Alabama Journal of Mathematics
39 (2015)

ISSN 2373-0404

Regression Analysis for Data Containing Outliers and High Leverage Points

Asim Kumer Dey
Department of Mathematics
Lamar University

Md. Amir Hossain
Institute of Statistical Research and Training
University of Dhaka

Kumer Pial Das
Department of Mathematics
Lamar University

The strong impact of outliers and leverage points on the ordinary least square (OLS) regression estimator is studied for a long time. Situations in which a relatively small percentage of the data has a significant impact on the model may not be acceptable to the user of the model. A vast literature has been developed to find robust estimators that cope with these "atypical" observations. Selection of proper methods of estimation in the presence of influential observations (either outliers or leverage or both) needs to be investigated in further details. This study is designed to find an appropriate method of estimation in regression analysis in the presence of these three different types of influential observation. A comparison has been made among different well known methods of estimation in each situation on the basis of data generated by Monte Carlo simulation.

Introduction

When the observations Y in the linear model

$$Y = X\beta + \epsilon$$

are normally distributed, the method of least squares is a good parameter estimation procedure in the sense that it produces an estimator of the parameter vector β that has good statistical properties. However, there are many situations where we have evidence that the distribution of the response variable is (considerably) non-Gaussian and/or there are outliers that affect the regression model. A case of considerable practical interest is one which the observations follow a distribution that has a longer or heavier tail than normal. These heavy-tail distributions tend to generate outliers, and these outliers may have a strong influence on the method of least squares in the sense that they "pull" the regression equation too much in their direction (Montgomery, Peck, & Vining, 2012). The purpose of this study is to determine the appropriate estimation methods of regression parameter for this case.

The paper is organized into four sections: the first introduces a brief description of regression model; the second describes outlier and leverage point,their diagnostic tests, and their effect on OLS estimation; the third discusses different robust estimation methods; the fourth gives a comparison among OLS method and different robust estimation methods for different situation through simulated data; and the fourth discusses the results of the study.

Regression Analysis

Regression Model

Regression analysis is concerned with the study of the dependence of one variable, the dependent variable, on one or more other variables, the explanatory variables, with a view of estimating and/or predicting the (population) mean or average value of the former in terms of the known or fixed (in repeated sampling) values of the later (Gujarati, 2003).

Suppose that we have a response variable y and a number of explanatory variables x_1,x_2,..., x_k that may be related to y. Then regression model for y can be written as

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} + \epsilon_i, \tag{1}$$

for all $i = 1, 2, ..., n$.

The β_0, β_1, $\cdots$, β_k are unknown parameters, known as regression coefficients, and ϵ_i are called the error term or disturbance term. This error term captures all other factors which influence the dependent variable y_i other than the explanatory variable x_i. ϵ_i is an independent random variable with zero mean and constant variance.

Equation 1 can be written in a vector form as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{2}$$

Corresponding Author: Kumer Pial Das, Ph.D., Email: kumer.das@lamar.edu where

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & a_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{bmatrix},$$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Here, $\mathbf{y}$ is an $n \times 1$ vector of the observations , $\mathbf{X}$ is an $n \times (k+1)$ matrix of the levels of the regressor variables, $\boldsymbol{\beta}$ is a $(k+1) \times 1$ vector of the regression coefficients, and $\boldsymbol{\epsilon}$ is an $n \times 1$ vector of random errors (Drapper & Smith, 1998).

Parameter Estimation

A number of procedures have been developed for parameter estimation and inference in linear regression. Among them ordinary least squares (OLS) is the simplest and very common method of estimation. The OLS estimator $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ is obtained by minimizing the error sum of square

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{n} \boldsymbol{\epsilon}_i^2 = \boldsymbol{\epsilon}'\boldsymbol{\epsilon} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}).$$

Then the model (2) can be estimated by using the estimated parameters $\hat{\boldsymbol{\beta}}$.

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \ldots + \hat{\beta}_p x_{ip} \tag{3}$$

The deviation of an estimated value of dependent variable from its observed value (actual value) is known as residual, which can be obtained as,

$$e_i = y_i - \hat{y}_i \tag{4}$$

Outlier and Leverage

In regression analysis, an outlier is an observation with large residual. In other words, it is an observation whose dependent variable value is unusual given its value on the predictor variables. On the other hand, an observation with an extreme value on a predictor variable is a point with high leverage. Leverage is a measure of how far an independent variable deviates from its mean (Rousseeuw, 1984).

A number of methods have been developed to diagnosis outliers. Among them studentized residuals are frequently used. Studentized residuals can be obtained as,

$$\text{Studentized residual} \quad = \quad \frac{e_i}{S_{(i)}\sqrt{1 - h_{ii}}} \tag{5}$$

Here $S_{(i)}$ is the standard deviation of the residuals where i^{th} observation is deleted and h_{ii}, leverage, is the i^{th} diagonal entry in the hat matrix, $H = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. If a Studentized Residual exceed +2 or −2, the observation is an outlier.

Cook's distance (or Cook's D) is another measure for diagnosing outlier. It can be defined as

$$\text{Cook's D} \quad = \quad \frac{{e_i}^2}{p\text{MSE}}\left(\frac{h_{ii}}{(1 - h_{ii})^2}\right) \tag{6}$$

where p is the number of parameter to be estimated and MSE is the mean square error of the regression model. If Cook's D $> \frac{4}{n}$, the observation is an outlier.

And leverage is measured by the diagonal elements, h_{ii}, of the hat matrix, $H = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. When a leverage $> 2p/n$ then it is a matter of concern.

Different graphical procedures such as plot of residuals versus fitted values, plot of leverages versus standardized residuals are also used to detect outliers and leverage points.

The method of ordinary least squares (OLS) is one of the most powerful and most popular estimation procedure in regression analysis because of its attractive statistical properties (e.g., best linear unbiased estimator (BLUE)) and mathematical simplicity. But when there are outliers in the data, these outliers have a strong influence on the method of OLS in the sense that these few data points change the path of the regression equation too much in their direction (Figure 1). As a result the values of the regression coefficients or summary statistics such as the t or F statistic, R^2, and the residual mean square of OLS estimation become very sensitive to these outliers.

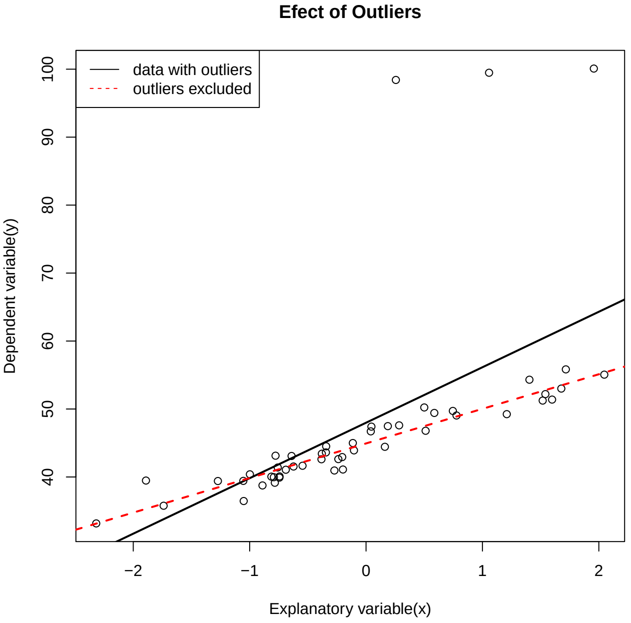

Figure 1. Outliers effect on regression model

One way to deal with this situation is to discard the extreme observations. This will produce a line that passes nicely through the rest of the data set and one that is more pleasing from a statistical standpoint. However, we are now discarding observations simply because it is expedient from a statistical modeling viewpoint, and this is not a good practice.

Robust Estimation

A robust estimation procedure is one that dampens the effect of observations that would be highly influential if least square are used. The idea of robust estimation is to weigh the observations differently based on how well behaved these observations are. A robust estimation procedure should produce essentially the same results as least squares when the underlying distribution in normal and there are no outliers. Roughly speaking, it is a form of weighted and reweighted least squares regression (Holland & Welsch, 1977). A number of robust estimation have been introduced in the last three decades. Among them M estimation and MM estimation are frequently used.

A robust estimator has two important properties, namely breakdown point and efficiency. Breakdown point of an estimator is the proportion of incorrect observations (i.e. arbitrarily large observations) an estimator can handle before giving an arbitrarily large result. The range of breakdown point is zero to 0.5 ($0 \leq$ breakdown point ≤ 0.5) (Rousseeuw & Leroy, 1987). The smallest possible breakdown point is $1/n$, that is, a single observation can distort the estimator so badly that it is of no practical use to regression model-builder. The breakdown point of OLS is $1/n$. The fraction of data that are contaminated by erroneous data typically varies between 1% to 10%. Therefore, we would generally want the breakdown point of an estimator to exceed 10%.

On the other hand, the efficiency of a robust estimator can be thought of as the residual mean square obtained from OLS divided by the residual mean square from the robust procedure.

M Estimation

M estimation introduced by (Huber, 1973) is the simplest approach both computationally and theoretically. Instead of minimizing sum of squares of the residuals, M estimator minimizes a sum of less rapidly increasing function (weight function) of the residuals. M estimation procedure solves this system by using iteratively reweighted least squares (IRLS).

M-estimators are defined to be robust against heavy-tailed error distribution and non-constant error variance and thus y outliers but they also implicitly assume that the model matrix X is measure without error. Under these conditions, M estimates are more efficient than OLS estimates. Under the Gauss-Markov assumptions, however, M estimates are 95% as efficient as OLS estimates. But this can be affected by high leverage points as an identical manner to OLS. Consequently, the breakdown point of the class of M-estimators is $1/n$.

MM estimation

MM estimation was introduced by (Yohai, 1987). It has simultaneously the following properties: (1) considering the errors have a normal distribution they are highly efficient and (2) they have high breakdown point.

MM estimator is based on the following two estimators: Least Trimmed Squares (LTS) Estimator and S Estimator. LTS Estimator (Rousseeuw, 1984) is a high breakdown value method. And S estimator (Rousseeuw & Yohai, 1984)) is a high breakdown value method with higher statistical efficiency than LTS estimation. As a result MM estimation can be defined by a two stage procedure:

1. The first step is to compute an initial (consistent) high breakdown value estimate which may not be efficient. The procedure provides two kinds of estimates as the initial estimate, the LTS estimate and the S estimate.

2. The second stage is to compute an M-Estimate of the error scale using the residuals from the initial (LTS/S) estimate.

A Comparison Among Different Estimation Methods

We have studied three well known methods of estimations: OLS estimation, M estimation and MM estimation. These three estimation procedure do not give similar results for different type of influential observation (either outliers or leverage or both), rather for a specific type of influential observation one gives better performance than others. Our objective is to identify the best method of estimation for the data with this specific type of influential observation.

We start by creating a data set of size $1,000$ by randomly generating three independent explanatory continuous variables (labeled x_1, x_2, x_3) and an error term (ϵ) from independent univariate normal distributions with zero mean and unit variance.

A y variable is then generated according to the formula

$$y_i = 10 + 5x_{1i} + 3x_{2i} + 1x_{3i} + \epsilon_i, \tag{7}$$

for $i = 1, 2, 3, ..., 1,000$. Here, $\beta_0 = 10, \beta_1 = 5, \beta_2 = 3, \beta_3 = 1$.

Having the true value of the parameters known, we can compare different methods on the basis of bias (i.e. deviation of the estimated values from their true values) and standard error of the estimated parameters.

We then contaminate 10% of dependent variables (which causes outliers) and use our three methods of estimation OLS, M, MM to estimate the parameters. For each method we iterate the procedure 1,000 times and in each iteration we calculate the bias and standard error of the estimates. We take an average of these 1,000 values. Finally, we make a comparison among OLS, M, and MM estimation procedures on the basis of bias and standard error of the estimates.

Similar analysis is also done for the following two situations (1) data with 1% contamination in one or more explanatory variable(s) which causes high leverage, and (2) data with both 10% contamination in dependent variables and 1% contamination in one or more independent variables.

OLS estimation for the original data

The Table 1 shows OLS estimation for the original data set, here R^2 is 0.9927 and adjusted R^2 is 0.9926.

Table 1
ANOVA for original data

Source	Df	SS	MS	F value	Pr(>F)
Model	3	33884	11294	4.9e+4	<0.01
Residual	996	250	0.25		
Total	999	3267474			

From Table 1 it can be observed that the overall model is highly significant, where R^2 and adjusted R^2 shows that almost 100% variation of y is explained by the model.

Table 2
OLS estimates for the original data

Coefficients (True value)	Estimated Parameter	Bias	Standard Error	Pr(>|t|)
$\beta_0(10)$	10.01	–0.001	0.02	< 0.01
$\beta_1(5)$	4.99	0.004	0.01	< 0.01
$\beta_2(3)$	2.99	0.001	0.01	< 0.01
$\beta_3(1)$	0.99	0.007	0.02	< 0.01

From Table 2 we can see that all the parameters are highly significant, they have negligible bias, and their standard errors are very small.Thus we can say that OLS estimation can produce very good estimates when data have no influential observation.

Data with 10% contamination in Y

We contaminate 10% of y values without modifying explanatory variables such that this contaminated values can cause outliers. Here original y values are taken from independent univariate normal distributions with zero mean and unit variance. And 10% values are generated from independent univariate normal distributions with mean 200 and variance 2. Obviously, these value produce outliers in the data. Our goal is to compare OLS, M, MM estimation procedures to find which method perform better in this situation.

Comparison is made on the basis of bias and standard error of the parameters, their significancy in the model (p-value of the t-test) as well as F statistic, and R^2 of the model.

Table 3
OLS estimates for 10% contamination in Y (ANOVA)

Source	df	SS	MS	F Value	Pr(>F)
Model	3	5453	1818	0.56	0.65
Residual	996	3262021	3275		
Total	999	3267474			

In this case when we use OLS estimation methods R^2 becomes 0.0017, which is very very small. Table 3 shows that a sudden drastic change is occurred in our model: the overall model become insignificant (p-value of the F test is very large) and R^2 is very small. From Table reftab4 we can see that all the variables are insignificant, and the standard errors and biases are very high.

Table 4
OLS Parameter Estimates

Variables	Estimated Parameter	Pr(>|t|)	Standard Error	Bias
Intercept	29.16	< 0.01	1.81	−19.00
x_1	1.31	0.46	1.78	0.491
x_2	–0.50	0.78	1.79	0.324
x_3	1.80	0.31	1.77	0.103

However when we use the M estimation to estimate the parameters the model become highly significant (p-value of the F test is very small) and R^2 becomes 0.7703, a reasonably good figure. Moreover, Table 5 shows that all the variables are significant, bias and standard error are very small.

Table 5
M estimates for 10% contamination in Y

Variables	Estimated Parameter	Pr(>|t|)	Standard Error	Bias
Intercept	10.22	< 0.01	0.02	–0.09
x_1	4.99	< 0.01	0.02	0.001
x_2	2.99	< 0.01	0.02	0.001
x_3	1.01	< 0.01	0.01	0.001

For the case of MM estimation, like M estimation, the model is highly significant and R^2 is 0.7612. Table 6 shows that all the variables are significant, bias and standard error are very small for the case when MM estimation is used.

Table 6
MM estimates for 10% contamination in Y

Variables	Estimated Parameter	Pr(>|t|)	Standard Error	Bias
Intercept	10.24	< 0.01	0.018	−0.001
x_1	4.99	< 0.01	0.018	0.001
x_2	2.99	< 0.01	0.018	0.001
x_3	1.01	< 0.01	0.017	0.001

Data with 1% Leverage Points

The above experiment has been replicated for data with 1% leverage values. This extreme extreme leverage points are obtained by generating two random samples of size 10 from two normal population with mean 160 and 170 and variance 1 and 4 respectively. Then we replace the values in first 10 position in x_{1i} and x_{2i}, without modifying y and x_{3i}. Then we want to compare OLS, M, MM estimation procedures for this situation.

When OLS estimation is used for data with leverage we get the overall model significant with very small R^2 (0.0991). From Table 7 we can see that although all the variables are significant the standard errors are very high.

Table 7
OLS estimates for data with 1% Leverage

Variables	Estimated Parameter	Pr(>|t|)	Standard Error	Bias
Intercept	10.35	< 0.01	0.173	0.009
x_1	1.02	< 0.01	0.121	4.164
x_2	−1.07	< 0.01	0.130	3.872
x_3	1.12	< 0.01	0.174	−0.005

We observe a small R^2 (0.0811) and higher standard error for the M estimation method too; however, the overall model and the parameters become significant in terms of p-value (Table 8).

Table 8
M estimation for data with 1% Leverage

Variables	Estimated Parameter	Pr(>|t|)	Standard Error	Bias
Intercept	10.36	< 0.01	0.18	0.03
x_1	0.99	< 0.01	0.13	4.15
x_2	−1.05	< 0.01	0.13	3.88
x_3	1.08	< 0.01	0.17	0.01

But when MM estimation is used in this situation not only the overall model and the parameters become highly significant but also R^2 become high (0.7642). Moreover, standard errors and bias of the parameters become smaller compare to OLS and M estimation (Table 9).

Table 9
MM estimation for data with 1% Leverage

Variables	Estimated Parameter	Pr(>|t|)	Standard Error	Bias
Intercept	10.03	< 0.01	0.017	0.001
x_1	5.01	< 0.01	0.018	−0.001
x_2	2.98	< 0.01	0.018	−0.001
x_3	0.98	< 0.01	0.017	0.001

Data with both 1% Leverage Points and 10% contamination in Y

This study also investigates the appropriate method for data with both leverage values and outliers. Then Extreme leverage points are obtained by generating two random samples of size 10 from two normal population with mean 160, 170 and variance 1 and 4 respectively. Then we replace the values in the first 10 position in x_{1i} and x_{2i}, without modifying x_{3i}. Again 10% values are generated from independent normal distribution with mean 200 and variance 2.

If OLS estimation is used for data with both outliers and leverage points we get all the variables except intercept are insignificant (Table 10). Again standard errors are very high and R^2 is very small (0.2140).

Table 10
OLS estimation for data with both leverage(1%) and outliers (10%)

Variables	Estimated Parameter	Pr(>|t|)	Standard Error	Bias
Intercept	19.037	< 0.01	0.873	−9.085
x_1	1.298	0.024	0.575	3.850
x_2	−0.493	0.423	0.615	3.329
x_3	1.091	0.228	0.904	0.093

If M estimation is used although all the variables becomes significant (Table 11), R^2 becomes very small (0.1734) and standard errors become very large.

But when MM estimation is used in this situation not only the overall model and the parameters become highly significant but also the R^2 become high (0.7642). Again standard errors and biases of the parameters are also very small in comparable to OLS and M estimation (Table 12).

Table 11
M estimation for data with both leverage (1%) and outliers (10%)

Variables	Estimated Parameter	Pr(>|t|)	Standard Error	Bias
Intercept	9.897	< 0.01	0.175	−0.980
x_1	1.380	< 0.01	0.115	3.708
x_2	−0.523	< 0.01	0.123	3.429
x_3	1.186	< 0.01	0.181	0.013

Table 12
MM estimation for data with both leverage(1%) and outliers (10%)

Variables	Estimated Parameter	Pr(>|t|)	Standard Error	Bias
Intercept	9.981	< 0.01	0.018	0.001
x_1	5.025	< 0.01	0.017	−0.001
x_2	3.021	< 0.01	0.019	−0.001
x_3	0.982	< 0.01	0.019	0.001

Conclusion

The primary objective of this study is to identify an appropriate estimation procedure for linear regression model that can deal with different types of influential observations. From the above discussion we can make the following three concluding remarks. First, when outliers present in data OLS estimation gives very misleading result. But M and MM estimation do better job in this situation and give proper result. Second, for data with leverage points OLS and M estimation gives misleading outputs, whereas, MM estimation gives expected results. Finally, OLS and M estimation gives deceptive results for data with both leverage points and outliers, but MM estimation gives proper results in this situation.

References

Drapper, N. R., & Smith, H. (1998). *Applied Regression Analysis*. John Wiley & Sons.

Gujarati, D. N. (2003). *Basic Econometrics. 4th*. New York: McGraw-Hill.

Holland, P. W., & Welsch, R. E. (1977). Robust regression using iteratively reweighted least-squares. *Commun. Statist. Theor. Meth.*, *6*(9), 813–827.

Huber, P. J. (1973). Robust regression: asymptotics, conjectures and Monte Carlo. *Annals of Statistics*, *1*, 799-821.

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). *Introduction to Linear Regression Analysis*. John Wiley & Sons.

Rousseeuw, P. J. (1984). Least median of squares regression. *Journal of the American Statistical Association*, *79*(388), 871-880.

Rousseeuw, P. J., & Leroy, A. M. (1987). *Robust Regression and Outlier Detection*. J. Wiley & Sons, New York.

Rousseeuw, P. J., & Yohai, V. J. (1984). *Robust regression by means of S-estimators*. Springer.

Yohai, V. J. (1987). High breakdown-point and high efficiency robust estimates for regression. *Annals of Statistics*, *15*, 642-656.