Scientia Iranica (2024) 31(15), 1330–1341

Sharif University of Technology
Scientia Iranica
Transactions E: Industrial Engineering
https://scientiairanica.sharif.edu

Research Note

A simulation study: Robust ratio double sampling estimator of finite population mean in the presence of outliers

T. Zaman[a,*] and H. Bulut[b]

a. *Faculty of Science, Department of Statistics, Çankırı Karatekin University,18100 Çankırı, Turkey.*
b. *Faculty of Science, Department of Statistics, Ondokuz Mayıs University, 55139 Samsun, Turkey.*

Received 17 April 2020; received in revised form 25 July 2021; accepted 4 October 2021

KEYWORDS
Ratio estimators;
Robust regression estimators;
Mean Square Error (MSE);
Efficiency;
Double sampling.

Abstract. In this study, we suggest a family of ratio estimators for the population mean parameter using various robust regression techniques. These robust regressions techniques are Huber MM, Last Trimmed Square (LTS), and Least Median Square (LMS) estimates. We evaluate the performance of estimators in terms of the Mean Square Error (MSE), and we compare the efficiency of our proposed robust-regression-ratio-type estimators with existing estimators under the optimal conditions. These comparisons show that our robust ratio-type estimators give more efficient results than the existing estimators under double sampling. In addition, the simulation and the empirical studies based on a data set that includes unusual observations show that our proposed estimators have a lower MSE than the existing estimators.

1. Introduction

In the random sampling setting, the auxiliary information is commonly used to improve estimates. The classical ratio estimator is the most common estimator of the population mean when the correlation between study and auxiliary variables is highly positive. The ratio and the regression estimators of the mean of the study variable are good examples. However, when there are extreme values in the data, the efficiency of classical estimators declines. Therefore, Kadilar et al. [1] suggested Huber-M estimator for ratio estimators and reduced the effect of the extreme values. Motivated by Kadilar et al. [1], Oral and Kadilar [2,3] introduced maximum likelihood estimators and incorporated modified maximum likelihood estimators into Kadilar and Cingi [4] estimators. Abid et al. [5] introduced different ratio estimators with the help of some robust measures. Then, Abid et al. [6] developed some new ratio estimators of variance based on robust measures. Zaman and Bulut [7] proposed robust ratio estimators based on the estimators given in Kadilar et al. [1]. Zaman [8] suggested combining estimators for the population mean using the estimators presented in Zaman and Bulut [7]. Subzar et al. [9] presented

*. *Corresponding author.*
E-mail addresses: zamantolga@gmail.com (T. Zaman); hasan.bulut@omu.edu.tr (H. Bulut)

To cite this article:
T. Zaman and H. Bulut“A simulation study: Robust ratio double sampling estimator of finite population mean in the presence of outliers”, *Scientia Iranica*, (2024) **31**(15), pp. 1330-1341
https://doi.org/10.24200/sci.2021.55813.4418

the robust regression ratio type estimators to estimate the mean of the study variable in outlier data. Zaman and Bulut [10] suggested robust regression-type estimators in stratified random sampling. Bulut and Zaman [11] extended Zaman and Bulut [7] for Minimum Covariance Determinant (MCD) estimates. Using Zaman and Bulut [7], Shahzad et al. [12] provided various estimators using robust regression and variance-covariance techniques. Naz et al. [13] presented ratio-type estimators for population variance using the information on the auxiliary variable's robust nonconventional location parameters. Subzar et al. [14] provided new ratio estimators of population mean utilizing some robust measures. Grover and Kaur [15] proposed robust ratio estimators to predict the mean in simple and stratified random sampling. Ali et al. [16] developed a class of robust-regression type estimators in the case of sensitive research. The ratio and the regression estimators are used if the population mean of the auxiliary variable is known, but this is not always the case. In double sampling, a good estimator of the population mean of the auxiliary variable requires the first-phase sample, and a second-phase sample is necessary to measure the study variable [17]. Neyman [18] was the first to introduce the concept of double sampling. Sukhatme [19] taught a class of estimators in double sampling. Following Kadilar et al. [1], Noor-ul-Amin et al. [20] applied the concept of Sukhatme [19] and provided the estimator of the mean using the Huber-M measure for double sampling. Singh et al. [21] presented various imputation methods to compensate for the missing data in estimating the population mean parameter for two-phase sampling. Guha and Chandra [22] proposed an improved chain-ratio estimator for the population total based on double sampling. Guha and Chandra [23] provided improved estimators for the population mean using two auxiliary variables comprise non-response in on two-phase sampling.

Let that the finite population consists of N distinct and identifiable units under study. A random sample of size n is drawn using Simple Random Sampling Without Replacement (SRSWOR). Let be the population mean of the study variable and the auxiliary variable $\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i$ and $\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$, respectively. The sample means for variables Y and X are indicated by $\overline{y}$ and $\overline{x}$, respectively.

If the population mean of the auxiliary variable is not known, double sampling is used to estimator the population mean of the study variable. Under the double sampling, the first phase sample of a fixed size $n_1\,(n_1 < N)$ is drawn to measure only x to formulate a good estimator of a population mean $\bar{X}$, the second phase sample of a fixed size $n_2\,(n_2 < n_1)$ is drawn to measure y.

To obtain the Mean Square Error (MSE) of the estimators, let $\overline{y} = \bar{Y}\left(1+\overline{e}_{y_2}\right)$, $x_1 = \bar{X}\left(1+\overline{e}_{x_1}\right)$ and $x_2 = \bar{X}\left(1+\overline{e}_{x_2}\right)$ such that:

$$E\left(\overline{e}_{y_2}\right) = E\left(\overline{e}_{x_1}\right) = E\left(\overline{e}_{x_2}\right) = 0,$$

$$E\left(\overline{e}_{y_2}^2\right) = \theta_2 C_y^2, E\left(\overline{e}_{x_1}^2\right) = \theta_1 C_x^2, E\left(\overline{e}_{x_2}^2\right) = \theta_2 C_x^2,$$

$$E\left(\overline{e}_{y_2}\overline{e}_{x_1}\right) = \theta_1\rho_{yx}C_yC_x, E\left(\overline{e}_{y_2}\overline{e}_{x_2}\right) = \theta_2\rho_{yx}C_yC_x, \quad (1)$$

where $C_y = S_y/\bar{Y}$, $C_x = S_x/\bar{X}$, and $\rho_{yx} = S_{yx}/(S_yS_x)$. Here,

$$S_y^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(y_i - \bar{Y}\right)^2,$$

$$S_x^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(x_i - \bar{X}\right)^2,$$

and:

$$S_{yx} = \frac{1}{N-1}\sum_{i=1}^{N}\left(y_i - \bar{Y}\right)\left(x_i - \bar{X}\right),$$

$$\theta_1 = \left(N - n_1\right)/Nn_1, \; \theta_2 = \left(N - n_2\right)/Nn_2.$$

Noor-ul-Amin et al. [20] obtained the slope coefficient of Kadilar and Cingi [4] estimators using the Huber-M estimator. Noor-ul-Amin et al. [20] adapted the Kadilar and Cingi [4] estimators to the double sampling design as follows:

$$\overline{y}_{1j} = \frac{\overline{y}_2 + b_j\left(\overline{x}_1 - \overline{x}_2\right)}{\overline{x}_2}\overline{x}_1, \quad (2)$$

$$\overline{y}_{2j} = \frac{\overline{y}_2 + b_j\left(\overline{x}_1 - \overline{x}_2\right)}{\overline{x}_2 + C_x}\left(\overline{x}_1 + C_x\right), \quad (3)$$

$$\overline{y}_{3j} = \frac{\overline{y}_2 + b_j\left(\overline{x}_1 - \overline{x}_2\right)}{\overline{x}_2 + \beta_2(x)}\left(\overline{x}_1 + \beta_2(x)\right), \quad (4)$$

where $j = 1$ represents Huber-M estimate. When there is an outlier in the dataset, they provided that b_j computed by Huber-M must be used instead of b computed by Ordinary Least Squares (OLS). The MSE expression of Noor-ul-Amin et al. [20] estimators obtain as below [20]:

$$\begin{aligned} MSE\left(\overline{y}_{i1}\right) = &\bar{Y}^2\Big[\theta_2 C_y^2 + k_i^2 + C_x^2 \\ &\left\{\theta_1(1-k_i)^2 - 2\theta_1\left(1-k_i^2\right) + \theta_2(1+k_i)^2\right\} \\ &- 2C_x^2 H_{yx}\left\{\theta_2 - \theta_1\left(1-k_i\right)\right\}\Big] \\ &+ C_x^2 B_1\bar{X}\bar{Y}\left(\theta_2 - \theta_1\right) \\ &\left(B_1\bar{X} - 2H_{yx} + 2\left(1+k_i\right)\right), \\ &\quad i = 1, 2, 3, \end{aligned} \quad (5)$$

where $k_1 = 0$, $k_2 = \frac{C_x}{\bar{X}}$, $k_3 = \frac{\beta_2(x)}{\bar{X}}$, and $H_{yx} = \rho_{yx}\frac{C_y}{C_x}$. B_1, is coefficients of slope obtained from Huber-M.

We improve the Noor-ul-Amin et al. [20] estimators by using Huber MM, Least Trimmed Squares (LTS), or Least Median Squares (LMS). We express MSE up to the first-order approximation. We compare the efficiencies of the estimators with that of the Noor ul Amin et al. [20] estimator and find a significantly lower MSE for double sampling. These robust regression methods are described below very briefly.

2. Robust regression methods

In linear regression, the OLS estimators are optimal when all of the regression assumptions are valid. However, it is well known that the OLS estimators are quite sensitive to outliers like other classic statistical methods. In the literature, many robust regression methods have been suggested to overcome this problem.

The objective function of OLS is to minimize the sum of squared residuals. Similarly, the LMS method aims to minimize the median of squared residuals [24]. In the LTS, the squared residuals are sorted, and the OLS method is performed on observations regarding the first (smallest) r residuals [25]. Generally, the M regression methods aim to minimize the ρ functions that are satisfied with some assumptions [26]. Accordingly, in literature, the M estimate is suggested by changing the ρ function by Huber [27]. This estimator is called Huber-M estimator. Finally, Yohai [28] proposed the MM regression method, which has high efficiency and breakdown point. Researchers can view more detailed information about robust regression estimates in Zaman and Bulut [7].

In this study, we use the R programming language for all calculations. According to this, we calculate Huber-M estimations by using the “rlm” function at the “MASS” package in R [29]. For Huber MM estimations, we use the “lmRob” function at the “robust” package in R [30]. Finally, we use the “lqs” function at the “MASS” package in R [29] for LTS and LMS estimations. We use the method=“lts” argument to obtain the LTS estimations, while LMS estimations are obtained using the method=“lms” argument in the function.

3. Suggested estimators

In this section, we propose a variety of ratio estimators considering some robust estimators instead of coefficients of slope in ratio estimators presented between Eqs. (2)-(4). We develop the following estimators:

$$\overline{y}_{1j} = \frac{\overline{y}_2 + b_j\left(\overline{x}_1 - \overline{x}_2\right)}{\overline{x}_2}\overline{x}_1, \tag{6}$$

$$\overline{y}_{2j} = \frac{\overline{y}_2 + b_j\left(\overline{x}_1 - \overline{x}_2\right)}{\overline{x}_2 + C_x}\left(\overline{x}_1 + C_x\right), \tag{7}$$

$$\overline{y}_{3j} = \frac{\overline{y}_2 + b_j\left(\overline{x}_1 - \overline{x}_2\right)}{\overline{x}_2 + \beta_2\left(x\right)}\left(\overline{x}_1 + \beta_2\left(x\right)\right), \tag{8}$$

where $\overline{y}_{ij}$; $i = 1, 2, 3$ and $j = 2, 3, 4$, where $j = 2$ represents Huber MM, $j = 3$ represents LTS and $j = 4$ represents LMS. b_j are the coefficients of slope computed by Huber MM, LTS, and LMS estimates, respectively.

The expressions of MSE for modified ratio estimators considering robust measures can be stated as Eq. (5). The main difference between the expressions of MSE is the usage of $B_j\,(j = 2, 3, 4)$ instead of B_1. The expressions of MSE for our suggested estimators belonging to robust regression estimates of interest are computed as follows.

To compute the MSE of the suggested estimators in Eqs. (6)-(8), we apply the notations (1) in Eqs. (6)-(8) as following the Noor-ul-Amin et al. [20] estimators, expressing the estimators, $\overline{y}_{ij}$, in terms of $\overline{e}_{y_2}\,\overline{e}_{x_i}$ $(i = 1, 2)$, we can write Eqs. (6)-(8) as:

$$\overline{y}_{ij} = \left[\bar{Y} + \bar{Y}\overline{e}_{y_2} + b_j\left(\bar{X}\overline{e}_{x_1} + \bar{X}\overline{e}_{x_2}\right)\right]$$
$$\left[\frac{1 + \overline{e}_{x_1} + k_i}{1 + \overline{e}_{x_2} + k_i}\right].$$

To the first degree of approximation for the Taylor series, we ignore the terms with power two or greater, and this expression is re-written as follows:

$$\overline{y}_{1j} - \bar{Y} \cong \bar{Y}\left[\left(\overline{e}_{y_2} - k_i\right) + \left(1 - k_i\right)\overline{e}_{x_1} - \left(1 + k_i\right)\overline{e}_{x2}\right]$$
$$+ b_j\bar{X}\left(\overline{e}_{x_1} + \overline{e}_{x_2}\right).$$

Taking square on both sides of this equation and applying expectations, the MSE equations of the estimators in Eqs. (6)-(8) is given by:

$$\begin{aligned} MSE\left(\overline{y}_{ij}\right) = & \bar{Y}^2\left[\theta_2 C_y^2 + k_i^2 + C_x^2\right. \\ & \left\{\theta_1(1-k_i)^2 - 2\theta_1\left(1-k_i^2\right) + \theta_2(1+k_i)^2\right\} \\ & \left. - 2C_x^2 H_{yx}\left\{\theta_2 - \theta_1\left(1 - k_i\right)\right\}\right] \\ & + C_x^2 B_j\bar{X}\bar{Y}\left(\theta_2 - \theta_1\right)\left(B_j\bar{X} - 2H_{yx}\right. \\ & \left. + 2\left(1 + k_i\right)\right), \\ & i = 1, 2, 3 \text{ and } j = 2, 3, 4, \end{aligned} \tag{9}$$

where, B_j are the coefficients of slope computed from Huber MM, LTS, and LMS estimators, respectively. The expressions of MSE of 4 different robust measures for each value i will be obtained.

4. Efficiency comparisons

In this section, we compare the MSE of the Noor-ul-Amin et al. [20] estimators, given in Eqs. (2)-(4), with the MSE of the suggested robust estimators, shown in Eqs. (6)-(8).

$$MSE\left(\overline{y}_{ij}\right) < MSE\left(\overline{y}_{i1}\right),$$

$$i=1,2,3 \quad \text{and} \quad j=2,3,4,$$

$$B_j\left(B_j\bar{X} - 2H_{yx} + 2\left(1+k_i\right)\right)$$

$$< B_1\left(B_1\bar{X} - 2H_{yx} + 2\left(1+k_i\right)\right),$$

$$\bar{X}\left(B_j - B_1\right)\left(B_j + B_1\right) - \left(B_j - B_1\right)$$

$$\left(2H_{yx} + 2\left(1+k_i\right)\right) < 0,$$

$$\left(B_j - B_1\right)\left[\bar{X}\left(B_j + B_1\right) - 2H_{yx} + 2\left(1+k_i\right)\right] < 0,$$

$$\text{For}\left(B_j - B_1\right) > 0; \text{that is } B_j > B_1 \text{ and}$$

$$\left(B_j + B_1\right) < \frac{2H_{yx} + 2\left(1+k_i\right)}{\bar{X}}. \tag{10}$$

Similarly, for $(B_j - B_1) < 0$, that is $B_j < B_1$ and:

$$\left(B_j + B_1\right) > \frac{2H_{yx} - 2\left(1+k_i\right)}{\bar{X}}. \tag{11}$$

When the condition (10) or (11) is satisfied, the MSE of the suggested robust ratio estimators is smaller than the Noor-ul-Amin et al. [20] estimators.

If B_1 is replaced with B above,

$$\left(B_j - B\right)\left[\bar{X}\left(B_j + B\right) - 2H_{yx} + 2\left(1+k_i\right)\right] < 0.$$

$$\text{For} \quad B_j - B > 0; \text{ that is } \quad B_j > B \text{ and}$$

$$B_j + B < \frac{2H_{yx} - 2\left(1+k_i\right)}{\bar{X}}. \tag{12}$$

Similarly, for $B_j - B < 0$; that is $B_j < B$ and:

$$B_j + B > \frac{2H_{yx} - 2\left(1+k_i\right)}{\bar{X}}. \tag{13}$$

The MSE of the suggested robust estimators is smaller than the usual ratio estimators for conditions (12) or (13).

5. Numerical example

In this section, we compare the performance of the suggested robust estimators with the estimators proposed by Noor-ul-Amin et al. [20] in the double sampling design using a real dataset. The population data is taken from Zaman and Bulut [7] and Zaman et al. [31].

Table 1. The statistics of data.

$N = 111$	$\beta_2(x) = 45.10873$	$B_{HubMM} = 0.0606$
$C_x = 1.538435$	$k_1 = 0$	$B_{LTS} = 0.0573$
$\rho_{yx} = 0.9487736$	$k_2 = 0.003427$	$B_{LMS} = 0.0562$
$\bar{Y} = 36.34234$	$k_3 = 0.100495$	$B_{HubM} = 0.06634$
$C_y = 2.131294$	$H_{yx} = 1.61437$	

This data consists of the number of teachers and students in each high school in 18 districts of Trabzon, a city in Turkey, for the 2011-2012 academic year. The statistics of the population are given in Table 1.

Following the Noor-ul-Amin et al. [20] estimators, to examine the sensitivity of sample sizes on suggested robust estimators in double sampling, we assume three different sample sizes at the first phase, $n_1 = 30, 40$, and 50. Then, from the first phase sample for each choice of n_1, we consider three different sample sizes, $n_2 = 10, 15$, and 20. To compare the proposed estimators with the Noor-ul-Amin et al. [20] estimators, we use the same sample sizes with Noor-ul-Amin et al. [20] study.

We obtained the MSE values of the suggested robust estimators and the Noor-ul-Amin et al. [20] estimators using the information in Table 1. The performance for each proposed estimator concerning the Noor-ul-Amin et al. [20] estimators are obtained as follows based on Eq. (14). The obtained MSE and RE values are presented in Tables 2 and 3, respectively.

$$RE\left(\overline{y}_{ij}\right) = \frac{MSE\left(\overline{y}_{ij}\right)}{MSE\left(\overline{y}_{i1}\right)},$$

$$i=1,2,3 \qquad \text{and} \qquad j=2,3,4, \tag{14}$$

where $MSE\left(\overline{y}_{ij}\right)$ is the mean square error for each estimator in Section 3 and $MSE\left(\overline{y}_{i1}\right)$ is the mean square error for each estimator presented in Noor-ul-Amin et al. [20].

The MSE of the Noor-ul-Amin et al. [20] and suggested ratio estimators are given in Table 2. The proposed robust estimators perform better than the Noor-ul-Amin et al. [20] estimators in terms of MSE. So the suggested estimators are more efficient.

The Relative Efficient (RE) values given in Table 3 are obtained using Eq. (14). If the relative efficiency is smaller than 1, the suggested robust estimators have a smaller MSE than the Noor-ul-Amin et al. [20] estimators. From Table 3, it is seen that the proposed robust estimators perform better than the in Noor-ul-Amin et al. [20] estimators. This situation is expected because the conditions presented in Eq. (11) are satisfied with the suggested robust-regression-ratio-type estimators. These results are apparent in Table 4.

Table 2. MSE values for real data application.

			Noor-ul-Amin et al. [20]	Proposed estimators based on		
n_1	n_2	Estimators	Huber M	Huber MM	LTS	LMS
30	10	$\overline{y}_{1j}$	5189.48	4347.98	3912.21	3782.31
		$\overline{y}_{2j}$	5191.42	4349.81	3913.99	3784.07
		$\overline{y}_{3j}$	5264.13	4419.63	3982.19	3851.77
	20	$\overline{y}_{1j}$	1406.82	1196.44	1087.50	1052.71
		$\overline{y}_{2j}$	1406.80	1196.40	1087.45	1052.65
		$\overline{y}_{3j}$	1422.73	1211.60	1102.24	1067.31
40	10	$\overline{y}_{1j}$	5769.93	4823.24	4328.66	4186.86
		$\overline{y}_{2j}$	5772.42	4825.61	4330.97	4189.16
		$\overline{y}_{3j}$	5860.20	4910.13	4413.66	4271.29
	20	$\overline{y}_{1j}$	1987.27	1671.70	1506.85	1457.61
		$\overline{y}_{2j}$	1987.81	1672.21	1507.33	1458.09
		$\overline{y}_{3j}$	2018.79	1702.10	1536.61	1487.18
	30	$\overline{y}_{1j}$	726.38	621.19	566.72	549.51
		$\overline{y}_{2j}$	726.27	621.07	566.60	549.38
		$\overline{y}_{3j}$	738.32	632.76	578.08	560.80
50	10	$\overline{y}_{1j}$	6118.20	5108.39	4591.38	4420.19
		$\overline{y}_{2j}$	6121.03	5111.10	4594.02	4422.81
		$\overline{y}_{3j}$	6217.84	5204.44	4685.44	4513.57
	20	$\overline{y}_{1j}$	2335.54	1956.86	1759.03	1699.95
		$\overline{y}_{2j}$	2336.42	1957.69	1759.84	1700.75
		$\overline{y}_{3j}$	2376.44	1996.41	1797.82	1738.50
	30	$\overline{y}_{1j}$	1074.65	906.35	816.09	791.65
		$\overline{y}_{2j}$	1074.88	906.56	816.29	791.84
		$\overline{y}_{3j}$	1095.97	927.06	836.46	811.92
	40	$\overline{y}_{1j}$	444.21	381.09	347.94	338.08
		$\overline{y}_{2j}$	444.11	380.99	347.83	337.97
		$\overline{y}_{3j}$	455.73	392.39	359.11	349.21

Table 3. Theoretical results for relative efficiencies of each proposed estimator according to Noor-ul-Amin et al. [20] estimators.

			Noor-ul-Amin et al. [20]	Proposed estimators based on		
n_1	n_2	Estimators	Huber M	Huber MM	LTS	LMS
30	10	$\overline{y}_{1j}$	1	0.838	0.754	0.729
		$\overline{y}_{2j}$	1	0.838	0.754	0.729
		$\overline{y}_{3j}$	1	0.840	0.756	0.732
	20	$\overline{y}_{1j}$	1	0.850	0.773	0.748
		$\overline{y}_{2j}$	1	0.850	0.773	0.748
		$\overline{y}_{3j}$	1	0.852	0.775	0.750
40	10	$\overline{y}_{1j}$	1	0.836	0.750	0.726
		$\overline{y}_{2j}$	1	0.836	0.750	0.726
		$\overline{y}_{3j}$	1	0.838	0.753	0.729
	20	$\overline{y}_{1j}$	1	0.841	0.758	0.733
		$\overline{y}_{2j}$	1	0.841	0.758	0.734
		$\overline{y}_{3j}$	1	0.843	0.761	0.737
	30	$\overline{y}_{1j}$	1	0.855	0.780	0.756
		$\overline{y}_{2j}$	1	0.855	0.780	0.756
		$\overline{y}_{3j}$	1	0.857	0.783	0.760
50	10	$\overline{y}_{1j}$	1	0.835	0.750	0.722
		$\overline{y}_{2j}$	1	0.835	0.751	0.723
		$\overline{y}_{3j}$	1	0.837	0.754	0.726
	20	$\overline{y}_{1j}$	1	0.838	0.753	0.728
		$\overline{y}_{2j}$	1	0.838	0.753	0.728
		$\overline{y}_{3j}$	1	0.840	0.757	0.732
	30	$\overline{y}_{1j}$	1	0.843	0.759	0.737
		$\overline{y}_{2j}$	1	0.843	0.759	0.737
		$\overline{y}_{3j}$	1	0.846	0.763	0.741
	40	$\overline{y}_{1j}$	1	0.858	0.783	0.761
		$\overline{y}_{2j}$	1	0.858	0.783	0.761
		$\overline{y}_{3j}$	1	0.861	0.788	0.766

Table 4. The results of condition in Eq. (10).

Method	β_j	$\beta_j+\beta_1$	Results for $\overline{y}_{1j}$	Results for $\overline{y}_{2j}$	Results for $\overline{y}_{3j}$
Huber-MM	0.0606	0.1269	True	True	True
LTS	0.0573	0.1236	True	True	True
LMS	0.0562	0.1225	True	True	True
Huber-M (β_1) :	0.06634	Condition limits:	0.0031	0.0031	0.0026

In Table 4, the methods with the highest beta value are Huber MM, LTS, and LMS, respectively. When the proposed estimators are examined according to these values, it is seen that the estimator with the smallest beta value is the most effective. Therefore, the results in Table 4 support Tables 2 and 3. In short, the real dataset results show that the robust-regression-ratio-type estimators are expected to be better than the existing estimators because there are unusual observations in the data. We see that these results are expected if we look at them more carefully because conditions (11) and (13) are satisfied with the suggested robust estimators. Also, the suggested robust regression-ratio-type estimator based on the LMS estimate has the best result among proposed robust ratio estimators.

6. Simulation study

A simulation study is carried out to calculate the MSE values by using proposed estimators and Noor-ul-Amin et al. [20] estimators. The datasets have been generated as follows:

$$Y_i = 2 + 3X_1 + \varepsilon_i, \tag{15}$$

where $X_1 \sim N(0,1)$ and $\varepsilon_i \sim N(0,1)$ for usual observations, $X_1 \sim N(25,1)$ and $\varepsilon_i \sim N(25,1)$ for unusual observations. We have guaranteed that there is an outlier in the dataset. For the simulation design;

We choose 10000 samples of the size sizes at the first phase $n_1 = 30$, 40, and 50 and from the first phase sample, for each choice sample size n_1, we chose different sample sizes in the second phase, $n_2 = 10, 20$, and 30.

Using the Eqs. (2)-(4) and (6)-(8), the value of $\widetilde{Y}_i$ in Eq. (16) is calculated 10000 times.

For each sample, we derived the expression of MSE of the existing and the suggested estimators are obtained by Eq. (16):

$$MSE\left(\widetilde{Y}_i\right) = \frac{1}{10000}\sum_{i=1}^{10000}\left(\widetilde{Y}_i - \bar{Y}\right)^2, \tag{16}$$

where $\bar{Y}$ shows the population mean parameter.

We give our R codes a better understanding of the simulation study in the supplementary file.

We assumed that the ratios of extreme values are 10%, 20%, and 30% and under the condition $n_2 < n_1$, sample sizes in the first phase, $n_1 = 30, 40$, and 50, then, for each choice of n_1, it is considered as sample sizes in the second phase, $n_2 = 10, 20$, and 30 in this study. In Tables 5, 6, and 7, our suggested robust estimators' MSE values and relative efficiency for each first phase and second phase sample sizes are given for outliers 10%, 20%, and 30%, respectively. The MSE values belonging to these estimators are calculated by Eq. (16). Tables 5–7 show that performances of all of the suggested robust-regression-ratio estimators perform better than the Noor-ul-Amin et al. [20] estimators. It is also noted that the values of efficiencies of the suggested estimators given in Tables 5–7 increased significantly, showing that the suggested estimators' performances would increase dramatically if there were more outliers in the data. In addition, there is an inverse relationship between the selected sample sizes to evaluate the performance of the suggested estimators. When the sample size of first phase sample (n_1) increases, the efficiencies of the suggested estimators also decrease; whereas, when the sample size of second phase sample (n_2) increases, the performances of the suggested estimators increase. These simulation findings support the results in Tables 2 and 3.

7. Conclusion

We extended Noor-ul-Amin et al. [20] estimators to robust regression-ratio-type estimators by utilizing Huber MM, Least Trimmed Square (LTS), and Least Median Square (LMS) estimators. Tables 2-7 show that the suggested robust regression-ratio-type estimators for estimating the population mean under double sampling is more efficient. The estimators in Eqs. (6)-(8) provide lower Mean Square Error (MSE) than the MSE of the Noor-ul-Amin et al. [20] estimators in Eqs. (2)-(4) under the double sampling. This means that

Table 5. The MSE and RE values of estimators in simulated data sets with 10% outliers.

			Huber M		Huber MM		LTS		LMS	
n_1	n_2	Estimator	MSE	RE	MSE	RE	MSE	RE	MSE	RE
30	10	$\overline{y}_{1j}$	2.400	1	1.781	0.742	1.803	0.751	1.765	0.735
		$\overline{y}_{2j}$	2.403	1	1.784	0.742	1.807	0.752	1.768	0.736
		$\overline{y}_{3j}$	2.403	1	1.784	0.742	1.806	0.752	1.768	0.736
	20	$\overline{y}_{1j}$	0.594	1	0.440	0.741	0.442	0.744	0.436	0.734
		$\overline{y}_{2j}$	0.594	1	0.440	0.741	0.442	0.744	0.436	0.734
		$\overline{y}_{3j}$	0.594	1	0.440	0.741	0.442	0.744	0.436	0.734
40	10	$\overline{y}_{1j}$	3.047	1	2.235	0.733	2.251	0.739	2.228	0.731
		$\overline{y}_{2j}$	3.048	1	2.235	0.734	2.252	0.739	2.229	0.732
		$\overline{y}_{3j}$	3.047	1	2.235	0.733	2.252	0.739	2.229	0.731
	20	$\overline{y}_{1j}$	0.888	1	0.646	0.727	0.648	0.730	0.640	0.721
		$\overline{y}_{2j}$	0.888	1	0.646	0.727	0.648	0.730	0.640	0.721
		$\overline{y}_{3j}$	0.888	1	0.646	0.727	0.648	0.730	0.640	0.721
	30	$\overline{y}_{1j}$	0.292	1	0.214	0.731	0.215	0.736	0.213	0.728
		$\overline{y}_{2j}$	0.292	1	0.214	0.731	0.215	0.736	0.213	0.728
		$\overline{y}_{3j}$	0.292	1	0.214	0.731	0.215	0.736	0.213	0.728
50	10	$\overline{y}_{1j}$	3.112	1	2.259	0.726	2.279	0.732	2.249	0.723
		$\overline{y}_{2j}$	3.113	1	2.259	0.726	2.280	0.732	2.250	0.723
		$\overline{y}_{3j}$	3.113	1	2.259	0.726	2.279	0.732	2.250	0.723
	20	$\overline{y}_{1j}$	1.057	1	0.761	0.720	0.762	0.721	0.757	0.716
		$\overline{y}_{2j}$	1.057	1	0.761	0.720	0.762	0.721	0.757	0.716
		$\overline{y}_{3j}$	1.057	1	0.761	0.720	0.762	0.721	0.757	0.716
	30	$\overline{y}_{1j}$	0.445	1	0.318	0.715	0.322	0.723	0.316	0.709
		$\overline{y}_{2j}$	0.445	1	0.318	0.715	0.322	0.723	0.316	0.709
		$\overline{y}_{3j}$	0.445	1	0.318	0.715	0.322	0.723	0.316	0.709
	40	$\overline{y}_{1j}$	0.184	1	0.133	0.720	0.132	0.719	0.132	0.716
		$\overline{y}_{2j}$	0.184	1	0.133	0.720	0.132	0.719	0.132	0.716
		$\overline{y}_{3j}$	0.184	1	0.133	0.720	0.132	0.719	0.132	0.716

Table 6. The MSE and RE values of estimators in simulated data sets with 20% outliers.

			Huber M		Huber MM		LTS		LMS	
n_1	n_2	Estimator	MSE	RE	MSE	RE	MSE	RE	MSE	RE
30	10	$\overline{y}_{1j}$	2.006	1	1.639	0.817	1.571	0.783	1.553	0.774
		$\overline{y}_{2j}$	2.006	1	1.639	0.817	1.571	0.783	1.553	0.774
		$\overline{y}_{3j}$	2.006	1	1.639	0.817	1.571	0.783	1.553	0.774
	20	$\overline{y}_{1j}$	0.505	1	0.417	0.824	0.406	0.802	0.397	0.785
		$\overline{y}_{2j}$	0.506	1	0.417	0.824	0.406	0.802	0.397	0.785
		$\overline{y}_{3j}$	0.506	1	0.417	0.824	0.406	0.803	0.397	0.785
40	10	$\overline{y}_{1j}$	2.270	1	1.811	0.798	1.772	0.780	1.748	0.770
		$\overline{y}_{2j}$	2.272	1	1.812	0.798	1.773	0.780	1.749	0.770
		$\overline{y}_{3j}$	2.273	1	1.813	0.798	1.774	0.780	1.750	0.770
	20	$\overline{y}_{1j}$	0.783	1	0.641	0.819	0.628	0.802	0.616	0.787
		$\overline{y}_{2j}$	0.783	1	0.641	0.819	0.628	0.802	0.616	0.787
		$\overline{y}_{3j}$	0.783	1	0.641	0.819	0.628	0.802	0.616	0.787
	30	$\overline{y}_{1j}$	0.250	1	0.202	0.806	0.198	0.789	0.196	0.781
		$\overline{y}_{2j}$	0.250	1	0.202	0.806	0.198	0.789	0.196	0.781
		$\overline{y}_{3j}$	0.250	1	0.202	0.806	0.198	0.789	0.196	0.781
50	10	$\overline{y}_{1j}$	2.677	1	2.187	0.817	2.151	0.804	2.123	0.793
		$\overline{y}_{2j}$	2.677	1	2.187	0.817	2.152	0.804	2.123	0.793
		$\overline{y}_{3j}$	2.677	1	2.187	0.817	2.151	0.804	2.123	0.793
	20	$\overline{y}_{1j}$	0.930	1	0.763	0.820	0.744	0.799	0.728	0.782
		$\overline{y}_{2j}$	0.931	1	0.763	0.820	0.744	0.799	0.728	0.782
		$\overline{y}_{3j}$	0.931	1	0.763	0.820	0.744	0.799	0.728	0.782
	30	$\overline{y}_{1j}$	0.408	1	0.327	0.803	0.322	0.789	0.318	0.780
		$\overline{y}_{2j}$	0.408	1	0.328	0.803	0.322	0.788	0.318	0.780
		$\overline{y}_{3j}$	0.408	1	0.328	0.803	0.322	0.788	0.319	0.780
	40	$\overline{y}_{1j}$	0.149	1	0.122	0.820	0.120	0.803	0.119	0.796
		$\overline{y}_{2j}$	0.149	1	0.122	0.820	0.120	0.803	0.119	0.796
		$\overline{y}_{3j}$	0.149	1	0.122	0.820	0.120	0.803	0.119	0.796

Table 7. The MSE and RE values of estimators in simulated data sets with 30% outliers.

			Huber M		Huber MM		LTS		LMS	
n_1	n_2	Estimator	MSE	RE	MSE	RE	MSE	RE	MSE	RE
30	10	$\overline{y}_{1j}$	2.041	1	1.919	0.940	1.822	0.892	1.791	0.877
		$\overline{y}_{2j}$	2.047	1	1.924	0.940	1.827	0.893	1.796	0.878
		$\overline{y}_{3j}$	2.052	1	1.930	0.940	1.832	0.893	1.802	0.878
	20	$\overline{y}_{1j}$	0.441	1	0.424	0.961	0.404	0.917	0.403	0.916
		$\overline{y}_{2j}$	0.441	1	0.424	0.961	0.404	0.917	0.404	0.916
		$\overline{y}_{3j}$	0.441	1	0.424	0.961	0.404	0.917	0.404	0.916
40	10	$\overline{y}_{1j}$	2.159	1	2.061	0.955	1.977	0.916	1.976	0.915
		$\overline{y}_{2j}$	2.162	1	2.065	0.955	1.981	0.916	1.979	0.915
		$\overline{y}_{3j}$	2.165	1	2.067	0.955	1.983	0.916	1.982	0.915
	20	$\overline{y}_{1j}$	0.669	1	0.636	0.952	0.607	0.908	0.605	0.904
		$\overline{y}_{2j}$	0.669	1	0.636	0.952	0.607	0.908	0.605	0.904
		$\overline{y}_{3j}$	0.669	1	0.637	0.952	0.607	0.908	0.605	0.904
	30	$\overline{y}_{1j}$	0.239	1	0.230	0.962	0.219	0.917	0.219	0.917
		$\overline{y}_{2j}$	0.239	1	0.230	0.962	0.220	0.917	0.220	0.917
		$\overline{y}_{3j}$	0.240	1	0.230	0.962	0.220	0.918	0.220	0.917
50	10	$\overline{y}_{1j}$	1.967	1	1.848	0.939	1.783	0.906	1.711	0.870
		$\overline{y}_{2j}$	1.985	1	1.865	0.940	1.800	0.907	1.728	0.871
		$\overline{y}_{3j}$	1.999	1	1.880	0.940	1.815	0.908	1.743	0.872
	20	$\overline{y}_{1j}$	0.981	1	0.899	0.916	0.878	0.895	0.876	0.893
		$\overline{y}_{2j}$	0.985	1	0.901	0.915	0.880	0.894	0.878	0.892
		$\overline{y}_{3j}$	0.987	1	0.903	0.915	0.882	0.894	0.880	0.892
	30	$\overline{y}_{1j}$	0.333	1	0.327	0.984	0.293	0.883	0.290	0.873
		$\overline{y}_{2j}$	0.338	1	0.332	0.984	0.299	0.884	0.293	0.868
		$\overline{y}_{3j}$	0.341	1	0.336	0.984	0.302	0.886	0.296	0.867
	40	$\overline{y}_{1j}$	0.163	1	0.160	0.976	0.154	0.944	0.153	0.933
		$\overline{y}_{2j}$	0.164	1	0.160	0.976	0.155	0.944	0.153	0.933
		$\overline{y}_{3j}$	0.164	1	0.161	0.976	0.155	0.945	0.154	0.933

the suggested estimators outperform the existing ratio estimators in terms of mean squared error. According to both real data and simulation studies, the best result is obtained using the estimators proposed based on the LMS estimate. It is recommended to use the suggested estimators in practice when there are outliers in the data set. In the forthcoming studies, we hope to improve new estimators based on robust regression techniques in other sampling designs.

Supplementry data is available at:
file:///C:/Users/Asus/AppData/Local/Temp/
Supplementary%20File.pdf

References

1. Kadilar, C., Candan, M., and Cingi, H. "Ratio estimators using robust regression", *Hacettepe Journal of Mathematics and Statistics*, **36**(2), pp. 181–188 (2007).
2. Oral, E. and Kadilar, C. "Robust ratio-type estimators in simple random sampling", *Journal of the Korean Statistical Society*, **(4)**, pp. 457–467 (2011a). DOI:10.1016/j.jkss.2011.04.001
3. Oral, E. and Kadilar, C., "Improved ratio estimators via modified maximum likelihood, Pakistan Journal of Statistics", **27**(3), pp. 269–282 (2011b).
4. Kadilar, C. and Cingi, H. "Ratio estimators in simple random sampling", *Applied Mathematics and Computation*, **151**(3), pp. 893–902 (2004). https://doi.org/10.1016/S0096-3003(03)00803-8
5. Abid, M., Nazir, H.Z., Riaz, M., et al. "Improved ratio estimators using some robust measures", *Hacettepe Journal of Mathematics and Statistics*, **47**(5), pp. 1375–1393 (2018).
6. Abid, M., Ahmed, S., Tahir, M., et al. "Improved ratio estimators of variance based on robust measures", *Scientia Iranica*, **26**(4), pp. 2484–2494 (2019). DOI: 10.24200/sci.2018.20604
7. Zaman, T. and Bulut, H. "Modified ratio estimators using robust regression methods", *Communications in Statistics-Theory and Methods*, **48**(8), pp. 2039–2048 (2019). https://doi.org/10.1080/03610926.2018.1441419
8. Zaman, T. "Improvement of modified ratio estimators using robust regression methods", *Applied Mathematics and Computation*, **348**, pp. 627–631 (2019). https://doi.org/10.1016/j.amc.2018.12.037
9. Subzar, M., Bouza, C.N., and Al-Omari, A.I. "Utilization of different robust regression techniques for estimation of finite population mean in srswor in case of presence of outliers through ratio method of estimation", *Investigacin Operacional*, **40**(5), pp. 600–609 (2019).
10. Zaman, T. and Bulut, H. "Modified regression estimators using robust regression methods and covariance matrices in stratified random sampling", *Communications in Statistics-Theory and Methods*, **49**(14), pp. 3407–3420 (2020). https://doi.org/10.1080/03610926.2019.1588324
11. Bulut, H. and Zaman, T. "An improved class of robust ratio estimators by using the minimum covariance determinant estimation", *Communications in Statistics-Simulation and Computation*, **51**(5), pp. 2457–2463 (2022). https://doi.org/10.1080/03610918.2019.1697818
12. Shahzad, U., Al-Noor, N.H., Hanif, M., et al. "Imputation based mean estimators in case of missing data utilizing robust regression and variance-covariance matrices", *Communications in Statistics-Simulation and Computation*, **51**(8), pp. 4276–4295 (2022). https://doi.org/10.1080/03610918.2020.1740266
13. Naz, F., Abid, M., Nawaz, T., et al. "Enhancing the efficiency of the ratio-type estimators of population variance with a blend of information on robust location measures", *Scientia Iranica*, **27**(4), pp. 2040–2056 (2019). DOI: 10.24200/sci.2019.5633.1385
14. Subzar, M., Al-Omari, A.I., and Alanzi, A.R. "The robust regression methods for estimating of finite population mean based on SRSWOR in case of outliers", *CMC-Computers Materials and Continua*, **65**(1), pp. 125–138 (2020). DOI:10.32604/cmc.2020.010230
15. Grover, L.K. and Kaur, A. "An improved regression type estimator of population mean with two auxiliary variables and its variant using robust regression method", *Journal of Computational and Applied Mathematics*, **382**, pp. 1–18 (2021). https://doi.org/10.1016/j.cam.2020.113072
16. Ali, N., Ahmad, I., Hanif, M., et al. "Robust-regression-type estimators for improving mean estimation of sensitive variables by using auxiliary information", *Communications in Statistics-Theory and Methods*, **50**(4), pp. 979–992 (2021). https://doi.org/10.1080/03610926.2019.1645857
17. Saleem, I., Sanaullah, A., and Hanif, M. "Double-sampling regression-cum-exponential estimator of the mean of a sensitive variable", *Mathematical Population Studies*, **26**(3), pp. 163–182 (2019). https://doi.org/10.1080/08898480.2019.1565273
18. Neyman, J. "Contribution to the theory of sampling human population", *Journal of the American Statistical Association*, **33**(201), pp. 101–16 (1938).
19. Sukhatme, B.V. "Some ratio-type estimators in two-phase sampling", *Journal of the American Statistical Association*, **57**(299), pp. 628–32 (1962).
20. Noor-ul-Amin, M., Shahbaz, M.Q., and Kadilar, C. "Ratio estimators for population mean using robust regression in double sampling", *Gazi University Journal of Science*, **29**(4), pp. 793–798 (2016).

21. Singh, G.N., Suman, S., and Kadilar, C. "On the use of imputation methods for missing data in estimation of population mean under two-phase sampling design", *Hacettepe Journal of Mathematics and Statistics*, **47**(6), pp. 1715–1729 (2018).

22. Guha, S. and Chandra, H. "Improved chain-ratio type estimator for population total in double sampling", *Mathematical Population Studies*, **27**(4), pp. 216–231 (2020). https://doi.org/10.1080/08898480.2019.1626635

23. Guha, S. and Chandra, H. "Improved estimation of finite population mean in two-phase sampling with subsampling of the nonrespondents", *Mathematical Population Studies*, **28**(1), pp. 24–44 (2021). https://doi.org/10.1080/08898480.2019.1694325

24. Rousseeuw, P.J. and Leroy, A.M. "Robust regression and outlier detection", *Wiley Series in Probability and Mathematical Statistics*, New York: Wiley (1987).

25. Li, L.M. "An algorithm for computing exact least-trimmed squares estimate of simple linear regression with constraints", *Computational Statistics and Data Analysis*, **48**(2005), pp. 717–734 (2004). https://doi.org/10.1016/j.csda.2004.04.003

26. Ergül, B. "Robust regression and applications", Master Thesis, Eskişehir Osmangazi University. 184031 (2006).

27. Huber, P.J. "Robust regression: asymptotics, conjectures and Monte Carlo", *The Annals of Statistics*, pp.799–821 (1973).

28. Yohai, V.J. "High breakdown-point and high efficiency robust estimates for regression", *The Annals of Statistics*, **15**(2), pp. 642–56 (1987).

29. Venables, W.N. and Ripley, B.D., *Modern Applied Statistics with S*, Fourth Edition, Springer, New York (2002).

30. Wang, J., Zamar, R., Marazzi, A., et al. "Robust: Port of the S+" Robust Library, *R Package Version* 0.5-0.0. https://CRAN.R-project.org/package=robust (2020).

31. Zaman, T., Sağlam, V., Sağ ir, M., et al. "Investigation of some estimators via taylor series approach and an application", *American Journal of Theoretical and Applied Statistics*, **3**(5), pp. 141–147 (2014). DOI: 10.11648/j.ajtas.20140305.14

Biographies

Tolga Zaman is an Associate Professor at the Department of Statistics in Cankiri Karatekin University, Cankiri, Turkey. He received his MS and PhD degrees in Statistics from Ondokuz Mayis University Samsun, Turkey in 2013 and 2017, respectively. His research interests are sampling theory, resampling methods, robust statistics, and statistical inference. He has published more than 60 research papers in international/national journals and conferences.

Hasan Bulut is working as Associate Professor in Department of Statistics at Ondokuz Mayıs University, where he received his Doctor degree in 2017 based on robust clustering and robust multivariate analyses. His main research interests have been the fields of socio- economic development, robust principal component analysis, robust clustering analysis, multivariate statistical methods, and applied statistics. He has papers published in journals like Socioeconomic Planning Sciences, the Journal of Applied Statistics, Communication in Statistics: Theory and Methods, and Communication in Statistics: Simulation and Computation. Moreover, he has a book about multivariate statistical methods with R applications.