Varying coefficient model of longitudinal data of dengue fever in Bandung city

Bertho Tantular\(^1,2\), Yudhie Andriyana\(^2\), Budi Nurani Ruchjana\(^3\)

\(^1\) Ph.D Student of Mathematics, Universitas Padjadjaran, Sumedang, Indonesia
\(^2\) Department of Statistics, Universitas Padjadjaran, Sumedang, Indonesia
\(^3\) Department of Mathematics, Universitas Padjadjaran, Indonesia

Corresponding email: bertho@unpad.ac.id

Abstract. Studies on infectious diseases, especially those transmitted through intermediaries such as dengue fever, are sometimes seen as growth patterns over time. The measurement is repeated over time and hence we have a longitudinal data structure. We are interested in analyzing factors that influencing the dengue fever rate. A Modelling technique is used to determine the effects of covariates, but there are variable where the effects of covariates may also change over time when the effects of covariates change over time. Varying coefficient model is an alternative to model such situation to determine these effects. In general, the model will form a pattern, both linear and non-linear. As the objective function, we propose to use a P-Splines Quantile Objective function such that a flexible dengue pattern can be developed. The proposed technique will be implemented to dengue fever cases in Bandung City. The density of Bandung City is a factor that causes the spread of infectious diseases. In addition, various risk factors for the spread of dengue cases in Bandung are also taken into account, such as rainfall, temperature or temperature and humidity. Another factor that can be a covariate is the lifestyle factor of the people of Bandung, such as clean and healthy living behaviour and healthy house. The result shows existence of a non-linear pattern, and the fluctuations in the incident value that changes over time. There are changes in the values of the intercept and slope coefficients over time for each level of the quantile.

Keywords: longitudinal data; varying coefficient model; P-splines; quantile regression

1. Introduction
Data obtained from repeated measurements has a longitudinal data structure. Longitudinal data are generally used when we want to be examined regarding the correlation of individual changes over time. In modelling for longitudinal data, responses are obtained when units are determined at different times. One of the things that can be obtained from longitudinal data is the chronology of a sequential response. The main purpose of longitudinal data analysis is to examine the effects of covariates from each level in general on responses and effects in response changes over time. One important benefit of longitudinal data analysis is that it is possible to separate the cross-sectional effect and the repeated
measurement effect. In addition, in the analysis of longitudinal data, it can also be observed the diversity between units and the diversity in changes between time. Diversity not obtained from observed variates results in dependencies between responses as well after the covariate is controlled. This violates the assumptions in the ordinary linear regression model and must be overcome to avoid mistakes in inference.

Studies on infectious diseases, especially those transmitted through intermediaries such as dengue fever are generally seen the development of patterns of spread over time. The measurements of this study were repeated so that the data structure formed from this study is a longitudinal data structure.

In general, longitudinal data can be divided into balanced data and unbalanced data. A longitudinal data is called balanced data if all units are measured at the same time \( t_{ij} \), with \( i = 1, 2, \ldots, n \) and \( j = 1,2, \ldots, T \) where \( n \) is the number of measurement units and \( T \) is the repeated measurement. Instead a data is called unbalanced data if units are measured at different times \( t_{ij} \), with \( i = 1, 2, \ldots, n_j \) and \( j = 1,2, \ldots, T_i \). In the case of balanced data, it can also be seen as multivariate data with responses treated as different variables over time (repeated measures design). Whereas in the case of unbalanced data the use of repeated measures design will be problematic. One approach that can be done on unbalanced data is to use a mixed model that can accommodate differences in time for longitudinal data [12].

Some researchers have made several approaches of longitudinal data analysis. A common approach is to use repeated measure design, generally used in experimental data. In the economic and social fields data is usually obtained through surveys so that the longitudinal data analysis approach uses panel data analysis. For both approaches, it must be defined in advance whether the fixed model or random model will be used in the analysis. A more general approach is the linear mixed model. In a linear model a mixture of both fixed and random effects is defined in one model [7]. [8] made an approach of longitudinal data analysis using a multilevel model. Multilevel models were introduced by [5] which mentioned that they can overcome various problems that arise from data with a hierarchical structure, including longitudinal data. In multilevel models, hierarchical structures are defined as levels. In general, the level used is not limited but in the longitudinal data the level used is only two, namely at the lowest level, namely time called Level 1 and a higher level, namely individuals called Level 2. Multilevel models in addition to determining diversity between groups can also show correlation between the two observations which in the other model are assumed to be absent. Multilevel models can also measure the interactions that might occur between variables at different levels.

Various approaches have been carried out for modelling of longitudinal data using linear model patterns. While some data actually show a pattern that is not linear even changes from time to time, as on the case of dengue fever in Bandung City which actually has a pattern that is not linear. The pattern of data on cases of dengue fever in Bandung City changes with time. This can occur due to changes in the coefficient value of the model built, in this case there is a change in the coefficient over time. Various approaches to longitudinal data analysis that have been explained in the background section assume that the coefficient of the model built is a constant. In certain circumstances this may not be appropriate because the coefficient of the model built may be a particular function of the variable under study. In the case of longitudinal data, the coefficient of the model may change with time or in other words is a function of the time variable for each unit of measurement repeated.

Varying coefficient modelling (VCM) introduced by [6] suggests that the effects of covariates can differ between the values of other variables. The special case in VCM is for longitudinal data the effects of covariates can differ between time variables.

2. Varying Coefficient Model

As explained in the previous section, longitudinal data analysis can be done through various approaches. The general model for longitudinal data structures is mixed-effect regression, with the following mathematical equation
\[ y_{ij} = \beta^{(0)} + \beta^{(1)}_i x_{ij}^{(1)} + \cdots + \beta^{(p)}_i x_{ij}^{(p)} + \varepsilon_{ij} \]  

(1)

where

\[ \beta^{(k)}_i = \beta^{(k)} + u_j \]

for \( k = 1, 2, \ldots, p \), \( y \) is response, \( x^{(k)}_i \) is explanatory variable, \( \beta^{(k)} \) is coefficient regression and \( u_j \) is the error term [7].

These approaches are parametric regression methods. The parametric approach has limitations on the relationship between the average longitudinal response and the covariate. This limitation results in the simplification of the model so that it is not flexible in the relationship between response rates and covariates in various longitudinal studies, [4].

As previously explained, the VCM can determine the effects of different covariates among the values of the time variable [6]. The general model equation for VCM is as follows

\[ y_{ij}(t_{ij}) = \beta^{(0)}(t_{ij}) + \beta^{(1)}_i(t_{ij}) x_{ij}^{(1)}(t_{ij}) + \cdots + \beta^{(p)}_i(t_{ij}) x_{ij}^{(p)}(t_{ij}) + \tilde{\varepsilon}_{ij}(t_{ij}) \]  

(2)

where \( Y(t_{ij}) \) is the observation of the response variable at time \( t_{ij} \), \( X^{(k)}(t_{ij}) \) is vector covariate at time \( t_{ij} \) and \( \beta^{(k)}(t_{ij}) \) is \( k \)-th parameter at time \( t_{ij} \). In this case time \( (T) \) is a variable.

In equation (2) there is an error which is a common error structure. In general, this error can be modelled as follows:

\[ \tilde{\varepsilon}_{ij}(t_{ij}) = V(X(t_{ij}), t_{ij}) \varepsilon_{ij}(t_{ij}) \]  

(3)

where \( V(X(t_{ij}), t_{ij}) \) is a non-negative function. Through this error model, VCM is divided into three criteria, namely: homoscedastic models [1], simple heteroscedastic models [2] and general heteroscedastic models [3] These three models use a non-parametric approach to estimating the parameters, which called P-splines quantile regression.

3. Quantile Regression

Model (2) can generally be solved by using the average approach of the responses. This approach requires very strict assumptions, so that some deviations related to data such as outliers or non-linear data patterns are difficult to accommodate. A more robust approach is Median regression [10]. In median regression data divided into two parts, half of which are above the median and the rest are below the median. A more general approach to median regression is quantile regression. In general, quantile regression will produce a more efficient estimator so that it will be better used when the data distribution is asymmetrical or solid at the end of the data distribution [2].

Model (1), when \( \varepsilon \) is assumed to have a distribution function denoted by \( F \), the \( \tau \)-th quantile of this error (\( \varepsilon \)) is

\[ F_{\tau}^{-1}(\tau) = \inf \{ u : P(\varepsilon \leq u) \geq \tau \} \]  

(4)

where \( u \) is a quantile point of the regression error model (2). The quantile curve equation for \( Y \) conditional to \( X \) can be written as follow:

\[ q_{\tau}(Y \mid X) = [\beta_0 + F_{\tau}^{-1}(\tau) + \beta_1 X^{(1)} + \cdots + \beta_p X^{(p)} = X^T \beta] \]  

(5)

where \( \beta^\tau = (\beta_0 + F_{\tau}^{-1}(\tau) + \beta_1 + \cdots + \beta_p)^T \). The estimator \( \hat{\beta}^\tau \) can be obtained by minimizing the empirical equation of this objective function:

\[ \frac{1}{n} \sum_{i=1}^{n} \rho_{\tau}(Y_i - X_i^T \hat{\beta}) \]  

(6)
where $X_i^T = (1, X_i^{(1)}, \ldots, X_i^{(p)})^T$ is the $i$-th row of matrix $X$ (in this case state $i$-th observation). After equation (6) is minimized an estimator will be obtained $\hat{\beta}^T$ then this estimator is used to obtain the following conditional quantile function estimator:

$$
\hat{q}_\tau(Y \mid X) = X^T \hat{\beta}
$$

(7)

In minimizing the objective function in equation (6) an explicit solution is not obtained to get an estimator $\hat{\beta}^T$. So to get the regression coefficient estimator must go through an optimization procedure. Based on the proposal of [11], Linear Programming is used to obtain the optimal regression coefficient estimates.

The data analysis process begins by building a regression model that is more flexible than the usual regression model. As explained in the previous section, using the varying coefficient model (VCM) as in equation (2) defined the response variable $Y(t)$ is the number of infectious disease cases and the covariate $X^{(k)}(t)$ for $k = 1, \ldots, p$ are variables that are seen to influence the spread of infectious diseases, such as risk factors for the spread of disease, environmental factors and other factors that are likely to continue to develop along with the data collection process.

To estimate the model, the P-splines quantile regression method is used. In general, the model coefficient estimators are:

$$
\beta_k(t_j) \approx \sum_{i=1}^{m_k} \alpha_{ki}B_k(t_j,v_k)
$$

(8)

where $B_k(t_j,v_k)$ is B-splines with degree $v_k$ and $\alpha_{ki}$ is an unknown coefficient that should be estimated.

There are several approaches to obtaining estimated model coefficients. In general, researchers use the L2-loss function to obtain a regression mean. However, this approach has several weaknesses, including tightness to assumptions and cannot provide a complete picture of the distribution of responses. Therefore, this research will use the use of the weighted L1-loss function, known as the pinball loss function or quantile loss function with the form:

$$
\rho_\tau(z) = \begin{cases} 
\tau z & \text{if } z > 0 \\
-(1-\tau)z & \text{otherwise}
\end{cases}
$$

(9)

where $\tau \in [0,1]$, in order to obtain the objective function of the quantile as follows

$$
\sum_{i=1}^{n} \sum_{j=1}^{N_i} \rho_\tau(Y(t_j) - \sum_{k=1}^{p} \sum_{i=1}^{m_k} \alpha_{ki}B_k(t_j,v_k)X_i^{(k)}(t_j))
$$

(10)

through the minimization of the quantile objective function, an estimated B-splines coefficient ($\alpha_{ki}$) will be obtained, so that:

$$
\hat{\beta}_k(t_j) \approx \sum_{i=1}^{m_k} \hat{\alpha}_{ki}B_k(t_j,v_k)
$$

(11)

and can be determined estimates for the conditional quantile of each level $\tau$ as follows:

$$
\hat{q}_\tau(Y(t) \mid X(t),t) = \hat{\beta}_0 + \sum_{k=1}^{p} \sum_{i=1}^{m_k} \hat{\alpha}_{ki}B_k(t,v_k)X_i^{(k)}(t_j)
$$

(12)

Furthermore, from these results can be described the pattern of spread of infectious diseases based on time ($t$).
4. Data Analysis and Result

The data used in the initial stages is data regarding cases of dengue fever in Bandung City. The data used are secondary data obtained from the Bandung City Health Office. The various covariates used are larvae free numbers, healthy houses, density, temperature, rainfall, humidity and wind speed. Data are measured monthly from 2012 to 2016 with the observation unit of 30 sub-districts in Bandung.

The response variable used in this study is the incident rate (IR) of the dengue fever case in Bandung, while the covariate consists of two parts: larva free numbers, healthy houses and densities measured annually for each sub-district and temperature, rainfall, humidity and wind speed measured monthly for the entire city of Bandung. For these circumstances, some covariates in their measurements vary by location (in this case sub-districts) whereas some measurements vary by time (in this case monthly).

4.1 Exploring the data

Before conducting data analysis, the exploration process is carried out first. In the first stage, data fluctuation plots were made for each observation unit, namely the sub-district:

![Figure 1. Scatterplots of data fluctuations by sub-district](image)

From these results it can be seen that the fluctuation of data for each sub-district is different, this shows that the effects of the sub-district must be considered for when analysed this data. Furthermore, if Figure 1 is combined into one it will look like in Figure 2. The results from Figure 2 show that there are annual patterns that can be analysed in the sense that in general for all districts in the city of Bandung, pattern fluctuations occur based on monthly measurements.

![Figure 2. General Line Plots](image)
Then the evolution of mean plot is made. This plot illustrates the profile of population development that is relevant over time. Figure 3 shows that there was a decrease in the dengue fever’s IR in Bandung over time.

![Figure 3. Evolution of Mean](image)

Besides evolution of mean to build a longitudinal model, evolution of variance is also needed. Evolution of variance describes the development of population variations over time. The results can be seen in Figure 4 below. It is clear that variance always changes with time.

![Figure 4. Variance profile](image)

Furthermore, monthly plots are made for 2012 to 2016 to see the general pattern of the data. If the plot is separated for each observation unit (district), a profile plot for each sub-district will be obtained, the results of which are presented in Figure 5 below:
Based on these exploration results, the Bandung dengue fever case data has a pattern that changes with time. In addition to modelling in this case one must pay attention to differences in measurement of the covariate. Furthermore, the data were analysed using a varying coefficient model using the P-splines quantile regression method.

5. Result and Discussion
Modelling the dengue fever case in Bandung using the varying coefficient model using the P-splines quantile regression method in the analysis using R software packages namely QregVCM [13]. P-splines quantile regression in VCM produces a regression line as much as the quantile level used (Andriyana, 2015). Patterns that are formed from various quantile regression lines divide the data into several groups (depending on how many quantile levels are used). In this study the number of knots used was 5 knots with a degree of splines of 3 (cubic splines). In the calculation process, the smoothing parameter value (λ) is optimized using the SIC value at each of the quantile levels.

The results of the interim analysis of dengue fever data in Bandung using only a covariate, namely a healthy house, with a quantile P-splines regression model can be seen in Figure 6 below:
Figure 6. Quantile Plot of VCM Dengue Fever Data in Bandung City

Figure 6 shows the estimated values of the incidence rate of dengue cases in Bandung from January to December for the 0.25th quantile, the 0.33th quantile, the 0.5th quantile, the 0.66 quantile and the 0.75th quantile. Based on these results, the highest incident value was in May while the lowest incident value was in November.

Furthermore, in VCM the regression coefficient changes with time. The following results show changes in the intercept and slope regression coefficients for the 25th quantile, the 50th quantile and the 75th quantile.

Figure 7. Intercept and Slope for quantile 0.25, 0.5 and 0.75

Based on the results in Figure 7, the estimated intercept and slope values for each quantile change over time. The pattern of the intercept and slope coefficients at the quantile to 0.25 and the quantile to
0.5 are relatively similar although there are differences in the values. But for the quantile to 0.75 the intercept and slope coefficients have different patterns from other quantile values.

Furthermore, the pattern of two observation units that have quite different values are taken, namely Andir and Buahbatu sub-districts. The results can be seen in Figure 8 below.

![Figure 8. Pattern of District Andir (Black) and Buahbatu (Orange)](image)

From the results in Figure 8 it can be seen that the Andir sub-district is always relatively between the 0.25th quantile and the 0.5th quantile except in March, April and June are below the 0.25 quantile, while December is above the 0.5th quantile although it is still below the quantile -0.75. This is different from Buahbatu Sub-district, from January to July and November always above the 0.75 quantile, while August to October and December are between the 0.5 quantile and the 0.75 quantile.

6. Conclusion and Discussion
Data regarding the incidence of dengue cases in Bandung City that are measured monthly from 2012 to 2016 with the observation unit is the district has a longitudinal data structure. Based on the results of exploratory data on the incidence of DHF cases in Bandung, it can be analysed by varying coefficient models using P-splines quantile regression. Preliminary results using the R software package namely QregVCM, SparseM and quantreg that the data show the existence of a non-linear pattern, and the fluctuations in the incident value that changes with time. The same thing can be seen in the changes in the values of the intercept and slope coefficients over time for each level of the quantile.

In this study only involved one covariate. Suggestions for further research are to use various covariates as explained before, so that the modelling will be richer with various information from each covariate involved.

7. Acknowledgement
The authors thank Rector of Universitas Padjadjaran and to the anonymous referee whose valuable checking has improved this paper. This work is funded by Riset Disertasi Doktor Unpad (RDDU) contract number: 5884/UN6.D/LT/2019.

References
[1] Andriyana, Y., Gijbels, I., Verhasselt, A. P-splines quantile regression estimation in varying coefficient models. *Test* 23 153–194 2014.
[2] Andriyana, Y., Gijbels, I., Verhasselt, A. Quantile regression in varying-coefficient models: non-crossing quantile curves and heteroscedasticity. *Statistical Paper* 59 1589-6121 2018.
[3] Andriyana, Y., and Gijbels, I. Quantile regression in heteroscedastic varying coefficient models, *Advances Statistical Analysis* 101, 151–176, 2017.

[4] Fitzmaurice, G. M., Laird N. M., Ware, J. H. *Applied Longitudinal Analysis*. John Wiley & Sons, Inc. 2004.

[5] Goldstein, H. *Multilevel Statistical Models*, 2nd Ed. E-Book of Arnold, London, 1999.

[6] Hastie, T. and Tibshirani, R. Varying-coefficient models. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 55, 4, 757-796. 1993.

[7] Hedeker, D. and Gibbons, R. D. *Longitudinal Data Analysis*. John Wiley & Sons, New Jersey, 2006.

[8] Snijder, T. A. B., Bosker, R. J. *Multilevel Analysis: An introduction to basic and advance multilevel modelling*. 2nd edition. London: SAGE Publications, 2012.

[9] Tantular, B., Jaya, I. G. N. M., Andriyana, Y. Longitudinal Data Exploration of Modelling of Upper Respiratory Tract Infections in Bandung City. *Journal of Physics: Conference Series*. 1265, 2019.

[10] Koenker, R. Quantile regression for longitudinal data. *Journal of Multivariate Analysis* 91, 74–89, 2004

[11] Koenker, R. *Quantile Regression*. Cambridge: Cambridge University Press, 2005.

[12] Diggle P J, Heagerty P, Liang K-Y, Zeger S L *Analysis of Longitudinal Data*. Oxford: University Press, 2002

[13] Andriyana, Y., Ibrahim, M.A., Gijbels, I. and Verhasselt, A. QRegVCM: Quantile Regression in Varying-Coefficient Models: *R Package version* 1.2. 2018