Quantile regression in varying coefficient model of upper respiratory tract infections in Bandung City

B Tantular¹,²*, Y Andriyana¹ and B N Ruchjana²

¹Department of Statistics, Universitas Padjadjaran, Jl. Raya Bandung Sumedang km 21 Jatinangor, Sumedang 45363, Indonesia
²Department of Mathematics, Universitas Padjadjaran, Jl. Raya Bandung Sumedang km 21 Jatinangor, Sumedang 45363, Indonesia

*bertho@unpad.ac.id

Abstract. Varying coefficient models are commonly used to obtain effects of covariates that vary over other variables. A special case of varying coefficient model is applied to longitudinal data where the covariates may vary over time. When the function is not easy to specify parametrically, we then need to work on a non-parametric regression technique. In this case, we approximate the function by B-splines. B-splines smoothing tends to overfit with increasing knots, then a penalty is added to the quantile objective function. This estimation procedure is called P-splines. As the objective function, we propose to use quantile loss function. The technique will be implemented to the upper respiratory tract infection data in Bandung City which was measured repeatedly from 30 sub district in Bandung City and hence we have a longitudinal data structure.

1. Introduction

Varying coefficient models are linear regression models that allow the regression coefficient vary with other variables. Varying coefficient models, introduced by [1], can be applied to regression coefficients that allowed vary over time. This method can determine the effect of different covariates between the values of the time variable \((t)\). Several researchers have applied time varying coefficient model including [2] who proposed the use of the two-step estimation method, [3] which uses base function expansion including variable selection and [4] which combines the P-splines method with non-negative garrote variable selection. Therefore varying coefficient models can be used to analyze longitudinal data, such as infectious disease data that was measured repeatedly.

Quantile regression, an extension of median regression, can detect effect of covariates to location, scale and shape of the response distribution. Therefore it gives more view about response distribution. In the context of quantile regression in varying coefficient model, [5] proposes a two-step estimation procedure and [6] uses the basis function approach while [7] used P-splines quantile objective functions. In this paper we will focused on use of P-splines quantile regression in varying coefficient model. This method will applied to upper respiratory tract infection data in Bandung City that has been explored in [8] with update time variable to 2019.

According to [7] time varying coefficient model has form as follow

\[
Y(t_{ij}) = \beta_0 + \sum_{p=1}^{P} \beta_p(t_{ij})X^{(p)}(t_{ij}) + \varepsilon(t_{ij})
\]  

(1)
where \( i = 1, 2, 3, \ldots, n \); \( j = 1, 2, 3, \ldots, N \), in this case \( n \) is the number of subject to be measured and \( N \) is the number of repeated measurement of subject \( i \). This \( N \) indicates that this model can work on unbalance design. In case of balance design \( N \) equal to \( N \) for all \( i \). \( Y(t) \) is the response variable at time \( t \), \( X_k(t) \) is the \( k \)-th covariate at time \( t \) and \( \beta_k(t) \) is the \( k \)-th parameter at time \( t \). In this setting time \( (T) \) is a variable. \( \epsilon(t) \) is the error of the model that independent to \( X(t) \) and has \( \tau \)-th quantile equal to zero in the context of quantile regression.

2. Method

The method used to estimate the parameters in the regression model is Ordinary Least Square (OLS). The ideal conditions in the regression model are often not fulfilled due to various data conditions, therefore the use of OLS for data have not normal distribution (non-Gaussian) will be problematic. For non normal data conditions, a robust method is needed, which is not affected by data conditions.

Measure of locations is a robust measure that can be used in non-normal data conditions. Median regression is an alternative model that can be used to replace linear regression for non-Gaussian data. The median estimator is an observation that minimizes the \( L_1 \)-loss function for the median location parameter. Although robust median regression, like mean regression, provides little information about the distribution of the data. Sometimes we need to know more about other parts of the distribution.

Koenker and Basset generalized this idea to obtain a regression estimator for each quantile [9]. Using the median regression analogy, the minimization problem is now based on an asymmetric-weighted absolute residual. Currently, there are quite a lot of literatures regarding quantile regression, both in the parametric approach and the nonparametric approach that use a more flexible form of quantile regression function. Based on [10] quantile regression can be interpreted as a generalization of the median regression. Quantile regression is a regression technique that describes the relationship between the response variable and the explanatory variable at various quantiles, not on the centralized size (mean or median) of the response variable alone. Furthermore, we will explain about quantile regression based on [11], where the quantile on univariate data is called unconditional quantile and the quantile in the regression model is called conditional quantile.

Let \( Y_1, \ldots, Y_n \) is a random sample of the response \( Y \) with size \( n \) and have identical independent distribution. Estimate of quantile \( q(\tau) \) can be optimize by minimizing the expected loss function as follow

\[
q_\tau(Y) = \arg\min_c \frac{1}{n} \sum_{i=1}^{n} \rho_\tau(Y_i - c)
\]  

(2)

where \( \rho_\tau(.) \) that called check-function or pinball-function is a non-differentiable function then its cannot be minimizing by ordinary method. The formula of \( \rho_\tau(.) \) is analogous to the squared loss function in the context of least squares regression [9]. General formula of \( \rho_\tau(.) \) is

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

(3)

This is an unconditional quantile expression. Minimizing of the objective function on unconditional quantiles can be used in regression models. Let \( Y \) and \( X^{(1)}, \ldots, X^{(p)} \) are random variables that involved in the model (1). Model (1) can be written in matrix notation as follow

\[
Y_{n \times 1} = X_{n \times (p+1)}^{T} \beta_{p \times 1} + \varepsilon_{n \times 1}
\]  

(4)

where \( \varepsilon \) is error term. Conditional quantile for model (4) that is a function of response \( Y \) given matrix \( X \) is

\[
q_\tau(Y|X) = X^{T} \beta_\tau
\]  

(5)

where:

\[ \beta_\tau = (\beta_0^\tau, \beta_1, \ldots, \beta_p)^T \]
\[ \beta_0^p = \beta_0 + F_{\varepsilon}^{-1}(\tau) \]

\[ F_{\varepsilon}^{-1}(\tau) = \inf \{ u : F_n(u) \geq \tau \} : \text{the } \tau\text{-th quantile value of error } \varepsilon \text{ that assumed independent to } X. \]

Therefore for \(0 \leq \tau_1 \leq \tau_2 \leq 1\) applies \(F_{\varepsilon}^{-1}(\tau_1) \leq F_{\varepsilon}^{-1}(\tau_2)\).

Estimate of parameter can be obtain by minimizing following objective function

\[ \min_{\beta} \mathbb{E} \left[ \rho_\tau(Y - X^T \beta^\tau) \right]. \tag{6} \]

Let random samples of size \(n\) which identically independently distribution are \((X_1^{(1)}, \ldots, X_1^{(p)}, Y_1), (X_2^{(1)}, \ldots, X_2^{(p)}, Y_2), \ldots, (X_n^{(1)}, \ldots, X_n^{(p)}, Y_n)\) then empirical objective function for (6) is

\[ \min_{\beta} \frac{1}{n} \sum_{i=1}^{n} \rho_\tau(Y_i - X_i^T \beta^\tau), \tag{7} \]

where \(X_i = (1, X_i^{(1)}, \ldots, X_i^{(p)})^T\) is observation \(i\)-th of \(X\). Minimizing (7) to \(\beta\) will get \(\hat{\beta}\) and the estimator of conditional quantile function (5) can be obtained.

Conditional quantile is a quantile for regression model (4) and also quantile for varying coefficient model. Conditional quantile of \(Y(T)\) given \((X(T), T) = (X_{ij}, t_{ij})\) based on (6) has a form following expression

\[ q_\tau(Y(t_{ij})|X_{ij}, t_{ij}) = \inf \{ y : P\{Y(T) \leq y|(X(T), T) = (X_{ij}, t_{ij})\} \geq \tau \} \tag{8} \]

where \(0 \leq \tau \leq 1\) with \(t_{ij}\) is the time point at \(ij\)-th of time variable \(T\). Therefore quantile regression on varying coefficient model (1) can be written as following

\[ q_\tau(Y(t)|X(t), t) = X^T(t) \beta^\tau(t) \tag{9} \]

where \(\beta_k(t_{ij}), \text{ for } k = 1, \ldots, p,\) is an approximation of normalized B-splines function with degree \(v_k\). This beta function is

\[ \beta_k(t_{ij}) \approx \sum_{l=1}^{m_k} \alpha_{kl} B_{kl}(t_{ij}; v_k) \tag{10} \]

\(B_{kl}(t_{ij}; v_k)\) function, at \(l = 1, \ldots, u_k + v_k\) is a basis B-splines with degree \(v_k\) and \(u_k + 1\) is an equidistance knots for \(k\)-th component, where \(m_k = u_k + v_k\).

Unknown coefficients \(\alpha = (\alpha_0^T, \ldots, \alpha_p^T)^T\) where \(\alpha_0^T = (\alpha_{k1}, \ldots, \alpha_{nk})^T\) should be estimate by minimizing the following objective function

\[ S(\alpha) = \sum_{i=1}^{n} \sum_{j=1}^{N_i} \rho_\tau \left( Y_{ij} - \sum_{k=0}^{p} \sum_{l=1}^{m_k} \alpha_{kl} B_{kl}(t_{ij}; v_k) X_{ij}^{(k)} \right) + \sum_{k=0}^{p} \sum_{l=1}^{m_k} \lambda_k |\Delta_k^d \alpha_{kl}|^\gamma \tag{11} \]

where \(\gamma > 0, \lambda_k > 0\) for \(k = 0, 1, \ldots, p\) is smoothing parameter that control goodness-of-fit and penalty. \(\Delta_k^d\) is a differencing operator with \(d\)-th order

\[ \Delta_k^d \alpha_{kl} = \sum_{t=0}^{d_k} (-1)^t \binom{d_k}{t} \alpha_{k(l-t)} \tag{12} \]

where \(d\) is a natural number. Penalty is added to B-splines to avoid overfitting that can be maintained many knots but limited the effect. This approach is called P-splines that a combination of B-splines and penalties [12].

Quantile objective function (11) contain \(\rho_\tau(.)\) is a check-function that was explained before. One of the methods can be used to maximize or minimize some linear function with linear constraints is linear programming problem. Constraints use in linear programming can be an equality or inequality. Based on [13] and [14], minimizing quantile objective function (10) can be done by using interior point methods on linear programming.
3. Result and Discussion

Upper respiratory tract infection is a serious problem for public health caused by boca virus. Both of children and adult can be infected by boca virus. Upper respiratory tract infection is not only health problem but also an economic problem. It has a cost to society for treatment and covers for absenteeism from school and works. Upper respiratory tract infection especially to toddler is one of the infectious diseases in Bandung City, West Java Indonesia that has become a serious concern of the government due to the rapid transmission and high numbers of incidence of the disease. In 2019 upper respiratory tract infection cases in Bandung City is higher than the other city in West Java, Indonesia. The high incidence of Upper respiratory tract infection in Bandung City is caused by the high population density and high mobility of the population. Develop an early warning system is needed to reduce the negative impact of Upper respiratory tract infection and controlling the spread of this disease.

Upper respiratory tract infection data was measured repeatedly over time from 2012 until 2019 under different conditions for different 30 sub district in Bandung City. This data have longitudinal structures, it means that repeated measurements are nested within the districts. This structures cause dependence and contain heteroscedasticity among observation. This dependencies may caused by a spatial structures of the data, because of the complexity the spatial dependence not noticed in this paper. The data used in this paper is data regarding incident rate of upper respiratory tract infection of toddler in Bandung City. The data used are secondary data based on the yearly report of Bandung City Health Office. The response variable used in this study is the incident rate of the upper respiratory tract infection case of toddler in Bandung City and the covariates are percentage of toddler who received exclusive breastfeeding (breastmilk), percentage of malnutrition and percentage of toddler who received vitamin a.

Exploration and estimation computing of the data using some function of R software. The estimation coefficients based on [7] using already package called QRegVCM. The package is working for varying coefficient model with P-splines quantile regression method both for homoscedastic or heteroscedastic error. Besides that, the package contain handling of non-crossing quantile curve. For more details, we can see R documentation of QRegVCM [15].

The following figures will be shown data exploration of this data. The first figure is a scatter plot of a response and a covariate with linear regression line of each year. In this article we only show a scatter plot of incident rates and percentage of toddler who received exclusive breastfeeding (breastmilk). The other scatter plots in general have same feature with this scatter plot. The second figure is plot of the response by repeated measure points. In this case is a plot between incident rate and year.

![Figure 1. Scatter plot of incident rates by breastmilk](image-url)
Figure 1 shows that scatter plot of upper respiratory tract infection incident rate of toddler in Bandung City by percentage of toddler who received exclusive breastfeeding (breastmilk). Fitted linear regression lines for each year were added on the scatter plot indicated by different type and colour. As seen at Figure 1, there is an intersection among the regression lines. This shows that there is a variation in the regression coefficient, both intercept and slope. According to [1], this condition is an indication of the varying coefficient model so that we analyze this data using varying coefficient model.

Next figure is a plot of upper respiratory tract infection incident rate of toddler in Bandung City by year. The figure shows the fluctuation of incident rates over years in general. As we can see on Figure 2 there is downward trend from this data, and the trend seem like not linear. Therefore quantile regression will be used in our model and P-splines approximation of quantile objective function to estimate the model parameters.

![Figure 2. Plot incident rates by years](image)

As we explained before, the model built for the incident rate of the upper respiratory tract infection data is a varying coefficient model using the P-splines quantile regression method. In this paper quantile regression is also used to divide data into four groups (low, medium, high, very high) thus there are three quantile levels, 25%, 50% and 75%. We were setting same knots for all coefficients equal to 2 and also same degrees for all coefficients equal to 2.

Figure 3 shows the estimated values of the incident rate upper respiratory tract infection in Bandung City from 2012 to 2019 for the 25th quantile (indicate by blue dash line), the 50th quantile (indicate by green dot line) and the 75th quantile (indicate by red dash dot line). The plot shows that in general the incident rate is decreasing over year. As can we seen on this figure, the incident rates data grouped by the quantile lines into four groups namely low, medium, high and very high. A Low group for the data less than 25th quantile line, medium for the data between 25th and 50th quantile line, high the data between 50th and 75th quantile line and very high for the data more than 75th quantile line. Other than that, there is a potential crossing line between 25th and 50th quantile at 2019. The crossing quantile curve not allowed in quantile regression, therefore a non crossing in quantile regression is a further discussion which can be seen at [11].
Figure 3. Quantile plot incident rate of upper respiratory tract infection in Bandung City

Estimation of varying coefficient model produce regression coefficients change over time. The following figures show the change in intercepts and slopes of regression coefficients for the 25th quantile, the 50th quantile and the 75th quantile.

Figure 4. Intercept and Slope for quantile 0.25, 0.5 and 0.75

The estimated intercepts and slopes values for each quantile change over time shows in Figure 4. The pattern of the intercept coefficient at 25th quantile decrease linearly over time and so it is with the
slope decrease over time. In other side, the intercept coefficient at 50\textsuperscript{th} quantile decrease over time but the slope in has increase s-shape over time. As same as 25\textsuperscript{th} quantile, the 75\textsuperscript{th} quantile has intercept and slope decrease over time.

4. Conclusion
Longitudinal data structure of incidence rate of upper respiratory tract infection of toddler in Bandung City was measured repeatedly from 2012 to 2019. The observation unit is sub district of Bandung City. Based on data exploration result the incidence rate of upper respiratory tract infection in Bandung, it is analysed by P-splines quantile regression in varying coefficient models. The results of R Package QregVCM that the data show the incident rate is decreasing over year. Other than that, there is a potential crossing curves line between quantile 25 and 50. Because the crossing quantile curve has not be allowed in quantile regression, then a non crossing in quantile regression will be discuss as further topic.

The results of estimated regression coefficients change over time. Every quantile shows a different pattern of the estimated regression coefficient changes over time. In general, the intercept coefficient patterns are decrease over time. However, the slope coefficients have a different pattern for the 50\textsuperscript{th} quantile which fluctuation increase over time, but for the 25\textsuperscript{th} and the 75\textsuperscript{th} quantile monotone decrease over time.

According to the results, incidence rate of upper respiratory tract infection of toddler data in Bandung City generally has a downward trend. Using 25\textsuperscript{th}, 50\textsuperscript{th} and 75\textsuperscript{th} quantiles, they show the division of incident rates data into four groups, namely low, medium, high and very high. These results provide an overview of district groups. These results can be taken into consideration in determining sub-district priorities in the policy of the Bandung City Health Service in the prevention and control of upper respiratory tract infection of toddler cases in Bandung City.

Acknowledgments
Authors wish to thank Rector of Universitas Padjadjaran and this work is funded by Unpad doctoral dissertation research grant program (RDDU) with contract number: 5884/UN6.D/LT/2019. The important and constructive recommendation to improve the quality of this paper is also gratefully acknowledged.

References
[1] Hastie T and Tibshirani R 1993 Varying-coefficient models Journal of the Royal Statistical Society: Series B (Methodology) 55 4 757–779
[2] Fan J and Zhang W 2008 Statistical methods with varying coefficient models Statistics and its Interface 1 1179
[3] Wang L, Li H, and Huang J Z 2008 Variable selection in nonparametric varying-coefficient models for analysis of repeated measurements Journal of American Statistical Association 103 1556–1569
[4] Antoniadis A, Gijbels I, and Verhasselt A 2012 Variable Selection in Varying-Coefficient Models Using P-Splines Journal of Computational and Graphical Statistics 21 3 638–661
[5] Mu Y and Wei Y 2009 A dynamic quantile regression transformation model for longitudinal data Statistica Sinica 19 1137–1153
[6] Wang H J, Zhu Z, and Zhou J 2009 Quantile regression in partially linear varying coefficient models The Annals of Statistics 37 6B 3841–3866
[7] Andriyana Y, Gijbels I, and Verhasselt A 2014 P-splines quantile regression estimation in varying coefficient models TEST 23 1 153–194
[8] Tantular B Jaya I G N M and Andriyana Y 2019 Longitudinal Data Exploration of Modelling of Upper Respiratory Tract Infections in Bandung City Journal of Physics: Conference Series 1265 012026 1–8
[9] Koenker R and Bassett J G 1978 Regression quantiles *Econometrica: Journal of the Econometric Society* 46 1 33–50

[10] Koenker R, 2005 *Quantile Regression* (New York: Cambridge University Press)

[11] Andriyana Y, Gijbels I, and Verhasselt A 2018 Quantile regression in varying-coefficient models: non-crossing quantile curves and heteroscedasticity *Statistical Papers* 59 1589–6121

[12] Eilers P H C and Marx B D 1996 Flexible smoothing with B-splines and penalties *Statistical Science* 11 2 89–121

[13] Portnoy S and Koenker R 1997 The Gaussian hare and the Laplacian tortoise: computability of squared-error versus absolute-error estimators *Statistical Science* 12 4 279–300

[14] Koenker R and Ng P 2005 A Frisch-Newton algorithm for sparse quantile regression *Acta Mathematicae Applicatae Sinica* 21 2 225–236

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