Research Journal of Mathematical and Statistical Sciences _____ISSN 2320-6047
Vol. **7(3),** 19-21, September **(2019)** *Res. J. Mathematical and Statistical Sci.*

Review Paper

Quantile regression versus regression diagnostics in presence of outliers

Samhita Pal
Department of Statistics, University of Calcutta, India
samhitapal3896@gmail.com

Available online at: www.iscamaths.com , www.isca.in , www.isca.me
Received 4th March 2019, revised 14th August 2019, accepted 5th September 2019

Abstract

Regression diagnostics are measures computed from the data in order to detect the influential points, following which the outliers can be corrected or removed and the ordinary least square regression may be fitted to the remaining observations. On the other hand, robust regression techniques try to devise estimators that are not so strongly affected by outliers by eliminating the effects of unusually high residuals due to the presence of outliers. This paper aims at comparing the two methods in a simulated data set containing few outlier values. Quantile regression serves as one of the many robust regression techniques.

Keywords: Quantile regression, goodness of fit, regression diagnostics, studentized.

Introduction

Outliers[1,2] are unusual observations in a data set which are either too large or too small compared to the rest of the data points. In the regression context (y_i , x_i), i=1(1)n, outliers can be of three types – i. outlier in y, ii. outlier in x (leverage points), iii. outlier in both x and y (influential points).

Least squares regression fails o properly explain a dataset when some points in the data have either excessively large or small values in comparison with the other data points. This is mainly because the OLS regression minimizes the sum of the squared errors and hence any point whose dependent value differs hugely from the other data points will have a disproportionately large effect on the resulting constants that are being solved for.

Let us consider the model $y_i=\beta_1+ \beta_2x_i+u_i$, where u_i is the deviation of response values from the PRF. u_i's are unobservable random variables taking +ve or –ve values. u_i's are called stochastic disturbance or error term. The SRF approximates the PRF based on the sample observations. It is given by $Y_i = \hat{\beta}_1 + \hat{\beta}_2 x_i$, where $\hat{\beta}_1$ and $\hat{\beta}_2$ are respectively the LS estimates of β_1 and β_2.

As can be seen from Figure-1, the LS regression line has shifted towards the outliers giving a poor fit of r^2=0.01219. As a result, the OLS regression fails to provide a good fit to the sample observations. This is mainly because in least square regression, we model the conditional mean and the mean is affected by outliers.

Regression Diagnostics[3,4]**:** The OLS gives poor fit to the data with outliers because the residuals for the 5th and 91st observations are very large and contribute hugely to the sum of squared errors. From here, we can make out that residual terms play a vital role in identifying outliers. Regression diagnostics deal with identification of the outliers and then either correcting or removing the outlier points that influence the OLS regression. However, the method of deleting the influential points and fitting LS regression to the new data works well only when there are one to two outliers.

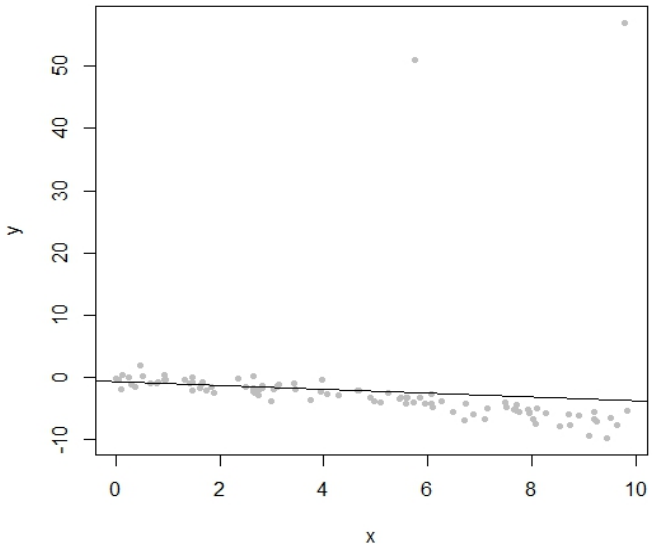

Figure-1: OLS regression line in presence of outliers.

A measure for detecting outliers is the Studentized residuals[5,6] $t_i^* = \frac{e_i}{\hat{\sigma}_i\sqrt{1-h_{ii}}}$, where e_i is the residual corresponding to the i[th] response, i.e. $y_i - (\widehat{\beta_1} + \widehat{\beta_2}x_i)$, $\hat{\sigma}_i = \frac{1}{n-1}\sum_{k\neq i}(y_k - \widehat{\beta_{1(i)}} - \widehat{\beta_{2(i)}}x_i)^2$, h_{ii} = i[th] diagonal element of the hat matrix $H=X(X^TX)^{-1}X^T$. Also $\widehat{\beta_{k(i)}}$, k=1,2 are the LS estimators obtained by

deleting the i^{th} observation. If $|t_i^*|$ exceeds 2, we conclude that the i^{th} observation is an outlier in y.

In our simulated data, studentized t value for i=5 is 3.4957783789 and for i=91 is 2.6771819796.

The r^2 value obtained by fitting y on x after removing the outliers, i.e. the 5^{th} and 91^{st} observations is given by 0.8364. The new regression line is y = 0.2471- 0.7511*x.

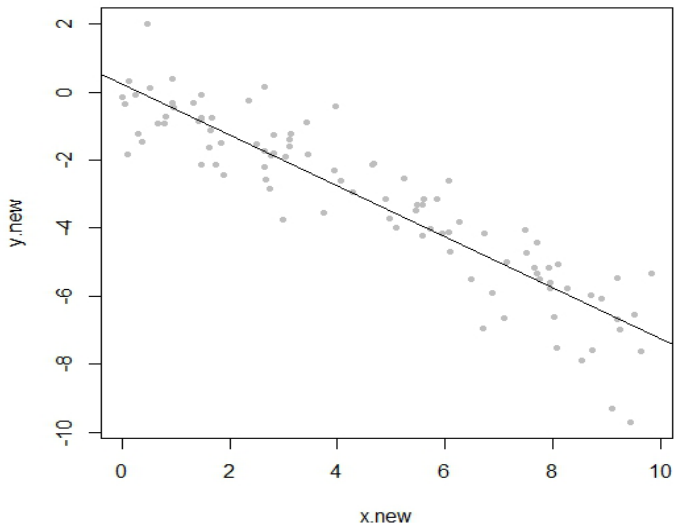

Figure-2: Scatter plot and regression line of the data after removal of the outliers.

Quantile Regression[7-10]**:** If e_i is the model prediction error, OLS minimizes $\sum_i e_i^2$, whereas median regression (also known as least absolute deviation/LAD regression[11] minimizes $\sum_i |e_i|$. This is also a robust regression technique which is unaffected by outliers. Quantile regression minimizes a sum that gives asymmetric penalties $(1-\tau)e_i$ for over-prediction and τe_i for under-prediction. The quantile regression model gives the relationship between the regressor or predictor and the conditional quantile function $Q_\tau(y|x) = \inf\{y \mid F_Y(y|x) \geq \tau\}$. Assuming linear relationship, we have $Q_\tau(y|x) = \beta_1 + \beta_2 x_i$ where β_1 and β_2 are unknown constants which are to be estimated by minimizing the asymmetric loss function[12]

$$\Rightarrow \min_{\hat{\beta}1,\hat{\beta}2}[\sum_{y_i \geq \hat{\beta}_1+\hat{\beta}_2 x_i} \tau\,|\, y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i \,|\, + \sum_{y_i < \hat{\beta}_1+\hat{\beta}_2 x_i} (1-\tau)\,|\, y_i - \hat{\beta}_1 - \beta$$

sub to $\underset{\sim}{Y} - \underset{\sim}{X}\,\underset{\sim}{\beta}$ = p-q. Solving the above simplex, we can find the estimates of β_1 and β_2.

The goodness of fit[13] in quantile regression is motivated by the r^2 of classical least square regression and is given by

$R(\tau) = 1 - \frac{\hat{V}(\tau)}{V(\tau)}$, where

$$\hat{V}(\tau) = \frac{1}{n}[\sum_{i=1}^{n} |\, y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i \,| + (2\tau - 1)\sum_{i=1}^{n} (y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i)]$$

& $V(\tau) = \frac{1}{n}[\sum_{i=1}^{n} |\, y_i - \xi_\tau \,| + (2\tau - 1)\sum_{i=1}^{n} (y_i - \xi_\tau)]$, ξ_τ is the τ^{th} sample quantile of y. $R(\tau)$ should lie in [0,1] as $\hat{V}(\tau)$ is always less than $V(\tau)$, where 1 would indicate a perfect fit since the numerator $\hat{V}(\tau)$, which represents the weighted sum of deviations would be zero. It is a local measure of fit for QRM as it depends upon τ, unlike the global measure r^2 in OLS.

We fit the quantile regression line to the same data with two outliers for the median. This is equivalent to the LAD regression. The sample quantile regression equation for modeling the population conditional median is given by y=0.221-0.71307 x with the goodness of fit measure given by R^2= 0.3922186 (sample median = - 2.59644).
From Figure-3, it is evident that the median regression fits the data far better as compared to mean regression.

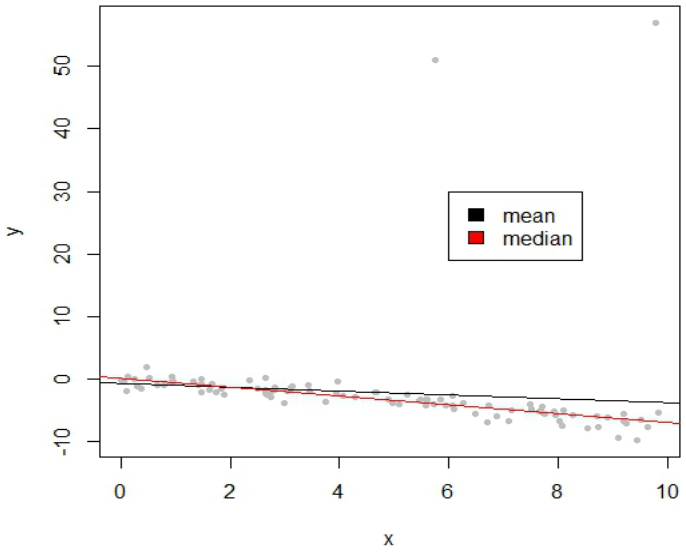

Figure-3: OLS regression and Quantile regression lines.

Results and discussions

The r-squared value of the OLS fitted to the original data containing two outlier points was as low as 0.01219. On identifying the outliers and removing them from the dataset, OLS was fitted again to the new data giving a much improved r-square value of 0.8364. On the other hand, on fitting the median regression (which is often referred to as robust regression as it is robust o outlier values), we find the goodness of fit of the fitted line is close to 0.4. Although apparently the OLS regression in the cleansed data seems to give a better fit, it should be kept in mind that two data points had to be eliminated to achieve this

fit. Had there been more outlier or influential points in the data, elimination would not have been the best option for analyzing the data set as outlier points might also contain important information regarding behavior of the variables. Quantile regression fitted for the 0.5 quantile does not require any data points to be removed, but still gives a moderate fit to the clustered data points. The outliers do not affect the estimation procedure of the model parameters and hence is said to be robust. The outlier values can later be studied individually.

Conclusion

From the comparative analysis, it is easy to see that robust regression such asquantile regression surpasses the techniques of regression diagnostics when the data contains outliers; more so when the number of outliers exceeds 2 or 3. No data points are lost when robust regression methods are followed which facilitates the future study of the influential points. However, if OLS is the required regression setup, then regression diagnostics surely prove to be very useful.

References

1. NG Damodar (2004). Heteroscedasticity. In: Basic Econometrics, Fourth edition, The McGraw-Hill Companies, US.

2. Vic Barnett Toby Lewis (1974). Outliers in Statistical data. Third edition, Wiley, US.

3. Galen Bollinger (1981). Book Review: Regression Diagnostics: Identifying Influential Data and Sources of Collinearity. *Journal of Marketing Research*, 18(3), 392-393.

4. R Dennis Cook and Sanford Weisberg (1989). Regression diagnostics with dynamic graphics. *Technometrics*, 31(3), 277-291.

5. Ranganai E. (2016). On studentized residuals in the quantile regression framework. *SpringerPlus*, 5(1), 1231.

6. Balasooriya U. and Tse Y.K. (2007). Outlier detection in linear models: a comparative study in simple linear regression. *Communications in Statistics- Theory and Methods*, 15(12), 3589-3597.

7. Roger Koenker (2005). Quantile Regression. Illinois, US: Cambridge University Press.

8. Roger Koenker and Kevin F Hallock (2001). Quantile Regression: An Introduction. *Journal of Economic Perspectives,* 15(4), 143-156.

9. Koenker R. and Bassett G. (1978). Regression quantiles. *Econometrika*, 46, 33-50.

10. Mata J. and Machado J.A. (1996). Firm start-up size: A conditional quantile approach. *European Economic Review*, 40(6), 1305-1323.

11. Kahraman U.M. and Iyit N. (2018). Performance of LAD Regression, M-Regression and Quantile Regression Methods in order to Investigate Stock Prices of the Banks in the BIST Bank Index. *International Journal of Scientific Research and Management*, 6(4), 266-273.

12. Christoffersen P.F. and Diebold F.X. (1997). Optimal prediction under asymmetric loss. *Econometric theory*, 13(6), 808-817.

13. Koenker R. and Machado J.A. (1999). Goodness of fit and related inference processes for quantile regression. *Journal of the american statistical association*, 94(448), 1296-1310.