Journal of Physics: Conference Series

Detection of different outlier scenarios in circular regression model using single-linkage method

To cite this article: N F M Di *et al* 2017 *J. Phys.: Conf. Ser.* **890** 012127

Detection of different outlier scenarios in circular regression model using single-linkage method

N F M Di, S Z Satari and R Zakaria

Faculty of Industrial Sciences & Technology, Universiti Malaysia Pahang, 26300 Gambang, Pahang, Malaysia

E-mail: nurfaraidah@gmail.com

Abstract. Outliers are the set of data that are significantly deviates or dissimilar from the rest of the data set. In circular regression model, the existence of outliers are well known to give a large effect on the parameter estimates and inferences. In this study, we proposed clustering-based method using single linkage to detect multiple outliers. Single-linkage is one of several clustering methods, where the distance between two clusters is determined by a single pair element that are closest to each other. We examined two outlier scenarios with a certain degree of contamination. The performance of proposed method on different outlier scenarios are compared and the best method for each outlier scenario is chosen.

1. Introduction

One of the most challenging tasks in detecting outliers from the circular model is to deal with the high dimensionality of the data. When the data come from a disperse distribution on the circle, a small level of contamination by outliers would be unnoticed and would effect the estimates of location or spread. In this study, we proposed a new clustering algorithm to detect multiple outliers in circular regression model based on Down and Mardia circular-circular regression model. The proposed method is an extension of clustering method proposed by Sebert et al. [1] and Satari [2]. Sebert et al. [1] introduced a new clustering-based approach for multiple outlier identification that utilizes the predicted and residual values obtained from a least square fit of linear data set with different outlier scenarios. Satari [2] extended Sebert et al. [1] method to detect outliers for circular data.

1.1 Down and Mardia circular-circular regression model

To understand the Down and Mardia (DM) model [3], let α and β be an angular location parameters and ω is a slope parameter in the closed interval [-1,1], where *u* and *v* are fixed independent angle and the dependent random angle, respectively. The DM model is given by

$$\tan\frac{1}{2}(v-\beta)=\omega\tan\frac{1}{2}(u-\alpha). \tag{1}$$

The probability density function $f(v)$ and the angular error denoted by e are given in the equations (2) and (3) respectively

$$f(v)=\frac{1}{2\pi I_0(\kappa)}\exp\{\kappa\cos(v-u)\}, \tag{2}$$

$$e = v - \mu(u : \alpha, \beta, \omega), \tag{3}$$

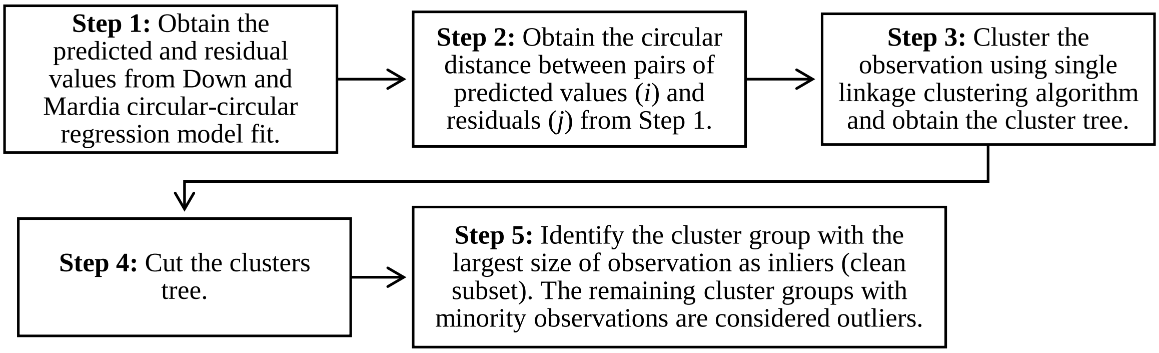
where μ is given by:

$$\mu = \beta + 2\tan^{-1}\{\omega \tan \frac{1}{2}(u - \alpha)\}. \tag{4}$$

The angular error has the von Mises distribution with the mean direction 0 and nonnegative error concentration parameter, κ.

2. Proposed method

The steps of proposed method are outlined in figure 1:



Step 1: Obtain the predicted and residual values from Down and Mardia circular-circular regression model fit.

Step 2: Obtain the circular distance between pairs of predicted values (*i*) and residuals (*j*) from Step 1.

Step 3: Cluster the observation using single linkage clustering algorithm and obtain the cluster tree.

Step 4: Cut the clusters tree.

Step 5: Identify the cluster group with the largest size of observation as inliers (clean subset). The remaining cluster groups with minority observations are considered outliers.

Figure 1. Steps of proposed method.

2.1 Proposed circular distance

The circular distance is defined as distance between observations that take the minimum value of two arc lengths between the points along the circumference [4], i.e., for any angles α and β,

$$d_0(\alpha, \beta) = \min(\alpha - \beta, 2\pi - (\alpha - \beta)) = \pi - |\pi - |\alpha - \beta||. \tag{5}$$

In this study, we proposed a new similarity measure for circular data based on Euclidean distance d_{ij} presented by

$$d_{ij} = \sqrt{\sum_{k=1}^{d}(X_{ik} - X_{jk})^2}, \tag{6}$$

where d_{ij} is the distance between observation *i* and *j*, *d* is the number of variables, X_{ik} is the value of the *k*th variable for the *i*th observation and X_{jk} is the value of the *k*th variable for the *j*th observation where *i* = 1, 2,…, *d* and *j* = 1, 2,…, *d*. Based on equation (5), the matrix of distance between all possible pairs of variables are calculated by using the Euclidean distance, equation (6). Therefore, the proposed circular distance denoted by d_{C-Euc} is given by

$$d_{C-Euc} = \sqrt{\sum_{k=1}^{p}(\pi - |\pi - |\theta_{ik} - \theta_{jk}||)^2} \tag{7}$$

where d_{C-Euc} is the distance between *i* and *j*, *p* is the number of variables, and θ_{ik} is the value of *k*th variable for the *i*th observation where *i* = 1, 2,…, *d* and *j* = 1, 2,…, *d*.

2.2 Single-linkage clustering algorithm

Clustering and outlier detection share a well-known complementary relationship. In clustering, the goal is to partition the data into dense subsets, whereas in outlier detection, the goal is to determine any data which do not seem to fit naturally in these dense subsets [5]. In this study, we use clustering method namely single-linkage method to detect outliers. Single-linkage builds a tree of clusters also known as dendogram by calculating the smallest distance between any single data point in the first cluster and any single data point in the second cluster. This method is widely used due to its sensitivity towards the presence of outliers.

2.3 Stopping rule

The single-linkage algorithm basically will produce clusters tree. To cut the clusters tree, we used stopping rule proposed by Satari [2]. The clusters tree are cut and form groups at a height of $\bar{h} + 2.06 s_h$ where $\bar{h}$ is the average of the cluster tree heights for all $N-1$ clusters, the circular standard deviation of the heights is $s_h = \sqrt{-2 \log \bar{R}_h}$ and $\bar{R}_h$ is the mean resultant length of the heights of the $N-1$ clusters.

3. Simulation study for different outlier scenarios

Simulation studies have been carry out using SPLUS package to access the efficiency of the proposed method in comparison with the existing method by Satari [2]. Satari [2] proposed circular distance based on City-block distance denoted by $d_{C-Satari}$ as presented below

$$d_{C-Satari} = \sum_{k=1}^{p} (\pi - |\pi - |\theta_{ik} - \theta_{jk}||). \tag{8}$$

We randomly simulate the data set to have outliers for two specified scenarios which are outlier scenarios in *v*-space and *u*-space. Outlier scenario is referred to the placement of the outlying observations relative to the inlying observations [1]. In each scenario, the outliers were placed away from inliers at specified distance with six levels of contamination (λ). We generated two sample of sizes $n = 30$ and $n = 120$ for independent circular variable (*u*) and circular error (*e*) from von Mises distribution. The values of *u* are assumed fixed and generated from $VM(\pi/2, 2)$ and the values of *e* are generated from $VM(0, \kappa)$ with concentration parameters of $\kappa = 5$ and $\kappa = 20$ relatively. Then, five contamination levels of λ = 0.0, 0.2, 0.4, 0.6, 0.8 and 1.0 were used to simulate the data. The power performance of the proposed method was examined using “success” probability (*pout*). The success probability is defined by

$$pout = \frac{"success"}{s} \tag{9}$$

where “*success*” is number of data set that the method successfully identified all the outlying observations, and *s* is the total number of simulations. The probability of getting the highest “success” will be achieved as the *pout* value approaching to 1.0. In this study, the simulation process is repeated 1000 times.

4. Results and discussion

The total of 24 conditions are investigated in order to see the performance of proposed methods in detecting multiple outliers using different number of sample sizes (*n*), concentration parameter (κ) and level of contamination (λ), with two different outlier scenarios. Table 1 shows the success probability values for outlier scenario in *v*-space. Note that SL-Euc is proposed method and SL-Satari is method proposed by Satari [2]. It can be seen that the *pout* values are gradually increased as the values of sample size (*n*) and concentration parameter (κ) are increased. The values of *pout* also increasing when the levels of contamination (λ) are increased for both fixed values of sample size (*n*) and concentration parameter (κ). At $\lambda = 1.0$, all the planted outliers are detected in all conditions for both methods.

Table 1. Success probability (*pout*) values for outlier scenario in *v-space*.

λ	n	SL-Euc		SL-Satari	
		κ = 5	κ = 20	κ = 5	κ = 20
0.0	30	0.000	0.083	0.000	0.208
	120	0.000	0.246	0.000	0.299
0.2	30	0.000	0.187	0.000	0.298
	120	0.062	0.340	0.047	0.388
0.4	30	0.092	0.622	0.112	0.592
	120	0.259	0.795	0.197	0.726
0.6	30	0.840	0.978	0.948	0.967
	120	0.992	0.998	0.987	0.978
0.8	30	0.972	0.994	0.993	0.996
	120	1.000	1.000	1.000	1.000
1.0	30	1.000	1.000	1.000	1.000
	120	1.000	1.000	1.000	1.000

The *pout* values for *v-space* outliers are illustrated in figure 2 for $n = 120$ and $\kappa = 20$. At fixed value of $n = 120$, the larger the value of κ, the faster the curve is approaching to one. These results indicated that both methods are abled to detect outliers at lower contamination level when concentration parameter κ is high as presented in figure 2(a). At $\lambda = 0.4$, the *pout* values of $\kappa = 20$ are higher as compared to $\kappa = 5$ for both methods. Alternatively, as illustrated in figure 2(b), at fixed value of $\kappa = 20$, when the value of *n* is high, the *pout* values are approaching to one faster.

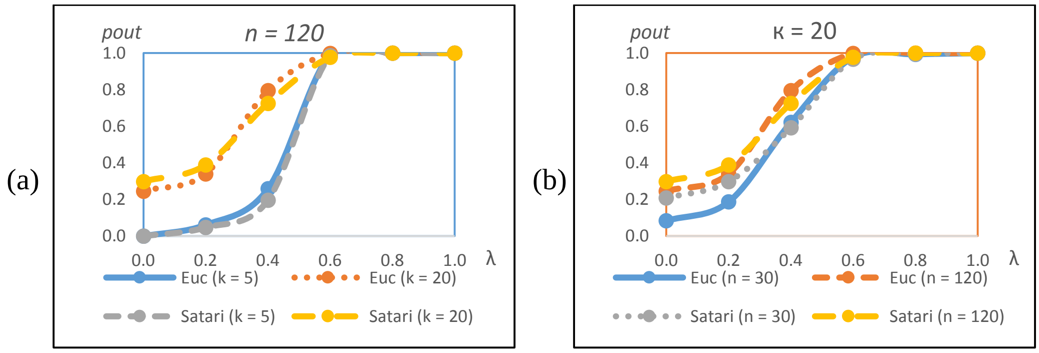


Figure 2. (a) Plot of *pout* versus λ with sample sizes, *n* = 120 for outlier scenario in *v*-space (b) Plot of *pout* versus λ with κ = 20 for outlier scenario in *v-space*.

Table 2. Success probability (*pout*) values for outlier scenario in *u*-space.

λ	n	SL-Euc		SL-Satari	
		κ = 5	κ = 20	κ = 5	κ = 20
0.0	30	0.000	0.098	0.000	0.086
	120	0.000	0.199	0.000	0.192
0.2	30	0.000	0.110	0.000	0.106
	120	0.031	0.213	0.038	0.212
0.4	30	0.094	0.204	0.122	0.222
	120	0.120	0.299	0.156	0.328
0.6	30	0.820	0.929	0.828	0.927
	120	0.865	0.993	0.867	0.992
0.8	30	0.893	0.981	0.919	0.995
	120	0.945	1.000	0.949	1.000
1.0	30	0.987	1.000	0.991	1.000
	120	1.000	1.000	1.000	1.000

Table 2 shows the "success" probability (*pout*) values for outlier scenario in *u*-space. Similar pattern can be seen, in which at higher value of sample size *n* and concentration parameter κ, the faster the *pout* values are approaching to one. At lower value of contamination level $(\lambda = 0.0)$ the *pout* values are equal to zero for $\kappa = 5$ indicated that both methods failed to detect outliers. However, when the contamination level is high, the *pout* values are approaching one, especially when the value of κ is higher. From figure 3 below, it can be seen that at any fixed values of *n* and κ, the *pout* values are increasing as the level of contamination increased. The proposed method appeared to have large *pout* values as compared to Satari [2] when sample size *n* is large.

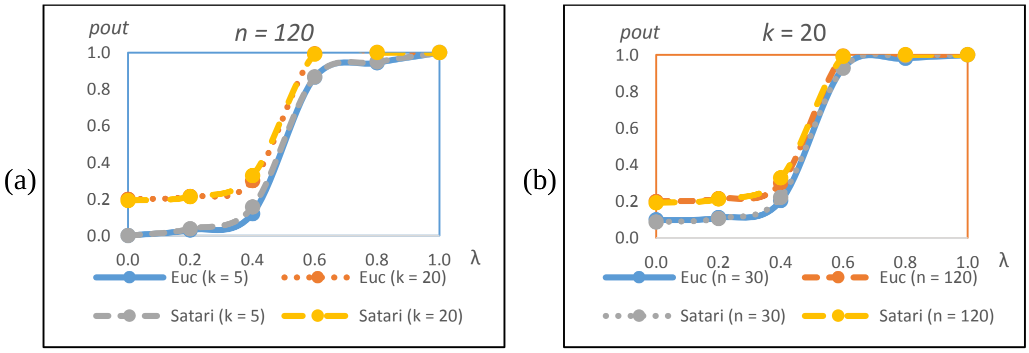


Figure 3. (a) Plot of *pout* versus λ with sample sizes, *n* = 120 for outlier scenario in *u*-space (b) Plot of *pout* versus λ with $\kappa = 20$ for outlier scenario in *u*-space.

5. Conclusion

Generally, the results suggest that outliers are more easily detected if the values of sample size (*n*) and concentration parameter (κ) are large. The results also indicated that as the level of contamination (λ) increased, the *pout* values are also increased as high as 1.0. The proposed method is proven at par with the existing method by Satari [2] especially when the concentration parameters are high. Table 3 below summarizes the best method for each sample size (*n*) and concentration parameter (κ).

Table 3. The best method for each condition.

	Outlier in *v*-space		Outlier in *u*-space	
	n = 30	*n* = 120	*n* = 30	*n* = 120
$\kappa = 5$	SL-Satari	SL-Euc	SL-Satari	SL-Satari /SL-Euc
$\kappa = 20$	SL-Satari	SL-Euc	SL-Satari	SL-Euc

The best method for each condition is chosen based on how fast is the *pout* values approaching one, where at lower number of sample size, existing method by Satari [2] perform better in detecting outlier in *v*-space and *u*-space outlier scenarios. On the other hand, when the sample size is large $(n = 120)$, the proposed methods appeared to have higher *pout* values for both concentration parameter (κ) values. In terms of outlier scenario, the detection of outlier for *v*-space gave higher *pout* values especially when $\lambda = 0.4$.

References

[1] Sebert D M, Montgomery D C and Rollier D A 1998 *Computational Statistics and Data Analysis* **27** p461-484
[2] Satari S Z 2014 Parameter Estimation and Outlier Detection for Some Types of Circular Model *PhD's Thesis* University of Malaya
[3] Down T D and Mardia K V 2002 *Biometrika* **89**(3) p683-697
[4] Jammalamadaka S R and SenGupta A 2001 *Topics in circular statistics* World Scientific Publishing
[5] Aggarwal C C and Yu P S 2001 Outlier detection for high dimensional data *ACM* **30**