World Academy of Science, Engineering and Technology
International Journal of Mathematical and Computational Sciences
Vol:3, No:5, 2009

Adjusted Ratio and Regression Type Estimators for Estimation of Population Mean when some Observations are missing

Nuanpan Nangsue

***Abstract*—**Ratio and regression type estimators have been used by previous authors to estimate a population mean for the principal variable from samples in which both auxiliary x and principal y variable data are available. However, missing data are a common problem in statistical analyses with real data. Ratio and regression type estimators have also been used for imputing values of missing y data. In this paper, six new ratio and regression type estimators are proposed for imputing values for any missing y data and estimating a population mean for y from samples with missing x and/or y data. A simulation study has been conducted to compare the six ratio and regression type estimators with a previous estimator of Rueda. Two population sizes N = 1,000 and 5,000 have been considered with sample sizes of 10% and 30% and with correlation coefficients between population variables X and Y of 0.5 and 0.8. In the simulations, 10 and 40 percent of sample y values and 10 and 40 percent of sample x values were randomly designated as missing. The new ratio and regression type estimators give similar mean absolute percentage errors that are smaller than the Rueda estimator for all cases. The new estimators give a large reduction in errors for the case of 40% missing y values and sampling fraction of 30%.

***Keywords*—**Auxiliary variable, missing data, ratio and regression type estimators.

I. INTRODUCTION

MISSING data is a common problem that statisticians must treat in statistical analyses of real data. In survey research, data is often missing due to nonresponse. There are three types of nonresponse in survey research: noncoverage, unit nonresponse and item nonresponse [1]. Noncoverage can occur if an important subpopulation of the target population is not included in the sample design. Unit nonresponse occurs if it is not possible to obtain any of the required survey data from a selected unit. Item non-response occurs if it is only possible to obtain some of the required data from a selected unit. A variety of methods have been developed to attempt to compensate for missing survey data. Weighting adjustments are commonly used to compensate for noncoverage and unit nonresponse, while imputation methods that assign values for missing responses are used to compensate for item nonresponses [2]. Missing data means that estimates created from the reduced size of the data set are less efficient. Also the standard complete data methods cannot be immediately used to analyze the data. Further, possible biases can exist because respondents and non-respondents may differ in systematic ways. These biases are difficult to eliminate since the precise reasons for nonresponse are usually unknown [3].

In sample surverys, there are many estimation techniques that require advanced knowledge of known auxiliary x data to improve the efficiency of the estimator of a population mean($\overline{Y}$) [4]. For example, ratio and regression type estimators require knowledge of auxiliary information. Several authors have used ratio and regression type estimators to estimate population means (see, e.g., [5], [6], [7], [8], [9]. In this paper six new ratio and regression type estimators are proposed for estimating a population mean for a principal y variable when x and/or y data are missing.

Nuanpan Nangsue is with the Department of Applied Statistics, Faculty of Applied Science, King Mongkut's University of Technology North Bangkok, 1518 Pibulsongkram Road, Bangsue, Bangkok, Thailand 10800 (phone: +662 913 2500 ext. 4923; fax: +662 585 6105; e-mail:nmt@kmutnb.ac.th).

This research was supported by a research grant from the Faculty of Applied Science, King Mongkut's University of Technology North Bangkok.

II. METHODS

In this section the Rueda [4] estimator and some ratio and regression type estimators are described.

Rueda assumes that Ω is a population of N units from which a random sample of fixed size n is drawn. Rueda assumes that $t = (n\text{-}p\text{-}q)$ of the n sample observations are complete, but that for p of the sample observations x values are known but y values are missing and that for q of the observations y values are known but x values are missing. p and q are assumed to be integer numbers satisfying $0<p,$

$q < n/2.$

The units in the sample are separated into three sets.

$$s_1 = \{i \in s \ / \ x_i, y_i \text{ are available}\}$$
$$s_2 = \{i \in s \ / \ x_i \text{ are available, but } y_i \text{ is not}\}$$
$$s_3 = \{i \in s \ / \ y_i \text{ are available, but } x_i \text{ is not}\}$$

Finally, let s_4 be the members of the population which are not included in the sample.

A. Rueda Estimator

Rueda et al [4] proposed the following post survey predictor for the population mean $\overline{Y}$.

$$T^* = f_{s1}\bar{y}_{s1} + f_{s2}U_2^* \quad f_{s3}\bar{y}_{s3} \quad (1 \quad f_s)U_4^* \qquad (1)+$$

where

$$f_{s1} = \frac{n-p-q}{n}, \quad f_{s2} = \frac{p}{n}, \quad f_{s3} = \frac{q}{n}, \quad f_s = \frac{n}{N}$$

The sample means $\bar{y}_{s1}$ and $\bar{y}_{s3}$ in (1) are known and Rueda et al used the generalized least squares theory to estimate the value of U_2^* for the missing *y* values in *s2*. They estimated the value of U_2^* by using the minimum variance linear unbiased estimator of the sample regression coefficient ($\hat{\beta}$) to estimate the population regression coefficient β. Then the predictor $U_2^* = \hat{\beta}\bar{x}_{s2}$ is a linear and unbiased estimator for the missing *y* values $\bar{y}_{s2}$. The term U_4^* represents the mean of *y* for the unsampled set s_4. However, the *x* and *y* values are not available for this set. Rueda et al have suggested that the mean $\bar{y}_{s1\cup s3}$ of all available sample *y* values can be used as an estimator for U_4^* . The final estimator of Rueda et al was then

$$T^* = k_1\bar{y}_{s1} + k_2\bar{y}_{s3} \quad k_3U_2^* \qquad + \qquad (2)$$

where $k_1 = \dfrac{n-p-q(N \quad p)-}{N(n-p)}$, $k_2 = \dfrac{q(N-p)}{N(n-p)}$, $k_3 = \dfrac{p}{N}$

and $U_2^* = \hat{\beta}\bar{x}_{s2}$ with $\hat{\beta} = b_1 = \dfrac{rS_y}{S_x}$ (3)

Rueda et al have shown that the estimator T^* is an asymptotically normal unbiased estimator for the population mean of the principal variable.

B. Ratio and Regression Type Estimators

A ratio estimator was first proposed by Cochran [10] as a sample estimator of the population mean for *y* when complete *(x,y)* data was available for a sample and the population mean $\bar{X}$ was known. A ratio estimator was used by Tracy and Osahan [11] as an estimator of a population mean when some sample ***y*** data was missing and $\bar{X}$ was known. To estimate the population mean $\bar{Y}$ they only used data from the set *s1* in which both *x* and *y* values are known. In terms of the sets defined in Section 2, their ratio estimator is:

$$\bar{y} = \bar{y}_{s1}\left(\frac{\bar{X}}{\bar{x}_{s1}}\right) \qquad (4)$$

where $\bar{X}$ is the population mean of *X* and the subscript *s1* means data for the set *s1*.

Singh [12] proposed two regression type estimators that also used only the data from the set *s1*. The first estimator assumes that the population mean $\bar{X}$ is known and is defined by:

$$\bar{y} = \bar{y}_{s1} + b_1(\bar{X} - \bar{x}_{s1}) \qquad (5)$$

where b_1 is the sample regression coefficient for the set *s1* defined in (3).

The second estimator assumes that the population mean $\bar{X}$ is not known, but must be estimated from the sample set of known *x* values. This estimator is defined by:

$$\bar{y} = \bar{y}_{s1} + b_1(\bar{x}_{s1} - \bar{x}\ ) \qquad (6)$$

In this paper six modified ratio and regression type estimators are proposed in which all available data are used. These new estimators are designed to give improved estimates for the mean of the missing *y* values in the set *s2*, i.e., improved estimates for U_2^*.

Adjusted ratio. In this paper it is assumed that *y* data is available for sets *s1* and *s3* and *x* data for the sets *s1* and *s2* and therefore it is proposed to use $\bar{y}_{s1s3}$ for the sample mean of *y* and $\bar{x}_{s1s2}$ for the sample mean of *x*. The new estimator, which will be called *R1*, is then:

$$(R1) \qquad \bar{y}^* = \bar{y}_{s1s3}\left(\frac{\bar{X}}{\bar{x}_{s1s2}}\right) \qquad (7)$$

where, for example, the subscript *s1s3* means that data from sets *s1* and *s3* are used.

Adjusted regression type 1. For this estimator, which will be called *R2*, it is proposed to modify (5) by using a new regression type estimator that estimates $\bar{y}$ from the *s1* and *s3* data and $\bar{x}$ from the *s1* and *s2* data. The new regression type estimator is:

$$(R2) \qquad \bar{y}^* = \bar{y}_{s1s3} + b_1(\bar{X} - \bar{x}_{s1s2}) \qquad (8)$$

Adjusted regression type 2. Similar to regression type 1, it is proposed to replace (6) by a new regression type estimator, which will be called *R3*, that includes all available data for *x* and *y*. The new regression type estimator is then:

$$(R3) \qquad \bar{y} = \bar{y}_{s1s3} + b_1(\bar{x}_{s1s2} - \bar{x}_{s1}) \qquad (9)$$

Adjusted ratio and regression type 1. For this estimator, the adjusted ratio estimator *R1* is combined with a regression formula of the kind shown in (3). Further, the sample regression coefficient from (3) is included. The new ratio and regression type estimator, which will be called *R4*, is then:

$$(R4) \qquad \bar{y}^* = \bar{y}_{s1s3}\left(\frac{\bar{X}}{\bar{x}_{s1s2}}\right)^{b_1} \qquad (10)$$

Adjusted ratio and regression type 2. In this type the estimator of (8) for $\bar{y}_{s1s3}$ is replaced by the estimator in (10). The new estimator, which we call *R5*, is then:

$$(R5) \qquad \bar{y}^* = \bar{y}^*_{s1s3} + b_1\left(\bar{X} - \bar{x}_{s1s2}\right)$$

$$= (\ \bar{y}_{s1s3}\left(\frac{\bar{X}}{\bar{x}_{s1s2}}\right)^{b_1} + b_1(\bar{X} - \bar{x}_{s1s2}) \qquad (11)$$

Adjusted ratio and regression type 3. In this type the estimator of (10) for $\bar{y}_{s1s3}$ is replaced by the estimator in (8). The new estimator, which we call *R6*, is then:

$$(R6) \qquad \bar{y}^* = \bar{y}^*_{s1s3}\left(\frac{\bar{X}}{\bar{x}_{s1s2}}\right)^{b_1}$$

$$= \left(\bar{y}_{s1s3} + b_1\left(\bar{X} - \bar{x}_{s1s2}\right)\right) * \left(\frac{\bar{X}}{\bar{x}_{s1s2}}\right)^{b_1} \qquad (12)$$

This paper has adjusted the Rueda estimator from (2) by estimating U_2^* with (7) to (12).

III. Simulation Results

A simulation study with 10,000 repetitions has been conducted to compare the six new ratio and regression type estimators with the Rueda estimator. Two population sizes *N*=1,000 and 5,000 have been considered for different sample sizes and correlation coefficients between *X* and *Y*. In a sample, 10 and 40 percent of y values and 10 and 40 percent of *x* values were randomly designated as missing. For each sample, mean absolute percentage errors (MAPE) were calculated for the seven different estimators. The results are presented in Tables I and II.

TABLE I
Mean Absolute Percentage Error (MAPE) from Simulation Results for Population of Size N = 1,000, Sampling Fractions are 10% and 30%, *p* and *q* are 10% and 40% and Correlation Coefficients are 0.5 and 0.8.

ρ	Sampling fraction	p	q	MAPE						
				Rueda	R1	R2	R3	R4	R5	R6
0.5	10%	10%	10%	3.79626	**3.73583**	3.73589	3.73590	3.73589	3.73589	3.73589
			40%	4.03462	3.99027	3.98959	**3.98950**	**3.98950**	3.98963	3.98963
		40%	10%	5.88427	**4.93782**	4.93899	4.93846	4.93867	4.93915	4.93915
			40%	5.91413	**4.94384**	4.94499	4.94547	4.94478	4.94516	4.94516
	30%	10%	10%	3.30034	2.04610	2.04585	**2.04582**	**2.04582**	2.04586	2.04586
			40%	3.33466	2.04944	2.04904	**2.04899**	2.04905	2.04903	2.04903
		40%	10%	***11.9295***	2.57428	2.57213	**2.57167**	2.57182	2.57227	2.57227
			40%	***11.9608***	2.56689	2.56519	**2.56439**	2.56450	2.56559	2.56559
0.8	10%	10%	10%	2.46415	**2.35909**	2.35922	2.35921	2.35922	2.35922	2.35922
			40%	2.69016	2.59113	2.59074	2.59077	2.59075	**2.59071**	2.59072
		40%	10%	4.58606	3.10955	**3.10681**	3.10687	**3.10681**	3.10682	3.10683
			40%	4.60303	3.12873	**3.12661**	3.12776	3.12670	**3.12661**	3.12662
	30%	10%	10%	3.02458	1.28720	1.28686	**1.28683**	1.28685	1.28688	1.28688
			40%	3.00956	1.28989	1.28936	**1.28933**	1.28934	1.28940	1.28940
		40%	10%	***12.0140***	1.62831	**1.62353**	1.62361	1.62355	1.62355	1.62355
			40%	***11.9723***	1.63943	1.63065	1.63070	**1.63057**	1.63099	1.63099

TABLE II
Mean Absolute Percentage Error (MAPE) from Simulation Results for Population of Size N = 5,000, Sampling Fractions are 10% and 30% , *p* and *q* are 10% and 40% and Correlation Coefficients are 0.5 and 0.8.

ρ	Sampling fraction	p	q	MAPE						
				Rueda	R1	R2	R3	R4	R5	R6
0.5	10%	10%	10%	1.94112	**1.77427**	1.77435	1.77435	1.77435	1.77435	1.77435
			40%	1.93400	1.75865	1.75819	**1.75814**	1.75818	1.75820	1.75820
		40%	10%	4.11612	2.16936	2.16877	2.16882	**2.16876**	2.16878	2.16878
			40%	4.16336	2.15997	2.15928	**2.15905**	2.15920	2.15933	2.15933
	30%	10%	10%	2.99602	**0.90738**	0.90753	0.90753	0.90753	0.90753	0.90753
			40%	2.99040	0.89368	0.89342	0.89345	0.89344	**0.89341**	**0.89341**
		40%	10%	***12.0064***	1.14637	1.14465	1.14472	1.14469	**1.14464**	**1.14464**
			40%	***12.0155***	1.14082	1.13874	**1.13862**	1.13865	1.13879	1.13879
0.8	10%	10%	10%	1.37751	**1.09816**	1.09829	1.09830	1.09829	1.09829	1.09829
			40%	1.37265	1.10435	1.10429	**1.10428**	1.10429	1.10430	1.10430
		40%	10%	4.01832	1.35970	1.35895	1.35899	1.35895	**1.35893**	**1.35893**
			40%	4.01160	1.34777	1.34730	**1.34714**	1.34727	1.34739	1.34739
	30%	10%	10%	3.01337	0.57456	0.57444	0.57444	0.57444	**0.57443**	**0.57443**
			40%	2.99377	0.57179	0.57152	0.57153	0.57152	**0.57151**	**0.57151**
		40%	10%	***11.9934***	0.75776	0.75532	**0.75530**	0.75532	0.75533	0.75533
			40%	***11.9950***	0.75572	**0.75153**	0.75164	**0.75153**	0.75155	0.75155

IV. Conclusion

The simulation results in Tables I and II show that all of the new ratio and regression type estimators give similar mean absolute percentage errors which are always smaller than the errors for the Rueda estimator. For missing *x* values of 10% and 40% and missing *y* values of 10% the differences in the errors are small. However, for missing *y* values of 40% and sampling fraction of 30% the new estimators give much smaller errors than the Rueda estimator. It is therefore recommended that the new estimators should be used when an appreciable percentage of *y* values are missing.

Acknowledgment

The author would like to acknowledge the assistance of Dr Elvin J Moore, Department of Mathematics, King Mongkut's University of Technology North Bangkok in the preparation of this paper.

References

[1] G. Kalton, and D. Kasprzyk, "Imputing for missing survey responses", in Proceedings Section of Survey Research Method. American Statistical Association, 1982, pp. 22-33.
[2] JM. Brick, and G. Kalton, "Handling missing data in survey research, "Statistical Methods in Medical Research, Vol. 5, No. 3, pp. 215- 238, 1996.
[3] D.B. Rubin, Multiple imputation for nonresponse in survey, New York: John Wiley and Sons, 1987.
[4] M. Rueda, S. Gonzalez, A. Arcos, Y. Roman,M.D. Martinez, and J.F. Munoz, "Estimation of the population mean using auxiliary information when some observations are missing," International symposium on applied stochastic models and data analysis, May 17-20, Brest France, 2005.
[5] Abu-Dayyeh, A. Walid, M.S. Ahmed, R.A. Ahmed, and Hassen A. Muttlak, "Some estimators of a finite population mean using auxiliary information," Applied Mathematics and Computation, vol. 139, pp. 287-298, 2003.
[6] C. Kadilar, and H. Cingi, "A new estimator using two auxiliary variables," Applied Mathematics and Computation, vol. 162, pp. 901-908, 2005.
[7] C. Kadilar, and H. Cingi, "Improvement in estimating the population mean in simple random sampling," Applied Mathematics Letters, vol. 19, pp. 75-79, 2006.
[8] C. Kadilar, M. Candan, and H. Cingi, "Ratio estimators using robust regression," Journal of Mathematics and Statistics, vol. 36, pp. 181 – 188, 2007.
[9] N. Nangsue, and J. Sappakitkamjorn, "Variance estimation of the ratio estimator of the population mean in simple random sampling," in Proceedings of Third Workshop on Statistics, Mathematics and Computation and First Portuguese-Polish Workshop on Biometry, Lisbon, Portugal, 21 – 22 July, 2008, p. 59.
[10] W.G. Cochran, Sampling Technique, 3rd Ed. New York: John Wiley and Sons, 1977.
[11] D.S. Tracy, and S.S. Osahan, "Random nonresponse on study variable versus on study as well as auxiliary variables," Statistica, vol. 54, pp.163 – 168, 1994.
[12] S. Singh, Advanced sampling theory with applications, vol. II. The Netherlands: Kluwer Academic Press Publishers, 2003.