Huang and Nguyen *Journal of Statistical Distributions and Applications* (2018) 5:3
https://doi.org/10.1186/s40488-018-0084-9

Journal of Statistical
Distributions and Applications

RESEARCH Open Access

A nonparametric approach for quantile regression

Mei Ling Huang[1]* and Christine Nguyen[2]

*Correspondence:
mhuang@brocku.ca
[1]Department of Mathematics & Statistics, Brock University, St. Catharines, Ontario L2S 3A1, Canada
Full list of author information is available at the end of the article

Abstract

Quantile regression estimates conditional quantiles and has wide applications in the real world. Estimating high conditional quantiles is an important problem. The regular quantile regression (QR) method often designs a linear or non-linear model, then estimates the coefficients to obtain the estimated conditional quantiles. This approach may be restricted by the linear model setting. To overcome this problem, this paper proposes a direct nonparametric quantile regression method with five-step algorithm. Monte Carlo simulations show good efficiency for the proposed direct QR estimator relative to the regular QR estimator. The paper also investigates two real-world examples of applications by using the proposed method. Studies of the simulations and the examples illustrate that the proposed direct nonparametric quantile regression model fits the data set better than the regular quantile regression method.

Keywords: Conditional quantile, Goodness-of-fit, Gumbel's second kind of bivariate exponential distribution, Nonparametric kernel density estimator, Nonparametric regression, Weighted loss function

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

1 Introduction

It is important to study quantile regression to estimate high conditional quantiles in real-world events Koenker (2005). Some extreme events can cause damages to society: stock market crashes, pipeline failures, large flooding, wildfires, pollution, earth quakes and hurricanes. We wish to estimate high conditional quantiles of a random variable y with cumulative distribution function (c.d.f.) $F(y)$ given a variable vector, $\mathbf{x} = (x_1,x_2,\ldots,x_d)$, and $\mathbf{x}_p = (1, x_1, x_2, \ldots, x_d)^T \in R^p$ where $p = d + 1$. The τth conditional linear quantile is defined by

$$Q_y(\tau|\mathbf{x}) = Q_y(\tau|x_1,x_2,\ldots,x_d) = F^{-1}(\tau|\mathbf{x}),\ 0 < \tau < 1. \tag{1}$$

The traditional quantile regression is concerned with the estimation of the τth conditional quantile regression (QR) of y for given $\mathbf{x}$ which often sets a linear model as

$$Q_y(\tau|\mathbf{x}) = \mathbf{x}_p^T\beta(\tau) = \beta_0(\tau) + \beta_1(\tau)x_1 + \cdots + \beta_d(\tau)x_d, 0 < \tau < 1, \tag{2}$$

where $\beta(\tau) = (\beta_0(\tau), \beta_1(\tau), \beta_2(\tau), \ldots, \beta_d(\tau))^T$.

For linear model *(2)*, we estimate the coefficient $\beta(\tau) = (\beta_0(\tau), \beta_1(\tau), \beta_2(\tau), \ldots, \beta_d(\tau))^T \in R^p$ from a random sample $\{(y_i, \mathbf{x}_i),\ i = 1, \ldots, n\}$,

© The Author(s). 2018 **Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

where $\mathbf{x}_{pi} = (1, x_{i1}, x_{i2}, \ldots, x_{id})^T$ 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 $\widehat{\beta}(\tau)$ by solving

$$\widehat{\beta}(\tau) = \arg \min_{\beta(\tau) \in R^p} \sum_{i=1}^{n} \rho_\tau(y_i - \mathbf{x}_{pi}^T \beta(\tau)),\ 0 < \tau < 1, \tag{3}$$

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

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

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 $n \times 1$ vectors with elements of $u_i, v_i,\ i = 1, \ldots, n$, respectively (Koenker, 2005).

In recent years, studies are looking for efficiency improvements of estimator *(3)* (Yu et al. 2003; Wang and Li 2013; Huang et al. 2015; Huang and Nguyen 2017). The regular linear quantile regression *(2)* needs the estimator $\widehat{\beta}(\tau)$ in *(3)* for the estimated conditional quantile curves. But this estimated conditional quantile curve may be restricted under the model setting.

Many studies have used nonparametric method of quantile regression in recent years, for example, Chaudhuri (2003), Yu and Jones (1991), Hall et al. (1999) and Yu et al. (2003). Chapter 7 in Keoker (2005) proposed a local polynomial quantile regression (LPQR), and other methods. Also we can see detailed discussions on theory, methodologies and applications in Li and Racine (2007) and Cai (2013).

In order to overcome the limitation of the model setting in *(2)* in this paer we propose a direct nonparametric quantile regression method 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, it also overcome thecrossing problem of estimating quantile curves. We like to see if the new method has an improvement relative to the regular linear quantile regression and other nonparametric quantile regression methods, we will do two studies in this paper:

1. Monte Carlo simulations will be performed to confirm the better efficiency of the new direct QR estimator relative to the regular QR estimator and a nonparametric LPQR.
2. The new proposed method will be applied to two real-world examples of extreme events and compared with the linear model in Huang and Nguyen (2017).

In Section 2, we propose a direct nonparametric quantile regression estimator. A relative measure of comparing goodness-of-fit for quantile models is given in Section 3. In Section 4, the results of Monte Carlo simulations generated from Gumbel's second kind of bivariate exponential distribution Gumbel (1960) show that the proposed direct method produces high efficiencies relative to existing linear QR and LPQR methods. In Section 5, the regular linear quantile regression and the proposed direct quantile regression are applied to two real-life examples: the Buffalo snowfall and CO_2 emission examples in Huang and Nguyen (2017). The study of these examples illustrate that the proposed direct nonparametric quantile regression model fits the data better than the existing linear quantile regression method.

2 Proposed direct nonparametric quantile regression

In this paper, for generality, we ignore the idea of the linear model *(2)*. We obtain a direct estimator for true conditional quantile in *(1)*:

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

by using local conditional quantile estimator $\xi_i(\tau|\mathbf{x}_i) = Q_y(\tau|\mathbf{x}_i)$ based the ith point of given random sample, $\left\{(y_i, \mathbf{x}_i), i = 1, \ldots, n\right\}$, for $\mathbf{x}_i = (x_{1i}, x_{2i}, \ldots, x_{di})^T$.

We construct the following a five-step algorithm of a direct nonparametric quantile regression:

Step 1: Estimate the conditional density of y for given $\mathbf{x} = (x_1, x_2, \ldots, x_d)$ using a kernel density estimation method (Silverman 1986; Scott 2015):

$$\widehat{f}(y|\mathbf{x}) = \frac{\widehat{f}(y, \mathbf{x})}{\widehat{g}(\mathbf{x})}, \tag{4}$$

where $\widehat{f}(y, \mathbf{x})$ is an estimator of the joint density of y and $\mathbf{x}$, and $\widehat{g}(\mathbf{x})$ is an estimator of the marginal density of $\mathbf{x}$.

A d-dimensional kernel density estimator from a random sample $\mathbf{X}_i = (X_{1i}, X_{2i}, \ldots, X_{di})$, $i = 1, 2, \ldots, n$, from a population $\mathbf{x} = (x_1, x_2, \ldots, x_d)$ for joint 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 $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_d)$ which satisfies $\int\limits_{R^d} K(\mathbf{x})d\mathbf{x} = 1$.

Fukunaga (1972) suggested using

$$\widehat{g}(\mathbf{x}) = \frac{(\det \mathbf{S})^{-1/2}}{nh^d} \sum_{i=1}^{n} k\left\{\frac{(\mathbf{x} - \mathbf{X}_i)^T \mathbf{S}^{-1} (\mathbf{x} - \mathbf{X}_i)}{h^2}\right\},$$

where $\mathbf{S}$ is the sample covariance matrix of the data, K is the normal kernel, the function k is

$$k(u) = \left(\frac{1}{2\pi}\right)^{d/2} \exp\left(-\frac{u}{2}\right), \quad k(\mathbf{x}^T\mathbf{x}) = K(\mathbf{x}) = (2\pi)^{-d/2} \exp\left(-\frac{1}{2}\mathbf{x}^T\mathbf{x}\right).$$

A plug-in selector of the bandwidth $h > 0$ will be given by (Silverman 1986, p. 85) as

$$h_{opt} = \left\{\int t^2 K(t)dt\right\}^{-2/(d+2)} \left\{\int K(t)^2 dt\right\}^{1/(d+4)} \left\{\int \left(\nabla^2 g(\mathbf{x})\right)^2 d\mathbf{x}\right\}^{-1/(d+4)} n^{-1/(d+4)}, \tag{5}$$

If a multivariate normal kernel is used for smoothing the normal distribution data with unit variance,

$$h_{opt} = \left\{\frac{4}{d+2}\right\}^{1/(d+4)} n^{-1/(d+4)}.$$

Step 2: Estimate the conditional c.d.f. of y given $\mathbf{x}$:

$$\widehat{F}(y|\mathbf{x}) = \int_{-\infty}^{y} \widehat{f}(y|\mathbf{x})dy.$$

Step 3: Estimate the local conditional quantile function $\xi(\tau|\mathbf{x})$ of y given $\mathbf{x}$ by inverting an estimated conditional c.d.f. $\widehat{F}(y|\mathbf{x})$.

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

It is difficult to compute a global inverse function $\widehat{\xi}(\tau|\mathbf{x})$ of the kernel estimated conditional c.d.f. $\widehat{F}(y|\mathbf{x})$ which has many terms. To avoid the the computational global difficulties, we estimate the local conditional quantile point $\xi_i(\tau|\mathbf{x}_i)$ of y given $\mathbf{x}_i$ by inverting $\widehat{F}(y|\mathbf{x}_i)$ at the ith data point $(y_i, \mathbf{x}_i)$:

$$\widehat{\xi}_i(\tau|\mathbf{x}_i) = \widehat{Q_y}(\tau|\mathbf{x}_i) = \inf\{y : \widehat{F}(y|\mathbf{x}_i) \geq \tau\} = \widehat{F}^{-1}(\tau|\mathbf{x}_i), \quad i = 1, 2, \ldots, n. \tag{6}$$

Thus, we have n points $\left(\mathbf{x}_i, \widehat{\xi}_i(\tau|\mathbf{x}_i)\right)$, $i = 1, 2, \ldots, n$.

Step 4: We propose a direct nonparametric quantile regression estimator for the τth conditional quantile curve of $\mathbf{x}$ by using Nadaraya-Watson (NW) nonparametric regression estimator (Scott, 2015, p. 242) on $\left(\mathbf{x}_i, \widehat{\xi}_i(\tau|\mathbf{x}_i)\right)$, $i = 1, 2, \ldots, n$:

$$Q_D(\tau|\mathbf{x}) = \widehat{\xi}(\tau|\mathbf{x}) = \frac{\sum_{i=1}^{n} K_{\mathbf{h}}\{\mathbf{x} - \mathbf{X}_i\}\,\widehat{\xi}_i(\tau|\mathbf{x}_i)}{\sum_{j=1}^{n} K_{\mathbf{h}}\left\{\mathbf{x} - \mathbf{X}_j\right\}} = \sum_{i=1}^{n} W_{h_{\mathbf{x}}}(\mathbf{x}, \mathbf{X}_i)\widehat{\xi}_i(\tau|\mathbf{x}_i), \quad 0 < \tau < 1, \tag{7}$$

where $W_{h_{\mathbf{x}}}(\mathbf{x}, \mathbf{X}_i)$ is called an equivalent kernel, and $\mathbf{h} = (h_1, \ldots, h_d)$,

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

where

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

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

The new point of *(7)* is that it uses Step 3's *(6)* numerical results: n points $\left(\mathbf{x}_i, \widehat{\xi}_i(\tau|\mathbf{x}_i)\right)$, $i = 1, 2, \ldots, n$, to estimate a conditional mean curve of the τth quantile function based on these n points, then smoothes these n points out.

In this paper, for the kernel regression, we use K which is the standard normal kernel. Similar as formula *(5)*, we use the optimal bandwidth for the jth dimension (Silverman 1986, p.40),

$$h_{j,opt} = \left\{\int t^2 K(t)dt\right\}^{-2/5} \left\{\int K(t)^2 dt\right\}^{1/5} \left\{\int \left(\nabla^2 \widehat{g}_j(x_j)\right)^2 d\mathbf{x}_j\right\}^{-1/5} n^{-1/5}, \quad j = 1, \ldots, d, \tag{8}$$

where $\widehat{g}_j(x_j)$ is the estimated the jth dimensional marginal density of x_j in $\mathbf{x} = (x_1, x_2, \ldots, x_d)$, n is the sample size of the random sample in *(4)*.

Step 5: Check all procedures, and make any necessary adjustments.

3 Comparison of goodness-of-fit on quantile regression models

In order to compare the regular QR estimator in *(3)* and the direct nonparametric QR estimator in *(7)*, we extend the idea of measuring goodness-of-fit by Koenker and Machado (1999). We suggest using a Relative $R(\tau)$, $0 < \tau < 1$, which is defined as

$$\textit{Relative } R(\tau) = 1 - \frac{V_D(\tau)}{V_R(\tau)}, \quad -1 \leq R(\tau) \leq 1, \quad \text{where} \tag{9}$$

$$V_D(\tau) = \sum_{y_i \geq Q_D(\tau|\mathbf{x}_i)} \frac{\tau}{n} \left| y_i - Q_D(\tau|\mathbf{x}_i) \right| + \sum_{y_i < Q_D(\tau|\mathbf{x}_i)} \frac{(1-\tau)}{n} \left| y_i - Q_D(\tau|\mathbf{x}_i) \right|,$$

where $Q_D(\tau|\mathbf{x}_i)$ is obtained by *(7)*, and

$$V_R(\tau) = \sum_{y_i \geq \mathbf{x}_i^T \widehat{\beta}(\tau)} \frac{\tau}{n} \left| y_i - \mathbf{x}_i^T \widehat{\beta}(\tau) \right| + \sum_{y_i < \mathbf{x}_i^T \widehat{\beta}(\tau)} \frac{(1-\tau)}{n} \left| y_i - \mathbf{x}_i^T \widehat{\beta}(\tau) \right|,$$

where $\widehat{\beta}(\tau)$ is given by *(3)*.

4 Simulations

For investigating the proposed direct nonparametric quantile regression estimator in *(7)*, in this Section, Monte Carlo simulations are performed. We generate m random samples with size n each from the second kind of Gumbel's bivariate exponential distribution Gumbel (1960) which has a non-linear conditional quantile function of y given x in *(11)*. It has c.d.f. $F(x,y)$ and density function $f(x,y)$ in *(10)*:

$$F(x,y) = (1-e^{-x})(1-e^{-y})(1+\alpha e^{-(x+y)}),\ x \geq 0,\ y \geq 0,\ \alpha > 0, \tag{10}$$

$$f(x,y) = e^{-(x+y)}(1+\alpha(2e^{-x}-1)(2e^{-y}-1)),\ x \geq 0,\ y \geq 0,\ \alpha > 0.$$

The conditional density of y for given x is

$$f(y|x) = e^{-y}(1+\alpha(2e^{-x}-1)(2e^{-y}-1)),\ x \geq 0,\ y \geq 0,\ \alpha > 0.$$

The conditional c.d.f. of y for given x is

$$F(y|x) = e^{-y}(\alpha(2e^{-x}-1)(1-e^{-y})-1)+1,\ x \geq 0,\ y \geq 0,\ \alpha > 0.$$

The true τth conditional quantile function of y given x of *(10)* is

$$\xi(\tau|x) = Q_y(\tau|x) = \ln\left(\frac{2\alpha(2e^{-x}-1)}{\alpha(2e^{-x}-1)-1+\sqrt{(\alpha(2e^{-x}-1)+1)^2-4\alpha\tau(2e^{-x}-1)}}\right), \tag{11}$$
$$x \geq 0,\ \alpha > 0,\ 0 < \tau < 1.$$

Letting $\alpha = 1$, the c.d.f. in *(10)* is in Fig. 1.

We use three quantile regression methods:

1. The regular quantile regression $Q_R(\tau|x)$ estimation based on *(3)*:

$$Q_R(\tau|x) = \widehat{\beta}_0(\tau) + \widehat{\beta}_1(\tau)x. \quad 0 < \tau < 1 \tag{12}$$

2. The first-order linear polynomials Quantile Regression (LPQR) $Q_{LP}(\tau|x)$ (Chaudhuri 1991, Keoker 2005, Yu and Jones 1998), for z in a neighborhood of x,

$$Q_{LP}(\tau|x) = \widehat{a}_0(\tau,x) + \widehat{a}_1(\tau,x)(z-x). \quad 0 < \tau < 1, \tag{13}$$

where

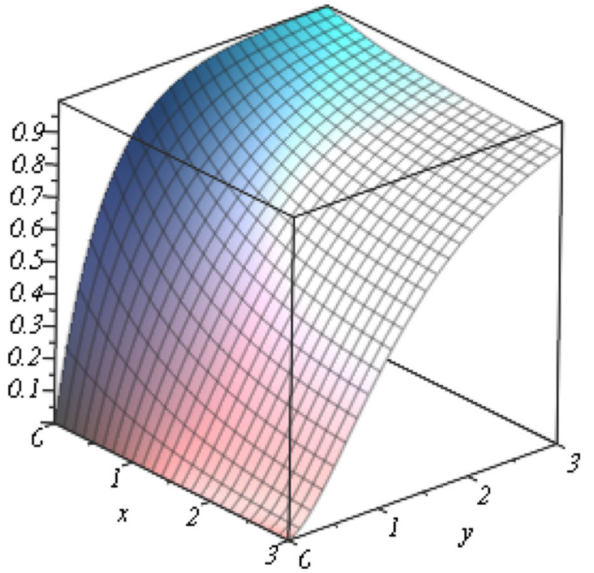

Fig. 1 The c.d.f. of Gumbel's Second kind of bivariate exponential distribution with $\alpha = 1$

$$\widehat{\mathbf{a}}(\tau, x) = \arg \min_{\beta(\tau)\in R^p} \sum_{i=1}^{n} \rho_\tau (y_i - a_0(\tau, x) - a_1(\tau, x)(x_i - x)) K\left(\frac{x - x_i}{h}\right), \quad 0 < \tau < 1,$$

here $\mathbf{a}(\tau, x) = (a_0(\tau, x), a_1(\tau, x))^T$, h and K are the bandwidth and kernel function. the LPQR can be computed by the *R* package 'quantreg' Koenker (2018).

3. The direct nonparametric quantile regression $Q_D(\tau|x)$ estimation based on *(7)*

$$Q_D(\tau|x) = \sum_{i=1}^{n} W_{h_{\mathbf{x}}}(\mathbf{x}, \mathbf{X}_i)\widehat{\xi}_i(\tau|x_i), \quad 0 < \tau < 1, \tag{14}$$

where $\widehat{\xi}_i(\tau|x_i)$ is obtained by *(6)*, $W_{h_{\mathbf{x}}}(\mathbf{x}, \mathbf{X}_i)$ is given by *(7)*.

For each method, we generate size $n = 100$, $m = 100$ samples. $Q_{R,i}(\tau|x)$, $Q_{LP,i}(\tau|x)$ and $Q_{D,i}(\tau|x)$, $i = 1, 2, \ldots, m$, are estimated in the ith sample. Let $\alpha = 1$ in *(11)*. Then the true τth conditional quantile is

$$\xi(\tau|x) = Q_y(\tau|x) = \ln\left(\frac{2e^{-x} - 1}{e^{-x} - 1 + \sqrt{e^{-2x} - \tau(2e^{-x} - 1)}}\right), \; x \ge 0, \; \alpha > 0, \; 0 < \tau < 1. \tag{15}$$

The simulation mean squared errors (SMSEs) of the estimators *(12)*, *(13)* and *(14)* are:

$$SMSE(Q_R(\tau|x)) = \frac{1}{m}\sum_{i=1}^{m}\int_0^N (Q_{R,i}(\tau|x) - Q_y(\tau|x))^2 dx; \tag{16}$$

$$SMSE(Q_{LP}(\tau|x)) = \frac{1}{m}\sum_{i=1}^{m}\int_0^N (Q_{LP,i}(\tau|x) - Q_y(\tau|x))^2 dx, \tag{17}$$

$$SMSE(Q_D(\tau|x)) = \frac{1}{m}\sum_{i=1}^{m}\int_0^N (Q_{D,i}(\tau|x) - Q_y(\tau|x))^2 dx, \tag{18}$$

where the true τth conditional quantile $Q_y(\tau|x)$ is defined in *(15)*. N is a finite x value such that the c.d.f. in *(10)* $F(N, N) \approx 1$. We take $N = 6$ and the simulation efficiencies (SEFFs) are given by

Table 1 Simulation Mean Square Errors (SMSEs) and Efficiencies (SEFFs) of Estimating $Q_y(\tau|x), m = 100, n = 100, N = 6$.

τ	0.95	0.96	0.97	0.98	0.99
$SMSE(Q_R(\tau\|x))$	22.091	26.632	28.982	42.725	73.340
$SMSE(Q_{LP}(\tau\|x))$	8.160	9.667	11.074	15.080	23.734
$SMSE(Q_D(\tau\|x))$	5.161	6.630	6.552	8.850	11.596
Efficiency					
$SEFF(Q_{LP}(\tau\|x))$	**2.7072**	**2.7449**	**2.6171**	**2.8332**	**3.0901**
$SEFF(Q_D(\tau\|x))$	**4.2804**	**4.0169**	**4.4234**	**4.8278**	**6.3246**

$$SEFF(Q_{LP}(\tau|x)) = \frac{SMSE(Q_R(\tau|x))}{SMSE(Q_{LP}(\tau|x))}, \quad SEFF(Q_D(\tau|x)) = \frac{SMSE(Q_R(\tau|x))}{SMSE(Q_D(\tau|x))},$$

where $SMSE(Q_R(\tau|x))$, $SMSE(Q_{LP}(\tau|x))$ and $SMSE(Q_D(\tau|x))$ are defined in *(16), (17)* and *(18),* respectively.

Table 1 shows that all of the $SEFF(Q_D(\tau|x))$ are larger than 1 when $\tau = 0.95, \ldots, 0.99$.

Figure 2 compares the $SMSE(Q_R(\tau|x)), SMSE(Q_{LP}(\tau|x))$ with the $SMSE(Q_D(\tau|x))$ for $\tau = 0.95, \ldots, 0.99$. It demonstrates that all $SMSE(Q_D(\tau|x))$ have smaller values than both $SMSE(Q_{LP}(\tau|x))$ and $SMSE(Q_R(\tau|x))$, thus, the simulation results show that the proposed estimator $Q_D(\tau|x)$ is more efficient relative to the regular linear estimator $Q_R(\tau|x)$ and nonparametric local polynomial estimator $Q_D(\tau|x)$.

Next, we compare $Q_D(\tau|x)$ and $Q_R(\tau|x)$ in Figs. 3 and 4.

Figure 3 shows the boxplots of $Q_R(\tau|x)$ and $Q_D(\tau|x)$ for $\tau = 0.95, 0.97$, and 0.99.(The true conditional quantiles are in blue line). The $Q_D(\tau|x)$ has much smaller variance than $Q_R(\tau|x)s$.

Figure 4 shows the average curves of the 100 estimated $\tau = 0.95$th quantile curves of $Q_R(\tau|x)$ (in blue dash line) and that of $Q_D(\tau|x)$ (in red solid). The average $Q_D(\tau|x)$ curve is much closer than $Q_R(\tau|x)$ to the true quantile curve (in green dash).

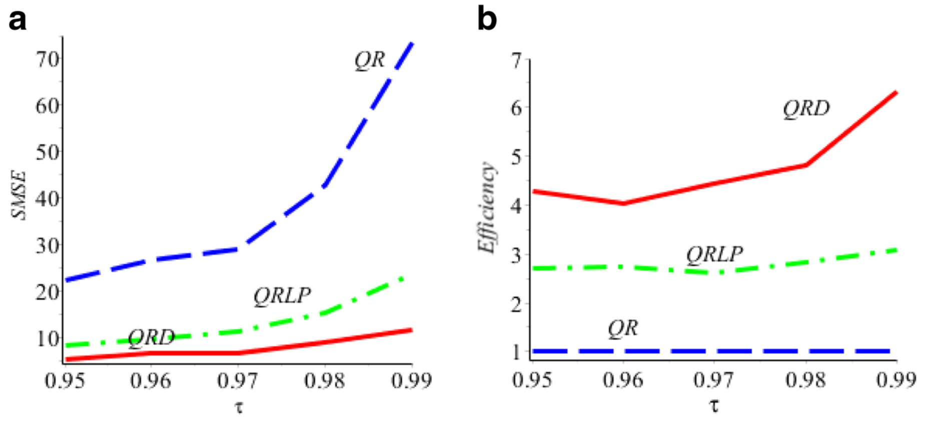

Fig. 2 **a** $SMSE(Q_D(\tau))$ is the red solid line, $SMSE(Q_{LP}(\tau))$ is the green dash-dot line, $SMSE(Q_R(\tau|x))$ is the blue dash line. **b** $SEFF(Q_D(\tau|x))$ is the red solid line, $SEFF(Q_{LP}(\tau|x))$ is the green dash-dot line,$SEFF(Q_R(\tau|x)) \equiv 1$ is blue dash line

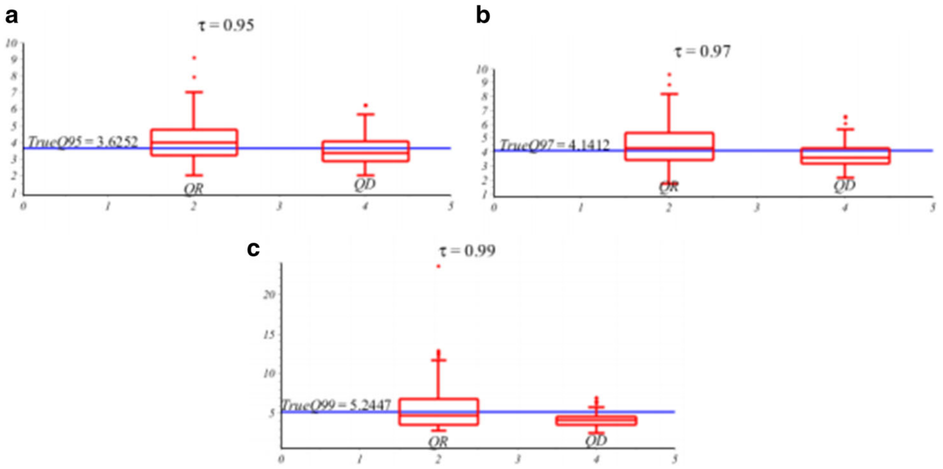

Fig. 3 Box plots for (**a**) $\tau = 0.95$th quantile curves; (**b**) $\tau = 0.97$th quantile curves; (**c**) $\tau = 0.99$th quantile curves. The true conditional quantile lines are in blue

From the overall results of the simulation, we can conclude that Table 1 and Figs. 2, 3, and 4 show that for $\tau = 0.95, \ldots, 0.99$, the proposed direct estimator $Q_D(\tau|x)$ in *(7)* is more efficient relative to the regular regression $Q_R(\tau|x)$ in *(2)* and a nonparametric LPQR in *(13)*.

5 Real examples of applications

In this section, we apply the following two regression models to the Buffalo snowfall and CO_2 emission examples in Huang and Nguyen (2017):

1. The regular quantile regression $Q_R(\tau|\mathbf{x})$ in model *(2)* using estimator $\widehat{\beta}(\tau)$ in *(3)*;
2. The direct nonparametric quantile regression $Q_D(\tau|\mathbf{x})$ in *(7)*.

5.1 Buffalo snowfall example

Huang and Nguyen (2017) used the following linear second order polynomial quantile regression model for this example (National Weather Service Forecast Office 2017):

$$Q_y(\tau|x) = \beta_0(\tau) + \beta_1(\tau)x + \beta_2(\tau)x^2,$$

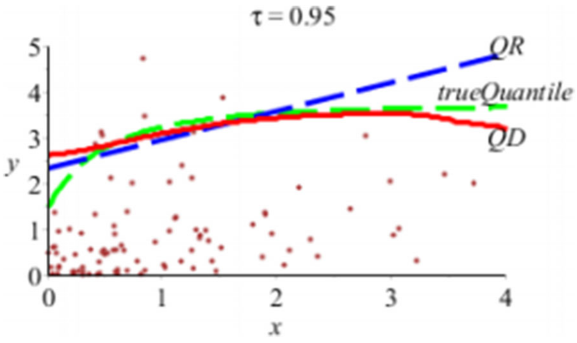

Fig. 4 In $n = 100$, $m = 100$, $\tau = 0.95$ simulations, the true Quantile-green dash; average regular QR-blue dash; average direct QD-red solid

Table 2 Buffalo Daily Snowfalls (cm) at High Quantiles Using Q_R and Q_D

	$\tau = 0.97$		$\tau = 0.99$	
Temperature (°C)	Q_R	Q_D	Q_R	Q_D
-15	37.38	25.49	105.46	60.64
-10	33.19	30.23	87.95	62.98
-5	30.98	33.33	72.08	56.54
0	30.73	29.89	57.86	54.56
5	32.47	33.27	45.29	52.39
10	36.17	37.34	34.36	43.04

where y represents the total snowfall (cm) and x represents the maximum temperature ($°C$).

In this paper we use the proposed five-step algorithm in Section 2 to obtain the new direct nonparametric quantile estimator $Q_D(\tau|\mathbf{x})$ in *(7).* We compare the new estimator $Q_D(\tau|\mathbf{x})$ with the regular quantile estimator $Q_R(\tau|\mathbf{x})$ in Huang and Nguyen (2017). Table 2 and Fig. 5 show the difference of values of two estimators. Figure 5a, b and c show the scatter plot of the daily snowfall vs. maximum temperature with the fitted Q_R, and Q_D

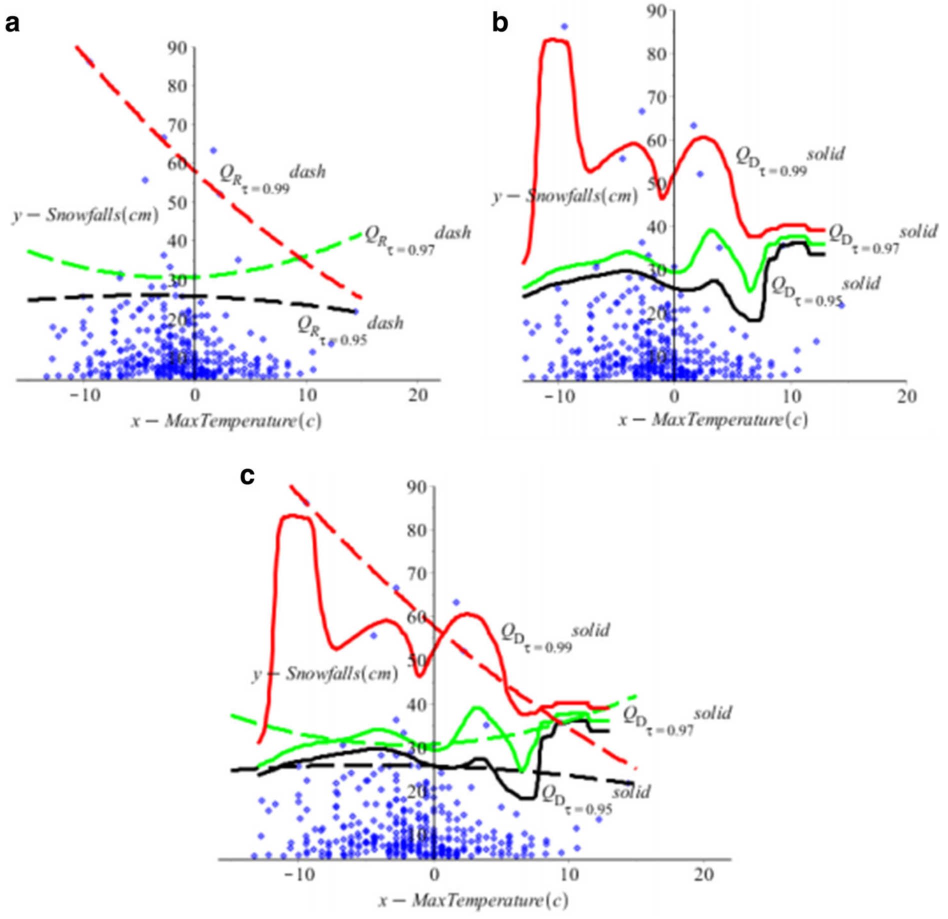

Fig. 5 For Buffalo snowfall example, data — blue, $n = 316$, (**a**) Regular Q_R— dash; (**b**) Direct Q_D— solid; (**c**) Both of the Regular Q_R and Direct Q_D in a plot at $\tau = 0.95$ in black, $\tau = 0.97$ in green and $\tau = 0.99$ in red

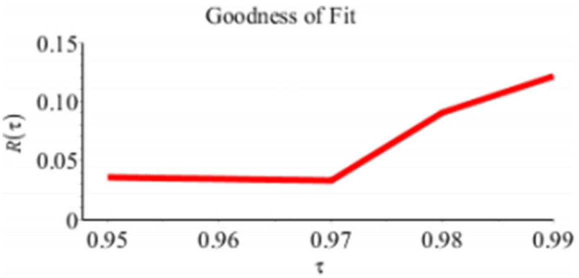

Fig. 6 Relative R_τ of Q_D relative to Q_R for the Buffalo snowfall example

quantile curves at $\tau = 0{,}95$, 0.97 and 0.99. It is interesting to see that the Q_D curves appear to follow the data patterns closer than the Q_R curves.

Table 2 lists the estimated Buffalo snowfall quantile values at a given maximum temperature for $\tau = 0.97$ and 0.99. It demonstrates that when quantiles are at high τ, the Q_D gives greater variety of snowfall predictions than the Q_R. The relationship of snowfall and max-temperature is not necessarily linear.

Figure 6 and Table 3 show the values of the Relative $R(\tau)$ in *(9)* for given $\tau = 0.95, \ldots, 0.99$. We note that $R(\tau) > 0$ which means that $V_D(\tau) < V_R(\tau)$ and Q_D is a better fit to the data than Q_R.

Figure 5c shows that the proposed direct nonparametric quantile regression Q_D predicts that for moderate temperatures, such as $5°C$ to $10°C$, it is likely to have smaller but varied snowfalls in Buffalo than the regular Q_D predicts. For temperature over $10°C$, the Q_D predicts a much higher value snow amount than the regular Q_R predicts. On another side, for very low temperatures, such as $-15°C$ to $0°C$, the Q_D and Q_R both predict more likely to have extreme heavy snowfalls that may cause damage. Thus prediction of heavy snowfalls is related to cold weather forecasts. But the prediction snowfalls related to temperature from the Q_D is not as a simple linear relationship as Q_R predicts. We also note that lots of snow occurred between -$5°C$ to $0°C$; the predictions form the Q_D are reflecting this fact and give varied predictions.

5.2 CO$_2$ emission example

Huang and Nguyen (2017) used the linear quantile regression model for this example:

$$Q_y(\tau|x_1, x_2) = \beta_0(\tau) + \beta_1(\tau)x_1 + \beta_2(\tau)x_2,$$

where y represents CO$_2$ emission (tonnes) per capita, x_1 represents ln of gross domestic product (GPD) (US \$), per capita and x_2 represents ln of electricity consumption (E.C.) (kilowatts) per capita (Carbon Dioxide Information Analysis Centre (2017)).

Similar as in the Buffalo Snowfall example in Subsection 5.1, we use the proposed five-step algorithm in Section 2 to obtain the new direct nonparametric quantile estimator

Table 3 Relative $R(\tau)$ Values for the Buffalo Snowfall Example

	$\tau = 0.95$	$\tau = 0.96$	$\tau = 0.97$	$\tau = 0.98$	$\tau = 0.99$
Relative $R(\tau)$	0.0359	0.0346	0.0324	0.0903	0.1206

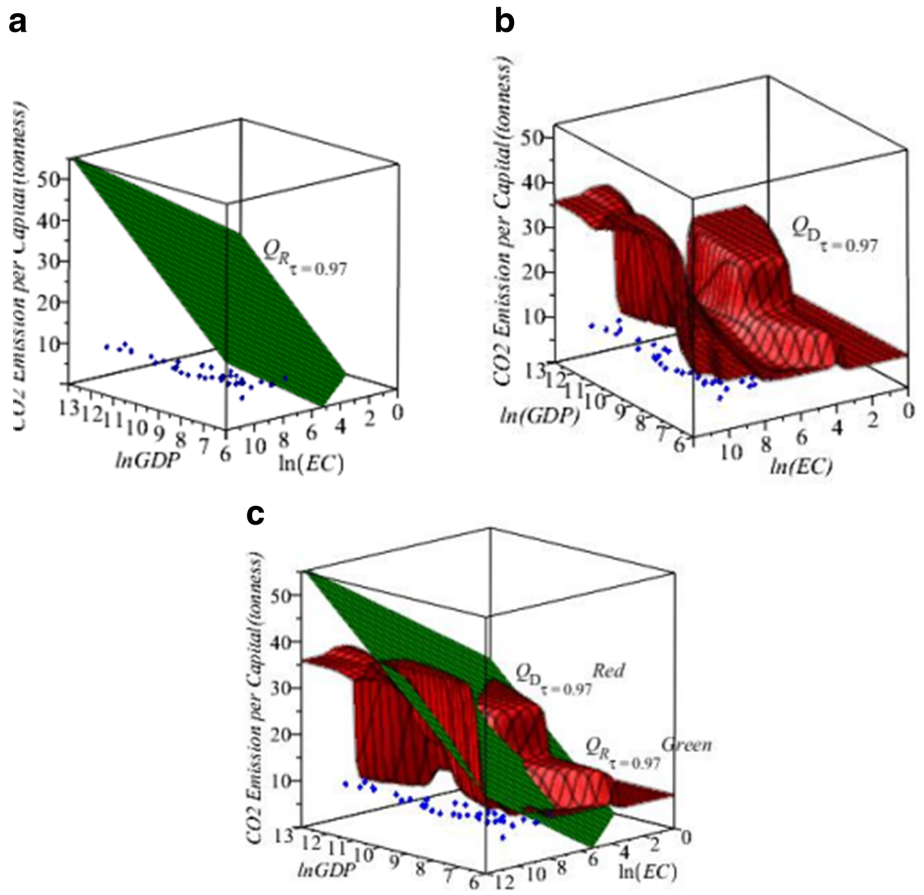

Fig. 7 3D Plots for CO_2 Emission, data — blue, $n = 123$, (**a**) Regular Q_R— green at $\tau = 0.97$; (**b**) Direct Q_D— red at $\tau = 0.97$; (**c**) Regular Q_R—green and Direct Q_D—red in a plot at $\tau = 0.97$

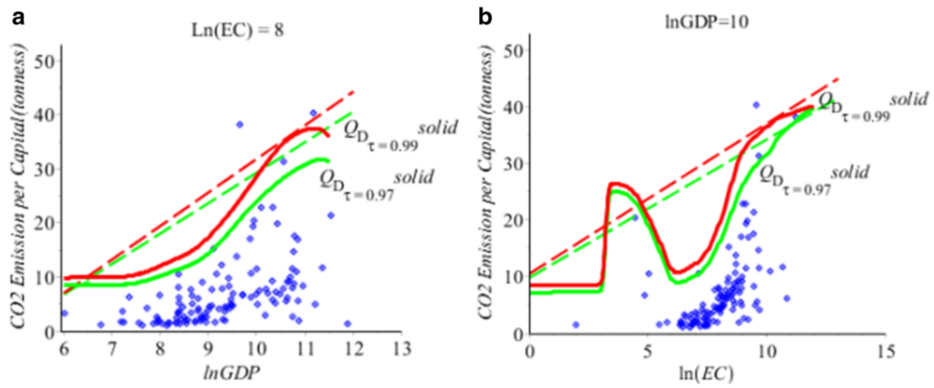

Fig. 8 2D plots for CO_2 Emission, data — blue, $n = 123$, (**a**) Regular Q_R (in dash) and direct Q_D (in solid) of the CO_2 emission vs ln(GDP) when the country's E.C. is 2980.96 kilowatts at $\tau = 0.97$ (green) and 0.99 (red). (**b**) Regular Q_R (in dash) and direct Q_D (in solid) of the CO_2 emission vs ln(E.C.) when the country's GDP is \$13,359.73 at $\tau = 0.97$ (green) and 0.99 (red)

Table 4 CO_2 Emission per capita at high quantiles given ln(GDP) estimators Q_R and Q_D

	$\tau = 0.97$	
ln of GDP per capita ($)	Q_R	Q_D
7.5	15.2181	8.8737
8	18.0437	10.1949
8.5	20.8693	11.7828
9	23.6950	14.4143
9.5	26.5206	19.0458
10	29.3462	24.0338
10.5	32.1718	27.9596
11	34.9975	31.1097
11.5	37.8231	30.7696
12	40.6487	31.2366

2980.96 Kilowatts of Electricity Consumed per capita

$Q_D(\tau|\mathbf{x})$ in *(7)*. We compare the new estimator $Q_D(\tau|\mathbf{x})$ with the regular quantile estimator $Q_R(\tau|\mathbf{x})$ in Huang and Nguyen (2017). Figures 7, 8 and Tables 4, 5 show the differences of the values of two estimators. Figure 7a shows the 3D scatter plot of CO_2 emission vs ln(GDP) and ln(EC) with the fitted regular Q_R surface at $\tau = 0.97$. Figure 7b shows the 3D scatter plot of CO_2 emission vs ln(GDP) and ln(EC) with the fitted direct Q_D surface at $\tau = 0.97$. Figure 7c shows the 3D scatter plot with both the regular Q_R (green) and direct Q_D (red) quantile surfaces of CO_2 emission vs the ln(GDP) and ln(E.C.) at $\tau = 0.97$. It is interesting to see the difference between the Q_R and Q_D quantile surfaces.

We may see the Q_R and Q_D quantile curves more cleanly in 2D plots. Figure 8a shows the 2D scatter plot of CO_2 emission vs ln(GDP) when the country's E.C. is 2980.96 kilowatts with the fitted regular Q_R and direct Q_D curves at at $\tau = 0.97$. Figure 8b shows the 2D scatter plot of CO_2 emission vs ln(E.C.) when the country's GDP is $13,359.73 with the fitted regular Q_R and direct Q_D curves at at $\tau = 0.97$. We note that the Q_R and Q_D quantile regression curves appear to fit the data. In general, the Q_D curves follow the data patterns closer than Q_R quantile lines, and the Q_D produces different estimated CO_2 emissions than the Q_R estimated at high quantiles. In Fig. 7, it is interesting to see that the Q_D conditional quantile surfaces are not linear as the linear planes of the Q_R.

Tables 4 and 5 provide details of the estimated high quantiles about countries' CO_2 emission at $\tau = 0.97$ when the countries consume 2980.96 kilowatts of electricity and have a GDP of $13,359.73, respectively.

Table 5 CO_2 emission per capita at high quantiles given ln(E.C.) estimators Q_R and Q_D

ln of Electricity Consumption	$\tau = 0.97$	
per capita (kilowatts)	Q_R	Q_D
0	6.9775	7.1919
2	11.8632	7.2759
4	16.7490	24.6924
6	21.6348	9.5560
8	26.5206	15.9569
10	31.4064	31.5634
12	36.2921	39.6481

GDP per capita of 13, 359.73

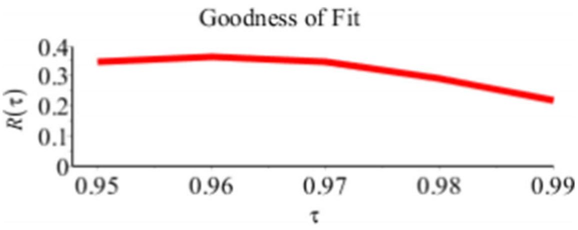

Fig. 9 Relative $R(\tau)$ of Q_D relative to Q_R for the CO_2 emission example

Figure 9 and Table 6 show the Relative $R(\tau)$ in *(9)*, for $\tau = 0.95, \ldots, 0.99$. All values of Relative $R(\tau)$ are larger than 0, which signifies that $V_D(\tau) < V_R(\tau)$ and it also suggests that the direct quantile regression estimator Q_D is a better fit to the CO_2 emission data than the regular quantile regression estimator Q_R.

Over all, it is interesting to see that the proposed direct estimator Q_D gave more variety of predictions than the Q_R on CO_2 emissions relative to gross domestic product and amounts of electricity produced. The relationships are not necessarily linear and model free. We expect that the predictions from Q_D may be more reasonable. The predictions may benefit prevention of further damages of CO_2 emissions to the environment.

6 Conclusions

After the above studies, we can conclude:

1. This paper proposes a new direct nonparametric quantile regression method which is model free. It uses nonparametric density estimation and nonparametric regression techniques to estimate high conditional quantiles. The paper provides a computational five-step algorithm which overcomes the limitations of the estimation in the linear quantile regression model and some other nonparametric quantile regression methods.

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 is different from the bivariate Pareto distribution which has a linear conditional quantile function, in Huang and Nguyen (2017). The simulation results confirm that the proposed new method is more efficient relative to the regular quantile regression estimators and a local polynomial nonparametric estimator.

3. The proposed new direct nonparametric quantile regression can be used to predict extreme values of snowfall and CO_2 emission examples in Huang and Nguyen (2017). The proposed direct quantile regression Q_D estimator gives a variety of predictions which fits data very well. The prediction of relationships are not simply just linear. We expect that the predictions from Q_D may be more reasonable than the regular quantile regression predictions. The new estimator may benefit prevention of further damages of the extreme events to human and the environment.

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

Table 6 Relative $R(\tau)$ values for CO_2 emission example

	$\tau = 0.95$	$\tau = 0.96$	$\tau = 0.97$	$\tau = 0.98$	$\tau = 0.99$
Relative $R(\tau)$	0.3480	0.3612	0.3494	0.2895	0.2151

Acknowledgements

We are grateful for the comments of the reviewers and editor. They have helped us to improve the paper. This research is supported by *the Natural Science and Engineering Research Council of Canada* (*NSERC*) *grant MLH, RGPIN-2014-04621*. We deeply appreciate the work and suggestions of Ramona Rat and Jenny Tieu which helped to improve the paper.

Authors' contributions

The authors MLH and CN carried out this work and drafted the manuscript together. Both authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details

[1]Department of Mathematics & Statistics, Brock University, St. Catharines, Ontario L2S 3A1, Canada. [2]Apotex Inc., Toronto, Ontario M9L 1T9, Canada.

Received: 12 September 2017 Accepted: 31 May 2018
Published online: 18 July 2018

References

Carbon Dioxide Information Analysis Center (2017). http://www.cdiac.ornl.gov. Accessed 20 Oct 2014

Cai, Z: Applied Nonparametric Econometrics. Wang Yanan Institute for Studies in Economics, Xiamen University, China (2013)

Chaudhuri, P: Nonparametric estimates of regression quantile and their local Bahadur representation. Ann. Stat. **2**, 760–777 (1991)

Fukunaga, K: Introduction to Statistical Pattern Recognition. Academic press, New York (1972)

Gumbel, EJ: Bivariate exponential distributions. J. Am. Stat. Assoc. **55**, 698–707 (1960)

Hall, P, Wolff, RCL, Yao, Q: Methods for estimating a conditional distribution. J. Am. Stat. Assoc. **94**, 154–163 (1999)

Huang, ML, Nguyen, C: High quantile regression for extreme events. J. Stat. Distrib. Appl. **4**(4), 1–20 (2017)

Huang, ML, Xu, X, Tashnev, D: A weighted linear quantile regression. J. Stat. Comput. Simul. **85**(13), 2596–2618 (2015)

Koenker, R: Quantile regression. Cambridge University Press, New York (2005)

Koenker, R. Package 'guantreg': Quantile Regression (2018). R Package, Version 5.35 (Available from https://www.r-project.org). Accessed 23 Apr 2018

Koenker, R, Bassett, GW: Regression Quantiles. Econometrica. **46**, 33–50 (1978)

Koenker, R, Machado, JAF: Goodness of fit and related inference processes for quantile regression. J. Am. Stat. Assoc. **96**(454), 1296–1311 (1999)

Li, Q, Racine, JS: Nonparametric Econometrics-Theory and Practice. Prinston University Press, Oxford (2007)

National Weather Service Forecast Office (2017). www.weather.gov/buf. Accessed 22 Sept 2014

Scott, DW: Multivariate Density Estimation, Theory, Practice and Visualization, second edition. John Wiley & Sons, New York (2015)

Silverman, BW: Density estimation for statistics and data analysis. Chapman & Hall, London (1986)

Wang, HJ, Li, D: Estimation of extreme conditional quantile through power transformation. J. Am. Stat. Assoc. **108**(503), 1062–1074 (2013)

Yu, K, Lu, Z, Stander, J: Quantile regression: applications and current research areas. Statistician. **52**(3), 331–350 (2003)

Yu, K, Jones, MC: Local linear regression quantile regression. J. Am. Stat. Assoc. **93**, 228–238 (1998)

Submit your manuscript to a SpringerOpen journal and benefit from:

► Convenient online submission
► Rigorous peer review
► Open access: articles freely available online
► High visibility within the field
► Retaining the copyright to your article

Submit your next manuscript at ► springeropen.com