Asymmetric Circular-linear Multivariate Regression Models with Applications to Environmental Data

ASHIS SENGUPTA[1] and FIDELIS I. UGWUOWO[2]

[1]*Applied Statistics Unit, Indian Statistical Institute, Kolkata, W.B., 700108, INDIA*
E-mail: ashis@isical.ac.in

[2]*Department of Statistics, University of Nigeria, Nsukka, Enugu State, NIGERIA*
E-mail: fiugwuowo@yahoo.com

We propose asymmetric angular-linear multivariate regression models, which were motivated by the need to predict some environmental characteristics based on some circular and linear predictors. A measure of fit is provided through the residual analysis. Some applications using data from solar energy radiation experiment and wind energy are given.

Keywords: Angular-linear dependency, environmental data, multivariate regression, solar energy, wind energy

---

1. Introduction

In recent times, great emphasis is being devoted to the development and possible use of alternative sources of energy. This is so because of the sure and eventual depletion of the conventional sources of energy, such as oil and natural gas and other fossil based fuels, coupled with their ever-rising cost of procurement and environmental hazards. It is in this light that the development of renewable energy resources in the world is very important. Some of the key renewable energy resources are the solar energy and wind energy. Due to the nature of the atmosphere, some solar energy fails to penetrate the atmosphere and a vast majority is reflected back by dust particles. Cloud covers prevent the earth's surface from being uniformly heated thus causing temperature differences on the surface and in the atmospheric masses situated near them. These temperature differences cause atmospheric pressure differences as well. The air attempts to equate these peaks and troughs by circulating across terrestrial surface in swirls and eddies. These air movements are further modified by the forces imposed by the earth's rotation as well as by the resistance to air movement near the ground presented by topographical features. Wind energy therefore, is defined as the kinetic energy of moving atmospheric air and which can be seen to be totally and directly caused by solar insulation and is infact a derivative of solar energy (Cheremisinoff, 1978).

We have observed a periodic phenomenon in the response variable for solar energy data, which spans for twelve hours in a day and this periodic phenomenon is not necessarily symmetric due to the atmospheric changes observed on daily basis. The wind

energy data, which spans for twenty-four hours, was observed for about one month and it could have possibly been asymmetric but our data did not reveal such hence we adopted a symmetric model.

Researchers especially in environmental studies have often encountered multivariate regression models involving some circular and linear predictors. Gould (1969) had earlier developed a regression model that predicted a circular response from a set of linear predictors by using a conditional von-Mises distribution on the set of linear covariates. The maximum likelihood estimators of the parameters for his model is however not unique. Ohta, Marita and Mizoguchi (1976) and De Wiest and Della Fiorentina (1975) suggested a measure of Air Quality Index (AQI) based on bivariate data comprising of the circular variable,(wind direction) and one linear variable, (level of pollutant). Johnson and Wehrly (1978) developed parametric models for angular-linear dependency based on maximum entropy conditional distribution, which have been generalized by SenGupta (2004) to encompass possibly asymmetric directional data on manifolds. The work by Fisher and Lee (1992) generalized the Johnson-Wehrly models by allowing the response variable to have a concentration parameter that is also a function of the linear covariates. Some expository literature on regression with directional data exist, e.g see the monographs by Jammalamadaka and SenGupta (2001), Mardia and Jupp (2000) and Batschelet (1981). However, there have been only limited attempts to model multivariate angular-linear data. Further, these do not involve the asymmetry when present, which is often so characteristic of circular data in practice. Here, we propose some asymmetric circular linear multivariate regression models and apply those to several real-life environmental data sets.

Our primary interest is in modeling multivariate multiple regression models involving a linear response, a circular predictor and a set of linear covariates. In section 2, we enumerate the forms of the models under consideration, the method of estimation and the criteria for model selection. Some discussions on model fitting are presented in section 3. The applications of our approach to two real-life environmental data sets, together with the residual analysis for checking the goodness-of-fit of the model are given in section 4. Some concluding remarks are presented in section 5.

2. Circular-linear regression models

In model (1), we consider a regression model that involves a simple cosine function given by

$$Y_i = M + \sum_{i=1}^{k} \beta_i x_i + A\cos\omega(t_i - t_0) + \varepsilon_i \qquad (1)$$

where Y_i is the linear response variable, M is the mean level, β_i is the regression coefficient, X_i is the linear independent variable, A is the amplitude, ω is the angular frequency, t_i is the circular independent variable (usually time) subject to a certain period T, t_0 is the acrophase and ε_i is the random error component.

The estimate of ω is given in radian or degree respectively as

$$\omega = 2\pi / T \quad \text{or} \quad 360^0 / T$$

We assume T and, hence, ω to be known.
Equation (1) can be written as

$$= M + \sum_{i=1}^{k} \beta_i x_i + A\{\cos\omega t_i \cos\phi + \sin\omega t_i \sin\phi\} + \varepsilon_i \qquad (2)$$
$$= M + \sum_{i=1}^{k} \beta_i x_i + C\cos\omega t_i + D\sin\omega t_i + \varepsilon_i$$

where $\phi = \omega t_0$, C = $A\cos\phi$, D = $A\sin\phi$

The estimates of A and ϕ can easily be found algebraically once those of C and D are obtained.

We next consider a trigonometric polynomial, which is a generalization of the cosine model. It contains the angular frequencies with a multiple of ω given by ωt, 2ωt,…, kωt. The function, in addition to the overall period T, contains smaller periods T/2, T/3, … which fit exactly into the overall period. The model is given as

$$Y_i = M + \sum_i \beta_i x_i + g(t_i) + \varepsilon_i$$

where $g(t_i) = A_1 \cos(\omega t - \phi_1) + A_2 \cos(2\omega t - \phi_2) + ... + A_k \cos(k\omega t - \phi_k)$ (3)

The terms $A_1 \cos(\omega t - \phi_1), A_2 \cos(2\omega t - \phi_2)$, etc. are called the first, second, etc. harmonics. This model is applied when there are multiple periods within the general period.

The non-linear periodic function could be applied when the oscillation pattern is complicated with major and minor peaks and troughs. The trigonometric polynomial shown in equation (3) may fit well but will involve too many terms and parameters. When the peaks and troughs do not follow each other, it implies that there is a skew and we propose the non-linear model for this situation to be given as model (2) by,

$$Y_i = M + \sum_{i=1}^{k} \beta_i x_i + A\cos(\psi + \nu\cos\psi) + \varepsilon_i \qquad (4)$$

where $\psi = \omega t - \phi$ and ν is the parameter of skewness. It has been shown that ν is a value that usually lies in an interval $-30^o \le \nu \le 30^o$ and we obtain the simple cosine function whenever $\nu = 0^o$.

When the oscillations are sharply peaked or flat-topped, we will consider yet another model given by,

$$Y_i = M_1 + \sum_{i=1}^{k} \beta_{1i} x_i + A_1 \cos(\psi + \nu_1 \sin\psi) + \varepsilon_i \qquad (5)$$

where ν_1 is the parameter of kurtosis and it indicates to what extent the shape differs from a sinusoidal oscillation.

3. Model fitting

Detecting the best parsimonious model is an important step in the analysis of data sets. The methods of linear and non-linear least squares were applied in parameter estimation. Several diagnostic tools were used to assess the goodness of the model fit. An overview of several approaches for assessing the effect of number of parameters in determining the best non-linear regression model can be found in the book by Seber and Wild (1988).

Going by some set of assumptions, the standardized residuals ought to resemble a sample from a standard normal distribution. A comparison of these residuals with a standard normal distribution allows an assessment of the distributional assumption. A plot of the residuals against fitted values allows an assessment of the variance structure. We will also consider the normal P-P plot and the Q-Q plot of the standardized residuals to assess the fit of the models.

4. Applications

4.1 *Solar energy data*

Solar energy is produced through nuclear fusion of the light elements that constitute the sun (hydrogen-Helium reaction). 4.7 x 10^5 tons/sec of elements reacting produces 3.8 x 10^{23} kw energy, out of which ½ billion reaches the earth. The sun supplies all forms of energy being used on earth both directly and indirectly and without which most activities would not proceed on a favorable note. Solar energy is the most dependable, abundant, and constant source of energy on earth and it is produced through the process of nuclear fusion of light elements, hydrogen, and helium during a spontaneous reaction of the atoms.

The measurements of the solar radiation can be obtained depending of course on the atmospheric weather at the point in time. The measurement could be made of diffuse radiation, direct radiation, or total radiation. The instrument conventionally used for total radiation is called Pyranometer. It can also be used for measuring diffuse radiation by shading it from direct radiation using a shade ring.

The experiment was conducted by engineers in Mechanical Engineering Department, University of Nigeria, Nsukka. It was performed for over a period of one year but a very reliable data was collection for a period of six days in the month of October and half-hourly record starting from 9 am to 5.30pm. This period was chosen due to the uninterrupted weather condition that provided sunshine throughout. We associated these hours with the angles 0^o, 20^o, 40^o,… respectively. The predictor variables considered are ambient temperature (X_{1i}), control temperature (X_{2i}), and the time in hours (t_i) to predict the absorber temperature (Y_i) in a well-constructed Thermosyphon solar water heater. A preliminary look at the data suggested model (1) as a reasonable choice. The estimates of the parameters of the regression curve were obtained by the method of least squares.

The equation of the fitted model (1) after deriving the amplitude and acrophase is

$$Y_i = -195.9700 + 5.7500x_{1i} + 2.0960x_{2i} + 2.7069\cos(15t_i - 179.5964) \qquad (6)$$

Finally, we convert the acrophase angle $\phi = 179.5964^o$ into hour of the day. Since 180^o stands for 12 noon, the acrophase is estimated to be 12 noon.

Table 1: Summary values.

Model	R	R^2	Adjusted R^2	Std. Error of the Estimate
1	0.939	0.883	0.878	5.6505

Table 2: The ANOVA table

Model	Sum of Squares	df	Mean Square	F	p-value
1 Regression	24721.674	4	6180.418	193.571	0.000
Residual	3288.636	103	31.929		
Total	28010.310	107			

Table 3: The Coefficients

Model 1	Unstd. Coeffs.		Std. Coeffs.		
Coefficients	Values(B)	Std. Error	Beta	t	p-value
(Constant)	-195.970	15.002		-13.063	.000
Amb. Temp.	5.730	.601	.457	9.536	.000
Cont. Temp.	2.096	.195	.536	10.760	.000
SinWT	1.063	.767	.049	1.387	.169
CosWT	-2.489	.869	-.102	-2.863	.005

The best fitted equation for predicting the absorbent temperature from ambient temperature, control temperature and time of the day for model (1) were determined using subroutine of SPSS and the related results are summarized in Tables 1,2 and 3. In Table 1, the coefficient of determination is very high with a value of 0.939 while Table 2 clearly shows that the multiple regression is significant. Table3 shows that both the circular and all the linear variables are highly significant.

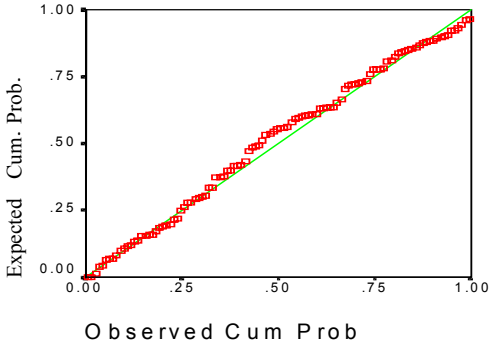

Figure 1: Normal p-p plot of regression standardized residuals

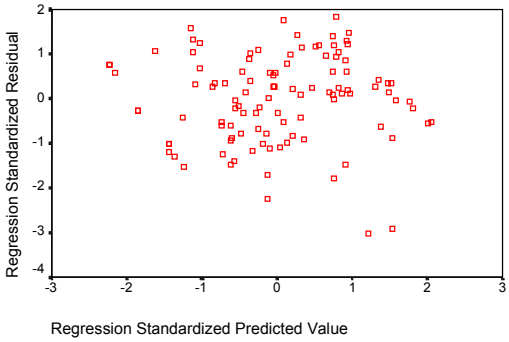

Figure 2: A scatter Plot for standardized predicted and residual values

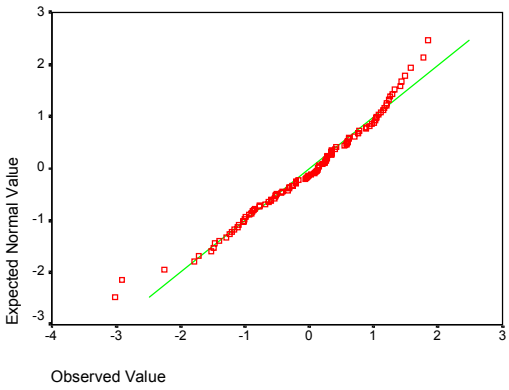

Figure 3: The normal Q-Q plot of standardized residuals

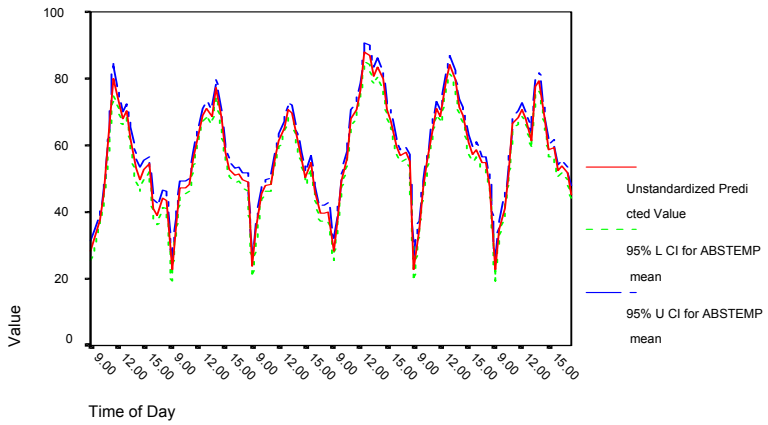

Figure 4: Plot of predicted absorbent temperature with 95% C.I.

The accuracy of the results and the distributional assumptions for model (1) were established using the plots in Figs. 1, 2 and 3. The fitted values of the regression curve are shown in Fig. 4, which exhibits an excellent fit.

4.2 Wind Energy

Among the renewable sources of energy, wind energy is not only freely and widely available, it is inexhaustible. With the use of routine meteorological records from Murtala Mohammed International Airport, Lagos Nigeria for the year 2003, data were obtained on the hourly, daily and monthly values of the wind direction and velocity. The records are primarily used for the determination of certain statistical characteristics of wind distribution. Other records include average daily wind energy, number of hours per month for which wind are adequate for windmill operation and the incidence and duration of calm period. Our interest is however on the prediction of wind energy generated (Y_i) given the velocity (x_{1i}), time (t_i) and wind direction (θ_i^o).

The fitted regression curve is due to the skewed dependency of wind energy on time as well known in environmental sciences and also exhibited in figure (5) using SPSS, the model (2) in equation (4) is enhanced hence,

$$Y_i = -110.542 + 40.185x_{1i} - 0.566t_i + 10.904\cos(\theta_i - 30\cos\theta_i) \qquad (7)$$

Finally, we convert the acrophase angle $\phi = 1.5178^o$ into hour of the day. Since 1.5^o stands for 6 minutes, the acrophase is estimated to be 6 minutes past midnight i.e. approximately midnight.

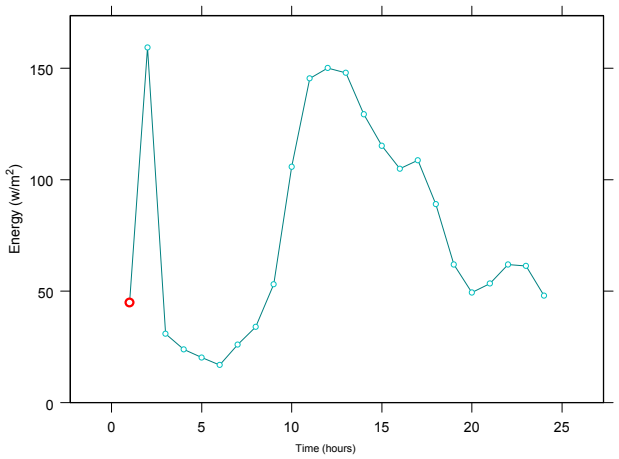

Figure 5: A plot of the wind energy vs time of the day

Table 4: Summary values.

Model	R	R^2	Adjusted R^2	Std. Error of the Estimate
2	0.992	0.984	0.982	6.2573

Table 5: The ANOVA table

Model 2	Sum of Squares	df	Mean Square	F	p-value
Regression	48685.561	3	16228.52	414.477	0.000
Residual	783.084	20	39.154		
Total	49468.645	23			

Table 6: The Coefficients

Model 2	Unstd.Coeffs.		Std. Coeffs.		
Coefficients	Values(B)	Std. Error	Beta	t	p-value
(Const.)	-110.542	5.763		-19.181	.000
Vel.(M/S)	40.185	1.193	0.991	33.683	.000
Time(hrs)	-0.566	0.192	-.086	-2.941	.008
Cos22wd	10.904	4.043	0.078	2.697	.014

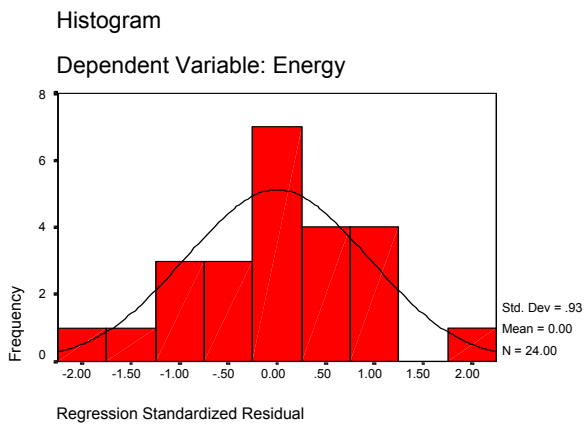

Figure 6: Histogram of regression standardized residual

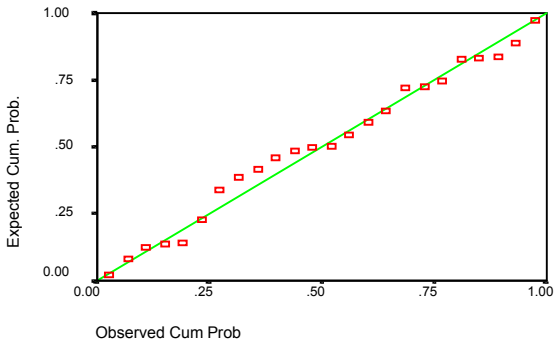

Figure 7: Normal p-p plot of regression standardized residual

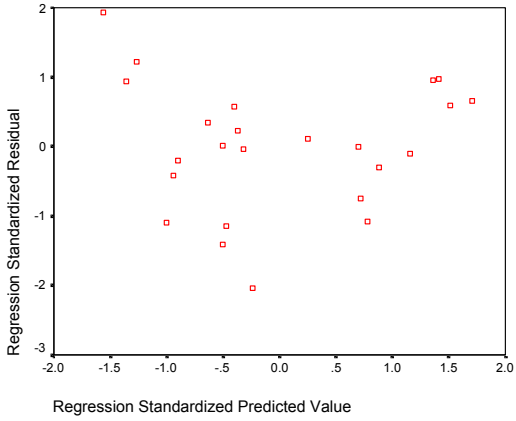

Figure 8: Scatter plot of standardized value and residual

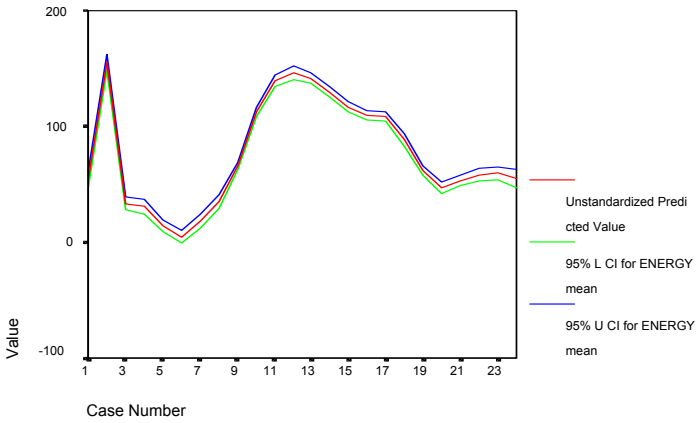

Figure 9: A plot of the fitted values with 95% C.I.

The accuracy of the results and the distributional assumptions for model (2) were determined using the plots in Figs 6, 7, and 8. The fitted values of the wind energy with 95% confidence interval are shown in Fig. .

5. Conclusions

The main emphasis of this work has been to fit regression models to possibly asymmetric angular linear variables. A simple cosine function was adopted for the solar energy data. However, an asymmetric model was selected for the wind energy data. Four diagnostic procedures were adopted to determine the nature of fit of the models. Those methods have shown that the models gave quite good fits to the data sets.

Acknowledgments

The contributions of TWAS-UNESCO and the host, Indian Statistical Institute, Kolkata, India, that enabled the second author to conduct this research work as a 2003-2006 TWAS associate are hereby acknowledged with thanks.

References

Batschelet, E. (1981) *Circular Statistics in Biology*, Academic Press, London.

Cheremisinoff, N.P. (1978) *Fundamentals of Wind Energy*, 1st edition, Science Publishers Inc., New York

De Wiest, F. and Della Fiorentina, H. (1975) Suggestion for a realistic definition of an air quality index relative to Hydro-Carbonaceous matter associated with airborne particles, *Atmospheric Environment*, 951-954.

Federov, V.V. (1972) *Theory of optimal experiments*, Academic Press, New York.

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

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

Jammalamadaka, S.R., and SenGupta, A. (2001) *Topics in Circular Statistics*, World Scientific Publishing, New Jersey.
Johnson, R.A and Wehrly, T. (1978) Some angular-linear distributions and related regression models. *Journal of the American Statistical Association*, **73**, 602-606.
Madia, K.V. and Jupp, P.E. (2000) *Directional Statistics*, J. Wiley, Chichester.
Laycock, P.J. (1975) Optimal Design: Regression models for directions, *Biometrika*, **62**, 305-311.
Ohta, T., Marita, M., and Mizoguchi, I. (1976) Local distribution of chlorinated Hydrocarbons in the ambient air in Tokyo, *Atmospheric Environment*, **10**, 557-560.
Seber G.A.F. and Wild C.J. (1988) *Nonlinear Regression*. John Wiley, New York.
SenGupta, A. (2004) On the constructions of probability distributions for directional data, *Bulletin of Indian Mathematical Society*, **96**, (2) 139-154.

Biographical sketches

Prof. Ashis SenGupta is a Professor at Indian Statistical Institute. He received his Ph.D. from Ohio State University, Columbus, in 1979 and has been a visiting faculty at various universities including Stanford University, University of Wisconsin-Madison, University of California-Santa Barbara, -Riverside, etc. He is President of IISA (India Chap.), Vice-President of FIM and Editor-in-Chief of JISPS. His recent research interests are mainly in the areas of multivariate analysis, directional data analysis, reliability and optimal multivariate inference.

Fidelis I. Ugwuowo is a senior lecturer in Department of Statistics, University of Nigeria, Nsukka. He was sponsored by World Bank to University of Ulster Coleraine, Northern Ireland, U.K to conduct his Ph.D. research, which was awarded by his University in 1996. He has been a visiting scientist at Indian Statistical Institute Kolkata, India under the auspices of TWAS-UNESCO Associateship scheme, which was awarded to him for 2000-2003 and later renewed for 2003-2006. His research interests include stochastic modeling of environmental, social and physical processes.