Journal of Official Statistics, Vol. 11. No. 2, 1995, pp. 181–204

Methodological Principles for a Generalized Estimation System at Statistics Canada

V. Estevao,[1] *M.A. Hidiroglou,*[1] *and C.E. Särndal*[2]

In this paper we present the methodological principles behind the development of the Generalized Estimation System (GES) at Statistics Canada. The GES allows the specification of an estimator from a wide group of estimators produced under a **g**eneral linear **reg**ression model. The resulting GREG estimators are characterized in the paper via three important concepts: model level, model groups and the model type. Familiar estimators such as the simple expansion, post-stratified and raking ratio estimators can be classified according to these concepts. But more generally, these concepts help to structure a wide class of possible estimators. The specification of a GREG model depends on the available auxiliary totals. This information is used to produce a set of g-factors to adjust the sample design weights. The resulting final weights have the property of producing estimates for the auxiliary variables which are equal to the known auxiliary totals. This consistency condition is appealing to most survey practitioners. Furthermore, efficient estimates are produced when the variable of interest is highly correlated with the auxiliary variables. The GES produces domain estimates for parameters such as domain size, totals, ratios of totals and means. This is done based on the sample design and the specified GREG model. We have shown that it is possible to extend the theory for the estimation of a total to handle any non-linear parameter. This is done through the usual Taylor approximation. Variance estimation is based on a formula suggested by Särndal, Swensson, and Wretman which incorporates both the g-weights and the residuals under the specified model.

Key words: Generalized regression estimator; auxiliary information; model groups; model level; model type.

1. Introduction

This paper presents an overview of the methodology used to construct Statistics Canada's Generalized Estimation System (GES). The idea of developing a GES has been present at Statistics Canada for several years. Documents that reflect the evolution of this idea include Choudhry (1988), Dumais and Carpenter (1988), Outrata and Chinnappa (1989), Särndal (1990), Lavallée and Leblond (1990), Hidiroglou (1991), and Estevao (1991). Software of a general purpose character, developed outside Statistics Canada, include LINWEIGHT (Bethlehem and Keller 1987), PC-CARP (Schnell, Kennedy, Sullivan, Park, and Fuller 1988), SUDAAN (Shah, Lavange, Barnwell, Killinger, and Wheeless 1989) and CLAN (Andersson

[1] Statistics Canada, Tunney's Pasture, Ottawa, Ontario K1A OT6, Canada.
[2] Université de Montréal, Département de mathématiques et de statistique, CP 6128, Succursale A, Montréal, Québec H3C 3J7, Canada.

and Nordberg 1994). The GES, and future extensions of it, produce domain point estimates and corresponding estimates of variance for parameters of the sampled finite population. This process considers the sampling design as well as any available auxiliary information.

In the current Version 3.1 of the GES, the parameters of interest are totals, ratios of totals, averages and proportions. The existing sampling designs include: (i) single-stage designs such as stratified simple random sampling with and without replacement (SRSWR and SRSWOR), and (ii) stratified cluster sampling and stratified probability proportional-to-size (PPS) sampling.

An important aspect of the GES is the use of auxiliary information in the form of known auxiliary variable totals. For single-stage element sampling, the known totals always refer to the population of elements, or to specified subgroups of this population. For single-stage cluster sampling and for sampling in two or more stages, auxiliary information can appear at different levels. These levels correspond to the different populations that can be distinguished. For example, in single-stage cluster sampling, we may have: (i) known auxiliary totals for the population of clusters (or for subgroups of it), or (ii) known auxiliary totals for the population of elements (or for subgroups of it).

A population subgroup with a known auxiliary total is called a *model group*. For each model group, a general linear regression model can be stated. The variable of interest is chosen as the dependent variable and the auxiliary variables as predictors. A special case is when the entire population defines the only existing model group. In the GES, the term GREG model refers to the collection of regressions fitted in the various groups.

The fit of the GREG model produces a generalized regression (GREG) estimator of the parameter in question. Simple special cases of the GREG estimator include the expansion, ratio and simple regression estimators. A more complex special case is the raking ratio estimator.

A principal function of the GES is to produce parameter estimates for one or more domains of interest. Common domains of interest are the individual strata and the entire population. But arbitrary subpopulations can be designated as domains. The domains may cut across the design strata and may be different from the model groups.

The sampling design and the GREG model are important components in the structure and function of a generalized estimation system. The specification of the sampling design should include: (i) the number of stages of selection, (ii) the sampling procedure at each stage, (iii) sample counts for each stage and (iv) the ultimate unit of selection. The GREG model requires: (i) identification of model group membership for each sampled unit, (ii) the corresponding model group totals at the population level and (iii) the identification of a model level. The model level specifies the population for which the auxiliary information is known. After the sampling design and GREG model have been specified, the GES is set up to produce parameter estimates for any domain of interest. This is reflected by three main modules in the GES. These are: (i) determination of the sampling weights, (ii) determination of the estimator factors (also called the g-factors), and (iii) calculation of point estimates and corresponding estimates of variance for the specified domains and parameters.

The sampling weights may be either provided directly or generated from the sampling design inputs. The g-factors, which may be read in directly, are calculated as a function of the auxiliary information and the sampling weights. The GES uses the sampling weights and the g-factors to produce the parameter estimates and corresponding estimates of precision. This is done for each parameter and domain specification.

The GES allows any number of model groups. An important requirement of the GES is that the model groups form a mutually exclusive and exhaustive partition of the population identified by the model level. This is required to generate a single g-factor for every unit and to allow estimation for any domain. A GREG estimator can be formed for the total of a given variable within any given model group. The GREG estimator of the entire survey population total is simply the sum of the estimators for the various model group totals.

This paper is organized as follows. Section 2 reviews the basic theory of the GREG estimator for a population total. Three important concepts used to characterize the GREG model, namely, model groups, model level and model type are discussed in Sections 3 and 4. Poststrata are shown to be an example of model groups but, as explained in Section 4, the concept has a much broader implication in the GES. Domain estimation is described in Section 5. The significance of the model level is made clear in Sections 6 and 7 for single-stage cluster designs and multistage designs, respectively. Section 8 extends the theory of the previous sections to the estimation of non-linear parameters. Throughout the paper, several examples are given to illustrate the principles used in the GES.

2. The Generalized Regression Estimator for Single-Stage Sampling Designs

Consider a finite population of elements $U = \{1, \ldots, k, \ldots, N\}$. A single-stage element sampling design is used to obtain a sample s from U. Let $p(s)$ denote the probability that s is realized. Let $\pi_k = \sum_{s \ni k} p(s)$ and $\pi_{k\ell} = \sum_{s \supset \{k,\ell\}} p(s)$ denote the inclusion probabilities for k and $\{k, \ell\}$ respectively. Note that when $\ell = k, \pi_{kk} = \pi_k$. The variable of interest is denoted by y and its value for the kth element is y_k. The objective is to estimate the population total $Y = \Sigma_U y_k$, with the aid of auxiliary information. Let $\mathbf{x} = (x_{1k}, \ldots, x_{jk}, \ldots, x_{Jk})'$ denote the value for the kth element of a J-dimensional auxiliary vector, $\mathbf{x}$. Suppose that data $(y_k, \mathbf{x}_k)$ are observed for each element $k \in s$. The population auxiliary totals present in the vector $\mathbf{X} = \Sigma_U \mathbf{x}_k = (\Sigma_U x_{1k}, \ldots, \Sigma_U x_{jk}, \ldots \Sigma_U x_{Jk})'$ are assumed to be known from one or more sources such as administrative registers or a census. We seek an estimator of Y that makes efficient use of this auxiliary information. An estimator that meets this objective is the generalized regression estimator (GREG)

$$\hat{Y}_{\text{GREG}} = \mathbf{X}'\hat{\mathbf{B}} + \Sigma_s a_k (y_k - \mathbf{x}_k'\hat{\mathbf{B}}) \tag{2.1}$$

where the vector $\hat{\mathbf{B}}$ is the solution of the sample based normal equations

$$\left(\Sigma_s a_k \mathbf{x}_k \mathbf{x}_k'/c_k\right)\hat{\mathbf{B}} = \Sigma_s a_k \mathbf{x}_k y_k / c_k \tag{2.2}$$

and c_k is defined in relation to the variance structure of the linear regression model

associated with the GREG estimator. This model, denoted by ξ, states that

$$y_k = \mathbf{x}_k'\boldsymbol{\beta} + \varepsilon_k \quad \text{for} \quad k \in U \tag{2.3}$$

where $E_\xi(\varepsilon_k) = 0$, $\text{Var}_\xi(\varepsilon_k) = c_k\sigma^2$ and $\text{Cov}_\xi(\varepsilon_k, \varepsilon_\ell) = 0$ for all $k \neq \ell$.

The GREG estimator provides a suitable basis for the development of the GES because: (i) most standard estimators that use auxiliary information can be obtained as special cases, (ii) the GREG estimator can be applied to any sampling design, and (iii) it can be used with any set of auxiliary variables for which the associated vector of totals, $\mathbf{X} = \sum_U \mathbf{x}_k$, is known. The GREG estimator (2.1) is linked to the model (2.3) in the following way: If the data $(y_k, \mathbf{x}_k)$ were observed for all N elements $k \in U$, a generalized least squares fit of the regression of y on x would require solving the census fit normal equations

$$(\Sigma_U \mathbf{x}_k \mathbf{x}'_k/c_k)\mathbf{B} = \Sigma_U \mathbf{x}_k y_k/c_k. \tag{2.4}$$

But $\mathbf{B}$ cannot be obtained since $(y_k, \mathbf{x}_k)$ are observed only for the sampled elements. We then replace (2.4) by the corresponding sample based normal equations (2.2) that can be solved for $\hat{\mathbf{B}}$ as a function of the sample data. This then allows calculation of the residuals $e_k = y_k - \mathbf{x}_k'\hat{\mathbf{B}}$ needed in (2.1).

It is useful to note two alternative expressions for $\hat{Y}_{\text{GREG}}$. The first is given by

$$\hat{Y}_{\text{GREG}} = \hat{Y} + (\mathbf{X} - \hat{\mathbf{X}})'\hat{\mathbf{B}} \tag{2.5}$$

where $\hat{Y} = \Sigma_s a_k y_k$ and $\hat{\mathbf{X}} = \Sigma_s a_k \mathbf{x}_k$ are the Horvitz-Thompson (HT) estimators of Y and $\mathbf{X}$, respectively. The term $(\mathbf{X} - \hat{\mathbf{X}})'\hat{\mathbf{B}}$ is a regression adjustment to the HT estimator $\hat{Y}$. Another useful expression for $\hat{Y}_{\text{GREG}}$ is given by

$$\hat{Y}_{\text{GREG}} = \Sigma_s a_k g_{ks} y_k \tag{2.6}$$

with

$$g_{ks} = 1 + (\mathbf{X} - \hat{\mathbf{X}})'(\Sigma_s a_k \mathbf{x}_k \mathbf{x}'_k/c_k)^{-1}\mathbf{x}_k/c_k. \tag{2.7}$$

Note that (2.6) shows the total weight of y_k as the product of the sampling weight a_k and the factor g_{ks}, called the estimator factor or g-factor. An important property of the GREG estimator is that the estimated population totals of the auxiliary variables equal the corresponding known population totals, i.e.,

$$\Sigma_s a_k g_{ks} \mathbf{x}_k = \mathbf{X}. \tag{2.8}$$

Furthermore, under general conditions, $\hat{Y}_{\text{GREG}}$ is a design consistent estimator of the target parameter Y for any configuration of the finite population values $y_1, y_2, \ldots, y_N$. This property does not depend on whether (2.3) is in some sense a "true model" or not. This has important implications for estimation of domain totals as discussed in later sections.

The regression residuals $e_k = y_k - \mathbf{x}_k'\hat{\mathbf{B}}$ are required in the calculation of the variance estimate corresponding to (2.1). The variance estimator is defined as

$$\widehat{\text{Var}}(\hat{Y}_{\text{GREG}}) = \Sigma_{(k,\ell)\in s}(\Delta_{k\ell}/\pi_{k\ell})(g_{ks}e_k/\pi_k)(g_{\ell s}e_\ell/\pi_\ell) \tag{2.9}$$

where $\Delta_{k\ell} = \pi_{k\ell} - \pi_k\pi_\ell$, and $\Sigma_{(k,\ell)\in s}$ stands for $\Sigma_{k\in s}\Sigma_{\ell\in s}$.

The traditional Taylor expansion variance estimator, which corresponds to taking

$g_k = 1$ for all k in (2.9), is known to give slight underestimation for small sample sizes. However, the presence of the g-factors in (2.9) reduces this underestimation, as evidenced by several simulation studies. Another reason to put the weights g_{ks} on the residuals in (2.9) is that the variance estimator becomes better for conditional inference. Särndal, Swensson, and Wretman (1989) showed that (2.9) is design consistent. It is also approximately unbiased under the model, conditionally on s. A similar desire to combine good design-based and good model-based features underlies the variance estimator of Kott (1990), who starts from (2.9) with $g_k = 1$ and attaches a ratio adjustment to achieve good model properties. A wider class of variance estimators for the GES could be created for a fixed size sampling design, as given by Rao's (1975) representation of a variance estimator for the HT estimator. Formula (2.9) is also implicit in the variance estimation suggested for special cases of $\hat{Y}_{\text{GREG}}$ covered by SUPER-CARP (Hidiroglou, Fuller, and Hickman 1976).

In summary, the calculations that are carried out in the GES are the following. Using the sample data and the sampling design specification, the GES calculates the sampling weights $a_k = 1/\pi_k$ needed for point estimation, as well as the quantities $\Delta_{k\ell}$ needed for variance estimation. Using the specifications of the sampling design, the known auxiliary totals $\mathbf{X} = \Sigma_U \mathbf{x}_k$ and the sample data $(y_k, \mathbf{x}_k)$ for $k \in s$, the GES further calculates: (i) the g-factors, g_{ks}, for $k \in s$, (ii) the point estimate, $\hat{Y}_{\text{GREG}} = \Sigma_s a_k g_{ks} y_k$, (iii) the residuals, $e_k = y_k - \mathbf{x}_k' \hat{\mathbf{B}}$ for $k \in s$, and (iv) the variance estimate (2.9).

3. Features of the Underlying Regression Models

The general linear model (2.3) includes many special cases of potential interest to the user of the GES. There are three important concepts that need to be discussed concerning a model: *model group*, *model level* and *model type*. The concept of model group is discussed in Section 4. The concept of model level relates to the type of unit used in the formulation of the model. A model is said to be at the element level if it is formulated in terms of auxiliary data on the individual elements, as in (2.3). The known auxiliary totals then refer to groups of elements. For the single-stage element designs discussed in Sections 3 to 5, a model is necessarily at the element level. For single-stage cluster designs, considered in Section 6, the model may be formulated either for elements or for clusters of elements. Thus, the model can be at the element level or at the cluster level. For multistage designs, considered in Section 7, several different model levels are possible. In element level models, the known auxiliary totals refer to groups of elements; in cluster level models, they refer to groups of clusters. In either case, the model is used to generate an estimator for characteristics of the elements, for example, the total $Y = \Sigma_U y_k$. Thus, the response variable in the model is always a function of data on elements.

The auxiliary variables specified in the model (2.3) characterize the model type. Simple auxiliary vectors and the associated model types are: (i) $\mathbf{x}_k = 1 = c_k$ for all $k \in U$, corresponding to the common mean model; (ii) $\mathbf{x}_k = x_k = c_k$ for all $k \in U$, where x_k is a single positive variable, corresponding to the ratio model; (iii) $\mathbf{x}_k = (1, x_k)'$ with $c_k = 1$ for all $k \in U$, corresponding to the simple regression

model with an intercept. These three formulations lead via (2.1) and (2.2) to well known GREG estimators of the population total. These are the expansion estimator, the ratio estimator, and the simple regression estimator.

For fixed size sampling designs, the HT estimator $\hat{Y} = \Sigma_s y_k/\pi_k$ can be obtained as a special case of the GREG given by (2.1), by taking $\mathbf{x}_k = \pi_k = c_k$ for all $k \in s$. We note that the auxiliary total $\mathbf{X} = \Sigma_U \pi_k = n$ is known since the sample size is fixed at n. However, the estimated variance (2.9) agrees with the traditional variance estimate only for stratified simple random sampling without replacement.

An important category of models, called group models in Särndal, Swenson, and Wretman (1992), arises when there exists a partition of U into subpopulations $U_1, \ldots, U_p, \ldots, U_P$. We use the term partition to mean a set of mutually exclusive subsets. The subpopulations may correspond, for example, to an age/sex classification of individuals or to an industry classification of business establishments. For every selected element $k \in s$, suppose we can observe the subpopulation to which the unit belongs and possibly other data. Examples of $\mathbf{x}_k$ vectors for such situations are $\mathbf{x}_k = (0, \ldots, 1, \ldots, 0)'$ and $\mathbf{x}_k = (0, \ldots, x_k, \ldots, 0)'$. In these vectors of dimension P, all entries except one are zero; the non-zero entry occurs in the position corresponding to the subpopulation to which k belongs. In the case of $\mathbf{x}_k = (0, \ldots, 1, \ldots, 0)'$ the model type can be characterized as one-way ANOVA; the auxiliary total that must be known is $\mathbf{X} = \Sigma_U \mathbf{x}_k = (N_1, \ldots, N_p, \ldots, N_P)'$ where N_p is the population count in the pth subpopulation. The corresponding GREG estimator is the classical post-stratified estimator (see Example 4.1). In the case of $\mathbf{x}_k = (0, \ldots, x_k, \ldots, 0)'$, the auxiliary total that must be known is $\mathbf{X} = \Sigma_U \mathbf{x}_k = (X_1, \ldots, X_p, \ldots, X_P)'$ where X_p is the total of x_k over U_p. The model types identified in this section, and the associated GREG estimators, are discussed in detail by Särndal, Swensson, and Wretman (1992).

A general expression for the $\mathbf{x}_k$ vector for a partitioned population is given by

$$\mathbf{x}_k = (\delta_{1k}\mathbf{x}'_{1k}, \ldots, \delta_{pk}\mathbf{x}'_{pk}, \ldots, \delta_{Pk}\mathbf{x}'_{Pk})' \tag{3.1}$$

where δ_{pk} is a subpopulation indicator whose value is $\delta_{pk} = 1$ if $k \in U_p$ and $\delta_{pk} = 0$ otherwise. The vector $\mathbf{x}_{pk}$ is composed of all the auxiliary variables available for element $k \in s_p$ where $s_p = U_p \cap s$ is the part of the sample s in U_p. Note that the set of auxiliary variables does not need to be the same in all subpopulations. For example, we could have $\mathbf{x}_{1k} = (1, w_k)'$ for the first subpopulation, $x_{2k} = z_k$ for the second subpopulation, and so on, where w and z are different variables. The requirement that $\mathbf{X} = \Sigma_U \mathbf{x}_k$ is known is equivalent to the requirement that the subtotal $\mathbf{X}_p = \Sigma_{U_p} \mathbf{x}_{pk}$ is known for $p = 1, \ldots, P$. When $\mathbf{x}_k$ is of the form (3.1), the general model specification (2.3) is equivalent to specifying a model separately for each subpopulation. The model implied for the pth subpopulation is given by (4.1) in the following section.

4. Model Groups

We continue to assume a single-stage element sampling design with a probability sample s drawn from the population U. A *model group* is defined to be a subpopulation U_p, $U_p \subseteq U$, such that:

i. a regression model with an auxiliary vector $\mathbf{x}_{pk}$ can be fitted separately within the group, and
ii. the group auxiliary total $\mathbf{X}_p = \Sigma_{U_p}\mathbf{x}_{pk}$ is known.

Let $\mathbf{x}_k$ be given by (3.1) and let $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \ldots, \boldsymbol{\beta}_p', \ldots, \boldsymbol{\beta}_P')'$. Then (2.3) implies a regression model for the model group U_p that can be stated as

$$y_k = \mathbf{x}_{pk}'\boldsymbol{\beta}_p + \varepsilon_k \quad \text{for} \quad k \in U_p \tag{4.1}$$

where $\mathrm{E}_\xi(\varepsilon_k) = 0, \mathrm{Var}_\xi(\varepsilon_k) = c_k\sigma^2$ and $\mathrm{Cov}_\xi(\varepsilon_k, \varepsilon_\ell) = 0$ for all $k \neq \ell$. This model can be fitted using the observed survey data from group U_p, that is, $(y_k, \mathbf{x}_{pk})$ for $k \in s_p$, where $s_p = U_p \cap s$. Then $\boldsymbol{\beta}_p$ is estimated by $\hat{\mathbf{B}}_p$, which satisfies the normal equation

$$(\Sigma_{s_p} a_k \mathbf{x}_{pk}\mathbf{x}_{pk}'/c_k)\hat{\mathbf{B}}_p = \Sigma_{s_p} a_k \mathbf{x}_{pk} y_k/c_k. \tag{4.2}$$

The predicted value of y for element $k \in U_p$ is $\mathbf{x}_{pk}'\hat{\mathbf{B}}_p$. A GREG estimator can be created for the group total $Y_p = \Sigma_{U_p} y_k$ as

$$\hat{Y}_{p,\mathrm{GREG}} = (\Sigma_{U_p}\mathbf{x}_{pk})'\hat{\mathbf{B}}_p + \Sigma_{s_p} a_k(y_k - \mathbf{x}_{pk}'\hat{\mathbf{B}}_p) \tag{4.3}$$

or, expressed in terms of g-factors, as

$$\hat{Y}_{p,\mathrm{GREG}} = \Sigma_{s_p} a_k g_{ks_p} y_k \tag{4.4}$$

where

$$g_{ks_p} = 1 + (\mathbf{X}_p - \hat{\mathbf{X}}_p)'(\Sigma_{s_p} a_k \mathbf{x}_{pk}\mathbf{x}_{pk}'/c_k)^{-1}\mathbf{x}_{pk}/c_k \tag{4.5}$$

with $\mathbf{X}_p = \Sigma_{U_p}\mathbf{x}_{pk}$ and $\hat{\mathbf{X}}_p = \Sigma_{s_p} a_k \mathbf{x}_{pk}$. That is, g-weights can be calculated according to (4.5) for the pth model group, for $p = 1, \ldots, P$.

Poststrata represent an important example of model groups because, by the customary definition, they are non-overlapping subpopulations with known population counts. Other subpopulations often considered in surveys are strata, domains of interest and clusters. They may or may not qualify as model groups. A domain of interest qualifies as a model group if there exists one or more auxiliary variables with known domain totals. For example, we might know the total number of elements in the domain. Strata qualify as model groups assuming the stratum sizes are known.

To produce estimates for any domain of the population of elements, GES requires the existence of a set of model groups defining a partition of U. Unless indicated otherwise, we assume from now on that the model groups satisfy this requirement.

The following result shows that the GREG estimator of the entire population total $Y = \sum_U y_k$ can be obtained by simply adding the GREG estimators for the model group totals.

Result. Let $U_1, \ldots, U_p, \ldots, U_P$ be a partition of U into P model groups. Let $\mathbf{x}_k$ in the global model (2.3) be given by (4.1) with a known vector of totals $\mathbf{X} = (\mathbf{X}_1', \ldots, \mathbf{X}_p', \ldots \mathbf{X}_P')'$ and let the model parameter vector be $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \ldots \boldsymbol{\beta}_p', \ldots \boldsymbol{\beta}_P')'$. Then, the GREG estimator for the entire population total $Y = \Sigma_U y_k$ is equal to the sum of the GREG estimators for the model groups totals.

That is

$$\hat{Y}_{\text{GREG}} = \sum_{p=1}^{P} \hat{Y}_{p,\text{GREG}} \tag{4.6}$$

where $\hat{Y}_{\text{GREG}}$ and $\hat{Y}_{p,\text{GREG}}$ denote the GREG estimators of Y and Y_p defined by (2.6) and (4.3), respectively.

Proof: The GREG estimator of the entire population total is derived by fitting the global model (2.3) with $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \ldots \boldsymbol{\beta}_p', \ldots, \boldsymbol{\beta}_P')'$ and with $\mathbf{x}_k$ given by (4.1). Then $\boldsymbol{\beta}$ is estimated by $\hat{\mathbf{B}}$, which is the solution of $(\Sigma_s a_k \mathbf{x}_k \mathbf{x}_k'/c_k)\hat{\mathbf{B}} = \Sigma_s a_k \mathbf{x}_k y_k/c_k$. Here, the matrix $(\Sigma_s a_k \mathbf{x}_k \mathbf{x}_k'/c_k)$ is block diagonal, and thus easy to invert. The predicted value of y for element k is obtained as $\mathbf{x}_k'\hat{\mathbf{B}} = \mathbf{x}_{pk}'\hat{\mathbf{B}}_p$ for $k \in U_p$, where $\hat{\mathbf{B}}_p$ satisfies the normal equation (4.2). The GREG estimator of the entire population total $Y = \Sigma_U y_k$ is therefore

$$\begin{aligned}\hat{Y}_{\text{GREG}} &= (\Sigma_U \mathbf{x}_k)'\hat{\mathbf{B}} + \Sigma_s a_k (y_k - \mathbf{x}_k'\hat{\mathbf{B}}) \\ &= \sum_{p=1}^{P} \left\{ \mathbf{X}_p'\hat{\mathbf{B}}_p + \Sigma_{s_p} a_k (y_k - \mathbf{x}_{pk}'\hat{\mathbf{B}}_p) \right\} = \sum_{p=1}^{P} \hat{Y}_{p,\text{GREG}}.\end{aligned} \tag{4.7}$$

This result gives a simple recipe for calculating the GREG estimator of $Y = \Sigma_U y_k$ for a population partitioned into model groups:

1. For each model group, calculate g-factors according to (4.5) and use these g-factors to produce $\hat{Y}_{p,\text{GREG}}$ as in (4.4).
2. Sum these estimates over the groups to produce the GREG estimator of the entire population total $Y = \Sigma_U y_k$ as in (4.6).

In the GES, the g-factors are produced by this group-by-group approach. Similarly the residuals for variance estimation are computed group-by-group. That is, the residuals

$$e_k = y_k - \mathbf{x}'_{pk}\hat{\mathbf{B}}_p \quad \text{for} \quad k \in s_p \tag{4.8}$$

are calculated for $p = 1, \ldots, P$ and then used in (2.9) for variance estimation. This approach has computational advantages and is intuitively appealing. The following four examples illustrate the use of model groups.

Example 4.1 Consider a population of individuals partitioned into P model groups $U_1, \ldots, U_p, \ldots, U_P$ corresponding, for example, to an age/sex classification. Let x be an auxiliary variable with value x_k for $k \in U_p$, whose auxiliary total $X_p = \sum_{U_p} x_k$ is known for each model group. Under the model $y_k = \beta_p x_k + \varepsilon_k$, $\text{E}_\xi(\varepsilon_k) = 0$, $\text{Var}_\xi(\varepsilon_k) = x_k\sigma^2$, $\text{Cov}_\xi(\varepsilon_k, \varepsilon_\ell) = 0$, for $k \neq \ell$, the GREG estimator of the model group total $Y_p = \sum_{U_p} y_k$ is $\hat{Y}_{p,\text{GREG}} = X_p\hat{B}_p$ where $\hat{B}_p = (\Sigma_{s_p} a_k y_k)/(\Sigma_{s_p} a_k x_k)$. Summing over the groups, the estimator of the entire population total is obtained as $\hat{Y}_{\text{GREG}} = \Sigma_{p=1}^{P} X_p\hat{B}_p$. This is the familiar poststratified ratio estimator. The classical poststratified estimator is obtained when $x_k = 1$ for all k, so that $X_p = N_p$ represents the known count in the pth group.

Example 4.2 Consider a survey with stratified sampling and poststratification estimation. The strata (indexed by $h = 1, 2, \ldots, H$) and the poststrata (indexed by $j = 1, 2, \ldots, J$) are based on different criteria. For example, in a survey of individuals, the strata may represent provinces and the individuals may be poststratified by age/sex groups. Suppose that a sample s is obtained by simple random sampling without replacement (SRSWOR) in each stratum, with n_h elements sampled from N_h in the hth stratum. Assume that an auxiliary variable x_k is observed for the elements $k \in s$. Depending on the auxiliary totals that are known, it is possible to define different sets of model groups. The model within a group is assumed to be of the form $y_k = \beta x_k + \varepsilon_k$ where β is an unknown slope parameter, $\mathrm{E}_\xi(\varepsilon_k) = 0$, $\mathrm{Var}_\xi(\varepsilon_k) = x_k\sigma^2$ and $\mathrm{Cov}_\xi(\varepsilon_k, \varepsilon_\ell) = 0$ for all $k \neq \ell$. The following three cases A, B and C illustrate different model group specifications.

Case A. The J poststrata are used as model groups. Here, the poststrata totals $X_j = \sum_{U_j} x_k$ must be known. The GREG estimator of Y is obtained as

$$\hat{Y}_{\mathrm{GREG}} = \sum_{j=1}^{J} X_j \hat{B}_j$$

with $\hat{B}_j = (\hat{Y}_j/\hat{X}_j)$, $\hat{Y}_j = \sum_{h=1}^{H}(N_h/n_h)\Sigma_{s_{hj}} y_k$ and $\hat{X}_j$ defined in the same manner, where $s_{hj} = s_h \cap s_j$ is the set of sampled elements in stratum h and poststratum j. This is a poststratified ratio estimator in which strata are combined to compute $\hat{B}_j$.

Case B. The $H \times J$ cells formed by crossclassifying the strata and poststrata are used as model groups. This implies that the cell totals $X_{hj} = \Sigma_{U_{hj}} x_k$ must be known. The GREG estimator of Y is

$$\hat{Y}_{\mathrm{GREG}} = \sum_{h=1}^{H}\sum_{j=1}^{J} X_{hj}\hat{B}_{hj}$$

where $\hat{B}_{hj} = \hat{Y}_{hj}/\hat{X}_{hj}$ with $\hat{Y}_{hj} = (N_h/n_h)\Sigma_{s_{hj}} y_k$ and $\hat{X}_{hj} = (N_h/n_h)\Sigma_{s_{hj}} x_k$. This can be described as a separate poststratified estimator, because strata are treated separately within each poststratum.

Case C. Familiar estimators were obtained with the model groups in cases A and B. They represent two extremes of the following intermediate case. Let the population cells $\{U_{hj}\}$ be combined in a specified fashion to form a partition of U into P model groups U_p for $p = 1, 2, \ldots, P$. That is, a typical group U_p is the union of a specified set of cells. The cells that make up U_p can be part of one or more of the strata or one or more of the poststrata. Supposing the group totals $X_p = \Sigma_{U_p} x_k$ are known, the GREG estimator of Y is given by

$$\hat{Y}_{\mathrm{GREG}} = \sum_{p=1}^{P} X_p \hat{B}_p$$

where $\hat{B}_p = \hat{Y}_p/\hat{X}_p$ with $\hat{Y}_p = \Sigma_{s_p} a_k y_k$, $\hat{X}_p = \Sigma_{s_p} a_k x_k$ and $a_k = N_h/n_h$ for each element k in stratum h.

In some situations, auxiliary information may be available for overlapping subpopulations. We cannot specify each of these subpopulations as a model group in GES, because they are not mutually exclusive. However, this need not cause a waste

of information. It is possible to make a complete use of the available information by: (i) specifying a model group as the union of subpopulations, and (ii) suitably redefining the vector of auxiliary variables. Example 4.3 shows how this is done.

Example 4.3 Suppose the population U consists of two overlapping subpopulations U_1 and U_2 such that the population count N_1 is known for U_1, whereas the total $X_2 = \Sigma_{U_2} x_k$ is known for U_2. Then, by considering $U = U_1 \cup U_2$ as the only model group, it is possible to profit from all of the available auxiliary information in deriving the GREG estimator by defining $\mathbf{x}_k$ for $k \in U$ in the following way

$$\mathbf{x}_k = \begin{cases} (1, x_k)' & \text{for } k \in U_1 \cap U_2 \\ (1, 0)' & \text{for } k \in U - U_2 \\ (0, x_k)' & \text{for } k \in U - U_1 \end{cases}$$

We assume that each sampled element can be classified as belonging to one of the three subpopulations used to define $\mathbf{x}_k$. It follows that $\mathbf{X} = \Sigma_U \mathbf{x}_k = (N_1, X_2)'$ is the required vector of auxiliary totals for the GREG estimator. Letting $s_i = s \cap U_i$, $i = 1, 2$, we have that the g-factors associated with this estimator produce the known totals as follows

$$\begin{pmatrix} N_1 \\ X_2 \end{pmatrix} = \Sigma_s a_k g_{ks} \mathbf{x}_k = \begin{pmatrix} \Sigma_{s_1} a_k g_{ks} \\ \Sigma_{s_2} a_k g_{ks} x_k \end{pmatrix}.$$

In the following example, which involves generalized raking, the only model group formulation permitting full use of available information is the entire population U.

Example 4.4 Consider a population of individuals U divided into $r \times c$ cells U_{ij} formed by crossclassifying age and occupation categories. The marginal counts $N_{i.}$ for $i = 1, 2, \ldots, r$, are known for the r age categories (rows). Similarly, the single auxiliary variable totals $X_{.j}$ for $j = 1, 2, \ldots, c$, are known for the c occupation categories (columns). The cell counts N_{ij} and the cell totals X_{ij} are, however, unknown. A model appropriate for this situation is

$$y_k = \alpha_i + \beta_j x_k + \varepsilon_k \quad \text{for} \quad k \in U_{ij}, \quad i = 1, \ldots, r, \quad j = 1, \ldots, c$$

which we can write in general form as

$$y_k = \mathbf{x}_k' \boldsymbol{\beta} + \varepsilon_k \quad \text{for} \quad k \in U \tag{4.9}$$

where both $\boldsymbol{\beta} = (\alpha_1, \ldots, \alpha_r, \beta_1, \ldots, \beta_c)'$ and $\mathbf{x}_k$ have the dimension $(r + c)$ and $\mathbf{x}_k$ is defined as

$$\mathbf{x}_k = (0, \ldots, 1, \ldots, 0, \quad 0, \ldots, x_k, \ldots, 0)'.$$

There is an entry equal to 1 in one of the first r positions corresponding to the row to which k belongs, and an entry with value x_k in one of the last c positions corresponding to the column to which k belongs. The other $(r + c - 2)$ entries have values of 0. Here $\mathbf{X} = \Sigma_U \mathbf{x}_k$ is a vector composed of the r known counts $N_{i.}$ for $i = 1, 2, \ldots, r$ and the c known totals $X_{.j}$ for $j = 1, 2 \ldots, c$. Here, the cells do not qualify as model groups because the N_{ij} and the X_{ij} are unknown. It would be possible to use either the age

categories or the occupation categories as model groups. But, in either case some auxiliary information must be disregarded. If the age categories are used as model groups then we cannot make use of the known total $X_{.j}$ for each occupation category. Similarly, some auxiliary information would have to be foregone if we used the occupation categories as model groups. The entire population is the only model group that can be defined to make use of all the auxiliary information. The GREG estimator is derived by fitting model (4.9) to the entire population U, assuming $\mathrm{E}_\xi(\varepsilon_k) = 0$, $\mathrm{Var}_\xi(\varepsilon_k) = c_k\sigma^2$ and $\mathrm{Cov}_\xi(\varepsilon_k, \varepsilon_\ell) = 0$ for all $k \neq \ell$ where c_k is a specified constant for each $k \in U$. The normal equations consist of a system of $(r + c)$ equations. Note that one equation is redundant when $x_k = 1$ for all $k \in s$, and must be deleted in solving the normal equations. They are similar to the equations solved by the CALMAR software described in Deville, Särndal, and Sautory (1993).

5. Estimators for Domains of the Finite Population

A *domain* is any subset of the population of elements U for which a separate estimate is required. Most large surveys require estimates for a variety of domains. Therefore, estimating characteristics of arbitrarily specified domains of the survey population is an important feature of the GES. The general notation for a domain is $U_{(d)}$. The sample s is drawn from U based on a sampling design with inclusion probabilities π_k and $\pi_{k\ell}$. As before, let $a_k = 1/\pi_k$ represent the sampling weight associated with element k. Also let $s_{(d)} = s \cap U_{(d)}$ denote the part of the sample s that falls in $U_{(d)}$. It is helpful to work with a domain variable of interest, $y_{(d)}$, whose value for the kth element is defined as

$$y_{(d)k} = \begin{cases} y_k & \text{if } k \in U_{(d)} \\ 0 & \text{if } k \notin U_{(d)} \end{cases}. \tag{5.1}$$

The domain total of y, denoted $Y_{(d)}$, can then be expressed as $Y_{(d)} = \Sigma_{U_{(d)}} y_k = \Sigma_U y_{(d)k}$.

In some survey applications the domains of interest form a partition of the survey population U and estimates are required for the D domain totals $Y_{(d)} = \Sigma_{U_{(d)}} y_k$, $d = 1, 2, \ldots, D$. Users often require these estimates to be additive over the domains. Estimates of the domain totals then add up to the estimate of the entire population total. The GES produces domain GREG estimators, $\hat{Y}_{(d),\mathrm{GREG}}$, $d = 1, \ldots, D$, with this additivity property, that is, $\sum_{d=1}^{D} \hat{Y}_{(d),\mathrm{GREG}} = \hat{Y}_{\mathrm{GREG}}$. A simple proof of this additivity property is given later in this section.

We assume, as before, that there exists P model groups $U_1, \ldots, U_p \ldots, U_P$ that form a partition of U. The domain of interest, $U_{(d)}$ can be related to the model groups in a variety of ways. We consider a general case and two special cases.

Special case 1. The domain is identical with a model group.
Special case 2. The domain is properly contained in a model group.
General case. The domain intersects one or more model groups.

Special case 1 implies that auxiliary totals are available for the domain itself. In practice, this is often not so and the general case (or special case 2) is more likely to prevail.

The general case is illustrated by the following situation. In a population of business establishments, the SIC code of a given business establishment may change from one year to another. The change may only be discovered when the unit is sampled and observed in the current survey. The estimation of the current SIC group total is an example of domain estimation as in the general case.

For simplicity, we consider first special case 1, that is, the domain of interest $U_{(d)}$ is identical with a model group. This implies that there exists a vector of known auxiliary totals $\mathbf{X}_{(d)} = \Sigma_{U_{(d)}} \mathbf{x}_k$ for the domain. We can then calculate g-factors based on this information and obtain the GREG estimator of the domain total according to (2.7) and (2.8)

$$\hat{Y}_{(d),\text{GREG}} = \Sigma_s a_k g_{ks_{(d)}} y_{(d)k} = \Sigma_{s_{(d)}} a_k g_{ks_{(d)}} y_k \tag{5.2}$$

where the g-weight for element k is given by

$$g_{ks_{(d)}} = 1 + (\mathbf{X}_{(d)} - \hat{\mathbf{X}}_{(d)})'(\Sigma_{s_{(d)}} a_k \mathbf{x}_k \mathbf{x}_k'/c_k)^{-1} \mathbf{x}_k/c_k \tag{5.3}$$

with $\mathbf{X}_{(d)} = \Sigma_{U_{(d)}} \mathbf{x}_k$ and $\hat{\mathbf{X}}_{(d)} = \Sigma_{s_{(d)}} a_k \mathbf{x}_k$. These expressions are illustrated by the following example.

Example 5.1 Consider a population partitioned into D domains, each being identical with a model group. Suppose that the total $X_{(d)} = \Sigma_{U_{(d)}} x_k$ is known for each domain $U_{(d)}$, $d = 1, \ldots, D$. Suppose the model for domain $U_{(d)}$ is $y_k = \beta_{(d)} x_k + \varepsilon_k$ where $\mathrm{E}_\xi(\varepsilon_k) = 0$, $\mathrm{Var}_\xi(\varepsilon_k) = x_k \sigma^2$ and $\mathrm{Cov}_\xi(\varepsilon_k, \varepsilon_\ell) = 0$ for all $k \neq \ell$. Applying (2.2) we get $\hat{B}_{(d)} = (\Sigma_{s_{(d)}} a_k y_k)/(\Sigma_{s_{(d)}} a_k x_k) = \hat{Y}_{(d)}/\hat{X}_{(d)}$. The GREG estimator (5.2) for the domain total $Y_{(d)}$ is

$$\hat{Y}_{(d),\text{GREG}} = \Sigma_{s_{(d)}} a_k g_{ks_{(d)}} y_k = X_{(d)} \hat{B}_{(d)} \tag{5.4}$$

where the g-factor is given by $g_{ks_{(d)}} = X_{(d)}/\hat{X}_{(d)}$ for all $k \in s_{(d)}$. This is a domain ratio estimator. The GREG estimator of the entire population total, obtained by adding the domain estimators, is

$$\hat{Y}_{\text{GREG}} = \sum_{d=1}^{D} X_{(d)} \hat{B}_{(d)} \tag{5.5}$$

which has the form of a poststratified ratio estimator if we consider the domains to be the poststrata. Here each domain should contain enough observations to avoid unstable slope estimates $\hat{B}_{(d)}$.

Consider now special case 2, that is, the domain is properly contained in one model group. For simplicity, we assume this model group to be the entire population. The vector of auxiliary totals $\mathbf{X} = \Sigma_U \mathbf{x}_k$ is assumed known. We can then calculate g-factors from (2.6) and produce a domain estimate by applying these g-weights to the domain variable values $y_{(d)k}$. Since these g-factors are calculated on auxiliary information at an aggregated level, that is, the entire population U, they do not require auxiliary information for the domain itself. The GREG estimator of $Y_{(d)} = \Sigma_{U_{(d)}} y_k$ is then

$$\hat{Y}_{(d),\text{GREG}} = \Sigma_s a_k g_{ks} y_{(d)k} = \Sigma_{s_{(d)}} a_k g_{ks} y_k \tag{5.6}$$

where g_{ks} is given by (2.7). We can also express this estimator as

$$\hat{Y}_{(d),\text{GREG}} = \hat{Y}_{(d)} + (\mathbf{X} - \hat{\mathbf{X}})'\hat{\mathbf{B}}_{(d)} \tag{5.7}$$

where $\hat{Y}_{(d)} = \Sigma_s a_k y_{(d)k}$ and $\hat{\mathbf{X}} = \Sigma_s a_k \mathbf{x}_k$ are the HT estimators of $Y_{(d)} = \Sigma_U y_{(d)k}$ and $\mathbf{X} = \Sigma_U \mathbf{x}_k$ and $\hat{\mathbf{B}}_{(d)}$ is the solution of the normal equation

$$(\Sigma_s a_k \mathbf{x}_k \mathbf{x}_k'/c_k)\hat{\boldsymbol{B}}_{(d)} = \Sigma_s a_k \mathbf{x}_k y_{(d)k}/c_k. \tag{5.8}$$

This normal equation arises formally from the fit of the regression of the domain variable $y_{(d)}$ on $\mathbf{x}$ through the model

$$y_{(d)k} = \mathbf{x}'_k\boldsymbol{\beta} + \varepsilon_k \quad \text{for} \quad k \in U \tag{5.9}$$

where $\mathrm{E}_\xi(\varepsilon_k) = 0$, $\mathrm{Var}_\xi(\varepsilon_k) = c_k\sigma^2$ and $\mathrm{Cov}_\xi(\varepsilon_k, \varepsilon_\ell) = 0$ for all $k \neq \ell$. This fit will sometimes be mediocre because of the special nature of $y_{(d)}$, which equals y inside the domain but is always equal to zero outside. Therefore, the residuals cannot be expected to fluctuate in close vicinity of zero. The residual for element k will usually have a substantial positive value or a substantial negative value depending on whether $y_{(d)k} = y_k$ or $y_{(d)k} = 0$.

For example, consider a population of business establishments where $x =$ *Gross Business Income* and $y =$ *Wages and Salaries*. Suppose the auxiliary variable x explains y well at the entire population level with a value of R^2 equal to about 0.90. The entire population estimator (2.5) will then realize important gains in precision due to regression, compared to the standard HT estimator. However, we can expect the regression of the domain variable $y_{(d)}$ on x to be much weaker. The residuals will be considerably larger. The domain estimator (5.7) may produce little or no gain due to regression. Here we are not primarily interested in the goodness of the fit. Instead the primary objective is to work with g-factors that: (i) produce additive domain estimates (an often required property), and (ii) are unchanged from one domain to another (which has some computational advantages).

Example 5.2 To illustrate special case 2, consider a population partitioned into D mutually exclusive and exhaustive domains $U_{(1)}, \ldots, U_{(d)}, \ldots, U_{(D)}$. Assume that the only model group is the entire population U and suppose there is a single positive auxiliary variable x for which the population total $X = \Sigma_U x_k$ is known. Consider the model stating that $y_k = \beta x_k + \varepsilon_k$ where $\mathrm{E}_\xi(\varepsilon_k) = 0$, $\mathrm{Var}_\xi(\varepsilon_k) = x_k\sigma^2$ and $\mathrm{Cov}_\xi(\varepsilon_k, \varepsilon_\ell) = 0$ for all $k \neq \ell$. Thus we have $\hat{B} = \hat{Y}/\hat{X}$ where $\hat{Y} = \Sigma_s a_k y_k$ and $\hat{X} = \Sigma_s a_k x_k$. The g-factors are given by $g_{ks} = X/\hat{X}$ for all $k \in s$. The estimator (5.7) of the domain total $Y_{(d)}$ is then given by

$$\hat{Y}_{(d),\text{GREG}} = (X/\hat{X})\Sigma_{s_{(d)}} a_k y_k = (X/\hat{X})\hat{Y}_{(d)}. \tag{5.10}$$

Summing over domains, we get the GREG estimator of the entire population total, which is the standard ratio estimator

$$\hat{Y}_{\text{GREG}} = \sum_{d=1}^{D} (X/\hat{X})\hat{Y}_d = (X/\hat{X})\hat{Y}. \tag{5.11}$$

The domain estimators defined by (5.10) for $d = 1, 2, \ldots, D$ can thus be viewed as a way to distribute the entire population ratio estimator (5.11) over the domains in an additive fashion. For the dth domain, (5.10) may be only slightly better than the HT estimator $\Sigma_{s(d)} a_k y_k$. For the entire population, (5.11) will be considerably better than the HT estimator $\Sigma_s a_k y_k$ if the model provides a strong fit for the entire population.

We turn now to the general case. That is, the domain of interest intersects one or more of the P model groups indexed $p = 1, \ldots, P$. For group U_p we have an auxiliary vector $\mathbf{x}_p$ with known total $\mathbf{X}_p = \Sigma_{U_p} \mathbf{x}_{pk}$. With this information, the g-factors g_{ks_p} defined by (4.5) are first calculated for each model group $U_p, p = 1, \ldots, P$. The GREG estimator of the domain total $Y_{(d)} = \Sigma_U y_{(d)k}$ is then obtained as

$$\hat{Y}_{(d),\mathrm{GREG}} = \sum_{p=1}^{P} \Sigma_{s_p} a_k g_{ks_p} y_{(d)k} \tag{5.12}$$

where $s_p = s \cap U_p$. In summary, the steps in the calculation of the domain estimator (5.12) are as follows.

1. For domain U_d, identify the intersecting model groups, that is, the groups U_p such that $U_d \cap U_p$ is non-empty.
2. If U_p is an intersecting model group, apply the combined weight $a_k g_{ks_p}$ to the value $y_{(d)k}$ and sum over the elements $k \in s_p$.
3. Sum over the intersecting model groups to obtain the point estimate (5.12) of the domain total $Y_{(d)}$.

In the GES, the computation of the estimator (5.12) and the corresponding variance estimator is easily handled. The variable y is replaced in the formulas in Sections 2 and 4 by the domain variable $y_{(d)}$. For the variance estimation, this implies that the residuals $e_k = y_k - \mathbf{x}'_{pk}\hat{\mathbf{B}}_p$ for $k \in s_p$ are replaced in (2.9) by

$$e_{(d)k} = y_{(d)k} - \mathbf{x}'_{pk}\hat{\mathbf{B}}_{(d)p} \tag{5.13}$$

for $k \in s_p$, where $\hat{B}_{(d)p}$ satisfies $(\Sigma_{s_p} a_k \mathbf{x}_{pk}\mathbf{x}'_{pk}/c_k)\hat{\mathbf{B}}_{(d)p} = \Sigma_{s_p} a_k \mathbf{x}_{pk} y_{(d)k}/c_k$. Here, (2.9) becomes

$$\widehat{\mathrm{Var}}(\hat{Y}_{(d),\mathrm{GREG}}) = \Sigma_{(k,\ell)\in s}(\Delta_{k\ell}/\pi_{k\ell})(g_{ks}e_{(d)k}/\pi_k)(g_{\ell s}e_{(d)\ell}/\pi_\ell). \tag{5.14}$$

Thus, for any model group U_p that intersects the domain of interest $U_{(d)}$, we have

$$e_{(d)k} = \begin{cases} y_k - \mathbf{x}'_{pk}\hat{\mathbf{B}}_{(d)p} & \text{if } k \in s_p \text{ and } k \in U_{(d)} \\ -\mathbf{x}'_{pk}\hat{\boldsymbol{B}}_{(d)p} & \text{if } k \in s_p \text{ and } k \notin U_{(d)}. \end{cases}$$

Further, $e_{(d)k} = 0$ for all sample elements belonging to non-intersecting model groups, which simplifies the calculation of the variance estimator (5.14).

The domain estimator (5.12) has two important properties: (i) design consistency and (ii) additivity over a set of domains forming a partition of the entire population U. The design consistency follows because (2.1), or equivalently (2.6), is a design consistent estimator for any configuration of values $y_1, y_2, \ldots, y_N$. Thus, it is design

consistent in particular for the configuration $y_{(d)1}, y_{(d)2}, \ldots, y_{(d)N}$. The additivity property,

$$\sum_{d=1}^{D} \hat{Y}_{(d),\mathrm{GREG}} = \hat{Y}_{\mathrm{GREG}}$$

where $\hat{Y}_{\mathrm{GREG}}$ and $\hat{Y}_{(d),\mathrm{GREG}}$ are defined by (4.6) and (5.12), follows easily since $\sum_{d=1}^{D} y_{(d)k} = y_k$ for all $k \in U$.

Example 5.3 Consider a domain that intersects the first two out of $P > 2$ model groups. Then the GREG estimator (5.12) of the domain total $Y_{(d)} = \Sigma_{U_{(d)}} y_k$ becomes

$$\hat{Y}_{(d),\mathrm{GREG}} = \sum_{p=1}^{2} \Sigma_{s_p} a_k g_{ks_p} y_{(d)k} = \sum_{p=1}^{2} \Sigma_{s_{(d)p}} a_k g_{ks_p} y_k$$

where $s_{(d)p} = s_p \cap U_{(d)} = s \cap U_p \cap U_{(d)}$ is the part of the sample s that falls in domain $U_{(d)}$ and group U_p for $p = 1, 2$. To be specific, suppose there is a known group count N_1 for the first model group, and a known total $X_2 = \Sigma_{U_2} x_k$ for variable x in the second model group. Then the GREG estimator of the domain total is

$$\hat{Y}_{(d),\mathrm{GREG}} = (N_1/\hat{N}_1)\hat{Y}_{(d)1} + (X_2/\hat{X}_2)\hat{Y}_{(d)2}$$

where $\hat{N}_1 = \Sigma_{s_1} 1/\pi_k$, $\hat{X}_2 = \Sigma_{s_2} x_k/\pi_k$, $\hat{Y}_{(d)1} = \Sigma_{s_{(d)1}} y_k/\pi_k$ and $\hat{Y}_{(d)2} = \Sigma_{s_{(d)2}} y_k/\pi_k$. The calculation of the variance estimate (5.14) will require the following residuals $e_{(d)k}$:

a. for sample elements in the first model group, $k \in s_1 = s \cap U_1$, we have

$$e_{(d)k} = \begin{cases} y_k - (\hat{Y}_{(d)1}/\hat{N}_1) & \text{if } k \in U_{(d)} \\ -(\hat{Y}_{(d)1}/\hat{N}_1) & \text{if } k \notin U_{(d)} \end{cases}$$

b. for sample elements in the second model group, $k \in s_2 = s \cap U_2$, we have

$$e_{(d)k} = \begin{cases} y_k - x_k(\hat{Y}_{(d)2}/\hat{X}_2) & \text{if } k \in U_{(d)} \\ -x_k(\hat{Y}_{(d)2}/\hat{X}_2) & \text{if } k \notin U_{(d)} \end{cases}$$

c. $e_{(d)k} = 0$ for all other sample elements, $k \in s - s_1 - s_2$.

6. Single-Stage Cluster Sampling

The previous sections dealt with the GREG estimators for single-stage element sampling designs. But, because of cost, administrative reasons or sampling efficiency, these types of designs are not used in many medium to large scale sample surveys. For those surveys, the sampling design generally involves cluster sampling in one or more stages, possibly with unequal probability selection at each stage. Clusters generally form natural groupings of the elements of the finite population. In single-stage cluster sampling, a probability sample of clusters is selected and all elements in these clusters are surveyed. This section discusses the use of GREG estimation in single-stage cluster sampling designs.

Let the population of elements $U = \{1, \ldots, k, \ldots, N\}$ be partitioned into N_{I} clusters. The population of clusters is denoted by $U_{\mathrm{I}} = \{1, \ldots, i, \ldots, N_{\mathrm{I}}\}$. This population may

be stratified. A sample s_I of clusters is selected from this population with associated probability of selection $p_I(s_I)$. The elements in the sampled clusters form the sample of elements $s = \cup_{i \in s_I} s_i$, where s_i denotes the set of sampled elements in the ith cluster. Note that in single-stage cluster sampling, s_i is composed of all elements in the ith cluster. The sampling weight for the ith cluster is denoted by $a_{Ii} = 1/\pi_{Ii}$ for $i \in U_I$, where π_{Ii} is the cluster inclusion probability. As in earlier sections, the sampling weight for the kth element is denoted by a_k. In single-stage cluster sampling designs, $a_k = a_{Ii}$ for every $k \in s_i$.

Let us consider the estimation of the population total $Y = \Sigma_U y_k$ and the domain total $Y_{(d)} = \Sigma_{U_{(d)}} y_k$. In cluster sampling designs, auxiliary information may be available: (i) for subgroups of the population of clusters (case A below) or (ii) for subgroups of the population of elements (case B below).

Case A. Model groups at the element level. Suppose the population of elements U is partitioned into the model groups $U_1, \ldots, U_p, \ldots, U_P$ for which we have known auxiliary totals $\mathbf{X}_p = \Sigma_{U_p} \mathbf{x}_{pk}$, $p = 1, \ldots, P$. The auxiliary variables may be different in each group. The part of the samples that falls in the pth model group is denoted as $s_p = s \cap U_p$. Within each model group we can formulate a regression model at the element level. The element level model for group U_p is given by (4.1), and the g-factor g_{ks_p} for the sampled element $k \in s_p$ is given by (4.5). The GREG estimator of the population total Y is then given by

$$\hat{Y}_{\text{GREG}} = \sum_{p=1}^{P} \Sigma_{s_p} a_k g_{ks_p} y_k. \tag{6.1}$$

The residuals produced by the model fit are given as

$$e_k = y_k - \mathbf{x}'_{pk} \hat{\mathbf{B}}_p \quad \text{for} \quad k \in s_p \tag{6.2}$$

where $\hat{\mathbf{B}}_p$ satisfies the normal equation (4.2).

The estimation of variance of $\hat{Y}_{\text{GREG}}$ involves a simple modification of formula (2.9).

The variance is estimated by

$$\widehat{Var}(\hat{Y}_{\text{GREG}}) = \Sigma_{(i,j) \in s_I} (\Delta_{Iij}/\pi_{Iij})(E_i/\pi_{Ii})(E_j/\pi_{Ij}) \tag{6.3}$$

where $E_i = \sum_{k \in s_i} g_{ks_p} e_k$ for $k \in s_p$, e_k is given by (6.2), and $\Delta_{Iij} = \pi_{Iij} - \pi_{Ii}\pi_{Ij}$.

It is a simple matter to produce estimates for domain totals. Replacing y_k by $y_{(d)k}$ given by (5.1), we obtain the GREG estimator of the domain total $Y_{(d)}$. For the corresponding variance estimation, E_i is replaced by $E_{(d)i} = \sum_{k \in s_i} g_{ks_p} e_{(d)k}$ where $e_{(d)k}$ is given by (5.13).

The following examples illustrate how familiar estimators may be obtained under the specification of an element level model.

Example 6.1 Suppose an auxiliary variable x_p is available for the group U_p with the known group total $X_p = \Sigma_{U_p} x_{pk}, p = 1, \ldots, P$. Within the group U_p, we specify the ratio model as $y_k = \beta x_{pk} + \varepsilon_k$ with $\text{Var}_\xi(\varepsilon_k) = \sigma^2 c_k = \sigma^2 x_{pk}$. We find that the g-factor for

each sampled element $k \in s_p$ is given by

$$g_{ks_p} = X_p/\hat{X}_p = X_p/\Sigma_{s_p} a_k x_{pk} \tag{6.4}$$

with $a_k = 1/\pi_{\mathrm{I}i}$ for all elements $k \in s_i$. The domain total $Y_{(d)}$ is then estimated by

$$\hat{Y}_{(d),\mathrm{GREG}} = \sum_{p=1}^{P} (X_p/\hat{X}_p)\hat{Y}_{(d)p} \tag{6.5}$$

where $\hat{Y}_{(d)p} = \sum_{s_p} a_k y_{(d)k}$. Note that the domains, the clusters and the model groups may cut across each other. Each may be based on a different classification criterion. Nested arrangements are also possible. Although (6.5) has the familiar appearance of a poststratified ratio estimator (with model groups as poststrata), it actually represents a variety of ratio type estimators. A special case of interest is when $x_{pk} = c_k = 1$ for each element $k \in U_p$. The g-factors are then $g_{ks_p} = N_p/\hat{N}_p$ for all $k \in s_p$, where N_p is the total number of elements in the group, $\hat{N}_p = \Sigma_{s_p} a_k$, and $a_k = 1/\pi_{\mathrm{I}i}$ for all $k \in s_i$.

Case B. Model groups at the cluster level. Suppose the population of clusters U_{I} can be partitioned into the model groups $U_{\mathrm{I}1}, \ldots, U_{\mathrm{I}p}, \ldots, U_{\mathrm{I}P}$ with auxiliary information available for each group. The auxiliary variable vector value $\mathbf{x}_{pi}$ is available for every cluster $i \in s_{\mathrm{I}p}$, where $s_{\mathrm{I}p} = s_{\mathrm{I}} \cap U_{\mathrm{I}P}$ is the set of sampled clusters falling in group $U_{\mathrm{I}p}$, and $\mathbf{X}_p = \Sigma U_{\mathrm{I}p} \mathbf{x}_{pi}$ is the known vector of group auxiliary totals. The cluster level model for group $U_{\mathrm{I}p}$ is

$$Y_i = \mathbf{x}'_{pi}\boldsymbol{\beta}_{\mathrm{I}p} + \varepsilon_i \quad \text{for} \quad i \in U_{\mathrm{I}p} \tag{6.6}$$

where $Y_i = \Sigma_{s_i} y_k$ is the total of y for the ith cluster. We assume $\mathrm{E}_\xi(\varepsilon_i) = 0$, $\mathrm{Var}_\xi(\varepsilon_i) = c_i \sigma_{\mathrm{I}}^2$ and $\mathrm{Cov}_\xi(\varepsilon_i, \varepsilon_j) = 0$ for all $i \neq j$. Note that the dependent variable in (6.6) is the cluster total Y_i. The GREG estimator of Y is given by

$$\hat{Y}_{\mathrm{GREG}} = \sum_{p=1}^{P} \Sigma_{s_{\mathrm{I}p}} a_{\mathrm{I}i} g_{is_{\mathrm{I}p}} Y_i$$

where $g_{is_{\mathrm{I}p}}$ is a cluster g-factor given by

$$g_{is_{\mathrm{I}p}} = 1 + (\mathbf{X}_p - \hat{\mathbf{X}}_p)'(\Sigma_{s_{\mathrm{I}p}} a_{\mathrm{I}i} \mathbf{x}_{pi} \mathbf{x}'_{pi}/c_i)^{-1} \mathbf{x}_{pi}/c_i \quad \text{for} \quad i \in s_{\mathrm{I}p} \tag{6.7}$$

where $\hat{\mathbf{X}}_p = \Sigma_{s_{\mathrm{I}p}} a_{\mathrm{I}i} \mathbf{x}_{pi}$. The sample based estimate $\hat{\mathbf{B}}_{\mathrm{I}p}$ of $\boldsymbol{\beta}_{\mathrm{I}p}$ is obtained as the solution of

$$(\Sigma_{s_{\mathrm{I}p}} a_{\mathrm{I}i} \mathbf{x}_{pi} \mathbf{x}'_{pi}/c_i)\hat{\mathbf{B}}_{\mathrm{I}p} = \Sigma_{s_{\mathrm{I}p}} a_{\mathrm{I}i} \mathbf{x}_{pi} Y_i/c_i \tag{6.8}$$

and the residuals obtained from the model fit are given as

$$e_i = Y_i - \mathbf{x}'_{pi}\hat{\mathbf{B}}_{\mathrm{I}p} \quad \text{for} \quad i \in s_{\mathrm{I}p}. \tag{6.9}$$

For cluster level models, the variance of $\hat{Y}_{\mathrm{GREG}}$ is estimated through a straightforward modification of (2.9),

$$\widehat{\mathrm{Var}}(\hat{Y}_{\mathrm{GREG}}) = \Sigma_{(i,j)\in s_{\mathrm{I}}} (\Delta_{\mathrm{I}ij}/\pi_{\mathrm{I}ij})(g_{is_{\mathrm{I}p}} e_i/\pi_i)(g_{js_{\mathrm{I}p}} e_j/\pi_j) \tag{6.10}$$

where e_i is given by (6.9). Again, estimates for domains require only a simple change of variable. The GREG estimator, $\hat{Y}_{(d),\mathrm{GREG}}$, of the domain total $Y_{(d)} = \Sigma_{U_{(d)}} y_k$ is

obtained by replacing Y_i by $Y_{(d)i} = \Sigma_{s_i} y_{(d)k}$. The appropriate residuals for the domain, $e_{(d)i}$, are calculated by replacing Y_i in (6.8) and (6.9) by $Y_{(d)i}$. The variance estimator, $\widehat{\text{Var}}(Y_{(d),\text{GREG}})$ is then obtained from (6.10) by replacing e_i by $e_{(d)i}$.

The ratio models in Example 6.1 were at the element level. It is interesting to compare it with the following Example 6.2, where the ratio models are placed instead at the cluster level.

Example 6.2 Suppose an auxiliary variable total $X_p = \Sigma_{U_{\text{I}p}} x_{pi}$ is available for each group of clusters $U_{\text{I}p}$, $p = 1, \ldots, P$. If we let $\mathbf{x}_{pi} = x_{pi}$ and $c_i = x_{pi}$ in the general regression model (6.6), we have a ratio model for the pth group. The g-factor for the sampled cluster $i \in s_{\text{I}p}$ is then obtained as

$$g_{is_{\text{I}p}} = (X_p / \hat{X}_p) = X_p / \Sigma_{s_{\text{I}p}} a_{\text{I}i} x_{pi} \tag{6.11}$$

where $\hat{X}_p = \Sigma_{s_{\text{I}p}} a_{\text{I}i} x_{pi}$. The resulting estimator of the domain total $Y_{(d)}$ is

$$\hat{Y}_{(d),\text{GREG}} = \sum_{p=1}^{P} (X_p / \hat{X}_p) \hat{Y}_{(d)p} \tag{6.12}$$

where $\hat{Y}_{(d)p} = \Sigma_{s_{\text{I}p}} a_{\text{I}i} Y_{(d)i}$. The model groups at the cluster level may correspond to strata of clusters or poststrata of clusters. Then (6.12) corresponds, respectively, to stratified and poststratified ratio estimators. The standard ratio estimator is obtained when the whole cluster population defines the only model group. Note that (6.12) is different from the element level ratio type estimator given by (6.5), although they have the same form. An interesting special case arises if in the model (6.6) we have $x_{pi} = 1$ and $c_i = 1$ for each cluster $i \in U_{\text{I}p}$. Then the g-factors for the sampled clusters are

$$g_{is_{\text{I}p}} = N_{\text{I}p} / \hat{N}_{\text{I}p} = N_{\text{I}p} / \Sigma_{s_{\text{I}p}} a_{\text{I}i} \quad \text{for} \quad i \in s_{\text{I}p} \tag{6.13}$$

where $N_{\text{I}p}$ is the number of clusters in $U_{\text{I}p}$. The estimator of the domain total $Y_{(d)}$ is then given by

$$\hat{Y}_{(d),\text{GREG}} = \sum_{p=1}^{P} \frac{N_{\text{I}p}}{\hat{N}_{\text{I}p}} \hat{Y}_{(d)p}. \tag{6.14}$$

A familiar estimator is produced when the sample selection consists of stratified SRSWOR of clusters and each stratum corresponds to a model group. If the strata are indexed $h = 1, \ldots, H$, we have $a_{\text{I}i} = N_{\text{I}h} / n_{\text{I}h}$ for each $i \in s_{\text{I}h}$ and the g-factor (6.13) is $g_{is_{\text{I}h}} = 1$ for each cluster $i \in s_{\text{I}h}$. Then (6.14) becomes the stratified expansion estimator $\sum_{h=1}^{H} (N_{\text{I}h} / n_{\text{I}h}) \Sigma_{s_{\text{I}h}} Y_{(d)i}$. The case of $H = 1$ implies that there is a single model group equal to the entire population of clusters. We obtain the simple expansion estimator $(N_{\text{I}} / n_{\text{I}}) \Sigma_{s_{\text{I}}} Y_{(d)i}$ where $N_{\text{I}} / n_{\text{I}}$ is the inverse of the cluster sampling rate.

7. Multistage Sampling

In *multistage sampling*, the population of elements is first partitioned into subpopulations called primary sampling units (PSUs) and a probability sample of PSUs is drawn. For the second stage, the set of units within each selected PSU is further

partitioned into second stage units (SSUs). A probability sample of SSUs is then drawn from the PSUs. For a sampling design that has r stages ($r \geqslant 2$), each selected sampling unit at the $(r-1)$th stage is further partitioned into rth stage sampling units (RSUs). Then, a probability sample of RSUs is drawn. The ultimate stage sampling units are not necessarily elements. They can also be clusters of elements and every population element in the selected ultimate stage clusters is then surveyed.

Let the population of elements $U = \{1, \ldots, k, \ldots, N\}$ be partitioned into N_{I} PSUs. A sample s_{I} of m_{I} PSUs is selected from the population of PSUs $U_{\mathrm{I}} = \{1, \ldots, i, \ldots, N_{\mathrm{I}}\}$, with probability $p_{\mathrm{I}}(s_{\mathrm{I}})$. Denote by $\pi_{\mathrm{I}i} = \sum_{s_{\mathrm{I}} \ni i} p_{\mathrm{I}}(s_{\mathrm{I}})$ the first-order PSU inclusion probability induced by the design $p_{\mathrm{I}}(\cdot)$. The sampling weight for the ith PSU is $a_{\mathrm{I}i} = 1/\pi_{\mathrm{I}i}$. The overall sampling weight for element k contained in the ith PSU is given by $a_k = a_{\mathrm{I}i} a_{k|i}$, where $a_{k|i}$ is the sampling weight of k resulting from $(r-1)$ stages of subsampling within the ith PSU. We denote the total realized sample as $s = \cup_{i \in s_{\mathrm{I}}} s_i$ where s_i is the sample of elements resulting from the $(r-1)$ stages of subsampling within the ith PSU.

To derive the variance estimator for the GREG estimator in multistage designs, we need some preliminary results. Let the parameter of interest be the population total $Y = \Sigma_U y_k$. An unbiased estimator of Y is

$$\hat{Y} = \Sigma_{s_{\mathrm{I}}} a_{\mathrm{I}i} \hat{Y}_i \tag{7.1}$$

where $\hat{Y}_i$ is a conditionally unbiased estimator of the ith PSU total. That is, $E(\hat{Y}_i | s_{\mathrm{I}}) = Y_i$ where the conditional expectation is taken over the $(r-1)$ remaining stages of selection, given s_{I}.

Let $f(Y_{s_{\mathrm{I}}})$, where $Y_{s_{\mathrm{I}}} = \{Y_i : i \in s_{\mathrm{I}}\}$, be an unbiased quadratic form estimator of the variance of the HT estimator $\Sigma_{s_{\mathrm{I}}} a_{\mathrm{I}i} Y_i$ under single-stage cluster sampling.

Also, let $V_{is_{\mathrm{I}}} = \mathrm{Var}(\hat{Y}_i | s_{\mathrm{I}})$ be the conditional variance due to the last $(r-1)$ stages of selection, and let $\nu_{is_{\mathrm{I}}}$ be conditionally unbiased estimator of this variance, that is, $E(\nu_{is_{\mathrm{I}}} | s_{\mathrm{I}}) = V_{is_{\mathrm{I}}}$. Note that $\nu_{is_{\mathrm{I}}}$ could depend on s_{I}. Rao (1975) provided a recursive unbiased variance estimator of $\hat{Y}$ given by

$$\widehat{\mathrm{Var}}(\hat{Y}) = f(\hat{Y}_{s_{\mathrm{I}}}) + \Sigma_{s_{\mathrm{I}}} (a_{\mathrm{I}i}^2 - b_{\mathrm{I}i}) \nu_{is_{\mathrm{I}}} \tag{7.2}$$

with

$$f(Y_{s_{\mathrm{I}}}) = \Sigma_{s_{\mathrm{I}}} b_{\mathrm{I}i} \hat{Y}_i^2 + \frac{1}{2} \sum_{i \neq j \in s_{\mathrm{I}}} \sum \gamma_{\mathrm{I}ij} \hat{Y}_i \hat{Y}_j \tag{7.3}$$

where the coefficients $b_{\mathrm{I}i}$ and $\gamma_{\mathrm{I}ij}$ are obtained from the expansion of the unbiased estimator of the variance of $\Sigma_{s_{\mathrm{I}}} a_{\mathrm{I}i} Y_i$ into the form (7.3). Formula (7.2) reduces to the one given by Raj (1966) if $V_{is_{\mathrm{I}}} = V_i$ for all s_{I}.

To illustrate how the coefficients in $f(Y_{s_{\mathrm{I}}})$, are derived, suppose that the first stage design consists of an SRSWOR selection of n_{I} from the N_{I} PSUs in the population. The estimated variance for $\hat{Y} = \Sigma_{s_{\mathrm{I}}} (N_{\mathrm{I}}/n_{\mathrm{I}}) Y_i$ is then given by

$$\widehat{\mathrm{Var}}(\hat{Y}) = N_{\mathrm{I}}^2 \left(\frac{1}{n_{\mathrm{I}}} - \frac{1}{N_{\mathrm{I}}} \right) \frac{1}{n_{\mathrm{I}} - 1} \Sigma_{s_{\mathrm{I}}} (Y_i - \bar{Y})^2$$

with $\bar{Y} = \Sigma_{s_\mathrm{I}} Y_i / n_\mathrm{I}$. Expanding this expression into the form (7.3) we obtain

$$b_{\mathrm{I}i} = \left(\frac{N_\mathrm{I}}{n_\mathrm{I}}\right)^2 \left(1 - \frac{n_\mathrm{I}}{N_\mathrm{I}}\right) \quad \text{and} \quad \gamma_{\mathrm{I}ij} = -2\left(\frac{N_\mathrm{I}}{n_\mathrm{I}}\right)^2 \left(1 - \frac{n_\mathrm{I}}{N_\mathrm{I}}\right) \frac{1}{(n_\mathrm{I} - 1)}.$$

In general, the use of (7.2) and (7.3) can then be described as follows. First, you express $f(Y_{s_\mathrm{I}})$ in the form (7.3). Then, you obtain a copy of it, $f(\hat{Y}_{s_I})$, by replacing each Y_i with its conditionally unbiased estimator $\hat{Y}_i$. In the estimated variance (7.2), it only remains to specify the conditional variance estimator ν_{is_I} for $i \in s_\mathrm{I}$. This can be done by recognizing that at each stage there is a conditional variance. The process can be repeated iteratively with a new function $f(\cdot)$ at each stage of selection. The estimated variance (7.2) can be computed using a procedure provided by Bellhouse (1985), which uses a general tree construction algorithm to represent the multistage design.

Surveys based on multistage sampling often use auxiliary data to improve the efficiency of their estimates. For instance, the Canadian Labour Force Survey uses a multistage stratified cluster design and calibrates its estimates using known age-sex and subprovincial counts. The use of auxiliary information can be translated into regression models that relate the variable of interest to a known auxiliary vector. Auxiliary information may be known for any of the populations at the various stages of selection. The partitioning of these populations into model groups can also occur at any of the stages, if the necessary auxiliary information exists for these model groups. Furthermore, the partitioning must be reasonable. In the following, the discussion is restricted to the use of model groups at the element level (case A) and at the PSU level (case B). An example of the use of model groups both at the PSU level and at the element level can be found in the weighting procedures used for the Canadian Labour Force Survey. PSUs are grouped into urban and rural clusters. Elements (individuals) are grouped by age-sex, by Census Metropolitan Area and by Economic Region.

Case A: Model groups at the element level. The features of this case are as described in Section 6. The model for group U_p is given by (4.1) and g-factors are calculated as in the cluster sampling case by (4.5). The GREG estimator is given by (6.1) and the associated residuals by (6.2) where $\hat{\mathbf{B}}_p$ satisfies (4.2). Recalling that $a_k = a_{\mathrm{I}i} a_{k|i}$, the variance estimator for $\hat{Y}_{\mathrm{GREG}}$ is now obtained with the aid of (7.2) and (7.3) as

$$\widehat{\mathrm{Var}}(\hat{Y}_{\mathrm{GREG}}) = f(\hat{E}_{s_\mathrm{I}}) + \Sigma_{s_\mathrm{I}} (a_{\mathrm{I}i}^2 - b_{\mathrm{I}i}) \nu^*_{is_\mathrm{I}} \tag{7.4}$$

where

$$f(\hat{E}_{s_\mathrm{I}}) = \Sigma_{s_\mathrm{I}} b_{\mathrm{I}i} \hat{E}_i^2 + \frac{1}{2} \sum_{i \neq j \in s_\mathrm{I}} \sum \gamma_{\mathrm{I}ij} \hat{E}_i \hat{E}_j$$

and $\hat{E}_{s_\mathrm{I}} = \{\hat{E}_i : i \in s_\mathrm{I}\}$ with $\hat{E}_i = \Sigma_{s_i} a_{k|i} z_k$, where

$$z_k = g_{ks_p} e_k = g_{ks_p} (y_k - \mathbf{x}_k' \hat{\mathbf{B}}_p) \quad \text{for} \quad k \in s_p. \tag{7.5}$$

The coefficients $b_{\mathrm{I}i}$ and $\gamma_{\mathrm{I}ij}$ in $f(\hat{E}_{s_\mathrm{I}})$ are identical with those that appear in $f(\hat{Y}_{s_\mathrm{I}})$ in equation (7.3). In (7.4), $\nu^*_{is_\mathrm{I}}$ is an approximately unbiased estimator of $V^*_{is_\mathrm{I}} = \mathrm{Var}(\hat{E}_i | s_\mathrm{I})$, which is obtained from ν_{is_I} in equation (7.2) by replacing y_k by z_k.

The GREG estimator for a domain total $Y_{(d)}$ is obtained by replacing y_k by $y_{(d)k}$ in (6.1). Note that this replacement is also carried out in (4.2), (7.4) and (7.5) to obtain the corresponding variance estimator.

Case B: Model groups at the PSU level. The features of this case are as specified in Section 6, except that a cluster is now called a PSU. The model for group $U_{\mathrm{I}p}$ is given by (6.6). However, unlike in Section 6, the PSU total Y_i is not known in this case, but must be estimated by $\hat{Y}_i = \Sigma_{s_i} a_{k|i} y_k$ as a result of the $(r-1)$ subsequent selection stages, so the GREG estimator of Y is now

$$\hat{Y}_{\mathrm{GREG}} = \sum_{p=1}^{P} \Sigma_{s_{\mathrm{I}p}} a_{\mathrm{I}i} g_{is_{\mathrm{I}p}} \hat{Y}_i. \tag{7.6}$$

The corresponding variance estimator is

$$\widehat{\mathrm{Var}}(\hat{Y}_{\mathrm{GREG}}) = f(\hat{Z}_{s_{\mathrm{I}}}) + \Sigma_{s_{\mathrm{I}}}(a_{\mathrm{I}i}^2 - b_{\mathrm{I}i})\nu_{is_{\mathrm{I}}} \tag{7.7}$$

where $f(\hat{Z}_{s_{\mathrm{I}}}) = \Sigma_{s_{\mathrm{I}}} b_{\mathrm{I}i} \hat{Z}_i^2 + \frac{1}{2} \sum_{i \neq j \in s_{\mathrm{I}}} \sum \gamma_{\mathrm{I}ij} \hat{Z}_i \hat{Z}_j$ and $\hat{Z}_{s_{\mathrm{I}}} = \{\hat{Z}_i : i \in s_{\mathrm{I}}\}$ with

$$\hat{Z}_i = g_{is_{\mathrm{I}p}}(\hat{Y}_i - \mathbf{x}'_{pi}\hat{\mathbf{B}}_{\mathrm{I}p}) \tag{7.8}$$

where $\hat{\mathbf{B}}_{\mathrm{I}p}$ is given by (6.8) provided Y_i is replaced by its estimate $\hat{Y}_i$. The coefficients $b_{\mathrm{I}i}$ and $\gamma_{\mathrm{I}ij}$ in $f(\hat{Z}_{s_{\mathrm{I}}})$ are identical with those given in $f(Y_{s_{\mathrm{I}}})$ in equation (7.3). Note that $\nu_{is_{\mathrm{I}}}$ is an unbiased estimator of the conditional variance $V_{is_{\mathrm{I}}} = \mathrm{Var}(\hat{Y}_i | s_{\mathrm{I}})$.

A regression estimator $\hat{Y}_{(d),\mathrm{GREG}}$ for a domain total $Y_{(d)}$ is obtained by replacing $\hat{Y}_i$ by $\hat{Y}_{(d)i} = \Sigma_{s_i} a_{k|i} y_{(d)k}$ in (7.6). The corresponding variance estimator, $\widehat{\mathrm{Var}}(\hat{Y}_{(d),\mathrm{GREG}})$, is obtained by carrying out the same replacement in (7.7) and (7.8), and by replacing functions of the y_k's in $\nu_{is_{\mathrm{I}}}$ by the same functions of the $y_{(d)k}$'s.

8. Estimation of Non-Linear Parameters

So far, only estimators of population totals have been discussed. By using the Taylor linearization method, the results for regression estimation and the corresponding variance estimation can be extended to parameters composed as non-linear functions of two or more totals. These include, for example, ratios of totals, regression coefficients, correlation coefficients, etc. Examples of the Taylor linearization method for non-linear parameters are given by Tepping (1968) and Woodruff (1971). In other words, we combine the GREG estimation technique with the Tepping-Woodruff procedure for non-linear parameters.

Suppose that the parameter to be estimated for population U is θ and that θ can be expressed as a function of the Q population totals in the vector $\mathbf{Y} = (Y_1, \ldots, Y_q, \ldots, Y_Q)$, where $Y_q = \Sigma_U y_{qk}$. That is,

$$\theta = F(\mathbf{Y}).$$

An estimator for θ that uses auxiliary information can be obtained by replacing each unknown total Y_q by its corresponding GREG estimator, $\hat{Y}_{q,\mathrm{GREG}}$. Letting

$$\hat{\mathbf{Y}}_{\mathrm{GREG}} = (\hat{Y}_{1,\mathrm{GREG}}, \ldots, \hat{Y}_{q,\mathrm{GREG}}, \ldots, \hat{Y}_{Q,\mathrm{GREG}})$$

then the resulting estimator of θ can be expressed as

$$\hat{\theta} = F(\hat{\mathbf{Y}}_{\text{GREG}}). \tag{8.1}$$

For the Taylor expansion of (8.1), we find the partial derivatives of F, evaluated at the approximate expected value point, that is, for $q = 1, \ldots, Q$,

$$D_q = \left. \frac{\partial F}{\partial \hat{Y}_{q,\text{GREG}}} \right|_{\hat{Y}_{\text{GREG}} = Y_{\text{GREG}}}$$

In D_q, replace each unknown total by its HT estimator to obtain $\hat{D}_q$. Then, supposing case A of Section 6 applies (modelling at the element level), we compute

$$\hat{u}_k = \sum_{q=1}^{Q} \hat{D}_q g_{ks_p} (y_{qk} - \mathbf{x}_k' \hat{\mathbf{B}}_{qp})$$

where p is the model group (poststratum) index, and g_{ks_p} and $\hat{\mathbf{B}}_{qp}$ denote the appropriate g-factors and regression vectors. To compute the variance estimator $\widehat{\text{Var}}(\hat{\theta}) = \widehat{\text{Var}}\{F(\hat{Y}_{\text{GREG}})\}$, we would now simply replace E_i in (6.3) by $E_i' = \Sigma_{k \in s_i} \hat{u}_k$.

9. Conclusions

In this paper we have presented the methodological principles that were used in the development of the Generalized Estimation System (GES) at Statistics Canada. GES was developed to produce domain estimates and corresponding estimates of variance for population parameters such as totals, averages, ratios, and proportions for a variety of sampling designs. The currently programmed specifications can handle stratified, single-stage sample designs such as stratified simple random sampling without replacement, stratified cluster sampling, and stratified probability-proportional-to-size, with and without replacement, (PPS) sampling. These sample designs are common at Statistics Canada. GES currently handles the estimation for over 20 major business surveys, and several social surveys, including the Canadian Labour Force Survey at Statistics Canada.

An important aspect of the GES is the use of auxiliary information in the form of known auxiliary variable totals (Särndal, Swensson, and Wretman 1992). For single-stage element sampling, the known totals always refer to the population of elements, or to specified subgroups of this population. For single-stage cluster sampling and for sampling in two or more stages, auxiliary information can appear at different levels corresponding to the different populations that can be distinguished. For example, in single-stage cluster sampling, we may have known auxiliary totals for the population of clusters (or for subgroups of it), or known auxiliary totals for the population of elements (or for subgroups of it). A population subgroup with a known auxiliary total is called a model group. For each model group, a general linear regression model can be stated and fitted with the variable of interest as criterion and the auxiliary variables as predictors. In the GES, the term GREG model refers to the collection of regressions fitted in the various groups.

The system is in modular form so that it can easily accommodate additional estimators and designs. It was developed in micro-SAS environment. The most recent

version (GES3.1) runs under SAS 6.08 (Beta) for Windows and needs SAS features such as BASE, AF, FSP, and IML. The system is menu driven. For example, the user is allowed to choose the estimator to be used for given parameters of interest. A model statement is used to define an estimator.

Hidiroglou and Särndal (1995) have extended this methodology to handle arbitrary two-phase sampling designs, where auxiliary information plays an important role. Also, as imputed data have an effect on variance estimation, we plan to include it using methods given in Särndal (1990), and Rancourt, Särndal, and Lee (1994).

10. References

Andersson, C. and Nordberg, L. (1994). A Method for Variance Estimation of Non-Linear Functions of Totals in Surveys – Theory and Software Implementation. Journal of Official Statistics, 10, 395–406.

Bellhouse, D.R. (1985). Computing Methods for Variance Estimation in Complex Surveys. Journal of Official Statistics, 1, 323–329.

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

Choudhry, G.H. (1988). Generalized Estimation and Variance Estimation System for Sub-Annual Surveys. Statistics Canada methodology report, April 1988.

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

Dumais, J. and Carpenter, R. (1988). Methodology of the Generalized Estimation System for Sub-Annual Business Surveys. Statistics Canada methodology report, September.

Estevao, V. (1991). Generalized Estimation System, Methodology Review. Statistics Canada informatics report, September.

Hidiroglou, M.A. (1991). Structure of the Generalized Estimation System (GES). Statistics Canada methodology report, September.

Hidiroglou, M.A., Fuller, W.A., and Hickman, R.D. (1976). SUPER CARP. Statistical Laboratory, Iowa State University, Ames, Iowa.

Hidiroglou, M.A., and Särndal, C.E. (1995). Use of Auxiliary Information for Two-Phase Sampling. Paper presented at the Annual American Statistical Association meetings held in Orlando, Florida.

Kott, P.S. (1990). Estimating the Conditional Variance of a Design Consistent Regression Estimator. Journal of Statistical Planning and Inference, 24, 287–296.

Lavallée, P. and Leblond, Y. (1990). Système Général d'Estimation, Spécifications. Statistics Canada methodology report, July 1990.

Outrata, E. and Chinnappa, B.N. (1989). General Survey Functions Design at Statistics Canada. Bulletin of the International Statistical Institute, 53: 2, 219–238.

Raj, D. (1966). Some Remarks on a Simple Procedure of Sampling without Replacement. Journal of the American Statistical Association, 61, 391–396.

Rancourt, E., Särndal, C.E., and Lee, H. (1994). Estimation of Variance in Presence

of Nearest Neighbour Imputation. Paper presented at the American Statistical Association meetings held in Toronto.

Rao, J.N.K. (1975). Unbiased Variance Estimation for Multistage Designs. Sankhyā C., 37, 133–139.

Särndal, C.E. (1990). Estimation in the Generalized Estimation System. Statistics Canada methodology report, January.

Särndal, C.E. (1990). Methods for Establishing the Precision of Survey Estimates when Imputation Has Been Used. Proceedings of Statistics Canada Symposium '90: Measurement and Improvement of Data Quality, 369–380.

Särndal, C.E., Swensson, B., and Wretman, J.H. (1989). The Weighted Residual Technique for Estimating the Variance of the General Regression Estimator of the Finite Population Total. Biometrika, 76, 527–537.

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

Shah, B.V., Lavange, L.M., Barnwell, B.G., Killinger, J.E., and Wheeless, S.C. (1989). SUDAAN: Procedures for Descriptive Statistics Users' Guide. Research Triangle Institute Report.

Schnell, D., Kennedy, W.J., Sullivan, G., Park, J.P., and Fuller, W.A. (1988). Personal Computer Variance Software for Complex Surveys. Survey Methodology, 14, 59–69.

Tepping, B.J. (1968). Variance Estimation in Complex Surveys. Proceedings of the Social Statistics Section, American Statistical Association, 11–18.

Woodruff, R.S. (1971). A Simple Method for Approximating the Variance of a Complicated Estimate. Journal of the American Statistical Association 66, 411–414.

Received June 1993
Revised May 1995