

REGRESSION TYPE ESTIMATORS FOR RANDOM NON-RESPONSE IN SURVEY SAMPLING

Sarjinder Singh, Anwar H. Joarder and D.S. Tracy

1. INTRODUCTION

Incompleteness- in the form of missingness- is a troubling feature of many datasets. Statisticians have recognised for some time that failure to account for the stochastic nature of incompleteness 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 are considering the effect of MAR on the well known ratio, product and regression estimators in survey sampling. Such estimators take the advantage of the correlation between the auxiliary variable x and study variable y . The ratio estimator is more efficient than the sample mean if x and y are highly positively correlated. The product estimator is useful if x and y are highly negatively correlated. But the regression estimator remains always more efficient than ratio, product and sample mean estimator. Tracy and Osahan (1994) studied the effect of random non-response on the usual ratio estimator of the population mean in two situations: (i) nonresponse in study as well as the auxiliary variable (Situation 1), and (ii) nonresponse in the study variable only (Situation 2).

In this paper, we suggest three regression type estimators in the presence of random non-response in different situations under an assumption that the number of sampling units on which information can not be obtained due to random non-response follow some distribution.

2. DISTRIBUTION OF RANDOM NON-RESPONSE AND SOME EXPECTED VALUES

Let $\Omega:(v_1, v_2, \dots, v_N)$ denote the population of N units from which a simple random sample of size n is drawn without replacement. If r ($r = 0, 1, \dots, (n-2)$) denotes 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 regression type estimators, therefore we are assuming that r should be less than $(n-1)$. We assume that if p denotes the probability of non-response among the $(n-2)$ possible values of non-response, then r has the following discrete distribution given by

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

where $q = 1 - p$ and $r = 0, 1, 2, \dots, (n-2)$. Let us define,

$$\varepsilon = \frac{\bar{y}_{n-r}}{\bar{Y}} - 1, \quad \delta = \frac{\bar{x}_{n-r}}{\bar{X}} - 1 \quad \text{and} \quad \eta = \frac{\bar{x}}{\bar{X}} - 1$$

where

$$\bar{y}_{n-r} = (n-r)^{-1} \sum_{i=1}^{n-r} y_i, \quad \bar{x}_{n-r} = (n-r)^{-1} \sum_{i=1}^{n-r} x_i \quad \text{and} \quad \bar{x} = n^{-1} \sum_{i=1}^n x_i \quad \text{have their usual meaning}$$

(Tracy and Osahan, 1994). The probability model defined at (2.1) is free from actual data values, hence can be considered as a model suitable for MAR situation. Then under the probability model given by (2.1), we have the following results:

$$E(\varepsilon) = E(\delta) = E(\eta) = 0, \quad E(\varepsilon^2) = \left[\frac{1}{(nq+2p)} - \frac{1}{N} \right] C_y^2, \quad E(\delta^2) = \left[\frac{1}{(nq+2p)} - \frac{1}{N} \right] C_x^2,$$

$$E(\eta^2) = \left(\frac{1}{n} - \frac{1}{N} \right) C_x^2, \quad E(\varepsilon\delta) = \left[\frac{1}{(nq+2p)} - \frac{1}{N} \right] \rho C_y C_x, \quad E(\varepsilon\eta) = \left(\frac{1}{n} - \frac{1}{N} \right) \rho C_y C_x \text{ and}$$

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

where $C_y = S_y / \bar{Y}$, $C_x = S_x / \bar{X}$ and $\rho = S_{xy} / (S_x S_y)$ have their usual meaning. It is interesting to note that under model (2.1), the above expected values are exact and hence makes valid comparison with the estimators in the absence of non-response. A logic and practical importance of the distribution defined at (2.1) has also been discussed in detail by Singh and Joarder (1998). In place of (2.1), we can also use another distribution (e.g. truncated binomial), but these distributions adds extra approximation in the expected values.

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 mean $\bar{X}$ of the auxiliary character is known. Thus we are considering a regression type estimator as

$$\bar{y}_1 = \bar{y}_{n-r} + \alpha_1 \left(\bar{X} - \bar{x}_{n-r} \right) \quad (3.1)$$

The minimum variance of the estimator at (3.1) is given by

$$\text{Min. } V(\bar{y}_1) = \left[\frac{1}{(nq+2p)} - \frac{1}{N} \right] S_y^2 (1 - \rho^2) \quad (3.2)$$

for the optimum value of α_1 given by

$$\alpha_1 = \frac{S_{xy}}{S_x^2} \quad (3.3)$$

If $p = 0$, then variance in (3.2) reduces to the variance of the usual linear regression estimator. In this situation, we can estimate α_1 by $\hat{\alpha}_1 = s_{xy}^* / s_x^{*2}$, where $s_{xy}^* = (n-r-1)^{-1} \sum_{i=1}^{n-r} (y_i - \bar{y}_{n-r}) (x_i - \bar{x}_{n-r})$ and $s_x^{*2} = (n-r-1) \sum_{i=1}^{n-r} (x_i - \bar{x}_{n-r})^2$ are the estimators of S_{xy} and S_x^2 respectively. Thus we have the following theorems:

Theorem 3.1: The regression estimator, if random non-response exists on both the study and auxiliary variables, is given by

$$\bar{y}_{llr} = \bar{y}_{n-r} + \frac{s_{xy}^*}{s_x^{*2}} (\bar{X} - \bar{x}_{n-r}) \quad (3.4)$$

Theorem 3.2: An estimator for estimating the variance of $\bar{y}_1$ is given by

$$\hat{V}(\bar{y}_1) = \left[\frac{1}{\left(\hat{nq} + 2\hat{p} \right)} - \frac{1}{N} \right] s_y^{*2} (1 - r^{*2}) \quad (3.5)$$

where $s_y^{*2} = (n-r-1)^{-1} \sum_{i=1}^{n-r} (y_i - \bar{y}_{n-r})^2$ and $\hat{p} = \frac{(n-1+r) - \sqrt{(n-1+r)^2 - 4nr(n-3)/(n-2)}}{2(n-3)}$ is a maximum likelihood estimator of p obtained from the distribution given by (2.1), $\hat{q} = 1 - \hat{p}$ and $r^* = s_{xy}^* / (s_x^* s_y^*)$. If $r = 0$ then $\hat{p} = 0$, and if $r = (n-2)$ then $\hat{p} = 1$ i.e. $\hat{p}$ is an admissible estimator of response probability p .

Strategy II: We consider the situation when information on variable y can not be obtained for r units while information on variable x is available and population mean $\bar{X}$ of the auxiliary variable is known. We propose the following estimator:

$$\bar{y}_2 = \bar{y}_{n-r} + \alpha_2 (\bar{X} - \bar{x}) \quad (3.6)$$

The minimum variance of the estimator $\bar{y}_2$ is given by

$$\text{Min. } V(\bar{y}_2) = \left[\frac{1}{(nq + 2p)} - \frac{1}{n} \right] S_y^2 + \left(\frac{1}{n} - \frac{1}{N} \right) S_y^2 (1 - \rho^2) \quad (3.7)$$

for the optimum value of α_2 given by

$$\alpha_2 = S_{xy} / S_x^2 \quad (3.8)$$

If $p = 0$ then the variance in (3.6) reduces to the variance of the usual linear regression estimator. We can estimate α_2 by $\hat{\alpha}_2 = s_{xy}^* / s_x^2$, where $s_x^2 = (n-1)^{-1} \sum_{i=1}^n (x_i - \bar{x})^2$. Thus we have the following theorems:

Theorem 3.3: The regression estimator, if non-response exists only on the auxiliary character, is given by

$$\bar{y}_{2lr} = \bar{y}_{n-r} + \frac{s_{xy}^*}{s_x^2} (\bar{X} - \bar{x}) \quad (3.9)$$

Theorem 3.4: An estimator for estimating the minimum variance of $\bar{y}_2$ is given by

$$\hat{V}(\bar{y}_2) = \left[\frac{1}{\left(\hat{nq} + 2\hat{p} \right)} - \frac{1}{n} \right] s_y^{*2} + \left(\frac{1}{n} - \frac{1}{N} \right) s_y^{*2} (1 - r^{*2}) \quad (3.10)$$

Strategy III: We are again considering the situation when information on variable y can 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 mean $\bar{X}$ of the auxiliary variable is not known. In this case we suggest another regression type estimator as

$$\bar{y}_3 = \bar{y}_{n-r} + \alpha_3 (\bar{x}_{n-r} - \bar{x}) \quad (3.11)$$

The minimum variance of $\bar{y}_3$ is given by

$$\text{Min. } V(\bar{y}_3) = \left[\frac{1}{(nq + 2p)} - \frac{1}{n} \right] S_y^2 (1 - \rho^2) + \left(\frac{1}{n} - \frac{1}{N} \right) S_y^2 \quad (3.12)$$

for the optimum value of $\alpha_3 = S_{xy} / S_x^2$. We can estimate α_3 by $\hat{\alpha}_2$. Thus we have the following theorems:

ORIGINAL DOCUMENT

Singh, S., Joarder, A.H. and Tracy, D.S (2000). Regression type estimators for random non-response in survey sampling. *Statistica*, LX, n.1, 39-44.\

Theorem 3.5: The regression estimator in the absence of population mean $\bar{X}$ but in the presence of non- response is given by

$$\bar{y}_{3lr} = \bar{y}_{n-r} + \frac{s_{xy}^*}{s_x^2} \left(\bar{x}_{n-r} - \bar{x} \right) \quad (3.13)$$

Theorem 3.6: An estimator of the minimum variance of the estimator $\bar{y}_3$ is given by

$$\hat{V}(\bar{y}_3) = \left[\frac{1}{\left(\hat{nq} + 2\hat{p} \right)} - \frac{1}{n} \right] s_y^{*2} (1 - r^{*2}) + \left(\frac{1}{n} - \frac{1}{N} \right) s_y^{*2} \quad (3.14)$$

Department of Mathematics and Statistics
University of Windsor, Windsor
Ontario, Canada, N9B 3P4

SARJINDER SINGH
& D.S. TRACY

Department of Mathematical Sciences
King Fahd University of Petroleum and Minerals
Dhahran 31261, Saudi Arabia

ANWAR H. JOARDER

ACKNOWLEDGEMENT

The authors are thankful to the referee for his valuable comments which helped a lot in improving the earlier draft of the manuscript.

REFERENCES

- D.B. RUBIN (1976). *Inference and missing data*. "Biometrika", 63(3), 581-592.
- D.S. TRACY, S.S. OSAHAN (1994). *Random non-response on study variable versus on study as well as auxiliary variables*. "Statistica", 54, 163-168.
- D.F. HEITJAN, S. BASU (1996). *Distinguishing 'Missing at Random' and 'Missing Completely at Random'*. "The American Statistician", 50, 207-213.
- SARJINDER SINGH, ANWAR H. JOARDER (1998). *Estimation of finite population variance using random non-response in survey sampling*. "Metrika", (In press).

ORIGINAL DOCUMENT

Singh, S., Joarder, A.H. and Tracy, D.S (2000). Regression type estimators for random non-response in survey sampling. *Statistica*, LX, n.1, 39-44.\

SUMMARY

In this paper, three regression type estimators are suggested for the estimation of finite population mean in two situations of random non-response considered by Tracy and Osahan (1994). The results obtained under the assumption that the number of sampling units on which information could not be obtained due to random non-response follows some distribution. Some suggestions are given for the estimation of the mean square error of the estimators proposed.