BEST: International Journal of Humanities, Arts, Medicine and Sciences (BEST: IJHAMS)
ISSN (P): 2348-0521, ISSN (E): 2454-4728
Vol. 4, Issue 4, Apr 2016, 165-170

COMPARATIVE STUDY TO DETECT OUTLIER VALUES IN SIMPLE LINEAR REGRESSION BETWEEN NACY AND THE PROPOSED METHOD

REEM ALI AL-JARAH[1] & AZHARSALIM AL- SAFFAR[2]

[1]Assistant Proph, Principle Dentistry Science Branch, Collage of Dentistry, Mosul University, Iraq

[2] Research, Statistic Department, Mathematical and Computer Science College, Mosul University, Iraq

ABSTRACT

In this paper I proposed a way to detect Outlier values in the simple linear regression model. Supported statistics Robust is the Median and Mean Absolute Deviation from the sample median. They altered the way NASI Nabil in (2001).

Use each of the Mean Squares Error (MSE) and Coefficient of Determination (R^2) compare the two, the proposed method has proved more efficient.

KEYWORDS: Mean Squares Error (MSE), Coefficient of Determination**,** NASI Nabil

INTRODUCTION

I had the problem of Outliers in data with great interest by researchers, especially in recent years. And knead because many scholars recognize that the existence of such observation in datasets would distort the results of statistical analysis used for those data and skew the results. Which leads to unrealistic results?

The problem of Outlier values came from a number of reasons, including measurement error, and refers to the use of incorrect measurements of natural and also error reading registry ... Etc. and that the problem of gays are in the form of linear dependent variable (y_i) or annotated variables (X_i) or both. For this reason, many researchers have resorted to several methods to detect outlier values, whether traditional methods or Robust. These methods have occupied a large area of statistical research. And this paper aims to propose a method for detecting Outlier values in the data. It is a variation of the NASI method where tools were to use statistics Robust are not sensitive to gays. It aims also to compare proposed method with the NASI method. In order to facilitate the application of the proposed method without talking about the data used and the period which took the example (page 46, 1987 (Rousseau and Leroy)) it is the same data adopted by nsens ne way.

BASIC CONCEPTS

The Concept of Outlier Values [1]

Outlier values are those the call as witness that seem illogical compared to other data shows a clear and significant departure from the rest of the components of the sample which I found those observation. They are the result of read errors and calculation or the result of the interpretation of a particular phenomenon.

Outlier Values [4], [3], [2]

Define outlier values as those observation that fall away from the regression equation and have very wrong compared to the rest of the other natural observation in datasets so that will have an impact on the form of linear and estimates, so that this impact be learned by analyzing the full data and compare the value of MSE and value (R^2_{adj}) in both

cases so that these values can be corrected by deleting it in case of being derived from the error in the log data.

Immunity [5]

(Box) in 1953, the first to give the item of immunity obviously statistical sense describing statistical method that Robust if sensitive data Benders, which estimated survival despite the lack of an efficient access to the data fit the assumptions made, and hippocampus aim to:

- Description the best model matches the data.
- Diagnosis of outlier values.
- Diagnosis and attention points that have an impact on the data.
- Handle models assumed on the basis that errors deviate distribution as a result of the presence of homosexuals.

WAY NACY'S METHOD

In this way, the emphasis was on the Ellipse -shaped to spread the data points for simple linear regression, either outlier values are detected through the following steps:

- The linear regression model for equaling data.
- Apply equation following ellipse**.**

$$\frac{((x-\bar{x})\cos\theta+(y-\bar{y})\sin\theta)^2}{a^2}+\frac{((y-\bar{y})\cos\theta-(x-\bar{x})\sin\theta)^2}{b^2}=1 \quad (1)$$

Where:

X: The independent variable.

Y: Representing the variable.

$a=\frac{\gamma_1 S_x}{\cos\theta}$, $b=\frac{\gamma_2 S_y}{\cos\theta}$

θ : is the angle between the ribs a ,$\gamma_1 . S_x$ and between the ribs,b ($\gamma_2 . S_y$) which can be extracted from the following relationship:

$$\hat{\beta} = \tan\theta \rightarrow \theta = \tan^{-1}\hat{\beta}$$

$\hat{\boldsymbol{\beta}}$ **:** represents the slope of the regression line.

$\boldsymbol{\gamma_1}$**:** Default value which specifies how the length of main axis (Major Axis) of the ellipse if ($\frac{\pi}{4}<\theta<-\frac{\pi}{4}$) and vice versa (γ_1) determine the length of the secondary axis if ($\frac{\pi}{2}<\theta<\frac{\pi}{4}$ او $\frac{\pi}{4}<\theta<-\frac{\pi}{2}$ **).**

$\boldsymbol{\gamma_2}$**:** represents a default value which specifies the length of the secondary axis (Minor Axis) of the ellipse if ($\frac{\pi}{4}<\theta<-\frac{\pi}{4}$) and vice versa (γ_2) determine the length of main axis if ($\frac{\pi}{2}<\theta<\frac{\pi}{4}$ او $\frac{\pi}{4}<\theta<-\frac{\pi}{2}$) . And identify refund as follows:

$$\gamma_1 = \frac{\cos\theta}{S_x(\hat{\beta}^2+1)}\sqrt{\hat{\beta}\,(y_1-\bar{y})+(x_1-\bar{x}))^2+(\hat{\beta}^2(y_1-\bar{y})+\hat{\beta}\,(x_1-\bar{x}))^2} \quad (2)$$

$$\gamma_2 = \frac{\cos\theta}{S_y(\hat{\beta}^2+1)}\sqrt{\hat{\beta}\,(y_2 - \bar{y}) + (x_2 - \bar{x}))^2 + (\hat{\beta}^2(y_2 - \bar{y}) + \hat{\beta}\,(x_2 - \bar{x}))^2} \quad (3)$$

Where:

S_x: is the standard deviation of the independent variable (x).

S_y: The standard deviation of the variable (y).

$\bar{y}$: The arithmetic mean of the variable observation adopted.

$\bar{x}$: The arithmetic mean of the covariate sightings.

$\hat{\beta}$: The slope of the regression line.

(y_1 ,x_1): Assuming its farthest point average ($\bar{x}$).

(y_2 ,x_2): Assuming its farthest point average ($\bar{y}$).

That decision if the value is outlier or not is by comparing the left end in the relationship (1), with one after compensation for each pair of observation (X_0,Y_0).

If equal to 1 point located on the perimeter of the ellipse and if smaller point located inside. Or if it is larger than 1, the point is outside the ellipse thus diagnosed this point as a watch freak.

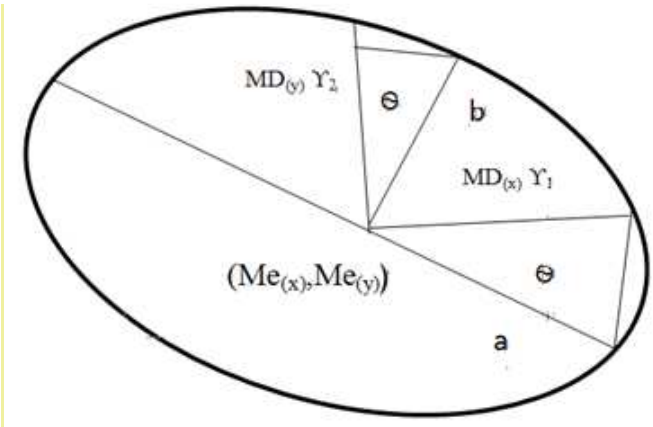
THE PROPOSED METHOD

After sharing the way Nacy's and other methods suggest circular Nacy's method and tools as follows:

- Use the Median and Mean Absolute Deviation from the sample median (MD) instead of the standard deviation for each of the approved and independent variables.
- use the mediator (Me) rather than the arithmetic mean for each of the approved and independent variables as:

$$MD = \frac{\sum_{i=1}^{n} |x_i - M_e|}{n} \quad (4)$$

- draw ellipse to be focused ($Me_{(x)}$, $Me_{(y)}$) and the main axis is a regression line, as shown in the chart (1) below:



Scheme 1: Illustrates the Ellipse

Is derived ellipse oblique equation recycles level (X-Y) to the level (X´-Y´) at an angle of (Θ) as Nacy's[4], and

after replacing Nacy's tools and a displacement process for the centre ellipse from point of origin to the point ($Me_{(x)}$,$Me_{(y)}$) are:

$$\frac{((x-Me_{(x)})\cos\theta+(y-Me_{(y)})\sin\theta)^2}{a^2}+\frac{((y-Me_{(y)})\cos\theta-(x-Me_{(x)})\sin\theta)^2}{b^2}=1 \quad (5)$$

Whereas:

$$a=\frac{\gamma_1 MD_{(x)}}{\cos\theta}, b=\frac{\gamma_2 MD_{(y)}}{\cos\theta}$$

$$\gamma_1=\frac{\cos\theta}{MD_{(x)}(\hat{\beta}^2+1)}\sqrt{\hat{\beta}\,(y_1-Me_{(y)})+(x_1-Me_{(x)}))^2+(\hat{\beta}^2(y_1-Me_{(y)})+\hat{\beta}\,(x_1-Me_{(x)}))^2} \quad (6)$$

$$\gamma_2=\frac{\cos\theta}{MD_{(y)}(\hat{\beta}^2+1)}\sqrt{\hat{\beta}\,(y_2-Me_{(y)})+(x_2-Me_{(x)}))^2+(\hat{\beta}^2(y_2-Me_{(y)})+\hat{\beta}\,(x_2-Me_{(x)}))^2} \quad (7)$$

PRACTICAL SIDE

We will discuss in this section the ability of the proposed method to detect outlier values in a simple linear model, and compared the results with the results reached by the way Nacy's. It has been selected the following variables:

Y: The growth rate represents the children.

X: Represents the average age when kids months.

They variables represent specific data for a group of children and which are shown in Table (1) below:

Table 1: Data on a Group of Children

التسلسل	Y	X
1	95	15
2	71	26
3	83	10
4	91	9
5	102	15
6	87	20
7	93	18
8	100	11
9	104	8
10	94	20
11	113	7
12	96	9
13	83	10
14	84	11
15	102	11
16	100	10
17	105	12
18	57	42
19	121	17
20	86	11
21	100	10

Source: From the Mickey el at (1967) as set forth Rousseeuw and Leroy at (1987) [7]

The Application of the Proposed Method

To facilitate the calculations (SPSS19 , -Excel-2010) where he found the statistics required for the application of

this method is the use of statistical programs, namely:

Me(x) = 11, Me(y) = 95

MD(X) = 4.809524, MD(Y) = 10.19048

$\hat{\beta} = -1.127$, $\hat{\alpha} = 109.87384$

Equation regression line$\hat{Y}$ = 109.87384 - 1.127 X ∴

$\theta = \tan\text{-}1\hat{\beta}$ = - 48.4169881

Cos θ = 0.6637044, Sin θ = -0.74799491

γ_1 7.246484355= a = $\frac{\gamma_1 \, MD(X)}{\cos\theta}$

b = $\frac{\gamma_2 \, MD(Y)}{\cos\theta}$= 15.35394228$\gamma_2$

Accordingly, the ellipse equation (5) of the data becomes:

$$\frac{(0.663704462\,X-0.74799491\,Y+63.75876737)^2}{(7.246484355\,\gamma_1)^2}+\frac{(0.663704462\,Y+0.74799491X-54.82397988)^2}{(15.35394228\gamma_2)^2}=1$$

Using equations (6) and (7) we find the values of (γ_1) and (γ_2) Altkadireeten as the Almhahd (18 and 19) are the farthest from each of the Me (x), Me(y) respectively, and the van:

(X_1, Y_1) = (42,57) , (X_2,Y_2) = (17,121)

Two by two points gay and thus be:$\gamma_1 = 0.2, \gamma_2$ = 0.5 .

Comparison of the Proposed Method and the Method of Nacy's after Deleting Outlier Observation

Table 2: Comparison of the Proposed Method and the Method of Nacy's

اسم الطريقة	MSE	R^2	method best
proposed method	77.991	0.270	proposed method
Method Nacy's	123.360	0.112	

We note from the table (2) above, that the proposed method gave anMean Squares Error (MSE=77.991), Less than one way Nacy's. We also note that the value of the coefficient of determination of the proposed method (R^2=0.270) is higher than the value of the coefficient of determination method Nacy's which shows that the proposed method best.

CONCLUSIONS AND RECOMMENDATIONS

In light of what has been presented in this paper we will summarize the conclusions and recommendations hill:

- Characterized the proposed use statistics robust.
- The proposed method showed sensitivity enough to detect outlier Observation.
- The proposed method demonstrated good efficiency Disclose Observation outlier.
- We recommend Alahtmta ways to identify outlier value in future studies and research so as to their importance and impact on the accuracy of the results.

REFERENCES

1. Altalb, Bashar Abdul Aziz Majid (1997) " the aiming programming of least squares regression to estimate the literal information compared to the" Master Thesis submitted to the Council of the Faculty of Administration and Economics - University of Mosul – Iraq.
2. Dabdoub, and Yunus (2005) the effect of outliers on the regression analysis with the application on preterm birth, "Faculty of Computer and Math / Science and Department of Statistics Amwalimatih, Rafidain magazine Science, Vol. 17 Issue 1:00 62-9-81-206).
3. Abdul Ahad and Manahel Daniel (2004) "appreciation of the hippocampus in the self-regression model of the first rank" Master Thesis submitted to the Council of the College of Computing and Mathematical Sciences. University of Mosul, Iraq.
4. Nancy, Nabil George (2001) "Assessment Methods for estimating the efficiency of the anomalous values of regression models," the Faculty of Administration and Economics, University of Mosul – Iraq.
5. 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. ISBN0-471-85233.
6. T. W. Anderson and J. D. Finn, (1996) “The New Statistical Analysis f Data. Springer-Verlag “New York, Inc. pp. 123-127.
7. Werner A. Stahel and Peter J. Rousseeuw, Elvezio M. Ronchetti Frank R. Hampel [3] (2005), Robust Statistics: The Approach Based on Influence Functions, John Wiley &Sons, Inc. pp.12-56.