Periodicals of Engineering and Natural Sciences ISSN 2303-4521
Vol. 8, No. 2, June 2020, pp.572-579

Comparison between the estimated of nonparametric methods by using the methodology of quantile regression models

Marwa Khalil Ibrahim[1], Qutab N. Nayef Al-Qazaz [2]
[1] Department of statistics, College of Administration and Economics, University of Baghdad
[2] Head of statistics department, College of Administration and Economics, University of Baghdad

ABSTRACT

This paper study two stratified quantile regression models of the marginal and the conditional varieties. We estimate the quantile functions of these models by using two nonparametric methods of smoothing spline (B-spline) and kernel regression (Nadaraya-Watson). The estimates can be obtained by solve nonparametric quantile regression problem which means minimizing the quantile regression objective functions and using the approach of varying coefficient models. The main goal is discussing the comparison between the estimators of the two nonparametric methods and adopting the best one between them.

Keywords: Stratified quantile regression, quantile function; marginal model, conditional model; B-spline, kernel regression (Nadaraya-Watson), varying coefficient models.

Corresponding Author:

Marwa Khalil Ibrahim

University of Baghdad
College of Administration and Economics
Email: marwakhalil.202024@yahoo.com

1. Introduction

Quantile regression was first proposed by Koenker and Bassett in 1978, to conduct inference about conditional quantile functions, and instead of depicting the behavior of E (Y | X). As a result, quantile regression is capable of providing a more complete statistical analysis of the stochastic relationships among random variables, and a more comprehensive picture than just one mean function. First, we must define the quantile function of a random variable Y which denoted as $Q_Y(q)$, where it is the inverse function of the cumulative distribution function $F_Y(y)$, that is:

$$F_Y(y) = F(y) = P(Y \leq y) = P\big((-\infty, y]\big) \quad (1)$$

$$Q_Y(q) = Q(q) = F_Y^{-1}(q) = inf\{y \in \mathbb{R}: F(y) \geq q\} \quad (2)$$

Where $F_Y(y): \mathbb{R} \to [0,1]$ of a random variable Y and $q \in (0,1)$. If F is continuous, strictly increasing on $[a, b]$, then Q is continuous, strictly increasing based on $[F(a), F(b)]$ and $Q_Y(q) = y$ if $F_Y(y) = q$.

Quantile regression is a statistical tool intended for estimate purposes. There are several various techniques with applications about the quantile regression. These techniques are classical linear quantile regression, nonparametric quantile regression and the modified quantile regression besides the applications such as examining the effects of growth trajectories over time for risk prediction [1], [2].

The success which achieved by using the univariate quantile pushed the researchers to start to expand it to multivariate quantiles. Nevertheless, they found serious problems because the lack of a natural basis for ordering multivariate data, that means not possible to generalize the natural order for $\mathbb{R}^p$ when $p \geq 2$. The

nonparametric method plays an important role to study the dependence of the quantiles of a multivariate response conditional on a set of covariates. The number of covariates varies according to study requirements [3], [4].

We study the varying coefficient models where response, covariates, and regression coefficients are allowed to vary with t. These models allow to extend the applications of local regression techniques from one-dimensional to multidimensional to allow the regression coefficients to vary in a smooth way with another variable, for instance time. Note that varying coefficient models are particularly useful in longitudinal analysis [5].

In this paper, we study the first stage of estimate the bivariate quantile function or (reference quantile contours) which represented by use the nonparametric method to estimate the quantile functions of two stratified quantile models (marginal and conditional models) Each of them separately, and comparison between the estimators of the two nonparametric methods, where the estimates can be obtained by solve nonparametric quantile regression problem. The nonparametric methods used here are B-splines which represent one type of smoothing splines and kernel regression (Nadaraya-Watson).

2. Theoretical and dynamic sides to estimate quantile function

Let $(\mathbf{Y}_{i,j}, \mathbf{X}_{i,j}, t_{i,j}); i = 1, \dots, n; j = 1, \dots, m_i$ a random sample consisting of n subjects. $\mathbf{Y}_{i,j} = (Y_{i,j,1}, Y_{i,j,2})^{\mathrm{I}}$ is the j^{th} pair of measurements of the i^{th} subject at time $t_{i,j}$ which have finite support $[0, \boldsymbol{T}]$, since, T some constant but not necessary evenly spaced for each i with $\mathbf{X}_{i,j} = (X_{i,j,1}, X_{i,j,2}, \dots, X_{i,j,d})^{\mathrm{I}}$. Where: $\mathbf{X}$ as d −dimensional associated covariates and $\mathbf{Y}$ as p −dimensional measurement vector [1].
Quantile regression estimates conditional quantiles without the requirement of normality, then the estimating of quantiles of some p −dimensional response $\mathbf{Y}$ Conditional on the values $\mathrm{x} \in \mathbb{R}^d$ of some covariates $\mathbf{X}$.
When $(p = 1)$ −dimensional (single-output case), where Y is used instead of $\mathbf{Y}$: for a (conditional) probability distribution $P^Y = P^{\mathrm{Y}}_{\mathbf{X}=\mathrm{x}}$ on $\mathbb{R}$, and when $(p \geq 2)$ −dimensional (multiple-output case) this means $\boldsymbol{Y} \in \mathbb{R}^{p \geq 2}$ [2], [3].
Assume we have two models of the quantile functions $\mathcal{g}_1$ and $\mathcal{g}_2$, these models are the marginal model $Q_{Y_1}(q)$ of $Y_{t,1} = y_1$ and the conditional model $Q_{Y_2}(q)$ of $Y_{t,2}$ given $Y_{t,1}$, the equations of the two models respectively is as the following:

$$Q_q(Y_{i,j,1}) = \mathcal{g}_1(q; t_{i,j}, \mathbf{X}_{i,j}) \quad (3)$$

$$Q_q(Y_{i,j,2}) = \mathcal{g}_2(q; t_{i,j}, \mathbf{X}_{i,j}, Y_{i,j,1}) \quad (4)$$

Where $\mathcal{g}_1(q; t, \mathbf{X})$ and $\mathcal{g}_2(q; t, \mathbf{X}, y_1)$ are the qth quantile functions of $Y_{t,1}$ given t and covariate $\mathbf{X}$, and the qth conditional quantile function of $Y_{t,2}$ given the value of $Y_{t,1}$ respectively.
Then, to estimate $\mathcal{g}_1$ and $\mathcal{g}_2$ we use the nonparametric methods by minimizing the quantile regression objective functions:

$$\hat{\mathcal{g}}_{n1}(q) = \operatorname{argmin} \sum_{i,j} \rho_q \left(Y_{i,j,1} - \mathcal{g}_1(t_{i,j}, \mathbf{X}_{i,j}) \right) \quad (5)$$

$$\hat{\mathcal{g}}_{n2}(q) = \operatorname{argmin} \sum_{i,j} \rho_q \left(Y_{i,j,2} - \mathcal{g}_2(t_{i,j}, \mathbf{X}_{i,j}, Y_{i,j,1}) \right) \quad (6)$$

We assume in both models that the quantiles of Y_2 are linear functions of Y_1 with time-varying coefficient (intercept and slope) functions, because we consider that have a problem of estimating the quantile functions $\mathcal{g}_1$ and $\mathcal{g}_2$, therefore we used the time-varying coefficient models which can transform the linear regression or

non-linear parametric regression to linear regression, since it describe the effect of variables as a constant correlation coefficient [1], [5].

Then, we written the models in (5) and (6) as:

$$Q_q(Y_{i,j,1}) = \alpha_1(q; t_{i,j}) + \sum_{k=1}^{d} \gamma_{1k}(q; t_{i,j}) \mathbf{X}_{i,j,k} \quad (7)$$

$$Q_q(Y_{i,j,2}) = \alpha_2(q; t_{i,j}) + \beta_2(q; t_{i,j}) Y_{i,j,1} + \sum_{k=1}^{d} \gamma_{2k}(q; t_{i,j}) \mathbf{X}_{i,j,k} \quad (8)$$

Where: α_1 and α_2 are the intercepts of $Q_{Y_1}(q)$ and $Q_{Y_2}(q)$

β_2 is the slope of $Q_{Y_2|Y_1}(q)$

γ_{1k} and γ_{2k} are the coefficients associated with $\mathbf{X}_k$ in both of models (7) and (8)

We can propose the general model for time-varying coefficient models for $(p > 2)$ depending on varying coefficient time series models with time trend and shown in the equation below:

$$Y_i = \alpha_0(t_i) + \sum_{j=1}^{m_i} \beta_j(t_i) \mathbf{X}_{i,j} + u_i \quad (9)$$

Where: u_i is a stationary random error process. Then, we can write the proposed general model for time-varying coefficient models as the following:

$$Y_{i,j,p} = \sum_{L=1}^{p} \boldsymbol{\beta}(q; t_{i,j}) Y_{i,j,p-1} + \sum_{k=1}^{d} \gamma_{pk}(q; t_{i,j}) \mathbf{X}_{i,j,k} \quad (10)$$

Where $\boldsymbol{\beta} = (\beta_{p1}, \beta_{p2}, \dots, \beta_{pp})$

Since, $L = 1, \dots, p$. if $L = 1$ that means the marginal model and if $L \geq 2$ that means the conditional model. The solution of the conditional sequence became more difficult in high dimensions [1], [5], [6].

3. Estimate quantile function

In this part we will study two nonparametric method: B-splines and kernel regression (Nadaraya-Watson) to estimate two quantile functions $\mathcal{G}_1$ and $\mathcal{G}_2$.

3.1. B-Splines

Splines are a smoothing technique used in regression analysis, since the degree of smoothness of the true coefficient functions depend on determines well by the number of knots. So, increasing the number of knots leads to obtain a more flexible curve [7], [8], [9].

In general case, the usefulness of B-splines lies in the fact that any spline function of order on a given set of knots can be expressed as a linear combination of B-splines:

$$S_{n,t}(x) = \sum_{i} \alpha_i \beta_{i,n}$$

For a spline function of degree k, the objective function for least squares minimization is

$$\hat{g}_{(BS)}(q) = \operatorname{argmin} \sum_{all\,x} \left\{ W(x) \left[y(x) - \sum_{i} \beta_{i,k,t}(x) \right] \right\}^2 \quad (11)$$

Then, according to the above formula we written the models in (5) and (6) as:

$$\hat{g}_{n1(BS)}(q) = \operatorname{argmin} \sum_{i=1}^{n} \boldsymbol{W}_i \sum_{j=1}^{m_i} \left(Y_{i,j,1} - (\alpha_1(q; t_{i,j}) + \sum_{k=1}^{d} \gamma_{1k}(q; t_{i,j}) \mathbf{X}_{i,j,k}) \right)^2 \quad (12)$$

$$\hat{g}_{n2(BS)}(q) = \operatorname{argmin} \sum_{i=1}^{n} \boldsymbol{W}_i \sum_{j=1}^{m_i} \left(Y_{i,j,2} - (\alpha_2(q; t_{i,j}) + \beta_2(q; t_{i,j}) Y_{i,j,1} + \sum_{k=1}^{d} \gamma_{2k}(q; t_{i,j}) \mathbf{X}_{i,j,k}) \right)^2 \quad (13)$$

$$\boldsymbol{W}_i = \frac{1}{m_i}$$

3.2. Kernel regression (Nadaraya-Watson)

In general, Kernel regression (Nadaraya-Watson) is a nonparametric technique used to estimate the conditional density function $f(y|x)$ [10], [11], [12].

$$f(y|x) = \frac{f(x,y)}{f_X(x)}$$

Using kernel estimators to estimate $f(y|x)$, then

$$\hat{f}(y|x) = \frac{\hat{f}(x,y)}{\hat{f}_X(x)} = \frac{\frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i)\, K_h(y - Y_i)}{\frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i)} \quad (14)$$

Where K is kernel function and h is bandwidth. The estimator of conditional distribution is given by:

$$\hat{F}(y|x) = \int_{-\infty}^{y} \hat{f}(u|x)du$$
$$\hat{F}_{NW}(y|x) = \sum_{i=1}^{n} W_i\, \mathrm{I}(Y_i \leq y)$$

Where: W_i is the nonzero weight function and $\mathrm{I}(Y_i \leq y)$ indicator function.

After incorporating the concept of conditional quantile and kernel estimation by using the Nadaraya-Watson we will get the formula:

$$\hat{g}_{(NW)}(q) = \operatorname{argmin} \sum_{i,j} \rho_q(Y_i - a)\, K\left(\frac{x - X_i}{h}\right) \quad (15)$$

Then, according to the above formula we written the models in (5) and (6) as:

$$\hat{g}_{n1(NW)}(q) = \operatorname{argmin} \sum_{ij} \rho_q \left(Y_{i,j,1} - (\alpha_1(q;t_{i,j}) + \sum_{k=1}^{d} \gamma_{1k}(q;t_{i,j})\mathbf{X}_{i,j,k}) \right) K\left(\frac{t-T}{h}\right) \quad (16)$$

$$\hat{g}_{n2(NW)}(q) = \operatorname{argmin} \sum_{ij} \rho_q \left(Y_{i,j,2} - (\alpha_1(q;t_{i,j}) + \beta_2(q;t_{i,j})Y_{i,j,1} + \sum_{k=1}^{d} \gamma_{2k}(q;t_{i,j})\mathbf{X}_{i,j,k}) \right) K\left(\frac{t-T}{h}\right) \quad (17)$$

4. Simulation study

In this part, we will explain the performance for the methodology that was followed above and shown it according to the following algorithm:

Algorithm

Step 1: Enter a value of **n** to represent the sample size.
Step 2: Simulate Y_1 according to the following formula

$Y_1 = 40t/(1+2t) + e_1$ Where $e_1 \sim N(0,1)$
Simulate Y_2 according to the following formula
$Y_2 = ln(1+t) + (1+.2t)Y_1 + e_2$ Where $e_2 \sim N(0,1)$
$t \sim uniform(0,3)$

Step 3: Find the estimators of B-Spline for Y_1 and $Y_2|Y_1$
Step 4: Find the estimators of kernel regression (Nadaraya-Watson) for Y_1 and $Y_2|Y_1$
Step 5: Find empirical distribution (ecdf) for step 2
Step 6: Find empirical distribution (ecdf) for step 3
Step 7: Find empirical distribution (ecdf) for step 4
Step 8: Find MSE between step 5 and step 6
Step 9: Find MSE between step 5 and step 7
Step 10: Compare between the results of step 8 and step 9 and choose the least result
Step 11: Go back to step 1
Step 12: Stop when you reach to the last size of n

The results obtained from the Simulation study were arranged according to the following table and figures:

Table 1. The MSE in both nonparametric method: B-splines with knots (0.5, 1, 1.5, 2 and 2.5) and kernel regression (Nadaraya-Watson) and sample sizes (25, 50, 75 and 100)

n	B-Spline	Nadaraya-Watson
25	0.072435	0.32665
50	0.052925	0.238141
75	0.064696	0.225085
100	0.060932	0.265386

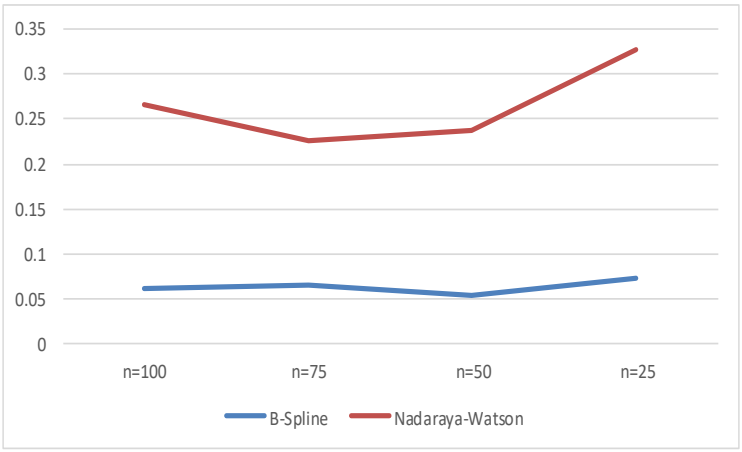

Figure 1. MSE chart (results from estimation Y_1 and $Y_2|Y_1$ by using nonparametric methods: B-splines with knots (0.5, 1, 1.5, 2, and 2.5) and kernel regression (Nadaraya-Watson) in sample sizes (25, 50, 75 and 100))

We draw the best method at a sample size $n = 50$ because it got less MSE for $Q_{Y_1}(q)$, $Q_{Y_2}(q)$ and $Q_{Y_2|Y_1}(q)$ as below:

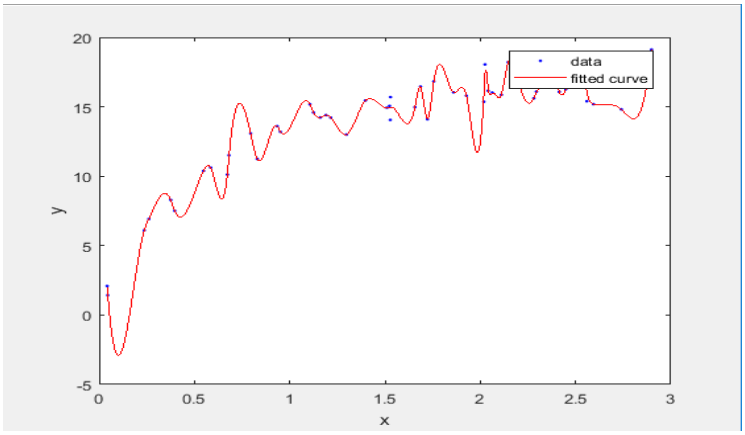

Figure 2. The estimation of quantile function of $Y_1\left(Q_{Y_1}(q)\right)$ using the best nonparametric methods: B-splines with knots (0.5, 1, 1.5, 2, and 2.5) at a sample size (50)

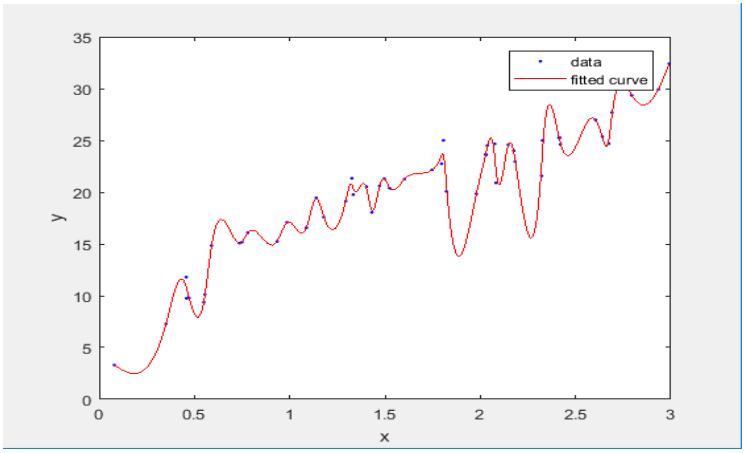

Figure 3. The estimation of quantile function of $Y_2\left(Q_{Y_2}(q)\right)$ using the best nonparametric methods: B-splines with knots (0.5, 1, 1.5, 2, and 2.5) at a sample size (50)

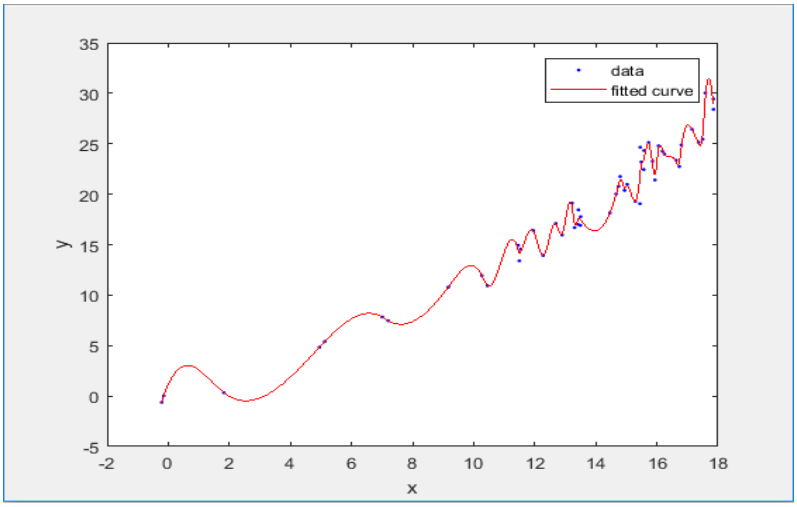

Figure 4. The estimation of quantile function of $Y_2|Y_1\left(Q_{Y_2|Y_1}(q)\right)$ using the best nonparametric methods: B-splines with knots (0.5, 1, 1.5, 2, and 2.5) at a sample size (50)

5. Conclusion

In this paper, we discussed two simulate models used to constructing the quantile functions of two stratified quantile models (marginal and conditional models) depending on time-varying coefficient models and then we estimated these models by using two nonparametric methods: smoothing spline (B-spline) and kernel regression (Nadaraya-Watson). The results of these methods were compared by using MSE.
Results have been obtained and displayed in Table 1. These results emphasize that estimates based on smoothing spline (B-spline) technique is better than the kernel regression (Nadaraya-Watson) because it gave the minimal MSE. Getting the best method was the first stage of estimate the quantile contours.

References

[1] Y., Wei, "an Approach to Multivariate Covariate-Dependent Quantile Contours with Application to Bivariate Conditional Growth Charts", Journal *of the American Statistical Association*, vol. 103, pp. 397-409, 2008.

[2] R., Koenker, and G., Bassett, "Regression Quantiles". *Econometrica*, vol. 46, no. 1, pp. 33–50, 1978.

[3] C., Davino, M., Furno and D., Vistocco, Quantile regression: theory and applications, Wiley series in probability and statistics, John Wiley & Sons, Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, United Kingdom, 2014.

[4] R., Koenker, V., Chernozhukov, X., He and L., Peng, Handbook of Quantile Regression, Taylor & Francis Group, LLC, 2018.

[5] J. Fan and W. Zhang, "Statistical methods with varying coefficient models", Statistics *and Its Interface*, vol.1, pp. 179–195, 2008.

[6] B., Chakraborty, "On multivariate quantile regression", Journal *of Statistical Planning and Inference*, vol. 110, pp. 109–132, 2003.

[7] H., Dette, S., Volgushev, "Non-Crossing Non-Parametric Estimates of quantile Curves", Journal *of the Royal Statistical Society*, vol. 70, part 3, pp. 609-627, 2008.

[8] K., Yu and M. C., Jones, “Local Linear Quantile Regression”, Journal *of the American Statistical Association*, vol. 93, no. 441, pp. 228–237, 1998.

[9] J., G., DE Gooijer, A., Gannoun and D., Zerom, "A Multivariate Quantile Predictor", *Communications in Statistics—Theory and Methods*, vol. 35, pp. 133–147, 2006.

[10] J., Z., Huang, C., O., Wu and L., Zhou, "Polynomial spline estimation and inference for varying coefficient models with longitudinal data", *Statistica Sinica*, vol. 14, pp. 763-788, 2004.

[11] C., O., Wu and C.-T., Chiang, "Kernel smoothing on varying coefficient models with longitudinal dependent variable ", statistica *sinica* , vol. 10, pp. 433-456, 2000.

[12] T., Zhang and W., B., Wu, "Inference of time-varying regression models", *The Annals of Statistics*, vol. 40, no. 3, pp.1376-1402, 2012.