Available online at **www.ijmrhs.com**

ISSN No: 2319-5886

International Journal of Medical Research & Health Sciences, 2017, 6(7): 50-56

Detection of Outliers in Regression Model for Medical Data

Stephen Raj S[1]* and Senthamarai Kannan K[2]

[1]Research Scholar, Department of Statistics, Manonmaniam Sundaranar University, Abhishekapatti, Tirunelveli, Tamil Nadu, India

[2]Professor, Department of Statistics, Manonmaniam Sundaranar University, Abhishekapatti, Tirunelveli, Tamil Nadu, India

**Corresponding e-mail: stephenstats17@gmail.com*

ABSTRACT

In regression analysis, an outlier is an observation for which the residual is large in magnitude compared to other observations in the data set. The detection of outliers and influential points is an important step of the regression analysis. Outlier detection methods have been used to detect and remove anomalous values from data. In this paper, we detect the presence of outliers in simple linear regression models for medical data set. Chatterjee and Hadi mentioned that the ordinary residuals are not appropriate for diagnostic purposes; a transformed version of them is preferable. First, we investigate the presence of outliers based on existing procedures of residuals and standardized residuals. Next, we have used the new approach of standardized scores for detecting outliers without the use of predicted values. The performance of the new approach was verified with the real-life data.

Keywords: Medical data, Outlier, Residual analysis, Regression and residual analysis

INTRODUCTION

Regression analysis is a statistical technique for analysing and modelling the relationship between dependent variable and one or more independent variables. This technique uses the mathematical equation to establish the relationship between variables. It is a predictive modelling technique used for forecasting and to find casual effect relationship between the variables. The applications of regression analysis were found in almost every field including physical and chemical sciences, engineering, economics, finance, pharmacology, life and biological sciences, social sciences, and other fields of study. In simple linear regression model, only one independent variable (x) is used to predict a single dependent variable (y). The scatter diagram is used to diagrammatically display the relationship between independent variable and dependent variable.

The equation of a straight line relating these two variables is given by Montgomery, et al. [1].

$$y = \beta_0 + \beta_1 x \tag{1}$$

where β_0 is the intercept and β_1 is the slope.

The difference between the observed value of y and the fitted straight line is a statistical error ε. It is a random variable that accounts for the failure of the model to fit the data exactly.

Hence the model is given by

$$y = \beta_0 + \beta_1 x + \varepsilon \tag{2}$$

which is called a 'simple linear regression model'.

The important objective of regression analysis is to estimate the unknown parameters β_0 and β_1 in the regression model. There are several techniques are available for estimating the unknown parameters, here we use the 'method of least squares'. In the method of least squares, we will estimate β_0 and β_1 so that the sum of the squares of the

differences between the observations y_i and the straight line is a minimum. The ordinary least squares (OLS) method has been used to fit the model and to estimate the parameter values. There are several assumptions that must be fulfilled for the OLS model to be valid. When the regression model does not satisfy the fundamental assumptions of the model, predictions and estimations based on the model, may be biased [2].

The least square estimator of the intercept β_0 is:

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} \tag{3}$$

The least square estimators of the slope β_1 is:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} y_i x_i - \frac{\left(\sum_{i=1}^{n} y_i\right)\left(\sum_{i=1}^{n} x_i\right)}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}} \tag{4}$$

Then the fitted simple linear regression model is given by:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \tag{5}$$

which gives a point estimate of the mean of y for a particular x.

The difference between the observed value y_i and the corresponding fitted value is called residual. Mathematically the i^{th} residual is given by Bipin et al. [3].

$$e_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i), \qquad i = 1, 2, ..., n \tag{6}$$

Residuals play an important role in investigating the adequacy of the fitted regression model and in detecting departures from the underlying assumptions.

After obtaining the least squares fit, we should check for the following:

• How well does this equation fit the data?

• Is the model likely to be useful as a predictor?

• Are any of the basic assumptions (such as constant variance and uncorrelated errors) violated?

All of these issues must be investigated before the model is finally adopted for use. Outliers/bad values can seriously disturb the least-squares fit. An observation falls far away from the line implied by the rest of the data. If this point is really an outlier, then the estimate of the intercept may be incorrect. On the other hand, the data point may not be a bad value and may be a highly useful piece of evidence concerning the process under investigation.

The major assumptions of the regression analysis are as follows: [4].

i. The relationship between the response y and the regressor's x is linear, at least approximately.

ii. The error term ε has zero mean.

iii. The error term ε has constant variance σ^2.

iv. The errors are uncorrelated.

v. The errors are normally distributed.

Assumptions (iv) and (v) implies that the errors are independent random variables. Assumption (v) is required for hypothesis testing and interval estimation.

The appropriateness of the model is studied and the quality of the fit is ascertained by model adequacy checking t or F-statistics or R^2.

Outliers are observations that appear inconsistent with the remainder of the data set [5]. Outliers may be mistakes, or else accurate but unexpected observations which could shed new light on the phenomenon under study [6]. In this study, we have concentrated on outlier detection methods on linear regression model. Specifically, we are concerned with observations that differ from the regression plane defined by the bulk of the data. It is important to identify these types of outliers in regression modelling because the observations, when undetected, can lead to erroneous parameter estimates and inferences from the model [7]. Identifying outliers in the real-world database is important for improving the quality of original data and for reducing the impact of outliers [8]. The standard outlier detection procedures are based on residuals, which require the predicted value. Hence, we have used a new approach without using residuals. The performance of the new approach was verified by using the real-life data set, based on medical data pertaining to the age and systolic blood pressure (mm Hg) of 30 people of different ages, was retrieved from the web site of Florida State University [9]. The increase in blood pressure with age is mostly associated with structural changes in the arteries and especially with large artery stiffness, which is associated with increased cardiovascular risk [10]. On average, systolic blood pressure increases with age, while diastolic blood pressure increases to age 50 and then decrease [2].

MATERIALS AND METHODS

Methods of outlier detection in regression

There are many methods already exists for the detection of outliers in linear regression. They may be classified into two groups, namely graphical and analytical methods [11,12].

Outliers were detected based on the following methods:

i. Residual analyses using standardized residuals, studentized residuals, jackknife residuals and predicted residuals;

ii. Residuals plots such as the graph of predicted residuals, the Williams graph, and the Rankit Q-Q plot;

iii. Scalar measures of influence statistics such as cook's Di (measures the change in the estimates that outcome of deleting each observation), DFFITSi (measures the change in the predicted value of the dependent variable when the current value is omitted from the calculations), DFBETASj (i) (measures the influence on regression coefficients), Atkinson measures, and the Covariance ratio (measure of model performance).

Diagnostics based on residual analysis

1. Residuals: The residual is defined as:

$$e_i = y_i - \hat{y}_i, \qquad i = 1, 2, ..., n \tag{7}$$

where yi is an observation (dependent variable) and is the corresponding fitted value. Since a residual may be viewed as the deviation between the data and the fit, it is a measure of the variability in the dependent variable not explained by the regression model.

2. Standardised residuals (Normalised): Chatterjee and Hadi discuss that the normal residuals are not appropriate for diagnostic purposes; a transformed version of them would be better. Transformations of residuals such as standardized residuals, studentised residuals, jackknife residuals and predicted residuals, are often preferred over raw residuals because they overcome some of the limitations of raw residuals.

A logical scaling for the residuals is the standardised residuals and is given by Pimpan et al. [13].

$$d_i = \frac{e_i}{\sqrt{MS_{RES}}}, i = 1, 2, 3, ..., n \tag{8}$$

where MSRES is the mean square residual. The standardized residuals have mean zero and approximately unit variance. Consequently, large standardized residuals (d_i>3) potentially indicate an outlier.

3. Standardized scores: In this paper, a new approach for outlier detection was used to detect the values in linear regression models. This method is based on the individual standardized scores of dependent variable (y) and independent variable (x).

$$x_i = \frac{x_i - \mu}{\sigma}, \quad i = 1, 2, ..., n \tag{9}$$

$$x_i = \frac{x_i - \mu}{\sigma}, \quad i = 1, 2, ..., n \tag{10}$$

RESULTS AND DISCUSSION

In this paper, the presence of outliers in blood pressure data based on residuals obtained from the fitted simple linear regression model have been studied and the relationship between age and SBP are investigated. Furthermore, we investigate the presence of outliers based on residuals and standardized residuals (Table 1).

Figures 1 and 2 shows the scatter plot, which suggests that there is a moderate statistical relationship between age versus SBP, and the tentative assumption of the straight-line model $y = \beta_0 + \beta_1 x + \varepsilon$ appears to be reasonable.

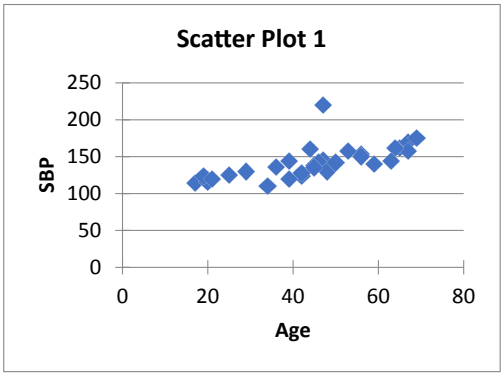

Figure 1 Scatter plot for the data set with outlier

Using MS-Excel, the following regression model is fitted to the medical data pertaining to the systolic blood pressure was measured for 30 people of different ages (n=30).

$$\hat{y} = 98.71 + 0.97x \tag{11}$$

with R_2=43% (where y is the systolic blood pressure and x is the age). The residuals (Column 5) and the standard residuals (Column 6) have been shown in Table 1 were taken from the excel output. Table 1 displays the observed values yi, the fitted values, residuals, standard residuals and standardized x_i and y_i scores.

From Table 1, we can observe that the residual e_2=75.65 is very large, the standardized residual d_2=4.45 exceeded the cut-off value of >3, standardized y_i score y_2=3.43 exceeded the cut-off value of >3; therefore, the observations at the data point 2 is considered as outliers. Table 2 displays the descriptive statistics and Table 3 shows that the model fitting information and summary statistics for the dependent and independent variables. The outliers detected by the method of residuals and standard residuals are similar to those detected by standardized score approach.

Table 1 Tabulation of residuals, standard residuals and difference method (n=30)

S. no	Age (x_i)	SBP (y_i)	Predicted SBP ($\hat{y}_i$)	Residuals (e_i)	Standard Residuals (d_i)	Standardized (x_i)	Standardized (y_i)
(1)	(2)	(3)	(4)	(5)	(6)	(8)	(9)
1	39	144	136.58	7.42	0.44	-0.4	0.06
2	47	220	144.35	**75.65**	**4.45**	0.12	**3.43**
3	45	138	142.4	-4.4	-0.26	-0.01	-0.2
4	47	145	144.35	0.65	0.04	0.12	0.11
5	65	162	161.82	0.18	0.01	1.3	0.86
6	46	142	143.37	-1.37	-0.08	0.06	-0.02
7	67	170	163.76	6.24	0.37	1.43	1.22
8	42	124	139.49	-15.49	-0.91	-0.2	-0.82
9	67	158	163.76	-5.76	-0.34	1.43	0.68
10	56	154	153.08	0.92	0.05	0.71	0.51

11	64	162	160.85	1.15	0.07	1.23	0.86
12	56	150	153.08	-3.08	-0.18	0.71	0.33
13	59	140	156	-16	-0.94	0.91	-0.11
14	34	110	131.72	-21.72	-1.28	-0.73	-1.44
15	42	128	139.49	-11.49	-0.68	-0.2	-0.64
16	48	130	145.32	-15.32	-0.9	0.19	-0.56
17	45	135	142.4	-7.4	-0.44	-0.01	-0.33
18	17	114	115.22	-1.22	-0.07	-1.84	-1.26
19	20	116	118.13	-2.13	-0.13	-1.64	-1.18
20	19	124	117.16	6.84	0.4	-1.71	-0.82
21	36	136	133.67	2.33	0.14	-0.6	-0.29
22	50	142	147.26	-5.26	-0.31	0.32	-0.02
23	39	120	136.58	-16.58	-0.97	-0.4	-1
24	21	120	119.1	0.9	0.05	-1.58	-1
25	44	160	141.43	18.57	1.09	-0.07	0.77
26	53	158	150.17	7.83	0.46	0.51	0.68
27	63	144	159.88	-15.88	-0.93	1.17	0.06
28	29	130	126.87	3.13	0.18	-1.05	-0.56
29	25	125	122.99	2.01	0.12	-1.32	-0.78
30	69	175	165.7	9.3	0.55	1.56	1.44

Table 2 Descriptive statistics for the data set

Variables	Age	SBP
Mean	45.13	142.53
Median	45.5	141
Mode	39	144
Standard Deviation	15.29	22.58
Range	52	110
Minimum	17	110
Maximum	69	220
Count	30	30

Table 3 Summary output for the data set

Regression Statistics	
Multiple R	0.66
R Square	0.43
Intercept	98.71
Slope (Age)	0.97

In regression analysis, the effect of the case can be studied by deleting the particular case from the data and analysing the rest of the population. Hence the results after deleting the 2nd observation with sample size (n=29) are shown below (Tables 4-6).

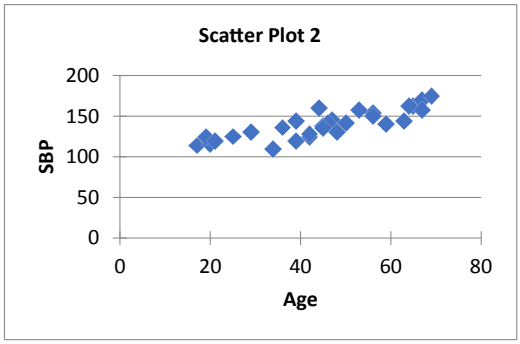

Figure 2 Scatter plot for the data set without outlier

Table 4 Tabulation of residuals, standard residuals and difference method (n=29)

S. no	Age (x_i)	SBP (y_i)	Predicted SBP ($\hat{y}_i$)	Residuals (e_i)	Standard Residuals (d_i)	Standardized (x_i)	Standardized (y_i)
(1)	(2)	(3)	(4)	(5)	(6)	(8)	(9)
1	39	144	134.1	9.9	1.05	-2.83	-1.99
2	45	138	139.8	-1.8	-0.19	-2.91	-2.68
3	47	145	141.7	3.3	0.35	-2.87	-2.39
4	65	162	158.78	3.22	0.34	-2.87	-2.45
5	46	142	140.75	1.25	0.13	-2.89	-2.51
6	67	170	160.68	9.32	0.99	-2.83	-2.11
7	42	124	136.95	-12.95	-1.38	-2.99	-3.31
8	67	158	160.68	-2.68	-0.29	-2.91	-2.79
9	56	154	150.24	3.76	0.4	-2.87	-2.39
10	64	162	157.83	4.17	0.44	-2.87	-2.39
11	56	150	150.24	-0.24	-0.03	-2.9	-2.62
12	59	140	153.09	-13.09	-1.39	-2.99	-3.36
13	34	110	129.35	-19.35	-2.06	-3.03	-3.65
14	42	128	136.95	-8.95	-0.95	-2.96	-3.08
15	48	130	142.64	-12.64	-1.35	-2.98	-3.31
16	45	135	139.8	-4.8	-0.51	-2.93	-2.85
17	17	114	113.22	0.78	0.08	-2.89	-2.45
18	20	116	116.06	-0.06	-0.01	-2.9	-2.51
19	19	124	115.11	8.89	0.95	-2.84	-1.99
20	36	136	131.25	4.75	0.51	-2.86	-2.28
21	50	142	144.54	-2.54	-0.27	-2.91	-2.73
22	39	120	134.1	-14.1	-1.5	-2.99	-3.36
23	21	120	117.01	2.99	0.32	-2.88	-2.33
24	44	160	138.85	21.15	2.25	-2.75	-1.36
25	53	158	147.39	10.61	1.13	-2.82	-1.99
26	63	144	156.88	-12.88	-1.37	-2.98	-3.36
27	29	130	124.61	5.39	0.57	-2.86	-2.22
28	25	125	120.81	4.19	0.45	-2.87	-2.28
29	69	175	162.58	12.42	1.32	-2.81	-1.93

Table 5 Descriptive statistics for the data set

Variables	Age	SBP
Mean	45.07	139.86
Median	45	140
Mode	39	144
Standard Deviation	15.56	17.5
Range	52	65
Minimum	17	110
Maximum	69	175
Count	29	29

Table 6 Summary output for the data set

Regression Statistics	
Multiple R	0.84
R Square	0.71
Intercept	97.08
Slope (Age)	0.95

CONCLUSION

In this paper, the detection of outliers in simple linear regression model have been discussed. A new approach for detecting outliers without the use of predicted values have been proposed. Which is quite useful in detecting outliers,

detects the outliers as same as the residual and standardized residual method. Hence, we suggest that in simple linear regression model, the difference method can be used for detecting outliers. Also by removing the influential point it is found that the model adequacy has been increased (from R_2=0.43 to R_2=0.71).

ACKNOWLEDGEMENT

The first author expresses his gratitude to the UGC for awarding the scheme of Basic Science Research Fellowship (BSRF) for providing financial support to carry out his work. The second author acknowledges the UGC for providing financial support to carry out this work under scheme UGC SAP (DRS-1).

REFERENCES

[1] Montgomery, Douglas C., Elizabeth A. Peck, and G. Geoffrey Vining. *Introduction to Linear Regression Analysis*. John Wiley & Sons, NY, USA. 2015.

[2] Leroy, Annick M., and Peter J. Rousseeuw. "Robust regression and outlier detection." *Wiley Series in Probability and Mathematical Statistics, John Wiley & Sons, NY, USA*. (1987).

[3] Bipin Gogoi and Mintu Kr. Das. Usage of graphical displays to detect outlying observations in linear regression. *Indian Journal of Applied Research* 5.5 (2015): 19-24.

[4] Framstad, Erik, Steinar Engen, and Nils Chr. "Regression analysis, residual analysis and missing variables in regression models." *Oikos* (1985): 319-323.

[5] Barnett, Vic, and Toby Lewis. *Outliers in Statistical Data*. Vol. 3. No. 1. New York: Wiley, 1994.

[6] Stefansky, Wilhelmine. "Rejecting outliers by maximum normed residual." *The Annals of Mathematical Statistics* 42.1 (1971): 35-45.

[7] Wisnowski, James W., Douglas C. Montgomery, and James R. Simpson. "A comparative analysis of multiple outlier detection procedures in the linear regression model." *Computational Statistics & Data Analysis* 36.3 (2001): 351-382.

[8] Rahman, SMA Khaleelur, M. Mohamed Sathik, and K. Senthamarai Kannan. "Multiple linear regression models in outlier detection." *International Journal of Research in Computer Science* 2.2 (2012): 23.

[9] Florida State University, Department of Scientific Computing. Available from: http://people.sc.fsu.edu/~jburkardt/datasets/regression/x03.txt.

[10] Pinto, Elisabete. "Blood pressure and ageing." *Postgraduate Medical Journal* 83.976 (2007): 109-114.

[11] Rockwood, Michael RH, and Susan E. Howlett. "Blood pressure in relation to age and frailty." *Canadian Geriatrics Journal: CGJ* 14.1 (2011): 2.

[12] Rajarathinam, A., and B. Vinoth. "Outlier detection in simple linear regression models and robust regression–A case study on wheat production data." *Statistics* 3.2(2014).

[13] Ampanthong, Pimpan, and Prachoom Suwattee. "A comparative study of outlier detection procedures in multiple linear regressions." *Proceedings of the International Multi-Conference of Engineers and Computer Scientists*. Vol. 1. 2009.