Metrika (2000) 51: 267–276

Metrika

Compromised imputation in survey sampling

Sarjinder Singh[1], **Stephen Horn**[2]

[1] Department of Mathematics and Statistics, University of Southern Maine, Portland, Maine, ME 04101-9300, USA (E-mail: sarjinder@yahoo.com)
[2] Department of Family and Community Services, Box 7788, Canberra Mail Centre, ACT 2610, Australia (E-mail: stephen.horn@facs.gov.au)

Received: July 1998

Abstract. In this paper, a compromised imputation procedure has been suggested. The estimator of mean obtained from compromised imputation remains better than the estimators obtained from ratio method of imputation and mean method of imputation. An idea to form "Warm Deck Method" of imputation has also been suggested.

Key words: Estimation of mean, missing data, imputation, ratio estimator, bias, mean squared error, design based approach.

1. Introduction

Missing data is a common problem in sample surveys and imputation is frequently used to substitute values for missing data. Statisticians have recognised for some time that failure to account for the stochastic nature of incompleteness in the form of missingness of data can spoil inference. A natural question arises what one needs to assume to justify ignoring the incomplete mechanism. Rubin (1976) addressed three concepts: missing at random (MAR), observed at random (OAR) and parameter distribution (PD). Rubin defined "The data are MAR if the probability of the observed missingness pattern, given the observed and unobserved data, does not depend on the value of the unobserved data". Heitzan and Basu (1996) have distinguished the meaning of missing at random (MAR) and missing completely at random (MCAR) in a very nice way. Following them, we implicitly assume MCAR in the present investigation. Let $\bar{Y} = N^{-1} \sum_{i=1}^{N} y_i$ be the mean of the finite population $\Omega = \{1, 2, \ldots, i, \ldots, N\}$. A simple random sample without replacement (SRSWOR), s, of size n is drawn from Ω to estimate $\bar{Y}$. Let r be the number of responding units out of sampled n units. Let the set of responding units be denoted by R and that of non-responding units be denoted by R^c. For every unit $i \in R$, the value y_i is observed. However for the units $i \in R^c$, the y_i values are missing and imputed values are derived. We assume that imputation is carried out with the aid of a quantitative auxiliary variable, x, such that x_i, the value of x for unit i, is known and positive for every $i \in s$. In other words, the data $x_s = \{x_i : i \in s\}$ are known. Following the notations of Lee et al. (1994), in the case of single value imputation, if the i^{th} unit requires imputation, the value $\hat{b}x_i$ is imputed, where $\hat{b} = \sum_{i \in R} y_i / \sum_{i \in R} x_i$. The data after imputation becomes

$$y_{\bullet i} = \begin{cases} y_i & \text{if } i \in R \\ \hat{b}x_i & \text{if } i \in R^c \end{cases} \tag{1.1}$$

This method of imputation is called the ratio method of imputation. Under this method of imputation, the point estimator of population mean given by,

$$\bar{y}_s = \frac{1}{n} \sum_{i \in s} y_{\bullet i} \tag{1.2}$$

becomes,

$$\bar{y}_{RAT} = \bar{y}_r \frac{\bar{x}_n}{\bar{x}_r} \tag{1.3}$$

where $\bar{x}_n = n^{-1} \sum_{i \in s} x_i$, $\bar{x}_r = r^{-1} \sum_{i \in R} x_i$ and $\bar{y}_r = r^{-1} \sum_{i \in R} y_i$.

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

$$y_{\bullet i} = \begin{cases} y_i & \text{if } i \in R \\ \bar{y}_r & \text{if } i \in R^c \end{cases} \tag{1.4}$$

and the point estimator (1.2) becomes

$$\bar{y}_m = \frac{1}{r} \sum_{i \in R} y_i = \bar{y}_r \tag{1.5}$$

Here we have considered the design based approach to compare the proposed strategy with the existing strategies. The next section has been devoted to define few notations and expectations which are useful to find the conditional bias and variance of the estimators at (1.3), (1.5) and the estimator resultant from the proposed compromised imputation procedure in Section 3.

2. Theory

Let us define,

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

Using the concept of two phase sampling following Rao and Sitter (1995) and the mechanism of MCAR, for given r and n, we have

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

and

$$E(\varepsilon^2) = \left(\frac{1}{r} - \frac{1}{N}\right) C_y^2, \quad E(\delta^2) = \left(\frac{1}{r} - \frac{1}{N}\right) C_x^2, \quad E(\varepsilon\delta) = \left(\frac{1}{r} - \frac{1}{N}\right) \rho C_y C_x$$

$$E(\eta^2) = \left(\frac{1}{n} - \frac{1}{N}\right) C_x^2, \quad E(\delta\eta) = \left(\frac{1}{n} - \frac{1}{N}\right) C_x^2, \quad E(\varepsilon\eta) = \left(\frac{1}{n} - \frac{1}{N}\right) \rho C_y C_x$$

where $C_y^2 = S_y^2/\overline{Y}^2$, $C_x^2 = S_x^2/\overline{X}^2$, $\rho = S_{xy}/(S_x S_y)$ and S_y^2, S_x^2 and S_{xy} have their usual meanings. Thus we have the following theorems:

Theorem 2.1: *The conditional bias of the estimator* $\bar{y}_{RAT}$ *is given by*

$$B(\bar{y}_{RAT}) \approx \left(\frac{1}{r} - \frac{1}{n}\right) \overline{Y}(C_x^2 - \rho C_y C_x) \tag{2.1}$$

Proof: The estimator $\bar{y}_{RAT}$ at (1.3) in terms of ε, δ and η can be written as

$$\bar{y}_{RAT} \approx \overline{Y}[1 + \varepsilon + \eta - \delta + \delta^2 + \varepsilon\eta - \varepsilon\delta - \delta\eta + O(\varepsilon^2)] \tag{2.2}$$

Taking expected value on both sides of (2.2) and its deviation from actual mean, we get (2.1). Hence the theorem.

Theorem 2.2: *The mean squared error of the estimator,* $\bar{y}_{RAT}$, *is given by*

$$MSE(\bar{y}_{RAT}) \approx \left(\frac{1}{n} - \frac{1}{N}\right) S_y^2 + \left(\frac{1}{r} - \frac{1}{n}\right) [S_y^2 + R_1^2 S_x^2 - 2R_1 S_{xy}] \tag{2.3}$$

where $R_1 = \overline{Y}/\overline{X}$.

Proof: We have, to the first order of approximation,

$$MSE(\bar{y}_{RAT}) = E[\bar{y}_{RAT} - \overline{Y}]^2 \approx \overline{Y}^2 E[\varepsilon + \eta - \delta]^2$$
$$= \overline{Y}^2 E[\varepsilon^2 + \eta^2 + \delta^2 + 2\varepsilon\eta - 2\varepsilon\delta - 2\delta\eta]$$

On putting the expected values, we get (2.3). Hence the theorem.

The variance of the estimator (1.5) obtained by the mean method of imputation is given by

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

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

$$R_1 < \frac{2S_{xy}}{S_x^2} = 2\beta \quad \text{(say)} \tag{2.5}$$

where $\beta = S_{xy}/S_x^2$. The condition (2.5) holds in most of the practical situations and the ratio method of imputation remains better than the mean method of imputation.

In the next section, we are suggesting a compromised imputation procedure. The estimator obtained from the proposed compromised imputation method has shown to remain better than the estimator obtained from the ratio method of imputation and hence the mean method of imputation.

3. Compromised imputation

In the case of compromised imputation procedure, the data take the form,

$$y_{\bullet i} = \begin{cases} \alpha n y_i/r + (1-\alpha)\hat{b}x_i & \text{if } i \in R \\ (1-\alpha)\hat{b}x_i & \text{if } i \in R^c \end{cases} \tag{3.1}$$

where α is a suitably chosen constant, such that the variance of the resultant estimator is minimum. Here, we are also using information from imputed values for the responding units in addition to non-responding units.

Thus we have the following theorem:

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

$$\bar{y}_{COMP} = \alpha\bar{y}_r + (1-\alpha)\bar{y}_r\frac{\bar{x}_n}{\bar{x}_r} \tag{3.2}$$

Proof: We have

$$\bar{y}_{COMP} = \frac{1}{n}\sum_{i\in s} y_{\bullet i} = \frac{1}{n}\left[\sum_{i\in R} y_{\bullet i} + \sum_{i\in R^c} y_{\bullet i}\right] \tag{3.3}$$

and using (3.1), we get (3.2). Hence the theorem.

The estimator at (3.2) is an analogue of the well known estimator of population mean proposed by Chakrabarty (1968), Vos (1980) and Adhavryu and Gupta (1983) as

$$\bar{y}_{CVAG} = \alpha\bar{y}_n + (1-\alpha)\bar{y}_n\frac{\overline{X}}{\bar{x}_n} \tag{3.4}$$

Now we have the following theorems:

Theorem 3.2: *The conditional bias of the proposed estimator $\bar{y}_{COMP}$ at (3.2) is given by*

$$B(\bar{y}_{COMP}) \approx (1-\alpha)\left(\frac{1}{r}-\frac{1}{n}\right)\bar{Y}(C_x^2-\rho C_y C_x) \tag{3.5}$$

Proof: The estimator $\bar{y}_{COMP}$ in terms of ε, δ and η can be written as

$$\bar{y}_{COMP} \approx \alpha\bar{Y}(1+\varepsilon)+(1-\alpha)\bar{Y}(1+\varepsilon+\eta-\delta+\varepsilon\eta-\varepsilon\delta-\delta\eta+\delta^2+O(\varepsilon^2)) \tag{3.6}$$

Taking expected value on both sides of (3.6) and its deviation from actual mean, we get (3.5). Hence the theorem.

Theorem 3.3: *The minimum mean squared error of the proposed estimator $\bar{y}_{COMP}$ is given by*

$$Min.MSE(\bar{y}_{COMP}) \approx MSE(\bar{y}_{RAT}) - \left(\frac{1}{r}-\frac{1}{n}\right)\left(1-\rho\frac{C_y}{C_x}\right)^2\bar{Y}^2C_x^2 \tag{3.7}$$

for the optimum value of α given by

$$\alpha = 1-\rho\frac{C_y}{C_x} \tag{3.8}$$

Proof: See Appendix 1.

4. Practicability

The main difficulty in using the proposed compromised imputation procedure is the choice of α. It is important to note that the optimum value of α depends only upon the well known parameter $K=\rho\frac{C_y}{C_x}$. The value of K is quite stable in the repeated surveys as shown by Reddy (1978). Thus if the value of K is known then the proposed method can be easily implemented in actual surveys. Some time the value of K is not known. In these situations, we are suggesting two estimators of α given by

$$\hat{\alpha}_1 = 1-\frac{\bar{x}_r s_{xy}^*}{\bar{y}_r s_x^{*2}} \tag{4.1}$$

where $s_{xy}^* = (r-1)^{-1}\sum_{i=1}^{r}(y_i-\bar{y}_r)(x_i-\bar{x}_r)$ and $s_x^{*2} = (r-1)^{-1}\sum_{i=1}^{r}(x_i-\bar{x}_r)^2$, and

$$\hat{\alpha}_2 = 1-\frac{\bar{x}_n s_{xy}^*}{\bar{y}_r s_x^2} \tag{4.2}$$

where $s_x^2 = (n-1)^{-1} \sum_{i=1}^{n} (x_i - \bar{x}_n)^2$. Although the choice between $\hat{\alpha}_1$ and $\hat{\alpha}_2$ is not very important for the infinite populations, because the asymptotic mean squared error of the resultant estimators of mean will remain same (Sampath, 1989). Also refer to Appendix-2. It is interesting to observe through simulation that the estimator $\hat{\alpha}_2$ remains slightly better than $\hat{\alpha}_1$ for the case of finite populations. Its reason may be that $\hat{\alpha}_2$ makes use of full auxiliary information $\bar{x}_n$ and s_x^2, whereas $\hat{\alpha}_1$ makes use of partial auxiliary information $\bar{x}_r$ and s_x^{*2} statistics. In general, the proposed compromised technique remains better that ratio or mean methods of imputation.

5. Recommendations

It is interesting to note that if a strong imputation variable, x_i, is available, so that $y_i = bx_i$ very nearly holds for all i. Then to very close approximation, $\hat{b} = b$ and $\hat{b}x_i = y_i$, that is, the imputed value is near perfect. Then the imputation rule (3.1) reduces to

$$y_{\bullet i} = \begin{cases} y_i[1 + \alpha(n/r - 1)] & \text{if } i \in R \\ (1-\alpha) y_i & \text{if } i \in R^c \end{cases} \tag{5.1}$$

Under such situations, the values of $\rho_{xy} -> 1$ and $C_y \approx C_x$, then the optimum value of α and hence its estimators $\hat{\alpha}_i, i = 1, 2$ will tend to zero. In other words, then the imputed values, using compromised technique, remain close to the true values in R^c. Also the actual values y_i's does not have any impact of imputation in R. It is remarkable that a bad guess of α may lead to bad results in the compromised imputation. Since the compromised imputation provides better estimator of population mean, therefore, it is recommended to use in future.

6. Few suggestions

This type of compromisation can also be done between other type of imputation methods. For example a compromisation between Hot deck and Cold deck methods of imputation may lead to "Warm Deck" method of imputation, defined as

$$\bar{y}_{WD} = \alpha \bar{y}_{CD} + (1-\alpha)\bar{y}_{HD} \tag{6.1}$$

For details of the Hot deck and Cold deck methods of imputation, one can refer to Rubin (1978). The correlation between values obtained via cold deck and hot deck methods is expected to be high and hence the resultant estimator (6.1) named as "Warm Deck" method of imputation is expected to be efficient for optimum value of α, given by

$$\alpha_0 = \frac{V(\bar{y}_{HD}) - Cov(\bar{y}_{HD}, \bar{y}_{CD})}{V(\bar{y}_{HD}) + V(\bar{y}_{CD}) - 2\,Cov(\bar{y}_{HD}, \bar{y}_{CD})} \tag{6.2}$$

A consistent estimator of α_0 at (6.2) can easily be obtained by replacing variance – covariance terms by their sample analogues.

A compromised method of imputation obtained by pooling Mean method and Nearest Neighbourhood (NN) method of imputation, given by

$$\bar{y}_{MN} = \alpha\bar{y}_r + (1 - \alpha)\bar{y}_{NN} \tag{6.3}$$

can be named as "Mean-cum-NN" method of imputation. In the same fashion, a linear combination of any of two or more imputations procedures can be used to make a compromised imputation procedure. To study the properties of the resultant imputation procedures is interesting, but algebra is tedious though straightforward.

7. Empirical study

For the purpose of empirical study, we considered a finite population of $N = 20$ units given by Horvitz and Thompson (1952). First, we selected all possible samples of $n = 7$ units, which results in $M \equiv \binom{20}{7} = 77520$ samples. First we dropped two units randomly from each sample corresponding to the study variable y. Then the dropped units were imputed with four methods:

1. Mean method, $\bar{y}_0$ (say)
2. Ratio method, $\bar{y}_1$ (say)
3. Proposed method with $\alpha = \hat{\alpha}_1$, say $\bar{y}_2$
4. Proposed method with $\alpha = \hat{\alpha}_2$, say $\bar{y}_3$

The relative efficiency,

$$RE = \frac{\sum_{s=1}^{M}[(\bar{y}_0)_s - \bar{Y}]^2}{\sum_{s=1}^{M}[(\bar{y}_j)_s - \bar{Y}]^2} \times 100, \quad j = 1, 2, 3$$

of the ratio and proposed method with respect to mean method of imputation is shown in Table 1. Same process was repeated with other finite populations as shown in Table 1.

8. Remarks

Following remarks are the consequences of the comments given by one of the referees.

Remark 8.1: If the data satisfy only the MAR, but not the MCAR assumption, i.e. if missingness in y_i may depend on x_i, imputation based on ratio method is still valid. Especially, it is now a method to remove a potential bias of $\bar{y}_r$. Hence one may expect, that the difference between mean imputation and imputation based on ratios may become even larger.

Table 1. Relative efficiency of the ratio and proposed methods of imputation with respect to mean method of imputation

Source	Description of the population	Relative efficiency $\bar{y}_1$	$\bar{y}_2$	$\bar{y}_3$
Horvitz and Thompson (1952) $N = 20$	y: no. of hh's on i^{th} block x: eye estimate of no. of hh's on i^{th} block	103.25	107.82	108.14
Gunst and Mason (1980). p. 358 $N = 33$	y: Height of black female applicants with the police department (cm) x: Foot length (cm)	114.36	119.45	121.45
Wang and Chow (1994). p. 349 $N = 31$	y: Volume x: Height	109.45	114.75	116.45
-do-	y: Volume x: Diameter	110.76	114.98	115.98

Remark 8.2: Here basic idea is to consider a weighted average of the (unbiased) complete case estimate and a more sophisticated (but slightly biased) estimate, which infact leads to the following result. Let $\hat{\theta}_1$ be an unbiased estimate of θ with variance v_1. Let $\hat{\theta}_2$ be a biased estimate of θ with variance $v_2 < v_1$. Then there exists a $w \in [0, 1]$ such that

$$MSE[w\hat{\theta}_1 + (1 - w)\hat{\theta}_2] < v_1 = MSE(\hat{\theta}_1) \tag{8.1}$$

Many shrinkage methods follow this idea, which has been especially considered for prediction problems by Houwelingen and Cessie (1990) and Copas (1983), but also the Stein-estimator is close to this idea.

Remark 8.3: From (3.7), the difference between the MSE's of the ratio and proposed method of imputation is given by

$$D = MSE(\bar{y}_{RAT}) - Min.MSE(\bar{y}_{COMP})$$
$$= \left(\frac{1}{r} - \frac{1}{n}\right)\left(1 - \rho\frac{C_y}{C_x}\right)^2 \bar{Y}^2 C_x^2 \tag{8.2}$$

which is always a positive quantity.

In the model $m: y_i = \beta x_i + e_i$ such that $E_m(e_i|x_i) = 0$ and $E_m(e_i e_j \mid x_i, x_j) = \begin{cases} \alpha x_i^g, & i = j \\ 0, & i \neq j \end{cases}$ if $\beta = \bar{Y}/\bar{X}$ then the difference D reduces to zero because $\rho\dfrac{C_y}{C_x} = 1$ under such situations. The method of assuming $\beta = \bar{Y}/\bar{X}$ is not its best choice even for ratio method of imputation.

On the other hand, if $\beta = \sum_{i \in \Omega} X_i Y_i / \sum_{i \in \Omega} X_i^2$, which is more realistic measure of regression coefficient because it minimises $\sum_{i \in \Omega} e_i^2$ under the constraint that intercept is zero. Under such situations, the difference D is always positive. It is also to be noted that the proposed method of imputation attains the minimum variance of regression type estimator and hence valid under the assumptions of the general linear model defined as $m_l : y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ with intercept β_0.

Appendix-1: To prove Theorem 3.1, we have

$$\bar{y}_{COMP} \approx \bar{Y} + \bar{Y}\varepsilon + (1-\alpha)\bar{Y}(\eta - \delta) + O(\varepsilon^2)$$

where $O(\varepsilon^2)$ indicates terms of higher orders of $\varepsilon, \delta, \eta$ etc. Thus we have

$$\begin{aligned} MSE(\bar{y}_{COMP}) &= E[\bar{y}_{COMP} - \bar{Y}]^2 \\ &\approx E[\bar{Y}\varepsilon + (1-\alpha)\bar{Y}(\eta-\delta)]^2 \\ &= \left(\frac{1}{r} - \frac{1}{N}\right)\bar{Y}^2 C_y^2 \\ &\quad + \left(\frac{1}{r} - \frac{1}{n}\right)\bar{Y}^2[(1-\alpha)^2 C_x^2 - 2(1-\alpha)\rho C_y C_x] \end{aligned} \tag{A.1}$$

On differentiating (A.1) with respect to $(1-\alpha)$ and equating to zero, we get

$$\alpha = 1 - \rho\frac{C_y}{C_x} \tag{A.2}$$

On substituting (A.2) in (A.1), we get the proof of Theorem 3.3.

Appendix-2: Following Singh and Joarder (1998), define $\psi_0 = \frac{s_x^{*2}}{S_x^2} - 1$, $\psi_1 = \frac{s_x^2}{S_x^2} - 1$ and $\psi_2 = \frac{s_{xy}^*}{S_{xy}} - 1$ such that $E(\psi_i) = 0$, $i = 0, 1, 2$ and $E(\psi_i^2) = O(r^{-1})$. Now, using $\hat{\alpha}_1$ as an estimate of α, the estimator $\bar{y}_{COMP}$ can be expressed as

$$\bar{y}_{COMP} \approx \bar{Y} + \bar{Y}\varepsilon + \alpha\bar{Y}(\eta - \delta) + O(\varepsilon^2) \tag{A.3}$$

One can easily see that up to the first order of approximation, replacement of α by $\hat{\alpha}_2$ in the estimator $\bar{y}_{COMP}$ will also yield the same expression (A.3).

Acknowledgements. The authors are grateful to the two referees for their valuable comments which helped a lot in bringing the paper to its present form. They are also thankful to the Prof. Dr. Ursula Gather for her kind efforts to bring this paper in the final form.

References

Adhavryu D, Gupta PC (1983) On some alternative sampling strategies using auxiliary information. *Metrika* 30:217–226

Chakrabarty RP (1968) *Contribution to the theory of ratio-type estimators*. Ph. D. Thesis, Texas A and M University

Copas JB (1983) Regression, prediction and shrinkage. *J. Roy. Statist. Soc.,* B, 45:311–354
Gunst RF, Mason RL (1980) *Regression Analysis and Its Application. A data-oriented approach.* New York, Marcel Dekker, Inc.
Heitjan DF, Basu S (1996) Distinguishing 'Missing at Random' and 'Missing Completely at Random'. *The American Statistician* 50:207–213.
Horvitz DG, Thompson DJ (1952) A generalisation of sampling without replacement from a finite universe. *J. Amer. Statist. Assoc.* 47:663–685
Houwelingen, van and Cessie, le (1990) Predictive value of statistical models. *Statistics in Medicine* 9:1303–1325
Lee H, Rancourt E, S*ä*rndal CE (1994) Experiments with variance estimation from survey data with imputed values. *J. Official Statist.* 10(3):231–243
Rao JNK, Sitter RR (1995) Variance estimation under two phase sampling with application to imputation for missing data. *Biometrika* 82:453–460
Reddy VN (1978) A study on the use of prior knowledge on certain population parameters in estimation. *Sankhyā*, C, 40:29–37
Rubin RB (1976) Inference and missing data. *Biometrika* 63(3):581–592
Rubin RB (1978) *Multiple imputation for nonresponse in surveys.* John Wiley, New York
Sampath S (1989) On the optimal choice of unknowns in ratio type estimators. *J. Indian Soc. Agril. Statist* 41:166–172
Singh S, Joarder AH (1998) Estimation of finite population variance in the presence of random non-response. *Metrika* 47:241–249
Vos JWE (1980) Mixing of direct, ratio and product method estimators. *Statistica Nearlandica* 34:209–213
Wang SG, Chow SC (1994) *Advanced Linear Models: Theory and Applications*. New York, Marcel Dekker, Inc.