Journal of Official Statistics
Vol. 6. No. 4, 1990, pp. 451–454

Miscellanea

Under the heading Miscellanea, essays will be published dealing with topics considered to be of general interest to the readers. All contributions will be refereed for their compatibility with this criterion.

Recently Proposed Variance Estimators for the Simple Regression Estimator

Phillip S. Kott[1]

Abstract: This paper contrasts two model-based variance estimators for the simple regression estimator recently proposed in the literature. Both may be good estimators of the model variance as well as design consistent estimators of the design mean squared error. A third variance estimator that combines the strengths of the original two is also discussed.

Key words: Design consistent; Design mean squared error; Nearly model unbiased.

1. Introduction

Recently, Kott (1990) and Särndal, Swensson, and Wretman (1989) have proposed different model-based variance estimators for a design consistent regression estimator. Both are design consistent estimators of design mean squared error under reasonable conditions. The estimator in Kott, v_K, is a model unbiased estimator of model variance when the unit variances are correctly specified up to a scaling factor. By contrast, the estimator in Särndal, Swensson, and Wretman, v_{SSW}, is not exactly a model unbiased estimator of model variance in most cases. Nevertheless, v_{SSW} can be more robust to the misspecification of unit variances than v_K, a point not noted by the original authors.

We will examine these issues for a simple regression estimator under simple random sampling. A third variance estimator will be suggested that combines the strengths of the other two.

2. The Simple Regression Estimator

Suppose we want to estimate the population mean, $\bar{y}_N$, for a population of N units based on a simple random sample, S, of n y_i values. Suppose further that we know x_i values for all the units in the population and believe that the population y_i values are fitted by the stochastic equation

$$y_i = \alpha + \beta x_i + v_i^{1/2} \varepsilon_i \tag{1}$$

where the ε_i are independent random vari-

Acknowledgement: The author would like to thank an anonymous referee for his/her helpful comments.

[1] Senior Mathematical Statistician, Survey Research Branch, National Agricultural Statistics Service, U.S. Department of Agriculture, Washington, DC 20250, U.S.A.

ables with mean zero and variance unity. It is often assumed that the v_i are known up to a scaling factor. It is more reasonable, however, to suppose that they are positive but otherwise unknown.

The simple regression estimator for $\bar{y}_N$ is

$$\hat{y}_R = \bar{y}_S + (\bar{x}_N - \bar{x}_S)\hat{\beta}$$

where $\bar{z}_S$ is the sample mean of z_i values ($z = y$ or x) and $\hat{\beta} = \Sigma_S(x_i - \bar{x}_S)y_i/\Sigma_S(x_i - \bar{x}_S)^2$. It is well known that $\hat{y}_R$ is a model unbiased estimator of $\bar{y}_N$ in the sense that $E_\varepsilon(\hat{y}_R - \bar{y}_N) = 0$. Under reasonable conditions, $\hat{y}_R$ is also design consistent; i.e., $\hat{y}_R - \bar{y}_N$ tends toward zero, almost surely, as n grows arbitrarily large irrespective of the accuracy of the model. The estimator is not, in general, design unbiased.

3. Variance Estimation

One well known estimator for the design mean squared error estimator for $\hat{y}_R$ is

$$v_D = (1 - f)[n(n - 1)]^{-1} \sum_{i \in S} e_i^2 \qquad (2)$$

where $f = n/N$, and $e_i = y_i - \bar{y}_S - \hat{\beta}(x_i - \bar{x}_S)$. This is simply the traditional variance estimator for the sample mean, $\bar{y}_S$, with the y_i replaced by e_i. Note that there is no need to subtract $n\bar{e}_S^2$ from $\Sigma_S e_i^2$ in (2) because $\bar{e}_S = 0$.

Kott (1990) proposed the following estimator for the model variance of $\hat{y}_R$ as an estimator of $\bar{y}_N$

$$v_K = Av_D$$

where $A = E_\varepsilon\{(\hat{y}_R - \bar{y}_N)^2\}/E_\varepsilon(v_D)$. This model variance estimator requires an assumption about the relative sizes of the v_i in (1). Alternatively, the v_i can be estimated from the sample. Whatever the choices for the v_i, v_K is a design consistent estimator of the design mean squared error of $\hat{y}_R$ under reasonable conditions because A is asymptotically unity. It is a trivial matter to show that v_K is an exactly model unbiased estimator of the model variance of $\hat{y}_R$ when the v_i are specified correctly up to a scaling factor.

An alternative model variance estimator was proposed by Särndal, Swensson, and Wretman (1989)

$$v_{SSW} = (1 - f)\{n(n - 1)\}^{-1} \sum_{i \in S} (g_i e_i)^2$$

where $g_i = 1 + n(\bar{x}_N - \bar{x}_S)(x_i - \bar{x}_S)/\Sigma_S(x_j - \bar{x}_S)^2$. For ease of exposition, let $w_i = (x_i - \bar{x}_S)/\Sigma_S(x_j - \bar{x}_S)^2$ from now on. It is reasonable to assume that the population has the following asymptotic structure. As n grows arbitrarily large, the x_i are bounded, the w_i are $O_p(n^{-1})$, and $(\bar{x}_N - \bar{x}_S)$ is $O_p(n^{-1/2})$, where the subscript p refers to the probability space generated by the sampling design. As a result, v_{SSW} is a design consistent estimator of the design mean squared error of $\hat{y}_R$ whenever v_D is.

It is not as easy to see that v_{SSW} is a nearly model unbiased estimator of model variance of $\hat{y}_R$. The model variance of $\hat{y}_R$ is

$$E_\varepsilon\{(\hat{y}_R - \bar{y}_N)^2\} = (1 - f)n^{-1}\bar{v}_S + N^{-1}(\bar{v}_N - \bar{v}_S) + 2(1 - f)n^{-1}(\bar{x}_N - \bar{x}_S) \sum_{i \in S} w_i v_i + (\bar{x}_N - \bar{x}_S)^2 \sum_{i \in S} w_i^2 v_i. \qquad (3)$$

Making extensive use of the fact that the $E_\varepsilon(e_i^2) = v_i + O_p(n^{-1})$, the model expectation of v_{SSW} can be seen to be

$$E_\varepsilon(v_{SSW}) = (1 - f)n^{-1}\bar{v}_S + 2(1 - f)n^{-1}(\bar{x}_N - \bar{x}_S) \times \sum_{i \in S} w_i v_i + O_p(n^{-2}).$$

Thus the model bias of v_{SSW} is of probability order n^{-2} under reasonable conditions when

N is large relative to n; formally, when N^{-1} is $O(n^{-3/2})$.

Observe that as an estimator of model variance, v_D has a model bias of probability order $n^{-3/2}$. The same holds true for v_K when the v_i are misspecified. To see this, suppose it is wrongly assumed that the v_i are all equal. The expression A would then be equal to $1 + O_p(n^{-1})$. It would fail to capture the $2(1-f)n^{-1}(\bar{x}_N - \bar{x}_S)\Sigma_S w_i v_i$ term in equation (3), which is nonzero when $\bar{x}_N \neq \bar{x}_S$ and v_i is an increasing function of x_i.

We have just seen that v_{SSW} can be said to be a nearly unbiased model variance estimator in situations where v_K, when based on preconceived v_i, is not. On the other hand, v_{SSW} is not exactly model unbiased for any particular set of v_i. Moreover, v_K can be rendered nearly model unbiased by estimating the v_i from the sample; see Kott (1990).

4. The Royall and Cumberland Approach

Royall and Cumberland (1978) were concerned only with model-based properties when they developed two variance estimators for the simple regression estimator. The simpler one computationally is (G_2 on p. 357)

$$v_{RC} = (1-f)^2\{n(n-1)\}^{-1} \sum_{i \in S} [(g_i e_i)^2 + f e_i^2/(1-f)]/(1-d_i)$$

where $d_i = 1 - n(x_i - \bar{x}_S)/[(n-1)\Sigma_S(x_j - \bar{x}_S)^2]$.

It is not difficult to see that v_{RC} and v_{SSW} are nearly equal (i.e., their difference is $O(n^{-2})$) when N^{-1} is $O(n^{-2})$. On the other hand, v_{RC} is not even a design consistent estimator of the design mean squared error of $\hat{y}_R$ when N^{-1} is $O(n^{-1})$. These properties are shared by Royall and Cumberland's other variance estimator (G_1 on p. 357).

This paper has been concerned with estimators of model variance that are model unbiased or nearly model unbiased when the model in (1) holds and N^{-1} is $O(n^{-3/2})$. Royall and Cumberland's estimators are exactly unbiased when the model holds and the v_i are equal. When the model holds the v_i are not all equal, however, their estimators are only nearly model unbiased when N^{-1} is $O(n^{-2})$.

5. A New Variance Estimator

There is no reason why the principal ideas behind v_K and v_{SSW} cannot be combined. The result would be

$$v_C = A^* v_{SSW} \tag{4}$$

where $A^* = E_\varepsilon\{(\hat{y}_R - \bar{y}_N)^2\}/E_\varepsilon(v_{SSW})$. This estimator is a design consistent estimator of the design mean squared error of $\hat{y}_R$ under reasonable conditions. It is a nearly model unbiased estimator of model variance when N is large relative to n, and is exactly unbiased when the specification of the v_i inherent in A^* is correct up to a scaling factor.

6. Discussion

Kott (1990) investigated the model variance of the ratio estimator under simple random sampling in some detail. He showed that when N is large relative to n, his adjusted Yates–Grundy design mean squared error estimator is a nearly model unbiased estimator of the model variance no matter what the specification of the v_i. Although this property can also be shown to hold for ratio estimators under more complex sampling designs, it does not hold for design consistent regression estimators in general, as we have seen.

There are two ways to modify a traditional design mean squared error estimator into a nearly model unbiased estimator of model variance when the unit variances are unknown. The method proposed by Kott (1990) requires estimating the v_i from the sample. The method proposed by Särndal, Swensson, and Wretman (1989) requires that N be large relative to n, but is much simpler to implement. (Note: the Royall and Cumberland variance estimation strategies do not directly involve the modification of design mean squared error estimators.)

As noted, the two methods can be combined, as they are in equation (4). Since replacing real v_i values with estimated ones usually causes a small bias, it makes sense to estimate the v_i from the sample only when N is *not* large relative to n.

For most practical applications, the nearly model unbiased v_{SSW} is good enough. The change resulting from the additional, often complicated, adjustment in (4) would be trivial. Still, there is something vaguely disturbing about the practice of letting the e_i^2 serve as estimates of the v_i without adjusting for their tendency to be biased downward. A tempting modification in the case of the simple regression estimator would be to multiply v_{SSW} by $(n - 1)/(n - 2)$. This at least renders it exactly model unbiased in the very special case when all the v_i are equal and $\bar{x}_S = \bar{x}_N$.

7. References

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, 73, 351–358.

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

Received December 1989
Revised July 1990