Available online at www.isas.org.in/jisas
JOURNAL OF THE INDIAN SOCIETY OF
AGRICULTURAL STATISTICS 74(3) 2020 181–186

Robust Regression in the Presence of Leverage: An Application to the Baseball Data

Amit Saha, K.N. Singh, Bishal Gurung, Santosha Rathod and Md. Yeasin

ICAR-Indian Agricultural Statistics Research Institute, New Delhi

Received 26 December 2017; Revised 14 May 2020; Accepted 14 July 2020

SUMMARY

It is well known that linear regression analysis is one of the commonly used statistical tools in various fields. The ordinary least squares (OLS) is generally adopted to estimate the parameters in the model provided all the necessary assumptions are satisfied. OLS is widely used because of its desirable properties like unbiasedness, minimum variance, consistency, asymptotic unbiasedness etc. However, outcomes of OLS may be affected if some of the assumptions do not hold properly. Presence of outliers is one of the main reason to deliver poor results in OLS. So, it is very much important to use a robust method for parameter estimation which is not much affected by outliers. Robust regression analysis is a statistical technique which is an improvement to least squares estimation to cope or to detect the outliers. In other words, the robust regression analysis performs well when the assumptions are not satisfied by the data. One can transform variables to deal with the data when some of the assumptions are substantially violated. But, the influence of outliers has often not been attenuated by the transformation. So it is better to use robust regression that is resistant to the influence of outliers. In this paper, OLS and two robust regression methods (M and S estimation) are discussed and applied to run the regression model on baseball data. It has been seen that S estimation method outperformed the OLS and M estimation.

Keywords: M-estimation, OLS estimation, Robust regression and S-estimation.

1. INTRODUCTION

Regression is one of the popular methods among statisticians. OLS method is popularly used to estimate the parameters of the regression model. But, OLS method is highly sensitive to outliers. Even a single outlier has potential to distort the significance of parameter estimates. On the other hand, robust regression curbs the influence of outliers. The principal aim of robust regression is to provide resistant results in the presence of outliers.

It is well known that one of the assumptions of linear regression is normality of the residuals. This normality assumption is breached due to the presence of outlier. So, an outlier has the potential to violate the assumption of linear regression. But, the solution is also available to attenuate the effect of outlier. One solution is to transform one or more explanatory variables or response variable. One may go through the adjustment of the model by adding higher order terms. However, these solutions may fail to provide a significant result. The problem will arise more when the error distribution is heavier tailed than normal. But, robust regression method is appropriate in this case. In reality, each observation is equally weighted in least square estimation whether it is normal or extreme observation. In case of robust regression method, weighting is given to observation unequally. Robust methods give lower weightage to the extreme observation. That's why, it can be effective to the outlier data appropriately.

Huber (1973,1981) who first introduced robust regression estimators which are also known as M regression estimators. A few commonly used M–estimators are L_1, L_2, $L_1 - L_2$, Huber, L_p, Cauchy, German-Maclure, Welsh estimator etc. However, they have many pros and cons. M estimator is robust only when the contamination occurs in the response direction but it is not robust with respect to leverage points. Leverage points are the extreme

Corresponding author: Amit Saha
E-mail address: amitsahauni@gmail.com

observations in the predictor space, whereas outliers are in the response space. In some situations, extreme observation is observed in both response and predictor space, which is known as influence point.

Hampel (1975) suggested an estimator whose breakdown point is 0.50, which means it can resist the effect of outliers up to 50%. The estimator is based on minimizing the median absolute deviation of the residuals. S estimator is nothing but the generalization of least median squares which was introduced by Rousseeuw and Yohai (1984). The reason behind the name of S estimator is that they are based on estimators of scale. S estimator also has some good robustness properties such as bounded influence, higher breakdown point (50%) etc.

2. LINEAR REGRESSION

In matrix notation, linear regression model can be written as

$$\boldsymbol{X} = \boldsymbol{Z\theta} + \boldsymbol{\varepsilon} \tag{1}$$

where, $\boldsymbol{X}$ is a $n \times 1$ vector of observed response values,

$\boldsymbol{Z}$ is a matrix of $n \times q$ order of the predictor variables,

$\boldsymbol{\theta}$ is a vector of parameters of order $q \times 1$ and $\boldsymbol{\varepsilon}$ is a vector of error of order $n \times 1$.

OLS is the most commonly used regression method to estimate the parameters. The OLS estimate is obtained by minimization of sum of squares of residuals (S) of the above mentioned model.

The parameters are determined by taking the partial derivatives of S with respect to $\boldsymbol{\theta}$, and setting the results equal to zero. The solution will be

$$\widehat{\boldsymbol{\theta}} = (\boldsymbol{Z}'\boldsymbol{Z})^{-1}(\boldsymbol{Z}'\boldsymbol{X}) \tag{2}$$

But, this estimator has been losing its reliability for lack of robustness in spite of its simplicity to compute. It has already been mentioned that one outlier can spoil various assumptions. Due to presence of outliers, values may not be identically distributed and the property of homoscedasticity may also be violated. These assumptions have to be fulfilled for the validation of OLS regression model. When the regression model does not meet the fundamental assumptions, the prediction and estimation of the model may become biased. Residuals, differences between the values predicted by the model and the real data that are very large can seriously distort the prediction. When these residuals are extremely large, they are called outliers. The outliers inflate the error variance, so that confidence interval becomes stretched and estimation cannot become asymptotically consistent.

3. ROBUST REGRESSION

Robust regression estimation is an alternative to OLS at the time of unfulfillment of fundamental assumption. The sole cause of unfulfilled assumption is the outliers. Robust regression dampens the effect of outliers (Draper and Smith, 2014) and gives a stable result as compared to OLS.

The important properties of robust regression are-

1. Breakdown point (BDP): BDP represents the ability of the estimator to withstand in the presence of outliers (Chen, 2002) which is usually expressed in percentage. Suppose, a robust estimator has 30% BDP. It means that robust regression estimation provides a stable or useable estimator when 30% of data comprises outlier. BDP of OLS is 0% which means that one outlier is sufficient to distort the estimation.
2. Efficiency: Efficiency of robust regression is measured as compared to OLS.

   Efficiency of Robust Estimator:-

   $\mathit{Residual\ Mean\ Square(OLS)/Residual\ Mean\ Square\ (Robust\ Method)}$
3. Bounded influence: To cope up the problem of estimator with leverage, the bounded influence is designed; Ex. - Generalized M or GM estimation, S-estimation. This property shows the ability of estimator to provide robust result in the presence of leverage.

   Robust model should have following characteristics-

   a. Model should be unbiased.

   b. Model should be efficient.

   c. If the assumptions are violated, then the performance of the model is not substantially affected much.

   d. It is asymptotically normal.

3.1 M-estimation

M-estimation is the most common robust regression estimation method. In fact, M-estimation is the generalization of maximum likelihood estimation. That's why, the name of this estimation is 'M-estimation' which based on the minimization the sum of a function $\tau(.)$ of the residuals. M estimators are obtained by minimizing the objective function over all θ as follows:

$$\sum_{j=1}^{n} \tau(e_j) = \sum_{j=1}^{n} \tau(x_j - \sum_{k=0}^{q} z_{jk}\theta_k) \quad (3)$$

where,

$e_j = x_j - \hat{x}_j$ is the jth residual;

$j = 1,2, \dots, n$ (observations) and

$k = 0,1,2, \dots, q$ (parameters).

The noteworthy properties of τ function are-

1. It is always nonnegative.
2. It is symmetric.
3. It is equal to zero when its argument is zero.
4. It is a monotonic function.

Least square is the special case of M-estimation when $\tau(e_j) = e^2$. This least square estimator satisfies all the above mentioned properties of τ function.

To estimate the M regression parameters, the estimator should be scale equivariant. It is necessary to go with the standardization of the residuals when it is not scale equivariant. Usually, a popular estimator μ is used to make scale equivariant which is known as rescaled MAD (Median Absolute Deviation). So. The re-scaled MAD is:

$$\mu = 1.4826 \times MAD \quad (4)$$

where, the formula for MAD (median absolute deviation) is:

$$\text{MAD} = Median\left|e_j - median\ (e_j)\right| \quad (5)$$

So, need to minimize,

$$\sum_{j=1}^{n} \tau(v_j) = \sum_{j=1}^{n} \tau\left({}^{e_j}/_{\mu_j}\right) = \sum_{j=1}^{n} \tau\left({}^{(x_j - \sum_{k=0}^{q} z_{jk}\theta_k)}/_{\mu_j}\right) \quad (6)$$

Partial derivative of $\tau(v)$ with respect to θ,

$$\sum_{j=1}^{n} z_{jk}\ \varphi\left({}^{(x_j - \sum_{k=0}^{q} z_{jk}\theta_k)}/_{\mu_j}\right) = 0 \quad (7)$$

Next step is to choose the τ function. Many functions have been defined in literature, viz. Huber's function, Tukey's bi-weight function etc. Here, Tukey's bi-weight objective function has been used which is more resistant to the outliers as compared to the Huber M-estimator (Andersen, 2008) and taken the tuning constant $d = 4.685$ to get 95% efficiency. So, the τ function is;

$$\tau(v_j) = \begin{cases} \frac{3{v_j}^2}{d^2} - \frac{3{v_j}^4}{d^4} + \frac{{v_j}^6}{d^6}, & |v_j| \le d \\ 1\quad , & |v_j| > d \end{cases} \quad (8)$$

$$\varphi(v_j) = \tau'(v_j) = \begin{cases} v_j\left[1 - \left(\frac{v_j}{d}\right)^2\right]^2, & |v_j| \le d \\ 0, & |v_j| > d \end{cases} \quad (9)$$

$$w_j = \begin{cases} \left[1 - \left({}^{v_j}/_{d}\right)^2\right]^2, & |v_j| \le d \\ 0, & |v_j| > d \end{cases} \quad (10)$$

Generally weight function is defined by,

$$w_j = \frac{\varphi\left({}^{(x_j - \sum_{k=0}^{q} z_{jk}\theta_k)}/_{\mu_j}\right)}{\left({}^{(x_j - \sum_{k=0}^{q} z_{jk}\theta_k)}/_{\mu_j}\right)} \quad (11)$$

Using equation (11), equation (7) can be written as (weight is used from Tukey's bi-weight function),

$$\sum_{j=1}^{n} z_{jk} w_j\left(x_j - \sum_{k=0}^{q} z_{jk}\theta_k\right) = 0 \quad (12)$$

After solving the equation (12), it gives an estimator for θ that is;

$$\hat{\boldsymbol{\theta}}_{j+1} = (\boldsymbol{Z}'\boldsymbol{W}_j\boldsymbol{Z})^{-1}(\boldsymbol{Z}'\boldsymbol{W}_j\boldsymbol{X}) \quad (13)$$

It is an iterative process. At each step, the estimate value of θ is obtained. To begin the iteration process, estimated parameters using OLS are used. Then, the value of residual is obtained. Next step is to calculate the value of μ_j, v_j, w_j respectively. Then, new estimate of θ will be found by the equation (13). At next iteration; the same process will be repeated and get the another estimate of θ. At each iteration, a new estimate of θ will be obtained. Iteration will be continuing until the consecutive estimates are sufficiently close to one another.

Convergence Criterion: $\frac{\|\hat{\theta}_{j+1} - \hat{\theta}_j\|}{\|\hat{\theta}_{j+1}\|} < \epsilon$, Here, $\epsilon = 1.E - 8$.

3.2 S-estimation

S-estimation is based on the residual scale of M-estimation. S-estimation is the solution to overcome the drawback of M-estimation. S-estimation too has robustness properties viz. BDP, efficiency and bounded influence. This method uses the residual standard deviation to overcome the weakness of median. S-estimation is the generalization of least median squares which was introduced by Rousseeuw and Yohai (1984). It is defined as the minimization of dispersion of residuals;

$$\hat{\theta} = \arg\min V(\theta) \tag{14}$$

where $V(\theta)$ is the dispersion of residuals. The dispersion $V(\theta)$ is the solution of,

$$\frac{1}{n-q}\sum_{j=1}^{n}\tau\left(\frac{x_j-\sum_{k=0}^{q} z_{jk}\theta_k}{V}\right)=\alpha \tag{15}$$

Two types of function for τ can be used which are-

1. Tukey's bi-square function
2. Yohai's optimal function

Here, Tukey's bi-square function has been used as τ function which satisfies all the properties as mentioned in M-estimation. That is;

$$\tau(v_j) = \begin{cases} \frac{3v_j^{\,2}}{k_0^{\,2}} - \frac{3v_j^{\,4}}{k_0^{\,4}} + \frac{v_j^{\,6}}{k_0^{\,6}}, & |v_j| \le k_0 \\ 1 \quad , & |v_j| > k_0 \end{cases} \tag{16}$$

$$\varphi(v_j) = \tau'(v_j) = \begin{cases} v_j\left[1-\left(\frac{v_j}{k_0}\right)^2\right]^2, & |v_j| \le k_0 \\ 0, & |v_j| > k_0 \end{cases} \tag{17}$$

$$w_j = \begin{cases} \left[1-\left({}^{v_j}/_{k_0}\right)^2\right]^2, & |v_j| \le k_0 \\ 0, \; |v_j| > k_0 \end{cases} \tag{18}$$

where, $k_0 = 1.547$ which is known as tuning constant. It controls breakdown value and efficiency with 50% breakdown value. The final solution of eq. (15) will come from an iterative procedure which was given by Marazzi (1993).

4. APPLICATION

4.1 Data description

A data of major baseball players has been taken from the Baseball, Encyclopedia (9th edition, Macmillan). Analysis has been done using SAS software. Here, one dependent variable and five explanatory variables have been used. They are:

Table 1. Data of major baseball players

X	Z_1	Z_2	Z_3	Z_4	Z_5
0.283	0.144	0.049	0.012	0.013	0.086
0.276	0.125	0.039	0.013	0.002	0.062
0.281	0.141	0.045	0.021	0.013	0.074
0.328	0.189	0.043	0.001	0.03	0.032
0.29	0.161	0.044	0.011	0.07	0.076
0.296	0.186	0.047	0.018	0.05	0.007
0.248	0.106	0.036	0.008	0.012	0.095
0.228	0.117	0.03	0.006	0.003	0.145
0.305	0.174	0.05	0.008	0.061	0.112
0.254	0.094	0.041	0.005	0.014	0.124
0.269	0.147	0.047	0.012	0.009	0.111
0.3	0.141	0.058	0.01	0.011	0.07
0.307	0.135	0.041	0.009	0.005	0.065
0.214	0.1	0.037	0.003	0.004	0.138
0.329	0.189	0.058	0.014	0.011	0.032
0.31	0.149	0.05	0.012	0.05	0.06
0.252	0.119	0.04	0.008	0.049	0.233
0.308	0.158	0.038	0.013	0.003	0.068
0.342	0.259	0.06	0.016	0.085	0.158
0.358	0.193	0.066	0.021	0.037	0.083
0.34	0.155	0.051	0.02	0.012	0.04
0.304	0.197	0.052	0.008	0.054	0.095
0.248	0.133	0.037	0.003	0.043	0.135
0.367	0.196	0.063	0.026	0.01	0.031
0.325	0.206	0.054	0.027	0.01	0.048
0.244	0.11	0.025	0.006	0	0.061
0.245	0.096	0.044	0.003	0.022	0.151
0.318	0.193	0.063	0.02	0.037	0.081
0.207	0.154	0.045	0.008	0	0.252
0.32	0.204	0.053	0.017	0.013	0.07
0.243	0.141	0.041	0.007	0.051	0.264
0.317	0.209	0.057	0.03	0.017	0.058
0.199	0.1	0.029	0.007	0.011	0.188
0.294	0.158	0.034	0.019	0.005	0.014
0.221	0.087	0.038	0.006	0.015	0.142
0.301	0.163	0.068	0.016	0.022	0.092
0.298	0.207	0.042	0.009	0.066	0.211
0.304	0.197	0.052	0.008	0.054	0.095
0.297	0.16	0.049	0.007	0.038	0.101
0.188	0.064	0.044	0.007	0.002	0.205
0.214	0.1	0.037	0.003	0.004	0.138
0.218	0.082	0.061	0.002	0.012	0.147
0.284	0.131	0.049	0.012	0.021	0.13
0.27	0.17	0.026	0.011	0.002	0
0.277	0.15	0.053	0.005	0.039	0.115

Dependent variable (X): Batting Average

Explanatory Variables(Z):

Z_1: Runs scored/times at bat

Z_2: Doubles/times at bat

Z_3: Triples/times at bat

Z_4: Home runs/times at bat

Z_5: Strike outs/times at bat

4.2 Test for Multicollinearity

While working with the multiple regression, multicollinearity is a major problem. Many assumptions of regression may be violated in the presence of multicollinearity. So, identification of multicollinearity is inevitable before going through the analysis. There are different methods for identification of multicollinearity. Most used methods are:

Variance inflation factor(VIF).

Eigen system analysis of $Z'Z$.

Here, VIF method have been utilized to identify multicollinearity. The thumb rule is:

If VIF > 4 - Need further investigation

VIF > 10 - There is serious multicollinearity.

Table 2. VIF values of variables

Variable	VIF
X	0
Z_1	3.05
Z_2	1.54
Z_3	2.35
Z_4	2.04
Z_5	1.53

From Table 2, it is concluded that there is no or very little multicollinearity because VIF values are in the range of 0-3.05 for all the variables which is less than 4. If there was multicollinearity, then one should go for principal component regression (PCR) or ridge regression or any other remedies to cope up with the multicollinearity.

4.3 Detection of Influential Observation and Leverage Points

Cook's distance (Cook, 1977) is widely used method to detect influential observations that affect the values of fitted model. Cook's distance involves with both response and explanatory variables' observation.

Table 3. Cook's D values to detect influential observation

Observation number	Cook's D(>0.10)
4	0.104
6	0.212
29	0.259
42	0.183
44	0.116

Here, Cook's distance has been used to detect the influential observation. From Table 3 and Fig. 1; It is seen that the influential observations are 4, 6, 29, 42, 44. It is also seen from table 4 that nine observations are leverage points.

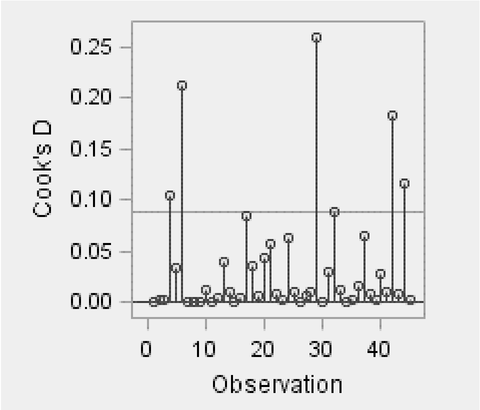

Fig. 1. Plot for the detection of influential observations

Table 4. Observations with leverage points

Observations	Maholanobis Distance	Leverage (*)
4	3.4498	*
17	2.6425	*
19	3.3339	*
29	3.8935	*
31	2.8293	*
37	3.0604	*
40	2.6257	*
42	3.3657	*
44	3.1269	*

4.4 Comparison of three methods

In this paper, the performance of S-estimation is compared with the OLS and M-estimation. Adjusted R^2 has been used to test the goodness of fit of the models. The best model will have largest adjusted R^2 value and smallest Mean Absolute Percentage Error (MAPE). From Table 5; it is observed that S-estimation has the largest adjusted R^2 and lowest MAPE than M-estimation and OLS method. So, S-estimation has performed better than the others two methods.

Table 5. Performance of the three methods

Methods	Adjusted R^2	MAPE	Significant Variables
OLS	0.8461	7.52	Z_1, Z_2, Z_5
M Estimation	0.7431	7.81	Z_1, Z_2, Z_5
S Estimation	0.8799	7.46	Z_1, Z_2, Z_5

The fitted regression model using OLS-

$$\hat{x} = 0.1831 + 0.4466z_1 + 0.9909z_2 + 0.62160z_3 + 0.2737z_4 - 0.2845z_5 \quad (19)$$

The fitted regression model using M-estimation-

$$\hat{x} = 0.1853 + 0.4343z_1 + 0.9678z_2 + 0.6524z_3 + 0.3230z_4 - 0.2939z_5 \quad (20)$$

The fitted regression model using S-estimation-

$$\hat{x} = 0.2097 + 0.3991z_1 + 1.1354z_2 - 0.4338z_3 + 0.1512z_4 - 0.4462z_5 \quad (21)$$

Table 6. Standard error of significant parameter estimates for all the models

METHODS	Standard error of parameter estimates			
	Intercept	Z_1	Z_2	Z_5
OLS	0.0171	0.1096	0.3130	0.0517
M Estimation	0.0179	0.1147	0.3276	0.0542
S Estimation	0.0112	0.0732	0.2342	0.0402

From Table 6; Standard errors of parameter estimates (significant variables) are meaningfully lower in case of S-estimation as compared to OLS and M-estimation. Again, the regression model has been estimated with Z_1, Z_2, Z_5 using S-estimation. The fitted regression model-

$$\hat{x} = 0.2052 + 0.4363z_1 + 1.019z_2 - 0.4281z_5 \quad (22)$$

The value of adjusted R^2 for the above estimated model is 0.8530. It means 85.30% variability of the response variable is explained by Z_1, Z_2 and Z_5. All the variables are found to be significant at 1% level of significance.

5. CONCLUSION

In this paper, OLS, M-estimation and S-estimation have been discussed briefly. From the results, it is seen that M estimator shows lower adjusted R^2 values and higher MAPE as compared to the OLS in spite of being a robust regression method. As discussed in earlier sections, M-estimator does not show its' robustness properties when it deals with the leverage points as this estimator does not have the property of bounded influence. So, even one bad leverage point can be the enough reason for entirely break down of the model fitting. In this data, nine leverage points are exist, so that M estimator shows very poor results, even inferior to OLS. That means, M estimator is not robust in all the cases especially in the presence of leverage and the performance of M estimator can be inferior to OLS in the presence of many leverage points. On the other hand, S estimator has shown its' worthy performance with higher adjusted R^2, lower MAPE and lower standard error values of parameter estimates as compared to both OLS and M estimator as it has all the robustness properties including the bounded influence. From the final result, it can observe that batting average depends upon run scored/times at bat, doubles/times at bat and strike out/times at bat. Eventually, it is concluded that S-estimation performs better than OLS and M-estimation in the presence of leverage points.

REFERENCES

Andersen, R. (2008). Modern Methods for Robust Regression. Thousand Oaks: SAGE Publications.

Chen, C. (2002). Robust Regression and Outlier Detection with the ROBUSTREG Procedure, *Statistics and Data Analysis*, paper 265-27, SAS Institute Inc., Cary, NC.

Cook, R.D. (1977). Detection of influential observations in linear regression. *Technometrics*, **19**, 15-18.

Draper, N.R. and Smith, H. (2014). *Applied regression analysis*. John Wiley & Sons.

Hampel, F.R. (1975). Beyond location parameters: Robust concepts and methods. *Bull. Int. Statist. Inst.*, ***46***, 375-382.

Huber, P.J. (1973). Robust regression: Asymtotics, conjectures and Monte carlo, *Ann. Stat.*, **1**, 799-821.

Huber, P.J. (1981). Robust Statistics, John Wiley and Sons, New York.

Marazzi, A. (1993), *Algorithm, Routines, and S Functions for Robust Statistics*, Pacific Grove, CA: Wadsworth and Brooks/Cole.

Rousseeuw, P.J. (1984). Least Median Squares Regression. *J. Amer. Statist. Assoc.*, **79**, 871-880.

Rousseeuw, P. and Yohai, V. (1984). Robust regression by means of S-estimators. *In Robust and nonlinear time series analysis*(pp. 256-272), Springer, New York, NY.

The Baseball Encyclopedia. Ninth edition by Macmillan publishing company (1993).