Robust Outlier Detection in a Multivariate Linear Regression Model

Onisokumen David[1], Ijomah Maxwell A.[2]

[1]*Department of Mathematics/Statistics, Ignatius Ajuru University of Education, Nigeria*
[2]*Department of Mathematics/Statistics, University of Port Harcourt, Nigeria*

***Abstract:* - Outlier detection has been extensively studied and has gained widespread popularity in the field of statistics. As a consequence, many methods for detecting outlying observations have been developed and studied. However, a number of these approaches developed are specific to certain application domain in the univariate case, while apparently robust and useful have not made their way into general practice. In this paper, we considered Mahalanobis Distance technique, k-mean clustering technique and Principal component Analysis technique using data on birth weight, birth height and head circumference at birth from 100 infants from 2016 to 2019.To determine robustness among the multivariate outlier detection techniques, among others are selected for analysis. The Akaike's, Schwarz's and Hannan-Quinn criterion as well as the R^2 were used to determine the most robust regression among the selected models. Findings indicates that the k-mean Clustering technique outperforms the other two technique in regression model.**

***Key words:* Outlier Detection; Mahalanobis Distance; K-Clustering; Principal Component Analysis;**

I. INTRODUCTION

Multiple regression models are widely used in applied statistical techniques to quantify the relationship between a response variable *Y* and multiple predictor variables X$_i$, and we utilize the relationship to predict the value of the response variable from a known level of predictor variable or variables.

The models take the general form

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots\dots\dots\dots\dots\dots + \beta_P X_P + \varepsilon \quad (1)$$

where β_0 = a constant β_1, β_2, ..., β_p = regression parameters ε = random error . When we have n observations on Y and X$_i$'s this equation can be represented as follows

$$Y = X\beta + \varepsilon$$

where $Y = (y_1, y_2, \dots\dots, y_n)'$ is a n-vector of responses

$$X_{nx(p+1)} = \begin{bmatrix} 1x_{11}x_{21}\dots\dots x_{p1} \\ : \\ 1x_{12}x_{22}\dots\dots x_{p2} \\ : \\ 1x_{1n}x_{2n}\dots\dots x_{pn} \end{bmatrix} (2)$$

and $\beta = (\beta_0, \beta_1, \dots\dots, \beta_p)'$ is a p+1-vector parameters and $\varepsilon = (\varepsilon_1, \varepsilon_2, \dots\dots\dots., \varepsilon_n)'$ is a n-vector of error terms. An ordinary least square solution to (1) is one in which the Euclidean norm of the vector $(Y - X\beta)$ is minimized. That is ,

$$\min \|Y - X\beta\|_2$$

By setting the gradient of the square of this norm, $(Y - X\beta)^1 (Y - X\beta),$ to zero with respect to the vector β, the necessary condition for the solution vector $\hat{\beta}$ is that $\hat{\beta}$ must be a solution to

$$X^1 X\beta = X^1 Y \quad (3)$$

In other symbols, the solution is

$$\hat{\beta} = (X^1 X)^{-1} X^1 Y \quad (4)$$

The symbol I denotes the transposition of a vector or matrix.

Two major assumptions of ordinary least squares are that the errors are independent random variables and are normally distributed. If these assumptions are violated, then outlier is likely to manifest. Outliers are extreme values that deviate from other observations on data, they may indicate a variability in a measurement, experimental errors or a novelty. In other words, an outlier is an observation that diverges from an overall pattern on a sample. In linear regression, extensive study on the problem of outliers and leverage points can be found in the literature (see [1,2,6].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 [5]. An outlier may appear in the data due to an abnormal behavior in the data generation process. Therefore, it often contains utile information concerning the abnormal characteristics of the systems and entities influencing the data generation process. So, detecting those unusual characteristics provides useful knowledge specific to the application [14]. Outlier can affect multivariate analysis given incorrect conclusions and makes modelling difficult. Marascuilo and Serlin, [8]; Osborne and

Overbay, [9] writing on the impact of outlier in a multivariate regression analysis observed that outliers can have a dramatic impact on the results of common multivariate statistical analyses. Pedhazur, [10], in agreement wrote and said, they can distort correlation coefficients and create problems in regression analysis, even leading to the presence of collinearity among the set of predictor variables in multiple regression. Distortions to the correlation may in turn lead to biased sample estimates, as outliers artificially impact the degree of linearity present between a pair of variables. 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 concepts to look out for in data analysis. 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. Identifying outliers in multivariate data pose challenges 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 χ2 distribution Rousseeuw & Zomeren [13]. However, also values larger than this critical value are not necessarily outliers, they could still belong to the data distribution. Barnett [1]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.

This paper focuses on multivariate outlier detection as the use of some common summary statistics such as the sample mean and variance to detect outliers can cause the analyst to reach a conclusion totally opposite to the case if outliers weren't present. Furthermore, 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 problems [6]. 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. The rest of the paper is as follows; in section 2 gives a brief methods of outlier detection and discusses its computation and its main properties. Section 3 deals with the material and methods while in section 4, analysis of the result was carried out. Section 5 concludes with pointer to the most robust among these techniques.

II. 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.

Mahalanobis Distance (MD_i)

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

$$MD_i = \sqrt{\left(x_i - \bar{x}\right)^T V^{-1} \left(x_i - \bar{x}\right)} \qquad (5)$$

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 .

Cook's Distance (Di)

Dennis Cook [3] 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.

$$D_i = \frac{\sum_{j=1}^{n} \left(\left(\hat{Y}_i - \hat{Y}_j\right)(i)\right)^2}{MSE} \qquad (6)$$

The following equation equally expressed as,

$$D_i = \frac{{C_i}^2}{\rho MSE}\left[\frac{h_{ii}}{\left(1 - h_{ii}\right)^2}\right] \qquad (7)$$

$$D_i = \frac{\left(\hat{\beta} - \hat{\beta}^{-1}\right)^T \left(X^T X\right)\left(\hat{\beta} - \hat{\beta}^{-1}\right)}{\left(1 + \rho\right) S^2} \qquad (8)$$

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

K-Mean Clustering

k-Means which is firstly proposed by MacQueen [6], is a well-known and widely used clustering algorithm. k-Means is one of the simplest clustering algorithms in machine learning which can be used to automatically recognize groups of similar instances/items/objects/points in data training. The algorithm classifies instances to a pre-defined number of clusters specified by the use (e.g. assume k clusters). The first important step is to choose a set of k instances as centroids (centres of the clusters) randomly, usually choose one for each cluster as far as possible from each other. Next, the algorithm continues to read each instance from the data set and assigns it to the nearest cluster. There are some methods to measure the distance between instance and the centroid but the most popular one is Euclidian distance. The cluster centroids are always recalculated after every instance insertion. This process is iterated until no more changes are made. The k-Means algorithm is explained in this following pseudocode. Finally, this algorithm aims at minimizing an objective function, in this case a squared error function. The objective function

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2 \qquad (9)$$

where $\left\| x_i^{(j)} - c_j \right\|^2$ is a chosen distance measure between a data point $x_i^{(j)}$ and the cluster centrec_j is an indicator of the distance of the n data points from their respective cluster centres.

Robust PCA

The goal of robust PCA methods is to obtain principal components that are not influenced much by outliers. A first group of methods is obtained by replacing the classical covariance matrix by a robust covariance estimator. Principal Component Analysis is a dimension reduction technique which follows

$$\sum\nolimits_{\boldsymbol{A}} \boldsymbol{U} = U\Lambda, \qquad (10)$$

Where $\sum_{A} = \frac{\mathbf{1}}{\boldsymbol{n}} \sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{n}} (\boldsymbol{X_i} - \boldsymbol{\mu})(\boldsymbol{X_i} - \boldsymbol{\mu})^{T}$ is the covariance matrix, and $\boldsymbol{\mu}$ is the global mean, U represents an eigenvector of Σ_A, and Λ is the diagonal entry associated with the eigenvalue.

Leverage Point(hi)

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,

$$h_{ii} = (H)_{ii} \qquad (11)$$

The i^th diagonal of the hat matrix $H = X\left(X^{'}X\right)^{-1}X^{'}$. 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.

III. MATERIALS AND METHODS

The materials for the study were secondary data on birth weight, birth height and head circumference at birth of 100 infants ranging from 2016 to 2019, collected from ten health facilities in Ahoada West Local Government Area of Rivers State, Nigeria. For the method, there are many methods already exists for the detection of outliers in linear regression as indicated earlier. They may be classified into two groups, namely graphical and analytical methods [11,12]. In this paper, we considered and compared three of the multivariate techniques (i.e. Mahanalobis Distance, K-clustering and the Principal Component Analysis methods). We start by subjecting the data to each technique to check for presence of outlier. For comparison purposes, thereafter deleting the outliers detected by each method we then run a regression analysis on data now free from outlier for the data from each technique to obtain a robust regression estimates based on each method using the Akaike info criterion (AIC), Schwarz criterion (SWC) and Hannan-Quinn criterion (HQC).

IV. RESULT

In this paper, the presence of outliers in 100 infants data based on residuals obtained from the fitted multiple linear regression model have been studied and the relationship between height, weight and HCF are investigated. Furthermore, we investigate the presence of outliers based on mahanalobis distance, K-clustering and Principal Component Analysis methods.

Case 1: Mahanalobis Distance Technique:

As indicated in the objectives, we first consider use of mahalanobis distance for outlier detection. The result is as shown below:

Table 1a: Residual Statistics from Mahalanobis Distance Technique

	Minim	Maxim	Mean	Std. Deviation	N
Mahal. Distance	.024	12.308	1.980	2.364	100
Cook's Distance	.000	1.129	.016	.113	100
Centered Leverage Value	.000	.124	.020	.024	100
a. Dependent Variable: hcf					

From the regression analysis, we obtain the above table which shows that the maximum Mahalanobis distance value is 12.308. Using the probability value (P-value), it was found that there were four outlying values from the data. In other to remove the outlier, we delete the row where these values appears and rerun the regression and the result below is obtained.

Table 1b: Regression from Mahalanobis Distance

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	33.38227	4.061290	8.219623	0.0000
WEIGHT	0.195759	0.515806	0.379520	0.7052
HEIGHT	0.020620	0.073162	0.281839	0.7787

The above result shows that both weight and height have direct relationship with head circumference (HCF). However, the variables were not significant at 5% levels.

Case 2: K-Mean Clustering Technique:

For the k-mean clustering technique, in order to detect the outlier(s), we obtained two clustering centre and result is as shown below.

Table 2a: Cluster centers from k-mean clustering technique

	Cluster	
	1	2
Weight Height HCF	2.58 48.00 47.50	3.14 49.73 34.69

Choosing cluster group 2 which is higher than that of cluster group 1, we substrate these values from the original data and thus obtain a new regression as shown below.

Table 2b: Regression analysis based on K-Mean Clustering

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	44.39613	0.067317	-659.5095	0.0000
WEIGHT	0.224517	0.110407	2.033538	0.0448
HEIGHT	0.019721	0.015681	1.257655	0.2116

The regression result above shows again that, both weight and height have direct relationship with head circumference (HCF) but only weight has a significant effect on HCF at 5% levels, indicating outlyingness.

Case 3: The Principal Component Analysis

To perform the principal component analysis for outlier, we obtained a component related space as shown in figure 1 below

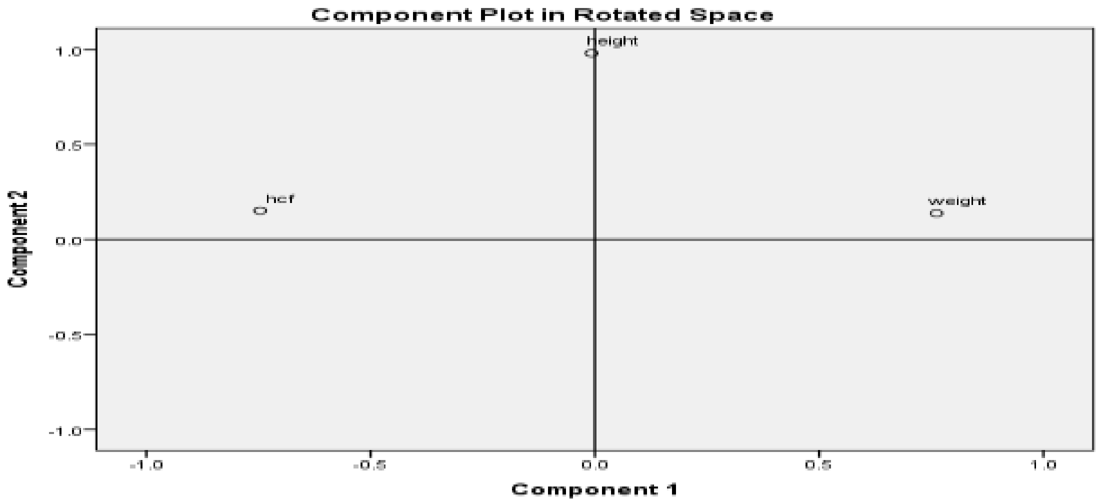

Figure 1

The plot indicates that there is a reasonable structure has been found, with most outlying observations seeming to belong to the height variable. A closer look at the figure 1 above reveals that outlier was detected in height based on this method. Therefore we choose height as our outlying variable and using the statistics below, we delete the outlier(s) and rerun a regression. The robust regression result is as shown in table 3 below.

Table 3: Regression analysis for Principal component analysis

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	34.72802	3.510917	9.891439	0.0000
WEIGHT	0.206577	0.477290	0.432812	0.6661
HEIGHT	-0.006334	0.064663	-0.097960	0.9222

The above regression result reveals that both explanatory variables were insignificant at 5% levels.

The result summary is as shown in table 4 below

Table 4: Summary Regression table for the three techniques

Variable	**Parameter estimate**	**Standard Error**	**t-value**	**Prob.**	**R^2**	**$F_{prob.}$**	**AIC**	**SWC**	HQC
Mahalanobis Distance									
Intercept	36.2859	4.0709	8.9134	0.0000	0.02	0.3688	5.1967	5.2753	5.2285
Weight	-0.7328	0.5285	-1.3864	0.1688					
Height	0.0242	0.0751	0.0751	0.7479					
K-Mean Clustering									
Intercept	-44.3961	0.0673	-659.50	0.0000	0.05	0.6693	2.0648	2.1434	2.0966
Weight	0.2245	0.1104	2.03354	0.0448					
Height	0.0197	0.0156	1.2577	0.2116					
Principal Component Analysis									
Intercept	34.7280	3.5109	9.8914	0.0000	0.02	2.7409	4.8849	4.9645	4.9171
Weight	0.2066	0.4773	0.4328	0.6661					
Height	-0.0063	0.0647	-0.0980	0.9222					

From the above summary result, it can be seen that, only the variable birth weight of the k-mean Clustering Model is significant at 5% level of significant. By model selection criterion, the preferred model is the model with smaller value of Akaike, Schwarz and Hannan-Quinn criterion. A close examination of the models indicates that, K-mean Clustering technique is more robust than the other two since its Akaike, Schwarz and Hannan criterion values are least compared to the value of the other two models. This is also confirmed in figure 3 below.

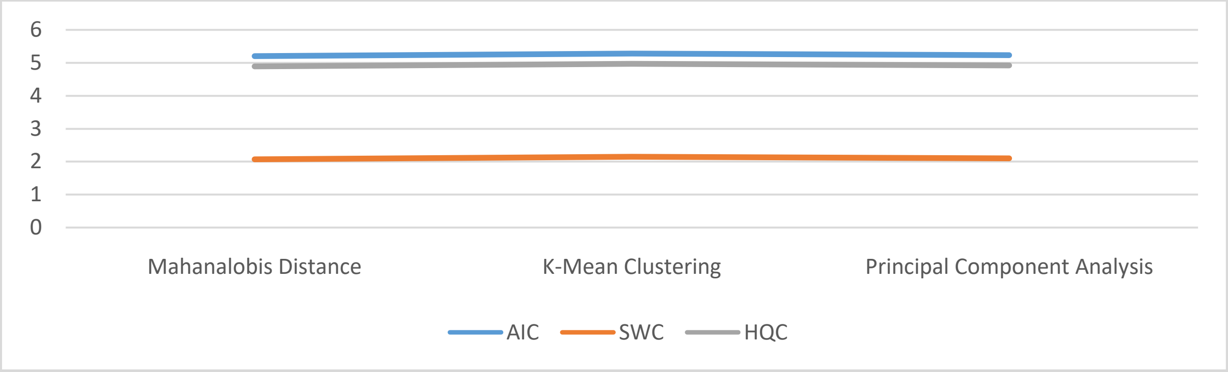

Fig.2: Comparison of model selection criteria between Mahanalobis Distance, K-Clustering and Principal Component Analysis methods

V. CONCLUSION

Robust Outlier detection methods in multiple linear regression are reviewed and investigated in this paper. Three of the robust Outlier detection methods were selected and compared using data of an infant at birth (i.e. weight, height and head circumference) from the year 2016 up until 2019 collected from ten health facilities to identify the most robust among them. Findings from the result shows that the K-mean Clustering outlier detection model is more robust than the other two methods, having the least selection criterion among the three methods. This results clearly reveals that K-mean Clustering is most sensitive to outlier than the other methods.

REFERENCES

[1] Barnett V and Lewis T (1984), “Outliers in statistical data,” John Wiley & Sons, New York.

[2] Belsley DA, Kuh E, and Welsch RE (1980) “Regression Diagnostic: Identifying influential data and sources of collinearity,” John Wiley & Sons, New York; Chichester.

[3] Cook, R. D. [Influential Observations, High Leverage Points, and Outliers in Linear Regression]: Comment. Statist. Sci. 1 (1986), no. 3, 393—397

[4] David M. Rocke and David L. Woodruff (1996): Identification of Outliers in Multivariate Data, Journal of the American Statistical Association, Vol. 91, No. 435 , pp. 1047-1061

[5] Kannan, K.S., and Manoj K., (2015); Outlier Detection in Multivariate Data, Applied Mathematical Sciences, Vol. 9, no. 47, 2317 - 2324

[6] Laycock PJ 1975 “Optimal regression: regression models for directions,” Biometrika, 62: 305–311.

[7] MacQueen, J. (1967) Some Methods for Classification and Analysis of Multivariate Observations. Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability, 1, 281-297.

[8] Marascuilo and Serlin, (2000); Efficient Algorithms for Mining Outliers from Large Data Sets,Proc. ACM SIDMOD Int. Conf. on Management of Data, 2000.

[9] Osborne and Overbay, (2004):”The New Jersey Data Reduction Report,” Data Eng. Bull., September, 2004.

[10] Pedhazur, (1997):”Testing for multisource contamination in location / scale families,” Communication in Statistics, Part A: Theory and Methods, 18, 1-34, 1997.

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

[12] Rockwood, Michael R.H, and Susan E. Howlett. “Blood pressure in relation to age and frailty.” Canadian Geriatrics Journal: CGJ 14.1 (2011): 2.

[13] Rousseeuw, P.J. and von Zomeren, B.C. (1990). Unmasking multivariate outliers and leverage points. Journal of the American Statistical Association, 85, 633-639.

[14] Thakkar,P. Vala. J.&prajapati, V (2016): Survey on outlier in data stream. International Journal of Computer Application, Vol. 136.