RESEARCH ARTICLE
doi: 10.2306/scienceasia1513-1874.2010.36.249

ScienceAsia 36 (2010): **249–253**

Asymptotic covariance and detection of influential observations in a linear functional relationship model for circular data with application to the measurements of wind directions

Abdul G. Hussin[a,*], **Ali Abuzaid**[b], **Faiz Zulkifili**[a], **Ibrahim Mohamed**[b]

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

*Corresponding author, e-mail: ghapor@um.edu.my

Received 4 May 2010
Accepted 30 Aug 2010

ABSTRACT: This paper discusses the asymptotic covariance and outlier detection procedure in a linear functional relationship model for an extended circular model proposed by Caires and Wyatt. We derive the asymptotic covariance matrix of the model via the Fisher information and use the results to detect influential observations in the model. Consequently, an influential observation detection procedure is developed based on the COVRATIO statistic which has been widely used for similar purposes in ordinary linear regression models. We show via simulation that the above procedure performs well in detecting influential observations. As an illustration, the procedure is applied to the real data of the wind direction measured by two different instruments.

KEYWORDS: circular variables, error of concentration parameters, maximum likelihood estimation

INTRODUCTION AND THE MODEL

Caires and Wyatt[1] suggested and derived a consistent parameter estimator for a linear functional relationship model for the case when both variables are circular. As an analogy to the linear functional relationship model for real data, we assume both sets of observations for circular variables $\boldsymbol{X}$ and $\boldsymbol{Y}$ are observed with errors[2]. Suppose x_i and y_i are the observed values of the circular variables $\boldsymbol{X}$ and $\boldsymbol{Y}$, respectively, $0 \leqslant x_i, y_i < 2\pi$, for $i = 1, \ldots, n$. For any fixed X_i we assume that the observations x_i and y_i are measured with errors δ_i and ε_i, respectively. We use the same notation here as in the linear functional relationship model (for continuous or real variables) and thus the full model as proposed by Caires and Wyatt[1] can be written as

$$\begin{aligned} &x_i = X_i + \delta_i, \quad y_i = Y_i + \varepsilon_i, \\ &Y_i = \alpha + X_i \pmod{2\pi}, \qquad i = 1, 2, \ldots, n. \end{aligned} \tag{1}$$

We also assume δ_i and ε_i are independently distributed with von Mises distributions, that is $\delta_i \sim \mathrm{VM}(0, k)$ and $\varepsilon_i \sim \mathrm{VM}(0, v)$. There are $n + 3$ parameters to be estimated. These are α, κ, v and the incidental parameters $X_1, \ldots, X_n$. The parameters are estimated by using the maximum likelihood estimation method. Assuming that the ratio of the error concentration parameters, i.e., $v/\kappa = \lambda$, is known, then the log likelihood function is given by

$$\begin{aligned} \log L(\alpha, \kappa, X_i; \lambda, x_i, y_i) = &-2n \log(2\pi) - n \log I_0(\kappa) \\ &- n \log I_0(\lambda\kappa) + \kappa \sum \cos \eta_i + \lambda\kappa \sum \cos \tau_i, \end{aligned}$$

where $\eta_i = x_i - X_i$ and $\tau_i = y_i - \alpha - X_i$. Differentiating the log likelihood function with respect to α, κ, and X_i, we obtain the likelihood equations for the parameters which may be solved iteratively. It can be shown that the estimates of α and X_i, which are $\hat{\alpha}$ and $\hat{X}_i$, respectively, are given by

$$\hat{\alpha} = \tan^{-1} \left\{ \frac{\sum \sin(y_i - \hat{X}_i)}{\sum \cos(y_i - \hat{X}_i)} \right\}, \tag{2}$$

$$\hat{X}_{i1} \approx \hat{X}_{i0} + \frac{\sin \hat{\eta}_{i0} + \lambda \sin \hat{\tau}_{i0}}{\cos \hat{\eta}_{i0} + \lambda \cos \hat{\tau}_{i0}} \tag{3}$$

where $\hat{\eta}_{i0} = x_i - \hat{X}_{i0}$, $\hat{\tau}_{i0} = y_i - \hat{\alpha} - \hat{X}_{i0}$ and $\hat{X}_{i1}$ is an improvement of $\hat{X}_{i0}$ by taking X_i as an initial value. We then can find an estimate of κ for any value of λ from the equation

$$A(\kappa) + \lambda A(\lambda\kappa) = w \equiv \frac{1}{n} \left\{ \sum \cos \hat{\eta}_i + \lambda \sum \cos \hat{\tau}_i \right\} \tag{4}$$

where

$$A(r) = \frac{I_1(r)}{I_0(r)} = 1 - \frac{1}{2r} - \frac{1}{8r^2} - \frac{1}{8r^3} + \ldots, \quad (5)$$

and $\hat{\eta}_i = x_i - \hat{X}_i$, $\hat{\tau}_i = y_i - \hat{\alpha} - \hat{X}_i$, and $I_0(r)$ and $I_1(r)$ are the asymptotic power series for the Bessel functions[3]. Simplifying (4) using (5), we have the approximate result

$$8(1+\lambda-w)\kappa^3 - 8\kappa^2 - \left(1+\frac{1}{\lambda}\right)\kappa - \left(1+\frac{1}{\lambda^2}\right) = 0. \quad (6)$$

Eq. (6) has one positive real root and two complex roots. The positive real root is taken as the estimate of κ, denoted by $\hat{\kappa}$. For $\lambda = 1$, $\hat{\kappa} = A^{-1}(w)$ where the inverse function of A is defined by Fisher[4] as follows:

$$A^{-1}(w) = \begin{cases} 2w + w^3 + \frac{5w^5}{6}, & w < p, \\ -0.4 + 1.39w + \frac{0.43}{1-w}, & p \leqslant w < q, \\ (w^3 - 4w^2 + 3w)^{-1}, & w \geqslant q, \end{cases}$$

where $p = 0.53$ and $q = 0.85$. Caires and Wyatt[1] applied the model to the problem of assessing radar measurements of wave data. The radar measurements were compared with wave model predictions and buoy measurements. The combined results gave a fairly good picture of the quality of the radar data. This model can be applied in many other areas of applied sciences whenever the objective is to look at the underlying relationship between two sets of circular data rather than to predict one variable from other.

It is of interest to derive the asymptotic covariance matrix of the parameters of the above model. The results can then be used in the identification of influential observations in the model. Influential observations are observations that are subjected to contamination by some unexpected events. The existence of influential observations in a data set may affect the parameter estimates and consequently lead to a wrong conclusion. Many procedures are available to identify influential observations in linear regression models[5,6] but fewer in functional relationship models[7]. Recently, Abuzaid et al[8] discussed the identification of single outliers which can be influential in simple circular regression models based on the circular residuals via graphical tools and numerical procedures. In this paper, the proposed influential observation detection procedure is based on the determinant of the asymptotic covariance matrix of the parameters of model (1). In the following section, we derive the asymptotic covariance matrix of the parameters of the model. The next two sections describe the proposed influential observation detection procedure in detail. Simulation studies are carried out to obtain the percentage points of the procedure and to investigate the performance of the procedure. We will then apply the procedure to wind direction data measured by two different instruments.

ASYMPTOTIC COVARIANCE OF PARAMETERS

In this section we derive the asymptotic covariance of parameters for a linear functional relationship model for circular data. We assume that the ratio of the error concentration parameters, denoted by κ, is known via the Fisher information matrix. By considering the first partial derivative and minus the expected value of the second partial derivative of the log likelihood function, we obtain the estimated Fisher information matrix, F, for $\hat{X}_i, \ldots, \hat{X}_n$, $\hat{\kappa}$ and $\hat{\alpha}$ given by

$$F = \begin{bmatrix} R & \mathbf{0} & W \\ \mathbf{0} & S & \mathbf{0} \\ W^T & \mathbf{0} & U \end{bmatrix}$$

where R is an $n \times n$ diagonal matrix with all diagonal elements equal to $\hat{\kappa}A(\hat{\kappa}) + \lambda\hat{\kappa}A(\lambda\hat{\kappa})$, W is an $n \times 1$ column vector with all elements equal to $\lambda\hat{\kappa}A(\lambda\hat{\kappa})$, $S = n\lambda^2 A'(\lambda\hat{\kappa}) + nA'(\hat{\kappa})$ where $A'(\kappa) = 1 - A^2(\kappa) - A(\kappa)/\kappa$, and $U = \lambda\hat{\kappa}nA(\lambda\hat{\kappa})$.

We are primarily interested in the bottom right minor of the inverse of F of order 2×2, which forms the asymptotic covariance matrix of $\hat{\kappa}$ and $\hat{\alpha}$. From the theory of partitioned matrices[9], the covariance matrix is given by

$$\mathrm{Var}\begin{bmatrix} \hat{\kappa} \\ \hat{\alpha} \end{bmatrix} = \begin{bmatrix} S^{-1} & \mathbf{0} \\ \mathbf{0} & (U - W^T R^{-1} W)^{-1} \end{bmatrix}.$$

It can be shown that

$$(U - W^T R^{-1} W)^{-1} = \frac{A(\hat{\kappa}) + \lambda A(\lambda\hat{\kappa})}{\lambda\hat{\kappa}A(\hat{\kappa})A(\lambda\hat{\kappa})}.$$

In particular, we have the following results:

$$\mathrm{Var}(\hat{\kappa}) = \frac{1}{n\lambda^2 A'(\lambda\hat{\kappa}) + nA'(\hat{\kappa})},$$

$$\mathrm{Var}(\hat{\alpha}) = \frac{[A(\hat{\kappa}) + \lambda A(\lambda\hat{\kappa})]}{\lambda\hat{\kappa}A(\hat{\kappa})A(\lambda\hat{\kappa})},$$

and $\mathrm{Cov}(\hat{\kappa}, \hat{\alpha}) = 0$. For $\lambda = 1$, $\mathrm{Var}(\hat{\kappa}) = 1/2nA'(\hat{\kappa})$ and $\mathrm{Var}(\hat{\alpha}) = 2/\hat{\kappa}A(\hat{\kappa})$, where

$$A(\kappa) = \frac{1}{2n}\left\{\sum \cos\hat{\eta}_i + \sum \cos\hat{\tau}_i\right\}$$

$$A'(\kappa) = 1 - A^2(\kappa) - \frac{A(\kappa)}{\kappa}.$$

These results will be used in the influential observation detection procedure discussed in the following section.

INFLUENTIAL OBSERVATION DETECTION

The COVRATIO statistic has long been used to identify influential observations in linear regression models via a row deletion approach[5]. In this section, we extend the usage of the statistic to detect influential observations in functional relationship models for circular data. The statistic is the ratio of the estimated covariance matrix of an estimated parameter using the full data and the estimated covariance matrix of the reduced data set when the ith observation is deleted. Belsely et al[5] suggested using the statistic to measure the effect of removing the observation based on the determinantal ratio given by

$$\mathrm{COVRATIO}_{(-i)} = \frac{|\mathrm{COV}_{(-i)}|}{|\mathrm{COV}|},$$

where $|\mathrm{COV}|$ is the determinant of covariance matrix for the full data set and $|\mathrm{COV}_{(-i)}|$ is for the reduced data set by excluding the ith row. If the ratio is close to unity then the ith observation is not an influential observation. For convenience, the $|\mathrm{COVRATIO}_{(-i)}-1|$ statistic is usually used in linear regression cases.

Using the same idea, the COVRATIO statistic for the functional relationship model for circular data (1) can be shown to be

$$\mathrm{COVRATIO}_{(-i)} = \frac{n(n-1)^{-1}\hat{\kappa}A(\hat{\kappa})A'(\hat{\kappa})}{\hat{\kappa}_{(-i)}A(\hat{\kappa}_{(-i)})A'(\hat{\kappa}_{(-i)})}, \quad (7)$$

where $\hat{\kappa}_{(-i)}$, $A(\hat{\kappa}_{(-i)})$, and $A'(\hat{\kappa}_{(-i)})$ are, respectively, the estimated concentration parameter, the ratio of the modified Bessel function for the first kind of order one and first kind of order zero, and its first derivative, for the reduced data set. The following subsections discuses the percentage points and the power of performance for the COVRATIO statistic (7).

The percentage points of the COVRATIO statistic

The Monte Carlo simulation method is used to obtain the percentage points of the COVRATIO statistic by considering five different sample sizes $n = 20$, 30, 50, 100, and 200. By assuming that the ratio of the error concentration parameters $\lambda = 1$, six values of the error concentration parameter $\kappa = 10$, 15, 30, 50, 70, and 100 are considered. For each combination of sample size n and error concentration parameter κ, two sets of random errors δ and ε are generated from the von Mises distribution with mean 0 and concentration κ, $\mathrm{VM}(0,\kappa)$. We generate $\boldsymbol{X}$ of size n from $\mathrm{VM}(\pi/4, 1.5)$ and fix the intercept parameter, α at 0. The response variable $\boldsymbol{Y}$ is obtained based on model (1). Subsequently, the generated data is fitted using model (1) and the $|\mathrm{COV}|$ is calculated. Then we exclude the ith row from the generated data, $i = 1, \ldots, n$, to obtain the $|\mathrm{COV}_{(-i)}|$ and $|\mathrm{COVRATIO}_{(-i)} - 1|$ statistics. The process is repeated 2000 times and the 10th, 5th and 1st upper percentiles of the maximum values of $|\mathrm{COVRATIO}_{(-i)} - 1|$ are calculated.

Table 1 Cut-off points for the null distribution of $|\mathrm{COVRATIO}_{(-i)} - 1|$ statistic and its standard error.

n	90%	95%	99%
20	0.568(0.008)	0.690(0.024)	1.024(0.058)
30	0.368(0.004)	0.432(0.006)	0.589(0.036)
50	0.223(0.006)	0.260(0.008)	0.346(0.018)
100	0.119(0.003)	0.136(0.004)	0.171(0.003)
200	0.064(0.001)	0.072(0.002)	0.090(0.003)

Simulation results show that the percentage points vary only slightly for each level of the concentration parameter, κ, and are not shown here. Thus the arithmetic mean of the simulated percentage points for each sample size n are considered as the cut-off points and are given in Table 1. The corresponding standard deviations of the percentage points are given in parentheses. Results show that the cut-off points are a decreasing function of the sample size n. The values of standard deviation are very small which indicates that the error concentration parameter κ does not vary much around the arithmetic mean.

Power of performance of the COVRATIO statistic

Simulation studies are carried out to examine the performance of the COVRATIO statistic for model (1). A similar procedure to that described in the previous subsection is used to generate the data. In addition, we contaminate an observation at position $[d]$ as follows:

$$Y^*_{[d]} = Y_{[d]} + \zeta\pi \pmod{2\pi},$$

where $Y^*_{[d]}$ is the value of $Y_{[d]}$ after contamination and ζ is the degree of contamination in the range $0 \leqslant \zeta \leqslant 1$. The generated data are then fitted by model (1) and the maximum of the $\mathrm{COVRATIO}_{(-i)} - 1$ statistic is specified. The process is repeated 2000 times and the power of performance is examined by computing the percentage of correct detection of the contaminated observation at position $[d]$.

Fig. 1(a) gives the plot of power of performances of the procedure for n=50 and various values κ. It can be seen that the performance is an increasing function of the level of contamination ζ and the

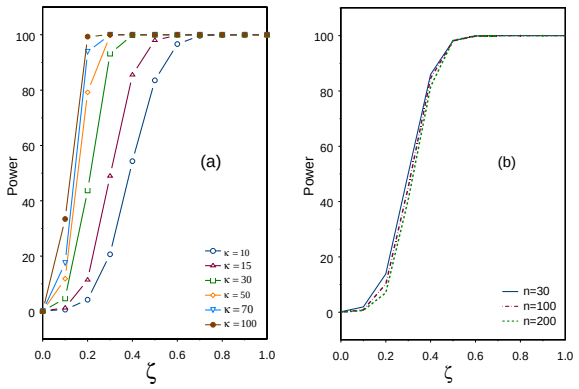

Fig. 1 The $|\mathrm{COVRATIO}_{(-i)} - 1|$ statistic of wind direction data for (a) $n = 50$, (b) $\kappa = 15$.

error concentration parameter κ. On the other hand, Fig. 1(b) gives the plot for κ=15 and different sample sizes. The power of performance decreases slightly as the sample size increases. For all cases, the power of performance is almost 100% for $\zeta > 0.6$.

NUMERICAL EXAMPLE

As an illustration, we consider 129 measurements of wind directions (in radians) recorded over the period of 22.7 days along the Holderness coastline (the Humberside coast of the North Sea, UK) by using two different instruments (HF radar system and anchored wave buoy [10]). Fig. 2 shows the scatter plot of wind direction data with the scale broken artificially at $0 = 2\pi$. Two points seem to be far from others at the left top of the plot. However, they are actually consistent with the rest of observations as they are close to other observation at the right top or left bottom due to the closed range property of the circular variable.

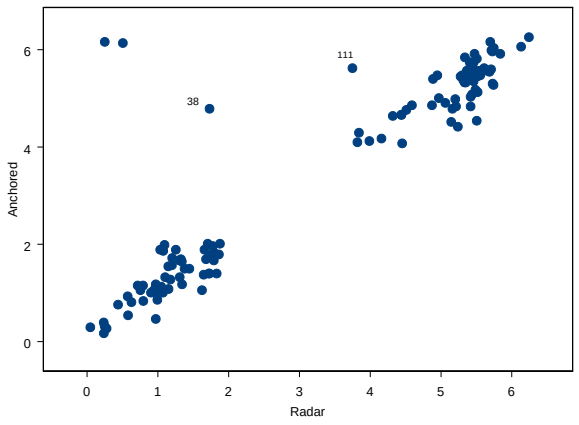

Fig. 2 Scatter plot of wind data measured by HF radar system and anchored wave buoy.

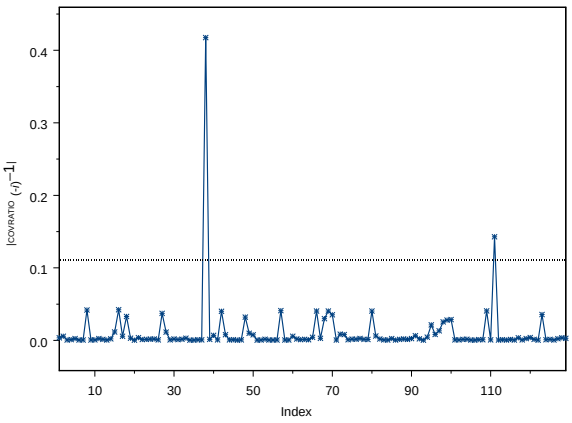

Fig. 3 Power of performance for $|\mathrm{COVRATIO}_{(-i)} - 1|$ statistic with cut-off point denoted by dashed line.

We fit the full data using model (1) giving

$$Y = 0.086 + X \pmod{2\pi}.$$

The estimates of α and κ are 0.0857 and 22.8584 with standard error 0.4280 and 1.0064, respectively. The determinant of the covariance matrix of model (1) based on the full data set is 0.1855. Since the sample size is 129, the cut-off point considered is 0.11 at 0.05 significance level as given in Table 1. The values of the $|\mathrm{COVRATIO}_{(-i)} - 1|$ statistic are plotted in Fig. 3. It is obvious that there are two points that exceed the cut-off points (dashed line). Thus observations 38 and 111 are identified as influential observations. The results agree with the findings in Abuzaid et al [8]. After removing these two points and reanalysing the reduced data set, the new estimates of α and κ are 0.0575 and 41.1326 with standard errors of 0.3157 and 1.8250, respectively. Compared to results before the removal of the two points, we note that the value of $\hat{\alpha}$ of the reduced data is closer to 0 and κ has almost doubled. This suggests that both observations are influential. Since the purpose of this model is to look at the underlying relationship between two circular variables, it is very important to identify influential observations in the data set. This can be achieved through the proposed procedure.

CONCLUSIONS

In this paper, we derive the asymptotic covariance matrix of the parameters of the linear functional relationship model for circular data. The determinant of the covariance matrix is used to identify possible influential observations via the COVRATIO statistic based on the row deletion approach. This procedure allows us to detect possible influential observations

in a given bivariate circular data set. Based on a simulation study, we obtain the cut-off points of the procedure for different sample sizes and three significance levels. The procedure has been shown to perform well in detecting influential observations. As an illustration, this procedure has been applied to the wind direction data and the results agree with the previous studies.

REFERENCES

1. Caires S, Wyatt LR (2003) A linear functional relationship model for circular data with an application to the assessment of ocean wave measurements. *J Agr Biol Environ Stat* **8**, 153–69.
2. Kendall MG, Stuart A (1973) *The Advanced Theory of Statistics*, Griffin, London.
3. Abramowitz M, Stegun IG (1965) *Handbook of Mathematical Functions*, Dover, New York.
4. Fisher NI (1993) *Statistical Analysis of Circular Data*, Cambridge Univ Press, Cambridge.
5. Belsley DA, Kuh E, Welsch RE (1980) *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*, Wiley, New York.
6. Chatterjee S, Hadi AS, Price B (2000) *Regression Analysis by Example*, Wiley, New York.
7. Zamar RH (1989) Robust estimation in the errors-in-variables model. *Biometrika* **76**, 149–60.
8. Abuzaid AH, Hussin AG, Mohamed IB (2008) Identifying single outlier in linear circular regression model based on circular distance. *J Appl Probab Stat* **3**, 107–17.
9. Graybill FA (1961) *Theory and Application of the Linear Model*, Duxbury Press, North Scituate, Massachusetts.
10. Sova MS (1995) The sampling variability and the validation of high frequency radar measurements of the sea surface. PhD thesis, Univ of Sheffield.