ESTIMATING THE MODEL VARIANCE OF A RANDOMIZATION-CONSISTENT REGRESSION ESTIMATOR

Phillip S. Kott and K.R.W. Brewer
National Agricultural Statistics Service, Fairfax, VA 20230, USA and
Australian National University, ACT 0200, Australia

KEY WORDS: Asymptotic; Calibration equation; Weighted-residual-mean-squared-error estimator

1. INTRODUCTION

Suppose we want to estimate a population total, $T = \sum_U y_k$, based an a sample, S, of n elements. Randomization-based theory tells us we can do that with a regression estimator of the form:

$$t = \sum_{k\in S} (y_k/\pi_k) + [\sum_{k\in U} \mathbf{x}_k - \sum_{k\in S} (\mathbf{x}_k/\pi_k)](\sum_{k\in S} \mathbf{x}_k'\mathbf{x}_k d_k/\pi_k)^{-1} \sum_{k\in S} \mathbf{x}_k' y_k d_k/\pi_k, \quad (1)$$

where π_k is the sample selection probability of element k, $\mathbf{x}_k = (x_{k1}, ..., x_{kP})$ is a row vector of values associated with element k, $\sum_U \mathbf{x}_k$ is known, and the d_k are arbitrary non-negative constants. Särndal, Swensson, and Wretman (1989) call t a "general regression estimator" or GREG. From a model-based point of view, however, t is not very general. That is why we do not use that name here.

The estimator t can be written as $t = \sum_S a_k y_k$, where $a_k = (1/\pi_k) + [\sum_U \mathbf{x}_i - \sum_S (\mathbf{x}_i/\pi_k)](\sum_S \mathbf{x}_i'\mathbf{x}_i d_i/\pi_i)^{-1} \mathbf{x}_k' d_k/\pi_k$. Often the d_k in equation (1) are chosen so that these a_k have desirable properties (e.g., being positive; see Brewer, 1999). The a_k have been constructed in such a way that the calibration equation (Deville and Särndal 1992), $\sum_S a_k \mathbf{x}_k = \sum_U \mathbf{x}_k$, is satisfied.

Under mild conditions, t is randomization consistent (see Isaki and Fuller, 1982, who use the synonymous term "design consistent;" Brewer, 1979, introduces a similar property). We will not be deeply interested in randomization-based properties here. Our focus, instead will be on the properties of t as an estimator for T under the linear model:

$$y_k = \mathbf{x}_k\beta + \epsilon_k,$$

where β is an unspecified P-member column vector, $E(\epsilon_k | \mathbf{z}_k) = 0$ for all $k \in U$, and $\mathbf{z}_k = (\mathbf{x}_k, \pi_k, d_k)$. It is easy to see that t is a unbiased estimator for T under the model in the sense that

$$E(t - T) = E(\sum_S a_k y_k - \sum_U y_k) = \sum_S a_k \mathbf{x}_k\beta - \sum_U \mathbf{x}_k\beta = 0 .$$

We concentrate on the model variance of t as an estimator for T (also called the "prediction variance of t") because evidence suggests that such a focus can produce variance estimators with better coverage properties (see Kott 1990). This phenomenon results from the model-based approach's attention to the realized sample and the repercussion of using that sample for inference. Randomization-based inference, by contrast, averages over all possible samples.

We will further assume that $E(\epsilon_k\epsilon_i|\mathbf{z}_k, \mathbf{z}_i,) = \delta_{ik}\sigma_k^2$, where σ_k^2 may be a function of $\mathbf{z}_k$. The variance of t as an estimator for T under the model we have specified is

$$\begin{aligned} E[(t - T)^2] &= E[(\sum_{k\in S} a_k y_k - \sum_{k\in U} y_k)^2] \\ &= E[(\sum_{k\in S} a_k\epsilon_k - \sum_{k\in U} \epsilon_k)^2] \\ &= \sum_{k\in S} a_k^2\sigma_k^2 - 2\sum_{k\in S} a_k\sigma_k^2 + \sum_{k\in U} \sigma_k^2. \end{aligned} \quad (2)$$

The weighted-residual-mean-squared-error estimator (Särndal Swensson, and Wretman 1989; p. 432, eq. (4.6)) for t under Poisson sampling is

$$v_R = \sum_S a_k^2(1 - \pi_k)r_k^2,$$

where

$r_k = y_k - \mathbf{x}_k\mathbf{b}$, and

$\mathbf{b} = (\sum_S \mathbf{x}_k'\mathbf{x}_k d_k/\pi_k)^{-1} \sum_S \mathbf{x}_k' y_k d_k/\pi_k$

is an unbiased estimator for β. We will be concerned here with adapting v_R to estimate the model variance expressed in equation (2). By starting with a variance/mean-squared-error estimator from randomization-based theory, we protect ourselves somewhat from model failure. Many of our results are given in a different context by Royall and Cumberland (1978). Sections 2 through 7 discuss alternativet asymptotic setups. Section 8 provides a summary and a discussion.

2. WHEN THE SAMPLE IS LARGE AND THE POPULATION IS LARGER

The simplest situation to discuss is when the sample size is large and the population is larger. By the former, we mean that terms of the same relative asymptotic order as 1/n can be ignored. By the later, we mean that terms of the same relative asymptotic order as n/N, where N is the size of U, can be ignored.

All the π_k are assumed here to be O(n/N), so that v_R can be approximated by

$$v_0 = \sum_S a_k^2 r_k^2$$

when relative O(n/N) terms are ignored. We will assume, not unreasonably, that both $\sum_S a_k\sigma_k^2 / \sum_S a_k^2\sigma_k^2$ and $\sum_U \sigma_k^2 / \sum_S a_k^2\sigma_k^2$ are O(n/N). Thus, when relative O(n/N) terms are ignorable, the model variance of t and an estimator of T from (2) is (approximately)

$$V_0 = \sum_S a_k^2\sigma_k^2.$$

Observe that

$$E(r_k^2) = E[(y_k - \mathbf{x}_k\mathbf{b})^2]$$

$$= E[(\epsilon_k - \mathbf{x}_k(\sum_S \mathbf{x}_i'\mathbf{x}_i d_i/\pi_i)^{-1}\sum_S \mathbf{x}_i'\epsilon_i d_i/\pi_i)^2]$$

$$= \sigma_k^2 - 2\,\mathbf{x}_k(\sum_S \mathbf{x}_i'\mathbf{x}_i d_i/\pi_i)^{-1}\mathbf{x}_k'(d_k/\pi_k)\sigma_k^2 +$$

$$\mathbf{x}_k(\sum_S \mathbf{x}_i'\mathbf{x}_i d_i/\pi_i)^{-1}[\sum_S \mathbf{x}_i'\mathbf{x}_i(d_i/\pi_i)^2\sigma_i^2](\sum_S \mathbf{x}_i'\mathbf{x}_i d_i/\pi_i)^{-1}\mathbf{x}_k'$$

We can reasonably assume that

$$\mathbf{x}_k(\sum_S \mathbf{x}_i'\mathbf{x}_i d_i/\pi_i)^{-1}\mathbf{x}_k'(d_k/\pi_k)\sigma_k^2 \text{ and}$$

$$\mathbf{x}_k(\sum_S \mathbf{x}_i'\mathbf{x}_i d_i/\pi_i)^{-1}[\sum_S \mathbf{x}_i'\mathbf{x}_i(d_i/\pi_i)^2\sigma_i^2](\sum_S \mathbf{x}_i'\mathbf{x}_i d_i/\pi_i)^{-1}\mathbf{x}_k'$$

are O(1/n) when P is fixed as n grows arbitrarily large. Consequently, v_0 is an (approximately) unbiased estimator for V_0 and, thus, the model variance of t as an estimator for T when both relative O(n/N) and O(1/n) terms are ignorable.

3. WHEN THE POPULATION IS LARGE AND σ_k^2 IS KNOWN UP TO A CONSTANT

In this section O(n/N) is again ignorably small, but O(1/n) may not be. If $\sigma_k^2 = kv_k$ for known v_k, then from equation (3):

$$E(r_k^2) = \sigma_k^2\{1 - 2\,\mathbf{x}_k(\sum_S \mathbf{x}_i'\mathbf{x}_i d_i/\pi_i)^{-1}\mathbf{x}_k'(d_k/\pi_k) +$$
$$\mathbf{x}_k(\sum_S \mathbf{x}_i'\mathbf{x}_i d_i/\pi_i)^{-1}[\sum_S \mathbf{x}_i'\mathbf{x}_i(d_i/\pi_i)^2 v_i](\sum_S \mathbf{x}_i'\mathbf{x}_i d_i/\pi_i)^{-1}\mathbf{x}_k'/v_k\}.$$

Thus, an (approximately) unbiased estimator for the model variance of t as an estimator for T is

$$v_{(2)} = \sum_S a_k^2 r_k^{(2)},$$

where

$$r_k^{(2)} = r_k^2/\{1 - 2\,\mathbf{x}_k(\sum_S \mathbf{x}_i'\mathbf{x}_i d_i/\pi_i)^{-1}\mathbf{x}_k'(d_k/\pi_k) +$$
$$\mathbf{x}_k(\sum_S \mathbf{x}_i'\mathbf{x}_i d_i/\pi_i)^{-1}[\sum_S \mathbf{x}_i'\mathbf{x}_i(d_i/\pi_i)^2 v_i](\sum_S \mathbf{x}_i'\mathbf{x}_i d_i/\pi_i)^{-1}\mathbf{x}_k'/v_k\}.$$

Note that $v_{(2)}$ remains approximately unbiased when O(1/n) is ignorably small even when $\sigma_k^2 = kv_k$ fails providing that

$$\mathbf{x}_k(\sum_S \mathbf{x}_i'\mathbf{x}_i d_i/\pi_i)^{-1}\mathbf{x}_k'(d_k/\pi_k) \text{ and}$$

$$\mathbf{x}_k(\sum_S \mathbf{x}_i'\mathbf{x}_i d_i/\pi_i)^{-1}[\sum_S \mathbf{x}_i'\mathbf{x}_i(d_i/\pi_i)^2 v_i](\sum_S \mathbf{x}_i'\mathbf{x}_i d_i/\pi_i)^{-1}\mathbf{x}_k$$

are O(1/n).

4. WHEN THE POPULATION IS LARGE AND σ_k^2 IS NOT KNOWN UP TO A CONSTANT

Let us rewrite $r_k = y_k - \mathbf{x}_k\mathbf{b}$ as

$$r_k = \epsilon_k - \mathbf{x}_k\mathbf{C}\epsilon = \epsilon_k - \mathbf{g}_k\epsilon,$$

where $(\sum_S \mathbf{x}_i'\mathbf{x}_i d_i/\pi_i)^{-1}\mathbf{x}_k'(d_k/\pi_k)$ is the kth column of

$\mathbf{C} = \{c_{pk}\}_{Pxn}$, $\epsilon = (\epsilon_1, ..., \epsilon_n)'$, and

$\mathbf{g}_k = \mathbf{x}_k\mathbf{C} = (g_{k1}, ..., g_{kn})'$.

Now $E(r_i^2) = (1 - 2g_{ii})\sigma_i^2 + \sum_k g_{ik}^2\sigma_k^2$ or

$$E[r_i^2/(1 - 2g_{ii})] = \sigma_i^2 + (1 - 2g_{ii})^{-1}\sum_{k\in S} g_{ik}^2\sigma_k^2$$

$$= \sigma_i^2 + (1 - 2g_{ii})^{-1}\sum_{k\in S}[\sum_{p=1}^{P} x_{ip}c_{pk}]^2\sigma_k^2 \qquad (4)$$

$$= \sigma_i^2 + (1 - 2g_{ii})^{-1}\sum[\sum_{p\geq p'}(2 - \delta_{pp'})x_{ip}c_{pk}x_{ip'}c_{p'k}]\sigma_k^2.$$

Equation (4) can also be expressed in matrix form as

$$E(\mathbf{r}^*) = \sigma^2 + \mathbf{Q}\mathbf{H}\sigma^2,$$

where $\mathbf{r}^* = (r_1^2/[1 - 2g_{11}], ..., r_n^2/[1 - 2g_{nn}])'$,
$\sigma^2 = (\sigma_1^2, ..., \sigma_n^2)$, $\mathbf{Q} = \{q_{if}\}_{nxF}$, F = P(P+1)/2,
$q_{if} = (1 - 2g_{ii})^{-1}(2 - \delta_{pp'})x_{ip}x_{ip'}$ (f corresponds to pp′),
$\mathbf{H} = \{h_{fk}\}_{Pxn}$, and $h_{fk} = c_{pk}c_{p'k}$.

Observe that if every $|c_{fk}| \leq O(1/n)$, then each $|h_{fk}| \leq O(1/n^2)$.

Letting $\mathbf{I}_{(n)}$ denote the nxn identity matrix, an unbiased estimator for σ^2 is

$$\mathbf{s}^2 = (\mathbf{I}_{(n)} + \mathbf{QH})^{-1}\mathbf{r}^*$$

when $(\mathbf{I}_{(n)} + \mathbf{QH})$ is invertible (see Chew 1970). If F < n, a more computationally convenient form is

$$\begin{aligned}\mathbf{s}^2 &= \{\mathbf{I}_{(n)} - \mathbf{QH} + (\mathbf{QH})^2 - (\mathbf{QH})^3 + \dots\}\,\mathbf{r}^* \\ &= \{\mathbf{I}_{(n)} - \mathbf{Q}[\mathbf{I}_{(F)} - \mathbf{HQ} + (\mathbf{HQ})^2 + \dots]\mathbf{H}\}\mathbf{r}^* \\ &= \{\mathbf{I}_{(n)} - \mathbf{Q}[\mathbf{I}_{(F)} + \mathbf{HQ}]^{-1}\mathbf{H}\}\mathbf{r}^*,\end{aligned}$$

which requires $(\mathbf{I}_{(F)} + \mathbf{HQ})$ to be invertible.

Thus,

$$v_s = \sum_S a_k^2 s_k^2,$$

where $\mathbf{s}^2 = (s_1^2, \dots, s_n^2)$, is an unbiased estimator for the model variance of t as an estimator for T when relative O(n/N) terms can be ignored.

5. A SIMPLE EXAMPLE COMPARING $r_k^{(2)}$ AND s_k^2

Consider the following simple example. Suppose P=1, so the vector $\mathbf{x}_k$ reduces to the scalar x_k. In addition, let $d_k = 1/x_k$, a popular formulation. The estimator t becomes the ratio estimator,

$$t = \sum_U x_k \left(\sum_S y_k/\pi_k\right) / \left(\sum_S y_k/\pi_k\right).$$

The element-k residual has the form:

$$r_k = y_k - x_k\left(\sum_S y_i/\pi_i\right) / \left(\sum_S y_i/\pi_i\right).$$

When we can assume $\sigma_k^2 = kv_k$, $r_k^{(2)}$ becomes

$$r_k^{(2)} = r_k^2 / \{1 - [2(x_k/\pi_k)/(\textstyle\sum_S x_i/\pi_i)] + (x_k^2/v_k)(\textstyle\sum_S v_i/\pi_i^2) / (\textstyle\sum_S x_i/\pi_i)^2\}.$$

Observe that when $\pi_k = x_k = \lambda v_k$, $r_k^{(2)} = nr_k/(n-1)$.

Deriving s_k^2 is greatly simplified because P = F = 1. The matrix $\mathbf{Q}$ becomes the column vector, $\mathbf{Q} = (x_1^2/(1 - 2\omega_1), \dots, x_n^2/(1-2\omega_n))'$, where $\omega_k = g_{kk} = (x_k/\pi_k) / \sum_S (x_i/\pi_i)$. Similarly, $\mathbf{H}$ becomes the row vector, $\mathbf{H} = ((\omega_1/x_1)^2, \dots, (\omega_n/x_n)^2)$. After some manipulation we get

$$s_k^2 = [r_k^2/(1-2\omega_k)] - [x_k^2/(1-2\omega_k)][\textstyle\sum_S(\omega_i r_i/x_i)^2/(1-2\omega_i)]/[1+\sum_S \omega_i^2/(1-2\omega_i)].$$

This is similar to what we would get from replacing the v_i in the above expression for $r_k^{(2)}$ by the corresponding r_i^2 (the difference is $O(1/n^2)$). That should not come as a surprise.

6. WHEN THE POPULATION IS NOT LARGE AND $\sigma_k^2 = \mathbf{x}_k\gamma$

Suppose we cannot assumed that O(n/N) terms are ignorably small. If the element variances have the form $\sigma_k^2 = \mathbf{x}_k\gamma$ for some not-necessarily-specified γ, then equation (2) can be rewritten as

$$\begin{aligned}E[(t - T)^2] &= \sum_{k\in S} a_k^2\sigma_k^2 - \sum_{k\in S} a_k\sigma_k^2 \\ &= \sum_{k\in S} (a_k^2 - a_k)\sigma_k^2, \qquad (5)\end{aligned}$$

because $\sum_S a_k\sigma_k^2 = \sum_S a_k\mathbf{x}_k\gamma = \sum_U \mathbf{x}_k\gamma = \sum_U \sigma_k^2$.

An (approximately) unbiased estimator for the variance of t as an estimator would simply replace the σ_k^2 in (4) with r_k^2, $r_k^{(2)}$, or s_k^2 depending on what other assumptions are being made.

Observe that even when n is large, and we choose $v_{m1} = \sum_S (a_k^2 - a_k)r_k^2$ as the variance estimator, it differs from the randomization estimator, v_R, when $a_k \neq 1/\pi_k$. The model-based and randomization variance estimators are asymptotically equivalent under mild conditions, however, because $a_k\pi_k = 1 + O_p(1/\sqrt{n})$, where P here denotes the probability space generated by the random sampling.

7. OTHER POSSIBILITIES

Suppose we can not assume that $\sigma_k^2 = \mathbf{x}_k\gamma$ for some γ. We can, however, assume that n is large. Under mild conditions, $\sum_S a_k\sigma_k^2 / \sum_U \sigma_k^2 = 1 + O_p(1/\sqrt{n})$. Although we are still interested exclusively in model expectations, we can nonetheless use this randomization-based equality to establish the relative size of terms when n is large.

This equality provides alternative justification for the variance estimators discussed in the last section. It may appear that the replacement of σ_k^2 in equation (5) by r_k^2 is the only sensible policy when n is large because it is computationally the easiest. Suppose, however, that O(n/N) = O(1/n), and we are willing to ignore relative bias term of probability order $n^{-3/2}$, but not of order 1/n. It then becomes more reasonable to use $r_k^{(2)}$ (when v_k is assumed known) or s_k^2 (otherwise).

Finally, suppose neither the sample nor the population is large, and we can *not* assume that $\sigma_k^2 = \mathbf{x}_k\gamma$ for some γ. This leave us no alternative better than assuming some model structure for all the σ_k^2, fitting that model with the in the sample, and then applying the results to estimate σ_k^2 for those elements not in the sample.

8. SUMMARY AND DISCUSSION

We have essentially proposed three estimators for the model variance of t and an estimator for T in most situations:

$$v_{m1} = \sum_S (a_k^2 - a_k) r_k^2,$$

$$v_{m2} = \sum_S (a_k^2 - a_k) r_k^{(2)}, \text{ and}$$

$$v_{m3} = \sum_S (a_k^2 - a_k) s_k^2.$$

The first is the simplest to compute, while the last is nearly unbiased under the broadest range of circumstances.

All three of these variance estimators have the same large-sample-size randomization-based properties as v_R when the sample is drawn using Poisson sampling. The randomization-mean-squared-error estimator, v_R, itself is nearly (i.e., large-n asymptotically) randomization unbiased under Poisson sampling. Given a more general design, the weighted-residual-mean-squared-error estimator of Särndal, Swensson, and Wretman (1989) is

$$v_R' = v_R + \sum_{k,i \in S; k \neq i} (1 - \pi_k \pi_i / \pi_{ik}) a_k r_k a_i r_i.$$

For simple random sampling, this reduces to

$$v_R' = [n/(n-1)]\{v_R - (1 - n/N)(\sum_S a_k r_k)^2/n\},$$

which is large-n asymptotically identical to v_R when $\sum_S a_k r_k < O_P(N)$. A similar argument can me made for broader range of designs satisfying

$$\pi_k \pi_i / \pi_{ik} \approx n/(n-1) + O(1/N)$$

when $O(n/N) = O(1/n)$.

Whatever general theoretical advantage v_R' offers over v_R in terms of a potentially reduced randomization-bias can be lost to increased randomization variance resulting from the as many as $n(n-1)/2$ distinct terms in v_R' but not in v_R. In quite a few practical situations the model will come close to holding, and the three model-based variance estimators proposed above will not only estimate the model variance of t as an estimator for T better than v_R', they will estimate the randomization mean squared error of t better as well.

From a purely model-based point of view, any model-unbiased estimator for T of the form $t = \sum_S a_i y_i$ will satisfy the calibration equation. The three estimators for the model variance of t as an estimator for T retain the same properties as those discussed in the text except that $a_k \pi_k = 1 + O_P(1/n)$ can not be assumed. In order to compute these model variance estimators, however, we need define the sample residuals. In Section 1, they are defined by

$$r_k = y_k - \mathbf{x}_k (\sum_S \mathbf{x}_i' \mathbf{x}_i d_i/\pi_i)^{-1} \sum_S \mathbf{x}_i' y_i d_i/\pi_i,$$

where the choice for the d_k imply the a_k. From a model-based point of view, the two need not be related. Indeed, given values for the a_k that satisfy the calibration equation, any choice for d_k will do to define the r_k.

REFERENCES

BREWER, K.R.W. (1979). A class of robust sampling designs for largescale surveys. *Journal of the American Statistical Association*, 74, 911-915.

BREWER, K.R.W. (1999). Cosmetic calibration with unequal probability sampling. *Survey Methodology,* 25, 205-212.

CHEW, V. (1970). Covariance matrix estimation in linear models. *Journal of the American Statistical Association*, 65, 173-181.

DEVILLE, J-C. and SÄRNDAL, C-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.

ISAKI, and FULLER, W.A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77, 89-96.

KOTT, P.S. (1990). Estimating the conditional variance of a design consistent regression estimator. *Journal of Statistical Planning and Inference*, 24, 287-296.

ROYALL, R.M. and CUMBERLAND, W.G. (1978). Variance estimation in finite population sampling, *Journal of the American Statistical Association*, 72, 351-358.

SÄRNDAL, C-E, SWENSSON, B., and WRETMAN, J. (1989). The weighted residual technique for estimating the variance of the general regression estimator of a finite population total. *Biometrika*, 76, 527-537,

SÄRNDAL, C-E, SWENSSON, B., and WRETMAN, J. (1992). *Model Assisted Survey Sampling.* New York: Springer.