Restricted Inference in Circular-Linear and Linear-Circular Regression

Thelge Buddika Peiris[1*] and Sungsu Kim[2]

1. Department of Mathematical Sciences, Worcester Polytechnic Institute
2. Division of Statistics, Northern Illinois University

[*]Corresponding Author: tbukp@yahoo.com

Received: 29th June 2015/ Revised: 24th December 2015/ Accepted: 11th February 2016

ABSTRACT

In this paper, we investigate restricted inference on two types of circular regression, called circular-linear and linear-circular. Our aim in this paper is to propose an alternative method which is necessary to apply where one observes a weak association between circular dependent and linear predictor variables, or between linear dependent and circular predictor variables, having clear knowledge about the sign of slope. We illustrate that restricted inference is particularly useful for those circular regressions, which is due to weak association. Comparison between our proposed restricted inference and the unrestricted inference are given by using two examples, one from ecological study and the other from environmental study.

Keywords: Air quality, Amplitude of tide, Circular data, Slope parameter.

1. Introduction

Circular variables are those that take any periodic measurements. Two typical examples are angle, which is periodic in 360 degrees, and the hourly time, which is periodic in 24 hours. Circular variables appear in many areas of research. Various examples can be found in Mardia and Jupp (1999). For example, when studying variables that influence the climate at a certain site, it is found that, from a meteorology point of view, most studies focus on wind direction and related variable such as rainfall (Carnicero, et al. 2011). Circular regression means any regression involving circular variables as response or predictor variables. In this paper, we investigate two circular regressions. One is called a linear-circular regression, which has a linear variable as response and a circular variable as predictor. The other is called a circular-linear regression, which has a circular response variable and a linear predictor variable.

In many applications, it is reasonable to assume that the regression function varies monotonically with the predictor variable in some region of interest. Order-restricted inference in a simple linear regression (Mukerjee and Tu, 1995) is only useful when the association between a response and predictor variables is weak in general. When the predictor or response variable is a circular variable, it is shown in this paper that the order restricted inference is useful, since a simple linear regression involving a circular variable and a linear variable tends to have a weak association. We show this phenomenon with a couple of examples in Section 4. The restricted inference is conservative and may produce a wider confidence interval than the unrestricted method, particularly when the magnitude of slope is large.

This paper is organized as follows. In Section 2, we explain the backgrounds that are useful for later sections. In Section 3, we propose order restricted circular regression models. In Section 4, we present data analysis examples, followed by the discussion and concluding remarks in Section 5.

2. Background

2.1 Circular Regression

A variable that is measured in the form of any periodic manner is called a circular variable. For a couple of examples, an angle is a circular variable having 2π period, and the time of a day is a circular variable having 24 hours as the period. In this section, we discuss circular regression models that take into account of the periodic nature of circular variable. Suppose one observes $(\theta_1, y_1), (\theta_2, y_2), \cdots, (\theta_n, y_n)$, where θ denotes circular measurements and y denotes linear measurements. Consider the following circular regression model (Kim and Sen Gupta, 2014)

$$y_i = \alpha + \beta \cos(\theta_i - \mu) + \varepsilon_i, \tag{1}$$

where α and β denote intercept and slope parameters, respectively, μ denotes the mean direction, and ε_i has zero mean and a constant variance σ^2. In this model, when $\beta > 0$, as θ goes away from μ, it is seen that y decreases, while as θ moves towards μ, y increases.

For $\beta < 0$, we have the opposite association between y and θ. Notice that this is a 2-to-1 mapping from θ to y. We claim that a 2-to-1 mapping such as (1) is necessary when mapping from θ to y, or from y to θ. As an example, suppose one attempts to model a monotonic association between y and θ, i.e. 1-to-1 mapping from θ to y. For example, consider the following regression model for y on θ.

$$y = \alpha + \beta \tan\left(\frac{\theta - \mu}{2}\right) + \varepsilon,$$

where μ denotes the mean direction of θ. Then, it is easily seen that y takes $-\infty$ and ∞, two opposite ends of the real line, when $\theta = \mu - \pi$ and $\theta = \mu + \pi$, respectively. However, due to the periodic nature of θ, $\theta = \mu - \pi$ and $\theta = \mu + \pi$ are the same angle, which immediately tells that an 1-to-1 mapping is not suitable between y and θ. On the other hand, in model (1), y takes the same value when $\theta = \mu - \pi$ and $\theta = \mu + \pi$. As an example, y can represent the seasonal unemployment rate and θ can represent the month. It is clear that the unemployment rate goes back to the January's rate after passing December in an yearly periodic cycle.

Likewise, when regressing θ on y instead of y on θ, one may consider the following model in the same notion as stated above.

$$\cos(\Theta - \mu) = \alpha + \beta y + \varepsilon,$$

where we assume that ε is normally distributed with mean 0 and variance σ^2 and σ^2 is small enough so that 3 times the standard deviation is less than 1.

2.2 Restricted Inference in Simple Linear Regression

In this section, we present the results from (Mukerjee and Tu, 1995) about the order restricted simple linear regression. Consider a simple linear regression given by the following:

$$y_i = \alpha + \beta x_i + \varepsilon_i.$$

Suppose one has a prior knowledge about the sign of β, the slope parameter. We consider the inferences with the restriction $\beta \geq 0$ and similarly the case for $\beta \leq 0$ can be obtained.

We choose the transformation corresponding to $\sum x_i = 0$ so that the unrestricted maximum likelihood estimators (MLE's) $\hat{\alpha}$ and $\hat{\beta}$ are independent. Then the restricted estimators are given by,

$$\alpha^* = \hat{\alpha} \quad \text{and} \quad \beta^* = \hat{\beta} = \max\{\hat{\beta}, 0\}.$$

Let $S_x^2 = \sum x_i^2$ and $S_y^2 = \sum (y_i - (\alpha - \beta x_i))^2 / \upsilon$ where $\upsilon = n - 2$. Then, an $(1-\alpha)100\%$ confidence interval $[L,U]$ for β can be obtained

$$L = (\hat{\beta} - t_{\upsilon,\alpha/2} S_y / S_x)^+ \quad \text{and} \quad U = (\hat{\beta} + t_{\upsilon,\alpha/2} S_y / S_x)^+,$$

where L , U denote the lower and upper bounds, respectively, $t_{\upsilon,\alpha/2}$ is the upper $\alpha/2$ quantile of a t - distribution with υ degrees of freedom. Deriving confidence intervals in the restricted case is difficult (Mukerjee and Tu 1995).

Considering inferences about the regression function at a given point x_0, an $(1-\alpha)100\%$ confidence interval $[L,U]$ for the mean response at $x_0 = 0$ is given by

$$L = (\hat{\alpha} - t_{\upsilon,\alpha/2} S_y / \sqrt{n}) \quad \text{and} \quad U = (\hat{\alpha} + t_{\upsilon,\alpha/2} S_y / \sqrt{n}),$$

and for $x_0 > 0$,

$$L = (\hat{\alpha} - C_{\alpha/2} S_y / \sqrt{n}) \quad if \quad \hat{\beta} \le 0,$$

$$L = (\hat{\alpha} - C_{\alpha/2} S_y / \sqrt{n}) \sqrt{1 - (S_x^2 \hat{\beta}^2)/(C_{\alpha/2}^2 S_y^2)}$$

$$if \;\; 0 < \hat{\beta} < C_{\alpha/2} S_y / \sqrt{S_x^2 + S_x^4 /(n x_0^2)},$$

$$L = \hat{\alpha} + \hat{\beta} x_0 - C_{\alpha/2} S_y \sqrt{1/n + (x_0 / S_x^2)}$$

$$if \;\; \hat{\beta} \ge C_{\alpha/2} S_y / \sqrt{S_x^2 + S_x^4 /(n x_0^2)},$$

$$U = \hat{\alpha} \;\; if \;\; \hat{\beta} x_0 \le t_{\upsilon,\alpha/2} S_y \sqrt{1/n + (x_0 / S_x^2)},$$

$$U = \hat{\alpha} + \hat{\beta} x_0 + t_{\upsilon,\alpha/2} S_y \sqrt{1/n + (x_0 / S_x^2)}, \;\; otherwise,$$

and for $x_0 < 0$,

$$L = \hat{\alpha} \ \ if \ \ \hat{\beta} x_0 \le -t_{\upsilon,\alpha/2} S_y \sqrt{1/n + (x_0 / S_x^2)},$$
$$L = \hat{\alpha} + \hat{\beta} x_0 - t_{\upsilon,\alpha/2} S_y \sqrt{1/n + (x_0 / S_x^2)}, \ \ otherwise,$$
$$U = (\hat{\alpha} + C_{\alpha/2} S_y / \sqrt{n}) \ \ if \ \ \hat{\beta} \le 0,$$
$$U = (\hat{\alpha} + C_{\alpha/2} S_y / \sqrt{n}) \sqrt{1 - (S_x^2 \hat{\beta}^2)/(C_{\alpha/2}^2 S_y^2)}$$
$$if \ \ 0 < \hat{\beta} < C_{\alpha/2} S_y / \sqrt{S_x^2 + S_x^4 /(n x_0^2)},$$
$$U = \hat{\alpha} + \hat{\beta} x_0 + C_{\alpha/2} S_y \sqrt{1/n + (x_0 / S_x^2)}$$
$$if \ \ \hat{\beta} \ge C_{\alpha/2} S_y / \sqrt{S_x^2 + S_x^4 /(n x_0^2)},$$

where $C_{\alpha/2}$ can be looked up from the table found in Mukerjee and Tu (1995).

3. Restricted Inference in Circular Regression

3.1 Circular Response Variable and Linear Predictor Variable

In this section, we study the case of regression having a circular variable Θ on a linear predictor variable Y. For example, Θ represents the spawning time of a particular fish and Y represents the tidal amplitude of the fish's environment. A particular numerical example is given later in the example section. We propose the following regression frame work:

$$\cos(\Theta - \mu) = \alpha + \beta y + \varepsilon, \qquad (2)$$

where ε represents a normally distributed error having zero mean and the small variance enough to cover the range of (-1, 1) using plus and minus of three times the standard deviation. In (2), $\beta > 0$ refers to the case where θ moves away from μ in both directions, clockwise and counter-clockwise as y decreases, while $\beta < 0$ means that θ moves towards μ in both sides as y decreases. The link function in (2) can be rewritten as shown below:

$$\Theta = \mu + \arccos(\alpha + \beta y). \qquad (3)$$

Using those formulas in Section 2, we can obtain an $(1-\alpha)100\%$ confidence interval for Θ as shown below, which is also displayed in Figure 1.

$$[\mu - \arccos(U), \ \mu - \arccos(L)] \cup [\mu + \arccos(L), \ \mu + \arccos(U)].$$

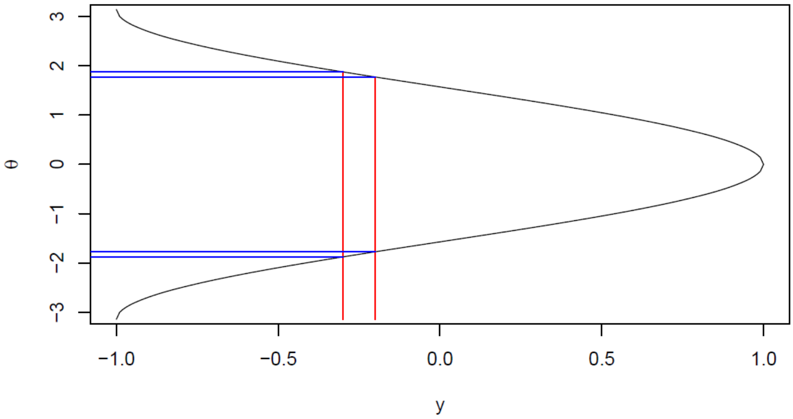

Figure 1: Sketch of the confidence interval for Θ , where $\mu = 0$.

3.2 Linear Response Variable and Circular Predictor Variable

In this section, we consider the case of regression having a linear variable Y on a circular predictor variable Θ . For example, Θ may represent the month of year and Y represents a seasonal linear variable. A particular numerical example is given later in the example section. In this paper, we study a type of regression frame work shown in (1).

From (1), $\beta > 0$ refers to the case where Y decreases as θ moves away from μ in both directions, while $\beta < 0$ means that Y decreases as θ moves towards μ in both sides of μ . When $\beta > 0$, a confidence interval for $\alpha + \beta \cos(\theta - \mu)$ is given in Section 2.

4. Numerical Examples

4.1 Circular-Linear Model

In a marine biology study by Robert T. Warner at University of California, Santa Barbara, data were gathered on the spawning time of a particular fish (Lund, 1999). It is hypothesized that the spawning time is affected by tidal characteristics of the fish's environment; For example as the amplitude goes down, the spawning time moves away from the mean direction time, which estimate is given by 13.42. In the following, we estimate the model (2) with the constraint $\beta \geq 0$. In Figure 2, a scatter plot of the response and the predictor variable based on 86 observations is shown. A weak association is evidently shown in the figure.

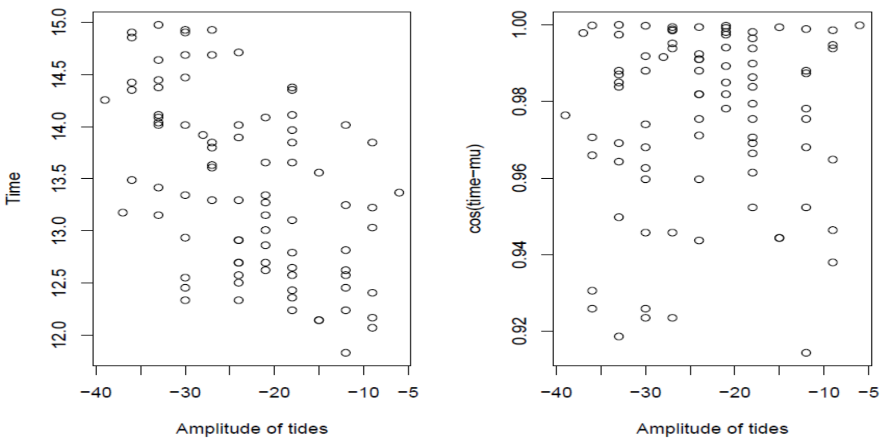

Figure 2: Plots for Time versus Amplitude (Left) and cos(time- μ) vs. Amplitude (Right)

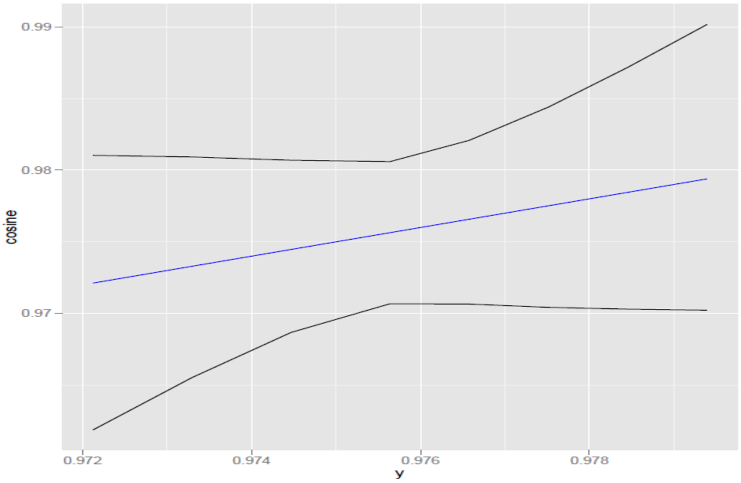

Figure 3: Confidence band with the mean line, where the 'cosine' and 'y' represent cosine of (spawning time - mean direction) and the amplitude of tides, respectively.

In Figure 3, we provide the confidence band with the mean line using the method in Section 2. In Table 1, those mean added values used in the plot are listed. Using (3), the corresponding “spawning times” for the mean added amplitude of tide equal to -20 is given by

$$(12.50, 12.67) \cup (14.18, 14.34),$$

whereas the unrestricted confidence interval is

$$(12.50, 12.69) \cup (14.15, 14.34).$$

In this section, we have shown that the restricted confidence interval is narrower than the unrestricted one.

Table 1: Mean values and confidence intervals of cos(spawning time - mean direction) for a range of the amplitude of tides.

Amplitude	**Lower bound**	**Mean value**	**Upper bound**
-38.1512	0.9619	0.9721	0.9810
-33.1512	0.9655	0.9733	0.9809
-28.1512	0.9687	0.9745	0.9807
-23.1512	0.9707	0.9756	0.9806
-19.1512	0.9707	0.9766	0.9821
-15.1512	0.9704	0.9775	0.9844
-11.1512	0.9703	0.9785	0.9872
-7.1512	0.9702	0.9794	0.9902

4.2 Linear-Circular Model

Air quality is defined as a measure of the condition of air relative to the requirements of one or more biotic species, and/or to any human need or purpose at a given location. As the air quality index (AQI) increases, an increasingly large percentage of the population is likely to experience increasingly severe adverse health effects. AQI varies by pollutant, and is different in various geographic locations. De Wiest and Della Fiorentina (1975) proposed a new air quality index, which is based on the experimental data from different sampling sites in Liege, Belgium. Using their data set (De Wiest and Della Fiorentina, 1975), we illustrate the order restricted inference of linear circular regression problem with AQI as the dependent variable and the wind direction as the predictor variable, by imposing an order restriction of $\beta \geq 0$ to the model given in (1). In Figure 4, a scatter plot of the response and the predictor variable based on 15 observations is shown.

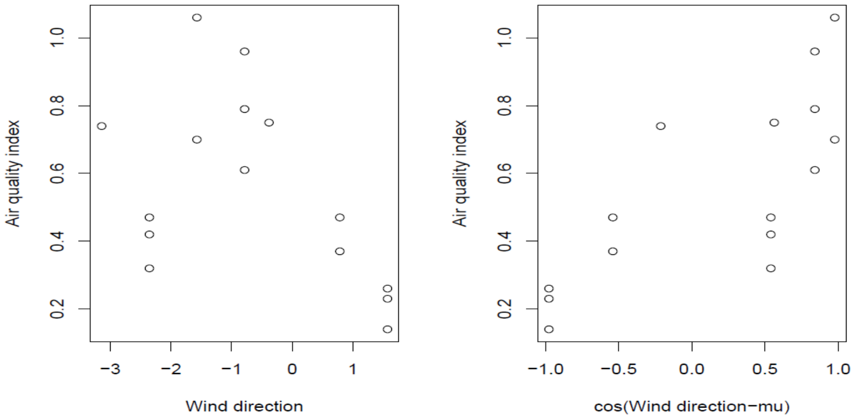

Figure 4: Plots for AQI vs. wind direction (Left) and AQI vs. cos(wind direction- μ) (Right).

In Figure 5, we provide the mean line for those values in the sampled range of wind direction, along with the confidence band. Those mean added values used in the plot are listed in Table 2.

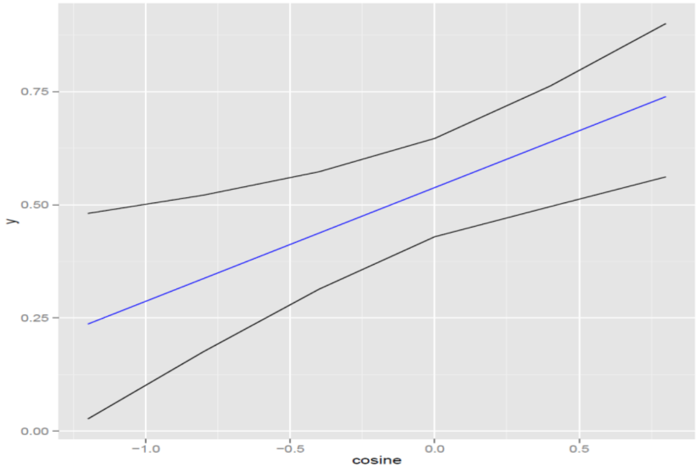

Figure 5: Mean line with the confidence band, where the 'y' and 'cosine' represent the air quality index and cosine of (wind direction - mean direction), respectively.

Using (1), a 95% confidence interval of the corresponding “air quality index” for the wind direction equal to 75 degrees is given by (0.4581, 0.6834), whereas the unrestricted confidence interval is (0.4623, 0.6834). In this section, it is shown that the restricted confidence interval can be wider than the unrestricted one when the magnitude of slope is large.

5. Discussion and Concluding Remark

In this paper, we present the order-restricted inference on two types of circular regressions, where we utilize one example from ecological study and the other example from environmental study. In both examples, plots for mean values with confidence band are provided.

Table 2: Mean values and confidence intervals of the air quality index within the sampled range of wind direction.

Wind direction	Lower bound	Mean value	Upper bound
170.8	0.0422	0.2467	0.4724
131.8	0.1756	0.3371	0.5214
106.3	0.3137	0.4376	0.5731
83.2	0.4295	0.5381	0.6467
58.05	0.4961	0.6386	0.7626
13.5	0.5618	0.7391	0.9006

Lacking pivotals, deriving confidence intervals in the restricted case is difficult. This constitutes perhaps the major deficiency of restricted inference at present as far as applications of the restricted methodology to applied problems are concerned. A systematic method consists of inverting one sided test of hypotheses (Lee 1984; Schoenfeld 1986; Williams 1977), and we used formulas derived using that method. As noted by Williams (1997) and Mukerjee and Tu (1995), these intervals are conservative, because least favourable distribution are used to compute the level of significance for composite null hypotheses in each of the one-sided test, and these distributions are usually different.

We emphasize that the restricted inference is necessary in a situation where one observes a weak association and has knowledge of subject matter about the sign of slope. On the other hand, since the definition of 'weak association' is somewhat indefinite, without clear knowledge about the sigh of slope, we recommend our readers to construct both restricted and unrestricted confidence intervals, and choose a better one to make an inference. Nonetheless, if one is clear about the sign, she/he needs to apply the restricted inference. We conclude this manuscript hoping that the models proposed in this paper benefit broad applications in diverse disciplines.

Acknowledgements

We are very grateful about valuable comments provided by reviewers that have improved our final version.

References

1. Carnicero J. A., Wiper M. P. and Ausin C. (2011). Non-parametric Methods for Circular-circular and Circular-linear Data Based on Bernstein Copulas. *Working Papers, Universidad Carlos III De Madrid*.

2. De Wiest F. and Della Fiorentina H. (1975). Suggestions for a Realistic Definition of an Air Quality Index Relative to Hydro-carbonaceous Matter Associated with Airborne Particles. *Atmospheric Environment, 9:951-954.* DOI: 10.1016/0004-6981(75)90105-5

3. Kim S. and SenGupta A. (2014). Inverse Circular-linear and Linear-circular Regression. *Communications in Statistics: Theory and Method.* DOI: 10.1007/s00362-012-0454-1

4. Lee C. I. C. (1984). Truncated Bayesian Confidence Region and Its Corresponding Simultaneous Intervals in Restricted Normal Models. Technical Report, Memorial University of Newfoundland, Dept. of Mathematics and Statistics.

5. Lund U. (1999). Least Circular Distance Regression for Directional Data. *Journal of Applied Statistics, 26:723-733.* DOI: 10.1080/02664769922160

6. Mardia K. V. and Jupp P. E. (1999). Directional Statistics. Wiley Series, N.Y.
DOI: 10.1002/9780470316979

7. Mukerjee S. E. and Tu I. R. (1995). Order-restricted Inferences in Linear Regression. *Journal of American Statistical Association, 90:717-728.*
DOI: 10.1080/01621459.1995.10476565

8. Rao J. S. and SenGupta A. (2001). Topics in Circular Regression. World Scientific, N.Y.

9. Schoenfeld D. (1986) Confidence Intervals for Normal Means Under Order Restrictions, with applications to Dose-Responses Curves, Toxicology Experiments, and Low-Dose Extrapolations. *Journal of American Statistical Association, 81:186-195.*
DOI: 10.1080/01621459.1986.10478258

10. Williams D. A. (1977). Some Inference procedures for Monotonically Ordered Normal Means. *Biometrika, 64:9-15.*
DOI: 10.1093/biomet/64.1.9