Modeling longitudinal data based on Fourier regression

Suparti¹, R Santoso¹, A Prahutama¹, A R Devi¹ and Sudargo²

¹Statistics Department, Universitas Diponegoro, Semarang, Indonesia
Jl. Prof. Soedharto, SH, Tembalang, Semarang 50275, Indonesia
²Universitas PGRI Semarang, Indonesia
E-mail: suparti702@gmail.com

Abstract. Regression analysis is an analysis in statistics to model the relationship between predictor variables and response variable. Regression analysis can be performed by two approaches; parametric and non-parametric models. Some commonly used estimators of nonparametric approaches are spline, local polynomial, kernel, wavelet, and Fourier. Fourier nonparametric regression is a regression based on Fourier series with cosine and sinus patterns. Regression analysis can be explored, not only for cross section data but also for longitudinal data. Longitudinal data is an observed data of some uncorrelated subjects, and each subject was observed for some periods. This research developed a nonparametric regression approach for longitudinal data by using Fourier series. One of the advantages of Fourier series is it combines additive that is able to overcome the data with recurrent and high fluctuations. The research use data with 3 subjects and 128 observations. The Fourier model by combining additive linear functions and cosine functions is more suitable for modeling repeated data that has an element of trend, as the 2nd sector.

1. Introduction

Regression analysis is an analysis in statistics to analyze predictor and response variable. Regression analysis can be carried out in parametric and nonparametric approach. Parametric analysis uses the Ordinary Least Square (OLS) approach [1], where the shape of the curve is known, and the residual is assumed to meet normal distribution, homoscedasticity, and non-autocorrelation. Meanwhile, the nonparametric approach is performed if the shape of the curve is unknown. Nonparametric model is more complex than parametric model. However, the nonparametric approach will produce a better estimation curve than parametric approach. Meanwhile, the parametric approach is easy to do but very strict with assumptions.

Regression analysis can be applied to cross sections, time series and longitudinal data. Time series data is data from a subject that is observed repetitively over time. While the cross section data is data from several subjects that only carried out one observation on each subject and mutually predictor. The combination of time series data and cross section forms longitudinal data [2]. Longitudinal data is data obtained from repeated observations of each subject at different time intervals. This data correlates on the same subjects and independent between different subjects. According to Wu and Zhang [2] there are several advantages to the study of longitudinal data such as being able to know individual changes, and requiring subjects that are not too many because of repeated observations. Moreover, the estimation is more efficient because it is performed together in all subjects and all observations.

Nonparametric regression approach can be done with several methods including kernel [3], local polynomial [4], spline [5], wavelet and Fourier[6]. The kernel approach and local polynomials are
bandwidth-based approaches, while the spline is a model that requires optimum knot points. The bandwidth and optimum knot is determined by using MSE (Mean Square Error), CV (Cross Validation) and GCV (Generalized Cross Validation) values [4]. In contrast to spline and local polynomials, determining the best model of Fourier series depends on constants in the sine or cosine variables [6].

Fourier regression model is a nonparametric regression approach with the Fourier series approach. Fourier series is a trigonometric polynomial approach that has flexibility in local data. Fourier series is very suitable for data patterns that have repetitive distributions or form sine and cosine curves [6]. In this research, nonparametric regression approach for longitudinal data will be studied by using Fourier series. The data used in this study are three sectors/groups of inflation expenditure in Indonesia, namely the foodstuffs group (1st sector); education and sports group (2nd sector); and transportation and financial services group (3rd sector).

2. Nonparametric regression
Regression is a method that describes the relationship between the predictor variable and the response variable [1]. The regression method can be approached in a nonparametric manner if the shape of regression curve is unknown. Nonparametric models provide high flexibility and require no assumptions like parametric models. Given the multiple linear regression model as follows [3]:

\[ y_i = \beta_0 + \beta_1 t_i + \beta_2 t_i^2 + K + \beta_p t_{ip} + \varepsilon_i \]  

(1)

with \( y_i \) is a response variable while \( t \) is a predictor variable. In general, equation (1) can be changed in form

\[ y_i = f(t_i) + \varepsilon_i \]  

(2)

\( f(t) \) is a regression function curve which its shape is unknown with \( x_i \) is a predictor variable. The function curve of \( f(t) \) is assumed to be smooth in certain function spaces.

3. Fourier regression Model
Fourier regression model is obtained through Fourier series in the form of sine and cosine functions. Bilodeau (1992) provides nonparametric regression modeling with a combination additive of linear function and cosine function [6]. This combination is expected to separate data trends and data fluctuations. By taking a combination additive of linear and cosine function as in the followings [7]:

\[ f(t) = \frac{1}{2} \alpha_0 + \gamma t + \sum_{k=1}^{K} \alpha_k \cos kt \]

(3)

If \( f(t) \) is a regression curve in equation 2, then the curve is approached by using the Fourier series as follows:

\[ f(t_i) = \frac{1}{2} \alpha_0 + \gamma t_i + (\alpha_{i1} \cos t_i + \alpha_{i2} \cos 2t_i + \cdots + \alpha_{iK} \cos Kt_i) \]  

(4)

If equation 4 is written in the form of matrix, it becomes

\[ y = f(t) + \varepsilon \]  

with \( f(t) = A \theta \)

\[ A = \begin{bmatrix} 1 & t_1 & \cos t_1 & \cos 2t_1 & \cdots & \cos Kt_1 \\ 1 & t_2 & \cos t_2 & \cos 2t_2 & \cdots & \cos Kt_2 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & t_n & \cos t_n & \cos 2t_n & \cdots & \cos Kt_n \end{bmatrix} \]

\[ \theta = [\phi \ \gamma \ \alpha_{i1} \ \alpha_{i2} \ \cdots \ \alpha_{iK}]^T \]  

with \( \phi = \frac{n}{2} \alpha_0 \)
By using ordinary least square, the Fourier regression estimator is

$$\hat{\theta} = (A^T A)^{-1} A^T y$$  \hspace{1cm} (5)

The equation (5) is a nonparametric regression estimator by using the Fourier series approach.

4. Longitudinal data

Longitudinal data are data with several subjects taken and repeated observation for each subject. The subjects in longitudinal data are assumed to be mutually independent, but the observations within subject are interdependent so that there is a correlation between their observations [2]. The structure of longitudinal data is presented in Table 1.

| i-th subject | j-th observations | ij-th response |
|--------------|-------------------|---------------|
| 1st subject  | 1                 | y_{1j}        |
|              | 2                 | y_{2j}        |
|              | \vdots            | \vdots        |
|              | n                 | y_{nj}        |
| 2nd subject  | 1                 | y_{1j}        |
|              | 2                 | y_{2j}        |
|              | \vdots            | \vdots        |
|              | n                 | y_{nj}        |
| \vdots       | \vdots            | \vdots        |
| m-th subject | 1                 | y_{mj}        |
|              | 2                 | y_{m2}        |
|              | \vdots            | \vdots        |
|              | n                 | y_{mn}        |

5. Result and discussion

Nonparametric regression modeling for longitudinal data is defined as following equation

$$y_{ij} = f(t_{ij}) + \varepsilon_{ij} \text{ for } i = 1, 2, \ldots, m \text{ and } j = 1, 2, \ldots, n$$  \hspace{1cm} (6)

with $y_{ij}$ is the response variable for the i-th subject and j-th observation, $t_{ij}$ are the i-th predictor variable and j-th observation, $\varepsilon_{ij}$ is the i-th subject residual and j-th observation. $f(t_{ij})$ is a nonparametric regression curve for longitudinal data with unknown shape. The $f(t_{ij})$ curve is approached by using Fourier series as follows:

$$f(t_{ij}) = \frac{1}{2} \alpha_{0i} + \gamma_i t_{ij} + \sum_{k=1}^{K} \alpha_{ki} \cos k t_{ij}$$

Equation 6 is written in the form of a matrix as:

If \( y = f(t) + \varepsilon \) then

$$f(t) = \begin{bmatrix} f(t_{11}) & f(t_{12}) & \ldots & f(t_{1n}) \\ f(t_{21}) & f(t_{22}) & \ldots & f(t_{2n}) \\ \vdots & \vdots & \ddots & \vdots \\ f(t_{m1}) & f(t_{m2}) & \ldots & f(t_{mn}) \end{bmatrix} \quad y = \begin{bmatrix} y_{11} \\ y_{12} \\ \vdots \\ y_{mn} \end{bmatrix}$$

$$f(t_{ij}) = \frac{1}{2} \alpha_{0i} + \gamma_i t_{ij} + (\alpha_{i1} \cos \theta_{ij} + \alpha_{i2} \cos 2t_{ij} + \alpha_{iK} \cos Kt_{ij})$$

If \( f(t) = A\theta + \varepsilon \) then \( Y = A\theta + \varepsilon \)

Where

\[ Y = A\theta + \varepsilon \]
\[
Y = \begin{bmatrix}
Y_1 \\
Y_2 \\
\vdots \\
Y_m
\end{bmatrix}
A = \begin{bmatrix}
A_1 & 0 & \cdots & 0 \\
0 & A_2 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & A_m
\end{bmatrix}
\theta = \begin{bmatrix}
\theta_1 \\
\theta_2 \\
\vdots \\
\theta_m
\end{bmatrix}
\varepsilon = \begin{bmatrix}
\varepsilon_1 \\
\varepsilon_2 \\
\vdots \\
\varepsilon_m
\end{bmatrix}
\]

or

\[
Y_i = A_i \theta_i + \varepsilon_i
\]

where

\[
\begin{bmatrix}
Y_{i1} \\
Y_{i2} \\
\vdots \\
Y_{in}
\end{bmatrix} = \begin{bmatrix}
x_{i1} \cos \theta_{i1} \cos 2 \theta_{i1} \cdots \cos K \theta_{i1} \\
x_{i2} \cos \theta_{i2} \cos 2 \theta_{i2} \cdots \cos K \theta_{i2} \\
\vdots \\
x_{in} \cos \theta_{in} \cos 2 \theta_{in} \cdots \cos K \theta_{in}
\end{bmatrix}
\]

The estimation of Fourier regression model use OLS (Ordinary Least Square) is \(\hat{\theta} = (A^T A)^{-1} A^T Y\) or \(\hat{\theta}_i = (A_i^T A_i)^{-1} A_i^T Y_i\) and \(\hat{Y}_i = A_i \hat{\theta}_i\).

This research studies the Fourier regression modeling for longitudinal data by using 3 sectors/groups of Indonesia year on year inflation expenditure data from January 2007 to August 2017 from Bank Indonesia. They are the foodstuffs group (1st sector 1); education and sports group (2nd sector); and transportation, communication, and financial services group (3rd sector). The scatter plot data are presented in figure 1, and the descriptive statistics of 3 sectors of inflation expenditure in Indonesia data are described in table 2.

**Figure 1. Scatter plot data**

**Table 2. Descriptive statistics of 3 sectors of Inflation in Indonesia**

| Sector   | Minimum | Maximum | Mean  | Variance |
|----------|---------|---------|-------|----------|
| 1st sector | 1.452   | 20.020  | 8.810 | 18.638   |
| 2nd sector | 2.747   | 10.410  | 4.870 | 3.237    |
| 3rd sector | -6.852  | 16.330  | 3.715 | 27.193   |

Based on table 2, the 3rd sector data have high variability with a minimum value of -6.852 and a maximum value of 27.193. They have the largest range than the others. The 2nd sector data have low variability with a minimum value of 2.747 and a maximum value of 3.237. They have the smallest range than the others.

The next step in Fourier regression modeling for longitudinal data is to determine the value of \(K\) optimum. The value of \(K\) optimum is obtained based on trial and error, where \(K\) is a positive integer. The \(K\) value for each subject is the same. Table 3 show the MSE (Mean Square Error) and R-square values for some \(K\) values.
As the value of K is greater, the value of MSE is smaller, and the value of R square is greater. If the model is partitioned for each subject, then each subject can be described in the estimation curve. Figure 2 shows the estimated curve of each subject for the Fourier regression of longitudinal data with K = 120 as follows:

![Figure 2. Scatter plot data and estimation Fourier model with K=120](image)

Based on Figure 2, all of sectors are estimated correctly by using Fourier regression for K = 120. With a value of K = 120, it means that the parameters used are 122 pieces. It is too many parameters. The following is the Fourier model for the three inflation sectors in Indonesia with K=120:

\[
\begin{align*}
\hat{y}_{1j} &= 6.145 + 0.064 \cdot t_{1j} - 1.160 \cos t_{1j} - 0.273 \cos 2t_{1j} + \ldots + 0.023 \cos 120t_{1j} \\
\hat{y}_{2j} &= 17.524 - 0.137 \cdot t_{2j} - 0.315 \cos t_{2j} - 0.377 \cos 2t_{2j} + \ldots + 0.092 \cos 120t_{2j} \\
\hat{y}_{3j} &= 39.470 - 0.431 \cdot t_{3j} + 0.281 \cos t_{3j} - 0.704 \cos 2t_{3j} + \ldots + 1.701 \cos 120t_{3j}
\end{align*}
\]

Models involving large K values are not effective because they will involve many parameters in the model so that the model is more complex. A good model has a small K with small MSE and a large of R square. In modeling this data, the criteria can be achieved at large K values. We choose the value of R square above 0.8, so the value of K = 110 in the first sector model, K = 50 in the second sector model and K= 90 in the third model. But the K value in each subject must be equal so that we choose K= 110. With the choice of a K value is 110 indicates that modeling with the Fourier method in this data is not efficient.

Of the three models, modeling data with an up and down pattern followed by a trend tends to be better in Fourier series with a combination of additive linear functions and cosine functions. As shown in the second sector data modeling. Figure 3 is the Fourier model for the three inflation sectors in Indonesia with K=50.
6. Conclusion
The nonparametric regression model for longitudinal data by using the Fourier series generated estimation curves like OLS, but the values in the predictor variables contain cosine elements. Based on data processing that forms fluctuations, spreads and extreme patterns, it showed that the estimation of Fourier regression for longitudinal data generated a parsimony curve and generated high R-square. Based on this research, the Fourier model by combining additive linear and cosine functions is more suitable for modeling repeated data that has an element of trend.

Acknowledgment
The author would like to thank the Directorate of Research and Community Service of Ministry of Research, Technology and Higher Education for funding this research with PTUPT scheme (Higher Education Applied Research) for the year 2018 with contract no: 101-124 / UN7.P4.3 / PP / 2018 on February 5, 2018. Thank you for the Statistics Department of Diponegoro University.

References
[1] Draper N R and Smith H 1992 *Applied regression analysis 2nd Edition* (New York: Marcel Dekker).
[2] Wu H and Zhang J T 2006 *Nonparametric regression methods for longitudinal data analysis* (New Jersey: A John-Wiley and Sons Inc).
[3] Hardle W 1990 *Applied Nonparametric Regression* (Cambridge University).
[4] Eubank R L 1988 *Spline Smoothing And Nonparametric Regression* (New York: Marcel Dekker).
[5] Wahba G 1990 *Spline Models For Observation Data* (Amidson Pennsylvania: University of Winsconsin).
[6] Bilodeau M 1992 *The Canadian Journal of Statistics* 3 257.
[7] Xu X and Shang X 2017 *Stat papers* 58 811.