CHAPTER 5

Outlier Detection in Multivariate Data

5.1 Introduction

Multivariate outlier detection is the important task of statistical analysis of multivariate data. Many methods have been proposed for univariate outlier detection. Identifying outliers in multivariate data pose challanges that univariate data do not. A multivariate outlier need not be an extreme in any of its components The idea of extremeness arises inevitably form some form of 'ordering' of the data. They are based on (robust) estimation of location and scatter, or on quantiles of the data. A major disadvantage is that these rules are independent from the sample size. The basis for multivariate outlier detection is the Mahalanobis distance. The standard method for multivariate outlier detection is robust estimation of the parameters in the Mahalanobis distance and the comparison with a critical value of the χ^2 distribution (Rousseeuw & Zomeren (1990)). However, also values larger than this critical value are not necessarily outliers, they could still belong to the data distribution.

Barnett (2003) had discussed the basic principles and problems of 'the ordering of multivariate data'. Interest in outliers in multivariate data remained the same as for the univariate case. Extreme values could again provide naturally interpretable measures of environmental concern in their own right and if they were not only extreme, but 'surprisingly' extreme or unrepresentative, they might again suggest that some unanticipated influence is present in the data source. So once more we encountered the notion of 'maverick' or 'rogue' extremes which we would term outliers.

Thus, as in a univariate sample, an extreme observation may 'stick out' so far from the others that it must be declared an outlier, and an appropriate test of discordancy may demonstrate that it is statistically unreasonable even when viewed as an 'extreme'. Such an outlier is said to be discordant and may lead us to infer that some contamination is present in the data. Discordancy may, as in univariate data, prompt us to reject the offending observation from the sample. But rather than rejection we will again wish to consider accommodation (outlier-robust) procedures which limit the effects of possible contamination or identification of the nature of the contamination as an interest in its own right.

The methods were 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 pulled the regression line towards itself. This could 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 outlier challenge is one of the earliest of statistical interests, and since nearly all data sets contain outliers of varying percentages, it continues to be one of the most important. Sometimes outliers can grossly distort the statistical analysis, at other times their influence may not be as noticeable. Statisticians have accordingly developed numerous algorithms for the detection and treatment of outliers, but most of these methods were developed for univariate data sets. This chapter focuses on multivariate outlier detection.

Especially when using some of the common summary statistics such as the sample mean and variance, outliers can cause the analyst to reach a conclusion totally opposite to the case if outliers weren't present. For example, a hypothesis might or might not be declared significant due to a handful of extreme outliers. In fitting a regression line via least squares, outliers can sufficiently alter the slope so as to induce a sign change.

Classical outlier detection methods are powerful when the data contain only one outlier. However, the powers of these methods decrease drastically if more than one outlying observations are present in the data. This loss of power is usually due to what are known as the masking porоblems. In addition, these methods do not always succeed in detecting outliers, simply because they are affected by the observations that they are supposed to identify. Therefore, a method which avoids these problems is needed.

5.2 The Regression Model

The Regression analysis to fit the equations to the observed variables. The usual methods of linear model is defined as

$$y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip} + e_i \quad for \quad i = 1, ..., n \tag{5.1}$$

where n is the sample size. y_i is the response variable and $x_i 1, ..., x_i p$ are called explanatory variables. e_i is the error term is to assumed as normally distributed by mean zero and standard deviation σ.

$$\beta = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} \tag{5.2}$$

from the data:

$$\begin{array}{cc} & Variables \\ Cases & \begin{bmatrix} x_{11} & \cdots & x_{1p} & y_1 \\ \vdots & & \vdots & \vdots \\ x_{i1} & \cdots & x_{ip} & y_i \\ \vdots & & \vdots & \vdots \\ x_{n1} & \cdots & x_{np} & y_n \end{bmatrix} \end{array} \tag{5.3}$$

Applying a regression estimator to such a data set yeilds

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_p \end{bmatrix} \tag{5.4}$$

where the estimates $\hat{\beta}_j$ are called the regression coefficients. Vectors and matrices will be denoted by boldface throughout. Although the actual β_j are unknown, one can multiply the explanatory variables with these $\hat{\beta}_j$ and obtain

$$\hat{y}_i = x_{i1}\hat{\beta}_1 + ... + x_{ip}\hat{\beta}_p. \tag{5.5}$$

where $\hat{y}_i$ is called the predicted or estimated value of y_i. The residual r_i of the i^{th} case is the difference between what is actually observed and what is estimated:

$$r_i = y_i - \hat{y}_i. \tag{5.6}$$

As an illustration, at the effect of outliers in the simple regression model

$$y_i = \beta_1 x_i + \beta 2 + e_i \tag{5.7}$$

in which the slope β_1 and the intercept β_2 are to be estimated. In the simple regression model, one can make a plot of the (x_i, y_i), which is some time called a scatter plot, in order to visualize the data structure.

5.3 Multivariate Outlier Detection

In the example below, a random number has been generated, which has two columns, x and y. After that, outliers are detected separately from x and y. outliers are taken in to those data which are identified outliers for both columns. In Figure 5.1 and 5.2, outliers are labeled with star symbol.

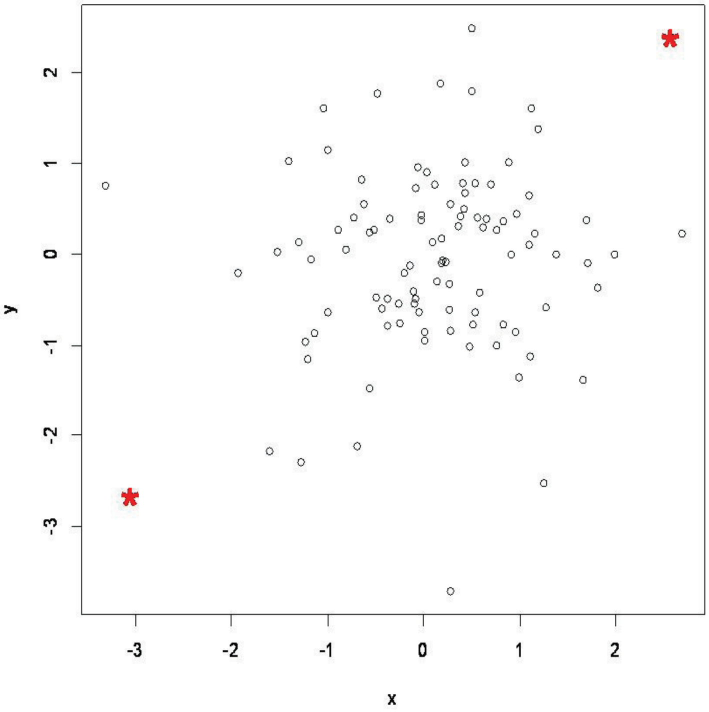

FIGURE 5.1: scatter plot for multivariate outlier detection - 1

When there are more than three variables in an application, a final list of outliers might be produced with majority of outliers detected from individual variables. This problem was illustrated with another real-world applications given as follows.

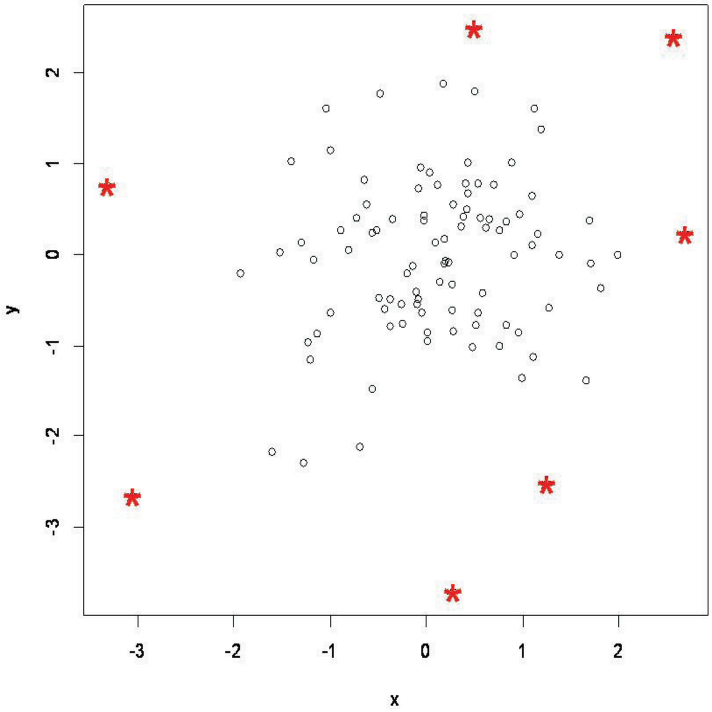

FIGURE 5.2: scatter plot for multivariate outlier detection - 2

5.4 Multivariate Normality Assumption

Multivariate normality is defined when each variable under consideration is normally distributed with respect to each other variable. Multivariate normal distributions take the form of symmetric three-dimensional bells when the x axis is the values of a given variable, the y axis is the count for each value of the x variable, and the z axis is the values of any other variable under consideration. Structural equation modeling and certain other procedures assume multivariate normality.

In Figure 5.3 and 5.4, respectively perspective and contour plots for identifying multivariate normality. The perspective plot were using binary data sets. In order to get a perspective plot continue with two variables that is bivariate normal distribution.

When the data is bivariate normal the perspective plot produces 3 dimensional

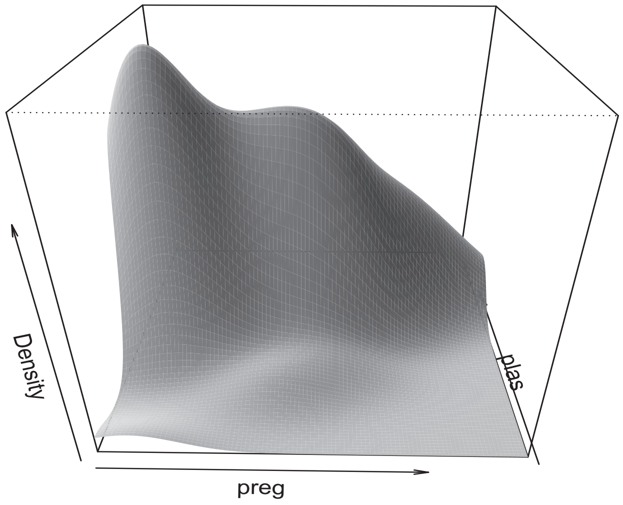

FIGURE 5.3: Perspective Plot for Normality

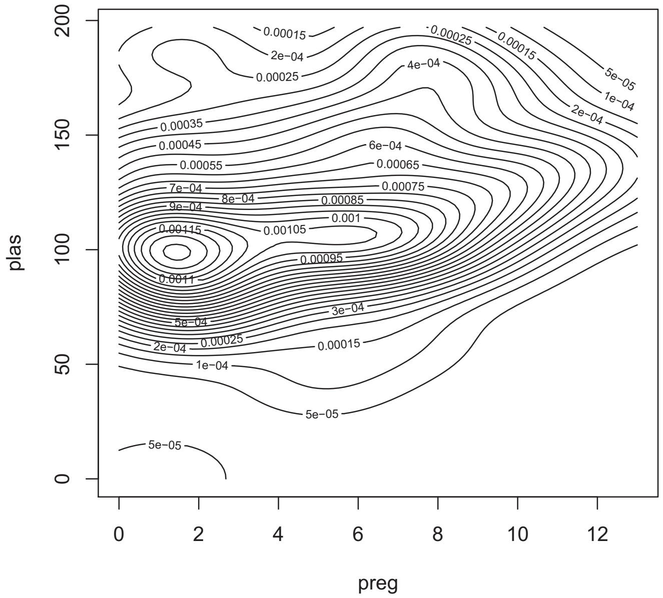

FIGURE 5.4: Contour plot for Normality

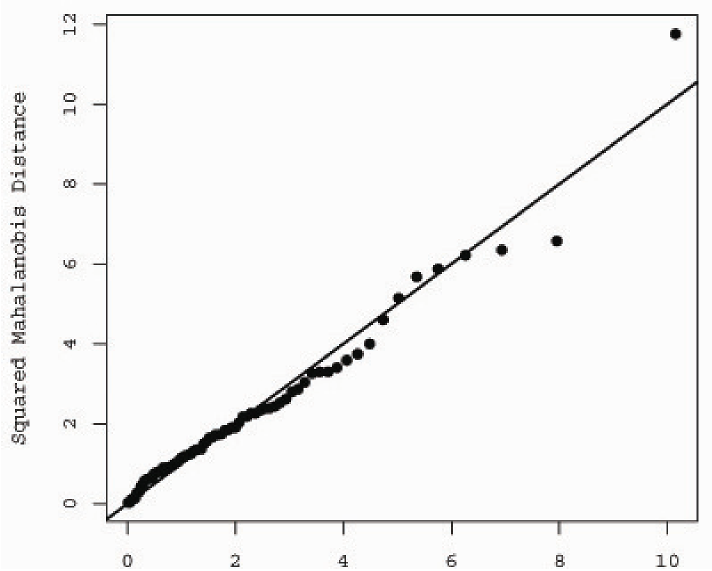

FIGURE 5.5: Q-Q Plot for Evaluating Multivariate Normality and Outliers

bell shaped in the graph. The contour plots are very useful since it gives the information about normality and correlation among pregnancy measures of flowers since contour lines lie around main diagonal.

5.4.1 Q-Q Plot for Outliers

The variable $d^2 = (x-\mu)^{'}\Sigma^{-1}(x-\mu)$ has a chi-square distribution with p degrees of freedom, and for "large" samples the observed Mahalanobis distances have an approximate chi-square distribution. This result can be used to evaluate (subjectively) whether a data point may be an outlier and whether the observed data may have a multivariate normal distribution.

A chi-square Q-Q plot can be used to whether there is any deviation from multivariate normality. Suppose the data has follows a multivariate normal distribution, the resulting plot should be roughly a straight line. We plot the ordered Mahalanobis distances versus estimated quantiles (percentiles) for a sample of size n from a chi-squared distribution with p degrees of freedom. This should resemble a straight-line for data from a multivariate normal distribution. Outliers will show up as points on the upper right side of the plot for which the Mahalanobis distance is notably greater than the chi-square quantile value.

5.5 Methods for Multivariate Outlier Detection

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

5.5.1 Mahalanobis Distance (MD_i)

Mahalanobis distance (MD) is a statistical measure of the extent to which cases are multivariate outliers, based on a chi-square distribution, assessed using $p < .001$. The shape and size of multivariate data are measured by the covariance matrix.

A familiar distance measure which takes into account the covariance matrix is the Mahalanobis distance. A classical approach for detecting outliers is to compute the Mahalanobis distance for each observation $x_i (i = 1, ..., n)$:

$$V = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})^T \tag{5.8}$$

Then

$$MD_i = \sqrt{(x_i - \bar{x})^T V^{-1} (x_i - \bar{x})} \tag{5.9}$$

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.

Hadi (1992) proposed an iterative Mahalanobis distance type of method for the detection of multiple outliers in multivariate data which tackles problems which can arise with method,

1. It will be refered to as method
2. Outlying observations may not necessarily have large values for D_i because masking may occure, i.e. one outlier masks the appearance of another outlier.

The critical chi-square values for 2 to 10 degrees of freedom at a critical alpha of .001 are shown below.

TABLE 5.1: Chi-square values for 2 to 10 degrees of freedom at a critical $\alpha = 0.001$

df	**Critical value**
2	13.82
3	16.27
4	18.47
5	20.52
6	22.46
7	24.32
8	26.13
9	27.88
10	29.59

For df > 10, refer to a complete table is provided in Appendix C.

A maximum MD_i is larger than the critical chi-square value for $df = k$ (the number of predictor variables in the model) at a critical alpha value of .001 indicates the presence of one or more multivariate outliers.

5.5.2 Cook's distance (Di)

Dennis Cook (1977) introduced a distance measure for commonly used estimates to study 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.

$$D_i = \frac{\sum_{j=1}^{n}(\hat{Y}_i - \hat{Y}_j(i))^2}{pMSE} \tag{5.10}$$

The following equation equally expressed as,

$$D_i = \frac{e_i^2}{pMSE}\left[\frac{h_{ii}}{(1-h_{ii})^2}\right] \tag{5.11}$$

$$D_i = \frac{(\hat{\beta} - \hat{\beta}^{-i})^T (X^T X)(\hat{\beta} - \hat{\beta}^{-i})}{(1+p)s^2} \tag{5.12}$$

where, $\hat{\beta}$ is the 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.

5.5.3 Leverage Point(h_i)

Outliers play an important role in regression. It is common practice to distinguish between two types of outliers. Outliers in the response variable represent model failure. Such observations are called outliers. Outliers with respect to the predictors are called leverage points. They can affect the regression model, too. Their response variables need not be outliers.

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 easy to detect. In linear regression model, the leverage score for i^{th} data unit is defined as,

$$h_{ii} = (H)_{ii} \tag{5.13}$$

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.

5.5.4 DFFITS

DFFITS is the diagnostics tool for statistical regression model shows that influence point. It is a standardized function of the difference between the predicted value for the observation when it is included in the dataset and when (only) it is excluded in the dataset. This quantity measures how much the regression function changes at the i^{th} observation when the i^{th} variable is deleted.

$$DFFITS_i = \frac{\hat{Y}_i - \hat{Y}_{i(i)}}{s_{(i)}\sqrt{v_{ii}}} \tag{5.14}$$

$$= \frac{v_{ii}}{1 - v_{ii}}\left(\frac{e_i}{s_{(i)}\sqrt{1 - v_{ii}}}\right) \tag{5.15}$$

$$(n-p'-1)s_{(i)}^2 = (n-p')s^2 - \frac{e_i^2}{1-v_{ii}}$$

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

5.6 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.

1. Initialize the data, and identify the normality assumption is to the given data set whether distributed normal or not.
2. If yes, compute the four distance measures i.e., Mahalanobis, Cook's distance, Leverage, DFFITS, for step by step procedures.
3. Then, simultaneously apply a multivariate outlier detection rule to each distance measures.
4. An observations is considered as outliers, if it is an outlier for every distances.
5. Whether the number of outliers has tabulated and visualized by bar chart. Finally the result will be given as follows.

The various four distances measures are computed by the statistical software (MiniTab.version.16, R, SPSS.version.21, and SAS.version.9.1).

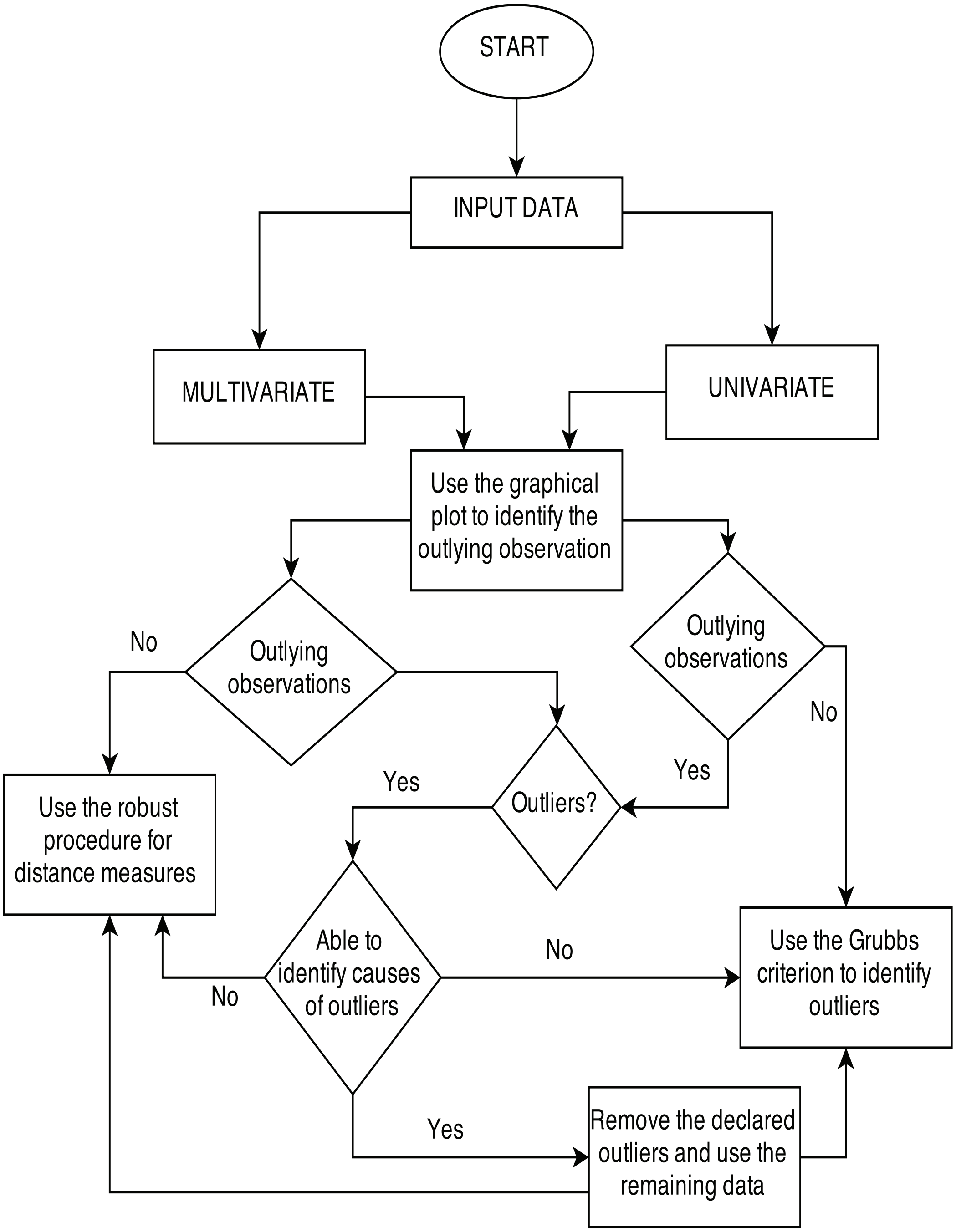
START
INPUT DATA
MULTIVARIATE
UNIVARIATE
Use the graphical plot to identify the outlying observation
No
Outlying observations
Outlying observations
No
Yes
Yes
Outliers?
Use the robust procedure for distance measures
Able to identify causes of outliers
No
No
Use the Grubbs criterion to identify outliers
Yes
Remove the declared outliers and use the remaining data

FIGURE 5.6: computational procedure for detecting multivariate outlier

5.7 Numerical Illustration

In this chapter, datasets are extracted from url 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 and Pediatric. This chapter is used for diabetes datasets and variable name detailed information is given in Appendix D.

In Table 5.2, 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.5.2) represents the outlier points which are identified by the various distance measures.

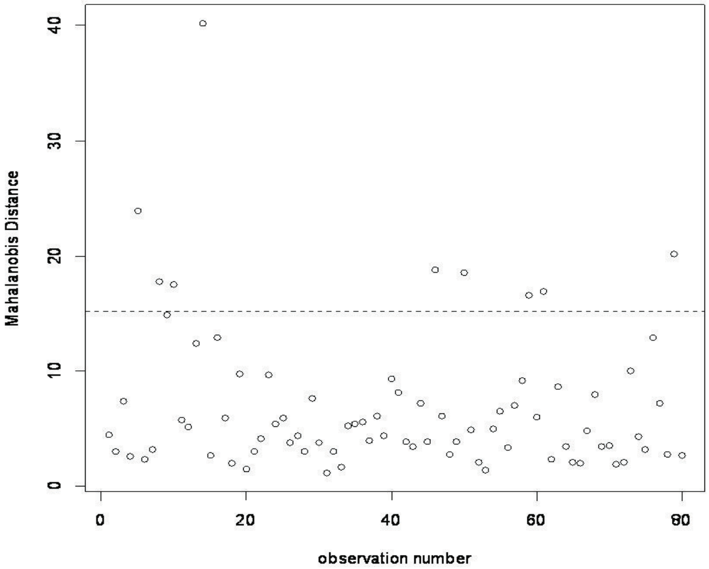

FIGURE 5.7: Plot for Mahalanobis Distance using outlier detection

TABLE 5.2: Outliers identified by Mahalanobis Distance(MD_i), Cook's distance (D_i), Leverage Point (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	$\underline{23.883}$	$\underline{.042}$	$\underline{.302}$	-0.581	45	3.871	.002	.049	0.117
6	2.301	.001	.029	-0.086	46	$\underline{18.796}$	$\underline{.065}$	$\underline{.238}$	-0.722
7	3.175	.000	.040	-0.034	47	6.072	.000	.077	0.059
8	$\underline{17.765}$	.007	$\underline{.225}$	-0.243	48	2.770	.004	.035	-0.167
9	14.880	$\underline{.081}$	.188	$\underline{0.813}$	49	3.845	.001	.049	-0.098
10	$\underline{17.505}$	$\underline{.141}$	$\underline{.222}$	$\underline{1.083}$	50	$\underline{18.503}$	.008	$\underline{.234}$	-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	$\underline{40.231}$	$\underline{.235}$	$\underline{.509}$	$\underline{1.379}$	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	$\underline{.050}$	.124	0.641	59	$\underline{16.547}$	.022	$\underline{.209}$	0.420
20	1.508	.000	.019	0.061	60	5.993	.002	.076	-0.115
21	3.061	.017	.039	0.523	61	$\underline{16.958}$	.004	$\underline{.215}$	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	$\underline{.088}$	.101	$\underline{0.865}$
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	$\underline{.036}$	.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	$\underline{.044}$	.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	$\underline{20.142}$	$\underline{.071}$	$\underline{.255}$	$\underline{0.758}$
40	9.317	$\underline{.058}$	.118	0.061	80	2.666	.000	.034	-0.058

TABLE 5.3: Number of Outliers detected by various distances

Distances Measures	Outliers Detected
Mahalanobis Distance(MD_i)	9
Cook's distance(D_i)	11
Leverage values(h_i)	9
DFFITS	5

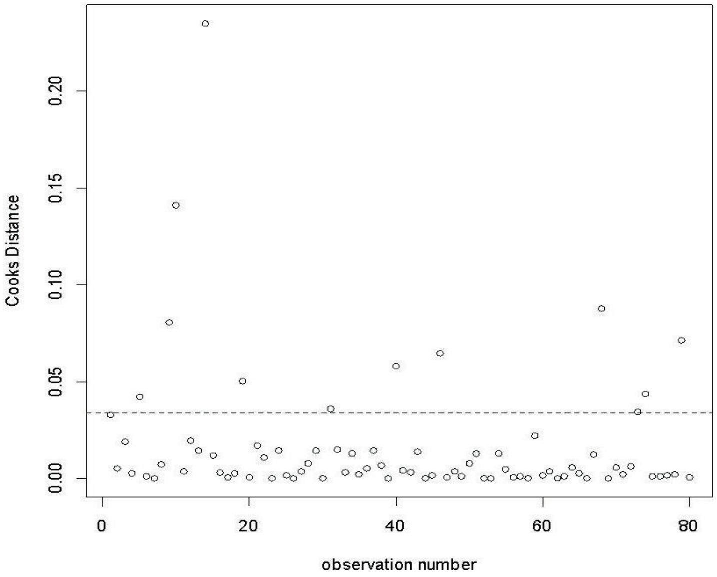

FIGURE 5.8: plot for Cook's Distance using outlier detection

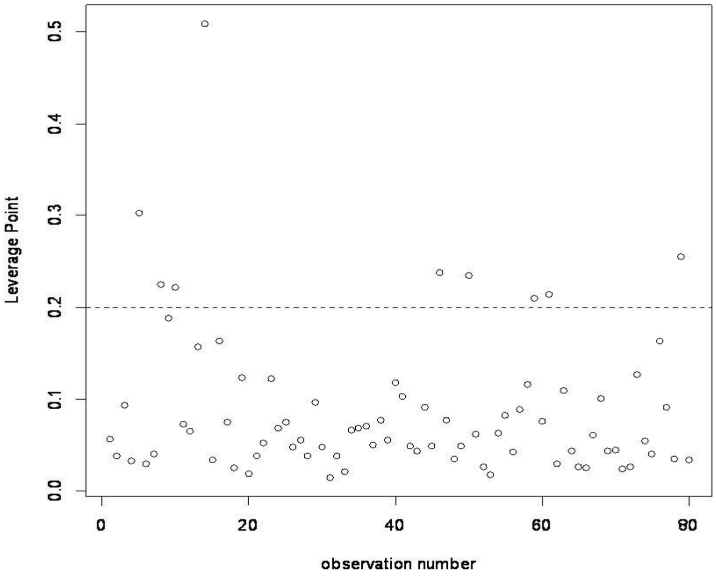

FIGURE 5.9: plot for leverage using outlier detection

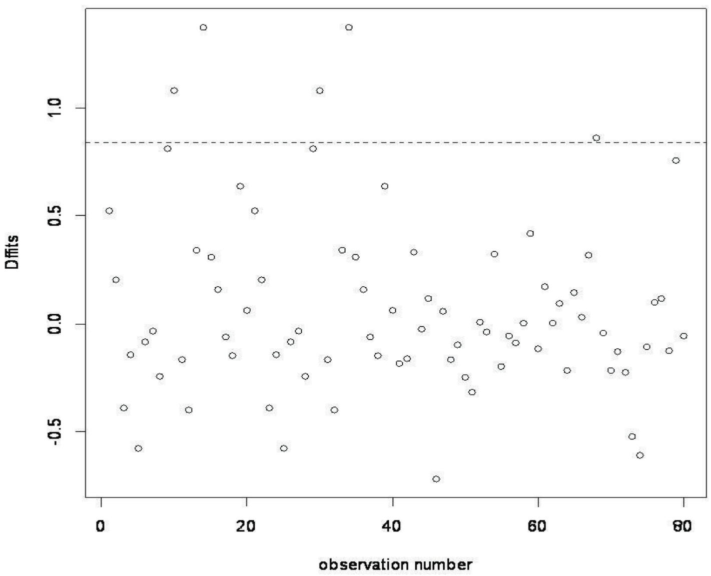

FIGURE 5.10: Plot for Dffit using outlier detection

5.8 Conclusion

This chapter deals with the procedure for computing the presence of outliers using various distance measures such as Mahalanobis Distance (MD_i), Cook's 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 Cook's Distance (D_i) is 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.