Journal of Official Statistics, Vol. 18, No. 4, 2002, pp. 577–589

Asymptotically Efficient Generalised Regression Estimators

G.E. Montanari and M.G. Ranalli[1]

In this article we introduce an enlarged class of generalised regression estimators of a finite population mean that includes as special case the optimal estimator. Theoretical results show that the latter can be seen as a generalised regression estimator based on a suitable superpopulation linear regression model. Then, an estimation procedure able to merge the large sample efficiency of the optimal estimator with the greater stability of the generalised regression estimator for samples of moderate size is proposed. A simulation study provides empirical evidence in support of the quoted theory.

Key words: Survey sampling; regression estimator; superpopulation model.

1. Introduction

Regression estimation is a powerful technique for estimating finite population means or totals of survey variables when the population means or totals of a set of auxiliary variables are known. Two well-known types of regression estimators have recently appeared in the literature, namely the Generalised Regression Estimator (GRE) and the Optimal Estimator (OPE). Till now, they have been studied separately (for a short review see Montanari 1998). In this article we explore connections between the two types of regression estimators and establish conditions under which the GRE is exactly or asymptotically equivalent to the OPE. To this end, an Extended Generalised Regression Estimator (EGRE) is proposed to allow regression estimation based on superpopulation models with a nondiagonal variance matrix. Theoretical results show that the OPE can be seen as an EGRE which incorporates the auxiliary variables used at the sampling design stage. Thus, the OPE uses a larger number of auxiliary variables than the GRE and, as a consequence, it may be less stable than the GRE when the sample size is not large. We then propose a new estimation procedure which allows for a more efficient use of the auxiliary information coming both from the sampling scheme and alternative sources. The core of the proposal is a compromise between the large sample efficiency of the OPE and the superior stability of the GRE for smaller sample sizes. The new strategy is shown through simulation studies to be robust over several different superpopulation models and on average more efficient than the OPE or GRE alone.

[1] Department of Statistical Sciences, University of Perugia, Via A. Pascoli, 06100 Perugia, Italy. Email: giorgio@stat.unipg.it

2. Background and Preliminary Definitions

Consider a finite population $U = \{u_1, u_2, \ldots, u_N\}$, where the i-th unit is represented by its label i. Let Y_i and $\mathbf{x}_i = (X_{1i}, X_{2i}, \ldots, X_{qi})'$ be the values of the survey variable y and of a q-dimensional auxiliary variable $\boldsymbol{x}$ associated with unit i. The population mean vector $\bar{\mathbf{x}} = \sum_{i=1}^{N} \mathbf{x}_i/N$ of $\boldsymbol{x}$ is assumed known, e.g., from administrative registers or census data. The unknown y variable mean, $\bar{Y} = \sum_{i=1}^{N} Y_i/N$, has to be estimated by means of a sample s of size n drawn from U according to a probabilistic sampling design and taking into account the knowledge of $\bar{\mathbf{x}}$. We assume here that $\mathbf{x}_i$ is known for units in the sample but not in the population and that, for consistency with external sources, the estimator of $\bar{Y}$ to be adopted must use the known values contained in $\bar{\mathbf{x}}$ when applied to the auxiliary variables. So, in our perspective the vector $\bar{\mathbf{x}}$ is taken as given, as well as the sampling design. In this article we are not going to discuss either the choice of subsets of elements of $\bar{\mathbf{x}}$ for estimation purposes or the design of the sampling scheme.

Let $\hat{\bar{Y}} = \sum_{i=1}^{n} Y_i/N\pi_i$ and $\hat{\bar{\mathbf{x}}} = \sum_{i=1}^{n} \mathbf{x}_i/N\pi_i$ be the design-unbiased Horvitz-Thompson estimators of $\bar{Y}$ and $\bar{\mathbf{x}}$, respectively, where π_i $(i = 1, \ldots, N)$ is the first order inclusion probability of the sampling design. The most common way of taking into account the knowledge of the auxiliary variable population means is to adopt the regression estimator $\hat{\bar{Y}}_R = \hat{\bar{Y}} + \hat{\boldsymbol{\beta}}'(\bar{\mathbf{x}} - \hat{\bar{\mathbf{x}}})$, where $\hat{\boldsymbol{\beta}}$ is a vector of regression coefficients, given by some function of sample data $\{(Y_i, \mathbf{x}_i); i = 1, \ldots, n\}$. The class of regression estimators contains well-known estimators, such as ratio and product estimators, ratio and product estimators with linearly transformed auxiliary variables, post-stratified and regression estimators. So, the main issue the statistician has to deal with is the definition of $\hat{\boldsymbol{\beta}}$.

Let us denote by $\mathbf{W}$ the $N \times N$-matrix whose ij-th entry is $(\pi_{ij} - \pi_i \pi_j)/N^2 \pi_i \pi_j$, where π_{ij} $(i, j = 1, \ldots, N)$ is the second order inclusion probability of the sampling design, and $\pi_{ii} = \pi_i$. Assembling the values of y and $\boldsymbol{x}$ into an N-vector $\mathbf{Y}$ and an $N \times q$-matrix $\mathbf{X}$ having $\mathbf{x}_i'$ on its i-th row, the variances of $\hat{\bar{Y}}$ and $\hat{\bar{\mathbf{x}}}$ and the covariance between them are given by $\mathrm{V}(\hat{\bar{Y}}) = \mathbf{Y}'\mathbf{W}\mathbf{Y}$, $\mathrm{V}(\hat{\bar{\mathbf{x}}}) = \mathbf{X}'\mathbf{W}\mathbf{X}$ and $\mathrm{C}(\hat{\bar{Y}}, \hat{\bar{\mathbf{x}}}) = \mathbf{X}'\mathbf{W}\mathbf{Y}$, respectively.

In the sequel, we assume that the sampling design and the population are such that nonlinear estimators converge in probability to their first order Taylor linear approximations when the sample size and the population size approach infinity. We will term as "asymptotic" any property that depends upon this convergence.

3. The Optimal Estimator

This type of regression estimator is obtained from the difference estimator, i.e., $\tilde{\bar{Y}}_O = \hat{\bar{Y}} + \mathbf{B}'(\bar{\mathbf{x}} - \hat{\bar{\mathbf{x}}})$, where $\mathbf{B}$ is a vector of constants, and $\tilde{\bar{Y}}_O$ is unbiased for $\bar{Y}$ and has a variance given by $\mathrm{V}(\tilde{\bar{Y}}_O) = \mathrm{V}(\hat{\bar{Y}}) + \mathbf{B}'\mathrm{V}(\hat{\bar{\mathbf{x}}})\mathbf{B} - 2\mathbf{B}'\mathrm{C}(\hat{\bar{Y}}, \hat{\bar{\mathbf{x}}})$. The latter is minimised by assuming $\mathbf{B} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{Y}$ (Montanari 1987). When $\mathbf{X}'\mathbf{W}\mathbf{X}$ is singular and its rank is $q' < q$, to define $\mathbf{B}$ one or more entries of $\boldsymbol{x}$, hence of $\bar{\mathbf{x}}$, have to be dropped in such a way as to obtain a $q' \times q'$ non-singular variance matrix $\mathbf{X}'\mathbf{W}\mathbf{X}$.

The optimum value of $\mathbf{B}$ can be estimated in many ways. For our purposes, if we take Horvitz-Thompson estimators of variances and covariances between unbiased estimators, under mild conditions on the sampling design, a consistent estimator of $\mathbf{B}$ is given by $\hat{\mathbf{B}} = (\mathbf{X}_s'\mathbf{W}_{ss}\mathbf{X}_s)^{-1}\mathbf{X}_s'\mathbf{W}_{ss}\mathbf{Y}_s$, where $\mathbf{W}_{ss} = \{(\pi_{ij} - \pi_i \pi_j)/N^2 \pi_i \pi_j \pi_{ij}\}_{i,j=1,\ldots,n}$; $\mathbf{X}_s = \{\mathbf{X}_i'\}_{i=1,\ldots,n}$; $\mathbf{Y}_s = \{Y_i\}_{i=1,\ldots,n}$. Then, replacing $\mathbf{B}$ by $\hat{\mathbf{B}}$ in $\tilde{\bar{Y}}_O$, we get

$\hat{\hat{Y}}_O = \hat{\hat{Y}} + \hat{\mathbf{B}}'(\bar{\mathbf{x}} - \hat{\hat{\mathbf{x}}})$, which has been called the Optimal Estimator by Rao (1994). This estimator shares in large samples the properties of $\tilde{\tilde{Y}}_O$, the latter being the first order Taylor linear approximation of the former (Montanari 1987). So, $\hat{\hat{Y}}_O$ is asymptotically design unbiased and has the minimum asymptotic variance among all regression estimators based on the same auxiliary information $\bar{\mathbf{x}}$.

Other properties of the OPE are: 1) the means of the auxiliary variables estimated through $\hat{\hat{Y}}_O$ equal the corresponding population means, i.e., $\hat{\hat{\mathbf{x}}}_O = \bar{\mathbf{x}}$; 2) when there is more than one survey variable, $\hat{\hat{Y}}_O$ can be expressed as a simple weighted estimator with the same weights applying to all variables (Montanari 1998); 3) $\hat{\hat{Y}}_O$ gives valid conditional inferences (Rao 1997). Note that the asymptotic optimality of the OPE is a strictly design based property.

The main drawback of $\hat{\hat{Y}}_O$ is that it is complex to compute and may be unstable in finite size samples, because it requires estimating sampling variances and covariances (Casady and Valliant 1993). However, if an adequate number of degrees of freedom g is available for estimating $\mathbf{B}$, the problem can be overcome. For example, for standard complex multistage sampling designs with replacement sampling at the first stage, g can be roughly taken as the number of sample first stage clusters minus the number of strata (Lehtonen and Pahkinen 1995, p. 181; for more elaboration on this topic see Eltinge and Jang 1996). A stable estimated $\mathbf{B}$ may be expected when g is large enough relative to the dimension q of the auxiliary variable $\boldsymbol{x}$.

4. The Generalised Regression Estimator

Most popular estimators of a finite population mean or total belong to the class of the generalised regression estimators. Such a class is described in Särndal, Swensson, and Wretman (1992; Chap. 6) and in Estevao, Hidiroglou, and Särndal (1995). A GRE is based on an underlying superpopulation linear regression model relating the survey variable to the auxiliary variables whose population means are known. Consider the model $\mathrm{E}_m(\mathbf{Y}) = \mathbf{X}'\boldsymbol{\beta}$, $\mathrm{V}_m(\mathbf{Y}) = \sigma^2\boldsymbol{\Sigma}$, where $\boldsymbol{\Sigma} = diag\{\eta_i\}_{i=1,\ldots,N}$ is a known matrix. Note that E_m, V_m and C_m denote the expected value, variance and covariance with respect to the model. Let $\hat{\boldsymbol{\beta}}_N = (\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}\boldsymbol{\Sigma}^{-1}\mathbf{Y}$ be the census weighted least squares regression estimator of $\boldsymbol{\beta}$. Then, replacing $\hat{\boldsymbol{\beta}}_N$ by the consistent estimator $\hat{\boldsymbol{\beta}}_n = (\mathbf{X}_s'\boldsymbol{\Sigma}_{ss}^{-1}\mathbf{X}_s)^{-1}\mathbf{X}_s\boldsymbol{\Sigma}_{ss}^{-1}\mathbf{Y}_s$, where $\boldsymbol{\Sigma}_{ss} = diag\{\eta_i\pi_i\}_{i=1,\ldots,n}$, the corresponding GRE is defined to be $\hat{\hat{Y}}_G = \hat{\hat{Y}} + \hat{\boldsymbol{\beta}}_n'(\bar{\mathbf{x}} - \hat{\hat{\mathbf{x}}})$. In the sequel we will call the model upon which the GRE is based the ‘‘working model’’.

The large sample properties of $\hat{\hat{Y}}_G$ can be established by means of its first order Taylor linear approximation $\tilde{\tilde{Y}}_G = \hat{\hat{Y}} + \hat{\boldsymbol{\beta}}_N'(\bar{\mathbf{x}} - \hat{\hat{\mathbf{x}}})$ (Särndal, Swensson, and Wretman 1992; p. 235). In particular, $\hat{\hat{Y}}_G$ is asymptotically design unbiased, and when the working model holds true it has the minimum expected asymptotic design variance with respect to the model, i.e., for any other design unbiased or approximately design unbiased estimator $\hat{\hat{Y}}^*$ of $\bar{Y}$, $\mathrm{E}_m\mathrm{V}(\tilde{\tilde{Y}}_G) \leq \mathrm{E}_m\mathrm{V}(\hat{\hat{Y}}^*)$, for all $\boldsymbol{\beta}$ (Wright 1983). Montanari (1998) proved that when the working model holds true, asymptotically $\mathrm{V}(\tilde{\tilde{Y}}_G) = \mathrm{V}(\tilde{\tilde{Y}}_O)$. In contrast, if the model is wrongly specified, the value of the asymptotic variance of $\hat{\hat{Y}}_G$ may be appreciably higher than that of $\hat{\hat{Y}}_O$ based on the same auxiliary information $\bar{\mathbf{x}}$. This event is not uncommon, as the specification of the model is limited by the availability of the population means of auxiliary variables.

Other properties of the GRE are: 1) the means of the auxiliary variables estimated through $\hat{\hat{Y}}_G$ equal the corresponding population means, i.e., $\hat{\hat{\mathbf{x}}}_G = \bar{\mathbf{x}}$; 2) when there is more than one survey variable, $\hat{\hat{Y}}_G$ can be expressed as a simple weighted estimator with the same weights applying to all variables; 3) $\hat{\hat{Y}}_G$ gives valid conditional inferences provided that the model holds true. Note that the asymptotic optimality of the GRE requires the model to be true, and concerns the average asymptotic variance over the finite populations that can be generated under the model. Hence, the GRE efficiency is vulnerable to model misspecifications.

5. An Extended Class of GRE's

In this section we explore the relationships between GRE's and the OPE. To this end, let us enlarge the GRE class to fit the non-diagonal variance matrix of the working model. Consider the model $\mathrm{E}_m(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$ and $\mathrm{V}_m(\mathbf{Y}) = \sigma^2\boldsymbol{\Sigma}$, where $\boldsymbol{\Sigma} = \{v_{ij}\}_{i,j=1,\dots,N}$ is now a positive definite symmetric $N \times N$-matrix. Let $\boldsymbol{\Sigma}_{ss}^{-1}$ be the symmetric matrix that has v^{ij}/π_{ij} as ij-th entry, where $i, j = 1, \dots, n$ and v^{ij} is the ij-th entry of $\boldsymbol{\Sigma}^{-1}$. Then, provided that the matrix $(\mathbf{X}'_s\boldsymbol{\Sigma}_{ss}^{-1}\mathbf{X}_s)^{-1}$ exists for all samples s, the corresponding Extended Generalised Regression Estimator (EGRE) can be written $\hat{\hat{Y}}_{EG} = \hat{\hat{Y}} + \hat{\boldsymbol{\beta}}'_n(\bar{\mathbf{x}} - \hat{\hat{\mathbf{x}}})$, where $\hat{\boldsymbol{\beta}}_n = (\mathbf{X}'_s\boldsymbol{\Sigma}_{ss}^{-1}\mathbf{X}_s)^{-1}\mathbf{X}'_s\boldsymbol{\Sigma}_{ss}^{-1}\mathbf{Y}_s$. Observe that the entries of $\mathbf{X}'_s\boldsymbol{\Sigma}_{ss}^{-1}\mathbf{X}_s$ and $\mathbf{X}'_s\boldsymbol{\Sigma}_{ss}^{-1}\mathbf{Y}_s$ are design unbiased estimators of the corresponding entries of $\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X}$ and $\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{Y}$, respectively, provided that $v^{ij} \neq 0$ implies $\pi_{ij} \neq 0$. Thus, under mild conditions on the second order inclusion probabilities, $\hat{\boldsymbol{\beta}}_n$ converges in probability to $\hat{\boldsymbol{\beta}}_N = (\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{Y}$, which is the census weighted least squares regression estimator of $\boldsymbol{\beta}$. Obviously, when $\boldsymbol{\Sigma}$ is a diagonal matrix, the EGRE, $\hat{\hat{Y}}_{EG}$, reduces to the customary GRE, $\hat{\hat{Y}}_G$.

Note that when $\mathbf{W}$ is non-singular, the OPE belongs to the EGRE class defined above setting $\boldsymbol{\Sigma} = \mathbf{W}^{-1}$. However, generally $\mathbf{W}$ is only non-negative definite, being singular for many sampling designs. But even in such a case, there are EGRE's that are asymptotically equivalent to the OPE, as we show in the next section.

6. Connecting EGRE's and OPE's

Let us by Design Balanced Variable (DBV) denote any non-null auxiliary variable z whose mean is estimated without error by the Horvitz-Thompson estimator, i.e., $\hat{\hat{Z}} = \sum_{i=1}^{n} Z_i/N\pi_i = \bar{Z}$. This type of variable plays a fundamental role in establishing conditions under which the OPE is equivalent to an EGRE. In fact, assembling the population values Z_i of a DBV variable into the N-vector $\mathbf{Z}$, we have the following theorems.

Theorem 1. A variable z is a DBV if and only if the vector $\mathbf{Z}$ belongs to the subspace orthogonal to that spanned by the columns of $\mathbf{W}$.

Proof. If $\mathbf{Z}$ is a DBV, it follows that $\mathbf{Z}'\mathbf{W}\mathbf{Z} = 0$ and $\mathbf{C}'\mathbf{W}\mathbf{Z} = 0$ for any vector $\mathbf{C}$, since the covariance between the unbiased estimators of the means of a DBV and any other variable is identically zero. Hence, $\mathbf{W}\mathbf{Z} = \mathbf{0}$. On the other hand, if $\mathbf{W}\mathbf{Z} = \mathbf{0}$ holds, then $\mathbf{Z}'\mathbf{W}\mathbf{Z} = 0$, i.e., $\mathbf{Z}$ is a DBV.

From Theorem 1 follows that the subspace spanned by the DBV's has the dimension $N - r(\mathbf{W})$, where $r(\cdot)$ denotes the rank of a matrix. So, $r(\mathbf{W}) = N - t$, where $t \geq 0$, implies that there are t linearly independent DBV's. Now, let $\mathbf{Z}$ be an $N \times t$ matrix containing t linearly independent DBV's and assume that $\mathbf{X}$ does not contain any DBV. Then, we have the following theorem.

Theorem 2. Consider the working model $\mathrm{E}_m(\mathbf{Y}) = (\mathbf{ZX})\boldsymbol{\beta}$ and $\mathrm{V}_m(\mathbf{Y}) = \sigma^2\boldsymbol{\Sigma}$ and the matrix $\mathbf{W}$ corresponding to the sampling design in use. If $\boldsymbol{\Sigma}$ is a variance matrix for which there exists a scalar such that $\boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}\mathbf{Z}(\mathbf{Z}'\boldsymbol{\Sigma}^{-1}\mathbf{Z})^{-1}\mathbf{Z}'\boldsymbol{\Sigma}^{-1} = \alpha\mathbf{W}$, then $\tilde{\tilde{Y}}_{EG} = \tilde{\tilde{Y}}_O$, i.e., the EGRE based on the assumed working model is asymptotically equivalent to the OPE based on the same auxiliary information $\bar{\mathbf{x}}$.

Proof. To prove the result it is sufficient to rewrite $(\mathbf{ZX})\boldsymbol{\beta} = \mathbf{Z}\boldsymbol{\beta}_z + \mathbf{X}\boldsymbol{\beta}_x$, where $\boldsymbol{\beta}_z$ and $\boldsymbol{\beta}_x$ are the vector of regression coefficients for the DBV's and the auxiliary variables, respectively. Then, after some algebra, the census weighted least squares estimator of $\boldsymbol{\beta}_x$ is given by $\hat{\boldsymbol{\beta}}_{xN} = (\mathbf{X}'\mathbf{AX})^{-1}\mathbf{X}'\mathbf{AY}$, where $\mathbf{A} = \boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}\mathbf{Z}(\mathbf{Z}'\boldsymbol{\Sigma}^{-1}\mathbf{Z})^{-1}\mathbf{Z}'\boldsymbol{\Sigma}^{-1}$. Since $r(\boldsymbol{\Sigma}) = N$, by well-known matrix algebra results we have $r(\mathbf{A}) = N - r\,(\mathbf{Z}) = N - t$. Thus, if there exists a scalar α such that $\alpha\mathbf{W} = \mathbf{A}$, it follows that $\hat{\boldsymbol{\beta}}_{xN} = \mathbf{B}$, hence $\tilde{\tilde{Y}}_{EG} = \tilde{\tilde{Y}}_O$ as $\tilde{\tilde{Y}}_{EG} - \tilde{\tilde{Y}}_O = (\hat{\boldsymbol{\beta}}_{xN} - \mathbf{B})(\bar{\mathbf{x}} - \hat{\hat{\mathbf{x}}})$.

Theorem 2 gives a sufficient condition for the asymptotic equality between an EGRE and the OPE that uses the same auxiliary variable $\boldsymbol{x}$, i.e., besides the auxiliary variables that are not DBV's, the working model should include a number $N - r(\mathbf{W})$ of linearly independent DBV's and the variance matrix $\sigma^2\boldsymbol{\Sigma}$ should be set so that $\mathbf{A}$ is a matrix proportional to $\mathbf{W}$ ($\mathbf{A} \propto \mathbf{W}$). The outcome is an asymptotic minimum variance estimator, irrespective of the working model goodness, given the amount of auxiliary information $\bar{\mathbf{x}}$.

Generally speaking, with finite sample size, the OPE is approximately equal to an EGRE based on a working model that includes the effect of any existing DBV's. Unfortunately, the theorem does not provide guidelines for determining the structure of the matrix $\boldsymbol{\Sigma}$ and the DBV's that correspond to the sampling design in use. In the next section, we will examine a number of case studies where solutions are available.

The next theorem assures the finite size sample identity between an EGRE and the OPE.

Theorem 3. If

$$\boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}\mathbf{Z}(\mathbf{Z}'\boldsymbol{\Sigma}^{-1}\mathbf{Z})^{-1}\mathbf{Z}'\boldsymbol{\Sigma}^{-1} \propto \mathbf{W}$$

implies

$$\boldsymbol{\Sigma}_{ss}^{-1} - \boldsymbol{\Sigma}_{ss}^{-1}\mathbf{Z}_s(\mathbf{Z}_s'\boldsymbol{\Sigma}_{ss}^{-1}\mathbf{Z}_s)^{-1}\mathbf{Z}_s'\boldsymbol{\Sigma}_{ss}^{-1} \propto \mathbf{W}_{ss}$$

for all possible samples s, then $\hat{\hat{Y}}_{EG} = \hat{\hat{Y}}_O$.

Proof. To prove the result it is sufficient to write $(\mathbf{Z}_s\mathbf{X}_s)\boldsymbol{\beta} = \mathbf{Z}_s\boldsymbol{\beta}_z + \mathbf{X}_s\boldsymbol{\beta}_x$ and to note that the sample weighted least squares estimator of $\boldsymbol{\beta}_x$ with respect to the weight matrix $\boldsymbol{\Sigma}_{ss}^{-1}$ is given by $\hat{\boldsymbol{\beta}}_x = (\mathbf{X}'\mathbf{A}_{ss}\mathbf{X})^{-1}\mathbf{X}'\mathbf{A}_{ss}\mathbf{Y}$, where $\mathbf{A}_{ss} = \boldsymbol{\Sigma}_{ss}^{-1} - \boldsymbol{\Sigma}_{ss}^{-1}\mathbf{Z}_s(\mathbf{Z}_s'\boldsymbol{\Sigma}_{ss}^{-1}\mathbf{Z}_s)^{-1}\mathbf{Z}_s'\boldsymbol{\Sigma}_{ss}^{-1}$. Hence, when $\mathbf{A}_{ss} \propto \mathbf{W}_{ss}$, we have $\hat{\boldsymbol{\beta}}_x = \hat{\mathbf{B}}$ and the result follows from $\hat{\hat{Y}}_{EG} - \hat{\hat{Y}}_O = (\hat{\boldsymbol{\beta}}_x - \hat{\mathbf{B}})(\bar{\mathbf{x}} - \hat{\hat{\mathbf{x}}})$.

7. Examples of Equivalencies Between EGRE's and OPE's

The starting point for deriving working models under which the corresponding EGRE is asymptotically equivalent to the OPE that uses the same auxiliary variable $\boldsymbol{x}$ is the structure of the matrix $\mathbf{W}$. When $r(\mathbf{W}) = N$ there is no DBV. So, setting $\boldsymbol{\Sigma} \propto \mathbf{W}^{-1}$, the corresponding EGRE is asymptotically equal to the OPE. Furthermore, because of Theorem 3, the EGRE is the OPE as well. As an example, consider Poisson sampling with size measure a_i, $i = 1, \ldots, N$. Let $\pi_i = na_i / A$ be the inclusion probabilities of population units, where n is the expected sample size and A is the total of a_i. In this case $\mathbf{W} = diag\{\pi_i^{-1} - 1\}$. Setting $\boldsymbol{\Sigma} = diag\{a_i/(A - na_i)\}$, the GRE is the OPE as well. This form of the variance matrix was also achieved by Särndal (1996) by minimising the asymptotic variance of a GRE for Poisson sampling.

Next we give examples where DBV's exist. As a rule of thumb, potential DBV's are variables proportional to the first order inclusion probabilities within subpopulations from which fixed size samples are selected. This is documented by the following examples.

Example 1. Consider simple random sampling of n units. In this case $r(\mathbf{W}) = N - 1$ and any vector proportional to the unit vector $\mathbf{1}$ is a DBV. Setting $\mathbf{Z} = \mathbf{1}$ and $\boldsymbol{\Sigma} = \mathbf{I}$, where $\mathbf{I}$ is the identity matrix, then $\mathbf{A} \propto \mathbf{W}$. Furthermore, because of Theorem 3, a GRE based on a homoscedastic linear regression model with an intercept term is an OPE as well.

Example 2. Consider stratified simple random sampling and denote by hi the i-th unit within the h-th stratum where hi ranges over the pairs $11, 12, \ldots, 1N_1$, $21, 22, \ldots, 2N_2, \ldots, H1, H2, \ldots, HN_H$. Let n_h be the sample size within stratum h. In this case $r(\mathbf{W}) = N - H$ and the indicator variables of stratum membership of population units are a set of linearly independent DBV's. Further, let $\mathbf{z}_l$ be the vector of the values of the l-th stratum membership indicator variable, whose entries are $Z_{lhi} = 1$, when $h = l$, and $Z_{lhi} = 0$, otherwise. Define $\mathbf{Z} = [\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_H]$. Then, setting $\boldsymbol{\Sigma} = diag\{\eta_{hi}\}$, where $\eta_{hi} = [n_h(N_h - 1)]/[N_h(N_h - n_h)]$, it follows that $\mathbf{A} \propto \mathbf{W}$. So, when the working model includes the stratum membership indicator variables of population units and the variance matrix is specified as above, the GRE is asymptotically equal to the OPE. Note that this form of the variance matrix was also achieved by Särndal (1996) by minimising the asymptotic variance of a GRE based on a working model with an intercept term for each stratum. When n_h is constant across strata, because of Theorem 3, the GRE is identically equal to the OPE.

Example 3. Consider stratified two-stage random sampling and let us denote by hij the j-th elementary unit within the i-th Primary Sampling Unit (PSU) of the h-th stratum with $h = 1, \ldots, H$; $i = 1, \ldots, N_h$; $j = 1, \ldots, M_{hi}$. Using simple random sampling without replacement in both stages, n_h PSU's are selected from each stratum and m_{hi} elementary units are drawn from each selected PSU. In this case $r(\mathbf{W}) = N - H$ and for each value of $l = 1, \ldots, H$, the vector $\mathbf{z}_l$, whose entries are $Z_{lhij} = M_h / N_h M_{hi}$ when $h = l$ and $Z_{lhij} = 0$ otherwise, where $M_h = \sum_{i=1}^{N_h} M_{hi}$, is a DBV. Thus, the matrix $\mathbf{Z} = [\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_H]$ contains H linearly independent DBV's. Inserting $\mathbf{Z}$ into the working model and setting $\boldsymbol{\Sigma} = diag\{\boldsymbol{\Sigma}_{hi}\}$, where $\boldsymbol{\Sigma}_{hi} = [a_{hi}\mathbf{I} + b_{hi}\mathbf{1}\mathbf{1}']^{-1}$ is the $M_{hi} \times M_{hi}$ matrix with

$$a_{hi} = \frac{N_h}{n_h} \frac{M_{hi}}{m_{hi}} \frac{M_{hi} - m_{hi}}{M_{hi} - 1} \quad \text{and} \quad b_{hi} = \frac{N_h}{n_h} \frac{M_{hi}}{m_{hi}} \frac{m_{hi} - 1}{M_{hi} - 1} - \frac{N_h}{n_h} \frac{n_h - 1}{N_h - 1}$$

it follows from Theorem 2 that $\mathbf{A} \propto \mathbf{W}$ and the EGRE is asymptotically equal to the OPE. Furthermore, when n_h is constant across strata, because of Theorem 3, the EGRE is equal to the OPE as well. Note that the structure of $\boldsymbol{\Sigma}$ is that of an equal correlation model within PSU's.

Now, two examples involving unequal probability sampling are presented. However, for simplicity, we assume sampling with replacement. In such a case, first and second order inclusion probabilities must be replaced by $\varphi_i = \mathrm{E}(\delta_i)$ and $\varphi_{ij} = \mathrm{E}(\delta_i \delta_j)$, where δ_i is the random variable defined to be the number of times the i-th unit has been selected in the sample s. As a consequence, Horvitz-Thompson estimators have to be replaced by Hansen-Hurvitz analogues and in the matrix $\mathbf{W}$ the inclusion probabilities π_i and π_{ij} are replaced by φ_i and φ_{ij}.

Example 4. Consider a with replacement unequal probability sampling design of fixed size n and with selection probabilities P_i , $i = 1, 2, \ldots, N$. The sampling variance of the unbiased Hansen-Hurvitz estimator is given by $\mathbf{Y}'\mathbf{W}\mathbf{Y}$, where $\mathbf{W} = n^{-1}N^{-2}(\mathbf{P}^{-1} - \mathbf{1}\mathbf{1}')$, and $\mathbf{P} = diag\{P_i\}$. In this case, $r(\mathbf{W}) = N - 1$ and the variable $Z_i = NP_i$ is a DBV. Inserting the latter into the working model and setting $\boldsymbol{\Sigma} = diag\{nP_i\}$, the GRE is asymptotically equal to the OPE. Furthermore, because of Theorem 3, the GRE is the OPE as well.

Example 5. Consider a stratified two-stage sampling as in Example 3, but now at the first stage n_h PSU's are selected from each stratum h using a with replacement unequal probability scheme with selection probabilities P_{hi}. In this case, $r(\mathbf{W}) = N - H$. The matrix $\mathbf{Z} = [\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_H]$, where $\mathbf{z}_l$ is the vector whose entries are $Z_{lhij} = M_h P_{hi}/M_{hi}$ when $h = l$ and $Z_{lhij} = 0$ otherwise, contains H linearly independent DBV's. Inserting $\mathbf{Z}$ into the working model and setting $\boldsymbol{\Sigma} = diag\{\boldsymbol{\Sigma}_{hi}\}$, where $\boldsymbol{\Sigma}_{hi} = [a_{hi}\mathbf{I} + b_{hi}\mathbf{1}\mathbf{1}']^{-1}$ is the $M_{hi} \times M_{hi}$ matrix with

$$a_{hi} = \frac{1}{n_h P_{hi}} \frac{M_{hi}}{m_{hi}} \frac{M_{hi} - m_{hi}}{M_{hi} - 1} \quad \text{and} \quad b_{hi} = \frac{1}{n_h P_{hi}} \frac{M_{hi}}{m_{hi}} \frac{m_{hi} - 1}{M_{hi} - 1}$$

the EGRE is asymptotically equal to the OPE. As before, when n_h is constant across strata, the EGRE is identically equal to the OPE.

Note that with replacement results are often used for approximating without replacement results.

8. Estimation Strategies

The examples examined in the previous section illustrate that, given an amount of auxiliary information $\bar{\mathbf{x}}$, generally speaking, the OPE is equivalent to the EGRE based on a working model which includes the maximum number of linearly independent DBV's and assumes a variance matrix that reflects the structure of the first and second order inclusion probabilities. Thus, since the OPE allows a better fit of the data, this explains its asymptotic superiority. However, in finite size samples the OPE estimator is exposed to instabilities due to a possible inadequate number of residual degrees of freedom available for estimating all the parameters. In particular, this concern may be relevant in stratified designs with a few observations per stratum, or multistage sampling designs with a few PSU's per stratum.

The analysis and the examples presented above suggest the following options for estimating a population mean taking into account a given amount of auxiliary information $\bar{\mathbf{x}}$. The first option consists in simply using the OPE. This choice assures the maximum asymptotic efficiency. However, if the stability of that estimator is of concern, as when the total sample size is not large enough or the number of strata is high compared to the number of observations, and we are confident enough about a more parsimonious working model, a second option is to use the EGRE based on it.

A further intermediate option, that we recommend when a reliable model is unavailable, consists in specifying a working model with a variance matrix set according to Theorem 2 and a suitably reduced number of DBV's. In particular, in the case of stratified samples, this may be accomplished by introducing into the model DBV's corresponding to superstrata obtained by collapsing original strata. Collapsing should be performed so that within each superstratum, strata effects can be considered negligible. For each superstratum, a DBV is obtained by adding the DBV's of the collapsed strata. By reducing the number of DBV's inserted into the model, we accept a lower level of asymptotic efficiency to better control the finite size sample variance of the estimator. The latter option has been implemented in the following simulation study.

9. An Empirical Study

The theory developed in the previous sections is asymptotic, being based on first order approximations. In this section we report results from an empirical study carried out to test the theory in the presence of finite size samples. In particular we refer to Example 2 above.

A finite population of 1,200 units partitioned into 20 strata of equal size was considered. The values of an auxiliary variable x were generated through a Chi-squared random variable with 8 degrees of freedom. They were assigned to the strata in two ways. In the first case they were randomly assigned to the strata, to simulate a stratification based on other characters independent of x (Stratification 1). In the second case the values of x were first ordered. Then the first 60 smallest values were assigned to the first stratum, the subsequent 60 smaller values were assigned to the second stratum and so forth, to simulate a stratification based on classes of x values (Stratification 2).

Given the values of x, six populations of y-values were generated, according to the following models:

P_1: $Y_{hi} = 100(1 + 0.15\varepsilon_{hi})$ (total independence between y, x and h);
P_2: $Y_{hi} = (10 + 3h)(1 + 0.15\varepsilon_{hi})$ (dependence between y and h);
P_3: $Y_{hi} = 3X_{hi}\ (1 + 0.15\varepsilon_{hi})$ (linear dependence between y and x);
P_4: $Y_{hi} = (12 + 4h + 5X_{hi})(1 + 0.15\varepsilon_{hi})$ (linear dependence between y, x and h);
P_5: $Y_{hi} = X_{hi}^{\ 2}\ (1 + 0.15\varepsilon_{hi})$ (quadratic dependence between y and x);
P_6: $Y_{hi} = 20h\sqrt{X_{hi}}(1 + 0.15\varepsilon_{hi})$ (non-linear dependence between y, x and h);

where Y_{hi} and X_{hi} are the values of y and x in the i-th unit ($i = 1, \ldots, 60$) of the h-th stratum ($h = 1, \ldots, 20$) and the ε_{hi}'s are independent observations from a standard normal distribution. All models are heteroscedastic with variances proportional to the squared expected values.

For each stratification and population, 10,000 proportional stratified random samples of size 40 (two units per stratum), 80 (four units per stratum) and 240 (twelve units per stratum) were selected. For this sampling scheme, the DBV's are the indicator variables of stratum membership of population units. The variance matrix that corresponds to the optimal estimator is the identity matrix, being n_h and N_h constant across strata (see Example 2).

For each sample, assuming $\bar{X}$ is known, the following estimators were computed:

$\bar{y}$ and $\bar{x}$, i.e., the sample means of y and x (Horvitz-Thompson estimators);
$\hat{\bar{Y}}_{G1}$, i.e., the GRE based on the model $\mathrm{E}_m(Y_{hi}) = \beta X_{hi}$, $\mathrm{V}_m(Y_{hi}) = X_{hi}\sigma^2$ (combined ratio estimator);
$\hat{\bar{Y}}_{G2}$, i.e., the GRE based on the model $\mathrm{E}_m(Y_{hi}) = \beta_1 + \beta_2 X_{hi}$, $\mathrm{V}_m(Y_{hi}) = X_{hi}^2\sigma^2$;
$\hat{\bar{Y}}_{G3}$, i.e., the GRE based on the model $\mathrm{E}_m(Y_{hi}) = \beta_1 + \beta_2 X_{hi}$, $\mathrm{V}_m(Y_{hi}) = \sigma^2$ (adopting the variance matrix that corresponds to the OPE);
$\hat{\bar{Y}}_{G4}$, i.e., the GRE based on the model $\mathrm{E}_m(Y_{hi}) = \sum_{l=1}^{5} \beta_l D_{4lhi} + \beta_6 X_{hi}$, $\mathrm{V}_m(Y_{hi}) = \sigma^2$, where D_{4l} is the variable obtained by adding the DBV's of the four strata $4(l-1)+1$, $4(l-1)+2$, $4(l-1)+3$, $4(l-1)+4$, for $l = 1, 2, 3, 4, 5$;
$\hat{\bar{Y}}_{G5}$, i.e., the GRE based on the model $\mathrm{E}_m(Y_{hi}) = \sum_{l=1}^{10} \beta_l D_{2lhi} + \beta_{11} X_{hi}$, $\mathrm{V}_m(Y_{hi}) = \sigma^2$, where D_{2l} is the variable obtained by adding the DBV's of the two original strata $2(l-1)+1$ and $2(l-1)+2$, for $l = 1, \ldots, 10$;
$\hat{\bar{Y}}_O$, i.e., the OPE that uses $\bar{X}$;
$\tilde{\bar{Y}}_O$, i.e., the first order linear approximation of the OPE.

Estimators $\hat{\bar{Y}}_{G3}$, $\hat{\bar{Y}}_{G4}$ and $\hat{\bar{Y}}_{G5}$ are based on working models that assume the variance matrix corresponding to the OPE but include a reduced number of DBV's, according to the collapsing strata technique (one DBV for $\hat{\bar{Y}}_{G3}$, five for $\hat{\bar{Y}}_{G4}$ and ten for $\hat{\bar{Y}}_{G5}$). Note that $\hat{\bar{Y}}_O$ corresponds to a working model with 21 parameters (twenty DBV's and one auxiliary variable) and instability can be expected, in particular for sample size $n = 40$.

The mean and the mean squared error across the 10,000 selected samples were computed for each estimator. Tables 1 and 2 report the scaled mean squared errors, having set that of $\tilde{\bar{Y}}_O$ equal to 100. Biases are not reported, since they were negligible in all cases.

First, note that $\tilde{\bar{Y}}_O$ is a gauge of the best we can expect from the OPE, as it shows its asymptotic behaviour; it is the most efficient among the computed estimators. On the other hand, the OPE is vulnerable to sampling fluctuations, in particular with Stratification 2 and $n = 40$. In fact, in the worst case the variance of $\hat{\bar{Y}}_O$ is 47.9% higher than that of $\tilde{\bar{Y}}_O$ (Table 2, Population P_5). With this stratification, the OPE performance is generally worse than that under Stratification 1. Most likely, the reason for this is the great diversity of stratum variances of x, because of its asymmetric distribution, coupled with the equal allocation of the sample: it results in a much more erratic estimated regression coefficient of $(\bar{X} - \bar{x})$ when the sample size is the smallest ($n = 40$) and there are two units per stratum.

Estimators $\bar{y}$, $\hat{\bar{Y}}_{G1}$, $\hat{\bar{Y}}_{G2}$ and $\hat{\bar{Y}}_{G3}$ are almost as efficient as $\tilde{\bar{Y}}_O$ when the model upon which they are based holds true, but they are quite inefficient in the presence of model failures, in particular with Stratification 1. For instance, when the variable y does not depend on x and h (Population P_1) or it depends only on h (Population P_2), the sample mean does well; but when y depends on x and the latter is not used for stratification, the sample mean suffers from not using any auxiliary information as the regression estimators do.

Table 1. Scaled mean squared error of estimators for Stratification 1 [$MSE(\hat{\tilde{Y}}_O) = 100$]

Population	n	$\bar{y}$	$\hat{\hat{Y}}_{G1}$	$\hat{\hat{Y}}_{G2}$	$\hat{\hat{Y}}_{G3}$	$\hat{\hat{Y}}_{G4}$	$\hat{\hat{Y}}_{G5}$	$\hat{\hat{Y}}_O$
P_1	40	100.0	1,304.6	119.4	103.0	104.0	104.5	107.9
$Y_{hi} = 100(1 + 0.15\varepsilon_{hi})$	80	100.0	1,177.2	109.1	101.3	101.4	101.6	101.9
	240	100.1	1,217.1	104.6	100.1	100.2	100.2	100.3
P_2	40	100.2	1,234.0	421.4	236.7	106.6	106.6	109.8
$Y_{hi} = (10 + 3h)(1 + 0.15\varepsilon_{hi})$	80	100.2	1,128.8	364.2	215.7	102.9	102.6	102.9
	240	100.2	1,103.8	333.8	210.9	101.1	100.6	100.6
P_3	40	982.6	100.0	101.3	103.4	105.3	105.8	108.8
$Y_{hi} = 3X_{hi}\ (1 + 0.15\varepsilon_{hi})$	80	924.2	100.2	100.7	102.0	102.6	102.9	103.3
	240	924.8	101.0	101.4	100.6	100.8	100.8	101.0
P_4	40	445.0	279.5	157.7	121.9	105.0	105.7	110.0
$Y_{hi} = (12 + 4h + 5X_{hi}) \times (1 + 0.15\varepsilon_{hi})$	80	422.4	271.0	148.9	119.0	102.2	102.2	102.7
	240	405.1	275.3	150.0	119.8	101.0	100.8	100.9
P_5	40	834.5	277.5	208.8	105.7	106.2	106.9	109.4
$Y_{hi} = X_{hi}^2\ (1 + 0.15\varepsilon_{hi})$	80	776.1	266.0	208.8	104.1	104.2	104.5	104.9
	240	772.2	256.5	205.2	99.6	99.9	100.0	100.1
P_6	40	293.6	229.6	236.1	205.8	110.9	110.5	113.2
$Y_{hi} = 20h\sqrt{X_{hi}}(1 + 0.15\varepsilon_{hi})$	80	267.8	234.7	226.3	200.0	104.3	104.0	104.2
	240	265.5	236.0	220.9	195.1	101.5	100.7	100.8
	40	**459.3**	**570.9**	**207.5**	**146.1**	**106.3**	**106.7**	**109.9**
Average	80	**431.8**	**529.7**	**193.0**	**140.4**	**102.9**	**103.0**	**103.3**
	240	**428.0**	**531.6**	**186.0**	**137.7**	**100.8**	**100.5**	**100.6**

Models upon which the GREG's are based

$\hat{\hat{Y}}_{G1}$: $E_m(Y_{hi}) = \beta X_{hi}$, $V_m(Y_{hi}) = X_{hi}\sigma^2$
$\hat{\hat{Y}}_{G2}$: $E_m(Y_{hi}) = \beta_1 + \beta_2 X_{hi}$, $V_m(Y_{hi}) = X_{hi}^2\sigma^2$
$\hat{\hat{Y}}_{G3}$: $E_m(Y_{hi}) = \beta_1 + \beta_2 X_{hi}$, $V_m(Y_{hi}) = \sigma^2$
$\hat{\hat{Y}}_{G4}$: $E_m(Y_{hi}) = \sum_{l=1}^{5} \beta_l D_{4lhi} + \beta_6 X_{hi}$, $V_m(Y_{hi}) = \sigma^2$
$\hat{\hat{Y}}_{G5}$: $E_m(Y_{hi}) = \sum_{l=1}^{10} \beta_l D_{2lhi} + \beta_{11} X_{hi}$, $V_m(Y_{hi}) = \sigma^2$

The combined ratio estimator $\hat{\hat{Y}}_{G1}$ works well only when the values of y are on average proportional to the values of x (Population P_3), whereas estimators $\hat{\hat{Y}}_{G2}$ and $\hat{\hat{Y}}_{G3}$ are inefficient when the relation between y and x is not linear. Observe that $\hat{\hat{Y}}_{G3}$ is almost always more efficient then $\hat{\hat{Y}}_{G2}$, even when the populations are heteroscedastic. It is also worth noting that Stratification 2 is based on x, and most of the information on y provided by x is captured by the stratification. Thus, the use of x at the estimation stage may be redundant and this explains why for this stratification the sample mean is a fairly efficient estimator apart from Population P_5. In fact, $\bar{y}$ is equivalent to a GRE based on a working model that uses the DBV's, i.e., the stratum membership indicators; however, this estimator does not take the known value of $\bar{X}$ when applied to the auxiliary variable as all other estimators do.

Finally, estimators $\hat{\hat{Y}}_{G4}$ and $\hat{\hat{Y}}_{G5}$, based on the collapsed stratum technique, are almost always more efficient than the OPE with sample sizes 40 and 80, because of the reduced number of DBV's inserted into the working model. Furthermore, when they are not the most efficient among all estimators, their scaled mean squared errors are not substantially larger than that of the best estimator for each population. Thus, the collapsed stratum technique seems a useful device to identify a stable approximated OPE which will be fairly efficiency robust with respect to model failures. In fact, $\hat{\hat{Y}}_{G4}$ and $\hat{\hat{Y}}_{G5}$ have the lowest

Table 2. Scaled mean squared error of estimators for Stratification 2 [$MSE(\tilde{\bar{Y}}_O) = 100$]

Population	n	$\bar{y}$	$\hat{\bar{Y}}_{G1}$	$\hat{\bar{Y}}_{G2}$	$\hat{\bar{Y}}_{G3}$	$\hat{\bar{Y}}_{G4}$	$\hat{\bar{Y}}_{G5}$	$\hat{\bar{Y}}_O$
P_1	*40*	100.0	118.4	100.4	99.9	99.9	100.3	109.9
$Y_{hi} = 100(1 + 0.15\varepsilon_{hi})$	*80*	100.0	118.7	100.1	100.0	99.9	100.1	103.0
	240	99.9	118.7	100.0	99.9	99.9	100.0	100.4
P_2	*40*	100.2	121.3	115.6	106.9	102.6	102.5	121.6
$Y_{hi} = (10 + 3h)(1 + 0.15\varepsilon_{hi})$	*80*	100.3	121.6	115.8	107.5	103.0	102.4	106.0
	240	99.9	119.7	113.8	105.9	101.6	100.6	101.1
P_3	*40*	108.8	100.4	100.4	100.1	100.4	101.7	131.0
$Y_{hi} = 3X_{hi}\,(1 + 0.15\varepsilon_{hi})$	*80*	106.9	101.0	101.1	100.8	101.2	101.8	108.6
	240	109.9	100.4	100.4	100.3	100.4	100.3	103.2
P_4	*40*	103.9	103.3	102.6	101.0	100.7	101.8	127.1
$Y_{hi} = (12 + 4h + 5X_{hi})(1 + 0.15\varepsilon_{hi})$	*80*	102.6	104.8	103.9	102.1	101.3	101.5	107.6
	240	105.3	102.7	102.0	100.7	100.3	100.5	102.6
P_5	*40*	164.9	128.5	129.5	110.4	101.2	102.6	147.9
$Y_{hi} = X_{hi}^2\,(1 + 0.15\varepsilon_{hi})$	*80*	157.4	124.3	125.5	108.3	100.5	101.7	114.5
	240	176.5	134.1	136.0	113.1	101.4	101.5	106.1
P_6	*40*	101.2	105.7	107.5	107.1	104.0	104.1	133.3
$Y_{hi} = 20h\sqrt{X_{hi}}(1 + 0.15\varepsilon_{hi})$	*80*	101.1	105.7	107.3	107.4	104.0	103.0	109.8
	240	102.0	104.7	106.1	106.5	102.9	101.2	102.1
	40	**113.2**	**112.9**	**109.3**	**104.2**	**101.5**	**102.2**	**128.5**
Average	*80*	**111.4**	**112.7**	**109.0**	**104.4**	**101.7**	**101.8**	**108.3**
	240	**115.6**	**113.4**	**109.7**	**104.4**	**101.1**	**100.7**	**102.6**

Models upon which the GREG's are based

$\hat{\bar{Y}}_{G1}$: $\mathrm{E}_m(Y_{hi}) = \beta X_{hi}$, $\mathrm{V}_m(Y_{hi}) = X_{hi}\sigma^2$
$\hat{\bar{Y}}_{G2}$: $\mathrm{E}_m(Y_{hi}) = \beta_1 + \beta_2 X_{hi}$, $\mathrm{V}_m(Y_{hi}) = X_{hi}^2\sigma^2$
$\hat{\bar{Y}}_{G3}$: $\mathrm{E}_m(Y_{hi}) = \beta_1 + \beta_2 X_{hi}$, $\mathrm{V}_m(Y_{hi}) = \sigma^2$
$\hat{\bar{Y}}_{G4}$: $\mathrm{E}_m(Y_{hi}) = \sum_{l=1}^{5} \beta_l D_{4lhi} + \beta_6 X_{hi}$, $\mathrm{V}_m(Y_{hi}) = \sigma^2$
$\hat{\bar{Y}}_{G5}$: $\mathrm{E}_m(Y_{hi}) = \sum_{l=1}^{10} \beta_l D_{2lhi} + \beta_{11} X_{hi}$, $\mathrm{V}_m(Y_{hi}) = \sigma^2$

averaged mean squared error across populations and stratifications (see the three rows at the bottom of the tables). This is particularly valuable in cases where there is uncertainty about a proper working model.

10. Final Remarks

Generally speaking, the OPE is approximately equal to an EGRE based on a working model that includes the effect of any existing DBV's. Unfortunately, Theorem 2 does not provide guidelines for determining the structure of the matrix $\mathbf{\Sigma}$ and the DBV's that correspond to the sampling design in use. However, for common designs, easy solutions can be found. Thus, given an amount of auxiliary information $\bar{\mathbf{x}}$, the OPE is approximately or exactly equal to an EGRE based on a working model which includes the maximum number of linearly independent DBV's and assumes a variance matrix that reflects the structure of the first and second order inclusion probabilities. Hence the OPE allows a better fit of the data, and this explains its asymptotic superiority. However, in finite size samples, the OPE is exposed to instabilities due to a likely inadequate number of residual degrees of freedom available for fitting the model. In particular, this concern may be relevant in stratified designs with a few observations per stratum, or multistage sampling designs with a few PSU's per stratum.

The above analysis suggests the following quasi-optimal estimation strategy. When a reliable model is lacking, use an EGRE based on a working model with a variance matrix set according to Theorem 2 and with a suitably reduced number of DBV's to achieve a sufficient number of residual degrees of freedom to fit the model. In particular, in the case of stratified samples, this may be accomplished by introducing into the model DBV's corresponding to superstrata obtained by collapsing original strata. Collapsing should be performed so that within each superstratum, strata effects can be considered negligible. For each superstratum, a DBV is obtained by adding the DBV's of the collapsed strata. By reducing the number of DBV's inserted into the model, we accept a lower level of asymptotic efficiency to better control the finite size sampling variance of the estimator.

A related topic of interest is that of variance estimation. The theory presented in the previous sections was developed using the Horvitz-Thompson estimator of the variances and covariances required to estimate $\mathbf{B}$ for the optimal estimator. But alternative ways for variance estimation are available, such as the Yates-Grundy formula or the use of resampling methods. Thus, it is uncertain which is the best variance estimation procedure to be used to better estimate $\mathbf{B}$. Variance estimators are also required to estimate the standard errors of regression estimators. In this respect, the usual way is to estimate the variance of the first order linear approximations of regression estimators, replacing the unknown regression coefficients with their sample estimates. The resulting estimator is usually somewhat negatively biased, especially with a larger number of auxiliary variables in the working model. That is also true for the OPE that uses the maximum number of auxiliary variables. Thus, the properties of variance estimators affect the coverage of confidence intervals, and further research is needed to explore this issue in order to single out a better estimation strategy for interval estimation.

So far we have restricted the analysis to the use of DBV's, since the vector $\bar{\mathbf{x}}$ was taken to be fixed. But clearly this point is part of the wider problem of selecting a subset of the available auxiliary variables for estimation purposes. For example, if the true model is linear, a quadratic term inserted into the working model, although asymptotically it would result in a more efficient estimator of the population mean, in finite size samples may give a less efficient estimator. How to select the best subset of available auxiliary variables for estimating a population mean is still an unsolved research problem, but the topic is beyond the scope of this article.

11. References

Casady, R.J. and Valliant, R. (1993). Conditional Properties of Post-stratified Estimators Under Normal Theory. Survey Methodology, 19, 183–192.

Eltinge, J.L. and Jang, D.S. (1996). Stability Measures for Variance Component Estimators Under a Stratified Multistage Design. Survey Methodology, 22, 157–165.

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

Lehtonen, R. and Pahkinen, E.J. (1995). Practical Methods for Survey Design. New York: J. Wiley.

Montanari, G.E. (1987). Post-sampling Efficient QR-prediction in Large-scale Surveys. International Statistical Review, 55, 191–202.

Montanari, G.E. (1998). On Regression Estimation of Finite Population Means. Survey Methodology, 24, 69–77.

Rao, J.N.K. (1994). Estimating Totals and Distribution Functions Using Auxiliary Information at the Estimation Stage. Journal of Official Statistics, 10, 153–165.

Rao, J.N.K. (1997). Developments in Sample Survey Theory: An Appraisal. Canadian Journal of Statistics, 25, 1–21.

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

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

Wright, R.L. (1983). Finite Population Sampling with Multivariate Auxiliary Information. Journal of the American Statistical Association, 78, 879–884.

Received ?????
Revised ?????