韓國數學敎育學會誌<數學敎育>
1984. 3. Vol. XXII, No. 2.

Journal of the Korea Society of Mathematical Education
Mar. 1984. Vol. XXII, No. 2.

Conditions under which the Ratio Estimator is a Best Linear Unbiased Estimator

by Jee Euen Sook
Kwangwoon University, Seoul, Korea

A well-known result in regression theory indicates the type of population under which the ratio estimate may be called the best among a wide class of estimates. The result was first proved for infinite populations. Brewer and Royall extended the result to finite populations. The result holds if two conditions are satisfied.

1. The relation between y_i and x_i is a straight line through the origin.
2. The variance of y_i about this line is proportional to x_i.

A "*best linear unbiased estimator*" is defined as follows. Consider all estimators $\hat{Y}$ of Y that are linear functions of the sample values y_i, that is, that are of the form

$$l_1y_1+l_2y_2+\cdots\cdots+l_ny_n$$

where the *l's* do not depend on the y_i, although they may be functions of the x_i. The choice of *l's* is restricted to those that give unbiased estimation of Y. The estimator with the smallest variance is called the *best linear unbiased estimator* (*BLUE*).

Formally, Brewer and Royall assume that the N population values (y_i, x_i) are a random sample from a superpopulation in which

$$y_i=\beta x_i+\varepsilon i \tag{1}$$

where the εi are independent of the x_i and $x_i>0$. In arrays in which x_i is fixed, εi has mean 0 and variance λx_i. The x_i $(i=1, 2, \ldots\ldots, N)$ are known.

In the randomization theory, the finite population total Y has been regarded as a fixed quantity. Under model (1), on the other hand, $Y=\beta X+\sum^{N}\varepsilon i$ is a random variable. In defining an unbiased estimator under this model, Brewer and Royall use a concept of unbiasedness which differs from that in randomization theory. They regard an estimator $\hat{Y}$ as unbiased if $E(\hat{Y})=E(Y)$ in repeated selections of the finite population and sample under the model. Such an estimator might be called *model-unbiased*.

Theorem. *Under model* (1) *the ratio estimator* $\hat{Y}_R=X\bar{y}/\bar{x}$ *is a best linear unbiased estimator for any sample, random or not, selected solely according to the values of the* x_i.

Proof. Since $E(\varepsilon i/x_i)=0$ in repeated sampling, it follows from (1) that

$$Y=\beta X+\sum^{N}\varepsilon i:\ E(Y)=\beta X \tag{2}$$

Furthermore, with the model (1) any linear estimator $\hat{Y}$ is of the form

$$\hat{Y}=\sum^{n}l_iy_i=\beta\sum^{n}l_ix_i+\sum^{n}l_i\varepsilon_i \tag{3}$$

If we keep the n sample values x_i fixed in repeated sampling under the model (1),

$$E(\hat{Y})=\beta\sum^{n} l_i x_i : V(\hat{Y})=\lambda\sum^{n} l_i^2 x_i \tag{4}$$

From (2) and (3), $\hat{Y}$ is clearly model-unbiased if $\sum^{n} l_i x_i \equiv X$. Minimizing $V(\hat{Y})$ under this condition by a Lagrange multiplier gives

$$2l_i x_i = c x_i : l_i = \text{constant} = X/n\bar{x} \tag{5}$$

The constant must have the value $X/n\bar{x}$ in order to satisfy the model-unbiased condition $l\sum^{n} x_i = X$. Hence the *BLUE estimator* $\hat{Y}$ is $n\bar{y}X/n\bar{x}=X\bar{y}/\bar{x}=\hat{Y}_R$, the usual ratio estimator. This completes the proof.

Furthermore, from (2) and (3), with $l=X/n\bar{x}$,

$$\hat{Y}_R - Y = \sum^{n} l_i \varepsilon_i - \sum^{N} \varepsilon i = (X/n\bar{x})\left(\sum^{n} \varepsilon i\right) - \sum^{N} \varepsilon i \tag{6}$$

$$= \frac{(X-n\bar{x})}{n\bar{x}} \sum^{n} \varepsilon i - \sum^{N-n} \varepsilon i \tag{7}$$

where $\sum^{N-n}$ denotes the sum over the $(N-n)$ population values that are ***not*** in the sample. Hence

$$V(\hat{Y}_R) = \frac{\lambda(X-n\bar{x})^2(n\bar{x})}{(n\bar{x})^2} + \lambda(X-n\bar{x}) = \frac{\lambda(X-n\bar{x})X}{n\bar{x}} \tag{8}$$

A model-unbiased estimator of λ from this sample is easily shown to be

$$\lambda = \sum^{n} \frac{1}{x_i}(y_i - \hat{R}x_i)^2/(n-1) \tag{9}$$

where $\hat{R}=\bar{y}/\bar{x}$, as usual. This value may be substituted in (8) to give a model-unbiased sample estimate of $V(\hat{Y}_R)$.

The practical relevance of these results is that they suggest the conditions under which the ratio estimator is superior not only to $\bar{y}$ but is the best of a whole class of estimators. When we are trying to decide what kind of estimate to use, a graph in which the sample values of y_i are plotted against those of x_i is helpful. If this graph shows a straight line relation passing through the origin and if the variance of the points y_i about the line seems roughly proportional to x_i, the ratio estimator will be hard to beat.

Sometimes the variance of the y_i in arrays in which x_i is fixed is not proportional to x_i. If this residual variance is of the form $\lambda v(x_i)$, $v(x_i)$ is known, Brewer and Royall showed that the *BLUE estimator* becomes

$$\hat{Y} = X \frac{\sum^{n} w_i y_i x_i}{\sum^{n} w_i x_i^2} \tag{10}$$

where $w_i=1/v(x_i)$. In a population sample of Greece, Jessen et al. (1947) judged that the residual variance increased roughly as x_i^2. This suggests a weighted regression with $w_i=1/x_i^2$, which gives

$$\hat{Y} = \frac{X(\sum^{n} w_i y_i x_i)}{\sum^{n}(w_i x_i^2)} = \frac{X}{n}\sum^{n}\left(\frac{y_i}{x_i}\right) \tag{11}$$

For a given population and given n, $V(\hat{Y}_R)$ in (8) is clearly minimized, given every $x_i>0$, when the sample consists of the n largest x_i in the population. In [16] small natural populations of the type to which ratio estimates have been applied, Royall (1970) found for samples having $n=2$ to

12 that selection of the n largest x_i usually increased the accuracy of $\hat{Y}_R$.

In summary, the Brewer-Royall results show that the assumption of a certain type of model leads to an unbiased ratio estimator and formulas for $V(\hat{Y}_R)$ and practice in cases where examination of the y, x pairs from the available data suggests that the model is reasonably correct. The variance formulas (8) and (9) appear to be sensitive to inaccuracy in the model, although this issue needs further study.

Further work by Royall and Herson (1973) discusses the type of sample distribuiton needed with respect to the x_i in order that $\hat{Y}_R$ remains unbiased when there is a polynomial regression of y_i on x_i.

References

1. Brewer, K.W.R. (1963) Ratio estimation in finite populations: Some results deducible from the assumption of an underlying stochastic process. *Australian Jour. Stat.*, *5*, 93-105.
2. Cochran, W.G. (1977). *Sampling Techniques.* John Wiley & Sons, New York, third edition.
3. Jessen, R.J., et al. (1947). On a population sample for Greece, *Jour. Amer. Stat. Assoc.*, *42*, 357-384.
4. Royall, R.M. (1970). On finite population sampling theory under certain linear regression models. *Biometrika*, *57*, 377-387.