The Weighted Residual Technique for Estimating the Variance of the General Regression Estimator

Carl-Erik Särndal, Bengt Swensson and
Jan H. Wretman

SCB
R & D Report
Statistics Sweden
Research - Methods - Development
1988:10

INLEDNING

TILL

R & D report : research, methods, development / Statistics Sweden. – Stockholm : Statistiska centralbyrån, 1988-2004. – Nr. 1988:1-2004:2.
Häri ingår Abstracts : sammanfattningar av metodrapporter från SCB med egen numrering.

Föregångare:

Metodinformation : preliminär rapport från Statistiska centralbyrån. – Stockholm : Statistiska centralbyrån. – 1984-1986. – Nr 1984:1-1986:8.

U/ADB / Statistics Sweden. – Stockholm : Statistiska centralbyrån, 1986-1987. – Nr E24-E26

R & D report : research, methods, development, U/STM / Statistics Sweden. – Stockholm : Statistiska centralbyrån, 1987. – Nr 29-41.

Efterföljare:

Research and development : methodology reports from Statistics Sweden. – Stockholm : Statistiska centralbyrån. – 2006-. – Nr 2006:1-.

R & D Report 1988:10. The weighted residual technique for estimating the variance of the general regression estimator / Carl-Erik Särndal m. fl.
Digitaliserad av Statistiska centralbyrån (SCB) 2016.

urn:nbn:se:scb-1988-X101OP8810

The Weighted Residual Technique for Estimating the Variance of the General Regression Estimator

Carl-Erik Särndal, Bengt Swensson and
Jan H. Wretman

SCB
R & D Report
Statistics Sweden
Research - Methods - Development
1988:10

Från trycket	Augusti 1988
Producent	Statistiska centralbyrån, Utvecklingsavdelningen
Ansvarig utgivare	Staffan Wahlström
Förfrågningar	Jan Wretman, tel. 08-7834961

ISSN 0283-8680
Printed in Sweden
Garnisonstryckeriet, Stockholm 1988

June, 1988

The Weighted Residual Technique for Estimating the Variance of the General Regression Estimator

by

Carl-Erik Särndal	Bengt Swensson	Jan H. Wretman
Statistics Sweden	University of Örebro	Statistics Sweden
Stockholm, Sweden	Örebro, Sweden	Stockholm, Sweden

Summary. Despite the advancement of computer-intensive methods for variance estimation for complex survey situations, closed form variance estimators will always have strong appeal. In this paper we construct a simple yet general method for estimating the variance of the general regression estimator. The method calls for weighting of the regression residuals when the variance estimator is calculated by the well-known Horwitz-Thompson formula. The weights are obtained in a simple way from the general regression estimator formula. A strong point in favour of the proposed technique is that it can be defended from design-based criteria as well as from model-based criteria, and that it works well for certain kinds of conditional inference. Illustrative examples of the technique are given, including consequences for the important practice of poststratification.

1. Introduction.

There is a considerable literature on variance estimation for survey estimates; a thorough account is given in the recent book by Wolter (1985). Some recent methods, such as the boot-strap, are computer intensive. Even if re-sampling is not involved, the calculation of a variance estimator is often heavy. Simple, closed form variance estimators, such as the one

examined in this paper, will always have strong appeal.

In survey sampling, the recent emphasis on linear statistical modeling has stimulated progress in the model-based approach to inference in surveys, as well as in the design-based (randomization theory) approach. Holt and Smith (1979), Royall and Cumberland (1978, 1981a, 1981b) and other advocates of the model-based point of view have made important remarks on variance estimation, some of which put the finger on weaknesses in the traditional techniques of variance estimation. Even for widely used estimators, such as the classical ratio estimator and the classical simple regression estimator, the issue of variance estimation has not been finally resolved. Studies by Wu (1982), Wu and Deng (1983), Deng and Wu (1987) show that it is hard to single out a "best" variance estimator; bestness depends on the performance criterion in use.

This paper examines a general technique for estimating the variance of the general regression estimator, which is used to estimate a finite population total or mean in the presence of auxiliary information, univariate or multivariate. The spirit of the paper is to point to agreement between design-based and model-based approaches. In particular, the work was motivated by the question: how does one construct a variance estimator for the general regression estimator that combines simplicity with generality and that has favourable properties under the sampling design as well as under an assumed regression model? In this paper we develop and analyze a variance estimation technique that meets these objectives. A different technique with roughly the same objectives was proposed by Kott (1987); see Section 5.

Let $U = \{1, ..., k, ..., N\}$ be a finite population, and let y_k be the value of the study variable y for the k:th population unit. We seek to estimate the population total $t = y_1 + ... + y_N$. For any given set A of population units ($A \subseteq U$), we shall write $\Sigma_A\, y_k$ for $\Sigma_{k \in A}\, y_k$, for example, $t = \Sigma_U\, y_k$.

The following example illustrates that uncritical use of the traditional approach to variance estimation may be controversial, but that the problem may be resolved by the approach of this paper:

Example 1.1: A simple random sample of size n is drawn from U, and then poststratified. The usual estimator of the population total t is

$$\hat{t} = \Sigma_{h=1}^{H}\, N_h\ \bar{y}_{sh}\ , \qquad (1.1)$$

where $\bar{y}_{sh} = \Sigma_{sh}\, y_k/n_h$ is the mean of the in the sample s_h composed of those n_h units that happen to fall in the h:th group (poststratum), N_h is the known size of the population group U_h, and $N = N_1 + ... + N_H$; $n = n_1 + ... + n_H$. Standard sampling texts (for example, Cochran (1977), p. 135) give the (unconditional) variance of (1.1), to first order approximation, as

$$V_u(\hat{t}) \doteq N^2\, \{(1 - f)/n\}\, \Sigma_{h=1}^{H}\, W_h\, S_{Uh}^2\ , \qquad (1.2)$$

where $f = n/N$, $W_h = N_h/N$, $S_{Uh}^2 = \Sigma_{Uh}\, (y_k - \bar{y}_{Uh})/(N_h - 1)$, with $\bar{y}_{Uh} = \Sigma_{Uh}\, y_k/N_h$. The standard technique for building a variance estimator is to use the variance as a starting point. Thus, replacing S_{Uh}^2 in (1.2) by its sample analogue $S_{sh}^2 = \Sigma_{sh}\, (y_k - \bar{y}_{sh})^2/(n_h - 1)$, we get

$$\hat{V}_u(\hat{t}) = N^2 \{(1 - f)/n\} \Sigma_{h=1}^{H} W_h S_{Sh}^2, \qquad (1.3)$$

whose average value (over repeated simple random samples with $n_h \geq 2$) equals (1.2), so the estimator is "correct" in design-based thinking, and many sampling statisticians will be satisfied with (1.3). However, others would argue, as do Holt and Smith(1979), that inference should be made conditionally on the realized sample configuration $\mathbf{n} = (n_1, \ldots, n_h, \ldots, n_H)'$. If such an outlook is adopted, one reason for dissatisfaction with (1.3) is that the contribution to $\hat{V}_u(\hat{t})$ from group h equals S_{Sh}^2 times a weight that is insensitive to n_h, namely, $N^2 W_h(1 - f)/n$. The statistician favouring the conditional outlook would argue: Suppose the h:th group happened to produce unusually few observations, so that the realized n_h falls considerably short of its expectation, $E(n_h) = n W_h$. The natural inclination then is to want a greater than average contribution to the variance estimator from the underrepresented poststratum. Now, (1.3) fails to achieve this, whereas an estimator such as

$$\hat{V}(\hat{t}) = (1 - f) \Sigma_{h=1}^{H} N_h^2 S_{Sh}^2/n_h \qquad (1.4)$$

meets the objective. In (1.4), which is the variance estimator obtained by the technique in this paper (see Example 4.1 below), the poststratum sample variance S_{Sh}^2 carries the weight $(1 - f) N_h^2/n_h$, whose tendency to drop with an increase in n_h makes good sense from the conditional point of view. Furthermore, in using (1.4) the advocate of design-based inference gives up none of <u>his</u> principles, since (1.4) is as correct as (1.3) in the sense that both have an average, over repeated simple random

samples, that is (at least approximately) equal to (1.2). We advocate (1.4) over (1.3) and make the point that randomization theory is flexible enough to accomodate the conditional point of view. □

2. The general regression estimator.

A probability sample, s, is drawn from U with a sampling design having the inclusion probabilities $\pi_k = \Pr(k \in s)$ and $\pi_{k\ell} = \Pr(k \text{ and } \ell \in s)$. Note that $\pi_{kk} = \pi_k$ for all k. The size of s, n_s, is permitted to be random. We examine the design-based statistical properties of estimators of t. That is, design expectation, design bias and design variance become important quantities to consider. The design must often be chosen more for reasons of practical and administrative necessity than to yield the highest possible precision of the resulting estimates. Unequal inclusion probability designs are extremely common. In design-based analysis, the sampling weights $1/\pi_k$ are required, as in the approximately design unbiased general regression estimator of t,

$$\hat{t} = \Sigma_s\, y_k/\pi_k + \{\Sigma_U\, \mathbf{x}_k - \Sigma_s\, \mathbf{x}_k/\pi_k\}'\, \hat{\mathbf{B}} \,, \qquad (2.1)$$

where $\hat{\mathbf{B}}$ is a q-vector of estimated regression coefficients, and $\Sigma_U\, \mathbf{x}_k$ is the known total of the auxiliary q-vectors $\mathbf{x}_1, \ldots, \mathbf{x}_N$. Estimators of this form were discussed by Cassel, Särndal and Wretman (1976, 1977), Särndal (1980, 1981, 1982), Isaki and Fuller (1982), Wright (1983), and others, but prior to these attempts at unified presentation, various special cases of (2.1) were in widespread use, for example, (1.1), (3.3) and (3.4). Now, (2.1) appeals to an underlying regression

$$y_k = \mathbf{x}_k'\boldsymbol{\beta} + \text{error} \,, \qquad (2.2)$$

about which the only assumption is that y is well explained by the vector $\mathbf{x}$, so that the average (error)^2 is small. (We do not assume (2.2) to be the "true (superpopulation) model" that generated the population values $y_1, \ldots, y_k, \ldots, y_N$.) Alternatively, we can write (2.1) as

$$\hat{t} = \Sigma_U \hat{y}_k + \Sigma_s e_{ks}/\pi_k \quad , \tag{2.3}$$

where

$$\hat{y}_k = \mathbf{x}_k'\hat{\mathbf{B}}$$

is the predicted value for the k:th unit, obtained from the regression fit described below, and the corresponding regression residual is

$$e_{ks} = y_k - \hat{y}_k = y_k - \mathbf{x}_k'\hat{\mathbf{B}} \quad . \tag{2.4}$$

Different suggestions are found in the literature on regression estimation as to the weighting of the observations when $\hat{\mathbf{B}}$ is calculated in (2.1), but the weighting is not of significant consequence for the large sample efficiency of $\hat{t}$; see Särndal (1980), Wright (1983).

Our definition of $\hat{\mathbf{B}}$ springs from wanting to see $\hat{\mathbf{B}}$ as the ordinary π-weighted estimator of the unknown finite population regression vector

$$\mathbf{B} = (\Sigma_U \mathbf{x}_k\mathbf{x}_k'/c_k)^{-1} \Sigma_U \mathbf{x}_k y_k/c_k \quad . \tag{2.5}$$

Clearly, $\mathbf{B}$ represents the result of a hypothetical weighted least squares regression fit of (2.2) to the whole finite population (a "census fit"), where the data point $(y_k, \mathbf{x}_k)$ carries the weight $1/c_k$, which is unrelated to the sampling weight $1/\pi_k$. Particularly simple and often used is the uniform weighting $1/c_k = 1$ for all k. The choice of $1/c_k$ is discussed below. In the design-based approach, an obvious estimator of $\mathbf{B}$ is the

sample weighted analogue

$$\hat{\mathbf{B}} = (\Sigma_s \mathbf{x}_k \mathbf{x}_k'/c_k\pi_k)^{-1} \Sigma_s \mathbf{x}_k y_k/c_k\pi_k \quad . \tag{2.6}$$

Now, $\hat{\mathbf{B}}$ is a function of a number of sample sums of the type $\Sigma_s x_{ik}x_{jk}/c_k\pi_k$ and $\Sigma_s x_{ik}y_k/c_k\pi_k$, where x_{ik} is the value of x_i for the k:th unit. Each of these sample sums is design unbiased and, under appropriate conditions on moments up to fourth order and on the inclusion probabilities, design consistent for its population counterpart. For example, $\Sigma_s x_{ik}x_{jk}/c_k\pi_k$ is design consistent for $\Sigma_U x_{ik}x_{jk}/c_k$. It follows that $\hat{\mathbf{B}}$ is design consistent for $\mathbf{B}$. The conditions, given, for example, in Isaki and Fuller (1982), will be referred to as the regularity conditions. In the rest of the paper we assume that (2.6) is the expression for $\hat{\mathbf{B}}$ in the estimator (2.3).

The weighting $1/c_k$ will be chosen here to achieve simplicity of form and to eliminate the possibility that the estimator (2.3) be unduly influenced by the simultaneous occurrence, for one or several units k, of a large residual e_{ks} and a large sampling weight $1/\pi_k$. We can, in fact, achieve that $\Sigma_s e_{ks}/\pi_k = 0$ for all s by restricting the weights $1/c_k$ to within a certain class of weights. As is easily verified, (2.3) reduces to "the simple projection estimator" (Särndal and Wright (1984)),

$$\hat{t} = \Sigma_U \hat{y}_k \, , \tag{2.7}$$

if the c_k in (2.6) are taken as

$$c_k = \boldsymbol{\lambda}' \mathbf{x}_k \, , \tag{2.8}$$

for any q-vector $\boldsymbol{\lambda}$ independent of k and such that $c_k > 0$ for all k.

That is, the simple form (2.7) obtains if c_k is any convenient positive linear combination of the available x-variable values. For a simple example, supposing that $\mathbf{x}_k$ contains the constant one, then $c_k = 1$ for all k is one choice that yields the simple form (2.7).

An alternative objective for the weights $1/c_k$ is, obviously, to try to choose them so as to minimize the variance of $\hat{t}$ under the given design. However, one can not hope to find a set of weights $1/c_k$ that is best, uniformly for all populations $y_1, \ldots, y_N$. Moreover, even if weights $1/c_k$ may be found that give reduced variance for <u>some</u> populations, the gain in design-based efficiency (compared to a "convenient" set of weights) would at best be modest. To select the strongest possible auxiliary variables x_i is a more important preoccupation in design-based thinking than to seek "optimum" weights $1/c_k$. The practical approach is to use weights that satisfy (2.8) and give the simple form (2.7); such a weighting is assumed from now on.

3. <u>The implied weights</u>

The estimator (2.7) can be expressed as a linear combination of the π-expanded y-values y_k/π_k:

$$\hat{t} = \Sigma_s \, g_{ks} \, y_k/\pi_k \, , \tag{3.1}$$

namely, if we define, for $k \in s$,

$$g_{ks} = (\Sigma_U \, \mathbf{x}_k)' \, (\Sigma_s \, \mathbf{x}_k \mathbf{x}_k'/c_k\pi_k)^{-1} \, \mathbf{x}_k/c_k \, , \tag{3.2}$$

called the g-weight of unit k, where c_k is of the form (2.8). The

g-weights are sample dependent and implied by the equivalence of (2.7) and (3.1). They are somewhat akin to the elements of the "hat matrix" in regression diagnostics, see Hoaglin and Welsch(1978).

Example 3.1. The ratio estimator arises from (2.7) if $\mathbf{x}_k = x_k$, a scalar, and $c_k \propto x_k$:

$$\hat{t} = (\Sigma_U x_k)\,(\Sigma_s y_k/\pi_k)/(\Sigma_s x_k/\pi_k)\,. \tag{3.3}$$

Here, the g-weight is the same for all k in a given s, namely,

$$g_{ks} = (\Sigma_U x_k) \,/\, (\Sigma_s x_k/\pi_k)\,. \qquad \square$$

Example 3.2. The simple regression estimator is obtained from (2.7) if $\mathbf{x}_k = (1, x_k)'$, and $c_k = c$, a constant, for all k:

$$\hat{t} = N\{\tilde{y}_s + \hat{B}(\overline{x}_U - \tilde{x}_s)\}, \tag{3.4}$$

where $\tilde{y}_s = (\Sigma_s y_k/\pi_k)/\hat{N}$; $\tilde{x}_s = (\Sigma_s x_k/\pi_k)/\hat{N}$; $\hat{N} = \Sigma_s\, 1/\pi_k$,

$\overline{x}_U = (\Sigma_U x_k)/N$, and

$$\hat{B} = \{\Sigma_s (x_k - \tilde{x}_s)(y_k - \tilde{y}_s)/\pi_k\}/\{\Sigma_s (x_k - \tilde{x}_s)^2/\pi_k\}\,.$$

From (3.4), we identify the g-weight of unit k as

$$g_{ks} = N\,[\hat{N}^{-1} + (\overline{x}_U - \tilde{x}_s)(x_k - \tilde{x}_s)/\{\Sigma_s (x_k - \tilde{x}_s)^2/\pi_k\}]\,; \tag{3.5}$$

which depends on k as well as on s. $\square$

We now show three properties of the g-weights that will be used later:

Property 1. The g-weights yield "perfect estimates" when applied to the $\mathbf{x}_k$-values:

$$\Sigma_s\; g_{ks}\, \mathbf{x}_k/\pi_k \;=\; \Sigma_U\, \mathbf{x}_k\,. \tag{3.6}$$

Property 2. For $m = 1$ and 2, and for any fixed s,

$$\Sigma_s (g_{ks})^m c_k/\pi_k = \Sigma_U (g_{ks})^{m-1} c_k , \tag{3.7}$$

where c_k is of the form (2.8).

Property 3. For each fixed k, g_{ks} can be viewed as a random variable, the random element being s, whose distribution is determined by the design, and for each k, g_{ks} converges in design probability to unity, under the regularity conditions: Letting

$$T_s = \Sigma_s x_k x_k'/c_k\pi_k , \quad T_U = \Sigma_U x_k x_k'/c_k ,$$

we have that $T_U T_s^{-1} \xrightarrow{P} I_{qxq}$, the qxq identity matrix, and so

$$g_{ks} = \lambda' T_U T_s^{-1} x_k/c_k \xrightarrow{P} \lambda' x_k/c_k = 1 ,$$

where $\xrightarrow{P}$ denotes convergence in design probability, and we have used (2.8). In large samples, g_{ks} may thus be approximated by unity.

4. The suggested variance estimator and its design-based properties.

The primary objective in this paper is to improve current practice for estimating the variance of (3.1). Our requirements for a variance estimator $\hat{V}(\hat{t})$ include all of the following:

(a) good properties with respect to the sampling design that dictates the sample selection;

(b) good properties with respect to an assumed regression model;

(c) simplicity of form and applicability in general, that is, for any design and any linear regression model.

In addition, it is a bonus if the variance estimator possesses

(d) sensible properties under a conditional inference outlook.

With respect to the design, a requirement on $\hat{V}(\hat{t})$ is that it lead to an approximate $100(1-\alpha)\%$ confidence level for t, calculated as

$$\hat{t} \pm z_{1-\alpha/2} \{\hat{V}(\hat{t})\}^{1/2} ,$$

where the constant $z_{1-\alpha/2}$ is exceeded with probability $\alpha/2$ by the unit normal variate. We therefore require that $\hat{V}(\hat{t})$ be design consistent for $V(\hat{t})$. However, many estimators have this property, so the choice is still wide. We look to an assumed model model for guidance in limiting the choice: In addition, our variance estimator should have a zero or very limited bias with respect to a formulated model for the y_k-values. This aspect is discussed in Section 5. (Optimality with respect to both model and design seems a remote hope for a variance estimator. Compromise is thus unavoidable.)

To motivate the variance estimator to be proposed, denote by E_k the "census fit residual":

$$E_k = y_k - \mathbf{x}_k'B , \tag{4.1}$$

where B is the population regression vector (2.5). In view of (2.8), $\Sigma_U E_k = 0$. Using also (3.6), we can express the error of (3.1) as

$$\hat{t} - t = \Sigma_s g_{ks} E_k/\pi_k . \tag{4.2}$$

The approximate variance of (3.1) is commonly given in the literature as

$$V(\hat{t}) = \Sigma\Sigma_U \Delta_{k\ell} E_k E_\ell , \tag{4.3}$$

where $\Delta_{k\ell} = \pi_{k\ell} - \pi_k\pi_\ell$, $E_k = E_k/\pi_k$. Note that (4.3) can be obtained by approximating g_{ks} by unity for all k in (4.2), a step justified in large

samples by Property 3 of the preceding section. The classical technique for turning (4.3) into a variance estimator is to replace the unknown E_k by its sample-based counterpart e_{ks}, which leads to

$$\hat{V}_1 = \hat{V}_1(\hat{t}) = \Sigma\Sigma_s \breve{\Delta}_{k\ell} \breve{e}_{ks}\breve{e}_{\ell s} \quad , \tag{4.4}$$

where $\breve{\Delta}_{k\ell} = \Delta_{k\ell}/\pi_{k\ell}$; $\breve{e}_{ks} = e_{ks}/\pi_k$, with e_{ks} given by (2.4), and $\Sigma\Sigma_s$ is shorthand for $\Sigma_{k\in s}\Sigma_{\ell\in s}$. The estimator (4.4) was considered in early work on the general regression estimator, for example, Särndal (1981, 1982), Särndal and Råbäck (1983). The Yates-Grundy version of (4.4), which applies if the design is of fixed size, is

$$\hat{V}_{YG} = -(1/2)\,\Sigma\Sigma_s \breve{\Delta}_{k\ell}(\breve{e}_{ks} - \breve{e}_{\ell s})^2 . \tag{4.5}$$

Kott (1987) gave conditions under which (4.5) is design consistent for (4.3).

Now, (4.4) and (4.5) are not ideal when the model considerations are added (see Examples 4.1 to 4.3 below). We seek improvement in this regard by the following simple modification of (4.4).

<u>The weighted residual variance estimator.</u> Modify (4.4) by attaching the g-weight g_{ks} to the residual e_{ks}:

$$\hat{V}_g = \hat{V}_g(\hat{t}) = \Sigma\Sigma_s \breve{\Delta}_{k\ell}(g_{ks}\breve{e}_{ks})(g_{\ell s}\breve{e}_{\ell s}) \quad . \tag{4.6}$$

The method was proposed in Särndal (1982). A Yates-Grundy version of (4.6) may also be considered, for the case when s is of fixed size.

Our first important conclusion is that when (4.4) is design consistent, so is (4.6), under the regularity conditions. This follows directly from Property 3 in the preceding section: Since g_{ks} converges in design

probability to unity for all k, the g-weighting of the residuals will not upset the design consistency. That is, from the design-based point of view, (4.4) and (4.6) are equally correct, and the g-weighting is a seemingly trivial step. However, from the model-based point of view the improvement is substantial, as Section 5 will show. Empirical work (not reported here) has shown that (4.6) works well for conficence intervals, even for modest sample sizes. Before proceeding, we show by some examples how g-weighting affects the variance estimator formula.

Example 4.1. With notation as in Example 1.1, the "general design" poststratified estimator is

$$\hat{t} = \Sigma_{h=1}^{H} N_h \tilde{y}_{sh} \tag{4.7}$$

with $\tilde{y}_{sh} = (\Sigma_{sh} y_k/\pi_k)/\hat{N}_h$, where $\hat{N}_h = \Sigma_{sh} 1/\pi_k$. This estimator derives from (2.7) if $\mathbf{x}_k$ is taken as an H-vector composed of $H-1$ entries zero and a single entry "one" indicating the population group to which k belongs, and $c_k = C_h$, a constant, for all units k in the h:th group. The g-weights needed for (4.6) are in this case $g_{ks} = N_h/\hat{N}_h$ for all $k \in s_h$.

Consider in particular simple random sampling (n units drawn from N; $f = n/N$), and let $A_h = \{(n_h - 1)/n_h\}\{n/(n-1)\}$, where n_h is the (random) size of s_h. Then we get from (4.6)

$$\begin{aligned} \hat{V}_g &= (1 - f) \Sigma_{h=1}^{H} A_h N_h^2 s_{sh}^2/n_h \\ &\doteq (1 - f) \Sigma_{h=1}^{H} N_h^2 s_{sh}^2/n_h , \end{aligned} \tag{4.8}$$

where $s_{sh}^2 = \Sigma_{sh} (y_k - \bar{y}_{sh})^2/(n_h-1)$, with $\bar{y}_{sh} = \Sigma_{sh} y_k/n_h$. This

estimator is excellent from the conditional point of view, as already discussed in Example 1.1. By contrast, (4.4) gives a variance estimator in which the weight attached to S_{Sh}^2 is proportional to n_h, instead of proportional to n_h^{-1} as is the case in (4.8). That is, if (4.4) is used, the variance contribution from the h:th group will increase, not decrease, as one would like, when the group's share of the sample gets larger. □

Example 4.2. For the Bernoulli sampling design, the inclusion or non-inclusion in the sample s of a unit k is determined by a Bernoulli experiment: $\pi_k = \Pr(k \in s) = \pi$; $\Pr(k \notin s) = 1 - \pi$ for all k, the experiments being independent. If we take $x_k = 1$, and $1/c_k = 1$ for all k, formula (2.7) leads to the expanded sample mean estimator

$$\hat{t} = N \Sigma_s y_k/n_s \ ,$$

so $g_{ks} = N\pi/n_s$, where the (random) sample size n_s is binomially distributed. Letting $S_s^2 = \Sigma_s (y_k - \overline{y}_s)^2/(n_s-1)$, we get from (4.6)

$$\hat{V}_g = \{(n_s-1)/n_s\} N^2 (1-\pi) S_s^2/n_s \qquad (4.9)$$

$$\doteq N^2 (1-\pi) S_s^2/n_s \ .$$

The factor $N^2 (1-\pi)/n_s$ decreases as n_s increases, which makes good sense, conditionally speaking. By contrast, (4.4) leads in this case to

$$\hat{V}_1 = N^2 (1-\pi) (n_s-1) S_s^2/\{E(n_s)\}^2 \ , \qquad (4.10)$$

where $E(n_s) = N\pi$. Here, the factor n_s-1 in the numerator is unfortunate. "Traditional reasoning" may alternatively lead to the variance estimator

$$\hat{V}_u = N^2 (1-\pi) S_s^2 / \{E(n_s)\} \ , \qquad (4.11)$$

which is also inappropriate, since independent of n_s. All of (4.9) to (4.11) are design consistent, but only (4.9) has definite appeal, conditionally on n_s. □

Example 4.3. We return to the classical ratio estimator (3.3) seen in Example 3.1. The weighted residual variance estimator (4.6) becomes

$$\hat{V}_g = \{(\Sigma_U x_k)/(\Sigma_s x_k/\pi_k)\}^2 \ \Sigma\Sigma_s \Delta_{k\ell} e_{ks} e_{\ell s} \ ,$$

where $e_{ks} = e_{ks}/\pi_k = (y_k - bx_k)/\pi_k$, with $b = (\Sigma_s y_k/\pi_k)/(\Sigma_s x_k/\pi_k)$. For simple random sampling, letting $\bar{x}_s = \Sigma_s x_k/n$; $\bar{y}_s = \Sigma_s x_k/n$ and

$$S_{es}^2 = \Sigma_s \{y_k - (\bar{y}_s/\bar{x}_s)x_k\}^2/(n-1) \ ,$$

we get

$$\hat{V}_g = (\bar{x}_U/\bar{x}_s)^2 N^2\{(1 - f)/n\} S_{es}^2, \qquad (4.12)$$

a variance estimator that has received much attention in recent literature on the ratio estimator, for example, in Wu (1982), Wu and Deng (1983). Royall and Cumberland (1981a) use an estimator that is only slightly different. As a result, (4.12) is now generally considered superior to the "traditional" formula (Cochran (1977), p.155)

$$\hat{V}_1 = N^2\{(1 - f)/n\} S_{es}^2. \qquad \square$$

Example 4.4. The g-weights that apply for the simple regression estimator (3.4) are given by (3.5). In the special case of simple random sampling, the g-weighted variance estimator (4.6) becomes simply

$$\hat{V}_g = N^2 \{(1 - f)/n\} \Sigma_s (g_{ks} e_{ks})^2/(n-1) \ ,$$

where $e_{ks} = y_k - \bar{y}_s - \hat{B}(x_k - \bar{x}_s)$; $g_{ks} = 1 + n(\bar{x}_U - \bar{x}_s)(x_k - \bar{x}_s)/A_{xx}$, with $A_{xx} = \Sigma_s (x_k - \bar{x}_s)^2$; $\hat{B} = \Sigma_s (y_k - \bar{y}_s)(x_k - \bar{x}_s)/A_{xx}$. This result agrees in essence with (but has simpler structure than) the variance estimators suggested by Royall and Cumberland (1978, 1981b) in their work on robust variance estimation in the model-based context. □

5. Properties under the model of the proposed variance estimator

We now examine the properties of the variance estimator (4.6) by assuming for $y_1, \ldots, y_N$ a regression model, denoted ξ and given by

$$y_k = \mathbf{x}_k'\beta + \varepsilon_k \quad ,$$

where the ε_k are independent under the model, and such that

$$E_\xi(\varepsilon_k) = 0; \quad V_\xi(\varepsilon_k) = \sigma_k^2 = \sigma^2 c_k = \sigma^2 \lambda' x_k \qquad (5.1)$$

for some choice of λ. We let E_ξ and V_ξ be the mean and variance operators, respectively, with respect to the model. In the image of (4.6), we consider

$$\hat{V}^* = \Sigma\Sigma_s \breve{\Delta}_{k\ell} (g_{ks}\breve{\varepsilon}_k)(g_{\ell s}\breve{\varepsilon}_\ell) \qquad (5.2)$$

where $\breve{\Delta}_{k\ell} = (\pi_{k\ell} - \pi_k\pi_\ell)/\pi_{k\ell}$, $\breve{\varepsilon}_k = \varepsilon_k/\pi_k$. Since the ε_k are unobservable model errors, $\hat{V}^*$ is obviously not a variance estimator, but serves as a tool for the argument. The "real" estimator, $\hat{V}_g$ given by (4.6), is the "sample copy" of $\hat{V}^*$, obtained by substituting the calculated residual $e_{ks} = y_k - \mathbf{x}_k'\hat{\mathbf{B}}$ for the unobservable ε_k. We have

$$E_\xi(e_{ks} - \varepsilon_k) = 0 \; ; \quad V_\xi(e_{ks} - \varepsilon_k) = \mathbf{x}_k' \, V_\xi(\hat{\mathbf{B}}) \, \mathbf{x}_k \quad .$$

Under general conditions, the difference $\hat{V}^* - \hat{V}_g$ converges in model probability to zero.

A principal concern in the model-based examination, as in Royall and Eberhardt (1975), Royall and Cumberland (1978), Kott (1987), is to see how well a variance estimator succeeds in predicting the model Mean Square Error, $MSE_\xi(\hat{t}) = E_\xi\{(\hat{t}-t)^2\}$. For $\hat{V}^*$ we obtain the following important conclusions:

For any given realized sample s,

(i) $$E_\xi(\hat{V}^*) = \Sigma_s (g_{ks}\sigma_k/\pi_k)^2 - \Sigma_U g_{ks}\sigma_k^2 , \qquad (5.3)$$

(ii) $$MSE_\xi(\hat{t}) = \Sigma_s (g_{ks}\sigma_k/\pi_k)^2 - \Sigma_U \sigma_k^2 , \qquad (5.4)$$

(iii) the Relative Model Bias (RMB) of $\hat{V}^*$ is

$$RMB_\xi(\hat{V}^*) = [E_\xi(\hat{V}^*) - MSE_\xi(\hat{t})]/MSE_\xi(\hat{t})$$

$$= -\Sigma_U (g_{ks}-1)\sigma_k^2 / \{\Sigma_s (g_{ks}\sigma_k/\pi_k)^2 - \Sigma_U \sigma_k^2\} . \qquad (5.5)$$

Here, (5.5) is an immediate consequence of (5.3) and (5.4), the proofs of which use that, for $m = 1$ and 2,

$$\Sigma_s (g_{ks})^m \sigma_k^2/\pi_k = \Sigma_U (g_{ks})^{m-1}\sigma_k^2 , \qquad (5.6)$$

which follows from (3.7), if we note that the model (5.1) assumes that $\sigma_k^2 \propto c_k = \lambda' x_k$. Now, $E_\xi(\varepsilon_k\varepsilon_\ell) = 0$ for all $k \neq \ell$, so

$$E_\xi(\hat{V}^*) = \Sigma_s (g_{ks}\sigma_k/\pi_k)^2 - \Sigma_s (g_{ks}\sigma_k)^2/\pi_k .$$

Here, use (5.6) with $m = 2$ to transform the negative term on the right hand side, which gives (5.3). To prove (5.4), use (3.6) to obtain

$$\hat{t} - t = \Sigma_s (\pi_k^{-1} g_{ks} - 1)\varepsilon_k + \Sigma_{U-s} \varepsilon_k .$$

Now, a straightforward evaluation of $E_\xi\{(\hat{t} - t)^2\} = MSE_\xi(\hat{t})$ leads to (5.4), by use of (5.6) with $m = 1$.

The results (5.3) to (5.5) prompt the following comments:

(1) $RMB_\xi(\hat{V}^*)$ is exactly zero for some important situations. An example is the situation of Example 4.1, where $g_{ks} = N_h/\hat{N}_h$ for all k in group h. Now, if the design is stratified simple random sampling, with the fixed sampling fraction $n_h/N_h = \pi_k$, for k in stratum h, then $g_{ks} = 1$ for all k, and thus $RMB_\xi(\hat{V}^*) = 0$.

(2) Even if not exactly zero, $RMB_\xi(\hat{V}^*)$ is often small. The reason is that in (5.3) and (5.4) the first (positive) term on the right hand side is common to the two expressions; moreover, since it involves π_k^{-2}, this term dominates the second (negative) term in both expressions. In particular, for simple random sampling, this amounts to saying that $RMB_\xi(\hat{V}^*)$ is negligible if the sampling fraction $f = n/N$ is negligible, as illustrated in Example 5.1 below.

(3) It is straightforward, in a case where $RMB_\xi(\hat{V}^*)$ is not already zero, to adjust $\hat{V}^*$ so as to remove its model bias. But in practice this hardly seems worth the effort, because (a) the numerical impact of the bias removal is ordinarily small, and (b) the bias removal may complicate the form of the variance estimator, and (c) the bias removal appeals to a model which, however well it fits the data, is only an assumption. The properties of $\hat{V}^*$ under the model are sufficiently good (that is, $RMB_\xi(\hat{V}^*)$

is sufficiently close to zero) in order that we should not be concerned about the minor model bias. The model is consultative, not normative, for design-based practice.

The comments are illustrated by the following example.

Example 5.1. We reconsider the case of simple random sampling (n units from N, so that $\pi_k = n/N = f$ for all k), and the ratio estimator

$$\hat{t} = N\bar{x}_U \bar{y}_s/\bar{x}_s \ .$$

Consider the "ratio model"

$$y_k = \beta x_k + \varepsilon_k \ , \tag{5.7}$$

with independent errors such that $E_\xi(\varepsilon_k) = 0$, $V_\xi(\varepsilon_k) = \sigma^2 x_k$. The variance structure satisfies (2.8). Since $g_{ks} = \bar{x}_U/\bar{x}_s$, (5.2) becomes

$$\hat{V}^* = (\bar{x}_U/\bar{x}_s)^2 N^2 \{(1 - f)/n\} S_{\varepsilon s}^2 \tag{5.8}$$

where $S_{\varepsilon s}^2 = \Sigma_s (\varepsilon_k - \bar{\varepsilon}_s)^2/(n-1)$, with $\bar{\varepsilon}_s = \Sigma_s \varepsilon_k/n$. A calculation gives

$$RMB_\xi(\hat{V}^*) = - \{f/(1 - f)\}\{(\bar{x}_U - \bar{x}_s)/\bar{x}_{U-s}\} \ ,$$

where $\bar{x}_{U-s} = (\Sigma_{U-s} x_k)/(N-n)$. That is, even if a realized sample is highly unbalanced, so that $(\bar{x}_U - \bar{x}_s)/\bar{x}_{U-s}$ deviates substantially from zero, $RMB_\xi(\hat{V}^*)$ will be near zero if the sampling fraction f is small. Substituting $e_{ks} = y_k - (\bar{y}_s/\bar{x}_s)x_k$ for ε_k in (5.8), we get the variance estimator (4.12) favoured in recent literature.

To continue the example, let us remove the model bias of $\hat{V}^*$ given by (5.8). Let $\hat{\sigma^2}$ be any model unbiased estimator of σ^2. Then

$$\hat{V}^{**} = \hat{V}^{*} + \{N \bar{x}_U(\bar{x}_U - \bar{x}_s)/\bar{x}_s)\}\hat{\sigma^2} \qquad (5.9)$$

is unbiased under the model, that is, $E_\xi(\hat{V}^{**}) = MSE_\xi(\hat{t})$. If in particular we choose $\hat{\sigma^2} = s_{\varepsilon s}^2/\bar{x}_s$, the model unbiased estimator (5.9) becomes

$$\hat{V}^{**} = \{\bar{x}_U\bar{x}_{U-s}/(\bar{x}_s)^2\}\, N^2\, \{(1 - f)/n\}\, s_{\varepsilon s}^2 .$$

Most practitioners would probably feel indifferent between $\hat{V}^{**}$ and $\hat{V}^{*}$. That is, they would consider rather unimportant whether $(\bar{x}_U/\bar{x}_s)^2$ or $(\bar{x}_U\bar{x}_{U-s})/(\bar{x}_s)^2$ be used as a multiplicative factor, but the presence of either factor will most likely be seen as important, because it augments the variance estimate when the realized sample happens to be one in which $\bar{x}_s$ is small compared to $\bar{x}_U$. □

Note finally that $\hat{V}^{*}$ given by (5.2) is made functional by putting the calculated residual $e_{ks} = y_k - \mathbf{x}_k'\hat{\mathbf{B}}$ in the place of ε_k. This step, which gives the weighted residual estimator (4.6), adds a minor model bias. Again, a bias removal term can be applied, but from a practical standpoint, the incentive to do so is not strong, since for modest to large samples, the numerical impact of such a term would be very small. The main objective of the model consideration is, as pointed out, to assist in the choice of a variance estimator.

Concluding remark. It is of interest to compare with the method of Kott (1987), which yields variance estimators for (2.6) that are (a) design consistent and (b) unbiased under an assumed regression model. He proposed to multiply the Yates-Grundy formula (4.5) by the "adjustment ratio"

$$AR(\hat{V}_{YG}) = E_\xi\{(\hat{t}-t)^2\}/E_\xi(\hat{V}_{YG}) .$$

The result, evidently, is to make the variance estimator

$$\hat{V}_{YG}{}^{0} = AR(\hat{V}_{YG})\ \hat{V}_{YG}$$

unbiased for $E_{\xi}\{(\hat{t}-t)^2\} = MSE_{\xi}(\hat{t})$, under the model ξ. This interesting suggestion leads to a correction factor that is often a rather complex expression, even for simple regression models, as shown in the examples of Kott(1987). □

REFERENCES

Cassel, C.M., Särndal, C.E. and Wretman, J.H. (1976). Some results on generalized difference estimation and generalized regression estimation for finite populations. Biometrika **63**, 615-620.

Cassel, C.M., Särndal, C.E. and Wretman, J.H. (1977). Foundations of Inference in Survey Sampling. New York: Wiley.

Cochran, W. G. (1977). Sampling Techniques, 3rd edition. New York: Wiley.

Deng, L. Y. and Wu, C. F. J. (1987). Estimation of variance of the regression estimator. J. Am. Statist. Assoc. **82**, 568-576.

Hoaglin, D. C. and Welsch, R. E. (1978). The hat matrix in regression and ANOVA. The American Statistician **32**, 17-22.

Holt, D. & Smith, T. M. F. (1979). Post Stratification. J. R. Statist. Soc. A **142**, 33-46.

Isaki, C. T. and Fuller, W. A. (1982). Survey design under the regression superpopulation model. J. Am. Statist. Assoc. **77**, 89-96.

Kott, P. S. (1987). The design unbiased regression estimator and its conditional variance. Manuscript seen by the courtesy of the author.

Royall, R. M. and Cumberland, W. G. (1978). Variance estimation in finite ppopulation sampling. J. Am. Statist. Assoc. **73**, 351-358.

Royall, R. M. and Cumberland, W. G. (1981a). An empirical study of the ratio estimator and estimators of its variance. J. Am. Statist. Assoc. **76**, 66-88.

Royall, R. M. and Cumberland, W. G. (1981b). The finite-population linear regression estimator and estimators of its variance - an empirical study. J. Am. Statist. Assoc. **76**, 924-930.

Royall, R. M. and Eberhardt, K. R. (1975). Variance estimates for the ratio estimator. Sankhya C, **37**, 43-52.

Särndal, C. E. (1980). On π-inverse weighting versus best linear weighting in probability sampling. Biometrika **67**, 639-650.

Särndal, C. E. (1981). Frameworks for inference in survey sampling with applications to small area estimation and adjustment for nonresponse. Bull. Int. Stat. Inst. **49**:1, 494-513.

Särndal, C. E. (1982). Implications of survey design for generalized regression estimation of klinear functions. J. Statist. Plan. Inf. **7**, 155-170.

Särndal, C. E. and Råbäck, G. (1983). Variance estimation and unbiasedness for small domain estimators. Statistical Review **1983**:5 (Essays in Honor of Tore E. Dalenius).

Särndal, C. E. and Wright, R. L. (1984). Cosmetic form of estimators in survey sampling. Scand. J. Statist. **11**, 146-156.

Wolter, K. M. (1985). Introduction to Variance Estimation. New York: Springer-Verlag.

Wright, R.L. (1983). Finite population sampling with multivariate auxiliary information. J. Am. Statist. Assoc. **78**, 879-884.

Wu, C. F. (1982). Estimation of variance of the ratio estimator. Biometrika **69**, 183-189.

Wu, C. F. J. and Deng, L. Y. (1983). Estimation of variance of the ratio estimator: an empirical study. In Scientific Inference, Data Analysis and Robustness, ed. by G. E. P. Box et al. New York: Academic Press, 245-277.

R & D Reports är en för U/ADB och U/STM gemensam publikationsserie som fr o m 1988-01-01 ersätter de tidigare "gula" och "gröna" serierna. I serien ingår även **Abstracts** (sammanfattning av metodrapporter från SCB).

R & D Reports, Statistics Sweden, are published by the Department of Research & Development within Statistics Sweden. Reports dealing with statistical methods have green (grön) covers. Reports dealing with EDP methods have yellow (gul) covers. In addition, abstracts are published three times a year (light brown (beige) covers).

Reports published earlier during 1988 are:

Nummer	Titel (författare)
1988:1 (beige)	Abstracts I - Sammanfattningar av metodrapporter från SCB
1988:2 (grön)	Coverage Probabilities for Confidence Intervals Based on Stratified Random Sampling (Jörgen Dalén)
1988:3 (gul)	Base Operators as a Tool for Systems Development (Bo Sundgren)
1988:4 (gul)	Development of Systems Design for National Household Surveys - Report from a short-term mission to Harare, Zimbabwe, 12th-28th January, 1988 (Birgitta Lagerlöf)
1988:5 (grön)	Några råd och synpunkter för rationalisering av produktionsmomentet granskning (Leopold Granquist)
1988:6 (grön)	Hur möta energianvändningen och dess utveckling - några alternativa beräkningar (Urban Aspén)
1988:7 (grön)	Bortfallsbarometer nr 3 (Peter Lundquist)
1988:8 (grön)	Generalized Linear Modeling of Sample Survey Data (Lennart Nordberg)
1988:9 (beige)	Abstracts II - Sammanfattningar av metodrapporter från SCB
1988:10 (grön)	The Weighted Residual Technique for Estimating the Variance of the General Regression Estimator (Carl-Erik Särndal, Bengt Swensson and Jan H. Wretman)

Kvarvarande BEIGE och GRÖNA exemplar av ovanstående promemorior kan rekvireras från Elisabet Klingberg, U/STM, SCB, 115 81 Stockholm, eller per telefon 08-7834178.

Dito GULA exemplar kan rekvireras från Ingvar Andersson, U/ADB, SCB, 115 81 Stockholm, eller per telefon 08-7834147.