Applied Mathematical Sciences, Vol. 3, 2009, no. 35, 1727 - 1737

Optimal Method of Imputation in Survey Sampling

Sarjinder Singh

Department of Mathematics
University of Texas at Brownsville and Texas Southmost College
Brownsville, Texas 78520, USA
sarjinder@yahoo.com

Sylvia R. Valdes

Department of Mathematics and Statistics
University of Southern Maine
Portland, Maine, ME 04101-9300, USA
sylviavl@usm.maine.edu

Abstract

In this paper, we propose an optimal method of imputation which leads to an estimator of population mean with minimum mean squared error in survey sampling when the data values are missing completely at random (MCAR). The mean, ratio and regression methods of imputation are shown to be special cases and less efficient than the developed optimal method of imputation. An analytical comparison shows that the first order mean square error approximation for the optimal method of imputation is always smaller than the one for regression method of imputation. The percent relative efficiency is found to be inversely proportional to the response rate.

Keywords: Estimation of mean; Missing data; Imputation; Ratio and regression type estimators; Bias; Mean Squared Error

1 INTRODUCTION

Incomplete data or non-response in the form of missingness, censoring or grouping is a troubling issue for many data sets. Statisticians have recognized for some time that failure to account for the stochastic nature of incompleteness or non-response can spoil the nature of data. There are several factors that affect the non-response rate in any particular inquiry. Hansen and Hurwitz (1946) were the first to deal with the problem of incomplete samples in mail surveys. Mail surveys or telephone surveys are commonly used by bureaucratic or business organizations because of their low cost. Rubin (1976) defined two key concepts: Missing at random (MAR) and Observed at random (OAR).

Missing at random (MAR): The data are MAR if the probability of the observed missingness pattern given the observed and unobserved data does not depend on the values of the unobserved data. It will therefore include cases where the enumerator is not able to contact the respondents only by chance and had he been able to contact, the data would have been collected. For example, when the information is kept on punched cards, the nonresponse due to the accidental loss of one or more cards is of the first category. Although this illusion is rather outdated in the world of modern computing but still there is a chance that some data files may get damaged due to virus attacks. This type of nonresponse is called random nonresponse.

Observed at random (OAR): The data are OAR, if for every possible value of the missing data, the probability of the observed missingness pattern, given the observed and unobserved data, does not depend on the values of the observed data. The combination of MAR and OAR is called MCAR. In other words, the MCAR can be defined as $f\,(A|D) = f(A)$ for all D where D is the data matrix and A is the missing data indicator matrix ($a_{ij} = 1$ if d_{ij} is reported, $a_{ij} = 0$ otherwise). Heitjan and Basu (1996) have also considered the problem of distinguishing on between MAR and MCAR. Note that the concept of OAR is vestige of Rubin (1976). Now a days people jump right from MAR to MCAR, which is a logical step and quite easy to follow. Let $\overline{Y} = N^{-1}\sum\limits_{i=1}^{N} y_i$ be the mean of the finite population $\Omega = (1, 2, ...i, ...N)$. A simple random sample without replacement (SRSWOR), s , of size n is drawn from Ω to estimate $\overline{Y}$. Let r be the number of responding units out of sampled n units. Let the set of responding units be denoted by A and that of non-responding units be denoted by $\overline{A}$. For every unit $i \in A$, the value y_i is observed. However for the units $i \in \overline{A}$, the y_i values are missing and imputed values are derived. We assume that imputation is carried out with the aid of an auxiliary variable, x, such that x_i, the value of x for unit i, is known and positive for every $i \in s = A \cup \overline{A}$. In other words, the data $x_s = \{x_i : i \in s\}$ are known. Following the notations of Lee, Rancourt and Särndal (1994), in the case of single value imputation, if the ith unit requires imputation, the value $\widehat{b}x_i$ is imputed, where $\widehat{b} = \frac{\overline{y}_r}{\overline{x}_r}$. Data after imputation becomes:

$$y_{\bullet i} = \begin{cases} y_i & \text{if} \quad i \in A \\ \widehat{b}x_i & \text{if} \quad i \in \overline{A} \end{cases} \tag{1.1}$$

where A and $\overline{A}$ denote the responding and non-responding units in the sample. This method of imputation is called the ratio method of imputation. Under this method of imputation, the point estimator of population mean given by:

$$\overline{y}_s = \frac{1}{n}\sum_{i=1}^{n} y_{\bullet i} \tag{1.2}$$

becomes:

$$\overline{y}_{\text{rat}} = \overline{y}_r \left(\frac{\overline{x}_n}{\overline{x}_r}\right), \tag{1.3}$$

where $\overline{x}_n = n^{-1}\sum\limits_{i=1}^{n} x_i$, $\overline{x}_r = r^{-1}\sum\limits_{i=1}^{r} x_i$ and $\overline{y}_r = r^{-1}\sum\limits_{i=1}^{r} y_i$. Note that the suffix *rat* in (1.3) stands for ratio estimator and the suffix s in (1.2) stands for sample mean based on entire sample information s.

Under mean method of imputation, the data after imputation take the form:

$$y_{\bullet i} = \begin{cases} y_i & \text{if} \quad i \in A \\ \overline{y}_r & \text{if} \quad i \in \overline{A} \end{cases} \tag{1.4}$$

and the point estimator (1.2) becomes:

$$\overline{y}_m = \frac{1}{r}\sum_{i=1}^{r} y_i = \overline{y}_r \tag{1.5}$$

Under regression method of imputation, the data after imputation take the form:

$$y_{\bullet i} = \begin{cases} y_i & \text{if} \quad i \in A \\ \overline{y}_r + \widehat{\beta}\,(x_i - \overline{x}_r) & \text{if} \quad i \in \overline{A} \end{cases} \tag{1.6}$$

where $\widehat{\beta} = \frac{s_{xy}}{s_x^2}$, with $s_{xy} = (r-1)^{-1} \sum\limits_{i=1}^{r} (x_i - \overline{x}_r)(y_i - \overline{y}_r)\,,\ s_x^2 = (r-1)^{-1} \sum\limits_{i=1}^{r} (x_i - \overline{x}_r)^2\,,$ and the point estimator (1.2) becomes:

$$\overline{y}_{\text{reg}} = \overline{y}_r + \widehat{\beta}\,(\overline{x}_n - \overline{x}_r) \tag{1.7}$$

where the suffix *reg* stands for the regression estimator.

The next section has been devoted to define notation and expectations which are useful to find the conditional bias and variance of the estimators at (1.3), (1.5), (1.6) and the estimator resultant from the proposed methods of imputation.

2 BACKGROUND THEORY

Let us define:

$$\epsilon = \frac{\overline{y}_r}{\overline{Y}} - 1, \quad \delta = \frac{\overline{x}_r}{\overline{X}} - 1, \quad \text{and} \quad \eta = \frac{\overline{x}_n}{\overline{X}} - 1$$

Assuming MCAR and using the concept of two phase sampling by following Rao and Sitter (1995), for given r and n, we have:

$$E(\epsilon) = E(\delta) = E(\eta) = 0$$

and

$$\begin{aligned} E\left(\epsilon^2\right) &= \left(\frac{1}{r} - \frac{1}{N}\right) C_y^2, \quad E\left(\delta^2\right) = \left(\frac{1}{r} - \frac{1}{N}\right) C_x^2, \quad E(\epsilon\delta) = \left(\frac{1}{r} - \frac{1}{N}\right) \rho_{xy} C_y C_x \\ E\left(\eta^2\right) &= \left(\frac{1}{r} - \frac{1}{N}\right) C_x^2, \quad E(\delta\eta) = \left(\frac{1}{n} - \frac{1}{N}\right) C_x^2, \text{ and } E(\epsilon\eta) = \left(\frac{1}{n} - \frac{1}{N}\right) \rho_{xy} C_y C_x \end{aligned}$$

where

$$C_y^2 = \frac{S_y^2}{\overline{Y}^2}, \quad C_x^2 = \frac{S_x^2}{\overline{X}^2}, \quad \rho_{xy} = \frac{S_{xy}}{S_x S_y}, \ S_x^2, \ S_y^2 \ \text{ and } S_{xy} \text{ have their usual meanings.}$$

Thus we have the following theorems:

Theorem 2.1. The conditional bias of the estimator $\overline{y}_{\text{rat}}$ is given by:

$$B\left(\overline{y}_{\text{rat}}\right) = \left(\frac{1}{r} - \frac{1}{n}\right)\overline{Y}\left(C_x^2 - \rho_{xy}C_yC_x\right) \tag{2.1}$$

and is valid for the given values (or conditional values) of r and n.
Proof. The estimator $\overline{y}_{\text{rat}}$ at (1.3) in terms of ϵ , δ and η can be written as:

$$\overline{y}_{\text{rat}} = \overline{Y}\ \left[1 + \epsilon + \eta - \delta + \delta^2 + \epsilon\eta - \epsilon\delta - \delta\eta + O\left(\epsilon^2\right)\right] \tag{2.2}$$

Taking expected value on both sides of (2.2) and its deviation from actual mean, we get (2.1).
Theorem 2.2. The mean squared error of the estimator, $\overline{y}_{\text{rat}}$, is:

$$\text{MSE}\left(\overline{y}_{\text{rat}}\right) = \left(\frac{1}{n} - \frac{1}{N}\right)S_y^2 + \left(\frac{1}{r} - \frac{1}{n}\right)\left[S_y^2 + R^2S_x^2 - 2RS_{xy}\right] \tag{2.3}$$

where $R = \frac{\overline{Y}}{\overline{X}}$.

Proof. To the first order of approximation,we have

$$\begin{aligned}\text{MSE}\left(\overline{y}_{\text{rat}}\right) &= E\left[\overline{y}_{\text{rat}} - \overline{Y}\right]^2\\ &\approx E\left[\epsilon + \eta - \delta\right]^2\\ &= E\left[\epsilon^2 + \eta^2 + \delta^2 + 2\epsilon\eta - 2\varepsilon\delta - 2\delta\eta\right]\end{aligned}$$

On putting the expected values, we get (2.3).
The variance of the estimator (1.5) obtained by the mean method of imputation is given by:

$$V\left(\overline{y}_m\right) = \left(\frac{1}{r} - \frac{1}{N}\right)S_y^2 \tag{2.4}$$

On comparing (2.3) with (2.4), one can easily see that the ratio method of imputation is better than mean method of imputation if:

$$R < 2\frac{S_{xy}}{S_x^2} = 2\beta \tag{2.5}$$

where $\beta = \frac{S_{xy}}{S_x^2}$.

The condition (2.5) holds in most practical situations and the ratio method of imputation remains better than the mean method of imputation.

The mean square error (MSE) of the estimator (1.7) obtained by the regression method of imputation is given by:

$$\text{MSE}\left(\overline{y}_{\text{reg}}\right) = \left(\frac{1}{n} - \frac{1}{N}\right) S_y^2 + \left(\frac{1}{r} - \frac{1}{n}\right) S_y^2 \left(1 - \rho_{xy}^2\right) \tag{2.6}$$

On comparing (2.6) with (2.3) and (2.4), one can see that the regression method of imputation remains always better than mean imputation as well as ratio imputation.

In the next section, we are suggesting an optimal method of imputation. The estimator obtained from the proposed method of imputation has shown to remain better than the estimator obtained from the mean, ratio and regression methods of imputation.

3 OPTIMAL METHOD OF IMPUTATION

In the proposed method of imputation, the data after imputation take the form:

$$y_{\bullet i} = \begin{cases} y_i & \text{if} \quad i \in A \\ \Phi + \Psi x_i & \text{if} \quad i \in \overline{A} \end{cases} \tag{3.1}$$

where Φ and Ψ are suitably chosen constants, such that either the mean square error of the resultant estimator is minimum, or are pre-decided values.

Special Cases:

(i) If $\Phi = \overline{y}_r$, $\Psi = 0$, then (3.1) reduces to the mean method of imputation.

(ii) If $\Phi = 0$, $\Psi = \widehat{b}$, then (3.1) reduces to the ratio method of imputation.

(iii) If $\Phi = \overline{y}_r - \widehat{\beta}\, \overline{x}_r$, $\Psi = \widehat{\beta}$, then (3.1) reduces to the regression method of imputation.

(iv) If $\Phi = 0$, $\Psi = \overline{y}_r \frac{\left\{n\left(\frac{\overline{x}_n}{\overline{x}_r}\right)^{\alpha} - r\right\}}{(n\overline{x}_n - r\overline{x}_r)}$, where α is a suitably chosen constant, then (3.1) reduces to the power transformation method of imputation suggested by Singh and Deo (2003).

Note that Singh and Deo (2003) have shown that the power transformation imputation method remains as good as the regression method of imputation, and methods suggested by Singh and Horn (2000) and Singh, Horn and Tracy (2001).

Theorem 3.1. The point estimator (1.2) of the population mean $\overline{Y}$ under the proposed method of imputation becomes:

$$\overline{y}_p = p\overline{y}_r + (1-p)\,\Phi + \Psi\,(\overline{x}_n - p\overline{x}_r) \tag{3.2}$$

where $p = \frac{r}{n}$ denotes the response rate.
Proof. We have

$$\overline{y}_p = \frac{1}{n}\sum_{i=1}^{n} y_{\bullet i} = \frac{1}{n}\left[\sum_{i\in A} y_{\bullet i} + \sum_{i\in \overline{A}} y_{\bullet i}\right] \tag{3.3}$$

and using (3.1), we get (3.2).
Now we have the following theorems:
Theorem 3.2. The conditional bias of the proposed estimator $\overline{y}_p$ at (3.2) is:

$$B\left(\overline{y}_p\right) = (1-p)\left(\Phi + \Psi\ \overline{X} - \overline{Y}\right) \tag{3.4}$$

Proof. Assuming $|\delta| < 1$ and $|\eta| < 1$, and neglecting higher order terms, the estimator $\overline{y}_p$ in terms of ϵ, δ and η can be written as:

$$\overline{y}_p = p\overline{Y}\,(1+\epsilon) + (1-p)\,\Phi + \Psi\overline{X}\,\{(1-p) + (\eta - p\delta)\} \tag{3.5}$$

Taking expected value on both sides of (3.5) and its deviation from actual mean, we get (3.4). Hence the theorem.
The optimal method of imputation will be unbiased if

$$\Phi + \Psi\ \overline{X} - \overline{Y} = 0$$

that is:

(i) if $\Phi = 0$, and $\Psi = \frac{\overline{Y}}{\overline{X}}$.
(ii) if $\Phi = \left(\overline{Y} - \Psi\overline{X}\right)$ for any choice of Ψ .
(iii) if $\Psi = \frac{\left(\overline{Y}-\Phi\right)}{\overline{X}}$ for any choice of Φ.

Theorem 3.3. The minimum mean squared error of the proposed estimator $\overline{y}_p$ is:

$$\text{Min.MSE}\left(\overline{y}_p\right) = p^2 S_y^2 \left[\left(\frac{1}{r} - \frac{1}{N}\right) - \Theta\rho_{xy}^2\right] \tag{3.6}$$

where

$$\Theta = \frac{\left\{\left(\frac{1}{n} - \frac{1}{N}\right) - p\left(\frac{1}{r} - \frac{1}{N}\right)\right\}^2}{\left\{\left(\frac{1}{n} - \frac{1}{N}\right) + p^2\left(\frac{1}{r} - \frac{1}{N}\right) - 2p\left(\frac{1}{n} - \frac{1}{N}\right)\right\}} \tag{3.7}$$

for the optimal values of Φ and Ψ given by:

$$\Phi = \overline{Y}\left(1 + p \bigtriangledown \rho_{xy}\frac{C_y}{C_x}\right) \tag{3.8}$$

and

$$\Psi = -\frac{\overline{Y}}{\overline{\overline{X}}}\ p\ \bigtriangledown\rho_{xy}\frac{C_y}{C_x} \tag{3.9}$$

where

$$\bigtriangledown = \frac{\Theta}{\left\{\left(\frac{1}{n} - \frac{1}{N}\right) - p\left(\frac{1}{r} - \frac{1}{N}\right)\right\}}$$

Proof. We have:

$$\begin{aligned} \text{MSE}\left(\overline{y}_p\right) &= E\left[\overline{y}_p - \overline{Y}\right]^2 \\ &= E\left[(p-1)\overline{Y} + p\overline{Y}\epsilon + (1-p)\Phi + \Psi\overline{X}\left\{(1-p) + (\eta - p\delta)\right\}\right]^2 \\ &= \zeta_1 + \zeta_2\Phi^2 + \zeta_3\Psi^2 - 2\zeta_4\Phi - 2\zeta_5\Psi + 2\zeta_6\Phi\Psi \end{aligned} \tag{3.10}$$

where

$$\begin{aligned}
\zeta_1 &= (1-p)^2\,\overline{Y}^2 + p^2\overline{Y}^2\left(\frac{1}{n}-\frac{1}{N}\right)C_y^2, \quad \zeta_2 = (1-p)^2\,, \\
\zeta_3 &= \overline{X}^2\left[(1-p)^2 + \left\{\left(\frac{1}{n}-\frac{1}{N}\right) + p^2\left(\frac{1}{r}-\frac{1}{N}\right) - 2p\left(\frac{1}{n}-\frac{1}{N}\right)\right\}C_x^2\right], \quad \zeta_4 = \overline{Y}\,(1-p)^2 \\
\zeta_5 &= \overline{X}\,\overline{Y}\left[(1-p)^2 - p\left\{\left(\frac{1}{n}-\frac{1}{N}\right) - p\left(\frac{1}{r}-\frac{1}{N}\right)\right\}\rho_{xy}C_yC_x\right], \text{ and } \zeta_6 = (1-p)^2\,\overline{X}
\end{aligned}$$

Now setting:

$$\frac{\partial \text{MSE}\left(\overline{y}_p\right)}{\partial\Phi} = 0 \quad \text{and} \quad \frac{\partial \text{MSE}\left(\overline{y}_p\right)}{\partial\Psi} = 0$$

we get

$$\Phi = \frac{\zeta_4\zeta_3 - \zeta_5\zeta_6}{\zeta_2\zeta_3 - \zeta_6^2} \quad \text{and} \quad \Psi = \frac{\zeta_2\zeta_5 - \zeta_4\zeta_6}{\zeta_2\zeta_3 - \zeta_6^2}$$

and substitution of these optimal values in (3.10) proves the theorem.

It can be shown using expression (3.7) that the proposed method of imputation remains always better than regression, ratio and mean methods of imputation.

Important Message for Imputator Statisticians: Note that the values of Φ and Ψ depend upon the value of population mean, $\overline{Y}$, and $\rho_{xy}\frac{C_y}{C_x}$, while this may not be a substantial limitation of the method suggested here. Note that in most of the cases the imputation is used to complete the missing data instead of estimating population mean $\overline{Y}$. Furthermore note that in many cases the population mean, $\overline{Y}$, or its good guess may be known. For example, consider the situation the total number of AIDS patients across the USA is known. Note that the number of AIDS patients depends upon number of HIV cases. If the number of HIV cases are known in a particular locality, then it can be used to impute the number of AIDS patients in that particular locality. Under such situation, the proposed method of imputation will be better than the existing methods.

4 RELATIVE EFFICIENCY

The percent relative efficiency of the proposed optimal method of imputation compare to the regression method of imputation is given by:

$$\text{RE} = \frac{\text{MSE}\left(\overline{y}_{\text{reg}}\right)}{\text{Min.MSE}\left(\overline{y}_{p}\right)} \times 100 = \frac{\left\{\left(\frac{1}{n} - \frac{1}{N}\right) + \left(\frac{1}{r} - \frac{1}{n}\right)\left(1 - \rho_{xy}^{2}\right)\right\} \times 100}{p^{2}\left\{\left(\frac{1}{r} - \frac{1}{N}\right) - \Theta\rho_{xy}^{2}\right\}} \tag{4.1}$$

which is an interesting expression. The percent relative efficiency expression given in (4.1) is a function of only four variables n, r, N and ρ_{xy}. Assuming that the population size $N = 100,000$ is large, and $n = 100$ moderate sample size, then the relative efficiency of the proposed estimator for different response rates between 40% to 90% and the values of the correlation coefficients between the range 0.1 to 0.9 is given in the following table.

Table 4.1. Relative efficiency of the proposed imputation method with respect to the regression method of imputation.

ρ_{xy}	$p = 0.40$	$p = 0.50$	$p = 0.60$	$p = 0.70$	$p = 0.80$	$p = 0.90$
0.1	612.2	398.0	276.7	203.5	155.9	123.3
0.3	519.2	382.0	267.8	198.6	153.4	122.3
0.5	531.2	350.0	250.0	188.8	148.4	120.4
0.7	441.2	302.0	223.3	174.1	140.9	117.4
0.9	321.1	237.9	187.7	154.5	130.9	113.4

5 DISCUSSION OF RESULTS

The percent relative efficiency of the proposed method of imputation ranges from 113.4% to 612.2% for response rate in the range 90% to 40% and correlation coefficient in the range 0.1 to 0.9 for moderate sample from a large population. We considered all the situations that can happen in real practice and hence our proposed method of imputation can be trusted.

Acknowledgements: The authors are thankful to the Editor for his comments on the original version of the manuscript.

References

[1] HANSEN, M.H., HURWITZ, W.N. (1946). The problem of non-response in sample surveys. *J. Amer. Statist. Assoc.*, 41, 517-529.

[2] HEITJAN, D.F., BASU, S. (1996) Distinguishing ‘Missing at Random’ and ‘Missing Completely at Random’. *Amer. Statist.* 50, 207-213.

[3] LEE, H. RANCOURT, E. SÄRNDAL, C.E. (1994). Experiments with variance estimation from survey data with imputed values. *J. Official Statist.*10 (3), 231-243.

[4] RAO, J.N.K., SITTER. R.R. (1995) Variance estimation under two phase sampling with application to imputation for missing data. *Biometrika*, 82, 453-460.

[5] RUBIN, D.B. (1976) Inference and missing data. *Biometrika* 63, 581-2.

[6] SINGH, S., DEO, B. (2003) Imputation by power transformation. *Statist. Papers*, 555-579.

[7] SINGH, S., HORN, S. (2000) Compromised imputation in survey sampling. *Metrika*, 51, 267-276.

[8] SINGH, S., HORN, S., TRACY, D.S. (2001) Hybrid of calibration and imputation in survey sampling. *Statistica*, LXI (1), 27-41

Received: November, 2008