A Modified Robust Estimator under Heteroscedasticity and Unusual Observations for Linear Regression Model

Shagufta Mubarik[1] (smc7824@yahoo.com)
College of Statistical and Actuarial Sciences, University of the Punjab, Lahore, Pakistan.

Dr. Maryam Ilyas (maryamilyas@hotmail.com)
College of Statistical and Actuarial Sciences, University of the Punjab, Lahore, Pakistan.

ABSTRACT

Regression estimators are not robust in the presence of unusual observations. Heteroscedasticity and unusual observations cannot proceed together therefore to address these problems robust versions of weighted least square are used. Yet, weighted least square estimators are also affected in the presence of unusual observations. Therefore, an estimator is required which perform well. A modified estimator is proposed in this study which is more outlier resistant. The comparative performance of modified estimator is evaluated by conducting two studies. The performance is investigated by Monte Carlo simulation and confirmed by real data. The results showed that the modified estimator outperformed.

Key Words: *Heteroscedasticity; High leverage points; influential observations; outliers; weighted least squares.*

JEL Classification: *C01, C15*

1. INTRODUCTION

Regression analysis is commonly used to model the relationship between two or more quantitative variables. One of the common assumptions of error term is that variance of error term is constant but in many practical situations it is violated therefore Ordinary Least Square (OLS) estimators become invalid. OLS estimators are also affected by unusual observations. Weighted Least Square (WLS) estimators are designed to counter the effects of heteroscedasticity and robust estimators are used to deal with heteroscedasticity and unusual observations. If the form of heteroscedasticity is known, then WLS estimators are used but for unknown form Fuller and Rao

1. Corresponding author

(1978) presented the Estimated Weighted Least Square (EWLS) estimator instead of robust versions of WLS. White (1980) presented a Heteroscedasticity Consistent Covariance Matrix (HCCM) estimator which is more consistent in the presence of heteroscedasticity.

Unusual observations in linear regression may occur in three ways that are outliers, high leverage points and influential point. The observations lying away from the majority of the data are called outliers. Whereas an observation that has an extreme value on the explanatory variable 'X' is said to be high leverage point. The existence of high leverage observations, makes it problematic the regression statistics to find a robust fit (see e.g. Billor, Chatterjee and Hadi, 2006). If a point has unusual x-value and y-value, then it is an influential point. Also an outlier that substantially affects the slope of regression line is said to be an influential point, (Imon, 2009).

In view of Acitas and Senoglu (2019), outliers may exist in the data for linear regression model. In the existence of outliers, the distribution of error terms may have longer tails than normal distribution. Actually long-tailed error distributions are more widespread in practical studies. There are various estimators which are used to deal with heteroscedasticity and outlying observations. Midi, Rana and Imon in (2009) used a WLS estimator for heteroscedastic data which they termed as Montgomery, Peck and Vining (MPV) estimator. If data contains outliers and heteroscedasticity both, WLS estimators are also affected. Thus Midi, Rana and Imon (2009) proposed a Robust Weighted Least Square (RWLS) estimator which is a better estimator. Some weaknesses existed in this estimator and need to improve it. Therefore, Midi, Rana and Imon in 2013 modified this estimator, termed as Two-Step Robust Weighted Least Square (TSRWLS) estimator which is more outlier resistant. There are many robust's weight functions that are commonly used such as Bisquare, Huber and Hampel in robust statistics see.e.g. Ghazali, Abd Halim & Jamidin, 2017. Hence, this study was implemented to investigate the performance of TSRWLS using different robust's weight functions (Cauchy and German-Maclure weight function (Bhar, 2010). Generally, the estimators available in literature to deal with heteroscedastic and outlying data are OLS, MPV, RWLS and TSRWLS. Only TSRWLS performed satisfactorily in the scenario of heteroscedasticity and outliers but for influential observations and heteroscedastic data these estimators are not discussed until now. Therefore, a new estimator, Modified Two-Step Robust Weighted Least Square (MTSRWLS) is proposed to explore the competitive performance of proposed and existing WLS estimators. Simulated data sets were generated by considering the effects of sample size, level of contamination and magnitude of outliers on the existing and the proposed WLS estimators under two

different scenarios; 1) heteroscedasticity and outliers, 2) heteroscedasticity and influential observations.

Remaining paper is organized as: Section 2 defines material & methods, robust versions of weighted least squares is presented in Section 3. Section 4 reports simulation results whereas in section 5 concluding remarks are presented.

2. MATERIALS AND METHODS

The General Linear Regression (GLR) model is considered as,

$$y= X\beta+\varepsilon \quad [1]$$

Where $y=(y_1,\ldots\ldots,y_n)'$ is the $n\times 1$ vector of observations, X is the fixed design matrix of 'k' predictor variables of order $n\times(k+1)$, $\beta=(\beta_0, \beta_1, \ldots, \beta_k)'$ is the vector of regression parameters of order $(k+1)\times 1$ and $\varepsilon =(\varepsilon_1,\ldots, \varepsilon_n)'$ is the $n\times 1$ vector of errors which assumes normal distribution having mean zero and constant variance σ^2, 'k' is the number of predictors and 'n' is the number of observations. The OLS estimator of regression parameter is

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{y})$$

and $var(\hat{\boldsymbol{\beta}}_{OLS}) = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$

If the error term has constant variance that is $\Psi = \sigma^2 I$, then $var(\hat{\boldsymbol{\beta}}_{OLS}) = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$ and the value of σ^2can be estimated as; $\hat{\sigma}^2 = \hat{\varepsilon}'\hat{\varepsilon}/(n-k-1)$ but if error terms are heteroscedastic then $\Psi = \sigma^2 V$ is the positive definite matrix. If $W=V^{-1}$ the estimated regression coefficient can be written as $\hat{\beta}_{WLS} = (\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{y})$ where W is the diagonal matrix having diagonal elements $w_1, w_2,\ldots\ldots,w_n$. The variance-covariance matrix of WLS is $\text{cov}(\hat{\beta}_{WLS}) = \sigma^2_{WLS}(X'WX)^{-1}$ where σ^2_{WLS} can be estimated by $\hat{\sigma}^2_{WLS} = \frac{\sum w_i\hat{\varepsilon}_i^2}{n-k-1}$.

2.1 Robust Versions of Weighted Least Square

Midi, Rana and Imon (2009) presented a WLS method for heteroscedastic data and used weight function by finding numerous near neighbors in the predictor variable and they named this estimator as MPV and for outliers along with heteroscedastic problem they presented RWLS

estimator. This estimator used fitted values of Least Trimmed Square (LTS) of median absolute deviation of response variable and median of predictor variable as initial weights and for final weights Huber robust weight function is used. Midi, Rana and Imon in (2013) proposed an improved version of this estimator i.e. TSRWLS which is more outlier resistant. This method used the inverse of the squared fitted values of scaled residuals as initial weights and for final weights Huber robust weight functions was used.

A modification of TSRWLS is suggested in this study i.e. Modified form of TSRWLS. This estimator is more efficient as compared to other estimators. The following steps are used to construct the weight matrix $\boldsymbol{W}$ used in MTSRWLS estimator.

1. By using least trimmed squares, find the fitted values, $\hat{y}_i$ and residuals $\hat{\varepsilon}_{LTS}$.
2. Cauchy weight function (Bhar, 2010) is used as initial weight i.e., $w_{1i} = 1/(1 + (\hat{\varepsilon}_i/2.385)^2)$, where 2.385 is the tuning constant and is the standardized residuals of the LTS
3. For final weighs, German-Maclure weight function (Bhar, 2010) is used i.e. $w_{2i} = 1/(1 + \hat{\varepsilon}_i^2)^2$ where $\hat{\varepsilon}_i$ is the scaled residual of LTS.
4. Multiply w_{1i} and w_{2i} to get the diagonal elements w_i of weight matrix $\boldsymbol{W}$.
5. Perform weighted least squares to obtain the regression coefficients.

In this study it is compared the following estimators: OLS, MPV, RWLS, TSRWLS and MTSRWLS in the existence of heteroscedastiy and unusual observations for Simple Linear Regression (SLR) model.

3. SIMULATION STUDIES

Simulation studies are presented in this section regarding the comparison of MTSRWLS with the existing estimators. For the comparisons two simulation scenarios are considered.

3.1 Heteroscedasticity and Outliers

To generate heteroscedastic data, the following model is used,

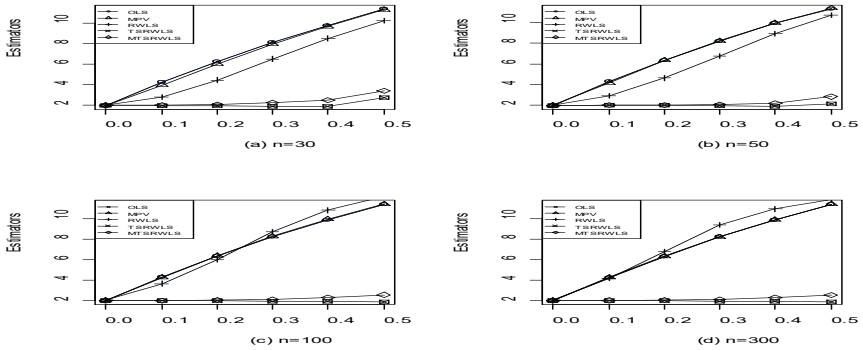
$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad [2]$$

Where β_0 =3 and β_1 =2, x_i is generated by dividing the sample size into five sub-samples that is x_i=(x_1, x_2, x_3, x_4, x_5) and x_1~ *Uniform*(n_1, 1,

9), x_2 ~ *Uniform*(n_2, 10, 19), x_3 ~ *Uniform*(n_3, 20, 29), x_4 ~ *Uniform*(n_4, 30, 39), x_5 ~ *Uniform*(n_5, 40, 49). Thus, *X* is a design matrix of fixed values. The ε_i are generated in a way to induce heteroscedasticity such as $\varepsilon_i = x_i \varepsilon_i^*$ where ε_i^* is normally distributed with mean 0 and variance 1 (Midi, Rana and Imon, 2009). Contaminated data is generated by using different magnitude of outliers by considering 12σ, 9σ & 6σ distances. The values lying well outside these distances are considered as outliers. For example, the distance for 12σ is generated as, $\varepsilon^*_{(cont)} = \bar{\varepsilon}^* \pm 12 s_{\varepsilon^*}$ where $\bar{\varepsilon}^*$is the mean and $s_{\varepsilon^*}$is the standard deviation of *n* residuals which are generated from $\varepsilon^* \sim N(0,1)$. The heteroscedastic errors with outliers are computed as $\varepsilon^*_{i(cont)} = x_i \varepsilon^*_{i(cont)}$. The Monte Carlo runs is set to be 10000. In order to approve that either generated data is heteroscedastic or not, single Monte Carlo run was used to confirm heteroscedasticity using Goldfeld-Quandt (GQ) test. The data generated by single Monte Carlo run showed for n=50, *GQ = 817.4653*, v_1 = 23, v_2 = 23, *p-value= 2.2e*$^{-16}$. It is clearly indicated the data contain heteroscedasticity as the null hypothesis of no heteroscedasticity is rejected (*p-value<0.05).* Heteroscedastic data with outliers was generated for sample sizes i.e. n = 30, 50, 100 and 300 with various percentages of outliers (0%, 10%, 20%, 30%, 40% and 50%) and three levels of magnitude of outliers i.e. 12$\boldsymbol{\sigma}$, 9$\boldsymbol{\sigma}$ and 6$\boldsymbol{\sigma}$.

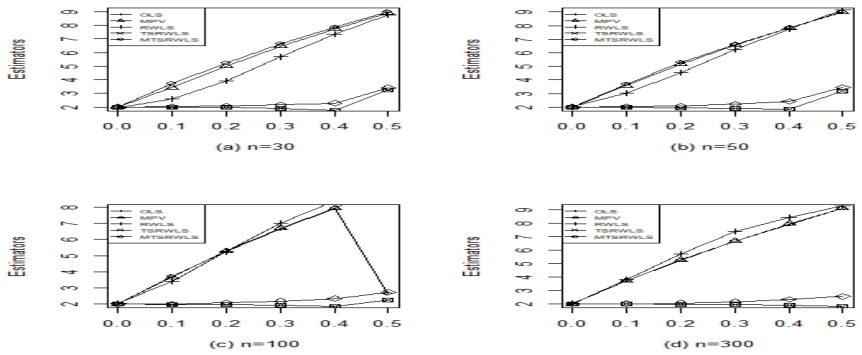
The comparisons of estimators for 12σ distance

Figure 1



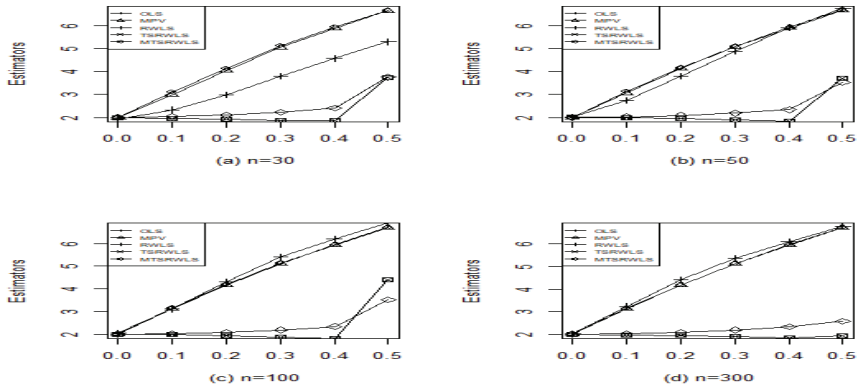
The comparisons of estimators for 9σ distance

Figure 2



The comparisons of estimators for 6σ distance

Figure 3



The comparisons of estimators in the presence of outliers and heteroscedasticity is presented through graphs (Figure 1, Figure 2 and Figure 3) by plotting the estimated value and delta (percentages of contamination) by considering 12σ, 9σ and 6σ distances for sample sizes; n=30, 50, 100 and 300.

For 12σ distance by considering all sample sizes, Figure 1 clearly display that as the percentage of contamination increased the OLS, MPV and RWLS estimators are far away from true value of parameter. However, the estimated value of MTSRWLS estimator is close to true value of parameter. Although TSRWLS estimator is also close to true value and give better performance than other estimators but not better than modified estimator. When

sample size is increased to 50 and 100 the value of modified estimator is more closed to true parameter. And for large sample size i.e. 300 the performance of MTSRWLS is outstanding. Hence for 12σ distance the performance of MTSRWLS is very good for all considering sample sizes.

When the magnitude of outliers is considered to 9σ for sample size 30 the estimated values of MTSRWLS and TSRWLS are very close to true value except at 50% contamination level but as the sample size increased to 50, 100 and 300 the performance of MTSRWLS is exceptional. Therefore, the performance of MTSRWLS is outstanding for small and large sample size as well.

For 6σ distance and n=30, except 50% contamination the estimated value of MTSRWLS estimator is very close to true value. Other estimators under consideration are far away from true value yet TSRWLS estimator give better performance but MTSRWLS outperform. Same pattern was observed for sample size 50 and 100 but for large sample size i.e. 300 the performance of MTSRWLS remain outstanding.

3.1.1 Real Data Application

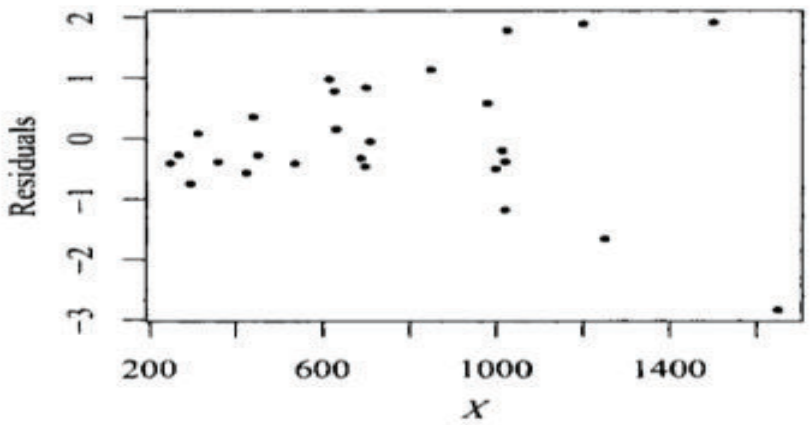
Results of simulation studies are confirmed by exploring the performance of the OLS, MPV, RWLS, TSRWLS, and MTSRWLS using real data sets.

Industrial Establishments Data

The data taken from (Chatterjee & Hadi, 2006, p. 164) consisting of the number of supervisors (y) and the number of supervised workers (X). The original data is heteroscedastic and two data points are changed to infuse outliers. To check whether data contains heteroscedasticity or not a residual plot is used. A plot (Figure 4) of the predictor variable and standardized reiduals clearly indicate the presence of heteroscedasticy.

Plot of the predictor variable and standardized residuals

Figure 4



To prove the simulation results the proposed and the existing estimators are applied to original and modified data set.

From the Table 1 results of original data are displayed, it is clear from the results that MTSRWLS and OLS estimators have estimated value approximately same but SE of MTSRWLS is smaller than OLS and all other estimators. The value of t-statistic is also high and RMSE is low. All estimators have high value of SE except MTSRWLS. Thus, the proposed estimator is not affected by heteroscedasticity.

To create outliers in original data, two data points are changed. To check whether modified data contain heteroscedasticity or not GQ test was applied and the result of test indicated that the *p-value* of test <0.05 so there is heteroscedasticity i.e. $GQ = 73.5134$, df1 = 12, df2 = 11, *p-value = 1.028e-08.* The results of Table 2 show that when heteroscedasticity and outliers are present in data the estimated value of OLS, MPV, RWLS and TSRWLS are affected by outliers and their estimated values are far away from the true value i.e 0.10. The estimated value of MTSRWLS estimator is not affected by outliers.

Thus, these numerical results signify that the proposed estimator does a superb job for both clean and contaminated data.

Summary Statistics of Heteroscedastic Data (Original Data)

Table 1

n=27	*OLS*	*MPV*	*RWLS*	*TSRWLS*	*MTSRWLS*
Estimate	0.10	0.12	0.10	0.10	**0.10**
SE	0.01	0.01	0.01	0.01	**0.00**
t-value	9.30	10.57	9.71	9.44	**21.05**
RMSE	9.56	6.27	8.20	8.79	**3.30**

Summary Statistics of Heteroscedastic and Outlying Data (Modified Data)

Table 2

n=27	*OLS*	*MPV*	*RWLS*	*TSRWLS*	*MTSRWLS*
Estimate	0.13	0.15	0.12	0.13	**0.10**
SE	0.02	0.02	0.02	0.01	**0.01**
t-value	5.68	7.68	6.85	9.63	**17.28**
RMSE	19.03	10.95	13.63	8.44	**3.95**

3.2 Heteroscedasticity and Influential Observations

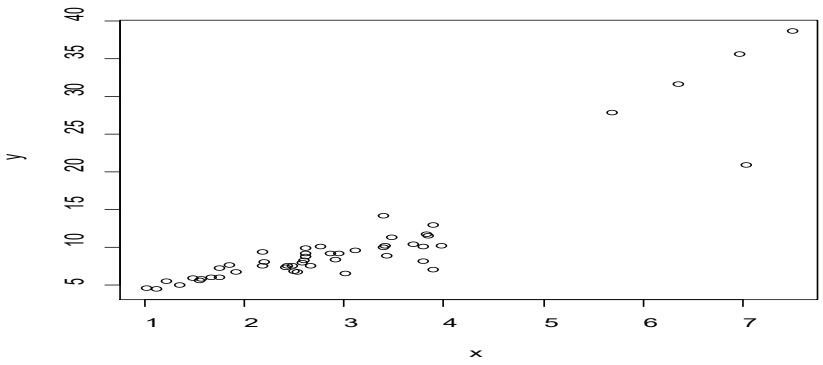
Monte Carlo simulation study is conducted to investigate the effect of large and small sample sizes and different percentages of contamination on simple linear regression model in the presence of heteroscedasticity and influential observations (Turkin & Toktamis, 2013).

For $\beta_0=3$ and $\beta_1=2$ the simple linear regression model (2) is used for simulation. The first $100(1-\delta)\%$ good observations of the predictor variable 'x_i' is simulated by *Uniform (1, 4)* distribution where $\boldsymbol{\varepsilon_1}$ is computed as $\varepsilon_1 \sim N(0,0.2)$. The heteroscedastic data with no influential data points, error term is simulated by using the following relation, $\varepsilon_i^* = x_i^* \varepsilon_1$ and then 'y_i'values for clean observations are computed. The remaining $100\ddot{a}\%$ observations are taken as influential observations. The contaminated observations for predictor variable x_c are generated as *N(7,0.5)*, 'y_c' values for contaminated observations are computed by using contaminated predictor variable and ε_c^* are generated in a way that they also transmit heteroscedasticity, according to this relation $\mathring{a}_c^* = x_c \times (\mathring{a}_c \sim N(2,0.5))$. Now $\boldsymbol{X}$ is the design matrix of fixed values of contaminated x_c and clean x_i observations and $\boldsymbol{y}$ is the vector of good and bad observations such as $\boldsymbol{y}=(y_i, y_c)$. The error term is the vector of contaminated error and uncontaminated error values i.e. $\boldsymbol{\varepsilon_i} = (\varepsilon_i, \varepsilon_c^*)$. For the detection of

heteroscedasticity and influential observations, the Goldfeld-Quandt test was used on single simulated data set to confirm the infused heteroscedasticity. Influential data points were detected graphically. The calculated value of GQ test for n=50, *GQ = 8.1821,* df1 = 23, df2 = 23, *p-value = 1.93e-06.* As the *p-value < 0.05* is the evidence of heteroscedasticity. To confirm whether the data contains influential data points or not a scatter plot was used. According to Chatterjee & Hadi in 1986, graphical methods which are based on the residuals would fail to detect these unusual data points therefore they used the scatter plot to detect the influential observations. It is clearly indicated from Figure 5 that influential observations exist in data. The fitted line will change if these observations are deleted.

Scatter plot of heteroscedastic and influential observations

Figure 5



Now in the presence of heteroscedasticity and influential observations, the comparative performance of different estimators in terms of estimated value, standard error, t-statistic and Root Mean Squares Error (RMSE) are obtained at different percentages of influential observations (0%, 10%, 20%, 30%, 40% and 50%) and for sample sizes i.e. (*n*=30, 50, 100, 300). From Table 3 to Table 6 display the summary statistics for true value of regression coefficient $\beta_1=2$. For simple linear regression model, Root Mean Square Error (RMSE) is calculated as; $\text{RMSE}=\sqrt{N\times\{MSE(\beta_0)+MSE(\beta_1)\}}$ **,** where 'N' is the Monte Carlo runs e.g. see Ahmed, Aslam and Pasha (2011).

Comparative Performance of Different Estimators for n=30

Table 3

		Percentages of contamination					
Estimators		**0%**	**10%**	**20%**	**30%**	**40%**	**50%**
OLS	Estimate	2.00	4.28	4.62	4.77	4.83	4.88
	SE	0.00	0.11	0.19	0.27	0.20	0.15
	t-value	7165.42	37.39	23.83	17.62	24.38	32.64
	RMSE	0.75	39.86	80.99	128.78	96.47	78.57
MPV	Estimate	2.00	3.32	3.74	4.08	4.31	4.48
	SE	0.25	0.41	0.36	0.39	0.31	0.52
	t-value	8.09	8.18	10.46	10.59	13.82	8.55
	RMSE	60.30	77.99	96.96	116.64	86.54	116.70
RWLS	Estimate	1.99	3.49	3.76	4.10	4.33	4.61
	SE	0.25	0.35	0.32	0.35	0.41	0.41
	t-value	7.95	9.93	11.68	11.73	10.66	11.26
	RMSE	67.56	88.68	111.30	171.18	94.49	93.18
TSRWLS	Estimate	2.00	2.47	2.72	3.03	3.44	4.40
	SE	0.22	0.27	0.37	0.34	0.29	0.22
	t-value	9.02	9.27	7.30	8.79	11.85	20.28
	RMSE	54.22	74.70	63.32	79.43	73.93	88.85
MTSRWLS	Estimate	**2.00**	**2.03**	**2.08**	**2.21**	**2.54**	**4.13**
	SE	0.13	0.10	0.18	0.25	0.12	0.13
	t-value	15.96	20.23	11.77	8.78	21.00	32.57
	RMSE	32.80	27.89	48.89	71.41	38.15	57.80

Comparative Performance of Different Estimators for n= 50

Table 4

		Percentages of contamination					
Estimators		**0%**	**10%**	**20%**	**30%**	**40%**	**50%**
OLS	Estimate	2.00	4.28	4.62	4.76	4.83	4.87
	SE	0.27	0.23	0.11	0.06	0.31	0.05
	t-value	7.34	18.46	41.02	79.56	15.77	99.20
	RMSE	76.39	79.71	47.13	27.15	152.71	26.27
MPV	Estimate	2.00	3.19	3.71	4.06	4.30	4.47
	SE	0.19	0.28	0.33	0.28	0.31	0.22
	t-value	10.37	11.33	11.31	14.42	13.84	19.97
	RMSE	34.46	59.42	92.24	69.31	90.63	63.84
RWLS	Estimate	2.00	3.22	3.71	4.05	4.27	4.53
	SE	0.19	0.33	0.27	0.24	0.24	0.22
	t-value	10.56	9.86	13.96	16.90	17.46	20.62
	RMSE	36.21	179.91	84.42	94.15	83.06	93.81
TSRWLS	Estimate	2.00	2.38	2.63	2.92	3.30	4.26
	SE	0.19	0.22	0.18	0.24	0.24	0.21
	t-value	10.78	10.63	14.64	12.38	13.76	20.71
	RMSE	45.19	59.38	43.91	62.65	81.48	107.71
MTSRWLS	Estimate	2.00	2.02	2.06	2.13	2.38	3.91
	SE	0.14	0.11	0.11	0.09	0.13	0.12
	t-value	14.39	18.78	18.30	22.52	18.07	32.69
	RMSE	35.28	28.41	29.84	26.50	43.37	59.87

Comparative Performance of Different Estimators for n=100

Table 5

Estimators		Percentages of contamination					
		0%	10%	20%	30%	40%	50%
OLS	Estimate	2.00	4.27	4.62	4.75	4.82	4.86
	SE	0.24	0.00	0.15	0.09	0.03	0.20
	t-value	8.48	54.99	30.88	50.11	156.37	24.08
	RMSE	66.26	0.00	60.70	44.21	15.76	110.36
MPV	Estimate	2.00	3.14	3.72	4.07	4.30	4.46
	SE	0.12	0.27	0.20	0.19	0.19	0.16
	t-value	16.06	11.71	18.83	20.91	22.67	28.20
	RMSE	23.72	65.09	47.97	50.36	53.32	42.06
RWLS	Estimate	2.00	3.04	3.57	3.88	4.10	4.39
	SE	0.13	0.17	0.33	0.18	0.19	0.13
	t-value	15.99	17.43	10.70	21.77	21.97	33.13
	RMSE	23.28	31.51	57.21	70.44	91.50	27.25
TSRWLS	Estimate	2.00	2.30	2.56	2.83	3.19	4.01
	SE	0.12	0.13	0.15	0.17	0.18	0.18
	t-value	17.15	17.73	17.32	16.30	17.46	22.84
	RMSE	29.09	29.96	32.36	45.40	52.53	85.38
MTSRWLS	Estimate	2.00	2.01	2.04	2.09	2.27	3.61
	SE	0.07	0.07	0.08	0.10	0.10	0.16
	t-value	27.95	28.23	24.72	20.78	23.79	22.59
	RMSE	18.71	17.91	20.71	27.72	28.13	78.64

Comparative Performance of Different Estimators for n=300

Table 6

Estimators		Percentages of contamination					
		0%	10%	20%	30%	40%	50%
OLS	Estimate	2.00	4.27	4.61	4.75	4.82	4.86
	SE	0.02	0.11	0.08	0.05	0.09	0.16
	t-value	111.45	38.27	56.12	96.41	55.27	30.59
	RMSE	4.95	38.97	33.64	22.44	43.30	84.56
MPV	Estimate	2.00	3.15	3.73	4.07	4.30	4.46
	SE	0.06	0.13	0.13	0.12	0.10	0.11
	t-value	32.24	24.89	27.91	35.07	43.28	40.46
	RMSE	14.63	30.46	35.91	30.96	26.51	27.13
RWLS	Estimate	2.00	2.78	3.25	3.59	3.89	4.25
	SE	0.06	0.14	0.11	0.22	0.10	0.08
	t-value	32.04	19.99	29.86	16.38	39.90	50.31
	RMSE	14.92	31.16	33.71	37.22	38.35	31.36
TSRWLS	Estimate	2.00	2.27	2.52	2.79	3.12	3.70
	SE	0.06	0.08	0.09	0.09	0.09	0.09
	t-value	34.41	29.63	27.80	30.46	34.56	41.93
	RMSE	15.29	18.96	25.00	22.77	25.58	27.13
MTSRWLS	Estimate	2.00	2.01	2.03	2.07	2.22	3.25
	SE	0.03	0.04	0.04	0.05	0.06	0.06
	t-value	67.11	56.46	54.37	43.45	38.49	57.64
	RMSE	7.76	9.32	10.39	13.04	17.06	18.81

When n=30 and 0% contamination (Table 3) the estimated values of regression parameters are close to the true value of parameter and their resultant SE are small for no influential observation. As the contamination level increased from 0 to 10% the values of the OLS, MPV and RWLS inflates the true value of parameter even TSRWLS estimators also inflates the true value slightly but MTSRWLS is the closest to the parameter value. The standard errors of all estimators are greater than the standard error of MTSRWLS and t-statistics of all estimators are relatively small except MTSRWLS at 10% contamination level. As the contamination level increases from 10% to 40%, the behaviour of all estimators are same as for 10% contamination except MTSRWLS. For 50% contamination level the value of MTSRWLS also inflates for small sample size(n=30). Overall performance of proposed estimator is much better among all existing estimators.

In Table 4 for n=50 and at 0% contamination level the estimated values of all considering estimators are close to true value and standard error of MTSRWLS is small among all estimators. When the percentage of contamination is increased the values of OLS, MPV and RWLS estimators move far away from the true parameter value. For higher percentage of contamination, the Modified estimator performs very well because the estimator has least SE, higher value of t-statistic and lower RMSE as compared to other estimators.

Similarly, for sample size 100 and 300 Table 5 and Table 6, clearly display that as the percentage of contamination increased, the OLS, MPV and RWLS estimators are inflate the true value of estimator yet TSRWLS estimator give better performance but MTSRWLS perform very well. So the performance of MTSRWLS is fabulous for considering all sample sizes.

Conclusions

The most widely used estimators available in literature are OLS, MPV, RWLS and TSRWLS. These estimators perform better in the presence of heteroscedasticity and outliers however these estimators are not used when heteroscedasticity and influential observations occur together. Therefore, an estimator which has potential to perform well in the presence of heteroscedasticity and unusual (outliers and influential) observations is required. Thus, comparison of proposed estimator with the existing estimators demonstrated in this study by using estimated regression coefficients their standard errors, t-test and efficiency by means of Root Mean Square Error (RMSE). The results obtained from simulated data for first study showed that the MTSRWLS estimator was the closest to true parameter value, has lowest Standard Error (SE) and Root Mean Square Error (RMSE) among all

considered existing estimators. To check the significance of first study a real data set is used. The analysis of real data approves the results of simulation study and showed that the performance of MTSRWLS is outstanding. The results of second study showed that the MTSRWLS perform very good for all sample sizes. Hence, the performance of MTSRWLS is found to be excellent in the presence of unusual observations along with heteroscedasticity in simple linear regression.

References

1. **Ahmed, M., Aslam, M. & Pasha, G.R.** (2011). *Inference under heteroscedasticity of unknown form using an adaptive estimator*. Communications in Statistics-Theory and Method, **40**(24): 4431– 4457.
2. **Acitas , S. & Senoglu, B.** (2019). *Ridge-type MML estimator in the linear regression model*. Iranian Journal of Science and Technology Transactions A, **43**:589-599.
3. **Bhar, L.** (2010). *Robust regression.* Advances in Data Analytical Techniques.**1**:70-78.
4. **Billor, N., Chatterjee, S. & Hadi, A. S.** (2006). *A re-weighted least squares method for robust regression estimation.* American journal of mathematical and management sciences, **26**(34), 229-252.
5. **Chatterjee, S., & Hadi, A. S.** (1986). *Influential observations, high leverage points, and outliers in linear regression.* Statistical Science, **1**(3): 379-393.
6. **Chatterjee, S., & Hadi, A. S.** (2006). *Regression analysis by example*. John Wiley & Sons, NY, USA. pp 164.
7. **Fuller, W. A., & Rao, J. N. K.** (1978). *Estimation for a linear regression model with unknown diagonal covariance matrix*. Annals of Statistics, **6**(5), 1149-1158.
8. **Ghazali, Z. M., Halim, M. S., & Jamidin, J. N.** (2017). *The performance comparison of two-step robust weighted least squares (TSRWLS) with different robust's weight functions.* International Journal of Advanced and Applied Sciences, 4(5), 44-47.
9. **Imon, R. A. H. M. (2009).** *Deletion residuals in the detection of heterogeneity of variances in linear regression*. Journal of Applied Statistics, **36**(3): 347-358.
10. **Midi, H., Rana, S. & Imon, A. H. M. R.** (2009). *The performance of robust weighted least squares in the presence of outliers and heteroscedastic errors*. WSEAS Transactions on Mathematics, **7**(8): 351-361.
11. **Midi, H., Rana, S. & Imon, A. H. M. R.** (2009). *Robust Estimation of Regression Parameters with Heteroscedastic Errors in the Presence of Outliers*. WSEAS Mathematics and Computers in Science and Engineering, (8). 128 – 134.
12. **Midi, H., S. Rana, & Imon, A. H. M. R.** (2013). *On a robust estimator in heteroscedastic regression model in the presence of outliers.* In Proceedings of the world congress on engineering, p: 280-285.
13. **Turkan, S., & Toktamis, O.** (2013). *Detection of influential observations in semi-parametric regression model*. Colombian Journal of Statistics, **36**(2): 271-284.
14. **White, H.** (1980). *A heteroscedasticity-consistent covariance matrix estimator and a direct test for heteroscedasticity*. Econometrica, **48**(4): 817–838.