Modeling Continuous Non-Linear Data with Lagged Fractional Polynomial Regression

Kazeem Kehinde Adesanya[1], Abass Ishola Taiwo[2, *], Adebayo Funmi Adedodun[3] & Timothy Olabisi Olatayo[4]

[1]Department of Health Information Management, Ogun State College of Health Technology
Ilese-Ijebu, Nigeria

[2] Department of Mathematical Sciences, Olabisi Onabanjo University
Ago-Iwoye, Nigeria

[3] Department of Mathematical Sciences, Olabisi Onabanjo University
Ago-Iwoye, Nigeria

[4] Department of Mathematical Sciences, Olabisi Onabanjo University
Ago-Iwoye, Nigeria

**Corresponding author's email: taiwo.abass [AT] oouagoiwoye.edu.ng*

ABSTRACT— ***Fractional Polynomial regression is a form of regression analysis in which the relationship between the independent variable and the dependent variable is modelled as a 1/nth degree polynomial. Thus, this work is used to propose an extension of Fractional Polynomial Regression (FPR) term Lagged Fractional Polynomial Regression (LFPR) which is an alternative method to traditional techniques of analysing the pattern and degree of relationship between two or more continuous non-linear data. The coefficients of the proposed method were estimate using Maximum Likelihood Estimation method. From the results, the LFPR model indicated that for a unit increase in Evaporation, Humidity and Temperature there will be an increase in the millimeter of rainfall series on yearly basis. The value of coefficient of variation (R^2) for the LFPR and FPR were 99% and 77%. While the value of adjusted Coefficient of Variation (R^2) for LFPR and FPR were 96% and 75% respectively. Hence, the proposed method outperformed and adequately explained the variation in the dependent variable better than Fractional Polynomial Regression based on the values (R^2) and adjusted (R^2).***

Keywords— Continuous data, Fractional Polynomial, Lagged, Regression, Maximum Likelihood Estimation

1. INTRODUCTION

Measurements on a response variable (y_t) collected over time (t) are called time series data. Such data often display periodic behaviour that repeats itself every s time periods. Time series arise in any situation in which data are collected periodically. They are common throughout science, technology and the humanities [1]. Time series regression is a statistical method for predicting a future response based on the response history and the transfer of dynamics from relevant predictors.

Regression is often used when discussing and analysing the relationship between two or more variables. This relationship can then be used for various computations like forecasting future values or for computing if there exists a relation amongst the various variables or not [2]. There are various methods of Regression Analysis, these are Simple Linear Regression, Multivariate Linear Regression, Polynomial Regression, Multivariate regression while the linear relationship between one response variable and two or more independent variables is multiple regression. The general form of the linear regression equation can be written as:

$$y = \beta_0 + \sum_{i=1}^{k} \beta_i X_i + e_t \qquad (1)$$

where y is independent variable, β is constant, x is dependent variable and e_t is the error term. However, difficulties arise when the assumption of linearity is found to be untenable and more appropriate model is required. The class of appropriate model suitable for analysing non-linear is polynomial regression. Polynomial regression is a model used when the response and regression variables are modelled as *nth* order polynomial equation.

Fractional Polynomials was introduced by [3] and it is an extension of polynomial models for determining the functional form of a continuous predictor. These models are suited for nonlinear data and it has been used in many researches like [4, 5 ,6, 7, 8, 9] and many more.

By transforming *t*, a continuous variable in a linear model, the first-order Fractional Polynomial model is

obtained:

$$y_t = \beta_0 + \sum_{i=1}^{k} \beta_k X_{t-1}^{pk} + \varepsilon_t \qquad (2)$$

where y_t is the dependent variable, β_k is the fractional coefficient, X_{t-1} is the independent variable and ε_t is error term. k^{th} Degree polynomial can be written as

$$Y_{t=}\beta_{0+}\ \beta_1 X_{t-1}^{p1} + \beta_2 X_{t-2}^{p2} + \ \beta_3 X_{t-3}^{p3} + \ldots + \beta_k X_{t-k}^{pk} + \ \varepsilon_t \qquad (3)$$

where $P_1 = \frac{1}{r_1}, P_2 = \frac{1}{r_2} \ldots P_k = \frac{1}{r_k}$

A distributed lag model is a model for time series data in which a regression equation is used to predict current values of a dependent variable based on both the current values of an explanatory variable and the lagged (past period) values of this explanatory variable [10]. The starting point for a distributed lag model is an assumed structure of the form:

$$y_t = \alpha + \omega_0 x_t + \omega_1 x_{t-1} + \omega_2 x_{t-2} + \cdots + \omega_n x_{t-n} + \varepsilon_t \qquad (4)$$

where y_t is the value at time period t of the dependent variable *y*, α is the intercept term to be estimated, and ω_i is called the lag weight (also to be estimated) placed on the value *i* periods previously of the explanatory variable *x*.

[11] used a statistical method based on Fractional Polynomials for the investigation of potential predictive factors and observed that analysis of continuous factors with Fractional Polynomials extract more information from such factors, to improve the statistical power to detect influential variables and their interaction with treatment. The analysis of the periodic component allowed us to increase our knowledge about the patterns of behaviour of a given set of data; it contributes to the construction of better forecast of the considered time series. Hence, this work will be used to discuss Fractional Polynomial regression (FPR) and a lagged form termed Lagged Fractional Polynomial Regression (LFPR) models with the objective of deriving the method of obtaining the coefficients using Maximum Likelihood method. The performance of the LFPR will be determined using the Coefficient of variation and Adjusted Coefficient of variation based on the analysis of the relationship between some continuous non-linear climatic variables. The stability of the residual will be determined using the value of Durbin Watson Statistic.

2. MATERIALS AND METHODS

2.1 Parameter Estimation for Fractional Polynomial Regression

Given the Fractional Polynomial Regression as:

$$\hat{y} = \beta_0 + \beta_1 X_{1i}^{p1} + \beta_2 X_{2i}^{p2} + \mathrm{U_i} \qquad (5)$$

where $\hat{y}$ is the dependent variable, β_0, β_1 and β_2 are the fractional coefficients, X_1 and X_2 is the independent variables, $P_1 = \frac{1}{r_1}, P_2 = \frac{1}{r_2}$ and U_i is the error term.

The likelihood and logarithm functions of (5):

$$= \frac{-n}{2} log 2\pi - \frac{n}{2} log \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} [\mathrm{y_i} - \beta_0 + \beta_1 X_{1i}^{p1} + \beta_2 X_{2i}^{p2})]^2 \qquad (6)$$

Given that P > 1, P = 2, 3, …, $P_1 = \frac{1}{2}, P_1 = \frac{1}{3}$ and differentiate equation (6) with respect to σ^2, β_0 , β_1 and β_2 and set all to zero where to obtain

$$\sigma^2 = \frac{1}{n} \Sigma_{i=1}^{n} [(\mathrm{y_i} - (\beta_0 + \beta_1 X_{1i}^{\frac{1}{2}} + \beta_2 X_{2i}^{\frac{1}{3}})]^2 \qquad (7)$$

$$\Sigma_{i=1}^{n} \mathrm{y_i} = n\beta_0 - \beta_1 \Sigma_{i=1}^{n} X_{1i}^{p1} - \beta_2 \Sigma_{i=1}^{n} X_{2i}^{p2} \qquad (8)$$

$$\sum_{i=1}^{n} X_{1i}^{p1} y_i - \beta_0 \sum_{i=1}^{n} X_{1i}^{p1} - \beta_1 X_{1i}^{p\overset{2}{1}} - \beta_2 \sum_{i=1}^{n} X_{1i}^{p1} X_{2i}^{p2} \qquad (9)$$

$$\sum_{i=1}^{n} X_{2i}^{p_2} y_i - \beta_0 \sum_{i=1}^{n} X_{2i}^{p_2} + \beta_1 X_{1i}^{p1} X_{2i}^{p_2} + \beta_2 \sum_{i=1}^{n} X_{2i}^{p_2} \qquad (10)$$

By expressing equations 8 - 10 in matrix form to obtain

$$\begin{bmatrix} N & \sum_{i=1}^{n} X_{1i}^{1/2} & \sum_{i=1}^{n} X_{2i}^{1/3} \\ \sum_{i=1}^{n} X_{1i}^{1/2} & \sum_{i=1}^{n} X_{1i} & \sum_{i=1}^{n} X_{1i}^{1/2} X_{2i}^{1/3} \\ \sum_{i=1}^{n} X_{2i}^{1/3} & \sum_{i=1}^{n} X_{1i}^{1/2} X_{2i}^{1/3} & \sum_{i=1}^{n} X_{2i}^{1/3} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} y_{1i} \\ \sum_{i=1}^{n} X_{1i}^{1/2} y_{1i} \\ \sum_{i=1}^{n} X_{2i}^{1/3} y_{1i} \end{bmatrix} \qquad (11)$$

2.2 Parameter Estimation for Lagged Fractional Polynomial Regression

The Lagged Fractional Polynomial is defined as:

$$y_t = \beta_0 + \beta_1 x_{t-1}{}^{p_1} + \beta_2 x_{t-2}{}^{p_2} + \epsilon_t \qquad (12)$$

where y_t is the dependant variable $\beta_0, \beta_1, \beta_2$ are fractional coefficients, x_{t-1}, x_{t-2} are the independent variables and ϵ_t is the error term.

By taking the likelihood and logarithm functions of equation (12)

$$= \frac{-n}{2} log 2\pi \frac{-n}{2} log\sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} [y_t - (\beta_0 + \beta_1 x_{t-1}{}^{p_1} + \beta_2 x_{t-2}{}^{p_2})]^2 \qquad (13)$$

By differentiating equation (13) with respect to σ^2 , β_0 , β_1 and β_2 and set all to zero, to obtain

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} [y_t - (\beta_0 + \beta_1 x_{t-1}{}^{p_1} + \beta_2 x_{t-2}{}^{p_2})]^2 \qquad (14)$$

$$\sum_{i=1}^{n} y_t = n\beta_0 + \beta_1 \sum_{i=1}^{n} x_{t-1}{}^{p_1} + \beta_2 \sum_{i=1}^{n} x_{t-2}{}^{p_2} \qquad (15)$$

$$\sum_{i=1}^{n} x_{t-1}{}^{p_1} y_t = \beta_0 \sum_{i=1}^{n} x_{t-1}{}^{p_1} + \beta_1 \sum_{i=1}^{n} (x_{t-1}{}^{p_1})^2 + \beta_2 \sum_{i=1}^{n} x_{t-1}{}^{p_1} x_{t-2}{}^{p_2} \qquad (16)$$

$$\sum_{i=2}^{n} x_{t-2}{}^{p_2} y_t = \beta_0 \sum_{i=2}^{n} x_{t-1}{}^{p_1} x_{t-2}{}^{p_2} + \beta_1 \sum_{i=2}^{n} x_{t-1}{}^{p_1} x_{t-2}{}^{p_2} + \beta_2 \sum_{i=2}^{n} (x_{t-2}{}^{p_2})^2 \qquad (17)$$

By expressing equations 15 - 17 in matrix, then we have;

$$\begin{bmatrix} n & \sum_{i=1}^{n} x_{t-1}{}^{p_1} & \sum_{i=1}^{n} x_{t-2}{}^{p_2} \\ \sum_{i=1}^{n} x_{t-1}{}^{p_1} & \sum_{i=1}^{n} (x_{t-1}{}^{p_1})^2 & \sum_{i=1}^{n} x_{t-1}{}^{p_1} x_{t-2}{}^{p_2} \\ \sum_{i=2}^{n} x_{t-1}{}^{p_1} x_{t-2}{}^{p_2} & \sum_{i=2}^{n} x_{t-1}{}^{p_1} x_{t-2}{}^{p_2} & \sum_{i=2}^{n} (x_{t-2}{}^{p_2})^2 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} \sum y_t \\ \sum x_{t-1} y_t \\ \sum x_{t-2} y_t \end{bmatrix} \qquad (18)$$

3. RESULT AND DISCUSSION

The data used in this research article was obtained from [12]. The data contain yearly rainfall, humidity, evaporation and Temperature from 1985 to 2015. The time plot of all the series in figure 1 – 4 exhibited a continuous non-linear pattern and this inform the used of LFPR.

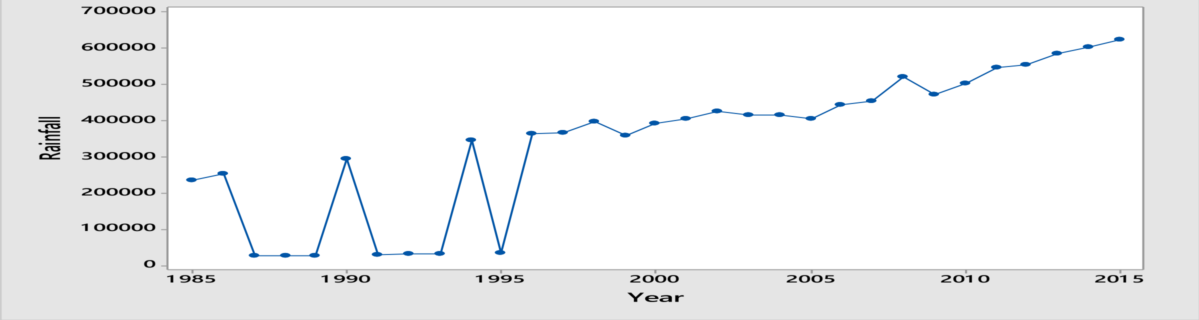

Figure 1. Time Series Plot of Rainfall from 1985 to 2015

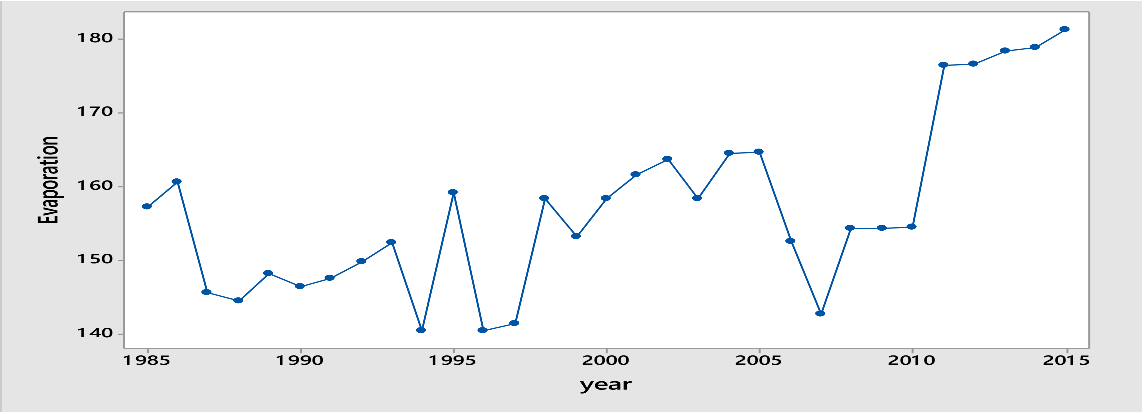

Figure 2. The Time Series Plot of Evaporation from 1985 to 2015

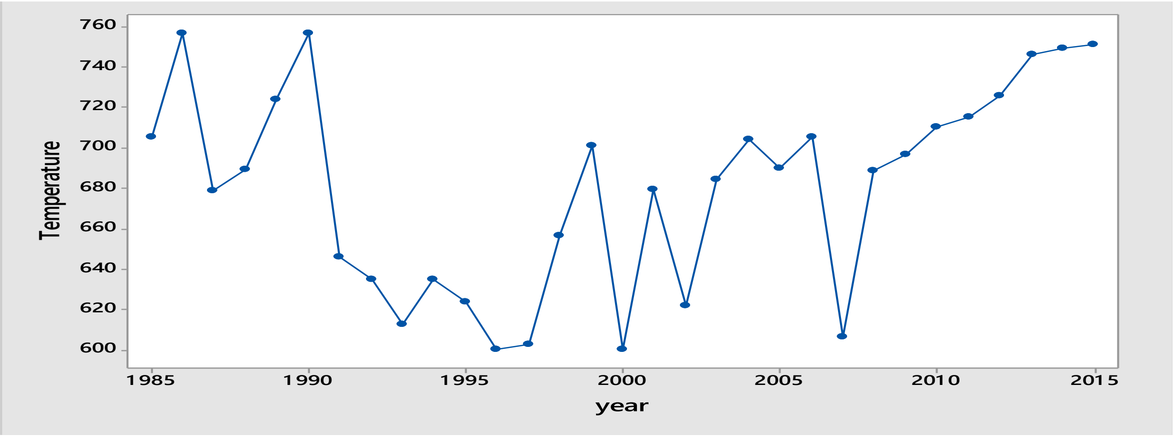

Figure 3. Time Series Plot of Temperature from 1985 to 2015

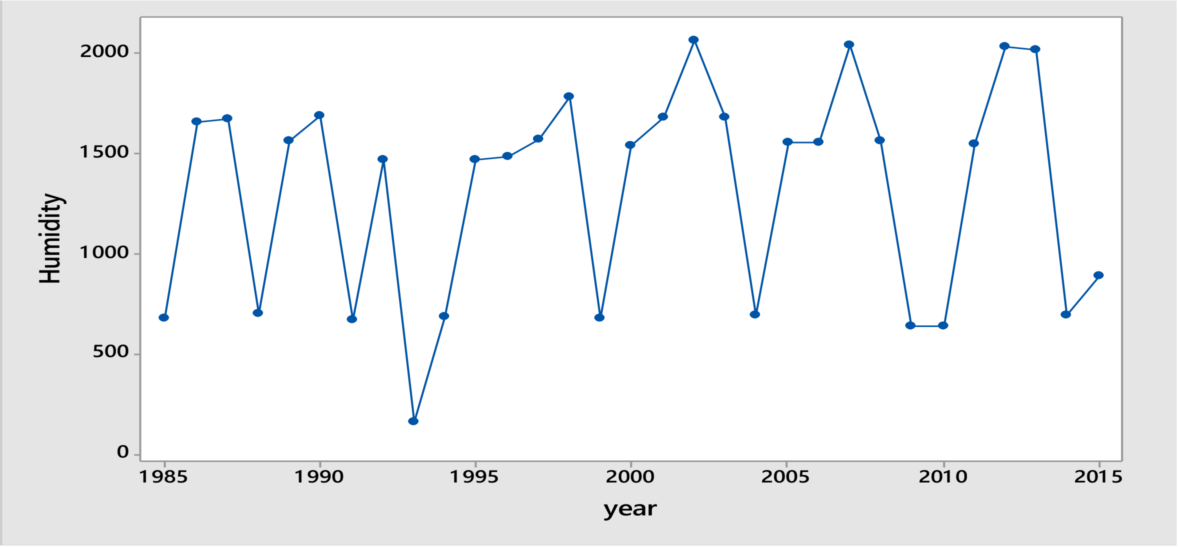

Figure 4. Time Series Plot of Humidity from 1985 to 2015

Given that; $P_1 = \frac{1}{r_1}$, $P_2 = \frac{1}{r_2}, \ldots, P_k = \frac{1}{r_k}$, the Fractional Polynomial Regression model used is;

$$\hat{y}_t = \beta_0 + \beta_1 X^{\frac{1}{r_1}} + \beta_2 X^{\frac{1}{r_2}} + \ldots + \beta_k X^{\frac{1}{r_k}}$$

where $r_1 = 2, r_2 = 3$

The Fractional Regression model obtained using Maximum Likelihood Estimation gives:

$$RAINFALL = -4367108.07073 + 800263.296473\, EVAPORATION + 43700.016154\, HUMIDITY + 5229.13782668\, TEMP$$

where $R^2 = 0.768255$,$Adjusted\ R^2 = 0.750610\ and\ Durbin\ Watson\ statistic = 1.66082.$

The Lagged Fractional Polynomial regression model is defined as above

$$\hat{Y}_t = \beta_0 + \beta_1 X_{t-1}^{\frac{1}{r_1}} + \beta_2 X_{t-2}^{\frac{1}{r_2}} + \ldots + \beta_k X_{t-k}^{\frac{1}{r_k}}$$

where $P_1 = \frac{1}{r_1}$, $P_2 = \frac{1}{r_2}, \ldots, P_k = \frac{1}{r_k}$ and $r_1 = 2, r_2 = 3$

The Lagged Fractional Polynomial Regression model obtained using Maximum Likelihood Estimation gives:

$$RAINFALL = 1011614 + 18814.90 EVAP_{t-1} + 43560.25\, HUM_{t-1} + 45319.73 TEMP_{t-1}$$

where $R^2 = 0.989245$, $Adjusted\ R^2 = 0.964310$ and $Durbin\ Watson\ statistic = 1.7635$.

The values of the Durbin Watson Statistic indicated the residual are not serially correlated and this implied the model is stable. The FPR and LFPR showed that the coefficients of Evaporation, Humidity and Temperature indicated that for every increase in Evaporation, Humidity and temperature there will be an increase in the millimeter of rainfall on yearly basis. In the FPR model, the value of R^2 showed that Evaporation, Humidity and temperature explained the variations in rainfall up to 77% and the value of $adjusted\ R^2$ showed the model is relatively a good fit. While in the LFPR model, the values of R^2 shows that Evaporation, Humidity and Temperature well explained the variations in rainfall up to 99% and the value of $adjusted\ R^2$ showed the model is a good fit with high level of predictive power. In hence, the LFPR model outperformed the FPR based on the values of $R^2\ and\ adjusted\ R^2$ when analysing the relationship between some continuous non-linear climatic variables.

4. CONCLUSION

This research article was used to discuss the performance of lagged extension of Fractional Polynomial regression. The coefficients of the models were estimated using Maximum likelihood method and the stability of the model was checked using Durbin Watson Statistic. The coefficients of FPR and LFPR model indicated that for a unit increase in the independent variables there will be an increase in rainfall. The LFPR model outperformed the FPR based on the values of $R^2\ and\ adjusted\ R^2$ when analysing the relationship between some continuous non-linear climatic variables. Conclusively, the LFPR indicated that the variations in the dependent was better explained by the predictors and even showed a better good fit with higher predictive power

5. REFERENCES

[1] Gasparrini, A. and Leone, M. “Attributable risk from distributed lag models.” BMC Medical Research Methodology vol. 14, pp. 55 - 60.

[2] Hubert G. “Statistical Analysis of Management Data” Second Edition. New York: Springer Publication, 2010.

[3] Royston, P. and Altman, D.G. “Regression using Fractional Polynomials of Continuous Covariates: Parsimonious Parametric Modelling”. Applied Statistics; vol. 43, pp. 429-467, 1994.

[4] Jansen, J.P. “Network Meta-Analysis of Survival Data with Fractional Polynomials”. BMC Medical Research Methodology. vol. 11, no 61, pp. 11-61, 2011.

[5] MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. “On the practice of dichotomization of quantitative variables”. Psychological Methods, vol. 7, 19–40, 2002.

[6] Royston, P, Altman, D. G. and Sauerbrei, W.” Dichotomizing continuous predictors in multiple regression:

A bad idea". Statistics in Medicine, vol. 25, no 1, pp.127-141, 2006.
[7] Royston, P. and Sauerbrei, W. "Multivariable Model-building: A Pragmatic Approach to Regression Analysis Based on Fractional Polynomials for Modelling Continuous Variables". Wiley; Chichesters, 2008.

[8] Hutcheson, G. D., Pampaka, M. and Williams. J. Enrolment, Achievement and Retention on Traditional' and 'Use of Mathematics' Pre-university Courses." Research in Mathematics Education vol. 13 no 2, pp. 147–168, 2011.

[9] Wainer, H. "14 Conversations about Three Things". Journal of Educational and Behavioral Statistics, vol. 35 no 1, pp. 5–25, 2010.

[10] Gasparrini A, Scheipl F., Armstrong B., and Kenward M.G. "A Penalized Framework for Distributed Lag Non-Linear Models". Biometrics vol. 73, pp. 938–94, 2017.

[11] Royston, P., and Sauerbrei. W. "A new measure of prognostic separation in survival data". Statistics in Medicine vol. 23, pp. 723–748, 2004.

[12] National Bureau of Statistics Annual Publication Published Reports (2016).