International Journal of Systems Science and Applied Mathematics
2020; 5(1): 4-11
http://www.sciencepublishinggroup.com/j/ijssam
doi: 10.11648/j.ijssam.20200501.12
ISSN: 2575-5838 (Print); ISSN: 2575-5803 (Online)

Review of Outlier Detection and Identifying Using Robust Regression Model

Getnet Bogale Begashaw[1, *], Yordanos Berihun Yohannes[2]

[1]Department of Statistics, College of Natural Science, Wollo University, Dessie, Ethiopia

[2]Department of Statistics, College of Natural and Computational Science, Salale University, Fitche, Ethiopia

Email address:

getnetbogale145@gmail.com (G. B. Begashaw), Yordanosberihun21@gmail.com (Y. B. Yohannes)

[*]Corresponding author

To cite this article:

Getnet Bogale Begashaw, Yordanos Berihun Yohannes. Review of Outlier Detection and Identifying Using Robust Regression Model. *International Journal of Systems Science and Applied Mathematics.* Vol. 5, No. 1, 2020, pp. 4-11. doi: 10.11648/j.ijssam.20200501.12

Received: October 25, 2019; **Accepted**: November 23, 2019; **Published**: April 13, 2020

Abstract: Outliers are observations that have extreme value relations. Herewith leverage is a measure of how an independent variable deviates from its mean. An observation with an extreme value on a predictor variable is a point with high leverage. The presence of outliers can lead to inflated error rates and substantial distortions of parameter and statistic estimates when using either parametric or nonparametric tests. Casual observation of the literature suggests that researchers rarely report checking for outliers of any sort and taking remedial measures for outliers. Outliers can have positive deleterious effects on statistical analyses. For instance, they serve to increase error variance and reduce the power of statistical tests; they can decrease normality, altering the odds of making both Type I and Type II errors for non- randomly distributed; and they can seriously bias or influence estimates that may be of substantive interest. These outliers are cased from incorrect recording data, intentional or motivated mis-reporting, sampling error and Outliers as legitimate cases sampled from the correct population. According to some literatures; Point outliers, Contextual Outliers and Collective Outliers are the three types of outliers. Robust regression estimators can be a powerful tool for detection and identifying outliers in complicated data sets. Robust regression, deals with the problem of outliers in a regression and produce different coefficient estimates than OLS does.

Keywords: Break Down Point, Leverage Points, M-estimation, Outlier, Robust Regression Model

1. Introduction

"Outliers" are unusual data values that occur almost in all research projects involving data collection. Outliers are observations that have extreme value relations. The term outlier is defined as follows:

1. …Data which are far away from the bulk of the data, or more generally, from the pattern set by the majority of the data. [5]
2. …Data point that is far outside the norm for a variable or population. [7]

An outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism. [10]

Beside this it is better to point out some definition about terms which is going hand by hand with outlier as follows. Leverage is a measure of how an independent variable deviates from its mean. An observation with an extreme value on a predictor variable is a point with high leverage. Influence An observation is said to be in influential if removing that observation substantially changes the estimation of the coefficients. A datum that is "influential" is one for which the regression estimate changes considerably if it is removed. Rejection Point is the point beyond which the influence function becomes zero. That is the contribution of the points beyond the rejection point to the final estimate is comparatively legible.

The presence of outliers can lead to inflated error rates and substantial distortions of parameter and statistic estimates when using either parametric or nonparametric tests. Casual observation of the literature suggests that researchers rarely report checking for outliers of any sort.

Outliers can have deleterious effects on statistical analyses. First, they generally serve to increase error variance and reduce the power of statistical tests. Second, if non-

randomly distributed they can decrease normality (and in multivariate analyses, violate as assumptions of sphericity and multivariate normality), altering the odds of making both Type I and Type II errors. Third, they can seriously bias or influence estimates that may be of substantive interest. [1]

There is a great deal of debate as to what to do with identified outliers. A thorough review of the various arguments is not possible here. We argue that what to do depends in large part on why an outlier is in the data in the first place. Where outliers are illegitimately included in the data, it is only commons ensue that those data points should be removed. Few should dis agree with that statement. When the outlier is either a legitimate part of the data or the cause is unclear, the issue becomes make several strong points for removal even in these cases in order to get the most honest estimate of population parameters possible.

However, not all researchers feel that way. [8] This is a case where researchers must use their training, intuition, reasoned argument, and thoughtful consideration in making decisions. Researchers sometimes use various "robust" procedures to protect their data from being distorted by the presence of outliers. These techniques "accommodate the outliers at no serious inconvenience or are robust against the presence of outliers." Certain parameter estimates, especially the mean and Least Squares estimations, are particularly vulnerable to outliers, or have "low break down" values.

For this reason, researchers turn to robust or "high break down" methods to provide alternative estimates for these important aspects of the data. The practical use of the outlier here considered methods is always the crucial point within this work. Due to this reason in this article we are emphasis this outlier detection using new regression model called Robust Regression Model. Not all outliers are illegitimate contaminants and not all illegitimate scores show up as outliers. [11]

2. Review About Outliers

2.1. Causes of Outliers

Outliers from incorrect recording data: Outliers are often caused by human error, such as errors in data collection, recording or entry. Data from an interview can be recorded incorrectly, or mistaken upon data entry. Errors of this nature can often be corrected by returning to the original documents or even the subjects if necessary and possible and entering the correct value.

Outliers from intentional or motivated mis-reporting: There are times when participants purposefully report incorrect data to experimenters or surveyors. A participant may make conscious effort to sabotage the research [16] or may be acting from other motives. Social desirability and self-presentation motives can be powerful. This can also happen for obvious reasons when data are sensitive.

Outliers from sampling error: Another cause of outliers is sampling. It is possible that a few members of a sample were inadvertently drawn from a different population than the rest of the sample. For-example: in education, in advert entry sampling academically gifted or mentally retorted students is a possibility and (depending on the goal of the study) might provide undesirable outliers. These cases should be removed as they do not reflect the target population.

Outliers as legitimate cases sampled from the correct population:

It is possible that an outlier can come from the population being sampled legitimately through random chance, it is important to note that sample size plays a role in the probability of outlying values. Within a normally distributed population, it is more probable that a given data point will be drawn from the most densely concentrated area of the distribution, rather than one of the tails. [9] As a researcher casts a wider net and the data set becomes larger, the more the sample resembles the population from which it was drawn and thus the likelihood of outlying values become greater. In other words, there is only about one percentage chance you will get an outlying data point from a normally distributed population, this means that, on the average, about one percentage of your subjects should be three standard deviations from the mean.

2.2. Types of Outliers

Outliers can be classified in to three categories based on its composition and relation to rest of the data.

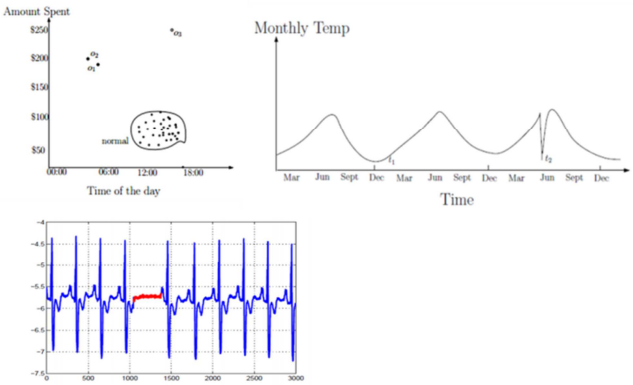

Figure 1. *Types of outlier.*

(A) Point outliers.
(B) Contextual Outliers.
(C) Collective Outliers: human ECG output corresponding to an Atrial Premature Contraction.

Point outliers: If an individual data point can be considered anomalous with respect to the rest of the data, then the datum is termed as a point outlier. This is the simplest type of outlier and it is the focus of the majority of research on outlier detection. A data instance is an outlier due to its attribute values which are inconsistent with values taken by normal instances.

From the above figure point o_1. o_2 and o_3 are considered to outlier.

Contextual outliers: These outliers are caused due to the occurrence of an individual data instance in a specific context

in the given data. Like point outliers, these outliers are also individual data instances. The difference is that a contextual outlier might not be an outlier in a different context. Thus contextual outliers are defined with respect to a context. contextual outliers satisfy two properties.

The underlying data has a spatial/ sequential nature: each data instance is defined using two sets of attributes, with contextual attributes and behavioral attributes. The contextual attributes define the position of an instance and are used to determine the context (or neighborhood) for that instance. For example, in spatial data sets, the longitude and latitude of a location are the contextual attributes.

The outlying behavior is determined using the values for the behavioral attributes within a specific context. A data instance might be a contextual outlier in a given context, but an identical data instance (in terms of behavioral attributes) could be considered normal in a different context. Contextual outliers have been most popularly explored in time-series data.

From figure above outlier t_2 in a temperature time series. Note that the temperature at time t_1 is same as that at time t_2 but occurs in a different context and hence is not considered as an outlier.

Collective outliers: If a collection of data points is anomalous with respect to the entire data set, it is termed as a collective outlier. The individual data points inside the collective outlier may not be outliers by themselves alone, but their occurrence together as a collection is anomalous. Collective outliers can occur only in data sets in which data points are somehow related. These outliers occur because a subset of data instances is outlying with respect to the entire data set. The individual data instances in a collective outlier are not outliers by themselves, but their occurrence together as a substructure is anomalous. collective outliers are meaningful only when the data has spatial or sequential nature. These outliers are either anomalous sub graphs or subsequences occurring in the data.

Principal options for dealing with Outlier:

Analyze the relevant data (i.e. removing outlier from the data) in order to get the most honest estimate of population parameters possible; [2]

On means of accommodating outliers is the use of transformations. By using transformation extreme scores can be kept in the data set, and the relative ranking of scores remains yet the skew and error variance present in the variable can be recorded [1].

Use various robust procedures to protect their data from being distorted by the presence of outliers.

Remark:

Option 1 evidence of outliers may produce type I or type II errors. Removal of outliers may tend to have a significant beneficial effect on error rates. Both correlations and t-tests may Show significant changes in statistics as a function of removal of outliers. It is advisable to fit the regression model with and without outliers. Then check the differences. Option 2 may not be appropriate for the model being tested or may affect its interpretation in undesirable ways. Taking the logarithms of a variable makes a distribution less skewed, but it also alters the relationship between the original variables in the model. [4] Option 3 accommodate the outliers at no serious inconvenience or are robust against the presence of outliers. [6] Certain parameter estimates, especially the mean and least square estimates, are particularly vulnerable to outliers, or have “low breakdown” values.

Effectively working with outliers in numerical data can be a rather difficult and frustrating experience. Neither ignoring nor deleting them at will is good solutions. If you do nothing, you will end up with a model that describes essentially none of the data, neither the bulk of the data nor the outliers.

Developing techniques to look for outliers and understanding how they impact data analysis are extremely important part of a thorough analysis, especially when statistical techniques are applied to the data. For example, in the procedure of outliers, any statistical test based on sample means and variances can be distorted. Estimated regression coefficients that minimize the sum of squares for error (SSE) are very sensitive to outliers.

2.3. Outliers and OLS

The method of Ordinary Least Squares (OLS) is the most frequently applied regression Technique. The application of this specific method requires several assumptions. Every researcher is aware of the fact that the OLS method performs poorly if these assumptions are not fulfilled. But in particular outlying observation observations within the data can cause violations of model assumptions and thereby can have huge impact on regression results. The intention of this article is to examine technically the effect of outliers on OLS Regression and to alternative Regression Techniques. The practical use of the outlier here considered methods is always the crucial point within this work.

The disturbance terms should be distributed independently and identically (i.i.d.), this distribution should be a normal one. This independent distribution requires independence among the disturbances (non-autocorrelation) and independence from the regressor variables. The disturbances are distributed identically as they follow a normal distribution with a common mean (zero) and a common variance (homoscedasticity). Violations of these assumptions can cause deviations in the underlying distribution, e.g. hetroscedasticity among the error terms becomes visible as “fat tails” in the underlying distribution. 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 will inflate the error variance. They inflate the standard errors. The confidence interval becomes stretched. The estimation cannot become asymptotically consistent.

One of the purposes of this work is it to point out the poor performance of OLS in the presence of outliers as a crucial error source in Regression Analysis. [7]

There are several assumptions that have to be fulfilled for the ordinary least squares regression model to be valid. When

the regression model does not meet the fundamental assumptions, the prediction and estimation of the model may become biased.

Limitations of the least squares estimator on outlier:

Although the least squares estimator is easy to calculate, it is also extremely sensitive to deviations from the model assumptions as a normal distribution is assumed for the errors. Hence, observations that are quite far from the majority of data can dramatically affect a least squares regression estimate. In the context of regression such gross errors are called `outliers'. It is important to classify the different ways in which data can be outlying, as each has different repercussions on the least squares estimator due to its asymmetry. Figure 2 highlights this: whilst in Figure 2 b the estimated regression line has been noticeably tilted, in Figure 2 d, as a result of the outlier, the line is approaching the normal to the original estimated line.

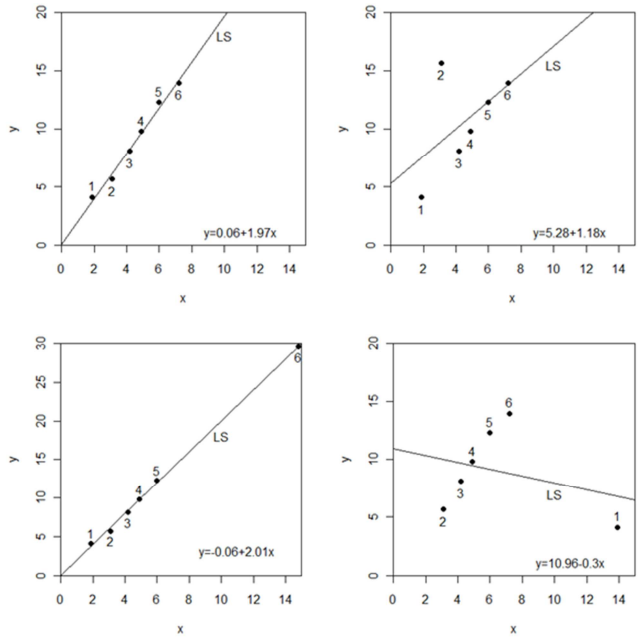

Figure 2. *limitation of OLS.*

(a) Six data with strong linear relationship.
(b) One datum replaced with an outlier in the y direction.
(c) One datum replaced with an outlier in the x direction.
(d) One datum replaced with a different sort of outlier in the x direction (a leverage point).

Outliers that bias the parameter estimates are those with leverage called bad leverage points.

Outliers that lie along the predicted line are those called good leverage points.

When outliers inflate the error variance, they sap the model of power to detect the outliers.

3. Outliers and Robust Regression Model

As the least squares method minimizes the average value of the squared residuals, this large residual is taken into consideration and has a strong influence on this average. The new regression line is strongly influenced by this residual and the new line of best fit is different from the original one. The new regression line tilts the large influence of the new residual and changes by that its original shape. But not only leverage points can be regression outliers. If the x-values but the y-value devotes in such a way, that the observation (x_i, y_i) is a regression outlier the according observation is called a "vertical" outlier or "outlier in the y direction". These outlying observations are influential on the least squares results as well, but with a less potential impact. Consequently, robust regression estimators can be a powerful tool for outlier detection in complicated data sets. [13, 14] Identifying multiple influential observations, even using very resistant regression estimators, becomes much harder due to two effects called `masking' and `swamping'. [15] Masking occurs when an outlying subset goes unnoticed because of the presence of another, whereas swamping refers to good observations being identified as outliers because of a remote subset of influential observations.

Robust regression, deals with the problem of outliers in a regression. Robust regression uses a weighting scheme that causes outliers to have less impact on the estimates of regression coefficients. Hence, robust regression generally will produce different coefficient estimates than OLS does.

3.1. Robust Regression Methods

Least absolute residuals (LAR) regression: is one of the most widely used robust regression procedures. It is insensitive to both outlying data values and inadequacies of the model employed. The method of least absolute residuals estimates the regression coefficients by minimizing the sum of the absolute deviations of the Y observations from their means. The criterion to be minimized, denoted by L_1, is

$$\sum_i^n |yi - (\beta 0 + \beta 1 Xi1 + \cdots \ldots \beta p - 1 Xip - 1)|. \qquad (1)$$

This method places less emphasis on outlying observations than does the method of least squares.

Iteratively Reweighted Least Squares (IRLS) robust regression: uses the weighted least squares procedures discussed in to dampen the influence of outlying observations. Instead of weights based on the error variances, IRLS robust regression uses weights based on how far outlying a case is, as measured by the residual for that case. The weights are revised with each iteration until a robust fit has been obtained.

Least Median of Squares (LMS) regression: minimizes the median squared residuals [12]. Since it focuses on the median residual, up to half of the observations can disagree without masking a model that fits the rest of the data. Replaces the sum of squared deviations in ordinary least squares by the median of the squared deviations, which is a robust estimator of location. The criterion for this procedure is to minimize the median squared deviation: median $\{[y_i-(\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\ldots\ldots+\beta_{p-1}x_{ip-1})]^2\}$ with respect to the regression coefficients. Thus, this procedure leads to estimated regression coefficients b_0, b_1, b_2….. b_{p-1} that minimizes the median of the squared residuals.

Other Robust Regression Procedures: There are many other robust regression procedures. Some involve trimming one or several of the extreme squared deviations before applying the least squares criterion; others are based on ranks. Many of the robust regression procedures require extensive computing.

3.2. Estimation on Robust Regression

Robust regression is a regression method that is used when the distribution of residual is not normal or there are some outliers that affect the model.

M-estimation: is the simplest approach both computationally and theoretically [7]. It is an extension of the maximum likelihood method. Although it is not robust with respect to leverage points, it is still used extensively in analyzing data for which it can be assumed that the contamination is mainly in the response direction.

Weakness: M-estimation is the lack of consideration on the data distribution and not a function of the overall data because only using the median as the weighted value.

S estimation: is a high breakdown value method [12]. With the same breakdown value, it has a higher statistical efficiency than LTS estimation. S estimation is based on residual scale of M-estimation. This method uses the residual standard deviation to overcome the weaknesses of median.

MM estimation: combines high breakdown value estimation and M-estimation [15]. It has both the high breakdown property and a higher statistical efficiency than S estimation. MM estimation procedure is to estimate the regression parameter using S estimation which minimize the scale of the residual from M-estimation and then proceed with M-estimation. MM estimation aims to obtain estimates that have a high breakdown value and more efficient.

Breakdown value is a common measure of the proportion of outliers that can be addressed before these observations affect the model.

4. Statistical Results and Discussions

Let's begin our discussion on robust regression with application on statistical software. Robust regression is applicable on different statistical software. In this section, we will show M-estimation with Huber and bisquare weighting on SAS and R. These two are very standard and are combined as the default weighting function in Stata's robust regression command. In Huber weighting, observations with small residuals get a weight of 1 and the larger the residual, the smaller the weight. With bisquare weighting, all cases with a non-zero residual get down-weighted at least a little.

4.1. Dealing with Robust Regression Using SAS

M-estimation: is a commonly used method for outlier detection and robust regression when contamination is mainly in the response direction. Proc robustreg in SAS command implements several versions of robust regression. The following example introduces the basic usage of the ROBUSTREG procedure. We used the following example to show how these robust techniques. The data is about national growth of 61 countries from 1960 to 1985.

Where the response variable is the GDP growth per worker (GDP) and the regressors are the constant term, labor force growth (LFG), relative GDP gap (GAP), equipment investment (EQP), and non-equipment investment (NEQ). The regression equation they used is

$$GDP = \beta_0 + \beta_1 LFG + \beta_2 GAP + \beta_3 EQP + \beta_4 NEQ + \varepsilon \quad (2)$$

By default, the procedure does M-estimation with the bisquare weight function, and it uses the median method for estimating the scale parameter. The MODEL statement specifies the covariate effects. The DIAGNOSTICS option requests a table for outlier diagnostics, and the LEVERAGE option adds leverage-point diagnostic results to this table for continuous covariate effects.

The following outputs are OLS procedure.

Table 1. *OLS estimate.*

Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	1	-0.01430	0.01028	-1.39	0.1697
LFG	1	-0.02981	0.19838	-0.15	0.8811
GAP	1	0.02026	0.00917	2.21	0.0313
EQP	1	0.26538	0.06529	4.06	0.0002
NEQ	1	0.06236	0.03482	1.79	0.0787

The OLS analysis of Table 1 indicates that GAP and EQP have a significant influence on GDP at the 5% level.

The following outputs are ROBUSTREG procedure with the default M estimation.

Table 2. *Model Fitting Information and Summary Statistics.*

Summary Statistics						
Variable	**Q1**	**Median**	**Q3**	**Mean**	**Standard Deviation**	**MAD**
LFG	0.0118	0.0239	0.0281	0.0211	0.00979	0.00949
GAP	0.5796	0.8015	0.8863	0.7258	0.2181	0.1778
EQP	0.0265	0.0433	0.0720	0.0523	0.0296	0.0325
NEQ	0.0956	0.1356	0.1812	0.1399	0.0570	0.0624
GDP	0.0121	0.0231	0.0310	0.0224	0.0155	0.0150

Remark:

The column labeled MAD provides a robust estimate of the univariate scale, which is computed as the standardized median absolute deviation (MAD). The columns labeled Mean and Standard Deviation provide the usual mean and standard deviation. A large difference between the standard deviation and the MAD for a variable indicates some extreme values for this variable.

Therefore the ROBUSTREG analysis of Table 2 indicates that there is no any variable that have extreme values on GDP at the 5% level.

Table 3. *M-estimates.*

Parameter Estimates							
Parameter	**DF**	**Estimate**	**Standard Error**	**95% Confidence Limits**		**Chi-Square**	**Pr > ChiSq**
Intercept	1	-0.0247	0.0097	-0.0437	-0.0058	6.53	0.0106
LFG	1	0.1040	0.1867	-0.2619	0.4699	0.31	0.5775
GAP	1	0.0250	0.0086	0.0080	0.0419	8.36	0.0038
EQP	1	0.2968	0.0614	0.1764	0.4172	23.33	<.0001
NEQ	1	0.0885	0.0328	0.0242	0.1527	7.29	0.0069
Scale	1	0.0099					

For the growth data, M estimation yields the fitted linear model:

$$\hat{Y} = -0.0247+0.1040X_1+0.0250X_2-0.2968X_3+0.0885X_4 \quad (3)$$

Table 4. *Diagnostics on M-estimation.*

Diagnostics					
Obs	**Mahalanobis Distance**	**Robust MCD Distance**	**Leverage**	**Standardized Robust Residual**	**Outlier**
1	2.6083	4.0639	*	-0.9424	
5	3.4351	6.7391	*	1.4200	
8	3.1876	4.6843	*	-0.1972	
9	3.6752	5.0599	*	-1.8784	
17	2.6024	3.8186	*	-1.7971	
23	2.1225	3.8238	*	1.7161	
27	2.6461	5.0336	*	0.0909	
31	2.9179	4.7140	*	0.0216	
53	2.2600	4.3193	*	-1.8082	
57	3.8701	5.4874	*	0.1448	
58	2.5953	3.9671	*	-0.0978	
59	2.9239	4.1663	*	0.3573	
60	1.8562	2.7135		-4.9798	*
61	1.9634	3.9128	*	-2.5959	

It also displays leverage points; however, there are no serious high leverage points. Beside this observation 60th is considered as an outlier on growth data on robust regression using M-estimation.

The ROBUSTREG Procedure

Table 5. *LTS estimates.*

LTS Profile	
Total Number of Observations	61
Number of Squares Minimized	33
Number of Coefficients	5
Highest Possible Breakdown Value	0.4590

LTS Parameter Estimates		
Parameter	**DF**	**Estimate**
Intercept	1	-0.0249
LFG	1	0.1123
GAP	1	0.0214
EQP	1	0.2669
NEQ	1	0.1110
Scale (sLTS)	0	0.0076
Scale (Wscale)	0	0.0109

Table 6. *Diagnostics.*

Diagnostics Profile		
Name	**percentage**	**Cutoff**
Outlier	0.0164	3.0000
Leverage	0.2131	3.3382

Based on the cutoff point for both outlier and leverage in Table 4 indicate that the 60th observation is outlier and some other

are considered as leverages.

Table 7. *Final Weighted LS estimates.*

Parameter Estimates for Final Weighted Least Squares Fit							
Parameter	**DF**	**Estimate**	**Standard Error**	**95% Confidence Limits**		**Chi-Square**	**Pr > ChiSq**
Intercept	1	-0.0222	0.0093	-0.0405	-0.0039	5.65	0.0175
LFG	1	0.0446	0.1771	-0.3026	0.3917	0.06	0.8013
GAP	1	0.0245	0.0082	0.0084	0.0406	8.89	0.0029
EQP	1	0.2824	0.0581	0.1685	0.3964	23.60	<.0001
NEQ	1	0.0849	0.0314	0.0233	0.1465	7.30	0.0069
Scale	0	0.0116					

$$\hat{Y} = -0.0222+0.0446X_1+0.0245X_{2+}0.2824X_3+0.0849X_4 \quad (4)$$

Equation 1: fitted model on LS estimate.

4.2. Dealing with Robust Regression Against Outlier Using R

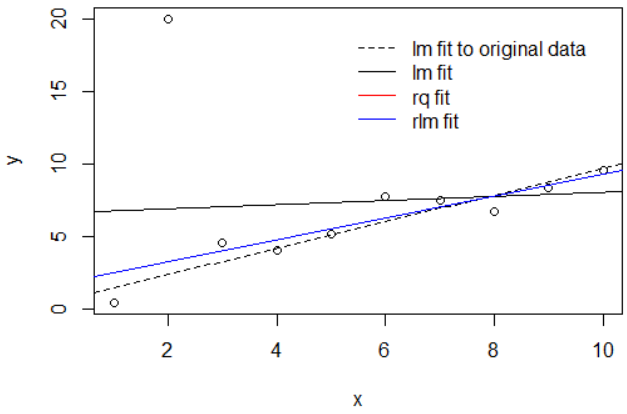

Figure 3. *Dealing with Robust Regression against outlier using R.*

Discussion: Based on the above graph: initially model fitting without outlier (Dash line) implies that fit the model without considering outlier (may be removed) but it has its own effect. While fitting model with outlier (Black line) implies that fitting model with considering many bad leverage points. It is risk for the given model, means that it needs some remedial needed to be taken. This leads to fit the model using robust regression model (Blue line). Robust doesn't consider influence of outlier (removal of outliers from the data), this leads that fitting model using robust regression is best way.

4.3. Discussions

Robust regression methods are not an option in most statistical software today. However, SAS, R (package is needed), PROC, NLIN etc can be used to implement iteratively reweighted least squares procedure. There are also Robust procedures available in S-Pluz. It is more advisable that using robust regression on outlier detection than removing or transforming the extreme data. More of all in this article we are dealing outlier on robust regression on SAS and R using M-estimation which is commonly popular.

5. Conclusions and Recommendations

5.1. Conclusions

One important fact to be noted is that Robust regression methods have much to offer a data analyst. They will be extremely helpful in locating outliers and highly influential observations. Whenever a least squares analysis is perfumed it would be useful to perform a robust fit also. If the results of both the fit are in substantial agreement, the use of Least Square Procedure offers a good estimation of the parameters. Robust regression techniques aiming to represent the majority of a sample can be extremely valuable in detecting data that would undermine the least squares estimator's performance. Increasingly sophisticated estimators have been proposed with ever more desirable properties. High breakdown point, high efficiency and bounded influence functions have been the main concerns. A key idea in the use of robust regression techniques is being able to rely on them not to be influenced by individual observations or subtends in the data, so that if the least squares estimator and the robust estimator coincide, the least squares estimator can be considered reliable. Furthermore, if the two estimates are different one needs to know if the robust estimate is actually representing the majority of the data or if it too may have been negatively influenced. This report draws the conclusion that in order to understand how and why aspects of a sample might be influencing an estimator, it is crucial to look critically at how the estimator performs in reality, as well as in theory. Without understanding the real-world, finite-sample properties of the estimator one cannot justifiably draw conclusions from the results of the robust parameter estimation.

5.2. Recommendations

I would like to recommend that the next generation of Robust estimators, which are called MM-estimators one can, observe a combination of the high asymptotic relative efficiency of M-estimators with the high breakdown point of the class of estimators known as the S-estimators. The 'MM' refers to the fact that multiple M-estimation procedures are carried out in the computation of the estimators. And perhaps, it is now the most commonly employed robust regression technique. I look forward qualified article on MM estimation and S-estimation who deal high breakdown point from the next generation.

References

[1] Betsabé Pérez, Isabel Molina and Daniel Peña, "Outlier detection and robust estimation in linear regression models with fixed group effects", *Journal of Statistical Computation and Simulation, 2014*.

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

[3] Catherine Stuart, "Robust Regression", 16th April, 2011.

[4] Ekezie Dan Dan And Ogu Agatha Ijeoma, "Statistical Nalysis/ Methods Of Detecting Outliers In A Univariate Data In A Regression Analysis Model", *Imo State University*, PMB 2000, Owerri Nigeria.

[5] Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. and Stahel, W. A. (1986), Robust Statistics, The Approach Based on Influence Functions, *John Wiley & Sons*, New York.

[6] Holland, P. and Welsch, R. (1977), "Robust regression using iteratively reweighted least-squares," Commun. Statist. *Theor. Meth*. 6, 813-827.

[7] Huber, P. J. (1981), Robust Statistics. John Wiley & Sons, New York.

[8] Johan COLLIEZ, Franck DUFRENOIS and Denis HAMAD, "Robust Regression and Outlier Detection with SVR: Application to Optic Flow Estimation", Laboratoire d'Analyse des Systemes du Littoral 50 rue Ferdinand Buisson, BP 699.

[9] Lianng Yuh and ViII A. Sullivan, "Robust Estimation Using Sas·Softvare", *Department of Biostatistics Merrell Dow Research Institute Cincinnati*, Ohio 45215 Mathsoft, Inc. Seattle, WA, 255-298.

[10] Ranjit Kumar Paul, "Some Methods Of Detection Of Outliers In Linear Regression Model", Iasri, *Library Avenue*, New Delhi-110012.

[11] Robert A. Yaffee, "Robust Regression Analysis: Some Popular Statistical Package Options", *Statistics, Social Science, and Mapping Group Academic Computing Services Information Technology Services* December 2002.

[12] Rousseeuw, P. J. and Leroy, A. M. (1987), Robust Regression and Outlier Detection, *Wiley Interscience, New York (Series in Applied Probability and Statistics)*, 329 pages. ISBN 0-471-85233-3.

[13] S- PLUS 2000 Modern Statistics and Advanced Graphics Guide to Statistics, Vol. 1 (1999).

[14] SAS on LineDoc. SAS Institute, Cary, NC: IML Robust Regression, http://v8doc.sas.com/sashtml/, March 26, 2002.

[15] Yohai V. J. (1987), "High Breakdown Point and High Efficiency Robust Estimates for Regression," Annals of Statistics, 15, 642-656.

[16] Yohai V. J., Stahel, W. A. and Zamar, R. H. (1991), "A Procedure for Robust Estimation and Inference in Linear Regression," in Stahel, W. (A. and Weisberg, S. W., Eds., Directions in Robust.

[17] Yuliana et al. 2014. M ESTIMATION, S ESTIMATION, AND MM ESTIMATION IN ROBUST REGRESSION. International Journal of Pure and Applied Mathematics, Volume 91 No. 3, 349-360. doi: http://dx.doi.org/10.12732/ijpam.v91i3.7