IRAQI STATISTICIANS JOURNAL

https://isj.edu.iq/index.php/isj

ISSN: 3007-1658 (Online)

Comparison of Weighted Estimation Methods for the Varying Coefficient Quantile Regression Model in the Case of Longitudinal Data

Ali Mohammed Farhan[1], Mohammed Sadiq Abdul Razzaq[2]

[1,2] Department of Statistics , College of Administration and Economics , University of Baghdad , Iraq

ARTICLE INFO

Article history:

Received	14/11/2024
Revised	16/11/2024
Accepted	14/1/2025
Available online	15/5/2025

Keywords:
Longitudinal data analysis,
Varying-coefficient,
Quantile regression,
Weighted local polynomial,
Weighted Spline

ABSTRACT

In this study, two weighted methods are proposed for estimating the varying coefficient quantile regression model in the case of longitudinal data. The first method is the Weighted Spline Method, and the second is the Weighted Local Polynomial Method. Both methods account for within-subject correlations, which are addressed by incorporating weights derived from the empirical likelihood of each method. Five levels of quantiles were examined, and a simulation study was conducted to compare the two methods under different conditions. The methods were applied to the success rates data of the third intermediate grade in Al-Diwaniyah Governorate, assessing the impact of four explanatory variables on the success rates of 337 middle and secondary schools over five years. Results indicated that the efficiency of the methods varied across quantile levels: the Weighted Spline Method was more efficient at high and low quantile levels, whereas the Weighted Local Polynomial Method proved more efficient at intermediate levels.

1. Introduction

Regression analysis is undeniably one of the most important topics in statistics, focusing on studying and analyzing the relationship between a response variable and one or more explanatory variables. It has wide applications in various fields such as medicine, economics, and sociology. Regression analysis comes in various forms, one of which is quantile regression proposed by (Koenker and Bassett, 1978). On the contrary ordinary regression, quantile regression measures the relationship between variables by estimating conditional quantile functions of the response variable. this provides researchers with insights into the relationship between variables across different conditional distributions, especially those at the beginning or end of the data. In recent years, researchers have shown significant interest in quantile regression for longitudinal data, highlighting the importance of longitudinal data for containing more information than time series or cross-sectional data. Longitudinal data integrates both types of data, providing richer information about the phenomenon under study. to estimate model parameters and analyze the relationship between variables, there are parametric, semi-parametric, and non-parametric methods. Parametric methods often make strict assumptions that may be challenging to apply in practice, whereas non-parametric methods have fewer assumptions. to increase flexibility, we use varying coefficient models, which allow the effect of explanatory variables on the response variable to vary based on the values of other variables, such as time or any other variable.

[*] Corresponding author: E-**mail** address: ali.mohammed1101a@coadec.uobaghdad.edu.iq

https://doi.org/10.62933/q04zh103

The problem addressed in this research revolves around the fact that when studying mean regression models or ordinary regression, these models provide the effect of explanatory variables on the response variable but do not offer a comprehensive view of this effect at different levels of the distribution. One of the challenges researcher's faces is that some data used in regression analysis do not meet the assumption of normal distribution, or they may not satisfy the assumptions regarding random error or suffer from skewness**.** Traditional regression models, such as ordinary least squares (OLS), are based on estimating the conditional mean of the response variable. While effective under certain conditions, these models become inadequate when data deviate from normality, contain outliers, or exhibit skewed distributions. Moreover, they fail to provide a comprehensive understanding of the relationship between explanatory variables and the response variable across different quantiles of the distribution, as they focus solely on the central tendency. The challenge becomes more complex when dealing with longitudinal data, which are characterized by repeated measurements over time and exhibit within-subject correlations. Ignoring these correlations may lead to inefficient and biased estimators. Therefore, there is a pressing need for more flexible models that can capture changes in the effects of explanatory variables over time or across other indexing variables, while properly accounting for the unique features of longitudinal data. This study addresses the problem of modeling variable relationships under such conditions, particularly when classical assumptions are violated and traditional regression approaches prove insufficient.

This study aims to investigate the quantile regression model with varying coefficients in the context of longitudinal data by:

1. **Estimating the model parameters** while respecting the structure and characteristics of longitudinal data.
2. **Taking within-subject correlations into account** to improve the efficiency and accuracy of the estimators.
3. **Comparing different weighted estimation methods** to assess their performance and effectiveness.
4. **Overcoming challenges arising from the violation of classical assumptions**, such as normality of errors or homoscedasticity.
5. **Providing a more comprehensive view** of how explanatory variables influence the response variable across different parts of its distribution, not just the mean.

Many previous studies have discussed this topic, among which we will review:

(Yu and Jones 1998) studied the nonparametric estimation of quantile regression using locally weighted linear methods. They proposed two estimators and concluded that both estimators are effective. (Wu and Chiang 2000) Proposed two types of kernel estimators based on the local two-stage least squares method to estimate time-varying coefficients for the varying coefficient regression model in the case of longitudinal response data and cross-sectional explanatory variables. (Karlsson 2007) developed a weighted version of quantile regression for nonlinear longitudinal data. (Kim 2007) Studied the varying coefficient regression model and presented a methodology for estimating the model using a Polynomial spline function. (Mu and Wei 2009) proposed a two-stage estimation method for estimating the nonparametric varying-coefficient quantile regression model in the case of longitudinal data. In this method, the nonparametric spline estimator was used for estimation. (Tang and Leng 2011) proposed a novel approach for estimating quantile regression with longitudinal data using the empirical likelihood function while considering within-subject correlation. (Wang and Zhu 2011) developed an approach for two empirical likelihood inference procedures for quantile regression in the case of longitudinal data. These proposed methods do not require the estimation of the unknown error density function and within-subject correlation. (Fu and Wang 2012) proposed a method for estimating a linear quantile regression model for longitudinal data, integrating within-subject and between-subject estimates, which includes the correlations between repeated measurements. (Saifalddin and Rasheed 2013) reviewed some nonparametric techniques for estimating time-varying coefficient functions in the context of the nonparametric varying-coefficient model for balanced longitudinal data. the techniques employed included local linear boundary regression and cubic smoothing spline techniques. (Rashed and Rasheed 2014) studied the varying coefficient model as well as the partial varying coefficient model. Both the varying coefficient models (VCM) and the partial varying coefficient model (PVCM) were estimated using nonparametric and semi-parametric estimation methods. (Badr 2016) Employed the nonparametric regression method to diagnose and estimate the longitudinal data model in cases where certain assumptions regarding the random error vector are not met, particularly in the problem of heteroscedasticity and autocorrelation problem. (Liu 2017) proposed a two-stage locally weighted estimation method. In the first stage, initial estimators are found using B-splines. In the second stage, the model is transformed into a varying-coefficient regression model, and the locally weighted likelihood method is applied to estimate the varying coefficient functions. (Kim and Cho 2018) proposed two types of weights for estimating the varying-coefficient regression model in order to address within-subject correlations. The first type is global weight, which incorporates all observations in its calculation, while the second type is local weight, which considers nearby observations. (Lin et al. 2020) developed a weighted approach to enhance the efficiency of the varying-coefficient autoregressive model for longitudinal data. They obtained the weights via empirical likelihood method, utilized spline method for obtaining smoothers.

2. Quantile regression varying-coefficient model in the case of longitudinal data

The quantile regression varying-coefficient model in the case of longitudinal data can be written as follows (Tang et al., 2013):

$$Y_{ij} = \boldsymbol{X}_{ij}^T \boldsymbol{\beta}_\tau(\tau, t_{ij}) + e_{ij}(\tau) \qquad (1)$$

Where:

Y_{ij}: The response variable, for the jth observation of the ith subject

X_{ij}^T: The explanatory variables

$\beta_\tau(\tau,)$:) Is the vector of unknown smooth functions at the quantile level τ

$e_{ij}(\tau)$: The random error

t_{ij}: Represent the factors that modify the coefficients of x_{ij}^Tthrough unspecified functions $\beta(\,)$. The dependency of $\beta(\,)$on t_{ij} involves a specific type of interaction between t_{ij} and x_{ij}^T in some cases; the variable t_{ij} might be time follows (Hoover, et al., 1998).

In this section, two methods for estimating the quantile regression varying-coefficient model in the case of longitudinal data will be discussed.

2.1 The Weighted spline method for estimating VCQR models (WSP)

We attempt to improve the efficiency of the estimators by adding weights that enhance their efficiency. These weights are obtained from the robust empirical likelihood method for varying-coefficient mean regression models (VCMR), not from quantile regression models, because integrating correlations with quantile regression is extremely challenging with longitudinal data. By excluding the quantile aspect to find the weights, we then proceed to estimate the quantile model as follows:

Assume that Y_{ij} is the response variable of the jth observation in the ith subject and $\boldsymbol{Z}_{ij} = (X_{ij,1}, \dots, X_{ij,p}, T_{ij})^T = (\boldsymbol{X}_{ij}^T, T_{ij})^T \in \mathbb{R}^{p+1}$represents the explanatory of the jth observation in the ith subject, with $= 1,2, \dots, n$ ،$j = 1,2, \dots, m_i$, where T_{ij} typically denotes time but can be any other index. Assume that $T_{ij} \in [0,1]$ At a given quantile level $\tau \in (0,1)$, the varying-coefficient quantile regression (VCQR) model is given by;(Lin et al.2020):

$$Q_\tau\left(Y_{ij} \mid \boldsymbol{Z}_i\right) = \boldsymbol{X}_{ij}^T \boldsymbol{\beta}_\tau\left(T_{ij}\right) \qquad (2)$$

Where:

$Q_\tau\left(Y_{ij} \mid Z_i\right)$: The conditional quantile function of Y_{ij} given Z_i at quantile level τ.

i: The index for subjects, ranging from 1 to n

j: The index for observations within a subject, ranging from 1 to m_i

$Z_i = \left(Z_{i1}, \dots, Z_{im_i}\right)^T$: The vector of explanatory variables for the jth observation in the ith subject, defined as $Z_{ij} = \left(X_{ij,1}, \dots, X_{ij,p}, T_{ij}\right)^T = \left(X_{ij}^T, T_{ij}\right)^T$, where Xij, k are the explanatory variables and T_{ij} is typically the time variable.

T_{ij}: Usually represents time but can be any other index. It is assumed to be in the interval [0,1].

$\boldsymbol{\beta}_\tau(t) = \left(\beta_{1,\tau}(t), \dots, \beta_{p,\tau}(t)\right)^T$: An unknown smooth vector-valued function of t, representing the varying coefficients at quantile level τ.

The error vector is defined as:

$$e_{ij}(\tau) = Y_{ij} - Q_\tau\left(Y_{ij} \mid Z_i\right) \qquad (3)$$

Assuming that the conditional quantile of order τ for $e_{ij}(\tau)$) given $\boldsymbol{Z}_i$ is zero, and assuming that the observations Y_{ij} are independent, and thus the errors $e_{ij}(\tau)$ are independent, by using the spline method, the nonparametric coefficient functions in equation (2) can be written as (Lin et al, 2020):

$$E\left(Y_{ij} \mid \boldsymbol{Z}_i\right) = \boldsymbol{X}_{ij}^T \boldsymbol{\beta}\left(T_{ij}\right) \qquad (4)$$

Where:

$\beta(t) = \left(\beta_1(t), \dots, \beta_p(t)\right)^T$: is the unknown vector of smooth functions.

The random errors are:

$$\epsilon_{ij} = Y_{ij} - \mathrm{E}\left(Y_{ij} \mid Z_i\right) \qquad (5)$$

And by replacing the model $\mathrm{E}\left(Y_{ij} \mid Z_i\right)$ we get:

$$\epsilon_{ij} = Y_{ij} - \boldsymbol{X}_{ij}^T \boldsymbol{\beta}\left(T_{ij}\right) \qquad (6)$$

and

$$E\,(\epsilon_{ij}/Z_i) = 0$$

Assuming that $E\,(\epsilon^2{}_{ij}/Z_i)$ depends only on T_{ij} , denoted by $\sigma^2(T_{ij})$ using spline functions to estimate $\boldsymbol{\beta}(t)$we get:

$$E\left(Y_{ij} \mid Z_i\right) = \sum_{k=1}^{p} X_{ij,k} \pi\left(T_{ij}\right)^T \boldsymbol{\alpha}_k = \Pi_{ij}{}^T \alpha \; (7)$$

Where:

$$\Pi_{ij}{}^T \alpha = (X_{ij,1}\pi\left(T_{ij}\right)^T, \dots, X_{ij,p}\pi\left(T_{ij}\right)^T)^T$$

$\pi(t) = \left(\beta_1(t)\ , \dots, \beta_{k_n{}^* + l^*}\right)$: are spline functions of order l with $k_n$knots, and $\alpha = \left(\alpha_1{}^T, \dots, \alpha_p{}^T\right)^T$: T are the spline coefficients.

To determine the within-subject correlation matrix, we use AR1 (autoregressive of order one) or CS (compound symmetry) or both:

$$R_i = corr\{(\varepsilon_{i1}, \dots, \varepsilon_{im_i})^T\} \qquad i = 1, \dots, n \; (8)$$

The idea is to approximate the inverse of the within-subject correlation matrix using the quadratic inference function (QIF) by modelling some basic matrices (which will be described later) as linear combinations (Qu and Li 2006; Qu et al, 2000):

$$R_i{}^{-1} = \sum_{i=1}^{s} a_{ij} M_{ij} \qquad (9)$$

Where:

Mi1: is the identity matrix.

Mij(i $\neq$ 1): are symmetric matrices.

s: an integer representing the number of basis matrices.

To select the basis matrices, one of the two commonly used schemes by researchers is employed (Song et al.2009; Kim and Cho2018):

1. Symmetric working matrices where S = 2:
$M1$ is the identity matrix.
$M2$ is a matrix with zeros on the main diagonal and ones elsewhere.

2. First-order autoregressive model where S = 3: $M1$: is the identity matrix.
$M2$: is a matrix with ones in the upper and lower diagonals around the main diagonal and zeros elsewhere.
$M3$: is a matrix with ones in the first and last positions of the main diagonal and zeros elsewhere.

To obtain estimates, we solve the following estimating function:

$$g_i(\boldsymbol{\alpha}) = \begin{pmatrix} \boldsymbol{\Pi}_i^T \boldsymbol{A}_i^{-\frac{1}{2}} \boldsymbol{M}_{i1} \boldsymbol{A}_i^{-\frac{1}{2}} (\boldsymbol{Y}_i - \Pi_i \boldsymbol{\alpha}) \\ \vdots \\ \boldsymbol{\Pi}_i^T \boldsymbol{A}_i^{-\frac{1}{2}} \boldsymbol{M}_{is} \boldsymbol{A}_i^{-\frac{1}{2}} (\boldsymbol{Y}_i - \Pi_i \boldsymbol{\alpha}) \end{pmatrix} \qquad (10)$$

Where:

$\boldsymbol{Y}_i = (\boldsymbol{Y}_{i1}, \dots, \boldsymbol{Y}_{im_i})^T$
$A_i = \mathrm{diag}(\sigma^2(t_{i1}), \dots, \sigma^2(t_{i1}))$
$\Pi_i = (\Pi_{i1}, \dots, \Pi_{im_i})^T$

The $\hat{\alpha}$ is the value that maximizes an empirical probability function via Empirical likelihood (EL) which is a statistical method used to estimate the probability function of data without assuming a specific distribution. Instead, this method relies on the available data to construct an empirical probability function and uses the data to determine the weights that provide the maximum possible likelihood for the sample, subject to certain constraints to ensure a correct estimate (Lv et al. 2019).

$$L(\boldsymbol{\alpha}) = \sup\left\{\prod_{i=1}^{n} \omega_i(\boldsymbol{\alpha})\right\} \qquad (11)$$

subject to the conditions:

$$1 - \omega_i(\boldsymbol{\alpha}) \geq 0$$

$$2 - \sum_{i=1}^{n} \omega_i(\boldsymbol{\alpha}) = 1$$

$$3 - \sum_{i=1}^{n} \omega_i(\boldsymbol{\alpha}) g_i(\boldsymbol{\alpha}) = \boldsymbol{0}$$

The empirical likelihood method does not require the estimation of variance, which can be complex in nonparametric or semiparametric regression models:

The empirical weights are obtained from:

$$\omega_i(\hat{\boldsymbol{\alpha}}) = n^{-1}\left\{1 + \lambda_{\hat{\alpha}}^T g_i(\hat{\boldsymbol{\alpha}})\right\}^{-1} \qquad (12)$$

where $\lambda_{\hat{\alpha}}$ و $\hat{\boldsymbol{\alpha}}$ are obtained by solving the following equations:

$$n^{-1} \sum_{i=1}^{n} \{\partial g_i(\hat{\boldsymbol{\alpha}})^T / \partial \hat{\boldsymbol{\alpha}}\} \lambda_{\hat{\alpha}} \left\{1 + \lambda_{\hat{\alpha}}^T g_i(\hat{\boldsymbol{\alpha}})\right\}^{-1} = \boldsymbol{0} \qquad (13)$$

and

$$n^{-1} \sum_{i=1}^{n} g_i(\hat{\boldsymbol{\alpha}}) \left\{1 + \lambda_{\hat{\alpha}}^T g_i(\hat{\boldsymbol{\alpha}})\right\}^{-1} = 0 \qquad (14)$$

After determining the weights, the estimator $\hat{\boldsymbol{\alpha}}_\tau$ is:

$$\hat{\boldsymbol{\alpha}}_{k,,\tau} = \underset{\boldsymbol{\alpha}_\tau \in \mathbb{R}^{pN}}{\mathrm{argmin}} \sum_{i=1}^{n} \omega_i(\hat{\boldsymbol{\alpha}}) \sum_{j=1}^{m_i} \rho_\tau\left(Y_{ij} - \boldsymbol{\Pi}_{ij}^T \boldsymbol{\alpha}_\tau\right) \qquad (15)$$

Where:

$$\hat{\boldsymbol{\alpha}}_{k,,\tau} = \left(\hat{\boldsymbol{\alpha}}_{1,\tau}{}^T, \dots, \hat{\boldsymbol{\alpha}}_{p,\tau}{}^T\right)^T \quad k = 1, \dots, p \qquad (16)$$

The varying coefficients $\beta_{k,\tau}$ are estimated from the following equation:

$$\widehat{\beta}_{k,\tau}(t) = \pi(t)^T \hat{\boldsymbol{\alpha}}_{k,,\tau} \ , \qquad k = 1, \dots, p \qquad (17)$$

2.2 Quadratic inference function and Basis matrices

The Quadratic Inference Function (QIF) is a statistical method for estimating models in the case of longitudinal data to improve estimates when observations are correlated within the same subject. Its idea relies on fundamental matrices that represent the correlation structure among these observations, where a quadratic estimation function is constructed based on these fundamental matrices.

(Qu et al. 2000) introduced the quadratic inference function (QIF) method and assumed that the inverse of the correlation structure can be estimated by the linear combination in equation (9):

where:

M_{i1}: is the identity matrix.

$M_{ij}(i \neq 1)$: are symmetric matrices.

s: an integer representing the number of basis matrices.

To select the basis matrices, one of the two commonly used schemes by researchers is employed (Song et al.2009; Kim and Cho2018):

1. Symmetric working matrices where $S = 2$:

M_1 is the identity matrix.

M_2 is a matrix with zeros on the main diagonal and ones elsewhere.

2. First-order autoregressive model where $S = 3$:

M_1 : is the identity matrix.

M_2: is a matrix with ones in the upper and lower diagonals around the main diagonal and zeros elsewhere.

M_3: is a matrix with ones in the first and last positions of the main diagonal and zeros elsewhere.

The advantage of this approach is that it does not require the estimation of linear coefficients (which can be considered nuisance parameters), as the generalized estimating equation is an approximate linear combination of the elements of the estimating function:

$$g_i(\alpha) = \begin{pmatrix} \mathrm{E}(Y_i)A_i^{-\frac{1}{2}}M_{i1}A_i^{-\frac{1}{2}}\left(Y_i - \mathrm{E}(Y_i)\right) \\ \vdots \\ \mathrm{E}(Y_i)A_i^{-\frac{1}{2}}M_{is}A_i^{-\frac{1}{2}}\left(Y_i - \mathrm{E}(Y_i)\right) \end{pmatrix} \quad (18)$$

Where:

$A_i = \mathrm{diag}(\sigma^2(t_{i1}), \dots, \sigma^2(t_{i1}))$

The QIF is characterized by its ability to handle complex correlation structures and provides less biased and more efficient estimates compared to other methods. It offers flexibility in the choice of fundamental matrices that can be used to enhance the estimation (Song et al. 2009**)**

2.3 interior knots

The selection of the number of knots in the spline method for estimation is of great importance, as the degree of smoothing depends on determining the number of knots. The knots allow for control over the coefficient function and play a crucial role in balancing the bias and variance components, which together form the Mean Squared Error (MSE). Therefore, increasing the number of knots leads to undesirable smoothing, and vice versa.

There are several methods for selecting knots, including simple methods such as choosing a knot at every point where there is a high density of data or mathematical methods that rely on empirical analysis, where the model's performance is evaluated with different numbers of knots. In this Dissertation, the Schwarz Criterion (SIC) will be used as follows (Schwarz 1978; Tang et al., 2013; Lin et al. 2020):

$$SIC(K) = \log\left[\sum_{i=1}^{n}\sum_{j=1}^{m} p_r\left(Y_{ij} - \mathbf{\Pi}_{ij}^{T}\hat{\boldsymbol{\alpha}}_{\tau(k)}\right)\right] + \frac{\log(n*m)}{2(n*m)}pN(K) \quad (19)$$

Where:

p: The number of explanatory variables

$N(K)$: $l + K$

l: The number of fixed parameters and

K : The number of interior knots

Where a number of models with different numbers of knots are constructed, and then the Schwarz Criterion is calculated for these models. The number of knots that achieves the lowest value of the Schwarz Criterion is selected.

2.4 The weighted local polynomial method for estimating VCQR models (WLP)

Equation (1) can be rewritten as follows:

$$Q_\tau\big(Y_{ij} \mid \boldsymbol{Z}_i\big) = \boldsymbol{X}_{ij}^T\boldsymbol{\beta}_\tau\big(T_{ij}\big) + e_{ij}(\tau) \qquad (20)$$

Under the assumption that the conditional quantile τ of $e_{ij}(\tau)$ given $\boldsymbol{Z}_i$ is equal to zero (Tang, and Leng, 2011) and that the observations Y_{ij} are independent, the errors $e_{ij}(\tau)$ are therefore independent. The estimated varying-coefficient regression model for longitudinal data can be expressed as follows in Equation (20) (Lin et al, 2020):

$$Q_\tau\big(\hat{Y}_{ij}/Z_i\big) = X_{ij}^T\beta_\tau\big(T_{ij}\big) \qquad (21)$$

The fundamental idea behind this method is to utilize weights in estimating the VCQR model, which can be summarized as follows: the coefficients vector is estimated according to the formula (Lin et al, 2020):

$$\hat{\boldsymbol{\alpha}}_\tau^{Wp} = \underset{\boldsymbol{\alpha}_\tau\in\mathbb{R}^{pN}}{\operatorname{argmin}} \sum_{i=1}^{n} \omega_i(\hat{\boldsymbol{\alpha}}) \sum_{j=1}^{m_i} \rho_\tau\big(Y_{ij} - \boldsymbol{\Lambda}_{ij}^T\boldsymbol{\alpha}_\tau\big) \qquad (22)$$

Where:

$$\Lambda_{ij} = \sum_{k=0}^{P} X_{ij}^T\,(T_{ij} - t_0)^k K_h\big(T_{ij} - t_0\big)$$

The weights are calculated through:

$$\omega_i(\hat{\boldsymbol{\alpha}}) = n^{-1}\big\{1 + \lambda_{\hat{\boldsymbol{\alpha}}}^T g_i(\hat{\boldsymbol{\alpha}})\big\}^{-1} \qquad (23)$$

$\lambda_{\hat{\boldsymbol{\alpha}}}$ and $\hat{\boldsymbol{\alpha}}$ are obtained by solving the following equations:

$$n^{-1}\sum_{i=1}^{n} \{\partial g_i(\hat{\boldsymbol{\alpha}})^T/\,\partial\hat{\boldsymbol{\alpha}}\}\lambda_{\hat{\alpha}}\big\{1 + \lambda_{\hat{\boldsymbol{\alpha}}}^T g_i(\hat{\boldsymbol{\alpha}})\big\}^{-1} = \boldsymbol{0} \qquad (24)$$

And

$$n^{-1}\sum_{i=1}^{n} g_i(\hat{\boldsymbol{\alpha}})\big\{1 + \lambda_{\hat{\boldsymbol{\alpha}}}^T g_i(\hat{\boldsymbol{\alpha}})\big\}^{-1} = \boldsymbol{0} \qquad (25)$$

Calculated through (α) $gi(\alpha)$ (Qu and Li , 2006)

$$g_i(\boldsymbol{\alpha}) = \begin{pmatrix} \boldsymbol{\Lambda}_i^T\boldsymbol{A}_i^{-\frac{1}{2}}\boldsymbol{M}_{i1}\boldsymbol{A}_i^{-\frac{1}{2}}\big(Y_i^k - \Lambda_i\boldsymbol{\alpha}\big) \\ \vdots \\ \boldsymbol{\Lambda}_i^T\boldsymbol{A}_i^{-\frac{1}{2}}\boldsymbol{M}_{is}\boldsymbol{A}_i^{-\frac{1}{2}}\big(Y_i^k - \Lambda_i\boldsymbol{\alpha}\big) \end{pmatrix} \qquad (26)$$

Where:

$Y_i^k = Y_i k_h\big(T_{ij} - t_0\big)$

$A_i = \operatorname{diag}(\sigma^2(t_{i1}), \dots, \sigma^2(t_{i1}))$

$\Lambda_i = (\Lambda_{i1}, \dots, \Lambda_{im})^T$

2.5 Smooth parameter h Selection

As it is well-known, the smoothing parameter (h) plays a crucial role in balancing bias and variance, the components from which the mean squared error is composed. Thus, the value of the mean squared error is influenced by the choice of this parameter. To select the smoothing parameter(h), we follow the following approach, where we rely on the nonparametric Akaike Information Criterion (AIC) (Cai and Xu, 2008), representing the corrected criterion for the bias of nonparametric regression models:

$$\mathrm{AIC}(h) = \{\hat{\sigma}_\tau^2\} + \frac{2(ph + 1)}{[nm - (ph + 2)]} \qquad (27)$$

Where:

$\hat{\sigma}_\tau^2$: is the estimated variance of the errors for quantile regression, calculated using the following formula:

$$\hat{\sigma}_\tau^2 = \sum_{i=1}^{n}\sum_{j=1}^{m} P_\tau\,\{Y_{ij} - X_{ij}^T\alpha_k(t_0)\}/nm \qquad (28)$$

ph : represents the nonparametric degrees of freedom, also known as the effective number of parameters, which depend on the trace of the hat matrix in the linear estimations in nonparametric quantile regression. There is no explicit expression for the hat matrix due to its non-linearity. However, we can use a first-order approximation to derive an explicit expression, which can be interpreted as an approximation of the hat matrix.

2.6 Comparison Criterion Mean Squared Error (MSE)

The comparison of estimation methods is a fundamental procedure in statistical research to identify the best estimation method. There are numerous comparison criteria, some for comparing parameters and others for the entire model. In general, the choice of the appropriate criterion depends on the nature of the data and the objectives of the statistical analysis. In this study , the comparison criteria Mean Squared Error (MSE) will be used .The idea of Mean Squared Error is summarized by calculating the squares of the differences between the true values of the response variable and the estimated values, then computing the mean of these squared differences. The formula for Mean Squared Error for a varying coefficient quantile regression model in the case of longitudinal data is as follows:

$$MSE = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m}\left(Y_{ij} - \hat{Y}_{ij}\right)^2}{n * m} \qquad (29)$$

Where:

$\hat{Y}_{ij}$ represents the estimated values of the response variable according to the estimation method.

MSE is advantageous because it is consistent, as it is used for comparing models in simulation experiments with different sample sizes. Additionally, it is sensitive to large differences between the true and estimated values because it squares these differences, making it useful for identifying models with poor estimates. Moreover, it is easy to calculate, understand, and interpret.

3. Results and discussion

3.1 Simulation:

In this section, we will provide a description of the simulation experiment used in this study, the statistical programming language R 4.3.1 was utilized to write the simulation program detailing all the different input scenarios according to the following varying coefficient quantile regression model:

$$Q_\tau\left(Y_{ij} \mid Z_{ij}\right) = \beta_{1,\tau}\left(T_{ij}\right)x_{ij1} + \beta_{2,\tau}\left(T_{ij}\right)x_{ij2} + \beta_{3,\tau}\left(T_{ij}\right)x_{ij3} + \beta_{4,\tau}\left(T_{ij}\right)x_{ij4} + \varepsilon_{ij}(\tau) \qquad (30)$$

Where:

Longitudinal Data:

Number of Subjects $i = 300$

Number of Observations within each Subject: $j = 5$

Number of explanatory variables $k = 4$

varying coefficient:

$\beta_1(t) = -\sin\left(\frac{\pi t^2}{3}\right)$, $\beta_2(t) = \left(\frac{\sqrt{t^3}}{2.4}\right)$

,$\beta_3(t) = -\cos\left(\frac{\pi\sqrt{t}}{2}\right)$, $\beta_4(t) = \left(\frac{(2t-6)^3}{5}\right)$

quantile levels:

five levels $\tau = 0.1 , 0.3 , 0.5 , 0.7 , 0.9$

Correlation coefficient:

$\rho = 0.5$

Correlation structure:

First-order Autoregressive $AR(1)\ and$ Compound symmetry CS

Generate time indicator:

are generated independently from Continuous uniform distribution $T_{ij} \sim U(0,1)$

Generate explanatory variables:

The explanatory variables were generated as follows:

$x_1 \sim Ber(0.5)$, $x_2 \sim Beta\ (0,1)$, $x_3 \sim N(13\ ,22)$, $x_4 \sim Bin(2,1/3)$

Generate a random error: three cases:

a. First case : Normal error $\varepsilon_i \sim N(0, R)$
b. Second case : Heteroscedastic symmetric error

$\varepsilon_i \sim N(0, V_i)$

$V_i = A_i^{-1} R A_i^{-1}$

$$A_i^{-1} = ding\left[0.7 - 0.3\left(\frac{T_i}{2}\right)^2\right]_{j=1}^{m}$$

c. Thread case : Heteroscedastic asymmetric error

$$\varepsilon_i = |e_i| - \sqrt{\frac{2}{\pi}(0.7 - 0.3\left(\frac{T_i}{2}\right)^2}\,)$$

$$T_i = (T_{i1}\,, \dots , T_{im})^T$$

$$e_i \sim N(0, V_i)$$

$$V_i = A_i^{-1} R A_i^{-1}$$

$$A_i^{-1} = ding\left[0.7 - 0.3\left(\frac{T_i}{2}\right)^2\right]_{j=1}^{m}$$

And R for the three cases above is:

1. $R(i,j) = \rho^{|i-j|}$ If AR(1)
2. $R(i,j) = \rho + (1-\rho).I(i=j)$ If CS , $I\ is\ inducer\ function:\ 1\ if\ \ i = j\,, 0\ \ if\ \ i \neq j$

Estimation methods:

a) The weighted spline method for estimating VCQR models (WSP)

b) The weighted local polynomial method for estimating VCQR models (WLP)

Comparison Criterion:

The Mean Squared Error (MSE) criteria were used to compare the effectiveness of the aforementioned estimation methods.

Number of repetitions:

The experiment was repeated 1000 times $r = 1000.$

Table (1) shows the MSE values for the Simulation through the results of the table 1 we find that When τ=0.1,0.9 we observe a general increase in the Mean Squared Error (MSE) across all methods, with the WSP method demonstrating the best performance at these levels, followed by the WLP method.

When τ=0.3,0.5,0.7 the Mean Squared Error (MSE) decrease for all methods compared to

the higher and lower quantile levels, with the WLP method demonstrating the best performance at these levels.

When errors are homogeneous, the Mean Squared Error (MSE) decrease for all methods across different quantile levels. However, in the case of heterogeneous errors, whether

symmetric or asymmetric, the Mean Squared Error (MSE) increase.

3.1 real data

The following varying coefficients quantile regression model for longitudinal data was assumed:

Table (1): the MSE values for the Simulation when($\rho = 0.5\,, n = 300, m = 5$)

τ	e_i	1	2	3	4	5	6	7	8	9	10	11	12
0.1	**WSP**	25.689	24.501	25.426	25.028	27.399	27.109	27.372	27.083	27.293	27.135	27.241	27.267
	WLP	26.773	25.535	26.499	26.084	28.555	28.253	28.528	28.226	28.445	28.281	28.391	28.418
0.3	**WSP**	22.745	21.693	22.512	22.16	24.259	24.002	24.236	23.979	24.166	24.026	24.119	24.142
	WLP	21.827	20.817	21.603	21.265	23.279	23.033	23.257	23.011	23.19	23.056	23.145	23.167

0.5	**WSP**	22.403	21.367	22.173	21.826	23.894	23.641	23.871	23.618	23.802	23.664	23.756	23.779
	WLP	21.605	20.606	21.384	21.049	23.043	22.8	23.021	22.777	22.955	22.822	22.91	22.933
0.7	**WSP**	25.711	24.522	25.448	25.05	27.423	27.133	27.396	27.106	27.317	27.159	27.265	27.291
	WLP	24.732	23.589	24.479	24.096	26.379	26.1	26.353	26.074	26.277	26.125	26.226	26.252
0.9	**WSP**	34.682	33.078	34.327	33.79	36.991	36.599	36.955	36.564	36.848	36.635	36.777	36.813
	WLP	35.639	33.991	35.273	34.722	38.011	37.609	37.974	37.572	37.865	37.645	37.792	37.828

e_i As follows :

1- Normalerror $R = CS, CS\ working\ correlation\ structure$

2- Normal error $R = AR(1), AR(1)\ working\ correlation\ structure$

3- Normal error $R = AR(1), CS\ working\ correlation\ structure$

4- Normal error $R = CS, AR(1)\ working\ correlation\ structure$

5- Heteroscedastic symmetric error $R = CS, CS\ working\ correlation\ structure$

6- Heteroscedastic symmetric error $R = AR(1), AR(1)\ working\ correlation\ structure$

7- Heteroscedastic symmetric error $R = AR(1), CS\ working\ correlation\ structure$

8- Heteroscedastic symmetric error $R = CS, AR(1)\ working\ correlation\ structure$

9- Heteroscedastic asymmetric error $CS, CS\ working\ correlation\ structure$

10- Heteroscedastic asymmetric error $R = AR(1), AR(1)\ working\ correlation\ structure$

11- Heteroscedastic asymmetric error $AR(1), CS\ working\ correlation\ structure$

12- Heteroscedastic asymmetric error $R = CS, AR(1)\ working\ correlation\ structure$

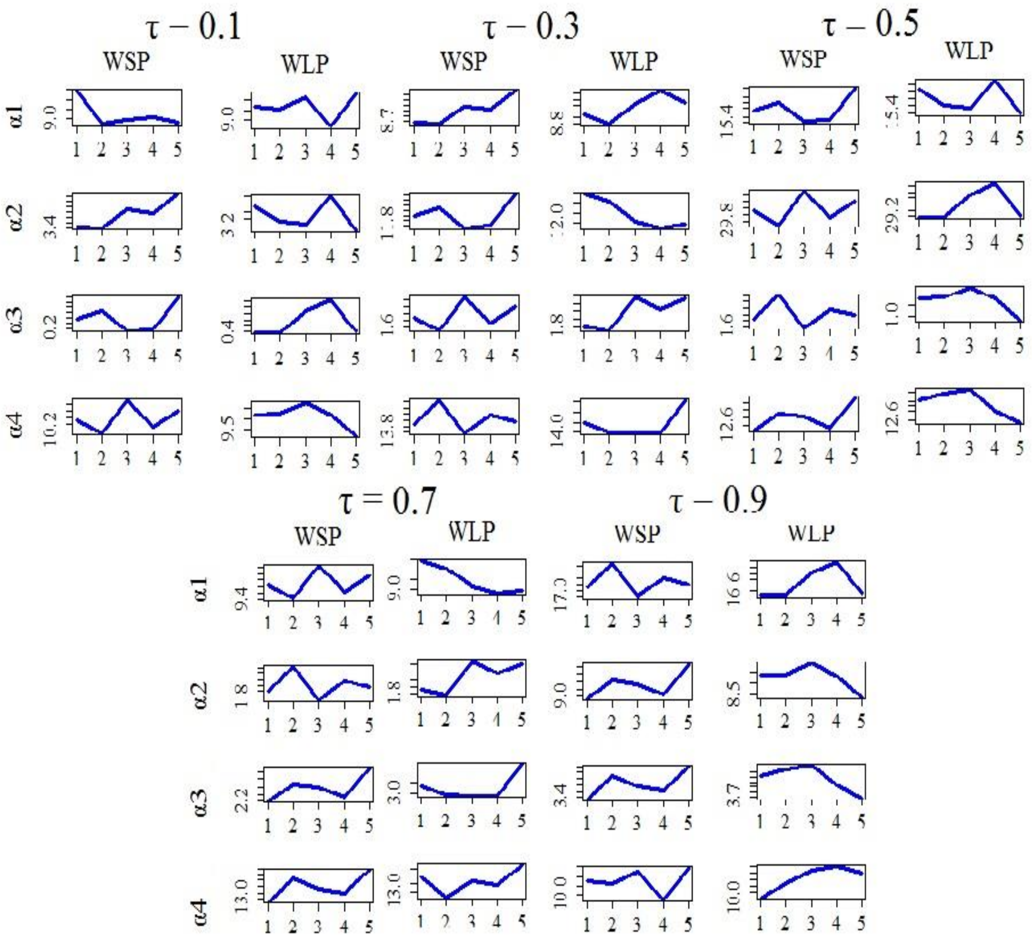

Figure 1. shows the estimated time-varying coefficient curves for real data

Table (2) : the MSE values for the real data

methods	**WSP**	**WLP**
τ		
0.1	23.5944	23.6098
0.3	20.2171	20.1122
0.5	19.0807	18.9135
0.7	22.2689	21.6039
0.9	28.1687	28.2496

$$Q_\tau(Y_{ij} \mid \boldsymbol{Z}_{ij}) = \alpha_{1,\tau}(T_{ij})x_{ij,1} + \alpha_{2,\tau}(T_{ij})x_{ij,2} + \alpha_{3,\tau}(T_{ij})x_{ij,3} + \alpha_{4,\tau}(T_{ij})x_{ij,4} \qquad (31)$$

Where:

$i = 1, \dots, 337$: represents the number of schools.

$j = 1, \dots, 5$: represents the number of observations for each school.

τ : represents the quantile level.

Y_{ij} : is the dependent variable representing the success rates for each school over five years.

$\boldsymbol{Z}_{ij}$: contains the covariate information.

T_{ij} : represents the measurement time in years, calculated by subtracting 2017 from the current year.

$x_{ij,1}$: is the first explanatory variable, a categorical variable indicating the type of school (0 if the school is private, 1 if the school is public).

$x_{ij,2}$: is the second explanatory variable representing the ratio of the number of teachers to the number of students in the same school over five years.

$x_{ij,3}$: is the third explanatory variable representing the average years of service for teachers in each school over five years.

$x_{ij,4}$: is the fourth explanatory variable, a categorical variable indicating the gender of the school's students (0 if the school is for boys, 1 if the school is coeducational, and 2 if the school is for girls).

$\alpha_{1,\tau}, \alpha_{2,\tau}, \alpha_{3,\tau}, \alpha_{4,\tau}$: are the time-varying coefficients representing the effect of the explanatory variables on the dependent variable. Through calculating the lag correlations, it was observed that the correlation structure is AR1.

From Figure1 which shows the estimated time-varying coefficient curves for real data and from Table (2) which shows: the MSE values for the real data we observe the following:

At the quantile level (0.1), the Weighted Spline method is more efficient than the Weighted Local Polynomial method,

At the quantile level (0.3), the Weighted Local Polynomial method is more efficient than the Weighted Spline method, the Spline method,

At the quantile level (0.5), the Weighted Local Polynomial method is more efficient than Weighted Spline method,

At the quantile level (0.7), the Weighted Local Polynomial method is more efficient than the Weighted Spline method.

At the quantile level (0.9), the Weighted Spline method is more efficient than the Weighted Local Polynomial method.

The Weighted Local Polynomial method is generally more efficient at most quantile levels, except at very low and very high quantile levels, where the Weighted Spline method proves to be more efficient. From Figure1, which represent the estimated functions of the time-varying coefficients for the estimation methods across five quantile levels, where the horizontal axis represents time and the vertical axis represents the possible values of the estimated functions, it was observed that there are positive effects of varying magnitudes for all explanatory variables on the response variable. The magnitude of the effect varies according to the quantile level and time, with the second and fourth explanatory variables having the greatest impact across different quantile levels, followed by the effects of the third and first variables, respectively.

4. Conclusions

From the theoretical, experimental, and applied aspects presented in the Dissertation, there are some important conclusions

can be summarized in the following points:

1. The weighted local polynomial method was the best among all methods at most quantile levels.

2. It is concluded that at very high and very low quantile levels, the weighted spline method was better than the other methods.

3. There is a varying positive effect of all explanatory variables in the model on the response variable.

4. While misspecification of the correlation structure weighted methods maintain their efficiency even when the correlation structure is misspecified.

5. The differences between the mean squared errors of the methods vary according to the types of errors used. In the first and second error formats, the differences are large and relatively decrease in the third error format.

6.The effect of the explanatory variables on the response variable varies according to the quantile level and the time point at which the relationship is estimated, highlighting the importance of the varying-coefficient model for predicting the future of the relationship. The effect of the explanatory variables on the response variable increases at the 0.5 quantile level.

References

[1] Badr, Duraid H. (2016). The diagnosis estimation of the nonparametric regression function of the panel data in Case some of its hypotheses are not verified. The Ph. D to the College of Administration and Economics, University of Baghdad.

[2] Cai, Z., & Xu, X. 2008. Nonparametric quantile estimations for dynamic smooth coefficient models. Journal of the American Statistical Association, 103(484), 1595-1608.

[3] Fu, L., & Wang, Y. G. (2012). Quantile regression for longitudinal data with a working correlation model. Computational Statistics & Data Analysis, 56(8), 2526-2538.

[4] Hoover, D. R., Rice, J. A., Wu, C. O., & Yang, L. P. (1998). Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. Biometrika, 85(4), 809-822.

[5] Karlsson, A. (2007). Nonlinear quantile regression estimation of longitudinal data. Communications in Statistics-Simulation and Computation, 37(1), 114-131.

[6] Kim, M. O. (2007). Quantile regression with varying coefficients .Ann. Statist. 35(1): 92-108.

[7] Kim, S., & Cho, H. R. (2018). Efficient estimation in the partially linear quantile regression model for longitudinal data. Electron. J. Statist. 12(1): 824-850

[8] Kim, S., & Cho, H. R. (2018). Efficient estimation in the partially linear quantile regression model for longitudinal data. Electron. J. Statist. 12(1): 824-850.

[9] Koenker, R., & Bassett Jr, G. (1978) . Regression quantiles. Econometrica: journal of the Econometric Society, 33-50.

[10] Lin, F., Tang, Y., & Zhu, Z. (2020). Weighted quantile regression in varying-coefficient model with longitudinal data. Computational Statistics & Data Analysis, 145, 106915.

[11] Liu, S. (2017). Efficient estimation of longitudinal data additive varying coefficient regression models. Acta Mathematicae Applicatae Sinica, English Series, 33, 529-550.

[12] Lv, J., Guo, C., & Wu, J. (2019). Smoothed empirical likelihood inference via the modified Cholesky decomposition for quantile varying coefficient models with longitudinal data. Test, 28, 999-1032.

[13] Mu, Y., & Wei, Y. (2009). A dynamic quantile regression transformation model for longitudinal data. Statistica Sinica, 1137-1153.

[14] Qu, A., & Li, R. (2006). Quadratic inference functions for varying-coefficient models with longitudinal data. Biometrics, 62(2), 379-391.

[15] Qu, A., Lindsay, B. G., & Li, B. (2000). Improving generalised estimating equations using quadratic inference functions. Biometrika, 87(4), 823-836.

[16] Rashed, Husam A. and Rasheed, Dhafir H. (2014). Comparison between the Local Polynomial Kernel and Penalized Spline to Estimating Varying Coefficient Model. Journal of Economics And Administrative Sciences, 20(78).

[17] Saifalddin, Ali. And Rasheed, Dhafir H. (2013). Comparison Robust M Estimate with Cubic Smoothing Splines for Time-Varying Coefficient Model for Balance Longitudinal Data. journal of Economics And Administrative Sciences, 19(73).

[18] Schwarz, G. (1978). Estimating the dimension of a model. The annals of statistics, 461-464.

[19] Song, P. X. K., Jiang, Z., Park, E., & Qu, A. (2009). Quadratic inference functions in marginal models for longitudinal data. Statistics in medicine, 28(29), 3683-3696.

[20] Tang, C. Y., & Leng, C.(2011) . Empirical likelihood and quantile regression in longitudinal data analysis. Biometrika, 98(4), 1001-1006.

[21] Tang, Y., Wang, H. J., & Zhu, Z. (2013). Variable selection in quantile varying coefficient models with longitudinal data. Computational Statistics & Data Analysis, 57(1), 435-449.

[22] Wang, H. J., & Zhu, Z. (2011). Empirical likelihood for quantile regression models with longitudinal data. Journal of statistical planning and inference, 141(4), 1603-1615.

[23] Wu, C. O., & Chiang, C. T. (2000). Kernel smoothing on varying coefficient models with longitudinal dependent variable. Statistica Sinica, 433-456.

[24] Yu, K., & Jones, M. (1998). Local linear quantile regression. Journal of the American statistical Association, 93(441), 228-237.