Gazi University Journal of Science
GU J Sci
29(4): 793-798 (2016)

Ratio Estimators for Population Mean Using Robust Regression in Double Sampling

Muhammad NOOR UL AMIN [1,♠], Muhammad QAISER SHAHBAZ [2], Cem KADILAR [3]

[1] *COMSATS Institute of Information & Technology, Lahore, Pakistan*
[2] *Department of Statistics, King Abdulaziz University, Jeddah, Saudi Arabia*
[3] *Hacettepe University, Department of Statistics, Ankara, Turkey*

Received: 04/05/2016 *Accepted: 07/10/2016*

ABSTRACT

In this article, ratio estimators for the population mean are suggested using the robust regression under the double sampling scheme. The mean squared error (MSE) expressions are obtained for the first degree of approximation. Theoretical comparisons show that the proposed estimators having the robust regression estimates are more efficient than the estimators using the least square method under the certain conditions. Theoretical findings are supported with the aid of a real life dataset in application and a simulation study is also conducted to evaluate the performance of the proposed estimators.

Keywords: *Auxiliary information, double sampling, robust regression.*

1. INTRODUCTION

Robust method of regression has become a familiar and dependable tool for most of the empirical problems in the presence of the outliers. The commonly used estimation technique in the robust regression is Huber M-estimation that bears the impact of outliers. The least square estimates are highly sensitive to the outliers that reduce the efficiency of the classical estimators.

Huber (1964) introduced the M-estimator for the robust regression by considering the following linear model

$$y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} + \varepsilon_i$$
$$= x_i^{'}\beta + \varepsilon_i$$

♠Corresponding author, e-mail: nooramin.stats@gmail.com

where $\beta' = [\beta_1, \beta_2, \ldots, \beta_k]$ is a vector whose elements are the population regression coefficients, ε is the residual term, y is the study variable and $x_i = [x_{i1}, x_{i2}, \ldots, x_{ik}]$ is k auxiliary variables.

For the sample of n observations, the model is given by

$$y_i = x_i'b + e_i$$

where $b' = [b_1, b_2, \ldots, b_k]$ is a vector whose elements are the sample regression coefficients and e_i is the residual term of the regression model.

The M-estimates are obtained by minimizing the following objective function with respect to the estimators

$$\sum_{i=1}^{n} \rho(y_i - x_i'b)^2$$

where the function ρ represents the contribution of each residual to the objective function.

Kadilar et al. (2007) suggested the following ratio type estimators for the population mean using the robust regression (b_{rob}) under the simple random sampling without replacement (SRSWOR):

$$t_1 = \frac{\bar{y} + b_{rob}(\bar{X} - \bar{x})}{\bar{x}} \bar{X},$$

$$t_2 = \frac{\bar{y} + b_{rob}(\bar{X} - \bar{x})}{\bar{x} + C_x}\left(\bar{X} + C_x\right),$$

$$t_3 = \frac{\bar{y} + b_{rob}(\bar{X} - \bar{x})}{\bar{x} + \beta_2(x)}\left[\bar{X} + \beta_2(x)\right].$$

The use of auxiliary variable improves the efficiency of the estimator. Ratio estimator may be used when the linear correlation between study and auxiliary variables is positive. Various authors have utilized the auxiliary information for the estimation of population mean, such as Grover et al. (2012), Noor ul amin and Hanif (2012) and Sanaullah et al. (2014) for recent studies. We also use the population parameters, such as coefficient of variation $\left(C_x\right)$ and coefficient of kurtosis $\left[\beta_2(x)\right]$, in the proposed estimators presented in Section 2.

Double sampling design is often considered when the information about population parameter of auxiliary variable is not known. Neyman (1938) was the first to use the method of double sampling to collect information on the strata sizes. Chand (1975), Kiregyera (1980, 1984), Srivenkataramana and Tracy (1989), Singh and Vishwakarma (2007), Choudhury and Singh (2012), Vishwakarma and Gangele (2014) have utilized the auxiliary information under the double sampling (or two phase sampling) method.

Consider the finite population $S = \{x_1, x_2, \ldots\ldots, x_N\}$ of size N. A first phase sample of size $n_1 (n_1 < N)$ is used to observe the auxiliary variable only while both study and auxiliary variables are studied on the second phase sample of size $n_2 (n_2 < n_1)$. The sample means for variables Y and X are denoted by $\bar{y}$ and $\bar{x}$, respectively. We assume that $\bar{x}_1$ is the sample mean of variable X for the first phase sample, $\bar{y}_2$ and $\bar{x}_2$ are the means of variables Y and X, respectively, for the sample at the second phase. The sample is drawn by SRSWOR at both phases.

We also use following notations:

$$\bar{e}_{y_2} = \frac{\bar{y}_2 - \bar{Y}}{\bar{Y}},\ \bar{e}_{x_i} = \frac{\bar{x}_i - \bar{X}}{\bar{X}}, \text{where } i = 1,2 \quad (1.1)$$

$$E\left(\bar{e}_{y_2}\right) = E\left(\bar{e}_{x_1}\right) = E\left(\bar{e}_{z_2}\right) = 0\ E\left(\bar{e}_{y_2}^2\right) = \theta_2 C_y^2, \quad E\left(\bar{e}_{x_i}^2\right) = \theta_i C_x^2,$$

$$E\left(\bar{e}_{y_2}\bar{e}_{x_i}\right) = \theta_i \rho_{xy} C_x C_y, E\left(\bar{e}_{y_2}\bar{e}_{z_i}\right) = \theta_i \rho_{yz} C_z C_y, \theta_i = \frac{1}{n_i} - \frac{1}{N},\ C_x = \frac{S_x}{\bar{X}},\ H_{yx} = \rho_{yx}\frac{C_y}{C_x}.$$

2. PROPOSED ESTIMATORS

If we adapt the estimators suggested by Kadilar and Cingi (2004) to the double sampling scheme, we can develop the following estimators:

$$t_{a1} = \frac{\bar{y}_2 + b(\bar{x}_1 - \bar{x}_2)}{\bar{x}_2}\bar{x}_1, \quad (2.1)$$

$$t_{a2} = \frac{\bar{y}_2 + b(\bar{x}_1 - \bar{x}_2)}{\bar{x}_2 + C_x}\left(\bar{x}_1 + C_x\right), \quad (2.2)$$

$$t_{a3} = \frac{\bar{y}_2 + b(\bar{x}_1 - \bar{x}_2)}{\bar{x}_2 + \beta_2(x)}\left[\bar{x}_1 + \beta_2(x)\right]. \tag{2.3}$$

Motivated by the adapted estimators of Kadilar et al. (2007), given in (2.1)-(2.3), we propose the estimators using the Huber M-estimator (b_{rob}) for the regression coefficient under the double sampling as follows:

$$t_{pr1} = \frac{\bar{y}_2 + b_{rob}(\bar{x}_1 - \bar{x}_2)}{\bar{x}_2}\bar{x}_1, \tag{2.4}$$

$$t_{pr2} = \frac{\bar{y}_2 + b_{rob}(\bar{x}_1 - \bar{x}_2)}{\bar{x}_2 + C_x}\left(\bar{x}_1 + C_x\right), \tag{2.5}$$

$$t_{pr3} = \frac{\bar{y}_2 + b_{rob}(\bar{x}_1 - \bar{x}_2)}{\bar{x}_2 + \beta_2(x)}\left[\bar{x}_1 + \beta_2(x)\right]. \tag{2.6}$$

In order to obtain the MSE of the proposed estimators in (2.4)-(2.6), we use the notations (1.1) in (2.4)-(2.6) as

$$\hat{\bar{y}}_{di} = \left[\bar{Y}\left(1+\bar{e}_{y_2}\right) + b_{rob}\bar{X}\left(\bar{e}_{x_1} + \bar{e}_{x_2}\right)\right]\left[\frac{1+\bar{e}_{x_1}+k_i}{1+\bar{e}_{x_2}+k_i}\right]. \tag{2.7}$$

To the first degree of approximation for Taylor series, we ignore the terms with power two or greater, the expression (2.7) is re-written by

$$\hat{\bar{y}}_{di} - \bar{Y} = \bar{Y}\left[\left(\bar{e}_{y_2} - k_i\right) + \left(1-k_i\right)\bar{e}_{x_1} - \left(1+k_i\right)\bar{e}_{x_2}\right] + b_{rob}\bar{X}\left(\bar{e}_{x_1} + \bar{e}_{x_2}\right) \tag{2.8}$$

Taking square on both sides of (2.8) and applying expectations, the MSE of the estimators in (2.4)-(2.6) is given by

$$MSE\left(\hat{\bar{y}}_{di}\right) = \bar{Y}^2\left[\theta_2 C_y^2 + k_i^2 + C_x^2\left\{\theta_1\left(1-k_i\right)^2 - 2\theta_1\left(1-k_i^2\right) + \theta_2\left(1+k_i\right)^2\right\} - 2C_x^2 H_{yx}\left\{\theta_2 - \theta_1\left(1-k_i\right)\right\}\right]$$
$$+ C_x^2 b_{rob}\bar{X}\bar{Y}\left(\theta_2 - \theta_1\right)\left(b_{rob}\bar{X} - 2H_{yx} + 2\left(1+k_i\right)\right), \tag{2.9}$$

where i = 1,2,3, $k_1 = 0,\ k_2 = \frac{C_x}{\bar{X}}$, and $k_3 = \frac{\beta_2(x)}{\bar{X}}$.

3. EFFICIENCY COMPARISONS

In this section, we compare the MSE of the adapted estimators, given in (2.1)-(2.3), with the MSE of proposed estimators, given in (2.4)-(2.6), to derive the conditions for which the proposed estimators will perform better than adapted estimators under the double sampling scheme.

$$MSE\left(t_{pri}\right) < MSE\left(t_{ai}\right), \qquad i = 1,2,3$$
$$\left(b_{rob} - b\right)\left[\bar{X}\left(b_{rob} + b\right) - 2\left(H_{yx} - 1\right) - 2k_i\right] < 0. \tag{3.1}$$

(3.1) is satisfied if the either of followings holds true:

$$\left(b_{rob} - b\right) > 0 \quad \text{and} \quad \left[\bar{X}\left(b_{rob} + b\right) - 2\left(H_{yx} - 1\right) - 2k_i\right] < 0, \tag{3.2}$$

or

$$\left(b_{rob} - b\right) < 0 \quad \text{and} \quad \left[\bar{X}\left(b_{rob} + b\right) - 2\left(H_{yx} - 1\right) - 2k_i\right] > 0, \tag{3.3}$$

From (3.2), we can write

$$b_{rob}+b>2b \text{ and } \left(b_{rob}+b\right)<2\frac{\left(H_{yx}-1\right)+k_i}{\bar{X}},$$

$$2b<b_{rob}+b<2\frac{\left(H_{yx}-1\right)+k_i}{\bar{X}}. \quad (3.4)$$

From (3.3), we have

$$\left(b_{rob}+b\right)<2b \text{ and } 2\frac{\left(H_{yx}-1\right)+k_i}{\bar{X}}<\left(b_{rob}+b\right),$$

$$2\frac{\left(H_{yx}-1\right)+k_i}{\bar{X}}<b_{rob}+b<2b. \quad (3.5)$$

Consequently, using (3.4) and (3.5), the overall efficient region for the proposed estimators is obtained by

$$\min\left(2b,\ 2\frac{\left(H_{yx}-1\right)+k_i}{\bar{X}}\right)<\left(b_{rob}+b\right)<\max\left(2b,\ 2\frac{\left(H_{yx}-1\right)+k_i}{\bar{X}}\right).$$

4. APPLICATION

In this section, we examine the performance of the proposed estimators, given in (2.4) – (2.6), with respect to the adapted estimators, given in (2.1) – (2.3), in double sampling design using a real data set given by Kadilar et al. (2007). This population consists of 106 observations. The two variables used for this analysis are level of apple production (in tons) as study variable (Y) and number of apple trees (1 unit =100 trees) as the auxiliary variable (X). The parameters of the population are as follows:

$$\bar{Y}=2212.59, \quad S_y=11551.53, \quad \rho_{yx}=0.86, \quad B=17.21$$
$$\bar{X}=274.22, \quad S_x=574.61, \quad N=106, \quad B_{rob}=5.02.$$

Further, in order to examine the sensitivity of sample sizes on different estimators, we consider three different sample sizes at the first phase, such as $n_1=30$, 40, and 50. Subsequently, from the first phase sample for each choice of n_1, we choose three different sample sizes, such as $n_2=$ 10, 15, and 20. The relative efficient results reported in Tables 1 and 2 are obtained using the MSE equation in (2.9) and the simulation study, respectively, with the aid of the following formula:

	$n_1=30$			$n_1=40$			$n_1=50$		
n_2	RE_1	RE_2	RE_3	RE_1	RE_2	RE_3	RE_1	RE_2	RE_3
10	2.06	1.18	1.25	2.04	1.17	1.24	2.06	1.18	1.24
15	1.78	1.12	1.16	1.85	1.12	1.18	1.99	1.16	1.20
20	1.53	1.07	1.10	1.65	1.08	1.13	1.76	1.11	1.15

$$RE_i=\frac{MSE(t_{ai})}{MSE(t_{pri})}; \quad i=1,2,3. \quad (4.1)$$

Table 1. Relative Efficiencies of Proposed Estimators with respect to Adapted Estimators Using the MSE Equations

5. SIMULATION STUDY

In this section, we present the results from the simulation study using the same population in Section 4 to evaluate the performance of three proposed estimators with respect to three adapted estimators, mentioned in Section 2. The mean square errors are computed by using the following formula:

$$MSE\left(\hat{\bar{Y}}\right)=\frac{1}{10000}\sum_{i=1}^{10000}\left(\hat{\bar{Y}}-\mu_Y\right)^2, \qquad (5.1)$$

where μ_Y is the mean of $\hat{\bar{Y}}$ from 10000 samples.

The simulation study is conducted using the following steps:

1. We select 10000 samples of three different sizes at first phase such as $n_1 = 30$, 40, and 50. From first phase sample, for each choice of n_1, we choose three different samples of sizes, $n_2 =$ 10, 15, and 20, as the second phase sample.
2. Using the samples obtained in Step 1, the 10000 values of t_{ai} and t_{pri}, separately, as $\hat{\bar{Y}}$ in (5.1), are obtained using (2.1)-(2.3) and (2.4)-(2.6), respectively.
3. For each sample, the MSE values of adapted and proposed estimators, given in (2.1)-(2.3) and (2.4)-(2.6), respectively, are computed using (5.1).
4. Using the MSE values found in Step 3, the values of relative efficiencies are obtained by (4.1) and reported in Table 2.

	$n_1 = 30$			$n_1 = 40$			$n_1 = 50$		
n_2	RE_1	RE_2	RE_3	RE_1	RE_2	RE_3	RE_1	RE_2	RE_3
10	5.37	5.23	3.75	6.67	6.52	4.84	7.39	7.20	5.10
15	2.61	2.63	2.08	3.62	3.54	2.70	3.77	3.68	2.75
20	1.72	1.70	1.50	2.36	2.32	1.91	2.64	2.59	2.05

Table 2. Relative Efficiencies of Proposed Estimators with respect to Adapted Estimators Using Simulation Study

The simulation results show that the greater efficiencies of the proposed estimators are always obtained as compared to the adapted estimators. Especially, when the first phase sample (n_1) increases, the performances of the proposed estimators also increase; whereas, when the second phase sample (n_2) decreases, the performances of the proposed estimators increase again. Further, the simulation results support the theoretical results from the MSE in (2.9) reported in Table 1.

6. CONCLUSION

After adapting the estimators of Kadilar and Cingi (2004) to double sampling design, we also modify the estimators of Kadilar et al. (2007) using the Huber M-estimator for double sampling design. The estimators are useful for positive linear relationship between auxiliary and study variables. The results in Tables 1 and 2 show that M estimation should be utilized for ratio estimators under double sampling design using the simple random sampling at both phases. The authors are still working to use the M-estimation for ratio estimators in other sampling designs.

CONFLICT OF INTEREST

No conflict of interest was declared by the authors.

REFERENCES

[1] Chand, L., "Some Ratio-Type Estimator Based on Two or More Auxiliary Variables", Unpublished Ph.D. dissertation, Iowa State University, Iowa. (1975).

[2] Choudhury, S. and Singh, B. K., "A class of chain ratio-product type estimators with two auxiliary variables under double sampling scheme". *Journal of the Korean Statistical Society*, 41(2); 247-256, (2012).

[3] Grover, L.K., Kaur. P. and Vishwakarma, G.K. "Product type exponential estimators of population mean under linear transformation of auxiliary variable in simple random sampling". Applied Mathematics and Computations, 219; 1937-1946, (2012).

[4] Huber, P. H., "Robust estimation of a location parameter", The Annals of Mathematical Statistics", 73-101, (1964).

[5] Kadilar, C. and Cingi, H., "Ratio estimators in simple random sampling", Applied Mathematics and Computation, 151; 893-902, (2004).

[6] Kadilar, C., Candan, M., and Cingi, H. "Ratio estimators using robust regression", Hacettepe Journal of Mathematics and Statistics, 36 (2); 181-188, (2007).

[7] Kiregyera, B., "A chain ratio-type estimator in finite population double sampling using two auxiliary variables", Metrika, 27;217–223, (1980).

[8] Kiregyera, B., "Regression-type estimators using two auxiliary variables and model of double sampling from finite populations", Metrika, 31; 215–226, (1984).

[9] Noor ul Amin, M. and Hanif, M., "Some exponential estimators in survey sampling", Pakistan Journal of Statistics, 28 (3); 367–374, (2012).

[10] Sanaullah, A., Ali H. A., Noor ul Amin, M., and Hanif, M., "Generalized exponential chain ratio estimators under stratified two-phase random sampling". Applied Mathematics and Computation, 226; 541– 547, (2014).

[11] Singh, H. P. and Vishwakarma, G. K., "Modified exponential ratio and product estimators for finite population mean in double sampling", Austrian Journal of Statistics, 36 (3); 217–225, (2007).

[12] Srivenkataramana, T. and Tracy, D. S., "Two-phase sampling for selection with probability proportional to size in sample surveys", Biometrika, 76(4); 818–821, (1989).

[13] Vishwakarma, G. K. and Gangele, R. K., "A class of chain ratio-type exponential estimators in double sampling using two auxiliary variates", Applied Mathematics and Computation, 227; 171-175, (2014).