Metrika (1998) 47:241–249

Metrika

Estimation of Finite Population Variance Using Random Non-Response in Survey Sampling*

SARJINDER SINGH

The Australian Bureau of Statistics, P.O. Box 10, Belconnen, ACT 2616, Australia
e-mail: sarjinder.singh@abs.gov.au

ANWAR H. JOARDER

Department of Econometrics, Monash University, Clayton 3168, Australia
e-mail: anwarj@dpc.kfupm.edu.sa

Abstract: In this paper, an estimator of finite population variance proposed by Isaki (1983) is studied under the two different situations of random non-response suggested by Tracy and Osahan (1994). A distribution is proposed for the number of sampling units on which information could not be obtained due to random non-response. The estimators for the mean square errors of the proposed strategies are also suggested.

Key Words: Auxiliary information, random non-response, finite population variance

1 Introduction

The use of auxiliary information in survey sampling has its own eminent role. The ratio estimator, product estimator and regression estimator are well known examples. Such estimators take advantage of the correlation between the auxiliary variable X and the study variable Y. Isaki (1983) showed that under suitable conditions efficient estimators of finite population variance exist in the presence of auxiliary information. Assuming that a random sample of size n is selected and $(y_i, x_i), i = 1, 2, \ldots, n$ are observed. Isaki (1983) considered the problem of estimating the finite population variance $S_y^2 = (N-1)^{-1} \sum_{i=1}^{N} (Y_i - \bar{Y})^2$, where $\bar{Y} = N^{-1} \sum_{i=1}^{N} Y_i$. Assuming also the population

* This paper was written while both authors were members of the Dept. of Econometrics, Monash University, Clayton 3168, Australia. This paper was presented on SISC – 1996, Sydney, Australia. The opinions and results discussed in this paper are of authors and not necessarily of their institutes.

variance of the auxiliary character, that is, $S_x^2 = (N-1)^{-1}\sum_{i=1}^{N}(X_i - \bar{X})^2$ is known, he considered a ratio estimator of S_y^2 as

$$s_I^2 = s_y^2 \frac{S_x^2}{s_x^2} \tag{1.1}$$

where $s_y^2 = (n-1)^{-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$ and $s_x^2 = (n-1)^{-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ are the unbiased estimators of S_y^2 and S_x^2 respectively. The bias and mean square error of the estimator s_I^2 are given by

$$B(s_I^2) = \left(\frac{1}{n} - \frac{1}{N}\right) S_y^2[\lambda_{04} - \lambda_{22}] \tag{1.2}$$

and

$$MSE(s_I^2) = \left(\frac{1}{n} - \frac{1}{N}\right) S_y^4[\lambda_{40} + \lambda_{04} - 2\lambda_{22}] \tag{1.3}$$

where $\lambda_{ls} = \dfrac{\mu_{ls}}{\mu_{20}^{l/2}\,\mu_{02}^{s/2}}$ and $\mu_{ls} = (N-1)^{-1}\sum_{i=1}^{N}(Y_i - \bar{Y})^l(X_i - \bar{X})^s$ have their usual meanings. Tracy and Osahan (1994) studied the effect of random non-response: (*i*) On the study as well as the auxiliary variable (Situation 1), and (*ii*) On the study variable only (Situation 2) on the usual ratio estimator of the population mean.

In this paper, we study the effect of random non-response on the study and auxiliary variables of several estimators of variance. For each estimator we drive approximate biases, mean square errors and estimators of the mean square errors.

2 Distribution of Random Non-Response and Some Expected Values

Let $\omega : (v_1, v_2, \ldots, v_N)$ denote the population of N units from which a simple random sample of size n is drawn without replacement. If $r(r = 0, 1, 2 \ldots (n-2))$ denote the number of sampling units on which information could not be obtained due to random non-response, then the remaining $(n-r)$ units in the sample can be treated as *srswor* sample from ω. Since we are considering the problem of unbiased estimation of finite population variance, therefore we are

assuming that r should be less than $(n-1)$ i.e. $0 \leq r \leq n-2$. We assume that if p denotes the probability of non-response among the $(n-2)$ possible values of non-responses, then r has the following discrete distribution given by

$$P(r) = \frac{(n-r)}{nq+2p}\,{}^{n-2}C_r p^r q^{n-2-r}\,, \tag{2.1}$$

where $q = 1-p$, $r = 0, 1, 2, \ldots, (n-2)$ and ${}^{n-2}C_r$, denote the total number of ways of r non-responses out of total possible $(n-2)$ responses. It is interesting that under this distribution of random non-response the exact bias and mean square error expressions, up to first order of approximation, exists for the proposed strategies and hence comparisons with the Isaki (1983) estimator are valid and meaningful.

Let us define

$$\varepsilon = \frac{s_y^{*2}}{S_y^2} - 1\,, \quad \delta = \frac{s_x^{*2}}{S_x^2} - 1 \text{ and } \eta = \frac{s_x^2}{S_x^2} - 1$$

where $s_y^{*2} = (n-r-1)^{-1}\sum_{i=1}^{n-r}(y_i - \bar{y}^*)^2$ and $s_x^{*2} = (n-r-1)^{-1}\sum_{i=1}^{n-r}(x_i - \bar{x}^*)^2$ are conditionally unbiased estimators of S_y^2 and S_x^2, respectively, and where $\bar{y}^* = (n-r)^{-1}\sum_{i=1}^{n-r} y_i$ and $\bar{x}^* = (n-r)^{-1}\sum_{i=1}^{n-r} x_i$. Thus under the probability model given by (2.1), we have the following results:

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

$$E(\varepsilon^2) = \left[\frac{1}{(nq+2p)} - \frac{1}{N}\right](\lambda_{40} - 1)\,, \quad E(\delta^2) = \left[\frac{1}{(nq+2p)} - \frac{1}{N}\right](\lambda_{04} - 1)\,,$$

$$E(\eta^2) = \left[\frac{1}{n} - \frac{1}{N}\right](\lambda_{04} - 1)\,, \quad E(\varepsilon\delta) = \left[\frac{1}{(nq+2p)} - \frac{1}{N}\right](\lambda_{22} - 1)\,,$$

$$E(\varepsilon\eta) = \left[\frac{1}{n} - \frac{1}{N}\right](\lambda_{22} - 1) \quad \text{and} \quad E(\delta\eta) = \left[\frac{1}{n} - \frac{1}{N}\right](\lambda_{22} - 1)\,.$$

It may be noted that if $p = 0$ i.e. if there is no non-response, the above expected values coincide with the usual results.

3 Proposed Strategies

Strategy I: We are considering the situation when random non-response exists on both the study variable y and the auxiliary variable x and population variance S_x^2 of the auxiliary character is known. Thus we are proposing an estimator of finite population variance as

$$\hat{v} = s_y^{*2} \frac{S_x^2}{s_x^{*2}} \tag{3.1}$$

Thus we have the following theorems.

Theorem 3.1: The bias in the proposed estimator $\hat{v}$, up to terms of order $O(n^{-1})$, is given by

$$B(\hat{v}) = \left[\frac{1}{(nq+2p)} - \frac{1}{N}\right] S_y^2(\lambda_{04} - \lambda_{22}) \tag{3.2}$$

Proof: The estimator $\hat{v}$ in terms of ε and δ can be written as

$$\hat{v} = S_y^2(1 + \varepsilon - \delta + \delta^2 - \varepsilon\delta + \cdots) \tag{3.3}$$

Taking expected value on both sides of (3.3) and using the results on the expectations from Section 2, we get (3.2). Hence the theorem.

Theorem 3.2: The mean square error, up to terms of order $O(n^{-1})$, of the proposed estimator $\hat{v}$ is given by

$$MSE(\hat{v}) = \left[\frac{1}{(nq+2p)} - \frac{1}{N}\right] S_y^4(\lambda_{40} + \lambda_{04} - 2\lambda_{22}) \tag{3.4}$$

Proof: It is easy to check that

$$MSE(\hat{v}) = E(\hat{v} - S_y^2)^2 = S_y^4 E(\varepsilon - \delta + \delta^2 - \varepsilon\delta)^2$$

$$= S_y^4 E(\varepsilon^2 + \delta^2 - 2\varepsilon\delta) = \left[\frac{1}{(nq+2p)} - \frac{1}{N}\right] S_y^4(\lambda_{40} + \lambda_{04} - 2\lambda_{22}) .$$

Hence the theorem.

Before obtaining an estimator of $MSE(\hat{v})$, we need the following lemma:

Lemma 1: A maximum likelihood estimator of the probability of non-response, p, is given by

$$\hat{p} = \frac{(n-1+r) - \sqrt{(n-1+r)^2 - \dfrac{4rn(n-3)}{(n-2)}}}{2(n-3)} \tag{3.5}$$

Proof: The proof of the lemma is obvious by setting $\dfrac{\partial log(P(r))}{\partial p} = 0$, which is quadratic in p and provides only one admissible estimator in the range of 0 to 1 as given in (3.5).

Theorem 3.3: An estimator of the $MSE(\hat{v})$ is given by

$$\widehat{MSE}(\hat{v}) = \left[\frac{1}{(n\hat{q}+2\hat{p})} - \frac{1}{N}\right] s_y^{*4}(\hat{\lambda}_{40}^* + \hat{\lambda}_{04}^* - 2\hat{\lambda}_{22}^*) \tag{3.6}$$

where $\hat{\lambda}_{ls}^* = \dfrac{\hat{\mu}_{ls}^*}{(\hat{\mu}_{20}^*)^{l/2}(\hat{\mu}_{02}^*)^{s/2}}$ and $\hat{\mu}_{ls}^* = (n-r-1)^{-1}\sum_{i=1}^{n-r}(y_i - \bar{y}^*)^l(x_i - \bar{x}^*)^s$ have their usual meaning.

Strategy II: Here we are considering the situation when information on variable y could not be obtained for r units while information on variable x is available and population variance S_x^2 of the auxiliary variable is known. First we propose the following estimator:

$$\hat{v}_1 = s_y^{*2}\frac{S_x^2}{s_x^2} \tag{3.7}$$

Thus we have the following theorems and their proofs are obvious.

Theorem 3.4: The bias in the estimator $\hat{v}_1$, up to terms of order $O(n^{-1})$, is given by

$$B(\hat{v}_1) = \left(\frac{1}{n} - \frac{1}{N}\right) S_y^2(\lambda_{04} - \lambda_{22}) \tag{3.8}$$

Theorem 3.5: The mean square error, up to terms of order $O(n^{-1})$, is given by

$$MSE(\hat{v}_1) = MSE(s_I^2) + \left[\frac{1}{(nq+2p)} - \frac{1}{N}\right] S_y^4 \tag{3.9}$$

Theorem 3.6: An estimator of mean square error of $\hat{v}_1$ is given by

$$\widehat{MSE}(\hat{v}_1) = \left(\frac{1}{n} - \frac{1}{N}\right) s_y^{*4}(\hat{\lambda}_{40}^* + \hat{\lambda}_{04} - 2\hat{\lambda}_{22}^*) + \left[\frac{1}{(n\hat{q}+2\hat{p})} - \frac{1}{N}\right] s_y^{*4} \tag{3.10}$$

where $\hat{\lambda}_{04} = \dfrac{\hat{\mu}_{04}}{\hat{\mu}_{02}^2}$ and $\hat{\mu}_{0s} = (n-1)^{-1} \sum_{i=1}^{n} (x_i - \bar{x})^s$.

If information on x is available for all the n units, then we can obtain both s_x^2 and s_x^{*2}. Using this information, we construct another estimator as

$$\hat{v}_2 = s_y^{*2} \frac{S_x^2}{s_x^2} + \alpha\left(\frac{s_x^{*2}}{S_x^2} - 1\right) \tag{3.11}$$

where α is a suitably chosen constant such that the mean square error of $\hat{v}_2$ is minimum. Thus we have the following two theorems.

Theorem 3.7: The bias in the estimator $\hat{v}_2$ is same as in the estimator $\hat{v}_1$.

Theorem 3.8: The minimum mean square error of the proposed estimator $\hat{v}_2$ is given by

$$Min.MSE(\hat{v}_2) = MSE(\hat{v}_1) - \frac{\left\{\dfrac{1}{(nq+2p)} - \dfrac{1}{n}\right\}^2 S_y^4(\lambda_{22}-1)^2}{\left\{\dfrac{1}{(nq+2p)} - \dfrac{1}{N}\right\}(\lambda_{04}-1)} \tag{3.12}$$

Proof: We have

$$\begin{aligned} MSE(\hat{v}_2) &= E(\hat{v}_2 - S_y^2)^2 = E[S_y^2(1+\varepsilon-\eta+\eta^2-\varepsilon\eta) + \alpha\delta - S_y^2]^2 \\ &= MSE(\hat{v}_1) + \alpha^2\left\{\frac{1}{(nq+2p)} - \frac{1}{N}\right\}(\lambda_{04}-1) \\ &\quad + 2\alpha S_y^2\left\{\frac{1}{(nq+2p)} - \frac{1}{n}\right\}(\lambda_{22}-1) \end{aligned} \tag{3.13}$$

On differentiating (3.13) with respect to α, we get

$$\alpha = -\frac{\left\{\frac{1}{(nq+2p)} - \frac{1}{n}\right\} S_y^2(\lambda_{22}-1)}{\left\{\frac{1}{(nq+2p)} - \frac{1}{N}\right\}(\lambda_{04}-1)} \tag{3.14}$$

and then putting the optimum value of α in (3.13), we get (3.12). Hence the theorem.

Theorem 3.9: An estimator of the $Min.MSE(\hat{v}_2)$ is given by

$$\widehat{Min.MSE}(\hat{v}_2) = \widehat{MSE}(\hat{v}_1) - \frac{\left\{\frac{1}{(n\hat{q}+2\hat{p})} - \frac{1}{n}\right\}^2 (\hat{\lambda}_{22}^* - 1)^2 s_y^{*4}}{\left\{\frac{1}{(n\hat{q}+2\hat{p})} - \frac{1}{N}\right\}(\hat{\lambda}_{04}-1)} \tag{3.15}$$

If optimum value of α is not known, then it is advisable to replace it with its estimator $\hat{\alpha}$ and then we get the following estimator, given by

$$\hat{v}_3 = s_y^{*2}\frac{S_x^2}{s_x^2} + \hat{\alpha}\left(\frac{s_x^{*2}}{S_x^2} - 1\right) \tag{3.16}$$

where $\hat{\alpha} = -\dfrac{\left\{\frac{1}{(n\hat{q}+2\hat{p})} - \frac{1}{n}\right\} s_y^{*2}(\hat{\lambda}_{22}^* - 1)}{\left\{\frac{1}{(n\hat{q}+2\hat{p})} - \frac{1}{N}\right\}(\hat{\lambda}_{04}-1)}$ denote a consistent estimator of α.

To find the mean square error of the estimator $\hat{v}_3$, let us define $\kappa = \dfrac{\hat{\alpha}}{\alpha} - 1$, where $E(\kappa) = O(n^{-1})$, then, MSE of $\hat{v}_3$ is given by

$$\begin{aligned} MSE(\hat{v}_3) &= E(\hat{v}_3 - S_y^2)^2 = E[S_y^2(1+\varepsilon-\eta+\eta^2-\varepsilon\eta) + \alpha(1+\kappa)\delta - S_y^2]^2 \\ &\approx MSE(\hat{v}_1) + \alpha^2\left\{\frac{1}{(nq+2p)} - \frac{1}{N}\right\}(\lambda_{04}-1) \\ &\quad + 2\alpha S_y^2\left\{\frac{1}{(nq+2p)} - \frac{1}{n}\right\}(\lambda_{22}-1) \end{aligned} \tag{3.17}$$

i.e. which is approximately same as $MSE(\hat{v}_2)$. It is remarkable here that estimators $\hat{v}_2$ and $\hat{v}_3$ may take inadmissible value, i.e., negative value. Thus an equally efficient alternative estimator has been suggested and is given below:

$$\hat{v}_4 = s_y^{*2} \frac{S_x^2}{s_x^2} \left(\frac{s_x^{*2}}{S_x^2}\right)^{\alpha} \tag{3.18}$$

for $\alpha \neq 1$. If $\alpha = 1$ then it leads to the following strategy.

Strategy III: Here we again consider the situation when information on variable y could not be obtained for r units while information on the variable x is obtained for all the sample units. But the difference is that the population variance S_x^2 of the auxiliary variable is not known. In this case we suggest another ratio estimator as

$$\hat{v}_5 = s_y^{*2} \frac{s_x^2}{s_x^{*2}} \tag{3.19}$$

Thus we have the following theorem.

Theorem 3.10: The approximate bias in the proposed estimator $\hat{v}_5$, up to terms of order $O(n^{-1})$, is given by

$$B(\hat{v}_5) = \left[\frac{1}{(nq+2p)} - \frac{1}{N}\right] S_y^2(\lambda_{04} - \lambda_{22}) \tag{3.20}$$

Theorem 3.11: The asymptotic mean square error of the proposed estimator $\hat{v}_5$, up to terms of order $O(n^{-1})$, is given by

$$MSE(\hat{v}_5) = \left[\left\{\frac{1}{(nq+2p)} - \frac{1}{N}\right\}(\lambda_{40} - 2\lambda_{22} + 1) + \left\{\frac{1}{n} + \frac{1}{(nq+2p)} - \frac{2}{N}\right\}(\lambda_{04} - 1)\right] S_y^4 \tag{3.21}$$

Theorem 3.12: An estimator of mean square error of $\hat{v}_5$ is given by

$$\widehat{MSE}(\hat{v}_5) = \left[\left\{\frac{1}{(n\hat{q}+2\hat{p})} - \frac{1}{N}\right\}(\hat{\lambda}^*_{40} - 2\hat{\lambda}^*_{22} + 1) + \left\{\frac{1}{n} + \frac{1}{(n\hat{q}+2\hat{p})} - \frac{2}{N}\right\}(\hat{\lambda}_{04} - 1)\right] s_y^{*4} \tag{3.22}$$

We acknowledge that it is worthy to study the properties of an another estimator suggested by the referee as:

$$\hat{v}_{Ref} = s_y^{*2} + \hat{\beta}^{*2}(S_x^2 - s_x^{*2}) \tag{3.23}$$

where $\hat{\beta}^* = s^*_{xy}/s_x^{*2}$ and it also remain non-negative.

Acknowledgements: The authors are thankful to the learned referee for fruitful comments which helped a lot to bring the original manuscript in the present form. They are also thankful to their institutes for providing financial assistance to present this paper on SISC – 1996, Sydney, Australia.

References

Isaki CT (1983) Variance estimation using auxiliary information. J. Amer. Statist. Assoc. 7:117–123

Tracy DS, Osahan SS (1994) Random non-response on study variable versus on study as well as auxiliary variables. Statistica 54:163–168

Received 21.02.96