Journal of Official Statistics, Vol. 16, No. 4, 2000, pp. 379–399

A Functional Form Approach to Calibration

Victor M. Estevao and Carl-Erik Särndal[1]

Calibration has become a widely used procedure for estimation in sample surveys. It uses *auxiliary information* to produce efficient estimates. Calibration requires that we know population totals (*control totals*) for one or more auxiliary variables (*x-variables*). The efficiency of the *calibration estimator* depends on how well the auxiliary variables explain the variability of y, the variable of interest. Traditionally, a *distance minimization approach* is used for building calibration estimators. The distance measures that have been proposed produce estimators that are nearly identical, so this approach does not provide much insight into the properties of different calibration estimators. In this article, we note that distance minimization is not the only possible starting point for calibration. We define and develop an alternative, the *functional form approach*. The *calibrated weights* are given a simple mathematical form that depends on two *parameters*. This defines a family of calibration estimators denoted by $\hat{Y}_{\text{CALF}}$. It includes the family of generalized regression (GREG) estimators $\hat{Y}_{\text{GREG}}$. We discuss the role of the auxiliary variables in the calibration. To do this, we assume a linear relation between y and the x-variables. In most surveys, the x-variables in the calibration are not the only ones that explain y. In the unlikely event that they do (except for random noise), we say that the model is *saturated* by the calibration. This case is not generally of interest because the resulting estimators have similar properties. Usually, other x-variables are significant in explaining y but are excluded from the calibration because they are either not observed in the sample or their control totals are unknown. In this case, the model is called *unsaturated*. We look at the unsaturated model and show that for some sample designs, we can find a $\hat{Y}_{\text{GREG}}$ which has minimum Taylor variance among the estimators in $\hat{Y}_{\text{CALF}}$. The Monte Carlo simulations at the end of the article illustrate these results.

Key words: Calibration estimators; GREG estimator; calibrated weights; auxiliary information.

1. Introduction

We denote the finite survey population as $U = \{1, \ldots, k, \ldots, N\}$. Let y be the variable of interest with value y_k for unit k in the population. We want to estimate the population total $Y = \sum_U y_k$ using y_k for the observed units $k \in s$, where s is a sample drawn from U according to a given sampling design. We write $\sum_A y_k$ for the sum $\sum_{k \in A} y_k$ where $A \subseteq U$ is a set of units from the population. The sampling weights under this design are denoted by $a_k = 1/\pi_k$, where $\pi_k = P(k \in s)$ is the inclusion probability of unit k. Research and practice in recent years have shown that it is fruitful to view estimation in surveys as a problem in linear weighting. We want to replace a_k with more efficient weights w_k determined from the available auxiliary information. We refer to Alexander (1987), Lemaître

[1] Statistics Canada, Tunney's Pasture, Ottawa, Ontario, K1A 0T6, Canada. E-mail: estevao@statcan.ca
Acknowledgments: We gratefully acknowledge the constructive comments from three referees and an excellent summary by the Associate Editor that helped bring about important clarifications in the article.

and Dufour (1987), Bethlehem and Keller (1987), Zieschang (1990), and to further references in these articles. The book by Särndal, Swensson and Wretman (1992) uses the linear weighting approach. Methods for restricting the weights so as to avoid negative or large values have been proposed by Huang and Fuller (1978), Bankier (1992), Deville and Särndal (1992), and Deville, Särndal and Sautory (1993). The Generalized Estimation System (GES) developed by Statistics Canada uses a computationally efficient algorithm to restrict the weights. This is described by Estevao (1994). The principles behind GES are presented in Estevao, Hidiroglou and Särndal (1995). Here, we examine calibration estimators of Y. These are linear weighting estimators given by

$$\hat{Y}_{\mathrm{CAL}} = \sum_s w_k y_k \tag{1.1}$$

where $\{w_k : k \in s\}$ is a *calibrated weight system*, determined from the auxiliary information available for the survey. The auxiliary vector denoted by $\mathbf{x}$ has dimension J and its value for unit k is $\mathbf{x}_k = (x_{1k}, x_{2k}, \ldots, x_{Jk})'$. The population vector total $\mathbf{X} = \sum_U \mathbf{x}_k$ is known. By definition, a weight system is calibrated if it satisfies the *calibration equation*

$$\sum_s w_k \mathbf{x}_k = \sum_U \mathbf{x}_k = \mathbf{X} \tag{1.2}$$

A calibrated weight system is *consistent* in that it produces an exact estimate of the population total of each of the J auxiliary variables. The known components of $\mathbf{X}$, taken from a reliable source, are often called control totals. In some surveys, this source is exterior to the survey, for example, a census or one or more matched administrative registers. In other surveys, the totals are computed from the survey frame itself, by adding the auxiliary values of the units on the frame, as for example, when a size measure (number of employees, say) is available for each enterprise listed on a business survey frame. One of the simplest and most frequently used applications occurs when the information consists of known counts, $N_j, j = 1, \ldots, J$, for a set of mutually exclusive and exhaustive population groups. In this case, $\mathbf{X} = (N_1, N_2, \ldots, N_J)'$ and we define the $\mathbf{x}_k$ vector as $\mathbf{x}_k = (\delta_{1k}, \delta_{2k}, \ldots, \delta_{Jk})'$, where $\delta_{jk} = 1$ if k belongs to group j, and $\delta_{jk} = 0$ otherwise. If the groups are the strata, the calibration leads to the separate expansion estimator. If the groups are not the strata, we obtain the poststratified estimator.

Many weight systems $\{w_k : k \in s\}$ satisfy (1.2). Some lead to accurate estimators, others give inefficient or implausible estimators. Clearly, we must restrict our attention to those weight systems satisfying (1.2) that are in some sense ''reasonable.'' With this goal in mind, several authors have constructed calibrated weight systems by means of a distance minimization approach. This approach has limited value in examining the properties of different calibration estimators, since all of the proposed distance measures lead to nearly identical estimators. In this article, we focus instead on a functional form approach, which leads to a more meaningful examination of different calibration estimators.

Before introducing the functional form approach, we recall the main steps of the distance minimization approach. First, we define a measure of distance between a given set of initial weights and the required calibrated weights. Then we minimize this distance subject to the calibration constraint (1.2) to obtain a set of calibrated weights. Here we use the sampling weights $\{a_k : k \in s\}$ as initial weights. Many distance measures are possible.

The Generalized Least Squares distance defined by

$$(1/2)\sum_s c_k(w_k - a_k)^2/a_k = (1/2)\sum_s c_k a_k(w_k/a_k - 1)^2 \tag{1.3}$$

is particularly important, where $\{c_k : k \in s\}$ is a specified (but arbitrary) set of positive constants. Minimizing (1.3) subject to constraint (1.2) leads to the calibrated weights $w_k = w_{k,\mathrm{GREG}}$, where

$$w_{k,\mathrm{GREG}} = a_k\{1 + (\mathbf{X} - \hat{\mathbf{X}})'\mathbf{T}_s^{-1}\mathbf{x}_k/c_k\} \tag{1.4}$$

with $\mathbf{T}_s = \sum_s a_k\mathbf{x}_k\mathbf{x}_k'/c_k$ and $\hat{\mathbf{X}} = \sum_s a_k\mathbf{x}_k$ is the Horvitz-Thompson (HT) estimator of $\mathbf{X}$. In the following sections, HT estimators are indicated by a superimposed ''^'' and no subscript.

The resulting calibration estimator $\hat{Y}_{\mathrm{GREG}} = \sum_s w_{k,\mathrm{GREG}}\, y_k$, is known to be asymptotically design unbiased (ADU). It is called the generalized regression estimator because it can be derived by a regression argument and expressed by adding a regression adjustment term to the HT estimator of Y, $\hat{Y} = \sum_s a_k\, y_k$. That is, we can write $\hat{Y}_{\mathrm{GREG}}$ as

$$\hat{Y}_{\mathrm{GREG}} = \hat{Y} + (\mathbf{X} - \hat{\mathbf{X}})'\hat{\mathbf{B}} \tag{1.5}$$

where

$$\hat{\mathbf{B}} = \mathbf{T}_s^{-1}\sum_s a_k\mathbf{x}_k\, y_k/c_k \tag{1.6}$$

In the distance measure approach, the c_k values moderate the importance of the terms in (1.3). A unit with a large value of c_k will have a calibrated weight w_k close to the initial design weight a_k. The c_k are relative in the sense that we can multiply them all by a positive constant without changing the resulting w_k. In the functional form approach outlined in Section 3, the c_k values are also relative but they are parameters that we can specify arbitrarily to define a GREG estimator. This allows us to create a family of regression estimators $\hat{Y}_{\mathrm{GREG}}$ corresponding to different choices of the values c_k.

An alternative measure of closeness in the distance measure approach is the Raking Ratio distance

$$\sum_s c_k a_k\{(w_k/a_k)\ln(w_k/a_k) - w_k/a_k + 1\} \tag{1.7}$$

discussed by Zieschang (1990) and Deville and Särndal (1992). Minimizing this measure subject to (1.2) also gives a family of calibrated weight systems. In general, to determine the calibrated weights $\{w_k : k \in s\}$ we must specify not only a distance measure but also the quantities c_k and the auxiliary variables we want to use for calibration. The sampling design, the calibrated weights $\{w_k : k \in s\}$ and the variable of interest y implicitly determine the estimator $\hat{Y}_{\mathrm{CAL}} = \sum_s w_k\, y_k$ and its properties.

In addition to (1.3) and (1.7), a number of other distance measures have been proposed and studied in the distance minimization approach. Deville and Särndal (1992) examined the calibration estimators produced by a group of distance measures including (1.3) and (1.7). Among these, $\hat{Y}_{\mathrm{GREG}}$ is a point of reference because of its simple, closed form (1.5). The calibration estimators derived from all of these distance measures share the same asymptotic variance, that of $\hat{Y}_{\mathrm{GREG}}$. This follows from the properties of the distance

measures. For example, comparing (1.3) and (1.7), we note that when $u_k = w_k/a_k$ is close to 1, $u_k \ln u_k - u_k + 1 \doteq (1/2)(u_k - 1)^2$, which is why Raking Ratio distance and Generalized Least Squares distance give asymptotically equivalent calibration estimators. Empirical studies have shown that even for modest sample sizes, there are only small differences in the calibrated estimates derived from different distance measures; see Singh and Mohl (1996); Stukel, Hidiroglou and Särndal (1996). Therefore, the distance minimization approach is not very useful for examining the properties of alternative calibration estimators. In contrast, we discover some interesting differences between calibration estimators in the functional form approach introduced in the next section.

The following result gives a useful expression for the difference between $\hat{Y}_{\text{GREG}}$ and any calibration estimator $\hat{Y}_{\text{CAL}}$ with weights satisfying (1.2).

Result 1.1. For each sample s, we have the following relationship between $\hat{Y}_{\text{GREG}}$ and $\hat{Y}_{\text{CAL}}$.

$$\hat{Y}_{\text{CAL}} = \hat{Y}_{\text{GREG}} + Z_s \tag{1.8}$$

with

$$Z_s = \sum_s (w_k - a_k) e_k \tag{1.9}$$

where e_k is the regression residual

$$e_k = y_k - \mathbf{x}_k' \hat{\mathbf{B}} \tag{1.10}$$

and $\hat{\mathbf{B}}$ is given by (1.6). □

The result follows from the definitions of $\hat{Y}_{\text{CAL}}$ and $\hat{Y}_{\text{GREG}}$ given by (1.1), (1.5) and (1.6), and by the calibration equation (1.2). We know that $\hat{Y}_{\text{GREG}}$ is ADU, but $\hat{Y}_{\text{CAL}}$ does not necessarily have this property. In the rest of the article, we restrict the weights w_k so that $\hat{Y}_{\text{CAL}}$ is also ADU. From (1.8), we have $V(\hat{Y}_{\text{CAL}}) = V(\hat{Y}_{\text{GREG}}) + V(Z_s) + 2Cov(\hat{Y}_{\text{GREG}}, Z_s)$. In general, the covariance can be positive, negative or zero. We are interested in finding conditions when the covariance is zero so we can make a statement on the optimality of $\hat{Y}_{\text{GREG}}$. To do this, we introduce an important subset of $\hat{Y}_{\text{CAL}}$ denoted by the functional form estimator $\hat{Y}_{\text{CALF}}$. We show that the family of regression estimators $\hat{Y}_{\text{GREG}}$ is part of $\hat{Y}_{\text{CALF}}$ and that for specific designs we can use Result 5.2 (in Section 5) to find one $\hat{Y}_{\text{GREG}}$ with minimum Taylor variance in $\hat{Y}_{\text{CALF}}$.

The article is organized as follows. In Section 2, we review the properties we would like to see in a system of calibrated weights. One of these is closeness to the sampling design weights a_k. In Section 3, we formulate our functional form approach. The calibrated weights are dependent on the values of two parameters. This defines a family of calibration estimators denoted by $\hat{Y}_{\text{CALF}}$ within $\hat{Y}_{\text{CAL}}$. The $\hat{Y}_{\text{CALF}}$ are shown to be asymptotically design unbiased. We note that the GREG estimators $\hat{Y}_{\text{GREG}}$ are part of $\hat{Y}_{\text{CALF}}$. Thus, for a given set of auxiliary variables, we have three nested families of estimators, $\hat{Y}_{\text{GREG}} \subseteq \hat{Y}_{\text{CALF}} \subseteq \hat{Y}_{\text{CAL}}$. In Section 4, we compare $\hat{Y}_{\text{CALF}}$ with $\hat{Y}_{\text{GREG}}$. Section 5 shows that for specific designs, we can find c_k and $\mathbf{x}_k$ that lead to an optimal estimator $\hat{Y}_{\text{GREG}}$ in $\hat{Y}_{\text{CALF}}$. In Section 6, we look at the difference $(\hat{Y}_{\text{CALF}} - \hat{Y}_{\text{GREG}})$ under different regressions between y and the x-variables in the population. The distinction between saturated and unsaturated models is useful in this discussion. Finally, Section 7 presents the results of a Monte Carlo study that illustrate some aspects of the preceding theory.

2. Why Calibrated Weights?

Many survey statisticians would agree on the following objectives for a calibrated weight system:

1. *Consistency*. A weight system that satisfies (1.2) is appealing, because it reproduces exactly the known population total for each auxiliary variable.
2. *Closeness to the basic weights*. The basic weights $a_k = 1/\pi_k$ have an attractive property in that they yield design unbiased estimates. Therefore, any departure from these weights should be small, to preserve the design unbiasedness, at least approximately or asymptotically.
3. *Control on auxiliary variable totals*. The more auxiliary totals we use in the calibration, the ''better'' we expect the resulting weight system to be. This intuitive statement is supported by theory, as shown in Section 6. The variance of a calibrated estimator tends to decrease as more variables and their known totals are brought into the calibration.

A calibrated weight system satisfies (by definition) the consistency objective 1. If derived by distance minimization, as described above, it also satisfies the closeness objective 2. Since the dimension of the $\mathbf{x}_k$ vector is arbitrary, the calibration can accommodate any number of auxiliary variables to meet objective 3. The following example illustrates how these objectives come into play in a simple case.

Example 2.1. Consider a survey involving a single, positive auxiliary variable x, with the known population total $X = \sum_U x_k$. The calibration equation is $\sum_s w_k x_k = X$. By substituting $\mathbf{x}_k = x_k$ and $c_k = x_k$ in (1.5) and (1.6), we obtain the traditional ratio estimator $\hat{Y}_{\mathrm{RAT}} = \hat{Y}(X/\hat{X})$ where $\hat{X} = \sum_s a_k x_k$. In this case, the calibrated weights are $w_k = a_k(X/\hat{X})$ for all $k \in s$.

From this example, it is easy to construct other weight systems. One possibility is through the family of estimators,

$$w_k = XR_k \tag{2.1}$$

with $R_k = q_k z_k / \sum_s q_k z_k x_k$, where $z_k = x_k^{p-1}$ and $p \geq 0$ and $q_k > 0$ are arbitrary constants. All calibrated weight systems in this family respect consistency objective 1. If the closeness objective 2 is to be maintained, then only one weight system is admissible for all sample sizes. This is the one with $q_k = a_k$ and $p = 1$, which gives the ratio estimator. For other values of p, (2.1) produces a design biased estimator with a mean squared error (MSE) that may greatly exceed that of the (unbiased) HT estimator, $\hat{Y} = \sum_s a_k y_k$. Thus, the calibrated weight systems produced by (2.1) are not generally suitable in the design-based perspective. Let us consider instead the family of estimators

$$w_k = a_k + (X - \hat{X})R_k \tag{2.2}$$

This produces many weight systems depending on how we specify p and the constants q_k. All of these satisfy objectives 1 and 2 since $(X - \hat{X})$ approaches zero in design probability as the sample size increases. Now, if the population size is also known, it is generally preferable to use the calibration vector $\mathbf{x}_k = (1, x_k)'$, with known control total $(N, X)'$. This meets objective 3 of using as much auxiliary information as possible.

Remark 2.1. In Example 2.1 and throughout the article, we work under the randomization distribution – the distribution induced by the sampling design with its inclusion probabilities π_k. Thus the notation $O_p(\cdot)$, used for the order in probability, applies to the randomization distribution. We use the fact that a difference between an HT estimator and its expected value divided by N is $O_p(n^{-1/2})$ under standard conditions. For example, $N^{-1}(\hat{X} - X)$ is $O_p(n^{-1/2})$. A discussion of asymptotics is found in Isaki and Fuller (1981).

3. Calibration Estimators Based on a Functional Form Approach

We now define a functional form approach for constructing calibration estimators. As before, let $\mathbf{x}_k = (x_{1k}, x_{2k}, \ldots, x_{Jk})'$ denote the auxiliary vector for unit k. The auxiliary information consists of the vector total $\mathbf{X} = \sum_U \mathbf{x}_k$, composed of the J known totals $\sum_U x_{jk}$ for $j = 1, \ldots, J$. In addition, we define a positive constant q_k and a vector $\mathbf{z}_k = (z_{1k}, z_{2k}, \ldots, z_{Jk})'$ for every $k \in s$, such that (a) $\dim(\mathbf{z}_k) = J = \dim(\mathbf{x}_k)$, and (b) the $J \times J$ matrix $\sum_s q_k \mathbf{z}_k \mathbf{x}_k'$ is nonsingular.

We now generate calibrated weights $\{w_k : k \in s\}$ by imposing the functional form

$$w_k = a_k + q_k \lambda_s' \mathbf{z}_k \tag{3.1}$$

where the parameters q_k and $\mathbf{z}_k$ are chosen to satisfy (a) and (b), and the vector λ_s, which depends on s but not on k, is implicitly determined by the calibration constraint (1.2). This yields $w_k = w_{k,\text{CALF}}$, where

$$w_{k,\text{CALF}} = a_k + (\mathbf{X} - \hat{\mathbf{X}})' \mathbf{R}_k \tag{3.2}$$

with

$$\mathbf{R}_k = \left(\sum_s q_k \mathbf{z}_k \mathbf{x}_k' \right)^{-1} q_k \mathbf{z}_k \tag{3.3}$$

We use the subscript CALF to identify the family of calibration estimators obtained by the functional form (3.1). This family of estimators includes the GREG estimators because the GREG weighting $w_{k,\text{GREG}}$ given by (1.4) is a special case of (3.2) when $q_k = a_k/c_k$ and $\mathbf{z}_k = \mathbf{x}_k$.

The components of $\mathbf{z}_k$ will usually be functions of the observed auxiliary data, $\{\mathbf{x}_k : k \in s\}$. When $x_{jk} > 0$ for $j = 1, \ldots, J$, one possible choice is $\mathbf{z}_k = (x_{1k}^{p-1}, x_{2k}^{p-1}, \ldots, x_{Jk}^{p-1})'$, where p is a positive constant. Taking $p = 2$ gives $\mathbf{z}_k = \mathbf{x}_k$. Another possibility is $z_{jk} = (x_{jk} - \tilde{x}_j)^{p-1}$, where $\tilde{x}_j$ is a weighted sample mean of the jth x-variable. Requirement (b) eliminates some choices, such as $\mathbf{z}_k = (1, 1, \ldots, 1)'$. The q_k are arbitrary positive coefficients. Multiplying all q_k by a positive value leaves the calibrated weights unchanged. In the simulations in Section 7, we consider $q_k = a_k$ and $q_k = \mu_k \sim \text{Uniform}(0, 1)$. It is surprising that the choice of the q_k has little effect on the variance of the calibrated estimator. Assigning random values to the q_k does not harm the efficiency of the estimators.

A reason for using form (3.1) is that we can control the change from the initial weight a_k to the calibrated weight w_k by appropriate choices of q_k and $\mathbf{z}_k$. To illustrate, consider the case $J = 1$ with $x_k > 0$. Let us take q_k to be constant for $k \in s$ and $z_k = x_k^{p-1}$, where $p > 0$. When $p > 1$, the relative difference between the calibrated weights w_k and the design

weights a_k is much greater for units with large values of x. The larger the value of p, the stronger this tendency. At $p = \infty$, the sample unit $k = (n)$ with the largest value of x (denoted by $x_{(n)}$) has a calibrated weight of $w_{(n)} = a_{(n)} + (X - \hat{X})/x_{(n)}$. The remaining sample units have the calibrated weight $w_k = a_k$. Thus the calibration creates a new weight for only one sample unit, the one with the largest value of x. Relatively large values of p are not without interest. For two of the populations ($K = -0.0001$ and $K = 0.0001$) in Simulation 1 of Section 7, the calibration estimator with $p = 10$ gives considerably lower variance than those corresponding to $p = 1$ and $p = 2$. Even $p = \infty$ gives lower variance than $p = 1$.

The computation of the weights (3.2) requires as inputs: (i) the auxiliary information $\mathbf{X} = \sum_U \mathbf{x}_k$, and (ii) the parameters q_k and $\mathbf{z}_k$. Since $N^{-1}(\mathbf{X} - \hat{\mathbf{X}})'\mathbf{R}_k$ is $\mathrm{O_p}(n^{-1/2})$, the calibrated weights (3.2) respect objective 2. They are ''asymptotically close.'' We have the following result for the functional form calibration estimator determined by (3.1).

Result 3.1. The calibration estimator generated by the functional form (3.1) is

$$\hat{Y}_{\mathrm{CALF}} = \sum_s w_{k,\mathrm{CALF}} y_k$$

where the weights $w_{k,\mathrm{CALF}}$ are given by (3.2). It can be written as the sum of the HT estimator $\hat{Y} = \sum_s a_k y_k$ and an adjustment term,

$$\hat{Y}_{\mathrm{CALF}} = \hat{Y} + (\mathbf{X} - \hat{\mathbf{X}})'\hat{\mathbf{Q}} \tag{3.4}$$

where

$$\hat{\mathbf{Q}} = \left(\sum_s q_k \mathbf{z}_k \mathbf{x}_k'\right)^{-1} \sum_s q_k \mathbf{z}_k y_k \tag{3.5}$$

Its bias is given by

$$\mathrm{Bias}(\hat{Y}_{\mathrm{CALF}}) = E(\hat{Y}_{\mathrm{CALF}}) - Y = -E\{(\hat{\mathbf{X}} - \mathbf{X})'(\hat{\mathbf{Q}} - \mathbf{Q})\} = \mathrm{O}(n^{-1}) \tag{3.6}$$

where $\mathbf{Q}$ is the population analogue of $\hat{\mathbf{Q}}$, obtained by replacing the two sums in (3.5) by their respective expected values. That is,

$$\mathbf{Q} = \left(\sum_U \pi_k q_k \mathbf{z}_k \mathbf{x}_k'\right)^{-1} \sum_U \pi_k q_k \mathbf{z}_k y_k \quad \square \tag{3.7}$$

Note that despite a certain similarity, $\hat{Y}_{\mathrm{CALF}}$ is not a regression estimator. The expression (3.5) for $\hat{\mathbf{Q}}$ reminds us of an instrumental variables regression, as also noted by Deville (1998).

Proof of Result 3.1. Expression (3.4) for $\hat{Y}_{\mathrm{CALF}}$ follows easily from (3.2). To verify expression (3.6) for the bias, we note that

$$\hat{Y}_{\mathrm{CALF}} - Y = \hat{Y} - Y - (\hat{\mathbf{X}} - \mathbf{X})'\mathbf{Q} - (\hat{\mathbf{X}} - \mathbf{X})'(\hat{\mathbf{Q}} - \mathbf{Q}) \tag{3.8}$$

Now take the expected value of both sides of (3.8) to find the bias of $\hat{Y}_{\mathrm{CALF}}$. Since $\mathbf{Q}$ is a population quantity, $\hat{Y} - Y$ and $-(\hat{\mathbf{X}} - \mathbf{X})'\mathbf{Q}$ have expected value zero. The bias expression (3.6) follows because $N^{-1}(\hat{\mathbf{X}} - \mathbf{X})'$ and $N^{-1}(\hat{\mathbf{Q}} - \mathbf{Q})$ are $\mathrm{O_P}(n^{-1/2})$, so the product $N^{-1}(\hat{\mathbf{X}} - \mathbf{X})'(\hat{\mathbf{Q}} - \mathbf{Q})$ is $\mathrm{O_p}(n^{-1})$.

Although the calibration estimator $\hat{Y}_{\text{CALF}}$ is not exactly unbiased, its bias is $O(n^{-1})$, which is usually small even for modest sample sizes. The squared bias is $O(n^{-2})$ and as n increases, it rapidly becomes negligible compared to the design-based variance, which is $O(n^{-1})$.

Several questions arise about the functional form approach used to produce $\hat{Y}_{\text{CALF}}$. In Sections 4 to 6, we examine the following:

1. Since we are free to choose the parameters q_k and $\mathbf{z}_k$ of $\hat{Y}_{\text{CALF}}$, how do the basic statistical properties of $\hat{Y}_{\text{CALF}}$, such as the bias and variance, depend on these choices? Is the variance, for example, highly sensitive to the choice of q_k and $\mathbf{z}_k$?
2. The GREG estimators $\hat{Y}_{\text{GREG}}$ are an important subset of the calibration estimators in $\hat{Y}_{\text{CALF}}$. Can we find conditions that allow us to identify one $\hat{Y}_{\text{GREG}}$ as a best estimator in $\hat{Y}_{\text{CALF}}$?

Remark 3.1. Weight system (3.2) may produce negative weights for some sample units. This causes no problem from a theoretical point of view. However, some users considers the occurrence of negative weights counterintuitive and a practical drawback.

Remark 3.2. We obtain an alternative calibrated weight system $\{w_k : k \in s\}$ by imposing the functional form

$$w_k = q_k \boldsymbol{\lambda}_s' \mathbf{z}_k \tag{3.9}$$

We determine $\boldsymbol{\lambda}_s$ to satisfy constraint (1.2). This generates the calibrated weight system

$$w_k = \mathbf{X}' \mathbf{R}_k \tag{3.10}$$

where $\mathbf{R}_k$ is given by (3.3). The weights (2.1) in Example 2.1 have this form. For any choice of q_k and $\mathbf{z}_k$, the calibrated weight system (3.10) satisfies the consistency objective 1 but not necessarily the closeness objective 2. The estimators built on the weights (3.10) ordinarily have a large design bias and hence large MSE. Therefore we ignore weight systems such as those given by (3.10).

4. Comparing the GREG Estimator with the CALF Estimator

Our objective is to compare $\hat{Y}_{\text{GREG}}$ with any other calibration estimator in the family $\hat{Y}_{\text{CALF}} = \sum_s w_{k,\text{CALF}}\, y_k$, where the weights $w_{k,\text{CALF}}$ are given by (3.2). From (1.5) and (3.4), the difference $Z_s = \hat{Y}_{\text{CALF}} - \hat{Y}_{\text{GREG}}$ can be written as

$$Z_s = (\hat{\mathbf{X}} - \mathbf{X})'(\hat{\mathbf{B}} - \hat{\mathbf{Q}}) \tag{4.1}$$

where $\hat{\mathbf{B}}$ and $\hat{\mathbf{Q}}$ are given by (1.6) and (3.5) respectively. The term Z_s is a complex, nonlinear statistic. We simplify the analysis of Z_s by writing it as the sum of a leading term which is a linear statistic and a remainder term of lower order. To do this, we need the population analogues of $\hat{\mathbf{B}}$ and $\hat{\mathbf{Q}}$. The analogue of $\hat{\mathbf{Q}}$ is $\mathbf{Q}$ given by (3.7). The analogue of $\hat{\mathbf{B}}$ is

$$\mathbf{B} = \left(\sum_U \mathbf{x}_k \mathbf{x}_k' / c_k \right)^{-1} \sum_U \mathbf{x}_k\, y_k / c_k \tag{4.2}$$

Now rewrite (4.1) as

$$Z_s = (\hat{\mathbf{X}} - \mathbf{X})'(\mathbf{B} - \mathbf{Q}) + (\hat{\mathbf{X}} - \mathbf{X})'\{(\hat{\mathbf{B}} - \mathbf{B}) - (\hat{\mathbf{Q}} - \mathbf{Q})\} \tag{4.3}$$

The first term $(\hat{\mathbf{X}} - \mathbf{X})'(\mathbf{B} - \mathbf{Q})$ is a linear statistic obtained by defining the following non-random, scalar quantity for every k:

$$L_k = \mathbf{x}_k'(\mathbf{B} - \mathbf{Q}) \tag{4.4}$$

Thus we have $(\hat{\mathbf{X}} - \mathbf{X})'(\mathbf{B} - \mathbf{Q}) = \sum_s a_k L_k - \sum_U L_k$ which is an HT estimator minus its expected value. Expression (4.3) becomes

$$Z_s = \sum_s a_k L_k - \sum_U L_k + (\hat{\mathbf{X}} - \mathbf{X})'\{(\hat{\mathbf{B}} - \mathbf{B}) - (\hat{\mathbf{Q}} - \mathbf{Q})\}$$

Here $N^{-1}Z_s$ and $N^{-1}(\sum_s a_k L_k - \sum_U L_k)$ are $\mathrm{O_p}(n^{-1/2})$. The remainder is of lower order because $N^{-1}(\hat{\mathbf{X}} - \mathbf{X})'\{(\hat{\mathbf{B}} - \mathbf{B}) - (\hat{\mathbf{Q}} - \mathbf{Q})\}$ is $\mathrm{O_p}(n^{-1})$. Therefore, to a first order approximation,

$$Z_s = \hat{Y}_{\mathrm{CALF}} - \hat{Y}_{\mathrm{GREG}} \cong \sum_s a_k L_k - \sum_U L_k \tag{4.5}$$

In (4.5), and later, the symbol $\cong$ is used with the following meaning. Let W be a statistic with first order approximation W° such that $N^{-1}W$ is $\mathrm{O_p}(n^{-1/2})$, $N^{-1}W^\circ$ is $\mathrm{O_p}(n^{-1/2})$ and $N^{-1}(W - W^\circ)$ is $\mathrm{O_p}(n^{-1})$. Then we write $W \cong W^\circ$. In the cases we consider, W° is the Taylor linearization of W. In (4.5), for example, Z_s corresponds to W and $\sum_s a_k L_k - \sum_U L_k$ corresponds to W°.

Remark 4.1. Suppose there is an exact linear relationship between y_k and $\mathbf{x}_k$ in the population, denoted by $y_k = \mathbf{x}_k'\boldsymbol{\beta}$ for $k = 1, \ldots, N$ and some vector $\boldsymbol{\beta}$. Then it follows from the calibration equation (1.2) and definitions (1.1), (1.5) and (3.4) that for any sample s, $\hat{Y}_{\mathrm{GREG}} = \hat{Y}_{\mathrm{CALF}} = \hat{Y}_{\mathrm{CAL}} = \mathbf{X}'\boldsymbol{\beta}$. In other words, all calibration estimators are identical. In practice, we do not usually have an exact linear relationship between y_k and $\mathbf{x}_k$. Nevertheless, this observation suggests that we examine the difference Z_s between $\hat{Y}_{\mathrm{CALF}}$ and $\hat{Y}_{\mathrm{GREG}}$ by using a linear regression between y_k and $\mathbf{x}_k$. This is done in Sections 6 and 7.

Remark 4.2. Since $N^{-1}Z_s$ and $N^{-1}(\hat{Y}_{\mathrm{GREG}} - Y)$ have the same order in probability, $\mathrm{O_p}(n^{-1/2})$, we cannot claim that Z_s is small or negligible compared to $(\hat{Y}_{\mathrm{GREG}} - Y)$, even in large samples. Since $\hat{Y}_{\mathrm{CALF}} = \hat{Y}_{\mathrm{GREG}} + Z_s$, can we find examples where a $\hat{Y}_{\mathrm{CALF}}$ has smaller variance than a $\hat{Y}_{\mathrm{GREG}}$? We show some examples in the simulations in Section 7.

Remark 4.3. The fact that $Z_s = \hat{Y}_{\mathrm{CALF}} - \hat{Y}_{\mathrm{GREG}}$ and $(\hat{Y}_{\mathrm{GREG}} - Y)$ are of the same order of magnitude may appear to conflict with the result given by Deville and Särndal (1992). They examined different distance measures and found that the resulting calibration estimators were asymptotically equivalent. There is no contradiction with their results. The distance measures that they considered were so similar that they generated essentially equivalent estimators. This was true even for modest sample sizes.

5. Approximating the Variances of the GREG Estimator and the CALF Estimator

By (3.6), the bias of $\hat{Y}_{\mathrm{CALF}}$ is $\mathrm{O}(n^{-1})$. Thus the squared bias is small compared to the variance, which is also $\mathrm{O}(n^{-1})$. We can focus on the variance, rather than on the MSE. We have

$$V(\hat{Y}_{\mathrm{CALF}}) = V(\hat{Y}_{\mathrm{GREG}}) + V(Z_s) + 2Cov(\hat{Y}_{\mathrm{GREG}}, Z_s) \tag{5.1}$$

It is difficult to get exact, closed form expressions for the terms of (5.1). Therefore, we carry out a first order approximation (Taylor linearization) of each of the statistics $(\hat{Y}_{\text{CALF}} - Y)$, $(\hat{Y}_{\text{GREG}} - Y)$ and Z_s, and then we use these results as first order approximations for the terms in (5.1). The linearization of Z_s was given by (4.5). We linearize $(\hat{Y}_{\text{GREG}} - Y)$ in a similar way. To do this, we define the population counterpart of the sample-based residual e_k given by (1.10), namely

$$E_k = y_k - \mathbf{x}_k'\mathbf{B} \tag{5.2}$$

where $\mathbf{B}$ is given by (4.2). Then we can write

$$\hat{Y}_{\text{GREG}} - Y = \sum_s a_k E_k - \sum_U E_k - (\hat{\mathbf{X}} - \mathbf{X})'(\hat{\mathbf{B}} - \mathbf{B}) \tag{5.3}$$

where $N^{-1}(\sum_s a_k E_k - \sum_U E_k)$ is $\text{O}_\text{p}(n^{-1/2})$ and $N^{-1}(\hat{\mathbf{X}} - \mathbf{X})'(\hat{\mathbf{B}} - \mathbf{B})$ is $\text{O}_\text{p}(n^{-1})$. A first order approximation is given by

$$\hat{Y}_{\text{GREG}} - Y \cong \sum_s a_k E_k - \sum_U E_k \tag{5.4}$$

Finally, for the linearization of $\hat{Y}_{\text{CALF}}$, we define

$$H_k = E_k + L_k = y_k - \mathbf{x}_k'\mathbf{B} + \mathbf{x}_k'(\mathbf{B} - \mathbf{Q}) = y_k - \mathbf{x}_k'\mathbf{Q} \tag{5.5}$$

Then, from (4.5) and (5.4),

$$\hat{Y}_{\text{CALF}} - Y = \hat{Y}_{\text{GREG}} - Y + Z_s \cong \sum_s a_k H_k - \sum_U H_k \tag{5.6}$$

The linearized statistics in (4.5), (5.4) and (5.6) all have the form of an HT estimator minus the corresponding expected value. Therefore, approximations of the four terms in (5.1) follow from the well-known results for design-based variance and covariance of HT estimators. These variances and covariances are based on the second order inclusion probabilities from the sampling design, $\pi_{kl} = P(k \,\&\, l \in s)$. We define $a_{kl} = 1/\pi_{kl}$, and $A_{kl} = a_k a_l / a_{kl} - 1$ for $k \neq l$. For $k = l$, we have $\pi_{kl} = \pi_{kk} = \pi_k$, $a_{kl} = a_{kk} = a_k$ and $A_{kl} = A_{kk} = a_k - 1$. The variance of the HT estimator, $\hat{Y} = \sum_s a_k y_k$, can then be written as $V(\hat{Y}) = \sum\sum_U A_{kl} y_k y_l$, a quadratic form in y_k, with the A_{kl} as coefficients. We use $\sum\sum_U$ as a short form for $\sum_{k\in U}\sum_{l\in U}$. We obtain the variances of the linearized variables in (4.5), (5.4) and (5.6) by replacing y_k by L_k, E_k and H_k respectively, in the quadratic form. For example, we have $V(\hat{Y}_{\text{CALF}}) \doteq V(\sum_s a_k H_k) = \sum\sum_U A_{kl} H_k H_l$. We denote such a ''Taylor variance'' by the subscript ''T'', so that $V_T(\hat{Y}_{\text{CALF}}) = \sum\sum_U A_{kl} H_k H_l$. Proceeding in a similar manner with the other terms of (5.1), we obtain the following result.

Result 5.1. The variance of $\hat{Y}_{\text{CALF}}$ is approximated by the following Taylor variance,

$$V(\hat{Y}_{\text{CALF}}) \doteq \sum\sum_U A_{kl} H_k H_l = V_T(\hat{Y}_{\text{CALF}}) \tag{5.7}$$

where A_{kl} is as defined above. The terms on the right-hand side of (5.1) have the approximate expressions

$$V(\hat{Y}_{\text{GREG}}) \doteq \sum\sum_U A_{kl} E_k E_l = V_T(\hat{Y}_{\text{GREG}}) \tag{5.8}$$

$$V(Z_s) \doteq \sum\sum_U A_{kl} L_k L_l = V_T(Z_s) \tag{5.9}$$

$$Cov(\hat{Y}_{\mathrm{GREG}}, Z_s) \doteq \sum\sum_U A_{kl} E_k L_l = Cov_T(\hat{Y}_{\mathrm{GREG}}, Z_s) \tag{5.10}$$

where H_k, E_k and L_k are given respectively by (5.5), (5.2) and (4.4). The Taylor approximations satisfy

$$V_T(\hat{Y}_{\mathrm{CALF}}) = V_T(\hat{Y}_{\mathrm{GREG}}) + V_T(Z_s) + 2Cov_T(\hat{Y}_{\mathrm{GREG}}, Z_s). \quad \square \tag{5.11}$$

The expressions (5.8) to (5.11) are general formulas for any sampling design. In particular, (5.8) is the well-known approximation of the variance of the GREG estimator; see, for example, Chapter 6 of Särndal, Swensson and Wretman (1992). Each particular design, such as simple random sampling without replacement (SRS), gives rise to particular coefficients A_{kl} and corresponding particular expressions for (5.8) to (5.11). Three designs (SRS, stratified SRS and Poisson sampling) are examined later in this section.

Equation (5.11) allows us to answer the question: For a fixed sampling design, are there conditions under which a GREG estimator $\hat{Y}_{\mathrm{GREG}}$ is better than all other members in $\hat{Y}_{\mathrm{CALF}}$? If we interpret ''better'' as ''having smaller Taylor variance,'' then the answer is given by the following result.

Result 5.2. Suppose that for each $k \in U$ there is a positive value c_k^* and a vector $\mathbf{x}_k^*$ of auxiliary variables with known total $\sum_U \mathbf{x}_k^* = \mathbf{X}^*$, such that $\sum\sum_U A_{kl} E_k^* \mathbf{x}_l^* = \mathbf{0}$ where $E_k^* = y_k - \mathbf{x}_k'^* \mathbf{B}^*$ with $\mathbf{B}^* = (\sum_U \mathbf{x}_k^* \mathbf{x}_k'^* / c_k^*)^{-1} (\sum_U \mathbf{x}_k^* y_k / c_k^*)$. Then the GREG estimator $\hat{Y}_{\mathrm{GREG}}^* = \hat{Y} + (\mathbf{X}^* - \hat{\mathbf{X}}^*)' \hat{\mathbf{B}}^*$ with $\hat{\mathbf{B}}^* = (\sum_s a_k \mathbf{x}_k^* \mathbf{x}_k'^* / c_k^*)^{-1} (\sum_s a_k \mathbf{x}_k^* y_k / c_k^*)$ has minimum Taylor variance among all estimators $\hat{Y}_{\mathrm{CALF}} = \sum_s w_{k,\mathrm{CALF}} y_k$ where $w_{k,\mathrm{CALF}} = a_k + (\mathbf{X}^* - \hat{\mathbf{X}}^*)' (\sum_s q_k \mathbf{z}_k \mathbf{x}_k'^*)^{-1} q_k \mathbf{z}_k$. In other words, for this set of auxiliary variables $\mathbf{x}_k^*$, $k \in U$,

$$V_T(\hat{Y}_{\mathrm{CALF}}) \geq V_T(\hat{Y}_{\mathrm{GREG}}^*)$$

over all estimators $\hat{Y}_{\mathrm{CALF}}$ obtained by valid choices of q_k and $\mathbf{z}_k$ as defined in Section 3. $\square$

The result follows immediately from (4.4), (5.10) and (5.11). Suppose $\sum\sum_U A_{kl} E_k^* \mathbf{x}_l^* = \mathbf{0}$ for some vector $\mathbf{x}_k^*$ with known total $\mathbf{X}^*$ and a set of values c_k^*. Then $\sum\sum_U A_{kl} E_k^* \mathbf{x}_l'^* (\mathbf{B} - \mathbf{Q}) = 0$ and using (4.4) and (5.10) we have $Cov_T(\hat{Y}_{\mathrm{GREG}}, Z_s) = \sum\sum_U A_{kl} E_k^* \mathbf{x}_l'^* (\mathbf{B} - \mathbf{Q}) = 0$. It follows from (5.11) that

$$V_T(\hat{Y}_{\mathrm{CALF}}) = V_T(\hat{Y}_{\mathrm{GREG}}^*) + V_T(Z_s) \geq V_T(\hat{Y}_{\mathrm{GREG}}^*)$$

where Z_s is given by $\hat{Y}_{\mathrm{CALF}} - \hat{Y}_{\mathrm{GREG}}^*$. We note that $\mathbf{x}_k^*$ and c_k^* determine the optimal GREG estimator, $\hat{Y}_{\mathrm{GREG}}^* = \hat{Y} + (\mathbf{X}^* - \hat{\mathbf{X}}^*)' \hat{\mathbf{B}}^*$, in this result. It is possible for other estimators in $\hat{Y}_{\mathrm{CALF}}$ to attain this minimum variance but we do not have any other way of identifying them.

Whether or not we can find $\mathbf{x}_k^*$ and c_k^* satisfying $\sum\sum_U A_{kl} E_k^* \mathbf{x}_l^* = \mathbf{0}$ depends on the sampling design. Let us look at three designs: SRS, stratified SRS and Poisson sampling.

Example 5.1. Consider the design SRS with sampling fraction $f = n/N$. Then (5.8) to (5.11) have the familiar forms: $V_T(\hat{Y}_{\mathrm{CALF}}) = N^2\{(1-f)/n\} S_{HU}^2$; $V_T(\hat{Y}_{\mathrm{GREG}}) = N^2\{(1-f)/n\} S_{EU}^2$; $V_T(Z_s) = N^2\{(1-f)/n\} S_{LU}^2$, where S_{HU}^2, S_{EU}^2 and S_{LU}^2 are the population variances of H_k, E_k and L_k respectively, and $Cov_T(\hat{Y}_{\mathrm{GREG}}, Z_s) = N^2\{(1-f)/n\} \sum_U (E_k - \bar{E}_U) L_k / (N-1)$, where $\bar{E}_U = \sum_U E_k / N$. This Taylor covariance is zero and Result 5.2 is satisfied, if

$$\sum_U (E_k - \bar{E}_U) \mathbf{x}_k = \mathbf{0} \tag{5.12}$$

We have $\sum_U E_k \mathbf{x}_k / c_k = \mathbf{0}$ because of the definitions of E_k and $\mathbf{B}$, given by (5.2) and (4.2) respectively. Suppose we have an auxiliary vector $\mathbf{x}_{0k}$ with known totals $\sum_U \mathbf{x}_{0k} = \mathbf{X}_0$. Then we can satisfy (5.12) by (i) defining $\mathbf{x}_k = (1, \mathbf{x}_{0k}')'$ and (ii) choosing c_k to have the same constant value, say $c_k = 1$ for all k. That is, we include an auxiliary variable with value 1 for every k, in addition to the other x-variables with known totals. This is possible since the population size $N = \sum_U 1$ is known for calibration. The inclusion of this special auxiliary variable corresponds to specifying an intercept in the regression. It follows from these conditions that $\bar{E}_U = 0$ and $\sum_U E_k \mathbf{x}_k = \mathbf{0}$, so (5.12) is satisfied. Consequently, $\hat{Y}_{\text{GREG}}$ with $c_k = 1$ and $\mathbf{x}_k = (1, \mathbf{x}_{0k}')'$ has minimum Taylor variance among all $\hat{Y}_{\text{CALF}}$ estimators formed from this set of auxiliary variables.

Example 5.2. Consider the stratified SRS design with n_h units sampled from the N_h units in stratum U_h, with the sampling fraction $f_h = n_h/N_h$ for $h = 1, \ldots, H$. We require that

$$\sum \sum_U A_{kl} E_k \mathbf{x}_l = \sum_{h=1}^{H} F_h \sum_{k \in U_h} (E_k - \bar{E}_{u_h}) \mathbf{x}_k = \mathbf{0} \tag{5.13}$$

where $\bar{E}_{U_h} = \sum_{U_h} E_k / N_h$ and $F_h = N_h^2 (1 - f_h)/(n_h(N_h - 1)) = (N_h/(N_h - 1))(1/f_h - 1)$. Again, using $\sum_U E_k \mathbf{x}_k / c_k = \mathbf{0}$, we can meet Result 5.2 with the following choices of $\mathbf{x}_k$ and c_k. Define $\mathbf{x}_k$ to include the stratum identifier $\boldsymbol{\delta}_k = (\delta_{1k}, \ldots, \delta_{hk}, \ldots, \delta_{Hk})'$, where $\delta_{hk} = 1$ if $k \in U_h$ and $\delta_{hk} = 0$ if $k \notin U_h$. This is permitted since the stratum sizes N_h are known. That is, we take $\mathbf{x}_k = (\boldsymbol{\delta}_k', \mathbf{x}_{0k}')'$, where $\mathbf{x}_{0k}$ consists of the available auxiliary variables other than the stratum identifer. Next, choose $1/c_k = F_h$ for all $k \in U_h$. Then, from $\sum_U E_k \mathbf{x}_k / c_k = \mathbf{0}$ it follows that $\sum_{h=1}^{H} F_h \sum_{k \in U_h} E_k \mathbf{x}_k = \mathbf{0}$ and $\sum_{k \in U_h} E_k = 0$ for $h = 1, \ldots, H$, so (5.13) is satisifed. Therefore, with $\mathbf{x}_k = (\boldsymbol{\delta}_k', \mathbf{x}_{0k}')'$ and $c_k = 1/F_h$, $\hat{Y}_{\text{GREG}}$ has minimum Taylor variance among all $\hat{Y}_{\text{CALF}}$ estimators based on this set of auxiliary variables $\mathbf{x}_k$. Here the best choice for c_k is not constant overall, but constant within strata. Also note that to close approximation, $1/c_k = a_k - 1$ since $a_k = 1/\pi_k = 1/f_h$ for all $k \in U_h$, and $N_h/(N_h - 1) \doteq 1$ under stratified SRS.

Example 5.3. Consider Poisson sampling, defined as follows: Unit k is selected or not selected depending on the outcome of a Bernoulli experiment with selection probability π_k proportional to a suitable size measure, for $k = 1, \ldots, N$, where the N experiments are independent. This is an exact probability proportional-to-size design, with a random sample size. The double sum in Result 5.2 simplifies into a single sum, namely, $\sum \sum_U A_{kl} E_k \mathbf{x}_l = \sum_U (a_k - 1) E_k \mathbf{x}_k = \mathbf{0}$. This holds if we specify $1/c_k = a_k - 1$, but not otherwise. Thus, $\hat{Y}_{\text{GREG}}$ with $c_k = 1/(a_k - 1)$ has minimum Taylor variance among all $\hat{Y}_{\text{CALF}}$ estimators based on the given $\mathbf{x}_k$. We note that the result is valid for any fixed set of auxiliary variables with known totals.

The practical recommendations from Result 5.2 are as follows. For any particular sampling design, we should try to find $\mathbf{x}_k$ and c_k so that $Cov_T(\hat{Y}_{\text{GREG}}, Z_s) = 0$ and use the estimator $\hat{Y}_{\text{GREG}}$ corresponding to these choices. It has minimum Taylor variance among all estimators $\hat{Y}_{\text{CALF}} = \sum_s w_{k,\text{CALF}} y_k$ with weights of the form (3.2). Ordinarily, the Taylor variance is a sufficiently close approximation to the exact variance. For stratified SRS and Poisson sampling, the preferred choice is $1/c_k = a_k - 1$. The common practice of taking $c_k = 1$ for all k is not supported by theory for these designs and probably not

for other designs either. The use of $1/c_k = a_k - 1$ is advantageous from another point of view, as noted in Särndal (1996).

6. Superpopulation Models and Linear Representations

Result 5.2 provides a condition to identify a $\hat{Y}_{\text{GREG}}$ with minimum Taylor variance in the family of estimators $\hat{Y}_{\text{CALF}}$. If no $\mathbf{x}_k^*$ and c_k^* exist to meet this condition, then we may find more efficient estimators in $\hat{Y}_{\text{CALF}}$ outside of $\hat{Y}_{\text{GREG}}$. This depends on whether the auxiliary variables in the calibration are ''sufficient'' to explain y, or whether there are other variables that are important but are excluded from the calibration because their totals are not known. We elaborate on this idea.

We can rewrite (5.2) as the linear representation

$$y_k = \mathbf{x}_k'\mathbf{B} + E_k \qquad k = 1, \ldots, N \tag{6.1}$$

where $\mathbf{x}_k$ is the auxiliary vector used in calibration and $\mathbf{B}$ is given by (4.2). We know by (5.8) that the residuals E_k are essential for determining the variance of $\hat{Y}_{\text{GREG}}$. In general, the more variables we can include in $\mathbf{x}_k$, the smaller the residuals E_k in the linear representation (6.1). This justifies objective 3 in Section 2. In calibration, our main concern is not whether or not there exists a ''true'' linear relationship between y_k and $\mathbf{x}_k$. We obtain an efficient calibration estimator if $\mathbf{x}_k$ explains a significant part of the variability of y_k. The degree to which this variability is explained by the calibration vector $\mathbf{x}_k$ varies from one variable of interest to another.

In a typical survey, the variables used for calibration will explain some but not all of the variation in the y_k values. There may exist other relevant explanatory variables than those present in the calibration vector $\mathbf{x}_k$, but absence of control totals for these additional variables prevents us from using them in the calibration. For unit k, let $\mathbf{x}_{k,\text{add}}$ (of dimension M) be the vector of those additional variable values. The population total, $\sum_U \mathbf{x}_{k,\text{add}} = \mathbf{X}_{\text{add}}$, is unknown. Denote the extended explanatory vector by $\mathbf{x}_{k,\text{ext}}' = (\mathbf{x}_k', \mathbf{x}_{k,\text{add}}')$, of dimension $J + M$. The vector $\mathbf{x}_{k,\text{ext}}$ may not explain all of the variation in y, but it may come closer than $\mathbf{x}_k$ alone.

We can then represent the same finite population as

$$y_k = \mathbf{x}_{k,\text{ext}}'\mathbf{B}_{\text{ext}} + E_{k,\text{ext}}, \qquad k = 1, \ldots, N \tag{6.2}$$

where $\mathbf{B}_{\text{ext}}$ is given by (4.2) if $\mathbf{x}_k$ is replaced by $\mathbf{x}_{k,\text{ext}}$ and $E_{k,\text{ext}}$ is the new residual.

The population-based residuals E_k and $E_{k,\text{ext}}$ in (6.1) and (6.2) cannot be calculated except in a census where $(y_k, \mathbf{x}_{k,\text{ext}})$ is observed for all N population units. If we had a census, we would find the mean squared residual $(1/N)\sum_U E_{k,\text{ext}}^2$ to be smaller than $(1/N)\sum_U E_k^2$, and considerably smaller if the fit is made substantially better by adding $x_{k,\text{add}}$ to the explanatory vector.

Here, as with many other issues in survey sampling, it is fruitful to view the finite population as a realization of a superpopulation model. The models that we now introduce are also important for the simulations in Section 7. Consider first the superpopulation model, denoted ξ, stating that

$$y_k = \mathbf{x}_k'\boldsymbol{\beta} + \varepsilon_k, \qquad k = 1, \ldots, N \tag{6.3}$$

where $E_\xi(\varepsilon_k) = 0$ for every k. The explanatory vector $\mathbf{x}_k$ consists of exactly those

x-variables for which control totals are available and used in the calibration equation (1.2), and no others. In this case (and more generally, when the model does not contain any variable not used in the calibration), we shall say that the model is saturated by the calibration.

In a survey, we cannot assume that the calibration vector $\mathbf{x}_k$ fully explains the study variable y. Other x-variables than those in the vector $\mathbf{x}_k$ are likely to be relevant. We must think in terms of an extended linear superpopulation model denoted by ξ_{ext} and given by

$$y_k = \mathbf{x}'_{k,\text{ext}}\boldsymbol{\beta}_{\text{ext}} + \varepsilon_{k,\text{ext}}, \qquad k = 1, \ldots, N \tag{6.4}$$

where $\boldsymbol{\beta}'_{\text{ext}} = (\boldsymbol{\beta}', \boldsymbol{\beta}'_{\text{add}})$, $\mathbf{x}'_{k,\text{ext}} = (\mathbf{x}'_k, \mathbf{x}'_{k,\text{add}})$ and $E_{\xi_{\text{ext}}}(\varepsilon_{k,\text{ext}}) = 0$ for every k. We can segment (6.4) as $y_k = \mathbf{x}'_k\boldsymbol{\beta} + \mathbf{x}'_{k,\text{add}}\boldsymbol{\beta}_{\text{add}} + \varepsilon_{k,\text{ext}}$, where $\mathbf{x}_k$ is the calibration vector and $\mathbf{x}_{k,\text{add}}$ consists of the x-variables not participating in the calibration. If $\boldsymbol{\beta}_{\text{add}} = \mathbf{0}$, (6.4) reduces to the saturated model ξ given by (6.3). But if $\boldsymbol{\beta}_{\text{add}} \neq \mathbf{0}$, we say that the model (6.4) is unsaturated, because it contains relevant variables other than those included in the calibration. If a finite population is generated by (6.4) with $\boldsymbol{\beta}_{\text{add}} \neq \mathbf{0}$, then (6.3) is an inappropriate model; in particular it is wrong to assume $E_\xi(\varepsilon_k) = 0$ for every k.

The distinction just made between saturated and unsaturated models has important implications for the difference between any members of $\hat{Y}_{\text{CALF}}$ and $\hat{Y}_{\text{GREG}}$, as noted by the following result.

Result 6.1. For any fixed sample s, the difference $Z_s = \hat{Y}_{\text{CALF}} - \hat{Y}_{\text{GREG}}$ has the following properties: $E_\xi(Z_s) = 0$ under the saturated model ξ, but $E_{\xi_{\text{ext}}}(Z_s) \neq 0$ under the unsaturated model ξ_{ext}. □

Result 6.1 states that $\hat{Y}_{\text{CALF}}$ and $\hat{Y}_{\text{GREG}}$ are equal in expectation under a saturated model, but not so under an unsaturated one. The corresponding frequency interpretation is that, for any fixed sample s, $\hat{Y}_{\text{CALF}}$ and $\hat{Y}_{\text{GREG}}$ are equal on the average over all finite populations generated by the saturated model ξ. It follows that $\hat{Y}_{\text{CALF}}$ and $\hat{Y}_{\text{GREG}}$ are equal when averaged over all samples as well as over all finite populations generated by ξ. This is confirmed by the Monte Carlo simulations in Section 7.

Proof of Result 6.1: From (6.3), and the definitions (1.6) and (3.5) for $\hat{\mathbf{B}}$ and $\hat{\mathbf{Q}}$, it follows that $E_\xi(\hat{\mathbf{B}}) = E_\xi(\hat{\mathbf{Q}}) = \boldsymbol{\beta}$, and therefore using (4.1) we have $E_\xi(Z_s) = 0$. By contrast, under the model (6.4) with $\boldsymbol{\beta}_{\text{add}} \neq \mathbf{0}$, a similar argument shows that $E_{\xi_{\text{ext}}}(Z_s) \neq 0$.

7. Monte Carlo Simulation

We carried out a Monte Carlo simulation to illustrate and confirm some of the findings in earlier sections. We generated repeated finite populations under the superpopulation model,

$$y_k = x_k + K\,h(x_k) + \varepsilon_k, \qquad k = 1, \ldots, N \tag{7.1}$$

where K is a constant and $\varepsilon_k \sim \text{Normal}(0, 25^2)$ is a random error term. We let $x_k \in (100, 300)$, and we took $h(x_k)$ to be the third degree polynomial $h(x_k) = (x_k - 100)(x_k - 200)(x_k - 300)$. A data set (y_k, x_k), $k = 1, \ldots, N$, generated by (7.1) can be thought of as corresponding to a stratum of units for which a measure of size, x_k, falls in the interval (100, 300). We assume that $X = \sum_U x_k$ is a known total used in the calibration.

For $K = 0$, (7.1) simplifies into $y_k = x_k + \varepsilon_k$. When $K \neq 0$, the term $K\,h(x_k)$ creates a wave curve about the line $y_k = x_k$, symmetric around $x_k = 200$. It has the effect of

weakening the correlation between y and x. In order to limit the amplitude of the wave, we used values of K close to zero. The presence of the term $K\,h(x_k)$ permits us to study the behaviour of $\hat{Y}_{\text{CALF}}$ in unsaturated cases, where we are not in a position to calibrate on all x-variables relevant for explaining y. Here, we do not calibrate on the two variables x_k^2 and x_k^3, because we assume that the totals $\sum_U x_k^2$ and $\sum_U x_k^3$ are unknown. The population size $N = \sum_U 1$ represents an additional total, which may or may not be used in the calibration. When $K \neq 0$, we can express (7.1) in the form (6.4) of an extended model ξ_{ext} where $\mathbf{x}_{k,\text{ext}} = (1, x_k, x_k^2, x_k^3)'$ and $\boldsymbol{\beta}_{\text{ext}}$ are vectors of dimension 4.

The simulations consisted of generating repeated finite populations using (7.1), and for each population, repeated SRS samples were drawn. More precisely, the steps were:

(a) For a fixed value of K, and for $k = 1, \ldots, 1{,}000$, draw $x_k \sim \text{Uniform}(100, 300)$; draw $\varepsilon_k \sim \text{Normal}(0, 25^2)$, independent of x_k; compute the corresponding value y_k from (7.1). The result is a finite population of $N = 1{,}000$ values (y_k, x_k), $k = 1, \ldots, 1{,}000$.

(b) From this finite population, select 1,000 independent samples s_j, each of size $n = 200$, using SRS. Each sample is replaced before the next is drawn. The values of x_k and y_k are recorded for $k \in s_j$, $j = 1, \ldots, 1{,}000$.

Step (a) was carried out 500 times. That is, we generated 500 finite populations, each of size $N = 1{,}000$, and for each of these, 1,000 SRS samples were drawn. For each value of K, the simulation was thus based on $500 \times 1{,}000 = 0.5 \times 10^6$ SRS samples. Three values of K were used: $K = 0$, $K = -0.0001$ and $K = 0.0001$.

For each value of K and for each of the 500 populations, the mean of the 1,000 y_k values is roughly 200, because by symmetry $h(x_k)$ is zero on average when $x_k \sim \text{Uniform}(100, 300)$. The variance of the 1,000 y_k values depends on K. For example, for $K = 0$, the variance is roughly $(1/12)(200^2) + (25^2) = 3958.3$, with a corresponding standard deviation of roughly $(3958.3)^{1/2} = 62.9$.

For each of the 0.5×10^6 samples, we computed several estimators $\hat{Y}_{\text{CALF}} = \sum_s w_{k,\text{CALF}}\, y_k$, corresponding to different parameters q_k and $\mathbf{z}_k$ in the weights $w_{k,\text{CALF}}$ given by (3.2). One choice was $q_k = a_k/c_k$ (with $c_k = 1$) and $\mathbf{z}_k = \mathbf{x}_k$ which gives a specific GREG estimator. Two different simulations were carried out, Simulation 1 and Simulation 2, as described below.

We computed summary Monte Carlo measures based on the 0.5×10^6 values obtained for each estimator under consideration. These were the average bias, the average squared bias, the average variance and the average mean squared error defined respectively, as

$$\text{AvBias} = \frac{1}{500}\sum_{i=1}^{500}(\bar{\hat{Y}}_i - Y_i)$$

$$\text{AvBias2} = \frac{1}{500}\sum_{i=1}^{500}(\bar{\hat{Y}}_i - Y_i)^2$$

$$\text{AvVar} = \frac{1}{500 \times 1000}\sum_{i=1}^{500}\sum_{j=1}^{1000}(\hat{Y}_{ij} - \bar{\hat{Y}}_i)^2$$

$$\text{AvMSE} = \frac{1}{500 \times 1000}\sum_{i=1}^{500}\sum_{j=1}^{1000}(\hat{Y}_{ij} - Y_i)^2$$

where $\bar{\hat{Y}}_i = \frac{1}{1000}\sum_{j=1}^{1000} \hat{Y}_{ij}$ and $\hat{Y}_{ij}$ is the value of the estimator $\hat{Y}$ for the jth sample drawn from the ith population with total Y_i. We also calculated the average variance of Z_s and the average covariance between Z_s and $\hat{Y}_{\text{GREG}}$ as

$$\text{AvVarZ} = \frac{1}{500 \times 1000}\sum_{i=1}^{500}\sum_{j=1}^{1000}(Z_{sij} - \bar{Z}_{si})^2$$

$$\text{AvCov} = \frac{1}{500 \times 1000}\sum_{i=1}^{500}\sum_{j=1}^{1000}(\hat{Y}_{\text{GREG}\,ij} - \bar{\hat{Y}}_{\text{GREG}\,i})(Z_{sij} - \bar{Z}_{si})$$

with $\bar{\hat{Y}}_{\text{GREG}\,i} = \frac{1}{1000}\sum_{j=1}^{1000} \hat{Y}_{\text{GREG}ij}$ and $\bar{Z}_{si} = \frac{1}{1000}\sum_{j=1}^{1000} Z_{sij}$ where $\hat{Y}_{\text{GREG}ij}$ and Z_{sij} are the respective values of $\hat{Y}_{\text{GREG}}$ and Z_s for the jth sample drawn from the ith population.

The results of Simulations 1 and 2 depend on the following factors, which are specified at the outset:

(i) The *calibration variables* that define the vector $\mathbf{x}_k$ (with known total $\mathbf{X} = \sum_U \mathbf{x}_k$) in the calibration.
(ii) The *parameters* q_k and $\mathbf{z}_k$ specified for the weights (3.2) for $\hat{Y}_{\text{CALF}}$.
(iii) The *constants* c_k in $\hat{Y}_{\text{GREG}}$.
(iv) The set of *x-variables in the superpopulation model*, and the overlap of these variables with the set of calibration variables specified in (i).

Simulation 1 was carried out with $\dim(\mathbf{x}_k) = \dim(\mathbf{z}_k) = 1$ and the following specifications:

(i) *Calibration variables*: $\mathbf{x}_k = x_k$ (scalar); $X = \sum_U x_k$ is the only total used for calibration.
(ii) *Parameters*: $\mathbf{z}_k = x_k^{p-1}$ for $p = 1, 2, \ldots$; for each value of p, two specifications for q_k: (1) $q_k = a_k = N/n$ for all k and (2) randomly generated q_k as $q_k = \mu_k \sim \text{Uniform}(0, 1)$. This extreme choice of coefficients is used to show that even randomly chosen q_k do not harm the efficiency.
(iii) The *constants* c_k : $c_k = 1$ for all k.
(iv) *Superpopulation models*: For $K = 0$, (7.1) is the saturated model $y_k = x_k + \varepsilon_k$. For $K = -0.0001$ or $K = 0.0001$, (7.1) gives two unsaturated models, $y_k = x_k'\beta + \mathbf{x}_{k,\text{add}}'\boldsymbol{\beta}_{\text{add}} + \varepsilon_{k,\text{ext}}$ with $\dim(\beta) = 1$, $\dim(\boldsymbol{\beta}_{\text{add}}) = \dim(\mathbf{x}_{k,\text{add}}) = 3$, $\mathbf{x}_{k,\text{add}} = (1, x_k^2, x_k^3)'$, $\beta = 1 + 110000K$ and $\boldsymbol{\beta}_{\text{add}} = (-6000000\,K, -600\,K, K)'$. This means that the three totals not used in the calibration are $N = \sum_U 1$, $\sum_U x_k^2$ and $\sum_U x_k^3$.

For Simulation 1, we have

$$\hat{\mathbf{B}} = \hat{B} = \sum_s x_k y_k \Big/ \sum_s x_k^2$$

and

$$\hat{\mathbf{Q}} = \hat{Q} = \sum_s q_k x_k^{p-1} y_k \Big/ \sum_s q_k x_k^p$$

In particular, for $p = 1$ and $q_k = a_k$, $\hat{Y}_{\text{CALF}}$ becomes the classical ratio estimator, $\hat{Y}_{\text{CALF}} = X(\hat{Y}/\hat{X})$. The results of Simulation 1, given in Table 1, lead to the following comments:

The effect of the exponent p: In the saturated case ($K = 0$, model $y_k = x_k + \varepsilon_k$), $\hat{B} - \hat{Q}$ has model expectation zero for any value of p. By Result 6.1, $Z_s = \hat{Y}_{\text{CALF}} - \hat{Y}_{\text{GREG}} = (\hat{X} - X)(\hat{B} - \hat{Q})$ is zero on the average over all populations and all samples. That is, $\hat{Y}_{\text{CALF}}$ and $\hat{Y}_{\text{GREG}}$ are identical on average. This is confirmed in Table 1. For all values of p, we have practically identical results for all tabulated quantities. In the unsaturated cases ($K = -0.0001$ and $K = 0.0001$), we cannot use Result 5.2 to identify $\hat{Y}_{\text{GREG}}$ as the best (minimum Taylor variance) estimator. Although $\sum_U x_k E_k = 0$, we do not have $\sum_U E_k = 0$. So even before the simulation, we might expect some other value of p to produce a smaller variance than $p = 2$. However, we have no means of predicting the optimal value of p. Table 1 shows that the variance (and the approximately equivalent MSE) of $\hat{Y}_{\text{CALF}}$ decreases as p increases, reaches a minimum $p = 8$ and starts to increase again for larger values of p. There is a strong negative covariance $Cov(\hat{Y}_{\text{GREG}}, Z_s)$ for values $p > 2$, which explains why $p = 2$ is not best. The decrease in the variance as p increases from $p = 1$ (ratio estimator) to $p = 8$ is considerable (around 15%), so there is considerable incentive in practice to prefer the calibration estimator with $p = 8$. The problem is that we cannot predict the optimal value of p. It depends on the population data values.

The effect of the coefficients q_k: For all values of K, we found that the effect of the q_k is neglibible. The coefficients $q_k = a_k = N/n$ and $q_k = \mu_k \sim \text{Uniform}(0, 1)$ give almost the same variance. In order to save space, we show $q_k = \mu_k \sim \text{Uniform}(0, 1)$ in Table 1 only for the case $K = 0.0001$, but similar results were found for other values of K.

Simulation 2 was carried out with $\dim(\mathbf{x}_k) = \dim(\mathbf{z}_k) = 2$ and the following specifications:

(i) *Calibration variables*: $\mathbf{x}_k = (1, x_k)'$; the vector total used for calibration is $\mathbf{X} = (N, X)'$, where $N = \sum_U 1$ and $X = \sum_U x_k$.
(ii) *Parameters*: $\mathbf{z}_k = (1, x_k^{p-1})'$ for $p = 2, 3, \ldots$ ($p = 1$ was excluded because the matrix $\sum_s q_k \mathbf{z}_k \mathbf{x}_k'$ is singular at $p = 1$); q_k as in Simulation 1 (two options).
(iii) The *constants* c_k : $c_k = 1$ for all k.
(iv) *Superpopulation models*: In the saturated case ($K = 0$), the model is $y_k = x_k + \varepsilon_k$. In the unsaturated cases ($K = -0.0001$; $K = 0.0001$), (7.1) can be written as $y_k = \mathbf{x}_k'\boldsymbol{\beta} + \mathbf{x}_{k,\text{add}}'\boldsymbol{\beta}_{\text{add}} + \varepsilon_{k,\text{ext}}$, where all of $\boldsymbol{\beta}$, $\mathbf{x}_k$, $\boldsymbol{\beta}_{\text{add}}$ and $\mathbf{x}_{k,\text{add}}$ have dimension 2; $\boldsymbol{\beta} = (-6000000\,K,\ 1 + 110000\,K)'$; $\mathbf{x}_k = (1, x_k)'$; $\boldsymbol{\beta}_{\text{add}} = (-600\,K, K)'$; $\mathbf{x}_{k,\text{add}} = (x_k^2, x_k^3)'$; the totals excluded from the calibration are $\sum_U x_k^2$ and $\sum_U x_k^3$.

The results of Simulation 2, also given in Table 1, lead to the following comments:

The effect of the exponent p: In the saturated case ($K = 0$, model $y_k = x_k + \varepsilon_k$), Result 6.1 states that $\hat{Y}_{\text{CALF}}$ and $\hat{Y}_{\text{GREG}}$ are identical on average over all populations and all samples. This is confirmed in Table 1. The tabulated quantities are practically identical for most values of p. There is no variance reduction (in fact, a slight increase) compared to Simulation 1. That is, adding N to the calibration totals gives no improvement over the calibration on X alone. In the unsaturated cases ($K = -0.0001$ and $K = 0.0001$), Result 5.2 shows that $\hat{Y}_{\text{GREG}}$ has minimum Taylor variance, because $\sum_U x_k E_k = \sum_U E_k = 0$.

Table 1. Averages based on 1,000 samples of size 200 within each of 500 populations of size 1,000 (actual values = displayed values $\times 10^3$)

a indicates: $q_k = a_k = N/n$

μ indicates: $q_k = \mu_k \sim$ Uniform (0, 1)

The data values for each population were generated under the model
$y = x + K(x - 100)(x - 200)(x - 300) + \varepsilon$ where $\varepsilon \sim$ Normal $(0, 25^2)$ and $x \sim$ Uniform (100, 300)

			Simulation 1: Calibration variable x						Simulation 2: Calibration variables				
			AvBias	AvBias2	AvVar	AvMSE	AvCov	AvVarZ	AvBias	AvBias2	AvVar	AvMSE	AvC
$K = 0$	a	p = 1	0.000	2.63	2502.55	2505.18	0.48	0.08	—	—	—	—	—
		p = 2	0.000	2.63	2501.99	2504.62	0.00	0.00	0.000	2.65	2511.09	2513.74	0.0
		p = 3	0.000	2.63	2501.62	2504.25	−0.42	0.05	0.000	2.65	2511.26	2513.91	−0.0
		p = 4	0.000	2.63	2501.41	2504.04	−0.73	0.15	0.000	2.65	2511.78	2514.43	−0.0
		p = 5	0.000	2.63	2501.31	2503.94	−0.97	0.28	0.000	2.65	2512.54	2515.19	−0.
		p = 6	0.000	2.63	2501.28	2503.91	−1.13	0.42	0.000	2.65	2513.47	2516.12	−0.
		p = 7	0.000	2.63	2501.31	2503.94	−1.25	0.57	0.000	2.65	2514.42	2517.17	−0.
		p = 8	0.000	2.63	2501.37	2504.00	−1.34	0.71	0.000	2.65	2515.65	2518.30	−0.
		p = 9	0.000	2.63	2501.45	2504.08	−1.41	0.86	0.000	2.65	2516.83	2519.48	−0.
		p = 10	0.000	2.63	2501.55	2504.18	−1.46	1.01	0.000	2.65	2518.06	2520.71	−0.
		p = 20	0.000	2.63	2502.86	2505.49	−1.68	2.55	0.000	2.67	2531.22	2533.89	−0.
		p = 50	0.000	2.64	2507.41	2510.05	−1.87	7.30	0.000	2.71	2573.09	2575.80	−0.
		p = ∞	0.000	2.71	2552.59	2555.30	−3.18	53.85	0.000	3.62	2991.20	2994.82	−3.
$K = -0.0001$	a	p = 1	−0.035	7.13	5547.28	5554.41	302.26	12.68	—	—	—	—	—
		p = 2	−0.029	6.37	5233.10	5239.47	0.00	0.00	−0.004	3.53	3425.38	3428.91	0.
		p = 3	−0.023	5.79	5017.02	5022.81	−223.64	6.97	0.003	3.52	3426.14	3429.66	0.
		p = 4	−0.017	5.41	4878.62	4884.03	−375.13	19.68	0.008	3.58	3428.24	3431.82	0.
		p = 5	−0.013	5.19	4795.31	4800.50	−470.00	31.02	0.013	3.69	3433.03	3436.72	1.
		p = 6	−0.011	5.07	4748.98	4754.05	−524.20	38.78	0.017	3.82	3442.16	3445.98	2.
		p = 7	−0.009	5.01	4726.85	4731.86	−550.65	43.03	0.020	3.94	3456.64	3460.58	3.
		p = 8	−0.007	4.98	4720.34	4725.32	−558.76	44.60	0.022	4.06	3476.68	3480.74	3.
		p = 9	−0.006	4.97	4723.86	4728.83	−554.98	44.34	0.024	4.16	3501.85	3506.01	4.
		p = 10	−0.006	4.97	4733.75	4738.72	−543.68	42.93	0.026	4.26	3531.37	3535.63	4.
		p = 20	−0.006	5.16	4905.98	4911.14	−349.69	21.36	0.032	5.05	3907.48	3912.53	4.
		p = 50	−0.007	5.52	5195.46	5200.98	−48.61	10.12	0.039	6.38	4646.32	4652.70	1.
		p = ∞	−0.007	5.92	5497.03	5502.95	202.33	61.15	0.047	8.98	5776.63	5785.61	2.

$K = 0.0001$	a	p = 1	0.027	5.99	5553.67	5559.66	302.40	12.70	—	—	—	—	—
		p = 2	0.023	5.39	5239.18	5244.57	0.00	0.00	0.000	3.13	3432.31	3435.44	0.0
		p = 3	0.016	4.94	5022.55	5027.49	−224.08	7.00	−0.007	3.17	3432.86	3436.03	0.
		p = 4	0.010	4.66	4883.56	4888.22	−376.12	19.78	−0.013	3.29	3434.66	3437.95	0.3
		p = 5	0.007	4.51	4799.73	4804.24	−471.54	31.21	−0.018	3.44	3439.00	3442.44	1.0
		p = 6	0.004	4.44	4752.96	4757.40	−526.22	39.06	−0.022	3.60	3447.55	3451.15	1.6
		p = 7	0.002	4.41	4730.46	4734.87	−553.08	43.38	−0.025	3.76	3461.37	3465.13	2.
		p = 8	0.000	4.40	4723.67	4728.07	−561.50	45.01	−0.027	3.90	3480.72	3484.62	2.3
		p = 9	−0.001	4.40	4726.96	4731.36	−557.98	44.77	−0.029	4.02	3505.22	3509.24	2.8
		p = 10	−0.001	4.42	4736.66	4741.08	−546.87	43.38	−0.030	4.14	3534.09	3538.23	3.0
		p = 20	−0.002	4.59	4908.04	4912.63	−353.61	21.68	−0.038	5.00	3905.48	3910.48	1.0
		p = 50	0.000	4.89	5196.12	5201.01	−53.57	10.01	−0.045	6.39	4636.50	4642.89	−3.4
		p = ∞	−0.001	5.26	5510.94	5516.20	210.16	61.48	−0.056	9.22	5803.38	5812.60	−6.3
	μ	p = 1	0.028	5.99	5555.32	5561.31	303.23	13.52	—	—	—	—	—
		p = 2	0.022	5.38	5240.81	5246.19	1.00	0.63	0.000	3.12	3439.37	3442.49	0.4
		p = 3	0.016	4.93	5023.95	5028.88	−223.18	7.50	−0.007	3.17	3440.05	3443.22	0.3
		p = 4	0.010	4.66	4884.67	4889.33	−375.49	20.25	−0.013	3.28	3442.11	3445.39	0.9
		p = 5	0.006	4.51	4800.54	4805.05	−471.26	31.74	−0.018	3.45	3446.75	3450.20	1.2
		p = 6	0.004	4.43	4753.50	4757.93	−526.32	39.68	−0.022	3.60	3455.59	3459.19	2.0
		p = 7	0.002	4.40	4730.75	4735.15	−553.54	44.13	−0.025	3.76	3469.65	3473.41	2.3
		p = 8	0.000	4.39	4723.74	4728.13	−562.31	45.87	−0.027	3.91	3489.18	3493.09	2.9
		p = 9	−0.001	4.40	4726.82	4731.22	−559.10	45.76	−0.029	4.04	3513.79	3517.83	3.2
		p = 10	−0.001	4.41	4736.34	4740.75	−548.28	44.47	−0.031	4.15	3542.75	3546.90	3.4
		p = 20	−0.002	4.59	4906.58	4911.17	−356.83	23.43	−0.038	5.03	3914.38	3919.41	1.8
		p = 50	−0.001	4.89	5193.76	5198.65	−58.67	12.76	−0.045	6.44	4649.65	4656.09	−3.6
		p = ∞	−0.001	5.25	5509.54	5514.79	206.60	63.62	−0.056	9.17	5818.33	5827.50	−8.

Now, the simulation results in Table 1 are affected not only by the first order term in the Taylor variance, but also by the higher order terms. However, these terms are small and Result 5.2 is clearly reflected in the simulation variances. Table 1 shows that the simulation variance attains a minimum for $\hat{Y}_{\text{GREG}}$ ($p = 2$). It then increases with p, but at a very gradual rate. There is a considerable variance reduction compared to Simulation 1, for most values of p. The simulation covariance AvCov is negligible compared to AvVar and small compared to AvVarZ. This is expected, because the Taylor covariance is exactly zero in this case.

The effect of the coefficients q_k: As in Simulation 1, the effect of the different coefficients q_k is negligible. The coefficients $q_k = a_k = N/n$ and $q_k = \mu_k \sim \text{Uniform}(0, 1)$ give nearly identical results for every value of p.

8. Concluding Discussion

We have argued in this article that the derivation of a calibration estimator does not have to start from a distance minimization argument. We have proposed instead a functional form approach. We have defined a family of calibration estimators $\hat{Y}_{\text{CALF}}$ by the parameters q_k and $\mathbf{z}_k$. The estimators in $\hat{Y}_{\text{CALF}}$ are all asymptotically design unbiased. We have noted that the generalized regression estimators $\hat{Y}_{\text{GREG}}$ are contained in $\hat{Y}_{\text{CALF}}$. We have looked at the difference $Z_s = \hat{Y}_{\text{CALF}} - \hat{Y}_{\text{GREG}}$ between any two estimators in these families and considered the Taylor variance of these statistics. We have derived a sufficient condition for the optimality one member of $\hat{Y}_{\text{GREG}}$ within $\hat{Y}_{\text{CALF}}$. For specific designs, we have showed how to determine the values of $\mathbf{x}_k^*$ and c_k^* for the optimal $\hat{Y}^*_{\text{GREG}}$.

The variance of $\hat{Y}_{\text{CALF}}$ depends on how well the auxiliary variables $\mathbf{x}_k$ explain the variability of y_k through the linear representation $y_k = \mathbf{x}_k'\mathbf{B} + E_k$, for $k = 1, \ldots, N$. We were led to make a distinction between saturated and unsaturated superpopulation models. The saturated models are in a sense less interesting, because in practice we cannot hope to be in a position to calibrate on all relevant x-variables. For a saturated model, there are essentially no differences between the calibration estimators. By contrast, the unsaturated models are more realistic because they allow for other explanatory variables than those used in the calibration.

The ideas in the article are illustrated by the results of Simulations 1 and 2. In both simulations, X is an important control total for the calibration. The population size N is also available and although it is not as useful by itself, it can be used as a second total. Should we calibrate only on X or on both N and X? The simulations gave a clear answer. We should calibrate on (N, X). In practice, we do not know if the model is saturated by the calibration variable x. If the model is saturated ($K = 0$), then it does not really matter whether we calibrate only on X or (N, X). We obtain similar estimators. If the model is unsaturated ($K \neq 0$), then Result 5.2 allows us to identify an optimal $\hat{Y}_{\text{GREG}}$ estimator in $\hat{Y}_{\text{CALF}}$ by calibrating on (N, X) and setting $c_k = 1$. In either case, we can calibrate on (N, X). When the model is unsaturated, we cannot identify an optimal estimator if only X is used. In general, our recommendation is to calibrate on as much available information as possible, and to try to use Result 5.2 to identify an optimal $\hat{Y}_{\text{GREG}}$. For certain designs, we showed how this objective is realized through suitable choices of the parameters c_k and $\mathbf{x}_k$.

9. References

Alexander, C.H. (1987). A Class of Methods for Using Person Controls in Household Weighting. Survey Methodology, 13, 183–198.

Bankier, M.D. (1992). Two-step Generalized Least Squares Estimation in the 1991 Canadian Census. Proceedings of the Survey Research Methods Section, American Statistical Association, 124–127.

Bethlehem, J.G. and Keller, W.J. (1987). Linear Weighting of Sample Survey Data. Journal of Official Statistics, 3, 141–153.

Deville, J.C. (1998). La correction de la non-réponse par calage ou par échantillonnage équilibré. Paper presented at the Congrès de l'ACFAS, Sherbrooke, Québec, May.

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., Särndal, C.E., and Sautory, O. (1993). Generalized Raking Procedures in Survey Sampling. Journal of the American Statistical Association, 88, 1013–1020.

Estevao, V. (1994). Calculation of G-Weights Under Calibration and Bound Constraints. Report, Statistics Canada.

Estevao, V., Hidiroglou, M.A., and Särndal, C.E. (1995). Methodological Principles for a Generalized Estimation System at Statistics Canada. Journal of Official Statistics, 11, 181–204.

Huang, E.G. and Fuller, W.A. (1978). Nonnegative Regression Estimation for Sample Survey Data. Proceedings of the Social Statistics Section, American Statistical Association, 300–305.

Isaki, C.T. and Fuller, W.A. (1981). Survey Design Under the Regression Superpopulation Model. Journal of the American Statistical Association, 77, 89–96.

Lemaître, G. and Dufour, J. (1987). An Integrated Method for Weighting Persons and Families. Survey Methodology, 13, 199–207.

Särndal, C.E. (1996). Efficient Estimators With Simple Variance in Unequal Probability Sampling. Journal of the American Statistical Association, 91, 1289–1300.

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

Singh, A.C. and Mohl, C.A. (1996). Understanding Calibration Estimators in Survey Sampling. Survey Methodology, 22, 107–115.

Stukel, D.M., Hidiroglou, M.A., and Särndal, C.E. (1996). Variance Estimation for Calibration Estimators: A Comparison of Jackknifing Versus Taylor Linearization. Survey Methodology, 22, 117–125.

Zieschang, K.D. (1990). Sample Weighting Methods and Estimation of Totals in the Consumer Expenditure Survey. Journal of the American Statistical Association, 85, 986–1001.

Received December 1998
Revised June 2000