ISSN (Online) : 2455 - 3662
SJIF Impact Factor :4.924

EPRA International Journal of

Multidisciplinary Research

Monthly Peer Reviewed & Indexed
International Online Journal

Volume: 4 Issue:10 October 2018

Published By :
EPRA Journals

Volume: 4 | Issue: 10 | October 2018 SJIF Impact Factor: 4.924 ISSN (Online): 2455-3662

EPRA International Journal of

Multidisciplinary Research (IJMR)

MODIFIED RATIO ESTIMATORS FOR POPULATION MEAN USING ROBUST REGRESSION BASED ON AUXILIARY ATTRIBUTE

Tolga Zaman[1]
[1]Çankırı Karatekin University,
Faculty of Science,
Department of Statistics,
Çankırı,
Turkey

Hasan Bulut[2]
[2]Ondokuz Mayıs University,
Faculty of Science,
Department of Statistics,
Samsun,
Turkey

ABSTRACT

There are several ratio estimators that estimate the population mean of study variable by using information about a population proportion possessing certain attributes. However when there are outliers in the data, the efficiency of the estimators decreases. For this reason, we adapt least median of squares (LMS) estimation to the proposed estimators by Singh et al. (Ratio Estimators in Simple Random Sampling Using Information on Auxiliary Attribute, Pak.J.stat.oper.res, 2008). Theoretically, we obtain the mean square error (MSE) for these estimators and we compare MSE equations of our suggested estimators and the proposed estimators by Singh et al. (2008). As a result of these comparisons, we observe that suggested estimates give more efficient results than estimates of Sing et. al. (2008) and these theoretical results are supported with the aid of a numerical example and simulation by basing on data that includes an outlier.

KEY WORDS: *Ratio-type estimators, Robust regression method, LMS estimation, Auxiliary attribute, Efficiency*

1. INTRODUCTION

If the relation between study variable y_i and auxiliary variable x_i in simple random sampling method can be shown using a linear equation and when there is a positive correlation between these two variable, ratio estimators are used to estimate population mean. We may want to estimate population mean of the study variable using information about population proportion possessing certain attributes in some researches. In this ratio estimators, population information of the auxiliary variable, such as the coefficient of the variation or the kurtosis, is often used to increase the efficiency of the estimation for a population mean.

Let y_i be ith characteristic of the population and ϕ_i is the case of possessing certain attributes. If ith unit has the desired characteristic, it takes the value 1, if not then the value 0. That is;

$\phi_i = \begin{cases} 1 & , \quad if\ ith\ unit\ of\ the\ population\ possesses\ attribute \\ 0 & , \quad otherwise \end{cases}$

Let $A = \sum_{i=1}^{N} \phi_i$ and $a = \sum_{i=1}^{n} \phi_i$ be the the total count of the units that possess certain attribute in population and sample, respectively. And $P = \frac{A}{N}$ and $p = \frac{a}{n}$ shows the ratio of these units, respectively.

However when outliers exist in the data, it is well known that classical estimators are affected from these outliers and their efficiencies decrease. Therefore, in this study, we suggest to use LMS estimate instead of OLS estimate in order to decrease the effect of outlier problem in data.

The remainder of the paper is organized as follows. First of all in Section 2, estimators and their MSE equations suggested by Singh et al. (2008) are presented for population mean estimate using information about population proportion possessing certain attributes in simple random sampling within the scope of the study. We propose ratio estimators based on LMS estimate and show their MSE equations in Section 3. Efficiency comparisons between the Singh et al. (2008) and the our suggested estimators, based on the MSE equations, are considered in Section 4. The results of numerical examples and simulation are reported in Section 5 and in Section 6, respectively. We arrive at a conclusion from these results in the last section.

2. TRADITIONAL RATION ESTIMATORS

In simple random sampling, Sing et al. (2008) suggested ratio estimators below in order to estimate population mean of study variable y, using information about population proportion possessing certain attributes;

$$t = \frac{\bar{y} + b_{\phi}(P - p)}{(m_1 p + m_2)}(m_1 P + m_2). \qquad (2.1)$$

where $m_1 \neq 0$ and m_2 is either real number or the functions of known parameters such as C_p, $\beta_2(\phi)$ and ρ_{pb}. A scheme is arranged in Table 1 for (2.1) equation based on m_1 ve m_2 constants (Singh et al, 2008).

Table 1: Estimators which were suggested by Sing et al.

Estimators	Values of	
	m_1	m_2
$\boldsymbol{t_1 = \frac{\bar{y} + b_{\phi}(P - p)}{p} P}$	1	0
$\boldsymbol{t_2 = \frac{\bar{y} + b_{\phi}(P - p)}{(p + \beta_2(\phi))}[P + \beta_2(\phi)]}$	1	$\beta_2(\phi)$
$\boldsymbol{t_3 = \frac{\bar{y} + b_{\phi}(P - p)}{(p + C_p)}[P + C_p]}$	1	C_p
$\boldsymbol{t_4 = \frac{\bar{y} + b_{\phi}(P - p)}{(p + \rho_{pb})}[P + \rho_{pb}]}$	1	ρ_{pb}
$\boldsymbol{t_5 = \frac{\bar{y} + b_{\phi}(P - p)}{(p\beta_2(\phi) + C_p)}[P\beta_2(\phi) + C_p]}$	$\beta_2(\phi)$	C_p
$\boldsymbol{t_6 = \frac{\bar{y} + b_{\phi}(P - p)}{(pC_p + \beta_2(\phi))}[PC_p + \beta_2(\phi)]}$	C_p	$\beta_2(\phi)$
$\boldsymbol{t_7 = \frac{\bar{y} + b_{\phi}(P - p)}{(pC_p + \rho_{pb})}[PC_p + \rho_{pb}]}$	C_p	ρ_{pb}
$\boldsymbol{t_8 = \frac{\bar{y} + b_{\phi}(P - p)}{(p\rho_{pb} + C_p)}[P\rho_{pb} + C_p]}$	ρ_{pb}	C_p
$\boldsymbol{t_9 = \frac{\bar{y} + b_{\phi}(P - p)}{(p\beta_2(\phi) + \rho_{pb})}[P\beta_2(\phi) + \rho_{pb}]}$	$\beta_2(\phi)$	ρ_{pb}
$\boldsymbol{t_{10} = \frac{\bar{y} + b_{\phi}(P - p)}{(p\rho_{pb} + \beta_2(\phi))}[P\rho_{pb} + \beta_2(\phi)]}$	ρ_{pb}	$\beta_2(\phi)$

In Table 1, C_p, $\beta_2(\phi)$ and ρ_{pb} are, respectively, coefficient of variation belonging to ratio of units possessing certain attributes, coefficient of population kurtosis and population correlation coefficient between ratio of units possessing certain attributes and study variable. $\bar{y}$ and p are, respectively, sample mean belonging to study variable and sample proportion possessing certain attributes. b_{ϕ} is the coefficient of the slope obtained by least squares method. It is calculated using $b = \frac{s_{\phi y}}{s_{\phi}^2}$ and is unbiased. s_{ϕ}^2 is the sample variance of unit ratios possessing

certain attributes and $s_{\phi y}$ is sample covariance between units ratio possessing certain attributes and study variable. Also, it assumed that population proportion possessing certain attribute is known in ratio estimators.
The MSE values of estimators given in Table 1 are obtained as below by using Taylor series approach (Singh et al, 2008);

$$MSE(t_i) \cong \frac{1-f}{n}\left[R_i^2 S_\phi^2 - B_\phi \rho_{pb} S_\phi S_y + S_y^2\right] \quad (2.2)$$

where $i = 1,2,\dots,10$; $B_\phi = \frac{S_{\phi y}}{S_\phi^2}$ and it is found by using least squares method. $\rho_{pb} = \frac{S_{\phi y}}{S_\phi S_y}$, is the point biserial correlation coefficient. $f = \frac{n}{N}$ is sample ratio; N is population size.

$R_1 = \frac{\bar{Y}}{P}$; $R_2 = \frac{\bar{Y}}{P+\beta_2(\phi)}$; $R_3 = \frac{\bar{Y}}{P+C_p}$; $R_4 = \frac{\bar{Y}}{P+\rho_{pb}}$, $R_5 = \frac{\bar{Y}\beta_2(\phi)}{P\beta_2(\phi)+C_p}$, $R_6 = \frac{\bar{Y}C_p}{PC_p+\beta_2(\phi)}$,
$R_7 = \frac{\bar{Y}C_p}{PC_p+\rho_{pb}}$, $R_8 = \frac{\bar{Y}\rho_{pb}}{P\rho_{pb}+C_p}$, $R_9 = \frac{\bar{Y}\beta_2(\phi)}{P\beta_2(\phi)+\rho_{pb}}$ and $R_{10} = \frac{\bar{Y}\rho_{pb}}{P\rho_{pb}+\beta_2(\phi)}$

Expressions above are population ratio. S_ϕ^2 is population variance of units ratio possesing certain attribute and S_y^2 is population variance of the study variable.
In this study, we suggest new ratio estimators based on LMS estimate as slope coefficient of estimators given in Table 1 instead of OLS estimate.

3. SUGGESTED ESTIMATORS

When there is an outlier in the data set, the efficiency of traditional methods decreases. In order to solve this problem, Kadilar et al. (2007) adapted Huber-M method which is only one of robust regression methods to ratio-type estimators and decreased the effect of outlier problem. Then, new ratio-type estimators proposed by considering Tukey-M, Hampel M, Huber MM, LTS, LMS and LAD robust methods by Zaman and Bulut (2018).

Thus, we propose to apply the following 10 ratio estimators for estimate population mean of study variable based on information about population proportion possessing certain attributes using robust regression, instead of ratio estimators presented in Table 1, to data which have outliers:

Table 2: Suggested Estimators

Estimators	Values of	
	m_1	m_2
$\boldsymbol{t_{pr1} = \frac{\bar{y} + b_{\phi(rob)}(P-p)}{p}P}$	1	0
$\boldsymbol{t_{pr2} = \frac{\bar{y} + b_{\phi(rob)}(P-p)}{(p+\beta_2(\phi))}[P+\beta_2(\phi)]}$	1	$\beta_2(\phi)$
$\boldsymbol{t_{pr3} = \frac{\bar{y} + b_{\phi(rob)}(P-p)}{(p+C_p)}[P+C_p]}$	1	C_p
$\boldsymbol{t_{pr4} = \frac{\bar{y} + b_{\phi(rob)}(P-p)}{(p+\rho_{pb})}[P+\rho_{pb}]}$	1	ρ_{pb}
$\boldsymbol{t_{pr5} = \frac{\bar{y} + b_{\phi(rob)}(P-p)}{(p\beta_2(\phi)+C_p)}[P\beta_2(\phi)+C_p]}$	$\beta_2(\phi)$	C_p
$\boldsymbol{t_{pr6} = \frac{\bar{y} + b_{\phi(rob)}(P-p)}{(pC_p+\beta_2(\phi))}[PC_p+\beta_2(\phi)]}$	C_p	$\beta_2(\phi)$
$\boldsymbol{t_{pr7} = \frac{\bar{y} + b_{\phi(rob)}(P-p)}{(pC_p+\rho_{pb})}[PC_p+\rho_{pb}]}$	C_p	ρ_{pb}
$\boldsymbol{t_{pr8} = \frac{\bar{y} + b_{\phi(rob)}(P-p)}{(p\rho_{pb}+C_p)}[P\rho_{pb}+C_p]}$	ρ_{pb}	C_p
$\boldsymbol{t_{pr9} = \frac{\bar{y} + b_{\phi(rob)}(P-p)}{(p\beta_2(\phi)+\rho_{pb})}[P\beta_2(\phi)+\rho_{pb}]}$	$\beta_2(\phi)$	ρ_{pb}
$\boldsymbol{t_{pr10} = \frac{\bar{y} + b_{\phi(rob)}(P-p)}{(p\rho_{pb}+\beta_2(\phi))}[P\rho_{pb}+\beta_2(\phi)]}$	ρ_{pb}	$\beta_2(\phi)$

In Table 2, $b_{\phi(rob)}$ is obtained by basing on least median of squares regression LMS estimate. The main advantage of LMS-estimate over OLS estimates is that they are not sensitive to outliers. Thus, when there are outliers in the data, LMS is more accurate than OLS estimation.

LMS was suggested by Rousseeuw (1984) and improved by Rousseeuw and Leroy (1987). The method has the idea that instead of the sum of error squares, the median of error squares is minimized. Function to minimize is given as;

$$\text{Min } median(\varepsilon_i{}^2) \qquad (3.1)$$

This estimator is robust against outliers in the direction of both x and y and its breakdown point is 0.5 [4]. For this reason, we prefer to use only LMS method in this study.

The algorithm of LMS is defined as following;

i. Regression coefficients are calculated for all pair of observations.

ii. Error terms belonging to n number of observation pairs are obtained for each calculated regression parameter value and median is calculated by squaring these calculated error terms.

iii. Regression estimate values corresponding to the least squared median value among obtained squared median values is taken and process is ended (Rousseeuw and Leroy, 1987).

Calculations belonging to LMS method are made using MASS package in R programming language (Venables and Ripley, 2002).

MSE equations of ratio-type estimators modified based on LMS estimates can be expressed as (2.2). The main difference between MSE equations is usage of $B_{\phi(rob)}$ instead of B_ϕ. MSE equations for all suggested estimators belonging to LMS estimates are obtained as below

$$MSE(t_{pri}) \cong \frac{1-f}{n}\left[R_i^2 S_\phi^2 - B_{\phi(rob)}\rho_{pb}S_\phi S_y + S_y^2\right] \quad ,i = 1,2,\dots,10 \qquad (3.2)$$

4. EFFICIENCY COMPARISONS

In this section, we compare the MSE of the suggested estimators given in (3.1) with the MSE of the ratio estimators given in (2.1).

$$MSE(t_{pri}) < MSE(t_i)$$

$$\frac{1-f}{n}\left[R_i^2 S_\phi^2 - B_{\phi(rob)}\rho_{pb}S_\phi S_y + S_y^2\right] < \frac{1-f}{n}\left[R_i^2 S_\phi^2 - B_\phi \rho_{pb}S_\phi S_y + S_y^2\right]$$

$$B_\phi \rho_{pb}S_\phi S_y - B_{\phi(rob)}\rho_{pb}S_\phi S_y < 0$$

$$\rho_{pb}S_\phi S_y\left(B_\phi - B_{\phi(rob)}\right) < 0$$

For $\rho_{pb}S_\phi S_y > 0$, that is $S_{\phi y} > 0$

$$B_\phi - B_{\phi(rob)} < 0$$

$$B_\phi < B_{\phi(rob)} \qquad (4.1)$$

Similarly, for $\rho_{pb}S_\phi S_y < 0$, that is $S_{\phi y} < 0$

$$B_\phi - B_{\phi(rob)} > 0$$

$$B_\phi > B_{\phi(rob)} \qquad (4.2)$$

When condition (4.1) or (4.2) is satisfied, the suggested estimators given in Table 2 are more efficient than the ratio estimator given in Table 1.

5. NUMERICAL ILLUSTRATIONS

We use the teacher and wdbc data sets to calculate efficiency of estimators which are given in Table 1 and Table 2.

Wdbc data consists of 30 variable and 569 observations. These 596 people have tumor which is type malignant or benign (Nagler, 2017). The first variable of data is taken as study variable y. And we add outlier to data by increasing values of last observation. For this data,

$$\phi_i = \begin{cases} 1 & , \quad if\ tumor\ is\ benign \\ 0 & , \quad if\ tumor\ is\ malignant \end{cases}$$

The population statistics of wdbc data are given in Table 3.

Table 3: wdbc Data Statistics

N:569	$\overline{\boldsymbol{Y}}$: 14.21207	$\boldsymbol{R_1}$: 36.4264369	$\boldsymbol{R_6}$: 0.2460919
n: 150	P: 0.3901582	R_2: 0.1528157	R_7: 16.1856913
$\boldsymbol{C_p}$: 1.6145352	S_y: 3.9277449	R_3: 7.0893994	R_8: 5.8252542
$\boldsymbol{\rho_{pb}}$: 0.788	S_ϕ: 0.6299241	R_4: 12.0656094	R_9: 35.6492429
$\boldsymbol{\beta_2(\phi)}$: 92.6112253	B_ϕ: 4.911776	R_5: 34.8684088	R_{10}: 0.1204865
$\boldsymbol{S_{\phi y}}$: 1.9490141	$B_{\phi(rob)}$: 6.73		

As second example, we use the teachers data which means the number of teachers working at school in Trabzon. (Directorate of National Education, Trabzon). The schools in Trabzon are taken as unit of population (Zaman et al., 2014). The data is defined as following;

$$y = the\ number\ of\ teachers$$

$$\phi_i = \begin{cases} 1 & , \quad if\ the\ number\ of\ teachers\ is\ more\ than\ 40 \\ 0 & , \quad otherwise \end{cases}$$

The population statistics of teachers data are given in Table 4. Similarly, we add outlier to data by increasing values of last observation

Table 4: Teacher Data Statistics

N: 111	$\overline{\boldsymbol{Y}}$: 31.8378378	$\boldsymbol{R_1}$: 114	$\boldsymbol{R_6}$: 1.5323264
n: 40	P: 0.2792793	R_2: 0.4276216	R_7: 61.0078513
$\boldsymbol{C_p}$: 3.6185675	S_y: 36.1852658	R_3: 8.1680578	R_8: 7.2333401
$\boldsymbol{\rho_{pb}}$: 0.878	S_ϕ: 1.0105909	R_4: 27.5154227	R_9: 109.3656218
$\boldsymbol{\beta_2(\phi)}$: 74.1740221	B_ϕ: 31.43095	R_5: 97.0476085	R_{10}: 0.3755432
$\boldsymbol{S_{\phi y}}$: 32.1002457	$B_{\phi(rob)}$: 35		

Both related data set include outliers. So, it is expected that performances of our suggested estimators are better than Singh et al. (2008) estimators under (4.1) or (4.2).

Using simple random sampling method, we assume wdbc data has $n = 150$ sample size and teacher data has $n = 40$ sample size. Here, coefficients of correlation for wdbc and teacher data are 0.788 and 0.878, respectively. MSE equations of estimators given in Table 1 and Table 2 are obtained in Section 2 and 3. By using these equations, we calculate relative efficiency values as below;

$$RE(\bar{y}_{pri}) = \frac{MSE(t_{pri})}{MSE(t_i)}\ ;\ ;\ i = 1,2,\dots,10 \qquad (5.1)$$

Tablo 5: Theoretical Results for the Relative Efficiencies of Suggested estimators with respect to Singh et al (2008) estimators

Estimator	$RE(.,t_i)$	
	Wdbc data	Teachers data
t_{pr1}	0.993197	0.991559
t_{pr2}	0.3823121	0.618895
t_{pr3}	0.859609	0.689157
t_{pr4}	0.9430732	0.893292
t_{pr5}	0.9925829	0.98845
t_{pr6}	0.3838637	0.621678
t_{pr7}	0.9670177	0.972068
t_{pr8}	0.812532	0.676243
t_{pr9}	0.9929006	0.990846
t_{pr10}	0.3819426	0.618841

We compute 10 relative efficiency values, as shown Table 5. If the Relative Efficiency (RE) value obtained from (5.1) is smaller than 1, then it is apparent that our suggested estimator is more efficiency than estimator proposed by Singh et al. (2008). In Table 5, we see that all of our suggested estimators are more efficiency than estimators which were suggested by Singh et al. (2008) both wdbc and teacher data sets, when data includes outliers. This is an expected case, because (4.1) or (4.2) is satisfied for all cases.

6. SIMULATION STUDY

A simulation study is conducted in order to calculated MSE values by using our suggested and Singh et al. (2008) estimators. Because we obtain similar results by using two data sets, wdbc data is only used for simulation study. The simulation design is as follows:

1. Firstly, n sample sized 10000 sample is drawn from the real data set without replacement with simple random sampling.
2. Then, t_i values are calculated 10000 times from 10000 samples based on sample size chosen to calculate t_i values.
3. Lastly, we computed the mean squared errors (MSE) as follows:

$$MSE = \frac{1}{10000}\sum_{i=1}^{10000}(t_i - \overline{Y})^2 \qquad (\mathbf{6.1})$$

where t_i represents the estimated mean for $i = 1,2,\dots,9999,10000$ and $\overline{Y}$ shows the population mean.

Tablo 6: Simulation Results for the Relative Efficiencies of Suggested estimators with respect to Singh et al (2008) estimators

	n	t_{pr1}	t_{pr2}	t_{pr3}	t_{pr4}	t_{pr5}	t_{pr6}	t_{pr7}	t_{pr8}	t_{pr9}	t_{pr10}
$RE(.,t_i)$	100	0.774	0.858	0.818	0.769	0.986	0.865	0.770	0.836	0.776	0.835
	150	0.791	0.839	0.814	0.789	0.930	0.843	0.789	0.825	0.792	0.824
	200	0.799	0.819	0.805	0.799	0.911	0.823	0.799	0.810	0.799	0.810
	250	0.781	0.792	0.785	0.780	0.825	0.793	0.780	0.788	0.781	0.788

In this simulation study, sample size is taken as $n = 100,150,200,250$. Efficiency values of suggested estimators are given relative to Singh et al. (2008) estimators for each n value in Table 6. These values are calculated with the help of equation (6.1). In Table 6, it is observed that suggested estimators are more efficient than Singh et al. (2008) estimators for all sample sizes.

Also, when there are outliers in data set, efficiencies of suggested estimators increases significantly compared to Singh et al. (2008) estimators. All of these results also support theoretical results in Table 5.

7. CONCLUSIONS

According to the theoretical discussion in Section 4 and the results of both the numerical examples and simulation, we infer that the suggested estimators are more efficient than the ratio estimators in Singh et al. (2008) when there are outliers in data. This article shows that LMS estimation can be used for the ratio estimators of the population mean in simple random sampling and that using LMS estimation improves the efficiency of Singh et al. (2008) estimators. In forthcoming studies, we hope to adapt the method presented here to estimators using two auxiliary attribute.

REFERENCES

1. *Kadılar, C.,Candan, M. and Çıngı, H. (2007). Ratio estimators using robust regression. Hacettepe Journal of Mathematics and Statistics, 36(2),181-188.*
2. *Nagler, T.,(2017). kdevine: Multivariate Kernel Density Estimation with Vine Copulas. R package version 0.4.1. https://CRAN.R-project.org/package=kdevine,*
3. *Rousseeuw, P. J. and Leroy, A. M. (1987). Robust regression and outlier detection. Wiley Series in Probability and Mathematical Statistics, New York: Wiley.*
4. *Singh, R., Chauhan, P., Sawan, N.,and Smarandache, F. (2008). Ratio estimators in simple random sampling using information on auxiliary attribute. Pak.j.stat.oper.res. 4(1), 47-53*
5. *Venables, W. N. and Ripley, B. D. (2002). Modern Applied Statistics with S. Fourth Edition. Springer, New York. ISBN 0-387-95457-0.*
6. *Zaman, T., Saglam, V., Sagir, M., Yucesoy, E., and Zobu, M. (2014). Investigation of some estimators via taylor series approach and an application. American Journal of Theoretical and Applied Statistics, 3(5), 141-147*
7. *Zaman, T. and Bulut, H. (2018). Modified ratio estimators using robust regression methods. Communications in Statistics - Theory and Methods, accepted., Doi. 10.1080/03610926.2018.1441419*