Sains Malaysiana 39(3)(2010): 491–494

A Complex Linear Regression Model

(Permodelan Regresi Linear Kompleks)

ABDUL GHAPOR HUSSIN*, NORLI ANIDA ABDULLAH & IBRAHIM MOHAMED

ABSTRACT

This paper gives a comprehensive discussion on complex regression model by extending the idea of regression model to circular variables. Various aspect have been considered such as the biasness of parameters, error assumptions and model checking. The advantage of this approach is that it allows the use of usual technique available in ordinary linear regression for the regression of circular variables. The quality of the estimates and the feasibility of the approach were illustrated via simulation. The model was then applied to the wave direction data.

Keywords: Circular variables; complex linear regression model; Kolmogorov-Smirnov test; ordinary regression model

ABSTRAK

Artikel ini membincangkan model regresi kompleks yang merupakan lanjutan daripada model regresi berarah. Pelbagai aspek dipertimbangkan seperti kepincangan parameter, andaian ralat dan pendiagnosisan kesahihan model tersebut. Kelebihan kaedah ini adalah ianya membenarkan penggunaan teknik sedia ada untuk model regresi linear biasa untuk regresi berarah. Kualiti anggaran parameter dan kebolehan kaedah ini disahkan dengan kajian simulasi. Model ini kemudiannya diterapkan kepada data arah ombak.

Kata kunci: Model linear regresi kompleks; model regresi biasa; ujian Kolmogorov-Smirnov; pembolehubah berarah

INTRODUCTION

The regression problem when both response and explanatory values are circular have not received enough attention. The circular variables are defined as one which takes value on the circumference at a circular, i.e. they are angles in range $[0, 2\pi)$ radian or $[0°, 360°)$. The data set for this circular variable are bounded closed space, for which the concept of origin is arbitrary or undefined. For this reason, the circular random variables must be analyzed by techniques differ from usual approach for linear or real line data set. Fisher and Lee (1992) suggested that when the distribution of these circular variables are not too dispersed, the regression problem can be handled satisfactorily by transforming the data to continuous linear variables. Fisher (1993) pointed out that one of the approaches is by using a linear function of $g(.) = 2\arctan(.)$. On the other hand, Hussin et al. (2004) proposed a simple regression model for circular variables X and Y given $y_i = \alpha\ \beta xi + \varepsilon i$ (mod 2π), where ε_i is circular random error having a von Mises distribution with mean circular 0 and concentration parameter κ. In this paper, we proposed an alternative approach in analyzing the regression of circular variables by transforming the circular data to continuous or real line data set via complex form. By doing so, we can use techniques available for ordinary linear regression model for the circular case.

Certain assumptions need to be satisfied in regression analysis before we can proceed with further work such as forecasting and outlier detection. The regression analysis for circular variables has this kind of assumption too. If the assumption is violated, it may cause distortion in the modeling and consequently affecting the parameter estimation, outlier detection and forecasting. Hence, the need for model checking and assumptions are also indispensable for regression in circular variables. It is of interest to explore some of the properties imposed in this model, such as biasness of parameters, error assumption and model checking.

FORMULATION OF THE MODEL

The two dimensional direction cosines of n observations $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ where $0 \le x_j, y_j < 2\pi$ and $j = 1, 2, \ldots, n$ can be written in complex form denoted by $(\cos x_1 + i \sin x_1, \cos y_1 + i \sin y_1), (\cos x_2 + i \text{ sub } x_2, \cos y_2 + i \sin y_2), \ldots, (\cos x_n + i \sin x_n, \cos y_n + i \sin y_n)$, respectively. Hence, the relationship between the two circular variables can now be described using the complex linear regression model which is given by

$$\cos y_j + i \sin y_j = a + b(\cos x_j + i \sin x_j) + \varepsilon_j, \qquad (1)$$

where ε_j has a bivariate Gaussian complex with mean $\begin{bmatrix}0\\0\end{bmatrix}$ and covariance $\sum = \begin{bmatrix}\sigma^2 & 0\\0 & \sigma^2\end{bmatrix}$, respectively (Goodman 1963). Taking the expectation of (1), we have

$$\cos \hat{y}_j + i \sin \hat{y}_j = \hat{a} + \hat{b}(\cos x_j + i \sin x_j) \quad (2)$$

Following Husin (1997), the log likelihood function is given by $\log L(a, b, \sigma^2; x_1, \ldots, x_n, y_1, \ldots, y_n) = -n \log(\pi\sigma^2) - \frac{1}{\sigma^2}\sum_j \left\{1 + a^2 + b^2 + 2a(\cos x_j - \cos y_j)\right\}$

$$+\frac{1}{\sigma^2}\sum_j \left\{2b\left(\cos y_j \cos x_j + \sin y_j \sin x_j\right)\right\}.$$

Differentiating $\log L$ with respect to parameters $\hat{a}$, $\hat{b}$ and $\hat{\sigma}^2$, we may obtain

$$\hat{a} = \frac{n\sum \cos y_j - \Psi \sum \cos x_j}{n^2 - \left(\sum \cos x_j\right)^2} \quad (3)$$

$$\hat{b} = \frac{1}{n}\left\{\Psi - \hat{a}\sum \cos x_j\right\}, \text{ and} \quad (4)$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum \left\{1 + \hat{a}^2 + \hat{b}^2 + 2\hat{a}(\hat{b}\cos x_j) - 2\hat{b}(\cos y_j \cos x_j + \sin y_j \sin x_j)\right\}, \quad (5)$$

where $\Psi = \sum \left\{\cos y_j \cos x_j + \sin y_j \sin x_j\right\}$. Further, by using Fisher information it can be shown that the variance of parameter $\hat{a}$, $\hat{b}$ and $\hat{\sigma}^2$ are given by

$$\text{var}(\hat{a}) = \text{var}(\hat{b}) = \frac{n\sigma^2}{2n^2 - 2\left(\sum \cos x_j\right)^2}, \quad (6)$$

and

$$\text{var}\left(\hat{\sigma}^2\right) = \frac{\sigma^4}{n}. \quad (7)$$

From (1), the complex error or residual can be divided into two parts which are real and imaginary. The real part is given as

$$\varepsilon_j^{\text{Re}} = \cos y_j - \hat{a} - \hat{b} \cos x_j, \quad (8)$$

and the imaginary part is

$$\varepsilon_j^{\text{Im}} = \sin y_j - \hat{b} \sin x_j. \quad (9)$$

Consequently, the normality of both real and imaginary residual can be tested using the Kolmogorov-Smirnov test, that is, we are testing for

$$H_0 : \varepsilon_j \sim \text{normal distribution.} \quad (10)$$

Other properties that will be considered in the next simulation section are the biasness of the parameters $\hat{a}$, $\hat{b}$ and $\hat{\sigma}^2$ of the model in (1).

Simulation Study

A simulation study is conducted to examine the normality of the error term in (8) and (9) as well as the biasness of MLE estimates $\hat{a}$, $\hat{b}$ and $\hat{\sigma}^2$. We provide detailed steps to investigate the normality of the error term as follows:

1. Generate an independent von Mises distribution for variable x_j of length n with mean $\mu = \frac{\pi}{4}$ and parameter concentration $\kappa = 3$.
2. Generate another independent von Mises distribution for variable ε_j (error of the dependent variable y_j) with mean $\mu = 0$ and parameter concentration κ.
3. The dependent variable

$$y_j = x_j + \varepsilon_j \quad (11)$$

is obtained by using the generated value in steps 1 and 2, where $0 \le y_j < 2\pi$. Note that from (1) and (11), without loss of generality, we may choose the true value for a as 0 and b as 1, respectively.
4. Estimate the parameters $\hat{a}$, $\hat{b}$ and $\hat{\sigma}^2$, followed by the Kolmogorov-Smirnov normality test for $\varepsilon_j^{\text{Re}}$ and $\varepsilon_j^{\text{Im}}$.
5. Repeat the above steps for 1,000 times. The following calculation were obtained:
   (a) Estimated bias of $a = \bar{\hat{a}} - 0$, where $\bar{\hat{a}} = \frac{1}{s}\sum_{s=1}^{1,000} \hat{a}_s$, since the true value of parameter a is equal to 0.
   (b) Estimated bias of $b = \bar{\hat{b}} - 1$, where $\bar{\hat{b}} = \frac{1}{s}\sum_{s=1}^{1,000} \hat{b}_s$, since the true value of parameter b is equal to 1.
   (c) Estimated mean, $\bar{\hat{\sigma}}^2 = \frac{1}{s}\sum_{s=1}^{1,000} \hat{\sigma}_s^2$.
6. The proportion of null hypothesis acceptance at 10%, 5% and 1% significant levels for $\varepsilon_j^{\text{Re}}$ and $\varepsilon_j^{\text{Im}}$ were calculated, respectively.

Table 1 gives the bias for $\hat{a}$, $\hat{b}$ and $\hat{\sigma}^2$ when we vary the sample size n and parameter concentration κ. Three main points can be observed from the results. Firstly, the estimated bias for a was consistently small for all different sample size n and parameter concentration κ. Secondly, the estimated bias for b decreased as parameter concentration κ increased. Note that in von Mises distribution, a large value of parameter concentration κ corresponds to small dispersion of circular data set, in contrast to normal distribution for linear or real line data set. The estimated mean of $\hat{\sigma}^2$ decreased as parameter concentration κ increased. A large parameter concentration κ signifies a small dispersion of the error ε_j. Thus, the estimated mean of $\hat{\sigma}^2$ is expected to be small if the dispersion of the error is also small.

The performance of the Kolmogorov-Smirnov test for the real and imaginary error parts are presented in Tables 2 and 3, respectively. The tables give the proportion of an accepted null hypothesis at three significant levels for different sample size n and parameter concentration κ. For instance, in the third row of Table 2, the proportion of accepted null hypothesis at 10% significant level is 0.998 when small sample size $n = 30$ is used.

This shows that 9% out of the 1,000 simulated error term follows normal distribution. Similar trend can be seen in both tables. The proportion of accepted null hypothesis increased as sample size n increased. Even at the lowest 1% significant level, almost 100% of the error terms follow normal distribution. Another noticeable feature is

TABLE 1. Simulation results for estimated bias of a, estimated bias of b and estimated mean of $\hat{\sigma}^2$

	Estimated bias of a			Estimated bias of b			Estimated bias of $\hat{\sigma}^2$		
	$n = 30$	$n = 50$	$n = 100$	$n = 30$	$n = 50$	$n = 100$	$n = 30$	$n = 50$	$n = 100$
$\kappa = 15$	-0.0002	-0.0012	-0.0006	-0.0341	-0.0334	-0.0333	0.0659	0.0661	0.0657
$\kappa = 30$	-0.0008	0.0004	0.0008	-0.0161	-0.0170	-0.0173	0.0323	0.0326	0.0331
$\kappa = 50$	0.0017	-0.0001	0.0001	-0.0111	-0.0100	-0.0101	0.0193	0.0198	0.0199

TABLE 2. Kolmogorov-Smirnov test for error term (real part)

	10%			5%			1%		
	$n = 30$	$n = 50$	$n = 100$	$n = 30$	$n = 50$	$n = 100$	$n = 30$	$n = 50$	$n = 100$
$\kappa = 15$	0.998	1.000	1.000	0.990	1.000	1.000	0.940	0.995	1.000
$\kappa = 15$	1.000	1.000	1.000	1.000	1.000	1.000	0.997	1.000	1.000
$\kappa = 15$	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

TABLE 3. Kolmogorov-Smirnov test for error term (imaginary part)

	10%			5%			1%		
	$n = 30$	$n = 50$	$n = 100$	$n = 30$	$n = 50$	$n = 100$	$n = 30$	$n = 50$	$n = 100$
$\kappa = 15$	0.998	1.000	1.000	0.992	1.000	1.000	0.969	0.998	1.000
$\kappa = 15$	1.000	1.000	1.000	0.999	1.000	1.000	0.995	1.000	1.000
$\kappa = 15$	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

that the proportion increases as parameter concentration κ increased. When $\kappa = 30$, all real and imaginary error parts follow normal distribution even though the sample size n is small.

By looking at the above results, both real and imaginary error terms follow a Gaussian complex distribution. Thus, simulation study agrees with normality assumption made on the error term in complex linear regression model.

APPLICATION OF THE MODEL

We used the wave direction data to illustrate the complex linear regression model. The data were collected along the Holderness Coastline of the North Sea, United Kingdom by using the radar and anchored wave buoy. There were 56 measurements in radian recorded over the period of 2 months (January and February 1994).

Our aim was to find a relationship between the radar (x) and anchored wave buoy (y) in measuring the wave direction. We fit the data to the model in (2). Table 4 gives the parameter estimation for the wave direction data together with their expected standard error. The estimated relationship for wave direction data is given as

$$\cos \hat{y} + i \sin \hat{y} = 0.08442 + 0.9789(\cos x + i \sin x). \qquad (12)$$

It is clear that the value of $\hat{a}$ and $\hat{b}$ are close to 0 and 1, respectively and the estimate of $\hat{\sigma}^2$ is 0.0351. This indicates that the direction taken using radar and anchor wave does not differ much. Further, the estimation for the three parameters appear to be good as they have small standard errors of 0.01772 for $\hat{a}$ and $\hat{b}$, and 0.00016 for $\hat{\sigma}^2$.

TABLE 4. Parameter estimation for wave direction data

Parameter	Estimate	Standard Error
a	0.08442	0.01772
b	0.9789	0.01772
$\hat{\sigma}^2$	0.0351	0.00016

Further, model checking is performed to verify the error assumption in (10). The quantile-quantile normal plot for real and imaginary error parts are given in Figures 1 and 2 respectively. Both plots suggest that the errors follow normal distribution. Hence, the assumptions are satisfied and the model in (12) is the regression model which best represent the relationship between wave direction that has been measured by radar and anchored wave buoy.

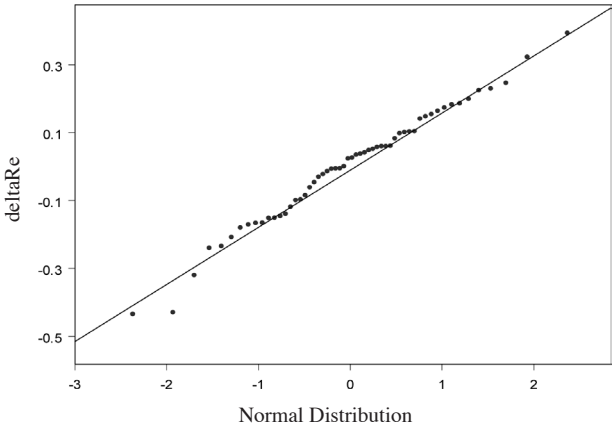

FIGURE 1. Quantile-quantile normal plot for the real residual part

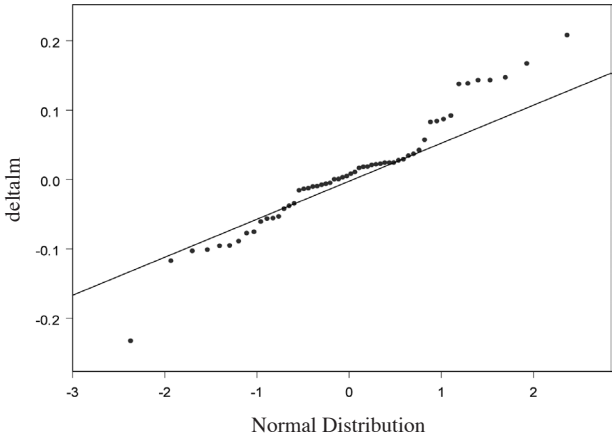

FIGURE 2. Quantile-quantile normal plot for the imaginary residual part

CONCLUSION

This paper suggests a regression of circular variables via complex form which allow us to use the available and usual techniques as in ordinary linear regression for regression of circular variables. It is shown that this method is simple and also give the closed-form expression for the parameter estimates.

Based on the simulation studies, we conclude that the proposed approach give a feasible and effective estimation of parameters and the imposed assumption. These are useful in particular for comparison studies involving circular data. We applied the proposed model to the problem of assessing radar measurements of wave data using two different techniques. The parameter estimation and model checking suggests a good picture at the quality of the complex linear regression model involving circular data.

REFERENCES

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

Fisher, N.I. 1993. *Statistical Analysis of Circular Data*. London: Cambridge University Press.

Goodman, N.R. 1963. Statistical analysis based on a certain multivariate complex Gaussian distribution (an introduction). *The Annals of Mathematical Statistics* 34: 152-165.

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 & 2): 1-6.

Abdul Ghapor Hussin*
Center for Foundation Studies in Sciences, Universiti Malaya
50603 Kuala Lumpur
Malaysia

Norli Anida Abdullah & Ibrahim Mohamed
Institute of Mathematical Sciences
Universiti Malaya
50603 Kuala Lumpur
Malaysia

*Corresponding author; email: ghapor@um.edu.my

Received: 16 June 2009
Accepted: 10 September 2009