Applied Mathematical Sciences, Vol. 9, 2015, no. 47, 2317 - 2324
HIKARI Ltd, www.m-hikari.com
http://dx.doi.org/10.12988/ams.2015.53213

Outlier Detection in Multivariate Data

K. Senthamarai Kannan and K. Manoj

Department of Statistics
Manonmaniam Sundaranar University
Tirunelveli - 627 012, Tamilnadu, India

Abstract

The objective of this research is detection of outliers in multivariate data employing various distance measure, particularly using robust regression diagnosis technique. Several classical outlier identification methods are based on the sample mean and covariance matrix in general. But they do not always yield better result, as they themselves are affected by the outliers. Sometimes one outlier point has hide the other outliers. To identify them, methods which have masking effect with outlier points are being used. An appropriate method is adopted to identify the unmasking outliers and also to compare the various distance measures.

Mathematics Subject Classification: 62H99, 62J05

Keywords: Outlier Detection, Mahalanobis Distance, Cooks, Leverage, Masking Effect, DFFITS

1 Introduction

Multivariate outlier detection is the important task of statistical analysis of multivariate data. The methods are applied to a set of data to illustrate the multiple outlier detection procedure in multivariate linear regression models. Outliers can mislead the regression results. When an outlier is involved in the study, it pulls the regression line towards itself. This can result in a solution

that is more precise for the outlier, but less precise for all of the other cases in the data set. The usual methods of linear model defined as

$$y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip} + e_i \quad for \quad i = 1, ..., n \tag{1}$$

where n is the sample size, y_i is the response variable and $x_{i1}, ..., x_{ip}$ are called explanatory variables. e_i is the error term is to assumed as normally distributed by mean zero and standard deviation σ. Rousseeuw[8] has described several identification of outliers in regression analysis deals with unmasking outliers and leverage points in multivariate data. David et al.[1] have discussed the difficulties occur in the multivariate outliers problem increase with dimension of data and also described significant difference of improvements in the methods for outlier detection with simulation technique with examples. Penny[5] has proposed a clinical trial based new treatment for identifying multivariate outliers. Gao et al.,[9] proposed a procedure for identifying multivariate outliers using Hawkins[2] data set. Max-Eigen Difference(MED) method was briefly discussed about theoretical aspect of procedures to compare with Mahalanobis Distance(MD) and Robust Distance(RD) with examples. Southworth[4] has discussed identifying outliers in clinical trial data, dealing with unmasking multivariate outliers. Dang et al.,[10] has introduced nonparametric multivariate outliers detection based on multivariate depth functions, also masking robustness against misidentification of outliers and non-outliers. In statistically analyzing data sets, whether check there are any outliers in the dataset.

Definition 1.1 *An outlier is an observation, which so much deviates from other observations as to arouse suspicions that it was generated by a different mechanism by Hawkins[3].*

For univariate case we have easy way to finding outliers and visualizing suitable plot. Normality assumption of the dataset should be checked before finding outliers. It is useful to simply weather outliers found or not in the data set.

2 Masking Effect

An observation is detected as inliers in the presence of the extreme observation and by deleting this extreme observation, the observations nearer to it are also found to be outliers. This phenomenon is considered as the masking effect. When masking is occurs the mean and covariance estimates are skewed towards a group of outliers, and the resulting gap of the outlier form the mean is small.

Example 2.1 *Let x be a univariate vector as, x=[3 4 5 7 9 12 15 22 38]. The use of Grubbs test of outlier detection, will just detect one outlier that is 38. But after deleting this outlier and again applying Grubbs test, 22 will be detected as outlier. So it can be said that 22 is masked by 38. As the mask (38) is removed 22 appears to be the outlier.*

3 Univariate Outlier Detection

Univariate data have an unusual value for a single variable. The results will be concerned with univariate outliers for the dependent variable in the data analysis. Manoj and Kannan[6] has identifying outliers in univariate data using

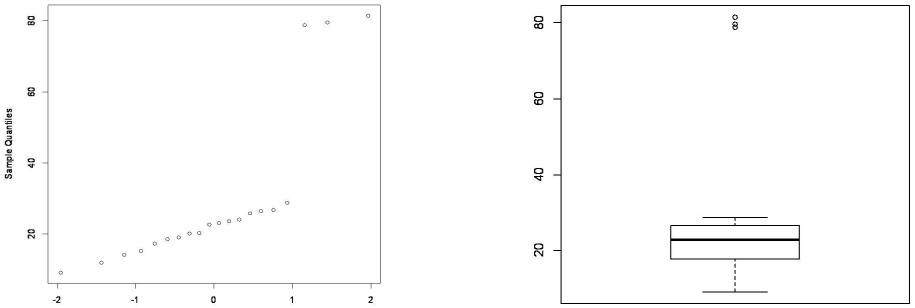

Figure 1: Univariate outlier detection using Scatter and Boxplot

various outlier identification methods and compare them with each others.These results of the plot represent three outliers visualized clearly. figure 1. shows the plots, where the three circles are outliers.

4 Multivariate outlier detection methods

Several methods are used to identify outliers in multivariate dataset. Among them, four of the outlier diagnostics methods of distance measures described in the following.

4.1 Mahalanobis Distance (MD_i)

A classical Approach for detecting outliers is to compute the Mahalanobis Distance (MD_i) for each observation x_i:

$$MD_i = \sqrt{(x_i - \bar{x})^T V^{-1} (x_i - \bar{x})} \tag{2}$$

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

4.2 Cook's Distance (Di)

Dennis Cook (1977) introduced distance measure for commonly used estimates of the influence of a data point when performing least squares regression analysis.

In practically the ordinary least squares analysis, Cook's distance Points with a large are considered to merit closer examination in the analysis.

$$D_i = \frac{\sum_{j=1}^{n}(\hat{Y}_i - \hat{Y}_j(i))^2}{pMSE} \tag{3}$$

The following equation equally expressed as,

$$D_i = \frac{e_i^2}{pMSE}\left[\frac{h_{ii}}{(1-h_{ii})^2}\right] \tag{4}$$

$$D_i = \frac{(\hat{\beta} - \hat{\beta}^{-i})^T (X^T X)(\hat{\beta} - \hat{\beta}^{-i})}{(1+p)s^2} \tag{5}$$

where, $\hat{\beta}$ is the least squares (LS) estimate of β , and $\hat{\beta}^{-i}$ is the LS estimate of β on the data set without case i. $\hat{Y}_j$ is the prediction from the full regression model for observation j; $\hat{Y}_{j(i)}$ is the prediction for observation j from a refitted regression model in which observation i has been omitted. h_{ii} is the i^{th} diagonal elements of the hat matrix $X(X^T X)^{-1}X^T$. MSE - is the mean square error, p is number of fitted parameters.

4.3 Leverage Point(h_i)

An observation with extreme value on a predictor variable is called a point with high leverage. In linear regression identification of leverage points may be quite to detect. In linear regression model, the leverage score for i^{th} data unit is defined as,

$$h_{ii} = (H)_{ii} \tag{6}$$

The i^{th} diagonal of the hat matrix $H = X(X'X)^{-1}X'$. Leverage values fall between 0 and 1. Investigate observations with leverage values greater than 3p/n, where p is the number of model terms (with constant) and n is the number of observations.

4.4 DFFITS

DFFITS is the diagnostics tool for statistical regression model shows that influence point. This quantity measures how much the regression function changes at the i^{th} observation when the i^{th} variable is deleted.

$$DFFITS_i = \frac{\hat{Y}_i - \hat{Y}_{i(i)}}{\sigma_{(i)}\sqrt{h_{ii}}} \tag{7}$$

For small samples datasets the values of 1 or greater values is considered as suspicious. In large samples of data sets values of $2\sqrt{p/n}$.

5 Computational Procedure

For the multivariate data set n observations with m variables, the basic idea of the methods can be described in the following steps.

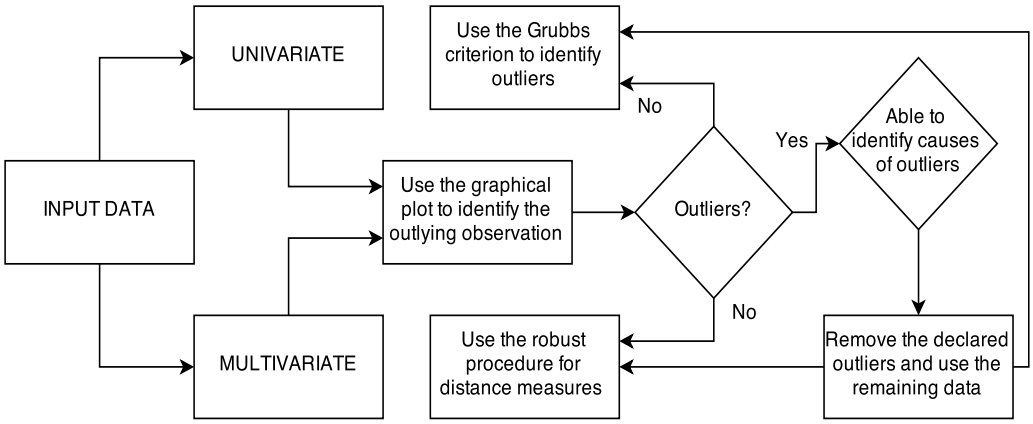

Figure 2: Diagram for computational procedure

6 Numerical Illustration

For the given example of datasets are extracted from url (UCI Repository) http://www.ics.uci.edu. which have 80 observations and 8 variables. The variables are described as Age, Pregnancy, Plasma, Pressure level, Skin cells, Insulin level, Body Mass Index(BMI) and Pediatric.

In Table 1 values are plotted as a scatter plot for identifying outlier points. The outlier points are identified by Mahalanobis Distance(MD_i) appeared to be same as it was observed in the leverage values(h_i). Though there are different methods for detecting outlier points, but it has been found that the maximum outlier can be detected by Cook's distance, While DFFITS can be used to detect the minimum outliers points. The tabulated data Table 2 represents the outlier points which are identified by the various distance measures.

7 Conclusion

This research deals with the procedure for computing the presence of outliers using various distance measures such as Mahalanobis Distance (MD_i), Cooks Distance(D_i), Leverage point(h_i) and DFFITS. From the diabetes dataset, the outlier identification level of Mahalanobis Distance (MD_i) and Leverage Point (h_i) are approximately the same, but DFFITS outlier detection sensitivity is very low and the outlier detection sensitivity using Cooks Distance (D_i) is

Table 1: Mahalanobis(MD_i), Cooks(D_i), Leverage(h_i) and DFFITS

i	MD_i	D_i	h_i	DFFITS	i	MD_i	D_i	h_i	DFFITS
1	4.460	.033	.056	0.523	41	8.147	.004	.103	-0.187
2	3.012	.005	.038	0.202	42	3.849	.003	.049	-0.164
3	7.376	.019	.093	-0.389	43	3.454	.014	.044	0.334
4	2.622	.003	.033	-0.145	44	7.191	.000	.091	-0.023
5	23.883	.042	.302	-0.581	45	3.871	.002	.049	0.117
6	2.301	.001	.029	-0.086	46	18.796	.065	.238	-0.722
7	3.175	.000	.040	-0.034	47	6.072	.000	.077	0.059
8	17.765	.007	.225	-0.243	48	2.770	.004	.035	-0.167
9	14.880	.081	.188	0.813	49	3.845	.001	.049	-0.098
10	17.505	.141	.222	1.083	50	18.503	.008	.234	-0.247
11	5.776	.003	.073	-0.166	51	4.905	.013	.062	-0.319
12	5.163	.020	.065	-0.401	52	2.118	.000	.027	0.006
13	12.365	.015	.157	0.340	53	1.411	.000	.018	-0.041
14	40.231	.235	.509	1.379	54	5.019	.013	.064	0.323
15	2.722	.012	.034	0.308	55	6.551	.005	.083	-0.198
16	12.940	.003	.164	0.159	56	3.408	.000	.043	-0.055
17	5.948	.001	.075	-0.063	57	7.004	.001	.089	-0.090
18	2.006	.003	.025	-0.149	58	9.171	.000	.116	0.001
19	9.771	.050	.124	0.641	59	16.547	.022	.209	0.420
20	1.508	.000	.019	0.061	60	5.993	.002	.076	-0.115
21	3.061	.017	.039	0.523	61	16.958	.004	.215	0.172
22	4.092	.011	.052	0.202	62	2.326	.000	.029	0.003
23	9.663	.000	.122	-0.389	63	8.672	.001	.110	0.093
24	5.374	.014	.068	-0.145	64	3.421	.006	.043	-0.216
25	5.959	.002	.075	-0.581	65	2.128	.003	.027	0.144
26	3.794	.000	.048	-0.086	66	1.979	.000	.025	0.028
27	4.353	.004	.055	-0.034	67	4.802	.013	.061	0.319
28	3.023	.008	.038	-0.243	68	7.955	.088	.101	0.865
29	7.628	.014	.097	0.813	69	3.491	.000	.044	-0.041
30	3.783	.000	.048	1.083	70	3.524	.006	.045	-0.216
31	1.130	.036	.014	-0.166	71	1.917	.002	.024	-0.130
32	3.056	.015	.039	-0.401	72	2.097	.006	.027	-0.227
33	1.645	.003	.021	0.340	73	10.048	.034	.127	-0.526
34	5.220	.013	.066	1.379	74	4.307	.044	.055	-0.609
35	5.405	.002	.068	0.308	75	3.168	.001	.040	-0.105
36	5.556	.005	.070	0.159	76	12.900	.001	.163	0.099
37	3.943	.014	.050	-0.063	77	7.221	.002	.091	0.117
38	6.081	.007	.077	-0.149	78	2.795	.002	.035	-0.128
39	4.366	.000	.055	0.641	79	20.142	.071	.255	0.758
40	9.317	.058	.118	0.061	80	2.666	.000	.034	-0.058

very high, since maximum number of outlier points are identified. This results clearly reveals that Cook's Distance identifies the maximum number of highly infected diabetes patients.

Acknowledgements The authors express his gratitude to the UGC for providing the financial support to carry out this work under scheme UGC SAP (DRS - I). The second author acknowledges the UGC for awarding the Scheme

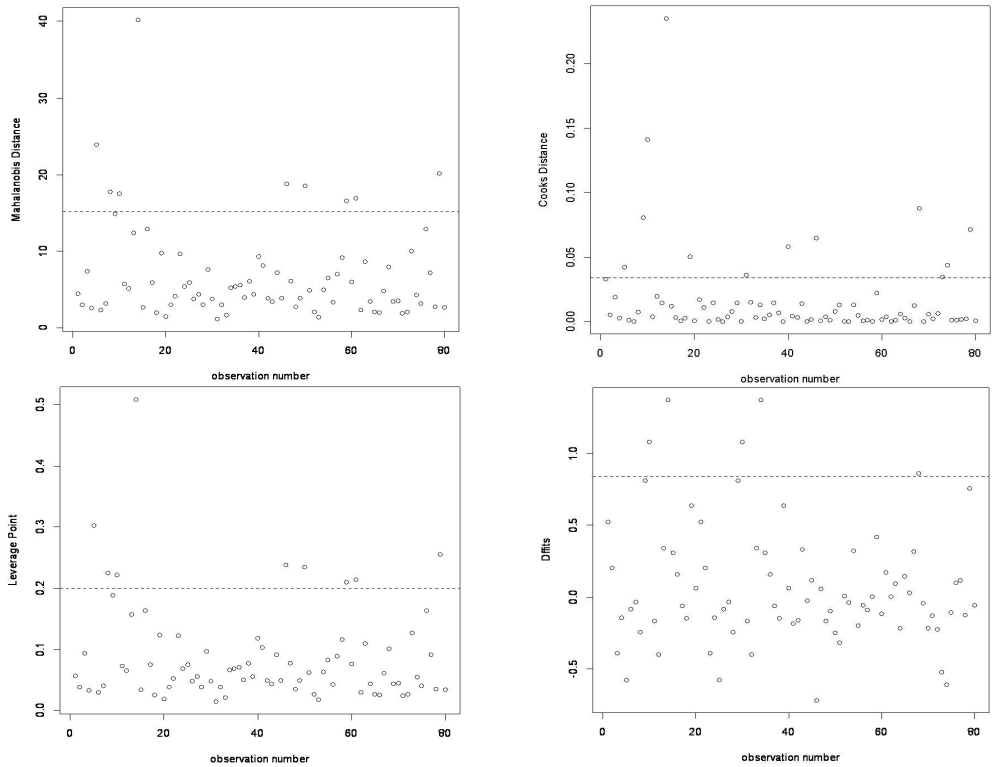

Figure 3: outlier points visualized by scatter plot for Mahalanobis Distance, Cooks Distances, Leverage and DFFITS

Table 2: Number of Outliers detected by variaous distances

Distances Measures	**Outliers Detected**
Mahalanobis Distance(MD_i)	9
Cooks distance(D_i)	11
Leverage values(h_i)	9
DFFITS	5

of Rajiv Gandhi National Fellowship (RGNF). This work is partly financial supported by UGC (RGNF).

References

[1] David M. Rocke, David L. Woodruff, Identification of Outliers in Multivariate Data, *Journal of the American Statistical Association,* **91(435)**(1996), 1047-1061. http://dx.doi.org/10.2307/2291724

[2] Hawkins, D. M., Bradu, D., Kass, G. V, Location of several outliers in

multiple regression data using elemental subsets. *Tecnometrics,***26**(1984), 197-208. http://dx.doi.org/10.1080/00401706.1984.10487956

[3] Hawkins, D. M.,*Identification of Outliers,* Chapman and Hall, New York, 1980. http://dx.doi.org/10.1007/978-94-015-3994-4

[4] Harry Southworth, Detecting outliers in multivariate laboratory data, *Journal of Biopharmaceutical Statistics,***18(6),**(2008), 1178-1183. http://dx.doi.org/10.1080/10543400802369046

[5] Kay I. Penny, Ian T. Jolliffe, Comparison of Multivariate Outlier detection methods for clinical laboratory safety data, *Journal of the Royal Statistical Society. Series D (The Statistician),***50(3),** (2001), 295-308. http://dx.doi.org/10.1111/1467-9884.00279

[6] Manoj, K., Senthamarai Kannan, K., Comparison of Methods for detecting Outliers, *International Journal of Scientific and Engineering Research,* **4(9),**(2013), 709-714.

[7] Peter Filzmoser, Anne Ruiz-Gazan, Christine Thomas-Agnan, Identification of local multivariate outliers, *Stat Papers,***55(1),**(2013), 29-47. http://dx.doi.org/10.1007/s00362-013-0524-z

[8] Rousseeuw, Peter J., and Annick M. Leroy.*Robust regression and outlier detection,* **589,** John Wiley Sons, 2005.

[9] Shaogen Gao, Guoying Li, Dongqian Wang, A New Approach for Detecting Multivariate Outliers, *Communications in Statistics - Theory and Methods,* **34(8)**(2005), 1857-1865. http://dx.doi.org/10.1081/sta-200066315

[10] Xin Dang, Robert Serflling., Nonparametric Depth-Based Multivariate Outliers Identifiers, and Masking Robustness Properties, *Journal of Statistical Planning and Inference,* USA. (2009).

Received: December 15, 2014; Published: March 21, 2015