Available online at www.isas.org.in/jisas

JOURNAL OF THE INDIAN SOCIETY OF AGRICULTURAL STATISTICS 66(3) 2012 381-390

Studies on Some Preliminary Test Estimators in Double Sampling

Phrangstone Khongji[1] and Gitasree Das[2]

[1]*Department of Basic Sciences and Social Sciences, School of Technology, North Eastern Hill University, Shillong, Meghalaya*

[2]*Department of Statistics, North Eastern Hill University, Shillong, Meghalaya*

Received 29 October 2010; Revised 08 June 2012; Accepted 11 June 2012

SUMMARY

It is known that in many of the large scale surveys, it is inevitable to adopt stratification for the purpose of preparing a frame from which the sample can be extracted. Cochran (1977) suggested a regression estimate in stratified sampling which he called a combined regression estimate. In the present study, situations will be considered where partial information about the mean of the auxiliary variable is available. In order to utilize the partial information, double sampling is used and a preliminary test is done to construct the combined regression preliminary test estimator. The bias, mean square error and relative efficiency are obtained for the suggested estimator. Apart from analytical results, these are also obtained by numerical techniques. The comparative study shows that the bias and mean square error function obtained by numerical methods depict similar pattern with that obtained by analytical methods. In order to judge the performance of the suggested estimator, besides analytical results, empirical work is also carried out with the help of both real life data as well as simulated data. Recommendation of the levels of the preliminary test and optimum allocation of sample sizes are given.

Keywords: Double sampling, Preliminary test estimator, Regression estimator.

1. INTRODUCTION

It is a well known fact that for estimating the population mean μ_y of a random variable Y, precision of the estimator can be increased when information on an auxiliary variable X, highly correlated with Y is readily available on all the units of the population. When the relationship between Y and X is found to be approximately linear but does not pass through the origin, linear regression estimate may be used, which is given by (Cochran 1977) as

$$t_1 = \bar{y} + b(\mu_x - \bar{x})$$

where b is an estimate of the change in y when x is increased by unity, $\bar{y}$ and $\bar{x}$ are the sample means of the variables Y and X respectively and μ_x is the population mean of X.

2. DOUBLE SAMPLING WITH PARTIAL INFORMATION ON AUXILIARY VARIABLES

To use the linear regression estimator t_1 it is usually assumed that population mean μ_x is known. However, in certain practical situation, μ_x is not known a prior, in which case the technique of double sampling is applied. Under double sampling the regression estimate t_1 becomes

$$t_2 = \bar{y}_n + b(\bar{x}_{n'} - \bar{x}_n)$$

where $\bar{x}_{n'}$ is the mean of X in the preliminary sample of size n' and $(\bar{y}_n, \bar{x}_n)$ are the means of Y and X from the sub sample of size n ($< n'$) and b is the least square regression coefficient of Y on X which can be computed from the sub sample.

*Corresponding author : Phrangstone Khongji
E-mail address : phrang2000@yahoo.com

Han (1973) described that the precision of an estimator can be improved if auxiliary variables are used in a regression estimator based on double sampling with partial information on auxiliary variable. Sometimes there are situations where we have partial information about the mean μ_x of the auxiliary variable *X*. In order to utilize the partial information, Han (1973) suggested the use of a preliminary test whereby he constructed a preliminary test estimator using double sampling with partial information on the auxiliary variable as follows

$$t_3 = \begin{cases} (\bar{y}_n - \rho\,\bar{x}_n) & \text{if} \quad |\bar{x}_{n'}| \le Z_\alpha/\sqrt{n'} \\ (\bar{y}_n + \rho(\bar{x}_{n'} - \bar{x}_n)) & \text{if} \quad |\bar{x}_{n'}| > Z_\alpha/\sqrt{n'} \end{cases}$$

where Z_α is the 100 $(1 - \alpha/2)$% point of $N(0, 1)$ and α is the level of significance of the preliminary test. The correlation coefficient ρ of the pair (X, Y) is assumed to be known.

Sisodia and Srivastava (1982), Das and Bez (1995), Kibria (1996), Das (2003) and others proposed some modified regression estimator with a preliminary test in double sampling, alternative to the usual regression estimator for the population mean.

The present work is aimed to proceed in accordance to further enhance the work done by Han (1973) and Das (1995) and several other authors to find an appropriate estimator through the use of preliminary test estimation and double sampling procedures.

It is known that stratified sampling consists of classifying the population units in a certain number of groups called strata and selecting samples independently from each group. The division of population into strata can be done in such a way that the values of the study variable are homogeneous within each stratum, in that the measurement varies little from one unit to another. A precise estimate of any stratum mean can be obtained from a small sample in that stratum. These estimates with best choices of sample sizes can be combined into a precise estimate for the whole population. When appropriately used, the variance of the estimated mean of the study variable *Y* under stratification is usually less than that of the variance under simple random sampling (Cochran 1977).

Cochran (1977) suggested a regression estimate in stratified sampling which he called a combined regression estimator and is given by

$$\bar{y}_{lrc} = \bar{y}_{st} + b(\mu_x - \bar{x}_{st}), \text{ where}$$

$$\bar{y}_{st} = \sum_h W_h \bar{y}_h, \qquad \bar{x}_{st} = \sum_h W_h \bar{x}_h$$

where $(\bar{y}_h, \bar{x}_h)$ are the strata means of the h^{th} stratum and $h = 1, 2 \ldots, L$.

In this estimate the whole population is stratified into different classes and samples are selected from each stratum by simple random sampling and the strata means are combined and used in a regression equation to obtain the desired mean. Here b is the estimate of combined regression coefficient and W_h is the stratum weight.

3. THE COMBINED REGRESSION PRELIMINARY TEST ESTIMATOR (CRPTE) IN DOUBLE SAMPLING

The combined linear regression estimator given by $\bar{y}_{lrc}$, can be utilized under three situations. Firstly when the population mean μ_x is known, as a consequence of which the study reduces to usual combined regression method of estimation. Secondly in certain practical situations μ_x is not known a prior, in which case the technique of double sampling can be applied wherein a preliminary sample is obtained to estimate μ_x and the estimator of μ_y is given by

$$t_4 = \bar{y}_{st} + b(\bar{x}_{n'} - \bar{x}_{st})$$

Here $\bar{x}_{n'}$ is the value of the mean of X obtained from the preliminary sample and is utilized to estimate μ_x. Thirdly when μ_x is partially known, then a preliminary test estimator using double sampling procedure can be used.

In the present study, the third case will be considered where partial information about the mean of the auxiliary variable will be used. The first sample is a stratified simple random sample of size n in which the pair (x_{hi}, y_{hi}) values are measured from n_h units drawn from each stratum and consequently estimating

of the pair $(\bar{x}_{st}, \bar{y}_{st})$, with $n = \sum_h n_h$. The second sample is a larger simple random sample of size $n'(= n + m)$ obtained by supplementing m more independent observations on X where only x_i is measured and evaluates $\bar{x}_{n'}$ which is utilized to estimate μ_x. In order to utilize the partial information, a preliminary test is done about the hypothesis

$$H_0 : \mu_x = \mu_0, \text{ against } H_1 : \mu_x \neq \mu_0$$

where μ_0 is the value obtained from the partial information.

If the null hypothesis H_0 is accepted, then μ_0 is used to replace μ_x in the regression estimator $\bar{y}_{lrc}$ and if H_0 is rejected then the sample mean $\bar{x}_{n'}$ based on the preliminary sample is used.

We assume that the auxiliary variable X and the study variable Y are jointly normally distributed with parameters given by $(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho)$. The marginal distributions which is the distribution of the study variable Y and the auxiliary variable X will also follow normal distribution given as $Y \sim N(\mu_y, \sigma_y^2)$ and $X \sim N(\mu_x, \sigma_x^2)$. The strata population (X_h, Y_h) being carved out from the parent population are also jointly assumed to follow the bivariate normal distribution with parameters written as $(\mu_{x_h}, \mu_{y_h}, \sigma_{x_h}^2, \sigma_{y_h}^2, \rho)$. Also since the relationship between the pair (X, Y) is always maintained even within the stratum the strata correlations are assumed to be equal to the population correlation coefficient ρ. The regression estimator depends on whether the covariance matrix is known or not. If known, one may let $\sigma_x^2 = \sigma_y^2 = 1$ without loss of generality.

Since the population is assumed to follow normal distribution, the preliminary sample utilized to collect information on the auxiliary variable for the estimation of $\bar{x}_{n'}$, also assumed to follow normal distribution and therefore $\bar{x}_{n'} \sim N(\mu_x, \sigma_x^2 / n')$ and under the assumption $\sigma_x^2 = \sigma_y^2 = 1$, we get $\bar{x}_{n'} \sim N(\mu_x, 1/n')$. Further marginal distributions of X_h and Y_h are also normal given as $X_h \sim N(\mu_{x_h}, \sigma_{x_h}^2)$ and $Y_n \sim N(\mu_{y_h}, \sigma_{y_h}^2)$.

For each stratum, the pair of variables (X_h, Y_h) for every h, follows a bivariate normal distribution with mean (μ_{x_h}, μ_{y_h}) and covariance matrix given by

$$\sum\nolimits_h = \begin{pmatrix} \sigma_{x_h}^2 & \rho\sigma_{x_h}\sigma_{y_h} \\ \rho\sigma_{x_h}\sigma_{y_h} & \sigma_{y_h}^2 \end{pmatrix}$$

The regression estimator depends on whether Σ_h is known or not. If Σ_h is known, one may let $\sigma_{x_h}^2 = \sigma_{y_h}^2 = 1$, (without loss of generality).

Also, the stratum means are given by

$$\bar{x}_h = \sum_{i=1}^{n_h} x_{hi} / n_h \quad \text{and} \quad \bar{y}_h = \sum_{i=1}^{n_h} y_{hi} / n_h$$

and are linear combinations of normally distributed random variables (X_h, Y_h).

Hence it can be easily observed that $\bar{x}_h$ and $\bar{y}_h$ also follow normal distribution with mean and variances given by

$$\bar{x}_h \sim N(\mu_{x_h}, \sigma_{x_h}^2 / n_h) \quad \text{and} \quad \bar{y}_h \sim N(\mu_{y_h}, \sigma_{y_h}^2 / n_h)$$

i.e. $\bar{x}_h \sim N(\mu_{x_h}, 1/n_h)$ and $\bar{y}_h \sim N(\mu_{y_h}, 1/n_h)$

The joint distribution of $(\bar{x}_h, \bar{y}_h)$ is bivariate normal with mean as (μ_{x_h}, μ_{y_h}) and covariance matrix of the sample means is given by

$$\sum\nolimits_c = \begin{pmatrix} \sigma_{x_h}^2 / n_h & \rho\sigma_{x_h}\sigma_{y_h} / n_h \\ \rho\sigma_{x_h}\sigma_{y_h} / n_h & \sigma_{y_h}^2 / n_h \end{pmatrix} = \begin{pmatrix} 1/n_h & \rho/n_h \\ \rho/n_h & 1/n_h \end{pmatrix}$$

Now when m_x is partially known, one can let $\mu_0 = 0$ without loss of generality), so that the hypothesis can be accepted, when

$$\left|(\bar{x}_{n'} - \mu_o) / SE(\bar{x}_{n'})\right| \leq Z_\alpha \quad \Rightarrow \quad \left|\bar{x}_{n'}\right| \leq Z_\alpha / \sqrt{n'}$$

Under the above assumptions the CRPTE in double sampling having partial information on the auxiliary variable X can be written as

$$t_5 = \begin{cases} (\bar{y}_{st} - \rho\bar{x}_{st}) & \text{if} \quad \left|\bar{x}_{n'}\right| \leq Z_\alpha / \sqrt{n'} \\ (\bar{y}_{st} + \rho(\bar{x}_{n'} - \bar{x}_{st})) & \text{if} \quad \left|\bar{x}_{n'}\right| > Z_\alpha / \sqrt{n'} \end{cases}$$

where, $b = \rho(\sigma_y / \sigma_x) = \rho$ under the above assumptions.

4. BIAS OF THE CRPTE IN DOUBLE SAMPLING

To evaluate the bias of t_5, we considered that the joint distribution of $(\bar{x}_{n'}, \bar{x}_{st}, \bar{y}_{st})$ which is a multivariate normal with mean (μ_x, μ_x, μ_y) and covariance matrix given by

$$\sum = \begin{pmatrix} \mathrm{Var}(\bar{x}_{n'}) & \mathrm{Cov}(\bar{x}_{n'}, \bar{x}_{st}) & \mathrm{Cov}(\bar{x}_{n'}, \bar{y}_{st}) \\ \mathrm{Cov}(\bar{x}_{st}, x_{n'}) & \mathrm{Var}(\bar{x}_{st}) & \mathrm{Cov}(\bar{x}_{st}, \bar{y}_{st}) \\ \mathrm{Cov}(\bar{y}_{st}, \bar{x}_{n'}) & \mathrm{Cov}(\bar{y}_{st}, \bar{x}_{st}) & \mathrm{Var}(\bar{y}_{st}) \end{pmatrix} \quad (1)$$

The derivation of the bias of the estimator t_5 involves conditional expectations, the condition being the acceptance or rejection of the hypothesis considered in the preliminary test. Further the expectations can be obtained from the integrals involving probability density functions which are assumed to be normal.

When the samples are selected with proportional allocation then the stratum weight is given by $W_h = (N_h/N) = (n_h/n)$

Thus,

$$\sum_h W_h^2 / n_h = \sum_h W_h^2 / n W_h = (1/n) \sum_h W_h = (1/n) \quad \left(\text{as } \sum_h W_h = 1 \right)$$

Therefore the above covariance matrix in (1) reduces to

$$\sum = \begin{pmatrix} 1/n' & 1/n' & \rho/n' \\ 1/n' & 1/n & \rho/n \\ \rho/n' & \rho/n & 1/n \end{pmatrix}$$

The Bias of an estimator is defined as

$$\mathrm{Bias}(t_5) = \mathrm{E}(t_5) - \mu_y$$

where $\mathrm{E}(\cdot)$ is the expectation

$$\mathrm{Bias}(t_5) = \mathrm{E}\left\{(\bar{y}_{st} - \rho\bar{x}_{st})\ldots \text{ if } |\bar{x}_{n'}| \le Z_\alpha / \sqrt{n'}\right\}$$

$$+\mathrm{E}\left\{(\bar{y}_{st} + \rho(\bar{x}_{n'} - \bar{x}_{st}))\ldots \text{if } |\bar{x}_{n'}| > Z_\alpha / \sqrt{n'}\right\} - \mu_y$$

$$\Rightarrow \mathrm{Bias}(t_5) = \sum W_h \mu_{y_h} - \rho \sum W_h \mu_{x_h}$$

$$+\mathrm{E}\left\{\rho\bar{x}_{n'} / |\bar{x}_{n'}| > Z_\alpha / \sqrt{n'}\right\} - \mu_y$$

It is given that

$$\sum W_h \mu_{x_h} = \mu_x \text{ and } \sum W_h \mu_{y_h} = \mu_y$$

(Cochran 1977), Thus

$$\mathrm{Bias}(t_5) = -\rho\mu_x + \rho\mathrm{E}\left\{\bar{x}_{n'} / |\bar{x}_{n'}| > Z_\alpha / \sqrt{n'}\right\}$$

After evaluating the integrals, we get

$$\mathrm{Bias}(t_5) = -\rho\mu_x\{\Phi(A) - \Phi(B)\}$$

$$+\rho(1/\sqrt{n'})\ \{\varphi(A) - \varphi(B)\} \quad (2)$$

where $\Phi(\cdot)$ is the cumulative distribution function of $N(0, 1)$ and $\varphi(\cdot)$is its density function and $A = Z_\alpha - \sqrt{n'}\ \mu_x$, $B = -Z_\alpha - \sqrt{n'}\ \mu_x$.

When the bias of the proposed estimator is computed for different values of the mean of the auxiliary variables μ_x, it is noticed that the behavior of the bias is symmetrical about $\mu_x = 0$. Thus it suffices to analyze the behavior of the bias for $\mu_x \ge 0$. The values of $\mathrm{Bias}(t_5)$ can be easily computed for different values of μ_x. In order to get an idea about the behavior of the bias function with respect to μ_x, $\mathrm{Bias}(t_5)$ is computed for a set of values of n, n', α and ρ which are represented in Fig. 1. It is found in general that $\mathrm{Bias}(t_5)$ has minimum value zero at $\mu_x = 0$. As μ_x increases, the $\mathrm{Bias}(t_5)$ increases to a maximum and then decreases to zero. Fig. 1 clearly shows that when the mean of the auxiliary variable is close to the hypothetical value, then bias is very close to 0. Also as μ_x moves away from the hypothetical value the bias increases, but after attaining maximum again reduces to zero. This establishes the utility of the present study that the utilization of partial information and preliminary test of the auxiliary variable reduces the bias of the proposed estimator.

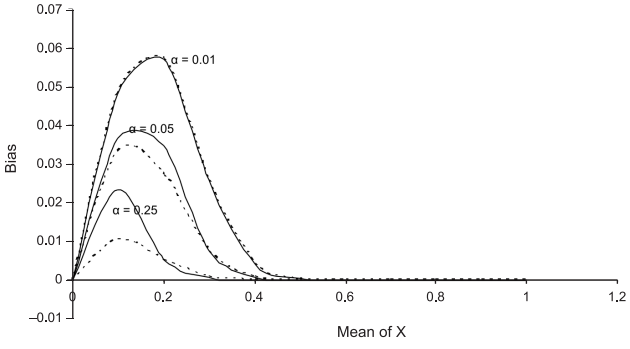

Fig. 1. Comparative behaviour of the $\mathrm{Bias}(t_5)$ with respect to μ_x for different values of α and for $\rho = 0.8$, $n = 100$, $n' = 200$, ___ analytical, ------ numerical .

It is also observed from eqn(2) that when the parameter α and ρ are fixed, then the bias is inversely proportional to the square root of the size of the preliminary sample n'. It can be concluded that with the increase in the sample size, the bias decreases. However, the bias is not affected by n, the size of the stratified random sample.

The above analytical method used for computing the bias the proposed estimator involves the evaluation of mathematical expectation of the random variables and consequently results in the computation of integrals. However, sometimes computation of integrals analytically may become cumbersome. Therefore an alternative method for the evaluation of the bias is also sought with the help of numerical techniques. In the present study, attempt is made to evaluate the bias of the constructed estimator t_5 using numerical integration with programmed written in Fortran 77. When the results obtained numerically and that by analytical methods are compared, it is found (Fig. 1) that the bias function shows a similar pattern. Also, we may observe, that the difference decreases with decrease in α, and when $\alpha = 0.01$, the bias obtained analytically and numerically coincide.

5. MEAN SQUARE ERROR (MSE) OF THE CRPTE IN DOUBLE SAMPLING

The derivation of MSE involves conditional expectations, the condition being the acceptance or rejection of the hypothesis considered in the preliminary test. Further the expectations can be obtained from the integrals involving probability density functions which are assumed to be normal.

To obtain MSE of t_5, we notice that

$$\begin{aligned}\mathrm{MSE}(t_5) &= \mathrm{var}(t_5) + \{\mathrm{Bias}(t_5)\}^2 \\ &= \mathrm{E}(t_5^2) - \{\mathrm{E}(t_5)\}^2 + \{\mathrm{Bias}(t_5)\}^2 \qquad (3)\end{aligned}$$

Now,

$$\begin{aligned}\mathrm{E}(t_5^2) &= \mathrm{E}(t_5^2 \text{ if } |\bar{x}_{n'}| \le Z_\alpha/\sqrt{n'}) \\ &\quad + \mathrm{E}(t_5^2 \text{ if } |\bar{x}_{n'}| > Z_\alpha/\sqrt{n'}) \\ &= \{\mathrm{E}(\bar{y}_{st} - \rho\bar{x}_{st})^2 \text{ if } |\bar{x}_{n'}| \le Z_\alpha/\sqrt{n'}\} \\ &\quad + [\mathrm{E}\{\bar{y}_{st} + \rho(\bar{x}_{n'} + \bar{x}_{st})\}^2 \text{ if } |\bar{x}_{n'}| > Z_\alpha/\sqrt{n'}]\end{aligned}$$

$$\begin{aligned}\mathrm{E}(t_5^2) &= \mathrm{E}(\bar{y}_{st}^2) - 2\rho\mathrm{E}(\bar{x}_{st}\bar{y}_{st}) + \rho^2\mathrm{E}(\bar{x}_{st}^2) \\ &\quad + \mathrm{E}\{(\rho^2\bar{x}_{n'}^2 - 2\rho^2\bar{x}_{n'}\bar{x}_{st} + 2\rho\bar{x}_{n'}\bar{y}_{st}) \text{ if } |\bar{x}_{n'}| > Z_\alpha/\sqrt{n'}\}\end{aligned}$$

The expectations on the right hand side involve computations of integrals of the bivariate normal probability density function. Moment generating function and differentiating under an integral sign is being utilized to simplify the above computations, hence

$$\begin{aligned}\mathrm{E}(t_5^2) &= (1-\rho^2)/n + \mu_y^2 + \rho^2/n' - \{2\rho\mu_x\mu_y - \rho^2\mu_x^2 \\ &\quad + (\rho^2/n')\}\{\Phi(A) - \Phi(B)\} + \{(2\rho\mu_y/\sqrt{n'}) \\ &\quad (\varphi(A) - \varphi(B) + (\rho^2/n')(A\varphi(A) - B\varphi(B)) \qquad (4)\end{aligned}$$

Therefore, substituting (2) and (4) in (3), we get

$$\mathrm{MSE}(t_5) = g_1 + h_1,$$

where $g_1 = \{(1 - \rho^2)/n + \rho^2/n'\}$

and $h_1 = (\rho^2/n')\{A\varphi(A) - B\varphi(B)\} - \rho^2(1/n' - \mu_x^2)\{\Phi(A) - \Phi(B)\}$

The values of $\mathrm{MSE}(t_5)$ can be easily computed for different values of μ_x. In order to get an idea about the behavior of the mean square error function with respect to μ_x, $\mathrm{MSE}(t_5)$ is computed for a set of values of n, n', α and ρ which are represented in Fig. 2. The figure shows that as μ_x is increases, the $\mathrm{MSE}(t_5)$ increases to a maximum and then decreases to a minimum and thereafter remains constant. The figure clearly shows that when the mean of the auxiliary variable is close to the hypothetical value, then the $\mathrm{MSE}(t_5)$ is minimum.

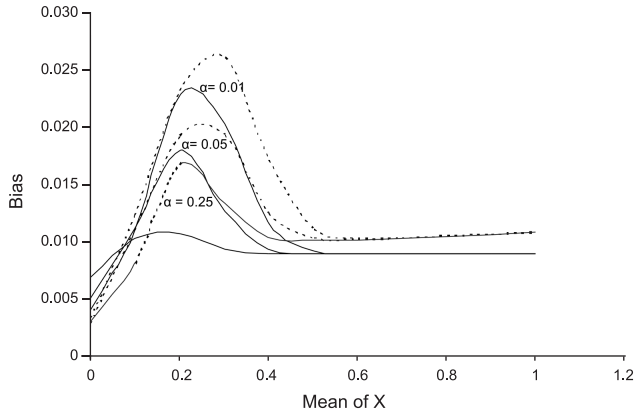

Fig. 2. Comparative behaviour of the $\mathrm{MSE}(t_5)$ with respect to μ_x for different values of α and for $\rho = 0.8$, $n = 100$, $n' = 200$, ___ analytical , -------- numerical.

Also as μ_x moves away from the hypothetical value the MSE(t_5) increases, but after attaining maximum again reduces to a constant. This establishes the utility of the present study that the utilization of partial information and preliminary test of the auxiliary variable reduces the MSE(t_5) of the proposed estimator.

It is seen that the analytical method of determining the MSE(t_5) involves evaluating the mathematical expectations of the random variables like $E(\bar{x}_{n'})$, $E(\bar{x}_{n'}^2)$, $E(\bar{x}_{n'}\ \bar{x}_{st})$ and $E(\bar{x}_{n'}\ \bar{x}_{st})$. The derivation of these expectation is done using moment generating function and this involves the application of single and double integration techniques. In the process of evaluation which involves bivariate frequency distributions, a tedious substitution of variables is necessary to simplify the integrals. The above expectation is finally obtained by differentiating under the integral sign. As a consequence of the complexity involved in analytical deductions and the availability of numerical techniques, the MSE(t_5) is evaluated using the numerical methods. In the numerical methods, the use of moment generating function and the substitutions involved can be avoided. The results of MSE(t_5) obtained numerically shows (Fig. 2) the similar pattern with that of the one derived by analytical methods for increasing values of the mean μ_x of the auxiliary variable. The differences in the values of MSE between analytical and numerical methods of computations are minimal.

6. RELATIVE EFFICIENCY

The study of the mean square error of an estimator will not be complete unless it is compared with other estimator(s). Without the use of real life data, relative efficiency of an estimator can be obtained analytically as the ratio of the variance or mean square error of one estimator to that of the mean square error of the proposed estimator. If the relative efficiency is greater than 1, it can be concluded that the proposed estimator is more efficient in comparison to the other estimator.

In the present study, the mean square error of the proposed estimator is compared with other estimator t_4 and conclusion is drawn through the relative efficiency. Under similar assumptions, derivation gives

$$\text{MSE}(t_4) = (1/n)(1 - \rho^2) + (1/n')\rho^2$$

Therefore the relative efficiency of t_5 to t_4 is given by

$$e(\alpha, \mu_x) = [\text{MSE}(t_4)]/[\text{MSE}(t_5)]$$

In order to get an idea about the behavior of the relative efficiency function with respect to μ_x, $e(\alpha, \mu_x)$ is computed for a set of values of n, n', α and ρ. Fig. 3 shows that, in general that $e(\alpha, \mu_x)$ has a maximum at $\mu_x = 0$. This establishes the utility of the present study that the utilization of partial information and preliminary test increases the efficiency of the estimator. Further Fig. 3 shows that as μ_x increases $e(\alpha, \mu_x)$ decreases to a minimum and then increases to unity. It is found that $e(\alpha, \mu_x)$ is very close to 1 at $\mu_x = 1$. Also, even though subjective, Fig. 3 can help to choose α values depending on μ_x so as to minimize the loss.

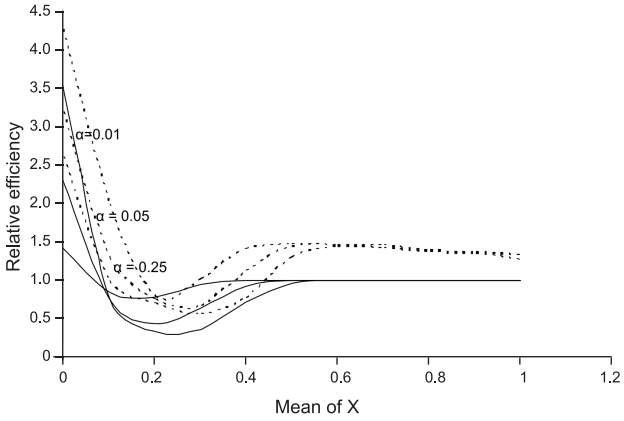

Fig. 3. Comparative behaviour of the relative efficiency of t_5 with respect to t_4 against μ_x for different values of α and for $\rho = 0.8$, $n = 100$, $n' = 200$. ___ analytical, -------- numerical.

7. OPTIMUM ALLOCATION

In planning of a sample survey, a stage is always reached at which an important decision must be made about the size of the sample. Too large a sample implies a waste of resources, and too small a sample diminishes the precision of the estimators. Thus an optimum size of the sample is required so as to balance precision and cost involved in the survey. The optimum allocation of sample sizes are attained either by minimizing precision against a given cost or minimizing cost against given precision. In obtaining optimum allocation of sample sizes for the proposed estimator, we consider a simple linear cost function C given by

$$C = c'n' + cn$$

where c is the cost per unit of observing the variable y and c' is the cost per unit of observing the variable X, assuming that the cost per unit is the same for all strata.

In general the values of μ_x are unknown, the experimenter has partial information about it. When $\mu_x = 0$, the mean square error of t_5 is least and the relative efficiency is largest. Thus it would be reasonable to let $\mu_x = 0$ in the MSE(t_5) and obtain the values of n and n' under the optimum situation of minimizing precision against a given cost.

For a specific cost C^*, routine mathematical derivation by the use of Lagrange's multipliers method gives

$$M_{opt}(t_5) = \left[\sqrt{Kc} + \sqrt{c'K'}\right]^2 / C^*$$

In a similar way the optimum allocation for the estimator t_4 is given by

$$M_{opt}(t_4) = (\sqrt{Kc} + \sqrt{K''c'}) / C^*$$

where $K = 1 - \rho^2$, $K' = \rho^2 \{\alpha + 2Z_\alpha \varphi(Z_\alpha)\}$ and $K'' = \rho^2$

Analytically it can be seen that $(\alpha + 2Z_\alpha \varphi(Z_\alpha))$ is a decreasing function of Z_α with a maximum equal to unity at $Z_\alpha = 0$. Therefore we can conclude that $M_{opt}(t_5) \leq \text{MSE}_{opt}(t_4)$ with equality holding for $Z_\alpha = 0$ in which case the two estimators coincide.

8. EMPIRICAL STUDIES AND CONCLUSION

In the present study, attempt is also made to study the performance of the proposed estimator through empirical data. Here, two sets of data were used : one a real life data and another, a simulated data. Real life data were extracted from Rapid household survey – Reproductive and Child health (RHS-RCH project, phase 1, 1998). The data provides district wise percentages of the demographic indicators for the Empowered Action Groups (EAG) states. From the data two distinct characteristics of the population were identified, namely complete child immunization (Y) and female literacy rate (X). The data, on the selected variables can be homogeneous within each state and heterogeneous between states. The variables X and Y being given in percentages can be considered to behave like the binomial variables and as a result, arcsine transformation is utilized to convert a binomial random variable into one that is nearly normal. The data corresponding to different states and also the combined data are tested for normality by using Shapiro Wilk's test and it was found that both X and Y follow normal distribution. Strata sample sizes obtained by proportional allocation for the EAG states is shown is Table 1.

Table 1. Strata sample sizes obtained by proportional allocation for the EAG states.

EAG States	Stratum	N_h	$W_h = (N_h/N)$	n_h
Bihar	1	30	0.19	6
Chhattisgarh	2	7	0.04	1
Jharkhand	3	13	0.08	2
Madhya Pradesh	4	38	0.24	7
Orissa	5	30	0.19	6
Rajasthan	6	30	0.19	6
Uttaranchal	7	10	0.06	2
Total		158	1	30

Source : Rapid household survey (RHS-RCH project, phase 1, 1998).

n_h is the stratum sample size

To utilize the proposed estimator t_5, it is assumed that the population mean μ_x of the auxiliary variable is partially known. When is unknown, alternative information about the mean of the auxiliary variable is obtained by the use of double sampling procedure and in the present case, this estimate of μ_x for female literacy rate is given by $\bar{x}_{n'} = 44.8\%$. Further suppose that we have partial information about μ_x and in this case the partial information on X is obtained from Census of India (1991) where the mean of female literacy rate is computed for the EAG states and given by $\mu_0 = 26.4\%$. In the present study the preliminary sample is utilized to test the hypothesis

$$H_0 : \mu_x = \mu_0 \text{ against } H_1 : \mu_x \neq \mu_0$$

If the hypothesis H_0 is accepted then μ_0 obtained from the partial information will be used in the proposed estimator; if H_0 is rejected, the sample mean $\bar{x}_{n'}$ based on the preliminary sample is used.

The mean of the auxiliary variable is considered to be partially known as μ_0. We can assume that $\mu_0 = 0$ without the loss of generality. The covariance matrix

Σ is also considered to be known, hence without loss of generality it can be assumed that $\sigma_x^2 = 1$ and $\sigma_y^2 = 1$. In order to compare μ_0 with the sample mean $\bar{x}_{n'}$ in the testing of hypothesis, the sample values of the auxiliary variable X is transformed both in origin and scale as $(x_i - \mu_0)/\sigma_X$, and the variable Y is transform to y_i/σ_Y, making sure that the assumptions on variances is unity.

Table 2 shows that for the proposed estimator, $\bar{x}_{st} = 1.497$, $\bar{y}_{st} = 3.347$ and $\bar{x}_{n'} = 1.438$ and in terms of the original unit $\bar{x}_{st} = 45.5\%$, $\bar{y}_{st} = 46.31\%$ and $\bar{x}_{n'} = 44.8\%$. The estimate of the combined regression preliminary test estimator t_5 in double sampling, for the percentage of complete child immunization for the EAG states is given by 45.29%.

Table 2. Computation of the strata means $\bar{x}_{st}$ and $\bar{y}_{st}$ for the EAG states.

EAG States	Stratum	W_h	$\bar{x}_h$	$\bar{y}_h$	$W_h\bar{x}_h$	$W_h\bar{y}_h$
Bihar	1	0.19	0.61	2.24	0.12	0.42
Chhattisgarh	2	0.04	0.61	3.72	0.03	0.16
Jharkhand	3	0.08	2.04	3.75	0.17	0.31
Madhya Pradesh	4	0.24	1.89	3.52	0.46	0.85
Orissa	5	0.19	1.58	3.97	0.30	0.75
Rajasthan	6	0.19	1.53	2.99	0.29	0.57
Uttaranchal	7	0.06	2.22	4.42	0.14	0.28
Total		1			$\bar{x}_{st}$ = 1.497	$\bar{y}_{st}$ = 3.347

Source : Rapid household survey (RHS-RCH project, phase 1,1998).

$(\bar{x}_h, \bar{y}_h)$ are the strata means.

When a reliable partial information on the mean of the auxiliary variable is not available, as in the present case, the hypothesis is rejected, as a result of which $\bar{x}_{n'}$ is utilized in the estimation of μ_x. Hence the proposed estimator t_5 reduces to the usual combined regression estimator under double sampling *i.e.* t_4. In the computation of MSE(t_5) in this case, the contribution of h_1 is highly negligible, hence the values of the mean square error is contributed mostly by g_1, the mean square error MSE(t_5) of the usual combined regression estimator under double sampling. Thus in this case the two estimators t_5 and t_5 are equally efficient.

In the present study an attempt is also made to compute t_5 by using simulated data set and also evaluate the mean square error of the proposed estimator to compare with other estimator through relative efficiency. In order to obtain simulated data following the assumptions of the present study, statistical software STATA 8.0(2003) is utilized. For simulation work, the strata population is taken as $N_1 = 35$, $N_2 = 40$, $N_3 = 50$ and $N_4 = 45$. By propoportional allocation, samples are selected (Table 3) with total sample size of $n = 35$. Bivariate normally distributed data is generated for the pair (X_h, Y_h) for $h = 1, 2, 3, 4$ and the corresponding input data for $(\mu_{x_h}, \mu_{y_h}, \sigma_{x_h}, \sigma_{y_h}, \rho)$ are N(80, 150, 17, 16, 0.8), N(75, 140, 15, 17, 0.8), N(70, 130, 16, 16, 0.8) and N(55, 110, 16, 16, 0.8). The stratification was done according to the mean of the value of the study variable Y and the stratum correlation coefficients are assumed to be constant and equal to the population correlation coefficient ρ. The selection of samples within each stratum is done by simple random sampling.

Table 3. Strata sample sizes obtained by proportional allocation for data generated by simulation.

Stratum	N_h	$W_h = (N_h/N)$	n_h
1	35	0.21	7
2	40	0.24	8
3	50	0.29	10
4	45	0.26	9
Total	170	1	35

Source : Data generated by STATA 8.0.

When μ_x is unknown, alternative information about the mean of the auxiliary variable is also obtained by the use of double sampling procedure and for the present data this estimate is given by $\bar{x}_{n'} = 73.69\%$. In the present study, it is considered that there exists partial information about the mean of the auxiliary variable and let us assume that the partial information so obtained given by $\mu_0 = 75.0$.

The stratum means in terms of the transformed values in X and Y are shown in the Table 4 which reveals that $\bar{x}_{st} = -0.1095$, $\bar{y}_{st} = 5.022$ and $\bar{x}_{n'} =$

Table 4. Computation of the strata means $\bar{x}_{st}$ and $\bar{y}_{st}$ for data generated by simulation.

Stratum	W_h	$\bar{x}_h$	$\bar{y}_h$	$W_h\bar{x}_h$	$W_h\bar{y}_h$
1	0.21	1.25	6.52	0.26	1.34
2	0.24	–0.07	5.15	–0.02	1.21
3	0.29	–0.38	4.81	–0.11	1.41
4	0.26	–0.90	3.98	–0.24	1.05
			$\bar{x}_{st}$ = –0.109	$\bar{y}_{st}$ = 5.022	

Source : Data generated by STATA 8.0.

–0.0602 and in terms of the previous origin and scale $\bar{x}_{st} = 72.62$, $\bar{y}_{st} = 132.93$ and $\bar{x}_{n'} = 73.69$.

The preliminary sample is utilized to test the hypothesis

$$H_0 : \mu_x = \mu_0 \text{ against } H_1 : \mu_x \neq \mu_0$$

When a reliable partial information of the mean of the auxiliary variable is available, as in present case, the hypothesis is accepted and the mean μ_0 is used in the estimator t_5. The combined regression preliminary test estimator t_5 in double sampling, the estimate of the mean of Y is given by 135.44. Further it is observed that for $\alpha = 0.01, 0.05, 0.25$, the MSE($t_5$) is smaller than the MSE(t_4) and consequently this increases the efficiency of the proposed estimator. Thus we see that the empirical study also supports the analytical work of the present study that under the stated assumptions the CRPTE in double sampling *i.e.*, t_5 is more efficient than the usual combined regression estimator, when reliable information about the mean of the auxiliary variable is available.

Table 5. Maximum and Minimum Values of $e(\alpha, 0)$, (e^* = maximum, e_0 = minimum)

α^*	e	ρ				ρ			
		0.3	0.5	0.7	0.9	0.3	0.5	0.7	0.9
			***n* = 10**	***n′* = 20**			***n* = 15**	***n′* = 20**	
0.25	e^*	1.01	1.04	1.10	1.24	1.02	1.06	1.13	1.27
	e_0	0.98	0.95	0.89	0.79	0.97	0.93	0.86	0.77
0.05	e^*	1.04	1.11	1.31	1.96	1.05	1.17	1.43	2.22
	e_0	0.94	0.83	0.68	0.50	0.91	0.77	0.62	0.47
0.01	e^*	1.05	1.15	1.42	2.66	1.07	1.22	1.62	3.31
	e_0	0.88	0.71	0.52	0.34	0.84	0.64	0.46	0.32
			***n* = 10**	***n′* = 30**			***n* = 15**	***n′* = 30**	
0.25	e^*	1.01	1.03	1.07	1.20	1.01	1.04	1.10	1.24
	e_0	0.99	0.96	0.92	0.82	0.98	0.95	0.89	0.79
0.05	e^*	1.02	1.08	1.21	1.73	1.04	1.11	1.31	1.96
	e_0	0.96	0.87	0.74	0.54	0.94	0.83	0.68	0.50
0.01	e^*	1.03	1.10	1.29	2.16	1.05	1.15	1.42	2.66
	e_0	0.92	0.78	0.60	0.38	0.88	0.72	0.53	0.35
			***n* = 10**	***n′* = 50**			***n* = 15**	***n′* = 50**	
0.25	e^*	1.01	1.02	1.05	1.15	1.01	1.03	1.07	1.19
	e_0	0.99	0.98	0.95	0.86	0.99	0.97	0.93	0.84
0.05	e^*	1.01	1.05	1.13	1.50	1.02	1.07	1.19	1.68
	e_0	0.97	0.92	0.81	0.60	0.96	0.88	0.75	0.55
0.01	e^*	1.02	1.06	1.17	1.73	1.03	1.09	1.26	2.06
	e_0	0.95	0.85	0.70	0.44	0.93	0.80	0.62	0.62

9. RECOMMENDATIONS FOR SELECTION OF ESTIMATORS

The experimenter usually wants to select an estimator with high efficiency. It would be ideal if the relative efficiency is always larger than unity. However, the relative efficiency exceeds unity in the neighborhood of $\mu_x = 0$, but it decreases below unity as μ_x increases (Fig. 3). To minimize the loss in efficiency, we shall use the criterion for selecting the α value (Han and Bancroft 1968) and we recommend the use of the following criterion:

If the experimenter does not know μ_x and is willing to accept an estimator which has a relative efficiency of no less than e_0, then among the set of estimators with $\alpha \in A$, where $A = \{\alpha : e(\alpha, \mu_x) \geq e_0$ for all $\mu_x\}$, the estimator is chosen to maximize $e(\alpha, \mu_x)$ over all α and μ_x. Since max $e(\alpha, \mu_x) = e(\alpha, 0)$, he selects the $\alpha \in A$ (say α^*) which maximizes $e(\alpha, 0)$ (say e^*). This criterion will guarantee that the relative efficiency of the chosen estimator is at least e_0 and it may become as large as e^*.

The Table 5 gives the values of $e(\alpha, 0)$ for $\rho = 0.3$, 0.5, 0.7, 0.9, and for different sets of values of n and n', the corresponding α^* to use and the maximum relative efficiency e^*. For given n and n', one enters the table and picks up e_0 which is the smallest relative efficiency he wishes to accept. The recommended α^* level is readily chosen. For example if $n = 10$, $n' = 50$, $\rho = 0.7$ and the experimenter wants to have an estimator which has a relative efficiency no less than $e_0 = 0.80$, then he would use $\alpha = 0.05$ because this maximizes $e(\alpha, 0)$ and the maximum relative efficiency he can obtain is 1.13.

REFERENCES

Cochran, W.G. (1977). *Sampling Techniques*, 3rd ed. Wiley and Sons, New York.

Das, G. (2003). A generalized study of preliminary test estimator in double sampling. *Assam Statist. Rev.*, **17,** 127-138.

Das, G. and Bez, K. (1995). Preliminary test estimators in double sampling with two auxiliary variables. *Comm. Statist.- Theory Methods*, **24,** 1211-1226.

Han, C.P. (1973). Double sampling with partial information on auxiliary variables. *J. Amer. Statist. Assoc.*, **68,** 914-918.

Han, C.P. and Bancroft, T.A. (1968). On pooling means when variance is unknown. *J. Amer. Statist. Assoc.*, **63,** 1333-1342.

Kibria, B.M.G. (1996). On preliminary test ridge regression estimators for linear restrictions in a regression model with non-normal disturbances. *Comm. Statist.-Theory Methods*, **25,** 2349-2369.

Reproductive and Child Health Report (2000): Rapid Household Survey. Phase I, 1998. International Institute for Population Sciences (IIPS), Mumbai.

Sisodia, B.V.S. and Srivastava, A.K.(1982). Modifying regression estimators with a preliminary test in double sampling. *Sankhya B*, **44,** 295-303.

APPENDIX

N_h	total number of units in the h^{th} stratum
n_h	number of units in the sample in the h^{th} stratum
y_{hi}	value of the study variable obtained for the i^{th} unit in the h^{th} stratum
$W_h = N_h/N$	stratum weight of the h^{th} stratum
$\sigma^2_{y_h}$	true variance in the h^{th} stratum
$\bar{x}_{n'}$	mean of the auxiliary variable in the preliminary sample of size n'
α	is the level of significance of the preliminary test.