Statistica Sinica **8**(1998), 1165-1173

A MODEL-BASED EVALUATION OF SEVERAL WELL-KNOWN VARIANCE ESTIMATORS FOR THE COMBINED RATIO ESTIMATOR

Phillip S. Kott

National Agricultural Statistics Service

Abstract: This note explores variance estimation of a combined ratio estimator from a purely model-based viewpoint. It shows that given a sample containing two distinct primary sampling units in every stratum, many of the standard randomization-based variance estimators are equally good estimators of model variance. In fact, model-based comparisons of four variance estimators, a form of the linearization variance estimator, the standard form of the jackknife, and two common forms of balanced half sampling parallel well-known randomization-based results. By contrast, a "textbook" version of the linearization variance estimator does not estimate model variance as well as these four. Part of the analysis can be extended to estimated linear regression coefficients and to regression estimators for population means expressible in projection form.

Key words and phrases: Balanced half sampling, jackknife, linearization, primary sampling unit.

1. Introduction

There is no longer much controversy about how to estimate the variance of a combined ratio estimator given sample data from two primary sampling units per stratum. Assuming that the finite population correction can be ignored, many variance estimators based on Taylor series linearization and replication have identical first-order randomization-based properties and well-known second-order properties (Rao and Wu (1985)). This note shows that one form of the linearization variance estimator, the standard form of the jackknife, and two common forms of balanced half sampling have model-based properties that closely parallel their randomization-based ones. Another version of the linearization variance estimator, by contrast, is not as effective at estimating model variance. The proofs of the near model unbiasedness of the four variance estimation techniques alluded to above can be extended to cover variance estimators for linear regression coefficients, and to regression estimators for population means expressible in projection form.

2. The Framework

In this note, we restrict our attention to a stratified single stage or multistage sampling design. Let h denote one of H strata, and assume that the first-stage sample consists of two distinct primary sampling units (PSU's) per stratum. The mechanism used to choose those units, and when appropriate the elements enumerated within the sampled PSU's, has little bearing on the purely model-based analysis to be conducted here.

The interested reader is directed to the works of Richard Royall for more background on the purely model-based view of survey sampling and inference taken here; for example, Royall (1970), Royall (1976), Royall and Cumberland (1978) and Royall and Cumberland (1981). This view, although useful for our present purpose, is sharply questioned in Hansen, Madow, and Tepping (1983).

Let $n = 2H$ be the total number of PSU's in the sample. In this note, we consider the variance of an estimator of the form:

$$t = \frac{\sum_{h=1}^{H}(y_{h1} + y_{h2})/n}{\sum_{h=1}^{H}(x_{h_1} + x_{h2})/n} = \frac{y}{x}, \tag{1}$$

where z (which can be y or x)$= \sum^{H}(z_{h1} + z_{h2})/n$. Each z_{hi} has a nonnegative value. Often, z_{hj} will be a survey-weighted aggregate derived from subsampled elements within PSU j of stratum h. The totals in both the numerator and denominator of equation (1) have been divided by n for subsequent asymptotic analyses.

The expression t in equation (1) is an unbiased estimator of b in the model:

$$y_{hj} = bx_{hj} + e_{hj}, \tag{2}$$

where the e_{hj} are independent of each other, $E(e_{hj}) = 0$, and $\mathrm{Var}(e_{hj}) = v_{hj}$. The (model) variance of t is $\mathrm{Var}(t) = \mathrm{Var}\{\sum^{H}(y_{h1} + y_{h2})/\sum^{H}(x_{h1} + x_{h2})\} = \mathrm{Var}\{\sum^{H}(e_{h1} + e_{h2})/\sum^{H}(x_{h1} + x_{h2})\} = (nx)^{-2}\sum(v_{h1} + v_{h2})$. If we assume that the v_{hj} are bounded from above as the number of strata — and thus n— grows arbitrarily large, $\mathrm{Var}(t) = O(1/n)$.

The model in equation (2) treats the x_{hj} as fixed. Treating the x_{hj} as random, which some argue is more natural in a survey sampling setting, has little affect on the analyses to be presented. They can simply be viewed as being conditioned on the realized values of the x_{hj}. Indeed, Royall uses the term "conditional" to refer to the purely model-based inferences made about estimators like t.

3. The Linearization Variance Estimator

We do not know the values of the e_{hj}. If we did, a potential unbiased estimator for $\mathrm{Var}(t)$ would be $(nx)^{-2}\sum(e_{h1} - e_{h2})^2$. An obvious way to estimate each

e_{hj} is with $y_{hj} - (y/x)x_{hj}$. This leads to one version of the so-called linearization variance estimator:

$$\text{var}_L(t) = (nx)^{-2} \sum_{h=1}^{H} \{[y_{h1} - (y/x)x_{h1}] - [y_{h2} - (y/x)x_{h2}]\}^2. \tag{3}$$

The name comes from the randomization-based process used to derive $\text{Var}_L(t)$.

Observe that $\text{Var}_L(t) = (nx)^{-2} \sum[(y_{h1} - y_{h2}) - (y/x)(x_{h1} - x_{h2})]^2 = (nx)^{-2} \sum[(e_{h1} - e_{h2}) - (e/x)(x_{h1} - x_{h2})]^2$. Taking the model expectation of $\text{var}_L(t)$ yields:

$$\begin{aligned} E[\text{var}_L(t)] = (nx)^{-2} \sum [(v_{h1} + v_{h2}) - 2(v_{h1} - v_{h2})(x_{h1} - x_{h2})/(nx) \\ + v(x_{h1} - x_{h2})^2/(nx^2)] = \text{Var}(t) + O(1/n^2), \end{aligned} \tag{4}$$

assuming that the v_{hj} and x_{hj} are bounded from above as H grows arbitrarily large. Thus, the relative model bias of $\text{var}_L(t)$ under mild conditions is $O(1/n)$. The twin assumptions that the v_{hj} and x_{hj} are bounded from above are stronger than we need to establish the asymptotic results in this note. We invoke them here because of their simplicity.

4. The Jackknife Variance Estimator

Although the literature contains several versions of the jackknife, we consider only the following one, which has become the standard in this context (see Rust (1985), p. 387):

$$\text{var}_J(t) = (1/2) \sum_{h=1}^{H} \sum_{j=1}^{2} [(y_{(hj)}/x_{(hj)}) - (y/x)]^2,$$

where $z_{(hj)} = (1/n)[\sum_{g \neq h}(z_{g1} + z_{g2}) + 2z_{hj'}]$, $j' = 2$ when $j = 1$, and $j' = 1$ when $j = 2$.

Observe that $z_{(hj)} - z = (z_{hj'} - z_{hj})/n$. A little manipulation reveals

$$\begin{aligned} \text{var}_J(t) &= (1/2) \sum \sum [(e_{(hj)}/x_{(hj)}) - (e/x)]^2 \\ &= (1/2) \sum \sum [(e_{(hj)} - e)/x + e_{(hj)}(x - x_{(hj)})/(x x_{(hj)})]^2 \\ &= (2nx)^{-1} \sum \sum [(e_{hj'} - e_{hj}) + (e_{(hj)}/x_{(hj)})(x_{hj} - x_{hj'})]^2 \\ &= (2nx)^{-1} \sum \sum [(e_{hj'} - e_{hj}) + (e_{(hj)}/x)(x_{hj} - x_{hj'}) \\ &\quad + e_{(hj)}(x_{hj} - x_{hj'})^2/(nx) + O(1/n^2)]^2. \end{aligned}$$

So

$$\begin{aligned} E[\text{var}_J(t)] = (nx)^{-2} \sum [(v_{h1} + v_{h2}) - 2(v_{h1} - v_{h2})(x_{h1} - x_{h2})/(nx) \\ + v(x_{h1} - x_{h2})^2/(nx^2) + O(1/n^2)], \end{aligned} \tag{5}$$

which is $E[\text{var}_L(t)] + O(1/n^3)$.

5. Balanced Half Sampling

Suppose one of the two sampled PSU's within each stratum were assigned to an entity named replicate r, and suppose there were R such replicates fully balanced (see Wolter (1985)) so that

$$\sum_{r=1}^{R}(x_r - x) = 0, \text{ and} \tag{6.1}$$

$$\sum_{r=1}^{R}(x_r - x)(z_r - z) = \sum_{h=1}^{H}(x_{h1} - x_{h2})(z_{h1} - z_{h2})/n^2, \tag{6.2}$$

where $z_r = (1/H)\sum_{hj\varepsilon r} z_{hj}$ (for our not-yet-revealed purposes, z can be either x or v). A set of replicates satisfying equation (6.2), but not necessarily (6.1) is said to be balanced.

One common balanced half sampling variance estimator for t is

$$\text{var}_{H1}(t) = (1/R)\sum_{r=1}^{R}[(y_r/x_r) - (y/x)]^2. \tag{7}$$

Let $v_{[r1]} = [(y_r/x_r) - (y/x)]^2 = [(e_r/x_r) - (e/x)]^2$, and observe that $E(e_r^2) = (2/n)v_r$, $E(e_r^2 e) = v_r/n$, and $E(v_{[r1]}) = E[(e_r - e)/x + e_r(x - x_r)/(xx_r)]^2 = (nx)^{-2}[(\sum\sum v_{hj}) - 2n(v_r/x_r)(x_r - x) + 2n(v_r/x_r^2)(x_r - x)^2]$.

The selection of PSU's for inclusion in replicate r has effectively been done at random, at least across strata. In what follows, we treat this selection as if it had been done randomly to make use of randomization-based property: $x_r - x$ is $O_p(1/\sqrt{n})$. This implies that $1/x_r = (1/x)\{1 - [x_r - x]/x + O_p(1/n)\}$. Thus, we can write

$$\begin{aligned}
E(v_{[r1]}) &= (nx)^{-2}[(\sum\sum v_{hj}) - 2n(v_r/x)(x_r - x) + 4n(v_r/x^2)(x_r - x)^2 + O_p(n^{-3/2})] \\
&= (nx)^{-2}[(\sum\sum v_{hj}) - 2n(v/x)(x_r - x) - 2n(v_r - v)(x_r - x)/x \\
&\quad + 4n(v/x^2)(x_r - x)^2 + O_p(n^{-3/2})] \\
&= (nx)^{-2}[(\sum\sum v_{hj}) - 2n(v/x)(x_r - x) + O_p(1/n)],
\end{aligned} \tag{8}$$

which is equal to $\text{Var}(t) + O_p(n^{-3/2})$. Consequently, the relative bias of each $v_{[r1]}$ term in $\text{var}_{H1}(t) = \sum^R v_{[r1]}/R$ is $O_p(1/\sqrt{n})$. The relative variance of $\text{var}_{H1}(t)$ itself, however, is $O_p(1/n)$ because equation (6.1) forces $2n(v/x)\sum^R(x_r - x) = 0$. In fact, from equations (8) and (6.2), we can deduce

$$\begin{aligned}
E[\text{var}_{H1}(t)] = (nx)^{-2}\sum[&(v_{h1} + v_{h2}) - 2(v_{h1} - v_{h2})(x_{h1} - x_{h2})/(nx) \\
&+ 4v(x_{h1} - x_{h2})^2/(nx^2) + O_p(n^{-3/2})],
\end{aligned} \tag{9}$$

which is equal to $\mathrm{Var}(t) + O_p(1/n^2)$.

Another common balanced half sampling variance estimator is

$$v_{H2}(t) = (1/[4R]) \sum_{r=1}^{R} [(y_r/x_r) - (y_{r'}/x_{r'})]^2, \tag{10}$$

where $z_{r'} = 2z - z_r$. Let $v_{[r2]} = (1/4)[(y_r/x_r) - (y_{r'}/x_{r'})]^2$. Let us again treat the allocation of PSU's to replicate r as if it were random. Thus, $1/x_r^2 = (1/x_2)\{1 - 2[x_r - x]/x + 3([x_r - x]/x)^2 + O_p(n^{-3/2})\}$; and

$$\begin{aligned} E(v_{[r2]}) &= (1/[2n])[v_r/x_r^2 + v_{r'}/x_{r'}^2] \\ &= (1/[2n])\{(v_r/x^2)(1 - 2(x_r - x)/x + 3[(x_r - x)/x_r]^2) \\ &\quad + (v_{r'}/x^2)(1 + 2(x_r - x)/x + 3[(x_r - x)/x_r]^2) + O_p(n^{-3/2})\} \qquad (11) \\ &= (nx)^{-2}\{\sum\sum v_{hj} - 2n(v_r - v)(x_r - x)/x + 3nv[(x_r - x)/x]^2 + O_p(n^{-3/2})\}, \end{aligned}$$

which is equal to $\mathrm{Var}(t) + O_p(1/n^2)$.

Unlike $v_{[r1]}$, the relative model variance of $v_{[r2]}$ is $O_p(1/n)$ as is that of $\mathrm{Var}_{H2}(t)$. In fact, continuing from equation (11), we have

$$\begin{aligned} E[\mathrm{var}_{H2}(t)] = (nx)^{-2} \sum [(v_{h1} + v_{h2}) - 2(v_{h1} - v_{h2})(x_{h1} - x_{h2})/(nx) \\ + 3v(x_{h1} - x_{h2})^2/(nx^2) + O_p(n^{-3/2})]. \qquad (12) \end{aligned}$$

6. Discussion

The linearization variance estimator, $\mathrm{var}_L(t)$, and the jackknife variance estimator, $\mathrm{Var}_J(t)$ are both nearly unbiased estimators for the model variance of t. That is to say, they have relative model biases of $O(1/n)$. In fact, equations (4) and (5) reveal that the model biases of the two variance estimators are $O(1/n^3)$. This result closely parallels equation (47) in Rao and Wu's (1985) strictly randomization-based analysis (assuming all $n_h = 2$ in their equation, each s_{e^2xh} becomes zero).

The two balanced half sampling variance estimators under discussion, $\mathrm{var}_{H1}(t)$ and $\mathrm{var}_{H2}(t)$, have relative model biases of $O_p(1/n)$. Thus, they too are nearly unbiased estimators of the model variance of t. Other versions of the balanced half sampling variance estimator have properties similar to those of the two versions discussed here. Explicit treatment of these other variations has been avoided for the sake of brevity.

The alert reader may have already observed that the large sample properties of random sampling were invoked to support the near model unbiasedness of the two balanced half sampling variance estimators. Nevertheless, it is the model

variance of t that is being estimated and the model expectation of the variance estimators that is being assessed. At no time are we averaging over potential samples as would be done in a randomization-based analysis.

The four main results of this note were captured in equations (4), (5), (9), and (12). Comparing them, it is easy to see that to $O_p(n^{-5/2})$, $E[\text{var}_{H1}(t)] \geq E[\text{var}_{H2}(t)] \geq E[\text{var}_L(t)] = E[\text{var}_J(t)]$. The first part of this is similar to equation (54) in Rao and Wu. In fact, the main differences between the four equations above and their analogues in Rao and Wu is that (1) the expressions here are conditioned on realized x_{hj} and v_{hj} values in typical model-based fashion and not averaged over potential realizations of those values, and (2) the Rao and Wu results for the two balanced half sampling estimators contain an additional nonnegative term which has a model expectation of zero (c in their equation (57) captures the square of the correlation, if any, between x_{hj} and $y_{hj} - x_{hj}[E_p(y)/E_p(x)]$, where E_p denotes expectation under the sampling design).

There is another important difference between the randomization-based analysis in Rao and Wu and the model-based analysis described here. In our set up, t is an estimator for the model parameter b in equation (2), while the sampled PSU's are assumed to be distinct and thus (model) independent. In Rao and Wu's framework t is an estimator for Y/X , where Y and X are the randomization expectations of y and x respectively. In addition, the PSU's are assumed to be sampled independently, which means with replacement sampling is used within strata. As a result, the sampled PSU's need not be distinct.

The relationship between the v_{hj} and x_{hj} in equations (4), (5), (9), and (12) will depend on the sampling design used to select the sample. For many self-weighting designs (in which every sampled element has an identical selection probability), x_{hj} is an estimate of the population size of stratum h. Often v_{hj} is roughly proportional to x_{hj}. If that proportional relationship were exact, then it is easy to see that $\text{Var}_L(t) \approx \text{Var}_J(t)$ would have a slight downward model bias, while $\text{Var}_{H1}(t)$ and $\text{var}_{H2}(t)$ would have a slight upward bias.

For a single stage design with a sole target variable of interest, it is good practice from a model-assisted viewpoint to select units with unequal selection probabilities in such a way that the v_{hj} are all roughly equal (see Brewer (1963)). If all the v_{hj} were, in fact, equal, then each of the four variance estimators under discussion would have a small positive model bias.

Finally, let us turn to the definition of the linearization variance estimator. Suppose X were known, and $(nX)^{-2}$ were used in place of $(nx)^{-2}$ in equation (3), as is recommended by most textbooks (for example, see Cochran (1977), pp. 169-171; Särndal, Swensson and Wretman (1992), is an exception). If we assume that the PSU's have been randomly selected so that $x - X$ is $O_p(1/\sqrt{n})$, then the variance estimator, $\text{var}_L^*(t) = (nX)^{-2} \sum\{[(y_{h1} - (y/x)x_{h1}] - [(y_{h2} - (y/x)x_{h2}]^2\}$,

would have a relative model bias of order $O_p(1/\sqrt{n})$, since $1/x^2 = (1/X^2)\{1 - 2[x - X]/X + O_p(1/n)\} = (1/X^2) + O_p(1/\sqrt{n})$.

The four variance estimators discussed in this note are all better estimators of the model variance of t than the "textbook" linearization variance estimator. Intuitively, this is because the four estimators contain realized sample values in their "denominators" and consequently estimate the variance of t conditioned on the realized x_{hj} better than the textbook estimator can.

7. Two Extensions

Under the sampling design we have been discussing, an estimated regression coefficient vector has the form $\mathbf{t} = (\sum^H \sum^2 \mathbf{z}_{hj}{}'\mathbf{z}_{hj}) \sum \sum \mathbf{z}_{hj}{}'\mathbf{q}_{hj}$, where $\mathbf{z}_{hj}$ is a row vector with K members. Let $\mathbf{y}_{hj} = \mathbf{z}_{hj}{}'\mathbf{q}_{hj}$, and $\mathbf{X}_{hj} = \mathbf{z}_{hj}{}'\mathbf{z}_{hj}$, which is symmetric. The estimator $\mathbf{t}$ can be rendered as $t = [n^{-1}\sum(\mathbf{X}_{h1} + \mathbf{X}_{h2})]^{-1}[n^{-1}\sum(\mathbf{y}_{h1} + \mathbf{y}_{h2})] = \mathbf{X}^{-1}\mathbf{y}$, which is simply equation (1) in matrix form.

The model in equation (2) generalizes to $\mathbf{y}_{hj} = \mathbf{X}_{hj}\mathbf{b} + \mathbf{e}_{hj}$, where the $\mathbf{e}_{hj}$ are uncorrelated with each other, $E(\mathbf{e}_{hj}) = \mathbf{0}_K$ and $\mathrm{Var}(\mathbf{e}_{hj}) = E(\mathbf{e}_{hj}\mathbf{e}_{hj}{}') = \mathbf{V}_{hj}$ is positive definite. Note that $\mathrm{Var}(\mathbf{e}_{hj}) = \mathrm{Var}(\mathbf{y}_{hj}) = \mathbf{z}_{hj}{}'\mathrm{Var}(q_{hj})\mathbf{z}_{hj}$.

The variance of $\mathbf{t}$ is $\mathrm{Var}(\mathbf{t}) = n^{-2}\mathbf{X}^{-1}\sum^H(\mathbf{V}_{h1} + \mathbf{V}_{h2})\mathbf{X}^{-1}$. Its linearization variance estimator is $\mathrm{var}_L(\mathbf{t}) = n^{-2}\mathbf{X}^{-1}\sum\{[\mathbf{y}_{h1} - \mathbf{X}_{h1}\mathbf{X}^{-1}\mathbf{y}] - [\mathbf{y}_{h2} - \mathbf{X}_{h2}\mathbf{X}^{-1}\mathbf{y}]\}^2 \mathbf{X}^{-1}$, where $\mathbf{z}^2$ denotes $\mathbf{z}\mathbf{z}'$. After some manipulation, $\mathrm{var}_L(\mathbf{t}) = n^{-2}\mathbf{X}^{-1}\sum\{[\mathbf{e}_{h1} - \mathbf{e}_{h2}] - [\mathbf{X}_{h1} - \mathbf{X}_{h2}]\mathbf{X}^{-1}\mathbf{e}\}^2\mathbf{X}^{-1}$, which has the following model expectation:

$$E[\mathrm{var}_L(\mathbf{t})] = n^{-2}\mathbf{X}^{-1}\sum_{h=1}^{H}\{[\mathbf{V}_{h1} + \mathbf{V}_{h2}] - [\mathbf{X}_{h1} - \mathbf{X}_{h2}]\mathbf{X}^{-1}[\mathbf{V}_{h1} - \mathbf{V}_{h2}] - [\mathbf{V}_{h1} - \mathbf{V}_{h2}]\mathbf{X}^{-1}[\mathbf{X}_{h1} - \mathbf{X}_{h2}] + [\mathbf{X}_{h1} - \mathbf{X}_{h2}]\mathbf{X}^{-1}\mathbf{V}\mathbf{X}^{-1}[\mathbf{X}_{h1} - \mathbf{X}_{h2}]\}\mathbf{X}^{-1}.$$

This is $\mathrm{Var}(\mathbf{t}) + \mathbf{O}(1/n^2)$, assuming that all the members of each $\mathbf{V}_{hj}$ and $\mathbf{X}_{hj}$ are bounded as H grows arbitrarily large with K fixed. Thus, the relative model bias of each term of the linearization variance estimator is $O(1/n)$. The asymptotic bounds on the members $\mathbf{V}_{hj}$ and $\mathbf{X}_{hj}$ are stronger assumptions than are needed here, but they have the advantage of simplicity.

Similar extensions of the analysis in Sections 4 and 5 can be applied to the jackknife variance estimator:

$$\mathrm{var}_J(\mathbf{t}) = (1/2)\sum_{h=1}^{H}\sum_{j=1}^{2}[\mathbf{X}_{(hj)}^{-1}\mathbf{y}_{(hj)} - \mathbf{X}^{-1}\mathbf{y}]^2,$$

where $\mathbf{Q}_{(hj)} = (1/n)[\sum_{g\neq h}(\mathbf{Q}_{g1} + \mathbf{Q}_{g2}) + 2\mathbf{Q}_{hj'}]$, $j' = 2$ when $j = 1$, and $j' = 1$ when $j = 2$, and to the two balanced half sampling estimators: $\mathrm{var}_{H1}(\mathbf{t}) =$

$(1/R)\sum[\mathbf{X}_r^{-1}\mathbf{y}_r - \mathbf{X}^{-1}\mathbf{y}]^2$, and $\text{var}_{H2}(\mathbf{t}) = (1/R)\sum[\mathbf{X}_r^{-1}\mathbf{y}_r - \mathbf{X}_{r'}^{-1}\mathbf{y}_{r'}]^2$, where $\mathbf{Q}_r = (1/H)\sum_{hj\in r}\mathbf{Q}_{hj}$ and $\mathbf{Q}_{r'} = 2\mathbf{Q} - \mathbf{Q}_r$. All three revolve around expressions of the form: $\mathbf{X}_+^{-1}\mathbf{y}_+ - \mathbf{X}^{-1}\mathbf{y} = \mathbf{X}_+^{-1}\mathbf{e}_+ - X^{-1}\mathbf{e} = \mathbf{X}^{-1}[\mathbf{X}\mathbf{X}_+^{-1}\mathbf{e}_+ - \mathbf{e}] = \mathbf{X}^{-1}[(\mathbf{e}_+ - \mathbf{e}) + (\mathbf{X}_+ - \mathbf{X})\mathbf{X}_+^{-1}\mathbf{e}_+]$, where the subscript "+" denotes "(hj)","r", or "r" as needed.

It is a straightforward exercise, closely paralleling the arguments in Sections 4 and 5, to show that the difference between the expected values of the linearization and jackknife variance estimators is $\mathbf{O}(n^{-3})$. Moreover, the difference between the expected values of either of the two balanced half sample estimators and the true variance of $\mathbf{t}$ is $\mathbf{O}_P(n^{-2})$, assuming full balance.

We have seen that the four variance estimation techniques under investigation provide variance estimates for $\mathbf{t}$ with bias of order $\mathbf{O}(1/n^2)$ or $\mathbf{O}_P(1/n^2)$. By contrast, the the standard error of all these variance estimators is dominated by the standard error of $n^{-2}\mathbf{X}^{-1}\sum^H(\mathbf{e}_{h1} - \mathbf{e}_{h2})^2\mathbf{X}^{-1}$, which is $\mathbf{O}_P(n^{-3/2})$. Thus, asymptotically at least, the standard error of these variances estimators play a most important role in inference than their biases. Kott (1994) proposes a method of estimating the effective degrees of freedom for the linearization variance estimator. That method, based on computing the effective degrees of freedom for $n^{-2}\mathbf{X}^{-1}\sum^H(\mathbf{e}_{h1} - \mathbf{e}_{h2})^2\mathbf{X}^{-1}$, applies equally well to the other three variance estimators.

Consider now an estimator of the form $t_G = \mathbf{pt}$, where $\mathbf{p}$ is a K member row vector. If the members of this vector are bounded from above and below by positive numbers as H grows arbitrarily large, then it is not hard to see that the four variance estimation techniques would provide variance estimates that are $O(1/n^2)$.

Recall that $\mathbf{t} = (\sum^H\sum^2\mathbf{z}_{hj}{}'\mathbf{z}_{hj})\sum\sum\mathbf{z}'_{hj}\mathbf{q}_{hj}$. If $\mathbf{p} = \sum\mathbf{z}_{hj}/\mathbf{M}$, where the summation is over all PSU's in the population and $\mathbf{M}$ is the number of elements in the population, then t_G would be a regression estimator for the population mean of q-values expressed in projection form. It should be noted that the variance of t_G is defined here with respect to the modified model parameter $\mathbf{pb}$ rather than the finite population total of q-values. Hence, one could argue that the linearization, jackknife, and half sample variance estimators in our discussion are natural estimators of model variance.

In a recent article, Valliant (1996) demonstrates, in seeming contradiction to the results presented here, that balanced half sampling can perform poorly from a model-based viewpoint. Valliant's model, however, has stratum effects, which our original model for the combined ratio estimator (equation (2)) lacks. It is of some interest to note that a separate ratio estimator, which implicitly assumes a linear relationship between y_{hj} and x_{hj} that may differ across strata, can be put in the form $t_G = \mathbf{pt}$, where the $K = H$ members of $\mathbf{p}$ are stratum population

totals for the x-value. Unfortunately, the asymptotic results of this section do not apply. They require that K be fixed as H grows arbitrarily large, which can not happen in this context.

References

Brewer, K. R. W. (1963). Ratio estimation and finite population: some results deducible from the assumption of an underlying stochastic process. *Austral. J. Statist.* **5**, 93-105.

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

Hansen, M. H., Madow, W. G. and Tepping, B. J. (1983). An evaluation of model-dependent and prpbability-sampling inference in sample surveys (with discussion). *J. Amer. Statist. Assoc.* **78**, 776-807.

Kott, P. S. (1994). A hypothesis test of linear regression coefficients with survey data. *Survey Methodology* **20**, 159-164.

Rao, J. N. K. and Wu C. F. J. (1985). Inferences from stratified samples: second-order analysis of three methods for nonlinear statistics. *J. Amer. Statist. Assoc.* **80**, 620-630.

Royall, R. M. (1970). On finite population sampling theory under certain linear regression models. *Biometrika* **57**, 377-387.

Royall, R. M. (1976). Current advances in sampling theory: implications for human observational studies. *Amer. J. Epidemiology* **104**, 463-474.

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

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

Rust, Keith (1985). Variance estimation for complex estimators in sample surveys. *J. Official Statist.* **1**, 381-397.

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

Valliant, Richard (1996). Limitations of balanced half sampling. *J. Official Statist.* **12**, 225-240.

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

Chief Research Statistician, National Agricultural Statistics Service, US Department of Agriculture, 3251 Old Lee Highway, Room 305, Fairfax VA 22030-1504, U.S.A.

E-mail: pkott@nass.usda.gov

(Received December 1996; accepted December 1997)