On Nonparametric Quantile Regression

Jenny Tieu and Mei Ling Huang*

[a]Department of Mathematics & Statistics, Brock University,
St. Catharines, Ontario, Canada L2S 3A1

September 16, 2018

Abstract

Quantile regression estimates conditional quantiles and has wide applications in the real world. Estimating high conditional quantiles is an important problem. The regular linear quantile regression (QR) method often sets a linear or non-linear model, then estimates the coefficients to obtain the estimated conditional quantile. This approach may be restricted by the model setting. To overcome this problem, this paper proposes a direct nonparametric quantile regression (QN) method. Monte Carlo simulations show good efficiency for the proposed QN estimator relative to the regular QR estimator. The paper also investigates a real-world example using the proposed QN method. Comparisons of the proposed QN method and existing QR methods are given.

Keywords: *Conditional quantile, extreme value distribution, Gumbel's second kind of bivariate exponential distribution, nonparametric regression, loss function.*

AMS 2010 Subject Classifications: primary: 62G32; secondary: 62J05

1. Introduction

Extreme value events occur in many fields such as financial markets, weather, industrial engineering, actuarial science, survival analysis, queueing networks, and other stochastic models. A random variable y in the extreme events is usually heavy-tailed distributed. It is important to estimate high conditional quantiles of y given a variable vector $\mathbf{x} = (1, x_1, x_2, ..., x_d)^T \in R^p$ and $p = d+1$. The linear quantile regression model entails the use of an L_1-loss function and the optimal solution of linear programming for estimating regression coefficients. Quantile regression obtains more comprehensive results than mean regression methods.

The mean linear regression is the estimation of the conditional expectation $E(y|\mathbf{x})$. The mean linear regression model assumes

$$\mu_{y|\mathbf{x}} = E\left(y|x_1, x_2, ..., x_d\right) = \mathbf{x}^T\boldsymbol{\beta} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_d.$$

*This research is supported by *the Natural Science and Engineering Research Council of Canada (NSERC) grant MLH, RGPIN-2014-04621.*

We estimate $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_d)^T \in R^p$ from a random sample $\{(y_i, \mathbf{x}_i), i = 1, ..., n\}$, where $\mathbf{x}_i = (1, x_{i1}, x_{i2}, ..., x_{id})^T$, is the p-dimensional design vector and y_i is the univariate response variable from a continuous distribution with cumulative distribution function (c.d.f.) $F(y)$. The least squares (LS) estimator $\widehat{\boldsymbol{\beta}}_{LS}$ is a solution to the following equation

$$\widehat{\boldsymbol{\beta}}_{LS} = \arg \min_{\boldsymbol{\beta} \in R^p} \sum_{i=1}^{n} (y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2, \tag{1}$$

that is, $\widehat{\boldsymbol{\beta}}_{LS}$ is obtained by minimizing the L_2-distance.

Example: Oral Glucose Tolerance Test Level (2013-2014)

As of 2015, diabetes is the seventh leading cause of disease and death in the United States. It is reported that 30.3 million Americans have diabetes and 84.1 million Americans have pre-diabetes. Often, untreated prediabetes leads to type 2 diabetes within five years (Centers for Disease Control and Prevention, 2017).

The National Health and Nutrition Examination Survey (NHANES) is a program by the Centers for Disease Control and Prevention (CDC). NHANES aims to assess the health and nutritional status of adults and children in the United States. We will examine the 2013-2014 NHANES data for two-hour oral glucose tolerance tests administered to $N = 1635$ adults between the ages of 18 to 65 (Centers for Disease Control and Prevention, 2014). People with high glucose level test results have impaired glucose tolerance, which indicate prediabetic or diabetic conditions. Since a glucose level less than 140 mg/dL is considered a normal glucose tolerance, then a threshold of 120 mg/dL is applied to distinguish subjects at an increased risk of diabetes. In this paper, we are interested in this group of subjects. After omitting subjects with less than or equal to 120 mg/dL glucose level, the data is reduced to $n = 509$ subjects.

In Figure 1, a column graph presents the glucose tolerance levels for the $N = 1635$ adults, and a 120 mg/dL threshold is indicated. The x-axis represents the subject order. The y-axis represents the glucose level (mg/dL) after a two-hour oral glucose tolerance test.

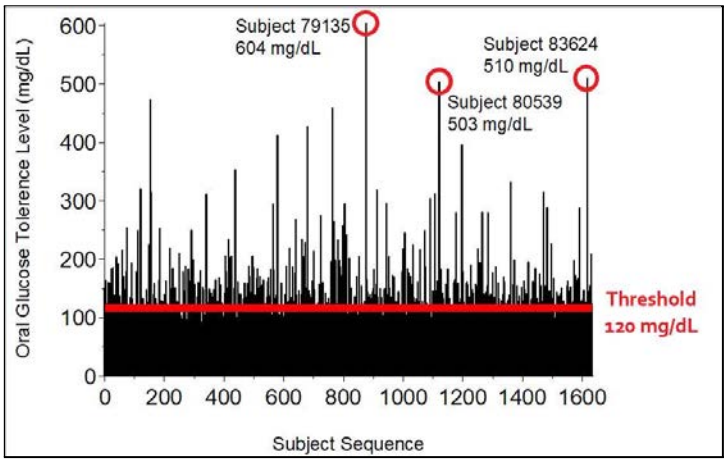


Figure 1. Oral glucose tolerance levels (mg/dL) of NHANES adults for 2013-2014 ($N = 1635$).

We consider glucose level as the response variable, and age and body mass index (BMI) as factors. We can employ a mean regression model to estimate the conditional mean on the subject's glucose tolerance level (mg/dL) y given their age x_1 and BMI x_2,

$$\mu_{y|(x_1,x_2)} = E(y|x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2.$$

Using the method of least squares estimator in *(1)*, we have

$$\widehat{\mu}_{LS} = \widehat{\mu}_{y|(x_1,x_2)} = 117.7407 + 0.5234x_1 + 0.8292x_2. \quad (2)$$

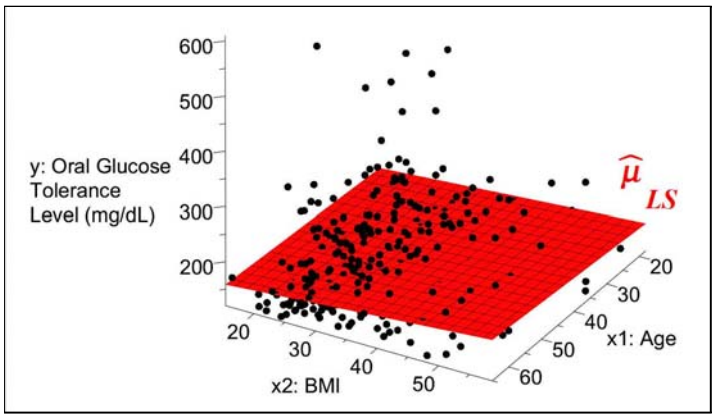


Figure 2. Oral glucose tolerance level y (mg/dL) vs. age x_1 and BMI x_2 for subjects with glucose levels higher than 120 mg/dL with the LS mean regression plane $\widehat{\mu}_{LS}-$ red in *(2)* ($n = 509$).

The least squares mean plane $\widehat{\mu}_{y|\mathbf{x}}$ only estimates the average glucose level given the subject's age and BMI. Our goal is to estimate the high conditional quantiles for extreme glucose levels which may pose as an indicator for diabetes. Figure 2 shows that the extreme high glucose levels are not captured well by the least squares mean plane.

In this paper, we use quantile regression methods to estimate extreme glucose levels. Quantile regression estimates high conditional quantiles of a random variable y with c.d.f. $F(y)$ given a variable vector, $\mathbf{x} = (x_1,x_2,...,x_d)$, and $\mathbf{x}_p = (1, x_1, x_2, ..., x_d)^T \in R^p$ where $p = d + 1$. The τth conditional linear quantile is defined by

$$Q_y(\tau|\mathbf{x}) = Q_y(\tau|x_1, x_2, ..., x_d) = F^{-1}(\tau|\mathbf{x}). \quad (3)$$

The traditional linear quantile regression is concerned with the estimation of the τth conditional linear quantile regression model of y for given $\mathbf{x}$ which is defined as

$$Q_y(\tau|\mathbf{x}) = \mathbf{x}_p^T \boldsymbol{\beta}(\tau) = \beta_0(\tau) + \beta_1(\tau)x_1 + \cdots + \beta_d(\tau)x_d, \quad 0 < \tau < 1. \quad (4)$$

We estimate the coefficient $\boldsymbol{\beta}(\tau) = (\beta_0(\tau), \beta_1(\tau), \beta_2(\tau), ..., \beta_d(\tau))^T \in R^p$ from a random sample $\{(Y_i, \mathbf{X}_i),\ i = 1, ..., n\}$, where $\mathbf{X}_i = (1, x_{i1}, x_{i2}, ..., x_{id})^T$ is the p-dimensional design vector and y_i is the univariate response variable from a continuous distribution with a c.d.f. $F(y)$.

Koenker and Bassett (1978) proposed an L_1-weighted loss function to obtain estimator $\widehat{\boldsymbol{\beta}}(\tau)$ by solving

$$\widehat{\boldsymbol{\beta}}(\tau) = \arg \min_{\boldsymbol{\beta}(\tau) \in R^p} \sum_{i=1}^{n} \rho_\tau(Y_i - \mathbf{X}_{pi}^T \boldsymbol{\beta}(\tau)), \quad 0 < \tau < 1, \tag{5}$$

where ρ_τ is a loss function, namely

$$\rho_\tau(u) = u(\tau - I(u < 0)) = \begin{cases} u(\tau - 1), & u < 0; \\ u\tau, & u \geq 0. \end{cases}$$

The linear quantile regression problem can be formulated as a linear program

$$\min_{(\boldsymbol{\beta}(\tau), \mathbf{u}, \mathbf{v}) \in R^p \times R_+^{2n}} \{\tau \mathbf{1}_n^T \mathbf{u} + (1 - \tau) \mathbf{1}_n^T \mathbf{v} | \mathbf{X}\boldsymbol{\beta}(\tau) + \mathbf{u} - \mathbf{v} = \mathbf{y}\},$$

where $\mathbf{1}_n^T$ is an n-vector of 1s, $\mathbf{X}$ denotes the $n \times p$ design matrix, and $\mathbf{u}, \mathbf{v}$ are n-vectors with elements of u_i, v_i respectively (Koenker, 2005).

In recent years, many studies search for efficiency improvements of estimator *(5)* (Hall, et al. 1999; Wang and Li, 2013; Huang et al. 2015; Huang and Nguyen, 2017). The regular linear quantile regression model in *(4)* needs the estimator in *(5)* for the high conditional quantile curves. The estimated very high or very low conditional quantile curves may be restricted under the model setting.

In order to overcome the limitation of the model setting in *(4)*, in this paper we propose a direct nonparametric quantile regression (QN) method in Section 2 which uses the ideas of nonparametric kernel density estimation and nonparametric kernel regression. The proposed method is not only different from most other existing nonparametric quantile regression methods, but it also overcomes the crossing problem of estimating quantile curves. We like to see if the new QN method has an improvement relative to the existing regular linear quantile regression (QR) in *(4)* and *(5)*. We will do two studies in this paper:

1. Monte Carlo simulations will be performed to show the efficiency of the new QN estimator relative to the regular QR estimator.

2. We will apply the newly proposed QN method to the oral glucose tolerance example and compare its result to the result of the regular QR method.

In Section 2, we propose a new direct nonparametric quantile regression estimator by using a nonparametric mean regression method. In Section 3, the results of Monte Carlo simulations generated from Gumbel's second kind of bivariate exponential distribution (Gumbel, 1960) show that the proposed QN method produces high efficiencies relative to existing regular QR method in *(4)* and *(5)*. Finally, in Section 4, we show that the new proposed QN model fits the glucose tolerance example data better than the regular QR model in *(4)*.

2. Proposed Algorithm of Direct Nonparametric Quantile Regression (QN)

We ignore the idea of the linear QR model *(4)* to obtain a direct estimator for the true conditional quantile in *(3)*:

$$\widehat{Q}_y(\tau|\mathbf{x}) = \widehat{Q}_y(\tau|x_1, x_2, ..., x_d) = \widehat{F}^{-1}(\tau|\mathbf{x}),$$

by using local conditional quantile estimator $\xi_i(\tau|\mathbf{x}) = Q_y(\tau|\mathbf{x}_i)$ based the ith point of a given random sample, $\{(Y_i, \mathbf{X}_i), i = 1, ..., n\}$, for $\mathbf{X}_i = (x_{i1}, x_{i2}, ..., x_{id})^T$.

We construct the following four-step algorithm of a direct nonparametric quantile regression estimator (QN):

Step 1: Estimate the conditional c.d.f. $F(y|\mathbf{x})$ of y for given $\mathbf{x} = (x_1,x_2, ...,x_d)$ using kernel estimation method (Hall et al. 1999; Silverman, 1986; Scott, 2015)

$$\widehat{F}(y|\mathbf{x}) = \frac{\frac{1}{n}\sum_{i=1}^{n} I(Y_i \leq y) K\left\{\frac{\mathbf{x}-\mathbf{X}_i}{h}\right\}}{\widehat{g}(\mathbf{x})}, \tag{6}$$

where $I(Y_i \leq y)$ is an indicator function, and $\widehat{g}(\mathbf{x})$ is an estimator of marginal density of $\mathbf{x}$.

Note that a d-dimensional multivariate kernel density estimator from a random sample $\mathbf{X}_i = (X_{1i},X_{2i},,...,X_{di})$, $i = 1, 2, ..., n$, from a population $\mathbf{x} = (x_1,x_2, ...,x_d)$ with density $g(\mathbf{x})$, is given by:

$$\widehat{g}(\mathbf{x}) = \frac{1}{nh^d}\sum_{i=1}^{n} K\left\{\frac{\mathbf{x}-\mathbf{X}_i}{h}\right\},$$

where $h > 0$ is the bandwidth and the kernel function $K(\mathbf{x})$ is a function defined for d-dimensional $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_d)$ which satisfies $\int_{R^d} K(\mathbf{x})d\mathbf{x} = 1$.

An estimator for the optimal bandwidth $h > 0$ will be given by:

$$\widehat{h}_{opt} = \left(\frac{4}{d+2}\right)^{1/(d+2)} n^{-1/(d+4)},$$

Step 2: Estimate the local conditional quantile function $\xi(\tau|\mathbf{x})$ of y given $\mathbf{x}$ by inverting the estimated conditional c.d.f. $\widehat{F}(y|\mathbf{x})$ in *(6)* from the Step 1:

$$\widehat{\xi}(\tau|\mathbf{x}) = \widehat{Q_y}(\tau|\mathbf{x}) = \inf\{y : \widehat{F}(y|\mathbf{x}) \geq \tau\} = \widehat{F}^{-1}(\tau|\mathbf{x}).$$

Additionally, to avoid the computational difficulties of $\widehat{\xi}(\tau|\mathbf{x})$, we estimate the local conditional quantile function $\xi_i(\tau|\mathbf{x}_i)$ of y given $\mathbf{x}_i$ by inverting an estimated conditional c.d.f. $\widehat{F}(y|\mathbf{x}_i)$ at the ith data point:

$$\widehat{\xi}_i(\tau|\mathbf{x}_i) = \widehat{Q_y}(\tau|\mathbf{x}_i) = \inf\{y : \widehat{F}(y|\mathbf{x}_i) \geq \tau\} = \widehat{F}^{-1}(\tau|\mathbf{x}_i), \quad i = 1, 2, ..., n. \tag{7}$$

Step 3: Propose a direct nonparametric quantile regression estimator Q_N for the τth conditional quantile curve of $\mathbf{x}$ by using the Nadaraya-Watson (NW) nonparametric regression estimator on $\left(\mathbf{x}_i, \widehat{\xi}_i(\tau|\mathbf{x}_i)\right)$, $i = 1, 2, ..., n$:

$$Q_N(\tau|\mathbf{x}) = \widehat{\xi}(\tau|\mathbf{x}) = \frac{\sum_{i=1}^{n} K_{\mathbf{h}}\{\mathbf{x}-\mathbf{X}_i\}\widehat{\xi}_i(\tau|\mathbf{x}_i)}{\sum_{j=1}^{n} K_{\mathbf{h}}\{\mathbf{x}-\mathbf{X}_j\}} = \sum_{i=1}^{n} W_{h_x}(\mathbf{x}, \mathbf{X}_i)\widehat{\xi}_i(\tau|\mathbf{x}_i), \quad 0 < \tau < 1, \tag{8}$$

where $W_{h_x}(\mathbf{x}, \mathbf{X}_i)$ is called an equivalent kernel,

$$W_{h_x}(\mathbf{x}, \mathbf{X}_i) = \frac{K_{\mathbf{h}}\{\mathbf{x} - \mathbf{X}_i\}}{\sum_{j=1}^{n} K_{\mathbf{h}}\{\mathbf{x} - \mathbf{X}_j\}}, \quad i = 1, 2, ..., n,$$

where

$$K_{\mathbf{h}}\{\mathbf{x} - \mathbf{X}_i\} = \frac{1}{nh_1...h_d} \prod_{j=1}^{d} K\left(\frac{x - x_{ij}}{h_j}\right), \quad i = 1, ..., n,$$

where K is the kernel function, and $h_j > 0$ is the bandwidth for the jth dimension.

Step 4: Check all procedures, and make any necessary adjustments.

3. Simulations

To investigate the efficiency of the proposed direct nonparametric quantile regression estimator QN in *(8),* Monte Carlo simulations are performed in this Section. We generate m random samples with size n each from the second kind of Gumbel's bivariate exponential distribution (Gumbel, 1960) with a non-linear conditional quantile function of y given x in *(10).* It has c.d.f.:

$$F(x, y) = (1 - e^{-x})(1 - e^{-y})(1 + \alpha e^{-(x+y)}),\ x \geq 0,\ y \geq 0,\ \alpha > 0, \tag{9}$$

The true τth conditional quantile function of y given x of *(9)* is

$$\xi(\tau|x) = Q_y(\tau|x) = \ln\left(\frac{2\alpha(2e^{-x} - 1)}{\alpha(2e^{-x} - 1) - 1 + \sqrt{(\alpha(2e^{-x} - 1) + 1)^2 - 4\alpha\tau(2e^{-x} - 1)}}\right), \tag{10}$$
$$x \geq 0,\ \alpha > 0, \quad 0 < \tau < 1.$$

We use two quantile regression methods to estimate the true conditional quantile in *(10)*:

1. The regular quantile regression $Q_R(\tau|x)$ estimation based on *(4) and (5):*

$$Q_R(\tau|x) = \widehat{\beta}_0(\tau) + \widehat{\beta}_1(\tau)x, \quad 0 < \tau < 1. \tag{11}$$

2. The direct nonparametric quantile regression $Q_N(\tau|x)$ estimation based on *(8)*

$$Q_N(\tau|x) = \sum_{i=1}^{n} W_{h_{\mathbf{x}}}(\mathbf{x}, \mathbf{X}_i)\widehat{\xi}_i(\tau|x_i), \quad 0 < \tau < 1,\ i = 1, 2, ..., n, \tag{12}$$

where $\widehat{\xi}_i(\tau|x_i)$ is obtained by *(7).*

For each method, we generate size $n = 300$, $m = 100$ samples. $Q_{R,i}(\tau|x)$ and $Q_{N,i}(\tau|x)$, $i = 1, 2, ..., m$, are estimated in the ith sample. Let $\alpha = 1$ in *(10).* Then, the true τth conditional quantile is

$$\xi(\tau|x) = Q_y(\tau|x) = \ln\left(\frac{2e^{-x} - 1}{e^{-x} - 1 + \sqrt{e^{-2x} - \tau(2e^{-x} - 1)}}\right), \quad x \geq 0,\ 0 < \tau < 1. \tag{13}$$

The simulation mean squared errors (SMSEs) of the estimators *(11)* and *(12)* are respectively:

$$SMSE(Q_R(\tau|x)) = \frac{1}{m}\sum_{i=1}^{m}\int_0^N (Q_{R,i}(\tau|x) - Q_y(\tau|x))^2 dx; \quad (14)$$

$$SMSE(Q_N(\tau|x)) = \frac{1}{m}\sum_{i=1}^{m}\int_0^N (Q_{N,i}(\tau|x) - Q_y(\tau|x))^2 dx, \quad (15)$$

where the true τth conditional quantile $Q_y(\tau|x)$ is defined in *(13)*. N is a finite x value such that the c.d.f. in *(9)* $F(N, N) \approx 1$. We take $N = 6$ and the simulation efficiencies (SEFFs) are given by

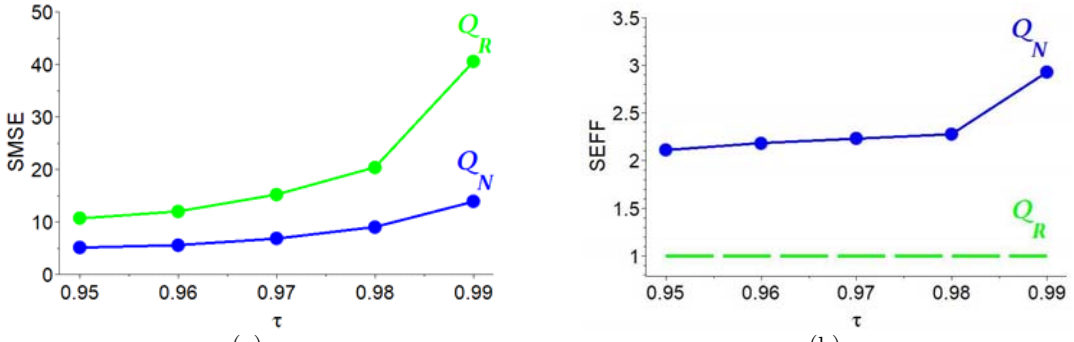
$$SEFF(Q_N(\tau|x)) = \frac{SMSE(Q_R(\tau|x))}{SMSE(Q_N(\tau|x))},$$

where $SMSE(Q_R(\tau|x))$ and $SMSE(Q_N(\tau|x))$ are defined in *(14)* and *(15)*, respectively.

Table 1. Simulation mean squared errors (SMSEs) and efficiencies (SEFFs) of estimating $Q_y(\tau|x), m = 100, n = 300, N = 6$.

τ	0.95	0.96	0.97	0.98	0.99
$SMSE(Q_R(\tau\|x))$	10.6577	11.9188	15.1286	20.2851	40.4940
$SMSE(Q_N(\tau\|x))$	5.0554	5.4794	6.7928	8.9271	13.8425
$SEFF(Q_N(\tau\|x))$	**2.1062**	**2.1752**	**2.2271**	**2.2723**	**2.9253**

Table 1 and Figure 3 show that all of the $SEFF(Q_N(\tau|x)) > 1$ when $\tau = 0.95,..., 0.99$. At these high quantiles, we can conclude that using the proposed direct nonparametric estimator $Q_N(\tau|x)$ in *(12)* is more efficient relative to the regular quantile regression estimator $Q_R(\tau|x)$ in *(11)*.



(a) (b)

Figure 3. (a) $SMSE(Q_N(\tau|x))-$ blue line, $SMSE(Q_R(\tau|x))-$ green line. (b) $SEFF(Q_N(\tau|x))-$ blue line, $SEFF(Q_R(\tau|x)) \equiv 1-$ green dash line.

4. Oral Glucose Tolerance Test Example

In this Section, we apply two quantile regression models to the oral glucose tolerance test example from Section 1:

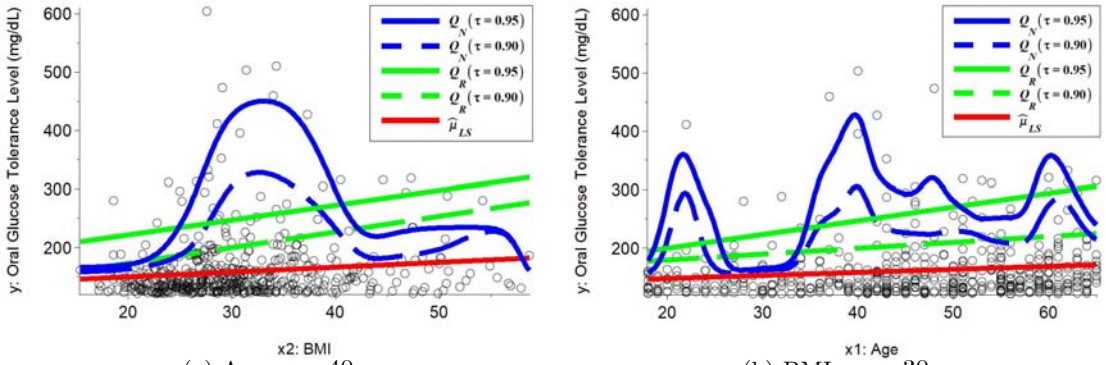
1. The regular quantile regression $Q_R(\tau|\mathbf{x})$ in model *(4)* using estimator $\widehat{\boldsymbol{\beta}}(\tau)$ in *(5)*;
2. The direct nonparametric quantile regression $Q_N(\tau|\mathbf{x})$ in *(8)*.

We also compare these two models with mean regression model.

At first, we use the following linear quantile regression model for this example:

$$Q_y(\tau|x) = \beta_0(\tau) + \beta_1(\tau)x_1 + \beta_2(\tau)x_2, \quad 0 < \tau < 1,$$

where y is the glucose level (mg/dL) given the subject's age x_1 and BMI x_2.



(a) Age $x_1 = 40$ (b) BMI $x_2 = 30$

Figure 4. For subjects, $n = 509$, with glucose level greater than 120 mg/dL, scatter plot with the LS mean regression line $\widehat{\mu}_{LS}$ - red, Q_R - green and Q_N - blue. (a) Oral glucose tolerance level (mg/dL) vs. BMI for age 40; (b) Oral glucose tolerance level (mg/dL) vs. age for BMI of 30.

We use the proposed four-step algorithm from Section 2 to obtain the new direct nonparametric quantile estimator $Q_N(\tau|\mathbf{x})$ in *(8)*. We compare the new estimator $Q_N(\tau|\mathbf{x})$ with the regular quantile estimator $Q_R(\tau|\mathbf{x})$ using *(5)*. Tables 2, 3 and Figure 4 show the difference of values of the two estimators. Figures 4(a), (b) show the scatter plot of the glucose level vs. age and BMI respectively with the fitted Q_R and Q_N quantile curves at $\tau = 0.90$ and 0.95. It is interesting to see that the Q_N curves appear to follow the data patterns closer than the Q_R curves.

Table 2 lists the estimated glucose level quantile values at a given BMI for 40 year old subjects for $\tau = 0.90$ and 0.95. Table 3 lists the estimated glucose level quantile values at a given age for BMI of 30 for $\tau = 0.90$ and 0.95. It demonstrates that when quantiles are at high τ, the Q_N gives greater variety of glucose level predictions than the Q_R. The relationship of glucose level and age or BMI is not necessarily linear.

In order to compare the fit of the regular Q_R estimator in *(5)* and the fit of the direct nonparametric Q_N estimator in *(8)*, we extend the idea of measuring goodness-of-fit by Koenker and Machado (1999) and suggest using a Relative $R(\tau)$ (Huang and Nguyen, 2017), $0 < \tau < 1$.

Figure 5 shows the values of the Relative $R(\tau)$ for given $\tau = 0.90, ..., 0.99$. We note that $R(\tau) > 0$ which means that Q_N is a better fit to the data than Q_R at the high quantiles.

Table 2. Predicted high quantiles of oral glucose tolerance level (mg/dL) given BMI for age 40 years old (Population with glucose levels greater than 120 mg/dL)

BMI	$\widehat{\mu}_{LS}$	$\tau = 0.90$		$\tau = 0.95$	
		Q_R	Q_N	Q_R	Q_N
20	149.27	172.39	162.73	220.99	170.35
25	153.41	185.86	186.72	233.73	237.05
30	157.56	199.32	304.01	246.46	424.63
35	161.70	212.79	314.95	259.19	440.83
40	165.85	226.26	232.51	271.92	297.71
45	170.00	239.72	181.71	281.65	220.96
50	174.14	253.19	196.48	297.39	233.11
55	178.29	266.65	227.16	310.12	232.42

Table 3. Predicted high quantiles of oral glucose tolerance level (mg/dL) given age for BMI of 30 (Population with glucose levels greater than 120 mg/dL)

Age	$\widehat{\mu}_{LS}$	$\tau = 0.90$		$\tau = 0.95$	
		Q_R	Q_N	Q_R	Q_N
25	149.70	184.21	166.13	211.35	236.77
30	152.32	189.25	162.16	223.05	164.46
35	154.94	194.29	211.66	234.76	262.81
40	157.56	199.32	304.01	246.46	424.63
45	160.17	204.36	225.70	258.16	292.25
50	162.79	209.40	228.03	269.87	280.54
55	165.41	214.44	209.70	281.57	251.38
60	168.02	219.48	273.45	293.27	356.56

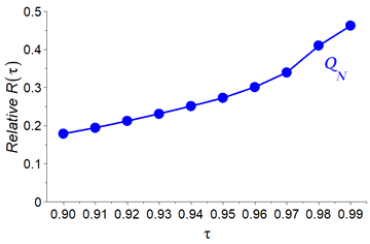


Figure 5. Relative $R(\tau)$ of Q_N to Q_R for glucose level example.

5. Conclusions and Suggestions

After the above studies, we can offer the following conclusions and suggestions:

1. This paper proposes a new direct nonparametric quantile regression method which is model free. It uses nonparametric regression techniques to estimate high conditional quantiles.

The paper provides a computational four-step algorithm which overcomes the limitations of the estimation in the linear quantile regression model.

2. The Monte Carlo simulation works on the second kind of Gumbel's bivariate exponential distribution which has a nonlinear conditional quantile function. The simulation results confirm that the proposed new direct nonparametric quantile regression estimator Q_N is more efficient relative to the regular linear quantile regression estimator Q_R.

3. The proposed new direct nonparametric quantile regression can be used to predict extreme values of glucose level for given age and BMI. The proposed Q_N estimator gives a variety of predictions which fits data very well. The prediction of relationships are not simply just linear. We expect that the Q_N predictions may be more reasonable than the Q_R predictions. The new estimator may benefit the identification and prevention of diabetes.

4. The proposed direct nonparametric quantile regression provides an alternative way for quantile regression. Further studies on the details of this method are suggested.

References

[1] Centers for Disease Control and Prevention. (2017). New CDC report: More than 100 million Americans have diabetes or prediabetes. Retrieved from https://www.cdc.gov/media/releases/2017/p0718-diabetes-report.html.

[2] Centers for Disease Control and Prevention. (2014). NHANES 2013-2014. Retrieved from https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx?BeginYear=2013.

[3] Gumbel, E. J. (1960). Bivariate exponential distributions, *Journal of the American Statistical Association,* 55, 698-707.

[4] Hall, P., Wolff, R.C.L. and Yao, Q. (1999). Methods for estimating a conditional distribution function. *Journal of the American Statistical Association, 94, 154-163.*

[5] Huang, M. L. and Nguyen, C. (2017). High quantile regression for extreme events, *Journal of Statistical Distributions and Applications*, 4(4), 1-20.

[6] Huang, M.L., Xu, X. and Tashnev, D. (2015). A weighted linear quantile regression. *Journal of Statistical Computation and Simulation,* 85(13), 2596-2618.

[7] Koenker, R. and Bassett, W.G. 1978. Regression Quantiles. *Econometrica*, 46. pp. 33-50.

[8] Koenker, R. (2005). *Quantile regression.* Cambridge University Press: New York.

[9] Koenker, R. and Machado, J.A.F. (1999). Goodness of fit and related inference processes for quantile regression. *Journal of the American Statistical Association,* 96(454), 1296-1311.

[10] Scott, D. W. (2015). *Multivariate Density Estimation, Theory, Practice and Visualization, second edition.* John Wiley & Sons: New York.

[11] Silverman, B.W. (1986). *Density estimation for statistics and data analysis.* Chapman & Hall: London, UK.

[12] Wang, H. J. and Li, D. (2013). "Estimation of extreme conditional quantile through power transformation", *Journal of the American Statistical Association,* 108(503), 1062-1074.