CHAPTER – 4

COMPROMISED ESTIMATION METHOD WITH MISSING DATA, IMPUTATION AND MEASUREMENT ERROR

4.1 INTRODUCTION

The problem to face with missing data is a common in a survey and imputation techniques are frequently used to deal with them. In other hand, the problem of measurement error is again a deep trouble for investigators when data deals with highly personal or prestigious information. These problems of non-response and measurement error are natural in today’s environment for any kind of survey sampling.

‘Compromised’ word is introduced by Singh and Horn (2000) in their research contribution. They proposed a method (called compromised method) of imputation where compromised stands for those values of auxiliary variable which are associated with the available values of study variable whereas entire values are available for auxiliary variable.

\[ \bar{y}_{comp} = \alpha \bar{y}_r + (1 - \alpha) \frac{\bar{x}_n}{\bar{x}_r} \]
Abu-Dayyeh et al. (2003) has studied some estimators of finite population mean using auxiliary information. They considered two classes of estimators to estimate the population mean for the variable of interest using two auxiliary variables. It turns out that the newly suggested estimators dominate all other well-known estimators in terms of mean square error and bias both. They presented how to extend the two classes of estimators if more than two auxiliary variables are available.

Rueda and González (2008) considered the problem where two variables can have missing values not necessarily for the same units and conclude that proposed method produces an important increase in efficiency when parameters such as the median, quartiles and variance are estimated. Shukla (2002) proposed some estimation strategies for estimating the population mean under this set up. The Factor-Type (F-T) estimator is considered throughout with the derivation of properties of suggested estimators.

Shukla et al. (2009) proposed mean estimation under imputation of missing data using factor-type estimator in two-phase sampling. Shukla et al. (2011) suggested linear combination based imputation method for missing data in sample. Shukla et al. (2012a) proposed estimation of population mean using two auxiliary sources in sample surveys. Shukla et al. (2012b) suggested an estimator for mean estimation in presence measurement error of the observations. Shukla et al. (2012c) presented a transformed estimator for estimation of population mean with missing data in sample-surveys. Thakur et al. (2011, 2012) suggested some imputation strategy under the setup of double sampling. Singh et al. (2007) proposed a method to reduce the bias appearing in the alternative to ratio-cum-product estimator for estimating finite population mean in sample surveys. Tailor and Sharma (2009) derived a

4.2 NOTATIONS AND ASSUMPTIONS

Assume given a set of information obtained through simple random sampling procedure on two characteristics X and Y. Suppose \( (x_i, y_i) \) be the pair of observational values and \( (X_i, Y_i) \) are corresponding true values on the characteristic \( (X,Y) \) respectively.

For the \( i^{th} \) unit \( (i=1, 2...n) \) in the same sample suppose the measurement error are as follow.

\[
U_i = (y_i - Y_i) \quad \ldots (4.1)
\]

\[
V_i = (x_i - X_i) \quad \ldots (4.2)
\]

Notations for the study are:

\( \bar{Y}, \bar{X} \) : Population Parameters

\( \bar{y} \) and \( \bar{x} \) : Mean per unit estimates for a simple random sample of size \( n \).

\( n \) : Sample size

\( N \) : Population size

\( \rho \) : Correlation coefficient between variables

\( f \) : Sampling friction
Let these measurement errors are stochastic in nature and are uncorrelated \([i.e. \rho(U_i, V_i) = 0]\) along with sum of measurement error is zero and variances are \(\sigma_i^2\) and \(\sigma_i^2\) respectively. The population means are \(\bar{X}\) and \(\bar{Y}\) for the true values with population variances are \(\sigma_X^2\) and \(\sigma_Y^2\) respectively. The \(\bar{Y}\) is unknown and mean of auxiliary information \(\bar{X}\) is known which is used as a source to estimate \(\bar{Y}\). From (4.1) we write

\[
\bar{y} = \frac{1}{n} \sum_{i=1}^{n} (U_i + Y_i)
\]  
... (4.3)

4.3 SOME EXISTING ESTIMATORS

Some well known imputation methods are given below:

4.3.1 RATIO METHOD OF IMPUTATION

For \(y_i\) and \(x_i\) define \(y_{*i}\) as

\[
y_{*i} = \begin{cases} 
  y_i & \text{if } i \in R \\
  \hat{b}x_i & \text{if } i \in R^C
\end{cases}
\]  
... (4.4)

Where \(\hat{b} = \frac{\sum_{i \in R} y_i}{\sum_{i \in R} x_i}\)

Using above, the imputation-based estimator of population mean is:

\[
\bar{y}_S = \frac{1}{n} \sum_{i \in S} \bar{y}_{*i} = \bar{y}_r \left( \frac{\bar{x}_n}{\bar{x}_r} \right) = \bar{y}_{RAT}
\]  
... (4.5)

where \(\bar{y}_r = \frac{1}{r} \sum_{i \in R} y_i\), \(\bar{x}_r = \frac{1}{r} \sum_{i \in R} x_i\) and \(\bar{x}_n = \frac{1}{n} \sum_{i \in S} x_i\)
4.3.2 MEAN METHOD OF IMPUTATION

For $y_i$ define $y_i^*$ as

$$y_i^* = \begin{cases} 
    y_i & \text{if } i \in R \\
    \bar{y}_r & \text{if } i \in R^C 
\end{cases} \quad \cdots (4.6)$$

Using above, the imputation-based estimator of population mean is

$$\bar{y}_m = \frac{1}{r} \sum_{i \in R} y_i = \bar{y}_r \quad \cdots (4.7)$$

4.4 PROPOSED ESTIMATOR(S)

The proposed estimators are

$$y_{1i} = \begin{cases} 
    y_i & \text{if } i \in R \\
    \frac{\bar{y}_r}{(1 - f_1)} [\phi_1 - f_1] & \text{if } i \in R^C 
\end{cases} \quad \cdots (4.14)$$

$$y_{2i} = \begin{cases} 
    y_i & \text{if } i \in R \\
    \frac{\bar{y}_r}{(1 - f_1)} [\phi_2 - f_1] & \text{if } i \in R^C 
\end{cases} \quad \cdots (4.15)$$

$$y_{3i} = \begin{cases} 
    y_i & \text{if } i \in R \\
    \frac{\bar{y}_r}{(1 - f_1)} [\phi_3 - f_1] & \text{if } i \in R^C 
\end{cases} \quad \cdots (4.16)$$

$$y_{4i} = \begin{cases} 
    y_i & \text{if } i \in R \\
    \frac{\bar{y}_r}{(1 - f_1)} [\phi_4 - f_1] & \text{if } i \in R^C 
\end{cases} \quad \cdots (4.17)$$
\[ y_{si} = \begin{cases} y_i & \text{if } i \in R \\ \frac{\bar{y}_r}{(1-f_i)} [\phi_S - f_i] & \text{if } i \in R^c \end{cases} \] ... (4.18)

The point estimates are

\[ \bar{y}_{DS1} = \xi \bar{y}_r + (1 - \xi) \frac{\bar{X}}{\bar{x}_r} \] ... (4.24)

\[ \bar{y}_{DS2} = \xi \bar{y}_r + (1 - \xi) \frac{\bar{X}}{\bar{x}_n} \] ... (4.25)

\[ \bar{y}_{DS3} = \xi \bar{y}_r + (1 - \xi) \frac{(A+C)\bar{X} + fB\bar{x}_r}{(A+fB)\bar{X} + C\bar{x}_r} \] ... (4.26)

\[ \bar{y}_{DS4} = \xi \bar{y}_r + (1 - \xi) \frac{(A+C)\bar{x}_n + fB\bar{x}_r}{(A+fB)\bar{x}_n + C\bar{x}_r} \] ... (4.27)

\[ \bar{y}_{DS5} = \xi \bar{y}_r + (1 - \xi) \frac{(A+C)\bar{x}_n + fB\bar{x}_r}{(A+fB)\bar{x}_n + C\bar{x}_r} \] ... (4.28)

### 4.5 Properties of Proposed Estimator(s)

**Theorem 4.1** For large sample approximation

[1]: The estimator \( \bar{y}_{DS1} \) could be expressed as:

\[ \bar{y}_{DS1} = \bar{Y} + \lambda_0 - \frac{\bar{y}_P}{\bar{X}} \lambda_1 + \frac{\bar{y}_P}{\bar{X}^2} \lambda_1^2 - \frac{P}{\bar{X}} \lambda_0 \lambda_1 \] ... (4.29)

[2]: Bias of \( \bar{y}_{DS1} \) is:

\[ E[\bar{y}_{DS1} - \bar{Y}] = \frac{\bar{y}_P}{r} C_X \left[ C_X \left\{ 1 + \frac{\sigma_Y^2}{\sigma_X^2} \right\} - \rho \sigma_Y \right] \] ... (4.30)
[3]: Mean squared error of $\bar{y}_{DS1}$ is:

$$E\left[ (\bar{y}_{DS1} - \bar{y})^2 \right] = \frac{1}{r} (\sigma_y^2 + \sigma_U^2) + \frac{P\bar{y}}{r} C_X \left[ PC \left\{ 1 + \frac{\sigma_y^2}{\sigma_X^2} \right\} - 2 \rho C_Y \right] \ldots (4.31)$$

THEOREM 4.2 For large sample approximation

[4]: The estimator $\bar{y}_{DS2}$ could be expressed as:

$$\bar{y}_{DS2} = \bar{y} + \lambda_0 - \frac{\bar{y} P_2}{X} \lambda_2 + \frac{\bar{y} P_2}{X} \lambda_2^2 - \frac{P_2}{X} \lambda_0 \lambda_2 \ldots (4.33)$$

[5]: Bias of $\bar{y}_{DS2}$ is:

$$E\left[ (\bar{y}_{DS2} - \bar{y}) \right] = \bar{y} P_2 \frac{1}{\sqrt{n}} C_X \left[ \frac{1}{\sqrt{n}} C_X \left\{ 1 + \frac{\sigma_y^2}{\sigma_X^2} \right\} - \frac{1}{\sqrt{r}} \rho C_Y \right] \ldots (4.34)$$

[6]: Mean squared error of $\bar{y}_{DS2}$ is:

$$E\left[ (\bar{y}_{DS2} - \bar{y})^2 \right] = \frac{1}{r} (\sigma_y^2 + \sigma_U^2) + \frac{\bar{y}}{\sqrt{n}} P_2 C_X \left[ \frac{1}{\sqrt{n}} P_2 C_X \left\{ 1 + \frac{\sigma_y^2}{\sigma_X^2} \right\} - \frac{2}{\sqrt{r}} \rho C_Y \right] \ldots (4.35)$$

THEOREM 4.3 For large sample approximation

[7]: The estimator $\bar{y}_{DS3}$ could be expressed as:

$$\bar{y}_{DS3} = \bar{y} + \lambda_0 + \frac{\bar{y} P_3 (\alpha - \beta)}{X} \lambda_2 + \frac{\bar{y} P_3 (\beta - \alpha)}{X^2} \lambda_2^2 + \frac{P_3 (\alpha - \beta)}{X} \lambda_0 \lambda_2 \ldots (4.37)$$
[8]: Bias of $\overline{y}_{DS3}$ is:

$$E[\overline{y}_{DS3} - \overline{y}] = \frac{\overline{y}\psi_1 C_X}{r} \left[ \rho \sigma_Y - \beta_1 C_X \left(1 + \frac{\sigma_Y^2}{\sigma_X^2}\right) \right] \quad \ldots (4.38)$$

[9]: Mean squared error of $\overline{y}_{DS3}$ is:

$$E[\overline{y}_{DS3} - \overline{y}]^2 = \frac{1}{r} \left[ (\sigma_Y^2 + \sigma_U^2) + \overline{y}^2 \psi_1 C_X \left\{ \psi_1 C_X \left(1 + \frac{\sigma_Y^2}{\sigma_X^2}\right) + 2 \rho C_Y \right\} \right] \quad \ldots (4.39)$$

**THEOREM 4.4** For large sample approximation

[10]: The estimator $\overline{y}_{DS4}$ could be expressed as:

$$\overline{y}_{DS4} = \overline{y} + \lambda_0 + \frac{\overline{y}P_4(\alpha_1 - \beta_1)}{X} \lambda_2 + \frac{\overline{y}P_4(\beta_1^2 - \alpha_1 \beta_1)}{X^2} \lambda_2^2 + \frac{P_4(\alpha_1 - \beta_1)}{X} \lambda_0 \lambda_2 \quad \ldots (4.41)$$

[11]: Bias of $\overline{y}_{DS4}$ is:

$$E[\overline{y}_{DS4} - \overline{y}] = \frac{\overline{y}\psi_2 C_X}{\sqrt{n}} \left[ \frac{1}{\sqrt{r}} \rho C_Y - \beta_1 \frac{1}{\sqrt{n}} C_X \left(1 + \frac{\sigma_Y^2}{\sigma_X^2}\right) \right] \quad \ldots (4.42)$$

[12]: Mean squared error of $\overline{y}_{DS4}$ is:

$$E[\overline{y}_{DS4} - \overline{y}]^2 = \frac{1}{r} (\sigma_Y^2 + \sigma_U^2) + \frac{\overline{y}^2 \psi_2 C_X}{\sqrt{n}} \left\{ \psi_2 C_X \left(1 + \frac{\sigma_Y^2}{\sigma_X^2}\right) + 2 \rho C_Y \right\} \quad \ldots (4.43)$$

**THEOREM 4.5** For large sample approximation

[13]: The estimator $\overline{y}_{DS5}$ could be expressed as:
\[ \overline{y}_{DS5} = \overline{Y} + \lambda_0 + \frac{\overline{Y} \eta}{\overline{X}} (\lambda_1 - \lambda_2) - \frac{\overline{Y} \eta}{\overline{X}^2} (\beta_1 \lambda_1^2 + h_1 \lambda_2^2) + \frac{\eta}{\overline{X}} (\lambda_0 \lambda_1 - \lambda_0 \lambda_2) - \frac{\overline{Y} \eta (\beta_1 - h_1)}{\overline{X}^2} \lambda_1 \lambda_2 \]

\[ \text{... (4.45)} \]

[14]: Bias of \( \overline{y}_{DS5} \) is:

\[ E[\overline{y}_{DS5} - \overline{Y}] = \overline{y}_{P_5} \left[ \frac{1}{r} (\alpha_1 - \beta_1) + (g_1 - h_1) \frac{1}{\sqrt{nr}} \rho C_X C \right] - \frac{1}{r} \beta_1 (\alpha_1 - \beta_1) \]

\[ + \frac{1}{n} h_1 (g_1 - h_1) + \frac{1}{\sqrt{nr}} \{ h_1 (\alpha_1 - \beta_1) + \beta_1 (g_1 - h_1) \} C_X^2 \left[ 1 + \frac{\sigma_Y^2}{\sigma_X^2} \right] \]

\[ \text{... (4.46)} \]

[15]: Mean squared error of \( \overline{y}_{DS5} \) is:

\[ E[\overline{y}_{DS5} - \overline{Y}]^2 = \frac{1}{r} (\sigma_Y^2 + \sigma_U^2) + \overline{Y}^2 \eta^2 (\sigma_X^2 + \sigma_Y^2) \left( \frac{1}{r} + \frac{1}{n} - 2 \frac{1}{\sqrt{nr}} \right) + 2 \overline{Y} \eta \rho \sigma_Y \sigma_X \left( \frac{1}{r} - \frac{1}{\sqrt{nr}} \right) \]

\[ \text{... (4.47)} \]

### 4.6 OPTIMAL CHOICES

For \( \overline{y}_{DS3}, \overline{y}_{DS4} \) and \( \overline{y}_{DS5} \), we have a wide choice to choose the scalar value in such a way that the resultant mean squared error becomes minimum. At different \( \xi^\prime \)'s we can have various mean squared errors from these estimators and hence for each mean squared error \( \psi_2, \psi_2 \) and \( \psi_3 \) exists. These \( \psi \) values are function of \( K \) solely and provide polynomial
of certain degrees. Roots of the polynomials are the optimal values of characterizing scalar to bear minimum mean squared error

4.7 EMPIRICAL STUDY

Population of size 250 (N=250) is given in appendix-I. The population parameters are displayed in Table 4.4.

Table -4.4 Population Parameters

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
<th>Parameter</th>
<th>Value</th>
<th>Parameter</th>
<th>Value</th>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\bar{Y}$</td>
<td>63.396</td>
<td>$n$</td>
<td>50</td>
<td>$C_0$</td>
<td>0.2899</td>
<td>$\rho_{01}$</td>
<td>0.8544</td>
</tr>
<tr>
<td>$\bar{X}_1$</td>
<td>48.136</td>
<td>$\sigma_Y$</td>
<td>18.3819</td>
<td>$C_1$</td>
<td>0.4637</td>
<td>$\rho_{02}$</td>
<td>0.8249</td>
</tr>
<tr>
<td>$\bar{X}_2$</td>
<td>56.364</td>
<td>$\sigma_{X_2}$</td>
<td>23.03</td>
<td>$C_2$</td>
<td>0.4085</td>
<td>$\rho_{12}$</td>
<td>0.8289</td>
</tr>
</tbody>
</table>

Table -4.5 Bias and Mean squared error for existing estimators

<table>
<thead>
<tr>
<th>Estimator</th>
<th>Bias*</th>
<th>Mean Squared Error*</th>
</tr>
</thead>
<tbody>
<tr>
<td>With Y and $X_1$ variables</td>
<td>With Y and $X_2$ variables</td>
<td>With Y and $X_1$ variables</td>
</tr>
<tr>
<td>$t_1 = \frac{y_r}{x_r} \left( \frac{\bar{X}}{\bar{X_n}} \right)^{\beta_1}$</td>
<td>0.760317</td>
<td>0.600939</td>
</tr>
<tr>
<td>$t_2 = \frac{y_r}{x_r} \left( \frac{\bar{X}_n}{\bar{X}_r} \right)^{\beta_2}$</td>
<td>0.040701</td>
<td>0.032144</td>
</tr>
<tr>
<td>$t_3 = \frac{y_r}{x_r} \left( \frac{\bar{X}}{\bar{X_r}} \right)^{\beta_3}$</td>
<td>0.793138</td>
<td>0.62638</td>
</tr>
</tbody>
</table>
### Table - 4.6 Bias and Mean squared error for suggested estimators

<table>
<thead>
<tr>
<th>Estimators</th>
<th>Optimum values</th>
<th>Bias</th>
<th>Mean Squared Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\bar{y}_{DS1}$</td>
<td>$P_1 = 0.1694$</td>
<td>0.13442</td>
<td>21.0626</td>
</tr>
<tr>
<td></td>
<td>$P_1 = 0.1866$</td>
<td>0.14804</td>
<td>21.0805</td>
</tr>
<tr>
<td>$\bar{y}_{DS2}$</td>
<td>$P_2 = 0.1786$</td>
<td>0.12611</td>
<td>21.0626</td>
</tr>
<tr>
<td></td>
<td>$P_2 = 0.1967$</td>
<td>0.13889</td>
<td>21.0805</td>
</tr>
<tr>
<td>$\bar{y}_{DS3}$</td>
<td>$\psi_1 = -0.1694$</td>
<td>0.11656</td>
<td>21.0626</td>
</tr>
<tr>
<td></td>
<td>$\psi_1 = -0.1866$</td>
<td>0.12999</td>
<td>21.0805</td>
</tr>
</tbody>
</table>

#### 4.8 DISCUSSION AND CONCLUSION

The basic compromised estimator was suggested by Singh and Horn (2000). In this chapter we have suggested five estimators as $\bar{y}_{DS1}$, $\bar{y}_{DS2}$, $\bar{y}_{DS3}$, $\bar{y}_{DS4}$ and $\bar{y}_{DS5}$ similar to Singh and Horn (2000) in terms of construction in structure and environment. One highlighted feature is that we have extended all these to the setup of measurement error also which was not considered earlier. All the suggested estimators, with imputation and measurement error, are more efficient than usual mean per unit estimator. Moreover, some of them are having higher efficiency to the compromised estimator of Singh and Horn (2000).