Jagannath University Journal of Science
Volume 07, Number II, June 2021, pp. 1–6
https://jnu.ac.bd/journal/portal/archives/science.jsp
ISSN 2224-1698

JnUJSci

Simple Linear Regression Model Using Different Slope Estimation Methods: A Simulation Study and Real Data Applications

Research Article

Mst. Romana Akter, Md Siddiqur Rahman*

Department of Statistics, Jagannath University, Dhaka-1100, Bangladesh

Received: 04 November 2020
Accepted: 20 January 2021

Abstract: There are many slope estimation methods in simple linear regression. The classical simple linear regression model constructed by the Least Squares (LS) method is better whenever the model's basic assumptions are entirely met. The LS method yields unsatisfactory results when data contain outliers and other contaminations. So we need a parameter estimation method which is robust and efficient. This study presents MM estimation, Least Median of Squares (LMS), and the Repeated Median (RM) methods in robust regression to determine the regression model. The performances of robust regression methods and the LS method are compared through a simulation study and a real data application for clean and contaminated data. It has been observed that robust regression methods have much better performance compared to the LS method. Among the existing robust regression methods, the performance of MM estimation is much better than any other robust methods.

Keywords: *Least Squares method • MM estimation • Least Median of Squares • Repeated Median method • Outlier*

1. Introduction

Sir Francis Galton first promoted regression analysis in the latter part of the 19th century. Galton had prepared the relation between heights of parents and children and famed that the heights of children of both tall and short parents revealed to change back or regress to the mean of the group. He developed a mathematical illustration of this tendency, the harbinger of today's regression models (Neter *et al.*, 1996). Zioutas observed linear regression models that are commonly used to explore data from multiple fields of study (Zioutas *et al.*, 2005).

LS method is often used to enumerate the slope and intercept of the best line through a set of data points. However, LS regression intercepts and slopes may be mistaken if the LS model's underlying assumptions are not met. Two factors in special that may result in incorrect LS regression coefficients are: (a) imprecision in the assessment of the independent (x-axis) variable and (b) insertion of outliers in the data analysis (Cornbleet and Gochman,1979). LS regression postulates an error-free x variable and a constant analytical imprecision of the y variable (also called "homoscedastic" variance), both of which are rarely met in practice (Stöckl *et al.*, 1998).

An outlier is an observation that appears to digress markedly from the other sample members in which it happens. Such extreme observations may be revealing some abnormality in the measured characteristics of a subject, or they may result from an error in the measurement or recording (Everitt, 2002).

LS method is facile for calculation, but it is sensitive to outliers. Hampel initiated alternative methods to LS, known as "Robust Regression" (Hampel, 2002). Robust regressions are needed because they can provide reliable results in the presence of outliers. This method is a momentous tool for analyzing the data, which is influenced by outliers. The resulting models are stout against outliers (Draper and Smith, 1998). There are many robust regression procedures

*Corresponding author: Md Siddiqur Rahman
E-mail: rsiddiq11@stat.jnu.ac.bd

estimation is the improvement of the M-estimation method (Susanti *et al.*, 2014), LMS estimates reduce the median of squared residuals (Siegel 1982). The RM method can still give good estimators when 50% of the data are contaminated by outliers (Siegel 1982).

2. Theory and Methodology

2.1 Method of Least Squares (LS)

Consider the simple linear regression model:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1,2,\dots,n \tag{1}$$

where y_i is the response variable in the *i*th trial, β_0 (intercept) and β_1 (slope) are parameters. x_i is a known constant, namely; the value of the predictor variable in the *i*th trial. ε_i is a random error term with mean zero and variance σ^2. The LS standard asserts that one figure out the sum of n squared deviations; this standard is denoted by $S(\beta_0, \beta_1)$.

$$S(\beta_0,\beta_1) = \sum_{i=1}^{n} \varepsilon_i{}^2 = \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2. \tag{2}$$

According to the LS method, the estimators of β_0 (intercept) and β_1(slope) are those values b_0, and b_1 respectively, that minimize the criterion $S(\beta_0, \beta_1)$ for the given sample observations $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ using the analytical approach. It can be shown that the estimated values of β_0 and β_1 are

$$b_1 = \frac{S_{xy}}{S_{xx}} \tag{3}$$

and

$$b_0 = \overline{y} - b_1\overline{x} \tag{4}$$

where

$$S_{xx} = \sum_{i=1}^{n}(x_i - \overline{x})^2 \text{ and}$$
$$S_{xy} = \sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}).$$

2.2 MM Estimator

MM estimation is an exceptional type of M-estimation (Yohai, 1987). It aims to find estimators that have a high breakdown value and more efficient. Breakdown value is a general measure of the proportion of outliers that can be addressed before these observations affect the model (Chen, 2002). The MM-estimates can be appeared by a three-stage procedure. In the first stage, compute a basic consistent estimator $\hat{\hat{\beta}}_0$ with high Breakdown Point (*BP*) but probably low normal efficiency. In the second stage, compute a strong M-estimator of scale $\hat{\sigma}$ of the residuals based on the initial estimator. In the third stage, find an M-estimator $\hat{\hat{\beta}}$ starting at $\hat{\hat{\beta}}_0$.

In practice, LMS or S-estimate with Huber or bi-square function is typically conducted as the initial estimator $\hat{\hat{\beta}}_0$. Let

$$\rho_0(r) = \rho_1(r/k_0), \tag{5}$$

$$\rho(r) = \rho_1(r/k_1), \tag{6}$$

and assume that each of the ρ- functions is restricted. The scale estimator $\hat{\sigma}$ satisfies

$$\frac{1}{n}\sum_{i=1}^{n}\rho_0\left(\frac{r_i(\hat{\hat{\beta}}_0)}{\hat{\sigma}}\right) = 0.5. \tag{7}$$

If the ρ-function is bi-weight, then $k_0 = 1.56$ confirms that the estimator has the asymptotic *BP* = 0.5. Note that an M-estimator minimizes

$$L(\beta) = \sum_{i=1}^{n}\rho\left(\frac{r_i(\hat{\hat{\beta}})}{\hat{\sigma}}\right). \tag{8}$$

Let ρ satisfy $\rho \le \rho_0$. Yohai (1087) exposed that $\hat{\hat{\beta}}$ satisfies $\mathrm{L}\left(\hat{\hat{\beta}}\right) \le \mathrm{L}\left(\hat{\hat{\beta}}_0\right)$, then $\hat{\hat{\beta}}$'s *BP* is not less than that of $\hat{\hat{\beta}}_0$. Besides, the *BP* of the MM-estimator depends only on k_0 and the asymptotic variance of the MM estimator depends only on k_1. We can select k_1 to achieve the desired normal efficiency without affecting its *BP*. For the sake of $\rho \le \rho_0$, we ought have $k_1 \ge k_0$; the bigger the k_1 is, the excessive efficiency this estimator can attain at the normal distribution.

The values of $k_1$with the corresponding efficiencies of the bi-weight ρ-function are provided by Maronna *et al.*, (2006). Please see the succeeding table for more detail.

Efficiency:	0.80	0.85	0.90	0.95
k_1:	3.14	3.44	3.88	4.68

However, Yohai (1987) reveals that MM-estimators with larger values of k_1 are more sensitive to outliers than the estimators corresponding to smaller values of k_1.

In practice, an MM-estimator with bi-square function and efficiency 0.85 (k_1= 3.44) starting from a bi-square S-estimator is advised.

2.3 Least Median of Squares (LMS) Estimator

Another required robust regression method is the LMS (Rousseeuw and Leroy, 1987) and (Rousseeuw, P. J. 1984). This scheme was first commenced in analytical chemistry by Massart and co-workers (Massart *et al.*, 1986). The fundamental concept behind this method is to introduce the median as inference in the LS method. Rousseeuw (1984) proposed to replace the LS with the LMS. The LMS estimator is given by

$$\text{Min } {}^{med}_{i}(r_i)^2, \text{ for } i = 1,2,\dots,n, \tag{9}$$

where the median is the $\lfloor n/2 \rfloor$ +1 th ranked value. In

conformity with the literature, the *BP* of LMS is

$$\varepsilon = \lim_{n\to\infty} \frac{\left\lfloor \frac{n}{2} \right\rfloor}{n} = 0.5 \quad (10)$$

where $\lfloor n/2 \rfloor$ illustrates the integer part of $n/2$. The *BP* is, therefore 50%.

Repeated Median (RM) Method

Siegel (1982) flourished repeated median (RM) method which is highly robust for fitting a regression line with a *BP* of 50%. The slope and intercept were then defined as

$$\beta_1 = \underset{i}{med}\ \underset{j\neq i}{med} \frac{(y_j - y_i)}{(x_j - x_i)} \quad (11)$$

$$\beta_0 = \underset{i}{med}\ \underset{j\neq i}{med} \frac{(x_j y_i - x_i y_j)}{(x_j - x_i)}. \quad (12)$$

First of all, the median of the slopes is computed for $(n-1)$ pairs between a point i and all other points j $(j \neq i)$,i.e., med $\beta_1(i,j)$. This is put through for all points i $(i = 1, \cdots, n)$. This n medians are attained and the median of these n medians med $\{\text{med}\beta_1(i,j)\}$, is then the repeated median method.

3. Simulation Study

This study considers the estimation of regression coefficients in a simple linear regression model. A simulation study is carried out for comparing the performances of classical regression method (LS) and robust regression methods with each other.

3.1 Design of the Simulation

The simulation is carried out by the steps listed below**:**

1. Generate the independent variables $x_i (i = 1, 2, \ldots, 500)$ from a standard normal distribution $N(0,1)$.
2. Generate the errors $\varepsilon_i (i = 1,2, \ldots, 500)$ independently from a standard normal distribution $N(0,1)$.
3. Generate the response variables y_i based on the model**:**

$$y_i = 10 + 6x_i + \varepsilon_i. \quad (13)$$

So the true intercept value is 10 and true slope value is 6. The process is repeated 10,000 times to obtain 500 independent samples of y_i and x_i. Then datasets are contaminated by 5%, 10% and 15% x-outliers, y-outliers and xy-outlieres.The datasets are contaminated with $N(50,1)$ (for x variable) and with $N(100,1)$ (for y variable). For each simulated data set, 5%, 10% and 15% trimmed means of Mean Squares Prediction Error (MSPE) and coefficients of determination (R^2) of the regression coefficients are recorded.

3.2 Simulation Results

At first, the performances of different regression methods from clean data to contaminated data (x-outlier) when true intercept value $(\beta_0 = 10)$ and true slope value $(\beta_1 = 6)$ are presented in **Table 1.** The estimated values of intercepts and slopes for all the methods considered almost coincide with the actual values for clean data. Each of the methods provides an *R*-squared of 98%, which reveals that 98% of the variation of the dependent variable can be explained by the independent variable. MSPE are also almost same. So each method produces a better fit for the model.

Table 1. Performance of different regression methods from clean data to contaminated data (*x*-outlier) when true intercept value (β_0 = 10) and true slope value (β_1 = 6)

Data	Methods	Intercept	Slope	MSPE	R-square
Clean	LS	9.9991	6.0003	0.4603	0.9874
	MM	9.9996	5.9994	0.4601	0.9875
	LMS	9.9975	5.9976	0.4890	0.9866
	RM	9.9991	6.0001	0.4622	0.9873
5% *x*-outlier	LS	9.8765	0.0501	16.7381	0.5445
	MM	9.9994	6.0004	0.4606	0.9874
	LMS	9.9996	6.0008	0.4851	0.9867
	RM	9.9998	5.8960	0.4685	0.9871
10% *x*-outlier	LS	9.8756	0.0245	16.8819	0.5426
	MM	9.9995	6.0004	0.4609	0.9874
	LMS	9.9967	6.0005	0.4850	0.9867
	RM	9.9997	5.7802	0.4881	0.9866
15% *x*-outlier	LS	9.8749	0.0161	16.9422	0.5410
	MM	9.9998	6.0003	0.4610	0.9874
	LMS	9.9979	5.9990	0.4841	0.9867
	RM	9.9986	5.6428	0.5300	0.9854

With 5%, 10% and 15% of x-outliers in the sample data, it can be seen that MM, LMS and RM methods give true

values of intercepts and slopes. The *BP* of each of these methods is 50%. The outlier has a large influence on the LS method. This method gives an almost true value of intercept, but the slope is far from the true one.

MM, LMS and RM estimations provide an *R*-squared of 98%, which reveals that 98% of the variation of the dependent variable is explained by the independent variable, while LS estimation provides an *R*-squared of 54%.The LS estimation also yields the higher average value of MSPE compared to other methods. So all the methods except LS produce a better fit for the model. The *BP* of LS is 0%.

Secondly, the performances of different regression methods for contaminated data (*y*-ouitlier) when true intercept value $\beta_0 = 10$) and true slope (β_1 = 6)are presented in **Table 2**.

Table 2. Performance of different regression methods from clean data to contaminated data (*y*-outlier) when true intercept value (β_0 = 10) and true slope value (β_1 = 6)

Data	Methods	Intercept	Slope	MSPE	R-square
5% *y*-outlier	LS	14.4898	5.7034	21.1884	0.4180
	MM	9.9995	6.0002	0.4606	0.9874
	LMS	9.9970	5.9995	0.4860	0.9867
	RM	10.0688	5.9967	0.4916	0.9872
10% *y*-outlier	LS	18.9779	5.3984	82.4698	−1.2664
	MM	9.9992	6.0001	0.4609	0.9874
	LMS	9.9973	5.9999	0.4847	0.9867
	RM	10.1463	5.9935	0.4741	0.9870
15% *y*-outlier	LS	23.4690	5.1100	184.0149	−4.0593
	MM	9.9995	6.0002	0.4612	0.9867
	LMS	9.9989	6.0002	0.4841	0.9874
	RM	10.2336	5.9875	0.4916	0.9870

In cases of 5%, 10% and 15% *y*-outliers in the sample data, it is observed that the robust regression methods produce satisfactory results, and they are resistant to outliers. On the other hand, the LS method is badly affected by outliers. It produces a very large MSPE. So we conclude that the LS estimation is inefficient and biased. It provides a negative *R*-square. *R*-square is negative solely when the approved model does not pursue the aptitude of the data, so fits poor than an aclinic line. This negative value also implies that the data are not narrated by the model.

A negative *R*-square is not a mathematical improbability or the symbol of a computer flaw. It merely means that the selected model (with its obligations) fits the data really sickly.

At last, the performances of different regression methods from clean data to contaminated data (*xy*-outliers) when true intercept value ($\beta_0 = 10$) and true slope value ($\beta_1 = 6$) are presented in **Table 3**. Our simulation results show that all the robust methods perform much better than the LS method. Among the existing robust methods, MM estimation method is the best and strongly resistant to outliers.

Table 3. Performance of different regression methods from clean data to contaminated data (*xy*-outliers) when true intercept value (β_0 = 10) and true slope value (β_1 = 6)

Data	Methods	Intercept	Slope	MSPE	R-square
5% *xy*-outliers	LS	9.9191	1.8344	8.4408	0.7713
	MM	10.0015	5.9994	0.4595	0.9874
	LMS	10.0031	5.9988	0.4857	0.9867
	RM	10.0119	5.8989	0.4691	0.9871
10% *xy*-outliers	LS	9.9221	1.8163	8.5138	0.7694
	MM	10.0015	5.9996	0.4598	0.9873
	LMS	10.0017	5.9992	0.4846	0.9867
	RM	10.0528	5.7842	0.4891	0.9866
15% *xy*-outliers	LS	9.9239	1.8104	8.5390	0.7687
	MM	10.0011	5.9995	0.4601	0.9873
	LMS	10.0002	6.0017	0.4831	0.9867

	RM	10.1216	5.6462	0.5330	0.9854

4. Real Data Applications

In this section, a real dataset is used to further illustrate the performance of different slope estimation methods in a simple linear regression model.

Machine Data

This dataset, titled Relative CPU Performance Data, has been collected from UCI Machine Learning Repository. This data set was introduced by Phillip Ein-Dor and Jacob Feldmesser (1978). The data set contains 209 instances and 9 attributes (6 predictive attributes, 2 non-predictive, one goal field, and the linear regression's guess). Here, we only considered two continuous variables (CHMAX: maximum channels in units (integer) and PRP: published relative performance (integer)).

Different regression methods are applied to this dataset, and **Tables 4** and **5** represent the intercept and slope of these methods from clean data to contaminated data.

If we change the value of the predictor variable (say 10th observation 15) by a large value 10000 and corresponding response variable (say 10th observation 13) by a large value 10000, then we see that in case of contaminated data, LS is rapidly affected by outliers. So it changes the values of intercepts and slopes. For LS, only one outlier is sufficient to cause such an effect. On the contrary, almost all the robust regression methods remain stable. Thus, we clearly observe that the robust regression methods are more stable than the LS method.

Table 4. Intercept and slope of different methods for clean data

Methods	Intercept	Slope
LS	104.8890	-0.0071
MM	33.2597	0.5830
LMS	22.8333	0.6667
RM	18.5000	2.4200

Table 5. Intercept and slope of different methods for contaminated data

Methods	Intercept	Slope
LS	203.2051	-0.0191
MM	34.2511	0.5908
LMS	22.1071	0.6429
RM	19.8000	2.2860

5. Conclusion

From this study, we estimate regression coefficients using different regression methods. A comparison is made among LS and robust regression methods using simulated datasets and real dataset. We compare the performances of robust regression methods and LS by contaminating the simulated datasets. We have observed that LS method and other robust regression methods provide almost the same results for clean data. But, in the case of contaminated data, robust regression methods have performed much better than the LS method. Among the existing robust regression methods (MM, LMS and RM), we observe that MM method produces the most efficient results in all contaminated cases. Although LMS and RM are strongly resistant to outliers, their efficiencies are low. In real datasets, when we replace some observations by outliers, LS changes immediately and provides poor results whether robust regressions are strongly resistant to outliers.

Limitations

Small departures from normality, LS method yields low power. It is not effective for determining and examining outliers (Wilcox R. R. (1998a) and Wilcox R. R. (1998b). We have observed that the performance of robust regression is better than LS both in a simulation study and real data applications. A disadvantage of the LMS method is its lack of efficiency because of its $n^{-\frac{1}{3}}$ convergence.

References

Chen, C. 2002. Robust Regression and Outlier Detection with the ROBUSTREG Procedure. *Statistics and Data Analysis*, 265-27.

Cornbleet, P. J., and Gochman, N. 1979. Incorrect least-squares regression coefficients in method-comparison analysis. *Clinical Chemistry, 25*, 432–438.

Draper, N. R., and Smith, H. 1998. Applied Regression Analysis (3rd edn). United States: Wiley-Interscience Publication.

Everitt, B. 2002. The Cambridge Dictionary of Statistics (2nd edition).Cambridge: Cambridge University Press.

Hampel, F. 2002. Robust Inference.*Encyclopedia of Environmetrics*,3, 1865–1885.

Maronna, R. A., Martin, R. D., and Yohai, V. J. 2006. Robust Statistics. John Wiley & Sons.

Massart, L., Kaufman, L., Rousseeuw, P. J., and Leroy, A. 1986. Least median of squares: a robust method for outlier and model error detection in regression and calibration. *Analytica Chimica Acta*, 187, 171-179.

Neter, J., Kutner, M., Nachtsheim, C., and Wasserman, W. 1996. *Applied Linear Statistical Models.* WCB McGraw-Hill, New York.

Rousseeuw, P. J. 1984. Least Median of Squares Regression. *Journal of the American Statistical*

Association, 79, 871-880.

Rousseeuw, P. J., and Leroy, A. M. 1987. Robust regression and outlier detection. New York: John Wiley & Sons.

Siegel, A. F. 1982. Robust regression using repeated medians. *Biometrika,* 69, 242–244.

Stöckl, D., Dewitte, K., and Thienpont, L. M. 1998. Validity of linear regression in method comparison studies: is it limited by the statistical model or the quality of the analytical input data? *Clinical Chemistry,* 44, 2340–2346.

Susanti, Y., Pratiwi, H., H., S. S., and Liana, T. 2014. M Estimation, S Estimation, And MM Estimation. *International Journal of Pure and Applied Mathematics,* 91, 349-360.

Wilcox, R. R. 1998. How many discoveries have been lost by ignoring modern statistical methods? *American Psychologist,* 53, 300–314.

Wilcox, R. R. 1998. The goals and strategies of robust methods. *British Journal of Mathematical and Statistical Psychology,* 51, 1-39.

Yohai, V. J. 1987. High Breakdown Point and High Efficiency Robust Estimates for Regression. *Annals of Statistics,* 15, 642-656.

Zioutas, G., Avramidis, A., and Pitsoulis, L. 2005. A Penalized Trimmed Squares Method for Deleting Outliers in Robust Regression, *Elsevier*.