Jour. Ind. Soc. Ag. Statistics
Vol. XLII. No. 2 (1990), pp. 206-217

SAMPLING SCHEMES PROVIDING UNBIASED GENERAL REGRESSION ESTIMATORS

PADAM SINGH* and H. V. L. BATHLA**
(Received : March, 1988)

SUMMARY

In this paper a sampling scheme has been proposed for which usual regression estimator becomes unbiased. The proposed sampling scheme is simpler than one proposed by Singh and Srivastava (1980) and in addition has the advantage of extending it to multivariate regression estimator and polynomial regression estimator. From efficiency point of view also the proposed sampling strategies are better than the conventional strategies under use.
Key words : Efficiency; Regression estimator; Multivariate regression estimator; Polynomial regression estimator; Ratio estimator; Sampling Scheme; Sampling strategy; Unbiased estimation.

Introduction

Singh and Srivastava (1980) proposed a sampling scheme for which the usual regression estimator is unbiased. The procedure suggested by them involves calculation of conditional probabilities if sample is to be selected unit by unit. In this paper an alternative sampling scheme has been suggested. The proposed sampling scheme has advantage of extension to polynomial regression estimator and multivariate regression estimators. The efficiencies of the proposed sampling strategies with those in vogue have also been compared.

*Institute for Research in Medical Statistics ICMR), New Delhi.
**Indian Agricultural Statistics Research Institute, New Delhi.

where N is the incidence matrix of the design D.

This matrix Ω gives the following three different variances between the adjusted treatments means :

(a) $[(4a+1)/6a]\sigma^2$, for the treatments which are I associates;

(b) $[(4a+2)/6a]\sigma^2$, for treatments which are II associates;

(c) $[(4a+3)/6a]\sigma^2$, for the treatments which are third associates.

The efficiency factor of the design D is

$$\text{E.F.} = [2\,(2a^2-1)/(4a^2+3a-5)],$$

which is quite high for $a > 3$.

ACKNOWLEDGEMENT

It is a great pleasure to acknowledge the financial assistance of the University Grants Commission, New Delhi for carrying out this project. We are also grateful to the referee for his valuable comments.

REFERENCES

[1] Chaudhary, S. S. and Singh, S. P. (1986) : A note on the dual of a PBIB design, *Indian J. Pure Appl. Math.* **17** (2) : 138-141.

[2] Mohan, R. N. (1983) : On the dual of a PBIB design, *Indian J. Pure. Appl. Math.* **14** : 13-17.

[3] Ramakrishna, C. S. (1956) : On the dual of a PBIB designs and new class designs with two replications, *Sankhya* **17** : 133-142.

2. Proposed Sampling Schemes

2.1 *Sampling Scheme Providing Unbiased Regression Estimator*

Suppose that the population under consideration consists of N distinct and identifiable units and a sample of size n is desired to be drawn from it. Further, suppose y is the study variable and x is the auxiliary variable. We assume that the information on the auxiliary variable is available for all the units of population The sampling scheme providing unbiased regression estimator is as under :

(a) Select a sample of size n by simple random sampling and without replacement. Calculate s_x^2 (the sample mean square) for the sample so selected, where

$$s_x^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2, \ \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

(b) Select a random number j between 0 and M where $M >$ Max (s_x^2) It is important to mention here that maximum of s_x^2 will generally be known since $x_i's$ are all known to us.

(c) If j is less than or equal to s_x^2 retain the sample selected at step (a) otherwise proceed to step (a) again for selecting a sample and continue the process till the sample is ultimately selected.

It is easy to see that for above sampling scheme the probability of selecting a sample 's' is proportional to the variability in the sample for the auxiliary variable i.e.

$$P_s = \frac{s_x^2}{\binom{N}{n} S_x^2}$$

It is assumed here that no combination of n units from N units of the population have same value of the auxiliary variable so that all samples have non-zero probabilities of selection This sampling scheme is simpler than that by Singh and Srivastava (980) in the sense that it does not involve calculation of conditional probability of success at the second draw. However it involves number of rejections. But because only one sample is to be selected it may not be a limitation.

We, now, extend the proposed sampling scheme for making polynomial regression estimator as well as multivariate regression estimator unbiased in the following sections.

2.2 *Sampling Scheme Providing Unbiased Polynomial Regression Estimator*

We shall consider the sampling scheme for which quadratic regression estimator is unbiased. The extension for a general kth degree polynomial regression estimator is straight-forward. The proposed sampling scheme consists of the following steps.

(a) Select a sample of size n by simple random sampling without replacement. For the selected sample, calculate

$$\lambda_s = m_{20}\, m_{40} - m_{20}^3 - m_{30}^2$$

where

$$m_{pq} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^p (y_i - \bar{y})^q$$

(b) Select a random number j from 0 to M, where,

$$M \geqslant \text{Max}\,(\lambda_s)$$

Here also M will be known generally since x_i are all known to us. However, for the extreme cases M can safely be taken as,

$$\binom{N}{n} \left[\mu_{20}\, \mu_{40} - \mu_{20}^3 - \mu_{30}^2 \right].$$

(c) if j is less than or equal to λ_i retain the sample otherwise proceed to step (a) for selecting another sample and continue the process till a sample is ultimately selected.

For this sampling scheme the probability of selecting a sample is gived by

$$P_s = \frac{m_{20}\, m_{40} - m_{20}^3 - m_{30}^2}{\binom{N}{n} T}$$

where

$$T = \frac{1}{\binom{N}{n}} \sum_{s=1}^{\binom{N}{n}} (m_{20}\, m_{40} - m_{20}^3 - m_{30}^2)$$

The estimator to be used in this situation is

$t_s^* = \bar{y} - b_1\, \bar{x}' - b_2\, \bar{x}^{2\prime}$, where $\bar{x}^2$, $\bar{X}^2$ are the means of (x^2)

for sample and population. $\bar{x}' = \bar{x} - \bar{X}$, $\bar{x}^{2\prime} = \bar{x}^2 - \bar{X}^2$ and b_1 and b_2 are the regression coefficients of the quadratic regression. given by

$$b_1 = \frac{m_{40}\, m_{11} + 2\, m_{30}\, m_{11}\, \bar{x} - m_{20}^2\, m_{11} - m_{30}\, m_{21} - 2\bar{x}\, m_{20}\, m_{21}}{m_{20}\, m_{40} - m_{20}^3 - m_{30}^2}$$

and

$$b_2 = \frac{m_{20}\, m_{21} - m_{30}\, m_{11}}{m_{20}\, m_{40} - m_{20}^3 - m_{30}^2}$$

Now

$$t_s^* = \bar{y} - \frac{m_{40}\, m_{11} + 2\, m_{30}\, m_{11}\bar{x} - m_{20}^2\, m_{11} - m_{30}\, m_{21} - 2\bar{x}\, m_{20}\, m_{21}}{m_{20}\, m_{40} - m_{20}^3 - m_{30}^2} \cdot \bar{x}'$$
$$- \frac{m_{20}\, m_{21} - m_{30}\, m_{11}}{m_{20}\, m_{40} - m_{20}^3 - m_{30}^2} \cdot \bar{x}^{2\prime}$$

$$\text{and } E(t_s^*) = \frac{1}{\binom{N}{n} T} \underset{s\in S}{\xi} \, [\, t_s^*] \, (m_{20}\, m_{40} - m_{20}^3 - m_{30}^2)$$
$$= \frac{1}{T}\, \xi \Big[\, \bar{y}\, (m_{20}\, m_{40} - m_{20}^3 - m_{30}^2) - (m_{40}\, m_{11}$$
$$+ 2\, m_{30}\, m_{11}\, \bar{x} - m_{20}^2\, m_{11} - m_{30}\, m_{21} - 2\bar{x}\, m_{20}\, m_{21})\, \bar{x}'$$
$$- (m_{20}\, m_{21} - m_{30}\, m_{11})\, \bar{x}^{2\prime} \Big]$$

where ξ donates the average over $\binom{N}{n}$ samples of the sample space S.

$$= \frac{1}{T}\,\xi\Big[(\bar{Y}+\bar{y}')\Big\{\sum_{r=1}^{n}(x_r-\bar{x})^2\sum_{r=1}^{n}(x_r-\bar{x})^4$$

$$-\Big(\sum_{r=1}^{n}(x_r-\bar{x})^2\Big)^3-\Big(\sum_{r=1}^{n}(x_r-\bar{x})^3\Big)^2\Big\}$$

$$-\Big\{\sum_{r=1}^{n}(x_r-\bar{x})^4\sum_{r=1}^{n}(x_r-\bar{x})(y_r-\bar{y})$$

$$+2\sum_{r=1}^{n}(x_r-\bar{x})(\bar{y}_r-\bar{y})-\sum_{r=1}^{n}(x_r-\bar{x})^3$$

$$\sum_{r=1}^{n}(x_r-\bar{x})^2.(y_r-\bar{y})$$

$$-2(\bar{X}+\bar{x}')\sum_{r=1}^{n}(x_r-\bar{x})^2\sum_{r=1}^{n}(x_r-\bar{x})^2(y_r-\bar{y})\Big\}\bar{x}'$$

$$-\Big\{\sum_{r=1}^{n}(x_r-\bar{x})^2\sum_{r=1}^{n}(x_r-\bar{x})^2(y_r-\bar{y})$$

$$-\sum_{r=1}^{n}(x_r-\bar{x})^3.\sum_{r=1}^{n}(x_r-\bar{x})(y_r-y)\Big\}\bar{x}^{2\prime}\Big]$$

$$=\bar{Y}+\frac{1}{N(N-1)T}\Big[\sum_{i=1}^{N}Y_i'\Big\{\sum_{i=1}^{N}X_i'^2$$

$$\sum_{i=1}^{N}X_i'^4-\Big(\sum_{i=1}^{N}X_i'^2\Big)^3$$

$$-\Big(\sum_{i=1}^{n}X_i'^3\Big)^2\Big\}-\Big\{\sum_{i=1}^{N}X_i'^4\sum_{i=1}^{N}X_i'\,Y_i'+2\sum_{i=1}^{N}X_i'^3$$

$$\sum_{i=1}^{N}X_i'\,Y_i'(\bar{X}+\bar{x}')-\Big(\sum_{i=1}^{N}X_i'^2\Big)^2\sum_{i=1}^{N}Y_i'\,X_i'$$

$$-\sum_{i=1}^{N}X_i'^3\sum_{i=1}^{N}X_i'^2\,Y_i'-2(\bar{X}+\bar{x}')\sum_{i=1}^{N}X_i'^2.\sum_{i=1}^{N}X_i'^2\,Y_i\Big\}$$

$$\sum_{i=1}^{N}x_i'-\Big\{\sum_{i=1}^{N}X_i'^2\sum_{i=1}^{N}X_i'^2\,Y_i'-\sum_{i=1}^{N}X_i'^3\sum_{i=1}^{N}X_i'\,Y_i'\Big\}$$

$$\sum_{i=1}^{N}X_i'^2\Big]$$

$$=\bar{Y}$$

Therefore, the usual quadratic regression estimator is unbiased for this sample scheme. The variance of this estimator to the first order of approximation is given by

$$V(t_s^*) = \frac{N-n}{n(N-1)} \mu_{02} (1 - \eta_{(2)}^2)$$

where $\eta_{(2)}^2$ is the correlation ratio. Thus for large samples the efficiency of the proposed sampling scheme is same as the efficiency of the usual quadratic regression estimator under simple random sampling without replacement.

2.3 *Sampling Scheme Providing Unbiased Multi-variate Regression Estimator*

Here we consider the sampling scheme for which 2–variate regression estimator is unbiased and the extension for a general k–variate regression estimator thereafter is straightforward. Let Y_i, X_{1i} and X_{2i} be the values of the study character (y), auxilliary characters x_1 and x_2 for ith unit of the population ($i=1, 2, \ldots N$). It is assumed that the auxilliary characters are known for all the population units. The sampling scheme then consists of the following steps:

(a) Select a samale of size n by simple random sampling without replacement. For the selected sample calculate

$$L_s = s_{x1}^2 \, s_{x_2}^2 - s_{x1x2}^2$$

(b) Select a random number j from 0 to M, where,

$$M \geqslant \text{Max.} (L_s)$$

(c) If $j \leq L_s$, retain the sample, otherwise proceed to step (a) for selecting another sample and continue the process till a sample is selected.

Now, it is clear that for the suggested sampling scheme, the probability of selecting a sample 's' is

$$P_s = \frac{s_{x1}^2 s_{x2}^2 - s_{x1x2}^2}{\binom{N}{n} T_1}$$

where, $$T_1 = \frac{1}{\binom{N}{n}} \sum^{\binom{N}{n}} (s_{x1}^2 s_{x2}^2 - s_{x1x2}^2)$$

Under the suggested sampling scheme the estimator is $t_s'' = \bar{y} - b_1 \bar{x}_1' - b_2 \bar{x}_2'$ where b_1 and b_2 are partial regression coefficients.

Now

$$t_s'' = \bar{y} - \frac{s_{x2}^2 s_{x1y} - s_{x1x2} s_{x2y}}{s_{x1}^2 s_{x2}^2 - s_{x1x2}^2} \bar{x}_1' - \frac{s_{x1}^2 s_{x2y} - s_{x1x2} s_{x1y}}{s_{x1}^2 s_{x2}^2 - s_{x1x2}^2} \bar{x}_2'$$

$$E(t_s'') = \frac{1}{\binom{N}{n} T_1} \underset{s \in S}{\xi}\ t_s'' (s_{x1}^2 s_{x2}^2 - s_{x1x2}^2)$$

$$= \frac{1}{T_1} \xi [\bar{y} (s_{x1}^2 s_{x2}^2 - s_{x1x2}^2) - (s_{x2}^2 s_{x1y} - s_{x1x2} s_{x2y}) \bar{x}_1'$$

$$- (s_{x1}^2 s_{x2y} - s_{x1x2} s_{x1y}) \bar{x}_2']$$

$$= \frac{1}{T_1} \xi \Big[(\bar{Y} + \bar{y}') \Big\{ \sum_{r=1}^{n} (x_{1r} - \bar{x}_1)^2 \sum_{r=1}^{n} (x_{2r} - \bar{x}_2)^2$$

$$- \Big(\sum_{r=1}^{n} x_{1r} (x_{2r} - \bar{x}_2) \Big)^2 \Big\}$$

$$- \bar{x}_1' \Big\{ \sum_{r=1}^{n} (x_{2r} - x_2)^2 \sum_{r=1}^{n} y_r (x_{1r} - \bar{x}_1) - \sum_{r=1}^{n} x_{1r} (x_{2r} - \bar{x}_2)$$

$$\sum_{r=1}^{n} y_r (x_{2r} - \bar{x}_2) \Big\} - \bar{x}_2' \Big\{ \sum_{r=1}^{n} (x_{1r} - \bar{x}_1)^2 \sum_{r=1}^{n} y_r (x_{2r} - \bar{x}_2)$$

$$- \sum_{r=1}^{n} x_{1r} (x_{2r} - x_2) \sum_{r=1}^{n} y_r (x_{1r} - \bar{x}_1) \Big\} \Big]$$

This on simplification reduces to

$$= \frac{\bar{Y}}{T_1} \xi \Big[\Big(\sum_{r=1}^{n} x_{1r}'^2 - n \bar{x}_1'^2 \Big) \Big(\sum_{r=1}^{n} x_{2r}'^2 - n \bar{x}_2'^2 \Big)$$

$$- \Big(\sum_{r=1}^{n} (x_{1r}' x_{2r}' - n \bar{x}_1' \bar{x}_2') \Big)^2$$

$$+ \frac{1}{T_1} \xi \Big[\bar{y}' \Big\{ \sum_{r=1}^{n} x_{1r}'^2 \sum_{r=1}^{n} x_{2r}'^2 - \Big(\sum_{r=1}^{n} x_{1r}' x_{2r}' \Big)^2 \Big\}$$

$$- \bar{x}_1' \Big\{ \sum_{r=1}^{n} x_{2r}'^2 \sum_{r=1}^{n} x_{1r}' y_r' - \sum_{r=1}^{n} x_{1r}' x_{2r}' \sum_{r=1}^{n} x_{2r}' y_r' \Big\}$$

$$- \bar{x}_2' \Big\{ \sum_{r=1}^{n} x_{1r}'^2 \sum_{r=1}^{n} x_{2r}' y_r' - \sum_{r=1}^{n} x_{1r}' x_{2r}' \sum_{r=1}^{n} x_{1r}' y_r' \Big\} \Big]$$

$$= \frac{\bar{Y}}{T_1} \cdot T_1 + \frac{1}{T_1} \cdot 0$$

$$= \bar{Y}$$

Therefore, the usual 2 – variate regression estimator is unbiased for this sampling scheme and the variance of this estimator to the first order of approximation is given by

$$V(t_s'') = \Big(\frac{1}{n} - \frac{1}{N}\Big) \mu_{02} \, (1 - R^2_{1.23})$$

where $R^2_{1.23}$ represents the multiple correlation coefficient of y with x_1 and x_2.

The proposed sampling schemes together with the estimators will hereafter be referred as sampling strategies.

3. Comparison

For large samples, it has been observed that the efficiency of the usual estimators under proposed sampling schemes is of the same order as that of the same estimators under simple random sampling without replacement. For comparing the efficiency of the proposed strategies for small samples, we consider the following empirical comparisons.

3.1 We consider 18 populations of size 15 each. These populations

TABLE 2—VARIANCES/MEAN SQUARE ERRORS OF DIFFERENT SAMPLING STRATEGIES FOR $n = 4$

Population	V_1	V_2	V_3	V_4	V_5	V_6
1.	0.1630×10^{0}	0.1646×10^{4}	0.2102×10^{2}	0.5621×10^{4}	0.1853×10^{2}	0.1996×10^{2}
2.	0.4949×10^{1}	0.1649×10^{4}	0.2577×10^{2}	0.5625×10^{4}	0.2209×10^{2}	0.2475×10^{2}
3.	0.4156×10^{3}	0.1960×10^{4}	0.4330×10^{3}	0.5077×10^{4}	0.3275×10^{8}	0.4354×10^{3}
4.	0.1630×10^{0}	0.3448×10^{4}	0.1389×10^{1}	0.1182×10^{5}	0.3803×10^{1}	0.1630×10^{0}
5.	0.6270×10^{1}	0.1458×10^{4}	$0,7828 \times 10^{1}$	0.3145×10^{4}	0.7363×10^{1}	0.7284×10^{1}
6.	0.4126×10^{3}	0.3754×10^{4}	0.4134×10^{3}	0.1217×10^{5}	0.3128×10^{3}	0.4156×10^{3}
7.	0.1630×10^{0}	0.2202×10^{4}	0.5071×10^{1}	0.7464×10^{4}	0.1485×10^{2}	0.1620×10^{0}
8.	0.5033×10^{1}	0.2201×10^{4}	0.9956×10^{1}	0.7552×10^{4}	0.1863×10^{2}	0.5033×10^{1}
9.	0.4156×10^{3}	0.2516×10^{4}	0.4171×10^{3}	0.7920×10^{4}	0.3238×10^{3}	0.4156×10^{3}
10.	0.1630×10^{0}	0.2730×10^{4}	0.1980×10^{2}	0.9346×10^{4}	0.1485×10^{2}	0.1996×10^{2}
11.	0.4949×10^{1}	0.2734×10^{4}	0.2454×10^{2}	0.9351×10^{4}	0.1841×10^{2}	0.2475×10^{2}
12.	0.4156×10^{3}	0.3044×10^{4}	0.4318×10^{3}	0.9702×10^{4}	0.3938×10^{3}	0.4354×10^{3}
13	0.1630×10^{0}	0.3754×10^{4}	0.5071×10^{1}	0.1289×10^{5}	0.3803×10^{1}	0.5113×10^{1}
14.	0.4949×10^{1}	0.3758×10^{4}	0.9817×10^{1}	0.1289×10^{5}	0.7363×10^{1}	0.9888×10^{1}
15.	0.4156×10^{3}	0.4068×10^{4}	0.4171×10^{3}	0.1924×10^{5}	0.3128×10^{3}	0.4205×10^{3}
16.	0.1630×10^{0}	0.1431×10^{4}	0.9980×10^{1}	0.4908×10^{4}	0.1853×10^{2}	0.5113×10^{1}
17.	0.4949×10^{1}	0.1435×10^{4}	0.1473×10^{2}	0.4813×10^{4}	0.2289×10^{2}	0.9899×10^{1}
18.	0.4156×10^{3}	0.1745×10^{4}	0.4220×10^{3}	0.5264×10^{4}	0.3275×10^{3}	0.4285×10^{3}

with $E(\epsilon_i \mid X_{1i}, X_{2i}) = 0$

and $E(\epsilon_i^2 \mid X_{1i}, X_{2i}) = A\,(X_{1i}, X_{2i})^g$

The strategies considered for comparison are as follows :

1. Suggested sampling strategy-3.
2. Ratio estimator under Midzuno-Sen sampling scheme.
3. Regression estimator under simple random sampling without replacement.
4. Ratio estimator under simple random sampling without replacement.
5. Simple mean under simple random sampling without replacement.
6. Suggested sampling strategy-1.

In strategies (2), (3), (4) and (6) the auxiliary variable used is x_1 only. Estimators of population mean or total obtained from (1), (2), (5) and (6) are unbiased, whereas, (3) and (4) provide biased estimators. The variances/mean square errors of the above strategies denoted by V_1, V_2, V_3, V_4, V_5 and V_6 for sample of size 4 have been computed and results presented in Table 2.

The following observations can be made from the above comparison:

(i) The suggested sampling strategy-3 is better than the usual regression estimator in all the populations.

(ii) The suggested strategy-3 is better than the ratio estimator under Midzuno-Sen sampling scheme in all the populations.

These results indicate that the performance of the suggested sampling strategy-3 is highly satisfactory and the gain in efficiency is quite considerable.

REFERENCE

Singh, P. and Srivastava, A.K. (1980): Sampling scheme providing unbiased regression estimator. *Biometrika*. 67, 205-209.

Jour. Ind. Soc. Ag. Statistics
Vol. XLII, No. 2 (1998), pp. 218-225

MILK PRODUCTION FUNCTIONS AND RESOURCE PRODUCTIVITY IN BOVINE

S. B. AGARWAL, R. K. PATEL and K. N. S. SHARMA
National Dairy Research Institute, Karnal (Haryana)
(Received : April, 1988)

SUMMARY

Feed-milk data on 32 Hariana cows and 30 Murrah buffaloes were collected both in the morning and evening by actual weighing throughout their lactation at a regular inlerval of one week to study milk production functions and resource productivity in bovine around Karnal under village conditions. Linear and log linear milk production functions were tried. The averge milk yield porday of lactation was estimated at 3.08 kg for cows and 3.75 kg for buffaloes. The intake of DCP per day of lactation was worked out to 0.26 kg and 0.32 kg whereas TDN 4.08 kg and 5.72 kg for cows and buffaloes respectively. The intake of DCP and TDN was less during dry period. Linear milk production functions were found more suitable compared to log-linear both in cows and buffaloes. Animals were given more nutrients during dry period than the requirement in relation to milk yield. The elasticities of inputs were gener ally higher for buffaloes compared to cows. The marginal value product of resources suggested that the milk producers would afford cost of DCP upto Rs. 8 for cows, and Rs. 15 for buffaloes for enhancing the milk productivity. Thus it was revealed that the reallocation of feed resources can play a significant role in increasing the milk production of both cows and buffaloes.

Introduction

Milk production is the net outcome of feed, breed, management and environmental effects. Feed alone accounts for about 60 percent of the total cost of milk production (Kuber Ram *et al.* [5]). Improved feeding practices and better management play a significant role in increasing the milk production of bovine (Agarwal *et al.* [1]). Feeds and fodders have