**COVRATIO Statistic for Simple Circular Regression Model**

Ali Abuzaid [a], Ibrahim Mohamed [a], Abdul G. Hussin*[b] and Adzhar Ramli [a]

[a] Institute of Mathematical Sciences, University of Malaya, 50603 Kuala Lumpur, Malaysia.
[b] Centre for Foundation Studies in Science, University of Malaya, 50603 Kuala Lumpur, Malaysia.

*Author for correspondence; e-mail: ghapor@um.edu.my

Received: 4 January 2011
Accepted: 1 June 2011

**Abstract**

Few studies have considered the modeling of a linear relationship between two circular variables or circular regression model. However, the problem of outlier detection in these models has not received enough consideration. This paper extends the COVRATIO statistic which is originally used to identify outliers in linear regression model. It is our aim to further exploit this approach of detecting outliers in circular regression model. The cut-off points for the statistic are obtained using simulation. It is found that the cut-off points are dependent to the sample size. We also show that the statistic has higher power of performance in detecting outliers for large sample size and large concentration parameter. A practical example is considered for illustration purposes.

**Keywords:** circular regression, COVRATIO, outlier, wind direction data.

**1. INTRODUCTION**

The difference in topological space between the linear and circular variables require the development of special models to study the relationship between circular variables. Discussion on circular regression when either response or covariate variable is circular has been dated back to Gould [1]. Several circular regression models have been proposed by Down and Mardia [2], Hussin et al. [3], Jammalamadaka and Sarma [4], Kato et al. [5], Laycock [6], Lund [7], and Mardia [8]. The regression line for the case when the dependent variable is linear and the independent variable is circular may be represented by using a sine or cosine curve and have been proposed as an example, by Khanabsakdi and Prapunrat [9]. In this paper, we consider the simple circular regression model proposed by Hussin et al. [3], where a linear relationship between the circular variables is assumed.

The analysis of circular regression may be subjected to the occurrence of outliers. Barnet and Lewis [10] and Belsley et al. [11] have discussed the problem of outliers in linear regression extensively. On the other hand, only recently Abuzaid et al. [12] discussed the identification of outliers in circular regression by using a new definition of circular residuals via different graphical and numerical methods.
Belsley et al. [11] used a row deletion approach to investigate the impact of deleting one row at a time on the estimated coefficients, fitted values and residuals of linear regression models. Any significant changes in the estimates suggest that the deleted observation is a candidate of outlier. In this paper we extend this approach to simple circular regression models. We investigate the impact of deleting one row from the circular regression data on the ratio of the determinants of the covariance matrix using all available observations to the determinants of the covariance matrix when the \( i \)th observation is deleted. The ratio will still be called the COVRATIO statistic.

### 2. CIRCULAR REGRESSION MODELS

#### 2.1 Regression of Circular Variables

The study on circular regression has started four decades ago. Gould [1] proposed a regression model to predict a circular response variable \( \Theta \) from a set of linear covariates, where \( \Theta \) follows von Mises distribution with mean \( \mu \) and concentration parameter \( \kappa \) denoted by \( \text{VM}(\mu, \kappa) \). The model is given by

\[
\mu = \mu_0 + \sum_{j=1}^{p} \beta_j x_j, \tag{1}
\]

where \( \mu_0 \) and \( \beta \) are unknown parameters and \( x_j \) is a linear covariate, for \( j = 1, \ldots, p \). Suppose that \( \theta_1, \theta_2, \ldots, \theta_n \) are independently distributed as the von Mises distributions with mean \( \mu \), concentration parameter \( \kappa \), Mardia [8] extended model (1) to give

\[
\mu_i = \mu_{0i} + \beta_i \omega, \tag{2}
\]

for some known numbers \( t_1, t_2, \ldots, t_n \) and unknown parameters \( \mu_{0i} \) and \( \beta_i \). Jammalamadaka and Sarma [4] proposed a regression model for two circular random variables \( X \) and \( Y \) in terms of the conditional expectation of the vector \( e^{i\theta} \) given \( x \) such that

\[
E(e^{i\theta} | x) = \rho(\theta|x)e^{i\mu(x)} = \varphi_1(\theta|x) + i\varphi_2(\theta|x),
\]

where \( \mu(x) \) is the conditional mean direction of \( y \) given \( x \) with conditional concentration \( 0 \leq \rho(\theta|x) \leq 1 \). Due to the difficulty of estimating \( \varphi_1(\theta|x) \) and \( \varphi_2(\theta|x) \), they are expressed instead as Fourier series.

In the case where \( X \) and \( Y \) are circular variables with mean directions \( \alpha \) and \( \beta \) respectively, Down and Mardia [2] applied the following mapping

\[
\tan \left( \frac{1}{2} (y - \beta) \tan((x - \alpha)/2) \right) = \alpha
\]

where \( \alpha \) is a slope parameter in the closed interval \([-1, 1]\). The mapping define a one-to-one relationship with a unique solution given by

\[
y = \beta + 2\tan^{-1} \left\{ \tan \left( \frac{x - \alpha}{2} \right) \right\}.
\]

They classified the regression model according to the nature of the parameters \( \alpha, \beta \) and \( \omega \). The maximum likelihood estimates are derived and the properties of the model are discussed with the applications to circadian biological rhythms and wind direction data.

On the other hand, Kato et al. [5] expressed the regression curve as a form of Mobius circle transformation. For circular random covariate \( X \) and circular response variable \( Y \), they proposed the regression curve

\[
y = \beta_0 + \frac{\beta_1 x + \beta_0}{1 + \beta_1 x}, \tag{3}
\]

where \( \beta_0 \) and \( \beta_1 \) are complex parameters with \( \beta_0 \in \Omega, \beta_1 \in C \), \( \beta_1 \) is the conjugate of \( \beta \), and \( \varepsilon \) follow a wrapped Cauchy distribution. For this model, the angular error is assumed to follow the wrapped Cauchy distribution, while Down and Mardia [2] assumed the angular error to follow the von Mises distribution. Due to the attractive properties of the wrapped Cauchy distribution, some desirable properties of the model were derived.
2.2 The Simple Regression Model for Circular Variables

Hussin et al. [3] extended model (2) for the case when both response and explanatory variables are circular. For any circular observations \((x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\) of circular variables \(X\) and \(Y\) with a linear relationship between them, they proposed a model given by

\[
y_i = \alpha + \beta x_i + \varepsilon_i \pmod{2\pi}
\]

where \(\varepsilon\) is circular random error having a von Mises distribution with circular mean 0 and concentration parameter \(\kappa\).

Some applications of model (3) are in the analysis of the wind and wave direction which are measured using two different techniques in order to compare an alternative instrument with the standard one. The log likelihood function of model (3) is given by

\[
\log L(\alpha, \beta, \kappa, x_1, ..., x_n, y_1, ..., y_n) = -n \log(2\pi) - n \log I_0(\kappa) + \kappa \sum \cos(y_i - \alpha - \beta x_i),
\]

where \(I_0(\kappa)\) is the modified Bessel function of order zero. The maximum likelihood estimates of model parameters are given by

\[
\hat{\alpha} = \begin{cases} \tan^{-1}(S/C) & S > 0, C > 0 \\ \tan^{-1}(S/C) + \pi & C < 0, C > 0 \\ \tan^{-1}(S/C) + 2\pi & S < 0, C > 0 \end{cases}
\]

where \(S = \sum \sin(y_i - \beta x_i)\) and \(C = \sum \cos(y_i - \beta x_i)\).

Due to the nonlinear nature of the first partial derivative of the log likelihood function, the parameter \(\beta\) can be estimated iteratively according to the formula

\[
\hat{\beta} \approx \hat{\beta}_0 + \frac{\sum x_i \sin(y_i - \hat{\alpha} - \hat{\beta}_0 x_i)}{\sum x_i^2 \cos(y_i - \hat{\alpha} - \hat{\beta}_0 x_i)}.
\]

The concentration parameter is estimated by

\[
\hat{\kappa} = A^{-1} \left(\frac{1}{n} \sum \cos(y_i - \hat{\alpha} - \hat{\beta}_0 x_i)\right),
\]

where the function \(A(\cdot)\) is the ratio of the modified Bessel function of the first kind of order one to the first kind of order zero. One of the inverses of function \(A(w)\) was approximated by Dobson [13]; it is given by

\[
A^{-1}(w) = \frac{9-8w+3w^2}{8(1-w)}. \tag{4}
\]

Hussin et al. [3] imposed a restriction on the values of the parameter to ensure the practicality of the model. Consider the following four points (in radian): \((0.10, 0.90), (2.00, 1.99), (4.30, 4.63)\) and \((6.25, 6.24)\) for illustration. The points are fairly close to the straight lines \(y = x\) and \(y = 4x \pmod{2\pi}\). By maximizing the log likelihood function of model (3), there is a clear maximum at \(\beta = 1\) in the interval \(0.5 < \beta < 1.5\). Other local maxima are observed at \(\beta = 4\) and \(\beta = 129.7\). However, there is no practical interpretation for the last two values. Thus, the value close to one would be a logical and simpler choice.

The asymptotic variances of the parameters have been derived and are given by

\[
\text{Var}(\hat{\alpha}) = \frac{\sum x_i^2}{\hat{\kappa}A(\hat{\kappa})(n\sum x_i^2 - (\sum x_i)^2)},
\]

\[
\text{Var}(\hat{\beta}) = \frac{n}{\hat{\kappa}A(\hat{\kappa})(n\sum x_i^2 - (\sum x_i)^2)},
\]

and

\[
\text{Var}(\hat{\kappa}) = \frac{1}{n(\hat{\kappa} - \hat{\kappa}A(\hat{\kappa}) - A(\hat{\kappa}))}.
\]

Further, the covariances of the parameters are given by

\[
\text{cov}(\hat{\alpha}, \hat{\beta}) = -\frac{\sum x_i}{\hat{\kappa}A(\hat{\kappa})(n\sum x_i^2 - (\sum x_i)^2)},
\]

\[
\text{cov}(\hat{\alpha}, \hat{\kappa}) = 0,
\]

\[
\text{cov}(\hat{\beta}, \hat{\kappa}) = 0,
\]

where \(A(\cdot) = \frac{1}{n} \sum \cos(y_i - \hat{\alpha} - \hat{\beta}_0 x_i)\), and \(\hat{\kappa}\) is the estimated concentration parameter of the circular random error.
Due to the close resemblance of model (3) to the linear regression model and its simple covariance matrix form, we investigate the possibility of utilizing a statistic for identifying outlier in the linear regression model to circular regression model. This is described in the next section.

3. COVRATIO STATISTIC FOR CIRCULAR REGRESSION

Belsley et al. [11] proposed a numerical statistic to identify outliers in linear regression models based on the determinantal ratio. It is given by
\[
\text{COVRATIO}_{(-i)} = \frac{|\text{COV}|}{|\text{COV}_{(-i)}|} = \frac{\kappa(k_i)}{\kappa(k_{i-1})} \left(\begin{array}{c}
\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2
\end{array}\right)
\]

where \(|\text{COV}|\) is the determinant of covariance matrix for full data set and \(|\text{COV}_{(-i)}|\) is that 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 them. In other words, the \(i^{th}\) observation is consistent with the other observations. Alternatively, if the value of \(|\text{COVRATIO}_{(-i)}| - 1| = 0\) exceeds the cut-off point obtained in the next section, then it indicates that the \(i^{th}\) observation is a candidate to be an outlier, where \(n\) is the sample size.

We now extend the idea to the circular regression case. The covariance matrix for the simple circular regression model (3) is given by
\[
\text{COV} = \frac{1}{\kappa(k)} \left(\begin{array}{c}
\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2
\end{array}\right)
\]

and the determinant of the covariance matrix is
\[
|\text{COV}| = \frac{1}{\kappa(k(i))}
\]

Thus, the COVRATIO statistic for the \(i^{th}\) observation is given by
\[
\text{COVRATIO}_{(-i)} = \frac{|\text{COV}|}{|\text{COV}_{(-i)}|} = \frac{\kappa(k_{i-1})}{\kappa(k_i)}
\]

The above statistic is simple and easy to be obtained. Hence, any observation with \(|\text{COVRATIO}_{(-i)}| - 1| > 0\) exceeding the cut-off point obtained in the next section will be identified as an outlier.

4. CUT-OFF POINTS OF COVRATIO STATISTIC

The cut-off points of the statistic are obtained by using Monte Carlo simulation method. Fifteen different sample sizes of \(n = 10, 20, ..., 150\) and six values of concentration parameter \(\kappa = 0.5, 1, 2, 5, 7\) and 10 are used. For each sample size \(n\) and \(\kappa\), a set of circular random errors is generated from von Mises distributions with mean direction 0 and concentration parameter \(\kappa\). Then, we continue with the following steps:

Step 1. Generate the response variable \(Y\) from \(\text{VM}(\pi/4, 10)\). The parameters of simple circular regression model are fixed at \(\alpha = 0\) and \(\beta = 1\).

Step 2. Calculate the observed values of the response variable \(Y\) based on model (3).

Step 3. Fit the generated circular data by using model (3).

Step 4. Calculate \(|\text{COV}|\) by using equation (4).

Step 5. Exclude the \(i^{th}\) row from the generated sample, where \(i = 1, ..., n\). Repeat steps 3-5 to obtain \(|\text{COV}_{(-i)}|\) for all \(i\).

Step 6. Compute \(\text{COVRATIO}_{(-i)}\) by using Eq. (5) and then obtain the value of \(|\text{COVRATIO}_{(-i)} - 1|\) for all \(i\).

Step 7. Specify the maximum value of \(|\text{COVRATIO}_{(-i)} - 1|\).

The process is carried out 2,000 times for each combination of sample size \(n\) and concentration parameter \(\kappa\). Then the 10%, 5% and 1% upper percentiles of the maximum values of \(|\text{COVRATIO}_{(-i)} - 1|\) are calculated.

Figure 1 plots the percentile values of \(|\text{COVRATIO}_{(-i)} - 1|\) against the concentration parameter \(\kappa\) for \(n = 50\). The results show that the values are independent of the
concentration parameter $\kappa$. Similar results are obtained for other sample size $n$. Consequently, we take the arithmetic mean of percentile values for different $\kappa$ as the cut-off point for each sample size $n$.

Table 1 gives the cut-off points and the corresponding standard deviations given in parentheses for various sample size $n$. The results show that the cut-off point is a decreasing functions of sample size $n$. The values of the standard deviations are also very small indicating the nondependency of the percentiles values on the concentration parameter.

For the linear case, Belsley et al. [11] had stated that $\left(\frac{6}{n}\right)$ is an appropriate cut-off point of $|\text{COVRATIO}_{(-i)}-1|$ statistic at 5% significant level. It is of interest to find such formula for the circular case. We found that the cut-off points for any sample size are very close to the value $\left(\frac{16}{n}\right)$ where $n$ is the sample size. The exact values of $\left(\frac{16}{n}\right)$ are given in the last row of Table 1, followed by the bias of the cut-off points at 5% significant level in parentheses. It is obvious that the biases in all cases are less than 0.022. Thus, the approximated value $\left(\frac{16}{n}\right)$ can be used as the cut-off points of the $|\text{COVRATIO}_{(-i)}-1|$ statistic at 5% significant level for any sample of size $n$. 

![Figure 1. The upper percentile points of $|\text{COVRATIO}_{(-i)}-1|$ for $n=50$.](image)
5. THE POWER OF PERFORMANCE OF COVRATIO STATISTIC

Monte Carlo simulation method is used to examine the performance of \(|COVRATIO_{(i-1)}|\) statistic for detecting outliers in the simple circular regression model. Samples of five different sample sizes \(n = 30, 50, 70, 100\) and 150 are used. We follow similar procedure described in Section 4 to generate the data. In addition, we let the observation at position \(d\), say \(y[d]\), be contaminated such that:

\[
y^*[d] = y[d] + \lambda \pi \pmod{2\pi},
\]

where \(y^*[d]\) is the value of \(y[d]\) after contamination and \(\lambda\) is the degree of contamination in the range \(0 \leq \lambda \leq 1\). The generated data of \(X\) and \(Y\) are then fitted by using model (3) and \(|COV|\) is calculated using Eq. (4). Consequently, by excluding the \(i^{th}\) row from the sample, for \(i = 1, \ldots, n\), and refitting the reduced data, we calculate the values of the \(COVRATIO_{(i-1)}\) by using Eq. (5). Finally, we specify the maximum value of the \(|COVRATIO_{(i-1)}|\). The process is repeated 2,000 times and the power of performance of \(|COVRATIO_{(i-1)}|\) statistic is examined by computing the percentage of correct detection of the contaminated observation at position \(d\). Three main factors are considered, namely the level of contamination \(\lambda\), concentration parameter \(\kappa\) and sample size \(n\). Figure 2 illustrates the power of performance of \(|COVRATIO_{(i-1)}|\) statistic for \(n = 70\) and four values of the concentration parameter \(\kappa = 2, 5, 7\) and 10. The following results are observed:

1. It can be seen that the performance is an increasing function of concentration parameter.

2. In the case when the concentration parameter is small \(\kappa \leq 2\), the percentages of correct detection are almost zero for any sample size and contamination level.

\[
\begin{array}{cccccc}
\hline
n & 10 & 20 & 30 & 40 & 50 \\
\hline
10\% & 1.530(0.2\times10^{-3}) & 0.729(0.1\times10^{-3}) & 1.469(0.1\times10^{-3}) & 0.337(0.2\times10^{-3}) & 0.287(0.2\times10^{-3}) \\
5\% & 1.609(0.4\times10^{-3}) & 0.806(0.9\times10^{-3}) & 0.554(0.3\times10^{-3}) & 0.416(0.4\times10^{-3}) & 0.332(0.2\times10^{-3}) \\
1\% & 1.783(0.1\times10^{-3}) & 0.973(0.3\times10^{-3}) & 0.714(0.4\times10^{-3}) & 0.591(0.1\times10^{-3}) & 0.420(0.2\times10^{-3}) \\
(16/n) & 1.600 (0.009)* & 0.800 (0.006)* & 0.533 (0.021)* & 0.400 (0.016)* & 0.320 (0.012)* \\
\hline
n & 60 & 70 & 80 & 90 & 100 \\
\hline
10\% & 0.230(0.1\times10^{-4}) & 0.205(0.4\times10^{-5}) & 0.177(0.4\times10^{-6}) & 0.158(0.3\times10^{-5}) & 0.149(0.7\times10^{-4}) \\
5\% & 0.272(0.3\times10^{-4}) & 0.231(0.4\times10^{-5}) & 0.206(0.1\times10^{-5}) & 0.183(0.2\times10^{-5}) & 0.167(0.8\times10^{-5}) \\
1\% & 0.360(0.3\times10^{-4}) & 0.301(0.5\times10^{-5}) & 0.270(0.2\times10^{-5}) & 0.250(0.6\times10^{-5}) & 0.217(0.6\times10^{-5}) \\
(16/n) & 0.267 (0.005)* & 0.227 (0.003)* & 0.200 (0.006)* & 0.178 (0.005)* & 0.160 (0.007)* \\
\hline
n & 110 & 120 & 130 & 140 & 150 \\
\hline
10\% & 0.134(0.7\times10^{-5}) & 0.122(0.7\times10^{-5}) & 0.112(0.7\times10^{-5}) & 0.103(0.6\times10^{-5}) & 0.104(0.1\times10^{-4}) \\
5\% & 0.152(0.9\times10^{-5}) & 0.141(0.2\times10^{-5}) & 0.131(0.2\times10^{-5}) & 0.121(0.1\times10^{-5}) & 0.113(0.4\times10^{-5}) \\
1\% & 0.203(0.9\times10^{-6}) & 0.190(0.4\times10^{-6}) & 0.181(0.9\times10^{-6}) & 0.171(0.4\times10^{-6}) & 0.147(0.4\times10^{-6}) \\
(16/n) & 0.145 (0.006)* & 0.133 (0.007)* & 0.123 (0.007)* & 0.114 (0.006)* & 0.107 (0.006)* \\
\end{array}
\]
(3) As the concentration parameter increases, the performance also increases accordingly.

(4) For large concentration parameter ($\kappa = 5, 7$ and 10), the performance highly depends on the level of contamination. When $\lambda < 0.3$, the performance is very weak and it increases rapidly up to $\lambda < 0.8$ before apparently being stable at $\lambda = 0.9$. On the other hand, the power of performance is an increasing function of sample size $n$ as shown in Figure 3.

6. PRACTICAL EXAMPLE

As an illustration, we consider 129 measurements of wind direction (in radians) recorded over a period of 22.7 days along the Holderness coastline (the Humberside coast of the North Sea, United Kingdom) by using two different instruments: an HF radar system and an anchored wave buoy. Figure 4 shows the scatter plot of the wind direction data with the scale broken artificially at $0 = 2\pi$. Two points seem to be outliers at the top left of the plot. However, they are actually consistent with the rest of the observations as they are close to other observations at the top right or bottom left due to the closed range property of the circular variable.

Further, the scatter plot in Figure 4 shows a linear relationship between the measurements of the HF radar system and the anchored wave buoy. Since both variables are circular, model (3) is used to fit the data. The maximum

![Figure 2. Power of performance for $|\text{COV-RATIO}_{\psi_i} - 1|$ when $n = 70.$](image)
Figure 3. Power of performance for $|\text{COV}\text{RATIO}_{\gamma} - 1|$ when $\kappa = 7$.

Figure 4. Scatter plot of the wind data.
likelihood estimates of the parameters are $\hat{\alpha} = 0.165$, $\beta = 0.973$ and $\kappa = 7.34$ giving

$$\hat{y} = 0.165 + 0.973x \pmod{2\pi}$$

The value of the determinant of the covariance matrix for full data set $|\text{COV}|$ is $2.89 \times 10^{-7}$ and the corresponding cut-off point is 0.124. The values of $|\text{COVRATIO}_{(-i)}$ are plotted in Figure 5. It can be seen that the values for observations number 38 and 111 exceed the cut-off point. Thus, the proposed COVRATIO statistic for simple circular regression model has successfully identified both observations as outliers as found by Abuzaid et al. [12].

![Figure 5](image.png)

**Figure 5.** Plot of $|\text{COVRATIO}_{(-i)}$ for wind direction data.

7. CONCLUSIONS

The extension of COVRATIO statistic to simple circular regression model shows consistent features as in the linear case. The cut-off points are found to be $\left( \frac{16}{n} \right)$ at 0.05 significant level compared to $\left( \frac{6}{n} \right)$ for the linear case. The statistic shows better performance for larger concentration parameter or larger sample size. Furthermore, it is able to detect similar set of outliers as found in Abuzaid et al. [12].

REFERENCES


