Journal of Modern Applied Statistical Methods

Volume 20 | Issue 1 Article 4

A Super Population Model Approach to Imputation for Estimating Population Mean

Priyanka Singh
Assistant Professor (Research), SRM Institute of Science and technology, India

Ajeet Kumar Singh
Assistant Professor, University of Rajasthan, India

V. K. Singh
Banaras Hindu University, India

Recommended Citation

Priyanka Singh, Ajeet Kumar Singh, V. K. Singh (2021). A Super Population Model Approach to Imputation for Estimating Population Mean. Journal of Modern Applied Statistical Methods, 20(1), https://doi.org/10.56801/Jmasm.V20.i1.4

Journal of Modern Applied Statistical Methods
2021, Vol. 20, No. 1,
Doi: 10.56801/Jmasm.V20.i1.4

ISSN 1538 – 9472

A Super Population Model Approach to Imputation for Estimating Population Mean

Priyanka Singh
SRM Institute of Science and technology, India

Ajeet Kumar Singh
University of Rajasthan, India

V. K. Singh
Banaras Hindu University, India

In this paper Super population model based study of some imputation strategies have been proposed for estimating population mean under non -response, considering population is of dynamic nature, a more practical and relevant approach and hence it has been assumed that each unit of the study variable is following Polynomial regression model. Bias's and MSE's have been obtained under model-based approach; also, empirical comparisons of these strategies have been done with some special cases of PRMs, so as to observe their performance over varying non-response rates, Robustness of the estimators have been checked under misspecification of the models.

Keywords: Non-response, Imputation, Polynomial regression model, balanced sample, misspecification of models.

1. Introduction

Non-response is a form of non-observation present in all kinds of surveys.Although non-response cannot be completely eliminated in practice, it could be overcome to a great extent by persuasion through developing special estimation techniques giving due consideration to the incomplete (or missing) data.Since a long time, various techniques for assessing and controlling non-sampling errors and especially non-response errors have been developed by several authors. A good deal of discussion, classification and illustrations of various types of errors have been given by Deming (1944, 1950), Mahalanobis (1940, 1944, 1946), Moser (1958), Zarkovich (1966) and Dalenius (1977a, 1977b, 1977c).

A deterministic model assumes that the population can be thought of comprising of two strata, namely, "response stratum" and "non-response stratum". However, besides its popular use by several authors, the model was considered as most simplistic and unrealistic. Later on, survey statisticians relied heavily on the conditional distribution of the response indicator R given the complete data X^* and accordingly defined a number of response models, such as, MAR, OAR, MCAR and NMAR (Rubin, 1976; Little and Rubin, 1987; Longford, 2005).

Imputation is one of the techniques of adjusting non-response which directly or indirectly assumes certain response models for its implementation. It is a technique of filling-in the missing (incomplete) values by some suitable known values in the sample or by some appropriate functions of these values in order to fill-up the "holes" in the incomplete data matrix, caused by missing values.Imputation is generally used to recreate a balanced design such that procedures used for analyzing complete data can be applied in many situations.

1.1 The Super Population Concept in Survey Sampling

In this type of approach, it is assumed that with each population unit is associated a random variable for which a stochastic structure is specified; the actual value associated with a population unit is treated as the outcome of this random variable. The Y-vector of population values is assumed to be generated from a distribution ξ, where ξ is known to be a member of class $C=\{\xi\}$. The class C is then called a super population model.

In the literature of model-based survey sampling, different kinds of super populations models were considered. For example, Brewer (1963a) and Royall (1970a, 1970b, 1971, 1976) adapted a linear model prediction theory to the finite population situation, while Cassel et al (1976, 1977) and Sarndal (1980b) proposed a generalized regression predictor that is asymptotically design unbiased (ADU). Isaki and Fuller (1982) proposed some ADU predictors involving several auxiliary variables. Wright (1983) examined a regression super population model and suggested a new class of predictors to link certain features of optimal design-unbiased and model-unbiased predictors.

1.2 Polynomial Regression Model (PRM)

A particular form of super population model, namely, **"Polynomial Regression Model (PRM)"** propounded by Royall and Herson (1973a, 1973b), described as follows:

$$Y_k = \delta_0\beta_0 + \delta_1\beta_1 x_k + \delta_2\beta_2 x_k^2 + \cdots + \delta_J\beta_J x_k^J + \varepsilon_k[v(x_k)]^{1/2}$$

$$= \sum_{j=0}^{J} \delta_j\beta_j\, x_k^j + \varepsilon_k[v(x_k)]^{1/2}; k = 1, 2, \ldots, N$$

$$= h(x_k) + \varepsilon_k[v(x_k)]^{1/2}; k = 1, 2, \ldots, N \qquad (1)$$

with $E_\xi[Y_k] = h(x_k) = \sum_{j=0}^{J} \delta_j\beta_j\, x_k^j$;

$Var[Y_k] = \sigma^2 v(x_k)$; $Cov(Y_r, Y_k) = 0, r \neq k$

where Y_k is the random variable associated with the k^{th} unit of the finite population of size N, x_k is the value of the k^{th} unit of the population on the known auxiliary variable X, typically referred to as their measures of size ($x_k > 0\ for\ k = 1, 2, \dots, N$), ϵ_k for $k = 1, 2, \dots, N$ are independent random variables each having mean zero and variance σ^2, δ_j ($j = 0,1, \dots J$) is zero or one according as the term is absent or present respectively in the model, $v(x_k)$is a known function of x-values and β_j ($j = 0,1, \dots J$) are unknown model parameters. Royall and Herson (1973a) denoted this model as $\xi[\delta_0\delta_1, \dots, \delta_J : v(x)]$. For example, $\xi[0,1:x]$ and $\xi[1,1,0,1:1]$ refer respectively to the models

$$Y_k = \beta_1 x_k + \epsilon_k x_k^{1/2} \quad (2)$$

and

$$Y_k = \beta_0 + \beta_1 x_k + \beta_3 x_k^3 + \epsilon_k \quad (3)$$

The importance of PRM can be understood due to the fact that the standard ratio estimator of population total becomes ξ–unbiased under the model $\xi[0,1:x]$for a given sample.

1.3 Initiation of the problem

Let a finite population of size N, denoted by Ω, consists of N_1 respondent and N_2 non-respondent units ($N = N_1 + N_2$) and a random sample of size n, denoted by s, drawn from the population Ω, consists of $n_1$respondents and $n_2$non-respondents such that $n = n_1 + n_2$. Further, let $\Omega = s \cup \bar{s}$, where s and $\bar{s}$be two disjoint sets of units, such that s represents the observed part of Ω in the form of sample drawn and $\bar{s}$ represents the non-observed part of the population. Similarly, let $s = s_1 \cup s_2$($s_1$and $s_2$being disjoint sets of units belonging to s), where s_1 is the set of units for which the information are observed and s_2 consists of missing data, that is, the units for which observations were not obtained. Let $\sum_p$stands for the summation over the set (subset) p of units.

Notations used:

(i) Population Values

Z: Study variable Y or auxiliary variable X,

z_k: the k^{th} value of Z,

$\bar{Z} = N^{-1} \sum_\Omega z_k$: the population mean of Z,

$S_Z^2 = (N-1)^{-1} \sum_\Omega (z_k - \bar{Z})^2$: Population mean square of Z,

$C_Z = S_Z/\bar{Z}$: Coefficient of variation of Z,

ρ_{YX}: correlation coefficient in the population between Y and X.

(ii) Sample Values

$\bar{z}_s = n^{-1}\sum_s z_k; \bar{z}_{\bar{s}} = (N-n)^{-1}\sum_{\bar{s}} z_k;\ \bar{z}_{s_i} = {n_i}^{-1}\sum_{s_i} z_k$ for $(i = 1, 2)$

(iii) Higher order moments

Further, let for the variable X and j=1, 2, 3, ...

$\bar{X}^{(j)} = N^{-1}\sum_\Omega x_k^j; \bar{x}^{(j)} = n^{-1}\sum_s x_k^j; \bar{x}_{\bar{s}}^{(j)} = (N-n)^{-1}\sum_{\bar{s}} x_k^j;$

$$\bar{x}_{s_i} = {n_i}^{-1}\sum_{s_i} x_k^j \text{ for } (i = 1, 2)$$

It has already been observed that

$\bar{X}^{(1)} = \bar{X};\ \bar{x}^{(1)} = \bar{x}_s; \bar{x}_{s_i}{}^{(1)} = \bar{x}_{s_i}$ for (i=1, 2) and $\bar{x}_{\bar{s}}^{(1)} = \bar{x}_{\bar{s}}$

Further, let $y_{.k}$ be the imputed value for the k^{th} unity$_k$ $(k = 1,2, \dots, n)$;

$\{B_\xi(T), M_\xi(T)\}$ be the bias and MSE of the estimator T under the model ξ respectively.

Imputation strategies considered

The following imputation strategies have been considered:

(1) Mean Method of Imputation

Under this method of imputation, the study variate after imputation takes the form

$$y_{.k} = \begin{cases} y_k & \text{if } k \in s_1 \\ \bar{y}_{s_1} & \text{if } k \in s_2 \end{cases} \quad (4)$$

Since, the sample mean $\bar{y}_s = \frac{1}{n}\{\sum_{s_1} y_k + \sum_{s_2} y_k\} = \frac{1}{n}\{n_1\bar{y}_{s_1} + n_2\bar{y}_{s_1}\} = \bar{y}_{s_1}$, therefore, the point estimator for estimating the population mean $\bar{Y}$, under this scheme would be $\bar{y}_M = \bar{y}_{s_1}$, which is the sample mean of the respondent group.

(2) Ratio Method of Imputation

In the ratio method of imputation, imputation is carried out with the aid of an auxiliary variable X, such that x_k, the value of X for unit k is known and positive for every $k \in s$. Following the notation of Lee et al (1994, 1995) in the case of single value imputation, if the k^{th} unit requires imputation, the value $\hat{b}x_k$ is imputed, where $\hat{b} = \frac{\sum_{s_1} y_k}{\sum_{s_1} x_k}$. Then the data after imputation becomes

$$y_{.k} = \begin{cases} y_k & \text{if } k \in s_1 \\ \hat{b}x_k & \text{if } k \in s_2 \end{cases} \quad (5)$$

Since the general point estimator of population mean takes the form

$$\bar{y}_s = \bar{y}_{s_1}\frac{\bar{x}_s}{\bar{x}_{s_1}} \quad (6)$$

Let us denote this point estimator by $\bar{y}_{RAT}$. Then, obviously $\bar{y}_{RAT} = \bar{y}_{s_1}\frac{\bar{x}_s}{\bar{x}_{s_1}}$ is a ratio estimator defined with the help of observed part of the sample and known sample mean for the variable X.

(3) Compromised Method of Imputation

Singh and Horn (2000) suggested a compromised imputation in which the data after imputation becomes

$$y_{.k} = \begin{cases} \alpha\frac{n}{n_1}y_k + (1-\alpha)\hat{b}x_k & \text{if } k \in s_1 \\ (1-\alpha)\hat{b}x_k & \text{if } k \in s_2 \end{cases} \quad (7)$$

Where α is a suitably chosen constant, such that the variance of the resultant estimator is minimum.

It can be seen that the resultant point estimator for estimating population mean is given by

$$\bar{y}_{COMP} = \alpha\bar{y}_{s_1} + (1-\alpha)\bar{y}_{s_1}\frac{\bar{x}_s}{\bar{x}_{s_1}} \quad (8)$$

2. Proposed Imputation Strategy (Exponential-Type Imputation Methods):

Recently, Asghar et al (2014) generalized the Bahl and Tuteja (1991) exponential type estimator as follows:

$$t = \lambda s_y^2 \exp\left[\alpha\left(\frac{\bar{X}-\bar{x}_s}{\bar{X}+(a-1)\bar{x}_s}\right)\right] \quad (9)$$

for estimating the finite population variance, where $0 < \lambda \leq 1$, $-\infty < \alpha < \infty$ and $a > 0$. Motivated by (9), the following imputation method has been proposed, for filling-in the missing data in order to estimate population mean $\bar{Y}$:

$$y_{.k} = \begin{cases} \lambda y_k & \text{if } k \in s_1 \\ \lambda\frac{\bar{y}_{s_1}}{n_2}\left[n\,\psi\left(\alpha, a, \bar{x}_{s_1}, \bar{X}\right) - n_1\right] & \text{if } k \in s_2 \end{cases} \quad (10)$$

where

$$\psi\left(\alpha, a, \bar{x}_{s_1}, \bar{X}\right) = \exp\left[\alpha\left(\frac{\bar{X}-\bar{x}_{s_1}}{\bar{X}+(a-1)\bar{x}_{s_1}}\right)\right] \quad (11)$$

It is easy to observe that the corresponding point estimator for population mean then would be

$$\bar{y}_e = \lambda \bar{y}_{s_1} \psi(\alpha, a, \bar{x}_{s_1}, \bar{X}) \tag{12}$$

2.1 ξ-Bias and ξ-MSE of point estimators $\bar{y}_M, \bar{y}_{RAT}, \bar{y}_{COMP}, \bar{y}_e$

Let us first consider the mean method of imputation and corresponding point estimator $\bar{y}_M$ for population mean. We denote by $[\xi, \bar{y}_M]$ the imputation strategy under the model ξ, which is PRM here.

We have the following theorems:

Theorem 1.: The ξ-bias, of the estimator $\bar{y}_M$ under the model ξ, denoted by $B_\xi[\bar{y}_M]$, is given by

$$B_\xi[\bar{y}_M] = \sum_{j=0}^{J} \delta_j \beta_j \left[\{\bar{x}_{s_1}{}^{(j)} - \bar{X}^{(j)}\}\right] \tag{13}$$

Proof: We have

$B_\xi[\bar{y}_M] = E_\xi[\bar{y}_M - \bar{Y}]$, by definition

Now using the ξ-model,

$$B_\xi[\bar{y}_M] = E_\xi[\bar{y}_{s_1} - \bar{Y}]$$

$$= E_\xi\left[\frac{1}{n_1}\sum_{s_1} y_k - \frac{1}{N}\sum_{\Omega} y_k\right]$$

$$= E_\xi\left[\frac{1}{n_1}\sum_{s_1}\left\{\sum_{j=0}^{J} \delta_j \beta_j x_k^j + \epsilon_k \left(v(x_k)\right)^{1/2}\right\} - \frac{1}{N}\sum_{\Omega}\left\{\sum_{j=0}^{J} \delta_j \beta_j x_k^j + \epsilon_k \left(v(x_k)\right)^{1/2}\right\}\right]$$

$$= \sum_{j=0}^{J} \delta_j \beta_j \frac{1}{n_1}\sum_{s_1} x_k^j - \sum_{j=0}^{J} \delta_j \beta_j \frac{1}{N}\sum_{\Omega} x_k^j$$

since $E_\xi(\epsilon_k) = 0$.

Now since $\frac{1}{n_1}\sum_{s_1} x_k^j = \bar{x}_{s_1}{}^{(j)}$ and $\frac{1}{N}\sum_{\Omega} x_k^j = \bar{X}^{(j)}$,

we have

$$B_\xi[\bar{y}_M] = \sum_{j=0}^{J} \delta_j \beta_j \left[\{\bar{x}_{s_1}{}^{(j)} - \bar{X}^{(j)}\}\right]$$

Thus, the expression follows.

Theorem 2.: The ξ-MSE of the estimator $\bar{y}_M$, denoted by $M_\xi[\bar{y}_M]$ under the model ξ is given by

$$M_\xi[\bar{y}_M] = \left[\sum_{j=0}^{J} \delta_j \beta_j \{\bar{x}_{s_1}{}^{(j)} - \bar{X}^{(j)}\}\right]^2 + \left(\frac{1}{n_1} - \frac{1}{N}\right)^2 \sigma^2 \sum_{s_1} v(x_k)$$

$$+ \frac{\sigma^2}{N^2}\left[\sum_{s_2} v(x_k) + \sum_{\bar{s}} v(x_k)\right] \tag{14}$$

Proof: As per definition of MSE, $M_\xi[\bar{y}_M] = E_\xi[\bar{y}_M - \bar{Y}]^2 = E_\xi\left[\frac{1}{n_1}\sum_{s_1} y_k - \frac{1}{N}\sum_{\Omega} y_k\right]^2$

$$= E_\xi\left[\frac{1}{n_1}\sum_{s_1}\left\{\sum_{j=0}^{J} \delta_j \beta_j x_k^j + \epsilon_k (v(x_k))^{1/2}\right\} - \frac{1}{N}\sum_{\Omega}\left\{\sum_{j=0}^{J} \delta_j \beta_j x_k^j + \epsilon_k (v(x_k))^{1/2}\right\}\right]^2$$

$$= E_\xi\left[\sum_{j=0}^{J} \delta_j \beta_j \left\{\frac{1}{n_1}\sum_{s_1} x_k^j - \frac{1}{N}\sum_{\Omega} x_k^j\right\} + (v(x_k))^{1/2}\left\{\frac{1}{n_1}\sum_{s_1} \varepsilon_k - \frac{1}{N}\sum_{\Omega} \varepsilon_k\right\}\right]^2$$

Now realizing that $E_\xi(\varepsilon_k^2) = \sigma^2$ and $E_\xi(\varepsilon_k, \varepsilon_r) = 0$; for $r \neq k$, expanding the term and taking ξ expectation for each term, we get the following expression.

$$M_\xi[\bar{y}_M] = \sum_{j=0}^{J} \delta_j \beta_j [\{\bar{x}_{s_1}{}^{(j)} - \bar{X}^{(j)}\}]^2 + \left(\frac{1}{n_1} - \frac{1}{N}\right)^2 \sigma^2 \sum_{s_1} v(x_k)$$

$$+ \frac{\sigma^2}{N^2}\left[\sum_{s_2} v(x_k) + \sum_{\bar{s}} v(x_k)\right]$$

Remark1.: The expression $M_\xi[\bar{y}_M]$ is a sum of two components, namely the term independent of variance function $\sigma^2$and the other dependent upon the varianceσ^2. The first component depends only upon the polynomial function$h(x_k)$.

The expressions of ξ-MSE of other imputation strategies, namely, $[\xi, \bar{y}_{RAT}]$, $[\xi, \bar{y}_{COMP}]$ and $[\xi, \bar{y}_e]$ can be obtained in similar manner

Theorem 3.: The ξ-bias and ξ-MSE of the strategy $[\xi, \bar{y}_{RAT}]$ are given by

$$B_\xi[\bar{y}_{RAT}] = \sum_{j=0}^{J} \delta_j \beta_j [\varphi(x)\bar{x}_{s_1}{}^{(j)} - \bar{X}^{(j)}] \tag{15}$$

where $\varphi(x) = \frac{\bar{x}_s}{\bar{x}_{s_1}}$

$$M_\xi[\bar{y}_{RAT}] = [B_\xi[\bar{y}_{RAT}]]^2 + \left(\frac{\varphi(x)}{n_1} - \frac{1}{N}\right)^2 \sigma^2 \sum_{s_1} v(x_k) + \frac{\sigma^2}{N^2}\left[\sum_{s_2} v(x_k) + \sum_{\bar{s}} v(x_k)\right] \quad (16)$$

Theorem 4.: The expressions of bias and MSE of the strategy $[\xi, \bar{y}_{COMP}]$ are

$$B_\xi[\bar{y}_{COMP}] = \sum_{j=0}^{J} \delta_j \beta_j \left[\bar{x}_{s_1}{}^{(j)}\{\alpha + (1-\alpha)\varphi(x)\} - \bar{X}^{(j)}\right] \quad (17)$$

$$M_\xi[\bar{y}_{COMP}] = \left[B_\xi[\bar{y}_{COMP}]\right]^2 + \left(\frac{\alpha}{n_1} + \frac{(1-\alpha)\varphi(x)}{n_1} - \frac{1}{N}\right)^2 \sigma^2 \sum_{s_1} v(x_k) + \frac{\sigma^2}{N^2}\left[\sum_{s_2} v(x_k) + \sum_{\bar{s}} v(x_k)\right] \quad (18)$$

Remark 2.: Using the expression (18), the optimum value of the parameter α can be obtained as

$$\alpha_0 = \frac{\left[\begin{array}{c}\sum_{j=0}^{J} \delta_j \beta_j \bar{x}_{s_1}{}^{(j)}\left\{\sum_{j=0}^{J} \delta_j \beta_j \bar{X}^{(j)} - \sum_{j=0}^{J} \delta_j \beta_j \varphi(x) \bar{x}_{s_1}{}^{(j)}\right\} - \\ \frac{1}{n_1}\left(\frac{\varphi(x)}{n_1} - \frac{1}{N}\right)\sigma^2 \sum_{s_1} v(x_k)\end{array}\right]}{\left[\{1-\varphi(x)\}\left\{\frac{\sigma^2}{n_1{}^2}\sum_{s_1} v(x_k) + \left(\sum_{j=0}^{J} \delta_j \beta_j \bar{x}_{s_1}{}^{(j)}\right)^2\right\}\right]} \quad (19)$$

Theorem 5.: The expressions of ξ-bias and ξ-MSE of the strategy$[\xi, \bar{y}_e]$are given by

$$B_\xi[\bar{y}_e] = \sum_{j=0}^{J} \delta_j \beta_j \left[\lambda\psi(\alpha, a, \bar{x}_{s_1}, \bar{X})\bar{x}_{s_1}{}^{(j)} - \bar{X}^{(j)}\right] \quad (20)$$

$$M_\xi[\bar{y}_e] = [B_\xi\{\bar{y}_e\}]^2 + \left(\frac{\lambda\psi(\alpha, a, \bar{x}_{s_1}, \bar{X})}{n_1} - \frac{1}{N}\right)^2 \sigma^2 \sum_{s_1} v(x_k) + \frac{\sigma^2}{N^2}\left[\sum_{s_2} v(x_k) + \sum_{\bar{s}} v(x_k)\right] \quad (21)$$

Remark 3.: The optimum values of the parameter λ, involved in $M_\xi[\bar{y}_e]$can be obtained respectively as

$$\lambda = \frac{\begin{bmatrix} \sum_{j=0}^{J} \delta_j \beta_j \, \bar{x}_{s_1}{}^{(j)} \left\{ \sum_{j=0}^{J} \delta_j \beta_j \, \bar{X}^{(j)} \right\} + \\ \frac{1}{N n_1} \sigma^2 \sum_{s_1} v(x_k) \end{bmatrix}}{\left[\left\{ \psi(\alpha, a, \bar{x}_{s_1}, \bar{X}) \right\} \left\{ \frac{\sigma^2}{{n_1}^2} \sum_{s_1} v(x_k) + (\sum_{j=0}^{J} \delta_j \beta_j \, \bar{x}_{s_1}{}^{(j)})^2 \right\} \right]} \tag{22}$$

2.2 Some specific cases of PRM and corresponding Bias and MSE of strategies

Due to involvement of a large number of super population parameters in the model, practically it is not possible to study the nature and salient characteristics of the estimators with such a generalized model. It is, therefore, sometimes necessary to consider in practice some simplified versions of PRM which might involve lesser number of parameters. With this view, as an example one may consider the following simple PRMs:

Model I: $\xi[0,1: x^g] \rightarrow Y_k = \beta_1 x_k + \epsilon_k x_k^{g/2}$ (23)

Model II: $\xi[1,1: x^g] \rightarrow Y_k = \beta_0 + \beta_1 x_k + \epsilon_k x_k^{g/2}$ (24)

Model III: $\xi[1,1,1: x^g] \rightarrow Y_k = \beta_0 + \beta_1 x_k + \beta_2 x_k^2 + \epsilon_k x_k^{g/2}$ (25)

Model IV: $\xi[1,1: x^2] \rightarrow Y_k = \beta_0 + \beta_1 x_k + \epsilon_k x_k^2$ (26)

Model V: $\xi[0,1: x^2] \rightarrow Y_k = \beta_1 x_k + \epsilon_k x_k^2$ (27)

where the constant g is usually unknown in practice. However, Cochran (1953) and Brewer (1963b) have shown that majority of the situations occurring in practice might be covered by assuming that $0 \leq g \leq 2$ (or perhaps even the narrower interval$1 \leq g \leq 2$). The ξ-bias and ξ-MSE expressions of the suggested estimators under model I for specific value of g=0 has been given:

2.3 Bias and MSE of model I: $\boldsymbol{\xi[0,1: x^g]}$ with g=0

$B_\xi[\bar{y}_M] = [\beta_1\{\bar{x}_{s_1} - \bar{X}\}]$ (28)

$M_\xi[\bar{y}_M] = [\beta_1\{\bar{x}_{s_1} - \bar{X}\}]^2 + \left(\frac{1}{n_1} - \frac{1}{N}\right)^2 \sigma^2 n_1 + \frac{\sigma^2}{N^2}[N - n_1]$ (29)

$B_\xi[\bar{y}_{RAT}] = [\beta_1\{\varphi(x)\bar{x}_{s_1} - \bar{X}\}]$ (30)

$M_\xi[\bar{y}_{RAT}] = [\beta_1\{\varphi(x)\bar{x}_{s_1} - \bar{X}\}]^2 + \left(\frac{\varphi(x)}{n_1} - \frac{1}{N}\right)^2 \sigma^2 n_1 + \frac{\sigma^2}{N^2}[N - n_1]$ (31)

$B_\xi[\bar{y}_{COMP}] = [\beta_1\{\alpha\bar{x}_{s_1} + (1-\alpha)\varphi(x)\bar{x}_{s_1} - \bar{X}\}]$ (32)

$M_\xi[\bar{y}_{COMP}] = \left[B_\xi[\bar{y}_{COMP}]\right]^2 + \left(\frac{\alpha}{n_1} + \frac{(1-\alpha)\varphi(x)}{n_1} - \frac{1}{N}\right)^2 \sigma^2 n_1$

$$+\frac{\sigma^2}{N^2}[N-n_1] \tag{33}$$

with

$$\alpha_0 = \frac{\beta_1 \bar{x}_{s_1}\{\beta_1 \bar{X} - \beta_1 \bar{x}_{s_1}\} - \frac{1}{n_1}\left(\frac{\varphi(x)}{n_1} - \frac{1}{N}\right)^2 \sigma^2 n_1}{\{1-\varphi(x)\}\left\{\frac{n_1\sigma^2}{n^2} + (\beta_1 \bar{x}_{s_1})^2\right\}} \tag{34}$$

$$B_\xi[\bar{y}_e] = [\beta_1\{\lambda\psi(\alpha, a, \bar{x}_{s_1}, \bar{X})\bar{x}_{s_1} - \bar{X}\}] \tag{35}$$

$$M_\xi[\bar{y}_e] = \left[B_\xi[\bar{y}_e]\right]^2 + \left(\frac{\lambda\psi(\alpha, a, \bar{x}_{s_1}, \bar{X})}{n_1} - \frac{1}{N}\right)^2 \sigma^2 n_1 + \frac{\sigma^2}{N^2}[N-n_1] \tag{36}$$

with

$$\lambda = \frac{\beta_1 \bar{x}_{s_1}\{\beta_1 \bar{X}\} + \frac{1}{Nn_1}\sigma^2 n_1}{\psi(\alpha, a, \bar{x}_{s_1}, \bar{X})\left\{\frac{\sigma^2}{n_1} + (\beta_1 \bar{x}_{s_1})^2\right\}} \tag{37}$$

2.4 Description of the Data:

For the analysis of the results empirically,a real population from Kish (1967) have been. The data has been presented in the Appendix E of Kish (1967). Let X be the number of dwellings, whereas Y be the dwelling occupied by renters.In order to generate a finite population for analysis purpose, we have considered only 90 pairs of values $(x_{k,}y_k)$ from $i = 232\ to\ 449$ excluding $i = 375, 377$ and 384 for which y_k values were zero. Accordingly, following population values have been obtained:

$N = 90,\ \bar{X} = 41.4556, \bar{Y} = 30.511,\ S_Y^2 = 598.237, S_X^2 = 747.854, \rho_{XY} = 0.982$

2.5 Selection of Samples

From the population, a random sample of size 20 (that is, $n = 20$) was selected. The sample of size 20 was then sampled again in order to get samples $s_1 and s_2$with 10%, 20%, and 30% non-response rates in the sample $s$that is, with (i) $n_1 = 18\ and n_2 = 2$ (ii) $n_1 = 16\ and n_2 = 4$ and (iii)$n_1 = 14\ and n_2 = 6$. These samples are referred as sample-I, sample-II, sample-III respectively. The configuration of these samples in respect to X- values are as follows:

s: 67, 61, 21, 14, 23, 45, 43, 30, 45, 5, 18, 89, 25, 61, 37, 89, 48, 20, 26, 24

Sample-I ($n_1 = 18\ and n_2 = 2$)

s_1 =67, 61, 21, 14, 23, 43, 30, 45, 5, 89, 25, 61, 37, 89, 48, 20, 26, 24

s_2 = 45, 18

Sample-II ($n_1 = 16\ and n_2 = 4$)

s_1 =61, 21, 14, 23, 45, 43, 30, 18, 89, 25, 61, 37, 89, 48, 26, 24

$s_2 = 67, 45, 5, 20$

Sample-III($n_1 = 14\ and n_2 = 6$)

s_1 =61, 45, 48, 89, 67, 37, 43, 14, 26, 20, 61, 21, 18, 30

$s_2 = 23, 45, 5, 89, 25, 24$

Empirical comparison of different imputation strategies

In addition to the configuration of the samples given, the following results under each sample plan have been obtained:

Sample I:$\sum_{s_1} x_k^2 = 39688, \sum_{s_2} x_k^2 = 2349, \sum_{\bar{s}} x_k^2 = 179196$

Sample II: $\sum_{s_1} x_k^2 = 35098, \sum_{s_2} x_k^2 = 6939, \sum_{\bar{s}} x_k^2 = 179196$

Sample III: $\sum_{s_1} x_k^2 = 30336, \sum_{s_2} x_k^2 = 11701, \sum_{\bar{s}} x_k^2 = 179196$

From Singh (2016), for the same data, we have$\beta_0 = 0.8787, \beta_1 = -4.9157$ and $\sigma^2 = 0.7998$. The comparison of suggested imputation strategies, namely, $[\xi, \bar{y}_M], [\xi, \bar{y}_{RAT}], [\xi, \bar{y}_{COMP}]$ and $[\xi, \bar{y}_e]$has been made under model I and model II, for $g = 0,1\ and\ 2$ for all the four samples, that is, when the non-response rates in the sample are $10\%,\ 20\%\ and\ 30\%$. Tables 1, 2 and 3 depict the absolute values of ξ-MSEs and PREs of the estimators with respect to the strategy $[\xi, \bar{y}_M]$for different non-response rates.

Table 1. Absolute Value of MSE and PRE (in parentheses) of Different Strategies

ξ-MSEs of Estimators	**(Non-Response Rate 10 %)**					
	Model I			**Model II**		
	g			**g**		
	0	**1**	**2**	**0**	**1**	**2**
$\bar{y}_M$	24.740 **(100.00)**	26.151 **(100.00)**	105.331 **(100.00)**	24.740 **(100.00)**	26.151 **(100.00)**	105.331 **(100.00)**
$\bar{y}_{RAT}$	87.777 **(28.185)**	89.127 **(29.341)**	164.952 **(63.855)**	87.414 **(28.302)**	88.764 **(29.461)**	164.588 **(63.997)**
$\bar{y}_{COMP}$	3.735*10^{-2} **(662.35*10^2)**	1.520 **(1720.460)**	84.576 **(124.540)**	3.735*10^{-2} **(662.35*10^2)**	1.520 **(1720.460)**	84.594 **(124.514)**
α_o	**2.130**	**2.130**	**2.121**	**2.135**	**2.135**	**2.126**
$\bar{y}_e$	3.735*10^{-2} **(662.35*10^2)**	1.520 **(1720.460)**	84.442 **(124.738)**	3.735*10^{-2} **(662.35*10^2)**	1.520 **(1720.460)**	84.459 **(124.713)**
λ	**1.000**	**1.000**	**0.998**	**1.000**	**1.000**	**0.998**

Table 2. Absolute Value of MSE and PRE (in parentheses) of Different Strategies

ξ-MSEs of Estimators	**(Non-Response Rate 20 %)**					
	Model I			**Model II**		
	g			**g**		
	0	**1**	**2**	**0**	**1**	**2**
$\bar{y}_M$	8.185 **(100.0)**	9.829 **(100.00)**	100.655 **(100.0)**	8.185 **(100.00)**	9.829 **(100.00)**	100.655 **(100.00)**
$\bar{y}_{RAT}$	87.782	89.322	174.523	87.249	88.789	173.990

	(9.324)	**(11.004)**	**(57.674)**	**(9.381)**	**(11.070)**	**(57.851)**
$\bar{y}_{COMP}$	$4.228*10^{-2}$ **$(193.590*10^{2})$**	1.733 **(567.166)**	95.053 **(105.893)**	$4.228*10^{-2}$ **$(193.590*10^{2})$**	1.733 **(567.167)**	95.064 **(105.881)**
α_o	**1.438**	**1.438**	**1.431**	**1.440**	**1.440**	**1.433**
$\bar{y}_e$	$4.228*10^{-2}$ **$(193.590*10^{2})$**	1.733 **(567.166)**	94.886 **(106.080)**	$4.228*10^{-2}$ **$(193.590*10^{2})$**	1.733 **(567.166)**	94.896 **(106.069)**
λ	**1.000**	**1.000**	**0.998**	**1.000**	**1.000**	**0.998**

Table 3. Absolute Value of MSE and PRE (in parentheses) of Different Strategies

ξ-MSEs of Estimators	**(Non-Response Rate 30 %)**					
	Model I			**Model II**		
	g			**g**		
	0	**1**	**2**	**0**	**1**	**2**
$\bar{y}_M$	$6.584*10^{-2}$ **(100.00)**	2.016 **(100.00)**	107.140 **(100.00)**	$6.584*10^{-2}$ **(100.00)**	2.016 **(100.00)**	107.140 **(100.00)**
$\bar{y}_{RAT}$	87.787 **(0.075)**	89.566 **(2.251)**	185.640 **(57.983)**	87.042 **(0.076)**	88.821 **(2.269)**	184.895 **(57.946)**
$\bar{y}_{COMP}$	$4.830*10^{-2}$ **(136.315)**	2.001 **(100.750)**	107.204 **(99.940)**	$4.831*10^{-2}$ **(136.315)**	2.001 **(100.750)**	107.204 **(136.315)**
α_o	**1.014**	**1.014**	**1.008**	**1.014**	**1.014**	**1.008**
$\bar{y}_e$	$4.830*10^{-2}$ **(136.315)**	2.00 **(100.750)**	106.995 **(100.136)**	$4.830*10^{-2}$ **(136.315)**	2.001 **(100.750)**	106.993 **(100.139)**
λ	**0.999**	**0.999**	**0.997**	**1.000**	**0.999**	**0.997**

3. Interpretation of Results from the Tables

(i) Since the ξ-MSE of all the estimators is function of sample values only and does not involve any of the finite population parameters which is contrary to the design-based approach and configuration of samples changes over non-response rates, the results can be assumed to be more realistic as these do not involve values of the unknown population parameters. This point makes model-based approach more practical.

(ii) For a fixed model, ξ-MSE of estimators has increasing trend with variation of the constant g where $0 \leq g \leq 2$. The trend is similar in both the models considered.

(iii) Amongst all the strategies, $[\xi, \bar{y}_{RAT}]$ seems to be the most inferior irrespective of values of g whether it is model I or model II.

(iv) Both the Strategies $[\xi, \bar{y}_{COMP}]$and$[\xi, \bar{y}_e]$under respective optimality conditions are almost equally efficient than other strategies. This result is uniformly true under both the models with a particular choice of g.

(v) The minimum ξ-MSE of the estimator $\bar{y}_e$is independent of the values of the parametersα and a. It is affected only with the value of the parameter λ.

(vi) As expected, the percent relative efficiency of the strategies as compared with the strategy$[\xi, \bar{y}_M]$ has a decreasing trend with increasing non-response rate.

(vii) The smallest ξ-MSE (or maximum efficiency) of all the strategies, except $[\xi, \bar{y}_{RAT}]$ is obtained for the models $\xi[0,1: 1]$ and $\xi[1,1: 1]$which is indicative

that most of the strategies might be most effective when $v(x_k) = x_k^0 = 1$, than when g = 1 or 2.

Recapitulating what have been discussed above, it can, therefore, be concluded that the proposed imputation strategy $[\xi, \bar{y}_e]$is uniformly better than other existing imputation strategies when compared under super population model approach. Further, it is as good as the compromised method of imputation under optimality conditions.

4. Robustness criteria of strategies

4.1 Misspecification of Models

Now it is well established that in making inferences to a finite population, whose form is motivated by a super population model, the model selected comes at the first place and hence, the results might be either equivalent or superior to the probability sampling approach if, in fact the model describes accurately the population being sampled. If the model is not fully realistic, the model-dependent approach may result in misleading inferences. Thus, the correct specification of the model is an important step in such an approach.

It is evident that PRM is comprised of two components namely, (i) polynomial regression$E_\xi(Y_k)$ and (ii) variance function$\sigma^2 v(x_k)$. It is, therefore, obvious that while choosing a particular PRM, there might be two types of misspecifications in regression models, namely,

(i) misspecification in selecting an appropriate variance function$v(x_k)$, and
(ii) misspecification in choosing the appropriate polynomial regression$h(x_k) = \sum_{j=0}^{J} \delta_j \beta_j x_k^j$.

It is, therefore, desirable to investigate the effect of both of these types of misspecification in the model on the efficiency of any estimator.

4.2 Robust Estimators

From the expressions of ξ-MSE of the estimators, it is clear that it is affected by the deviation of polynomial regression $h(x_k)$ and the function$v(x_k)$, while the ξ-bias is affected only by the function $h(x_k)$ and is totally independent of the function$v(x_k)$. So, it is desirable to observe the effect of misspecification of both the functions on the variance of ξ-MSE.

Royall and Herson (1973a,1973b), therefore, considered an estimator **"robust"**, if there is nominal change in the amount of ξ-MSE due to the deviation (misspecification) of the model, or, in other words, if the optimality of the estimator vitiates slightly under the misspecification of the model.

4.3 Examining Robustness of the Suggested Strategies

As we have obtained the ξ-MSE of a number of imputation strategies under model I and II with$g = 0,1\ and\ 2$, it can be stated that for fixed g-value, the models change

only in respect of polynomial function$h(x_k)$, while for a given model; change in the value of g produces a change in the function$v(x_k)$ only. So a comparison of strategies under these changes of the two functions would reveal the facts about the robustness property of the considered Strategies.

For the purpose, we have considered the absolute differences (A.D.) of ξ-MSE of two strategies, defined as

$$A.D.[T] = \left|M_\xi\{T\}_I - M_\xi\{T\}_{II}\right| \tag{38}$$

where $M_\xi\{T\}_I$ and $M_\xi\{T\}_{II}$ stand for the ξ-MSE of the estimator T under Models I and II respectively.

Table.4 shows the A.D. values of the strategies under different non-response rates and g-values

Table 4. Absolute Difference of the Strategies

Strategy	**g**	**Non-response rates (%)**			
		10	**15**	**20**	**30**
$[\xi, \bar{y}_M]$	0	0.00	0.00	0.00	0.00
	1	0.00	0.00	0.00	0.00
	2	0.00	0.00	0.00	0.00
$[\xi, \bar{y}_{RAT}]$	0	0.363	0.467	0.533	0.745
	1	0.363	0.467	0.533	0.745
	2	0.364	0.467	0.533	0.745
$[\xi, \bar{y}_{COMP}]$	0	0.00	0.00	0.0001	0.00
	1	0.00	0.00	0.00	0.00
	2	0.018	0.014	0.011	0.00
$[\xi, \bar{y}_e]$	0	0.00	0.00	0.00	0.00
	1	0.00	0.00	0.00	0.00
	2	0.017	0.001	0.001	0.002

Remark4.: The figures in the Table 4 reveal the fact that except the strategy $[\xi, \bar{y}_{RAT}]$ all other strategies are fairly robust against the misspecification of the model under all the non-response rates, even if the rate is tripled. Thus, it concludes that the suggested family of estimators$\bar{y}_e$ may be considered to be robust under the misspecification of the model.

Royall and Herson (1973a, 1973b) pointed out that by selecting a **"balanced sample"**, the problem of misspecification could be resolved. The condition for a

balanced sample $s$is$\bar{x}_s^{(j)} = \bar{X}^{(j)} for j = 1,2, \dots, J$. If a sample is exactly balance or approximately balance, then the estimator would be robust. In this case, we observe that $\bar{X} = 41.4556$, whereas$\bar{x}_s = 39.6$, thus,$\bar{x}_s \approx \bar{X}$, that is our sample is approximately balanced. Perhaps this might be reason that the strategies considered are almost robust.

5. Conclusion

The paper discussed the behavior of some of the existing imputation methods and one newly proposed imputation method, namely, exponential type imputation method for the estimation of population mean under model-based approach. Some of the specific cases of PRM are considered for comparing the performance of these imputation strategies and it is observed that the strategy$[\xi, \bar{y}_e]$ performs better than other strategies, irrespective of the rate of non-response in the population and in the sample. This implies that the suggested strategy may be looked upon as an advancement over other imputation strategies which already exist in the literature. Moreover, almost all the imputation methods were observed to be robust enough guarantying that these are least affected by the model misspecifications.

References

Asghar, A., Sanaullah, A. and Hanif, M. (2014): Generalized exponential type estimator for population variance in survey sampling, RevistaColombiana de Estadistica, 37(1), 213-224.

Bahl, S. Tuteja, R. K. (1991): Ratio and product type-exponential estimator, Informationand Optimization Sciences, XII, I, 159-163.

Brewer, K. R. W. (1963a): A model of systematic sampling with unequal probabilities, Australian Journal of Statistics, 5, 5-13.

Brewer, K. R. W. (1963b): Ratio estimation and finite populations: Some results deducible from the assumption of an underlying stochastic process, Australian Journal of Statistics, 5, 93-105.

Cassel, C. M., Sarndal, C. E. and Wretman, J. H. (1976): Some results on generalized difference estimation and generalized regression estimation for finite populations, Biometrika, 63, 615-620.

Chambers, R.L. (1986): Outlier robust finite population estimation, Journal of the American Statistical Association, 81(396), 1063-1069.

Cochran, W. G. (1953): Sampling Techniques, John Wiley and Sons, Inc., New York, I Edition.

Deming, W. E. (1944): On errors in surveys, American Sociological Review, 9, 359-369.

Deming, W. E. (1950): Some Theory of Sampling, John Wiley and Sons, New York.

Dalenius, T. (1977a): Bibilography of Non-sampling errors in surveys, I (A-G), International Statistical Review. 3, 71-89.

Dalenius, T. (1977b): Bibilography of Non-sampling errors in surveys, II (H-Q), International Statistical Review. 45, 181-197.

Dalenius, T. (1977c): Bibilography of Non-sampling errors in surveys, III (R-Z), International Statistical Review. 45, 313-317.

Dillman, D., Eltinge, J., Groves, R. M. and Little, R. (2002): Survey non-response in design, data collection and analysis, In Survey Non-response, 3-26, New York: Wiley.

Isaki, C. T. and Fuller, W. A. (1982): Survey design under a regression superpopulation model, Journal of the American Statistical Association, 77, 89-96.

Kish, L. (1967): Survey Sampling. John Wiley and Sons, Inc., New York, II Edition.

Longford, N., T. (2005): Missing Data and Small Area Estimation: Modern Analytical Equipment for the Survey Statisticians, Springer Science, Inc., New York.

Little, R. J. A. and Rubin, D. B. (1987): Statistical Analysis with Missing Data. New York: John Wiley & Sons, Inc

Mahalanobis, P. C. (1940): A sample survey of the acreage under jute in Bengal, Sankhya, 4, 511-530.

Mahalanobis, P. C. (1944): On large-scale sample surveys, Philosophical Trasactionsof the Royal Society of London, 231 (B), 329-451.

Mahalanobis, P. C. (1946): Recent experiments in statistical sampling in the Indian Statistical Institute, Journal of the Royal Statistical Society, 109A, 325-378.

Moser, C. A. (1958): Survey Methods in Social Investigation, Heinemann, London.

Royall, R. M. (1970a): On finite population sampling theory under certain linear regression models, Biometrika, 57, 377-387.

Royall, R. M. (1970b): Finite population sampling: On labels in estimation, Annals of Mathematical Statistics, 41, 1774-1779.

Royall, R. M. (1971a): Linear regression models in finite population sampling theory, in Foundations of Statistical Inference, V. P. Godambe and D. A. Sprott (eds.), Toronto: Holt, Rinehart and Winston, 259-274.

Royall, R. M. (1971b): Discussion of paper by D. Basu, in Foundations of Statistical Inference, V. P. Godambeand D. A. Sprott (eds.), Toronto : Holt, Rinehart and Winston, 238-239.

Royall, R. M. (1971): Linear regression models in finite population sampling theory, in Foundations of Statistical Inference, V. P. Godambe and D. A. Sprott (eds.), Toronto: Holt, Rinehart and Winston, 259-274.

Royall, R. M. and Herson, J. (1973a): Robust estimation in finite populations I, Journal of the American Statistical Association, 68(344), 880-889.

Royall, R. M. and Herson, J. (1973b): Robust estimation in finite populations II: Stratification on a size variable, Journal of the American Statistical Association, 68(344), 890-893.

Rubin, D. B. (1976): Inference and missing data. Biometrika, 63, 581-592.

Singh,A.K.,Singh,Priyanka,andSingh.V.K.(2016):ModelBasedStudyofFamilies of Exponential Type Estimators in Presence of Non-Response, in Communications in Statistics – Theory andMethods, Vol -46,13,6478-6490.

Singh, S. and Horn, S. (2000).: Compromised imputation in survey sampling, Metrika, 51,267-276.

Smith, T. M. F. (1976): The foundation of Survey Sampling: A review, Journal of the Royal Statistical Society, 139 A, 183-204

Zarkovich, S. S. (1966): Quality of Statistical Data, Food and Agricultural Organization of the United Nations, Rome.