J. Ind. Soc. Agril. Statist.
59(2), 2005 : 118-125

On the Estimation of Finite Population Regression Coefficient

M. Memita Devi, H.V.L. Bathla[1], U.C. Sud[1] and I.C. Sethi[1]
Institute of Nuclear Medicine & Allied Sciences, DRDO, Delhi
(Received : November, 2004)

SUMMARY

A double sampling based estimator of finite population regression coefficient has been developed. The estimator developed performs better than the ordinary least squares (OLS) estimator. Performance of the double sampling based estimator is studied vis-a-vis the ordinary least square estimator under a suitable cost function.

Key words: Double sampling, Finite population, Regression coefficient, Linear cost function.

1. INTRODUCTION

Sample surveys are generally multivariate in nature. Many a times the interest is to establish the pattern of relationship between variables than estimation of simple parameters like means or totals. When the variables are quantitative and knowledge of the subject matter suggests causal relationship then regression analysis may be an appropriate method.

When the object of inference is the regression coefficient there can be two approaches possible i.e. descriptive or analytic. Descriptive inferences can be design based or model based. Design based inferences are based on the distribution generated by random sampling. Relevant details can be found in a standard textbook such as Cochran (1977). On the other hand, in the model based approach models are postulated to represent the population structure. Inferences in this approach are based on the probability distribution specified in the model i.e. the so-called ξ-distribution. Important references in this context are Smith (1978) and Sarndal (1978).

The object of inference in this paper is the finite population regression coefficient. The finite population regression coefficient denoted by 'B' is obtained by minimizing the residual sum of squares over all the 'N' units in the population. When the data in the question is obtained through a sample survey, the ordinary least squares (OLS) approach gives misleading inferences. This is due to the fact that the data is obtained through complex sampling design involving stratification, clustering or units may be selected with varying probability. The assumptions accompanying the OLS approach are therefore unlikely to be satisfied in the context of survey data. Alternative estimators using probability weights are proposed in the literature for estimation of finite population regression coefficient, see for instance Kish and Frankel (1974).

[1] *Indian Agricultural Statistics Research Institute, New Delhi-110012*

2. MAXIMUM LIKELIHOOD AND PROBABILITY WEIGHTED ESTIMATORS

The maximum likelihood estimator of the parameter of superpopulation say β was developed by Demets and Halperin (1977). They assume that information on 'design' variable X_3 is available on all the units of the population. Assuming trivariate normality between the study variable X_1, explanatory variable X_2 and the design variable X_3 the maximum likelihood estimator is given by

$$\hat{\beta}_{12} = \frac{s_{12} + \left(s_{13}s_{23}/s_3^2\right)\left(\hat{\sigma}_3^2/s_3^2 - 1\right)}{s_2^2 + \left(s_{23}^2/s_3^2\right)\left(\hat{\sigma}_3^2/s_3^2 - 1\right)}$$

where $\hat{\sigma}_3^2 = \frac{1}{N}\sum_{\alpha=1}^{N}(x_{3\alpha} - \bar{x}_3)^2$

$$s_{ij} = \frac{1}{n}\sum_{\alpha=1}^{n}(x_{i\alpha} - x_i)(x_{j\alpha} - x_j), s_i^2 = \frac{1}{n}\sum_{\alpha=1}^{n}(x_{i\alpha} - \bar{x}_i)^2$$

$i, j = 1, 2, 3$

Holt *et al.* (1980), through a simulation study, demonstrated improved performance of $\hat{\beta}_{12}$ over $\hat{b}_{12}$ in terms of the criteria of mean squared errors of $\hat{b}_{12}$ and the standard error of $\hat{\beta}_{12}$ where $\hat{b}_{12} = \frac{s_{12}}{s_2^2}$

The π-weighted analogues of the corresponding OLS and MLE, as proposed by Nathan and Holt (1980) are as follows

$$\hat{b}_{12}^* = \frac{s_{12}^*}{s_2^{*2}}$$

and $$\hat{\beta}_{12}^* = \frac{s_{12}^* + \left(s_{13}^* s_{23}^* / s_3^{*2}\right)\left(\hat{\sigma}_3^2 / s_3^{*2} - 1\right)}{s_2^{*2} + \left(s_{23}^{*2} / s_3^{*2}\right)\left(\hat{\sigma}_3^2 / s_3^{*2} - 1\right)}$$

with $$s_{ij}^* = \sum_{\alpha=1}^{n}\frac{x_{i\alpha}x_{j\alpha}}{N\pi_\alpha} - \frac{\bar{x}_i^*\bar{x}_j^*}{\sum_{\alpha=1}^{n}\frac{1}{N\pi_\alpha}}, \; s_i^{*2} = s_{ii}^* \text{ and}$$

$$\bar{x}_i^* = \sum_{\alpha=1}^{n}\frac{x_{i\alpha}}{N\pi_\alpha}$$

It may be noted that both $\hat{b}_{12}^*$ and $\hat{\beta}_{12}^*$ are asymptotically unbiased.

The variances of $\hat{b}_{12}^*$ and $\hat{\beta}_{12}^*$ can be obtained by the Taylor series linearization approach. Estimates of variances of the estimators can be obtained by substituting the sampled analogues of the population parameters in the variance expressions.

Estimates of variance of $\hat{b}_{12}^*$ and $\hat{\beta}_{12}^*$ in case of SRSWOR $\left(\pi_\alpha = \frac{n}{N}\right)$ are given by

$$\hat{V}\left(\hat{b}_{12}^*\right) = \frac{\left(\frac{1}{n} - \frac{1}{N}\right)}{\left(s_2^2\right)^2}\frac{1}{n-1}\sum_{\alpha=1}^{n}\left[(x_{2\alpha} - \bar{x}_2)\right.$$

$$\left.\left\{(x_{1\alpha} - \bar{x}_1) - \frac{s_{12}}{s_2^2}(x_{2\alpha} - \bar{x}_2)\right\}\right]^2$$

$$\hat{V}(\hat{\beta}_{12}^*) = \left(\frac{s_1}{s_2}\right)^2\left(\frac{1}{n} - \frac{1}{N}\right)\frac{1}{n-1}\sum_{\alpha=1}^{N}(e_{1\alpha} - \bar{e}_1)^2$$

where

$$e_{1\alpha} = \left[\left(\frac{x_{1\alpha} - \bar{x}_1}{s_1}\right) - \frac{s_{12}}{s_1 s_2}\left(\frac{x_{2\alpha} - \bar{x}_2}{s_2}\right)\right.$$

$$\left. - \frac{s_{23}}{s_2 s_3}\left\{\frac{s_{23}}{s_2 s_3} - \frac{s_{12}}{s_1 s_2}\frac{s_{23}}{s_2 s_3}\right\}\left(\frac{x_{3\alpha} - \bar{x}_3}{s_3}\right)^2\right]$$

For stratified sampling, $\pi_\alpha = \frac{n_h}{N_h}$, h = 1, 2, ... k (assuming that there are 'k' strata). An estimator of variance of $\hat{b}_{12}^*$ say, $\hat{b}_{12st}^*$ in case of stratified sampling is given by

$$\hat{V}\left(\hat{b}_{12st}^*\right) = \frac{1}{\left(\sum_{h=1}^{k}\frac{N_h}{N}s_{2h}^2\right)^2}\sum_{h}^{k}\frac{N_h^2}{N^2}\left(\frac{1}{n_h} - \frac{1}{N_h}\right)^2$$

$$\frac{1}{n_h - 1}\sum_{i=1}^{n_h}\left[(x_{2h\alpha} - \bar{x}_{2h})\right.$$

$$\left.\left\{(x_{1h\alpha} - \bar{x}_{1h}) - \frac{\sum_{h=1}^{k}\frac{N_h}{N}s_{12h}}{\sum_{h=1}^{k}\frac{N_h}{N}s_{2h}^2}(x_{2h\alpha} - \bar{x}_{2h})\right\}\right]$$

where

$$s_{ijh} = \frac{1}{n_h - 1}\sum_{\alpha=1}^{n_h}(x_{ih\alpha} - \bar{x}_{ih})(x_{jh\alpha} - \bar{x}_{jh}),\ s_{ih}^2 = s_{iih}$$

$$\bar{x}_{ih} = \frac{1}{n_h}\sum_{\alpha=1}^{n_h} x_{ih\alpha},\ i, j = 1, 2, 3$$

Similarly, an estimator of variance of $\hat{\beta}_{12}^*$ say, $V(\hat{\beta}_{12st}^*)$ in case of stratified sampling is given by

$$\hat{V}\left(\hat{\beta}_{12st}^*\right) = \frac{1}{\left(\sum_{h=1}^{k}\frac{N_h}{N}s_{2h}^2\right)^2}\sum_{h=1}^{k}\left(\frac{N_h}{N}\right)^2\left(\frac{1}{n_h} - \frac{1}{N_h}\right)\left(\frac{1}{n_h - 1}\right)\sum_{\alpha=1}^{n_h}(e_{1h\alpha} - \bar{e}_{1h})^2$$

where

$$e_{1h\alpha} = \Bigg[(x_{1hi\alpha} - \bar{x}_{1h})(x_{2h\alpha} - \bar{x}_{2h}) - \frac{\sum_{h=1}^{k}\frac{N_h}{N}s_{12h}}{\sum_{h=1}^{k}\frac{N_h}{N}s_{2h}^2}(x_{2h\alpha} - \bar{x}_{2h}) - \frac{\sum_{h=1}^{k}\frac{N_h}{N}s_{13h}\sum_{h=1}^{k}\frac{N_h}{N}s_{23h}}{\left(\sum_{h=1}^{k}\frac{N_h}{N}s_{3h}^2\right)^2}(x_{3h\alpha} - \bar{x}_{3h})^2 + \frac{\sum_{h=1}^{k}\frac{N_h}{N}s_{12h}\left(\sum_{h=1}^{k}\frac{N_h}{N}s_{23h}\right)^2}{\left(\sum_{h=1}^{k}\frac{N_h}{N}s_{2h}^2\right)^2\left(\sum_{h=1}^{k}\frac{N_h}{N}s_{3h}^2\right)^2}(x_{3h\alpha} - \bar{x}_{3h})^2\Bigg]$$

$$\bar{e}_{1h} = \frac{1}{n_h}\sum_{\alpha=1}^{n_h} e_{1h\alpha}$$

2.1 A Simulation Study

Theoretical comparison of variances of $\hat{b}_{12}^*$ and $\hat{\beta}_{12}^*$ is not possible due to complex expressions of variances of the estimators. Accordingly, a simulation study was carried out for comparison purpose. Accordingly, a multivariate normal population of size 1000 was generated using algorithm of Ahrnes and Deiter (1972) and multivariate normal generator of Scheuer and Stoller (1962).

The initial parameters for this simulation study were based on those of, a data of Chiru district of Rajasthan (District Handbook of Census). The variables chosen were total area (in ha), unirrigated area (in ha), and cultivated waste in different villages. The cultivated waste was used as the design variable due to the fact that Rajasthan has mostly unirrigated area.

Table 2.1. Correlation between different characters

	X1	X2	X3
X1	1	0.988	0.6095
X2	0.988	1	0.4921
X3	0.6095	0.478	1

(Chiru Data)

	X1	X2	X3
X1	1	0.991	0.587
X2	0.991	1	0.5018
X3	0.587	0.5018	1

(Simulated Data)

The empirical study was based on, 100 independent samples of size 100 each, drawn from the finite population of size 1000. The following, survey designs were used.

A: Simple random sampling

B: Stratified sampling with proportional allocation

C: Stratified sampling with equal allocation

D: Stratified sampling with increasing allocation

E: Stratified sampling with U-shape allocation . (In a U-shaped allocation the number of selected

units from the various strata first increase and then decrease)

For the case of stratified sampling, the above generated population, was stratified into five strata, on the basis of the design variable.

Table 2.2. Sample sizes drawn from each strata with different sampling designs

Sampling design	Strata size				
	$N_1 = 150$	$N_2 = 20$	$N_3 = 350$	$N_4 = 251$	$N_5 = 44$
A	7	30	40	20	3
B	20	20	20	20	20
C	16	18	20	22	24
D	22	20	16	20	22
E	18	21	22	21	18

Table 2.3. Variances of $\hat{b}^{*}_{12}$ and $\hat{\beta}^{*}_{12}$ under different survey designs

Survey design	$V(\hat{b}^{*}_{12})$	$V(\hat{\beta}^{*}_{12})$
A	0.28144	0.001872
B	0.26933	0.002902
C	0.29387	0.002902
D	0.30589	0.003179
E	0.23225	0.002666

From the results of the empirical as given in Table 2.3 it can be seen that $\hat{\beta}^{*}_{12}$ performs better than $\hat{b}^{*}_{12}$ in terms of standard errors of these estimators.

3. DOUBLE SAMPLING BASED MAXIMUM LIKELIHOOD ESTIMATOR

The improved performance of $\hat{\beta}^{*}_{12}$ over $\hat{b}^{*}_{12}$ provides scope for using a double sampling based estimator, wherein $\hat{\sigma}^{2}_{3}$ is replaced by s'^{2}_{3} estimated on the basis of a large sample of size n'.

Accordingly, we propose the double sampling based π – weighted maximum likelihood estimator as double sampling based maximum likelihood estimator of regression coefficient, which is given by

$$\hat{\beta}'_{12d} = \frac{\left[\sum_{h=1}^{k}\frac{n'_h}{n'}\left[\frac{1}{n_h}\sum_{\alpha=1}^{n_h}x_{1h\alpha}x_{2h\alpha} - \bar{x}_{1h}\bar{x}_{2h}\right] + \frac{\sum_{h=1}^{k}\frac{n'_h}{n'}\left[\frac{1}{n_h}\sum_{\alpha=1}^{n_h}x_{1h\alpha}x_{3h\alpha} - \bar{x}_{1h}\bar{x}_{3h}\right]}{\sum_{h=1}^{k}\frac{n'_h}{n'}\left[\frac{1}{n_h}\sum_{\alpha=1}^{n_h}x^2_{3h\alpha} - \bar{x}^2_{3h}\right]}\right]}{\left[\sum_{h=1}^{k}\frac{n'_h}{n'}\left[\frac{1}{n_h}\sum_{\alpha=1}^{n_h}x^2_{2h\alpha} - \bar{x}^2_{2h}\right] + \frac{\sum_{h=1}^{k}\frac{n'_h}{n'}\left[\frac{1}{n_h}\sum_{\alpha=1}^{n_h}x_{2h\alpha}x_{3h\alpha} - \bar{x}_{2h}\bar{x}_{3h}\right]}{\sum_{h=1}^{k}\frac{n'_h}{n'}\left[\frac{1}{n_h}\sum_{\alpha=1}^{n_h}x^2_{3h\alpha} - \bar{x}^2_{3h}\right]}\right]}$$

$$\left[\sum_{h=1}^{k}\frac{n'_h}{n'}\left[\frac{1}{n_h}\sum_{\alpha=1}^{n_h}x_{2h\alpha}x_{3h\alpha} - \bar{x}_{2h}\bar{x}_{3h}\right]\left[\frac{\sum_{h=1}^{k}\frac{n'_h}{n'}\left[\frac{1}{n'_h}\sum_{\alpha=1}^{n'_h}x^2_{3h\alpha} - \bar{x}'^2_{3h}\right]}{\sum_{h'=1}^{n'_h}\frac{n'_h}{n'}\left[\frac{1}{n_h}\sum_{\alpha=1}^{n_h}x^2_{3h\alpha} - \bar{x}^2_{3h}\right]} - 1\right]\right]$$

$$\left[\frac{\sum_{h=1}^{k}\frac{n'_h}{n'}\left[\frac{1}{n'_h}\sum_{\alpha=1}^{n'_h}x^2_{3h\alpha} - \bar{x}'^2_{3h}\right]}{\sum_{h'=1}^{n'_h}\frac{n'_h}{n'}\left[\frac{1}{n_h}\sum_{\alpha=1}^{n_h}x^2_{3h\alpha} - \bar{x}^2_{3h}\right]} - 1\right]$$

It can be seen that to the first order of approximation $\hat{\beta}'_{12d}$ asymptotically unbiased

$$E_1E_2E_3\left(\hat{\beta}'_{12d}\right)=\beta_{12}$$

when E_2 and E_3 are conditional expectation given $\frac{n'_h}{n'}$ is fixed.

The variance of $\hat{\beta}'_{12d}$ to the first order of approximation is given by

$$V\left(\hat{\beta}'_{12d}\right)=\frac{A}{n'}-\frac{B}{n'}+\sum_h^n c_h/n_h$$

where

$$A=\sum_{h=1}^{k}\frac{N_h}{N}\frac{1}{N_h-1}\sum_{\alpha=1}^{N_h}\left(E'_{h\alpha}-\bar{E}'_h\right)^2$$

$$B=\sum_{h=1}^{k}\frac{N_h}{N}\frac{1}{N_h-1}\sum_{\alpha=1}^{N_h}\left(e'_{h\alpha}-\bar{e}'_h\right)^2$$

$$c_h=\sum_{h=1}^{k}\left(\frac{N_h}{N}\right)^2\frac{1}{N_h-1}\sum_{\alpha=1}^{N_h}\left(e'_{h\alpha}-\bar{e}'_h\right)$$

$$E'_{h\alpha}=\left(X_{1h\alpha}-\bar{X}''_1\right)\left(X_{2h\alpha}-\bar{X}''_2\right)-\frac{\sum_{h=1}^{k}\frac{N_h}{N}S''_{12h}}{\sum_{h=1}^{k}\frac{N_h}{N}S''^2_{2h}}\left(X_{2h\alpha}-\bar{X}''_2\right)^2$$

$$e_{h\alpha}=\left(X_{1h\alpha}-\bar{X}''_1\right)\left(X_{2h\alpha}-\bar{X}''_2\right)-\frac{\sum_{h=1}^{k}\frac{N_h}{N}S''_{12h}}{\sum_{h=1}^{k}\frac{N_h}{N}S''^2_{2h}}\left(X_{2h\alpha}-\bar{X}''_2\right)^2$$

$$-\frac{\sum_{h=1}^{k}\frac{N_h}{N}S''_{13h}\sum_{h=1}^{k}\frac{N_h}{N}S''_{23h}}{\left(\sum_{h=1}^{k}\frac{N_h}{N}S''^2_{3h}\right)^2}\left(X_{3h\alpha}-\bar{X}''_3\right)^2$$

$$+\frac{\sum_{h=1}^{k}\frac{N_h}{N}S''_{12h}\left(\sum_{h=1}^{k}\frac{N_h}{N}S''_{23h}\right)^2}{\left(\sum_{h=1}^{k}\frac{N_h}{N}S''^2_{2h}\right)\left(\sum_{h=1}^{k}\frac{N_h}{N}S''^2_{3h}\right)^2}\left(X_{3h\alpha}-\bar{X}''_3\right)^2$$

with

$$S''_{ijh}=\frac{1}{N_h}\sum_{\alpha=1}^{N_h}\left(X_{ih\alpha}-\bar{X}''_i\right)\left(X_{jh\alpha}-\bar{X}''_j\right),\ S''^2_{ih}=S''_{iih}$$

$$\bar{E}'_h=\frac{1}{N_h}\sum_{\alpha=1}^{N_h}E'_{h\alpha},\ \bar{e}'_h=\frac{1}{N_h}\sum_{\alpha=1}^{N_h}e'_{h\alpha}$$

$$\bar{X}''_i=\frac{1}{N}\sum_{h=1}^{k}N_h\sum_{h=1}^{k}X_{ih\alpha},\ i,j=1,2,3$$

An estimate of variance of double sampling based MLE of regression co-efficient is given as

$$\hat{V}\left(\hat{\beta}'_{12d}\right)=\frac{1}{n'\left(\sum_{h=1}^{k}N_hs'^2_{2h}\right)^2}\sum_{h=1}^{k}\frac{n'^2_h}{n'^2}\frac{1}{(n'_h-1)}\sum_{\alpha=1}^{n'_h}\left(e_{2h\alpha}-\bar{e}_{2h}\right)^2+\frac{1}{n'}\sum_{h=1}^{k}\frac{n_h/n}{\left(\sum_{h=1}^{k}\frac{n_h}{n}s^2_{2h}\right)^2}\frac{1}{n'_h-1}\sum_{\alpha=1}^{n_h}\left(E_{h\alpha}-\bar{E}_h\right)^2$$

$$-\sum_{h=1}^{k}\frac{N_h/N^2}{\left(\sum_{h=1}^{k}\frac{n_h}{n}s^2_{2h}\right)^2}\frac{1}{n_h-1}\sum_{\alpha=1}^{n_h}\left(E_{h\alpha}-\bar{E}_h\right)^2$$

where

$$e_{2h_\alpha} = (x_{1h\alpha} - \bar{x}_1')(x_{2h\alpha} - \bar{x}_2') - \frac{\sum_{h=1}^{k} \frac{n_h'}{n'} s_{12h}}{\sum_{h=1}^{k} \frac{n_h'}{n'} s_{2h}^2}(x_{2h\alpha} - \bar{x}_2')^2$$

$$- \frac{\sum_{h=1}^{k} \frac{n_h'}{n'} s_{13h} \sum_{h=1}^{k} \frac{n_h'}{n'} s_{23h}}{\left(\sum_{h=1}^{k} \frac{n_h'}{n'} s_{3h}^2\right)^2}(x_{3h\alpha} - \bar{x}_3')^2$$

$$+ \frac{\sum_{h=1}^{k} \frac{n_h'}{n'} s_{12h} \left(\sum_{h=1}^{k} \frac{n_h'}{n'} s_{23h}\right)^2}{\left(\sum_{h=1}^{k} \frac{n_h'}{n'} s_{2h}^2\right)\left(\sum_{h=1}^{k} \frac{n_h'}{n'} s_{3h}^2\right)^2}(x_{3h\alpha} - \bar{x}_3')^2$$

and

$$\bar{x}_i' = \frac{1}{n'} \sum_{h=1}^{k} \frac{n_h'}{n_h} \sum_{\alpha=1}^{n_h} x_{ih\alpha}$$

$$s_{ijh} = \frac{1}{n_h} \sum_{\alpha=1}^{n_h} (x_{ih\alpha} - \bar{x}_i')(x_{jh\alpha} - \bar{x}_j')$$

3.1 Comparison of $\hat{\beta}_{12d}$ and $\hat{b}_{12}^*$

From the variance expression derived for the double sampling based MLE it can be seen that the theoretical comparison between the double sampling based MLE and the OLS estimator is not possible. To empirically compare the two estimators a multivariate normal population of size 1000 was generated using the algorithm of Ahrens and Deiter (1972) as modeled for multivariate normal population by Scheuer and Stoller (1962).

For the purpose of empirical comparison 100 independent samples of size 100 each were drawn from the population. These samples were drawn for different sampling designs with values of n_h' and n_h are presented in Table 3.1.

It can be seen from the table that π-weighted double sampling based MLE($\hat{\beta}_{12d}'$) estimator scores over the π-weighted OLS estimator (b_{12}^*) in terms of the variance criterion for all the sampling design considered. Based on the simulation study we can say that

$$V\left(\hat{\beta}_{12d}'\right) < V\left(\hat{b}_{12}^*\right).$$

Table 3.1 Sample sizes drawn from each strata by double sampling with different sampling designs

Sampling design		Strata size $N_1 = 150$	$N_2 = 205$	$N_3 = 350$	$N_4 = 251$	$N_5 = 44$
A	n_h'	21	90	120	60	9
	n_h	7	30	40	20	3
B	n_h'	60	60	93	60	27
	n_h	20	20	20	20	20
C	n_h'	48	54	105	66	27
	n_h	16	18	20	22	24
D	n_h'	66	60	87	60	27
	n_h	22	20	16	20	22
E	n_h'	54	63	93	63	27
	n_h	18	21	22	21	18

Table 3.2 Variance of β_{12d}' and $\hat{b}_{12}^*$

Survey design	$V(\hat{b}_{12}^*)$	$V(\hat{\beta}_{12d}^*)$
A	0.28144	0.06460
B	0.26933	0.05545
C	0.29387	0.05603
D	0.30589	0.05359
E	0.23225	0.05462

3.2 Efficiency Comparison of $\hat{\beta}_{12d}$ over $\hat{b}_{12}^*$

To examine if the gain due to double sampling is worth the extra expenditure required for collecting the information on design variable we make efficiency comparison of maximum likelihood estimator vis-à-vis OLS estimator under a suitable cost function.

For this purpose we consider the cost function

$$C_0 = c_1 n' + \sum_{h=1}^{k} c_h n_h$$

where c_1 is the cost per unit for collecting information on the design variable (x_3) in the first phase sample and c_h is the cost per unit for collecting information on X_1 and X_2 in the h^{th} stratum of second phase sample and $h = 1, 2, \ldots, k$.

The optimum values of sample sizes on minimization of $V\left(\hat{\beta}'_{12d}\right)$ by fixing the cost are given by

$$n' = \frac{C_0}{\sqrt{c_1}\sqrt{A-B} + \sqrt{c_2}\sum_{h=1}^{k}\sqrt{c_h}}\sqrt{\frac{A-B}{c_1}}$$

$$n_h = \frac{C_0}{\sqrt{c_1}\sqrt{A-B} + \sqrt{c_2}\sum_{h=1}^{k}\sqrt{c_h}}\sqrt{\frac{c_h}{c_2}}$$

Substituting the optimum values of n' and n_h in the expression for variance of $\hat{\beta}'_{12d}$, we obtain the optimum variance of $\hat{\beta}'_{12d}$ as

$$V\left(\hat{\beta}'_{12}\right)_{opt} = \frac{1}{C_0}\left[\sqrt{c_1}\sqrt{A-B} + \sqrt{c_2}\sum_{h=1}^{k}\sqrt{c_h}\right]^2$$

In case of OLS estimator the appropriate cost function is given by

$$C_0 = c_2 n$$

$$\Rightarrow n_{opt} = C_0/c_2 = n_1$$

Then the resultant optimum variance of the OLS estimator becomes

$$V\left(\hat{b}^*_{12}\right)_{opt} = \frac{c_2}{C_o}\frac{1}{\left(S_2^2\right)^2}\frac{1}{N-1}\sum_{\alpha=1}^{N}\left(X_{2\alpha}-\bar{X}_2\right)^2 \left[\left(X_{1\alpha}-\bar{X}_1\right)-\frac{S_{12}}{S_2^2}\left(X_{2\alpha}-\bar{X}_2\right)\right]^2$$

The relative efficiency of $\hat{\beta}'_{12d}$ vis-à-vis $\hat{b}^*_{12}$ is, therefore given by

$$\text{R.E.} = \frac{c_2\frac{1}{\left(S_2^2\right)^2}\frac{1}{N-1}\sum_{\alpha=1}^{N}\left(X_{2\alpha}-\bar{X}_2\right)^2 \left[\left(X_{1\alpha}-\bar{X}_1\right)-\frac{S_{12}}{S_2^2}\left(X_{2\alpha}-\bar{X}_2\right)\right]^2}{\left[\sqrt{c_1}\sqrt{A-B} + \sqrt{c_2}\sum_{h=1}^{k}\sqrt{c_h}\right]^2}$$

3.3 A Cost Study

To work out the relative efficiency in numerical terms, we take following values for C_0, c_1 and c_2 (by assuming $c_h = c_2$ i.e. cost per unit in each stratum of the second phase sample as equal)

$C_0 = 4500$, $c_1 = 15$, $c_2 = 20$

$c_1 = 12$, $c_2 = 20$

$c_1 = 10$, $c_2 = 20$

$c_1 = 8$, $c_2 = 20$

From the table we can see that the relative efficiency values are greater than 1 for different combinations of c_1 and c_2 suggesting there by that the MLE based regression estimator is an improvement over the OLS estimator. Also we can see that as the ratio c_1/c_2 decreases relative efficiency of $\hat{\beta}'_{12}$ over $\hat{b}^*_{12}$ increases. In other words as the cost of collecting the information on design variable X_3 becomes cheaper the relative efficiency of the proposed estimator increases.

Table 3.3 Relative efficiency of the double sampling based MLE over the OLS estimator for different values of c_1 and c_2

c_1	c_2	c_1/c_2	n'_1	n'_2	n'_3	n'_4	n'_5	n_1	n_2	n_3	n_4	n_5	R.E.
15	20	0.75	16	72	104	48	16	2	9	13	6	2	1.90
12	20	0.6	16	75	111	52	8	4	19	28	13	2	1.99
10	20	0.5	18	77	112	53	6	6	26	38	18	2	2.06
8	20	0.4	19	78	114	54	7	8	33	48	23	3	2.15

REFERENCES

Ahrnes, J.H. and Deiter, U. (1972). Computer methods for sampling from the exponential and normal distributions. *Communications of the ACM,* **15**, 873-882.

Cochran, W.G. (1977). *Sampling Techniques*. Third Edition, Wiley & Sons, New York, London.

Demets, D. and Halperin, M. (1977). Estimation of a single regression coefficient in samples arising from a sub-sampling procedure. *Biometrics,* **33**, 47-56.

Holt, D., Smith, T.M.F. and Winter, P.D. (1980). Regression analysis of data from complex surveys. *J. Roy. Statist. Soc.*, **A143**, 474-487.

Kish, L. and Frankel, M.R. (1974). Inference from complex samples. *J. Roy. Statist. Soc.*, **B36**, 1-37.

Nathan, G. and Holt, D. (1980). The effect of survey design on regression analysis. *J. Roy. Statist. Soc.*, **B42**, 377-386.

Sarndal, C.E. (1978). Design based and model based inference in survey sampling. *Scan. J. Statist.*, **5**, 27-52.

Scheuer, E.M. and Stoller, D.S. (1962). On the generation of normal random vectors. *Technometrics*, **4**, 278-281.

Smith,T.M.F. (1978). A model building approach to survey analysis. European Meeting of Statisticians, Oslo.