Jour. Ind. Soc. Ag. Statistics
53(2), 2000 : 115-124

Regression Estimators from Survey Data for Small Sample Sizes

Y.K. Sharma, Randhir Singh[1], Anil Rai[1] and S.S. Verma
Defence Institute of Physiology and Allied Sciences,
Lucknow Road, Delhi-110054
(Received : May, 1997)

SUMMARY

The technique of regression analysis may give misleading inferences when applied to complex survey data and sample size is small. In recent past, number of attempts have been made to find out suitable statistical procedure for regression analysis of survey data under different inferential frame-work. In this article, an attempt has been made to study the performance of different regression estimators of survey data for small sample sizes. The performance of p-weighted estimators were found to be satisfactory for small sample sizes as they are robust against the failure of model assumptions.

Key words : Regression analysis, Survey data, Simulation, Small sample.

1. Introduction

Regression analysis is frequently applied to the data from comlex surveys. It is well known that survey data collected by imposing complex survey design does not satisfy the implicit, identically and independently distributed assumptions of error in regression model. As a consequence of this such results may be misleading.

There are basically two inferential framework to get an estimator of regression coefficient for survey data. First is based on the assumptions of the linear regression model which is optimum if these assumptions are satisfied (Royall [14], Royall and Herson ([15], 16]), Nathan and Holt [10], Nordberg [11], Dumochel and Duncan [4], Demits and Halprin [3]). Second is based on randomisation by which data is collected. The later approach provides p-weighted consistent estimator of regression coefficient by giving suitable weights to the terms of estimator obtained with the help of models (Nathan and Holt [10], Fuller [8], Holt *et al.* [9], Scott and Holt [19]). Christensen ([1], [2]) attempted to obtain the estimator of regression coefficient by incorporating the population structure in the error of the regression model itself.

1 IASRI, Library Avenue, New Delhi-110012

The use of jackknifing technique in the complex surveys has been found to be widespread. In the complex surveys, it becomes difficult to estimate the variance especially when the parameter under consideration is non-linear like correlation coefficient, regression coefficient etc. Therefore, in such situations jackknifing technique has been considered for estimation of variance (Quenouille ([12], [13]), Tuckey [20], Durbin [5], Efron ([6], [7]), and Wu [21]).

In this article, an attempt has been made to compare the properties of ordinary least square (OLS) estimator, maximum likelihood estimator (MLE), p-weighted OLS, p-weighted MLE and their corresponding jackknifed estimators under different sampling designs for estimating regression coefficient and its variance for small sample sizes in case of complex survey designs.

2. *Estimators*

Suppose a finite population U consists of N identifiable units and a sample of n units is drawn with probabilities p(s). For a sampling design p, let π_i (>0) denote the inclusion probability of the i-th unit (i = 1, ..., N). Let Y, X, Z denote dependent, independent and design variables respectively. Let our parameter of interest be finite population regression coefficient B.

$$B = \frac{N\Sigma_U X_i Y_i - \Sigma_U X_i \Sigma_U Y_i}{N\Sigma_U X_i^2 - (\Sigma_U X_i)^2}$$

where Σ_U denotes the summation over all the units of population U.

The OLS estimator of B and its variance is

$$M_1 = \frac{n\Sigma_s x_i y_i - \Sigma_s x_i \Sigma_s y_i}{n\Sigma_s x_i^2 - (\Sigma_s x_i)^2} \tag{2.1}$$

$$V(M_1) = \frac{\sigma^2}{n\Sigma x_i^2 - (\Sigma x_i)^2} \tag{2.2}$$

where, Σ_s denotes the summation over all the units of the sample s and σ^2 is model error variance.

The design consistent estimator of B is

$$M_2 = \frac{\Sigma_s \frac{1}{\pi_i} \Sigma_s \frac{y_i x_i}{\pi_i} - \Sigma_s \frac{y_i}{\pi_i} \Sigma_s \frac{x_i}{\pi_i}}{\Sigma_s \frac{1}{\pi_i} \Sigma_s \frac{x_i^2}{\pi_i} - (\Sigma_s \frac{x_i}{\pi_i})^2} \tag{2.3}$$

The p-weighted variance of M_2 can be obtained by using Taylor series linearization technique as

$$V(M_2) = \frac{\Sigma \Sigma_U \Delta_{ij} \frac{(X_i - \bar{X}_U)}{\pi_i} E_i \frac{(X_j - \bar{X}_U)}{\pi_j} E_j}{\left[\Sigma_U \frac{(X_i - \bar{X}_U)^2}{\pi_i}\right]^2} \tag{2.4}$$

where, X_U is the mean of all units belonging to population U and

$$\Delta_{ij} = \pi_{ij} - \pi_i \pi_j \text{ and } E_i = Y_i - BX_i$$

Demits and Halprin [3] derived MLE estimator of B under multivariate normality assumption of the target population as

$$M_3 = \frac{s_{yx} + \left(\frac{s_{yz}\, s_{xz}}{s_z^2}\right)\left(\frac{\hat{\sigma}_z^2}{s_z^2} - 1\right)}{s_x^2 + \left(\frac{s_{xz}^2}{s_z^2}\right)\left(\frac{\hat{\sigma}_z^2}{s_z^2} - 1\right)} \tag{2.5}$$

The sample statistics

$$\bar{x}_i\,,\ s_i^2\,,\ s_{ij}\,,\ \rho_{ij}\,,\ \rho_{ij.k} \qquad (i, j, k = y, x, z)$$

are defined in the usual way analogous to the corresponding distribution parameters

$$\mu_i\,,\ \sigma_i^2\,,\ \sigma_{ij}\,,\ \sigma_{ij.k}$$

The variance of M_3 is given by

$$V(M_3) = (1 - \rho_{yx}^2)\frac{\sigma_y^2}{n\,\sigma_x^2}\left[(1 - \rho_{yz.x}^2)\left\{1 - \rho_{xy}^2\left(\frac{1}{Q} - 1\right)\right\} + \frac{1}{Q}\rho_{yz.x}^2(1 - 2\rho_{xz}^2)^2 + (1 - \rho_{xz}^2)\,\rho_{xz}^2\,\rho_{yz.x}^2\left\{\frac{\mu_4(x \mid z)}{\sigma_{x.z}^4} - 1 + \frac{n}{n}\left(\frac{\mu_4(z)}{\sigma_z^4} - 1\right)\right\}\right] \tag{2.6}$$

where, $\mu_4(z)$ is the fourth moment of z variable.

The corresponding p-weighted estimator of B as given in Nathan and Holt [10] is

$$M_4 = \frac{s^*_{yx} + \left(\frac{s^*_{yz}\, s^*_{xz}}{s^{*2}_z}\right)\left(\frac{\hat{\sigma}^2}{s^{*2}_z} - 1\right)}{s^{*2}_x + \left(\frac{s^*_{xz}}{s^{*2}_z}\right)\left(\frac{\hat{\sigma}^{*2}}{s^{*2}_z} - 1\right)} \tag{2.7}$$

where

$$\bar{x}^*_{ii} = \frac{\Sigma_{\alpha \varepsilon s}\, x_{i\alpha}}{\pi_\alpha N}$$

$$s^*_{ij} = \frac{\Sigma_{\alpha \varepsilon s}\, x_{i\alpha}\, x_{j\alpha}}{N \pi_\alpha} - \frac{\bar{x}^*_i\, \bar{x}^*_j}{\Sigma \frac{1}{N \pi_\alpha}}$$

$$s^{*2} = s^*_{ii} \quad (i, j = y, x, z \; ; \; \alpha = 1, \ldots, N)$$

$$\pi_\alpha = \text{prob}\ (\alpha \varepsilon s \mid z) > 0$$

Variance of M_4 is given by

$$V(M_4) = \frac{\sigma^2_y}{\sigma^2_x}(1 - \rho^2_{yx}) \Bigg[(1 - \rho^2_{xz})(1 - \rho^2_{yz\,.\,x}) + [\, \rho^2_{xz}(1 - \rho^2_{yz\,.\,x}) + \rho^2_{yz\,.\,x}(1 - 2\rho^2_{xz})\,]\, E_\alpha (\Sigma_{z_\alpha^2} \mid N p_\alpha) + \rho^2_{xz}\, \rho^2_{yz\,.\,x}(1 - \rho^2_{xz}) \left(\frac{\mu_4 (x \mid z)}{\sigma^2_{x\,.\,z}}\right) \Bigg] \tag{2.8}$$

The variance of M_3 and M_4 are obtained by Demits and Halprin [3] under trivariate normality assumption of target population, whereas, it is derived by Nathan and Holts [10] under less stringent conditions.

The technique of jackknifing is one of the most important variance estimation method for non-linear statistics in complex surveys. It is easy to obtain corresponding jackknifed estimator of regression coefficient and its estimate of variance for M_1, M_2, M_3 and M_4. Let the estimator of regression coefficient corresponding to the estimators M_1, M_2, M_3 and M_4 respectively based on sample of size n be given by

$$\hat{\theta} = M_{i(j)} \quad (i = 1, 2, 3, 4) \tag{2.9}$$

Let $\hat{\theta}_{(i)}$ be the estimate of regression coefficient obtained from recomputing $\hat{\theta}$ with i^{th} pair (y_i, x_i) deleted from the sample. Therefore, the pseudo-values are

$$\hat{\theta}_i = n\hat{\theta} - (n-1)\hat{\theta}_{(i)}$$

The ordinary jackknife point estimator of θ is then given by

$$\hat{\bar{\theta}} = m_{k(J)} = \frac{1}{n}\Sigma_i^n \hat{\theta}_i \quad (k = 1, 2, 3, 4) \tag{2.10}$$

and jackknife variance estimate of $\hat{\bar{\theta}}$ is

$$\hat{V}(\hat{\bar{\theta}}) = \frac{1}{n(n-1)}\Sigma_i^n (\hat{\theta}_i - \hat{\bar{\theta}})(\hat{\theta}_i - \hat{\bar{\theta}})' \tag{2.11}$$

3. Simulation

A multivariate normal population of size 5000 is simulated with the help of a data obtained from Fisheries survey of West Bengal using the algorithm given by Schewer and Stoller [18]. The total catch from a pond (y) is a dependent variable, total quantity of seed used in a pond (x) is an independent variable and area of the pond (z) is design variable. The various sampling design considered in this study on the basis of z are (i) simple random sampling with replacement (srswr) (ii) simple random sampling without replacement (srswor) (iii) probability proportional to size with replacement (ppswr) (iv) probability proportional to size without replacement (ppswor). For each sampling design eight different estimators are given by (2.1), (2.3), (2.5), (2.7) and (2.10) are considered for this emperical study. The bias, mean square error (m.s.e) and bias ratio (B.R) for each of the eight estimators under four sampling designs were obtained based on 1000 samples of equal sizes and properties of the above statistics were studied. The samples are selected by the algorithm given by Sampford [17].

The results of bias, m.s.e., and B.R. for various estimators under different sampling designs are presented in Table 1 and Table 2. The following conclusions may be drawn regarding the bias of the estimators. (1) In case of small sample sizes, the bias of the estimators under all the sampling design is negative (*i.e.*, underestimate) and with the increase in sample size bias decreases and becomes almost zero, which implies that all the estimators are consistent. (2) In case of srswr and srswor designs, the estimators M_1 and M_2 are identical and are less biased than other estimators. (3) In case of ppswr and ppswor, M_1 and M_2 are less biased and M_4 compares favourably with M_1

Table 1 : Bias, mean square error and bias ratio of various estimators for srswr and srswor designs for different sample sizes

		Estimators							
n	Design	M_1	M_2	M_3	M_4	$M_{1(i)}$	$M_{2(i)}$	$M_{3(i)}$	$M_{4(i)}$
10	wr	–0.01	–0.01	1.02	–0.01	–0.07	–0.07	–0.06	1.61
		(0.33)	(0.33)	(1.32)	(0.34)	(4.73)	(4.73)	(6.21)	(18.92)
		2.36*	2.36*	89.20*	2.50*	3.57*	3.57*	2.68*	37.00*
10	wor	–0.02	–0.02	0.99	–0.01	– 0.01	– 0.01	–0.07	–1.85
		(0.35)	(0.35)	(1.34)	(0.38)	(4.34)	(4.34)	(6.04)	(1.88)
		3.66*	3.67*	86.03*	3.21*	0.69*	0.69*	3.01*	135.10*
20	wr	–0.00	–0.00	1.00	0.00	–0.08	–0.08	–0.17	–3.57
		(0.14)	(0.14)	(11.09)	(0.14)	(3.00)	(3.00)	(3.19)	(1.65)
		0.25*	0.25*	95.72*	0.13*	4.90*	4.89*	9.71*	277.86*
20	wor	–0.00	–0.00	1.01	–0.00	–0.00	–0.00	–0.08	–3.50
		(0.13)	(0.13)	(1.08)	(0.13)	(2.79)	(2.79)	(2.96)	(2.85)
		2.03*	2.03*	97.11*	1.90*	0.13*	0.52*	4.91*	207.48*
30	wr	0.01	0.01	0.07	0.02	– 0.07	0.07	–0.17	–4.65
		(0.10)	(0.10)	(0.10)	(0.10)	(2.66)	(2.66)	2.78	(1.55)
		5.04*	5.04*	22.69*	6.47*	4.59*	4.59*	10.23*	372.19*
30	wor	0.01	0.01	1.06	0.01	0.04	0.04	–0.06	–4.76
		(0.09)	(0.09)	(0.09)	(0.09)	(2.76)	(2.76)	(2.86)	(1.59)
		3.31*	3.31*	20.45*	3.66*	2.62*	2.61*	3.68*	347.40*
50	wr	0.00	0.00	0.07	0.00	0.00	0.00	–0.13	–7.76
		(0.05)	(0.05)	(0.05)	(0.05)	(2.57)	(2.57)	(2.62)	(1.61)
		2.75*	2.61*	3.04*	3.09*	0.53*	0.52*	8.42*	610.69*
50	wor	0.01	0.02	0.07	0.02	–0.19	–0.04	–0.16	–7.62
		(0.05)	(0.05)	(0.04)	(0.05)	(2.60)	(2.60)	(2.64)	(1.65)
		8.09*	9.60*	36.50*	9.90*	12.38*	2.53*	10.17*	592.98*

n = Sample size

In the table first entry represents bias, Quantity in the brackets ( ) represent m.s.e.,

Quantity with asterisk (*) represent the bias ratio.

Table 2 : Bias, mean square error and bias ratio of various estimators for ppswr and ppswor designs for different sample sizes

		Estimators							
n	Design	M_1	M_2	M_3	M_4	$M_{1(j)}$	$M_{2(j)}$	$M_{3(j)}$	$M_{4(j)}$
10	wr	–0.00	0.00	1.01	0.00	–0.19	– 0.02	-0.00	–1.95
		(0.49)	(0.46)	(1.40)	(0.47)	(6.56)	(6.56)	(8.30)	(1.26)
		1.33*	1.23*	86.02*	1.22*	7.74*	0.80*	0.15*	173.90*
10	wor	0.03	0.50	1.05	0.46	0.12	0.27	0.26	–1.75
		(0.44)	(0.40)	(1.41)	(0.42)	(5.68)	(5.68)	(7.12)	(1.30)
		5.17*	7.84*	88.60*	7.09*	5.16*	11.57*	10.04*	153.59*
20	wr	–0.01	0.00	1.00	0.00	–0.12	0.10	0.81	–3.73
		(0.22)	0.18	(1.16)	(0.18)	(3.81)	(3.81)	(4.07)	(1.35)
		2.23*	0.19*	94.54*	0.92*	6.46*	5.13*	4.03*	320.99*
20	wor	0.00	0.02	1.02	0.02	–0.33	0.01	0.00	–3.72
		(0.22)	(0.19)	(1.14)	(0.19)	(3.80)	(3.80)	(4.05)	(1.33)
		1.56*	5.07*	95.99*	5.53*	16.98*	1.01*	0.17	322.73*
30	wr	0.02	0.27	0.07	0.02	0.07	0.24	0.17	–5.29
		(0.14)	(0.12)	(0.13)	(0.12)	(3.46)	(3.46)	(3.56)	(1.35)
		5.78*	7.92*	20.66*	7.63*	3.97*	12.90*	9.19*	454.90*
30	wor	0.02	0.04	0.06	0.04	-0.69	0.09	0.01	–5.56
		(0.13)	(0.11)	(0.12)	(0.11)	(3.45)	(3.45)	(3.53)	(1.36)
		5.14*	13.51*	19.38*	13.63*	37.33*	5.34*	0.60*	476.20*
50	wr	0.03	0.04	0.11	0.04	–0.47	0.09	–0.04	–7.970
		(0.09)	(0.07)	(0.09)	(0.07)	(3.39)	(3.39)	(3.42)	(1.36)
		10.10*	15.79*	38.08*	15.64*	25.90*	5.16*	2.25*	682.70*
50	wor	0.01	0.03	0.07	0.03	–0.81	0.19	0.11	–8.95
		(0.08)	(0.06)	(0.10)	(0.06)	(3.49)	(3.49)	(3.51)	(1.40)
		2.24*	12.77*	23.50*	12.60*	44.03*	10.50*	5.94*	756.30*

n = Sample size

In the table first entry represent bias, Quantity in the brackets ( ) represent m.s.e.,
Quantity with asterisk (*) represent the bias ratio.

and M_2. (4) M_4 is better than M_3 in case of all the sampling designs as it takes care of the probability structure of the design. (5) Jackknifed estimators do not lead to reduction in the bias as compared to non-jackknifed estimators, as these estimators are independent of the sampling structure of the design and some form of weighting is required.

An examination of the Tables 1-2 leads to following conclusions with regards to m.s.e. (1) In case of equal probability sampling design, the estimators M_1 and M_2 have m.s.e of same order and are better than other estimators except M_4 which compares favourably with M_1 and M_2 as it is π-weighted maximum likelihood estimator. (2) In case of ppswr and ppswor, the estimator M_4 is superior than other estimators because it takes into account the sample selection in some way. (3) For small samples M_4 has slightly higher m.s.e as compared to M_1 and M_2. (4) For large samples, M_3 compares favourably with other estimators. (5) The performance of jackknife estimators is not satisfactory as compared to other estimators as they are not weighted according to sampling design used.

The B.R. is obtained to see the relation between bias and m.s.e. This relationship is important for confidence interval to be valid. The conclusions which follow from Tables 1-2 are : (1) The B.R. of the estimators M_1, M_2, M_4, $M_{2(j)}$ and $M_{3(j)}$ are in general better than other estimators for all the designs considered in this study. For particular design some conclusions are (i) for srswr M_1, M_2, $M_{2(j)}$ and $M_{3(j)}$ are similar. (ii) for srswor M_1, M_2, $M_{1(j)}$ and $M_{2(j)}$ appear superior compared to others. (iii) for ppswr M_1 and $M_{3(j)}$ have comparatively less B.R. (iv) for ppswor M_1, M_4 and $M_{3(j)}$ are better than rest of the estimators. However, from the results based on B.R., it is observed that no particular estimator is found to be better.

Combining the results obtained above, we conclude that for self weighting designs (srswr and srswor) estimators M_1 and M_2 are almost same in terms of both bias m.s.e and B.R. In case of unequal probability sampling design (ppswr and ppswor), the estimator M_4 is superior to other estimators in respect of its bias and m.s.e as target population is multivariate normal satisfying the assumptions required for this estimator. Also, it takes into account the inclusion probabilities of the units.

In terms of B.R., M_4 compares favourably with other estimators particularly jackknifed estimators $M_{2(j)}$ and $M_{3(j)}$, but in terms of m.s.e., these jackknife estimators are less efficient than other estimators. However, jackknife

variance estimators have the advantage that they are simple, noninvasive and computationally attractive and can be improved further by suitable adjustments for given situations.

ACKNOWLEDGEMENT

The authors are grateful to Dr. W. Selvamurthy, Director, Defence Institute of Physiology and Allied Sciences, for his keen interest and encouragement.

REFERENCES

[1] Christensen, R. (1984). A note on ordinary least squares methods for two-stage sampling. *J. Amer. Statist. Assoc.*, **79,** 720-721.

[2] Christensen, R. (1987). The analysis of two-stage sampling data by ordinary least squares. *J. Amer. Statist. Assoc.*, **82,** 492-498.

[3] Demits, D. and Halprin, M. (1977). Estimation of single regression coefficient in samples arising from a sub-sampling procedure. *Biometrics,* **33,** 47-56.

[4] DuMouchel, W.H. and Duncan, G.J., (1983). Using sample survey weights in multiple regression analysis of stratified samples. *J. Amer. Statist. Assoc.*, **78,** 535-543.

[5] Durbin, J. (1959). A note on application of Quenouille's method of bias reduction to the estimation of ratios. *Biometrika,* **46,** 477-480.

[6] Efron, B. (1979). Bootstrap method : Another look at the jackknife. *Ann. Statist.*, **7,** 1-26.

[7] Efron, B. (1982). The jackknife, the bootstrap and other resampling plans. *SIAM,* Philadelphia.

[8] Fuller, W.A. (1975). Regression analysis for sample survey. *Sankhya C: The Indian Journal of Statistics,* **37,** 3, 117-132.

[9] Holt, D., Smith T.M.F. and Winter, P.D. (1980). Regression analysis of data from complex surveys. *J. Roy. Statist. Soc. A.* **143,** 474-487.

[10] Nathan, G. and Holt, D. (1980). The effect of survey design on regression analysis. *J. Roy. Statist. Soc. B,* **42,** 377-386.

[11] Nordberg, L. (1989). Generalized linear modelling of sampling survey data. *Journal of Official Statistics,* **5,** 223-239.

[12] Quenouille, M.H. (1949). Problems in plane sampling. *Annals of Mathematical Statistics,* **20,** 355-375.

[13] Quenouille, M.H. (1956). Notes on bias in estimation. *Biometrika,* **43,** 353-360.

[14] Royall, R.M. (1970a). On finite population sampling theory under certain linear regression models. *Biometrika,* **57,** 377-389.

[15] Royall, R.M. and Herson, J. (1973a). Robust estimation in finite populations-I. *J. Amer. Statist. Assoc.*, **68,** 880-890.

[16] Royal, R.M. and Herson, J. (1973 b). Robust estimation in finite populations-II. Stratification on the size variable. *J. Amer. Statist. Assoc.*, **68,** 890-893.

[17] Sampford, M.R. (1967). On sampling without replacement with unequal probability of selection. *Biometrika,* **54,** 499-513.

[18] Schewer, E.M. and Stoller, D.S., (1962). On generation of normal random vectors. *Technometrics,* 278-281.

[19] Scott, A.J. and Holt, D. (1982). The effect of two-stage sampling on ordinary least squares method. *J. Amer. Statist. Assoc.*, **77,** 844-854.

[20] Tuckey, J.W. (1958). Bias and confidence in not quite large samples. *Annals of Mathematical Statistics,* **29,** 614.

[21] Wu, C.F.J. (1986). Jackknife, bootstrap and other resampling methods in regression analysis. *The Annals of Statistics,* **4,** 1261-1295.