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 \( U = \{1, 2, \ldots, k, \ldots, N\} \) from which a probability sample \( s \) is drawn. We denote by \( \pi_k \) 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 \( Y = \sum_{k \in s} y_k \).

We denote by \( x \) an auxiliary vector of dimension \( J > 1 \), and by \( x_k \) its value for unit \( k \). We assume that we have the following auxiliary information:

(i) The population vector total \( X = \sum_{k} x_k \) is known.
(ii) The vector value \( x_k \) is known for every \( k \in s \).

Here, \( X \) is assumed known from an outside source such as a census. If we know the value \( x_k \) for every \( k \in U \), as when \( x_k \) 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 \( X \) as \( \hat{X} = \sum_{k} a_k x_k \). Under general conditions \( N^{-1}(\hat{X} - X) \) is \( O_p(n^{-1/2}) \).

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

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

$$w_k = a_k (1 + \hat{\lambda}^T_k \mathbf{z}_k), \quad k \in s$$  \hspace{1cm} (2.1)

where $\hat{\lambda}_k = (\mathbf{X} - \hat{\mathbf{X}})^T (\sum_i a_i \mathbf{z}_k \mathbf{x}_k^T)^{-1}$. 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 $(\sum_i a_i \mathbf{z}_k \mathbf{x}_k^T)$ 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 $\sum_k w_k \mathbf{x}_k = \mathbf{X}$, and the estimator can be written as

$$\hat{\mathbf{Y}}_{\text{CAL}} = \sum_k w_k \mathbf{y}_k = \hat{\mathbf{Y}} + (\mathbf{X}-\hat{\mathbf{X}})^T \hat{\mathbf{B}},$$

where $\hat{\mathbf{B}} = (\sum_i a_i \mathbf{z}_k \mathbf{x}_k^T)^{-1} \sum_i a_i \mathbf{z}_k \mathbf{y}_k$.

Here $\hat{\mathbf{B}}$ is a nonlinear design-weighted statistic, thus, it is not what we call a HT statistic. Although $\hat{\mathbf{Y}}$ is a linear statistic, the term $(\mathbf{X}-\hat{\mathbf{X}})^T \hat{\mathbf{B}}$ makes $\hat{\mathbf{Y}}_{\text{CAL}}$ a nonlinear estimator. This causes no problem for point estimation since $\hat{\mathbf{Y}}_{\text{CAL}}$ can be readily computed. But the nonlinear form of the estimator creates a problem for obtaining a simple exact expression for the variance of $\hat{\mathbf{Y}}_{\text{CAL}}$ 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 $\hat{\mathbf{Y}}_{\text{CAL}}$ 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, $\hat{\mathbf{Y}}_{\text{CAL,lin}}$, 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 $\hat{\mathbf{Y}}_{\text{CAL,lin}}$ to be a “sufficiently good” linear approximation to $\hat{\mathbf{Y}}_{\text{CAL}}$. Under general conditions,

$$N^{-1}(\hat{\mathbf{Y}}_{\text{CAL}} - \hat{\mathbf{Y}}_{\text{CAL,lin}}) \quad \text{is} \quad O_p(n^{-1/2}), \quad \text{not just} \quad O_p(n^{-1/2}),$$

permitting the easily derived variance of $\hat{\mathbf{Y}}_{\text{CAL,lin}}$ to be used as an accurate approximation of the variance of $\hat{\mathbf{Y}}_{\text{CAL}}$, 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 $\hat{\mathbf{Y}}_{\text{CAL,lin}}$. Automated linearization has two steps:

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

$$\hat{\mathbf{Y}}_{\text{CAL}} = \hat{\mathbf{Y}} + (\mathbf{X} - \hat{\mathbf{X}})^T \hat{\mathbf{B}} - (\mathbf{X} - \mathbf{X})^T (\hat{\mathbf{B}} - \mathbf{B})$$  \hspace{1cm} (2.2)

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

**Step 2.** Rewrite (2.2) as

$$\hat{\mathbf{Y}}_{\text{CAL}} = (\hat{\mathbf{Y}} - \hat{\mathbf{X}}^T \mathbf{B}) + \mathbf{X}^T \mathbf{B} - (\hat{\mathbf{X}} - \mathbf{X})^T (\hat{\mathbf{B}} - \mathbf{B})$$  \hspace{1cm} (2.3)

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

$$\hat{\mathbf{Y}}_{\text{CAL,lin}} = (\hat{\mathbf{Y}} - \hat{\mathbf{X}}^T \mathbf{B}) + \mathbf{X}^T \mathbf{B} = \sum_i a_i e_k + \mathbf{X}^T \mathbf{B}$$  \hspace{1cm} (2.4)

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

1347
Our point estimator of $Y$ is $\hat{Y}_{CAL}$. It has a small bias, $E(\hat{Y}_{CAL}) - Y = -E\{ (\hat{X} - X)^T (\hat{B} - B) \}$, since $-N^{-1} E\{ (\hat{X} - X)^T (\hat{B} - B) \}$ is of order $O(n^{-1})$. Therefore, the variance of $\hat{Y}_{CAL}$ is approximately the variance of the linearized statistic $\hat{Y}_{CAL}$. Since $X'B$ is a constant, the use of auxiliary information reduces the variance of the estimator from $\text{Var}(\sum a_i y_i)$ to approximately $\text{Var}(\sum a_i e_k)$. It is important to note that the $e_k$ are fixed but unknown values and that $\sum a_i e_k$ 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 $\sum a_i e_k$ is a HT statistic, we obtain immediately,

$\text{Var}(\hat{Y}_{CAL}) \equiv \text{Var}(\hat{Y}_{CAL,lin}) = \sum \sum F_{kl} \hat{e}_k \hat{e}_l$  

(2.5)

where $F_{kl} = \frac{a_i a_j}{a_k a_l} - 1$ for $\ell \not= k$, $F_{kl} = F_{lk} = a_k - 1$ for $\ell = k$, with $a_k = 1/\pi_k$, where $\pi_k$ is the joint inclusion probability of $k$ and $\ell$. We use $\sum \sum l$ as shorthand for the double sum $\sum_{k \in U \cap l \in U}$. To estimate the variance of $\hat{Y}_{CAL}$, we use the sample-based analogue of (2.5),

$\hat{\text{Var}}(\hat{Y}_{CAL}) = \sum \sum (a_i a_j - a_k) \hat{e}_i \hat{e}_j$  

(2.6)

where $\hat{e}_k = y_k - \tilde{x}_k^T \hat{B}$ and $\sum \sum s$ stands for $\sum_{k \in U \cap l \in U}$.

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

$(\sum_k F_{kl} x_k x_l) B^a = \sum_k F_{kl} x_k y_k$.  

(2.7)

A comparison with the general form $(\sum_k z_k x_k^T) B = \sum_k z_k y_k$ defining $B$, suggests that an optimal instrument vector is $z_k = z_k^0$, where $z_k^0 = \sum_{i \in U} F_{ki} x_i$. The result agrees with Montanari's (1987) determination of $B$ so as to minimize the variance of the unbiased difference estimator $\hat{y} + (X - \hat{X})^T B$.

To see the features of the weights, let us write them as $w_k = a_k \{ 1 + (\sum x_i - \sum a_i x_i)^T (\sum a_i z_i x_i^T)^{-1} z_i \}$. 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 $x_k$, the instrument vector values $z_k$, and the known auxiliary vector of totals $\sum x_k$.

(ii) The $a_k$ are fixed by the design.

(iii) We are free to choose the $z_k$ as long as $\sum a_i z_i x_i^T$ is invertible.

(iv) The weights $w_k$ calibrate to the known totals $\sum x_k$ for any valid instrument vector $z_k$.

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

Some choices of $z_k$ are “better” than others. The optimal choice, as noted above, is $z_k = z_k^0 = \sum a_i F_{ki} x_i$. It makes sense that the optimal choice depends on the sampling design. The sample-based choice corresponding to $z_k^0$ is $z_k = z_k^* = a^{-1} \sum a_i F_{ki} x_i$. 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 $z_k$ 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 $x$. In the application of this approach it does not matter whether there exists a linear relationship between $y$ and $x$. 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 $x_k = (1, x_k)^T$, where $x_k$ is a scalar variable value. The required population information is $\sum x_k = (N, \sum x_k)^T$. Then the optimal instrument is found to be $z_k = z_k^* = \sum a_i F_{ki} x_i = \sum a_i F_{ki} x_i$.
In step one, starting from \( \bar{x}_i = \sum_i x_i / N \). The corresponding sample based choice is \( \bar{z}_i = \sum_i z_i / n \). However, both \( z_i^* \) and \( z_i^\prime* \) are invalid because the first component of these vectors is always zero, leading to a singular matrix \( \sum_i a_i z_i x_i^T \). We drop the first auxiliary variable with the known total \( N \) and work instead with the vector \( x_1 = x_2 \). This gives \( z_k = z_k^* = n - n - 1 (1 - 1)(x_k - \bar{x}_i) \). The result is the familiar \( \hat{Y}_{\text{CAL}} = N \{ y_k + (x_k - \bar{x}_i) y_k \} \) with \( b = (\sum_i (x_i - \bar{x}_i) y_i) / \sum_i (x_i - \bar{x}_i)^2 \). As is easily verified, this estimator gives \( \hat{Y}_{\text{CAL}} = N \) when \( y_k = 1 \) for all \( k \), and \( \hat{Y}_{\text{CAL}} = N \{ y_k + (x_k - \bar{x}_i) y_k \} \) when \( y_k = x_k \). That is, even though we must reduce the auxiliary vector from \( x_k = (1, x_k)^T \) to \( x_k = x_1 \), the resulting set of weights still reproduce the two known quantities \( N \) and \( \sum_i x_k \). No loss of information is incurred from the non-invertibility of \( \sum_i a_i x_i x_i^T \) with \( x_k = (1, x_k)^T \).

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 \( x_1 \)-vector with known totals.

Step 2 Point estimation: Specify a valid \( z_k \), 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 \( U = \{ 1, 2, \ldots, k, \ldots, N \} \), a large probability sample, \( s \), is drawn with known first-phase inclusion probabilities \( \pi_{1k} \). The first-phase sampling weights are \( a_{1k} = 1/\pi_{1k} \) for \( k \in s \). One or more variables are observed for \( k \in s \). Then, from \( s \), a sub-sample, \( s \), is drawn with known conditional probabilities \( \pi_{2k} \). The second-phase sampling weights are \( a_{2k} = 1/\pi_{2k} \), conditionally on the realized \( s \). 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 \( Y = \sum_{i} y_i \) than the two-phase double expansion estimator \( \hat{Y} = \sum_{i} a_i y_i \).

We need to consider two auxiliary vectors for each unit \( k \). We denote these by \( x_1 \) and \( x_2 \), with \( x_{1k} \) and \( x_{2k} \) representing their respective values for unit \( k \). Their dimensions are \( J_1 \geq 1 \) and \( J_2 \geq 1 \) respectively. The auxiliary information for \( x_1 \) and \( x_2 \) is as follows:

(i) The population vector total \( x_1 = \sum_i x_{1k} \) is known while the population vector total \( x_2 = \sum_i x_{2k} \) is not known.

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

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

The information given by (i), (ii) and (iii) is used to compute the weights for the calibration estimator \( \hat{Y}_{\text{CAL}} = \sum w_k y_k \) in an effort to improve on \( \hat{Y} = \sum a_i y_i \). 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 \( \sum_i w_k x_k = \left( \sum_i a_{1k} x_{1k} \right) / \sum_{i} a_{1i} x_{2i} \), with \( x_k = \left( x_{1k}, x_{2k} \right) \) 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 \( \sum_{k \in s} w_{1k} x_{1k} = \sum_{i} a_{1i} x_{1i} \). 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
\[ \sum w_i x_i = \sum w_i x_i^k \]. The final weights satisfy
\[ \sum w_i x_i = \left( \sum w_i x_i^k \right) \]. 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 \( w_i \) 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 \( x_{ik} \) for \( k \in s \). This may or may not lead to an increase in efficiency, depending on the relation between \( x_{ik} \) and \( v_k \). These questions are discussed in Estevao and Särndal (2002). We can use automated linearization to form the final 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
\[ w_{ik} = a_{ik} \left( 1 + \lambda_{ik}^T z_{ik} \right) \]
with
\[ \lambda_{ik} = \left( \sum a_{ik} x_{ik} - \sum a_{ik} x_{ik}^T (\sum a_{ik} z_{ik} x_{ik}^T)^{-1} \right) \]
for some valid instrument vector \( z_{ik} \). The \( w_{ik} \) are then used as input to compute the final calibrated weights as
\[ w_k = a_k \left( 1 + \lambda_k^T z_k \right) \]
where
\[ \lambda_k = \left( \sum w_{ik} x_{ik} - \sum a_k x_{ik}^T (\sum a_k z_k x_{ik}^T)^{-1} \right) \]
where \( x_k = \begin{pmatrix} x_{ik} \\ x_{ik}^T \end{pmatrix} \) and \( z_k \) is another valid instrument vector. We derive the variance by using automated linearization. First, we insert into \( \hat{Y}_{CAL} = \sum w_{ik} y_k \) the expression for \( w_k \). Then, we define \( \hat{B} = \begin{pmatrix} \hat{B}_1 \\ \hat{B}_2 \end{pmatrix} = \begin{pmatrix} \sum a_{ik} z_{ik} x_{ik}^T \sum a_{ik} z_{ik} x_{ik}^T \end{pmatrix} \) and center it on its population counterpart, the non-random vector \( \hat{B}^* = \begin{pmatrix} \hat{B}_1^* \\ \hat{B}_2^* \end{pmatrix} = \left( \sum a_{ik} z_{ik} x_{ik}^T \right)^{-1} \left( \sum a_{ik} z_{ik} x_{ik}^T \right) \). After some algebra, we then define the statistic \( \hat{B}^* = \left( \sum a_{ik} z_{ik} x_{ik}^T \right)^{-1} \left( \sum a_{ik} z_{ik} x_{ik}^T \right) \), which we center on its population counterpart \( \hat{B}^* = \left( \sum a_{ik} z_{ik} x_{ik}^T \right)^{-1} \left( \sum a_{ik} z_{ik} x_{ik}^T \right) \). The result is
\[ \hat{Y}_{CAL} = \sum \beta_k \left( y_k - x_k^T \hat{B} \right) \]
where \( R \) is the lower order term given by
\[ R = \begin{pmatrix} \hat{X}_1 - \hat{X}_3^T \end{pmatrix} \left( \hat{B}_1 - \hat{B}_1^* \right) \left( \hat{X}_1 - \hat{X}_3^T \right)^T \]
with \( \hat{X}_j = \sum a_{ik} x_{ik} \) and \( \hat{X}_j = \sum a_{ik} x_{ik} \), \( j = 1, 2 \). The term of main interest is the linear statistic
\[ \tilde{Y}_{CAL,lin} = \sum a_k e_k + \sum a_k e_k^* + x_k^T \hat{B}^*_1 \]
where \( x_k^T \hat{B}^*_1 = \left( \sum a_{ik} x_{ik} \right)^T \hat{B}^*_1 \) is a constant, and the residuals in the two random terms are
\[ e_k = y_k - x_k^T \hat{B} = y_k - x_k^T \hat{B}_1 \]
and
\[ e_k^* = x_k^T \hat{B} - x_k^T \hat{B}^* \].

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
\[ \text{Var}(\hat{Y}_{CAL}) = \text{Var}(\tilde{Y}_{CAL,lin}) = V_c \left( \hat{E}_c \right) + E_c \left( V_c \right) \]
where \( E_c = \sum a_{ik} \left( y_k - x_k^T \hat{B}_1^* \right) \) and \( V_c \) is the conditional variance of \( \sum a_k e_k \), given \( s_1 \). The expressions for \( V_c \left( \hat{E}_c \right) \) and \( E_c \left( V_c \right) \), 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 \( \hat{B} \) and \( \hat{B}_1^* \) instead of \( B \) and \( B_1^* \). Note that \( E_c \left( V_c \right) = 0 \) if sample selection stops after the first phase.

The first term, \( V_c \left( \hat{E}_c \right) \), is reduced by the presence in \( e_k^* \) of the regressor \( x_{ik} \) only, whereas the second term, \( E_c \left( V_c \right) \), gets reduced by both regressors, \( x_{ik} \) and \( x_{ik}^T \). 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, \( z_{ik} \) and \( z_k \). The simple standard choices are \( z_{ik} = x_{ik} \) and \( z_k = x_k \).

We comment briefly on the automated linearization in case (a). The outcome is also a variance of the form \( V_c \left( \hat{E}_c \right) + E_c \left( V_c \right) \), with one residual for the first
component, \( V_i(E_x) \), and another residual for the second, \( E_i(V_x) \). But these residuals are somewhat different in (a) and (b). For (b), we stated earlier in this section the first component residuals as 
\[ e_k = y_k - x_k^T B \cdot x_k^T B, \]
showing a removal of the influence of both \( x_{ik} \) and \( x_{ik} \), and those of the second component as 
\[ e_k' = x_k^T B \cdot x_k^T B, \]
showing a removal of the influence on \( x_k^T B \) (rather than on \( y_k \)) of \( x_{ik} \) alone. The same pattern holds for (a), in that both \( x_{ik} \) and \( x_{ik} \) are removed in the first residual, and \( x_{ik} \) 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 \( \hat{Y}_{CM} \) estimators. This is confirmed by the simulations in Estevao and Särndal (2002). For unusual relationships between \( y_k \), \( x_{ik} \) and \( x_{ik} \), 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 sampling. But conditions (iii) and (iv) are precluded single stage sampling and forces us to use two-stage sampling. We want to estimate the total \( Y = \sum_i y_i \), and more efficiently than with the simple unbiased \( \hat{Y} = \sum_i a_i y_i = \sum_i a_i (\sum_j a_{ij} y_j) \).

In general, auxiliary information exists for both the units and the clusters. We denote by \( x_{ii} \), an auxiliary vector value associated with cluster \( i \), and by \( x_k \), an auxiliary vector value associated with unit \( k \). We consider the following information to be available:

(i) The cluster population vector total \( \sum_i x_{ii} \) is known.
(ii) For every \( i \in s_i \), the cluster vector \( x_{ii} \) is known.
(iii) The unit population vector total \( \sum_k x_{kk} \) is known.
(iv) For every \( k \in s \), the unit vector \( x_k \) is known.

If \( x_{ii} \) is known for every \( i \in U_1 \), 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 \( x_k \) 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 \( x_{kk} \) is recorded for all sampled units and the total \( \sum_k x_{kk} \) 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 \( x_k \) for every unit \( k \in U \), where \( U \) 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 \( t_{s_i} = \sum_i x_k \) for \( i \in s_i \) can then be entered into \( x_{ii} \) in (ii), assuming \( \sum_i t_{s_i} = \sum i x_k \) 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 \( U \) (cluster statistics). Then, we observe both the value of a cluster
variable of interest, $y_i$, for $i \in s_1$, and the value of a unit variable of interest, $y_k$ for $k \in s$. For example, if households are clusters, $y_i$ 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 $Y = \sum y_i$ for statistics on household income, and $Y = \sum y_k$ for statistics on individuals’ employment. We thus examine the calibration estimators $\hat{Y}_{CAL} = \sum w_i y_i$ and $\hat{Y}_{CAL} = \sum w_k y_k$ with cluster weights $w_i$ satisfying

\[ \sum w_i x_i = \sum x_i \]  

and unit weights $w_k$ satisfying

\[ \sum w_k y_k = \sum y_k \]  

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 $w_i$, 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. $\sum w_i = N_i w_i$ for every $i \in s_1$, where $N_i$ is the known size of cluster $i$.
2. $w_k = a_{ik} w_i$ for the selected units $k$ in cluster $i \in s_1$.

Each option imposes a simple relationship between the $w_k$ and the $w_i$. Depending on the option selected, we can write (4.1) and (4.2) in terms of either $w_k$ or $w_i$. 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, $a_{ik} = 1$ and (2) implies $w_k = w_i$. 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 $x_k$ is replaced by one and the same constructed value, $\sum x_i / N_i$, for every $k$ in cluster $i$. By contrast, we keep the individual $x_k$ and use one of the integrated weighting options to set up the calibration problem, leading to integrated $w_i$ and $w_k$. The calculation of the weights of the calibration estimators $\hat{Y}_{CAL} = \sum w_i y_i$ and $\hat{Y}_{CAL} = \sum w_k y_k$ is described below.

(a) Non-integrated calibration: Starting from $a_{ij}$, compute cluster weights $w_i$ for $i \in s_1$, calibrated to the cluster information in the manner of (4.1); in an independent second calibration, starting from $a_k = a_i a_{ik}$, 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 $w_i$ by $\sum w_i / N_i$, making that equation a function of the $w_i$. Assign the “equal shares” value $x_{ik} = x_i / N_i$ to every selected unit $k$ in cluster $i$. Then starting from $a_k = a_i a_{ik}$, compute unit weights $w_k$ for $k \in s$, calibrated to satisfy

\[ \sum w_k \left( \frac{x_{ik}}{x_k} \right) = \left( \sum \frac{x_i}{\sum x_k} \right) \]

Then compute the cluster weights as $w_i = \sum w_k / N_i$.

(c) Calibration with integration option (2): In (4.2), replace $w_i$ by $a_{ik} w_i$, making that equation a function of the $w_i$. Starting from $a_{ik}$, compute cluster weights $w_i$ for $i \in s_1$, calibrated to satisfy
are thus 623 cluster statistics and unit statistics are examined. There
We have three cases, (a) to (c), and for each, both
section.
one for each stage of selection, as shown later in the
complex than in single-stage unit sampling. A software
the point estimation does not become any more
sampling. That is, despite the two stages of sampling,
in selected clusters are observed (single stage cluster
component of variance. These residuals are summarized
in Table 2.
Consider case (b) for unit statistics. The total to
estimate is \( Y = \sum d_i y_k \). The weights for \( \hat{Y}_{CAL} = \sum w_i y_k \)
are computed for \( k \in s \) as \( w_i = a_i (1 + x_i^T z_k) \) with
where \( z_k \) is any valid instrument, and \( x_i \) is selected unit in cluster. The estimator of the
total for units, is \( \hat{Y}_{CAL} = \sum w_i y_k \). Automated
linearization gives \( \hat{Y}_{CAL} = \hat{Y}_{CAL,lin} + R \), where \( R \) is the
lower order term
\[
R = -\left( \sum a_i x_k - \sum a_i x_i \right)^T (\hat{B} - B)
\]
with \( \hat{B} = \begin{pmatrix} \hat{B}_1 \\ \hat{B}_2 \end{pmatrix} = \begin{pmatrix} \sum a_i z_k (x_k) \end{pmatrix}^T (\sum a_i z_i y_k) \), and
the linearized statistic is
\[
\hat{Y}_{CAL,lin} = \sum a_i y_k + \begin{pmatrix} \sum a_i x_i \end{pmatrix}^T \begin{pmatrix} \hat{B}_1 \\ \hat{B}_2 \end{pmatrix}.
\] (4.3)

The second term on the right hand side is a
constant, and the preceding random term, \( \sum a_i e_k = \sum a_i (\sum a_i e_i) \), has the residuals

<table>
<thead>
<tr>
<th>Case</th>
<th>Integrated Weighting Option</th>
<th>Method</th>
<th>Calibration Equation(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a)</td>
<td>None</td>
<td>Using ( a_i ) as starting weights, compute ( w_i ) to satisfy (4.1). Independently, using ( a_k ) as starting weights, compute ( w_k ) to satisfy (4.2).</td>
<td></td>
</tr>
<tr>
<td>(b)</td>
<td>( \sum a_i w_i = N_i w_i )</td>
<td>In (4.1), replace ( w_i ) by ( \sum a_i w_i / N_i ). With ( a_i ) as starting weights, compute ( w_k ) to satisfy both (4.1) and (4.2). Then compute the ( w_i ).</td>
<td></td>
</tr>
<tr>
<td>(c)</td>
<td>( w_k = a_i w_i )</td>
<td>In (4.2), replace ( w_k ) by ( a_i w_i ). With ( a_i ) as starting weights, compute ( w_i ) to satisfy both (4.1) and (4.2). Then compute the ( w_k ).</td>
<td></td>
</tr>
</tbody>
</table>

Table 1. Summary of Calibrated Weighting Methods for Two-Stage Estimation.
where single stage cluster sampling, 0V equations (4.3) to (4.5). We replace variance, we simply change the variable of interest in other quantities intact. Denote by result of replacing of unit values the first variance component remains.

By conditioning on \( s_i \), and using \( \text{Var}(\hat{Y}_{CAL,lin}) = V_i(E_c) + E_i(V_c) \), we find \( \text{Var}(\hat{Y}_{CAL}) = \text{Var}(\hat{Y}_{CAL,lin}) \)

\[
\begin{align*}
    \text{Var}(\hat{Y}_{CAL}) &= \sum \sum F_{ij}^T e_i e_j + \sum a_i V_i \nosep
    e_j = y_j - \left(x_j^T \right) \left( B_2 \right)^{-1} \left( x_j \right)
    B &= \left( \begin{array}{c}
        B_1 \\
        B_2
    \end{array} \right) = \left( \sum x_k \left( x_k \right)^T \right) \left( \sum x_k y_k \right) \nosep
    &\quad \text{with} \\
    B = B^0 \quad \text{that minimizes (4.5), with a corresponding optimal } z_k = z_k^0. \text{ Some algebra shows that } z_k = a_i \sum j F_{ij}^T \left( x_j^T \right) \left( x_j \right) + \sum j F_{ij}^T \left( x_j \right) \right) \text{ for } k \text{ in cluster } i. \text{ It is no surprise that } z_k^0 \text{ depends on the sampling design at both stages. Future work will examine when } z_k = z_k^0 \text{ is a valid instrument and whether } z_k = z_k^0 \text{ gives any appreciable variance advantage over the simple standard } z_k = x_k. \text{ If this advantage is minimal, the preferred choice in practice is the simple } z_k = x_k. \\

\begin{align*}
    e_j &= y_j - \left( x_j^T \right) \left( B_2 \right) \quad \text{with} \\
    B_1 &= \left( \begin{array}{c}
        B_1^0 \\
        B_2^0
    \end{array} \right) \quad \text{and} \\
    B_2 &= \left( \begin{array}{c}
        B_1^0 \\
        B_2^0
    \end{array} \right) \quad \text{the residuals for unit and cluster statistics, are summarized in Table 2.} \\
    e_j &= y_j - \left( x_j^T \right) \left( B_2^0 \right) \quad \text{with} \\
    B_1 &= \left( \begin{array}{c}
        B_1^0 \\
        B_2^0
    \end{array} \right) \quad \text{and} \\
    B_2 &= \left( \begin{array}{c}
        B_1^0 \\
        B_2^0
    \end{array} \right) \quad \text{The residuals for unit and cluster statistics, are summarized in Table 2.}
\end{align*}

Consider now case (b) for cluster statistics. The calibrated estimator of the cluster total \( Y_i = \sum y_i \) is \( \hat{Y}_{CAL} = \sum w_i y_i \), with the cluster weights \( w_i = a_i (1 + \lambda_i) z_i^T \). We can write \( \hat{Y}_{CAL} \) as a sum of unit values \( Y_{CAL} = \sum w_i y_i \), if we define \( y_{ik} = y_i / N_i \) 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 \( y_i \) by \( y_{ik} \), keeping other quantities intact. Denote by \( B^{(c)} = \left( \begin{array}{c}
        B_1^{(c)} \\
        B_2^{(c)}
    \end{array} \right) \quad \text{the result of replacing } y_i \text{ by } y_{ik} \text{ in } B \) of (4.4). The approximation to \( \text{Var}(\hat{Y}_{CAL}) \) is then given by (4.5) with \( e_k = y_{ik} - \left( x_{ik}^T \right) \left( B_2^{(c)} \right) \quad \text{and} \\
\begin{align*}
    e_k &= y_{ik} - \left( x_{ik}^T \right) \left( B_2^{(c)} \right) \quad \text{the residuals for unit and cluster statistics, are summarized in Table 2.} \\
    B_1 &= \left( \begin{array}{c}
        B_1^{(c)} \\
        B_2^{(c)}
    \end{array} \right) \quad \text{and} \\
    B_2 &= \left( \begin{array}{c}
        B_1^{(c)} \\
        B_2^{(c)}
    \end{array} \right) \quad \text{The residuals for unit and cluster statistics, are summarized in Table 2.}
\end{align*}

Another issue of interest in case (b) is the choice of the instrument \( z_k \). The standard choice is to take \( z_k = x_k \) for \( k \) in cluster \( i \). But one can derive a
automated linearization of \( \hat{Y}_{CM} = \sum w_i y_i \) produces \( B = \left( \sum z_i x_i^T \right)^{-1} \left( \sum z_i y_i \right) \), and the residuals in Table 2 follow from case (b) for unit statistics by setting \( x_{i1} = 0 \) and \( x_{ik} = 0 \) 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 \( \hat{Y}_{CM} = \sum w_i y_i \) leads to \( B_i = \left( \sum z_i x_i^T \right)^{-1} \left( \sum z_i y_i \right) \), and the residuals follow from case (c) for cluster statistics by setting \( x_i = 0 \) and \( t_{xi} = 0 \) 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 \( e_{ij} \) are adjusted for both \( x_{ij} \) and \( t_{xij} \), but in (a) they are only adjusted for \( t_{xij} \). Thus (b) and (c) are better than (a) for the first component. The residual \( e_i \) is adjusted for both auxiliaries in (b), but not in (a) and (c), where it is only adjusted for \( x_{ik} \). 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 \( e_{ik} \) is adjusted for both \( x_{ik} \) and \( t_{wik} \), but in (a) it is only adjusted for \( x_{ik} \). By design, the residual \( e_{ik} \) is always zero in (a). A particularly unfavourable situation for the second variance component arises for case (c), where the residual is \( x_{i1}^T B_{12} \). 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 $y_i = x_i^T B + \epsilon_i$ where it is assumed that $E(\epsilon_i) = 0$ and $Var(\epsilon_i) = c_i \sigma^2$ with $c_i > 0$. 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 $z_i$ 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 $w_i$ 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


