Simultaneous Calibration Estimators for Two-Phase Samples

Stephen Ash
Bureau of the Census[1], 4700 Silver Hill Road, Room 3723-3, Washington D.C. 20233-8700

Keywords: Auxiliary information; Regression Estimator.

1. Introduction

This paper discusses the use of auxiliary information with estimators of two-phase samples. More specifically, we discuss calibration estimators for two-phase sample designs. The goal of calibration estimators is to use the available auxiliary information to create weights that improve upon the original design based weights. Here the improvement is with respect to design mean squared error. Calibration weights do this because their solution aims to make the sample weighted totals consistent with the universe totals of the auxiliary information.

Hidirouglou and Särndal (1998) suggested a two-phase calibration estimator that calculates the calibration weights in two steps. Estevao and Särndal (2003) expanded the notion of calibrating two-phase samples using two steps. They considered ten different possible calibration estimators for a two-phase sample design. For each of these estimators they suggested a two-step solution. Dupont (1995) also suggested several two-phase regression estimators, each with two steps.

This paper considers using a one-step calibration solution instead of the two-step solution. Within this one step we calculate the two calibration weights – one for each phase – with two different sets of auxiliary information. In this way both sets of auxiliary information are made to work together at once.

Because we solve for the two weights at once, both weights turn-out to be functions of both sets of auxiliary information. We call these estimators Simultaneous Calibration Estimators (SCE). These estimators differ from the estimators based on the two-step approach in that we put both steps that would be minimized separately into one minimization. To illustrate the general idea of a SCE we consider three cases where the following types of auxiliary information are available:

- ▹ first phase and overall
- ▹ second phase and overall
- ▹ first and second phase.

Here overall refers to auxiliary information that we can use across both the first and second phase samples.

For all three cases the calibration constraints we use also define estimators similar to QR-estimators (Wright 1983). We make this connection in a manner analogous to Deville and Särndal (1992). We use the general QR-estimator version of the SCE to examine two special cases: the ratio and regression estimator. Lastly we consider variance estimators for all three of the SCEs.

For our discussion we will only consider the Generalized Least Squares distance function. See Deville and Särndal (1992), Huang and Fuller (1978), and Singh and Mohl (1996) for alternative distance functions that can be applied to calibrating sample weights. For all of the estimators in this paper we do not consider the problems of being overconstrainted which may produce negative weights or whether the inverses exist. For our discussion we assume that all of the matrices for which inverses are required are nonsingular and that negative weights are not produced.

1.1. Notation

Let the universe and sample of the population be denoted as U and s, respectively. For the first phase, we index the units as i and sometimes refer to them as Primary Sampling Units (PSUs). Let the universe and sample of the first phase sample units be denoted as U_1 and s_1, respectively. Likewise let the second phase universe and samples within a given PSU i be denoted as U_{2i} and s_{2i}, respectively. We index the units of the second phase as k.

Let d_i and d_k represent the first and second phase design weights $d_k = P(k \in s_1)^{-1}$ and $d_{k/i} = p_{k/i}^{-1} = P\left(k \in s_{2i}\right)^{-1}$. More simply, we say the design weights are equal to the inverse of the selection probabilities. We write k instead of k / i to simplify the notation, although it is always assumed. Similarly we define w_i and w_k as the first and second phase calibration weights, respectively.

Let y_k be the variable of interest and $\mathbf{x}_k$ be a vector of known auxiliary variables. The total of the variable of interest is $T_y = \Sigma_{U_1} \Sigma_{U_{2i}} y_k$ and can be estimated with

[1] This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a Census Bureau review more limited in scope than that given to official Census Bureau publications. This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress.

the design weights as $\hat{T}_y = \Sigma_{s_1} \Sigma_{s_{2i}} d_i d_k y_k$. Using the auxiliary variable $\mathbf{x}_k$ note that we will sometimes express the total of PSU i as $\mathbf{x}_i = \Sigma_{U_{2i}} \mathbf{x}_k$ and its estimator as as $\hat{\mathbf{x}}_i = \Sigma_{s_{2i}} d_k \mathbf{x}_k$. For estimators of a PSU total involving two or more terms, we will use a $\hat{T}$ with an additional i subscript, e.g., $\hat{\mathbf{T}}_{\mathbf{x}yq,i} = \Sigma_{s_{2i}} d_k \mathbf{x}_k y_k q_k$. We use a tilda (~) on top of the T to denote estimators of the first phase sample units only, e.g., $\tilde{\mathbf{T}}_{\hat{\mathbf{x}}\mathbf{z}} = \Sigma_{s_1} d_i \hat{\mathbf{x}}_i \mathbf{z}_i$.

1.2. Review of Calibration Weights for a Uni-stage Sample Design

The idea of calibration weights is that we have some auxiliary information that we know for all units in the universe. It would then be desirable to have weights whose sample estimate of the total for the auxiliary information is exactly equivalent to the universe total, i.e., for a uni-stage sample design we want $\sum_s w_k \mathbf{x}_k = \sum_U \mathbf{x}_k$. We narrow the choice of w_k by necessitating that the weights be not much different than the original design weights. Calibration weights are then the solution that minimizes some distance between w_k and d_k subject to the constraint that the weighted totals equal the unweighted totals for the auxiliary information. We build the constraint of equal sample and universe totals into the minimization using Lagrange multipliers.

For all of the estimators we consider, we express our calibration equations more generally by including a term q_k that is consistent with QR-estimators (Wright 1983) and (Deville and Särndal 1992). We will apply a different q to each of the two sets of auxiliary information.

For background, note that the QR-estimator for an uni-stage sample design can be defined as

$$\hat{T}_{y,QR} = \mathbf{T_x}\hat{b} + \Sigma_s \left(y_k - \mathbf{x}_k \hat{b}\right) / r_k \qquad (1)$$

$$= \sum_s \left[1/r_k + \left(\sum_s \mathbf{x}_k / r_k - \sum_U \mathbf{x}_k\right)\hat{b}\right] y_k$$

$$= \sum_s w_k y_k$$

where $\hat{b} = \left(\Sigma_s q_k \mathbf{x}_k \mathbf{x}_k'\right)^{-1} \left(\Sigma_s q_k \mathbf{x}_k y_k\right)$.

The QR-estimator can be derived as a calibration estimator since $\hat{b}$, which in turn defines w_k, is the solution that minimizes $\sum_s \left(w_k - d_k\right)^2 / q_k$ subject to the constraint want $\sum_s w_k \mathbf{x}_k = \sum_U \mathbf{x}_k$. Here q_k is used to weight the term (w_k - d_k)^2 for each unit in (1).

Since we are interested in design based estimators, we assume that 1 / r_k = d_k and q_k always involves the product of d_k and some other quantity – in fact we write d_k separately from q_k , i.e., q'_k = q_k d_k. Including q_k will be useful when we discuss two special cases of each of the SCEs: the ratio and regression estimator.

1.3. Auxiliary Variables for Two-phase Samples

For our discussion we say that we can have three types of known auxiliary variables. We define them by the set of units for which the auxiliary variable is known. The three types are:

(a) $\mathbf{x}_k$ for all $k \in U$
(b) $\mathbf{x}_i$ for all $i \in U_1$
(c) $\mathbf{x}_k$ for all $k \in U_{2i}$

Case (a) is what we have referred to as the overall. Here we can sum the variables over the entire universe and the first and second phases samples. Case (b) is when we only have the first phase auxiliary information available. For case (c) we only have the auxiliary information for the units of the second phase sample. This may arise when there's a wealth of within PSU data available but obtaining it is prohibitive because of cost or effort. So instead of obtaining the auxiliary information for all PSUs, we only obtain it for the units of the second phase sample.

When we have two different sets of auxiliary information, we will use $\mathbf{x}_k$ and $\mathbf{z}_i$ to explicitly denote each. The auxiliary variables of $\mathbf{x}_k$ and $\mathbf{z}_i$ are allowed to have different dimensions and also different variables. Note that if a given variable is used for both the first phase and the overall, we include the values in both vectors $\mathbf{x}_k$ and $\mathbf{z}_i$.

2. Simultaneous Calibration Estimators for Two-phase Samples

We now discuss SCEs that consider (or constrain) any combination of the two parts of the two-phase sample design. Proofs of all results are available upon request.

Proposition 2.1. A SCE for Calibrating Overall and First Phase Auxiliary Information

Let $\mathbf{x}_k$ be a vector of values that we know overall and $\mathbf{z}_i$ be a vector of values that we know for all units of the first phase. We suggest that the estimator with weights that minimize f_1 of Table 1 also calibrate the totals for the overall and first phase simultaneously, i.e.,

$$w_k^* = d_i d_k \times$$

$$\left[\begin{array}{l} 1+\left(\mathbf{T_x}-\hat{\mathbf{T}}_\mathbf{x}\right)' \hat{\mathbf{C}}_\mathbf{1}^{-1}\left(\left(\hat{\mathbf{x}}_i-\tilde{\mathbf{T}}_{\hat{\mathbf{x}}\mathbf{z}'q}\tilde{\mathbf{T}}_{\mathbf{zz}'q}^{-1}\mathbf{z}_i\right)q_i+\mathbf{x}_k q_k\right) \\ +\left(\mathbf{T_z}-\tilde{\mathbf{T}}_\mathbf{z}\right)'\tilde{\mathbf{T}}_{\mathbf{zz}'q}^{-1}\left[\begin{array}{l} \mathbf{z}_i q_i-\tilde{\mathbf{T}}_{\mathbf{z}\hat{\mathbf{x}}'q}\hat{\mathbf{C}}_\mathbf{1}^{-1}\times \\ \left(\left(\hat{\mathbf{x}}_i-\tilde{\mathbf{T}}_{\hat{\mathbf{x}}\mathbf{z}'q}\tilde{\mathbf{T}}_{\mathbf{zz}'q}^{-1}\mathbf{z}_i\right)q_i+\mathbf{x}_k q_k\right)\end{array}\right]\end{array}\right]$$

$$w_i = d_i \times$$

$$\left[\begin{array}{l} 1+\left(\mathbf{T_x}-\hat{\mathbf{T}}_\mathbf{x}\right)'\hat{\mathbf{C}}_\mathbf{1}^{-1}\left(\hat{\mathbf{x}}_i-\tilde{\mathbf{T}}_{\hat{\mathbf{x}}\mathbf{z}'q}\tilde{\mathbf{T}}_{\mathbf{zz}'q}^{-1}\mathbf{z}_i\right)q_i + \\ \left(\mathbf{T_z}-\tilde{\mathbf{T}}_\mathbf{z}\right)'\tilde{\mathbf{T}}_{\mathbf{zz}'q}^{-1}\left[\mathbf{z}_i-\tilde{\mathbf{T}}_{\mathbf{z}\hat{\mathbf{x}}'q}\hat{\mathbf{C}}_\mathbf{1}^{-1}\left(\hat{\mathbf{x}}_i-\tilde{\mathbf{T}}_{\hat{\mathbf{x}}\mathbf{z}'q}\tilde{\mathbf{T}}_{\mathbf{zz}'q}^{-1}\mathbf{z}_i\right)q_i\right]\end{array}\right]$$

where $\hat{\mathbf{C}}_\mathbf{1}=\left[\tilde{\mathbf{T}}_{\hat{\mathbf{x}}\hat{\mathbf{x}}'q}+\hat{\mathbf{T}}_{\mathbf{xx}'q}-\tilde{\mathbf{T}}_{\hat{\mathbf{x}}\mathbf{z}'q}\tilde{\mathbf{T}}_{\mathbf{zz}'q}^{-1}\tilde{\mathbf{T}}_{\mathbf{z}\hat{\mathbf{x}}'q}\right]^{-1}$.

Here w_i is the weight for the first phase and w_k^* is the overall weight for unit k – representing both the first and second phase, i.e., $w_k^* = w_i\, w_k$. In Table 1, l and g are vectors of Lagrange multipliers

The third and fourth terms of f_1 formalize that we want the sample estimates to equal the universe totals for each of the sets of auxiliary information. The second term requires that the first phase weight be close to the original first phase design weight. Although the first term is not strictly conditional on the first phase, it requires that the overall weight be close to the product of the first phase weight we are minimizing in the second term and the original design weight of the second phase.

We also see that both weights are functions of both sets of auxiliary information. This is different than the Hidirouglou and Särndal (1998) weights for which the first step weights are only a function of the first set of auxiliary information that is calibrated.

Proposition 2.2. A SCE for Calibrating Overall and Second Phase Auxiliary Information

Let $\mathbf{x}_k$ be a vector of values that we know for the overall and $\mathbf{z}_k$ be a vector of values that we know for all units of the second phase. We suggest that the estimator with weights that minimize f_2 of Table 1 also calibrate the totals for the overall and second phase simultaneously, i.e.,

$$w_k^* = d_i d_k \times$$

$$\left[\begin{array}{l} 1+\left(\mathbf{T_x}-\hat{\mathbf{T}}_\mathbf{x}\right)'\mathbf{C}_2^{-1}\left(\mathbf{x}_k q_k^*+\mathbf{x}_k q_k-\hat{\mathbf{T}}_{\mathbf{xz}'q,i}\hat{\mathbf{T}}_{\mathbf{zz}'q,i}^{-1}\mathbf{z}_k\right) \\ +\left(\mathbf{z}_i-\hat{\mathbf{z}}_i\right)'\hat{\mathbf{T}}_{\mathbf{zz}',i}^{-1}\mathbf{z}_k \\ -\left[\sum_{s_1} d_{i^*}\left(\mathbf{z}_{i^*}-\hat{\mathbf{z}}_{i^*}\right)'\hat{\mathbf{T}}_{\mathbf{zz}',i^*}^{-1}\hat{\mathbf{T}}_{\mathbf{zx}',i^*}\right]\times \\ \mathbf{C}_2^{-1}\left(\mathbf{x}_k q_k^*+\mathbf{x}_k q_k-\hat{\mathbf{T}}_{\mathbf{xz}'q,i}\hat{\mathbf{T}}_{\mathbf{zz}'q,i}^{-1}\mathbf{z}_k\right)\end{array}\right]$$

Table 1: Constraints for Simultaneous Calibration Estimators

Phases of Sample Design Calibrated	Constraint of the Simultaneous Calibration Estimators
Overall and First	$f_1=\frac{1}{2}\Sigma_{s_1}\Sigma_{s_{2i}}\left[\left(w_k^*-w_i d_k\right)^2/d_i d_k q_k\right]+\frac{1}{2}\Sigma_{s_1}\left[\left(w_i-d_i\right)^2/d_i q_i\right]$ $-\left(\Sigma_{s_1}\Sigma_{s_{2i}}w_k^*\mathbf{x}_k-\Sigma_{U_1}\Sigma_{U_{2i}}\mathbf{x}_k\right)' l-\left(\Sigma_{s_1}w_i\mathbf{z}_i-\Sigma_{U_1}\mathbf{z}_i\right)' g$
Overall and Second	$f_2=\frac{1}{2}\Sigma_{s_1}\Sigma_{s_{2i}}\left[\left(w_k^*-d_i w_k\right)^2/d_i d_k q_k^*\right]+\frac{1}{2}\Sigma_{s_1}\left[d_i\Sigma_{s_{2i}}\left[\left(w_k-d_k\right)^2/d_k q_k\right]\right]$ $-\left(\Sigma_{s_1}\Sigma_{s_{2i}}w_k^*\mathbf{x}_k-\Sigma_{U_1}\Sigma_{U_{2i}}\mathbf{x}_k-\right)' l-\Sigma_{s_1}\left[d_i\left(\Sigma_{s_{2i}}w_k\mathbf{z}_k-\mathbf{z}_i\right)' g_i\right]$
First and Second	$f_3=\frac{1}{2}\Sigma_{s_1}\left[\left(w_i-d_i\right)^2/d_i q_i\right]+\frac{1}{2}\Sigma_{s_1}\Sigma_{s_{2i}}\left[\left(w_k^*-w_i d_k\right)^2/d_i d_k q_k\right]$ $-\left(\Sigma_{s_1}w_i\mathbf{x}_i-\Sigma_{U_1}\mathbf{x}_i\right)' l-\left(\Sigma_{s_1}\Sigma_{s_{2i}}w_k^*\mathbf{z}_k-\Sigma_{s_1}w_i\Sigma_{U_{2i}}\mathbf{z}_k\right)' g$

$$w_k = d_k \times \left[\begin{array}{l} 1+\left(\mathbf{T}_\mathbf{x}-\hat{\mathbf{T}}_\mathbf{x}\right)'\hat{\mathbf{C}}_2^{-1}\left(\mathbf{x}_k-\hat{\mathbf{T}}_{\mathbf{xz}'q,i}\hat{\mathbf{T}}^{-1}_{\mathbf{zz}'q,i}\mathbf{z}_k\right)q_k \\ +\left(\mathbf{z}_i-\hat{\mathbf{z}}_i\right)'\hat{\mathbf{T}}^{-1}_{\mathbf{zz}'q,i\mathbf{i}}\mathbf{z}_k q_k \\ -\left[\sum_{s_1} d_{i^*}\left(\mathbf{z}_{i^*}-\hat{\mathbf{z}}_{i^*}\right)\hat{\mathbf{T}}^{-1}_{\mathbf{zz}'q,i^*}\hat{\mathbf{T}}_{\mathbf{zx}'q,i^*}\right]\times \\ \hat{\mathbf{C}}_2^{-1}\left(\mathbf{x}_k-\hat{\mathbf{T}}_{\mathbf{xz}'q,i}\hat{\mathbf{T}}^{-1}_{\mathbf{zz}'q,i}\mathbf{z}_k\right)q_k \end{array}\right]$$

where

$$\hat{\mathbf{C}}_2 = \left[\left(\hat{\mathbf{T}}_{\mathbf{xx}'q^*}+\hat{\mathbf{T}}_{\mathbf{xx}'q}\right)-\Sigma_{s_1} d_i\hat{\mathbf{T}}_{\mathbf{xz}'q,i}\hat{\mathbf{T}}^{-\mathbf{1}}_{\mathbf{zz}'q,i}\hat{\mathbf{T}}_{\mathbf{zx}'q,i}\right].$$

The third and fourth terms formalize that we want weights that are consistent for the totals of the overall and the second phase auxiliary information, respectively. The second term requires that the second phase weights be close to the original second phase design weight. Here each PSU's difference is weighted by it's original first phase design weight. The first term requires that the overall weights be close to the product of the second phase weight that's minimized in the second term and the original first phase design weight. This first term links the minimization of the overall and second phase calibrations.

Here we again see how both sets of auxiliary information contribute to each of the weights. The third term of w_k^* is interesting since it adjusts the overall weight for a given unit according to the difference in the estimated and universe PSU totals for all PSUs.

Proposition 2.3. A SCE for Calibrating First and Second Phase Auxiliary Information

Let $\mathbf{x}_k$ be a vector of values that we know for all units of the first phase and $\mathbf{z}_k$ be a vector of values that we know for all of the units for the second phase. We suggest that the estimator with weights that minimize f_2 of Table 1 also calibrate the totals for the first and second phase simultaneously, i.e.,

$$w_k^* = d_i d_k \times \left[\begin{array}{l} 1+\left(\mathbf{T}_\mathbf{x}-\tilde{\mathbf{T}}_\mathbf{x}\right)'\tilde{\mathbf{T}}^{-\mathbf{1}}_{\mathbf{xx}'q}\times \\ \left[\mathbf{x}_i q_i-\tilde{\mathbf{T}}_{\mathbf{x}\breve{\mathbf{z}}'q}\hat{\mathbf{C}}_3^{-\mathbf{1}}\times\left(q_i\breve{\mathbf{z}}_i+q_k\mathbf{z}_k-q_i\tilde{\mathbf{T}}_{\breve{\mathbf{z}}\mathbf{x}'q}\tilde{\mathbf{T}}^{-\mathbf{1}}_{\mathbf{xx}'q}\mathbf{x}_i\right)\right] \\ +\left(\tilde{\mathbf{T}}_\mathbf{z}-\hat{\mathbf{T}}_\mathbf{z}\right)\hat{\mathbf{C}}_3^{-\mathbf{1}}\left(q_i\breve{\mathbf{z}}_i+q_k\mathbf{z}_k-q_i\tilde{\mathbf{T}}_{\breve{\mathbf{z}}\mathbf{x}'q}\tilde{\mathbf{T}}^{-\mathbf{1}}_{\mathbf{xx}'q}\mathbf{x}_i\right) \end{array}\right]$$

$$w_i = d_i \times \left[\begin{array}{l} 1+\left(\mathbf{T}_\mathbf{x}-\hat{\mathbf{T}}_\mathbf{x}\right)'\times \\ \left(\tilde{\mathbf{T}}^{-1}_{\mathbf{xx}'q}\mathbf{x}_i-\tilde{\mathbf{T}}^{-1}_{\mathbf{xx}'q}\tilde{\mathbf{T}}_{\mathbf{x}\breve{\mathbf{z}}q}\mathbf{C}_3^{-1}\left(\breve{\mathbf{z}}_i-\tilde{\mathbf{T}}_{\breve{\mathbf{z}}\mathbf{x}'q}\tilde{\mathbf{T}}^{-1}_{\mathbf{xx}'q}\mathbf{x}_i\right)\right)q_i \\ +\left(\tilde{\mathbf{T}}_\mathbf{z}-\hat{\mathbf{T}}_\mathbf{z}\right)'\mathbf{C}_3^{-1}\left(\breve{\mathbf{z}}_i-\tilde{\mathbf{T}}_{\breve{\mathbf{z}}\mathbf{x}'q}\tilde{\mathbf{T}}^{-1}_{\mathbf{xx}'q}\mathbf{x}_i\right)q_i \end{array}\right]$$

where $\breve{\mathbf{z}}_i=\hat{\mathbf{z}}_i-\mathbf{z}_i$ and

$$\hat{\mathbf{C}}_3 = \left[\hat{\mathbf{T}}_{\mathbf{zz}'q}-\tilde{\mathbf{T}}_{\breve{\mathbf{z}}\breve{\mathbf{z}}'q}-\tilde{\mathbf{T}}_{\breve{\mathbf{z}}\mathbf{x}'q}\tilde{\mathbf{T}}^{-\mathbf{1}}_{\mathbf{xx}'q}\tilde{\mathbf{T}}_{\mathbf{x}\breve{\mathbf{z}}'q}\right].$$

3. Calibration Estimators as Regression Estimators

Deville and Särndal (1992) showed how regression estimators can be derived from calibration estimators for uni-stage sample designs. We can similarly derive QR estimators from calibration estimators. We now interpret our calibration estimators as QR-estimators. We get the regression estimator when we let the *q*-term equal 1.0. The analogous regression estimator for all of our SCEs then is the same as in Results 3.1, 3.3, and 3.4, respectively, where the *q*-term is simply dropped from the expressions for **B**.

Result 3.1. The Overall and First Phase Sample SCE as a Regression Estimator

The regression estimator associated with f_1 can be expressed as

$$\begin{aligned}\hat{T}_{y,f_1} &= \Sigma_{s_1}\Sigma_{s_{2i}} w_k^* y_k \\ &= \hat{T}_y+\left(\mathbf{T}_\mathbf{x}-\hat{\mathbf{T}}_\mathbf{x}\right)'\hat{\mathbf{B}}_\mathbf{x}+\left(\mathbf{T}_\mathbf{z}-\tilde{\mathbf{T}}_\mathbf{z}\right)'\hat{\mathbf{B}}_\mathbf{z}\end{aligned} \qquad (1)$$

where

$$\hat{\mathbf{B}}_\mathbf{x} = \hat{\mathbf{C}}_\mathbf{1}^{-\mathbf{1}}\left(\tilde{\mathbf{T}}_{\hat{\mathbf{x}}\hat{y}q}+\hat{\mathbf{T}}_{\mathbf{x}yq}-\tilde{\mathbf{T}}_{\hat{\mathbf{x}}\mathbf{z}'q}\tilde{\mathbf{T}}^{-\mathbf{1}}_{\mathbf{zz}'q}\tilde{\mathbf{T}}_{\mathbf{z}\hat{y}q}\right) \qquad (2)$$

$$\hat{\mathbf{B}}_\mathbf{z} = \tilde{\mathbf{T}}^{-\mathbf{1}}_{\mathbf{zz}'q}\left[\begin{array}{l}\tilde{\mathbf{T}}_{\mathbf{z}\hat{y}q}-\tilde{\mathbf{T}}_{\mathbf{z}\hat{\mathbf{x}}'q}\hat{\mathbf{C}}_\mathbf{1}^{-\mathbf{1}}\times \\ \left(\tilde{\mathbf{T}}_{\hat{\mathbf{x}}\hat{y}q}-\tilde{\mathbf{T}}_{\hat{\mathbf{x}}\mathbf{z}'q}\tilde{\mathbf{T}}^{-\mathbf{1}}_{\mathbf{zz}'q}\tilde{\mathbf{T}}_{\mathbf{z}\hat{y}q}-\hat{\mathbf{T}}_{\mathbf{x}yq}\right)\end{array}\right]. \qquad (3)$$

Note that Result 3.1 follows directly from Proposition 2.1. Similarly we will see that Result 3.3 and 3.5 follow directly from Propositions 2.2 and 2.3, respectively.

We say that (2) and (3) are regression estimators because they have a form that is similar to the uni-stage regression estimator, i.e., $\hat{\mathbf{B}}=\hat{\mathbf{T}}^{-\mathbf{1}}_{\mathbf{xx}'}\hat{\mathbf{T}}_{\mathbf{x}y}$. In (2) we have $\hat{\mathbf{C}}_\mathbf{1}$ analogous to $\hat{\mathbf{T}}_{\mathbf{xx}'}$ and the expression in parentheses analogous to $\hat{\mathbf{T}}_{\mathbf{x}y}$. We can similarly argue that (3) also has a form that is analogous to the solution of the uni-stage regression coefficient.

An important special case of the QR estimator is the ratio estimator. The ratio estimator arises when the auxiliary information is univariate and the q-term is equal to the inverse of the auxiliary information. The next result discusses the estimator $\hat{\mathbf{T}}_{y,f_1}$ under analogous uni-stage ratio estimator assumptions for the ratio estimator.

Result 3.2. The Overall and First Phase Sample SCE as a Ratio Estimator

Let $q_i = z_i^{-1}$, $q_k = x_k^{-1}$, $\mathbf{x}_k = x_k$, and $\mathbf{z}_k = z_k$, i.e., let x and z both be univariate. After some simplification we can express (1) as

$$\hat{T}_{y,f_1} = \hat{T}_y + \left(T_x - \hat{T}_x\right)\hat{C}_1^{-1}\left[\hat{T}_y + \left(\tilde{T}_{\hat{x}\hat{y}/z} - \hat{T}_x\hat{T}_y / \tilde{T}_z\right)\right]$$

$$+\left(T_z - \tilde{T}_z\right)\tilde{T}_z^{-1}\left[\hat{T}_y - \left(\hat{T}_x / \hat{C}_1\right)\left[\hat{T}_y + \left(\tilde{T}_{\hat{x}\hat{y}/z} - \hat{T}_x\hat{T}_y / \tilde{T}_z\right)\right]\right]$$

where $\hat{C}_1 = \left[\hat{T}_x + \left(\tilde{T}_{\hat{x}^2/z} - \tilde{T}_x^2 / \tilde{T}_z\right)\right]$.

If additionally the ratio and product estimators involving both sets of auxiliary information are equal, i.e., $\tilde{T}_{\hat{x}\hat{y}/z} = \hat{T}_x\hat{T}_y / \tilde{T}_z$ and $\tilde{T}_{\hat{x}^2/z} = \hat{T}_x^2 / \tilde{T}_z$, then

$$\hat{T}_{y,f_1} = \left(T_x / \hat{T}_x\right)\hat{T}_y .$$

We therefore conclude that if there is no difference in ratio versus the product estimates, then the estimator reduces to a simple ratio estimator, only adjustmenting for the overall auxiliary information.

Result 3.3. The Overall and Second Phase Sample SCE as a Regression Estimator

The regression estimator associated with f_2 can be expressed as

$$\hat{T}_{y,f_2} = \Sigma_{s_1}\Sigma_{s_{2i}} w_k^* y_k$$

$$= \hat{T}_y + \left(\mathbf{T}_\mathbf{x} - \hat{\mathbf{T}}_\mathbf{x}\right)'\hat{\mathbf{B}}_\mathbf{x} + \Sigma_{s_1} d_i\left(\mathbf{z}_i - \hat{\mathbf{z}}_i\right)'\hat{\mathbf{B}}_{\mathbf{z}\mathbf{1},i}$$

$$- \Sigma_{s_1} d_i\left(\mathbf{z}_i - \hat{\mathbf{z}}_i\right)'\hat{\mathbf{B}}_{\mathbf{z2,i}}$$

where

$$\hat{\mathbf{B}}_\mathbf{x} = \hat{\mathbf{C}}_2^{-\mathbf{1}}\left[\hat{\mathbf{T}}_{\mathbf{x}yq} + \hat{\mathbf{T}}_{\mathbf{x}yq^*} - \Sigma_{s_1} d_i \hat{\mathbf{T}}_{\mathbf{xz}'q,i}\hat{\mathbf{T}}_{\mathbf{zz}'q,i}^{-\mathbf{1}}\hat{\mathbf{T}}_{\mathbf{z}yq,i}\right]$$

$$\hat{\mathbf{B}}_{\mathbf{z}1,i} = \hat{\mathbf{T}}_{\mathbf{zz}'q,i}^{-\mathbf{1}}\hat{\mathbf{T}}_{\mathbf{z}yq,i}$$

$$\hat{\mathbf{B}}_{\mathbf{z}2,i} = \hat{\mathbf{T}}_{\mathbf{zz}'q,i}^{-\mathbf{1}}\hat{\mathbf{T}}_{\mathbf{zx}'q,i}\hat{\mathbf{C}}_2^{-\mathbf{1}} \times$$

$$\left[\hat{\mathbf{T}}_{\mathbf{x}yq} + \hat{\mathbf{T}}_{\mathbf{x}yq^*} - \Sigma_{s_1} d_i \hat{\mathbf{T}}_{\mathbf{xz}'q,i^*}\hat{\mathbf{T}}_{\mathbf{zz}'q,i^*}^{-\mathbf{1}}\hat{\mathbf{T}}_{\mathbf{z}yq,i^*}\right]$$

Here i^* indexes the summation over the first phase sample for the third term of both (1) and (2).

Result 3.4. The First and Second Phase Sample SCE as a Regression Estimator

The regression estimator associated with the SCE associated with f_3, can be expressed as

$$\hat{T}_{y,f_3} = \Sigma_{s_1}\Sigma_{s_{2i}} w_k^* y_k$$

$$= \hat{T}_y + \left(\mathbf{T}_\mathbf{x} - \hat{\mathbf{T}}_\mathbf{x}\right)'\hat{\mathbf{B}}_\mathbf{x} + \left(\tilde{\mathbf{T}}_\mathbf{z} - \hat{\mathbf{T}}_\mathbf{z}\right)'\hat{\mathbf{B}}_\mathbf{z} \qquad (4)$$

where

$$\hat{\mathbf{B}}_\mathbf{z} = \mathbf{C}_3^{-\mathbf{1}}\left(\tilde{\mathbf{T}}_{\breve{\mathbf{z}}\hat{y}q} + \hat{\mathbf{T}}_{\mathbf{z}yq} - \tilde{\mathbf{T}}_{\breve{\mathbf{z}}\mathbf{x}'q}\tilde{\mathbf{T}}_{\mathbf{xx}'q}^{-\mathbf{1}}\tilde{\mathbf{T}}_{\mathbf{x}\hat{y}q}\right)$$

$$\hat{\mathbf{B}}_\mathbf{x} = \tilde{\mathbf{T}}_{\mathbf{xx}'q}^{-\mathbf{1}}\left[\tilde{\mathbf{T}}_{\mathbf{x}\hat{y}q} - \tilde{\mathbf{T}}_{\mathbf{x}\breve{\mathbf{z}}'q}\mathbf{C}_3^{-\mathbf{1}}\left(\tilde{\mathbf{T}}_{\breve{\mathbf{z}}\hat{y}q} + \hat{\mathbf{T}}_{\mathbf{z}yq} - \tilde{\mathbf{T}}_{\breve{\mathbf{z}}\mathbf{x}'q}\tilde{\mathbf{T}}_{\mathbf{xx}'q}^{-\mathbf{1}}\tilde{\mathbf{T}}_{\mathbf{x}\hat{y}q}\right)\right]$$

Result 3.5. The First and Second Phase Sample SCE as a Ratio Estimator

Let $q_k = x_k^{-1}$, $q_k = z_k^{-1}$, $\mathbf{x}_k = x_k$, and $\mathbf{z}_k = z_k$, i.e., let x and z be univariate. After some simplification we can express (4) as

$$\hat{T}_{y,f3} = \hat{T}_y + \left(T_x - \tilde{T}_x\right)\frac{1}{\tilde{T}_x}\left[\hat{T}_y - \frac{\tilde{T}_{\breve{z}}}{C_3}\left[\hat{T}_y + \tilde{T}_{\frac{\breve{z}\hat{y}}{x}} - \frac{\tilde{T}_{\breve{z}}\hat{T}_y}{\tilde{T}_x}\right]\right]$$

$$+\left(\tilde{T}_z - \hat{T}_z\right)\frac{1}{C_3}\left[\hat{T}_y + \tilde{T}_{\frac{\breve{z}\hat{y}}{x}} - \frac{\tilde{T}_{\breve{z}}\hat{T}_y}{\tilde{T}_x}\right]$$

where $\hat{C}_3 = \left[\hat{T}_z - \tilde{T}_{\breve{z}^2/x} + \tilde{T}_{\breve{z}}^2 / \tilde{T}_x\right]$.

If additionally the ratio and product estimators involving both sets of auxiliary information are equal, i.e., $\tilde{T}_{\hat{z}y/x} = \left(\tilde{T}_{\breve{z}}\hat{T}_y\right) / \tilde{T}_x$ and $\tilde{T}_{\breve{z}^2/x} = \tilde{T}_{\breve{z}}^2 / \tilde{T}_x$, we see that $\hat{T}_{y,f_3} = \hat{T}_y\left(T_x / \tilde{T}_x\right)\left(\tilde{T}_z / \hat{T}_z\right)$.

We therefore conclude that if there is no difference in ratio versus the product estimates, the estimator reduces to a simple ratio estimator with two adjustments – one for the first phase and another for the second phase.

4. Variance Estimation

To estimate the design variances we suggest using the two-phase variance estimator for regression estimators as suggested by Hidirouglou and Särndal (1998), i.e.,

$$\hat{v}\left(\hat{T}_{y,reg}\right) = \Sigma\Sigma_{s_1}\left(1 - p_{1i}p_{1j}p_{1ij}^{-1}\right)w_{1i}\hat{e}_{1i}w_{1j}\hat{e}_{1j}$$

$$+ \Sigma_{s_1} p_{1i}^{-1}\left[\Sigma\Sigma_{s_{2i}}\left(1 - p_{2k}p_{2\ell}p_{2k\ell}^{-1}\right)w_{2k}\hat{e}_{2k}w_{2\ell}\hat{e}_{2\ell}\right]$$

Here p_{1ij} denotes the first phase second order selection probability, i.e., $p_{1ij} = P\left(i, j \in s_1\right)$. Similarly $p_{2k\ell}$ denotes the second phase second order selection probability, i.e., $p_{2k\ell} = P\left(k, \ell \in s_{2i}\right)$.

The w_{1i} and w_{1j} are the first phase weights and the w_{2k} and $w_{2\ell}$ are the second phase weights as defined in §2. The $\hat{e}_{1i}$ and $\hat{e}_{1j}$ are the first phase residuals and the $\hat{e}_{2k}$ and $\hat{e}_{2\ell}$ are the second phase residuals.

For the estimator associated with f_1 we suggest defining the residuals as

$$\hat{e}_{1i} = \hat{y}_i - \left(\hat{\mathbf{x}}'_i \hat{\mathbf{B}}_{\mathbf{x}} + \mathbf{z}'_i \hat{\mathbf{B}}_{\mathbf{z},i}\right) \tag{5}$$

$$\hat{e}_{2k} = y_k - \left(\mathbf{x}'_k \hat{\mathbf{B}}_{\mathbf{x}} + \mathbf{z}'_k \hat{\mathbf{B}}_{\mathbf{z},i}\right) \tag{6}$$

and for the estimator associated with f_2 we suggest defining the residuals as

$$\hat{e}_{1i} = \hat{y}_i - \left(\hat{\mathbf{x}}'_i \hat{\mathbf{B}}_{\mathbf{x}} + \Sigma_{s_1} d_{i^*} \mathbf{z}'_{i^*} \hat{\mathbf{B}}_{\mathbf{z}1,i^*} - \Sigma_{s_1} d_{i^*} \mathbf{z}'_{i^*} \hat{\mathbf{B}}_{\mathbf{z2},i^*}\right) \tag{7}$$

$$\hat{e}_{2k} = y_k - \left(\mathbf{x}'_k \hat{\mathbf{B}}_{\mathbf{x}} + \Sigma_{s_1} d_{i^*} \mathbf{z}'_{i^*} \hat{\mathbf{B}}_{\mathbf{z}1,i^*} - \Sigma_{s_1} d_{i^*} \mathbf{z}'_{i^*} \hat{\mathbf{B}}_{\mathbf{z2},i^*}\right) \tag{8}$$

For the estimator associated with f_3 we suggest defining the residuals as

$$\hat{e}_{1i} = \hat{y}_i - \left(\hat{\mathbf{x}}'_i \hat{\mathbf{B}}_{\mathbf{x}} + \hat{\mathbf{z}}'_i \hat{\mathbf{B}}_{\mathbf{z}}\right) \tag{9}$$

$$\hat{e}_{2k} = y_k - \left(\mathbf{x}'_i \hat{\mathbf{B}}_{\mathbf{x}} + \mathbf{z}'_k \hat{\mathbf{B}}_{\mathbf{z}}\right) \tag{10}$$

We define the regression coefficients of (5)-(10) as we did in §3.

A problem with the variance of the regression estimator associated with f_2 is that we don't have an expression for w_i. We only have w_k^* and w_k, where we assume that w_i is implicitly included for w_k^*. An unsatisfying solution to this problem is to use w_i^+ and w_k^+ calculated from w_k^* and w_k to estimate variances where w_i^+ = median{ w_k^* / w_k } and $w_k^+ = w_k^* / w_i^+$.

6. Summary / Discussion

We think it is important that the solution makes both sets of auxiliary information work together. We see this because both sets of weights are functions of both sets of auxiliary information. This is in contrast to the two-step approach where the first weight only accounts for one of the sets of auxiliary information and the second weight accounts for the second set of auxiliary information, conditional on the first weight having already been calculated.

Although we believe that the SCE will make better use of both sets of auxiliary data than the two-step approach, at this time we do not know the conditions under which this would be true. It is a difficult question to answer because several conditions impact both of the estimators including:

- ▹ the associations between the variable of interest y_k and each of the auxiliary variables $\mathbf{x}_k$ and $\mathbf{z}_k$
- ▹ the association between $\mathbf{x}_k$ and $\mathbf{z}_k$
- ▹ the sample designs for each of the two phases.

More research is needed to clarify this point.

References

Deville, J.-C. and Särndal, C.-E. (1992) “Calibration Estimators in Survey Sampling,” Journal of the American Statistical Association, 87, 376-382.

Dupont, F. (1995).“Alternative Adjustments Where There Are Several Levels of Auxiliary Information,” Survey Methodology, 24, 11-20.

Estevao, V.M., and Särndal, C.-E. (2003). “The Ten Cases of Auxiliary Informtion for Calibration in Two-Phase Sampling,” Journal of Official Statistics, 18, 233-255.

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

Hidroglou, M.A. and Särndal, C.-E. (1998). “Use of Auxiliary Information for Two-Phase Sampling,” Survey Methodology, 24, 11-20.

Singh, A.C. and Mohl, C.A. (1996). “Understanding calibration estimators in survey sampling,” Survey Methodology, 22, 107-115.

Wright, R.L. (1983). “Finite Population Sampling With Multivariate Auxiliary Information,” Journal of the American Statistical Association, 78, 879-884.