CORE
Metadata, citation and similar papers at core.ac.uk
Provided by UKM Journal Article Repository

Sains Malaysiana 42(6)(2013): 869–874

Detection of Outliers in the Complex Linear Regression Model

(Pengesanan Nilai Tersisih dalam Model Regresi Linear Kompleks)

A.G. Hussin*, A.H. Abuzaid, A.I.N. Ibrahim & A. Rambli

ABSTRACT

The existence of outliers in any type of data affects the estimation of models' parameters. To date there are very few literatures on outlier detection tests in circular regression and it motivated us to propose simple techniques to detect any outliers. This paper considered the complex linear regression model to fit circular data. The complex residuals of complex linear regression model were expressed in two different ways in order to detect possible outliers. Numerical example of the wind direction data was used to illustrate the efficiency of proposed procedures. The results were very much in agreement with the results obtained by using the circular residuals of the simple regression model for circular variables.

Keywords: Circular variables; complex linear regression model; outlier

ABSTRAK

Kewujudan nilai tersisih dalam mana-mana jenis data mempengaruhi anggaran parameter model. Sehingga kini sangat sedikit kajian dijalankan mengenai ujian pengesanan nilai tersisih dalam regresi bulatan dan ini mendorong kami untuk mencadangkan teknik mudah untuk mengesan sebarang nilai tersisih. Kajian ini mempertimbangkan penggunaan model regresi linear kompleks untuk menyuaikan data bulatan. Reja kompleks daripada model regresi linear kompleks dinyatakan dalam dua cara yang berbeza untuk mengesan nilai tersisih yang mungkin. Contoh berangka iaitu data arah angin digunakan untuk menggambarkan kecekapan prosedur yang dicadangkan. Keputusan yang diperoleh amat bersetuju dengan keputusan yang diperoleh dengan menggunakan reja bulatan daripada model regresi mudah untuk pemboleh ubah bulatan.

Kata kunci: Model regresi linear kompleks; nilai tersisih; pemboleh ubah bulatan

Introduction

The bounded close range of circular random variables causes the difficulties of studying the relationship among these variables. There are few forms of circular regression models (Downs & Mardia 2002; Fisher & Lee 1992; Hussin et al. 2004; Kato et al. 2008). Circular data as any other types of data are subjected to contaminate with some unexpected observations which are known as 'outliers' and there are some numerical statistics proposed to identify outliers in univariate circular data (Abuzaid et al. 2009, 2012a, 2012b; Collett 1980; Rambli et al. 2012). Most of these models suffer from the complicity and the absence of the close form of the maximum likelihood estimates. However, up to 2007 there is no relevant literature on the outliers in circular regression. Recently Abuzaid et al. (2008, 2011) discussed the identification of outliers in simple circular regression model based on angular residuals and COVRATIO statistic by using different numerical and graphical tests.

This paper proposed an alternative approach to detect outliers in circular data by using the complex linear regression model (Hussin et al. 2010), where the circular data can be expressed in the form of direction cosines or complex form, as an example, a circular observation θ is written in the form of $\cos\theta + i\sin\theta$. Based on this, the complex residuals may be calculated and written in two different forms i.e. the Cartesians coordinates and polar form to detect any possible outliers. This is another alternative model to fit the circular data especially with the existence of the close form of the models' parameters estimates.

This paper is organized as follows: the following section discusses the simple regression model for circular variables. In the subsequent section, we describe the complex linear regression model and the maximum likelihood estimates of its parameters, followed by the proposed ways to obtain and express the complex residuals. We then compare the identification of outliers in the wind direction data between both considered models.

The Regression Model for Circular Variables

The first attempt to predict a circular response variable Θ from a set of linear covariates have been done by Gould (1969) and Laycock (1975) where Θ has von Mises distribution with mean $E(\Theta) = \mu$ and concentration parameter κ. The proposed model is given by:

$$\mu = \mu_0 + \Sigma\beta_j x_j, \quad (1)$$

where μ_0 and β's are unknown parameters.

Mardia (1972) extended model (1), by assuming that there is a set of circular independent and identically observations $\theta_1, \theta_2, \ldots, \theta_n$, from von Mises distribution with mean directions $\mu_1, \mu_2, \ldots, \mu_n$ and unknown concentration parameter κ, where,

$$\mu_j = \mu_0 + \beta t_j, \tag{2}$$

for some known numbers $t_1, t_2, \ldots, t_n$, while μ_0 and β are unknown parameters.

Hussin et al. (2004), extended model (2) to the case when the response and the explanatory variables are circular, where the t_j s are circular too by taking into consideration that $\theta|t = 0$ and $\theta|t = 2\pi$ should be the same. Suppose for any circular observations $(x_1,y_1), (x_2,y_2), \ldots, (x_n,y_n)$ of circular variables X and Y where the true relationship between X and Y is linear. They proposed the following model known as the simple regression model for circular variables,

$$y_j = \alpha + \beta x_j + \varepsilon_j \,(\mathrm{mod}\, 2\pi) \tag{3}$$

where ε_j is circular random error having a von Mises distribution with mean circular 0 and concentration parameter κ. They make a restriction on the model parameters, especially on β, to be close to the unity. Some of the applications are in the relation between two different instruments for measurements of the wind direction. The log likelihood function for model (3) is given by:

$$\log L(\alpha,\beta,\kappa; x_1,\ldots,x_n, y_1,\ldots,y_n) = -n\log(2\pi) - n\log I_0(\kappa) + \kappa\Sigma\cos(y_1 - \alpha - \beta x_j)$$

where $I_0(\kappa)$ is….

The MLE of the model's parameters α, β and κ are given by:

$$\hat{\alpha} = \begin{cases} \tan^{-1}\left(\frac{S}{C}\right) & S>0, C>0, \\ \tan^{-1}\left(\frac{S}{C}\right)+\pi & C<0, \\ \tan^{-1}\left(\frac{S}{C}\right)+2\pi & S<0, C>0, \end{cases}$$

where $S = \Sigma\sin(y_j - \hat{\beta}x_j)$ and $C = \Sigma\cos(y_j - \hat{\beta}x_j)$. Due to the nonlinear nature of the first partial derivative with respect to β of the log L, parameter β can be estimated iteratively using $\beta_1 \approx \beta_0 + \frac{\sum x_j \sin\left(y_j - \hat{\alpha} - \beta_0 x_j\right)}{\sum x_j^2 \cos\left(y_j - \hat{\alpha} - \beta_0 x_j\right)}$, by giving some initial values of α_0 and β_0. The estimation of concentration parameter is given by $\hat{\kappa} = A^{-1}\left(\frac{1}{n}\sum\cos\left(y_j - \hat{\alpha} - \hat{\beta}x_j\right)\right)$, where $A(\omega)$ is the ratio of the modified Bessel function of the first kind of order one and first kind of order zero. The inverse of function $A(\omega)$ can be estimated by $A^{-1}(\omega) \approx \frac{9 - 8\omega + 3\omega^2}{8(1-\omega)}$ as suggested by Dobson (1978).

Abuzaid et al. (2008) used model (3) and proposed the identification of outliers based on the circular residuals by using different numerical and graphical techniques.

COMPLEX LINEAR REGRESSION MODEL

The complex linear regression model was first introduced by Hussin et al. (2010) to fit circular data. For n observations in two dimensions $(x_1,y_1), (x_2,y_2), \ldots, (x_n,y_n)$ where $0 \le x_j, y_j < 2\pi$ and $j = 1,2,..,n$, the direction cosines can be written as $(\cos x_1 + i\sin x_1, \cos y_1 + i\sin y_1)$, $(\cos x_2 + i\sin x_2, \cos y_2 + i\sin y_2), \ldots, (\cos x_n + i\sin x_n, \cos y_n + i\sin y_n)$.

Hence, the complex linear regression model is given by:

$$\cos y_j + i\sin y_j = \alpha + \beta(\cos x_j + i\sin x_j) + \varepsilon_j, \text{ for } j=1,\ldots,n, \tag{4}$$

where ε_j has a univariate Gaussian complex distribution (Laycock 1975). Thus the expectation of response variable Y is given by:

$$E[\cos(Y_j) + i\sin(Y_j] = \hat{\alpha} + \hat{\beta}(\cos x_j - i\sin x_j),$$

alternatively,

$$\cos \hat{y}_j + i\sin \hat{y}_j = \hat{\alpha} + \hat{\beta}(\cos x_j - i\sin x_j).$$

The log likelihood function is given by:

$$\log L(\alpha,\beta,\sigma^2; x_1, x_2, \ldots, x_n, y_1, y_2, \ldots, y_n) = -n\log\left(\pi\sigma^2\right) \frac{1}{\sigma^2}\sum_j \left(1 + \alpha^2 + \beta^2 + 2\alpha\left(\beta\cos x_j - \cos y_j\right)\right) + \frac{2\beta}{\sigma^2}\sum\left(\cos y_1 \cos x_j + \sin y_j \sin x_j\right).$$

The MLE of model's parameters are given by:

$$\hat{\alpha} = \frac{n\sum\cos y_j - \psi\sum\cos x_j}{n^2 - \left(\sum\cos x_j\right)^2} \quad \hat{\beta} = \frac{1}{n}\left(\psi - \hat{\alpha}\sum\cos x_j\right)$$

and $\hat{\sigma}^2 = \frac{1}{n}\Sigma(1 + \hat{\alpha}^2 + \hat{\beta}^2 + 2\hat{\alpha}(\hat{\beta}\cos x_j - \cos y_j) - 2\hat{\beta}(\cos y_j \cos y_j + \sin y_j \sin x_j))$ where $\Psi = \Sigma(\cos y_j \cos x_j + \sin y_j \cos x_j)$.

Thus, the complex residuals can be calculated as:

$$\varepsilon_j = (\cos \hat{y}_j - \cos y_j) + i(\sin \hat{y}_j - \sin y_j). \tag{5}$$

It is obvious that the obtained residuals of the complex regression models are also a complex random variable and it can be represented by two different ways as follows: On the Cartesian coordinate plane, where the complex residuals can be represented as an order pair of the real

and imaginary part and in the polar form, where each of the complex residuals $\varepsilon_j = (\cos \hat{y}_j - \cos \hat{y}_j) + i(\sin \hat{y}_j - \sin y_j)$ can be written in the form of $\varepsilon_j = R_j(\cos \theta_j + \sin \theta_j)$ where $Rj = \sqrt{\left(\cos \hat{y}_j - \cos y_j\right)^2 + \left(\sin \hat{y}_j - \sin y_j\right)^2}$ is the modulus. θ_j is the angle of the vector ε_j and called an argument of ε_j. It is denoted by:

$$\theta_j = Arg(\varepsilon_j) \text{ or } \theta_j = \tan^{-1}\left\{\frac{\left(\sin \hat{y}_j - \sin y_j\right)}{\left(\cos \hat{y}_j - \cos y_j\right)}\right\},$$

The argument is not unique since cos θ_j and sin θ_j are 2π-periodic. Thus, we consider the principle argument which is unique and represented by $\theta_j = Arg(\varepsilon_j)$ where $-\pi \leq Arg(\varepsilon_j) < \pi$.

The complex residuals will be used in the following section to detect possible outliers and compare the obtained results with recent works by Abuzaid et al. (2008).

NUMERICAL EXAMPLE

This section considers the wind directions data which have been considered by Abuzaid et al. (2008) and Hussin et al. (2004). A total of 129 measurements were recorded over the period of 22.7 days along the Holderness coastline (the Humberside coast of North Sea, United Kingdom) by using two different techniques which are HF radar system and anchored wave buoy. The scatter plot of wind direction data is shown in Figure 1 where the scale is broken artificially at 0 = 2π, which suggests that there is a linear relationship between HF radar system and anchored wave measurements.

Since both variables are circular, model (3) is used to fit the data. Alternatively, we use the complex linear regression model in (4) to fit the same data set. The MLEs of both models parameter and the associated standard error are given in Table 1.

The fitted complex linear regression and the simple linear regression model for circular variables are given in (6) and (7), respectively,

$$\cos\hat{y}_j + i\sin \hat{y}_j = -2.59\times10^{-3} + 0.927(\cos x_j + i\sin x_j). \quad (6)$$

$$\hat{y}_j = 0.165 + 0.973x_i \text{ (mod } 2\pi). \quad (7)$$

By looking at Figure 1 there are two points which are apparent to be outliers at the top left of the scatter plot. However, they are actually consistent with the rest of the observations as they are close to other observations at the top right or left bottom due to the closed range property of the circular variables.

For both models (6) and (7) we obtain the relevant residuals in order to detect possible outliers in the

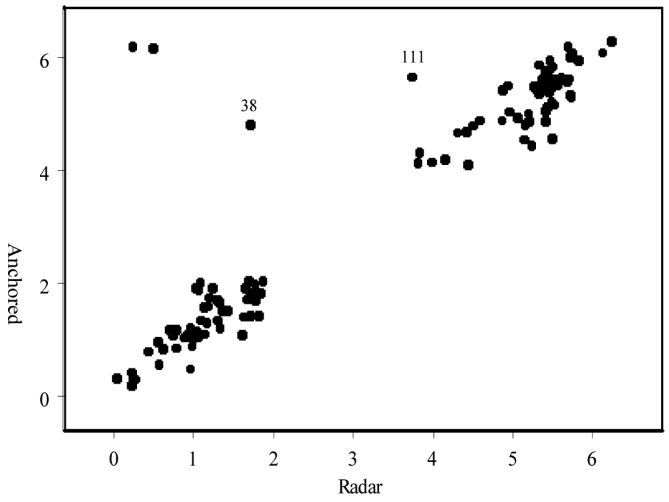

FIGURE 1. Scatter plot of wind data measured by HF radar system and anchored wave buoy

TABLE 1. Parameter estimates for wind direction data

Model	Complex linear regression model		Simple regression for circular variables	
Parameter	Estimate	St. error	Estimate	St. error
α	-2.59×10^{-3}	2.619×10^{-2}	0.165	0.064
β	0.927	2.619×10^{-2}	0.973	0.0159
σ^2	0.143	1.259×10^{-2}	-	-
κ	-	-	7.34	0.8816

regression models. Abuzaid et al. (2008) identified two outliers based on circular residuals by using different numerical and graphical tools, which are observation numbers 38 and 111.

The complex residuals of models (6) are calculated using (5) and may be represented in the following forms:

The Cartesian Coordinates: Figures 2 and 3 shows the complex residuals for the wind direction data. There are two points far from the rest of the observation which are observation numbers 38 and 111. The real part of the complex residuals for observation number 38 is consistent with the rest of the observation while the imaginary is very large. The other way around is noticed for the complex residuals for observation number 111.

The polar form: Figure 4 shows the complex residuals in the polar form and also there are two observations with numbers 38 and 111 which are not consistent with the rest of the observations.

It is obvious that both observation number 38 and 111 are candidate to be outliers. The obtained results are very much agreed with the results obtained by Abuzaid et al. (2008) which indicated the suitability of using the complex residuals to identify possible outliers in the complex linear regression for circular variable as an alternative approach to use the simple circular regression models which are rather complicated.

The removal of observations numbers 38 and 111 affect the estimation of model (3) and (4) parameters as given in Table 2.

It is noticeable that the removal of outlier decreases the variance of residuals for complex linear regression from 0.143 to 0.096 and increases the concentration parameter for simple regression model for circular variables from 7.34 to 11.01. It indicates a goodness of fit of the models after removing the outliers and the obtained residuals for both models are relatively consistent.

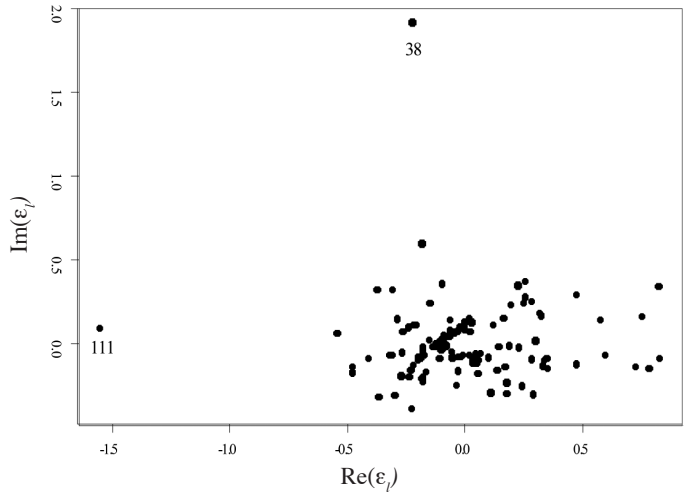

FIGURE 2. The complex residuals of wind direction data on coordinate plane

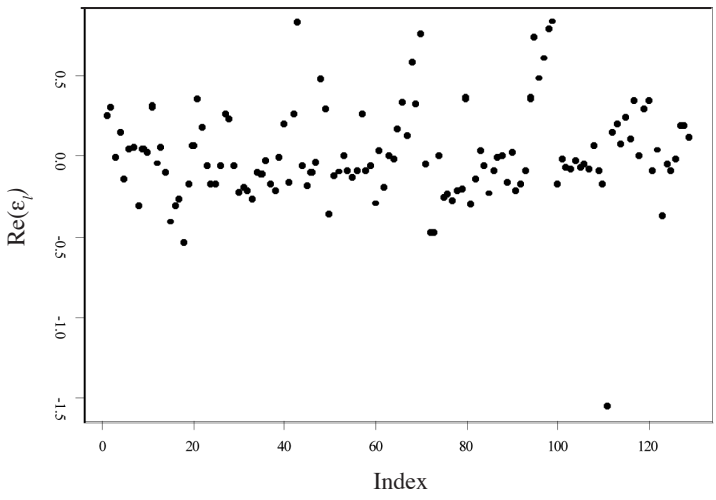

FIGURE 3. The complex residuals component for each observation
(a) The real part of the complex residuals for each observation

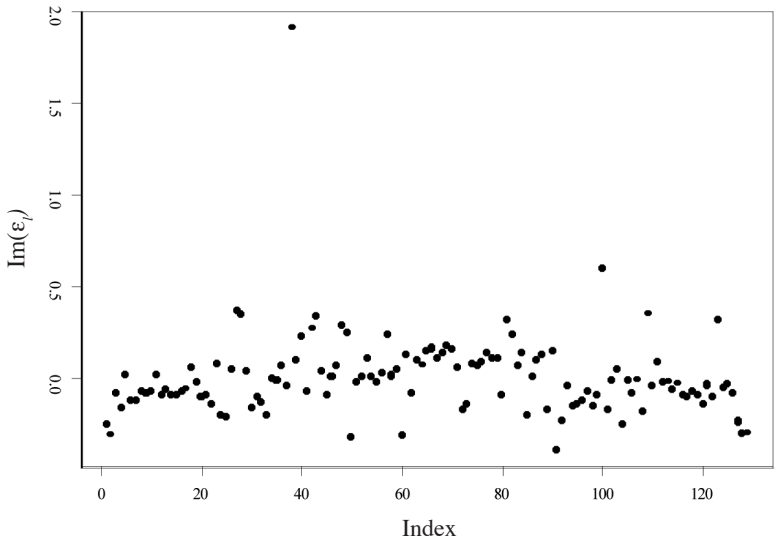

FIGURE 3(b). The imaginary part of the complex residuals for each observation

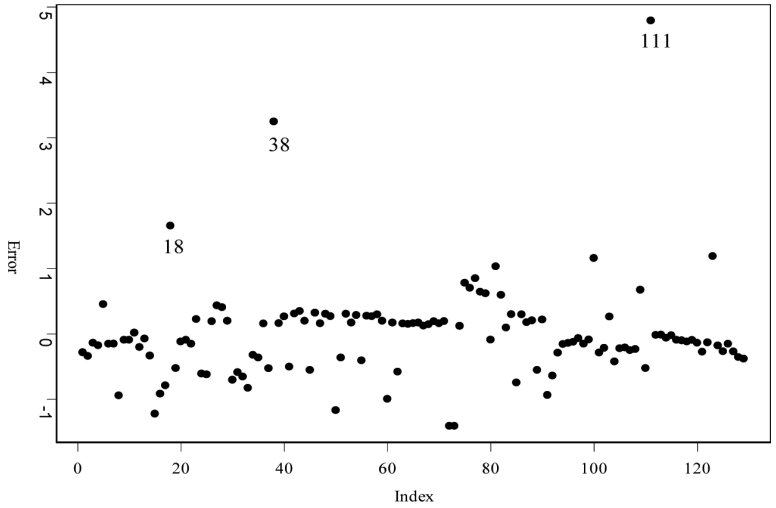

FIGURE 4. The complex residuals of wind direction data on polar form

TABLE 2. Parameter estimates for wind direction data excluding observation number 38 and 111

Model	Complex linear regression model		Simple regression for circular variables	
Parameter	Estimate	St. error	Estimate	St. error
α	-0.032	2.619×10^{-2}	0.153	0.0525
β	0.965	2.619×10^{-2}	0.974	0.0129
σ^2	0.096	1.259×10^{-2}	-	-
κ	-	-	11.010	1.3497

CONCLUSIONS

This paper proposed the complex linear regression model to fit the circular data and using the complex residuals to identify any possible outliers. It is noticeable that, for complex residuals if either the real or imaginary part of any residuals is inconsistent with the other residuals values, then these residuals are candidates to be outliers. The obtained results also agreed with the conclusions as given by Abuzaid et al. (2008).

REFERENCES

Abuzaid, A.H., Hussin, A.G. & Mohamed, I.B. 2008. Identifying single outlier in linear circular regression model based on circular distance. *Journal of Applied Probability and Statistics* 3(1): 107-117.

Abuzaid, A.H., Mohamed, I.B. & Hussin, A.G. 2009. A new test of discordancy in circular data. *Communications in Statistics - Simulation and Computation* 38(4): 682-691.

Abuzaid, A.H., Mohamed, I., Hussin, A.G. & Rambli, A. 2011. COVRATIO statistic for simple circular regression model. *Chiang Mai International Journal of Science and Technology* 38(3): 321-330.

Abuzaid, A.H., Mohamed, I. & Hussin, A.G. 2012a. Circular Boxplot. *Computational Statistics* 27(3): 381-392.

Abuzaid, A.H., Hussin, A.G., Rambli, A. & Mohamed, I.B. 2012b. Statistics for a new test of discordance in circular data. *Communication in Statistics: Simulation and Computation* 41(10): 1882-1890.

Collett, D. 1980. Outliers in circular data. *Applied Statistics* 29(1): 50-57.

Dobson, A.J. 1978. Simple approximation for the concentration parameters for the von Mises concentration parameter. *Applied Statistics* 27: 345-347.

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

Fisher, N.I. & Lee, A.J. 1992. Regression models for an angular response. *Biometrics* 48: 665-677.

Gould, A.L. 1969. A regression technique for angular response. *Biometrics* 25: 683-700.

Hussin, A.G., Abdullah, N.A. & Mohamed, I. 2010. A complex linear regression model. *Sains Malaysiana* 39(3): 491-494.

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 & Technology* 8(1 & 2): 1-6.

Kato, S., Shimizu, K. & Shieh, G.S. 2008. A circular-circular regression model. *Statistica Sinica* 18: 633-643.

Laycock, P.J. 1975. Optimal design: Regression model for directions. *Biometrika* 62: 305-311.

Mardia, K.V. 1972. *Statistics of Directional Data*. London: Academic Press.

Rambli, A., Ibrahim, S., Abdullah, M.I., Mohamed, I. & Hussin, A.G. 2012. On discordance test for the wrapped normal data. *Sains Malaysiana* 41(6): 769-778.

Abdul Ghapor Hussin*
Faculty of Science and Defence Technology
National Defence University of Malaysia
57000 Kuala Lumpur
Malaysia

Ali H.M. Abu Zaid
Faculty of Science
Al-Azhar University-Gaza
Palestine

Adriana Irawaty Nur Ibrahim & Adzhar Rambli
Institute of Mathematical Sciences
University of Malaya
50603 Kuala Lumpur
Malaysia

*Corresponding author; email: abdulghapor@gmail.com

Received: 10 August 2012
Accepted: 20 October 2012