Jahangirnagar University Journal of Statistical Studies
ISSN 1022-4734

http://www.juniv.edu/department/stat

JUJSS, Vol. 36, 2022, pp. 211-226

Model Selection and Testing Regression Coefficients for Contaminated Data

Imran Hossain Sumon*
Department of Statistics, Jahangirnagar University, Savar, Dhaka-1342, Bangladesh

Ajit Kumar Majumder
Department of Statistics, Jahangirnagar University, Savar, Dhaka-1342, Bangladesh

Abstract

Statistical analysis is better when the data are truly representative. However, extraneous stuff such as contamination, irrelevant data, missing observation, distributional faulty in the data cause inaccurate estimation of model parameters. In case of regression analysis, if we remove the extraneous stuff of such data, then we can lose necessary information. On the other hand, traditional estimation and classical test of regression coefficients may fail completely if we retain such type of data. This study aimed to provide an appropriate choice of estimation method, model selection criteria and introducing appropriate testing procedure for regression coefficient in case of contaminated data. Results of this study established that for contaminated and skewed data, robust approach outperforms classical approach to accurate estimation of parameters and identified that outliers are responsible for the inflated residual sum of squares that results in an incorrect solution of model selection and the test of regression coefficients. Findings from this study suggested that we can use the weighted residual sum of squares instead of the residual sum of squares to avoid such an inflated problem.

Key Words: Outliers; Skewness; Robust regression; Model selection; Testing Coefficients

1. Introduction

Regression analysis is the most widely used statistical technique for fitting models to real-life data and is regularly applied to most sciences. Regression analysis goes forward with several basic steps including model building, model selection, and hypothesis testing. The hypothesis test ensures whether underlying assumptions and methods of estimation are suitable for data or not. The fitted model is used to predict the future behavior of variables. However, before going to forecast by a model, we have to justify whether this model is adequate or not. In regression analysis, one of the most important tasks is to control extreme observations (outliers) because they are sensible for certain types of models. For example, in the presence of outliers and skewed data, the ordinary least square (OLS) is unable to produce robust estimates. To fix this problem, two methods are suggested in the literature (Rousseeuw & Leroy, 1987). The first one is regression diagnostics (Rousseeuw & Leroy, 1987), however, in the presence of multiple outliers, it is unfortunately much more difficult to diagnose them. The other approach is robust regression (Rousseeuw & Leroy, 1987). Because in the case of outliers and contaminated data, some of the OLS assumptions (e. g., error follows the normal distribution with constant variance) are violated, and for that reason, OLS does not provide appropriate result. In OLS, we

* Corresponding author: imranhossainsumon38@gmail.com

minimize the residual sum of squares that is extremely affected by outliers and contaminated data. Moreover, previous studies pointed out that in the presence outlier, weighted (robust) mean (Hossain, 2016) and variance (Hossain, 2017) provide more accurate estimates than their classical version. For that reason, we should use the robust regression method such as LAD, LMS, and M-estimation to obtain better results in such situations.

Model selection is another important part of regression analysis. In this procedure, firstly, we must fit several models and then among them, we select the most appropriate model by using a numerical summary of their goodness-of-fit, properties or combination of both for prediction (Määttä et al., 2016). Sequential testing, allowing variables to be added or deleted at each step, has often been employed. However, such p-value-based testing techniques only evaluated two nested models and have been widely criticized, as hypothesis tests are a poor basis for model selection in general. Cross-validation and its variation have been suggested and discussed as a useful model selection method (Akaike, 1974). The adjusted coefficient and Mallow's C_p statistic is also widely used in classical regression analysis and provide a ranking for all considerable models. Moreover, several model selection criterion, such as R^2 Criterion, Root mean square deviation $(RMSD)$, Adjusted R^2, Criterion Akaike information criterion (AIC), Akaike information criterion with small sample correction $(AICc)$, Bayesian information criterion (BIC), Schwartz's information criterion (SIC), Hannan-Quinn information criterion (HQC), Mean absolute error (MAE) and their variations will also be considered (Draper & Smith, 1981).

The alternative to OLS regression in the case of outliers is robust regression. Robust regression modeling has been applied in several studies. The introduction of Least Median of Squares approach comes from Rousseeuw (1984). In order to obtain LAD estimators, Charnes et al. (1954) addressed linear programming. Portnoy & Koenker (1997) provide a comprehensive summary of the LAD approach. Koenker & Bassett (1978) and Pollard (1991) obtained large sample properties for the LAD estimates. Alma (2011) concluded that the S-estimator would increase its effectiveness in case of 10% breakdown (Almetwally & Almongy, 2018). MM-estimation works best in contrast to a wide set of extrinsic condition (Almetwally & Almongy, 2018). From the descriptive point of view, we can say that the criteria and usual hypothesis testing for ordinary square regression may fail miserably even for a large sample. Khan & Majumder, (2012) introduced weighted model selection criteria to identify the best-fitted model under the OLS, LAD, and LMS regression in case of contaminated data. However, to the author's knowledge, no scholarly article has been developed on hypothesis testing in the case of the contaminated and skewed data under these estimation approaches. Therefore, this study attempts to fulfill this methodological gap. This research aims to compare the performance of OLS, LAD, and LMS methods and test the regression coefficients by using the proposed WRSS based test statistics in case of contaminated data.

2. Methods and Materials

Most well-known regression estimator is the ordinary least squares (OLS) estimator, $\hat{\beta}_{LS}$ which is defined as the vector that minimizes residual sum of squares (Gujarati, 2004).

$$SSE = \sum_{i=1}^{n} \left(y_i - \hat{y}_i \right)^2 \tag{2.1}$$

But the major shortcoming of the traditional approaches to linear regression has always been the fact that in the regression analysis one assumes that the errors are truly random or accidental. Blunders, clerical errors, misprints in published or electronically stored data are simply ignored in the analysis (Giloni & Padberg, 2002). "Whereas", this type of data has a large influence on the classical approach. The classical approach is also highly influenced by outliers and contaminated data. Sometimes outliers are the result of a systematic problem with either our data collection techniques or our model (McCann, 2005). The outliers occurring with extreme values of the regressor variables can be especially disruptive (Thanoon, 2015). Contamination can, of course, take many forms. It may be recording or reading errors. In this case, correction or rejection might be the only possibility. Alternatively, it might reflect low incidence mixing of x with another random variable whose source and manifesto are uninteresting. For this problem, McCann (2005) proposed a well-earned approach for dealing with outliers and contamination. These are the robust (or resistant) methods such as least median square (LMS), least absolute deviation (LAD), least trimmed squares (LTS), M-estimation, S-estimation etc.

LAD represents the method of solving an over-defined system of (linear) algebraic equations mentioned by Taylor (1974) is related to KF Gauss and PS Laplace. Mathematicians suggest to minimizes the sum of absolute residuals in the equations (Dasgupta & Mishra, 2011).

$$\underset{\hat{\beta}}{minimize} \sum_{i=1}^{n} \left| \varepsilon_i \right| \tag{2.2}$$

This is also called L_1 regression (or L_1 - norm regression) whereas least square is L_2 (Draper & Smith, 1981). The least squares regression is very far from the optimal in many non-Gaussian situations, especially when the errors follow distributions with longer tails (Thanoon, 2015). Unlike the *LS* method, the *LAD* method is not sensitive to outliers and produces robust estimates (Chen et al., 2008).

A different approach, least median squares *(LMS)* is introduced which minimizes the median of the squared residuals (Rousseeuw, 1984). That is replacing "sum" in OLS by median yield the *LMS* estimator of the parameter.

The least median of squares *(LMS)* estimator minimizes the objective function,

$$\underset{\hat{\beta}}{minimize}\ \underset{i}{med}\left(y_i-\hat{y}_i\right)^2=\underset{\hat{\beta}}{minimize}\ \underset{i}{med}\,\varepsilon_i^{\,2} \tag{2.3}$$

ere, $\varepsilon_i^{\,2}\left(i=1,2,\ldots,n\right)$ are the residual squares. The solution is that β that produces the minimum such median (Draper & Smith, 1981). In the case of least squares, the notion of the breakdown point $\grave{o}^*$ is zero. That is,

$$\grave{o}^*=0$$

But whereas the breakdown point of the univariate median is as high as 50% (Hampel, 1971).

2.1 Weighted Model Selection Criteria

We know that the residual sum of squares is used in all the classical model selection criteria though RSS is very much affected by outlier. To this problem, we have to use the weighted residual sum of squares s (WRSS) instead of RSS, which is given by

$$WRSS=\sum_{i=1}^{n}w_i(y_i-\hat{y}_i)^2 \tag{2.4}$$

Where,

w_i is the weight for i^{th} observation

y_i denotes the true value for i^{th} trial of the regressand variable,

$\hat{y}_i$ is the predicted value for i^{th} trial of the regressand variable.

Now w_i can be computed as

$$w_i=\begin{cases}1 & if\ \left|\dfrac{\varepsilon_i}{s^0}\right|<2.5\\ 0 & otherwise\end{cases}$$

(Rousseeuw, 1987, pp-44)

And the scale estimate is denoted by s^0 is defined as

$$s^0=1.4826\left(1+\frac{5}{n-p}\right)\sqrt{\underset{i}{med}\ \varepsilon_i^{\,2}}$$

(Rousseeuw, 1987, pp-44)

We propose to use the weighted residual sum of squares (WRSS) in place of residual sum of squares (RSS) in all the classical model selection criteria to obtain better fitted model in case of contaminated data (Khan & Majumder, 2012).

2.2 Proposed Test for Overall Regression Coefficients

In regression analysis firstly performing test is often a test of whether the regressor variables have any significant effect on the dependent variable. Let us consider the general form of the linear regression model

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon \tag{2.5}$$

Suppose we want to test the following hypothesis

$H_0 : \beta_1 = \beta_2 = \ldots = \beta_p$

H_1 : At least one β_j 's is not equal to zero, $j = 1,2,\ldots, p$

Under the null hypothesis the reduced model becomes

$$y = \beta_0 + \varepsilon^*$$

Now the proposed test statistics is,

$$F = \frac{\left(WSSR_{reduced} - WSSR_{full}\right) \Big/ (p-1)}{WSSR_{full} \Big/ (n-p)} \tag{2.6}$$

Assuming the normality of the distribution of random errors F has an F distribution with numerator and de-numerator degrees of freedom $p-1$ and $n-p$ respectively. Where, k is the number of explanatory variables including intercept.

At $\alpha\%$ level of significance the critical value of F with numerator and de-numerator degrees of freedom $p-1$ and $n-p$ respectively is $F_c = F_{p-1, n-p}\left(\alpha\right)$. If the calculated value of F is greater than the critical value of F, then we may reject the null hypothesis. That is, there are significant effects of the explanatory variables on the dependent variable. Otherwise, we may not reject the null hypothesis.

2.3 Proposed Test for Individual Regression Coefficient

Let us consider the general linear regression of equation (2.5). Consider that the weighted residual-based F test provides information that there are significant effects of the explanatory variables on the dependent variable (i. e., null rejected). Then we eager to specify the explanatory variable for which there is a significant effect on the dependent variable.

Now let us consider the following hypothesis

$$H_0 : \beta_j = 0 \qquad \text{against}$$
$$H_1 : \beta_j > 0 , \ j = 1,2,\ldots, p$$

The test statistics under the null hypothesis is given by

$$\left| t_j \right| = \frac{\left| \hat{\beta}_j \right|}{SD\left(\hat{\beta}_j \right)} \tag{2.7}$$

To obtain the test statistics, we have to calculate an estimate of the standard deviation of $\hat{\beta}_j$ by substituting $\hat{\sigma}$ for σ by using the following formula,

$$SD\left(\hat{\beta}_j \right) = \frac{\hat{\sigma}}{\sqrt{\sum_{i=1}^{n} \left(x_{ij} - \bar{x}_{.j} \right)^2}}$$

Where,

$$\hat{\sigma} = \sqrt{\frac{WSSR}{n - p - 1}}$$

x_{ij} is the value of the i^{th} observation for j^{th} explanatory variable

$\bar{x}_{.j}$ is the mean value of j^{th} explanatory variable

If we specify the shape to be normal (bell-shaped), that is, if we assume the normal linear regression model, then the resulting distribution of the test statistics t is still close to the t distribution with $n-2$ degrees of freedom. Assuming the normality of the error term, the critical value of t at $\alpha\%$ level of significance is $t_c = t_{(n-2)}\left(\alpha / 2 \right)$. If the calculated value of t_j is greater than the t_c, then we may reject the null hypothesis. That is, there is significant effect of the explanatory variable x_j on the dependent variable. Otherwise, we may not reject the null hypothesis.

2.4 Data Source

This paper uses GNP data from Table B-1, p. 232; third measure money stock data from Table B-61, p. 303 of the economic report of the president (1985) for scrutinizing the performance of different estimations methods in simple regression by proposed model selection criteria. The explanatory variable represents the third measure of money stock data, and the response variable is gross national product (GNP). The data was also used by Gujarati (2004), and others.

3. Results

Consider four different case to make comparison of the methods OLS, LMS and LAD. In first case, the comparison is made in simple regression for mentioned data without outlier. Afterward a single outlier is taken in the response variable Gross national product ($i.e.$, in y direction) and presented in this section as second case. Single outlier is taken in the explanatory variable money stock measure ($i.e.$, in x direction) and presented in this section as third case. Outliers are taken in both the dependent variable ($i.e.$, in y direction) and independent variable ($i.e.$, in x direction) in the fourth case. The results obtained in each case are presented below:

Uncontaminated Data

The classical regression estimation procedure ($i.e.$, OLS) is as usual. On the other hand, the objective functions of LMS and LAD methods are not straightforward that's why the estimation is not unique. Thus, an iterative procedure is adopted in estimating the fitted models for both methods. The fitted models by the method of OLS, LAD and LMS are respectively given in the next:

$$\hat{y}_{OLS} = 158.788 + 1.204X \tag{3.1}$$

$$\hat{y}_{LAD} = 140.968 + 1.207X \tag{3.2}$$

$$\hat{y}_{LMS} = 169.737 + 1.181X \tag{3.3}$$

From the figure 1, we observe that the OLS, LAD, and LMS methods give almost the same fitted line for the uncontaminated data. Also, classical model selection criteria (see Table 1 in appendix) show that all three methods perform very closely in order to fit the approximate model. Although the result of the model selection criterion is almost same, but all the criteria suggest that OLS is better than other methods in case of uncontaminated data. Weighted model selection criteria (see Table 1 in Appendix) shows that LAD is better than other methods. Classical F-test (see Table 1 in Appendix) and proposed test (see Table 1 in Appendix) indicate that linear regression model equation 3.1, 3.2, 3.3 provides a better fit to the data than a model that contains no independent variables. Both the classical t-test (see Table 1 in Appendix) and proposed t-test (see Table 1 in Appendix) indicate that money stock data have a significant association with GNP. Whereas, intercept coefficient is also significant

Contamination in y-Direction

In this section, a single outlier is taken in GNP of the same data used in immediate previous case. The fitted models by the method of OLS, LAD and LMS are respectively given in the next:

$$\hat{y}_{OLS} = -7585.5 + 7.718X \tag{3.4}$$

$$\hat{y}_{LAD} = 40.0467 + 1.302X \tag{3.5}$$

$$\hat{y}_{LMS} = 169.737 + 1.181X \tag{3.6}$$

All the fitted models from equation 3.1 to equation 3.3 indicate that the intercept is positive. But if we consider outlier in the y-direction of the data then OLS fitted model shows a negative intercept (see equation 3.4), which designates a failure of the OLS method to estimate the actual model. Whereas, the LMS and LAD fitted models computed the actual signs of the relationship.

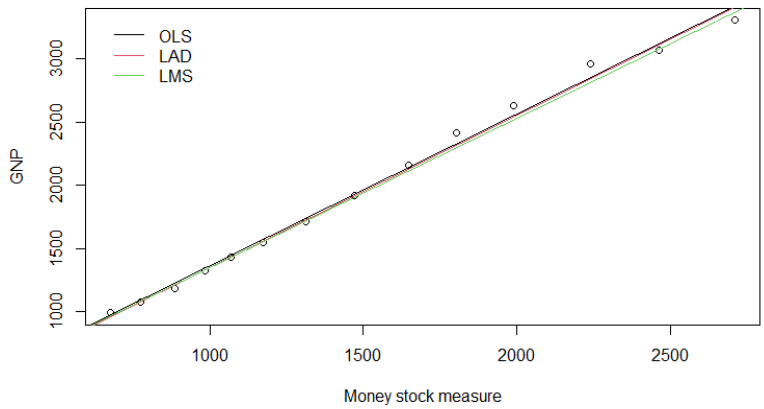

Figure 1. The Observed and Fitted y against Observed x for Uncontaminated Data

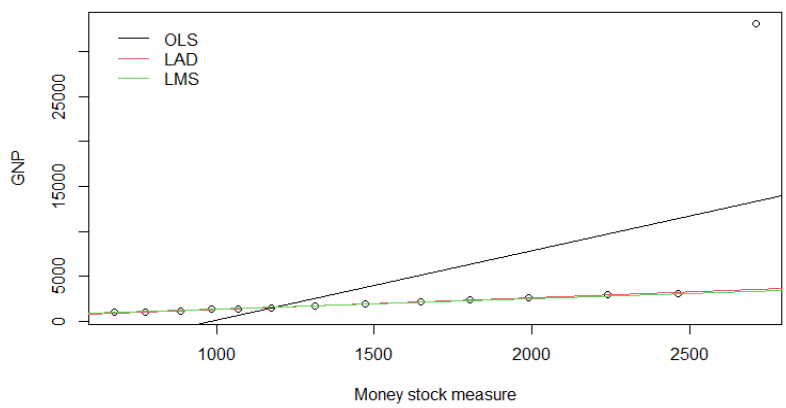

Figure 2. The Observed and Fitted y against Observed x for Contaminated Data in y-Direction

Graphical presentation of the fitted GNP against Money stock measure ($i.e.$, presented in equation 3.4 through equation 3.6) is exhibited in the Figure 2. Figure 2 reflects that the LMS and LAD methods yield a good fit for contaminated data in the y-direction. On the other hand, the OLS method has been affected by the outlier data, and therefore the intercept is so much different from the original one (see Figure 1). But all the classical model selection criteria (see Table 2 in Appendix) indicate that the estimated OLS model is better than the estimated LAD and LMS models, which is contradictory to the graphical presentation. On the other hand, examining all the weighted model selection criteria (see Table 2 in appendix) reveals that the fitted LMS and LAD models are better than the fitted OLS model which is exactly similar to the graphical representation and the actual relationship that exist between GNP and money stock measure. It is also revealed that LAD method is better than LMS method. Classical F-test (see Table 2 in Appendix) unveils that fitted OLS line (see equation 4.4) is better than a model that contains no independent variables whereas the conclusion is different under LAD and LMS approaches. The proposed weighted F test (see Table 2 in Appendix) reflects that fitted LAD and LMS provide a better fit than a model with the only intercept but not for the fitted OLS line. The classical t-test (see Table 2 in appendix) indicates that there is no significant association between money stock measure and GNP under the LAD and LMS methods but there is a significant association between money stock measure and GNP under the OLS method whereas the intercept coefficient is not significant for all the method. The proposed weighted t-test (see Table 2 in appendix) unveils that there is a significant association between money stock measure and GNP under all methods whereas the intercept coefficient is significant for LAD and LMS methods but not for the OLS method.

Contamination in x-direction

In this section, a single outlier is taken in money stock measure of the same data used in uncontaminated data. The fitted models by the method of OLS, LAD and LMS are respectively given in the next:

$$\hat{y}_{OLS} = 1777.10 + 0.0630X \tag{3.7}$$

$$\hat{y}_{LAD} = 1637.28 + 0.0615X \tag{3.8}$$

$$\hat{y}_{LMS} = 169.737 + 1.181X \tag{3.9}$$

Figure 3. The Observed and Fitted y against Observed x for Contaminated Data in x-Direction

Figure 3 shows that the LMS method yield a good fit than other two methods for contaminated data in the x-direction. Fitted models (see equation 3.7 and 3.8) for the contaminated data in x-direction under OLS and LAD methods are different from the models (see equation 3.1 and 3.2) for original data whereas the fitted model (see equation 3.9) for the contaminated data in x-direction under LMS methods are different from the model (see equation 3.3) for original data. It is evident from the classical model selection criteria (see Table 3 in Appendix) that OLS method is better compare to LAD and LMS methods for the contaminated data in X -direction which is aberrant to the graphical presentation. That signify the classical model selection criteria fail to identify the best method that fit to the majority of the data. According to all the weighted model selection criteria (see Table 3 in Appendix) LMS method is better than the OLS and LAD methods. Classical F-test unveils that the fitted model under only the OLS method is better than the model with the only intercept coefficient. According to the proposed weighted F-test fitted model under the all the methods is better than the model with the only intercept coefficient. The classical t-test (see Table 3 in appendix) indicates that there is no significant association between money stock measure and GNP under the LAD and LMS methods but there is a significant association between money stock measure and GNP under the OLS method whereas the intercept coefficient is not significant for LMS method. The proposed weighted t-test (see Table 3 in appendix) unveils that there is a significant association between money stock measure and GNP under all methods whereas the intercept coefficient is also significant for all methods.

Contaminated Data both in X-Direction and y-Direction

The fitted models by OLS, LMS and LAD method after considering the outliers in both the x-direction and y- direction are as like as:

$$\hat{y}_{OLS} = 2937.88 + 0.2040X \tag{3.10}$$

$$\hat{y}_{LAD} = 2090.22 + 0.0447X \tag{3.11}$$

$$\hat{y}_{LMS} = 31.933 + 1.302X \tag{3.12}$$

Figure 4 reflects that LMS provide better fit than LAD and OLS. By examining all the classical model selection criteria (see Table 4 in Appendix), it may conclude that the OLS method is better than the LMS and LAD method which is aberrant to the graphical presentation. It signifies that the classical model selection criteria fail to identify the best-fitted model for contaminated data in both directions. On the other hand, all the weighted model selection criteria (see Table 4 in appendix) indicate that the LMS fitted line are better than the OLS and LAD fitted lines which is exactly similar to the graphical representation. Classical F-test (see Table 4 in appendix) unveils that the model with the only intercept coefficient is better than the fitted model under the OLS, LAD, and LMS methods. According to the proposed weighted F-test (see Table 4 in appendix), the fitted model under the OLS and LMS methods is better than the model with the only intercept coefficient. The proposed weighted t-test (see Table 4 in appendix) unveils that there is a significant association between money stock measure and GNP under OLS and LMS methods whereas the intercept coefficient is significant for all methods.

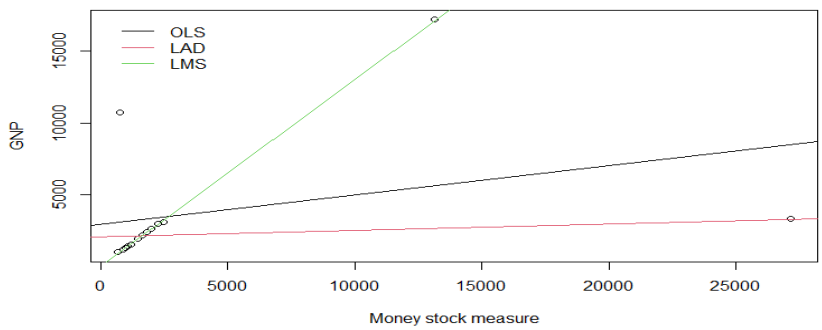

Figure 4. The Observed and Fitted y against Observed x for Contaminated Data in both x-Direction and y-Direction

5. Conclusion

This paper develops weighted test statistics in order to test the regression coefficients and applied the existing weighted model selection criteria for the purpose of comparison of the different estimation methods of the linear regression and it also develops weighted test statistics in order to test the regression coefficients. Analysis tells us, the weighted model selection criteria based on WRSS perform

well for contaminated data, while the usual model selection criteria fail to identify the best method in fitting the regression model. Another important fact is that weighted model selection criteria fail to identify the best-fitted regression model in the case of uncontaminated data. According to weighted model selection criteria, the LMS method gives a better fit than other two methods for contaminated data in x-direction and y-direction. It is also palpable from simple regression for contaminated data both in X-direction and y-direction that the LMS method is the best among the three compared methods. It is evident from the analysis part is that the weighted test statistics based on WRSS perform well for contaminated data, while the usual test statistics fail to test regression coefficients.

References

Akaike, H. (1974). A New Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723. https://doi.org/10.1109/TAC.1974.1100705.

Alma, Ö. G. (2011). Comparison of Robust Regression Methods in Linear Regression. *Int. J. Contemp. Math. Sciences*, *6*(9), 409–421.

Almetwally, E. M., & Almongy, H. M. (2018). Comparison Between M-estimation , S-estimation , And MM Estimation. *International Journal of Mathematics Archive*, *9*(11).

Charnes, A., Cooper, W. W., & Henderson, A. (1954). *An introduction to linear programming*. John Wiley & Sons.

Chen, K., Ying, Z., Zhang, H., & Zhao, L. (2008). Analysis of least absolute deviation. *Biometrika*, *95*(1), 107–122. https://doi.org/10.1093/biomet/asm082.

Dasgupta, M., & Mishra, S. K. (2011). Least Absolute Deviation Estimation of Linear Econometric Models: A Literature Review. *SSRN Electronic Journal*, *July 2004*. https://doi.org/10.2139/ssrn.552502.

Draper, N. R., & Smith, H. (1981). *Applied Regression Analysis* (2nd ed.). John Wiley & Sons.

Giloni, A., & Padberg, M. (2002). Least trimmed squares regression, least median squares regression, and mathematical programming. *Mathematical and Computer Modelling*, *35*(9–10), 1043–1060. https://doi.org/10.1016/S0895-7177(02)00069-9.

Gujarati, D. N. (2004). *Basic Econometrics* (4th ed.). McGraw Hill.

Hampel, F. R. (1971). A General Qualitative Definition of Robustness. *The Annals of Mathematical Statistics*, *42*(6), 1887–1896. https://doi.org/10.1214/aoms/1177693054.

Hossain, M. M. (2016). Proposed Mean (Robust) in the Presence of Outlier. *Journal of Statistics Applications & Probability Letters*, *3*(3), 103–107. https://doi.org/10.18576/jsapl/030301.

Hossain, M. M. (2017). Variance in the Presence of Outlier: Weighted Variance. *Journal of Statistics Applications & Probability Letters*, *4*(2), 57–59. https://doi.org/10.18576/jsapl/040203.

Khan, M. T. F., & Majumder, A. K. (2012). Comparing the Estimation Methods of OLS , LMS and LAD by Proposed Model Selection Criteria for Contaminated Data. *IJASETR*, *1*(4), 9–18.

Koenker, R., & Bassett, G. (1978). Regression quantiles Econometrical. In *Econometrica* (Vol. 46, Issue 1, pp. 33–50).

Määttä, J., Schmidt, D. F., & Roos, T. (2016). Subset Selection in Linear Regression using Sequentially Normalized Least Squares: Asymptotic Theory. *Scandinavian Journal of Statistics*, *43*(2), 382–395. https://doi.org/10.1111/sjos.12181.

McCann, L. (2005). Robust Model Selection and Outlier Detection in Linear Regression by. *Journal of the American Statistical Association*, *100*(472), 1297–1310.

Pollard, D. (1991). Asymptotics for Least Absolute Deviation Regression. *Cambridge University Press*, *7*(2), 186–199.

Portnoy, S., & Koenker, R. (1997). The gaussian hare and the laplacian tortoise: Computability of squared-error versus absolute-error estimators. *Statistical Science*, *12*(4), 279–296. https://doi.org/10.1214/ss/1030037960.

Rousseeuw, P. J. (1984). *Least Median of Squares Regression*.

Rousseeuw, P. J., & Leroy, A. M. (1987). *Robust regression and outlier detection*. Wiley.

Thanoon, F. H. (2015). *Robust Regression by Least Absolute Deviations Method*. *5*(3), 109–112.

Appendix

Table 1. Classical and Weighted model selection criteria of simple linear regression for uncontaminated data

Criterion	Classical	Weighted	Classical	Weighted	Classical	Weighted
	OLS		LAD		LMS	
RMSD	56.816	13.815	58.341	7.1154	62.915	11.262
AIC	4295.6	253.969	4529.3	67.373	5267.32	168.787
AICc	4296.6	255.059	4530.4	68.464	5268.41	169.878
SIC	4706.2	278.246	4962.3	73.813	5770.83	184.921
R^2	0.9943	0.9997	0.9940	0.9999	0.9930	0.9997
Adjusted R^2	0.9938	0.9996	0.9935	0.9999	0.9924	0.9997
MAE	44.417	10.240	40.533	3.108	41.619	6.8023
FPE	4304.0	234.894	4538.3	67.505	4871.72	169.119
Hypothesis testing by using classical and weighted test F statistics						
F **-statistics (p-value)**	2099.17 (<0.001)	22632.8 (<0.001)	2007.72 (<0.001)	19266.5 (<0.001)	2900.07 (<0.001)	27860.6 (<0.001)
P value of the corresponding coefficient by using classical and weighted test t statistics						
p_0	0.0031	(<0.001	0.0132	(<0.001	0.0014	(<0.001
p_1	<0.001	(<0.001	(<0.001	(<0.001	(<0.001	(<0.001

Table 2. Classical and Weighted model selection criteria of simple linear regression for contaminated data in y-direction

Criterion	Classical	Weighted	Classical	Weighted	Classical	Weighted
	OLS		LAD		LMS	
RMSD	6450.89	203.34	7876.62	54.42	7929.82	60.37
AIC	55376364	55022	82558825	3941.5	83677806	4850.7
AICc	55376365	55023	82558826	3942.6	83677807	4851.8
SIC	60669844	60282	90450702	4318.4	91676648	5314.3
R^2	0.35816	0.9994	0.0431	1	0.0301	0.9999

Adjusted R^2	0.3047	0.9993	0.0000	1	0.0000	0.9999
MAE	4257.216	59.247	2135.66	30.59	2156.16	36.89
FPE	55485440	50890	82721443	3949.4	77393195	4860.2
Hypothesis testing by using classical and weighted test F statistics						
***F* -statistics (p-value)**	6.6961 (0.0238)	1.3199 (0.2730)	1.4719 (0.2483)	14794 (<0.001)	1.8166 (0.2026)	27861 (<0.001)
P value of the corresponding coefficient by using classical and weighted test t statistics						
p_0	0.1466	0.2403	0.9893	(<0.001	0.7867	(<0.001
p_1	0.0238	(<0.001	0.4763	(<0.001	0.5442	(<0.001

Table 3. Classical and Weighted model selection criteria of simple linear regression for contaminated data in x-direction

Criterion	**Classical**	**Weighted**	**Classical**	**Weighted**	**Classical**	**Weighted**
	OLS		**LAD**		**LMS**	
RMSD	234.594	127.036	643.443	97.938	7716.98	60.375
AIC	57280.1	21475.1	550939.3	12764.1	79246201	4850.7
AICc	57280.2	21476.2	550940.4	12765.2	79246202	4851.8
SIC	60343.7	23527.9	603604.2	13984.2	86821421	5314.3
R^2	0.3078	0.9716	0.2709	0.9831	0.0000	0.9935
Adjusted R^2	0.3007	0.9692	0.2102	0.9817	0.0000	0.9930
MAE	75.33	68.207	515.77	44.262	2099.276	36.891
FPE	57280.4	19862.2	552024.5	12789.2	73294425	4860.2
Hypothesis testing by using classical and weighted test F statistics						
***F* -statistics (p-value)**	5.336 (0.0395)	10.664 (0.0068)	4.604 (0.0530)	14794 (<0.001)	4.734 (0.0513)	27861 (<0.001)
P value of the corresponding coefficient by using classical and weighted test t statistics						
p_0	(<0.001	(<0.001	(<0.001	(<0.001	0.0674	(<0.001
p_1	0.0395	0.0215	0.1387	0.0078	0.1534	(<0.001

Table 4. Classical and Weighted model selection criteria of simple linear regression for contaminated data in both x and y direction

Criterion	Classical	Weighted	Classical	Weighted	Classical	Weighted
	OLS		LAD		LMS	
RMSD	1549.92	1241.39	617.00	97.938	8952.34	56.233
AIC	2500288	2050707	506593	12764.1	106649118	4207.91
AICc	2500289	2050708	506594	12765.2	106649118	4209.0
SIC	2634016	2246736	555019	13984.2	116843808	4610.15
R^2	0.1087	0.9199	0.9802	0.9831	0.0002	0.9998
Adjusted R^2	0.0993	0.9132	0.9786	0.9817	0.0000	0.9998
MAE	383.73	1001.15	482.19	44.262	3014.235	33.063
FPE	2500302	1896689	507590	12789.2	98639249	4216.20
Hypothesis testing by using classical and weighted test F statistics						
F **-statistics (p-value)**	1.4585 (0.2504)	11.052 (0.0614)	0.2353 (0.6364)	6.7251 (<0.2357)	0. 3961 (0.0513)	7482.53 (<0.001)
P value of the corresponding coefficient by using classical and weighted test t statistics						
p_0	0.055	(<0.001	0.1125	(<0.001	0.1867	(<0.001
p_1	0.250	(<0.001	0.7810	0.1007	0.8442	(<0.001