Sains Malaysiana 44(7)(2015): 1027–1032

Outlier Detection in a Circular Regression Model

(Pengesanan Terpencil dalam Model Regresi Berkeliling)

ADZHAR RAMBLI*, ROSSITA MOHAMAD YUNUS, IBRAHIM MOHAMED & ABDUL GHAPOR HUSSIN

ABSTRACT

Recently, there is strong interest on the subject of outlier problem in circular data. In this paper, we focus on detecting outliers in a circular regression model proposed by Down and Mardia. The basic properties of the model are available including the exact form of covariance matrix of the parameters. Hence, we intend to identify outliers in the model by looking at the effect of the outliers on the covariance matrix. The method resembles closely the COVRATIO *statistic for the case of linear regression problem. The corresponding critical values and the performance of the outlier detection procedure are studied via simulations. For illustration, we apply the procedure on the wind data set.*

Keywords: Circular; circular regression; COVRATIO; *influential observation; outlier*

ABSTRAK

Pada masa ini, terdapat minat yang mendalam pada subjek masalah terpencil dalam data berkeliling. Dalam kertas ini, kami menumpukan untuk mengesan pencilan dalam satu model regresi berkeliling yang dicadangkan oleh Down dan Mardia. Sifat asas model yang disediakan termasuk parameter bentuk matriks kovarians yang tepat. Oleh itu, kami berhasrat untuk mengenal pasti pencilan dalam model ini dengan melihat kesan daripada pencilan dalam matriks kovarians. Kaedah ini hampir menyerupai statistik COVRATIO *bagi kes masalah regresi linear. Nilai kritikal sepadan dan prestasi prosedur pengesanan pencilan dikaji melalui simulasi. Untuk ilustrasi, kami menggunakan prosedur set data angin.*

Kata kunci: Berkeliling; regresi berkeliling; COVRATIO; *pemerhatian berpengaruh; terpencil*

INTRODUCTION

The study of outliers has been widely carried out in different areas of statistics. Their occurrence may be due to error or part of the phenomena under study. The earliest work includes that of Beckman and Cook (1983) who reviewed different approaches in dealing with outliers. Extensive studies on the subject can be found in linear regressions (Barnett & Lewis 1984; Belsley et al. 1980). On the other hand, only few studies of outliers in circular regression can be found in the literature. Abuzaid et al. (2013, 2011) and Ibrahim et al. (2013) explored the problem in two types of circular regression model using row deletion approach while Hussin et al. (2010) discussed the detection of influential observation in a linear functional relationship model for circular data.

Circular regression attempts to model the relationship between a circular dependent and a set of circular independent variables. The circular variables can be envisaged as being distributed on the circumference of a unit circle in the range $[0, 2\pi]$ radian. They are commonly found in many scientific fields including meteorology and biology. Laycock (1975) proposed a circular regression model of two circular variables u and v mimicking the complex linear regression. Later, Rivest (1997) proposed another model to predict the v-direction based on the rotation of the decentered u-angle. On the other hand, Jammalamadaka and Sarma (1993) considered the conditional expectation of the vector $e^{(iv)}$ given u which are further expressed in terms of Fourier series expansions with errors are assumed to follow a normal distribution. Meanwhile Hussin et al. (2004) proposed another simple model with a specific application on measuring the linear relationship between two circular variables. As in the linear case, the occurrence of outliers in bivariate circular data may affect the parameter estimation and forecasting accuracy. Abuzaid et al. (2011) identified outliers in a simple circular regression model by looking at the changes caused by removing one observation at a time on the covariance matrix of parameters. Later, Ibrahim et al. (2013) used similar idea with the covariance matrix of parameters is replaced by the sample covariance matrix when modeled using the Jammalamadaka and Sarma's circular regression. In this paper, we extend the idea to another circular regression model proposed by Downs and Mardia (2002) as the model has an exact form of covariance matrix of parameters to be utilized in the approach. We shall use the term 'DM circular regression model' to mean this model in the rest of the paper.

With that view in mind, this paper is organized as follows: Section 2 reviews the DM circular regression models and the method of estimating the parameters. Section 3 presents the derivation of covariance matrix of

the model. Section 4 presents the definition of *COVRATIO* statistic to be used in identifying influential observations in DM circular regression models. We then study the sampling behavior and performance of the procedure of detecting outliers in Sections 5 and 6, respectively. Finally, we apply the procedure on the wind data set as given in Hussin et al. (2004).

DM CIRCULAR REGRESSION MODEL

We consider the circular regression model proposed by Downs and Mardia (2002). Assume that v is the dependent random circular variable with circular location β, u is the fixed independent circular variable with circular location α and ω is a slope parameter in the closed interval [-1, 1]. The model is given by

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

which has a unique solution given by

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

The model has three functionally independent parameters α, β and ω, which can be estimated using the maximum likelihood estimation method. Given a random sample of $(u_j, v_j), j = 1, 2, \ldots, n$, the log-likelihood function can be obtained such that

$$\begin{aligned} l(\alpha,\beta,\omega,v_1,\ldots,v_n) = &-n\log I_0(\kappa) \\ &+\kappa\Sigma_j\cos(v_j-\beta-v(u_j-\alpha;\omega)) \\ &+\text{constant}, \end{aligned} \tag{3}$$

where $v(u_j-\alpha;\omega)=2\tan^{-1}\left\{\omega\tan\frac{1}{2}(u_j-\alpha)\right\}$. Since the first term on the right-hand side is constant, we may work with the precision parameter $\hat{\rho}$ such that

$$\hat{p}(\alpha,\beta,\omega)=\frac{1}{n}\sum\nolimits_j\cos\left(v_j-\beta-v\left(u_j-\alpha;\omega\right)\right). \tag{4}$$

As a special case, we may set α and β to have a relationship such that $\alpha \pm \beta = 0$ or $\beta = \pm\alpha$. Hence, the log-likelihood functions of (3) and maximum likelihood estimator $\hat{\rho}$ of Eq. (4) are changed accordingly.

We employ an iterative method of obtaining the estimates of (α, β, ω), say $\left(\hat{\alpha},\hat{\beta},\hat{\omega}\right)$, which maximize (3). This can be done by using the MS function available in S-Plus software. The function requires the determination of initial values α_0, β_0 and ω_0. These initial values correspond to values which give maximum precision parameter $\hat{\rho}$ in (4) for all possible pairs (α, β, ω) in a pre-specified sets. In our case, the following sets of parameter values are considered; $\alpha = [-\pi, \pi]$, $\beta = [-\pi, \pi]$ and $\omega = [-1, 1]$. Then using those initial values, we obtain the estimates iteratively for the three parameters of the model.

COVARIANCE MATRIX OF DM CIRCULAR REGRESSION MODEL

Downs and Mardia (2002) provided the information matrix for DM circular regression model using the log-likelihood function with known parameters $(\alpha, \beta, \omega, \kappa)$:

$$l = const - n\log I_0(\kappa) + \kappa\Sigma\cos(v_i-\mu_i),$$

where

$$\mu_i=\beta+2\tan^{-1}\left\{\omega\tan\frac{1}{2}(u_i-\alpha)\right\}.$$

Using the facts that

$$E\{\cos(v_i-\mu_i)\}=A(\kappa),\ E\{\sin(v_i-\mu_i)]=0,$$

the Fisher information matrix $\boldsymbol{I}=(I_{ij})$ for $\theta T=(\beta,\alpha,\omega,\kappa)=(\theta_1,\theta_2,\theta_3,\theta_4)$. Given that, $I_{14}=I_{24}=I_{34}=0$ so that $\hat{\theta}_1,\hat{\theta}_2$, and $\hat{\theta}_3$ are independent of $\hat{\theta}_4$ as expected, asymptotically, we have

$$I=\begin{bmatrix} C_{11} & 0 \\ 0 & C_{22} \end{bmatrix},$$

where $\boldsymbol{0}^T=(0,0,0)$, C_{11} is 3×3 and C_{22} is a scalar. Then, $C_{22}=I_{44}=nA'(\kappa)$, $C_{11}=\kappa A(\kappa)\boldsymbol{B}(\theta_1,\theta_2,\theta_3)$ where the elements of the matrix $\boldsymbol{B}(\beta,\alpha,\omega)$ are

$$\begin{aligned} &b_{11}=n, \qquad b_{12}=\sum(\mu_i), \qquad b_{13}=\sum(\mu_i)_\omega \\ &b_{22}=\sum(\mu_i)_\alpha^2,\ b_{23}=\sum(\mu_i)_\alpha(\mu_i)_\omega,\ b_{33}=\sum(\mu_i)_\omega^2, \end{aligned}$$

with

$$(\mu_i)_\omega=\frac{2\tan\frac{1}{2}(u_i-\alpha)}{1+\omega^2\tan^2\frac{1}{2}(u_i-\alpha)}$$

and

$$(\mu_i)_\alpha=-\frac{\omega\sec^2\frac{1}{2}(u_i-\alpha)}{1+\omega^2\tan^2\frac{1}{2}(u_i-\alpha)}.$$

Thus, the covariance matrix is given by

$$\begin{aligned} \operatorname{cov}\left(\hat{\beta},\hat{\alpha},\hat{\omega}\right) &= \left\{\hat{\kappa}A\left(\hat{\kappa}\right)\right\}^{-1}\left\{\boldsymbol{B}\left(\hat{\beta},\hat{\alpha},\hat{\omega}\right)\right\} \\ \operatorname{cov}\left(\hat{\beta},\hat{\kappa}\right) &= \operatorname{cov}\left(\hat{\alpha},\hat{\kappa}\right) \\ &= \operatorname{cov}\left(\hat{\omega},\hat{\kappa}\right)=0. \end{aligned} \tag{5}$$

We will use the covariance matrix in developing the outlier detection procedure as described in the next section.

OUTLIER DETECTION PROCEDURE

We develop an outlier detection procedure for DM circular regression using row deletion approach. If an outlier exists in the data, it is expected to affect the parameter of interest of the regression model such as parameter estimates, variance of residuals and covariance matrix. In particular, we look at the effect of removing an observation on the covariance matrix of the model by following work by Belsley et al. (1980) on the linear case. Let $|COV|$ and $|COV_{(-i)}|$ be the determinant of the covariance matrix for the full data set and the reduced data set after removing the ith row, respectively. We then define the *COVRATIO* statistic as

$$COVRATIO_{(-i)} = \frac{|COV_{(-i)}|}{|COV|}. \quad (8)$$

If the ratio is not close to one, then the ith observation is a candidate of being an outlier. In this paper, we use the test statistic $|COVRATIO_{(-i)}|$ with the covariance matrix (5) to identify outliers in the DM circular regression model. The critical values of the test were obtained via simulation in the following section.

CRITICAL VALUES OF THE TEST STATISTIC

We perform a simulation study to investigate the sampling behavior of the test statistic $|COVRATIO_{(-i)}|$. Sets of circular random errors are generated with mean direction $\mu = 0$ and different values of concentration parameter $\kappa = 5$, 10, 30 and 50 from the von Mises distribution. Then, we generate the values of the independent circular variable μ 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 with fixed values of $\alpha = 1.5$, $\beta = 1.5$, and $\omega = 0.5$.

Upon fitting the DM regression model on the full and reduced simulated data set, we obtain the value of $|COVRATIO_{(-i)}|$ for $i = 1, 2, \ldots, n$ and consequently the maximum value of $|COVRATIO_{(-i)}|$. The process is carried out 500 times for each combination of sample size and concentration parameter. We then computed the 1, 5 and 10% upper percentiles of the maximum values of $|COVRATIO_{(-i)}|$ which will be considered as the critical values of the test statistic.

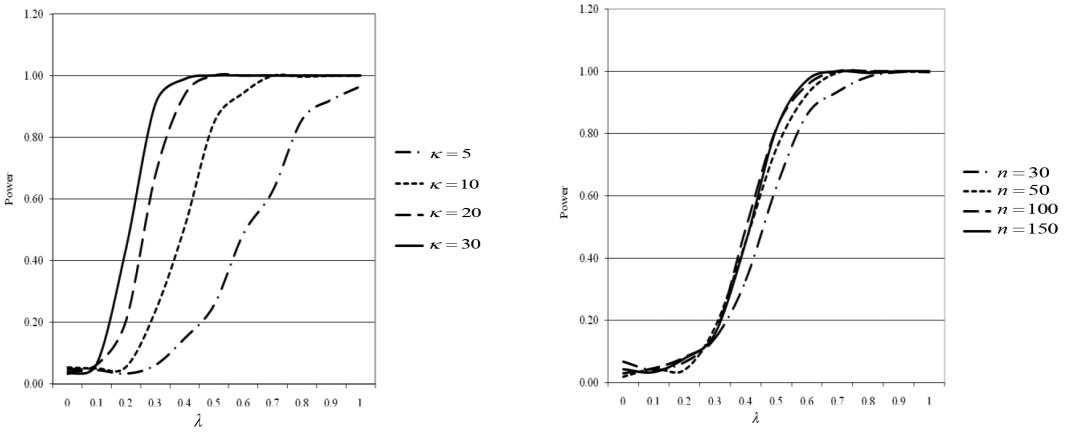
In general, for all κ and percentile levels considered, the critical values get smaller as n gets larger. On the other hand, there is no consistent pattern of the critical values observed as k increases, though they achieve their maximum when $\kappa = 3$ to 5. Note that the critical values described above are only for the case when $\omega = 0.5$. In fact, the critical values do not depend on the parameter values α and β, but depend on ω. When ω gets closer to 1, the critical values get smaller. However, when ω gets closer to 0, the $|COVRATIO_{(-i)}|$ statistic fails to give reasonable set of critical values. This failure occurred due to the DM model property which is the only model that has a unique solution as given in (2) only if the value of $\omega \neq 0$. The critical values can be obtained from the authors upon request.

POWER OF PERFORMANCE OF TEST STATISTIC

To investigate the power of performance of test statistic, several sample sizes are considered. We generate the data using similar steps employed in the previous section. We contaminated the d-th observation v_d such that

$$v_d^* = v_d + \lambda\pi \mod(2\pi),$$

where v_d^* is the contaminated observation at position d and λ is the contamination level, $0 \leq \lambda \leq 1$. The generated data are fitted using (1). Then, we calculate the maximum value of test statistic for each simulated data set. The process is repeated 500 times. By comparing the values with the corresponding critival values, we calculate the percentage of correct detection of the contaminated observation at position d.



(a) (b)

FIGURE 1. Power of performance of $|COVRATIO_{(-i)} - 1|$ statistics, for (a) n=70, (b) κ =10

Figure 1(a) gives the plot of power of performance of the procedure for $n = 70$ and various value of κ. It can be seen that the power function is an increasing function of the concentration parameter κ. This is expected as the data is more concentrated for larger concentration and more dispersed around the unit circle for smaller concentration. On the other hand, Figure 1(b) gives the plot of power of performance of the procedure 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 than the others for $n = 30$. Similar results are observed for the other cases.

APPLICATION

Here we consider the ocean wind direction data obtained from Hussin et al. (2004). The data is the direction of the local wind which blows across the sea surface and along the coast where the HF radar system and anchored wave buoy are deployed. There were a total of 129 measurements recorded by both instruments. Several plots can be used to show the distributions of both measurements. In general, from Figures 2 and 3, both sets of measurement follow the same distribution. It can be seen that there is a high frequency in the second quadrant for both sets of measurements. From Figure 4, there are two points located at the top of the Q-Q plot. These points might correspond to observations which are candidates to be outliers. Some of the descriptive statistics for the ocean wind direction data are given in Table 1. The summary statistics of the HF radar and anchored wave buoy are almost similar including the concentration parameter with the value less than one.

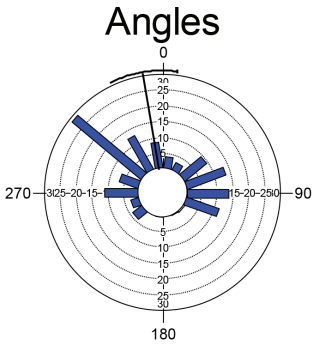


FIGURE 2. Circular histogram for HF

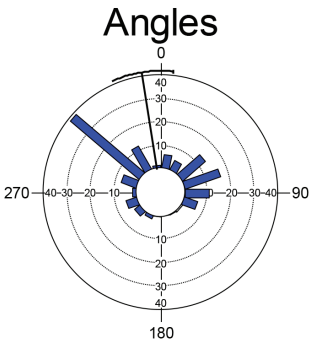


FIGURE 3. Circular histogram for AB

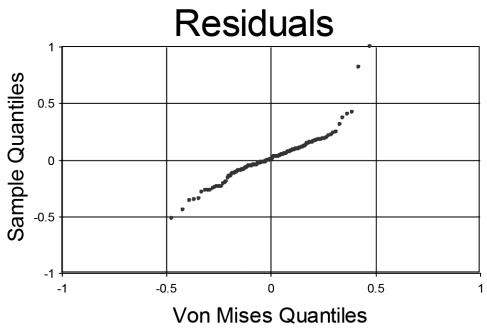


FIGURE 4. Q-Q plot for residuals

Using the data set, we calculate the precision parameters in the pre-specified sets as described in previous. The initial values of each parameter correspond to the highest point observed in (4) giving $\alpha_o = 126°$, $\beta_o = 126°$ and $\omega_o = 0.9$. Thus, using these initial values, the final estimated parameter values are obtained by maximizing the log likelihood function given by equation (5); $\hat{\alpha} = 65.39°$, $\hat{\beta} = 71.82°$ and $\hat{\omega} = 0.91$.

TABLE 1. Descriptive statistics for ocean wind direction data

Variable	HF	AB
Mean Direction	350.43°	351.06°
Mean Resultant Length	0.41	0.44
Circular Std Dev	76.374°	73°
Median Direction	334.72°	327.33°
Concentration parameter	0.902	0.99

Figure 5 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 $|COVRATIO_{(-i)}|$ statistics we identify that observations 38 and 111 are candidates for influential observations. This can be further verified by looking at Figure 6 whereby there are two observations with high value of $|COVRATIO_{(-i)}|$ denoted by p. Here, we have $n = 129$ and κ = 6.84. By considering the critical value corresponding to n = 100 and κ = 7, Table 2 gives the test value and the decision for each observation.

Based on the results, in the first iteration, we identify observations $n = 38$ as influential observation because the test values exceed the critical value 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 6 and 7, respectively. Furthermore, we investigate the effect of these two observations on the parameter estimates. After removing observations 38 and 111 from the data set, we noticed that $\hat{\alpha}$ and $\hat{\beta}$ decrease by a large value which is $\hat{\alpha} = 31.20°$, $\hat{\beta} = 35.57°$ and $\hat{\omega} = 0.94$ as shown in Table 3. 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.

Conclusion

The extension $|COVRATIO_{(-i)}|$ statistic to the model of interest shows consistent. The critical values and the performance of the procedure are obtained via simulation.

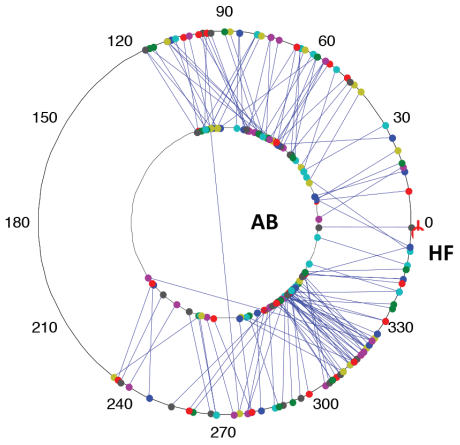


FIGURE 5. Spoke plot of wind data

TABLE 2. Result based on *COVRATIO* statistics

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

TABLE 3. 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 38th observation	65.68°	71.84°	0.91
Without the 111th observation	33.81°	38.42°	0.94
Without both observations	31.20°	35.57°	0.94

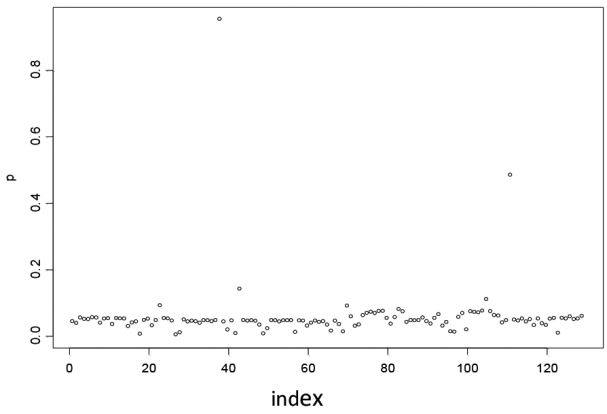


FIGURE 6. Plot of p versus index for 1st iteration

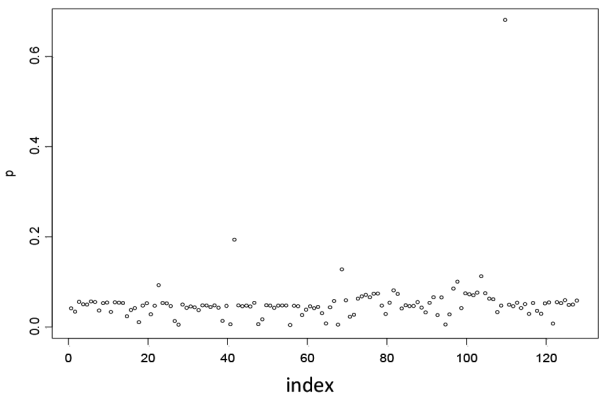


FIGURE 7. Plot of p versus index for 2nd iteration

The statistic shows better performance for large sample size and high concentration parameter. Finally, as an example, the $|COVRATIO_{(-i)}|$ statistic for circular bivariate data has successfully detected two outliers in the ocean wind direction data set which are observations number 8 and 111. This result shows that it is able to identify the presence of outliers in the DM circular regression model.

ACKNOWLEDGEMENTS

We are most grateful and would like to thank the reviewers for their valuable suggestions, which led to an improvement of the article. This research is financially supported by the Research Grant (No: *RP009C-13AFR*) from the University of Malaya.

REFERENCES

Abuzaid, A.H., Hussin, A.G. & Mohamed, I.B. 2013. Detection of outliers in simple regression model using mean circular error statistic. *Journal of Statistical Computation and Simulation* 83(2): 269-277.

Abuzaid, A.H., Mohamed, I.B., Hussin, A.G. & Rambli, A. 2011. Covratio statistic for simple circular regression model. *Chiang Mai J. Sci*. 38(3): 321-330.

Barnett, V. & Lewis, T. 1984. *Outliers in Statistical Data*. New York: John Wiley & Sons.

Beckman, R.J. & Cook, R.D. 1983. Outlier.....s. *Technometrics* 25(2): 119-149.

Belsley, D.A., Kuh, E. & Welsch, R.E. 1980. *Regression Diagnostic: Identifying Influential Data and Sources of Collinearity*. New York: John Wiley & Sons.

Downs, T.D. & Mardia, K.V. 2002. Circular regression. *Biometrika* 89(3): 683-697.

Hussin, A.G., Abuzaid, A.H., Zulkifli, F. & Mohamed, I.B. 2010. Asymptotic covariance and detection of influential observation in a linear functional relationship model for circular data with an application to the measurements of winds directions. *SCIENCEASIA* 36: 249-253.

Hussin, A.G., Fieller, N.R.J. & Stillman, E.C. 2004. Linear regression for circular variables with application to directional data. *Journal of Applied Science and Technology* 8: 1-6.

Ibrahim, S., Rambli, A., Hussin, A.G. & Mohamed, I. 2013. Outlier detection in a circular regression model using COVRATIO statistic. *Communications in Statistics-Simulation and Computation* 42(10): 2272-2280.

Jammalamadaka, S.R. & Sarma, Y.R. 1993. Circular regression. In *Statistical Sciences and Data Analysis*, edited by Matusita, K. Utrecht: VSP. pp. 109-128.

Laycock, P.J. 1975. Optimal regression: Regression models for directions. *Biometrika* 62(168): 305-311.

Rivest, L.P. 1997. A decentred predictor for circular-circular regression. *Biometrika* 84(3): 717-726.

Adzhar Rambli*, Rossita Mohamad Yunus & Ibrahim Mohamed
Institute of Mathematical Sciences
University of Malaya, 59100 Kuala Lumpur
Malaysia

Abdul Ghapor Hussin
Centre for Defence Foundation Studies
National Defence University of Malaysia
Kem Sungai Besi, 57000 Kuala Lumpur
Malaysia

*Corresponding author; email: adzfranc@yahoo.com

Received: 19 September 2014
Accepted: 6 February 2015