{pmatrix}\mathbf{x}_{ik}\\ \mathbf{x}_k\end{pmatrix} = \begin{pmatrix}\sum_{U_1}\mathbf{x}_{1i}\\ \sum_U \mathbf{x}_k\end{pmatrix}$
(c)	$w_k = a_{k\mid i} w_{1i}$	In (4.2), replace w_k by $a_{k\mid i} w_{1i}$. With a_{1i} as starting weights, compute w_{1i} to satisfy both (4.1) and (4.2). Then compute the w_k.	$\sum_{s_1} w_{1i}\begin{pmatrix}\mathbf{x}_{1i}\\ \hat{\mathbf{t}}_{\mathbf{x}i}\end{pmatrix} = \begin{pmatrix}\sum_{U_1}\mathbf{x}_{1i}\\ \sum_U \mathbf{x}_k\end{pmatrix}$

Table 1. Summary of Calibrated Weighting Methods for Two-Stage Estimation.

$\sum_{s_1} w_{1i}\begin{pmatrix}\mathbf{x}_{1i}\\ \hat{\mathbf{t}}_{\mathbf{x}i}\end{pmatrix} = \begin{pmatrix}\sum_{U_1}\mathbf{x}_{1i}\\ \sum_U \mathbf{x}_k\end{pmatrix}$, where $\hat{\mathbf{t}}_{\mathbf{x}i} = \sum_{s_i} a_{k\mid i}\mathbf{x}_k$ is the sample-weighted, unbiased estimator of $\mathbf{t}_{\mathbf{x}i} = \sum_{U_i}\mathbf{x}_k$. Then compute the unit weights as $w_k = a_{k\mid i} w_{1i}$.

The procedures for (a), (b) and (c) are summarized in Table 1. All three cases reduce to a weight calculation of the form (2.1), just as in single-stage unit sampling. That is, despite the two stages of sampling, the point estimation does not become any more complex than in single-stage unit sampling. A software programmed to compute formula (2.1), such as CLAN97 or GES, can be used to compute the calibration estimators in (a), (b) and (c). However, the two-stage design leads to a more complicated variance than in section 2. The variance has two components, one for each stage of selection, as shown later in the section.

We have three cases, (a) to (c), and for each, both cluster statistics and unit statistics are examined. There are thus $3\times 2 = 6$ situations to examine. For each of these, the approximate variance of the calibrated estimator (which equals the variance of the linearized statistic) has the form $\mathrm{V}_1(\mathrm{E}_c) + \mathrm{E}_1(\mathrm{V}_c)$, where $\mathrm{V}_1(\mathrm{E}_c)$ is the first stage variance component and $\mathrm{E}_1(\mathrm{V}_c)$ the second stage variance component. The latter is zero if there is no sampling at the second stage, that is, all units in selected clusters are observed (single stage cluster sampling). It is straightforward to carry out the automated linearization in the 6 situations. This leads to the expression for the residuals, one for each component of variance. These residuals are summarized in Table 2.

Consider case (b) for unit statistics. The total to estimate is $Y = \sum_U y_k$. The weights for $\hat{Y}_{CAL} = \sum_s w_k y_k$ are computed for $k \in s$ as $w_k = a_k(1 + \lambda_s^T \mathbf{z}_k)$ with

$$\lambda_s^T = \left(\begin{pmatrix}\sum_{U_1}\mathbf{x}_{1i}\\ \sum_U \mathbf{x}_k\end{pmatrix} - \begin{pmatrix}\sum_s a_k\mathbf{x}_{ik}\\ \sum_s a_k\mathbf{x}_k\end{pmatrix}\right)^T \left(\sum_s a_k\mathbf{z}_k\begin{pmatrix}\mathbf{x}_{ik}\\ \mathbf{x}_k\end{pmatrix}^T\right)^{-1},$$

where $\mathbf{z}_k$ is any valid instrument, and $\mathbf{x}_{ik} = \mathbf{x}_{1i}/N_i$ for every selected unit k in cluster. The estimator of the total for units, is $\hat{Y}_{CAL} = \sum_s w_k y_k$. Automated linearization gives $\hat{Y}_{CAL} = \hat{Y}_{CAL,lin} + R$, where R is the lower order term

$$R = -\begin{pmatrix}\sum_s a_k\mathbf{x}_{ik} - \sum_{U_1}\mathbf{x}_{1i}\\ \sum_s a_k\mathbf{x}_k - \sum_U \mathbf{x}_k\end{pmatrix}^T (\hat{\mathbf{B}} - \mathbf{B})$$

with $\hat{\mathbf{B}} = \begin{pmatrix}\hat{\mathbf{B}}_1\\ \hat{\mathbf{B}}_2\end{pmatrix} = \left(\sum_s a_k\mathbf{z}_k\begin{pmatrix}\mathbf{x}_{ik}\\ \mathbf{x}_k\end{pmatrix}^T\right)^{-1}\left(\sum_s a_k\mathbf{z}_k y_k\right)$, and the linearized statistic is

$$\hat{Y}_{CAL,lin} = \sum_s a_k e_k + \begin{pmatrix}\sum_{U_1}\mathbf{x}_{1i}\\ \sum_U \mathbf{x}_k\end{pmatrix}^T \begin{pmatrix}\mathbf{B}_1\\ \mathbf{B}_2\end{pmatrix}. \qquad (4.3)$$

The second term on the right hand side is a constant, and the preceding random term, $\sum_s a_k e_k = \sum_{s_1} a_{1i}\left(\sum_{s_i} a_{k\mid i} e_k\right)$, has the residuals

$e_k = y_k - \begin{pmatrix} \mathbf{x}_{ik} \\ \mathbf{x}_k \end{pmatrix}^T \begin{pmatrix} \mathbf{B}_1 \\ \mathbf{B}_2 \end{pmatrix}$ with

$$\mathbf{B} = \begin{pmatrix} \mathbf{B}_1 \\ \mathbf{B}_2 \end{pmatrix} = \left(\sum_U \mathbf{z}_k \begin{pmatrix} \mathbf{x}_{ik} \\ \mathbf{x}_k \end{pmatrix} \right)^{-1} \left(\sum_U \mathbf{z}_k y_k \right). \quad (4.4)$$

By conditioning on s_{I}, and using $\mathrm{Var}(\hat{Y}_{CAL,lin}) = \mathrm{V}_1(\mathrm{E}_c) + \mathrm{E}_1(\mathrm{V}_c)$, we find

$$\begin{aligned} \mathrm{Var}(\hat{Y}_{CAL}) &\approx \mathrm{Var}(\hat{Y}_{CAL,lin}) \\ &= \sum\sum_{U_1} F_{1ij} e_{1i} e_{1j} + \sum_{U_1} a_{1i} \mathrm{V}_i \end{aligned} \quad (4.5)$$

where $e_{1i} = \sum_{U_i} e_k = t_{yi} - \begin{pmatrix} \mathbf{x}_{1i} \\ \mathbf{t}_{\mathbf{x}i} \end{pmatrix}^T \begin{pmatrix} \mathbf{B}_1 \\ \mathbf{B}_2 \end{pmatrix}$

and $\mathrm{V}_i = \sum\sum_{U_i} F_{k\ell|i} e_k e_\ell$ with $F_{k\ell|i} = \dfrac{a_{k|i} a_{\ell|i}}{a_{k\ell|i}} - 1$ and

$F_{1ij} = \dfrac{a_{1i} a_{1j}}{a_{1ij}} - 1$.

The residual e_{1i} in the first component $\mathrm{V}_1(\mathrm{E}_c) = \sum\sum_{U_1} F_{1ij} e_{1i} e_{1j}$ equals the cluster total of the residuals e_k in the second component $\mathrm{E}_1(\mathrm{V}_c) = \sum_{U_1} a_{1i} \mathrm{V}_i$. Both e_{1i} and e_k have their magnitude reduced by both the cluster auxiliary and the unit auxiliary. The regressor is $(\mathbf{x}_{1i}^T, \mathbf{t}_{\mathbf{x}i}^T)$ in e_{1i} and $(\mathbf{x}_{ik}^T, \mathbf{x}_k^T)$ in e_k. In particular, in single stage cluster sampling, $\mathrm{V}_i = 0$ for all i, and only the first variance component remains.

Consider now case (b) for cluster statistics. The total $Y_{\mathrm{I}} = \sum_{U_1} y_{1i}$ is estimated by $\hat{Y}_{\mathrm{I},CAL} = \sum_{s_1} w_{1i} y_{1i}$, where the w_{1i} are computed from the already available w_k as $w_{1i} = \sum_{s_i} w_k / N_i$. We can write $\hat{Y}_{\mathrm{I},CAL}$ as a sum of unit values $\hat{Y}_{\mathrm{I},CAL} = \sum_s w_k y_{ik}$, if we define $y_{ik} = y_{1i} / N_i$ for all k in cluster i. To obtain its variance, we simply change the variable of interest in equations (4.3) to (4.5). We replace y_k by y_{ik}, keeping other quantities intact. Denote by $\mathbf{B}^{(c)} = \begin{pmatrix} \mathbf{B}_1^{(c)} \\ \mathbf{B}_2^{(c)} \end{pmatrix}$ the result of replacing y_k by y_{ik} in $\mathbf{B}$ of (4.4). The approximation to $\mathrm{Var}(\hat{Y}_{\mathrm{I},CAL})$ is then given by (4.5) with $e_k = y_{ik} - \begin{pmatrix} \mathbf{x}_{ik} \\ \mathbf{x}_k \end{pmatrix}^T \begin{pmatrix} \mathbf{B}_1^{(c)} \\ \mathbf{B}_2^{(c)} \end{pmatrix}$ and $e_{1i} = y_{1i} - \begin{pmatrix} \mathbf{x}_{1i} \\ \mathbf{t}_{\mathbf{x}i} \end{pmatrix}^T \begin{pmatrix} \mathbf{B}_1^{(c)} \\ \mathbf{B}_2^{(c)} \end{pmatrix}$. The residuals for unit and cluster statistics, are summarized in Table 2.

Another issue of interest in case (b) is the choice of the instrument $\mathbf{z}_k$. The standard choice is to take $\mathbf{z}_k = \begin{pmatrix} \mathbf{x}_{ik} \\ \mathbf{x}_k \end{pmatrix}$ for k in cluster i. But one can derive a $\mathbf{B} = \mathbf{B}^0$ that minimizes (4.5), with a corresponding optimal $\mathbf{z}_k = \mathbf{z}_k^0$. Some algebra shows that $\mathbf{z}_k^0 = a_{1i} \sum_{\ell \in U_i} F_{k\ell|i} \begin{pmatrix} \mathbf{x}_{i\ell} \\ \mathbf{x}_\ell \end{pmatrix} + \sum_{j \in U_1} F_{1ij} \begin{pmatrix} \mathbf{x}_{1j} \\ \mathbf{t}_{\mathbf{x}j} \end{pmatrix}$ for k in cluster i. It is no surprise that $\mathbf{z}_k^0$ depends on the sampling design at both stages. Future work will examine when $\mathbf{z}_k = \mathbf{z}_k^0$ is a valid instrument and whether $\mathbf{z}_k = \mathbf{z}_k^0$ gives any appreciable variance advantage over the simple standard $\mathbf{z}_k = \mathbf{x}_k$. If this advantage is minimal, the preferred choice in practice is the simple $\mathbf{z}_k = \mathbf{x}_k$.

Consider now case (c). The calibrated estimator of the cluster total $Y_{\mathrm{I}} = \sum_{U_1} y_{1i}$ is $\hat{Y}_{\mathrm{I},CAL} = \sum_{s_1} w_{1i} y_{1i}$, with the cluster weights $w_{1i} = a_{1i}(1 + \lambda_{s_1}^T \mathbf{z}_i)$ for $i \in s_{\mathrm{I}}$, where

$$\lambda_{s_1}^T = \left(\begin{pmatrix} \sum_{U_1} \mathbf{x}_{1i} \\ \sum_{U_1} \mathbf{t}_{\mathbf{x}i} \end{pmatrix} - \begin{pmatrix} \sum_{s_1} a_{1i} \mathbf{x}_{1i} \\ \sum_{s_1} a_{1i} \hat{\mathbf{t}}_{\mathbf{x}i} \end{pmatrix} \right)^T \left(\sum_{s_1} a_{1i} \mathbf{z}_i \begin{pmatrix} \mathbf{x}_{1i} \\ \hat{\mathbf{t}}_{\mathbf{x}i} \end{pmatrix} \right)^{-1}$$

and $\hat{\mathbf{t}}_{\mathbf{x}i} = \sum_{s_i} a_{k|i} \mathbf{x}_k$. The calibrated estimator of the unit total $Y = \sum_U y_k$ is $\hat{Y}_{CAL} = \sum_s w_k y_k$, where the integrated unit weights for k in cluster i are $w_k = a_{k|i} w_{1i}$, using the computed w_{1i}. We can use automated linearization on $\hat{Y}_{\mathrm{I},CAL}$ and $\hat{Y}_{CAL}$ to obtain the linearized statistic and the residuals that determine the two components of the approximate variance. The details of the derivations are omitted. The residuals, given in Table 2, are expressed in terms of the vectors

$$\mathbf{B}_{\mathrm{I}} = \begin{pmatrix} \mathbf{B}_{\mathrm{I}1} \\ \mathbf{B}_{\mathrm{I}2} \end{pmatrix} = \left(\sum_{U_1} \mathbf{z}_i \begin{pmatrix} \mathbf{x}_{1i} \\ \mathbf{t}_{\mathbf{x}i} \end{pmatrix}^T \right)^{-1} \left(\sum_{U_1} \mathbf{z}_i y_{1i} \right) \quad \text{and}$$

$\mathbf{B}_{\mathrm{I}}^{(u)} = \begin{pmatrix} \mathbf{B}_{\mathrm{I}1}^{(u)} \\ \mathbf{B}_{\mathrm{I}2}^{(u)} \end{pmatrix}$, where $\mathbf{B}_{\mathrm{I}}^{(u)}$ is obtained by replacing y_{1i} in $\mathbf{B}_{\mathrm{I}}$ by $t_{yi} = \sum_{U_i} y_k$.

The residuals given in Table 2 for case (a) are simple to explain. In case (a) for unit statistics, the

Case	Integrated Weighting Option	Estimation of a total for	Residual $e_{\mathrm{I}i}$	Residual e_k
(a)	None	Units	$t_{yi} - \mathbf{t}_{\mathbf{x}i}^T \mathbf{B}$	$y_k - \mathbf{x}_k^T \mathbf{B}$
		Clusters	$y_{\mathrm{I}i} - \mathbf{x}_{\mathrm{I}i}^T \mathbf{B}_{\mathrm{I}}$	0
(b)	$\sum_{s_i} w_k = N_i w_{\mathrm{I}i}$	Units	$t_{yi} - \mathbf{x}_{\mathrm{I}i}^T \mathbf{B}_1 - \mathbf{t}_{\mathbf{x}i}^T \mathbf{B}_2$	$y_k - \mathbf{x}_{ik}^T \mathbf{B}_1 - \mathbf{x}_k^T \mathbf{B}_2$
		Clusters	$y_{\mathrm{I}i} - \mathbf{x}_{\mathrm{I}i}^T \mathbf{B}_1^{(c)} - \mathbf{t}_{\mathbf{x}i}^T \mathbf{B}_2^{(c)}$	$y_{ik} - \mathbf{x}_{ik}^T \mathbf{B}_1^{(c)} - \mathbf{x}_k^T \mathbf{B}_2^{(c)}$
(c)	$w_k = a_{k\|i} w_{\mathrm{I}i}$	Units	$t_{yi} - \mathbf{x}_{\mathrm{I}i}^T \mathbf{B}_{\mathrm{I}1}^{(u)} - \mathbf{t}_{\mathbf{x}i}^T \mathbf{B}_{\mathrm{I}2}^{(u)}$	$y_k - \mathbf{x}_k^T \mathbf{B}_{\mathrm{I}2}^{(u)}$
		Clusters	$y_{\mathrm{I}i} - \mathbf{x}_{\mathrm{I}i}^T \mathbf{B}_{\mathrm{I}1} - \mathbf{t}_{\mathbf{x}i}^T \mathbf{B}_{\mathrm{I}2}$	$\mathbf{x}_k^T \mathbf{B}_{\mathrm{I}2}$

Table 2. Summary of residuals in the Components of the Variance (4.5) for Two-Stage Sampling and Estimation. The notation is explained in the text.

automated linearization of $\hat{Y}_{CAL} = \sum_s w_k y_k$ produces $\mathbf{B} = \left(\sum_U \mathbf{z}_k \mathbf{x}_k^T\right)^{-1} \left(\sum_U \mathbf{z}_k y_k\right)$, and the residuals in Table 2 follow from case (b) for unit statistics by setting $\mathbf{x}_{\mathrm{I}i} = 0$ and $\mathbf{x}_{ik} = 0$ for all *i* and *k*, because case (a) involves no cluster related information in estimating for units. For case (a) for cluster statistics, the automated linearization of $\hat{Y}_{\mathrm{I},CAL} = \sum_{s_{\mathrm{I}}} w_{\mathrm{I}i} y_{\mathrm{I}i}$ leads to $\mathbf{B}_{\mathrm{I}} = \left(\sum_U \mathbf{z}_i \mathbf{x}_{\mathrm{I}i}^T\right)^{-1} \left(\sum_U \mathbf{z}_i y_{\mathrm{I}i}\right)$, and the residuals follow from case (c) for cluster statistics by setting $\mathbf{x}_k = 0$ and $\mathbf{t}_{\mathbf{x}i} = 0$ for all *i* and *k*, because case (a) uses no unit related information in estimating for clusters.

An examination of the residuals in Table 2 leads to some interesting conclusions. Let us first compare the residuals for unit statistics. In (b) and (c), the residuals $e_{\mathrm{I}i}$ are adjusted for both $\mathbf{x}_{\mathrm{I}i}$ and $\mathbf{t}_{\mathbf{x}i}$, but in (a) they are only adjusted for $\mathbf{t}_{\mathbf{x}i}$. Thus (b) and (c) are better than (a) for the first component. The residual e_k is adjusted for both auxiliaries in (b), but not in (a) and (c), where it is only adjusted for $\mathbf{x}_k$. Thus (b) has the best potential for efficient estimation of unit statistics. Compare now the residuals for cluster statistics. In (b) and (c), the residual $e_{\mathrm{I}i}$ is adjusted for both $\mathbf{x}_{\mathrm{I}i}$ and $\mathbf{t}_{\mathbf{x}i}$, but in (a) it is only adjusted for $\mathbf{x}_{\mathrm{I}i}$. By design, the residual e_k is always zero in (a). A particularly unfavourable situation for the second variance component arises for case (c), where the residual is $\mathbf{x}_k^T \mathbf{B}_{\mathrm{I}2}$. Thus (a) or possibly (b) has the best potential for efficient estimation of cluster statistics.

5 Summary and discussion

The question of efficient weighting of the observed values has always been important in survey sampling theory. An important step was the formulation in 1952 of the HT estimator, prescribing that the weight of each unit equals the inverse of the probability of its inclusion in the sample. Thus, in stratified simple random sampling (STSRS), the weight given to all units sampled from a stratum equals the inverse of the sampling rate in the stratum. Neyman's convincing results in 1934 on optimal estimation under STSRS laid the foundation of what is now commonly called the design-based theory of estimation. Another important principle embodied in HT estimation is that the same weight system applies to all *y*-variables of interest in a multi-purpose survey. This preserves the design unbiasedness for every *y*-variable. Assuming no non-response, the sampling design alone determines once and for all the weighting and the construction of the point estimator.

The principle of a single weight system extends to the calibration estimators in this paper. However, unlike the sampling weights a_k, the calibrated weights w_k are calculated only after drawing the sample. They are

usually more efficient (give a smaller variance) than the a_k for every single *y*-variable and they produce estimators with a negligible bias.

The literature on calibration has been based on a model oriented construction of these estimators. Both the *model assisted* and *model dependent* approaches to calibration involve an explicit assumption of a linear superpopulation model between **x** and *y*. This model is of the form $y_k = \mathbf{x}_k^T\mathbf{B} + \varepsilon_k$ where it is assumed that $E(\varepsilon_k) = 0$ and $Var(\varepsilon_k) = c_k\sigma^2$ with $c_k > 0$. In practice however, this model is often invalid.

In our approach, the use of auxiliary information is not linked to model fitting. We define a parameterization of the calibration weights involving the instrument vector $\mathbf{z}_k$ and then apply the method of automated linearization to obtain a linear approximation of the calibration estimator. This linear approximation is a design-based function of a set of fixed but unknown population residuals determined implicitly without any modelling. The w_k are calculated using all or part of the available auxiliary information. We have shown how to do this for different designs including one-phase and two-stage designs. It is important to note that the construction of the point estimator has nothing to do with the *y*-variables; the same weights apply to all *y*-variables as is the case for the HT estimator. However, the calibration estimator can be considerably more efficient for some *y*-variables than others. This depends on the resulting population residuals.

References

Andersson, C. (1997). Continuous labour force surveys: performance analysis of a single weight procedure. Internal report, Statistical Methodology Unit, Statistics Sweden.

Binder, D. (1996). The right and wrong ways to Taylor linearize for single phase and two phase samples: a cookbook approach. *Survey Methodology* **22**, 17-22.

Binder, D. and Kovačević, M.S. (1995). Estimating some measures of income inequality from survey data: and application of the estimating equation approach. *Survey Methodology* **21**, 137-145.

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. (1999). Variance estimation for complex statistics and estimators: Linearization and residual techniques. *Survey Methodology* **25**, 193-203

Deville, J.C. (2002). Correction for non-response by generalized calibration. Report, ENSAI, France.

Estevao, V. and Särndal, C.E. (2002). The ten cases of auxiliary information for calibration in two-phase sampling. *Journal of Official Statistics* **18**, 233-255.

Huang, E.T. and Fuller, W.A. (1978). Nonnegative regression estimation for sample survey data. Proceedings Social Statistics Section, American Statistical Association, 300-305.

Lemaître, G.E. and Dufour, J. (1987). An integrated method for weighting persons and families. *Survey Methodology* **13**, 199-207.

Le Guennec, J. and Sautory, O. (2002). Application of generalized calibration to the correction of non-response: An experiment. Report, ENSAI, France.

Montanari, G.E. (1987). Post-sampling efficient QR-prediction in large-sample surveys. *International Statistical Review* **55**, 191-202.

Nieuwenbroek, N.J. (1993). An integrated method for weighting characteristics of persons and households using the linear regression estimator. Internal report, Central Bureau of Statistics, The Netherlands.

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

Théberge, A. (1999). Extensions of Calibration Estimators in Survey Sampling. *Journal of the American Statistical Association* **94**, 635-644.

Woodruff, R.S. (1971). A simple method for approximating the variance of a complicated estimate. *Journal of the American Statistical Association* **66**, 411-414.