Transnational University Limburg is a unique collaboration of two universities in two countries: the University of Hasselt and Maastricht University.

UHASSELT
KNOWLEDGE IN ACTION

www.uhasselt.be
Universiteit Hasselt
Campus Hasselt:
Martelarenlaan 42 | 3500 Hasselt
Campus Diepenbeek:
Agoralaan Gebouw D | 3590 Diepenbeek

UHASSELT
KNOWLEDGE IN ACTION

Maastricht University

Faculty of Sciences
School for Information Technology

Master of Statistics

Master's thesis

Varying Coefficient Model for Dengue Fevers using P-Splines Quantile Regression

Robyn Irawan
Thesis presented in fulfillment of the requirements for the degree of Master of Statistics, specialization Biostatistics

SUPERVISOR :
Prof. dr. Anneleen VERHASSELT

SUPERVISOR :
Prof.Dr. Yudhie ANDRIYANA

2019
2020

Maastricht University

Faculty of Sciences
School for Information Technology

Master of Statistics

Master's thesis

Varying Coefficient Model for Dengue Fevers using P-Splines Quantile Regression

Robyn Irawan

Thesis presented in fulfillment of the requirements for the degree of Master of Statistics, specialization Biostatistics

SUPERVISOR :

Prof. dr. Anneleen VERHASSELT

SUPERVISOR :

Prof.Dr. Yudhie ANDRIYANA

ACKNOWLEDGEMENT

First and foremost, praises and thanks to Buddha, Guanyin Bodhisattva (观世音菩萨) and devas, for the showers of blessings to complete this master thesis successfully.

I would like to express my deep and sincere gratitude to my internal supervisor, Professor dr. Anneleen Verhasselt of Master of Statistics at Universiteit Hasselt, for her guidance, patience, and time through each stage of the whole working process of this master thesis. Her dynamism, sincerity, and generosity are at their best for letting me to have a lot of valuable meetings and discussions. Her guidance has brought me to be able to finish my thesis on time during my down condition due to the pandemic.

My gratitude also to my external supervisor, Dr. Yudhie Andriyana from Department of Statistics, Universitas Padjadjaran, Bandung, Indonesia, for his constant guidance and support. It has been a great opportunity that has been given by him to me for applying his research and providing the data for my thesis. He has motivated me to face this challenge with confidence, persistence and enthusiasm when I was having a terrible self-doubt.

I am extremely grateful to my beloved mother for her love, prayers, caring and sacrifices for educating and preparing me for my future (我非常感谢亲爱的母亲的爱，祈祷，关心和牺牲，为我的未来教育和准备). Also, I express my thanks to my brothers and the big families for their supports and valuable prayers. It has been a roller coaster and emotional life during my study and to be far from home, but I am grateful that I am still able to support my family and be the pride of my hometown, Tanjung Balai Karimun, Indonesia.

My special thanks to all my friends both in Indonesia and Belgium for the great positivity and support. My gratitude to Wara, Neilshan, MJ, Hanh, Njeri, Mel, Mbak Murih, Osan and Connie for the material and non-material support during this thesis progress. Also to Indonesian Student Association in Belgium (PPI Belgia) community that provides a lot of activities to bring the joy together with the other Indonesians in Belgium.

Last but not least, I would like to express my gratitude to Vlaamse Interuniversitaire Raad - Universitaire Ontwikkelingssamenwerking (VLIR-UOS) for granting me full scholarships to attend Master of Statistics International Course Programme at Hasselt University, Belgium. It has been the greatest life achievement for me to have a chance to study abroad and attain a Master degree, such a dream come true.

May all beings be happy.

Diepenbeek, June 2020
Robyn Irawan

Abstract

Bandung, one of the biggest cities in Indonesia, has a serious problem with dengue fever. Dengue virus is mostly transmitted by Aedes mosquitoes. The distribution of dengue rate varies over times. One may be interested to investigate the dengue rate level every month (e.g., low, medium low, medium high and high). In order to investigate the dengue rate distribution, we propose a quantile regression technique with several quantile levels. We obtain not only some conditional quantile values for dengue rate, but also the information of time. Hence, we need to build a flexible modeling technique involving not only some covariates but also the information of time. Therefore, we propose a (time) varying-coefficient model (VCM) where the coefficients vary over time. In VCM, we consider the coefficients as an unknown function of time variable. Those coefficients can be approximated by a B-splines function. The quantile objective function itself is penalized by a difference operator on the coefficients of the basis B-splines, which we call P-splines quantile objective function. The tuning parameter of the penalty term is chosen in a data driven way, in this case, we propose to use Schwarz Information Criteria. We consider two models for the variability in the VCM: homoscedastic and simple heteroscedastic models. Year 2017 is shown to have a consistent low level of dengue rate over the months.

Key Words: Quantile regression; P-spline; VCM; dengue; heteroscedasticity;

Contents

List of Figures

List of Tables

1 Introduction

In tropical regions, dengue fever is one of the most dangerous disease and it may cause death. This disease is caused by dengue virus, with 4 different stereotypes, which is transmitted by female Aedes mosquitoes (i.e. aegypti and albopticus). Symptomatic dengue virus infections were grouped into three categories: undifferentiated fever, dengue fever (DF) and dengue haemorrhagic fever (DHF). DHF was further classified into four severity grades, with grades III and IV being defined as dengue shock syndrome (DSS) with the highest possibility which can cause death [1]. In this thesis, we will not divide the analysis on each categories but just to be a disease called dengue fever.

The cycle of dengue virus transmission can be seen in Figure 1. Firstly, the female aedes mosquito that bring the dengue virus bites a healthy human and that human becomes infected. This human will catch dengue fever with any of the categories mentioned before. The transmission to other human will be done when another female aedes mosquito, that does not bring the virus, bites the infected human and will bring the virus. By that, the cycle will then be repeated.

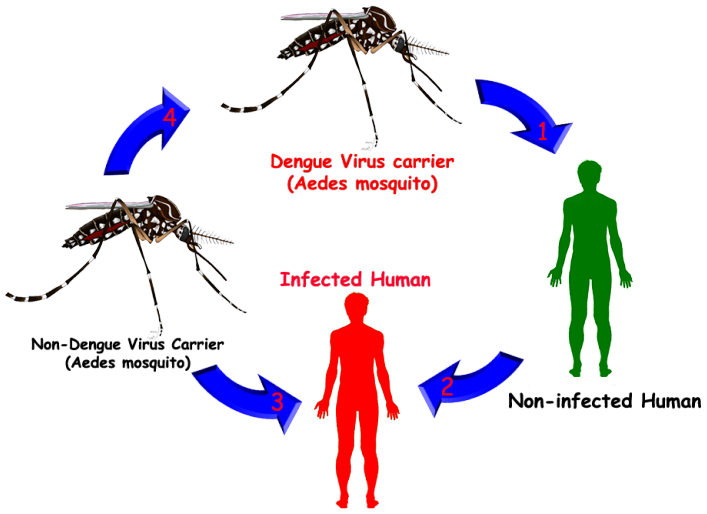


Figure 1: Cycle of dengue virus transmission

The number of dengue cases in the worldwide that are reported annually are increasing every year. Nearly 75% of cases, the disease burden due to dengue were reported from South-East Asia Region and Western Pacific Region. The epidemic of the dengue is a major public health problem in some countries in the southern part of Asia such as Indonesia, Myanmar, Sri Lanka, Thailand, and Timor-Leste. In Indonesia, 20% of the total number of the cases in the whole country reported are from both Jakarta and West Java province [1].

Bandung is one of the most populous cities in Indonesia and geographically located on the highland part of West Java province. This city has a high level of humidity with the temperature of 18.5°C - 30.1°C and is located 791 meters above sea level. Aedes aegypti, the vector of dengue fever transmission in Indonesia, can live in a place with the temperature above 16°C and under 1,000 meters above sea level. Therefore, Bandung is one of the cities that is suitable for the Aedes mosquito to live at and breed [2].

It is our interest to detect the monthly pattern of the dengue incidence rate and to detect the years with certain risk level of dengue fever. One way to do this, is by building a statistical model such as regression model. Concerning dengue fever is a vector-borne disease, normally it has a majority of the low incidence rate. By this situation, the incidence rate will likely to be right-skewed and not bell-shaped [3]. Therefore, the simple linear regression might not be the best option for this case. One way to tackle this condition is by applying quantile regression, applying the regression on each quantile, such as at the tails and the median of the data.

In addition, when the data is recorded as repeated measurements, a model needs to take them into account. In this thesis, we use a varying coefficient model, where each parameter to be varies. The estimation of these parameters will be done by using P-splines which give more accurate curve estimation for nonlinear trend changes [4].

2 Methodology

2.1 Data Description

In this thesis, the data is gathered from 2013-2017 with recordings every month in Bandung city, in total 60 months. The data contains the monthly dengue rate, precipitation and temperature for 5 years. The response variable is the dengue rate: number of recorded cases out of a certain number of people. The estimators that are taken into the study are the monthly precipitation level in mm(s) and temperature in degree Celcius.

2.2 Statistical Tools

The methodology consists of 3 statistical tools: quantile regression, P-splines and varying coefficient model (VCM).

2.2.1 Quantile Regression

Quantile regression is the generalization of median regression into τ-th level quantile regression. This idea is used mainly for the non-Gaussian setting data which has skewed distribution instead of bell-shaped. This setting does not perform good in simple linear regression [5]. Basically, quantile regression is an extension of linear regression that provides greater flexibility than other regression methods to identify differing relationships at different parts of the distribution of the dependent variable [6].

Let Y be the random variable of the interest and μ is the mean of Y which is obtained by minimizing the expected square deviation:

$$\mu = \text{argmin}_c \text{E}(Y - c)^2.$$

The estimator of μ, called $\hat{\mu}$, is obtained from a sample of Y.

Meanwhile, quantiles are particular locations of the distribution. For $\tau \in [0, 1]$ and $F_Y(.)$ is the cumulative distribution function of Y, the τ-th level quantile of Y ($q_Y(\tau)$) is defined as [7]:

$$q_Y(\tau) = F_Y^{-1}(\tau) = \inf\{c : F_Y(c) \geq \tau\}. \tag{1}$$

This term can be re-expressed as a minimization argument of the expected weighted asymmetric deviation,

$$q_Y(\tau) = \underset{c \in \mathbb{R}}{\text{argmin}}\, \text{E}[\rho_\tau(Y - c)], \tag{2}$$

where ρ_τ is called "check-function" that can be expressed as:

$$\rho_\tau(z) = \begin{cases} \tau z & \text{if } z > 0 \\ -(1-\tau)z & \text{otherwise.} \end{cases} \tag{3}$$

Taking the independent and identically distributed (i.i.d) random sample of $Y_1, \ldots, Y_n$ from Y, then the empirical cumulative distribution function is [7]

$$F_n(Y) = \frac{1}{n}\sum_{i=1}^{n} I(Y_i \leq y),$$

where $I(K)$ denote the indicator of a set K, i.e. $I(K) = 1$, if K holds and $I(K) = 0$, when K does not hold. An estimator of $F_Y^{-1}(\tau)$, the objective function by [5] is defined as:

$$\begin{aligned} \hat{q}_Y(\tau) &= \underset{c\in\mathbb{R}}{\operatorname{argmin}} \int_{-\infty}^{\infty} \rho_\tau(Y-c) dF_n(y) = \underset{c\in\mathbb{R}}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n} \rho_\tau(Y_i - c) \\ &= \underset{c\in\mathbb{R}}{\operatorname{argmin}} \frac{1}{n}\left[\sum_{i\in\{i:Y_i\geq c\}} \tau|Y_i - c| + \sum_{i\in\{i:Y_i<c\}} (1-\tau)|Y_i - c| \right]. \end{aligned} \tag{4}$$

Suppose we have $\mathbf{X} = (1, X^{(1)}, \ldots, X^{(p)})^T$, as a vector of the predictor variables for response variable Y. Consider a multiple linear regression for this case as:

$$\begin{aligned} Y &= \beta_0 + \beta_1 X^{(1)} + \ldots + \beta_p X^{(p)} + \varepsilon \\ &= \mathbf{X}^T\boldsymbol{\beta} + \varepsilon, \end{aligned} \tag{5}$$

with $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)^T$ as the vector of regression coefficients and ε as the error term with mean of zero. Taking n i.i.d observations from $\big(\mathbf{X}, Y\big)$, the coefficient vector can be estimated by:

$$\min_{\beta_0,\ldots,\beta_p} \sum_{i=1}^{n} (Y_i - \mathbf{X}_i^T\boldsymbol{\beta})^2. \tag{6}$$

Defined a^τ as the τ-th level quantile of error ε. The conditional quantile curves of response Y given $\mathbf{X}$ can be expressed by:

$$\begin{aligned} q_Y(\tau|\mathbf{X}) &= [\beta_0 + a^\tau] + \beta_1 X^{(1)} + \ldots + \beta_p X^{(p)} \\ &= \beta_0^\tau + \beta_1 X^{(1)} + \ldots + \beta_p X^{(p)} \\ &= \mathbf{X}^T\boldsymbol{\beta} \end{aligned} \tag{7}$$

where $\boldsymbol{\beta} = (\beta_0^\tau, \ldots, \beta_p)^T$ with $\beta_0^\tau = \beta_0 + a^\tau$.

The conditional quantile function is monotonely increasing in the argument $\tau \in [0, 1]$, where for $0 \leq \tau_1 < \tau_2 \leq 1$ will infer to $q_Y(\tau_1|\mathbf{X}) \leq q_Y(\tau_2|\mathbf{X})$, for all $\mathbf{X}$. One condition that should not be violated is that each quantile lines should not cross each other, this condition is called as non-crossingness. Taking n i.i.d observations from $\big(X^{(0)}, X^{(1)}, X^{(2)}, \ldots, X^{(p)}\big)$, with $X^{(0)} \equiv 1$, the non-crossingness condition can only be guaranteed when $\mathbf{X}{=}\bar{\mathbf{X}} = (1, \bar{X}^{(1)}, \bar{X}^{(2)}, \ldots, \bar{X}^{(p)})$ with $\bar{X}^{(j)} = n^{-1}\sum_{i=1}^{n} X_i^{(j)}$ [7].

Assuming $a^\tau = 0$, causing $\beta_0^\tau = \beta_0$, the coefficient $\boldsymbol{\beta}$ is found by solving the minimization problem of

$$\min_{\boldsymbol{\beta}} \mathrm{E}[\rho_\tau(Y - \mathbf{X}^T\boldsymbol{\beta})].$$

Suppose that we have n i.i.d observations from $\left(X^{(1)}, X^{(2)}, \ldots, X^{(p)}, Y\right)$:

$$\left(X_1^{(1)}, X_1^{(2)}, \ldots, X_1^{(p)}, Y_1\right), \ldots, \left(X_n^{(1)}, X_n^{(2)}, \ldots, X_n^{(p)}, Y_n\right).$$

The empirical objective function then is

$$\frac{1}{n}\sum_{i=1}^{n} \rho_\tau(Y_i - \mathbf{X}_i^T\boldsymbol{\beta}), \tag{8}$$

where $\mathbf{X}_i = (1, X_i^{(1)}, \ldots, X_i^{(p)})^T$. Then, by minimizing objective function (8) with respect to $\boldsymbol{\beta}$ we will then have an estimator $\hat{\boldsymbol{\beta}}$. Eventually, the estimator for the τ-th conditional quantile function is

$$\hat{q}_Y(\tau|\mathbf{X}) = \mathbf{X}^T\hat{\boldsymbol{\beta}}. \tag{9}$$

The method to estimate $\boldsymbol{\beta}$ with the objective function (8) is different from multiple linear regression such as solving (6). This is caused by the non-differentiability problem as showed in Figure 2 where the red line ($\rho_{0.5}(z)$) and blue line ($\rho_{0.25}(z)$) are not differentiable in the middle point ($z = 0$). Several methods are proposed to solve this minimization problem and one of the most popular method is linear programming optimization algorithm [Koenker].

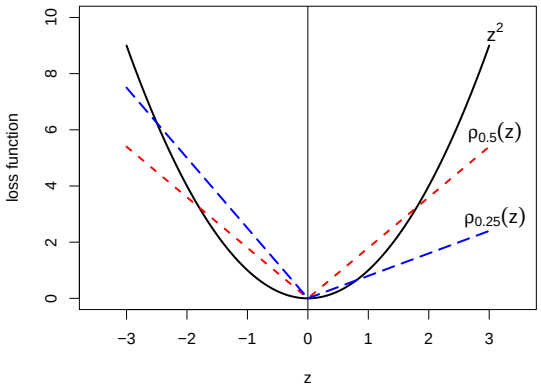


Figure 2: Square and quantiles loss function

2.2.2 P-Splines

P-splines were introduced as the combination of B-splines and the penalties applied on B-splines' adjacent coefficients [8]. These penalties are aimed to impose smoothness to avoid overfitting.

B-splines are the piecewise polynomial functions that have given support with respect to the degree and domain partition [9]. The spline function will be expressed as a linear

combination of B-splines basis function of the given degree. Given the non-decreasing knot sequence of $t_0, \ldots, t_m$, the i-th B-spline of degree ν, denoted by $B_i(x;\nu)$, is define recursively as:

$$\begin{aligned} B_i(x;0) &= \begin{cases} 1 & \text{if } t_i \leq x < t_{i+1} \\ 0 & \text{otherwise,} \end{cases} \\ B_i(x;\nu) &= \frac{x - t_i}{t_{i+\nu} - t_i} B_i(x;\nu-1) + \Big(1 - \frac{x - t_{i+1}}{t_{i+\nu+1} - t_{i+1}}\Big) B_{i+1}(x;\nu-1). \end{aligned}$$

For every x between the first and the last knot, there is a constrain of:

$$\sum_i B_i(x;\nu) = 1.$$

Figure 3 shows an example of B-splines of degree 1 with equidistant knots $t_0, \ldots, t_{10}$. The left B-spline (B_0) of this plot consists of two linear pieces with the knots of t_0, t_1 and t_2; one from t_0 to t_1, the other from t_1 to t_2. To the left of t_0 and right of t_2 this B-spline is zero. Three more B-splines of degree 1 (B_5, B_6 and B_7) are shown with each one based on three knots. A large set of B-splines can be constructed by introducing more knots.

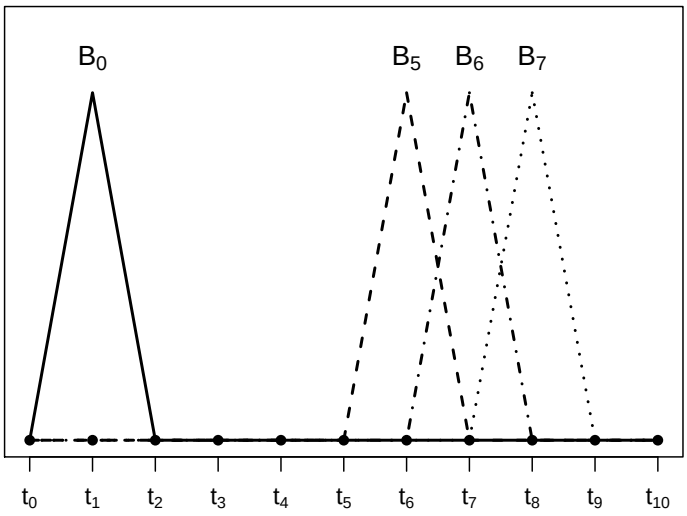


Figure 3: Illustrations of one isolated B-spline of degree 1 and several overlapping ones

Figure 4 shows several B-splines of degree 2, with equidistant knots $t_0, \ldots, t_{10}$, which each consists of three quadratic pieces and joined at two knots (an interval of $[t_i, t_{i+1})$). At the joining points not only the polynomial pieces match, but also their first derivatives are equal. The left B-spline, B_0, is based on 4 adjacent knots: $t_0, \ldots, t_3$. In the right part, three more B-splines of degree 2 are shown.

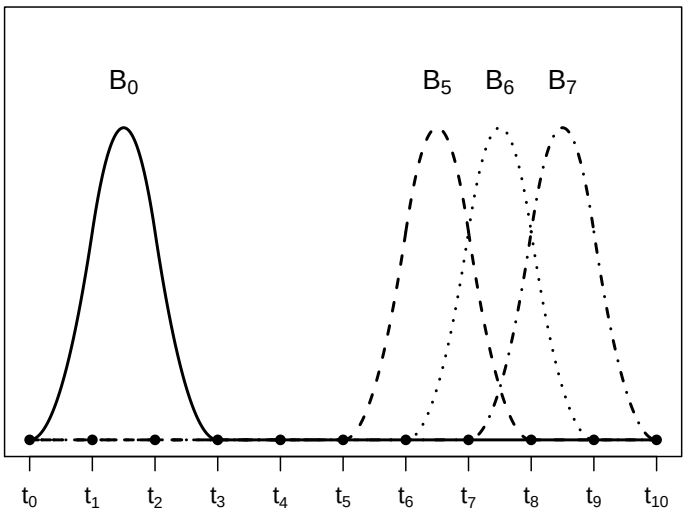


Figure 4: Illustrations of one isolated B-spline of degree 2 and several overlapping ones

Taking the domain from $[x_{min},\ x_{max}]$ and divided into m equal length intervals by $m' = m + 1$ knots. Each interval will be filled by $\nu + 1$ B-splines of degree ν, so that the construction of the B-splines will need $m' + 2\nu$ knots. The number of B-splines in the regression will be $k = m + \nu$.

From these examples, there are several general properties of B-splines of degree ν [8]:

- it consists of $\nu + 1$ polynomial pieces, each of degree ν;
- the polynomial pieces join at ν inner knots;
- at the joining points, derivatives up to order $\nu - 1$ are continuous;
- the B-spline is positive on a domain spanned by $\nu + 2$ knots; everywhere else it is zero;
- except at the boundaries, it overlaps with 2ν polynomial pieces of its neighbors;
- at a given x, $\nu + 1$ B-splines are nonzero.

Consider a regression model of n i.i.d observations:

$$Y_j = f(x_j) + \varepsilon_j,\ j = 1, \ldots, n$$

where Y_j is the response variable of observation j and $f(.)$ is the unknown function given covariate x_j and ε_j is the error term with $\varepsilon_j \sim N(0, \sigma^2)$. This function, $f(.)$, can be

approximated by the linear combination of the B-splines as:

$$f(x) \approx \sum_{i=1}^{k} \alpha_{i-1} B_{i-1}(x;\nu),$$

where $\boldsymbol{\alpha} = (\alpha_0, \ldots, \alpha_{n-1})$ is the associated coefficient vector and $B_i(.;\nu)$ is the basis of B-splines of degree ν with $m+1$ equidistant knots with $i = 1, \ldots, k = m + \nu$.

Now, assuming $q_Y(\tau|x) = f^\tau(x)$ in quantile regression setting, the approximation in the basis of B-splines is applied to $f^\tau(x)$ where $\boldsymbol{\alpha}$ is τ-dependent. Then, by adapting the P-splines least square objective function from [8] to quantile regression setting for minimization term, we have:

$$\min_{\boldsymbol{\alpha}} \left\{ \sum_{j=1}^{n} \rho_\tau \left(Y_j - \sum_{i=1}^{k} \alpha_{i-1} B_{i-1}(x_j;\nu) \right) + \lambda \sum_{i=d+1}^{k} |\Delta^d \alpha_{i-1}|^\gamma \right\}, \tag{10}$$

where $\lambda > 0$ is the smoothing parameter and Δ^d is the d-th order differencing operator where $\Delta^d \alpha_{i-1} = \sum_{t=0}^{d} (-1)^t \binom{d}{t} \alpha_{i-t-1}$ with $d \in \mathbb{N}$. For instance, when $d = 1$ we have $\Delta^1 \alpha_{i-1} = \alpha_{i-1} - \alpha_{i-2}$ and when $d = 2$ we have $\Delta^2 \alpha_{i-1} = \alpha_{i-1} - 2\alpha_{i-2} + \alpha_{i-3}$. The power $\gamma > 0$ is a general penalty term and since this minimization will be translated into linear programming problem, it is restricted into $\gamma = 1$.

2.2.3 Varying Coefficient Model

In several cases, the usage of linear regression models lack flexibility since these models strongly assume their regression coefficient, $\boldsymbol{\beta}$, to be constant. One of the generalization suggested by [10] is by allowing the coefficients to change smoothly with the value of other variable. This suggestion might be very useful to solve some modeling situations, such as longitudinal studies which regression coefficients may change as a function of the variable time of the repeated measurements of each subject. These models are called as varying coefficient models (VCM). These models allow when the effects of one or more covariates to vary smoothly over the values of other variables. One example of special case of VCM is assuming the covariates vary with a variable "time" in term of repeated measurements.

In this study we are focusing on the longitudinal data setting which we consider having repeated observations on $(Y(T), (X^{(1)}(T), ..., X^{(p)}(T)), T)$ with T denotes the variable of time (with domain $\mathbb{T}$), $Y(T)$ is the response variable at time T, and $(X^{(1)}(T), ..., X^{(p)}(T))$ is the vector of covariates at time T. In this term, we assume the measurements to be independent between subject, but can be correlated within the measurement times for each subject.

For $\forall t \in \mathbb{T}$, VCM can be expressed as:

$$\begin{aligned} Y(t) &= \beta_0(t)X^{(0)}(t) + \beta_1(t)X^{(1)}(t) + \ldots + \beta_p(t)X^{(p)}(t) + \tilde{\varepsilon}(t) \\ &= \mathbf{X}^T(t)\boldsymbol{\beta}(t) + \tilde{\varepsilon}(t), \end{aligned} \tag{11}$$

where $\mathbf{X}(t) = (X^{(0)}(t), \ldots, X^{(p)}(t))^T$ with $X^{(0)}(t) \equiv 1$ for all $t \in \mathbb{T}$ as the vector of predictor variables at time t and $\boldsymbol{\beta}(t) = (\beta_0(t), \beta_1(t), \ldots, \beta_p(t))^T$ as the vector of regression coefficient functions at time t.

In longitudinal study, suppose we have observations $(Y_{ij}, \mathbf{X}_{ij}, t_{ij})$ of $(Y(T), \mathbf{X}(T), T)$, for $i = 1, \ldots, n$ and $j = 1, \ldots, N_i$, at j-th time point of measurement t_{ij} for i-th subject and N_i is the number of repeated measurements of subject i. Then, Y_{ij} and $\mathbf{X}_{ij} = (X_{ij}^{(0)}, \ldots, X_{ij}^{(p)})^T$ are the i-th subject's response and covariates at t_{ij}, such that $Y_{ij} = Y(t_{ij})$ and $\mathbf{X}_{ij} = (X^{(0)}(t_{ij}), \ldots, X^{(p)}(t_{ij}))^T$ with $t_{ij} \in \mathbb{T}$. The design when the number of repeated measurements for each subject, N_i, are different among subject, it will be called unbalanced design. Otherwise, if it is restricted to have the same number of repeated measurements among subjects, then it will be called as balanced design.

Another important term in VCM is the error term $\tilde{\varepsilon}(t)$. Assume that the error term can be modeled as [11]:

$$\tilde{\varepsilon}(t) = V(\mathbf{X}(t), t)\varepsilon(t),$$

where $V(\mathbf{X}(t), t)$ is a non-negative function, $E(\varepsilon | X^{(1)}, \ldots, X^{(p)}) = 0$ and $\text{Var}(\varepsilon | X^{(1)}, \ldots, X^{(p)}) = 1$. There are several possible structures of $V(\mathbf{X}(t), t)$, they are:

1. Homoscedastic model, with $V(\mathbf{X}(t), t) = V \in \mathbb{R}^+$, such as:

$$Y(t) = \mathbf{X}^T(t)\boldsymbol{\beta}(t) + V\varepsilon(t); \tag{12}$$

2. Simple heteroscedastic model, with $V(\mathbf{X}(t), t) = V(t)$, such as:

$$Y(t) = \mathbf{X}^T(t)\boldsymbol{\beta}(t) + V(t)\varepsilon(t); \tag{13}$$

2.3 Homoscedastic Varying Coefficient Model

There are several studies about quantile regression in VCMs applied to longitudinal data. In this study, we build the model of quantile regression in VCMs for longitudinal data using the flexible P-splines estimation method. Based on the homoscedastic VCM defined in equation (12) with $V = 1$, it can be rewritten as:

$$Y(t) = \beta_0(t)X^{(0)}(t) + \beta_1(t)X^{(1)}(t) + \ldots + \beta_p(t)X^{(p)}(t) + \varepsilon(t), \ t \in \mathbb{T},$$

where $\varepsilon(t)$ is assumed to have τ-th level quantile value of 0 ($a^\tau(t) = 0$) and independent of $(\mathbf{X}(t), t)$. Suppose we have repeated observations $(Y_{ij}, \mathbf{X}_{ij}, t_{ij})$ of $(Y(T), \mathbf{X}(T), T)$, the τ-th level quantile of $Y(T)$ given $(\mathbf{X}(T), T = t_{ij}) = (\mathbf{X}_{ij}, t_{ij})$ is defined as $q_{Y_{ij}}(\tau | X_{ij}, t_{ij}) = \inf\{y : P\{Y(T) \leq y | (\mathbf{X}(T), T) = (\mathbf{X}_{ij}, t_{ij})\} \geq \tau\}$ for $0 \leq \tau \leq 1$.

The general form of the homoscedastic model for quantile regression in VCM for longitudinal data can be written as [12]:

$$\begin{aligned} q_Y(\tau|\mathbf{X}(t),t) &= \beta_0(t)X^{(0)}(t) + \beta_1(t)X^{(1)}(t) + \ldots + \beta_p(t)X^{(p)}(t) \\ &= \beta_0(t) + \sum_{r=1}^{p} X^{(r)}(t)\beta_r(t) \\ &= \mathbf{X}^T(t)\boldsymbol{\beta}(t). \end{aligned} \tag{14}$$

Each regression coefficient functions can be can be approximated by normalized B-splines with different degree of smoothness. The B-splines of degree ν_r used in coefficient function $\beta_r(t)$ approximation, for $r = 0, \ldots, p$, is:

$$\begin{aligned} \beta_r(t_{ij}) &\approx \alpha_{r0}B_{r0}(t_{ij};\nu_r) + \ldots + \alpha_{r(k_r-1)}B_{r(k_r-1)}(t_{ij};\nu_r) \\ &= \sum_{l=1}^{k_r} \alpha_{r(l-1)}B_{r(l-1)}(t_{ij};\nu_r) \\ &= \boldsymbol{\alpha}_r^T\mathbf{B}_r(t_{ij};\nu_r), \end{aligned} \tag{15}$$

where $\boldsymbol{\alpha}_r = (\alpha_{r0}, \ldots, \alpha_{r(k_r-1)})^T$ as the associated coefficient vector and $\mathbf{B}_r(t_{ij};\nu_r) = (B_{r0}(t_{ij};\nu_r) \ , \ldots, B_{r(k_r-1)}(t_{ij};\nu_r))^T$. The B-spline basis functions of degree ν_r denoted by $B_{r(l-1)}(t_{ij};\nu_r)$, $l = 1, \ldots, m_r + \nu_r$, with $m_r + 1$ equidistant knots for the r-th regression coefficient and $k_r = m_r + \nu_r$.

The estimation of the unknown coefficients $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_0^T, \ldots, \boldsymbol{\alpha}_p^T)$, there might be a case of overfitting when the number of basis functions are large [12]. It is proposed to use the penalty by [8] that penalize for too large differences between coefficients of adjacent B-splines. The objective function in this case, adapting from equation (10), can be written as:

$$S(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \frac{1}{N_i} \sum_{j=1}^{N_i} \rho_\tau \left(Y_{ij} - \sum_{r=0}^{p} X_{ij}^{(r)} \sum_{l=1}^{k_r} \alpha_{r(l-1)}B_{r(l-1)}(t_{ij};\nu_r) \right) + \sum_{r=0}^{p} \sum_{l=d_r+1}^{k_r} \lambda_r |\Delta^{d_r}\alpha_{r(l-1)}|^\gamma \tag{16}$$

with N_i is the number of repeated measurements from observation i, $\gamma > 0$ and $\lambda_r > 0$, $r = 0, \ldots, p$ are the smoothing parameters that control the trade-off between the goodness-of-fit and the penalty term [12]. Δ^{d_r} is the d_r-th order differencing operator of the r-th variable, such as $\Delta^{d_r}\alpha_{r(l-1)} = \sum_{t=0}^{d_r}(-1)^t\binom{d_r}{t}\alpha_{r(l-1-t)}$ with $d_r \in \mathbb{N}$.

This objective function will be translated into linear programming problem in order to find the estimates $\hat{\boldsymbol{\alpha}} = (\hat{\boldsymbol{\alpha}}_0^T, \ldots, \hat{\boldsymbol{\alpha}}_p^T)$ for $\boldsymbol{\alpha}$. This term is referred as P-splines quantile estimator. So, the estimated τ-th level quantile of Y_{ij} given $(\mathbf{X}_{ij}, t_{ij})$ can be written as:

$$\begin{aligned} \hat{q}_{Y_{ij}}(\tau|\mathbf{X}_{ij},t_{ij}) &= \mathbf{X}_{ij}^T(t)\hat{\boldsymbol{\beta}}(t) \\ &= \sum_{r=0}^{p} X_{ij}^{(r)} \sum_{l=1}^{k_r} \hat{\alpha}_{r(l-1)}B_{r(l-1)}(t_{ij};\nu_r). \end{aligned} \tag{17}$$

In addition, minimization of the objective function (16) with respect to $\boldsymbol{\alpha}$ involves the choice of the smoothing parameters, i.e. $\lambda_0, \ldots, \lambda_p$. The selection of the smoothing parameter is important to tackle overfitting [12].

First, consider that all λs are equal, such that $\lambda_0 = \ldots = \lambda_p = \lambda$. In quantile regression context, it is suggested to use Schwarz Information Criterion (SIC) [13], since the associated minimization problem is fairly easy [12]. In our context of multiple quantile regression,the Schwarz Information Criterion can be expressed as:

$$\text{SIC}(\lambda) = \log\left(\frac{1}{n}\sum_{i=1}^{n}\frac{1}{N_i}\sum_{j=1}^{N_i}\rho_\tau(Y_{ij} - \hat{q}_{Y_{ij}}(\tau|\mathbf{X}_{ij}, t_{ij}))\right) + \frac{\log(N)}{2N}p_\lambda, \tag{18}$$

where $N = \sum_{i=1}^{n} N_i$ and the quantity p_λ is the effective degrees of freedom of fitted model. In [14], it is stated that this quantity is similar as computing the number of zero residuals for the fitted model. The quantity p_λ is equal to the size of the elbow set ξ_λ [12], where:

$$\xi_\lambda = \{(i,j) : Y_{ij} - \hat{q}_{Y_{ij}}(\tau|\mathbf{X}_{ij}, t_{ij}) = 0\}.$$

The optimal value of λ can be chosen by minimizing the SIC value from equation (18).

2.4 Simple Heteroscedastic Varying Coefficient Model

In this model, the variability function $V(t)$ is not a constant and varies with variable t. Refer to simple heteroscedastic model (13) for multiple linear regression, the τ-th level conditional quantile of the response variable $Y(T)$ given $(\mathbf{X}(T), T = t)$ is written as:

$$\begin{aligned} q_{Y(t)}(\tau|\mathbf{X}(t), t) &= \beta_0(t)X^{(0)}(t) + \beta_1(t)X^{(1)}(t) + \ldots + \beta_p(t)X^{(p)}(t) + V(t)a^\tau(t) \\ &= [\beta_0(t) + V(t)a^\tau(t)]X^{(0)}(t) + \beta_1(t)X^{(1)}(t) + \ldots + \beta_p(t)X^{(p)}(t) \\ &= \mathbf{X}^T(t)\boldsymbol{\beta}^\tau(t), \end{aligned} \tag{19}$$

where $\boldsymbol{\beta}^\tau(t) = (\beta_0^\tau(t), \beta_1(t), \ldots, \beta_p(t))^T$ with $\beta_0^\tau(t) = \beta_0(t) + V(t)a^\tau(t)$.

Now, it is not only desired to estimate the parameters by P-spline technique, but also dealing with the crossingness issue. There are several methods proposed to accommodate these both, such as stated in [11]. One of the proposed method that will be used in this study is called adaptation of He's approach. This method adapts the approach that is suggested by [15] to overcome crossingness problem in quantile regression that we will apply in varying coefficient model.

There are two crucial assumptions on the error structure stated by [11] by adapting [15]:

1. The (conditional) median quantile of the error term $\varepsilon(t)$ equals zero: $q_{\varepsilon(t)}(0.5) = 0$, such that $a^{0.5}(t) = 0$.

2. The (conditional) median quantile of the absolute value of the error term $\varepsilon(t)$ equals one: $q_{|\varepsilon(t)|}(0.5) = 1$.

The estimation procedure consists of three steps: (i) using the first assumption, estimate the median quantile function; (ii) based on the second assumption and using the median quantile function's estimator from first step will allow us to estimate $V(t)$; (iii) using the estimation values from both previous steps, estimate the τ-th level conditional quantile of the error term $\varepsilon(t)$ and will infer to the quantile regression estimates.

In the first step, under the first assumption, the median quantile function under model (19) can be written as:

$$\begin{aligned} q_{Y(t)}(0.5|\mathbf{X}(t),t) &= \beta_0(t)X^{(0)}(t) + \beta_1(t)X^{(1)}(t) + \ldots + \beta_p(t)X^{(p)}(t) \\ &= \mathbf{X}^T(t)\boldsymbol{\beta}(t), \end{aligned}$$

where $\boldsymbol{\beta}(t) = (\beta_0(t), \ldots, \beta_p(t))^T$. Then, by using P-splines estimation method with $\tau = 0.5$, such by minimizing the objective function (16) of homoscedastic model in section 2.3, we obtain the estimated regression coefficient functions $\hat{\beta}_0(t), \ldots, \hat{\beta}_p(t)$. By this step, we have the estimated median regression value:

$$\begin{aligned} \hat{q}_{Y(t)}(0.5|\mathbf{X}(t),t) &= \hat{\beta}_0(t)X^{(0)}(t) + \hat{\beta}_1(t)X^{(1)}(t) + \ldots + \hat{\beta}_p(t)X^{(p)}(t) \\ &= \mathbf{X}^T(t)\hat{\boldsymbol{\beta}}(t). \end{aligned}$$

The next step is estimating the variability function $V(t)$ by using model (13). Leaving only the variability function and error term on the right hand side of the equation and applying absolute value for both side given $V(t) \geq 0$, we have:

$$|Y(t) - \mathbf{X}^T(t)\boldsymbol{\beta}(t)| = V(t)|\varepsilon(t)|.$$

Then, in term of median quantile regression and by the second assumption, we have:

$$q_{|Y(t)-\mathbf{X}^T(t)\boldsymbol{\beta}(t)|}(0.5|t) = V(t).$$

Based on the estimated median regression coefficient function $\hat{\boldsymbol{\beta}}(t)$ from the first step and using P-splines estimation technique, variability function $V(\cdot)$ can be estimated. Taking repeated observations $(Y_{ij}, \mathbf{X}_{ij}, t_{ij})$ of $(Y(T), \mathbf{X}(T), T)$, we have $|Y(t_{ij}) - \mathbf{X}^T(t_{ij})\hat{\boldsymbol{\beta}}(t_{ij})|$, for $i = 1, \ldots, n$ and $j = 1, \ldots, N$. The approximation of the variability function $V(\cdot)$ will be done by P-splines, we denote B-splines basis of degree ν^v with m^v equal length intervals of $m^v + 1$ equidistant knots, by $B^v_{l-1}(\cdot; \nu^v)$, $l = 1, \ldots, k^v = \nu^v + m^v$. The superscript v is used to mark that this is the B-splines basis to estimate $V(\cdot)$. With $(\alpha^v_0, \ldots, \alpha^v_{k^v-1})$ as the associated coefficient vector, the variability function $V(t_{ij})$ can be estimated as

$$V(t_{ij}) \approx \sum_{l=1}^{k^v} \alpha^v_{l-1} B^v_{l-1}(t_{ij}; \nu^v).$$

The estimators for the coefficient $(\hat{\alpha}^v_0, \ldots, \hat{\alpha}^v_{k^v-1})$ can be obtained by minimizing [11]

$$\sum_{i=1}^{n} \frac{1}{N_i} \sum_{j=1}^{N_i} \rho_{0.5}\left(\left| Y(t_{ij}) - \mathbf{X}^T(t_{ij})\hat{\boldsymbol{\beta}}(t_{ij}) \right| - \sum_{l=1}^{k^v} \alpha^v_{l-1} B^v_{l-1}(t_{ij}; \nu^v) \right) + \sum_{l=d^v+1}^{k^v} \lambda^v |\Delta^{d^v} \alpha^v_{l-1}|, \quad (20)$$

with respect to $\boldsymbol{\alpha}^v = (\alpha_0^v, \ldots, \alpha_{k^v-1}^v)$. $\lambda^v > 0$ is the penalization parameter and d^v is the order of the differencing operator. Then, the estimator $\hat{V}(t)$ is given by

$$\hat{V}(t) = \sum_{l=1}^{k^v} \hat{\alpha}_{l-1}^v B_{l-1}^v(t; \nu^v). \tag{21}$$

Remark that the estimator $\hat{V}(t)$ is not necessarily positive. One possibility, so that $\hat{V}(t) \in \mathbb{R}^+$, is by approximating $\log(V(t))$ by B-splines instead of $V(t)$.

In the final step, recall from model (13), by leaving only the variability function and error term on the right hand side

$$Y(t) - \mathbf{X}^T(t)\boldsymbol{\beta}(t) = V(t)\varepsilon(t),$$

then substitute $\boldsymbol{\beta}(t)$ with $\hat{\boldsymbol{\beta}}(t)$ from the first step and $V(t)$ with $\hat{V}(t)$ from the second step. The τ-th level quantile of $Y(t) - \mathbf{X}^T(t)\hat{\boldsymbol{\beta}}(t)$ can be expressed as

$$q_{Y(t)-\mathbf{X}^T(t)\hat{\boldsymbol{\beta}}(t)}(\tau|t) = \hat{V}(t)a^\tau(t).$$

The unknown (conditional) quantile $a^\tau(t)$ of the error term $\varepsilon(t)$ is estimated by k^h B-spline basis functions of degree ν^h. In other words,

$$a^\tau(t_{ij}) \approx \sum_{l=1}^{k^h} \alpha_{l-1}^h B_{l-1}^h(t_{ij}; \nu^h).$$

The superscript h is used to draw attention that this is the B-splines basis for estimation of $a^\tau(\cdot)$.

The coefficients $\boldsymbol{\alpha}^h = (\alpha_0^h, \ldots, \alpha_{k^h-1}^h)$ are obtained by P-splines approximation. The minimization of the following objective function

$$\sum_{i=1}^{n} \frac{1}{N_i} \sum_{j=1}^{N_i} \rho_\tau \left(Y(t_{ij}) - \mathbf{X}^T(t_{ij})\hat{\boldsymbol{\beta}}(t_{ij}) - \hat{V}(t_{ij}) \sum_{l=1}^{k^h} \alpha_{l-1}^h B_{l-1}^h(t_{ij}; \nu^h) \right) + \sum_{l=d^h+1}^{k^h} \lambda^h |\Delta^{d^h} \alpha_{l-1}^h|, \tag{22}$$

with respect to $\boldsymbol{\alpha}^h$, with $\lambda^h > 0$ is the penalization parameter and d^h is the order of the differencing operator in the penalty term for this quantile regression. Resulting with the estimator $\hat{\boldsymbol{\alpha}}^h$ for $\boldsymbol{\alpha}^h$, the estimator of $a^\tau(t)$ is

$$\hat{a}^\tau(t) = \sum_{l=1}^{k^h} \hat{\alpha}_{l-1}^h B_{l-1}^h(t; \nu^h).$$

Eventually, the estimated τ-th level conditional quantile is given by:

$$\hat{q}_{Y(t)}(\tau|\mathbf{X}(t), t) = \mathbf{X}^T(t)\hat{\boldsymbol{\beta}}(t) + \hat{V}(t)\hat{a}^\tau(t). \tag{23}$$

2.5 Statistical Software

The whole programming needs in this study are done by using R version 3.6.1 and viewed in RStudio version 1.2.5019. Library `QRegVCM` that provided the codes for both homoscedastic and heteroscedastic VCMs is used in this study. The outputs from this software include the graphs, parameter estimates and desired certain values.

3 Results and Analysis

3.1 Exploratory Data Analysis

Before going to the models, the behavior of the variables will first be checked. Figure 5 shows the density of the response variable and showing that the response is clearly right-skewed. In this case, it is better to model using quantile regression than Gaussian regression model.

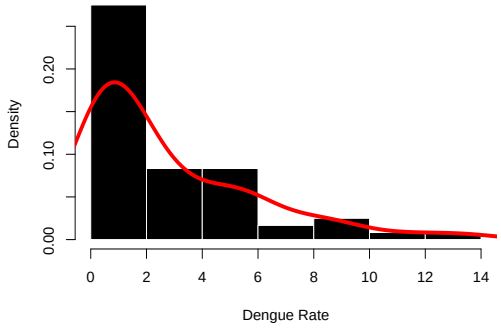
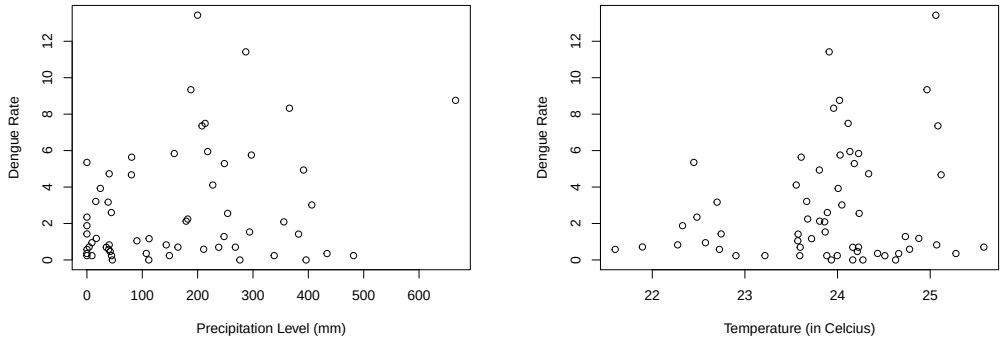


Figure 5: Histogram of dengue rate density

Relationship between dengue rate and the independent variables are shown in Figure 6. Figure 6a shows the higher the rainfall, the dengue rate becomes more variable. Even though the dengue rate shows a positive linear relationship with the rainfall, but there are still quite lot number of data with low dengue rate when they have high rainfall. The same goes for temperature as it has positive linear relationship with dengue rate, but quite lot of data are still have low dengue rate with higher temperature.

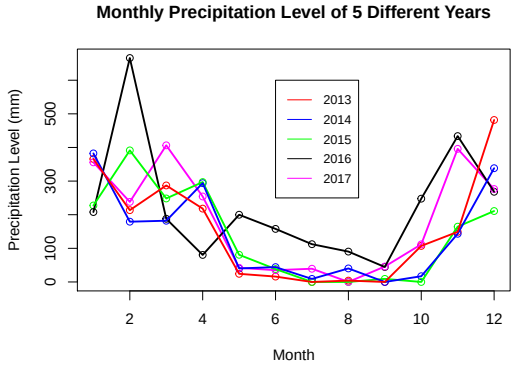


(a) Plot of dengue rate vs. precipitation rate (b) Plot of dengue rate vs. temperature

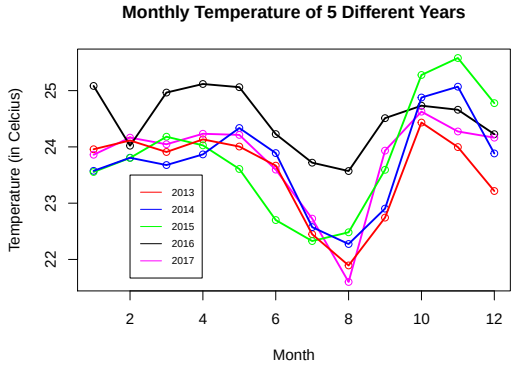
Figure 6: Scatter plots between dependent vs independent variables

The monthly characteristics for both dependent and independent variables are checked. The plots are shown in Figure 7, with each years line and different color. There seems such a high rainfall in the beginning and the end of year (See Figure 7a). This is due to the fact that the climate in Bandung which has dry and wet seasons. Rain season is normally in October to March, while dry season is happening in the other months.

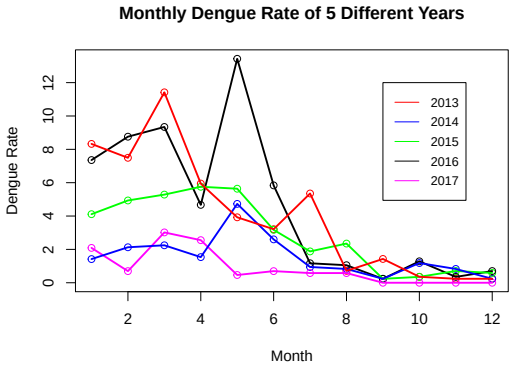
The temperature seems to have the same pattern as the precipitation level. In Figure 7b, the lowest temperature in most of the years fall in august during dry season and tend to have the hotter temperature during the beginning and end of the year. The relationship is shown in Figure 8. This happens due to water evaporation during wet season. The heat of the sunlight causes high humidity and yielding hotter climate in the city.



(a) Monthly plot of precipitation level



(b) Monthly plot of temperature



(c) Monthly plot of dengue rate

Figure 7: Monthly plots of independent and dependent variables for every year

The monthly dengue rate is shown in Figure 7c. In all years, it is clear that the dengue rate's trend is high in the beginning to the middle of the year and very low during the end of the year. During the high rainfall in the beginning of the year, the mosquitoes tend to breed faster and spread in high numbers. Meanwhile, in the end of the year, even though

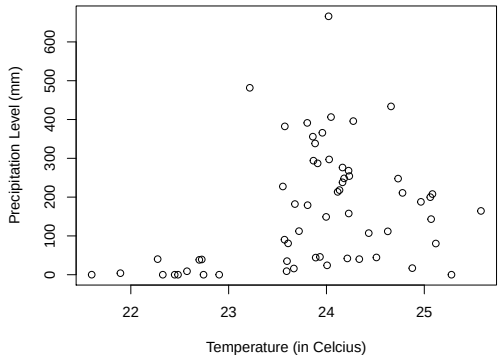


Figure 8: Scatter plot of rainfall vs temperature

the precipitation level is high, the dengue rate is low. This might be due to another factor that is not measured in this study.

3.2 Homoscedastic Varying Coefficient Model

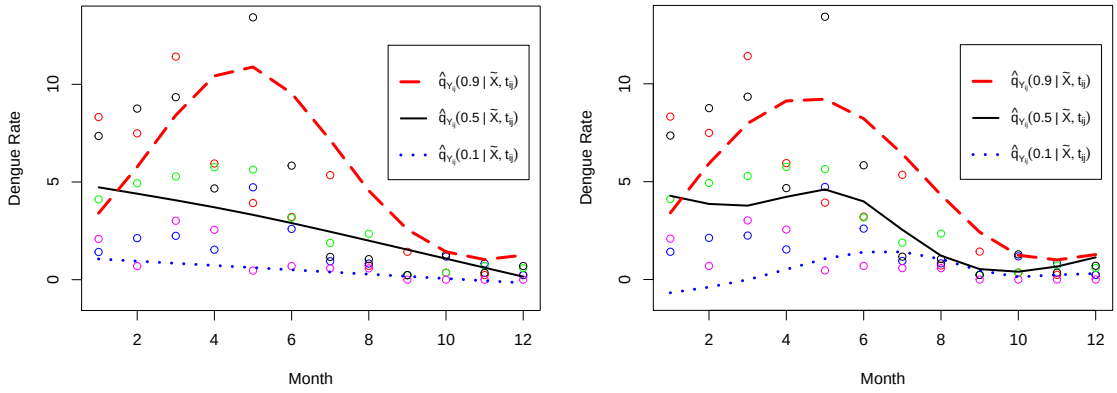
In our case, we take the year as the subject i and month as the time of measurement from that subject (t_{ij}). Denote Y_{ij} as the dengue rate of year i in j-th month (t_{ij}), the τ-th level quantile of dengue rate given both precipitation level and the temperature can be expressed as:

$$q_{Y_{ij}}(\tau|\text{rain}_{ij}, \text{temp}_{ij}, t_{ij}) = \beta_0(t_{ij}) + \beta_1(t_{ij})\text{rain}_{ij} + \beta_2(t_{ij})\text{temp}_{ij}.$$

The estimation of $\beta_0(t_{ij})$, $\beta_1(t_{ij})$ and $\beta_2(t_{ij})$ will be done by using P-splines. The parameters used in this model are as follows:

- $\tau = \{0.1, 0.5, 0.9\}$;
- Number of subjects $n = 5$ and equal number of repeated measurements $N_i = 12$;
- B-spline of degree 3 (cubic spline), $\nu = 3$;
- Using 3 and 5 knots, such that having $m = 2$ and $m = 4$ equal length intervals, respectively;
- Number of B-splines is then $k = 5$ and $k = 7$;
- $d = 2$ and $\gamma = 1$;
- Choosing an optimal smoothing parameter from $\lambda = \{10^{-3}, 10^{-2.75}, 10^{-2.5}, \ldots, 10^{0.75}, 10\}$.

The monthly estimates of the τ-th level quantiles of the dengue rate given the medians of the precipitation level and temperature, denote $\tilde{\mathbf{X}} = (1, \text{Med(rain)}, \text{Med(temp)})$, for both 3 knots and 5 knots are presented in Figure 9. The colored dots in red, blue, green, black and magenta represent the observed dengue rate in year 2013, 2014, 2015, 2016 and 2017, respectively. The optimal λs for 3 knots case with $\tau = \{0.1, 0.5, 0.9\}$ are $\{10, 10, 10^{-1}\}$ respectively, while for 5 knots case are $\{10^{-0.25}, 10^{0.5}, 1\}$ respectively. From both graphs in Figure 9a and 9b, there is a violation of quantile regression condition: non-crossingness problem. The median quantile and 0.9 quantile are both crossing each other around the the beginning of the year. This concludes that homoscedastic model is not suitable for our data.



(a) $\hat{q}_{Y_{ij}}(0.1|\tilde{\mathbf{X}}, t_{ij})$, $\hat{q}_{Y_{ij}}(0.5|\tilde{\mathbf{X}}, t_{ij})$, and $\hat{q}_{Y_{ij}}(0.9|\tilde{\mathbf{X}}, t_{ij})$ with 3 knots

(b) $\hat{q}_{Y_{ij}}(0.1|\tilde{\mathbf{X}}, t_{ij})$, $\hat{q}_{Y_{ij}}(0.5|\tilde{\mathbf{X}}, t_{ij})$, and $\hat{q}_{Y_{ij}}(0.9|\tilde{\mathbf{X}}, t_{ij})$ with 5 knots

Figure 9: Plots of monthly dengue rate and τ-th level quantile estimates for homoscedastic VCM

3.3 Simple Heteroscedastic Varying Coefficient Model

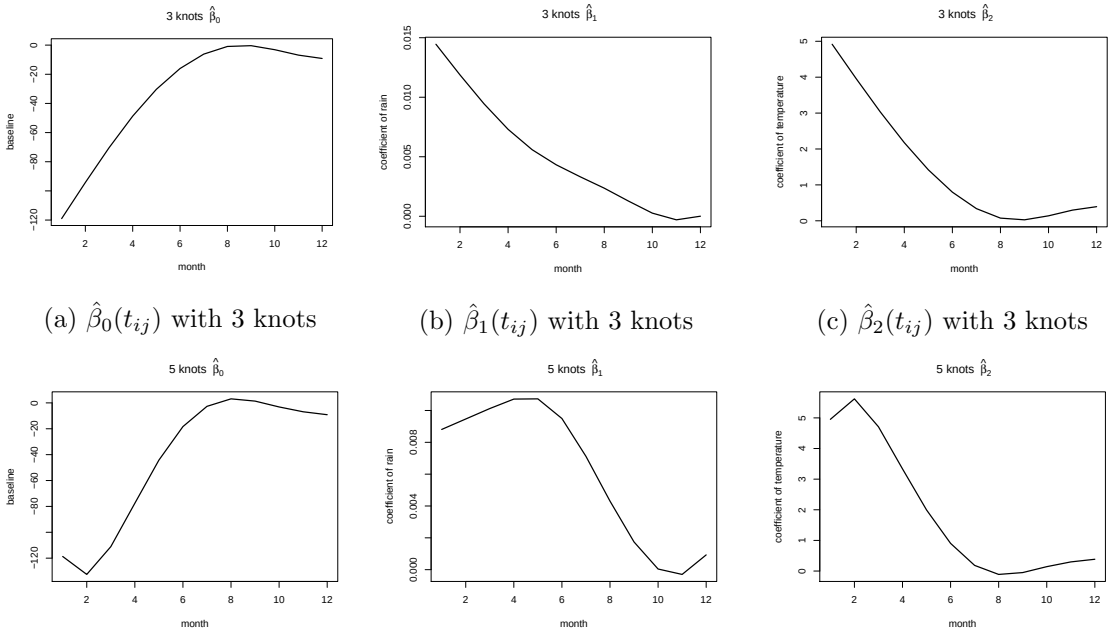
In this model, the τ-th level quantile of the dengue rate given rainfall and temperature is written as:

$$q_{Y_{ij}}(\tau|\text{rain}_{ij}, \text{temp}_{ij}, t_{ij}) = \beta_0(t_{ij}) + \beta_1(t_{ij})\text{rain}_{ij} + \beta_2(t_{ij})\text{temp}_{ij} + V(t_{ij})a^{\tau}(t_{ij}).$$

We use the same settings as homoscedastic model and adaptation of He's approach.

In adapted He's approach, the estimates of parameter functions are done with optimal smoothing parameter λ, for both 3 knots and 5 knots settings, only for median quantile are the same as homoscedastic model, they are 10 and $10^{0.5}$ respectively. Figure 10 shows the estimates of the coefficient functions $\hat{\beta}_0(t)$, $\hat{\beta}_1(t)$, and $\hat{\beta}_2(t)$ for both 3 and 5 knots. For

$\hat{\beta}_0$ on Figure 10a and Figure 10d, the baseline of the median quantile of the dengue rate are rapidly increasing from month to month towards zero. Figure 10b and 10e show that the rainfall effect on the median quantile of the dengue rate is decreasing over the months, although the effect is increasing by a little bit until month 5 when using 5 knots. Lastly, the temperature shows an exponentially decreasing effect on the median quantile of dengue rate over the months based on Figure 10c and 10f.



(a) $\hat{\beta}_0(t_{ij})$ with 3 knots (b) $\hat{\beta}_1(t_{ij})$ with 3 knots (c) $\hat{\beta}_2(t_{ij})$ with 3 knots

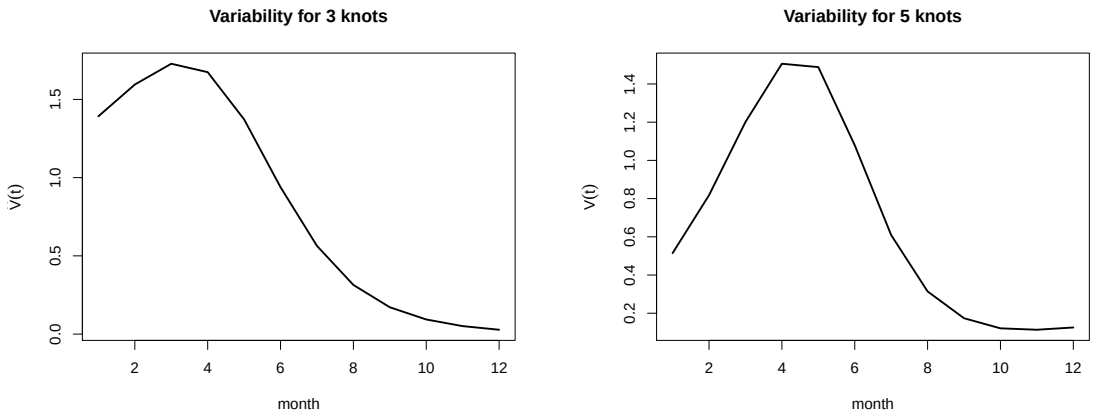
(d) $\hat{\beta}_0(t_{ij})$ with 5 knots (e) $\hat{\beta}_1(t_{ij})$ with 5 knots (f) $\hat{\beta}_2(t_{ij})$ with 5 knots

Figure 10: Regression coefficient functions estimates for 3 and 5 knots

Another term in this model is the variability function, which the estimates are shown in Figure 11. Both Figure 11a and 11b show the variability is increasing at the beginning of the year and exponentially decreasing until the end of the year. This due to the fact that the variability in the dengue rate is higher around the beginning until the middle of the year and less at the end of the year as shown in Figure 7c.

The τ-th level quantile of the dengue rate estimates, given the median of precipitation level and temperature, are shown in Figure 12. The estimation is done simply by inputting the median of the covariates value to equation (23). The estimates using 3 knots seem to be smoother than 5 knots, which is more wiggly. The SIC of 3 knots model is -0.0773, while 5 knots model is -0.0752. Based on the SIC value and smoothness, also since there is no violoation of non-crossingness, heteroscedatic VCM with 3 knots is chosen as the best model in this study.

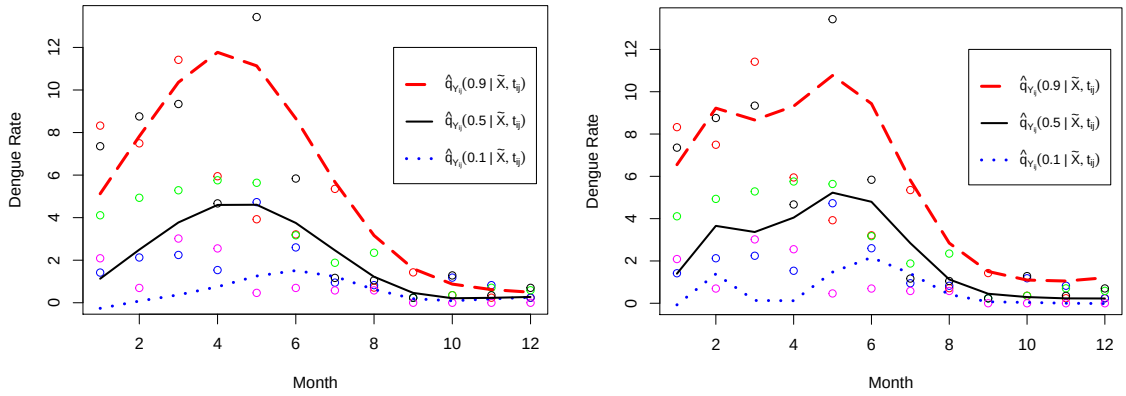
Now, based on the best model, the pattern of the dengue rate will be discussed. For the



(a) $\hat{V}(t_{ij})$ with 3 Knots

(b) $\hat{V}(t_{ij})$ with 5 Knots

Figure 11: Variability function estimates



(a) $\hat{q}_{Y_{ij}}(0.1|\tilde{\mathbf{X}}, t_{ij})$, $\hat{q}_{Y_{ij}}(0.5|\tilde{\mathbf{X}}, t_{ij})$, and $\hat{q}_{Y_{ij}}(0.9|\tilde{\mathbf{X}}, t_{ij})$ with 3 knots

(b) $\hat{q}_{Y_{ij}}(0.1|\tilde{\mathbf{X}}, t_{ij})$, $\hat{q}_{Y_{ij}}(0.5|\tilde{\mathbf{X}}, t_{ij})$, and $\hat{q}_{Y_{ij}}(0.9|\tilde{\mathbf{X}}, t_{ij})$ with 5 knots

Figure 12: Plots of monthly dengue rate and τ-th level quantile estimates for homoscedastic VCM

0.1 quantile, the estimated dengue rates has an almost flat pattern over the months and a slight increase in the middle of the year. The estimated dengue rate lies around 2 to 4 at the beginning to middle of the year and decrease towards around 0 in median quantile. Lastly, the 0.9 quantile of the estimated dengue rates lies around 5 to 12 from January to July and rapidly decreasing towards 1. Based on these lines, it is to confirmed that the dengue rate in Bandung reaches the peak around April and May since they are the transition season months.

The estimated τ-th level quantile of the dengue rate in a certain month can be predicted when we have the specific rainfall and temperature of Bandung. Note that the graph in Figure 12a is representing the median of each covariates. We can say that when the rainfall and the temperature values are the median of the data, we can predict that at the end of the year, it has a 0.9 quantile of the dengue rate around 1 and median together with 0.1 quantile will have around 0 of dengue rate. Another example is when the rainfall and the temperature values are the maximum values from our data, the pattern of 0.1, median and 0.9 quantiles will be formed as shown in Figure 13 below. We can see that at the beginning of the year, the dengue rate are very high, such that the 0.9 quantile is falling around 20 and at the end of the year, the 0.1, median and 0.9 quantiles of dengue rate are falling around 2.

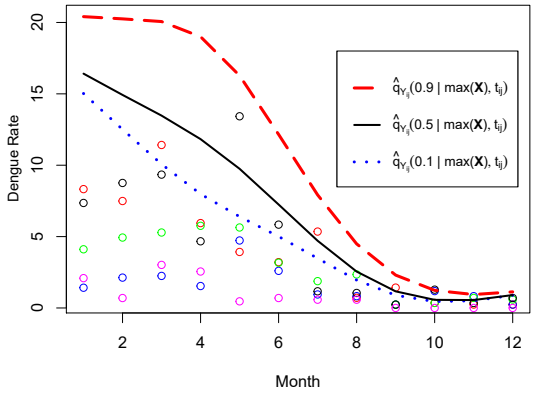


Figure 13: $\hat{q}_{Y_{ij}}(0.1|\max(\mathbf{X}), t_{ij})$, $\hat{q}_{Y_{ij}}(0.5|\max(\mathbf{X}), t_{ij})$, and $\hat{q}_{Y_{ij}}(0.9|\max(\mathbf{X}), t_{ij})$ for 3 knots simple heteroscedastic VCM with max($\mathbf{X}$)= (1,max(rain),max(temp))

It is also interesting to define the monthly levels of the dengue rate for every years. We will use the condition when the rainfall and temperature are both at the median value of the data. There are 4 levels that will be used such that:

1. if $Y_{ij} < \hat{q}_{Y_{ij}}(0.1|\tilde{\mathbf{X}}, t_{ij})$, low level (L) of dengue rate;
2. if $\hat{q}_{Y_{ij}}(0.1|\tilde{\mathbf{X}}, t_{ij}) \leq Y_{ij} < \hat{q}_{Y_{ij}}(0.5|\tilde{\mathbf{X}}, t_{ij})$, medium low level (ML) of dengue rate;
3. if $\hat{q}_{Y_{ij}}(0.5|\tilde{\mathbf{X}}, t_{ij}) \leq Y_{ij} < \hat{q}_{Y_{ij}}(0.9|\tilde{\mathbf{X}}, t_{ij})$, medium high level (MH) of dengue rate;
4. if $Y_{ij} \geq \hat{q}_{Y_{ij}}(0.9|\tilde{\mathbf{X}}, t_{ij})$, high level (H) of dengue rate.

Based on Figure 12a, the summary of the monthly dengue rate levels for 5 years can be seen in Table 1.

Based on the summary from the Table 1 above, as the latest year, 2017 seems to have dengue rate relatively lower than the others. Figure 14 shows the observations for dengue

Table 1: Summary of dengue rate levels from 5 different years over 12 months given the median of precipitation level and temperature of the data

Month \ Year	2013	2014	2015	2016	2017
January	H	MH	MH	H	MH
February	MH	ML	MH	H	ML
March	H	ML	MH	MH	ML
April	MH	ML	MH	MH	ML
May	ML	MH	MH	H	L
June	ML	ML	ML	H	L
July	MH	L	ML	L	L
August	ML	ML	MH	ML	L
September	MH	ML	ML	ML	L
October	MH	H	MH	H	L
November	MH	H	H	MH	L
December	L	L	H	H	L

rate only in 2017 as well as the three estimated conditional quantiles. As seen, the dengue rate is below the estimated conditional median line in most of the months. This confirms across the months of 2017, the worst dengue rate level is medium high in January, even though the highest dengue rate is in March. Since the data also shows the dengue rates in September to December are zero, they will surely have low level of dengue rate.

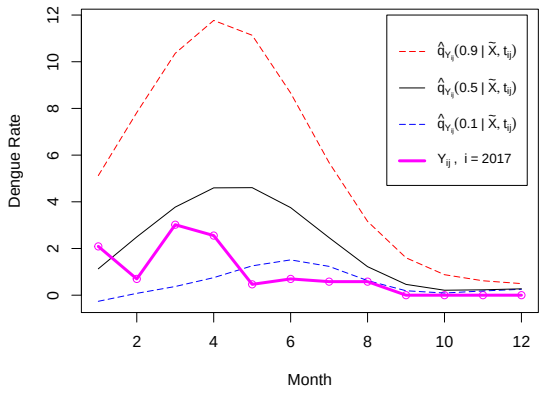


Figure 14: Plot of monthly dengue rate in 2017 and estimated conditional quantile curves

4 Discussion

In this study, the goal is to investigate the pattern of the dengue rate over the months in Bandung. To answer this research question, the analysis was done by considering the dengue rate as the outcomes, taking into account the precipitation level and temperature. There are five different years as the subject for this study with monthly dengue rate for each years. Overall, the dengue rate has right-skewed distributed by the fact that dengue fever is a vector-borne disease.

To build a robust model for non-Gaussian data setting, one of the statistical tool that can be used is quantile regression. Given the dengue rate is measured monthly for five years, this is a longitudinal data with year as the subject. One way to take longitudinal setting into account in regression is by allowing the covariates to vary with a time variable, this tool is called varying coefficient model (VCM). Lastly, it is desired to have an approach that will give a good estimate to the regression coefficients functions and preferably more flexible than polynomial regression. P-spline is chosen as the method to overcome the flexibility and stability issues in polynomial regression.

These three statistical tools (quantile regression, VCM, and P-splines) are used to model our dengue fever data with two different structures of the variability: constant and varying with a time variable. The constant variability structure is then called as homoscedastic VCM, while the other one is known as simple heteroscedastic VCM.

Homoscedastic VCM

Using the assumption of $a^{\tau}(t) = 0$, the τ-th level quantile of the response fully depends on the regression coefficient functions. The estimated regression coefficient functions for each τ-th level quantile are done with the objective function. The optimal smoothing parameter λ, as the penalty to impose smoothness, is chosen by minimizing Schwarz Information Criterion (SIC) value.

In the modeling of our data, B-splines of degree 3 with three and five knots are set in order to estimate 0.1, 0.5 and 0.9 quantiles of the dengue rate. The results show there is a violation in non-crossingness condition. Thus, this model is not suitable for our data.

Simple Heteroscedastic VCM

This model assumes the variability function to depend on time variable ($V(\mathbf{X}(t), t) = V(t)$). Thus, the τ-th level quantile of the response depends on the regression coefficients, variability, and τ-th level quantile of the error term. Using the adapted He's approach, the estimation of the τ-the order conditional quantile ($\hat{q}_{Y(t)}(\tau|\mathbf{X}(t), t)$) is yielded by three steps with two assumptions. These steps include estimations of median regression coefficient function ($\hat{\boldsymbol{\beta}}(t)$), variability function ($\hat{V}(t)$), followed by τ-th level quantile of the error term ($\hat{a}^{\tau}(t)$), respectively.

The result by applying this model to the dengue rate data shows both precipitation

level and temperature effects on the median quantile of dengue rate will decrease over the months. The estimated variability functions illustrate that the dengue rate tend to widely vary around the beginning to the middle of the year, but less vary at the end of the year. The best model acquired by this model is when using three knots on P-splines approach with $\lambda = 10$. The latest year from the data, 2017, consistently has low level of dengue rate around the middle to the end of the year based on the best model.

5 Conclusion

Overall, the estimation technique gives the flexible estimations to the conditional τ-th level quantile of the non-Gaussian setting response, even though there are some restrictions on the parameters in this study. Based on the results from homoscedastic and simple heteroscedastic models, only simple heteroscedastic model is suitable for our dengue rate data. This is caused by homoscedastic model estimations are only based on the individual objective functions that are not tackling the crossingness problem. The results of estimated τ-th level quantile showed in this study might be different with different settings, such as the choice of number of knots and degree of the splines. The choice of τ will affect the levels of the dengue rate due to the change of the quantile curves' shape and position.

6 Future Research Possibility

Some suggestions for future research are using better smoothing parameter selection method, different values of γ, degree of B-splines and number of knots. Another model that can be explored is a model that assumes the variability function to depend on the covariates together with the time variable, this model is called general heteroscedastic VCM. In addition, since dengue fever is a vector-borne disease, it will be better if the monthly mosquitoes or larva characteristic is taken into account

References

[1] World Health Organization, Special Programme for Research, Training in Tropical Diseases, World Health Organization. Department of Control of Neglected Tropical Diseases, World Health Organization. Epidemic, and Pandemic Alert. *Dengue: guidelines for diagnosis, treatment, prevention and control*. World Health Organization, 2009.

[2] R Irawan, B Yong, and F Kristiani. Non-spatial analysis of relative risk of dengue disease in bandung using poisson-gamma and log-normal models: A case study of dengue data from santo borromeus hospital in 2013. In *Journal of Physics: Conference Series*, volume 812, page 012034. IOP Publishing, 2017.

[3] Guilhem Rascalou, Dominique Pontier, Frédéric Menu, and Sébastien Gourbière. Emergence and prevalence of human vector-borne diseases in sink vector populations. *PloS one*, 7(5), 2012.

[4] Carsten Heuer. Modeling of time trends and interactions in vital rates using restricted regression splines. *Biometrics*, pages 161–177, 1997.

[5] Roger Koenker and Gilbert Bassett Jr. Regression quantiles. *Econometrica: Journal of the Econometric Society*, pages 33–50, 1978.

[6] Benjamin Lê Cook and Willard G Manning. Thinking beyond the mean: a practical guide for using quantile regression methods for health services research. *Shanghai Archives of Psychiatry*, 25(1):55, 2013.

[7] Koenker. *Quantile Regression (Econometric Society Monographs; No. 38)*. Cambridge university press, 2005.

[8] Paul HC Eilers and Brian D Marx. Flexible smoothing with b-splines and penalties. *Statistical science*, pages 89–102, 1996.

[9] Carl De Boor. B (asic)-spline basics. Technical report, Wisconsin Univ-Madison Mathematics Research Center, 1986.

[10] Trevor Hastie and Robert Tibshirani. Varying-coefficient models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 55(4):757–779, 1993.

[11] Yudhie Andriyana, Irène Gijbels, and Anneleen Verhasselt. Quantile regression in varying-coefficient models: non-crossing quantile curves and heteroscedasticity. *Statistical Papers*, 59(4):1589–1621, 2018.

[12] Yudhie Andriyana, Irène Gijbels, and Anneleen Verhasselt. P-splines quantile regression estimation in varying coefficient models. *Test*, 23(1):153–194, 2014.

[13] Roger Koenker, Pin Ng, and Stephen Portnoy. Quantile smoothing splines. *Biometrika*, 81(4):673–680, 1994.

[14] Roger Koenker et al. Additive models for quantile regression: Model selection and confidence bandaids. *Brazilian Journal of Probability and Statistics*, 25(3):239–262, 2011.

[15] Xuming He. Quantile curves without crossing. *The American Statistician*, 51(2):186–192, 1997.