International Journal of Scientific Research in Mathematical and Statistical Sciences

Research Paper

Volume-5, Issue-4, pp.95-99, August (2018)

E-ISSN: 2348-4519

Regression Analysis involving Circular Response and Circular Explanatory Variable

S. Bhattacharjee[1*], K.K. Das[2]

[1,2] Department of Statistics, Gauhati University, Guwahati-781014, India

Corresponding Author: sahana.bhattacharjee@hotmail.com, Tel.: +91-95086-60795

Available online at: www.isroset.org

Accepted 15/Aug/2018, Online 30/Aug/2018

***Abstract*—** In this paper, two regression models predicting a circular response from a circular predictor are discussed and a new measure for model comparison is introduced. A circular observation is the one which arises in terms of angles. In the first model, the expected value of the response is modeled in terms of Fourier series expansion of the circular predictor whereas the second model consists in minimizing the least circular distance (LCD) between the actual and predicted values of the response. The application of the regression models is exhibited through a data set on wind directions, measured during morning and evening at Silchar Meteorological observatory, Assam, in the Post Monsoon season, wherein the wind direction measured during the evening is modeled as a function of the wind direction measured during the morning. It is found that the wind direction during the evening is not changing significantly with respect to that during the morning. Also, the proposed measure for model comparison shows that the conditional regression model is a better fitting comparison to the LCD regression model.

Keywords *— Regression, circular response, circular predictor, model comparison*

I. INTRODUCTION

A researcher may often come across bivariate (or multivariate) data where it becomes important to study the inter-relationship between some or all the variables involved in the data set. This purpose can be served through the correlation and regression analysis of the data. Often, there arise situations where both the explanatory variable and the response variable are circular in nature. For example, a biologist may be interested in observing the direction from which the birds come as well as the direction of their return and study for any possible association between them or model the direction of their return as a function of the direction of their arrival [1]. Here, both the directions are measured in terms of angles and so, they are circular random variables (r.v). A circular r.v is the one which arise in terms of angles, is periodic in nature and takes values in the interval (0, 2π). A circular observation is represented as a point on the circumference of the unit circle [2]. Since the mathematical treatment required for a circular r.v is different from that required for a linear r.v [3], the regression model for formulating a circular response as a function of a circular explanatory variable is different from the usual regression models involving linear response and covariates. By linear r.v, is meant the variable which takes values on the real line or any subset of it. For example, [4] applied a regression model to predict the direction of earthquake displacement in terms of the direction of steepest descent. Also, [5] expressed the regression curve in the form of a Möbius transformation and applied it to data on wind direction and circadian rhythms. [6] regressed a circular response on circular predictor by expressing the regression curve as a form of Möbius circle transformation, where the angular error was Wrapped Cauchy distributed.

In this paper, two regression models involving a circular response variable and a circular predictor, termed as circular-circular regression model, are discussed and a statistical measure for comparing the goodness-of-fit of the two regression models is proposed. Finally, the two models are applied to a data set of wind directions measured during morning and evening at a meteorological station (observatory) located in Assam, India. The paper is organized as follows- Section I contains the introduction to circular random variable, circular regression models and a brief review of the literature on circular regression; Section II gives an account of the circular-circular regression models, the newly proposed model comparison measure and other related measures, data set used for showing application; Section III contains the results obtained on the basis of the data set considered and the discussion of the results; Section IV concludes the research work done in the paper with mention of the limitations and future scope of the same.

I. MATERIALS AND METHODS

This section gives an account of the two regression models for modeling a circular response as a function of a circular explanatory variable (i.e. circular-circular regression models) and introduces a statistical measure for comparing the goodness-of-fit of the two models.

II.1 Circular-Circular regression based on the conditional expectation of one circular variable given another

In this regression model [7], the first trigonometric moment of the circular response variable, say β, is conditional on the circular explanatory variable, say α. The conditional expectation of the vector $(\cos\beta + i\sin\beta)$given α is

$$E\{(\cos\beta + i\sin\beta) \mid \alpha\} = E(\cos\beta \mid \alpha) + iE(\sin\beta \mid \alpha)$$

$$= g_1(\alpha) + ig_2(\alpha) = \rho(\alpha)e^{i\mu(\alpha)} \qquad (1)$$

Evidently, $\mu(\alpha)$ represents the mean direction of $(\cos\beta + i\sin\beta)$ given α and $\rho(\alpha)$ denotes its concentration towards $\mu(\alpha)$ [1]. Therefore,

$$\mu(\alpha) = \hat{\beta} = \arctan^{*}\left(\frac{g_2(\alpha)}{g_1(\alpha)}\right) \qquad (2)$$

Where *arctan** *(.)* represents the quadrant specific inverse of the tangent.

Estimation of β requires estimation of $g_1(\alpha)$ and $g_2(\alpha)$ from the data. In the absence of specifications about their structure, they are approximated by their Fourier series expansions. Thus, the estimates of $g_1(\alpha)$ and $g_2(\alpha)$ are

$$\left.\begin{aligned} g_1(\alpha) &\approx \sum_{k=0}^{m}(A_k\cos k\alpha + B_k\sin k\alpha) \\ g_2(\alpha) &\approx \sum_{k=0}^{m}(C_k\cos k\alpha + D_k\sin k\alpha) \end{aligned}\right\} \qquad (3)$$

Or

$$\left.\begin{aligned} \cos\beta &\approx \sum_{k=0}^{m}(A_k\cos k\alpha + B_k\sin k\alpha) + \varepsilon_1 \\ \sin\beta &\approx \sum_{k=0}^{m}(C_k\cos k\alpha + D_k\sin k\alpha) + \varepsilon_2 \end{aligned}\right\} \qquad (4)$$

where $(\varepsilon_1, \varepsilon_2)$ is the error vector with zero mean and unknown covariance matrix. The unknown constants $(A_k, B_k, C_k, D_k), k = 0,1,2,...,m$ are estimated using the method of least squares.

From (4), it is clear that $\cos\beta$ and $\sin\beta$ are approximated by trigonometric polynomials of degree *m* and so, determination of *m* becomes important. If the reduction in error when the $(m+1)th$ degree polynomial is used is found to be significant, the $(m+1)th$ term is included in both the approximations. Otherwise, the m^{th} order model is fitted only if neither of the above two models require the $(m+1)th$ term [3]. The test of the significance of the reduction of error with inclusion of the $(m+1)th$ term for large values of *n* (number of observations) can be found in [1].

II.2 Least Circular Distance (LCD) regression: Rotational model with no covariates involved

This regression model is due to Lund (1999) and is used to model situations when both the response and the predictor is circular and also, there is a set of linear covariates. Least square regression, if applied to circular response, can yield erroneous results. So, a regression technique based on an alternative measure of distance for a circular variable, called the "Circular Distance'' is introduced, which, for two circular variables α and β, is defined as

$$D(\alpha, \beta) = \frac{1}{2}[1 - \cos(\alpha - \beta)] \qquad (5)$$

The regression technique carried out by minimizing $D(\alpha, \beta)$ is termed as the Least Circular Distance (LCD) regression and is the circular analog to the least square regression for linear response and linear covariates. The LCD regression model for circular response θ, circular predictor ϕ and a set of linear covariates x is defined as

$$\theta = \mu(\phi, x, \beta_1, \beta_2) + e \qquad (6)$$

where β_1, β_2 are vectors of parameters and *e* is the random circular error with zero mean direction. If there are *n* observations on θ, ϕ and *x* and the relationship in (6) holds good, the estimates of β_1 and β_2 are obtained so as to minimize

$$\sum_{i=1}^{n}[D(\theta_i, \mu(\phi_i, x_i, \beta_1, \beta_2))]$$

where $D(.,.)$is as defined in (5). The estimates thus obtained are called the Least Circular Distance estimates of β_1 andβ_2. Assuming that ϕ and *x* have additive effects on θ, [8] assigned the following form to μ:

$$\mu(\phi_i, x_i, \beta_1, \beta_2) = g_1(\phi_i, \beta_1) + g_2(\beta_2 x_i) \qquad (7)$$

$g_2(.)$ is the link function that maps the real line onto the unit circle. In the absence of any prior information, a general form of $g_1(.)$ is obtained by observing that it is a periodic function of ϕ, with period 2π. This leads to the infinite Fourier series representation of $g_1(.)$ identical to that in (3).

The order *m* of the expansion is decided upon by the same procedure as in the previous regression model.

Since linear covariates are absent in this research work, $g_2(\beta_2 x_i)=0$, and thus, the model reduces to the conditional circular-circular regression model discussed in section (II.1). Hence, the conditional circular-circular regression model can be thought of as the particular case of the LCD regression model, when prior information about the relationship between θ and ϕ is unknown and covariates are absent.

Assuming that there exists a rotational relationship between θ and ϕ [9], i.e. a relationship of the form $\theta \pm \phi = \theta_0 \bmod(2\pi)$, $g_1(.)$ is assigned the form $g_1(\phi_i,\beta_1)=\phi_i+\beta_1$. The rotational dependence between θ and ϕ can be checked by testing the significance of the sample circular-circular correlation coefficient $r_{c,n}$ defined by [10] as

$$r_{c,n}=\frac{\sum_{i=1}^{n}\sin(\theta_i-\bar{\theta})\sin(\phi_i-\bar{\phi})}{\sqrt{\sum_{i=1}^{n}\sin^2(\theta_i-\bar{\theta})\sum_{i=1}^{n}\sin^2(\phi_i-\bar{\phi})}}$$

where $\bar{\theta}$ and $\bar{\phi}$ are the sample mean directions of θ and ϕ respectively. Under the hypothesis H_0: ρ_0 (Population circular-circular correlation coefficient) =0, $r_{c,n}$ is asymptotically normally distributed.

Thus, the LCD regression model reduces to

$$\theta_i=\phi_i+\beta_1+e_i;i=1,2,...,n \tag{8}$$

The LCD estimate of β_1 is given by

$$\hat{\beta}_1=\frac{\sum_{i=1}^{n}\sin(\theta_i-\phi_i)}{\sum_{i=1}^{n}\cos(\theta_i-\phi_i)} \tag{9}$$

This renders the estimated i^{th} value of the response variable θ as

$$\hat{\theta}_i=\phi_i+\frac{\sum_{i=1}^{n}\sin(\theta_i-\phi_i)}{\sum_{i=1}^{n}\cos(\theta_i-\phi_i)};i=1,2,...,n \tag{10}$$

II.3 Goodness-of-fit measure

Once each of the above mentioned circular-circular regression models is fitted to the data, the next step consists in comparing the goodness-of-fit of the two models. In other words, the performance of the two models in modeling the data is required to be measured and compared. A statistical measure for assessing the goodness-of-fit of the two models is proposed in the next subsection.

II.3.1 Mean Circular Distance (MCD)

The Mean Circular Distance (MCD) between a set of points is defined as the simple arithmetic mean of the circular distance between every pair of observed and estimated value of the response. The MCD between the set of *n* values of the response variable *θ*, say $\theta_1,\theta_2,...,\theta_n$ and their estimated values $\hat{\theta}_1,\hat{\theta}_2,...,\hat{\theta}_n$ obtained by fitting the circular-circular regression model is given by

$$MCD=\frac{1}{2n}\sum_{i=1}^{n}\left[1-\cos(\theta_i-\hat{\theta}_i)\right]$$

Clearly, $\theta\in[0,1]$. The closer the value of MCD is to 0, higher is the agreement between the observed and the estimated values of the response variable and vice versa.

II.4 Data

For showing the application of the two regression models, the data on wind direction measured during 08:30 a.m. and 05:30 p.m. in the post-monsoon season (October – December) for the years 2012 and 2013 at the Silchar Meteorological Observatory, located in Assam, India is considered. Here, the wind direction measured at 05:30 p.m. is analyzed as a function of the wind direction measured at 08:30 a.m. The data is procured from the Regional Meteorological Centre, Guwahati. The analysis is carried out using the *R* software, version 3.4.0, with the help of the user contributed packages *CircStats* [11] and *circular* [12].

I. Results and Discussion

This section contains the results of the analysis of the data set on wind directions considered for study.

Table 1 shows the estimated coefficients of the circular-circular regression model based on the conditional expectation of the wind direction measured during 05:30 p.m. (β) given the wind direction measured at 08:30 a.m. (α) and the *p*-value of the test of the hypothesis

H_0: The third order terms in the polynomial used to approximate cos(β)|α and sin(β)|α is not significantly different from zero

at 5% level of significance.

Table 1: Estimated coefficient of the circular-circular regression model and *p*-value

Coefficient	**cos(β)|α**	**sin(β)|α**
Intercept	0.5360	-0.0513
First Order Terms		
cos(α)	-0.1315	0.0146
sin(α)	0.3142	0.4748
Second Order Terms		
cos(α)	0.3543	0.0559
sin(α)	-0.0331	-0.1485
p-value for third order	0.3253	0.3672

Since both the *p*-values for the third order are > 0.05, we accept H_0 and conclude that the terms up to the second order are to be retained for approximating both cos(β)|α and sin(β)|α.

Thus, the circular-circular model fitted to the data is

$$E\{(\cos\beta + i\sin\beta) \mid \alpha\} = 0.5360 - 0.1315\cos\alpha$$
$$+ 0.3142\sin\alpha + 0.3543\cos 2\alpha - 0.0331\sin 2\alpha +$$
$$i\begin{pmatrix} -0.0513 + 0.0146\cos\alpha + \\ 0.4748\sin\alpha + 0.0559\cos 2\alpha - 0.1485\sin 2\alpha \end{pmatrix}$$

Now, to see if the wind direction measured during 05:30 p.m. (θ) is rotationally dependent on the wind direction measured at 08:30 a.m. (ϕ), the sample circular-circular correlation coefficient is calculated and the test of its significance is carried out. Here, the calculated value of $r_{c,n} = -0.542$ and the *p*-value of the test of its significance at 5% significance level is 0.00. Therefore, it is evident that there is a rotational relationship between θ and ϕ. Since there is prior information about the rotational dependence of θ on ϕ and no linear covariate in the data considered for study, so, the two regression models are not identical. Consequently, the LCD regression model in (8) is fitted to the data.

The estimated value of the regression coefficient β_1 is found to be -0.4393 and the LCD regression model fitted to the data is

$$\theta = \phi - 0.4393$$

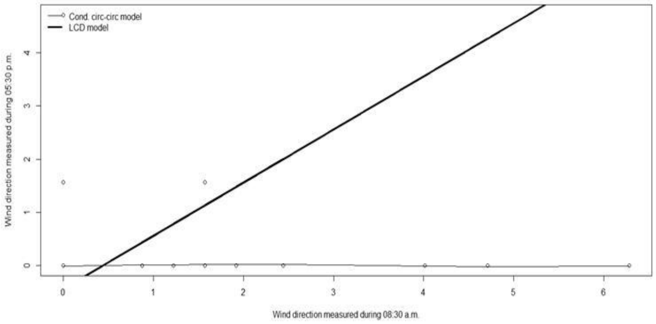
Table 2 contains the MCD between the observed and estimated values for the two regression models.

Table 2: MCD values for the two regression models

Model	**MCD**
Conditional circular-circular regression model	0.0108
LCD regression model	0.1366

It is observed that the conditional circular-circular model has a lesser MCD between the observed and estimated values as compared to the LCD regression model. Therefore, the conditional circular-circular regression model is a better fit to the data under consideration. This is also evident from Figure 1, which shows both the regression models fitted to the data under consideration.

Figure 1: Conditional circular-circular regression model and LCD regression model fitted to the data on wind directions



It is clear from figure 1 that the wind direction measured during the evening is not changing in any specific pattern with a change in the wind direction measured during the morning. Hence, the expectation of the circular response approximated using Fourier series expansion of the circular predictor models the data better than the one obtained by assuming a rotational dependence.

IV. CONCLUSION

This paper discusses two regression models predicting circular response from circular predictor and introduces a new measure called Mean Circular Distance for model comparison. The models are applied to data on wind directions measured during morning and evening collected for two successive years. The wind direction measured during the evening is not seen to change in any specific pattern with change in that measured during the morning. Further, the conditional regression model is a better fit to the data than the LCD regression model. One of the limitations of this paper is that the models discussed here do not take into consideration, any covariates. As a future scope of the present study, regression models by incorporating linear covariates can be explored and applications can be showed using real life data sets.

ACKNOWLEDGMENT

The first author acknowledges the financial support provided by the Department of Science and Technology (DST), Government of India, to the first author through the INSPIRE programme for carrying out this research work.

References

[1] J.S. Rao, A. Sengupta, "*Topics in Circular Statistics*", World Scientific Publishing Co. Pte. Ltd, Singapore, pp.187-195, 2001.

[2] K.V. Mardia, P.E. Jupp, "*Directional Statistics*", John Wiley and Sons Ltd., Chichester, pp.13, 2000.

[3] S.R. Jammalamadaka, U.L. Lund, "*The effect of wind direction on ozone levels: a case study*". Environmental and Ecological Statistics, Vol.13, pp.287–298, 2006.

[4] L.P. Rivest, "*A decentred predictor for circular–circular regression"*. Biometrika, Vol.84, pp.717-726 , 1997.

[5] T.D. Downs, K.V. Mardia, "*Circular regression*". Biometrika, Vol.89, pp.683-697, 2002.

[6] S. Kato, K. Shimizu, G.S. Shieh, "*A circular-circular regression model*", Statistica Sinica, Vol.18, pp.633-645, 2008.

[7] S.R. Jammalamadaka, Y.R. Sarma, "*Circular Regression*". In: Matusita, K. (Ed.), Statistical Sciences and Data Analysis. VSP, Utrecht, pp.109-128, 1993.

[8] U. Lund, "*Least circular distance regression for directional data*", Journal of Applied Statistics, Vol.26, Issue.6, pp.723–733, 1999.

[9] N. Fisher, "*Statistical Ananlysis of Circular Data*", Cambridge University Press, Cambridge, pp.151, 1993.

[10] S.R. Jammalamadaka, Y.R. Sarma, "*A correlation coefficient for angular variables*", Statistical Theory and Data Analysis, Vol.2, pp. 349–364, 1988.

[11] U. Lund, C. Agostinelli, "*Circstats: circular statistics, from "topics in circular statistics"*". R package version 0.2–4. Available from http://www.cran.r-project.org/package=circstats, 2012.

[12] U. Lund, C. Agostinelli, "*Circular: circular statistics*". R package version 0.4–7, Available from http://www.cran.r-project.org/package=circular, 2013.

AUTHORS PROFILE

Dr. Sahana Bhattacharjee did her B.Sc from Cotton College, M.Sc from Gauhati University and Ph.D. in Statistics from the Department of Statistics, Gauhati University under the guidance of Professor Kishore Kumar Das. She is currently working as an Assistant Professor in the Department of Statistics, Gauhati University. Her research interest lies in Circular statistics, Distribution theory and other related fields.

Professor Kishore Kumar Das did his B.Sc from Arya Vidyapeeth College, M.Sc from Gauhati University and Ph.D. in Statistics from the Department of Statistics, Gauhati University under the guidance of Professor S.B. Nandi. He is currently working as a Professor in the Department of Statistics, Gauhati University. He has guided more than 15 students for their doctoral research and has completed several research projects as the principal investigator. He has been in the teaching profession since the last 25 years or so and has published more than 60 papers in journals of national and international repute. His research interest lies in Distribution theory, Circular statistics, Bio-statistics, Actuarial Statistics and other related fields.