Ilorin Journal of Science
Volume 4, Number 1, 2017, pp. 130–138 (Printed in Nigeria)
ISSN: 2408 – 4840 © 2017 Faculty of Physical Sciences, University of Ilorin
https://doi.org/10.54908/iljs.2017.04.01.009

ILORIN
JOURNAL OF SCEINCE

ILJS-16-025

Comparisons of Some Outlier Detection Methods in Linear Regression Model

Oyeyemi*, G. M., Oluwaseun, O. B. and Adeleke, M. O.

Department of Statistics, University of Ilorin, Ilorin, Nigeria.

Abstract

Empirical evidence suggests unusual or outlying observations in data sets are much more prevalent than one might expect and therefore this paper addresses multiple outliers in linear regression model. Although reliable for a single or a few outliers, standard diagnostic techniques from an Ordinary Least Squares (OLS) fit can fail to identify multiple outliers. The parameter estimates, diagnostic quantities and model inferences from the contaminated data set can be significantly different from those obtained with the clean data. A regression outlier is an observation that has an unusual value of the dependent variable Y, conditional on its value of the independent variable X. Four procedures for detecting outliers in linear regression were compared; the Cook's, DFFITS, DFBETAS, and Mahalanobi's distances. DFBETAS is most efficient in outlier detection for small sample and small percentage of outliers but has low sensitivity when the sample size is large. Mahalanobi has more power of detection of small percentage of outliers regardless of sample size.

Keyword: Mahalanobis distance, Cooks' distance, Masking Effect, DFBETAS, DFFITS.

1. Introduction

Data collected by research workers commonly contain outliers and it is important that these outliers be identified in the course of a thorough and correct statistical analysis (Hadi and Simonoff, 1993). A number of procedures have been proposed in recent years for detecting outliers in linear regression, yet their detection still may be difficult, especially when there are multiple outliers in the data. Two common problems in multiple-outlier detection are masking and swamping (Hoaglin and Welsch, 1978, Atkinson, 1994). Masking occurs when one outlier obscures the existence of another while swamping occurs when a non-outlier is wrongly included in a group of observations thought to be outliers. Along with the proliferation of outlier-detection methodologies over the years, philosophical changes have occurred.

*Corresponding Author: Oyeyemi, G. M.
Email: gmoyeyemi@gmail.com

In the early years, many statisticians and practitioners viewed outlier-identification methodologies largely as ways to legitimize deleting observations which, though not necessarily erroneous, fell outside the pattern seen in the bulk of the data and were perhaps troublesome in the analysis (Barnett and Lewis, 1994, She and Owen, 2011).

Nowadays, outlier identification is viewed more broadly. It is widely recognized that, in some applications, outliers are of interest in their own right and may be the most important observations in the data set; identifying them may help chart future research (Hossain and Naik, 1991, Chandola *et al.*, 2007). The literature on influential observations has expanded our understanding of the need to identify certain points as candidates for special treatment lest they warp our impression of relationships in the body of the data. The focus has moved away from viewing these procedures as providing support for automatic deletion of points, and toward seeing them as aids in identifying points for more careful scrutiny. The latter perspective suggests interest in identifying moderate as well as extreme outliers.

Several factors can affect the efficiency of the analyzed methods. In particular, the method depends on: whether or not the data is multivariate normal; the dimension of the dataset; the type of outliers; the proportion of outliers in the dataset; and the outlier's degree of contamination (Markatou and He, 1994, Tang *et al.*, 2006, Alanamu and Oyeyemi, 2018). An observation that is substantially different from all other ones can make a large difference in the results of regression analysis**.** Outliers occur very frequently in real data, and they often go unnoticed because nowadays much data is processed by computers, without careful inspection or screening (Rahman *et al.*, 2012, Oyeyemi *et al.*, 2015). Outliers may be a result of keypunch errors, misplaced decimal points, recording or transmission errors, exceptional phenomena such as earthquakes or strikes, or members of a different population slipping into the sample.

Several decades have passed since linear regression analysis became a widely employed statistical methodology that utilizes the relation between quantitative response and quantitative and qualitative covariates to make predictions and inferences. When a regression model is considered for an application, researchers and analysts usually are not certain in advance whether a particular form of model is appropriate, especially with social science or epidemiological data. It is therefore natural to raise questions before making inferences based on the particular data at hand. A general question is: what type of model is appropriate – linear or nonlinear? A more specific question is whether the fitted model is unduly affected by

unusual points (Davies and Gather, 1993, Yu and Yao, 2017). If so, what features of the data explain this effect? Do collinear relationships exist among the data series used as predictors? Do such problems degrade the parameter estimation? Diagnostic techniques were gradually developed to find problems in model-fitting and to assess the quality and reliability of regression estimates. These concerns turned into an important area in regression theory intended to explore the characteristics of a fitted regression model for a given data set.

Discussion of diagnostics for linear regression models are often indispensable chapters or sections in most of the statistical textbooks on linear models and are also included as standard options in many statistical packages. Although techniques for regression diagnostics have been developed theoretically and methodologically for conventional linear regression models, diagnostics have not been extensively studied in survey sampling. The diagnostic tools provided by current popular software are generally based on ordinary or weighted least squares regression.

2. Materials and Methods

Data analysts routinely encounter data sets which potentially contain one or more outliers. When, as is usually the case, there is no a priori reason to suspect that particular observations are the outliers, an outlier test based on the sequential (perhaps better called "repeated") application of a single-outlier test statistic is commonly used. This sequential approach is especially prone to masking in the presence of multiple outliers.

Several methods for outlier detections are highlighted here ranging from Grubbs test, Dixon test, Rosner test, Cochran test, Tietjen-moor test, Inter Quartile Range, Boxplot, Histogram, Z-score, etc., but for the purpose of this project work, only four procedures will be considered

2.1 Cook's Distance

Cook (1977) introduced distance measure for commonly used estimates of the influence of a data point when performing least squares regression analysis. Cook's distance points with a large value are considered to merit closer examination in the analysis (Cook and Weisberg, 1982). Cook's Distance provides an overall measure of the combined impact of an observation on all of the estimated regression coefficients (b). It can be derived from the confidence region of β, which at level $100(1-\alpha)\%$ is given by those values b^* satisfying

$\frac{(b^*-b)'(b^*-b)}{ps^2} \leq F(1-\alpha; p, n-p)$.

Using this structure, Cook's Distance measure D_i was proposed as

$$D_i = \frac{(b_{(i)}-b)' X'X (b_{(i)}-b)}{ps^2}. \quad (1)$$

This is a measure of the distance $b_{(i)} to\ b$. If $b_{(i)} and\ b$ are relatively far from each other, this means that unit i has a substantial effect on the full sample estimate. Large values of D_i indicate observations that are influential on joint inferences about the parameters in the linear model. It has been found useful to relate D_i to the percentile values of $F_{(1-\alpha;\ p,\ n-p)}$ distribution to make the judgement on influence. For example, if the percentile value is less than about 20 percent, the unit has little apparent influence on the regression coefficients. On the other hand, if the percentile value is near 50 percent or more, the influence is partially important.

A more convenient form for D_i**,** without fitting a new regression function for each deletion, follows from substitution of DFBETA and it is expressed as;

$$D_i = \frac{e_i^2 h_i}{ps^2(1-h_i)^2} = \frac{r_i^2 h_i}{p(1-h_i)}, \quad (2)$$

where $r_i = \frac{e_i}{s\sqrt{(1-h_i)}}$ is the internally studentized residual. Note that from the expression in equation (2), D_i depends on the size of the studentized residual and the leverage value

2.2 Influence on fitted values: DFFIT and DFFITS

DFFITS is a diagnostics tool for statistical regression model which shows the influence point (Paul and Fung, 1991). This statistic summarizes how much the regression function change in predicted values when an observation is deleted, with the advantage that it does not depend on the particular coordinate system used to form the regression model. Rescaling DFFIT by the estimated deviation of the predicted value, with the sample standard error s replaced by the delete-one version $s_{(i)}$**,** DFFITS can be expressed as the product of a t-distributed random variable and a function of the leverage:

$$\text{DFFITS} = \frac{\hat{Y}_i - \hat{Y}_i(i)}{s_{(i)}\sqrt{h_i}} = \frac{X_i'(b-b_{(i)})}{s_{(i)}\sqrt{h_i}} = \frac{e_i}{s_{(i)}\sqrt{1-h_i}}\sqrt{\frac{h_i}{1-h_i}}. \quad (3)$$

A large value of DFFITS indicates that the observation is very influential in its neighborhood of the X space. A general cutoff to consider is 2; a size adjusted cutoff recommended by Belsley *et al* (1980) is $2\sqrt{\frac{p}{n}}$, where p is the number of independent variable and $\frac{p}{n}$ is the mean leverage.

2.3 Influence on regression coefficients: DFBETA and DFBETAS

Recall that an influential observation is one that combines discrepancy with leverage. Therefore, examine how regression coefficients change if outliers are omitted from the model. We can use D_{ij} (often termed $DFBETAS_{ij}$) to do so: DFBETA, the change in parameter estimates after deleting the ith observation can be formulated and rewritten as DFBETA = b-$b_{(i)} = \frac{A^{-1}x_i e_i}{1-h_i}$, where A = X′X. If we let $C = (X'X)^{-1}X' = (C_{ji})_{p\times n}$, then the jth element of the DFBETA vector is $b_j - b_{j(i)} = \frac{C_{ji}e_i}{1-e_i}$. Belsley *et al* (1980) suggest that changes in the estimated regression coefficients are often mostly usefully assessed relative to the variance of b. A scaled measure of the change can be defined as the following:

$$DFBETAS_{ij} = \frac{b_j - b_{j(i)}}{s_{(i)}\sqrt{(X'X)^{-1}_{jj}}} = \frac{C_{ji}}{\sqrt{\sum_{k=1}^{n} C_{jk}^2}} = \frac{e_i}{s_{(i)}\sqrt{1-h_i}}\frac{1}{\sqrt{1-h_i}}, \quad (4)$$

where $(X'X)^{-1}_{jj}$ is the (jj)th element of $(X'X)^{-1}$. The denominators of $DFBETAS_{ij}$ is analogous to the estimated standard error of 'b' with the sample standard error '**s**' replaced by delete-one version '$s_{(i)}$'.

The DFBETAS statistic is the product of a quantity of order $n^{-1/2}$, a t-distributed random variable, and a quantity that approaches 1 (assuming $h_i = 0$). Belsley *et al.* (1980) proposed a cutoff point of $\frac{2}{\sqrt{n}}$ to identifying influential cases. Thus, if all the observations in the sample follow an underlying normal model, the X's are bounded, and the leverages are small, roughly 95% of the observations will have a DFBETAS lesser than $\frac{2}{\sqrt{n}}$ in absolute value. In some samples, especially small or moderate size ones, this statement is less precise since h_i may not be negligible and the term involving C_{ji} may not be near $\frac{1}{\sqrt{n}}$.

DFBETAS is somewhat cumbersome to work with because an analyst must examine pn values, for each observation, there are p DFBETAS for each parameter.

2.4 Mahalanobi's Distance

One classical method to identify leverage points is to inspect the use of the Mahalanobi's Distance MD_i to find outliers x_i

$$MD_i = \sqrt{(x_i - \bar{x})C^{-1}(x_i - \bar{x})'} \ , \quad (5)$$

where $\bar{\boldsymbol{x}}$ and C are the sample mean and classical sample covariance matrix of the data set X, respectively. The distance tells us how far is x from the center of the cloud taking into account the shape of the cloud. It is well known that this approach suffers from the masking effect by which multiple outliers do not necessarily have a large MD_i.

In classical linear regression, the diagonal element h_i of the hat matrix $H=X(X'X)^{-1}X'$ are used to identify leverage points, the ith leverage $h_i = H_i$ is the ith diagonal element of the hat matrix H**.** Rousseeuw and Van Zomeren (1990) reported the following monotone relationship between the h_i and MD_i

$$h_i = [((MD_i)^2)/(N-1)] + [1/n] \ . \quad (6)$$

It points out that neither the MD_i nor the h_i is entirely save for detecting leverage points reliably. Multiple outliers do not necessarily have large MD_i values because of the masking effect. Rousseeuw and Leroy (1987) suggest using $h_i > 2p/n$ and $MD_i^2 > \varkappa^2_{p;1-0.95}$ as benchmarks for leverages and Mahalanobi's Distance.

3. Data Analysis and Discussion

The data of concern in this paper were simulated from R-statistical package and various outlier detection methods were carried out to verify how efficient each test could detect outliers. Furthermore, outliers at 10%, 20% and 30% were injected to various sample sizes of 10, 40 and 100 and then iterated 500 times and the simulated data were analyzed using the four test statistics expressed in equations (1) to (4). Data were simulated from Normal distribution with mean 2 and standard deviation 1, outliers were injected from a poisson distribution with mean 7 using R-statistical package.

Table 1: Frequency and probability of correctly detecting actual percentage number of injected outliers

Sample Size	% of Outlier	X			
		Cook's distance	DFFITS distance	DFBETAS distance	Mahalanobi's distance
10	10	217 (0.434)	186 (0.372)	257 (0.514)	201 (0.402)
	20	164 (0.328)	192 (0.384)	32 (0.064)	1 (0.002)
	30	104 (0.208)	202 (0.404)	0 (0.000)	0 (0.000)
40	10	29 (0.058)	97 (0.194)	73 (0.146)	9 (0.018)
	20	0 (0.000)	1 (0.002)	0 (0.000)	0 (0.000)
	30	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)
100	10	0 (0.000)	48 (0.096)	4 (0.008)	3 (0.006)
	20	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)
	30	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)

Table 2: Frequency and probability of correctly detecting percentage number of injected outliers

Sample Size	% of Outlier	X			
		Cook's distance	DFFITS distance	DFBETAS distance	Mahalanobi's distance
10	10	222 (0.444)	217 (0.434)	297 (0.594)	202 (0.404)
	20	174 (0.348)	220 (0.440)	32 (0.064)	1 (0.002)
	30	110 (0.220)	224 (0.448)	0 (0.000)	0 (0.000)
40	10	32 (0.064)	182 (0.364)	105 (0.210)	16 (0.032)
	20	0 (0.000)	1 (0.002)	0 (0.000)	0 (0.000)
	30	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)
100	10	0 (0.000)	113 (0.226)	5 (0.010)	9 (0.018)
	20	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)
	30	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)

Table 3: Frequency and probability of detecting more than number of injected outliers

Sample Size	% of Outlier	X			
		Cook's distance	DFFITS distance	DFBETAS distance	Mahalanobi's distance
10	10	5 (0.010)	31 (0.062)	40 (0.080)	1 (0.002)
	20	10 (0.020	28 (0.056)	0 (0.000)	0 (0.000)
	30	6 (0.012)	22 (0.044)	0 (0.000)	0 (0.000)
40	10	3 (0.006)	85 (0.170)	32 (0.064)	7 (0.014)
	20	0 (0.000)	0 (0.002)	0 (0.000)	0 (0.000)
	30	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)
100	10	0 (0.000)	65 (0.013)	1 (0.002)	6 (0.012)
	20	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)
	30	0 (0.000)	0 (0.000)	0 (0.000)	0 (0.000)

Table 4: Frequency and probability of detecting less than number of injected outliers

Sample Size	% of Outlier	X			
		Cook's distance	DFITTS distance	DFBETAS distance	Mahalanobi's distance
10	10	278 (0.556)	283 (0.566)	203 (0.406)	298 (0.596)
	20	326 (0.652)	280 (0.560)	468 (0.936)	499 (0.998)
	30	390 (0.780)	276 (0.552)	500 (1.000)	500 (1.000)
40	10	468 (0.936)	318 (0.636)	395 (0.790)	484 (0.968)
	20	500 (1.000)	499 (0.998)	500 (1.000)	500 (1.000)
	30	500 (1.000)	500 (1.000)	500 (1.000)	500 (1.000)
100	10	500 (1.000)	387 (0.774)	495 (0.990)	491 (0.982)
	20	500 (1.000)	500 (1.000)	500 (1.000)	500 (1.000)
	30	500 (1.000)	500 (1.000)	500 (1.000)	500 (1.000)

4. Discussion of Results and Conclusion

Based on the analysis presented above, the four outlier detection methods have similar pattern. DFBETAS has more power of detecting outlier when the sample size is small and when the number of injected outliers is also small. Furthermore, detection of outliers achieved high level of sensitivity, when the sample size is moderate but low sensitivity was the case when the sample size becomes large. Cooks Distance can only detect outliers when the sample size is small and when outlier injected is small. Mahalanobis do detect outliers when injected outliers are small regardless of the sample size. DFFITS among the four methods has more power of detecting outliers better at large sample size and it detects larger number of injected outliers at small sample size. It is therefore recommended that for small sample size and small number of outliers, DFBETAS should be employed for its accuracy of detecting outliers. While for moderate and large sample sizes and small numbers of outliers the DFFITS preferred.

Reference

Alanamu, T. and Oyeyemi, G. M. (2018): A New Robust Method of Estimating Linear Regression Model in the Presence of Outliers. *Pacific Journal of Science and Technology*. **19** (1), 125–132.

Atkinson, A. C. (1994): Fast very robust methods for the detection of multiple outliers. *Journal of the American Statistical Association*. **89**, 1329-1339.

Barnett, V. and Lewis, T. (1994): *Outliers in statistical data*, 3rd edn., John Wiley& Sons, New York.

Basley, D. A., Kuh, E., and Welsch, R. (1980): *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. John Wiley, New York.

Chandola, V., Banerjee, A. and Kumar, V. (2007): *Outlier detection: A survey, ACM Computing Surveys*. 1-83.

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

Cook, R. D. and Weisberg, S. (1982): *Residual and Influence in Regression.* Chapman & Hall Ltd., London.

Davies, L. and Gather, U. (1993). The Identification of multiple outliers. *Journal of the American Statistical Association.* **88** (423), 782-795.

Hadi, A. S. and Simonoff, J. S. (1993): Procedures for the identification of multiple outliers

in linear models. *Journal of the American Statistical Association.* **88**, 1264- 1272.

Hoaglin, D. C. and Welsch, R. E. (1978): The hat matrix in regression and ANOVA. *American Statistician.* **32**, 17-22.

Hossain, A. and Naik, D. N. (1991): A comparative study on detection of influential observations in linear regression Received. *Statistical Papers*. **32**, 55-69.

Markatou, M. and He, X. (1994): Bounded influence and high breakdown point testing procedures in linear models. *Journal of the American Statistical Association.* **89**, 543-549.

Oyeyemi, G. M., Bukoye, A. and Akeyede, I. (2015): Comparison of Outlier Detection Procedures in Multiple Linear Regression. *American Journal of Mathematics and Statistics.* **5**(1), 37-41.

Paul, S. R and Fung, K. Y. (1991): A generalized extreme studentized residual multiple outlier detection procedure in linear regression. *Technometrics.* **33**, 339-348.

Rahman, S. M. A, Sathik, M. M. and Kannan, K. S. (2012): Multiple Linear Regression Models in Outlier Detection. *International Journal of Research in Computer Science.* **2** (2), 23 – 28.

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

Rousseeuw, P. J. and Van Zomeren, B. C (1990): Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association.* **83**, 633-651.

She, Y. and A.B. Owen. (2011): Outlier Detection Using Nonconvex Penalized Regression. *Journal of American Statistical Association.* **106**, 626-639.

Tang, J., Chen, Z., Fu, A. W. and Cheung, D.W. (2006): Capabilities of outlier detection Schemes in large datasets, framework and methodology. *Knowledge and Information System.* **11** (1), 45-84.

Yu, C. and W. Yao. (2017): Robust Linear Regression: A Review and Comparison. *Communications in Statistics-Simulation and Computation.* **3**, 1-22.