The Journal of Applied Science
วารสารวิทยาศาสตร์ประยุกต์
ISSN 1513-7805 Printed in Thailand

Vol. 19 No. 1: 51-58 [2020]
doi: 10.14416/j.appsci.2020.01.005
ISSN 2586-9663 (Online)

Research Article

New ratio estimators for population mean in simple random sampling using robust regression

Nuanpan Lawson [1]*

[1] Department of Applied Statistics, Faculty of Applied Science, King Mongkut's University of Technology North Bangkok, Bangkok, Thailand.

***E-mail**: nuanpan.n@sci.kmutnb.ac.th

Received: 29/03/2020; Revised: 27/04/2020; Accepted: 11/05/2020

Abstract

Using the traditional least square regression estimator we can validate the assumption of normality of the residual error when outliers occur in the data. In this paper, alternative ratio estimators for estimating population mean using robust regression are proposed for use in the case where data is contaminated with outliers, under simple random sampling. The bias and mean square error of the proposed estimators have been investigated. A simulation study has been conducted to compare the efficiency of the proposed estimators with traditional estimators. The results show that the proposed estimators perform better than the existing estimators.

Keywords: ratio estimator, simple random sampling, robust regression, Taylor series approximation, Huber M-estimate.

Introduction

Auxiliary information could be utilised for estimating population mean of the variable of interest if the auxiliary variable is related to the variable of interest. Ratio and regression estimators use the benefit of related auxiliary variables and the variable of interest in order to increase the efficiency of the population mean estimators. Several researchers developed ratio and regression estimators in their study using known parameters of auxiliary variables. For example, Sisodia & Dwivedi (1981) proposed to use the benefit of a correlation coefficient to develop a new ratio estimator for estimating population mean. Later Singh & Tailor (2003) also improve the ratio estimator by adding the correlation coefficient into the traditional one. (see also Subramani & Kumarapandiyan, 2012; Jaroengeratikun & Lawson, 2019; Lawson, 2019). Some researchers proposed to use the regression coefficient to develop new ratio estimators. For instance, Nangsue (2009) proposed new ratio and regression estimators in the case of missing data. One of Nangsue's estimators utilizing the regression coefficient is as follows.

$$\hat{\bar{Y}}_1 = \left[\bar{y} + b(\bar{X} - \bar{x})\right]\left(\frac{\bar{X}}{\bar{x}}\right)^b , \qquad (1)$$

where $\bar{y}$ and $\bar{x}$ are the sample means of the study variable X and auxiliary variable Y respectively. $\bar{X}$ is the population mean of auxiliary variableis the sample b and X regression coefficient.

Later, Soponviwatkul & Lawson (2017) proposed new ratio estimators following Sisodia & Dwivedi (1981), Singh & Tailor (2003) and Nangsue (2009). Soponviwatkul & Lawson (2017) estimators performs well when conditions are satisfied.

However, estimation of regression coefficients using an ordinary least square method is not suitable for the data that is contaminated with outliers. The outliers will affect the normal distribution assumption of the traditional least square regression method. Robust regression is an alternative method of estimation dealing with data with outliers in order to increase the efficiency of the estimators using an idea of weighting method by assigning different weights to observations. Some researches proposed to use robust regression in the estimation of the regression coefficients in order to increase the efficiency when data is contaminated by outliers. For example, Kadilar et al. (2007) proposed using robust regression in the estimation of population mean using ratio estimators in the case of where data is contaminated under simple random sampling. They considered the ratio estimators proposed by Sisodia & Dwivedi (1981), Singh & Kakran (1993), Upadhyaya & Singh (1999) and Kadilar & Cingi (2004) in their study. Kadilar et al. (2007) applied robust regression called Huber M- estimator to estimate the coefficient in ratio estimator. The Huber M- estimator assigns different weights to the observations dependent on the residuals, observations with larger residuals get smaller weights. Kadilar et al. (2007) estimators are as follows.

$$\hat{\bar{Y}}_{pr1} = \left[\bar{y} + b_{rob}(\bar{X} - \bar{x})\right]\left(\frac{\bar{X}}{\bar{x}}\right), \tag{2}$$

$$\hat{\bar{Y}}_{pr2} = \left[\bar{y} + b_{rob}(\bar{X} - \bar{x})\right]\left(\frac{\bar{X} + C_x}{\bar{x} + C_x}\right), \tag{3}$$

$$\hat{\bar{Y}}_{pr3} = \left[\bar{y} + b_{rob}(\bar{X} - \bar{x})\right]\left(\frac{\bar{X} + \beta_{2(x)}}{\bar{x} + \beta_{2(x)}}\right), \tag{4}$$

$$\hat{\bar{Y}}_{pr4} = \left[\bar{y} + b_{rob}(\bar{X} - \bar{x})\right]\left(\frac{\beta_{2(x)}\bar{X} + C_x}{\beta_{2(x)}\bar{x} + C_x}\right), \tag{5}$$

$$\hat{\bar{Y}}_{pr5} = \left[\bar{y} + b_{rob}(\bar{X} - \bar{x})\right]\left(\frac{C_x\bar{X} + \beta_{2(x)}}{C_x\bar{x} + \beta_{2(x)}}\right), \tag{6}$$

where C_x is the coefficient of variation of X, $\beta_{2(x)}$ is the coefficient of kurtosis of X and b_{rob} is the sample regression coefficient estimated by the Huber M-estimator. Huber (1981) proposed to minimize the summation of outlier function. The objective function is $\min\sum_{i=1}^{n}\rho(e_i)$

and Huber's function (ρ) is given by $\rho(e)=\begin{cases}\dfrac{e^2}{2} & ,|e|\le k\\ k|e|-\dfrac{k^2}{2},|e|>k\end{cases}$

The mean square error of the Kadilar et al. (2007) estimators are given by

$$MSE(\hat{\bar{Y}}_{pri}) = \frac{(1-f)}{n}\bar{Y}^2\left[W_i^2C_X^2 + C_Y^2 + b_{rob}^2K^2C_X^2 - 2W_i\rho C_X C_Y + 2W_i b_{rob}KC_X^2 - 2b_{rob}K\rho C_X C_Y\right], \tag{7}$$

where $W_1=\frac{\bar{X}}{\bar{x}}$, $W_1=\frac{\bar{X}}{\bar{x}+C_x}$, $W_3=\frac{\bar{X}}{\bar{x}+\beta_{2(x)}}$, $W_4=\frac{\beta_{2(x)}\bar{X}}{\beta_{2(x)}\bar{x}+C_x}$,

$W_5=\frac{C_x\bar{X}}{C_x\bar{x}+\beta_{2(x)}}$ and $K=\frac{\bar{X}}{\bar{Y}}$, $i=1,2,...,5$

Later Zaman & Bulut (2018) compared the efficiency of the estimators proposed by Kadilar et al. (2007) using the Huber M- estimator with other estimators using different methods of robust regression with each. Recently, Zaman (2019) proposed a combined estimator which combined the estimators proposed by Zaman & Bulut (2018) when data is contaminated with an outlier and it is found that the new estimators are more efficient than the old ones under certain conditions.

In this paper, the alternative ratio estimators for estimating population mean using robust regression in the case that data is contaminated with outliers under simple random sampling, following the work of Nangsue (2009) and Kadilar et al. (2007), have been proposed. The proposed population mean estimators are discussed in Section 2. A simulation study is shown in Section 3 to compare the efficiency of the proposed estimators with the traditional estimators. Finally conclusion and discussion are given in section 4.

Methods

Motivated by Nangsue (2009) and Kadilar et. al (2007), we proposed new ratio estimators for use in the case of the existence of contaminated data by using a robust regression method to estimate the sample regression coefficient under simple random sampling. The proposed estimators are as follows.

$$\hat{\bar{Y}}_{N1}=\left[\bar{y}+b_{rob}(\bar{X}-\bar{x})\right]\left(\frac{\bar{X}}{\bar{x}}\right)^{b_{rob}}, \tag{8}$$

$$\hat{\bar{Y}}_{N2}=\left[\bar{y}+b_{rob}(\bar{X}-\bar{x})\right]\left(\frac{\bar{X}+C_x}{\bar{x}+C_x}\right)^{b_{rob}}, \tag{9}$$

$$\hat{\bar{Y}}_{N3} = \left[\bar{y} + b_{rob}(\bar{X} - \bar{x})\right]\left(\frac{\bar{X} + \beta_{2(x)}}{\bar{x} + \beta_{2(x)}}\right)^{b_{rob}}, \tag{10}$$

$$\hat{\bar{Y}}_{N4} = \left[\bar{y} + b_{rob}(\bar{X} - \bar{x})\right]\left(\frac{\beta_{2(x)}\bar{X} + C_x}{\beta_{2(x)}\bar{x} + C_x}\right)^{b_{rob}}, \tag{11}$$

$$\hat{\bar{Y}}_{N5} = \left[\bar{y} + b_{rob}(\bar{X} - \bar{x})\right]\left(\frac{C_x\bar{X} + \beta_{2(x)}}{C_x\bar{x} + \beta_{2(x)}}\right)^{b_{rob}}, \tag{12}$$

From equations (8) to (12), we can write a general form for the proposed estimators as follows.

$$\hat{\bar{Y}}_{N} = \left[\bar{y} + b_{rob}(\bar{X} - \bar{x})\right]\left(\frac{A\bar{X} + D}{A\bar{x} + D}\right)^{b_{rob}}, \tag{13}$$

where C_x is the coefficient of variation of X, $\beta_{2(x)}$ is the coefficient of the kurtosis of X, A and D are parameters of an auxiliary variable X and b_{rob} is the sample regression coefficient estimated by the Huber M-estimator.

Note that: where there are no outliers (an outlier is usually a data point that is larger than the third quartile or smaller than the first quartile), the proposed family of estimators in equation (13) can also be used by replacing b_{rob} with b, a sample regression coefficient estimated by the ordinary least square method.

The bias and mean square error of the proposed estimators are investigated under simple random sampling without replacement using a Taylor series approximation.

Let $e_0 = \frac{\bar{y} - \bar{Y}}{\bar{Y}}$ then $\bar{y} = \bar{Y}(1+e_0)$ and let $e_1 = \frac{\bar{x} - \bar{X}}{\bar{X}}$ then $\bar{x} = \bar{X}(1+e_1)$, such that $E(e_0) = E(e_1) = 0$, $E(e_0^2) = \frac{1-f}{n}C_y^2$, $E(e_1^2) = \frac{1-f}{n}C_x^2$ and $E(e_0e_1) = \frac{1-f}{n}C_{xy} = \frac{1-f}{n}\rho C_y C_x$.

Expressing equation (13) in terms of e’s, we have

$$\begin{aligned}
\hat{\bar{Y}}_{N} &= \left[\bar{Y}(1+e_0) + b_{rob}(\bar{X} - \bar{X}(1+e_1))\right]\left(\frac{A\bar{X} + D}{A\bar{X}(1+e_1) + D}\right)^{b_{rob}}, \\
&= \left[\bar{Y}(1+e_0) + b_{rob}(\bar{X} - \bar{X}(1+e_1))\right](1+We_1)^{-b_{rob}}, \\
&= \bar{Y} - Wb_{rob}e_1\bar{Y} + \frac{b_{rob}(b_{rob}+1)}{2}W^2e_1^2\bar{Y} + \bar{Y}e_0 - Wb_{rob}\bar{Y}e_0e_1 + \frac{b_{rob}(b_{rob}+1)}{2}W^2e_1^2\bar{Y}e_0 \\
&\quad -b_{rob}\bar{X}e_1 + b_{rob}^2\bar{X}We_1^2.
\end{aligned}$$

Up to the first degree of approximation using a Taylor series , the approximate bias of the proposed estimator is given by

$$
\begin{aligned}
\mathrm{Bias}(\hat{\bar{Y}}_N) &\approx (\hat{\bar{Y}}_N - \bar{Y}) \\
&= E(-Wb_{rob}e_1\bar{Y} + \frac{b_{rob}(b_{rob}+1)}{2}W^2e_1^2\bar{Y} + e_0\bar{Y} - Wb_{rob}e_0e_1\bar{Y} + \frac{b_{rob}(b_{rob}+1)}{2}W^2e_0e_1^2\bar{Y} \\
&\quad - b_{rob}e_1\bar{X} + b_{rob}^2We_1^2\bar{X}) \\
&= \frac{(1-f)}{n}\bar{Y}\left[\left(\frac{b_{rob}(b_{rob}+1)}{2}W^2 + b_{rob}^2KW\right)C_x^2 - Wb_{rob}\rho C_xC_y\right],
\end{aligned} \tag{14}
$$

The approximate MSE of the proposed estimator $\hat{\bar{Y}}_N$ to the first order of approximation is given by

$$
\begin{aligned}
MSE(\hat{\bar{Y}}_N) &\approx E(\hat{\bar{Y}}_N - \bar{Y})^2 \\
&= E(W^2b_{rob}^2e_1^2\bar{Y}^2 + e_0^2\bar{Y}^2 + b_{rob}^2e_1^2\bar{X}^2 - 2Wb_{rob}e_0e_1\bar{Y}^2 + 2Wb_{rob}^2e_1^2\bar{X}\bar{Y} - 2b_{rob}e_0e_1\bar{X}\bar{Y}) \\
&= \frac{(1-f)}{n}\bar{Y}^2\left[W^2b_{rob}^2C_X^2 + C_Y^2 + b_{rob}^2K^2C_X^2 - 2Wb_{rob}\rho C_XC_Y + 2Wb_{rob}^2KC_X^2 - 2b_{rob}K\rho C_XC_Y\right] \\
&= \frac{(1-f)}{n}\bar{Y}^2\left[C_Y^2 + \left(W^2b_{rob}^2 + b_{rob}^2K^2 + 2Wb_{rob}^2K\right)C_X^2 - 2\left(Wb_{rob} + b_{rob}K\right)\rho C_XC_Y\right],
\end{aligned} \tag{15}
$$

where $W = \frac{A\bar{X}}{A\bar{x}+D}$ and $K = \frac{\bar{X}}{\bar{Y}}$.

To see the performance of the proposed estimators, the mean square error of the proposed estimators in equation (15) will be compared with the mean square error of the Kadilar et al. (2007) estimators in equation (7) . The proposed estimators are more efficient than the estimators of Kadilar et al. (2007) if the conditions below are satisfied. The details are as follows:

The proposed general estimator $\hat{\bar{Y}}_N$ is more efficient than the estimator $\hat{\bar{Y}}_{pri}$, if

$$MSE(\hat{\bar{Y}}_N) < MSE(\hat{\bar{Y}}_{pri})$$

$$
\begin{aligned}
&W^2b_{rob}^2C_X^2 + C_Y^2 + b_{rob}^2K^2C_X^2 - 2Wb_{rob}\rho C_XC_Y + 2Wb_{rob}^2KC_X^2 - 2b_{rob}K\rho C_XC_Y < \\
&W^2C_X^2 + C_Y^2 + b_{rob}^2K^2C_X^2 - 2W\rho C_XC_Y + 2W b_{rob}KC_X^2 - 2b_{rob}K\rho C_XC_Y
\end{aligned}
$$

$$
\begin{aligned}
&W^2C_X^2\left(b_{rob}^2 - 1\right) - 2W\rho C_XC_Y\left(b_{rob} - 1\right) + 2WKC_X^2\left(b_{rob}^2 - b_{rob}\right) < 0 \\
&\left(b_{rob} - 1\right)\left[W^2C_X^2\left(b_{rob} + 1\right) - 2W\rho C_XC_Y + 2WKC_X^2b_{rob}\right] < 0
\end{aligned}
$$

For $b_{rob} - 1 > 0$ that is $b_{rob} > 1$ then

$$W^2C_X^2\left(b_{rob} + 1\right) - 2W\rho C_XC_Y + 2WKC_X^2b_{rob} < 0$$

$$b_{rob} < \frac{2W\rho C_X C_Y - W^2 C_X^2}{2WKC_X^2 + W^2 C_X^2}$$

Similarly, for $b_{rob} - 1 < 0$ that is $b_{rob} < 1$then

$$b_{rob} > \frac{2W\rho C_X C_Y - W^2 C_X^2}{2WKC_X^2 + W^2 C_X^2},$$

where $W = \dfrac{A\bar{X}}{A\bar{x} + D}$ and $K = \dfrac{\bar{X}}{\bar{Y}}$.

Results and discussion

The proposed estimators are used in simulation data in order to see the performance of the proposed estimators against the estimators of Kadilar et al. (2007). A population of (X,Y) is generated by the bivariate normal distribution with mean equal to 500 and 40 respectively. The coefficient of variation of X and Y are equal to 0.13 and 0.25 respectively. The correlation coefficient between X and Y is equal to 0.9. A sample size $n = 5,\ 20,\ 30,\ 50$ is selected using simple random sampling without replacement from the population size $N = 100$. There is an outlier in the sample set of data. The simulation is repeated 10,000 times using the R program. The bias and mean square error for all estimators is calculated. The results are presented in Table 1 below.

Table 1 The bias and mean square error of the proposed estimators and existing estimators.

Estimators	$n=5$		$n=20$		$n=30$		$n=50$	
	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE
$\hat{\bar{Y}}_{pr1}$	6.767	46.366	3.237	11.581	1.540	2.892	0.628	0.593
$\hat{\bar{Y}}_{pr2}$	7.053	50.117	2.980	9.991	1.398	2.494	0.568	0.530
$\hat{\bar{Y}}_{pr3}$	6.822	47.063	3.186	11.254	1.511	2.809	0.616	0.580
$\hat{\bar{Y}}_{pr4}$	6.777	46.488	3.228	11.523	1.535	2.877	0.626	0.590
$\hat{\bar{Y}}_{pr5}$	6.767	46.367	3.237	11.580	1.539	2.892	0.629	0.593
$\hat{\bar{Y}}_{N1}$	0.970	8.492	0.307	1.413	0.226	0.783	0.137	0.324
$\hat{\bar{Y}}_{N2}$	**0.405**	**7.952**	**0.167**	**1.393**	**0.141**	**0.779**	**0.099**	0.326
$\hat{\bar{Y}}_{N3}$	0.862	8.338	0.279	1.405	0.209	0.781	0.129	0.324
$\hat{\bar{Y}}_{N4}$	0.951	8.463	0.302	1.411	0.223	0.783	0.136	0.324
$\hat{\bar{Y}}_{N5}$	0.969	8.492	0.307	1.413	0.226	0.783	0.137	0.324

From Table 1 we can see that all proposed estimators perform a lot better than the estimators of Kadilar et al. (2007) in both bias and mean square error. The mean square errors of all proposed estimators are similar. The bias and mean square error are smaller when sample size increases. The $\hat{\hat{Y}}_{N2}$ seems to perform the best in term of bias and mean square error in this scenario. The bias and mean square errors for the proposed estimators are similar when sample size increased.

Conclusion

Outliers can affect the interpretation of results if they are not deal with before analysis. In this paper new ratio estimators have been proposed using the benefit of robust regression to solve a problem of contaminated data. We also presented the proposed estimators in a general form that can be useful with some known auxiliary variables. We have shown in theory that the mean square error of the proposed estimators are smaller than the existing estimators in certain conditions. A simulation study was also used to support the theory that the alternative ratio estimators for estimating population mean and the proposed estimators performed well when compare to the existing estimators. The Huber M-estimator has been used to estimate the regression coefficient for the robust regression method. The other robust regression methods, e.g., the Hampel (1971), Turkey (1977) methods can also be applied in the proposed estimators when data is contaminated with outliers. The alternative estimators may be beneficial in estimating population mean data with outliers.

Acknowledgement

This work was funded by The Faculty of Applied Science, King Mongkut's University of Technology North Bangkok, Thailand. I would like to thank the unknown referees for offering useful suggestions for improving this paper.

References

Hampel, F. R. (1971). A general qualitative definition of robustness. *The Annals of Mathematical Statistics,* 42 (6):1887–96. doi:10.1214/aoms/1177693054

Huber, P. (1981). Robust Statistics, Wiley, New York.

Jaroengeratikun, U. & Lawson, N. (2019). A combined family of ratio estimators for population mean using an auxiliary variable in simple random sampling, *Journal of Mathematical and Fundamental Sciences,* 51(1), 1-12. DOI: 10.5614/j.math.fund.sci.2019.51.1.1

Kadilar, C., & H. Çıngı. (2004). Ratio estimators in simple random sampling. *Applied Mathematics and Computation,* 151 (3), 893–902. doi:10.1016/S0096-3003(03)00803-8

Kadılar, C., M. Candan & H. Çıngı. (2007). Ratio estimators using robust regression. *Hacettepe Journal of Mathematics and Statistics* 36 (2):181–88.

Lawson, N. (2019). Ratio estimators of population means using quartile function of auxiliary variable using double sampling, *Songklanakarin Journal of Science and Technology*, 41(1), 117-122.

Nangsue, N. (2009). Adjusted ratio and regression type estimators for estimation of population mean when some observations are missing. *International Scholarly and Scientific Research & Innovation*, 3(5), 334-337.

Sisodia, B. V. & Dwivedi, V. K. (1981). A modified ratio estimator using coefficient of variation of auxiliary variable. Journal Indian Society of Agricultural Statistics, 33(2), 13-18.

Singh, H. P., & Kakran, M. S. (1993). A modified ratio estimator using known coefficient of kurtosis of an auxiliary character. (Unpublished manuscript).

Singh, H.P. & Tailor, R. (2003). Use of known correlation coefficient in estimating the finite population mean, *Statistics in Transition*, 6, pp. 555-560.

Soponviwatkul, K. & Lawson, N. (2017). New ratio estimators for estimating population mean in simple random sampling using a coefficient of variation, correlation coefficient and a regression coefficient. Gazi University Journal of Science, 30(4), 610-621.

Subramani, J. & Kumarapandiyan, G. (2012). Modified ratio estimators for population mean using function of quartiles of auxiliary variable. *Bonfring International Journal of Industrial Engineering and Management Science*, *2* (2): 19-23.

Tukey, J. W. (1977). *Exploratory data analysis*. MA: Addison-Wesley.

Upadhyaya, L. N. & Singh, H. P. (1999). Use of transformed auxiliary variable in estimating the finite population mean. *Biometrical Journal* 41 (5):627–36. doi:10.1002/(SICI)1521-4036(199909)41:5%3c627::AID-BIMJ627%3e3.0.CO;2-W

Zaman, T. (2019). Improvement of modified ratio estimators using robust regression methods. *Applied Mathematics and Computation*, Elsevier, 348(C), 627-631. https://doi.org/10.1016/j.amc.2018.12.037

Zaman, T. & Bulut, H. (2018). Modified ratio estimtors using robust regression methods, *Communications in Statistics -Theory and Methods,* 0(0), 1-10. https://doi.org/10.1080/03610926.2018.1441419