MAXIMIZING THE USE OF AUXILIARY INFORMATION FOR CALIBRATION AND REGRESSION ESTIMATION

V. Estevao, Z. Patak, Statistics Canada
and C. E. Särndal, Université de Montréal
Carl Erik Särndal, Département de mathématiques et de statistique,
Université de Montréal, CP6128, succursale A, Montréal, H3C 3J7

Key Words: Domains, Auxiliary information, Calibration groups, Regression groups

1. Introduction

The Generalized Estimation System (GES) being developed at Statistics Canada produces domain estimates from sample surveys using auxiliary information. This may involve knowledge of the auxiliary variable values $\boldsymbol{x}_k$ for each unit k in the population U, but it is sufficient to know only the auxiliary population totals, $\boldsymbol{X} = \sum_U \boldsymbol{x}_k$.

We let s denote a probability sample drawn from the finite population $U=\{1,...,k,...,N\}$ according to a given sampling design with known, strictly positive inclusion probabilities. The variable of interest is given by y. Estimates are wanted for various parameters, including the y-total for the whole population, $Y = \sum_U y_k$ and the y-total $Y_d = \sum_{U_d} y_k$ for an arbitrarily specified subpopulation or domain $U_d \subseteq U$.

From the selected sample s, we have the observed survey data $\{(y_k, \boldsymbol{x}_k), k \in s\}$. In the current GES, the observed value y_k is given a total weight $w_k = a_k g_k$, calculated as the product of

(i) the sampling weight $a_k = 1/\pi_k$, where π_k is the inclusion probability of unit k, and
(ii) the g-weight g_k, calculated with the aid of the known vector total $\boldsymbol{X}$ as

$$g_k = 1 + \lambda' \boldsymbol{x}_k / c_k \qquad (1.1)$$

with

$$\lambda' = \left(\boldsymbol{X} - \hat{\boldsymbol{X}}_\pi\right)' \left(\sum_s a_k \boldsymbol{x}_k \boldsymbol{x}_k' / c_k\right)^{-1}$$

where c_k are specified constants and $\hat{\boldsymbol{X}}_\pi = \sum_s a_k \boldsymbol{x}_k$.

The GES produces estimates using the final weights w_k. The entire population y-total $Y = \sum_U y_k$ is estimated as

$$\hat{Y}_{GREG} = \sum_s w_k y_k = \sum_s a_k g_k y_k \qquad (1.2)$$

Furthermore, if $s_d = s \cap U_d$ denotes the part of the sample s falling in the specified domain U_d, the domain y-total $Y_d = \sum_{U_d} y_k$ is estimated as

$$\hat{Y}_{d\,GREG} = \sum_{s_d} w_k y_k = \sum_{s_d} a_k g_k y_k \qquad (1.3)$$

Under general conditions, $\hat{Y}_{GREG}$ and $\hat{Y}_{d\,GREG}$ are design consistent estimators and corresponding design consistent variance estimates are easily calculated following the theory of regression estimators as described in Estevao, Hidiroglou and Särndal (1994).

In survey practice, estimates are usually required for many different domains which may or may not overlap. Of frequent interest is the case where a set of domains forms a partition of U - the domains are mutually exclusive and exhaust U. If a set of domains $U_1, \ldots, U_D$ forms a partition of U, then the domain estimates defined by (1.3) satisfy $\sum_d \hat{Y}_{d\,GREG} = \hat{Y}_{GREG}$, where $\hat{Y}_{GREG}$ is given by (1.2) - the domain estimates add up to the estimate made for the entire population. This is a desirable property for most surveys. In the following presentation, we need only concentrate on the estimation of the total Y_d for a single domain U_d.

In general, the weights g_k in (1.2) and (1.3) can be derived by minimizing a measure of distance between the set of final weights $w_k = a_k g_k$ and the initial sampling weights a_k subject to the constraint

$$\sum_s a_k g_k \boldsymbol{x}_k = \boldsymbol{X} \qquad (1.4)$$

Weights derived in this manner are called calibrated weights by Deville and Särndal (1992) who examined several calibration methods corresponding to different distance measures. The weights given by (1.1) are calibrated since they satisfy (1.4). They are obtained by minimizing the distance

$$\sum_s c_k (w_k - a_k)^2 / a_k \qquad (1.5)$$

subject to (1.4). In practice, some of the weights w_k may be negative while others may be unduly large. To constrain the values of these weights, Estevao (1994) suggested a computationally efficient algorithm for minimizing (1.5) subject to (1.4) and bounds $l_k \le w_k \le u_k$ on the individual weights.

We are interested in looking beyond the current GES and exploring other estimators which use known auxiliary totals to improve estimation through calibration and regression fitting. This paper proposes an extension of the class of estimators defined by (1.2) and (1.3) and examines some properties of the estimators in this class.

2. An Extended Class of Estimators

Consider a domain U_d where $U_d \subseteq U$. We want to create a class of design consistent estimators of the domain total Y_d. Let us express Y_d as the sum of a prediction component and a residual component,

$$Y_d = Y_{d\,PRED} + Y_{d\,RES} \qquad (2.1)$$

Here, $Y_{d\,PRED} = \boldsymbol{Z}_d'\boldsymbol{B}_z$ with $\boldsymbol{Z}_d = \sum_{U_d} \boldsymbol{z}_k$ and the vector $\boldsymbol{B}_z$ is given by $\boldsymbol{B}_z = \left(\sum_U \boldsymbol{z}_k \boldsymbol{z}_k'/c_k^*\right)^{-1} \sum_U \boldsymbol{z}_k y_k/c_k^*$, where the c_k^* are specified constants. In addition, we have $Y_{d\,RES} = \sum_{U_d} E_k$ with $E_k = y_k - \boldsymbol{z}_k'\boldsymbol{B}_z$.

Now (2.1) has an obvious interpretation in terms of regression analysis. Suppose the whole population is observed so that the data $\{(y_k, \boldsymbol{z}_k), k \in U\}$ are available for fitting a census regression fit. This fit produces $\boldsymbol{B}_z$ and the corresponding predicted values $\hat{y}_k = \boldsymbol{z}_k'\boldsymbol{B}_z$ for $k \in U$. If y is predicted well by $\boldsymbol{z}$ in the domain U_d then $Y_{d\,PRED}$ is close to Y_d. Furthermore, the remaining error $Y_{d\,RES}$ is small relative to $Y_{d\,PRED}$.

Now in practice only units in the sample s are observed. This means that $\boldsymbol{Z}_d$, $\boldsymbol{B}_z$ and $Y_{d\,RES}$ are unknowns requiring estimation. Many articles on domain estimation assume that auxiliary information is available for the domain so that $\boldsymbol{Z}_d$ is known, but in practice there is usually no such information at the domain level. Hence $\boldsymbol{Z}_d$ is unknown and must be estimated. Since $Y_{d\,PRED}$ is close to Y_d under the regression model, the central issue is to accurately estimate $Y_{d\,PRED}$. To obtain maximum efficiency in the estimation of $Y_{d\,RES}$ is of secondary importance. However, we need to obtain a design consistent estimator of $Y_{d\,RES}$, otherwise the estimator of Y_d will be design biased.

We suppose that the data $\{(y_k, \boldsymbol{z}_k), k \in s\}$ are available for a regression fit. We estimate $\boldsymbol{Z}_d$ as $\hat{\boldsymbol{Z}}_d = \sum_{s_d} w_k \boldsymbol{z}_k$, where w_k are the weights from the calibration $\sum_s w_k \boldsymbol{x}_k = \boldsymbol{X} = \sum_U \boldsymbol{x}_k$ on the vector $\boldsymbol{x}_k$ as described in section 1. The regression vector $\boldsymbol{z}_k$ may be the same as the vector $\boldsymbol{x}_k$ but not necessarily. We consider a class of estimators of the domain total Y_d defined as

$$\hat{Y}_d = \hat{Y}_{d\,PRED} + \hat{Y}_{d\,RES} \qquad (2.2)$$

where

$$\hat{Y}_{d\,PRED} = \hat{\boldsymbol{Z}}_d'\hat{\boldsymbol{B}}_z$$

$$\hat{\boldsymbol{Z}}_d = \sum_{s_d} w_k \boldsymbol{z}_k$$

$$\hat{\boldsymbol{B}}_z = \left(\sum_s w_k^* \boldsymbol{z}_k \boldsymbol{z}_k'/c_k^*\right)^{-1} \sum_s w_k^* \boldsymbol{z}_k y_k/c_k^*$$

and

$$\hat{Y}_{d\,RES} = \sum_{s_d} w_k^0 \left(y_k - \boldsymbol{z}_k'\hat{\boldsymbol{B}}_z\right)$$

The weights w_k^0, w_k^* and c_k^* are defined in an appropriate manner. Some typical choices are discussed in section 3. We note that the two sums in $\hat{\boldsymbol{B}}_z$ extend over the entire sample s, whereas the sums in $\hat{\boldsymbol{Z}}_d$ and $\hat{Y}_{d\,RES}$ extend only over s_d, the domain part of the sample. Thus, the regression fit "borrows strength" from the whole sample.

The corresponding estimator of the entire population Y is defined by (2.2) by letting the sums in $\hat{\boldsymbol{Z}}_d$ and $\hat{Y}_{d\,RES}$ extend over all of s instead of just s_d. Thus the sum of the $\hat{Y}_d$ over a set of domains that partition U automatically adds up to the estimate made for the entire population. This satisfies the important requirement of additivity.

3. Computational Steps and Properties of the New Class

The computation of $\hat{Y}_d = \hat{Y}_{d\,PRED} + \hat{Y}_{d\,RES}$ defined by (2.2) can be viewed as a procedure involving three steps.

Step 1 The calibration step consists of the calculation of the weights w_k used in $\hat{\boldsymbol{Z}}_d$. They are of the form $w_k = a_k g_k$ where the g_k are given by (1.1) or more generally, by any calibration method satisfying (1.4). The calibration is based on the vector of known totals $\boldsymbol{X} = \sum_U \boldsymbol{x}_k$. In a survey, a vector of totals may be known at the entire population level or at some subpopulation level such as strata or post-strata. The subpopulations for which these totals are known are called calibration groups. They are assumed to form a partition of U. The associated variables are called calibration variables. Our notation $\boldsymbol{X} = \sum_U \boldsymbol{x}_k$ incorporates both the definition of the calibration

variables and the calibration groups. For example, suppose there are P calibration groups U_p, for $p=1,2,...,P$ and a single variable x. Then $\boldsymbol{x}_k$ is the vector given by $\boldsymbol{x}_k = (\delta_{1k}x_k,...,\delta_{pk}x_k,...,\delta_{Pk}x_k)'$ where δ_{pk} is the group identifier such that $\delta_{pk}=1$ if $k \in U_p$ and $\delta_{pk}=0$ if $k \notin U_p$. It follows that $\boldsymbol{X}$ is the corresponding vector of the P known group totals of the calibration variable x. More generally, we can have different sets of calibration variables in the different groups. For example, we can specify x_1 and x_2 in the first group, x_1 and x_3 in the second group and so on, as long as the corresponding totals are known for the variables in each group. In this case $\boldsymbol{x}_k$ is given by $\boldsymbol{x}_k = (\delta_{1k}(x_{1k},x_{2k})',\delta_{2k}(x_{1k},x_{3k})'...)'$.

Step 2 In the regression fit step, we calculate $\hat{\boldsymbol{B}}_z$ using the predictor vector $\boldsymbol{z}_k$. The specification of regression groups and predictors is reflected in the definition of $\boldsymbol{z}_k$. When there is more than one regression group, this means a separate regression fit in each group. Different sets of predictors could be used in the different groups. For example, with three groups we can define $\boldsymbol{z}_k$ as $\boldsymbol{z}_k = (\delta_{1k}\boldsymbol{z}_{1k}',\delta_{2k}\boldsymbol{z}_{2k}',\delta_{3k}\boldsymbol{z}_{3k}')'$ where δ_{rk} is the regression group identifier for unit k and $\boldsymbol{z}_1 = (\boldsymbol{z}_1,\boldsymbol{z}_3,\boldsymbol{z}_4)'$, $\boldsymbol{z}_2 = (\boldsymbol{z}_1,\boldsymbol{z}_2)'$ and $\boldsymbol{z}_3 = \boldsymbol{z}_1$.

Also requiring specification in this step are the weights w_k^* and c_k^*. A standard choice is $w_k^* = a_k$ and $c_k^* = 1$ for all $k \in s$. Note that $\boldsymbol{z}_k$ need not be identical to the calibration vector $\boldsymbol{x}_k$ in step 1. However, $\boldsymbol{z}_k = \boldsymbol{x}_k$ is a special case of interest. In this case, the regression groups coincide with the calibration groups and in each group, the predictor variables are also the calibration variables.

Step 3 The residual estimation step consists of calculation of $Y_{d\,RES}$. Since $\hat{\boldsymbol{B}}_z$ is obtained from the regression fit in step 2, the only additional specification is the set of weights w_k^0. Typical choices include $w_k^0 = a_k$ and $w_k^0 = w_k = a_k g_k$.

We now illustrate estimator (2.2) by a series of remarks in this section and by examples in sections 4 and 5.

Remark 3.1 We can write $\hat{Y}_d = \sum_{s_d} w_k y_k + R_d$ where $R_d = \sum_{s_d}(w_k^0 - w_k)(y_k - \boldsymbol{z}_k'\hat{\boldsymbol{B}}_z)$. Here we note two things.

(1) If we choose $w_k^0 = w_k$ then $R_d = 0$ and $\hat{Y}_d = \sum_{s_d} w_k y_k = \sum_{s_d} a_k g_k y_k$, which is the estimator used in the current GES assuming g_k is given by (1.1). For this choice of w_k^0, the regression step is superfluous.

(2) If $w_k^0 \neq w_k$ then under general conditions $N^{-1}\left\{\sum_{s_d}(w_k^0 - w_k)(y_k - \boldsymbol{z}_k'\hat{\boldsymbol{B}}_z)\right\}$ is of order $O_p(n^{-1/2})$ which is the same order as $N^{-1}\left\{\sum_{s_d} w_k y_k - Y_d\right\}$. So although R_d is an estimator of 0, it cannot be ignored on the grounds that it converges more rapidly than the first term, $\sum_{s_d} w_k y_k$. However, the variance contribution of R_d is in many situations modest or insignificant compared to that of $\sum_{s_d} w_k y_k$.

Remark 3.2 Suppose that the domain of interest U_d coincides with a calibration group and $\boldsymbol{z}_k = \boldsymbol{x}_k$. Then $\boldsymbol{X}_d = \sum_{U_d} \boldsymbol{x}_k$ is a known total and by calibration $\hat{\boldsymbol{Z}}_d = \sum_{s_d} w_k \boldsymbol{x}_k = \boldsymbol{X}_d$. In this case (2.2) becomes

$$\hat{Y}_d = \boldsymbol{X}_d'\hat{\boldsymbol{B}}_x + \sum_{s_d} w_k^0 (y_k - \boldsymbol{x}_k'\hat{\boldsymbol{B}}_x)$$

The particular choice $w_k^* = w_k^0 = a_k$ leads to estimators discussed in Chapter 10 of Särndal, Swensson and Wretman (1992).

Remark 3.3 Suppose that the domain of interest U_d coincides with a calibration group. Then $\boldsymbol{X}_d = \sum_{U_d} \boldsymbol{x}_k$ is a known quantity and the weights $w_k = a_k g_k$ are calibrated to satisfy $\sum_{s_d} w_k \boldsymbol{x}_k = \boldsymbol{X}_d$. In particular, if x denotes a single calibration variable in the vector $\boldsymbol{x}$, then $\sum_{s_d} w_k x_k = X_d$. Now suppose we use (2.2) to estimate the corresponding domain total X_d of variable x. We hope to obtain $\hat{X}_d = X_d$ since the estimate should be the same as the known value. In fact (2.2) has this property provided that x_k is contained in the predictor vector $\boldsymbol{z}_k$. To show this, we put $y_k = x_k$ in (2.2). Then the regression fit gives $\hat{\boldsymbol{B}}_z = (0,...,1,...,0)'$, a vector in which all entries are 0 except for a single entry of 1 in the position that x occupies in the vector $\boldsymbol{z}$. This gives $y_k - \boldsymbol{z}_k'\hat{\boldsymbol{B}}_z = x_k - x_k = 0$ for all $k \in s_d$. Therefore, using (2.2) to estimate X_d, we get the result $\hat{X}_d = \sum_{s_d} a_k g_k \boldsymbol{z}_k'\hat{\boldsymbol{B}}_z = \sum_{s_d} a_k g_k x_k = X_d$, as we wanted

to show. Thus, if U_d is a calibration group, estimator (2.2) reproduces the known total of any variable used both as a calibration variable and as a regression predictor. This holds regardless of how the weights w_k^0, w_k^* and c_k^* are specified.

To ensure that (2.2) reproduces all of the known calibration totals, we must include all of the calibration variables as regression predictors in the definition of z_k. Of course, we can include additional variables.

Remark 3.4 The estimator of the entire population total Y can be obtained by setting $s_d = s$ in (2.2). This estimator can be written as

$$\hat{Y} = \sum_s w_k^0 y_k + \hat{\boldsymbol{B}}_z' \sum_s \left(w_k - w_k^0\right) z_k$$

Now let us make the following two assumptions:

(i) Suppose that $\left\{w_k^0, k \in s\right\}$ are calibrated weights calculated based on the entire population as a single calibration group. We have $\sum_s w_k^0 z_k = \boldsymbol{Z}$, where $\boldsymbol{Z} = \sum_U z_k$ is the vector of known totals for the whole population.

(ii) Suppose that $\left\{w_k, k \in s\right\}$ are calibrated to z-totals at a lower level of the population given by the partition, $U_1, ..., U_p, ..., U_P$. For each subpopulation U_p, we have $\sum_{s_p} w_k z_k = \boldsymbol{Z}_p$ where $\boldsymbol{Z}_p = \sum_{U_p} z_k$ is a vector of known totals for the subpopulation and $s_p = s \cap U_p$.

In view of assumptions (i) and (ii), it follows that $\sum_s \left(w_k - w_k^0\right) z_k = 0$ and $\hat{Y} = \sum_s w_k^0 y_k$. The practical significance of this is as follows. Suppose for timeliness or other administrative reasons, it was decided to release $\hat{Y} = \sum_s w_k^0 y_k$ as the estimate for the entire population. Later, it is possible to produce "on demand" domain estimates given by (2.2) which automatically benchmark to the already released estimate $\hat{Y} = \sum_s w_k^0 y_k$ for the entire population. In addition, these domain estimates are strengthened by using the more detailed auxiliary information $\boldsymbol{Z}_p$.

4. Domains with Known Auxiliary Totals

Domains of interest in a survey often represent different "levels" of the population. For example, in a business survey, estimates may be wanted for domains corresponding to industry groups at different levels of the Standard Industrial Classification - two, three or four digit level (SIC2, SIC3, SIC4) groups. At lower levels, we have fewer sample units. This makes it more important to use auxiliary information to the fullest extent possible. At higher levels, there is usually sufficient data to guarantee good precision.

Suppose x is a positive auxiliary variable related to the study variable y by a strong linear relationship roughly through the origin. Then it is appropriate to use x as an auxiliary variable through some form of ratio estimation. The example in this section assumes that auxiliary totals are available directly for the domains of interest. Section 5 gives examples where this is not the case, corresponding better to what is usually found in practice.

We assume a probability sample s from U from which we obtain the data $\left\{(y_k, x_k), k \in s\right\}$. Let us consider the estimation of the y-total for domain U_0, where $U_0 \subseteq U$. The part of the sample falling in this domain is given by $s_0 = s \cap U_0$. We examine the following cases, depending on the level of availability of the auxiliary totals.

Case 1 Suppose the auxiliary total $X_0 = \sum_{U_0} x_k$ is known for the domain. We can compute the g-weights $g_k = X_0 / \hat{X}_{0\pi}$ for $k \in s_0$ by calibrating at the domain level so $\sum_{s_0} a_k g_k x_k = X_0$. Using these in (1.3) allows us to produce an estimate for the domain total $Y_0 = \sum_{U_0} y_k$ as the ratio estimator

$$\hat{Y}_0 = \left(X_0 / \hat{X}_{0\pi}\right) \hat{Y}_{0\pi}$$

Now suppose we need to produce estimates for a set of subdomains U_{0d} for $d = 1, 2, ... D$, which partition the domain U_0. These estimates may be part of the regular survey requirements or they may be special requests at a later date. If we do not have auxiliary information for the subdomains, we use the previous g-weights to estimate the subdomain total $Y_{0d} = \sum_{U_{0d}} y_k$ as

$$\hat{Y}_{0d} = \left(X_0 / \hat{X}_{0\pi}\right) \hat{Y}_{0d\pi}$$

Although these weights yield estimates of good precision for U_0, they may give estimates of rather poor precision for the subdomains U_{0d}. Here, $\hat{Y}_{0d}$ suffers from a lack of detailed auxiliary information since the weights $g_k = X_0 / \hat{X}_{0\pi}$ are calibrated on an x-total known only at a level above the subdomain of interest. The strength of the auxiliary information is reduced when estimates are made for subdomains. However, using the auxiliary information X_0 for calibration is better than having nothing at all since

$\hat{Y}_{0d} = \left(X_0 / \hat{X}_{0\pi}\right)\hat{Y}_{0d\pi}$ is normally a more precise estimator than the ordinary Horvitz-Thompson estimator $\hat{Y}_{0d\pi}$.

Let us now consider case 2 where more detailed auxiliary information is available through known x-totals for the subdomains U_{0d}. In this case, we can use the subdomains as calibration groups, x is the calibration variable and the group totals $X_{0d} = \sum_{U_{0d}} x_k$ are known for $d = 1, 2, ...D$. This information produces the weights $w_k = a_k g_k$ with $g_k = X_{0d} / \hat{X}_{0d\pi}$ for $k \in s_{0d} = s \cap U_{0d}$. Let us look at the form of estimators $\hat{Y}_0$ and $\hat{Y}_{0d}$ for specific types of regression groups and choices of w_k^0. We examine cases 2a, 2b and 2c shown below, assuming throughout that $w_k^* = a_k$ and $c_k^* = c_k = x_k$.

Case 2a We define $w_k^0 = w_k$. In this case, the regression fit is superfluous as noted in Remark 3.1.

Case 2b We define $w_k^0 = a_k X_0 / \hat{X}_{0\pi}$. Let us consider the subdomains as the regression groups and x as the regression predictor in each group. The fitted slope for U_{0d} is $\hat{B}_{0d} = \hat{Y}_{0d\pi} / \hat{X}_{0d\pi}$.

Case 2c We again define $w_k^0 = a_k X_0 / \hat{X}_{0\pi}$. Let us consider the domain U_0 as the only regression group and x as the regression predictor. The fitted slope in U_0 is $\hat{B}_0 = \hat{Y}_{0\pi} / \hat{X}_{0\pi}$.

Using (2.2) we obtain the following design consistent estimators of Y_0 and Y_{0d}. Interestingly, cases 2a and 2b lead to the same estimators for both Y_0 and Y_{0d}.

Case	Level	Estimator
1	domain	$\hat{Y}_0^{(1)} = \left(X_0 / \hat{X}_{0\pi}\right)\hat{Y}_{0\pi}$
	subdomain	$\hat{Y}_{0d}^{(1)} = \left(X_0 / \hat{X}_{0\pi}\right)\hat{Y}_{0d\pi}$
2a=2b	domain	$\hat{Y}_0^{(2a)} = \sum_{d=1}^{D}\left(X_{0d} / \hat{X}_{0d\pi}\right)\hat{Y}_{0d\pi}$
	subdomain	$\hat{Y}_{0d}^{(2a)} = \left(X_{0d} / \hat{X}_{0d\pi}\right)\hat{Y}_{0d\pi}$
2c	domain	$\hat{Y}_0^{(2c)} = \left(X_0 / \hat{X}_{0\pi}\right)\hat{Y}_{0\pi}$
	subdomain	$\hat{Y}_{0d}^{(2c)} = X_{0d}\hat{B}_0 + \left(X_0 / \hat{X}_{0\pi}\right)\left(\hat{Y}_{0d\pi} - \hat{X}_{0d\pi}\hat{B}_0\right)$

The five different ratio estimators shown in the table prompt the following comments.

Remark 4.1 For the domain U_0, we do not expect a great difference in the variances of $\hat{Y}_0^{(1)} = \hat{Y}_0^{(2c)}$ and $\hat{Y}_0^{(2a)}$ unless the subdomain slopes Y_{0d}/X_{0d} differ greatly. In most practical situations, the principal gain in precision is realized because x is a strong overall covariate, not because the slopes differ in the subdomains.

Remark 4.2 Letting $W_{0d} = X_{0d} - \left(X_0 / \hat{X}_{0\pi}\right)\hat{X}_{0d\pi}$, we can write the subdomain estimators $\hat{Y}_{0d}^{(2a)}$ and $\hat{Y}_{0d}^{(2c)}$ as $\hat{Y}_{0d}^{(2a)} = \hat{Y}_{0d}^{(1)} + \hat{B}_{0d} W_{0d}$ and $\hat{Y}_{0d}^{(2c)} = \hat{Y}_{0d}^{(1)} + \hat{B}_0 W_{0d}$. In other words, we can represent $\hat{Y}_{0d}^{(2a)}$ and $\hat{Y}_{0d}^{(2c)}$ as the simple estimator $\hat{Y}_{0d}^{(1)}$ plus either the regression adjustment $\hat{B}_{0d} W_d$ or $\hat{B}_0 W_d$. The slight difference between $\hat{Y}_{0d}^{(2a)}$ and $\hat{Y}_{0d}^{(2c)}$ lies in the slope estimates applied to the regression adjustment. Using $\hat{B}_0$ rather than $\hat{B}_{0d}$ has little impact on the variance, so there will be little to choose between $\hat{Y}_{0d}^{(2a)}$ and $\hat{Y}_{0d}^{(2c)}$. By contrast, both $\hat{Y}_{0d}^{(2a)}$ and $\hat{Y}_{0d}^{(2c)}$ may improve significantly on $\hat{Y}_{0d}^{(1)}$ because of regression adjustments correlated negatively with $\hat{Y}_{0d}^{(1)}$. This illustrates that precision at a lower level can be improved considerably by having access to auxiliary totals at the more detailed level.

Remark 4.3 Suppose we have already released an estimate for domain U_0 as $\hat{Y}_0^{(1)} = \left(X_0 / \hat{X}_{0\pi}\right)\hat{Y}_{0\pi}$. We now want to use auxiliary information for the subdomains to provide subdomain estimates that add up to the released domain estimate. The subdomain estimates $\hat{Y}_{0d}^{(2c)}$ satisfy this requirement. They are automatically benchmarked to agree with the original domain estimate since $\hat{Y}_0^{(2c)} = \hat{Y}_0^{(1)} = \left(X_0 / \hat{X}_{0\pi}\right)\hat{Y}_{0\pi}$. This is an illustration of remark 3.4.

5. Auxiliary Information for Groups other than Domains

In many surveys, the auxiliary information extracted from the frame or an administrative source, does not relate to the domains of interest but to other similar subpopulations. To illustrate, it frequently happens in business surveys that the Standard Industrial Classification (SIC) code recorded on the sampling frame is outdated for some business establishment because of coding errors or change of business activity. The current or actual SIC code of an establishment may be different from the code recorded on the frame. The

codes on the frame can only be updated for the units in the current sample, after a review of the information reported by these units.

The codes on the frame define the administrative SIC groups. Auxiliary totals are often available for the population in these groups. If so, we can define these as calibration groups. On the other hand, we have the actual SIC groups identified by the updated code observed only for the sampled units. These groups are the domains of interest. Often in practice, there are no auxiliary totals available for these domains.

To increase the precision of the domain estimates, it is important to use the auxiliary totals known for the administrative SIC groups. This is particularly critical for small SIC groups which may contain few sampled units. Although the domains do not coincide with the administrative groups, we can use (2.2) to produce design consistent estimates that maximize the use of auxiliary information.

For example, suppose the domains are defined at the four digit level of the Standard Industrial Classification (SIC4 groups). Let d indicate the current or actual SIC4 group or domain for $d=1,2,...,D$ and let j indicate the administrative SIC4 group for $j=1,2,...,J$ where $J=D$. That is, j denotes the administrative or old SIC4 code and d the actual or new SIC4 code. The population U can be viewed as being partitioned into administrative groups U_j, $j=1,2,...,J$ for which auxiliary information is available and into domains of interest U_d, $d=1,2,...,D$ for which there is no auxiliary information. Although both groupings are at the SIC4 level, the U_j and U_d represent different partitions of the units in U.

Let $j = j(d)$ be the old code for a sampling unit with new code d. Typically, old and new codes agree for a majority of units. For example, 80% of the units in domain U_d may be also in the corresponding administrative group $U_{j(d)}$ and the remaining 20% may be part of other administrative groups.

To illustrate our estimator for this type of problem, suppose x is a scalar positive auxiliary variable with a strong linear relation to y roughly through the origin. Let us consider two simple cases. In case 1 below, auxiliary information is only available for the entire population. For domain estimation, this is not as strong as having auxiliary information at a lower level. To improve on this, case 2 assumes information available is available only for the administrative domains.

Case 1 The only known auxiliary total is $X = \sum_U x_k$. With $c_k = x_k$ in (1.1) the g-weights are $g_k = X/\hat{X}_\pi$ for all k. To estimate the total Y of the whole population of establishments, these weights give the classical ratio estimator

$$\hat{Y}^{(1)} = \left(X/\hat{X}_\pi\right)\hat{Y}_\pi$$

At this level, we can count on high precision because of the strong linear relationship between y and x. For the domain total Y_d, the same weights lead to the estimator

$$\hat{Y}_d^{(1)} = \left(X/\hat{X}_\pi\right)\hat{Y}_{d\pi}$$

as the estimator of the total Y_d of the domain. Its precision is often inadequate.

Case 2 To improve on the domain estimates, suppose we use the administrative groups to define J calibration groups U_j with known totals $X_j = \sum_{U_j} x_k$, $j=1,2,...,J$. In (1.1), we define $\boldsymbol{x}_k = \left(\delta_{1k}\, x_k, .., \delta_{jk}\, x_k, .., \delta_{Jk}\, x_k\right)'$ where δ_{jk} is the calibration group identifier. With $c_k = x_k$, the calibration produces the weights $w_k = a_k\, g_k$ with $g_k = X_j/\hat{X}_{j\pi}$ for all $k \in s_j$ where $s_j = s \cap U_j$. The part of the sample s falling in domain U_d is given by $s_d = s \cap U_d$. Now s_d is further subdivided into the sample cells $s_{dj} = s_d \cap U_j$. For cell s_{dj}, we use the notation $\hat{Y}_{dj\pi} = \sum_{s_{dj}} a_k y_k$ and $\hat{X}_{dj\pi} = \sum_{s_{dj}} a_k x_k$. Now different design consistent estimators arise from (2.2) depending on how we specify the regression groups and w_k^0. We assume throughout that $w_k^* = a_k$ and $c_k^* = c_k = x_k$.

Case 2a Let $w_k^0 = w_k$. Then the regression fit is superfluous. The estimator of the domain total Y_d obtained from (2.2) is

$$\hat{Y}_d^{(2a)} = \sum_{j=1}^{J}\left(X_j/\hat{X}_{j\pi}\right)\hat{Y}_{dj\pi}$$

and the corresponding estimator of the entire population total Y is

$$\hat{Y}^{(2a)} = \sum_{j=1}^{J}\left(X_j/\hat{X}_{j\pi}\right)\hat{Y}_{j\pi}$$

which is usually described in the literature as a post-stratified ratio estimator, the calibration groups being viewed as the post-strata.

Case 2b Let $w_k^0 = a_k\, X/\hat{X}_\pi$. Let each domain U_d be a regression group. The fitted slope for U_d is $\hat{B}_d = \hat{Y}_{d\pi}/\hat{X}_{d\pi}$. From (2.2), Y_d is estimated as

$$\hat{Y}_d^{(2b)} = \left\{\sum_{j=1}^{J} \left(X_j / \hat{X}_{j\pi}\right)\hat{X}_{dj\pi}\right\}\hat{B}_d \qquad (5.1)$$

The corresponding estimator of the entire population total Y is

$$\hat{Y}^{(2b)} = \sum_{d=1}^{D}\left\{\sum_{j=1}^{J} \left(X_j / \hat{X}_{j\pi}\right)\hat{X}_{dj\pi}\right\}\hat{B}_d$$

Case 2c Let $w_k^0 = a_k\left(X/\hat{X}_\pi\right)$. Let there be a single regression group equal to the entire population U. We get the pooled slope estimate $\hat{B} = \hat{Y}_\pi / \hat{X}_\pi$. Then from (2.2), Y_d is estimated by

$$\hat{Y}_d^{(2c)} = \left\{\sum_{j=1}^{J} \left(X_j / \hat{X}_{j\pi}\right)\hat{X}_{dj\pi}\right\}\hat{B} + \left(X/\hat{X}_\pi\right)\left(\hat{Y}_{d\pi} - \hat{B}\,\hat{X}_{d\pi}\right)$$

and the corresponding estimator of the entire population total Y is the classical ratio estimator

$$\hat{Y}^{(2c)} = \left(X/\hat{X}_\pi\right)\hat{Y}_\pi$$

Other possibilities exist for defining the regression groups. One would be to use each of the calibration groups U_j as a regression group. We make some observations on these results.

Remark 5.1 Since the calibration in case 2 takes place at the level of the domains of interest, all of $\hat{Y}_d^{(2a)}$, $\hat{Y}_d^{(2b)}$ and $\hat{Y}_d^{(2c)}$ should give considerably improved precision compared to $\hat{Y}_d^{(1)}$ for which calibration is at a higher level. An interesting feature of $\hat{Y}_d^{(2c)}$ is that although it capitalizes on the detailed information X_j, the corresponding entire population estimate $\hat{Y}^{(2c)}$ depends neither on the calibration groups nor on the domains of interest. By contrast, $Y^{(2b)}$ depends on the particular set of domains U_d, $d=1,2,...,D$.

Remark 5.2 One expects only modest differences in precision between $\hat{Y}_d^{(2a)}$, $\hat{Y}_d^{(2b)}$ and $\hat{Y}_d^{(2c)}$ because in all three cases calibration is done at the same level.

Remark 5.3 Note that case 1 and case 2c estimators agree at the population level so that $\hat{Y}^{(2c)} = \hat{Y}^{(1)} = \left(X/\hat{X}_\pi\right)\hat{Y}_\pi$, although they disagree at the domain level, where $\hat{Y}_d^{(2c)}$ is usually a considerable improvement on $\hat{Y}_d^{(1)} = \left(X/\hat{X}_\pi\right)\hat{Y}_{d\pi}$.

Remark 5.4 There are alternatives to $\hat{Y}_d^{(2a)}$, $\hat{Y}_d^{(2b)}$ and $\hat{Y}_d^{(2c)}$ that a survey analyst may be lead to consider.

(i) The direct approach is to use the Horvitz-Thompson estimator for the domain U_d,

$$\hat{Y}_{d\,HT} = \sum_{s_d} a_k y_k = \hat{Y}_{d\pi}$$

This is a simple unbiased approach but not very resourceful. It does not exploit the considerable information given by the known totals for the groups U_j.

(ii) The analyst who senses that the information for the groups is important might reason as follows. Let us identify the administrative SIC4 group $j(d)$ corresponding to domain d, and let us use the known total for that group, $X_{j(d)} = \sum_{j(d)} x_k$, to form the regression estimator

$$\hat{Y}_d = X_{j(d)}\hat{B}_d + \sum_{s_d} w_k^0\left(y_k - x_k\hat{B}_d\right)$$

where $\hat{B}_d$ is as in case 2b. The motivation here is to use $X_{j(d)}$ as a proxy for X_d if one is willing to ignore the fact that group $U_{j(d)}$ does not exactly coincide with the domain of interest U_d. For example, with $w_k^0 = a_k$, this leads to

$$\hat{Y}_d = X_{j(d)}\hat{B}_d \qquad (5.2)$$

which is a synthetic estimator. However, it is a biased estimator of Y_d because $X_{j(d)}$ and X_d differ by an unknown amount which may be quite large. In most situations, the administrative group total $X_{j(d)}$ cannot be taken as a proxy for the domain total X_d. If we compare the mean square error of (5.1) and (5.2), the reduction in variance realized by (5.2) is more than offset by the increase in the square of the bias resulting from the difference between $X_{j(d)}$ and X_d.

References

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

Estevao, V. (1994). Notes on Calculation of G-Weights. Report, Statistics Canada.

Estevao, V., Hidiroglou, M. and Särndal, C.E. (1993). Methodological Principles for a Generalized Estimation System at Statistics Canada. To appear, Journal of Official Statistics.

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