Outlier Robustness of the Estimators of Variance of the Ratio Estimator

R. O. Otieno[1] and C. Wafula[2]

[1]Mathematics Department, Jomo Kenyatta University of Agriculture and Technology, P.O. Box 62000, Nairobi, KENYA; [2]Mathematics Department, Kenyatta University, P.O. Box 43844, Nairobi, KENYA

ABSTRACT

This paper investigates the sensitivities of the variance estimators for the ratio estimator. The model-based estimators V_D and V_L are found to be more sensitive to outliers than the rest of the estimators while the jack-knife variance estimator is the best on this criterion.

KEYWORDS: Ratio, estimator, variance, outliers, robustness

1.0 INTRODUCTION

The problem of the estimation of the variance of the ratio estimator has recently received a lot of attention. Within the past three decades a number of variance estimators for the ratio estimator have been suggested in the literature. This has left a practitioner in some sort of dilemma as to which estimator to use in practice. To help a practitioner make a choice, a number of comparative studies have been carried out. The criterion that has been used in these studies is that of bias-robustness (Royall and Cumberland 1978). On this criterion the jack-knife variance estimator and the bias robust variance estimator, v_D of Royall and Cumberland (1978) have emerged winners. But an estimator of choice between these two estimators has not been resolved in the literature.

Bias-robustness is not the only criterion to use to compare estimators. Another criterion, which has rarely been used in the literature, is the sensitivities of the estimators to outliers (Hampel 1974). We use this criterion in this paper. On this criterion, the jack-knife variance estimator is more robust than the bias-robust variance estimator V_D.

2.0 MEASURE OF INFLUENCE

There are two main ways of assessing the sensitivity of an estimator to outlying values. One is based on some form of theoretical influence function and the other on case deletion.

The concept of an influence function is due to Hampel (1974). It measures the rate of change of an estimator with variation in the specification of the data. Let F be the distribution from which the data are generated. Then if T(F) is some form of functional of interest the rate of change of T(F) to some perturbation in F say G= (1- ε) F +εd x at a point x, is given by the function:

$$IF(F) = \lim_{\varepsilon \to 0} \left(\frac{T(G) - T(F)}{\varepsilon} \right)$$

The case deletion method of assessing sensitivity is as follows. Let $\hat{\theta}$ be an estimator, calculated from the sample (y_1, y_2,............,y_n), of some parameter θ. Let $\hat{\theta}_{(j)}$ be the corresponding estimator calculated with the j-th case of the sample excluded. Then two possible measures of influence of the j-th case of the estimator θ are

$$\text{SIF}(\hat{\theta}) = \hat{\theta}_{(j)} - \theta$$

$$I_s(\hat{\theta}) = \frac{SIF(\hat{\theta}) * 100}{\hat{\theta}}$$

where term SIF ($\hat{\theta}$) the sample influence function.

Deriving theoretical influence functions of the variance estimators can be a formidable task (Hampel 1974).

Because of this we shall be content with obtaining the sample influence functions in this paper.

3.0 THE RATIO ESTIMATOR AND ITS VARIANCE ESTIMATORS

A population consisting of N identifiable units with values (y_i, x_i), where x_i>0 (i = 1,2,....,N) was considered. Denote the population means of y and x by $\overline{Y}$ and $\overline{X}$ respectively.

To estimate $\overline{Y}$, it is customary to take a simple random sample of size n and to use the ratio estimator

$$Y_R = \overline{y}\ \overline{X} / \overline{x},$$

Where $\bar{y}$ and $\bar{x}$ are, respectively, the sample means of y and x. The following variance estimators for the ratio estimator have been suggested in the litterature (Royall and Cumberland 1978).

$$V_O = \left(\frac{1-f}{n}\right)\frac{\sum_1^n \hat{e}_i^{\,2}}{n-1}$$

$$V_2 = \left(\frac{\bar{X}}{\bar{x}}\right)^2 V_O$$

$$V_D = \frac{1-f}{n}\,\frac{\bar{x}_r\,\bar{X}}{\bar{x}^2}\,\frac{1}{n}\sum_1^n \hat{e}_i^{\,2}/(1-k_i)$$

$$V_L = \frac{1-f}{n}\,\frac{\bar{x}_r\bar{X}}{\bar{x}^2}\,\frac{1}{n-1}\sum_1^n \hat{e}_i^{\,2}/x_i \qquad \text{.......................... 1}$$

$$V_J = (1-f)\,\bar{X}^2\,\frac{n-1}{n}\sum_1^n D^2(j)$$

where: $\hat{e}_i = y_i - rx_i$, $r = \frac{\bar{y}}{\bar{x}}$, $k_i = x_i/n\bar{x}$,

$\bar{x}_r$ represents the mean of x's in the non-sampled units and $D_{(j)}$ is the difference between the ratio $(n\bar{y} - y_j)/(n\bar{x} - x_j)$ and the average of these n ratios.

4.0 SAMPLE INFLUENCE FUNCTIONS

The sample influence functions of V_0, V_2, V_D, V_j, and V_L was obtained. Since the algebra involved is straight forward, it is not included. Simplification gives the results as follows (Odhiambo 1991, Wafula 1988).

$$SIF(V_o) = \left(\frac{1-f}{n^2}\right)\left\{\frac{\hat{e}_i^2}{(1-k_i)^2}\left(\sum_1^n k_i^2 - 1\right) + \frac{2\hat{e}_i}{1-k_i}\sum_1^n k_i \hat{e}_i\right\} \quad \ldots\ldots 2$$

$$SIF(V_2) = \left(\frac{\bar{X}}{\bar{x}(1-k_i)}\right)^2 \left\{\frac{1-f}{n^2}\sum_1^n \hat{e}_i^2 + SIF(V_o)\right\} \quad \ldots\ldots 3$$

$$SIF(V_D) = \left(\frac{I-f}{n\bar{x}(1-k_i)^2}\right)\left\{\bar{x}_r + \frac{x_i}{N-n}\right\}$$

$$\left[\sum_{1=i}^n \hat{e}_j^2 - \frac{\hat{e}_i^2}{(1-k_i)^2} + \frac{2\hat{e}_i}{n\bar{x}(1-k_i)}\sum_1^n \hat{e}_j x_j^2 + \frac{\hat{e}_{i^2}\sum x_j^2}{n\bar{x}(1-k_i)^2}\right] \quad \ldots\ldots 4$$

$$SIF(V_J) = \frac{1-f}{n^2}\left(\frac{\bar{X}}{\bar{x}}\right)^2 \left\{\sum_1^n (\hat{e}_j/(1-k_j))^2 \frac{(k_j+k_i)(2-k_i-k_j)}{(1-k_i-k_j)^2} - \right.$$

$$\left. \frac{1}{n}\sum_{j=1}^n \left\{\frac{\hat{e}_j(2-k_i-k_j)}{(1-k_j)(1-k_i-k_j)}\sum_1^n \frac{\hat{e}_j(k_j+k_i)}{(1-k_j)(1-k_i-k_j)} - \frac{e_j^2}{(1-2k_i)^2(1-k_i)^2}\right\}\right\} \quad \ldots\ldots 5$$

$$SIF(V_L) = \left(\frac{1-f}{N-n}\right)\frac{\bar{X}}{n^3\bar{x}^{-2}(1-k_i)}\left[N\bar{X}x_i\sum \hat{e}_j^2/x_j - \left(\frac{N\bar{X}-n\bar{x}+x_i}{k_i(1-k_i)}\right)\hat{e}_i^2\right] \quad \ldots\ldots 6$$

One obvious observation from equations 2 - 6 is that the influence of a sample point on the variance estimators depends on two main factors:

(i) The residual of the point and; (ii) The leverage of the point.

From equation 2 we note that a point with a large residual will have a large influence on Vo. The first term in the curly brackets in equation 2 is negative and for a large residual this term is larger, in magnitude than the second. Hence in this case the change in

Vo will be negative. The change will be much larger if the residual for the i-th point is negative and

$$\Sigma x_j e_j > 0 .$$

Noting that $(1-k_i)^{-1}>1$ it follows that a high leverage point will inflate both terms in the curly brackets of equation 2 but this time the second term could as well be larger than the first. Hence in this case the change in Vo can be negative or positive. A point which is both an outlier and high leverage point will inflate both terms of equation 2. The second term will be smaller than the first. Hence the change in Vo will be negative. [For detailed proofs see Odhiambo (1991), Wafula (1988)].

From equation 3, if SIF (Vo) > 0 then SIF(V_2) >0 i.e a positive change in Vo will imply a positive change in V_2 . Further, if SIF (Vo) > 0 and $\overline{X} > \overline{x}\,(1-k_i)^2$ then V_2 will be more sensitive than Vo . This result is confirmed in our empirical study in the next section.

If f is negligible, and the sample is balanced then SIF(V_2)>SIF(V_J). It is also clear that both SIF (V_D) and SIF (V_L) are directly proportional to X_is and are more influenced by the leverage points than the rest of the estimators. The empirical results are in 4 sets of data.

EMPIRICAL STUDY

Results on the sensitivities of the above variance estimators to influential points in four populations are given in Table 1.

Table 1. Study populations

Population	Source	X	Y
1.	Cochran (1977) p. 152	Size of city in U.S in 1920	Size of city in U.S in 1930
2.	Olkin (1958)	Size of city in U.S in 1940	Size of city in U.S in 1950
3.	Olkin (1958)	Size of city in U.S in 1930	Size of city in U.S in 1940
4.	Ministry of Finance and Economic Planning (Kenya) and Earnings (1971)	Number of people employed in town in 1963	Number of people in town in 1966.

When a simple regression model is fitted in these populations the following points are flagged as unusual: 5, 10, 18, 26 and 35 in population 1; 1, 4, 12, 23, 33 and 38 in population 2; 1, 4, 5, 12, 23, 33 and 38 in population 3 and 1, 2 in population 4. Some characteristics of these points are given in Table 2. The characteristics include their standardised residuals, their leverages, and whether the influence is due to its residual, x value or both.

Table 2. Some characteristics of the influential points

Population	Point	Standardised residual	K_i	Influence due to*
1	5	0.25	0.09	X
	10	0.29	0.09	X
	18	1.14	0.12	X
	26	2.29	0.01	R
	35	2.99	0.01	R
2	1	0.40	0.096	X
	4	3.93	0.055	R
	12	0.14	0.125	X
	23	2.46	0.029	R
	33	4.34	0.11	RX
	38	0.6	0.084	X
3	1	-0.20	0.098	X
	4	4.73	0.043	R
	5	2.16	0.019	R
	12	-1.79	0.132	X
	23	2.78	0.022	R
	33	-0.96	0.115	X
	38	0.23	0.085	X
4	1	4.60	0.527	RX
	2	-5.03	0.207	RX

X indicates influence due to X value; Y indicates influence due to residual; RX indicates influence due to both x value and residual.

The samples obtained in these populations as follows. In population 1, 2 and 3 samples of size 40 were used which were obtained by dropping the last nine points of population 1 and the last ten points of populations 2 and 3. In population 4, a sample of size 30 obtained by dropping the last 4 points of the population was used. When the simple regression model was fitted in the four samples the same points as those for the populations were flagged as being influential.

In each sample we calculation of the variance estimates was done using the complete data and when each of the influential points is removed, followed by calculation of the sensitivities using I_s (.). The results are given in Table 3.

Table 3. Sensitivities of the estimators

POPULATION 1

SENSITIVITIES

POINT ESTIMATOR	5	10	18	26	35
V_O	17.2	16.3	16.3	-11.7	-12.3
V_2	41.4	40.7	49.9	-0.09	-11.0
V_D	79.5	79.4	110.8	-16.9	-20.8
V_J	34.8	33.8	43.0	-13.8	-15.1
V_L	67.0	68.0	91.1	71.1	-5.3

POPULATION 2

POINT ESTIMATOR	1	4	12	23	33	38
V_O	15.0	-26.5	15.8	-3.3	-43.0	16.0
V_2	40.7	-17.8	51.3	2.6	-28.1	38.1
V_D	80.3	-8.9	113.3	1.8	-6.1	69.3
V_J	36.7	-20.0	47.7	-1.3	-33.8	33.4
V_L	63.2	-5.6	87.1	-9.0	20.2	57.7

POPULATION 3

POINT ESTIMATOR	1	4	5	12	23	33	38
V_O	15.7	-47.8	-0.29	1.5	-9.1	11.1	16.5
V_2	42.4	-43.1	3.6	34.8	-5.0	41.7	39.1
V_D	83.3	-40.7	-2.2	90.6	-9.2	92.0	62.9
V_J	37.6	-46.1	1.5	27.2	-9.5	36.5	37.6
V_L	66.8	-17.5	-5.2	85.3	-13.3	76.1	58.1

POPULATION 4

POINT ESTIMATOR	1	2
V_O	-47.5	-83.1
V_2	134.3	-73.2
V_D	35653.8	1166.0
V_J	137.7	-84.0
V_L	37047.1	5694.1

The results are summarised as follows:

1. For points that are influential due to their leverage, Vo is the least sensitive while V_D is the most sensitive. The theoretical comparison between Vo and V_2 that were made in the last section hold well for these points. It was noted that if $\overline{X} \geq \overline{x}(1-k_i)^2$ and SIF (Vo) ≥ 0 then V_2 is more sensitive to the high leverage point than Vo is. Indeed this is the case for all the high leverage points in our empirical study.
2. No single estimator is a clear winner when a point is influential due to its large residual. The same is true for points that are both outliers and high leverage points. However, in this case the poor performances of V_D and V_L in population 4 are evident.
3. On average the randomisation estimator V_0, V_2 and the Gechurufe variance estimator and V_j were more robust to all types of outlying points than the model based estimators V_D and V_L in our empirical study.

3.0 Conclusion

From bias-robustness point of view, previous comparative studies of the variance estimators of the ratio estimator have favoured the estimators V_J and V_D (Odhiambo 1991, Wafula 1988). These studies have also shown that V_L is non robust and hence recommended that V_L should be used in practice with care.

On the other hand our limited empirical study points to a tentative conclusion that the model based estimators V_D and V_L may not be robust in the sense that they are sensitive to certain types of influential points. On the whole, V_j, was more robust than V_D. However, no firm conclusion can be drawn from a single empirical study and so more empirical studies are needed especially careful theoretical study of the influence functions of these variance estimators.

References

Cochran W.G. (1977) Sampling Techniques (3rd Edition). Wiley New York.

Hampel F.R. (1974) The influence curve and its role in robust estimators. *JASA* **69**, 383-393.

Olkin I. (1958) Multivariate ratio estimation for finite populations. *Biometrika* **45**, 145-165.

Odhiambo R.O (1991) Unpublished MSc Dissertation, Kenyatta University.

Royall R.M. and W. G. Cumberland (1978) Variance Estimation in Finite population sampling. *JASA* **73**, 351-358.

Wafula C. (1988) Some contributions to variance estimation in Sample Surveys. Unpublished Ph.D thesis, University of Kent at Cantebury, U.K.