Chapter 4

Relative efficiency and optimum allocation of the CRPTE in double sampling with partial information on the auxiliary variable.
4.1 Introduction

The study of the mean square error of an estimator will not be completed unless it is compared with other estimators. Without the use of real data, precision between estimators can be compared by analytical derivation and conclusion can be drawn from an inequality showing mean square expression of the two estimators on both sides. Another useful criteria for the comparison of the precision of any estimator is the relative efficiency defined 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 chapter, the mean square error of the proposed combined regression preliminary test estimator (CRPTE) in double sampling is compared with other estimator and conclusion is drawn through the relative efficiency. The relative efficiency of the proposed estimator is also simulated for different sample sizes, correlation and level of significance and the values are plotted for depicting its behaviour corresponding to various values of the mean of the auxiliary variable.

4.2 Relative efficiency of the CRPTE

The Combined regression estimator under double sampling is given by

\[ t_4 = \bar{y}_n + b_{yx} (\bar{x}_n - \bar{x}_n) \]

where \( \bar{y}_n = \sum \frac{W_k}{n} \bar{y}_h \) and \( \bar{x}_n = \sum \frac{W_k}{n} \bar{x}_h \)
The $\text{MSE}(t_4)$ can be derived as follows,

Let

$$
\overline{y}_u = \mu_y + \varepsilon_1, \quad \overline{x}_u = \mu_x + \varepsilon_2, \quad \overline{x}_s = \mu_x + \varepsilon_3, \quad
$$

$$
b_{ys} = \frac{S_{ys}}{S_x^2}, \quad S_{ys} = S_{ys} + \varepsilon_4, \quad S_x^2 = S_x^2 + \varepsilon_5$

where $\varepsilon_i$'s are errors such that $E(\varepsilon_i) = 0$ for $i = 1, 2, 3, 4$ and $5$ and $S_x^2$, $S_y^2$ and $S_{ys}$ are population variances and covariance respectively.

Now,

$$
\text{MSE}(t_4) = E\{(t_4 - E(t_4))^2\}
$$

$$
= E\{\overline{y}_u + b_{ys}(\overline{x}_s - \overline{x}_u) - \mu_y\}^2
$$

$$
= E\{(\mu_y + \varepsilon_1) + (s_{ys} / s_x^2)(\varepsilon_3 - \varepsilon_2) - \mu_y\}^2
$$

$$
= E[\varepsilon_1 + ((S_{ys} + \varepsilon_4)/(S_x^2 + \varepsilon_5))(\varepsilon_3 - \varepsilon_2)]^2
$$

Assuming

$$
|\varepsilon_4 / S_x^2| < 1
$$

we get,

$$
\text{MSE}(t_4) = E[\varepsilon_1 + (S_{ys} / S_x^2)[1 + (\varepsilon_4 / S_{ys})\{- (\varepsilon_3 / S_y^2) + \ldots \} (\varepsilon_3 - \varepsilon_2)]^2
$$

$$
= E[\varepsilon_1 + \{ (S_{ys} / S_x^2)[1 + (\varepsilon_4 / S_{ys})\{- (\varepsilon_3 / S_y^2) - \varepsilon_2 + (\varepsilon_4 / S_y^2)]\}^2
$$

$$
= E[\varepsilon_1 + (S_{ys} / S_x^2)\{- (\varepsilon_3 / S_y^2) - \varepsilon_2 + (\varepsilon_4 / S_y^2) + (\varepsilon_4 / S_{ys}) - (\varepsilon_4 / S_{ys})\}^2
$$

which upto second order of approximation, gives

$$
\text{MSE}(t_4) = E[\varepsilon_1 + (S_{ys} / S_x^2)(\varepsilon_3 - \varepsilon_2)]^2
$$
$$= E[\varepsilon_1^2 + (S_{ys}^2 / S_s^4)(\varepsilon_3^2 - 2\varepsilon_3\varepsilon_2 + \varepsilon_2^2) + 2(S_{ys} / S_s^2)(\varepsilon_1\varepsilon_3 - \varepsilon_1\varepsilon_2)]$$

$$= \text{Var} (\bar{y}_u) + (S_{ys}^2 / S_s^4)\text{Var} (\bar{x}_u) + (S_{ys} / S_s^2)\text{Var} (\bar{x}_u')$$

$$- 2(S_{ys}^2 / S_s^4)\text{Cov} (\bar{x}_u', \bar{x}_u) + 2(S_{ys} / S_s^2)\{\text{Cov} (\bar{x}_u', \bar{y}_u) - \text{Cov} (\bar{y}_u', \bar{x}_u')\}$$

Now, the variance of $\bar{x}_u$ and $\bar{y}_u$ is given by

$$\text{Var}(\bar{y}_u) = \sum(W_h^2 S_{ys}^2 / n_h)(1 - f_h) \quad \text{Var}(\bar{x}_u) = \sum(W_h^2 S_{xs}^2 / n_h)(1 - f_h)$$

(Cochran, 1977)

(Under the consideration that the sampling fraction $n_h / N_h$ are negligible.)

$s_{y_u}^2$ and $s_{x_u}^2$ are variance of $Y$ and $X$ in the $h^{th}$ stratum.

Also the variance of $\bar{x}_u'$ is given by

$$\text{Var}(\bar{x}_u') = \{(1 / n') - (1 / N)\} s_x^2$$

$$= s_x^2 / n'$$

(Cochran, 1977)

(for large population size $N$, $1 / N$ is negligible)

The pair $(X, Y)$ is considered to be a bivariate random variable with mean $(\mu_x, \mu_y)$ and covariance matrix $\sum_{(\alpha, \beta)}$, in which the variances are denoted by $S_x^2$ and $S_y^2$ and the correlation coefficient by $\rho$. The regression estimator depends on whether the covariance matrix is known or not. If $\sum_{(\alpha, \beta)}$ is known, we may let $S_x^2 = S_y^2 = 1$ without the loss of generality. The strata population $(X_u, Y_u)$ can also be considered as a bivariate random variable for every $h$, with mean $(\mu_{x_h}, \mu_{y_h})$. If the covariance matrix $\sum_{(x_h, y_h)}$ of the pair $(X_h, Y_h)$ is known, we can let $S_{x_h}^2 = S_{y_h}^2 = 1$ (WLOG).
When the samples are selected with proportional allocation then the strata weights are given by

\[ W_h = \frac{N_h}{N} = \left( \frac{n_h}{n} \right) \]  
(Cochran, 1977)

Thus \( \sum_h W_h^2/n_h = \sum_h W_h^2/nW_h = (1/n) \sum_h W_h = (1/n) \) (as \( \sum W_h = 1 \))

Hence the covariance matrix of \( \left( \bar{x}_{n'}, \bar{x}_n, \bar{y}_n \right) \) reduces to

\[
\Sigma = \begin{pmatrix}
\text{Var} (\bar{x}_{n'}) & \text{Cov}(\bar{x}_{n'}, \bar{x}_n) & \text{Cov}(\bar{x}_{n'}, \bar{y}_n) \\
\text{Cov}(\bar{x}_n, \bar{x}_{n'}) & \text{Var} (\bar{x}_n) & \text{Cov}(\bar{x}_n, \bar{y}_n) \\
\text{Cov}(\bar{y}_n, \bar{x}_{n'}) & \text{Cov}(\bar{y}_n, \bar{x}_n) & \text{Var} (\bar{y}_n)
\end{pmatrix}
\]

i.e., \( \sum = \begin{pmatrix} 1/n' & 1/n' & \rho/n' \\
1/n' & 1/n & \rho/n \\
\rho/n' & \rho/n & 1/n \end{pmatrix} \)

Hence,

\[
\text{MSE}(t_4) = \frac{1}{n} + \frac{\rho^2}{n} + \frac{\rho^2}{n'} - 2(\rho^2/n') + 2\rho(\rho/n' - \rho/n)
= \frac{1}{n}(1 - \rho^2) + \frac{1}{n'} \rho^2 \quad \text{(4.1)}
\]

Now, \( \text{MSE}(t_5) \) is given by (3.9) as follows

\[
\text{MSE}(t_5) = \{(1 - \rho^2)/n + \rho^2/n'\} + (\rho^2/n')\{A \varphi(A) - B \varphi(B)\}
- \rho^2(1/n' - \mu_x^2)\{\Phi(A) - \Phi(B)\}
\]

or \( \text{MSE}(t_5) = g_1 + h_1 \), where

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

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

Therefore, the relative efficiency of \( t_5 \) to \( t_4 \) is given by

\[
e_1 = \left[ \text{MSE} (t_4) \right] / \left[ \text{MSE} (t_5) \right] = g_1 / (g_1 + h_1)
\]
The behavior of relative efficiency can be observed for different values of $\mu_x$. As $e_1$ is symmetric about $\mu_x = 0$, we need to consider $e_1$ only for $\mu_x \geq 0$. The above behavior can be studied for selected values of the level of significance $\alpha$ and correlation coefficient $\rho$ respectively and by fixing some hypothetical values for $n$ and $n'$ (Table 4.1, Table 4.2; Fig 4.1, 4.2).

4.3 Discussion

In order to get an idea about the behavior of the relative efficiency function of $t$, with respect to $t_x$ for different values of $\mu_x$, $e_1$ can be computed for a set of values of $n$, $n'$, $\alpha$ and $\rho$. Table 4.1, 4.2 and Figure 4.1, 4.2 show that in general that $e_1$ has a maximum at $\mu_x = 0$. As $\mu_x$ increases, $e_1$ decreases to a minimum and then again increases to unity and remains constant thereafter. It is found that $e_1$ is very close to 1 at $\mu_x = 1$.

The Figures clearly show that when the mean of the auxiliary variable is close to the hypothetical value, then relative efficiency is maximum. Also as $\mu_x$ moves away from the hypothetical value the relative efficiency decreases, but after attaining minimum again increases to unity and remains constant thereafter.

Relative efficiency is high when $\mu_x$ is close to the hypothetical value, i.e., when the hypothesis considered in the study is likely to be accepted and as a result the mean of the partial information of $X$ is used in the proposed estimator. Thus the proposed estimator reduces to the usual combined regression estimator which as we know has efficiency higher than the combined regression
estimator under double sampling. On the other hand when \( \mu_x \) is far away from the hypothetical value of the mean of the auxiliary variable, then the hypothesis is likely to be rejected. In this case the mean of \( X \) from the preliminary sample is used for the proposed estimator, so that the estimator \( t_x \) reduces to the combined regression estimator under double sampling which is same as \( t_x \) and hence \( e_t \) is equals to unity for values of \( \mu_x \) far away from the hypothetical value of \( X \). For the intermediate values of \( \mu_x \), i.e., under situations when one is not certain as whether to accept or reject the hypothesis, then the CRPTE has a lesser efficiency as compared to the combined regression estimator under double sampling. This establishes the utility of the present study of preliminary test estimator using reliable partial information on the auxiliary variable.

4.4 Optimum allocation

In planning of a sample survey, a stage is always reached at which a decision must be made about the size of the sample. This decision is important. Too large a sample involves utilization of more time and resources and too small a sample diminishes the precision of the results. 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 the sampling scheme of the proposed estimator, the samples are extracted by double sampling in which the first sample is a stratified simple
random sample of size $n$ in which the pair $(x_{sh}, y_{sh})$ values are measured from $n_h$ units drawn from each stratum and consequently estimating the pair $\left(\bar{x}_{sh}, \bar{y}_{sh}\right)$, 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 $x_{sh}$ is measured and evaluates $\bar{x}_{n'}$ which is utilized to estimate $\mu_x$.

In order to obtain optimum allocation of sample sizes for the suggested estimator, let us consider simple linear cost function $C$ given by

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

(4.2)

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.

The CRPTE in double sampling constructed in the present study is given by

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

where $\bar{y}_{sh} = \sum_{h} W_h \bar{y}_h$ and $\bar{x}_{sh} = \sum_{h} W_h \bar{x}_h$.

In the present study the values of the sample sizes $n$ and $n'$ will be obtained by minimizing the $MSE(t_3)$ for a specific cost $C^*$. In order to evaluate the optimum $MSE$ for the estimator $t$, under the above mentioned constraint, we proceed as follows;
The mean square error of the proposed estimator is given by (3.9) as

\[ MSE(t_s) = \frac{(1 - \rho^2)}{n} + \frac{\rho^2}{n'} + (\rho^2 / n')\{A \varphi(A) - B \varphi(B)\} - \rho^2 (1 / n' - \mu_s^2)\{\Phi(A) - \Phi(B)\} \]

In general the values of \( \mu_s \) are unknown, the experimenter has partial information about it. When \( \mu_s = 0 \), the mean square error of \( t_s \) is least and the relative efficiency is largest. Thus it would be reasonable to let \( \mu_s = 0 \) in \( MSE(t_s) \) and obtain the values of \( n \) and \( n' \) under the optimum situation.

When \( \mu_s = 0 \), \( A = Z_a \) and \( B = -Z_a \), which further implies that

\[ \Phi(A) = 1 - (\alpha / 2) \quad \text{and} \quad \Phi(B) = \alpha / 2 \]

Hence,

\[ MSE(t_s) = \frac{(1 - \rho^2)}{n} + \frac{\rho^2}{n'} + \frac{\rho^2}{n'}\{Z_a \varphi(Z_a) - (-Z_a)\varphi(-Z_a)\} - (\rho^2 / n')\{1 - \alpha / 2 - \alpha / 2\} \]

\[ = (1 - \rho^2) / n + \frac{\rho^2}{n'}\{\alpha + 2Z_a \varphi(Z_a)\} / n' \]

\[ = (K / n) + (K' / n') \] ............................(4.3)

where \( K = 1 - \rho^2 \quad K' = \rho^2\{\alpha + 2Z_a \varphi(Z_a)\} \)

We minimize the MSE\((t_s)\) subject to a specific cost constraint \( C^* \), given by

\[ C^* = c'n' + cn \] ............................(4.4)

Lagrange’s multipliers method is used to minimize of the MSE\((t_s)\) in eqn(4.3) subject to the cost constraint \( C^* \) in eqn(4.4). The Langrange’s equation can be written as follows
\[ L(n', n, \lambda) = \left( K / n + K' / n' \right) - \lambda(C^* - c'n' - cn) \] 

\[ (4.5) \]

where \( \lambda \) is a constant or the Langrange multiplier.

The necessary conditions for a minimum of \( \text{MSE}(t_s) = (K / n) + (K' / n') \) subject to \( C^* - c'n' - cn = 0 \) are given by

\[
\frac{\partial L}{\partial n'} = 0, \quad \frac{\partial L}{\partial n} = 0 \quad \text{and} \quad \frac{\partial L}{\partial \lambda} = 0
\]

Hence,

\[-(K'/n'^2) + \lambda c' = 0\]

\[-(K/n^2) + \lambda c = 0\]

and

\[C^* - c'n' - cn = 0\]

Thus,

\[ (K'/n'^2) = c' \lambda \quad \text{and} \quad (K/n^2) = c \lambda\]

\[\Rightarrow (n'^2 c' / K') = (n^2 c / K) = 1 / \lambda\]

\[\Rightarrow (n'^2 / K'c) = (n^2 / K c') = (1 / \lambda c c') = L \quad \text{(say)}\]

\[\Rightarrow n' = \sqrt{L} \sqrt{K'c} \quad \text{and} \quad n = \sqrt{L} \sqrt{Kc'} \] 

\[\text{..........................}(4.6)\]

Substituting the value of \( n \) and \( n' \) in the cost function \( C^* \), we get

\[ C^* = c'n' + cn \]

\[ C^* = (c' \sqrt{K'c} + c \sqrt{Kc'}) \sqrt{L} \]

\[ \sqrt{L} = C^*/(c' \sqrt{K'c} + c \sqrt{Kc'}) \]

\[ \sqrt{L} = C^*/(\sqrt{(Kc + \sqrt{K'c'})} \]

\[\text{..........................}(4.7)\]
Thus the optimum value of $MSE(t_4)$ is obtained by substituting the value of $n$ and $n'$ from eqn (4.6) and eqn (4.7) in eqn (4.3)

$$\therefore M_{opt}(t_4) = \left[ \frac{K}{\sqrt{Kc'}} + \frac{K'}{\sqrt{K'c'}} \right] (1/\sqrt{L})$$

$$= \left[ \frac{\sqrt{K/c'} + \sqrt{K'/c'}}{\sqrt{cc'}} \right] \sqrt{(Kc + \sqrt{K'c'})}$$

Thus

$$M_{opt}(t_4) = \left[ \frac{\sqrt{Kc + \sqrt{cc'}}}{C^*} \right]$$

which gives the required optimum mean square error of the proposed estimator.

4.5 Comparison of the CRPTE with combined regression estimator under optimum condition

The Combined regression estimator under double sampling is given by

$$t_4 = y_{st} + b_{st} (x_n' - x_s)$$

The Mean square error of $t_4$ is given by

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

where $K = (1 - \rho^2)$ and $K'' = \rho^2$

and the cost function given by eqn (4.2)

Again by proceeding as in Section 4.4 by applying Lagrange's multipliers method, we minimize $MSE(t_4)$ for a specific cost $C^*$ and we get
\[ M_{\text{opt}}(t_4) = \left( \sqrt{Kc} + \sqrt{K''c'} \right)^2 / C^* \]

\[ \text{.................(4.9)} \]

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

4.6 Discussion

Thus we have proved that under optimum conditions, mean square error of the CRPTE in double sampling with an auxiliary variable is smaller than the mean square error of combined regression estimator under double sampling. Therefore under the stated assumptions, the proposed estimator is more efficient than the combined regression estimator under double sampling.
Table 4.1 Behaviour of relative efficiency of $t_5$ to $t_4$ with respect to $\mu \chi$ for different values of $\alpha$ and for $n' = 200, n = 100, \rho = 0.8$

<table>
<thead>
<tr>
<th>$\alpha$</th>
<th>$\mu \chi$</th>
<th>0</th>
<th>0.1</th>
<th>0.2</th>
<th>0.3</th>
<th>0.4</th>
<th>0.5</th>
<th>0.6</th>
<th>0.7</th>
<th>0.8</th>
<th>0.9</th>
<th>1.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.01</td>
<td></td>
<td>1.76</td>
<td>0.66</td>
<td>0.44</td>
<td>0.76</td>
<td>0.99</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0.05</td>
<td></td>
<td>1.51</td>
<td>0.7</td>
<td>0.66</td>
<td>0.95</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0.25</td>
<td></td>
<td>1.15</td>
<td>0.86</td>
<td>0.93</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 4.2 Behaviour of relative efficiency of $t_5$ to $t_4$ with respect to $\mu \chi$ for different values of $\rho$ and for $n' = 200, n = 100, \alpha = 0.05$

<table>
<thead>
<tr>
<th>$\rho$</th>
<th>$\mu \chi$</th>
<th>0</th>
<th>0.1</th>
<th>0.2</th>
<th>0.3</th>
<th>0.4</th>
<th>0.5</th>
<th>0.6</th>
<th>0.7</th>
<th>0.8</th>
<th>0.9</th>
<th>1.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.7</td>
<td></td>
<td>1.31</td>
<td>0.77</td>
<td>0.73</td>
<td>0.96</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0.8</td>
<td></td>
<td>1.51</td>
<td>0.7</td>
<td>0.66</td>
<td>0.95</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0.9</td>
<td></td>
<td>1.96</td>
<td>0.62</td>
<td>0.57</td>
<td>0.92</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
**Figure 4.1** Behaviour of relative efficiency of $t_5$ to $t_4$ with respect to $\mu_x$
for different values of $\alpha$ and for $n' = 200$, $n = 100$, $\rho = 0.8$

![Graph](image1)

**Figure 4.2** Behaviour of relative efficiency of $t_5$ to $t_4$ with respect to $\mu_x$
for different values of $\rho$ and for $n' = 200$, $n = 100$, $\alpha = 0.05$

![Graph](image2)