Journal of the Korean Data & Information Science Society
2011, **22**(5), 961–966

한국데이터정보과학회지

Circular regression using geodesic lines[†]

Sungsu Kim[1]

[1]Department of Statistics, Kyungpook National University

Received 10 July 2011, revised 24 July 2011, accepted 10 August 2011

Abstract

Circular variables are those that have a period in its range. Their examples include direction of animal migration, and time of drug administration, just to mention a few. Statistical analysis of circular variables is quite different from that of linear variables due to its periodic nature. In this paper, the author proposes new circular regression models using geodesic lines on the surface of the sample space of the response and the predictor variables.

Keywords: Circular regression, geodesics lines, sample space.

1. Introduction

Circular regression in this paper refers to a regression that includes a circular variable. In the following, we denote a circular regression as dependent-independent form. For example, a circular regression that has a circular variable and a linear variable as dependent and independent variables, respectively, is denoted as circular-linear regression. In the following, some background of circular regression is presented.

The Gould (1969) paper is considered to be the earliest appearance of a circular regression, where he proposes a circular-linear regression model. It is discussed that his model produces non-unique maximum likelihood (ML) estimates due to having the linear mean link. In Johnson and Wehrly (1978) and Fisher and Lee (1992), readers can find their methods on improving the Gould's model. A circular-(circular, linear) regression is discussed in Lund (1999). In his paper, it is shown that the least circular distance estimators are the same as the ML estimators when the circular dependent variable is assumed to follow a von Mises distribution. In this paper, however, it is shown that predicting a circular variable from a linear predictor is not feasible when their bivariate sample space is solely considered. A linear-(linear, circular) regression is proposed in SenGupta and Ugwuowo (2006), where a trigonometric polynomial is used for a circular predictor. A new linear-circular model is proposed in this paper, after examining their bivariate sample space, which is a cylinder.

Circular-circular regression is discussed in Jammalamadaka and Sarma (1993) and Downs and Mardia (2002). In Jammalamadaka and Sarma (1993), their circular model is translated

[†] This research was supported by the 2010 Kyungpook National University research fund.
[1] Assistant professor, Department of Statistics, Kyungpook National University, Daegu 702-701, Korea. E-mail: dr.sungsu@gmail.com

into a linear model by using sine and cosine trigonometric polynomials of an independent circular variable for each of sine and cosine terms of a dependent circular variable. In Rivest (1997), Downs and Mardia (2002), Kato *et al.* (2008) and SenGupta and Kim (2011), the Möbius transformation that maps the unit disc to itself are used. In this paper, a new model is proposed after examining their bivariate sample space, which is a torus.

The motivation for the works in this paper is to propose circular regression models in a new perspective, i.e. in a topological point of view. When a sample space is a plane for two linear variables, it is often preferred to fitting a straight line to describe their relationship, which has the least complex form among all other possible models. A straight line on a plane is also a geodesic line. In this paper, new models are established by fitting the best geodesic line for data points in each sample space of response and predictor variables. In this paper, we are not interested in comparing the new models with other existing models since our purpose is solely to introduce the new perspective of modeling when sample spaces are considered.

2. Methodology

2.1. Method for a linear and a circular variables

When two linear variables are studied, their sample space is a plane and bivariate observations are presented as scattered points on a plane. A straight line is often fitted on a plane based on those bivariate observations due to its simplest form. When bivariate observations are a linear and a circular measurements, their sample space is the surface of a cylinder. Therefore, we can visualize n bivariate observations as n scattered points on the surface of a cylinder. The analogous to a straight line on a plane is a geodesic line on a surface (O'Neill, 1997). Since a cylinder and a plane are locally isometric, i.e. there is a distance-preserving isometry mapping between them, the geodesic line on the surface of a cylinder is the same as that of a plane, i.e. a straight line. In this point of view, we consider the least square line on a plane for n bivariate measurements on a circular and a linear variables.

The following plots show 16 observations on air quality index (linear) and wind direction (circular).

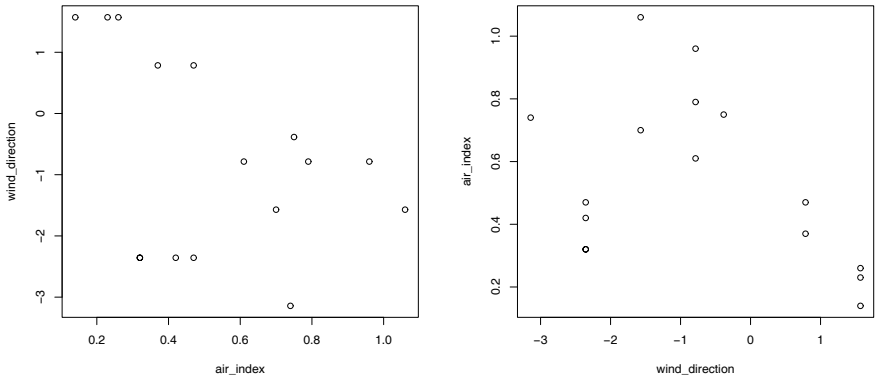

Figure 2.1 Scatter plots of wind direction and air index

The left and the right in Figure 2.1 have the circular variable and the linear variable in the y-axis, as the dependent variable, respectively. It is shown that a regression between a linear and a circular variables is only possible when a linear variable is the dependent variable, i.e. in y-axis, and a circular variable is the independent variable, i.e. in x-axis, but not in the other direction. When a circular variable is the dependent variable, as evidently shown in the left plot in Figure 2.1, there can be two values for the circular variable for each given value of a linear variable. This means that the model is not suitable for the purpose of predicting one corresponding value for a given value of a linear variable. Therefore, we do not discuss a circular-linear regression in this paper.

Having a circular dependent variable, as shown in the right of Figure 2.1, it is necessary to have more than one straight line for prediction of the air index, measured in parts per million, given a wind direction, measured in radians, due to the periodic nature of a circular variable. In Figure 2.1, one least square line can be fitted between $-\pi$ radians and -0.8 radians, and the other least square line can be fitted between -0.8 radians and 1.6 radians. This can be done by using a spline regression model (Marsh, 2002), which is given by

$$\text{AirIndex}_i = a_0 + b_1\text{WindDirection}_i + b_2 D_i(\text{WindDirection}_i + 0.8) + \epsilon_i, \quad i = 1, ..., 16,$$

where a_0, b_1 and b_2 are parameters, D_i is a dummy variable that takes 1 for wind direction more than -0.8, and 0 otherwise, and ϵ_i's are iid random variables with zero mean. The fitted equation is given by

$$\widehat{\text{AirIndex}}_i = -0.89 + 0.003\text{WindDirection}_i - 0.42 D_i(\text{WindDirection}_i + 0.8).$$

The left plot in Figure 2.2 shows the fitted line over the scatter plot shown in the right of Figure 2.1. Another linear-circular example is presented using the data set of a marine biology study, which is also used in Lund (1999). It is chosen that the dependent variable and independent variable are the amplitude of low tide, measured in inches, and the time of low tide, measured in hours, respectively. The fitted line over the scatter plot is shown in the right plot of Figure 2.2. A model for the data set is given by

$$\text{Amplitude}_i = a_0 + b_1\text{TimeLow}_i + b_2 D1_i(\text{TimeLow}_i - 9) + b_3 D2_i(\text{TimeLow}_i - 11) + \epsilon_i$$

for $i = 1, ..., 86$, where a_0, b_1, b_2 and b_3 are parameters, $D1_i$ takes 1 for the time of low tide more than 9 and 0 otherwise, $D2_i$ takes 1 for the time of low tide more than 11 and 0 otherwise, and ϵ_i's are iid random variables with zero mean. The fitted line equation is given by

$$\widehat{\text{Amplitude}}_i = -25.21 + 0.73 * \text{LowTide}_i - 7.6 * D1_i * (\text{LowTide}_i - 9) + 9.92 * D2_i * (\text{LowTide}_i - 11).$$

It is well known that the least square estimators, $\hat{a}$, $\hat{b}_1$, $\hat{b}_2$ and $\hat{b}_3$ are asymptotically normally distributed. Readers can find about the goodness of fit of using a spline regression in Eubank (1988). It is found that the optimal integrated risk in the spline regression model is $O\left(n^{-4/5}\right)$. This method of linear-circular regression, however, requires to choose the turning points properly as done in the above examples. This may be done by, first choosing an appropriate interval that a turning point belongs to, then trying many values in the interval in order to obtain the best least square fit among those models, which can be performed using a computer program. Alternatively, one can try to estimate the turning points as parameters in each model.

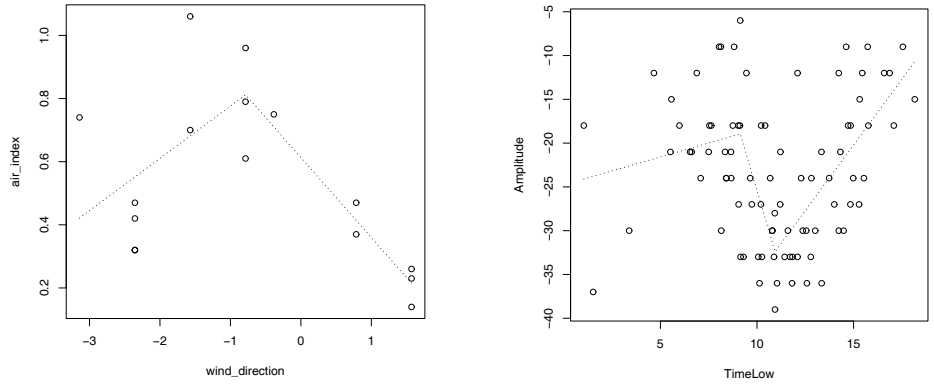

Figure 2.2 Fitted lines of wind direction and air index (left) and low tide and amplitude (right)

2.2. Method for two circular variables

Suppose (θ_1, ϕ_1),... ,(θ_n, ϕ_n) are n bivariate observations on two circular variables both ranging from -π to π. Their sample space is constructed as the surface on a torus. A parametrization of the surface, which is known as a Clairaut parametrization, is given by

$$x = (M + m\cos(\phi))\cos(\theta), y = (M + m\cos(\phi))\sin(\theta), z = a\sin(\phi),$$

where M and m are major and minor axis of a torus, respectively. Then, the geodesic equation for a torus (Irons, 2005) is given by

$$\theta = \theta_0 + h\int_{\phi_0}^{\phi} \frac{d\nu}{(M + m\cos(\nu))\sqrt{[(M + m\cos(\nu))^2 - h^2]}} + \epsilon, \tag{2.1}$$

where h is the geodesic slant, which is an angle between the geodesic line and the θ coordinate of the tangent vector, ϕ_0 is a starting point of ϕ, θ_0 is the conditional mean direction of θ given that $\phi = \phi_0$, and ϵ follows a circular distribution with zero mean direction. In terms of the absolute value of the h, we need to require that the absolute value of h is greater than 0 and less than $M - m$. Only in such a case, geodesic lines can pass through all points on the surface of torus, that is, they can alternatively cross both inner and outer equators on the surface, which is called 'unbounded geodesics. The integral in (2.1) does not have a closed form solution, but can be numerically fitted using the constrOptim function in R with three parameters θ_0, ϕ_0, and h.

The distance from an angle A to an angle B can be measured in two different ways depending on the direction of measurement around the circle. For example, the distance from 30° to 330° is measured as 300° or 60° when going counterclockwise or clockwise, respectively. Hence, a more proper distance measure between two angles, A and B, is given in the following definition (Jammalamadaka and SenGupta, 2001).

Definition 2.1 Circular distance (cd) between two points A and B on a circumference is defined to be

$$cd(A, B) = 1 - \cos(\alpha - \beta),$$

where α and β represent the angles measured from a specified origin corresponding to the points, A and B, respectively. If $\alpha - \beta = 0$ and $|\alpha - \beta| = \pi$, then $cd(A, B) = 0$ and $cd(A, B) = 2$, respectively.

When the dependent variable is a circular variable, instead of considering the sum of squared distances, we minimize the sum of the circular distances, yielding the least circular distance estimators of θ_0, ϕ_0, and h.

Fitting the equation is performed by the least circular distance method, where M and m are arbitrary chosen as 3 and 1, respectively. Hence, we have boundary conditions for θ_0 and ϕ_0 from $-\pi$ to π and for h from 0 to 2. After enforcing a restriction that $\hat{\theta}_{\phi=-\pi} = \hat{\theta}_{\phi=\pi}$, the objective function to be minimized is given by

$$Q = \sum_{i=1}^{n} \cos\left(\theta_i - \theta_0 + h\int_{\phi_0}^{\phi_i} \frac{d\nu}{(M + m\cos(\nu))\sqrt{[(M + m\cos(\nu))^2 - h^2]}}\right) - \lambda\left(\int_{-\pi}^{\pi} \frac{d\nu}{(M + m\cos(\nu))\sqrt{[(M + m\cos(\nu))^2 - h^2]}}\right), \tag{2.2}$$

where λ denotes the Lagrange multiplier. The first order equations are shown below

$$\frac{\partial Q}{\partial \theta_0} = \sum_{i=1}^{n} \sin\left(\theta_i - \theta_0 + h\int_{\phi_0}^{\phi_i} \frac{d\nu}{(M + m\cos(\nu))\sqrt{[(M + m\cos(\nu))^2 - h^2]}}\right) = 0$$

$$\frac{\partial Q}{\partial \phi_0} = \sum_{i=1}^{n} \sin\left(\theta_i - \theta_0 + h\int_{\phi_0}^{\phi_i} \frac{d\nu}{(M + m\cos(\nu))\sqrt{[(M + m\cos(\nu))^2 - h^2]}}\right) \times \phi_0 \int_{\phi_0}^{\phi_i} \frac{d\nu}{(M + m\cos(\nu))\sqrt{[(M + m\cos(\nu))^2 - \phi_0^2]}} = 0$$

$$\frac{\partial Q}{\partial h} = \sum_{i=1}^{n} \sin\left(\theta_i - \theta_0 + h\int_{\phi_0}^{\phi_i} \frac{d\nu}{(M + m\cos(\nu))\sqrt{[(M + m\cos(\nu))^2 - h^2]}}\right) \times [1 - h^2]\left\{\int_{\phi_0}^{\phi_i} \frac{d\nu}{(M + m\cos(\nu))\sqrt{[(M + m\cos(\nu))^2 - h^2]}}\right\} = 0$$

$$\frac{\partial Q}{\partial \lambda} = \int_{-\pi}^{\pi} \frac{d\nu}{(M + m\cos(\nu))\sqrt{[(M + m\cos(\nu))^2 - h^2]}} = 0,$$

where we used the Leibnitz's Rule (Casella and Berger, 2002). Using the bivariate wind direction data (Johnson and Wehrly, 1978), the fitted equation is shown below,

$$\hat{\theta} = -0.11 - 0.28\int_{-3.03}^{\phi} \frac{d\nu}{(3 + \cos(\nu))\sqrt{[(3 + \cos(\nu))^2 - 0.08]}}.$$

A linear estimator from the least circular distance estimation (LCDE) is asymptotically normally distributed (Kim, 2009). The asymptotic variance-covariance matrix is given by the inverse of the Hessian matrix of (2.2). For example,

$$\hat{h} \sim N(h, H[3, 3]^{-1}),$$

where $H[3,3]$ is the third diagonal element in the Hessian matrix of (2.2).

3. Summary

Circular regression methods using geodesic lines on the sample spaces are discussed in this paper. For linear-circular bivariate data, their sample space is described as the surface of a cylinder. For circular-circular bivariate data, it is described as the surface of a torus. Geodesic lines on a surface are fitted as the simplest form of model, as analogous to a straight line on a plane. Using the isometry between the surface of a cylinder and a plane, a linear spline regression model is fitted for a linear-circular case. For a circular-circular case, the geodesic line can be fitted numerically using the LCDE method. It is also shown that a circular-linear regression is not feasible when the sample space is taken into consideration.

References

Casella, G. and Berger, R. L. (2002). *Statistical inference*, Thomson Learning Inc., Pacific Grove, CA.

Downs, T. and Mardia, K. (2002). Circular regression. *Biometrika*, **89**, 683-698.

Eubank, R. (1988). *Spline smoothing and nonparametric regression*, Marcel Dekker, Inc., New York.

Fisher, N. and Lee, A. (1992). Regression models for an angular responses. *Biometrics,* **48**, 665-677.

Gould, A. (1969). A regression technique for angular variate. *Biometrics,* **25**, 683-700.

Irons, M. (2005). *The curvature and geodesics of the torus.* http://www.rdrop.com/ half/math/torus/index.xhtml.

Jammalamadaka, R. and Sarma, Y. (1993). Circular regression. *Statistical Sciences and Data Analysis*, 109-128.

Johnson, R and Wehrly, T. (1978). Some angular-linear distributions and related regression models. *Journal of the American Statistical Association,* **73**, 602-606.

Kato, S., Shimizu, K. and Shieh, G. S. (2008). A circular-circular regression model. *Statistica Sinica*, **18**, 633-645.

Lund, U. (1999). Least circular distance regression for directional data. *Journal of Applied Statistics*, **26**, 723-733.

Marsh, L. (2002). *Spline regression models*, Sage Publications, Inc., Thousand Oaks, CA.

O'Neill, B. (1997). *Elementary differential geometry*, 2nd Ed., Academic Press, San Diego.

Rivest, L. P. (1997). A de-centered predictor for circular-circular regression. *Biometrika*, **84**, 717-726.

SenGupta, A. and Kim, S. (2011). Statistical inference for homologous gene pairs of two circular genomes: A new circular-circular regression model. submitted to *Statistical Papers*.

SenGupta, A. and Ugwuowo, F. (2006). Asymmetric circular-linear multivariate regression models with applications to environmental data. *Environmental and Ecological Statistics,* **13**, 299-309.