Some new aspects on imputation in sampling

Diwakar Shukla¹, Narendra Singh Thakur² and Sharad Pathak¹

¹Department of Mathematics and Statistics, Dr. H.S. Gour University of Sagar, Sagar (M. P.), 470003, India.
²Centre for Mathematical Sciences (CMS), Banasthali University, Rajasthan, 304022, India.

Accepted 15 January, 2013

The number of causes that affect the quality of survey and missing data is one of such that keeps sample incomplete. Many imputation methods are available in literature, and are used to replace missing observations like Mean method, Ratio method, Compromised method, Ahmed’s method, Factor-type (F-T) method etc. This paper suggests some new estimation aspects in imputation theory using both, F-T estimator and Compromised estimator. All these are derived from existing procedure of usual Compromised method and found efficient, bias controlled, having many properties. In support of derived results, an empirical study is incorporated over artificial data set showing the efficiency in favour of the suggested methods.

Key words: Estimation, missing data, imputation, bias, mean squared error (m.s.e.), compromised estimator, factor-type (F-T) estimator, factor-type compromised imputation (FTCI).

INTRODUCTION

Missing data is one of the most common problems in sample surveys and various imputation methodologies are frequently used to substitute these values. Statisticians have recognized that failure to account for the stochastic nature of incompleteness in the form of absence of data can spoil inference. There are three major incompleteness concepts: Missing at random (MAR), observed at random (OAR), and parametric distribution (PD). In what follows in this paper, missing completely at random (MCAR) is used. Some well known imputation methods in literature are: deductive imputation, mean imputation overall (MO), random imputation overall (RO), mean imputation within classes (MC), random imputation within classes (RC), hot-deck imputation, flexible matching imputation, predicted regression imputation (PR), random regression imputation (RR), distance function matching etc.

In a contribution, Shukla (2002) suggested Factor-type (F-T) estimator in two-phase sampling which incorporates a parameter k combining known auxiliary information survey components. But k used therein is in the form of factors (k−1)(k−2); (k−2)(k−3)(k−4) which ultimately produces a noble property of getting multiple choices of k-values maintaining certain standard level of mean squared error (m.s.e). As suggested, best k among all is that also optimizing other characteristic of estimator. Singh and Horn (2000) proposed a compromised imputation procedure having a constant α in linear combination of “available main information” and “imputed auxiliary information”. The α bears an optimum selection at the minimum level of m.s.e. but this choice does not have a control over bias factor. This paper derives motivation from this source and, using properties of F-T estimator, presents a new F-T compromised imputation procedure for survey sampling. Special feature of the suggested is the same parameter k used twice, in the imputation part for missing observations as well as in the linear combination with the known auxiliary part. Some other useful contributions are due to Rueda and Gonzalez (2008), Singh et al. (2009) etc. over imputation techniques.

Let \( \bar{Y} \) be mean of a finite population for desired estimation \( \bar{Y} = N^{-1} \sum_{i=1}^{N} Y_i \). A random sample S of size n drawn without replacement (SRSWOR) from population \( \Omega = \{1,2,3,\ldots,N\} \) to estimate \( \bar{Y} \). Sample S of n units contains r responding units \( r < n \) forming a set R and \( (n-r) \) non-responding units with the sub-space \( (n-r) \) having symbol \( \kappa^c \). The variable Y is of main interest and
For every unit $i \in R$, the value $y_i$ observed is available. For $i \in R^C$, $y_i$ values are missing and imputed values need to be derived. The $i^{th}$ value $x_i$ of auxiliary variate $X$ is used as a source of imputation for missing data when $i \in R^C$. Assume for sample $S$, the data $x_s = \{x_i : i \in S\}$ are known and $S = R \cup R^C$. Under this setup, some well-known imputation methods in survey sampling literature are:

**Mean method of imputation**

For $y_i$, define $y^*_{i}$ as

$$y^*_{i} = \begin{cases} y_i & \text{if } i \in R \\ \bar{y}_r & \text{if } i \in R^C \end{cases}$$

(1)

The imputation-based estimator of population mean $\bar{Y}$ is:

$$\bar{y}_n = \frac{1}{n} \sum_{i \in R} y_i = \frac{1}{r} \sum_{i \in R} y_i$$

(2)

**Ratio method of imputation**

For $y_i$ and $x_i$, define $y^*_{i}$ as

$$y^*_{i} = \begin{cases} y_i & \text{if } i \in R \\ \hat{b}x_i & \text{if } i \in R^C \end{cases}$$

(3)

Where $\hat{b} = \sum_{i \in R} y_i / \sum_{i \in R} x_i$

The imputation-based estimator of $\bar{Y}$ is:

$$\bar{y}_S = \frac{1}{n} \sum_{i \in S} y^*_{i} = \bar{y}_r \left( \frac{x_n}{x_r} \right) = \bar{y}_{RAT}$$

(4)

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 $\bar{x}_n = \frac{1}{n} \sum_{i \in S} x_i$

**Compromised method**

Singh and Horn (2000) proposed compromised imputation procedure

$$y^*_{i} = \begin{cases} (\alpha n/r)y_i + (1-\alpha)\hat{b}x_i & \text{if } i \in R \\ (1-\alpha)\hat{b}x_i & \text{if } i \in R^C \end{cases}$$

(5)

The imputation-based estimator, in this case, is

$$\bar{y}_{COMP} = \left[ \alpha \bar{y}_r + (1-\alpha)\bar{y}_n \right] \frac{x_n}{x_r}$$

(6)

Where $\alpha$ is a suitable constant value such that the resultant variance is minimum.

**Lemma 1**

The bias, m.s.e. and minimum m.s.e. of $\bar{y}_{COMP}$ is (Singh and Horn, 2000):

(i) $B(\bar{y}_{COMP}) = (1-\alpha) \left( \frac{1}{r} - \frac{1}{n} \right) \bar{Y} (C_x^2 + \rho C_x C_Y)$

(7)

(ii) $M(\bar{y}_{COMP}) = \left( \frac{1}{r} - \frac{1}{n} \right) \bar{Y}^2 C_x^2 + \left( \frac{1}{r} - \frac{1}{n} \right) \bar{Y}^2 \left[ (1-\alpha)^2 C_x^2 - 2(1-\alpha)\rho C_x C_Y \right]$ 

(8)

(iii) For optimum $\alpha = \left( 1 - \frac{\rho C_Y}{C_x} \right)$, the minimum m.s.e. of $\bar{y}_{COMP}$ is given by the expression

$$M(\bar{y}_{COMP})_{\min} = \left[ \left( \frac{1}{r} - \frac{1}{N} \right) - \left( \frac{1}{r} - \frac{1}{n} \right) \rho^2 \right] S_1^2$$

(9)

**Ahmed methods**

For case where $y_{ji}$ denotes $j^{th}$ available observation for the $i^{th}$ imputation method, Ahmed et al. (2006) suggested:

(A) $y^*_{ji} = \begin{cases} y_{ji} & \text{if } i \in R \\ \frac{1}{(n-r)} \left[ n \bar{y}_r \left( \frac{\bar{X}}{\bar{x}_n} \right)^{\beta_x} - r \bar{y}_r \right] & \text{if } i \in R^C \end{cases}$

(10)
Under this, the point estimator of \( \bar{Y} \) is

\[
t_1 = y_r \left( \frac{X}{x_r} \right)^{\beta_1}
\]

(B) \( y_{2i} = \)

\[
\frac{1}{(n-r)} \left[ n y_i \left( \frac{x_n}{x_r} \right)^{\beta_1} \right] - r y_r, \text{ if } i \in R^c
\]

Under this, the point estimator of \( \bar{Y} \) is

\[
t_2 = y_r \left( \frac{x_n}{x_r} \right)^{\beta_2}
\]

(C) \( y_{3i} = \)

\[
\frac{1}{(n-r)} \left[ n y_i \left( \frac{x_n}{x_r} \right)^{\beta_2} \right] - r y_r, \text{ if } i \in R^c
\]

Under this, the point estimator of \( \bar{Y} \) is

\[
t_3 = y_r \left( \frac{X}{x_r} \right)^{\beta_3}
\]

Terms \( \beta_1, \beta_2 \) and \( \beta_3 \) are suitably chosen constants, so as to keep the variance minimum.

As special cases:

When \( \beta_3 = 1 \), then

\[
t_{Ratio} = y_r \left( \frac{X}{x_r} \right)
\]

and when \( \beta_3 = -1 \), then

\[
t_{Product} = y_r \left( \frac{x_r}{X} \right)
\]

This is natural analogue of ratio estimator called the product estimator used when an auxiliary variate \( x \) has negative correlation with \( y \).

**Factor-type methods of imputation**

Shukla and Thakur (2008) have suggested F-T imputation procedure. For this case:

(D) \( (y_{ft1})_i = \)

\[
y_i, \text{ if } i \in R
\]

\[
y_r \left( \frac{n y_i}{(n-r)} \right) \left[ n \phi_i (k) - r \right], \text{ if } i \in R^c
\]

(E) \( (y_{ft2})_i = \)

\[
y_i, \text{ if } i \in R
\]

\[
y_r \left( \frac{n y_i}{(n-r)} \right) \left[ n \phi_i (k) - r \right], \text{ if } i \in R^c
\]

(F) \( (y_{ft3})_i = \)

\[
y_i, \text{ if } i \in R
\]

\[
y_r \left( \frac{n y_i}{(n-r)} \right) \left[ n \phi_i (k) - r \right], \text{ if } i \in R^c
\]

Where

\[
\phi_i (k) = \left[ (A + C) \frac{X}{X} + fBx_n \right]
\]

\[
\phi_2 (k) = \left[ (A + fB) \frac{X}{X} + Cx_r \right]
\]

\[
\phi_3 (k) = \left[ (A + C) \frac{X}{X} + fBx_r \right]
\]

\[
h = \frac{n}{N}, 0 \leq k \leq \infty
\]

Under Equations 15, 16 and 17, point estimators of \( \bar{Y} \) is

\[
T_{FT1} = y_r \phi_i (k)
\]

\[
T_{FT2} = y_r \phi_2 (k)
\]

\[
T_{FT3} = y_r \phi_3 (k)
\]

As special cases, when \( k = 1, \beta_1 = 1 \) then \( T_{FT1} = t_l \)
When \( k = 2, \beta_l = -1 \) then \( T_{F_l} = t_l \)

When \( k = 4, \beta_l = 0 \) then \( T_{F_l} = t_l = \bar{y}_l; \quad (l = 1, 2, 3) \)

**PROPOSED ESTIMATOR**

Using Singh and Horn (2000), F-T compromised imputation (FTCI) estimator is (\( j = 1, 2, 3 \)):

\[
(y_n) = \begin{cases} 
\left(\frac{k}{r}\right)y_i + (1 - k)\phi_j(k) & \text{if } i \in R \\
(1 - k)\phi_j(k) & \text{if } i \in R^C
\end{cases}
\]

(19)

Where, \( \phi_1(k) = \bar{y}_r \left( \frac{(A + C)\bar{X} + fB\bar{x}_n}{(A + fB)\bar{x}_r + C\bar{x}_n} \right) \);

\( \phi_2(k) = \bar{y}_r \left( \frac{(A + C)\bar{X} + fB\bar{x}_r}{(A + fB)\bar{x}_r + C\bar{x}_r} \right) \);

\( \phi_3(k) = \bar{y}_r \left( \frac{(A + C)\bar{X} + fB\bar{x}_r}{(A + fB)\bar{x}_r + C\bar{x}_r} \right) \);

\( A = (k - 1)(k - 2); \quad B = (k - 1)(k - 4); \)

\( C = (k - 2)(k - 3)(k - 4); \quad 0 \leq k \leq \infty. \)

**Remark 1**

Singh et al. (2001) have suggested a hybrid of calibration and imputation method in the presence of random non-response in survey sampling. Estimation of the population mean associated with the proposed hybrid method remains unbiased under a design based approach. Liu et al. (2005, 2006) have presented new imputation methods developed for the treatment of missing data, which remove the bias of usual ratio imputation. Singh (2009) has a useful contribution in the area of imputation techniques for missing data by proposing a new method of imputation in survey sampling.

**Remark 2**

In the contribution of Singh and Horn (2000), a point is raised by Singh and Deo (2003) that the adjustment in responding units \( y_i \) is not allowed. The present paper also bears this point. Some arguments in favour of proposed techniques are in Remark 3.

**Remark 3**

The available information \( y_i \) in sample of size \( n \) may be affected by some serious non-sampling errors like:

(i) Biasness in recording by investigator,
(ii) Under estimation or over estimation in reporting facts by respondents,
(iii) Partial unwillingness or lack of interest in answering question by respondent,
(iv) Deliberate miss reporting by respondents,
(v) False response/answer recording etc.

If available data is affected by these non-sampling errors, then there in need to also adjust available \( y_i \) values.

Sometimes, auxiliary information correlated to main variable \( Y \) is almost error free, for example, in an economic survey of expenditure (\( Y \)) of military soldiers, the income (\( X \)) record is correctly available through salary pay roll but expenditure in houses may be affected by serious non-sampling errors even if response is made. So, one can adjust partially the available expenditure data by the auxiliary income data to get better picture of available sample. There are many smoothing statistical techniques available in the literature useful over available sample data. Moreover, it is a well proved fact that sample mean of available data is also highly affected by extreme observations, outliers which need to be tackled by correlated information. In light of these, one can also think of adjusting the responding units. The estimator of Singh and Horn (2000) has dual property that it smoothens the outliers in available data as well as imputes the missing observations.

**Remark 4**

The proposed estimator is in the setup of SRSWOR. One can think of extending the same in the general setup of sampling design with multivariate structure.

**Properties of \( \phi_j(k) \)**

(i) At \( k = 1; \quad A = 0; \quad B = 0; \quad C = -6 \)

\[ \phi_1(1) = \bar{y}_r \frac{\bar{X}}{x_n}; \quad \phi_2(1) = \bar{y}_r \frac{x_n}{\bar{X}}; \quad \phi_3(1) = \bar{y}_r \frac{X}{x_r} \]

(ii) At \( k = 2; \quad A = 0; \quad B = -2; \quad C = 0 \)

\[ \phi_1(2) = \bar{y}_r \frac{\bar{X}}{x_n}; \quad \phi_2(2) = \bar{y}_r \frac{x_n}{\bar{X}}; \quad \phi_3(2) = \bar{y}_r \frac{X}{x_r} \]

(iii) At \( k = 3; \quad A = 2; \quad B = -2; \quad C = 0 \)
Table 1. Some special cases of FTCI.

<table>
<thead>
<tr>
<th>k</th>
<th>( \left(\bar{y}_{FTCI}^1\right)_j )</th>
<th>( \left(\bar{y}_{FTCI}^2\right)_j )</th>
<th>( \left(\bar{y}_{FTCI}^3\right)_j )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( k = 1 )</td>
<td>( \bar{y}_r )</td>
<td>( \bar{y}_r )</td>
<td>( \bar{y}_r )</td>
</tr>
<tr>
<td>( k = 2 )</td>
<td>( \bar{y}_r \left(2 - \frac{x_n}{X}\right) )</td>
<td>( \bar{y}_r \left(2 - \frac{x_r}{X}\right) )</td>
<td>( \bar{y}_r \left(2 - \frac{x_r}{X}\right) )</td>
</tr>
<tr>
<td>( k = 3 )</td>
<td>( \bar{y}_r \left(3 - \frac{2(X - f \bar{x}_n)}{(1-f)X}\right) )</td>
<td>( \bar{y}_r \left(3 - \frac{2(X - f \bar{x}_r)}{(1-f)X}\right) )</td>
<td>( \bar{y}_r \left(3 - \frac{2(X - f \bar{x}_r)}{(1-f)X}\right) )</td>
</tr>
<tr>
<td>( k = 4 )</td>
<td>( \bar{y}_r )</td>
<td>( \bar{y}_r )</td>
<td>( \bar{y}_r )</td>
</tr>
</tbody>
</table>

\[ 
\phi(3) = \bar{y}_r \left[ \frac{X - f \bar{x}_n}{(1-f)X} \right]; \phi_2(3) = \bar{y}_r \left[ \frac{x_n - f \bar{x}_r}{(1-f)x_n} \right]; \\
\phi(3) = \bar{y}_r \left[ \frac{X - f \bar{x}_r}{(1-f)X} \right] \\
\]

(iv) At \( k = 4; A = 6; B = 0; C = 0 \)

\[ 
\phi_1(4) = \phi_2(4) = \phi_3(4) = \bar{y}_r \text{ (Table 1)} 
\]

Theorem 1

The point estimator of \( \bar{Y} \) under FTCI is:

\[ 
\left(\bar{y}_{FTCI}^1\right)_j = \left[ k \bar{y}_r + (1-k)\phi_j(k) \right], \quad j = 1,2,3 
\]

Proof

\[ 
\left(\bar{y}_{FTCI}^1\right)_j = \left(\bar{y}_S^1\right)_j = \frac{1}{n} \sum_{i \in S} (y^i)_j 
\]

\[ 
= \frac{1}{n} \left[ \sum_{i \in R} (y^i)_j + \sum_{i \in R^c} (y^i)_j \right] 
\]

\[ 
= \frac{1}{n} \left[ \sum_{i \in R} \left( \frac{kn}{r} y_i + (1-k)\phi_j(k) \right) + \sum_{i \in R^c} (1-k)\phi_j(k) \right] 
\]

\[ 
= \frac{1}{n} \left[ \frac{kn}{r} \bar{y}_r + (1-k)\phi_j(k) \right] 
\]

BIAS AND MEAN SQUARED ERROR

Let \( B(.) \) and \( M(.) \) denote, respectively the bias and m.s.e. of an estimator under a given sampling design. Under the large sample approximations, let

\[ 
\bar{y}_r = \bar{y}(1+\varepsilon), \bar{x}_r = \bar{x}(1+\delta), \bar{x}_n = \bar{x}(1+\eta) 
\]

Using the concept of two-phase sampling following, Cochran (2005), Heitjan and Basu (1996), and Rao and Sitter (1995), and with the mechanism of MCAR, for given \( r \) and \( n \), we have

\[ 
E(\varepsilon) = E(\delta) = E(\eta) = 0, \\
E(\varepsilon^2) = M_1C^2, \quad E(\delta^2) = M_1C^2, \quad E(\eta^2) = M_2C^2; \\
E(\varepsilon\delta) = M_1\rho C_1C_X, \quad E(\delta\eta) = M_2C^2, \quad E(\varepsilon\eta) = M_3C_1C_X 
\]
Where
\[ M_1 = \left( \frac{1}{r} - \frac{1}{N} \right), \quad M_2 = \left( \frac{1}{n} - \frac{1}{N} \right), \quad M = (M_1 - M_2) \]

**Remark 5**

Some specific symbols
\[ \alpha_1 = \frac{fB}{A + fB + C}, \quad \alpha_2 = \frac{C}{A + fB + C}, \]
\[ \alpha_3 = \frac{A + C}{A + fB + C}, \quad \alpha_4 = \frac{A + fB}{A + fB + C}, \]
\[ P = (\alpha_1 \delta + \alpha_3 \eta), \quad Q = (\alpha_2 \delta + \alpha_4 \eta), \]
\[ (\alpha_1 + \alpha_3) = (\alpha_2 + \alpha_4) = 1, \]
\[ \alpha = (\alpha_1 - \alpha_2) = -(\alpha_3 - \alpha_4), \theta = (1 - k) \alpha, \]
\[ V = \rho \frac{C_y}{C_x} . \]

**Theorem 2**

\[ [a_1]: \text{The estimator } (\widehat{y}_{FTCI}) \text{ in terms of } \varepsilon, \delta \text{ and } \eta \text{ up to first order of approximation is:} \]
\[ (\widehat{y}_{FTCI}) = \overline{Y}[1 + \varepsilon + \theta(\eta + \alpha \eta - \alpha_2 \eta^2)] \quad (21) \]

**Proof**
\[ (\widehat{y}_{FTCI}) = k \overline{y}_r + (1 - k) \phi_1(k) \]
\[ = \left( k \overline{y}_r + (1 - k) \overline{y}_r \frac{(A + C)\overline{X} + fB\overline{x}_n}{(A + fB)\overline{X} + C\overline{x}_n} \right) \]
\[ = \overline{Y}(1 + \varepsilon)\left[ k + (1 - k) \frac{(A + C)\overline{X} + fB\overline{x}_n}{(A + fB)\overline{X} + C\overline{x}_n} \right] \]
\[ = \overline{Y}(1 + \varepsilon)\left[ k + (1 - k) \frac{1 + \alpha \eta}{1 + \alpha_2 \eta} \right] \]
\[ = \overline{Y}(1 + \varepsilon)\left[ k + (1 - k)(1 + \alpha \eta)(1 + \alpha_2 \eta)^{-1} \right] \]
\[ = \overline{Y}(1 + \varepsilon)\left[ k + (1 - k)(1 + \alpha \eta)(1 - \alpha_2 \eta + \alpha_2^2 \eta^2 - \alpha_2^3 \eta^3 + ...) \right] \]

(by first order only)
\[ = \overline{Y}(1 + \varepsilon)\left[ k + (1 - k)(1 - \alpha_2 \eta + \alpha_2^2 \eta^2 + \alpha_2^3 \eta^3 + ...) \right] \]

\[ [a_2]: \text{Bias of } (\widehat{y}_{FTCI}) \text{ is up to first order approximation is:} \]
\[ B[\widehat{y}_{FTCI}] = -\overline{Y}0M_2 (\alpha_2 C_x^2 - \rho C_y C_x) \quad (22) \]

**Proof**
\[ B[\widehat{y}_{FTCI}] = E[\widehat{y}_{FTCI}] - \overline{Y} \]

Using Equation 21 and taking expectations
\[ B[\widehat{y}_{FTCI}] = \left[ \overline{Y}[1 + \varepsilon + \theta(\eta + \alpha \eta - \alpha_2 \eta^2)] - \overline{Y} \right] \]
\[ = -\overline{Y}0M_2 (\alpha_2 C_x^2 - \rho C_y C_x) \]

\[ [a_3]: \text{The m.s.e of } (\widehat{y}_{FTCI}) \text{ up to some level of approximations:} \]
\[ M[\widehat{y}_{FTCI}] = \overline{Y}^2 \left[ M_1 C_Y^2 + M_2 \{ \theta^2 C_X^2 + 20\rho C_Y C_X \} \right] \quad (23) \]

**Proof**
\[ M[\widehat{y}_{FTCI}] = E[\widehat{y}_{FTCI}] - \overline{Y}^2 \]
\[ = E[\overline{Y}[1 + \varepsilon + \theta(\eta + \alpha \eta - \alpha_2 \eta^2)] - \overline{Y}^2] \]
\[ = \overline{Y}^2 E[\varepsilon + \theta(\eta + \alpha \eta - \alpha_2 \eta^2)]^2 \]
\[ = \overline{Y}^2 E[\varepsilon + \theta \eta]^2 \]
\[
\bar{Y}^2 \left[ M_1 C_Y^2 + M_2 \left\{ \theta^2 C_X^2 + 20pC_Y C_X \right\} \right]
\]

**Theorem 3**

\( [a_5] \): The estimator \( \bar{y}_{FTCI} \) in terms of \( \varepsilon, \delta \) and \( \eta \) up to first order of approximation is:

\[
\bar{y}_{FTCI} = \bar{Y} \left[ 1 + \varepsilon + \theta \left( -\eta + \omega \delta - \omega \delta - \alpha_2 - \alpha_4 \right) \right]
\]  \( (25) \)

**Proof**

\[
\frac{d}{d\theta} \left[ M \left( \bar{y}_{FTCI} \right) \right] = 0
\]

\[ \Rightarrow \theta = -\rho \frac{C_Y}{C_X} \]

Substituting \( \theta = -\rho \frac{C_Y}{C_X} \) in Equation 23, we get

\[
M \left[ \bar{y}_{FTCI} \right]_{\text{min}} = \left( M_1 - M_2 p^2 \right) S^2_Y
\]  \( (24) \)

\( [a_6] \): Bias of \( \bar{y}_{FTCI} \) is

\[
B \left[ \bar{y}_{FTCI} \right] = -\bar{Y} \theta M \left( \alpha_2 C_X^2 - p C_Y C_X \right)
\]  \( (26) \)

**Remark**

\( [a_7] \): The m.s.e of \( \bar{y}_{FTCI} \) is

\[
M \left[ \bar{y}_{FTCI} \right] = \bar{Y}^2 \left[ M_1 C_Y^2 + M_2 \theta^2 C_X^2 + 20pC_Y C_X \right]
\]  \( (27) \)

\( [a_8] \): Minimum m.s.e of the estimator \( \bar{y}_{FTCI} \) is on

\[
\theta = \left( -\rho \frac{C_Y}{C_X} \right) = -\rho \Rightarrow (1-k)(\alpha_1 - \alpha_2) = -\rho
\]

\[ \Rightarrow (1-k) \left( \frac{fB-C}{A+fB+C} \right) = -\rho \]

On simplification, we get polynomial of degree four:

\[
k^4 - (f - V)k^3 - [(f + 15) - (f - 8)V]k^2 + [(f - 10) - (5f - 23)V]k
\]
\[ + [(f - 24) + (4f - 22)V] = 0 \]  \( (33) \)

**Remark 7**

Reddy (1978) has shown that quantity \( V = \rho \frac{C_Y}{C_X} \) is stable over moderate length time period and could be initially known or guessed by past data. Therefore, pair \((f, V)\) be treated as known and Equation 33 generates
maximum of four roots (some may imaginary) on which optimum level of m.s.e. will be attained.

**Remark 8**

The Equation 33 has only unknown k in the power of order four, other V and f by Reddy (1978) are known in advance. We can solve Equation 33 for obtaining optimal estimator in the suggested class.

**COMPARISON OF THE ESTIMATORS**

1. Let $D_1 = M[\bar{y}_{FTCI 1}]_{\min} - M[\bar{y}_{FTCI 2}]_{\min}$

\[
= (M_1 - M_2 \rho^2)S_Y^2 - (M_1 - M\rho^2)S_Y^2
\]

(34)

\[
\bar{y}_{FTCI 1} \text{ is better than } \bar{y}_{FTCI 2} \text{ if } D_1 > 0 \Rightarrow r < \frac{n}{2 - f}
\]

If population $N$ is large, $f$ is small then working rule is; $D_1 > 0$ if $r < \left(\frac{n}{2}\right)$ which usually happens and high chance for $D_1 > 0$.

2. Let $D_2 = M[\bar{y}_{FTCI 1}]_{\min} - M[\bar{y}_{FTCI 3}]_{\min}$

\[
D_2 > 0 \Rightarrow M\rho^2S_Y^2 > 0
\]

(35)

Which is always positive. So, $\bar{y}_{FTCI 3}$ is always better than $\bar{y}_{FTCI 1}$ until $n = r$.

3. Let $D_3 = M[\bar{y}_{FTCI 2}]_{\min} - M[\bar{y}_{FTCI 3}]_{\min}$

\[
D_3 > 0 \Rightarrow N > n
\]

(36)

Which is always true. So, $\bar{y}_{FTCI 3}$ is always better than $\bar{y}_{FTCI 2}$ until $n = N$. But $\bar{y}_{FTCI 2}$ is better than $\bar{y}_{FTCI 1}$ therefore estimation strategy $\bar{y}_{FTCI 3}$ is most preferable over other two.

4. $D_4 = M[\bar{y}_{COMP}]_{\min} - M[\bar{y}_{FTCI 1}]_{\min}$

$D_4 > 0 \Rightarrow r > \frac{n}{2 - f} \Rightarrow r > \frac{n}{2}$ for $0 < f < 1$ (37)

This generally does not happen until huge non-response occurs.

5. $D_5 = M[\bar{y}_{COMP}]_{\min} - M[\bar{y}_{FTCI 1}]_{\min}$

\[
= (M_1 - M\rho^2)S_Y^2 - (M_1 - M\rho^2)S_Y^2 = 0
\]

(38)

Therefore, both are equally efficient.

6. $D_6 = M[\bar{y}_{COMP}]_{\min} - M[\bar{y}_{FTCI 3}]_{\min}$

\[
D_6 > 0 \Rightarrow M\rho^2S_Y^2 > 0
\]

(39)

which is always true.

It seems $\bar{y}_{FTCI 1}$ and $\bar{y}_{FTCI 3}$ are better offer over compromised imputation procedure and $\bar{y}_{FTCI 3}$ is best.

**Remark 9**

One can think of comparing the proposed with other existing like Hot-deck, Cold-deck, Nearest neighbour, and Regression imputation methods. But all these are based on replacing by single unit to single missing unit of sample. The proposed is on replacing single unit by an average of random sub-group units of auxiliary variate. The comparison environment in proposed case is different.

**Remark 10**

It may be interesting to derive the Horvitz-Thompson estimator based on imputation data and compare with the proposed. It is an open problem to extend the content further.

**EMPIRICAL STUDY**

The attached Appendix A has generated artificial population of size $N = 200$ containing values of main variable $Y$ and auxiliary variable $X$. Parameter of these are given as follows:

\[
\bar{Y} = 42.485; \quad \bar{X} = 18.515; \quad S_Y^2 = 199.0598; \quad S_X^2 = 48.5375; \quad \rho = 0.8652; \quad C_X = 0.3763; \quad C_Y = 0.3321.
\]

Using random sample SRSWOR of size, $n = 30$; $r = 22$;
Table 2. Bias and optimum m.s.e. at $k = k_i (i=1,2)$.

<table>
<thead>
<tr>
<th>Estimator</th>
<th>Bias (.)</th>
<th>M(.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\bar{y}_{COMP}$</td>
<td>0.0150</td>
<td>6.2504 (at $\alpha_{opt} = 0.2365$)</td>
</tr>
<tr>
<td>$\tilde{y}_{FTCI}$</td>
<td>0.0079</td>
<td>3.8254</td>
</tr>
<tr>
<td>$\tilde{y}_{FTCI}$</td>
<td>0.0109</td>
<td>3.8254</td>
</tr>
<tr>
<td>$\tilde{y}_{FTCI}$</td>
<td>0.0034</td>
<td>6.2504</td>
</tr>
<tr>
<td>$\tilde{y}_{FTCI}$</td>
<td>0.0046</td>
<td>6.2504</td>
</tr>
<tr>
<td>$\tilde{y}_{FTCI}$</td>
<td>0.0113</td>
<td>2.0225</td>
</tr>
<tr>
<td>$\tilde{y}_{FTCI}$</td>
<td>0.0156</td>
<td>2.0225</td>
</tr>
</tbody>
</table>

$t = 0.15$, $\alpha = 0.2365$. Solving optimum condition $\theta = -V$ (Equation 33), the equation of power four in $k$ provides only two real values $k_1 = 0.8350$; $k_2 = 4.1043$. Rest other two roots appear imaginary (Table 2).

**ALMOST UNBIASED FTCI ESTIMATOR:**

If $B(\tilde{y}_{FTCI}) = 0$

$\Rightarrow -\bar{Y}0M_2(\alpha_2C_x^2 - \rho C_yC_x) = 0$ Using Equation 22

When $\theta = 0 \Rightarrow (1-k)\alpha = 0$ (40)

Case (i)

$(1-k) = 0 \Rightarrow k = 1 = k_1$

Case (ii)

$\alpha = 0 \Rightarrow \alpha_1 - \alpha_2 = 0$

$\Rightarrow \frac{fB}{A + fB + C} - \frac{C}{A + fB + C} = 0$

$\Rightarrow k = k_2 = 4$ [since $(k - 4) = 0$]

$k = k_i = \frac{1}{2}(s + f) + \sqrt{f^2 + 6f + 1}$

and $k = k_i = \frac{1}{2}(s + f) - \sqrt{f^2 + 6f + 1}$

Case (iii)

When $(\alpha_2C_x^2 - \rho C_yC_x) = 0 \Rightarrow \alpha_2 = \rho \frac{C_y}{C_x} = V$

$\Rightarrow [AV + fBV + (V - 1)C] = 0$ (41)

Equation 41 is a cubic equation in $k$, solvable for fix $f$ and guess value $V$, and one can obtain three values $k = k_1^\prime$; $k = k_2^\prime$ and $k = k_3^\prime$ to make the proposed estimator imputed unbiased to the first order of approximation.

At seven values $k_1^\prime, k_2^\prime, k_3^\prime, k_4^\prime, k_1^\prime, k_2^\prime$ and $k_3^\prime$, bias is almost zero or completely zero. Using data set in Appendix A, the $k$-values are compared in Table 3. The best choice is that having the lowest m.s.e.

**DISCUSSION**

The estimator $\bar{y}_{COMP}$ of Singh and Horn (2000) bears bias 0.0150, m.s.e. 6.2504 at $\alpha_{opt} = 0.2365$. But, bias is not at minimal level. Proposed estimators have shown two choices $k_1, k_2$ on a data set for known $(f, V)$ at which m.s.e is minimum and there are open options to choose that optimum $k$-value having the lowest bias. These two choices may extend to the maximum four on some data sets. So, FTCI estimators have an upper hand over Compromised estimator of Singh and Horn (2000) in
Table 3. Almost unbiased FTCI.

<table>
<thead>
<tr>
<th>k-Values</th>
<th>$\bar{y}_{FTCI}$</th>
<th>$\bar{y}_{y}^{1}$</th>
<th>$\bar{y}_{y}^{2}$</th>
<th>$\bar{y}_{FTCI}$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Bias (.)</td>
<td>M (.)</td>
<td>Bias (.)</td>
<td>M (.)</td>
</tr>
<tr>
<td>$k_{1}$ = 1.0000</td>
<td>0</td>
<td>8.0424</td>
<td>0</td>
<td>8.0424</td>
</tr>
<tr>
<td>$k_{2}$ = 4.0000</td>
<td>0</td>
<td>8.0424</td>
<td>0</td>
<td>8.0424</td>
</tr>
<tr>
<td>$k_{3}$ = 3.2682</td>
<td>0</td>
<td>8.0424</td>
<td>0</td>
<td>8.0424</td>
</tr>
<tr>
<td>$k_{4}$ = 1.8819</td>
<td>0</td>
<td>8.0424</td>
<td>0</td>
<td>8.0424</td>
</tr>
<tr>
<td>$k_{1}^{*}$ = 1.5743</td>
<td>-0.0004</td>
<td>13.1265</td>
<td>-0.0004</td>
<td>10.1586</td>
</tr>
<tr>
<td>$k_{2}^{*}$ = 2.3355</td>
<td>-0.0004</td>
<td>56.7904</td>
<td>-0.0001</td>
<td>28.8276</td>
</tr>
<tr>
<td>$k_{3}^{*}$ = 8.8231</td>
<td>-0.0005</td>
<td>315.6901</td>
<td>-0.0002</td>
<td>139.5245</td>
</tr>
</tbody>
</table>

terms of efficiency comparison. The $\bar{y}_{FTCI}$ is equally efficient to $\bar{y}_{COMP}$ in terms of m.s.e. but has unique option of lower bias. Moreover, estimator $\bar{y}_{FTCI}$ is more efficient than all others used herein. Another observation is that, FTCI could be made almost unbiased also over seven different choices of $k$-values, the best is that having lowest m.s.e. Thus, FTCI has wider prospects of superiority over Singh and Horn (2000). One more interesting feature is that FTCI contains only one parameter $k$, same as used by Singh and Horn (2000), but mixed up in mathematical structure in such a way so as to generate more efficient properties.

ACKNOWLEDGEMENT

Authors are thankful to the referee for the critical comments and useful suggestions which has improved the quality of the manuscript.

REFERENCES


Appendix A. Artificial population (N = 200).

<table>
<thead>
<tr>
<th>$Y_i$</th>
<th>45</th>
<th>50</th>
<th>39</th>
<th>60</th>
<th>42</th>
<th>38</th>
<th>28</th>
<th>42</th>
<th>38</th>
<th>35</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_i$</td>
<td>15</td>
<td>20</td>
<td>23</td>
<td>35</td>
<td>18</td>
<td>12</td>
<td>8</td>
<td>15</td>
<td>17</td>
<td>13</td>
</tr>
<tr>
<td>$Y_i$</td>
<td>40</td>
<td>55</td>
<td>45</td>
<td>36</td>
<td>40</td>
<td>58</td>
<td>56</td>
<td>62</td>
<td>58</td>
<td>46</td>
</tr>
<tr>
<td>$X_i$</td>
<td>29</td>
<td>35</td>
<td>20</td>
<td>14</td>
<td>18</td>
<td>25</td>
<td>28</td>
<td>21</td>
<td>19</td>
<td>18</td>
</tr>
<tr>
<td>$Y_i$</td>
<td>36</td>
<td>43</td>
<td>68</td>
<td>70</td>
<td>50</td>
<td>56</td>
<td>45</td>
<td>32</td>
<td>30</td>
<td>38</td>
</tr>
<tr>
<td>$X_i$</td>
<td>15</td>
<td>20</td>
<td>38</td>
<td>23</td>
<td>25</td>
<td>18</td>
<td>11</td>
<td>09</td>
<td>17</td>
<td></td>
</tr>
<tr>
<td>$Y_i$</td>
<td>35</td>
<td>41</td>
<td>45</td>
<td>65</td>
<td>30</td>
<td>28</td>
<td>32</td>
<td>38</td>
<td>61</td>
<td>58</td>
</tr>
<tr>
<td>$X_i$</td>
<td>13</td>
<td>15</td>
<td>18</td>
<td>25</td>
<td>09</td>
<td>08</td>
<td>11</td>
<td>13</td>
<td>23</td>
<td>21</td>
</tr>
<tr>
<td>$Y_i$</td>
<td>65</td>
<td>62</td>
<td>68</td>
<td>85</td>
<td>40</td>
<td>32</td>
<td>60</td>
<td>57</td>
<td>47</td>
<td>55</td>
</tr>
<tr>
<td>$X_i$</td>
<td>27</td>
<td>25</td>
<td>30</td>
<td>45</td>
<td>15</td>
<td>12</td>
<td>22</td>
<td>19</td>
<td>17</td>
<td>21</td>
</tr>
<tr>
<td>$Y_i$</td>
<td>67</td>
<td>70</td>
<td>60</td>
<td>40</td>
<td>35</td>
<td>30</td>
<td>25</td>
<td>38</td>
<td>23</td>
<td>55</td>
</tr>
<tr>
<td>$X_i$</td>
<td>25</td>
<td>30</td>
<td>27</td>
<td>21</td>
<td>15</td>
<td>17</td>
<td>09</td>
<td>15</td>
<td>11</td>
<td>21</td>
</tr>
<tr>
<td>$Y_i$</td>
<td>50</td>
<td>69</td>
<td>53</td>
<td>55</td>
<td>71</td>
<td>74</td>
<td>55</td>
<td>39</td>
<td>43</td>
<td>45</td>
</tr>
<tr>
<td>$X_i$</td>
<td>15</td>
<td>23</td>
<td>29</td>
<td>30</td>
<td>33</td>
<td>31</td>
<td>17</td>
<td>14</td>
<td>17</td>
<td>19</td>
</tr>
<tr>
<td>$Y_i$</td>
<td>61</td>
<td>72</td>
<td>65</td>
<td>39</td>
<td>43</td>
<td>57</td>
<td>37</td>
<td>71</td>
<td>71</td>
<td>70</td>
</tr>
<tr>
<td>$X_i$</td>
<td>25</td>
<td>31</td>
<td>30</td>
<td>19</td>
<td>21</td>
<td>23</td>
<td>15</td>
<td>30</td>
<td>32</td>
<td>29</td>
</tr>
<tr>
<td>$Y_i$</td>
<td>73</td>
<td>63</td>
<td>67</td>
<td>47</td>
<td>53</td>
<td>51</td>
<td>54</td>
<td>57</td>
<td>59</td>
<td>39</td>
</tr>
<tr>
<td>$X_i$</td>
<td>28</td>
<td>23</td>
<td>23</td>
<td>17</td>
<td>19</td>
<td>17</td>
<td>18</td>
<td>21</td>
<td>23</td>
<td>20</td>
</tr>
<tr>
<td>$Y_i$</td>
<td>23</td>
<td>25</td>
<td>35</td>
<td>30</td>
<td>38</td>
<td>60</td>
<td>60</td>
<td>40</td>
<td>47</td>
<td>30</td>
</tr>
<tr>
<td>$X_i$</td>
<td>07</td>
<td>09</td>
<td>15</td>
<td>11</td>
<td>13</td>
<td>25</td>
<td>27</td>
<td>15</td>
<td>17</td>
<td>11</td>
</tr>
<tr>
<td>$Y_i$</td>
<td>57</td>
<td>54</td>
<td>60</td>
<td>51</td>
<td>26</td>
<td>32</td>
<td>30</td>
<td>45</td>
<td>55</td>
<td>54</td>
</tr>
<tr>
<td>$X_i$</td>
<td>31</td>
<td>23</td>
<td>25</td>
<td>17</td>
<td>09</td>
<td>11</td>
<td>13</td>
<td>19</td>
<td>25</td>
<td>27</td>
</tr>
<tr>
<td>$Y_i$</td>
<td>33</td>
<td>33</td>
<td>20</td>
<td>25</td>
<td>28</td>
<td>40</td>
<td>33</td>
<td>38</td>
<td>41</td>
<td>33</td>
</tr>
<tr>
<td>$X_i$</td>
<td>13</td>
<td>11</td>
<td>07</td>
<td>09</td>
<td>13</td>
<td>15</td>
<td>13</td>
<td>17</td>
<td>15</td>
<td>13</td>
</tr>
<tr>
<td>$Y_i$</td>
<td>30</td>
<td>35</td>
<td>20</td>
<td>18</td>
<td>20</td>
<td>27</td>
<td>23</td>
<td>42</td>
<td>37</td>
<td>45</td>
</tr>
<tr>
<td>$X_i$</td>
<td>11</td>
<td>15</td>
<td>08</td>
<td>07</td>
<td>09</td>
<td>13</td>
<td>12</td>
<td>25</td>
<td>21</td>
<td>22</td>
</tr>
<tr>
<td>$Y_i$</td>
<td>37</td>
<td>37</td>
<td>37</td>
<td>34</td>
<td>41</td>
<td>35</td>
<td>39</td>
<td>45</td>
<td>24</td>
<td>27</td>
</tr>
<tr>
<td>$X_i$</td>
<td>15</td>
<td>16</td>
<td>17</td>
<td>13</td>
<td>20</td>
<td>15</td>
<td>21</td>
<td>25</td>
<td>11</td>
<td>13</td>
</tr>
<tr>
<td>$Y_i$</td>
<td>23</td>
<td>20</td>
<td>26</td>
<td>26</td>
<td>40</td>
<td>56</td>
<td>41</td>
<td>47</td>
<td>43</td>
<td>33</td>
</tr>
<tr>
<td>$X_i$</td>
<td>09</td>
<td>08</td>
<td>11</td>
<td>12</td>
<td>15</td>
<td>25</td>
<td>15</td>
<td>25</td>
<td>21</td>
<td>15</td>
</tr>
<tr>
<td>$Y_i$</td>
<td>37</td>
<td>27</td>
<td>21</td>
<td>23</td>
<td>24</td>
<td>21</td>
<td>39</td>
<td>33</td>
<td>25</td>
<td>35</td>
</tr>
<tr>
<td>$X_i$</td>
<td>17</td>
<td>13</td>
<td>11</td>
<td>11</td>
<td>09</td>
<td>08</td>
<td>15</td>
<td>17</td>
<td>11</td>
<td>19</td>
</tr>
<tr>
<td>$Y_i$</td>
<td>45</td>
<td>40</td>
<td>31</td>
<td>20</td>
<td>40</td>
<td>50</td>
<td>45</td>
<td>35</td>
<td>30</td>
<td>35</td>
</tr>
<tr>
<td>$X_i$</td>
<td>21</td>
<td>23</td>
<td>15</td>
<td>11</td>
<td>20</td>
<td>25</td>
<td>23</td>
<td>17</td>
<td>16</td>
<td>18</td>
</tr>
<tr>
<td>$Y_i$</td>
<td>32</td>
<td>27</td>
<td>30</td>
<td>33</td>
<td>31</td>
<td>47</td>
<td>43</td>
<td>35</td>
<td>30</td>
<td>40</td>
</tr>
<tr>
<td>$X_i$</td>
<td>15</td>
<td>13</td>
<td>14</td>
<td>17</td>
<td>15</td>
<td>25</td>
<td>23</td>
<td>17</td>
<td>16</td>
<td>19</td>
</tr>
<tr>
<td>$Y_i$</td>
<td>35</td>
<td>35</td>
<td>46</td>
<td>39</td>
<td>35</td>
<td>30</td>
<td>31</td>
<td>53</td>
<td>63</td>
<td>41</td>
</tr>
<tr>
<td>$X_i$</td>
<td>19</td>
<td>19</td>
<td>23</td>
<td>15</td>
<td>17</td>
<td>13</td>
<td>19</td>
<td>25</td>
<td>35</td>
<td>21</td>
</tr>
<tr>
<td>$Y_i$</td>
<td>52</td>
<td>43</td>
<td>39</td>
<td>37</td>
<td>20</td>
<td>23</td>
<td>35</td>
<td>39</td>
<td>45</td>
<td>37</td>
</tr>
<tr>
<td>$X_i$</td>
<td>25</td>
<td>19</td>
<td>18</td>
<td>17</td>
<td>11</td>
<td>09</td>
<td>15</td>
<td>17</td>
<td>19</td>
<td>19</td>
</tr>
</tbody>
</table>