CHAPTER FIVE

OUTLIERS IN DM CIRCULAR REGRESSION MODELS

5.1 Introduction

Outliers may give a significant impact on estimating the parameters of DM circular regression models especially with the classical estimation methods such as least squares (LS) method and maximum likelihood estimation (MLE) method. In this chapter, we investigate the robustness of MLE method for DM circular regression models in the presence of outliers. Outliers which affect the parameter estimates of interest are usually called influential observations. In this chapter, we use a row deletion approach in identifying such influential observations by investigating its effect on the covariance matrix of the parameters of the DM circular regression models.

5.2 Robustness of Maximum Likelihood Estimation Method

In this section, we intend to study the effect of outliers on the MLE estimate of the DM models as given in Eq. 4.5. Let u be a circular independent variable with values generated from $VM(\mu = 1, \kappa = 3, n = 50)$ and e be a circular random error with values generated from $VM(\mu = 0, \kappa = 20, n = 50)$. To test the robustness of MLE in this model, we introduce an outlier in the original dependent variable at the i th observation, $v[i]$, giving $v^*[i]$ such that

$$v^*[i] = v[i] + \lambda\pi\ , \qquad 0 \le \lambda \le 1, \quad i = 1,2,\ldots,50.$$

We calculate the estimated parameters with and without outliers using the maximum likelihood estimation method. We carry out the simulation 3000 times. The difference

of the estimated parameters and the true values for α and β are measured by the circular distance as given by Jammalamadaka and SenGupta (2001). That is, the circular distance between the estimated circular parameter $\hat{\theta}_{MLE} = (\hat{\alpha}_{MLE}, \hat{\beta}_{MLE})$ and $\theta_{true} = (\alpha_{true}, \beta_{true})$ is defined as

$$d(\theta) = \pi - \left| \pi - \left| \hat{\theta}_{MLE} - \theta_{true} \right| \right|$$

which is also known as circular bias. Further, the difference between the estimate parameter and the true value for ω is measured by linear bias, $Bias(\omega) = \hat{\omega}_{MLE} - \omega_{true}$.

The estimation of the parameter that exclude the outlier is when the value of contaminated point $\lambda = 0$. The results for three cases are tabulated in Tables 5.1-5.3. In all cases, as the value of contamination point λ increases, the bias for the three parameters get larger except when λ get closer to 1. This is because the contaminated point will be further away from one tail of the data but closer to the other tail. We conclude that the presence of outliers in circular data does affect the estimation of the parameters of DM circular regression models.

Table 5.1: Estimate of parameters and Biasness (True value α =0.5, β =0.5, ω =0.2)

λ	$\hat{\alpha}$	Circular bias, $d(\alpha)$	$\hat{\beta}$	Circular bias, $d(\beta)$	$\hat{\omega}$	Bias, $Bias(\omega)$
0	0.6550	0.1550	0.5821	0.0821	0.1737	0.0263
0.1	0.6272	0.1272	0.6621	0.1621	0.1790	0.0210
0.2	0.6854	0.1854	0.7409	0.2409	0.1718	0.0282
0.3	0.7416	0.2416	0.8353	0.3353	0.1815	0.0185
0.4	0.7951	0.2951	0.9297	0.4297	0.1699	0.0301
0.5	0.9215	0.4215	1.0130	0.5130	0.1800	0.0200
0.6	0.9711	0.4711	1.0204	0.5204	0.1705	0.0295
0.7	0.9900	0.4900	1.0106	0.5106	0.1499	0.0501
0.8	1.0302	0.5302	0.9266	0.4266	0.1489	0.0511
0.9	0.9844	0.4844	0.8312	0.3312	0.1627	0.0373
1	0.9825	0.4825	0.8004	0.3004	0.1688	0.0312

Table 5.2: Estimate of parameters and Biasness (True value α =2.5, β =2.5, ω =0.5)

λ	$\hat{\alpha}$	Circular bias, $d(\alpha)$	$\hat{\beta}$	Circular bias, $d(\beta)$	$\hat{\omega}$	Bias, $Bias(\omega)$
0	2.0246	0.4754	2.0495	0.4505	0.4884	0.3116
0.1	1.9311	0.5689	2.0768	0.4232	0.4837	0.3163
0.2	1.8957	0.6043	2.0432	0.4568	0.4504	0.3496
0.3	1.8427	0.6573	1.9987	0.5013	0.3942	0.4058
0.4	1.7074	0.7926	1.9220	0.5780	0.3593	0.4407
0.5	1.6241	0.8759	1.9120	0.5880	0.3474	0.4526
0.6	1.6970	0.8030	1.8915	0.6085	0.3314	0.4686
0.7	1.7541	0.7459	1.8121	0.6879	0.3065	0.4935
0.8	1.7987	0.7013	1.8823	0.6177	0.3574	0.4426
0.9	1.5669	0.9331	1.9526	0.5474	0.3869	0.4131
1	1.6852	0.8148	1.9396	0.5604	0.3856	0.4144

Table 5.3: Estimate of parameters and Biasness (True value α =1.5, β =1.5, ω =0.4)

λ	$\hat{\alpha}$	Circular bias, $d(\alpha)$	$\hat{\beta}$	Circular bias, $d(\beta)$	$\hat{\omega}$	Bias, $Bias(\omega)$
0	1.5737	0.0029	1.4452	0.1255	0.4534	0.0466
0.1	1.5972	0.0264	1.4675	0.1033	0.4501	0.0499
0.2	1.6097	0.0389	1.4535	0.1172	0.4438	0.0561
0.3	1.6157	0.0449	1.4644	0.1063	0.4503	0.0496
0.4	1.6127	0.0419	1.4551	0.1156	0.4445	0.0554
0.5	1.6141	0.0433	1.4493	0.1214	0.4476	0.0523
0.6	1.6309	0.0601	1.4595	0.1112	0.4435	0.0564
0.7	1.6317	0.0609	1.4527	0.1180	0.4375	0.0624
0.8	1.6400	0.0692	1.4896	0.0811	0.4380	0.0619
0.9	1.6238	0.0530	1.4566	0.1142	0.4447	0.0553
1	1.6093	0.0385	1.4474	0.1233	0.4514	0.0486

5.3 Graphical Techniques

Graphical techniques not only can be used for describing the circular data but also for diagnostic checking purposes in circular regression models. As an example, Figures 5.1-5.3 gives the circular histogram, index plot and Q-Q plot of the circular residuals obtained from a DM circular regression model given a simulated data set. Clearly, from these plots, most of the circular residuals are concentrated around 0°, except two points which are further away from the rest. Disregarding these two points from the Q-Q plot of a von Mises distribution, we can say that the residuals follow a *VM* distribution, which is an assumption made for the DM regression models. In addition, Abuzaid *et al.* (2010) proposed a circular boxplot which is an analogue of linear boxplot. The plot can be used to see the way the circular data are distributed around a unit circle and, more importantly, to identify possible outliers that might occur in the data. From Figure 5.4, we can see that most points lie around zero but there are two points which lie outside the lower/upper inner fences of the circular boxplot. Such points are identified as outliers. Meanwhile, Fakhrul *et al.* (2008) proposed a new plot called "Spoke plot" which is very useful to demonstrate the relationship between two circular variables. The plot constitute 2 circles referred as the inner and outer circles as shown in Figure 5.5. The values of an independent variable u are placed in the inner circle while the values of a dependent variable v in the outer circle. The blue lines connect the corresponding values of u and v. If lesser number of the blue lines are observed crossing the interior of the inner circle, then we expect there exist strong relationship between the two variables. At the same time, the plot can be used to recognice outliers in the data. It can be seen in the plot that a pair of (u, v) is not close to the other pairs and becomes a candidate to be outlier. The plot will also suggest that the relationships between the variables are quite strong with value closer to 1.

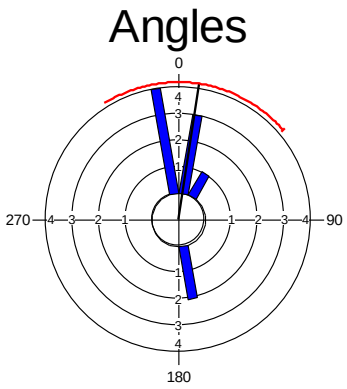

Figure 5.1: Scatter plot

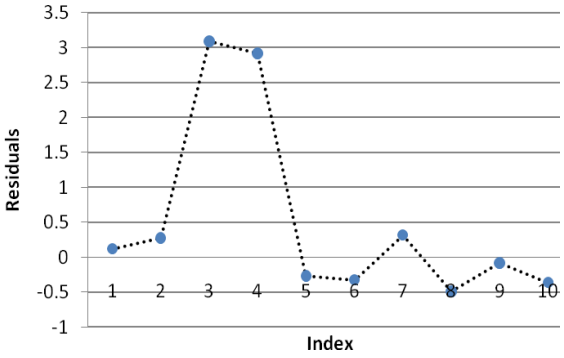

Figure 5.2: Circular residuals versus index

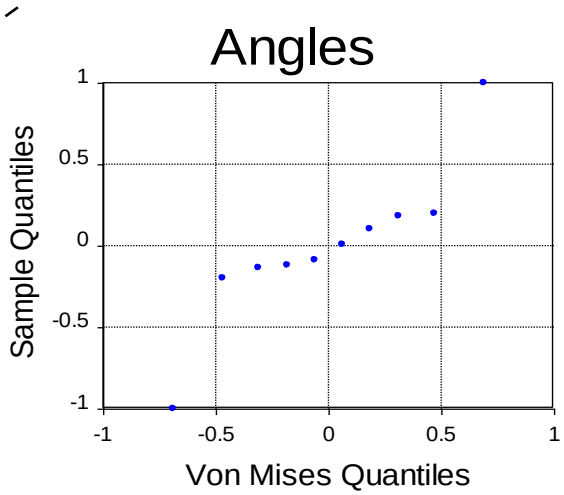

Figure 5.3: Q-Q plot for circular residuals

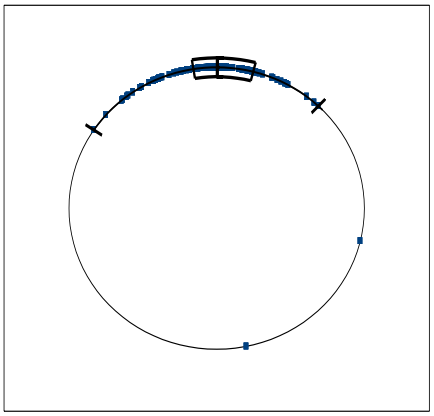

Figure 5.4: Circular boxplot

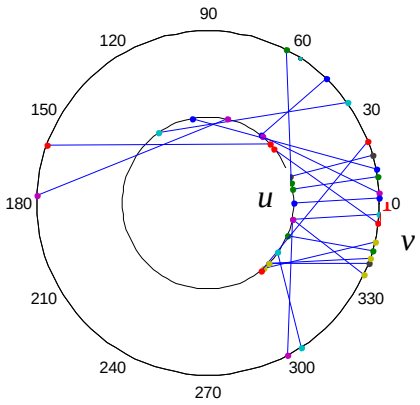

Figure 5.5: A sample of Spoke plot for variables *u* and *v*

5.4 *COVRATIO* Statistic

The row deletion approach has been widely used for linear regression models in identifying outliers which affect the parameter of interest such as parameter estimates, variance of residuals and also covariance matrix. These outliers are usually called influential observations. Here, we consider a statistic that has been used to identify influential observation in linear regression models, the *COVRATIO* statistic. The statistic employs the row deletion approach by looking at the effect of influential observations on the covariance matrix of the linear regression model. Belsley *et al.* (1980), in his paper, suggested the *COVRATIO* statistic based on the determinantial ratio of the covariance matrix given by

$$COVRATIO(-i) = \frac{\left|COV(-i)\right|}{\left|COV\right|}$$

where $\left|COV\right|$ is the determinant of the covariance matrix as given in Section 4.4 for the full data set and $\left|COV(-i)\right|$ is for the reduced data set by excluding the *i*th row. If the ratio is close to unity, then there is no significant difference between the covariance matrices, that is, the *i*th observation is consistent with the other observations. It has been shown further that, if the value of $\left|COVRATIO(-i)-1\right|$ is larger than (3*p*/*n*), then the *i*th observation is a candidate for influential observation, where *p* is the number of estimated coefficients and *n* is the sample size. In this study, we extend the use of this statistic to identify the presence of influential observation in the DM circular regression model. We will find the cut-off points of the statistic and investigate its performance via simulation in the following sections. The procedure will detect only a single influential observation at a time. The procedure can be repeated until no influential observation is identified.

5.5 Sampling Behaviour of the *COVRATIO* Statistic

We generate various sets of circular random error from the von Mises distribution with mean direction μ=0 and the concentration parameter κ=5, 10, 30 and 50. Then, we generate the values of the independent circular variable u from $VM(\pi/2,3)$ for a given sample size n. Using the above information, the observed values of the response variable v are then calculated using the DM circular regression model as given by Eq. (4.6) with fixed values of α =1.5, β =1.5, and $\omega = 0.5$.

Upon fitting the DM regression model on the simulated data, we obtain the fitted values $\hat{v}$. Then, we compute the value of $COVRATIO_{(-i)}$ for all $i = 1,2,\ldots,n$ and consequently obtain the maximum value of $\left|COVRATIO_{(-i)} - 1\right|$. The process is carried out 500 times for each combination of sample size and concentration parameter. We then calculated the 1%, 5% and 10% upper percentiles of the maximum values of $\left|COVRATIO_{(-i)} - 1\right|$ which will be considered as the cut-points of the $\left|COVRATIO_{(-i)} - 1\right|$ statistic.

The cut-off points are tabulated in Table 5.4. In general, for all κ and percentile levels considered, the cut-off points get smaller as n gets larger. On the other hand, there is no consistent pattern of the cut-off points observed as κ increases, though they achieve their maximum when κ = 3 to 5.

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

Table 5.4: Cut-off point of *COVRATIO* statistic

n	Level of percentile	κ							
		2	3	4	5	7	10	20	30
30	1%	2.59	1.51	1.66	1.01	1.09	1.06	1.03	1.01
	5%	1.11	0.79	0.89	0.74	0.76	0.72	0.79	0.67
	10%	0.83	0.71	0.73	0.67	0.66	0.66	0.69	0.62
40	1%	0.75	0.69	0.72	0.71	0.67	0.66	0.62	0.72
	5%	0.53	0.58	0.60	0.56	0.55	0.54	0.50	0.55
	10%	0.47	0.52	0.51	0.48	0.50	0.49	0.45	0.48
50	1%	0.57	0.53	0.61	0.64	0.54	0.56	0.64	0.57
	5%	0.41	0.46	0.50	0.50	0.45	0.48	0.48	0.46
	10%	0.38	0.43	0.44	0.44	0.39	0.42	0.44	0.42
60	1%	0.48	0.48	0.47	0.48	0.51	0.49	0.52	0.50
	5%	0.43	0.42	0.42	0.43	0.41	0.42	0.43	0.42
	10%	0.36	0.37	0.37	0.37	0.35	0.37	0.39	0.35
70	1%	0.31	0.41	0.46	0.48	0.47	0.46	0.43	0.46
	5%	0.29	0.36	0.37	0.37	0.37	0.36	0.36	0.37
	10%	0.28	0.34	0.34	0.33	0.33	0.33	0.32	0.33
80	1%	0.28	0.36	0.42	0.45	0.45	0.44	0.41	0.43
	5%	0.26	0.31	0.36	0.37	0.38	0.36	0.34	0.34
	10%	0.25	0.29	0.32	0.32	0.31	0.31	0.32	0.31
90	1%	0.26	0.32	0.37	0.39	0.38	0.36	0.38	0.42
	5%	0.23	0.28	0.31	0.34	0.29	0.32	0.31	0.29
	10%	0.22	0.26	0.28	0.29	0.25	0.26	0.29	0.26
100	1%	0.22	0.30	0.34	0.35	0.32	0.33	0.34	0.41
	5%	0.21	0.27	0.28	0.29	0.28	0.29	0.28	0.27
	10%	0.20	0.25	0.26	0.26	0.24	0.25	0.25	0.23
120	1%	0.19	0.24	0.31	0.32	0.34	0.28	0.29	0.26
	5%	0.18	0.23	0.26	0.25	0.24	0.23	0.24	0.23
	10%	0.17	0.22	0.23	0.23	0.21	0.21	0.22	0.21
140	1%	0.15	0.22	0.25	0.27	0.25	0.26	0.24	0.28
	5%	0.14	0.20	0.24	0.21	0.21	0.21	0.20	0.21
	10%	0.14	0.19	0.22	0.20	0.19	0.18	0.18	0.19
150	1%	0.15	0.21	0.24	0.28	0.26	0.24	0.28	0.24
	5%	0.14	0.19	0.21	0.22	0.21	0.21	0.22	0.20
	10%	0.14	0.17	0.19	0.18	0.19	0.19	0.18	0.18

smaller. Partial results are given in Appendix 7. However, when ω gets closer to 0, the $\left|COVRATIO_{(-i)}-1\right|$ statistic fails to give reasonable set of cut-off points.

5.6 Power of Performance of *COVRATIO* Statistic

To investigate the power of performance of $\left|COVRATIO_{(-i)}-1\right|$ statistic, several sample sizes are considered. We generate the data using similar steps employed in the previous section. 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 $0 \le \lambda \le 1$. The generated data are fitted using Eq. (4.1) and we obtain the fitted values $\hat{v}$. Then, we calculate the maximum value of $\left|COVRATIO_{(-i)}-1\right|$ statistic for each simulated data set. The power of performance is examined by computing the percentage of correct detection of the contaminated observation at position $v[d]$. We provide the result for $\alpha = 1.5$, β =1.5, and $\omega = 0.5$.

Figure 5.6 gives the plot of power of performance of $\left|COVRATIO_{(-i)}-1\right|$ statistic for n = 70 and various value of κ. It can be seen that the power is an increasing function of the concentration parameter κ. For large values of concentration parameter κ = 20 and 30, the statistic is able to detect an outlier at lower contamination levels. Whereas, for the smallest value of concentration parameter κ = 5, the statistic is only able to detect an outlier which is really far from the rest of the observation.

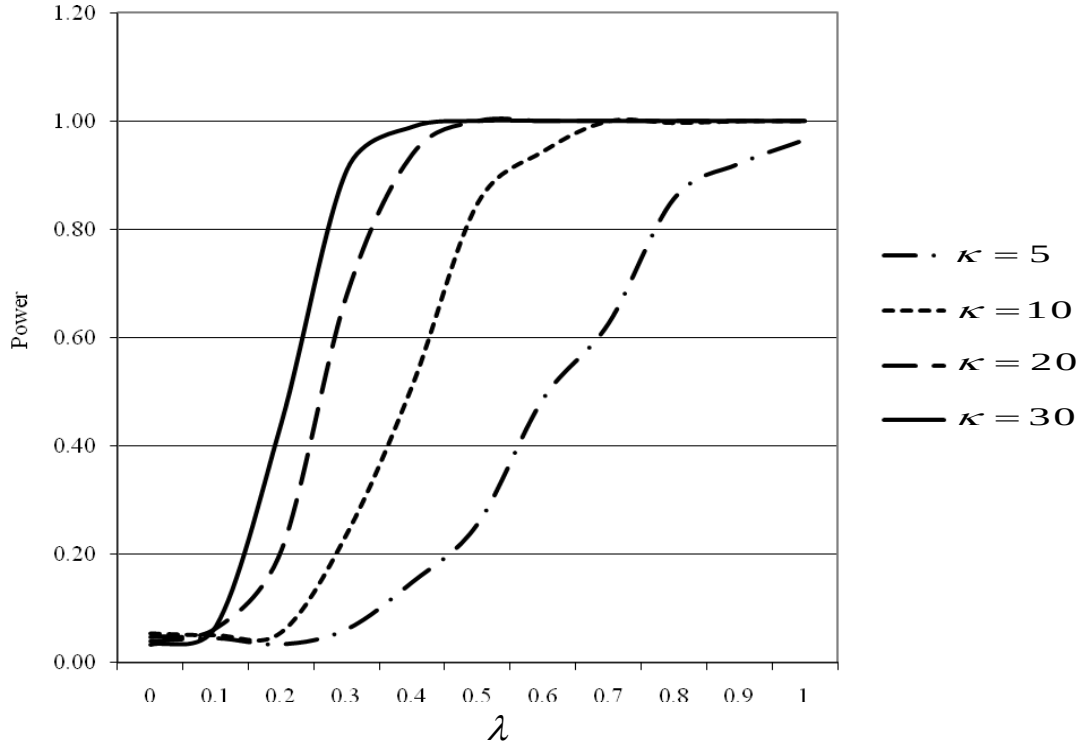

Figure 5.6: Power of performance of $\left|COVRATIO_{(-i)} - 1\right|$ statistic, for *n*=70

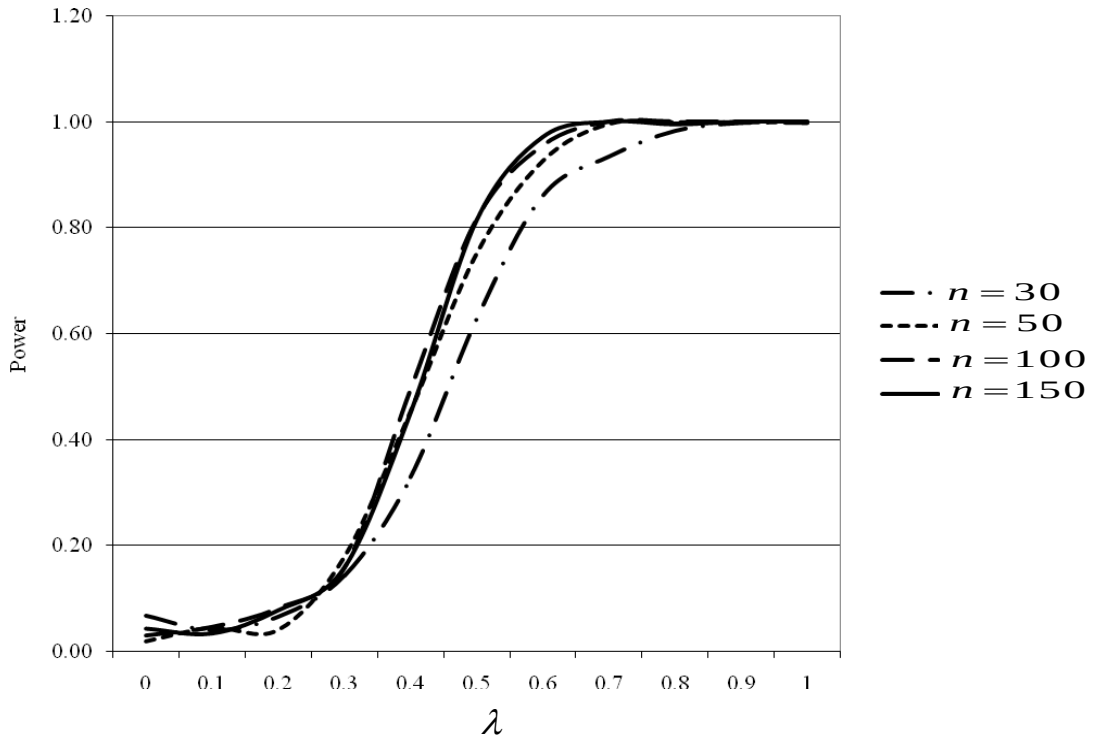

Figure 5.7: Power of performance of $\left|COVRATIO_{(-i)} - 1\right|$ statistic, for κ=10

On the other hand, Figure 5.7 gives the plot of power of performance of $\left|COVRATIO_{(-i)}-1\right|$ for κ=10 and various value of *n*. It can be seen that the power curves are very close to each other for *n* = 50, 100 and 150 while the power curve is lower for *n* = 30. Similar results are observed for the other cases.

5.7 Practical Example

As an illustration, we use the data set as described in Section 4.5.1. Figure 5.8 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 anchored wave buoy AB while the outer ring for high frequency radar HF. The lines connecting points on outer and inner rings correspond to the observed values of AB and HF respectively for the same individual/item. There are only two lines crossing the inner ring. Further, by using the $\left|COVRATIO_{(-i)}-1\right|$ statistics we identify that observations number 38 and 111 are candidates for influential observations. This can be further verified by looking at Figure 5.9 whereby there are two observations with high value of $\left|COVRATIO_{(-i)}-1\right|$ denoted by *p*.

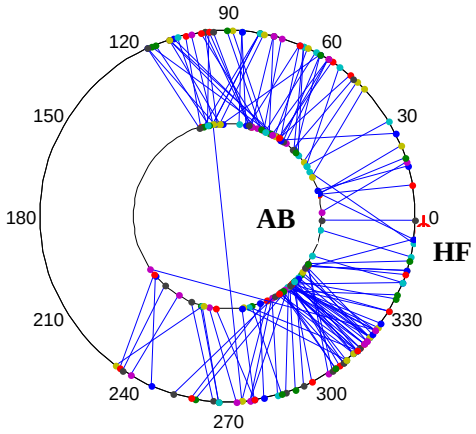

Figure 5.8: Spoke plot of wind data

Here, we have $n = 129$ and $\kappa = 6.84$. By considering the value of cut-off point corresponding to $n = 100$ and $\kappa = 7$, Table 5.5 gives the test value *p* and the decision for each observation.

Table 5.5: Result based on *COVRATIO* statistic

Iteration	Observation	Test value	Cut-off point	Decision
1	38	0.95	0.28	Outlier
2	111	0.68	0.28	Outlier

Based on the results, in the first iteration, we identify observations $n = 38$ as influential observation because the test values exceed the cut-off point of the statistic which is 0.28. In the second iteration, we identify $n = 111$ as influential observation. The plots *p* versus index are given in Figures 5.9-5.10 repectively. Further, we investigate the effect of these two observations on the parameter estimates. After removing observations 38 from the data set, we noticed that $\hat{\alpha}$ and $\hat{\beta}$ decrease by a large value which is $\hat{\alpha} = 33.81°$, $\hat{\beta} = 38.42°$ and $\hat{\omega} = 0.94$ as shown in Table 5.6. However, not much change is observed when observation 111 is removed. Hence, it is important to investigate the observations identified as influential observations in both measurements of ocean wind direction, and the information might be useful for further investigation.

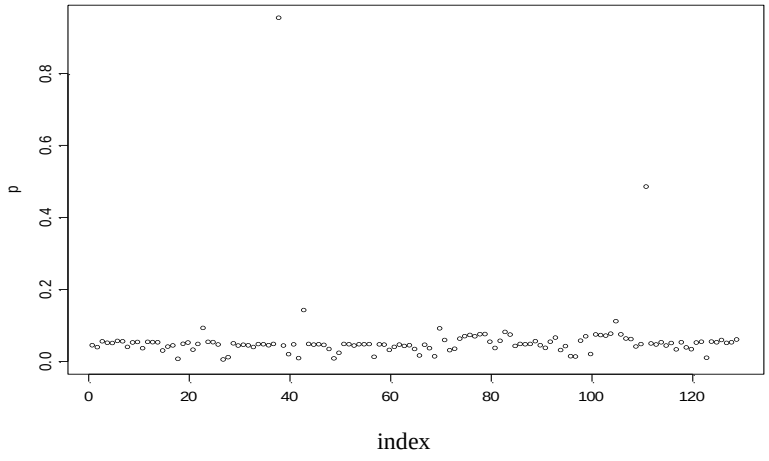

Figure 5.9: Plot of *p* versus index for 1^{st} iteration

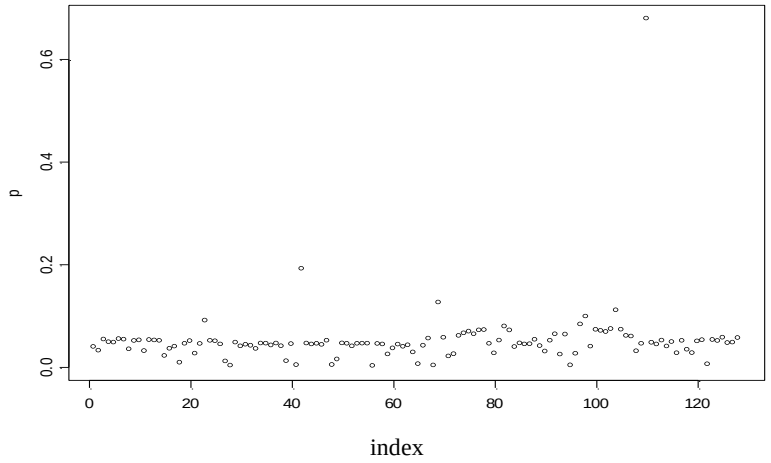

Figure 5.10: Plot of *p* versus index for 2^{nd} iteration

Table 5.6: Effect of influential observation on parameter estimates

Data	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\omega}$
Full data set	65.39°	71.82°	0.91
Without the 111^{th} observation	65.68°	71.84°	0.91
Without the 38^{th} observation	33.81°	38.42°	0.94
Without both observations	31.20°	35.57°	0.94

5.8 Summary

In this study, we have considered the problem of detecting influential observations in the DM circular regression model. We then extend the use of the $\left|COVRATIO_{(-i)}-1\right|$ statistic to the model of interest. The cut-off points and the performance of the procedure are obtained via simulation. Finally, as an illustration, the $\left|COVRATIO_{(-i)}-1\right|$ statistic for circular bivariate data has successfully detected observations number 8 and 111 as outliers in the ocean wind direction data set.