CHAPTER SIX

RESIDUAL ANALYSIS FOR OUTLIER DETECTION IN DM CIRCULAR REGRESSION MODELS

6.1 Introduction

A standard approach employed in the study of outlying observations in linear and nonlinear regression models is by performing residual analysis. This approach is applicable for the case of circular regression models. However, a different type of residuals from the linear case is required due to the bounded property of the circular variables. We consider the most recent definition of circular residuals found in the literature which utilize the formula of circular distance between circular residuals. Our focus is to identify outliers which affect the residuals of the DM circular regression model by employing a row deletion method. As in the last chapter, these outliers are usually called influential observations.

6.2 Circular residuals

It is important to study the residuals resulting from any regression modeling in order to check the model adequacy. In the case of linear regression, errors are assumed to be random, independent, identically and normally distributed with mean zero and constant variance. The standard definition of residuals for a linear regression model given by $e_i = y_i - \hat{y}_i$, where y_i and $\hat{y}_i$ are observed and predicted values respectively, cannot be used for circular regression models. For instance, let $y_i = 345°$ and $\hat{y}_i = 5°$. Then $e_i = 345° - 5° = 340°$ which is a total contrast to the actual circular residual, $20°$.

Few definitions of circular residuals can be found in the literature. Mardia (1972) defined the circular residual for the *i*th observation as

$$e_i^* = 1 - \cos(y_i - \hat{y}_i).$$

Here, e_i^* is linear and is bounded within the interval $[0,2]$. Thus, we are not able to use this residual to investigate the assumption of error that follows a specific circular distribution such as the *VM* distribution. Abuzaid (2008) proposed a new definition of circular residual based on circular distance as follows

$$r_{A_i} = \begin{cases} (\pi - |\pi - |y_i - \hat{y}_i||), & \text{if} \quad \hat{y}_i \leqslant y_i,\ y_i - \hat{y}_i \leqslant \pi \quad \text{or} \quad \hat{y}_i > y_i,\ \hat{y}_i - y_i > \pi \\ -(\pi - |\pi - |y_i - \hat{y}_i||), & \text{if} \quad \hat{y}_i \leqslant y_i,\ y_i - \hat{y}_i > \pi \quad \text{or} \quad \hat{y}_i > y_i,\ \hat{y}_i - y_i \leqslant \pi \end{cases}$$

This can be written in a simpler form as follows:

$$r_{A_i} = y_i - \hat{y}_i \,(\mathrm{mod}\, 2\pi).$$

The new residuals r_{A_i} are in the range $[-\pi, \pi]$. These residuals have been shown to be useful in investigating the goodness-of-fit of simple linear regression models (see Abuzaid *et al.* (2008)). Numerical and simulation studies were carried out to show that the circular residuals r_{A_i}, i=1,2,…,n are uncorrelated and follow a von Mises distribution with circular mean 0 and concentration parameter κ. In the next section, we look at a statistic developed based on the definition of circular distance that can be used to detect influential observations in DM circular regression models.

6.3 Mean Circular Error

Rao (1969) defined the circular distance between two circular observations θ_i and θ_j as $d_{ij} = 1 - \cos(\theta_i - \theta_j)$, $d_{ij} \in [0,2]$. We will use this statistic for detecting influential observations in the DM circular regression model by using the row deletion

approach. In this study, the DM circular regression model assumes the errors to follow a *VM* distribution with circular mean $0°$ and concentration parameter κ. Abuzaid (2010) defined a statistic known as mean circular error (*MCEs*) given by

$$MCEs = \frac{1}{n}\sum_{i=1}^{n}\sin\left(\frac{d_i}{2}\right) \tag{6.1}$$

where $d_i = \pi - \left|\pi - \left|y_i - \hat{y}\right|\right|$ is the circular distance between y_i and $\hat{y}_i$, n is the sample size and $MCEs \in [0,1]$.

We intend to use a row deletion approach to see the effect on the values of *MCEs* by removing an observation from the data set. The effect can be measured by looking at the maximum absolute difference between the value of the statistics for full and reduced data sets, denoted by *DMCEs*, such that

$$DMCEs = \max_i \left\{\left|MCEs - MCEs_{(-i)}\right|\right\} \tag{6.2}$$

where *MCEs* is the value for the full data set and $MCEs_{(-i)}$ is the value of *MCEs* when the *i*th observation is removed from the data. *MCEs* statistics are considered as a sort of arithmetic means which is not robust to the existence of influential observation. Thus, *DMCEs* can be used to detect possible influential observations in DM circular regression models. Any observation will be identified as an influential observation if its $MCEs_{(-i)}$ value gives rise to high value of *DMCEs* and the *DMCEs* exceeds a pre-specified cut-off point.

6.4 Sampling Behavior of the *DMCEs* Statistic

We perform a simulation study to investigate the sampling behavior of the *DMCEs* statistic. A set of circular random errors are generated from a von Mises

distribution with mean direction $\mu = 0$ and various values of concentration parameter κ = 5, 10, 30 and 50. We also generate the values of the independent circular random u from $VM(\pi/2,3)$ of size *n*. Observed value of the response variable v are then calculated based on the DM circular regression model with fixed values of α =1.5, β =1.5, and $\omega = 0.5$. Upon fitting the simulated data, we obtain the fitted values $\hat{v}$ of the DM circular regression model. Then we compute the values of the *MCEs* statistic for the full data set and the $MCEs_{(-i)}$ statistics for reduced data set, i=1,2,…,n . Hence, we may then find the value of the *DMCEs* statistic.

The process is carried out 2000 times for each combination of sample size and concentration parameter. We then calculate the 1%, 5% and 10% upper percentiles of the *DMCEs* statistic and the results are tabulated in Table 6.1. The results will be used as the cut-off point in the hypothesis testing to determine whether an observation is an influential observation or not.

In general, for all *n* and percentile levels, the value of the cut-off point decreases as the concentration parameter κ increases. Similarly, as the sample size increases, the cut-off points decrease for all percentile levels and concentration parameter κ.

Note that the cut-off points described above are only for the case when $\omega = 0.5$. Further investigation shows that the cut-off points do not depend on the parameter values α and β, but depend on ω. When ω gets closer to 1, the cut-off points get larger. Unlike the $\left|COVRATIO_{(-i)} - 1\right|$ statistic, we are able to obtain the cut-off points of *DMCEs* statistic for small ω. The cut-off points get smaller as ω gets closer to 0. Partial results are given in Appendix 8.

Table 6.1: Cut-off points of the *DMCEs* statistic

n	Level of percentiles	κ				
		5	10	20	30	50
10	10%	0.0855	0.0697	0.0589	0.0508	0.0401
	5%	0.0940	0.0818	0.0716	0.0630	0.0538
	1%	0.1000	0.0985	0.0964	0.0899	0.0791
20	10%	0.0400	0.0298	0.0170	0.0126	0.0096
	5%	0.0457	0.0376	0.0283	0.0151	0.0106
	1%	0.0500	0.0479	0.0428	0.0347	0.0301
30	10%	0.0245	0.0162	0.0109	0.0087	0.0066
	5%	0.0281	0.0195	0.0118	0.0095	0.0070
	1%	0.0330	0.0295	0.0212	0.0169	0.0081
40	10%	0.0178	0.0118	0.0082	0.0066	0.0051
	5%	0.0200	0.0130	0.0089	0.0071	0.0055
	1%	0.0247	0.0206	0.0104	0.0086	0.0065
50	10%	0.0142	0.0098	0.0068	0.0056	0.0043
	5%	0.0154	0.0105	0.0073	0.0060	0.0045
	1%	0.0193	0.0131	0.0084	0.0068	0.0052
60	10%	0.0119	0.0082	0.0058	0.0047	0.0036
	5%	0.0131	0.0088	0.0061	0.0050	0.0039
	1%	0.0156	0.0103	0.0068	0.0056	0.0044
70	10%	0.0102	0.0072	0.0050	0.0041	0.0032
	5%	0.0113	0.0076	0.0054	0.0043	0.0034
	1%	0.0136	0.0089	0.0060	0.0047	0.0039
80	10%	0.0092	0.0065	0.0045	0.0036	0.0028
	5%	0.0097	0.0068	0.0047	0.0039	0.0030
	1%	0.0112	0.0078	0.0052	0.0043	0.0034
90	10%	0.0082	0.0057	0.0040	0.0033	0.0025
	5%	0.0087	0.0060	0.0043	0.0035	0.0027
	1%	0.0101	0.0070	0.0050	0.0039	0.0030
100	10%	0.0074	0.0051	0.0036	0.0029	0.0023
	5%	0.0079	0.0055	0.0038	0.0031	0.0024
	1%	0.0090	0.0059	0.0043	0.0036	0.0027
120	10%	0.0062	0.0044	0.0031	0.0025	0.0019
	5%	0.0066	0.0046	0.0033	0.0027	0.0021
	1%	0.0076	0.0052	0.0037	0.0030	0.0022
150	10%	0.0051	0.0036	0.0025	0.0021	0.0016
	5%	0.0054	0.0038	0.0027	0.0022	0.0017
	1%	0.0062	0.0042	0.0029	0.0024	0.0019

6.5 Power of Performance of *DMCEs* Statistic

We will now investigate the power of performance of the *DMCEs* statistic via a simulation study. We follow the same scheme used in Section 6.3 to generate the simulated data. In addition, at point [d] of the response variable v, the observation $v[d]$ is contaminated as follows

$$v^*[d] = v[d] + \lambda\pi \mod(2\pi)$$

where $v^*[d]$ is the contaminated observation at position [d] and λ is the degree of contamination in the range of $0 \le \lambda \le 1$. When $\lambda = 0$, there is no contamination at position [d], whereas when $\lambda = 1$, the observation $y^*[d]$ is located at the anti mode of its initial location.

The generated data are fitted using Eq. (4.6) and consequently we obtain the fitted values $\hat{v}$. Then, we calculate the value of *DMCEs* for each simulated data set. The power of performances of *DMCEs* statistic is investigated by computing the percentage of correctly detecting the outlier at position [d]. We provide the result for $\alpha = 1.5$, β =1.5, and $\omega = 0.5$.

Figure 6.1 give the plot of the power of performance of the *DMCEs* statistic for κ = 10 and various sample sizes. We observe that the power of performance is an increasing function of sample size *n*. The *DMCEs* statistic performs better for larger sample size. On the other hand, Figure 6.2 shows the performance of *DMCEs* for *n* = 70 and various values of κ. When larger values are used, the performance is almost similar, but clearly better than that for small κ. Similar results are observed for other cases.

6.6 Practical Example

We consider the Circadian data described in Chapter 4. Figure 6.3 gives the spoke plot of the data. By taking the horizontal axis in the right direction as $0°$, the inner ring places the observations of S1 and the outer ring for S2. The lines connecting points on outer and inner rings means the blood pressure measurements correspond to the same student.

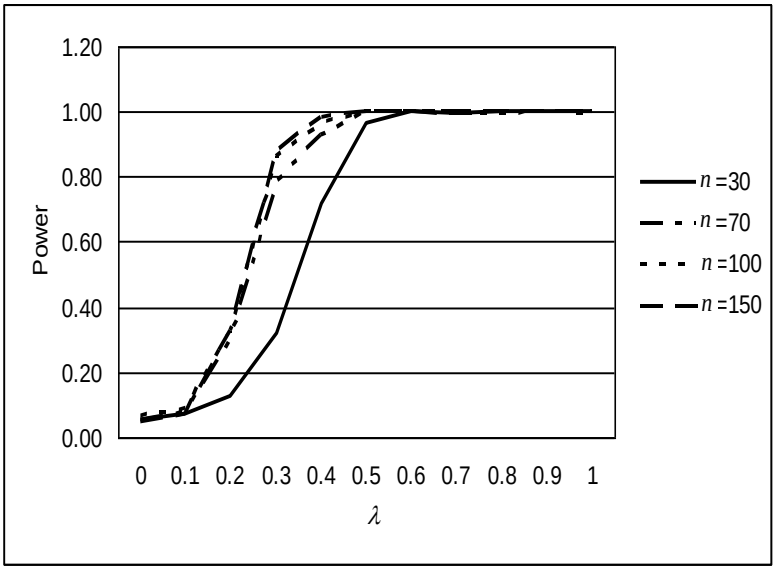

Figure 6.1: Power of performance of *DMCEs* statistics, for κ =10

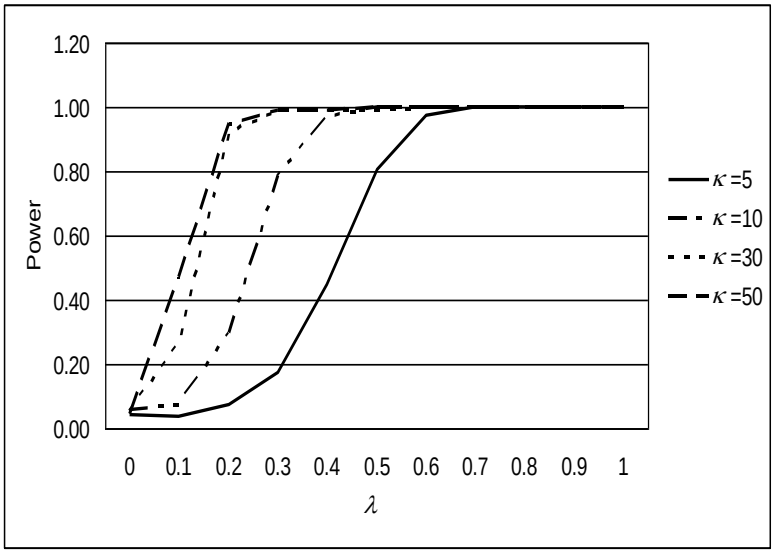

Figure 6.2: Power of performance of *DMCEs* statistics, for *n*=70

It can be seen that the line corresponding to student number 8 on the left hand side of the plot lies a distance away from the others. The student is flag as a candidate of outlier. By employing the *DMCEs* statistic which use the row deletion approach, such outlier is also known as an influential observation. The data is of size n = 10 with the concentration parameter $\kappa = 17.64$. Thus, from Table 6.1, the cut-off point to be used is 0.07. Upon calculating the *DMCEs* for the data, we have *DMCEs* = 0.09 which is greater than the cut-off point and conclude that student number 8 is an influential observation.

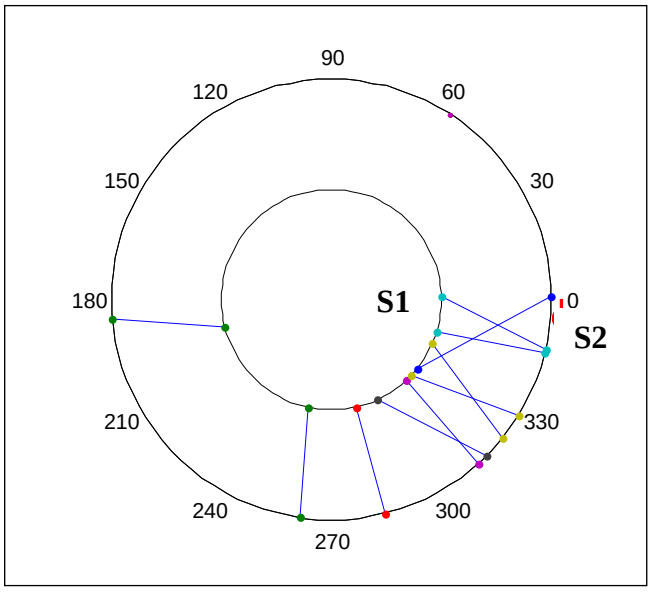

Figure 6.3: Spoke plot of Circadian data

Table 6.2: Effect of influential observation on parameter estimates

Data	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\omega}$
With the 8th observation	16.57°	5.74°	0.67
Without the 8th observation	51.02°	39.98°	0.82

Further, we investigate the effect of the influential observation on the parameter estimates as tabulated in Table 6.2. It can be seen that, when removing student number

8 from the data, $\hat{\alpha}$ and $\hat{\beta}$ increase by a large value in degree while $\hat{\omega}$ also changes from 0.669 to 0.820. Therefore, it is important to investigate student number 8 further which might give useful information to the investigators.

6.7 Summary

In this chapter, we have considered the problem of detecting influential observations in the DM circular regression models via a row deletion method. We use the *DMCEs* statistic for the purpose. The cut-off points are obtained via a simulation study. The *DMCEs* statistic is shown to be able to detect influential observation better for larger sample size and larger concentration parameter. When applied to the Circadian data, the *DMCEs* statistics is able to detect the observation number 8 as influential observation.