Journal of Statistical Theory and Practice (2023) 17:32
https://doi.org/10.1007/s42519-023-00325-8

ORIGINAL ARTICLE

An Algorithm of Nonparametric Quantile Regression

Mei Ling Huang[1] · **Yansan Han**[1] · **William Marshall**[1]

Accepted: 25 January 2023 / Published online: 29 March 2023

Abstract
Extreme events, such as earthquakes, tsunamis, and market crashes, can have substantial impact on social and ecological systems. Quantile regression can be used for predicting these extreme events, making it an important problem that has applications in many fields. Estimating high conditional quantiles is a difficult problem. Regular linear quantile regression uses an L_1 loss function [Koenker in Quantile regression, Cambridge University Press, Cambridge, 2005], and the optimal solution of linear programming for estimating coefficients of regression. A problem with linear quantile regression is that the estimated curves for different quantiles can cross, a result that is logically inconsistent. To overcome the curves crossing problem, and to improve high quantile estimation in the nonlinear case, this paper proposes a nonparametric quantile regression method to estimate high conditional quantiles. A three-step computational algorithm is given, and the asymptotic properties of the proposed estimator are derived. Monte Carlo simulations show that the proposed method is more efficient than linear quantile regression method. Furthermore, this paper investigates COVID-19 and blood pressure real-world examples of extreme events by using the proposed method.

Keywords Conditional quantile · Extreme value distribution · Generalized Pareto distribution · Kernel estimation · Linear programming · Nonparametric quantile regression

✉ Mei Ling Huang
mhuang@brocku.ca

Yansan Han
yh13dc@brocku.ca

William Marshall
wmarshall@brocku.ca

[1] Department of Mathematics and Statistics, Brock University, St. Catharines, ON L2S 3A1, Canada

1 Introduction

Extreme events are rare, unusual occurrences such as earthquakes, tsunamis, and market crashes. These events usually have the potential to substantially impact social, and ecological systems. Therefore, understanding and predicting extreme events is of interest in many fields such as earth sciences, traffic prediction, survival analysis and financial markets. To estimating such events' probability requires a model that focuses on the high conditional quantile of heavy-tailed distribution [7]. Therefore, more sophisticated quantile regression methods are used instead of traditional mean regression. Linear quantile regression uses the least absolute deviation (L_1) loss function, and optimization of this loss function is done using linear programming methods. With quantile regression, we can obtain the relationship between variables in high conditional quantiles.

The linear mean regression model is used to estimate the conditional mean of random variable y based on given $\boldsymbol{x} = (1, x_1, x_2, ..., x_k)^T$.

$$\mu_{(y|\boldsymbol{x})} = E(y|x_1, x_2, ..., x_k) = \boldsymbol{x}^T\boldsymbol{\beta} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k, \tag{1}$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, ..., \beta_k) \in R^p,\ p = k + 1$.

Let $y_i, i = 1, 2, ..., n$, be the response variable from a continuous distribution, which is explained by p-dimensional design vector $\mathbf{x}_i$. Then $\boldsymbol{\beta}$ can be estimated from a random sample $\{(y_i, \boldsymbol{x}_i),\ i = 1, 2, ..., n\}$ by applying the method of least squares. By minimizing its L_2-squared distance, we can obtain the least-square (LS) estimators $\widehat{\boldsymbol{\beta}}_{LS}$. This can be represented by the following equation

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

When the response variable is normally distributed, this model has appealing attributes such as computational tractability and accurate conditional mean estimation. However, the measurement of the central location would be significantly affected if there are outliers in the data. Moreover, for extreme events, the response variable usually has a heavy-tailed distribution such as the extreme value distribution [6], and the focus is on the high quantile curves rather than central location. In this circumstance, the mean regression model is inefficient at capturing the critical information required to predict extreme events. Research results show that quantile regression methods are better to apply.

A real-valued random variable y has right-continuous cumulative distribution function (CDF) $F_y(y)$. The τ^{th} quantile of such y is given by

$$Q(\tau) = F^{-1}(\tau) = \inf\{y : F(y) \geq \tau\},\ 0 < \tau < 1.$$

If a random univariate y can be explained by $\mathbf{x} = (1, x_1, x_2, \ldots, x_k)^T \in R^p$. The conditional τ^{th} quantile of y given $\mathbf{x}$ is defined as

$$Q_y(\tau|\mathbf{x}) = F^{-1}(\tau|\mathbf{x}) = \inf\{y : F(y|\mathbf{x}) \geq \tau\}. \tag{3}$$

Then the regular τ^{th} quantile linear regression (QR) of y is defined as [20].

$$Q_R(\tau|x) = \beta_0(\tau) + \beta_1(\tau)x_1 + \beta_2(\tau)x_2 + ... + \beta_k(\tau)x_k,\ 0 < \tau < 1, \tag{4}$$

where $\boldsymbol{\beta}(\tau) = (\beta_0(\tau), \beta_1(\tau), \beta_2(\tau), ..., \beta_k(\tau))^T$. Its estimator $\widehat{\boldsymbol{\beta}}(\tau)$ is obtained by solving the following equation

$$\widehat{\boldsymbol{\beta}}(\tau) = \arg\min_{\boldsymbol{\beta}(\tau)\in R^p} \sum_{i=1}^{n} \rho_\tau\Big(y_i - \mathbf{x}_i^T\boldsymbol{\beta}(\tau)\Big),\ 0 < \tau < 1, \tag{5}$$

where ρ_τ is a quantile-weighted L_1-loss function which is not differentiable,

$$\rho_\tau(u) = u(\tau - I(u < 0)) = \begin{cases} u(\tau - 1), & \text{if } u < 0; \\ u\tau, & \text{if } u \geq 0. \end{cases} \tag{6}$$

In other words, minimizing the expected loss to obtain estimator $\widehat{\boldsymbol{\beta}}(\tau)$. Furthermore, the quantile regression problem can be reformulated as a linear program

$$\min_{(\boldsymbol{\beta}(\tau),\boldsymbol{u},\boldsymbol{v})\in R\times R_+^{2n}} \Big\{\tau 1_n^T\boldsymbol{u} + (1-\tau)1_n^T\boldsymbol{v}|\boldsymbol{X}\boldsymbol{\beta}(\tau) + \boldsymbol{u} - \boldsymbol{v} = \boldsymbol{y}\Big\}, \tag{7}$$

where $\mathbf{X}$ is an $n \times p$ regression design matrix and $\mathbf{u}$, $\mathbf{v}$ are two $n \times 1$ vectors with elements of u_i, v_i respectively.

In the literature, there are other quantile regression methods. Bayesian approaches provide convenient alternative inference tools for quantile regression. A working likelihood is needed to carry out Bayesian analysis [21]. It is interesting to explore the Bayesian quantile regression methods. In Sect. 7.2, we compare a type of Bayesian quantile regression with this paper which proposes direct nonparametric quantile regression.

In this paper, two real-world datasets are analyzed using linear mean regression, linear quantile regression and nonparametric quantile regression. The first dataset is a tri-variate example based on the status of COVID-19 cases in Ontario that comes from the Ontario Government. The second example's data comes from National Health and Nutrition Examination Survey (NHANES) about systolic blood pressures.

1.1 Example 1. Number of Hospitalized COVID-19 Patients in Ontario, Canada (April 19, 2020–June 30, 2021)

COVID-19, the coronavirus is a contagious disease spread worldwide and causing the current ongoing pandemic. Based on United States Centers for Disease Control

and Prevention (CDC)'s report in March 2020, COVID-19 has already done more damage to the world than the SARS pandemic [4], which appeared in 2002. COVID-19's average mortality rate is estimated to be around 3.4% by the WHO. However, in Washington State USA, 67% of critically ill patients died [25]. Furthermore, since there are no specific coronavirus treatments right now [16]. People with severe COVID-19 symptoms must be hospitalized. Ensuring sufficient human and physical resources is essential to minimize the mortality rate.

The virus was confirmed in Canada on January 27, 2020, and in March 2020, as cases of community transmission were confirmed, all of Canada's provinces and territories declared states of emergency. Provinces and territories have, to varying degrees, implemented school and daycare closures, prohibitions on gatherings, closures of non-essential businesses and restrictions on entry. By mid to late summer of 2020, the country saw a steady decline in active cases until the beginning of late summer. Through autumn, the country saw a resurgence of cases in all provinces and territories. On September 23, 2020, Canada declared a "second wave" of the virus. Nation-wide cases, hospitalizations and deaths spiked preceding and following December 2020 and January 2021. Following Health Canada's approval of the Pfizer–BioNTech, mRNA-1273 and the Oxford–AstraZeneca vaccine for use, and on March 5, 2021, they additionally approved the Janssen COVID-19 vaccine for a total of four approved vaccines in the nation [14].

Ontario Canada, in late summer 2021, the province began preparing for a fourth wave of the virus, which was largely affecting unvaccinated individuals. In January 2022, there were changes in the policy regarding testing, such that the reported number of new positive cases no longer reflects the true number of new positive cases. In this paper, we focus on data from April 19, 2020–June 30, 2021, Ontario COVID-19 cases daily data collected for $n^* = 438$ days [13]. We focus on high numbers of hospitalized COVID-19 patient's relationship with percent positive tests last day and the number of new cases. Percent positive tests last day is the percent of COVID-19 tests that were positive in the last day. The response variable is the number of hospitalized COVID-19 patients. To focus on high numbers of hospitalized COVID-19 patients, its upper quartile (75%) 1010 patients, will be used as a threshold. After applying the threshold of 1010, the data was reduced to $n = 103$ days. Figure 1 is a chart plot shows $n^* = 438$ days of number of hospitalized COVID-19 patients. Table 1 shows the top 5 daily number of hospitalized COVID-19 patients in Ontario, Canada.

We also note that there were three waves of COVID-19 during April 19, 2020–June 30, 2021. The top three values are 1043 patients on May-04-2020, 1674 patients on Jan-13-2021, and 2360 patients on April 20, 2021. Our goal is to create a statistical model to analyze the current Ontario hospitalized COVID-19 patients' number and predict future extreme events.

We are interested in the relationship between the response variable y (the number of hospitalize COVID-19 patients) with x_1 (percent positive test last day) and x_2 (the number of new cases). By employing a least-square mean regression model in (1), we can model it using

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

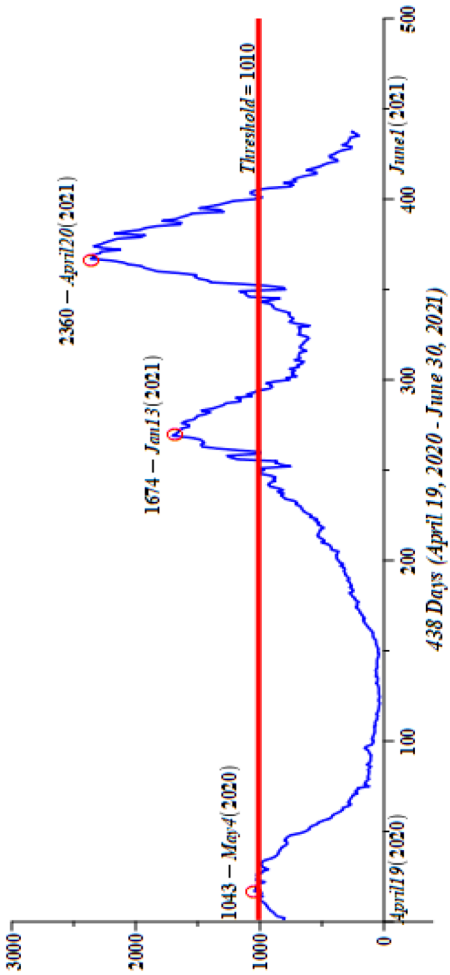


Fig. 1 The number of hospitalized COVID-19 patients in 438 days ($n^* = 438$) between April 19, 2020 and June 30, 2021, with a threshold of 1010 hospitalized COVID-19 patients. The number of days with data value above 1010 threshold is 103 days. ($n = 103$)

Table 1 Top 5 daily data of the number hospitalized of COVID-19 patients in Ontario, Canada (April 19, 2020–June 30, 2021)

Date	The number of hospitalized COVID-19 patients y	Percent positive test last day x_1 (%)	The number of new cases x_2
2021-04-20	2360	10	3469
2021-04-22	2350	7.8	3682
2021-04-27	2336	10.2	3265
2021-04-21	2335	7.9	4212
2021-04-23	2287	8.8	4505

Using the least-square method to estimate $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)^T$. The least-square plane function is

$$\widehat{\mu}_{LS}(x_1, x_2) = 650.2139 + 67.9667x_1 + 0.1731x_2.$$

The 0.95th quantile plane by the regular linear quantile regression (QR) in (5) is given by

$$\begin{aligned}\widehat{Q_R}(0.95|x_1, x_2) &= \widehat{\beta}_0(0.95) + \widehat{\beta}_1(0.95)x_1 + \widehat{\beta}_2(0.95)x_2 \\ &= 640.5811 + 89.8649x_1 + 0.2539x_2.\end{aligned}$$

Figure 2 is a 3D plot of the LS mean regression plane (in Green) and a linear quantile regression plane (in blue) which shows that there is a strong positive relation between number of hospitalize COVID-19 patients and its regressors. The average number of hospitalized patients increases as percent positive test last day increase. Similarly, the average number of hospitalized patients increases as the number of new cases per day increases.

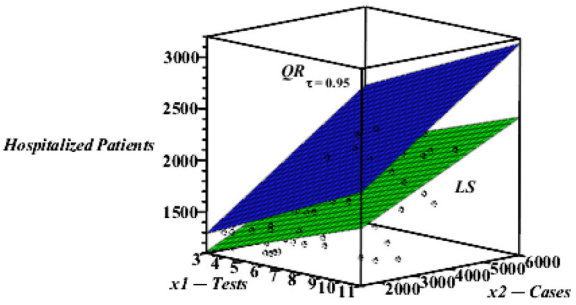


Fig. 2 A 3D LS mean regression plane $\widehat{\mu}_{LS}(x_1, x_2)$ (in green) and the 0.95th linear quantile regression plane $\widehat{Q_R}(0.95|x_1, x_2)$ (in blue) of the number of hospitalized COVID-19 patients y with a threshold of 1010 vs the percent positive test last day x_1 and the number of new cases x_2 ($n = 103$). Data are in black dots. We note that about 95% data below the $\widehat{Q_R}(0.95|x_1, x_2)$ plane

For building a 2D relationship of y—the number of hospitalized COVID-19 patients and x_1—the percent positive test last day. We fix x_2—the number of new cases. Let $x_2 = 3442$ which is the 0.75th quantile of x_2. Then the least square mean regression line and the 0.95th linear quantile regression line of y and x_1 are

$$\widehat{\mu}_{LS}(x_1)\big|_{x_2=3442} = 1246.1294 + 67.9667x_1, \text{ when } x_2 = 3442 \text{ (the 0.75th quantile of } x_2).$$

$$\widehat{Q_R}(0.95|x_1)\big|_{x_2=3442} = 1514.521 + 89.8649x_1, \text{ when } x_2 = 3442 \text{ (the 0.75th quantile of } x_2).$$

Figure 3a provides a scatter plot of x_1—percent positive test last day versus y—the number of hospitalized COVID-19 patients with its least-squares mean regression line (in green) and 0.95th linear quantile regression line (in blue).

Similarly, for building a 2D relationship of y—the number of hospitalized COVID-19 patients and x_2—the number of new cases. We fix x_1—the percent positive test last day. Let $x_1 = 7.8(\%)$ which is the 0.75th quantile of x_1. We have the least square mean regression line and the 0.95th linear quantile regression line of line of y and x_2 are

$$\widehat{\mu}_{LS}(x_2)\big|_{x_1=7.8} = 1180.3541 + 0.1731x_2, \text{ when } x_1 = 7.8 \text{ (the 0.75th quantile of } x_1).$$

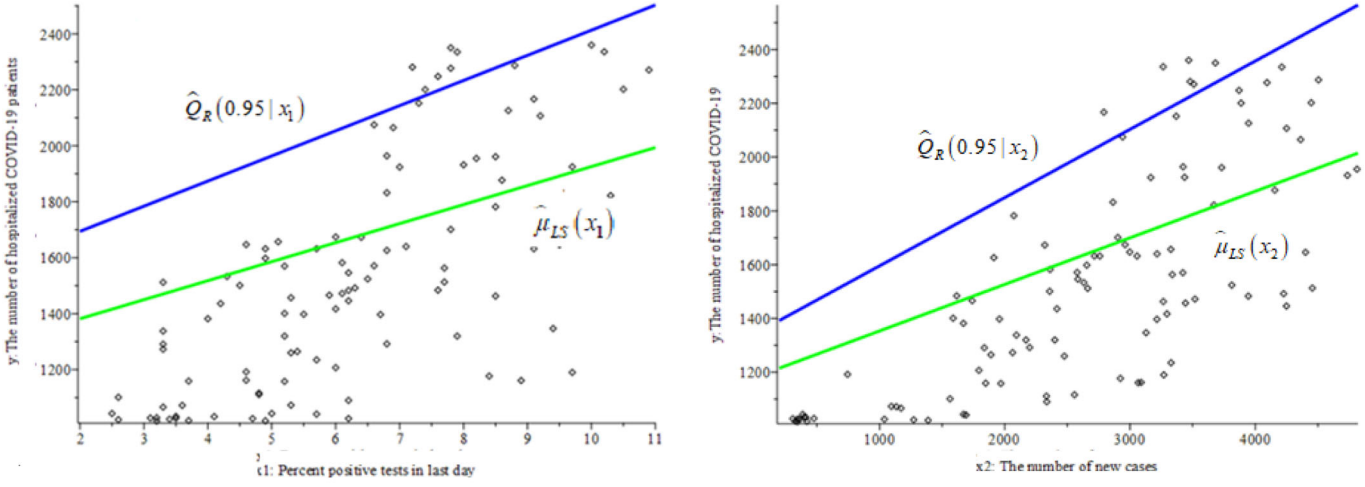
$$\widehat{Q_R}(0.95|x_2)\big|_{x_1=7.8} = 1341.5277 + 0.2539x_2, \text{ when } x_1 = 7.8 \text{ (the 0.75th quantile of } x_1).$$

Figure 3b provides a 2D plots of scatter plot of x_2—the number of new cases versus the y—number of hospitalized COVID-19 patients with its least-squares mean regression lines (in green) and 0.95th linear quantile regression lines (in blue).

From Figs. 2 and 3, we observed that the least-squares mean estimator $\widehat{\mu}_{LS}$ estimates the average number of hospitalized COVID-19 patients in Ontario, but it does not represent the extreme high values of the patients in hospital data pattern. It can not estimate the critical situation when the hospitals are crowded with patients. The 0.95th linear Quantile regression $\widehat{Q_R}(0.95|x_1, x_2)$ indicates the 95% number of hospitalized COVID-19 patients in Ontario under the plane or lines. But we note that this model does not capture nonlinear patterns in the high quantiles. In this paper, we proposed a new quantile regression estimator to improve estimating the high quantile plane for extreme values of numbers of hospitalized COVID-19 patients. We will use new quantile regression model to analysis this example in Sect. 6.

1.2 Example 2: Systolic Blood Pressures (January 2017–December 2018)

Blood pressure is expressed as a measurement with two numbers: systolic blood pressure and diastolic blood pressure. The systolic blood pressure refers to the amount of pressure the blood exerts against the artery walls during the heart contraction. The diastolic blood pressure represents the pressure when the heart rests between beats. High blood pressure, also referred to as hypertension, is blood pressure that is higher than normal. Hypertension may cause complications such as heart attack, stroke, and aneurysm [5, 23].



(a) $\widehat{Q_R}(0.95|x_1)\big|_{x_2=3442}$ and $\widehat{\mu}_{LS}(x_1)\big|_{x_2=3442}$

(b) $\widehat{Q_R}(0.95|x_2)\big|_{x_1=7.8}$ and $\widehat{\mu}_{LS}(x_2)\big|_{x_1=7.8}$

Fig. 3 Data are black dots, $n = 103$. **a** A 2D LS mean regression line $\widehat{\mu}_{LS}(x_1)$ (in green) and a 0.95th linear quantile regression line $\widehat{Q_R}(0.95|x_1)$ (in blue) of the number of hospitalized COVID-19 patients y versus percent positive test last day x_1 when $x_2 = 3442$; **b** A 2D LS mean regression line $\widehat{\mu}_{LS}(x_2)$ (in green) and a 0.95th linear quantile regression line $\widehat{Q_R}(0.95|x_2)$ (in blue) of the number of hospitalized COVID-19 patients y versus the number of new cases x_2 when $x_1 = 7.8$

Table 2 Classification of blood pressure, Centers for Disease Control and Prevention, 2020, USA

BP Classification	SBP (mmHg)	DBP (mmHg)
Normal	Between 90–120	Between 60–80
Prehypertension	120–139	or 80–89
Stage 1 hypertension	140–159	or 90–99
Stage 2 hypertension	≥ 160	or ≥ 100

Based on existing studies, high systolic pressures pose a greater risk of heart disease than elevated diastolic pressure. As a result, the response variable of this example is systolic blood pressure (SBP). The CDC is the national public agency of the United States. The CDC categorizes blood pressure in adults into 4 groups: normal, elevated (prehypertension), hypertension stage 1, and hypertension stage 2 [15]. These classifications are given in Table 2. A millimeter of mercury is a manometric unit of pressure, formerly defined as the extra pressure generated by a column of mercury one millimeter high and currently defined as exactly 133.322387415 pascals. It is denoted mmHg.

NHANES or the National Health and Nutrition Examination Survey, is a program by the CDC that aims to assess the health and nutritional status of adults and children in the USA. We will examine NHANES's 2017–2018 data which consists of $n^* = 6240$ subjects between weights 18.6 kg to 219.6 kg [5]. For this study, a threshold of 160 mmHg is applied since people with SBP higher than 160 mmHg are at high risk for coronary heart disease, which can lead to a heart attack or stroke [3]. After omitting subjects with SBP less than 160 mmHg, the data is reduced to $n** = 261$ subjects.

One common cause of hypertension is obesity. Being overweight increases the chance of developing high blood pressure [29]. We set response variable y – SBP (mmHg) vs regressor x – weight (kg). We treat 12 subjects whose weight $x > 215$ kg as outliers, leaving the $n = 249$ subjects with weight between 18.6–125 kg, Table 3 shows the top 5 data of SBP. Figure 4 presents the SBP for the $n = 249$ subjects, and a threshold of 160 mmHg is indicated. Let the response variable y be SPB and explanatory variable x be the subject's weight. Then, we can employ a least-squares mean regression model to estimate the conditional mean of SBP mmHg (y) given subject's weight kg (x).

Table 3 Top 5 data for SBP of subjects in USA (January 2017–December 2018) with weight between 18.6 and 125 kg

Subject ID	SBP (mmHg) y	Weight (kg) x
100,389	224	78.7
96,586	216	84.4
96,003	216	73.5
101,273	216	66.1
96,234	210	121.2

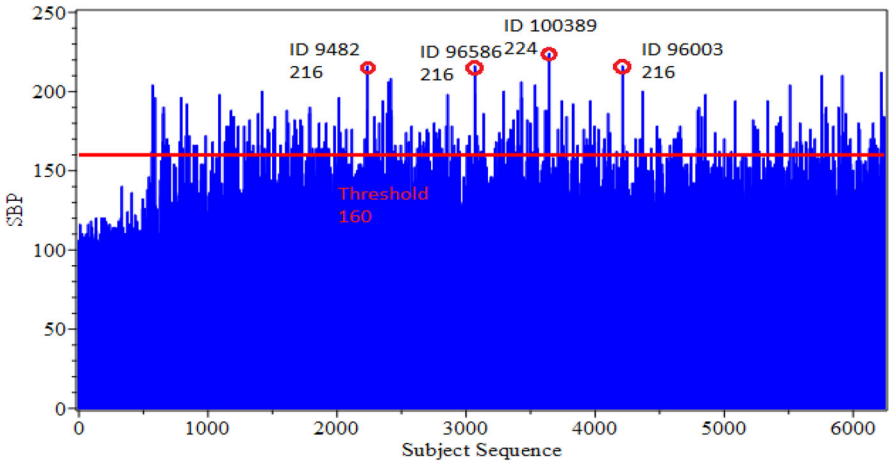


Fig. 4 Systolic blood pressures (SBP mmHg) of subjects ($n^* = 6240$) tested by the Centers for Disease Control and Prevention between Month, United States of American, January 2017—December 2018. The number of subjects who have SBP over 160 mmHg is $n^{**} = 261$

Similar as Example 1, we obtained the least-squares liner regression line and the 0.95th liner quantile regression line as (see Fig. 5)

$$\widehat{\mu}_{LS}(x) = 173.9923 + 0.0191x \quad \text{and} \quad \widehat{Q}_R(0.95|x) = 188.2925 + 0.1791x.$$

Figure 5 shows that the least-squares mean line $\widehat{\mu}_{LS}(x)$ can only estimate the mean values of systolic blood pressure relate to the weight. Also, Fig. 5 also shows about 95% subjects with systolic blood pressure data under line of the 0.95th quantile regression $\widehat{Q_R}(0.95|x)$. However, both $\widehat{\mu}_{LS}(x)$ and $\widehat{Q_R}(0.95|x)$ lines do not catch the relation well between very high values of systolic blood pressures related to weight. This paper, we propose a new direct nonparametric quantile regression method with 3 steps computational algorithm in Sect. 3, and we will discuss Example 2 by using the proposed new quantile regression method in Sect. 6.

In this paper, notation is introduced in Sect. 2. We propose a direct nonparametric quantile regression method with a three-step computer algorithm in Sect. 3. Section 4 gives asymptotic properties of proposed direct nonparametric quantile regression. The results of Monte Carlo simulation are in Sect. 5. Section 6 compares the proposed direct nonparametric quantile regression with the regular quantile regression and mean regression for two examples. Finally, Sect. 7 gives conclusions and discussions..

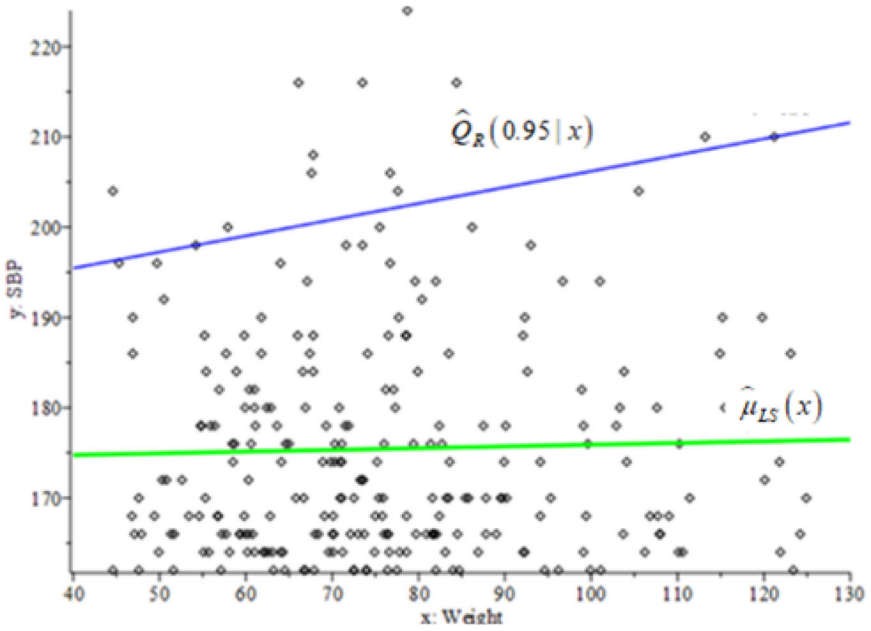


Fig. 5 Data are black dotes, $n = 249$. The $\widehat{\mu}_{LS}(x)$ line and the $\widehat{Q_R}(0.95|x)$ line for Systolic blood pressures (SBP mmHg) y versus weight x for subject with SBP greater than 160 mmHg and weight between 18.6 and 125 kg

2 Notation

2.1 Extreme Value Distribution

Extreme value theory (EVT) is used to study the probability of extreme observations especially from heavy-tailed probability distributions. It helps us find possible limit distribution for sample maxima/minima of independent and identically distributed (i.i.d.) random variables. [7] de Haan & Ferreira (2006) and [24] etc. developed theory; experts also applied extreme value models to many fields, i.e. [6, 9, 17] and [19].

Let $X_1, X_2, ..., X_n$ be i.i.d. random variables. Extreme value theory finds the possible limiting distribution of the sample extreme $\max(X_1, X_2, ..., X_n)$ or $\min(X_1, X_2, ..., X_n)$ as $n \to \infty$.

Definition 1 (Fisher & Tippett, 1928; Gnedenko, 1943) *The c.d.f. of any extreme value distribution is denoted by $G_\gamma(ax + b)$ for any constants $a > 0$, $b \in R$,*

$$G_\gamma(x) = \begin{cases} \exp\left(-(1+\gamma x)^{-\frac{1}{\gamma}}\right), & 1+\gamma x > 0 \text{ and } \gamma \neq 0: \\ \exp\left(-e^{-x}\right), & \gamma = 0,\ x > 0. \end{cases} \tag{8}$$

where the real parameter γ is called the extreme value index.

In this work, we are interested in predicting extreme events by estimating the high conditional quantile curves of heavy-tailed distributions ($\gamma > 0$).

2.2 Generalized Pareto Distribution

A conditional extreme value distribution exceeding a threshold has a generalized Pareto distribution (GPD) ([24] Pickands, 1975, and [8]).

Definition 2 *The c.d.f.* $H_\gamma(x)$ *of the two parameter GPD*(γ, σ) *with shape parameter* γ, *location parameter* μ *and scale parameter* σ *for random variable X is given by*

$$H_\gamma(x) = 1 - \left(1 + \gamma \frac{(x-\mu)}{\sigma}\right)^{-1/\gamma}, \ \gamma \neq 0, 0 < \mu < \infty, \ \sigma > 0, \ \mu < x < \infty. \quad (9)$$

3 New Estimation Method Proposed: A Direct Nonparametric Quantile Regression

This paper proposes a new direct nonparametric quantile regression method. We will ignore the idea of the linear model (4) for generality. Instead, a direct estimator of the true conditional quantile

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

will be obtained by using local conditional quantile estimator $\xi_i(\tau|\boldsymbol{x}) = \widehat{Q}_y(\tau|\boldsymbol{x}_i)$ based on the ith point of given random sample, $\{(y_i, \boldsymbol{x}_i), \ i = 1, ..., n\}$, for $\boldsymbol{x}_i = (x_{1i}, x_{2i}, ..., x_{di})^T$, $i = 1,2,...,n$..

This direct nonparametric quantile regression algorithm has three steps shown as follows:

Step 1 Estimate Conditional c.d.f.

First estimate the conditional c.d.f. $F(y|\boldsymbol{x})$ of y for given $\boldsymbol{x} = (x_1, x_2, ..., x_d)$ using kernel estimation method [26, 27]

$$\widehat{F}(y|\boldsymbol{x}) = \frac{\frac{1}{n}\sum_{i=1}^{n} I(Y_i \leq y) K\left\{\frac{\boldsymbol{x}-X_i}{h}\right\}}{\widehat{g}(\boldsymbol{x})}, \quad (10)$$

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

To estimate the marginal density of x**,** we use a kernel density estimator. Consider a d-dimensional random sample $\boldsymbol{X}_i = (x_{1i}, x_{2i}, ..., x_{di}), \ i = 1, 2, ..., n$ from a population $\boldsymbol{x} = (x_1, x_2, ..., x_d)$ with density $g(x)$. The kernel density estimator for $g(x)$ is given by

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

where $h > 0$ is the bandwidth and the kernel function $K(\boldsymbol{x})$ is a function defined for d-dimensional $\boldsymbol{x} = (x_1, x_2, ..., x_d)$ which satisfies $\int_{R^d} K(\boldsymbol{x})dx = 1$, [11] suggested using

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

where **S** is the sample covariance matrix of the data, and the function k is

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

An estimator for the bandwidth $h > 0$ will be given by [27, p. 85].

$$\widehat{h}_{opt} = A(K)n^{-1/(d+4)},$$

where $A(K) = \left(\frac{4}{d+1}\right)^{1/(d+4)}$ if a multivariate normal kernel is used for smoothing the normal distribution data with unit variance.

Step 2 Estimate the Local Conditional Quantile Function

Ideally, one would like to estimate the conditional quantile function $\xi(\tau|\boldsymbol{x})$ of y given **x** by inverting the estimated conditional c.d.f. in (10) from step 1

$$\widehat{\xi}(\tau|\boldsymbol{x}) = \widehat{Q}_y(\tau|\boldsymbol{x}) = \inf\left\{y : \widehat{F}(y|\boldsymbol{x}) \geq \tau\right\} = \widehat{F}^{-1}(\tau|\boldsymbol{x}).$$

However, since the kernel estimated conditional c.d.f. $\hat{F}(y|\boldsymbol{x})$ has many terms, it is challenging to compute its global inverse function $\hat{\xi}(\tau|\boldsymbol{x})$. To bypass the computational difficulties, we invert the estimated conditional c.d.f. (10) at the ith data point estimates the local conditional quantile point $\widehat{\xi}_i(\tau|\boldsymbol{x}_i)$

$$\widehat{\xi}_i(\tau|\boldsymbol{x}_i) = \widehat{Q}_y(\tau|\boldsymbol{x}_i) = \inf\left\{y : \widehat{F}(y|\boldsymbol{x}_i) \geq \tau\right\} = \widehat{F}^{-1}(\tau|\boldsymbol{x}_i), \;\; i = 1, 2, ..., n.$$

Step 3 Propose a Nonparametric Direct Quantile Regression

This paper proposes a direct nonparametric quantile regression estimator for the τ^{th} conditional quantile curve of **x** using Nadaraya-Watson (NW) nonparametric regression estimator [28] on $\left(\boldsymbol{x}_i, \widehat{\xi}_i(\tau|\boldsymbol{x}_i)\right)$ for $i = 1, 2, ..., n$. The proposed direct nonparametric quantile regression is given by

$$\widehat{Q_N}(\tau|\boldsymbol{x}) = \widehat{\xi}(\tau|\boldsymbol{x}) = \frac{\sum_{i=1}^{n} K_{\boldsymbol{h}}\{\boldsymbol{x} - \boldsymbol{X}_i\}\widehat{\xi}_i(\tau|\boldsymbol{x}_i)}{\sum_{j=1}^{n} K\left\{\boldsymbol{x} - \boldsymbol{X}_j\right\}} = \sum_{i=1}^{n} W_{h_x}(\boldsymbol{x}, \boldsymbol{X}_i)\widehat{\xi}_i(\tau|\boldsymbol{x}_i), \; 0 < \tau < 1 \tag{11}$$

where the equivalent kernel $W_{h_x}(\boldsymbol{x},\ \boldsymbol{X}_i)$ is

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

where

$$K_{\boldsymbol{h}}\{\boldsymbol{x} - \boldsymbol{X}_i\} = \frac{1}{nh_1...h_d}\prod_{j=1}^{d} K\left(\frac{x - x_{ij}}{h_j}\right),\ i = 1,\ 2,\ ...,\ n,$$

where K is the kernel function and $h_j > 0$ is the bandwidth for the jth dimension, and $\boldsymbol{h} = (h_1,\ h_2,\ ...,\ h_d)$.

We will use the standard normal kernel K for the kernel estimation. The optional bandwidth will also be considered [27],

$$h_{j,opt} = \left\{\int t^2 K(t)dt\right\}^{-2/5}\left\{\int K(t)^2 dt\right\}^{1/5}\left\{\int \left(\nabla^2(\boldsymbol{x})\right)^2 d\boldsymbol{x}\right\}^{-1/5} n^{-1/5},\ j = 1,\ 2,\ ...,\ d,$$

where n is the sample size of the random sample.

4 Asymptotic Properties of the Propose Direct Nonparametric Quantile Regression

In this section, we derive the asymptotic distribution of the proposed direct nonparametric quantile regression estimator $\widehat{\xi}(\tau|\boldsymbol{x})$ in (11). Let the following conditions hold:

Condition 1 (C1). In (10both the estimated conditional c.d.f. $\widehat{F}(y|\boldsymbol{x})$ of y given $\boldsymbol{x} = (x_1,\ x_2,\ ...,\ x_d)^T$ and p.d.f. $g(\boldsymbol{x})$ of $\boldsymbol{x}$ have continuous second-order derivatives with respect to $\boldsymbol{x}$. Also, $K(\bullet)$ is a symmetric, bounded, and compactly supported probability density function.

Condition 2 (C2). In (11), the product $nh_1...h_d \to \infty$ as $h_j \to \infty$ for all $j = 1,\ 2,\ ...,\ d$.

Condition 3 (C3). The estimated conditional c.d.f. $\widehat{F}(y|\boldsymbol{x})$ of y given $\boldsymbol{x} = (x_1,\ x_2,\ ...,\ x_d)^T$ in (8) has a conditional p.d.f. $f(y|\boldsymbol{x})$ of y given $\boldsymbol{x}$ that is continuous in $\mathbb{R}^d$ and $f(\xi(\tau|\boldsymbol{x})) > 0$.

The main asymptotic result for $\widehat{\xi}(\tau|\boldsymbol{x})$ given in Theorem 1.

Theorem 1 *Asymptotic properties of proposed direct nonparametric quantile regression, under Conditions C1, C2, and C3,*

$$(nh_1,\ h_2,\ ...,\ h_d)^{1/2}\left(\widehat{\xi}(\tau|\boldsymbol{x}) - \xi(\tau|\boldsymbol{x}) - \sum_{j=1}^{d} h_j^2 B_{\tau,\,j}(x)\right) \xrightarrow{D} N(0,\ V_\tau(\boldsymbol{x})),\ \text{as } n \to \infty \tag{12}$$

where $\widehat{\xi}(\tau|\boldsymbol{x})$ *is defined in (9),* $\xi(\tau|\boldsymbol{x})$ *is the true conditional quantile, and*

$$B_{\tau,j} = \frac{B_j(y,\boldsymbol{x})}{f(\xi(\tau|\boldsymbol{x}))},$$

where

$$B_{\tau,j}(x) = \frac{1}{2}\int v^2 K(v)dv\left[F''_{jj}(y|\boldsymbol{x}) + \frac{2g'_j F'_j(y|\boldsymbol{x})}{g(\boldsymbol{x})}\right],\ j = 1, 2..., d.$$

Here, for a real function $H(y,\boldsymbol{x}),\ j = 1, 2..., d,\ H'_j(y,\boldsymbol{x})$ *and* $H''_{jj}(y,\boldsymbol{x})$ *are the first and second derivatives of* $H(y,\boldsymbol{x})$ *with respect to* x_j*, respectively. For* $j = 0$*,* $H'_0(y,\boldsymbol{x})$ *and* $H''_{00}(y,\boldsymbol{x})$ *are the first and second derivatives of* $H(y,\boldsymbol{x})$ *with respect to y, respectively. Also,*

$$V_\tau(\boldsymbol{x}) = \frac{\tau\left(1-\tau\prod_{j=1}^{d}\left[\int K^2(v_j)dv_j\right]\right)}{f^2(\xi(\tau|\boldsymbol{x}))g(\xi(\tau|\boldsymbol{x}))}.$$

Proof. It is similar to Theorem 2 in [18].

5 Computer Simulations

Monte Carlo Simulation is used in this Section to compare efficiencies of the new proposed direct nonparametric quantile regression estimator $\widehat{Q_N}(\tau|x)$ (QN) in (11) against the regular linear quantile regression estimator $\widehat{Q_R}(\tau|x)$ (QR) by using (4) and (5). We generate $m = 1000$ random samples of size $n = 200$ from the Fisk distribution (the Burr XII distribution) [10] for this simulation.

The Fisk distribution has the c.d.f. with γ as the tail index.

$$F(y,\gamma) = \frac{1}{1+y^{-1/\gamma}},\ \gamma > 0,\ \ y > 0. \tag{13}$$

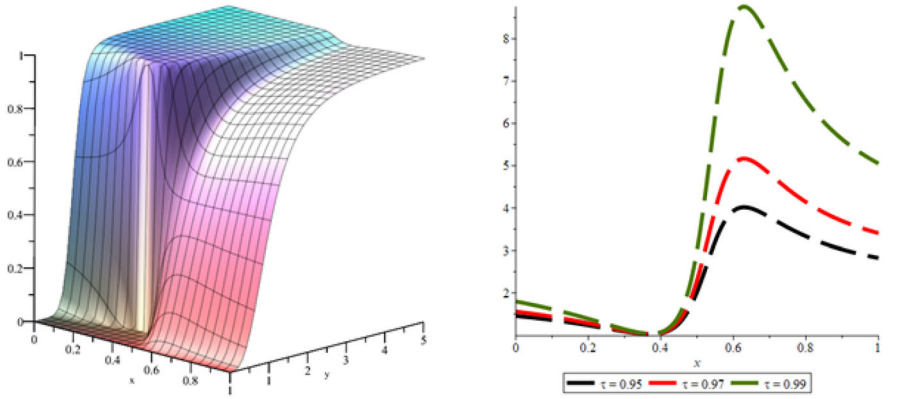
Let X be a one-dimensional and uniformly distributed on [0, 1], and Y given $X = x$ has the Fisk distribution with the conditional c.d.f.

$$F_{Fisk}(y|x,\gamma(x)) = \frac{1}{1+y^{-1/\gamma(x)}},\ \gamma(x) > 0,\ 0 \le x \le 1,\ \ y > 0 \tag{14}$$

and a conditional tail index given by

$$\gamma_{Fisk}(x) = \frac{3}{100}\left(\frac{120x^2-90x+17}{15x^2-15x+4}\right),\ 0 \le x \le 1. \tag{15}$$

The joint distribution of Fisk y and uniform x is given in Fig. 6. The true τ^{th}



(a) Joint Distribution (b) Conditional Quantile Curves

Fig. 6 a The joint distribution of Fisk y and uniform x. **b** Fisk distributed y's conditional quantile given x at $\tau = 0.95$, $\tau = 0.97$ and $\tau = 0.99$

conditional quantile function of y for given x is

$$Q_{YFisk}(\tau|x) = Q_y(\tau|x) = \left(\frac{\tau}{1-\tau}\right)^{\gamma(x)}, \ 0 \le x \le 1, \ \gamma(x) > 0, \ 0 < \tau < 1. \tag{16}$$

Figure 6 shows the true conditional quantiles of Fisk distributed y given Uniform x is not linear. We note that the traditional quantile regression has the assumption of linear model, We expect that the proposed direct nonparametric quantile regression estimator $\widehat{Q_N}(\tau|x)$(QN) should outperform the regular quantile linear regression estimator $\widehat{Q_R}(\tau|x)$ (QR) because $\widehat{Q_N}(\tau|x)$ does not have the linear model limitation. For comparison, we will examine the average simulation plots, box plots, simulation mean squared errors, and simulation efficiencies in the following section.

We use two quantile regression methods are going to be used to estimate the true conditional quantile of the Fisk distribution,

1. The regular linear quantile regression method (QR) $\widehat{Q_R}(\tau|x)$ based on (4),

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

2. The proposed direct nonparametric quantile regression method (QN) $\widehat{Q_N}(\tau|x)$ based on (11)

$$\widehat{Q_N}(\tau|x) = \sum_{i=1}^{n} W_{h_x}(x, X_i)\widehat{\xi}_i(\tau|x_i), \ 0 < \tau < 1.$$

For each method, we generate size $n = 200$, $m = 1000$ samples, $\widehat{Q_{R,i}}(\tau|x)$ and $\widehat{Q_{N,i}}(\tau|x)$ are estimated from the ith sample, where $i = 1, 2, ..., n$. Then we compare the $\widehat{Q_R}(\tau|x)$ and $\widehat{Q_N}(\tau|x)$ with the true quantile computed base on function (16) $Q_{YFisk}(\tau|x)$ quantiles for $\tau = 0.95,\ 0.97,$ and 0.99.

5.1 Average Simulation Plots

The average of $m = 1000$, $n = 200$ estimated curves obtained by QR and QN methods are compared with the true τ^{th} conditional quantile $Q_y(\tau|x)$ at $\tau = 0.95$, 0.97, and 0.99 in Fig. 7. It shows that the proposed direct nonparametric quantile regression estimator $\widehat{Q_N}(\tau|x)$ in red curves follow the true $Q_{YFisk}(\tau|x)$ in (16) in black dash closer than the regular linear quantile regression estimator $\widehat{Q_R}(\tau|x)$ straight lines at high quantiles.

Figure 8 compares the box plots for the estimates for y given $x = 0.4$ of $\widehat{Q_R}(\tau|x)$ and $\widehat{Q_N}(\tau|x)$ at (a) $\tau = 0.95$, (b) $\tau = 0.97$ and (c) $\tau = 0.99$. The result shows that the $\widehat{Q_N}(\tau|x)$ are more concentrated expected values and smaller variance to the true conditional quantile values $Q_y(\tau|x)$ in red straight lines. The proposed $\widehat{Q_N}(\tau|x)$ has a better performance than the regular linear quantile regression $\widehat{Q_R}(\tau|x)$.

5.2 Simulation Mean Squared Errors and Simulation Efficiencies

The simulation mean squared error (SMSE) of the QR estimator and QN estimator are defined as follows

$$\text{SMSE}\big(\widehat{Q_R}(\tau)\big) = \frac{1}{m}\sum_{i=1}^{m}\int_0^1 \big(\widehat{Q_{R,i}}(\tau|x) - Q_{YFisk}(\tau|x)\big)^2\, dx; \tag{17}$$

$$\text{SMSE}\big(\widehat{Q_N}(\tau)\big) = \frac{1}{m}\sum_{i=1}^{m}\int_0^1 \big(\widehat{Q}_{N,i}(\tau|x) - Q_{YFisk}(\tau|x)\big)^2\, dx, \tag{18}$$

where the true τ^{th} conditional quantile $Q_{YFisk}(\tau|x)$ is computed base on function (16), the $\widehat{Q_{R,i}}(\tau|x)$ and $\widehat{Q_{N,i}}(\tau|x)$ are estimated from the ith sample, $i = 1,2,\ldots,n$. from (4) and (11) respectively.

Then the simulation efficiency (SEFF) of QN estimator relative to QR estimator is defined as

$$\text{SEFF}_{Q_R}\big(\widehat{Q_N}(\tau)\big) = \frac{\text{SMSE}\big(\widehat{Q_R}(\tau)\big)}{\text{SMSE}\big(\widehat{Q_N}(\tau)\big)} \tag{19}$$

where the $\text{SMSE}\big(\widehat{Q_R}(\tau)\big)$ and $\text{SMSE}\big(\widehat{Q_N}(\tau)\big)$ of the QR and QN estimator is given above.

A summary of the SMSEs and SEFFs for the two estimation methods is provided in Table 4.

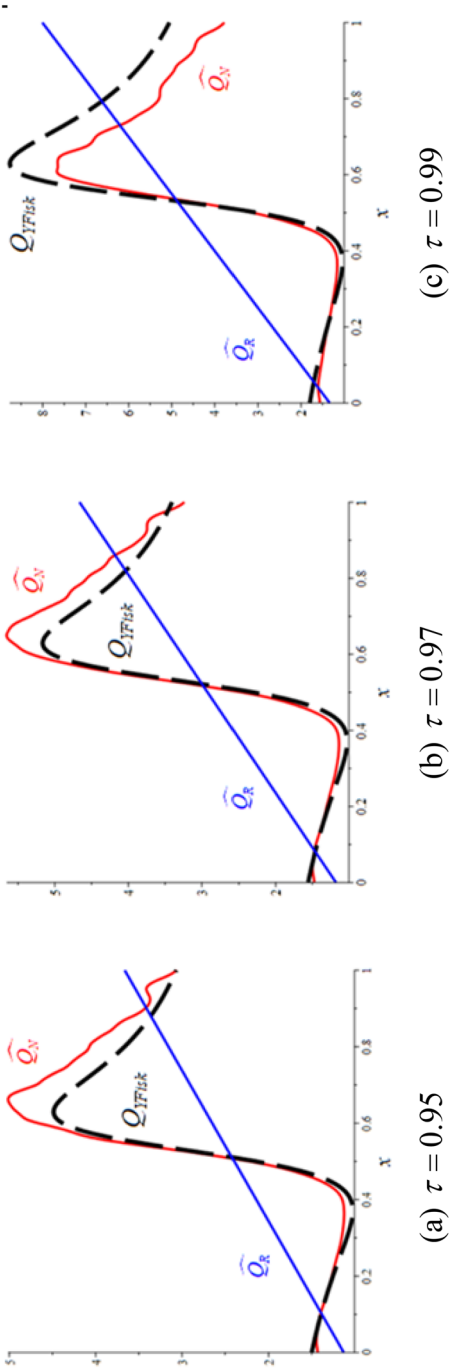


Fig. 7 Plots of simulation average $\widehat{Q}_R(\tau|x)$ (blue) and $\widehat{Q_N(\tau|x)}$ (red) for $m = 1000$ samples with size $n = 200$ versus the true conditional quantile $Q_{YFisk}(\tau|x)$ (black dash) at **a** $\tau = 0.95$, **b** $\tau = 0.97$, **c** $\tau = 0.99$

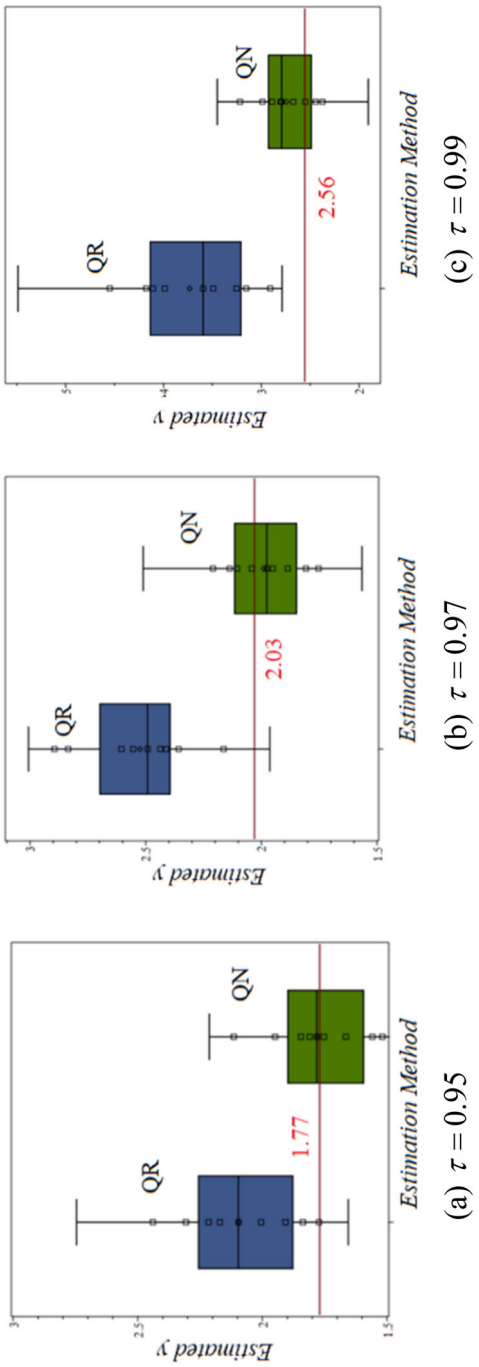


Fig. 8 Box plots of the estimates $\widehat{Q_R(\tau|\boldsymbol{x})}$ (blue) and $\widehat{Q_N(\tau|x)}$ (green) when $x = 0.4$. The true $Q_y(\tau|x)$ values are the red lines, with $m = 1000$ samples at **a** $\tau = 0.95$, **b** $\tau = 0.97$, **c** $\tau = 0.99$

Table 4 Simulation mean squared errors (SMSE) and simulation efficiencies (SEFF) for estimating $\widehat{Q_R}(\tau|x)$ and $\widehat{Q_N}(\tau|x)$ using $m = 1000$ samples of size $n = 200$

τ	0.95	0.96	0.97	0.98	0.99
$\text{SMSE}(\widehat{Q_R}(\tau))$	0.509745	0.677769	1.015396	1.620814	3.750400
$\text{SMSE}\left(\widehat{Q_N}(\tau)\right)$	0.441579	0.531421	0.760052	0.763920	1.507929
$\text{SEFF}_{QR}\left(\widehat{Q_N}(\tau)\right)$	**1.154370**	**1.275391**	**1.335956**	**2.121707**	**2.487120**

Values on the last row are bold which means that $\text{SEFF}_{QR}\left(\widehat{Q_N}(\tau)\right) > 1$

The SMSE of $\widehat{Q}_R(\tau|x)$ and $\widehat{Q_N}(\tau|x)$ are in Table 4. At each τ levels, the $\text{SMSE}\left(\widehat{Q_N}(\tau)\right)$ value is smaller than $\text{SMSE}\left(\widehat{Q}_R(\tau)\right)$ value, then the SEFF of $\widehat{Q_N}(\tau|x)$ estimator relative to $\widehat{Q_R}(\tau|x)$ estimator is greater than 1 at every τ levels. This indicates that the proposed direct nonparametric quantile regression is more efficient than the regular quantile regression for estimating high conditional quantiles of the Fisk distribution $Q_{YFisk}(\tau|x)$ in (16).

The SMSE for $\widehat{Q_R}(\tau|x)$ and proposed $\widehat{Q}_N(\tau|x)$ at different τ levels are plotted in Fig. 9a, which shows $\text{SMSE}\left(\widehat{Q_R}(\tau)\right) > \text{SMSE}\left(\widehat{Q_N}(\tau)\right)$. Figure 9b shows that $\text{SEFF}\left(\widehat{Q_N}(\tau)\right) > 1$ relative to $\widehat{Q_R}(\tau)$.

6 Applications

In this Section, we apply the proposed nonparametric QN methods in (11) and regular QR method in (4) to two real-world examples in Sect. 1.

Example 1. Hospitalized COVID-19 Patients in Ontario, Canada (April 19, 2020–June 30, 2021)

Let us recall the Example 1 in Sect. 1, the Ontario, Canada Hospitalized COVID-19 patients' example. The purpose of this example is to predict the high conditional quantile curves of the number of hospitalized COVID-19 patients, extreme values of which may cause significant strain on the health system. A threshold of 1010 (hospitalized patients) is applied. After the threshold is applied, $n = 103$ data are retained.

Goodness-of-fit Tests for Heavy Tailed Distribution

We will determine if the reduced data set, $n = 103$, with greater than 1010 hospitalized COVID-19 patients is from a generalized Pareto distribution (GPD) [8], with

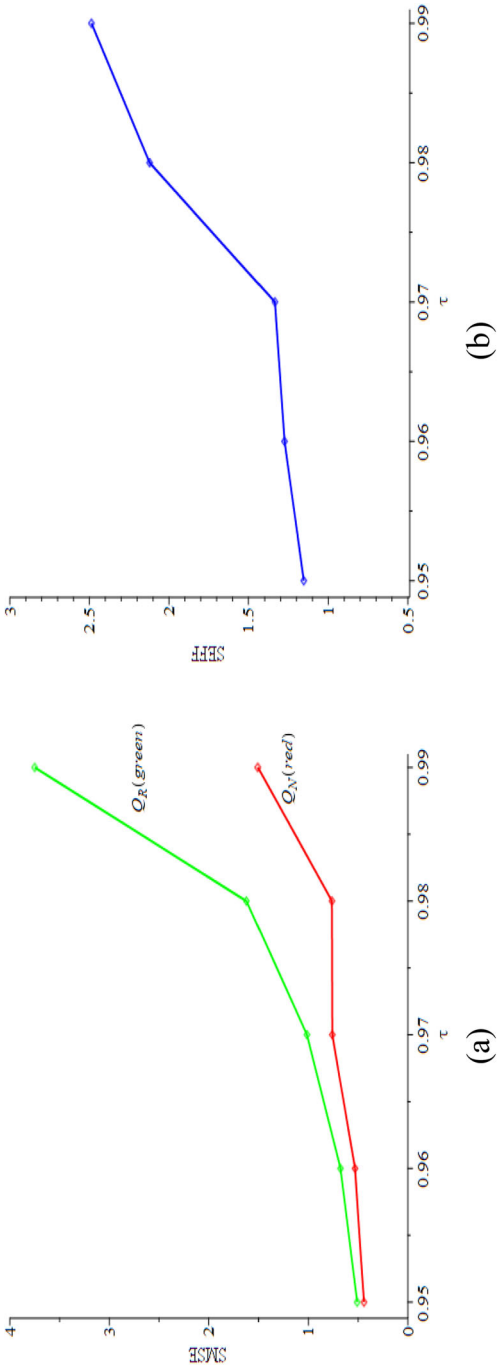


Fig. 9 a Simulation mean squared errors $\mathrm{SMSE}\left(\widehat{Q_R(\tau)}\right)$ (green line) and $\mathrm{SMSE}\left(\widehat{Q_N(\tau)}\right)$ (red line); **b** (blue) $\mathrm{SEFF}\left(\widehat{Q_N(\tau)}\right)$ relative to $\widehat{Q_R(\tau)}$ (blue line)

Table 5 Test statistic values and p-values of three goodness-of-fit tests using MLE for the y—number of hospitalized COVID-19 patients

	K-S	A-D	C-v-M
Test Statistic	0.1097	3.2526	0.3248
p-value	0.1495	0.0204	0.1152

probability density function (pdf.)

$$f_{\gamma,\mu,\sigma}(x) = \frac{1}{\sigma}\left(1+\gamma\frac{x-\mu}{\sigma}\right)^{-\frac{1}{\gamma}-1}, \ \gamma \neq 0, \ \sigma > 0, 0 < \mu < \infty, \ \mu \leq x < \infty, \tag{20}$$

where γ is the shape parameter, μ is a location parameter (this example takes $\mu =$ threshold 1010 patients), and σ ($\sigma > 0$) is a scale parameter. When $\gamma > 0$, the GPD describes a heavy tailed distribution which is what we are interested in. We would like to fit the GPD given in (20) to the $n = 103$ data and the maximum likelihood estimates are $\widehat{\gamma}_{MLE} = -0.6456$, $\widehat{\sigma}_{MLE} = 912.7147$.

Table 5 shows three Goodness-of-fit tests. the Kolmogorov–Smirnov test (K–S) [22], Anderson–Darling (A-D) test [2], and Cramer-von-Mises test (C-v-M) [1], are verified how well the estimated GPD model fits the data. All three tests showed no evidence to reject hypothesis that the data is from a generalized Pareto distribution at $\alpha = 0.01$ significance level. This paper emphasizes that high conditional quantiles estimation is important for heavy-tailed distribution since they are related to extremes. In general, quantile regression method applies data with any distribution.

Proposed Nonparametric Quantile Regression

In Sect. 1, we applied the least-squares mean regression and linear quantile regression method to the COVID-19 data. In this section, we will use proposed quantile regression method to estimate the high quantile planes of the extreme number of hospitalized COVID-19 patients and compare with previous two models.

The QN method in (11) does not need the linear model assumption. Following the steps in Sect. 3 to apply this method, we first use Gaussian kernel and obtain bandwidths of $h_1 = 0.8744$ and $h_2 = 498.5987$ to estimate the conditional c.d.f. of y given $\mathbf{x}$. Then, we compute the estimated local conditional quantile function. Lastly, we use a Gaussian kernel and bandwidth h_1, h_2 in the Nadaraya-Watson estimator to estimate the quantile surfaces. The proposed nonparametric Q_N quantile regression surfaces at $\tau = 0.95$ and 0.99 are provided in Fig. 10.

Next, we will check the estimation quantile regression curves with the number of new case number fixed at 3442 and then percent positive test last day fixed at 7.8%.

Regression of y on x_1 when $x_2 = 3442$ (the th Quantile of x_2)

We substitute $x_2 = 3442$ into $\widehat{Q_R}(\tau|x_1)$ and $\widehat{Q_N}(\tau|x_1)$ surfaces to explore the high conditional quantile curves. Figure 11 gives a scatter plot of the x_1- percent positive test last day vs. the y—number of hospitalized COVID-19 patients with $\widehat{\mu}_{LS}(x_1)$,

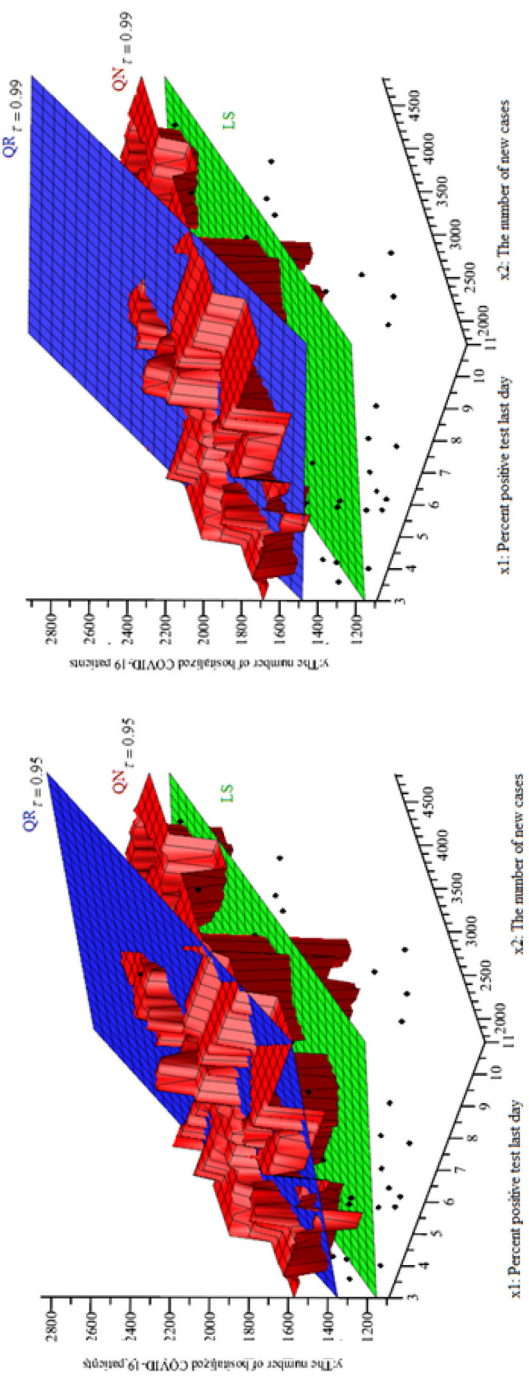
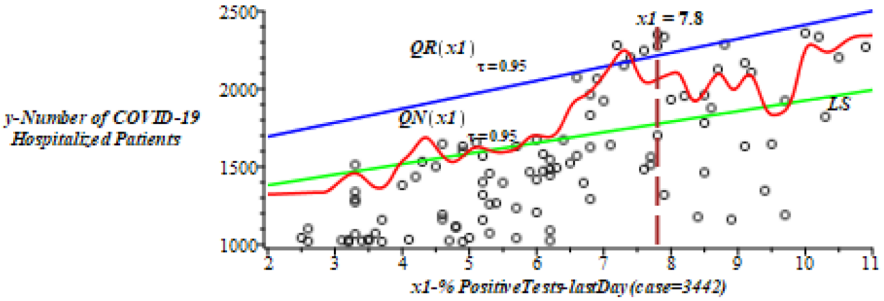
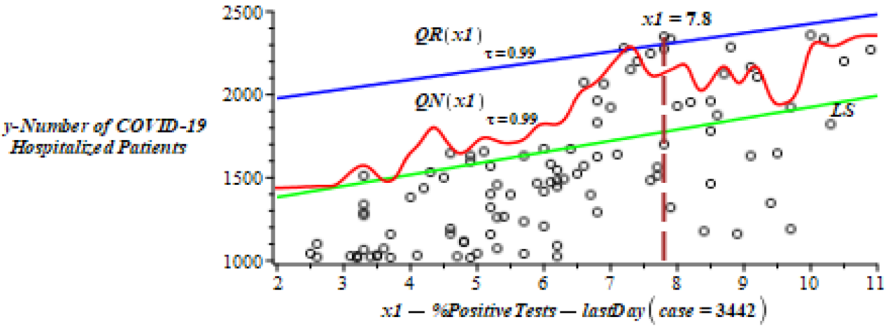


Fig. 10 The 3D plots of $\widehat{\mu}_{LS}(x_1, x_2)$ (green), $\widehat{Q_R}(\tau|x_1, x_2)$(blue) and $\widehat{Q_N}(\tau|x_1, x_2)$ (red) surfaces at **a** $\tau = 0.95$ and **b** 0.99 for the x_1- percent positive test last day and the x_2- number of new cases vs. the y—number of hospitalized COVID-19 patients greater than 1010 and the $n = 103$, data is in black dots



(a) $\tau = 0.95$, when $x_2 = 3442$.



(b) $\tau = 0.99$, when $x_2 = 3442$.

Fig. 11 The 2D plots of the $\widehat{\mu}_{LS}(x_1)$ (green), $\widehat{Q_R}(\tau|x_1)$ (blue) and $\widehat{Q_N}(\tau|x_1)$ (red) quantile curves at **a** $\tau = 0.95$ and **b** $\tau = 0.99$ for the x_1- percent positive test last day versus the y—number of hospitalized COVID-19 patients. Data is in black dots, $n = 103$

$\widehat{Q_R}(\tau|x_1)$ and $\widehat{Q_N}(\tau|x_1)$ curves for $\tau = 0.95$ and 0.99. Table 6 gives the mean and the conditional quantile estimates.

Figure 11 suggested that the $\widehat{Q_R}(\tau|x_1)$ predicts the y—number of hospitalized COVID-19 patients will increase indefinitely as the x_1 percent positive test last day increases. Conversely, the $\widehat{Q_N}(\tau|x_1)$ curve shows that there is maximum value of number of hospitalized COVID-19 patients can reach. The $\widehat{Q_N}(\tau|x_1)$ predicts reveal the reality that the available space in hospitals has a limit. Furthermore, the $\widehat{Q_N}(\tau|x_1)$ curves seem to fit the data better than the $\widehat{Q_R}(\tau|x_1)$ lines. Especially, at lower values of x_1.

From Table 6 we can see that $\widehat{Q_N}(\tau|x_1)$ predictions fit the data better than $\widehat{Q_R}(\tau|x_1)$ predictions. $\widehat{Q_N}(\tau|x_1)$ predictions can capture when $x_1 < 7\%$ the data y—number of hospitalized COVID-19 patients slowly increases from a relative low value of x_1. $\widehat{Q_N}(\tau|x_1)$ predicts also capture when $x_1 = 8\%$, the data y—number of hospitalized patients is high. Also, since the $\widehat{Q_R}(\tau|x_1)$ model is linear, it is over predicting the y values, where resulted in the predictions made by $\widehat{Q_R}(\tau|x_1)$ predicts are higher than $\widehat{Q_N}(\tau|x_1)$ predictions at $x_1 = 4 \sim 6\%$. On the other hand, the $\widehat{\mu}_{LS}(x_1)$ predicts can only represent the mean value of the y—number of hospitalized COVID-19 patients.

Table 6 The number of hospitalized COVID-19 patients (y) related to the % positive test last day (x_1) when the number of new cases $x_2 = 3442$

x_1- positive test last day (%)	$\widehat{\mu}_{LS}(x_1)$	$\tau = 0.95$		$\tau = 0.99$	
		$\widehat{Q_R}(\tau\|x_1)$	$\widehat{Q_N}(\tau\|x_1)$	$\widehat{Q_R}(\tau\|x_1)$	$\widehat{Q_N}(\tau\|x_1)$
4	1518.00	1873.98	1535.69	2090.87	1650.52
5	1585.96	1963.85	1605.34	2147.06	1720.26
6	1653.93	2053.71	1702.08	2203.26	1824.84
7.8 (75% quartile)	1776.27	2215.47	2063.05	2304.42	2131.77
9	1857.83	2323.31	2016.54	2371.86	2095.37

Regression of y on x_1 when $x_2 = 7.8$ (the 0.75th Quantile of x_1)

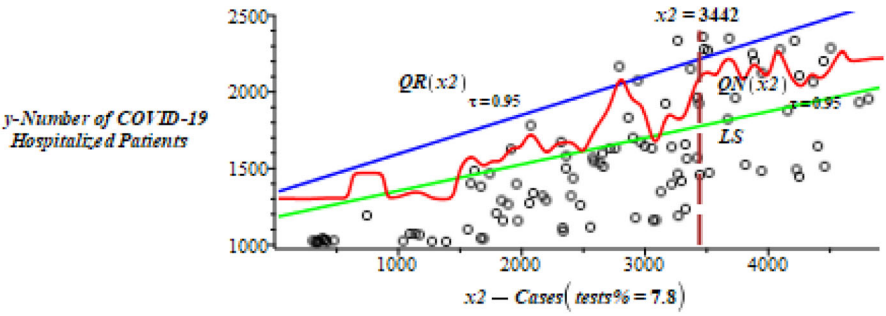
Substitute $x_1 = 7.8$ into $\widehat{Q_R}(\tau|x_2)$ and $\widehat{Q_N}(\tau|x_2)$ to explore the high conditional quantile curves for percent positive test last day equals 7.8 (%). Figure 12 gives a scatter plot of the number of new cases vs the number of hospitalized COVID-19 patients with $\widehat{\mu}_{LS}(x_2)$, $\widehat{Q_R}(\tau|x_2)$ and $\widehat{Q_N}(\tau|x_2)$ curves for $\tau = 0.95$ and 0.99. Table 7 gives the mean and the conditional quantile estimates.

At first, Table 7 shows that the QR method gives quantile curves crossing errors: when $x_2 = 1250$, $\widehat{Q_R}(\tau = 0.95|x_2) = 1658.91 > \widehat{Q_R}(\tau = 0.99|x_2) = 1595.70$ (bold numbers), and when $x_2 = 1750$, $\widehat{Q_R}(\tau = 0.95|x_2) = 1785.86 > \widehat{Q_R}(\tau = 0.99|x_2) = 1757.36$ (bold numbers). Thus, QR estimator gives not reasonable results. QN method never has crossing error result since algorithm in Sect. 3 and $\widehat{F}(y|\boldsymbol{x})$ in (10) has monotonicity which guarantee if $\tau_1 < \tau_2$ then $\widehat{Q_N}(\tau_1|\boldsymbol{x}) < \widehat{Q_N}(\tau_2|\boldsymbol{x})$ in (11) at same $\boldsymbol{x}$.

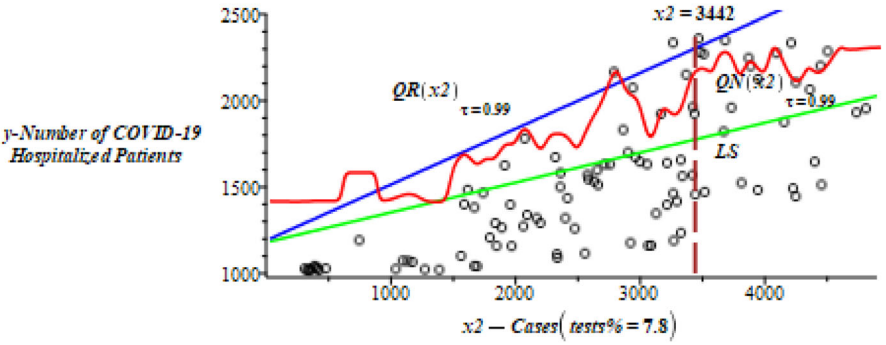
Based on Fig. 12 and Table 7, for $\tau = 0.95$ and 0.99, the $\widehat{Q_N}(\tau|x_2)$ predictions never exceed a value around 2400 hospitalized COVID-19 patients, consistent with the reality that Ontario's hospital resource are limited. Contrarily, the $\widehat{Q_R}(\tau|x_2)$ predictions increase indefinitely. Thus, the proposed $\widehat{Q_N}(\tau|x_2)$ predictions are more realistic. For example, we also observed that before the $x_2 = 1750$ the data y—number of hospitalized COVID-19 patients is not high. Accordingly, when the $x_2 = 1250$, $\widehat{Q_N}(\tau|x_2)$ prediction is lower than $\widehat{Q_R}(\tau|x_2)$ predict. Similarly, around $x_2 = 2750$ new cases, data has a surge of the y—number of hospitalized COVID-19 patients. This time, the $\widehat{Q_N}(\tau|x_2)$ predictions is higher than $\widehat{Q_R}(\tau|x_2)$ predictions. Overall, the proposed $\widehat{Q_N}(\tau|x_2)$ predicts fit the data better than $\widehat{Q_R}(\tau|x_2)$ predicts. At 75% quantile of the number of new cases daily $x_2 = 3442$, seems the $\widehat{Q_R}(\tau|x_2)$ is over predicted.

We extend the measure error method [18] to compute the Relative $R_N(\tau)$ of $\widehat{Q_N}(\tau|\mathbf{x})$ relative to $\widehat{Q_R}(\tau|\mathbf{x})$. It is measured by

$$\text{Relative } R_N(\tau) = 1 - \frac{V_N(\tau)}{V_R(\tau)},\ 0 \le R(\tau) \le 1, \text{ where} \tag{21}$$



(a) $\tau = 0.95$, when $x_1 = 7.8$



(b) $\tau = 0.99$, when $x_1 = 7.8$

Fig. 12 The 2D plots of the $\widehat{\mu}_{LS}(x_2)$ (green), $\widehat{Q_R}(\tau|x_2)$ (blue) and $\widehat{Q_N}(\tau|x_2)$ (red) quantile curves at **a** $\tau = 0.95$ and **b** $\tau = 0.99$ for the x_2- number of new cases vs. the y—number of hospitalized COVID-19 patients. Data is in black dots, $n = 103$

Table 7 The number of hospitalized COVID-19 patients (y) relative to the number of new cases (x_2) when the % positive test on last day $x_1 = 7.8(\%)$

x_2- number of new cases	$\widehat{\mu}_{LS}(x_2)$	$\tau = 0.95$		$\tau = 0.99$	
		$\widehat{Q_R}(\tau\|x_2)$	$\widehat{Q_N}(\tau\|x_2)$	$\widehat{Q_R}(\tau\|x_2)$	$\widehat{Q_N}(\tau\|x_2)$
1250	1396.77	**1658.91**	1315.33	**1595.70**	1430.16
1750	1483.33	**1785.86**	1541.83	**1757.36**	1657.57
2250	1569.90	1912.81	1634.02	1919.02	1766.55
2750	1656.46	2039.77	1998.27	2080.68	2102.94
3224 (75% quantile)	1857.83	2215.47	2095.45	2304.42	2167.54

Values of $\widehat{Q_R}(\tau|x_2)$ at x = 1250 and 1760 are bold, which indicate there are crossing errors at $\tau = 0.95$ and $\tau = 0.99$

Table 8 Relative $R_N(\tau)$ of $\widehat{Q_N}(\tau|\mathbf{x})$ relative to $\widehat{Q_R}(\tau|\mathbf{x})$ for example 1, n = 103

τ	0.95	0.96	0.97	0.98	0.99
Relative $R_N(\tau)$	0.3696	0.3545	0.3183	0.2644	0.1728

$$V_R(\tau) = \sum_{\substack{y_i \geq \boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}}(\tau) \\ i=1}}^{n} \frac{\tau}{n} \left| y_i - \boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}}(\tau) \right| + \sum_{\substack{y_i \geq \boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}}(\tau) \\ i=1}}^{n} \frac{1-\tau}{n} \left| y_i - \boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}}(\tau) \right| \tag{22}$$

where $\widehat{\boldsymbol{\beta}}(\tau)$ is obtained by Eq. (6) and

$$V_N(\tau) = \sum_{\substack{y_i \geq \widehat{Q_N}(\tau|\boldsymbol{x}_i) \\ i=1}}^{n} \frac{\tau}{n} \left| y_i - \widehat{Q_N}(\tau|\boldsymbol{x}_i) \right| + \sum_{\substack{y_i < \widehat{Q_N}(\tau|\boldsymbol{x}_i) \\ i=1}}^{n} \frac{1-\tau}{n} \left| y_i - \widehat{Q_N}(\tau|\boldsymbol{x}_i) \right|, \tag{23}$$

where $\widehat{Q_N}(\tau|\boldsymbol{x}_i)$ is obtained by Eq. (11).

Relative $R_N(\tau)$ given by Eq. (21) computes one minus the ratio of error losses between $V_N(\tau)$ and $V_R(\tau)$. If the value is greater than 0, it means the $\widehat{Q_N}(\tau|\boldsymbol{x}_i)$ model fits the data better than $\widehat{Q_R}(\tau|\boldsymbol{x}_i)$ model. Table 8 provides the relative $R_N(\tau)$ values for quantiles $\tau \in [0.95, 0.99]$.

Based on the results in Table 8, all relative $R_N(\tau)$ of $\widehat{Q_N}(\tau|\mathbf{x})$ relative to $\widehat{Q_R}(\tau|\mathbf{x})$ values are greater than 0 for various high quantiles $\tau \in [0.95, 0.99]$.

Above study shows that the proposed $\widehat{Q_N}(\tau|\mathbf{x})$ model fits the data better than $\widehat{Q_R}(\tau|\mathbf{x})$ model. The main conclusions for Example 1 analysis are:

1. The $\widehat{Q_N}(\tau|\mathbf{x})$ predictions of extreme values of the number of hospitalized COVID-19 patients are more accurate and realistic than the $\widehat{Q_R}(\tau|\boldsymbol{x}_i)$ method's predictions.
2. The $\widehat{Q_N}(\tau|\mathbf{x})$ estimates surfaces fit the data better than the $\widehat{Q_R}(\tau|\mathbf{x})$ planes for high conditional quantiles $\tau \in [0.95, 0.99]$.

This application of quantile regression for the y—number of hospitalized COVID-19 patients can help the government of Ontario allocate scarce medical resources during this pandemic to avoid overwhelming hospitals.

Example 2. Systolic Blood Pressures (January 2017–December 2018)

This section will revisit the NHANES 2017–2018 systolic blood pressures data from Sect. 1. We study the high conditional quantile curves of extreme high systolic blood pressures (SBP), using weight as regressor, as these variables have been closely linked with coronary heart disease. Normal systolic blood pressure is between 90 and 120 mmHg. For this example, SBP higher than 160 mmHg threshold is set to focus on subjects who are in stage 2 hypertension. These subjects have a high risk of having a heart attack, heart disease, stroke, brain problems, and kidney disease (Centers for Disease Control and Prevention, 2021).

Table 9 Test statistics and p-values of the goodness-of-fit tests using MLE for the y—systolic blood pressures data

	K–S	A-D	C-v-M
Test Statistic	0.1023	2.631	0.3051
p-value	0.0101	0.0424	0.1308

In Sect. 1, Example 2, an estimation was made using the least-squares mean regression method. Figure 5 shows the scatter diagram that compares the SBP (mmHg) with weight (kg) for the reduced data set of $n = 249$ subjects (after using a threshold of 160 mmHg with weights less than 125 kg). The least-squares regression line and linear quantile regression line are limited by linearity assumptions. In this Section, we use proposed new nonparametric quantile regression to study the relationship between the extreme value of SBP and weight.

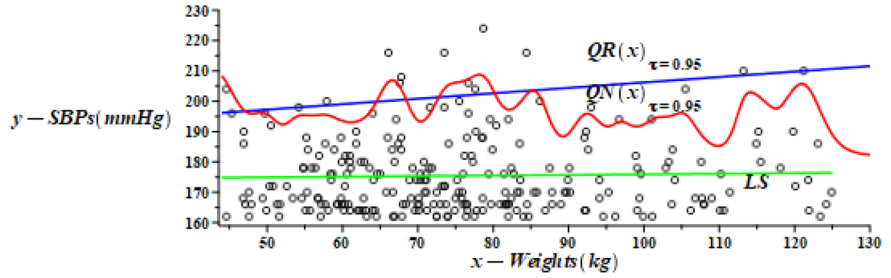
Similar as Example 1, we determine if the reduced data with SBP higher than 160 mmHg is from an extreme value distribution by checking whether it follows a generalized Pareto distribution. We fit the three-parameter GPD as given in Eq. (20) to the $n = 249$ data and use a threshold of 160 mmHg to obtain $\mu = 160$. The maximum likelihood estimates are $\widehat{\gamma}_{MLE} = -0.2279$ and $\widehat{\sigma}_{MLE} = 19.0375$. The three tests are the Kolmogorov–Smirnov test (K–S), Anderson–Darling (A-D) test, and Cramer-von-Mises test (C-v-M). Table 9 provide the test statistics and each tests' p-value.

Table 9 shows the p-value of the three tests. All three tests showed no evidence to reject hypothesis that the data is from a generalized Pareto distribution at $\alpha = 0.01$ significance level. The response y data in example 2 likely follows an estimated heavy-tailed GPD distribution. Thus, the high quantile regression is important for analysis this example's extremes.

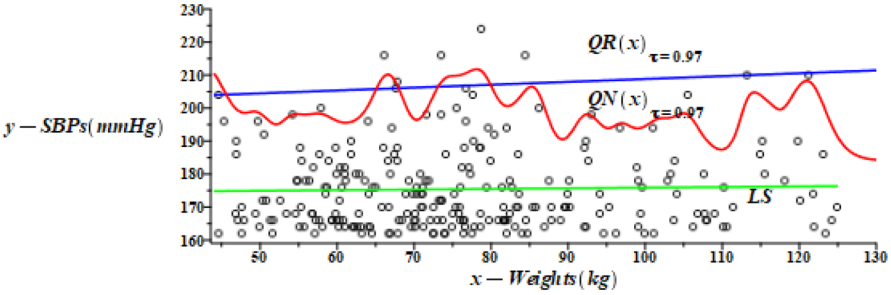
The assumption of a linear model is not required for the proposed nonparametric quantile regression method. Following steps given in Sect. 3, we first estimate the conditional c.d.f. $F(y|\boldsymbol{x})$. A Gaussian kernel and bandwidth of $h_{opt} = 8.5263$ is used. For the Nadaraya-Watson estimator, a Gaussian kernel and bandwidth $h = 3.5821$ is used. The proposed nonparametric $\widehat{Q_N}(\tau|x)$ quantile regression curves at $\tau = 0.95$ and 0.97 are provided in Fig. 13.

Figure 13 shows the LS lines $\widehat{\mu}_{LS}(x)$, QR lines $\widehat{Q_R}(\tau|x)$ and QN curves $\widehat{Q_N}(\tau|x)$ at $\tau = 0.95$ and 0.97. We observe that the high data pattern is followed closely by the $\widehat{Q_N}(\tau|x)$ curve. Thus, proposed $\widehat{Q_N}(\tau|x)$ predictions fit the data better than the $\widehat{Q_R}(\tau|x)$ predictions. For example, subjects around weight $x = 100$ kg have a relatively lower SBP than subjects with weight $x =$ 90–20 kg. The $\widehat{Q_N}(\tau|x)$ curves have a concave around $x = 100$ kg to represent this fact. Contrarily, the linear $\widehat{Q_R}(\tau|x)$ lines could not capture this scene and its predictions are higher than $\widehat{Q_N}(\tau|x)$'s predictions at weight $x = 100$ kg.

Table 10 provides those high conditional quantiles predictions for SBP (mmHg) given weight $x < 125$ kg. We can see that $\widehat{Q_R}(\tau|x)$ gives higher SBP predictions than $\widehat{Q_N}(\tau|x)$ predictions. Is $\widehat{Q_R}(\tau|x)$ over predicted? For subjects' weight $x > 90$ kg, the data shows lower SBP values, which leads us to think $\widehat{Q_N}(\tau|x)$ SBP predicts are more reasonable.



(a) $\tau = 0.95$



(b) $\tau = 0.97$

Fig. 13 Plot of the $\widehat{\mu}_{LS}(x)$ (green), $\widehat{Q_R}(\tau|x)$ (blue) and $\widehat{Q_N}(\tau|x)$ (red) at **a** $\tau = 0.95$ and **b** 0.97 for the y—SBP (mmHg) versus the x—weight (kg) (< 125 kg) for SBP greater than (160 mmHg). Data is in black dots, $n = 249$

Table 10 SBP y (mmHg) relative to weight x (kg) (< 125 kg) estimates $\widehat{\mu}_{LS}$, $\widehat{Q_R}(\tau|x)$ and $\widehat{Q_N}(\tau|x)$

x Weight (kg)	$\widehat{\mu}_{LS}$	$\tau = 0.95$		$\tau = 0.97$	
		$\widehat{Q_R}(\tau\|x)$	$\widehat{Q_N}(\tau\|x)$	$\widehat{Q_R}(\tau\|x)$	$\widehat{Q_N}(\tau\|x)$
60	175.14	199.04	193.52	205.34	196.11
70	175.33	200.83	193.63	206.21	196.63
80	175.52	202.62	203.56	207.08	206.69
90	175.71	204.41	189.35	207.95	191.47
100	175.90	206.20	192.37	208.82	194.56

Use function (21) to compute the Relative $R_N(\tau)$ of $\widehat{Q_N}(\tau|x)$ relative to $\widehat{Q_R}(\tau|x)$ for compare the nonparametric $\widehat{Q_N}(\tau|x)$ and regular linear quantile regression $\widehat{Q_R}(\tau|x)$ methods. (see Table 11).

Base on the graphs provided in Fig. 13 and all Relative $R_N(\tau)$ of $\widehat{Q_N}(\tau|x)$ relative to $\widehat{Q_R}(\tau|x)$ in Table 11 values are all greater than 0 for various high quantiles $\tau \in [0.95, 0.99]$. We can conclude that the proposed $\widehat{Q_N}(\tau|x)$ model fits the Systolic Blood Pressures data with a threshold of 160 (mmHg) when weight $x < 125$ kg, $n = 249$

Table 11 Relative $R_N(\tau)$ of $\widehat{Q_N}(\tau|x)$ relative to $\widehat{Q_R}(\tau|x)$ for the systolic blood pressures data, $n = 249$

τ	0.95	0.96	0.97	0.98	0.99
Relative $R_N(\tau)$	0.3228	0.3194	0.3125	0.3157	0.2968

data better than the regular $\widehat{Q_R}(\tau|x)$ linear quantile regression model. Following are the main results for Example 2 analysis are:

1. Figure 13 shows that the $\widehat{Q_N}(\tau|x)$ predictions fit the data better than $\widehat{Q_R}(\tau|x)$ predictions as all the Relative $R_N(\tau)$ of $\widehat{Q_N}(\tau|x)$ relative to $\widehat{Q_R}(\tau|x)$ are positive.
2. The proposed $\widehat{Q_N}(\tau|x)$ predictions result captured the fact that there is a higher chance for people whose weights are between 70 and 90 kg to have high SBPs. On the other hand, the $\widehat{Q_R}(\tau|x)$ method assumes the relationship between weight and SBP is linear. As a result, its model could not extract the same information.

This application of quantile regression for extreme SBP can be used to identify patients with high risk to have stage 2 hypertension and prevent coronary heart disease.

7 Overall Conclusions and Discussion

7.1 Conclusions

When researchers are interested in extreme events and need to estimate high conditional quantiles, the traditional least-squares estimator used for estimating the conditional mean is not suitable since the extreme events usually follow the heavy-tailed distribution. On the other hand, the quantile regression estimator uses a quantile-weighted L_1 - loss function and therefore can estimate the high quantiles.

The regular linear quantile regression method (QR) estimates high quantile curves with a linear model. However, very often, the parametric conditions can not be met in real-world data. In this situation, we should use the nonparametric quantile regression methods to obtain better estimates. In this paper, we proposed a new direct nonparametric quantile regression estimation method (QN). Three studies were performed to compare the proposed method with the regular quantile regression method to check the new method's capability:

1. In Sect. 4, mathematical properties of the proposed direct nonparametric quantile regression method (QN) were examined.
2. In Sect. 5, the Monte Carlo simulation was performed to compare the efficiencies of the QN and QR. The QN had better simulation efficiencies than QR.
3. In Sect. 6, the QR and QN were used in two applications. The QN had better fits to the data than the QR. The QN method avoided the quantile curve crossing problem of the QR method.

To conclude, the proposed direct nonparametric quantile regression estimator outperforms the regular quantile regression method when estimate high quantile curves of heavy-tailed distributed data. Several recommendations for future research are given

take different approach to estimate the c.d.f., use cross validation procedure to find optimal smoothing parameters, and implement the program in other languages with better algorithm and data structures to improve the program's execution efficiency.

7.2 Discussions: Comparing with Other Quantile Regression Methods

As we mention in Sect. 1. This Section will explore Bayesian quantile regression (B-QR) method. At first, we explore the Laplace likelihood B-QR by using Markov Chain Monte Carlo (MCMC) with Metropolis–Hastings algorithm. [12, 21, 30, 31]. We apply this MCMC method to Example 1 in this paper. We use [31] Yu and Moyeed (2001) R package, the results are in Fig. 14.

Next, we do comparison of MCMC B-QR results to the propose direct nonparametric quantile Regression method on Example 1 Hospitalised COVID-19 Hospitalized Patients.

Bayesian Quantile regression by MCMC method in Fig. 14 is given by

$$\widehat{Q_B}(0.95|x_1,\, x_2) = \widehat{\beta}_0(0.95) + \widehat{\beta}_1(0.95)x_1 + \widehat{\beta}_2(0.95)x_2 = 577.531 + 113.893x_1 + 0.2190x_2.$$

Figure 15 show comparisons of 2D B-QR with other quantile regression curves in Example 1 Hospitalized COVID-19 Patients in Sect. 6 Figs. 11a and 12a results at $\tau = 0.95$ level, where

$$\widehat{Q_R}(0.95|x_1)\big|_{x_2=3442} = 1514.521 + 89.8649x_1, \text{ when } x_2 = 3442 \text{ (the 0.75th quantile of } x_2)$$

$$\widehat{Q_B}(0.95|x_1)\big|_{x_2=3442} = 1331.329 + 113.893x_1, \text{ when } x_2 = 3442 \text{ (the 0.75th quantile of } x_2)$$

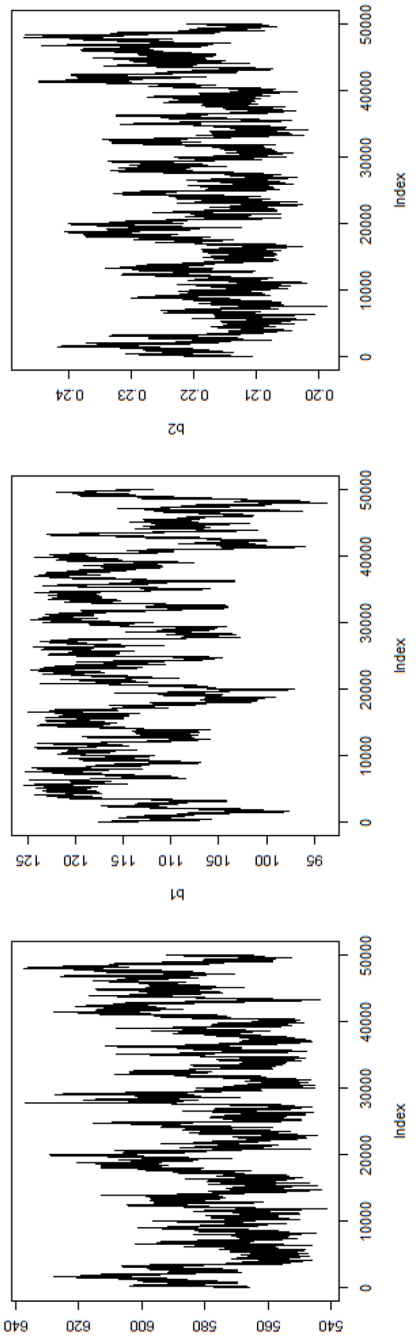
$$\widehat{Q_R}(0.95|x_2)\big|_{x_1=7.8} = 1341.5277 + 0.2539x_2, \text{ when } x_1 = 7.8 \text{ (the 0.75th quantile of } x_1)$$

$$\widehat{Q_B}(0.95|x_2)\big|_{x_1=7.8} = 1463.8964 + 0.2190x_2, \text{ when } x_1 = 7.8 \text{ (the 0.75th quantile of } x_1)$$

Note that $\widehat{Q_N}(0.95|x_1)$ and $\widehat{Q_N}(0.95|x_2)$ are not linear, in a direct nonparametric form.

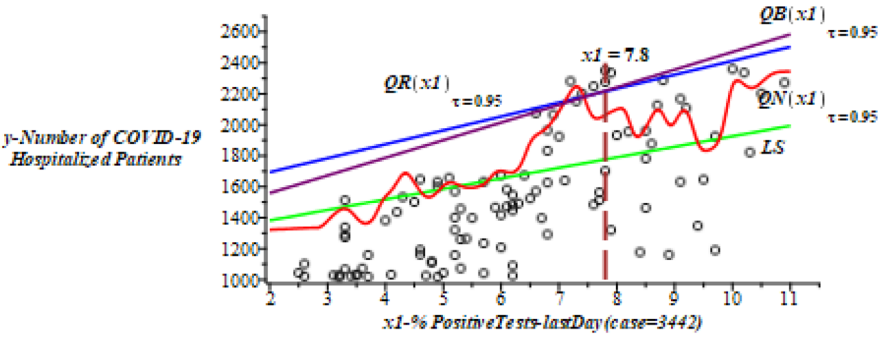
We note that in Fig. 15, Bayesian $\widehat{Q_B}(0.95|x_1)$ and $\widehat{Q_B}(0.95|x_2)$ curves (in purple) are close to $\widehat{Q_R}(0.95|x_1)$ and $\widehat{Q_R}(0.95|x_2)$ curves (in blue), respectively. These are reasonable since they are set as linear parametric models. We may conclude that:

1. We notice that the Bayesian quantile regression using parametric Laplace likelihood by MCMC Metropolis-Hasting algorithm obtains the estimated quantile regression to be linear which are close to the regular linear quantile regression curves. The MCMC method heavily depends on the Markov chain stationary distribution and M–H algorithm proposal distribution. Further studies may need.
2. The proposed direct nonparametric quantile regression represents the data pattern well. In the future, we may further study nonparametric Bayesian quantile regression methods.
3. There are other developing quantile regression methods. We will continue to explore and study them to improve current existing quantile regression methods.

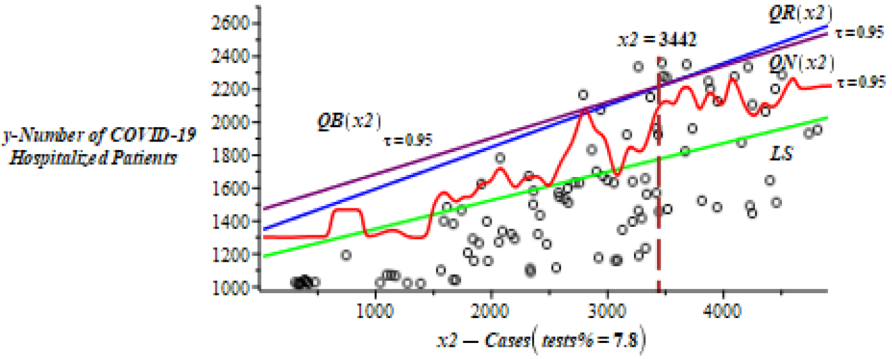



```
> output_summary

Type of dependent variable: continuous
Lasso variable selection: yes
Normal approximation of posterior: yes
Estimated quantile:  0.95
Lower credible bound:  0.025
Upper credible bound:  0.975
Number of burnin draws:  5000
Number of retained draws:  50000

Summary of the estimated beta:

            Bayes Estimate   lower   upper adj.lower adj.upper
(Intercept)        577.531 551.439 614.277  -7789.35   8944.41
x1                 113.893 100.946 122.962  -3352.26   3580.05
x2                   0.219   0.207   0.237     -6.52      6.96
```

Fig. 14 Bayesian Quantile Regression using the MCMC by Metropolis–Hastings algorithm. We draws 55,000 samples, and then burned away the first 5000 samples before getting the estimates: $\hat{\beta}_0(\tau = 0.95) = 577.531$; $\hat{\beta}_1(\tau = 0.95) = 113.893$; $\hat{\beta}_2(\tau = 0.95) = 0.2190$



(a) Bayesian Quantile Regression $\widehat{Q_B}\left(0.95\,|\,x_1\right)$ in purple colour.



(b) Bayesian Quantile Regression $\widehat{Q_B}\left(0.95\,|\,x_2\right)$ in purple colour.

Fig. 15 MCMC Bayesian quantile regression curves in purple. **a** $\widehat{Q_B}(0.95|x_1)$. **b** $\widehat{Q_B}(0.95|x_2)$

Acknowledgements We are grateful for the comments and suggestions of the Editor and reviewers. They have helped us to improve the paper.

Funding This research is supported by NSERC (Natural Science and Engineering Council of Canada) grant numbers RGPIN-2022-04799-MLH and RGPIN-2019-05418-WM.

Declarations

Conflict of interest The authors declare no conflict of interest.

References

1. Anderson TW (1962) On the distribution of the two-sample Cramer-von Mises criterion. Ann Math Stat 33(3):1148–1159. https://doi.org/10.1214/aoms/1177704477
2. Anderson TW, Darling DA (1952) Asymptotic theory of certain “goodness-of-fit" criteria based on stochastic processes”. Ann Math Stat 23:193–212
3. BASS URGENT CARE (2019) The 4 stages of hypertension. Retrieved from https://www.bassadvancedurgentcare.com/post/the-4-stages-of-hypertension

4. Center for Disease Control and Prevention (2017) SARS basics fact sheet. Retrieved from https://www.cdc.gov/sars/about/fs-sars.html.
5. Centers for Disease Control and Prevention (2020) NHANES 2017–2018. Retrieved from CDC.gov: https://wwwn.cdc.gov/nchs/nhanes/Search/DataPage.aspx?Component=Examination&CycleBeginYear=2017.
6. Das K, Sams C, Singh V (2016) Characterization of tail of river flow data by generalized pareto distribution. J Statist Res 48–50(2):55–70
7. de Haan L, Ferreira A (2006) Extreme value theory: an introduction. Springer, New York
8. de Zea Bermudez P, Kotz S (2010) Parameter estimation of the generalized Pareto distribution - Part I. J Statist Plann Inference 140(6):1353–1373
9. Dey A, Edwards A, Das K (2020) Determinants of high crude oil price: a nonstationary extreme value approach. J Statist Theory Practice 14(1):1–14
10. Fisk PR (1961) The graduation of income distributions. Econometrica 29(2):171–185
11. Fukunaga K (1972) Introduction to statistical pattern recognition. Academic Press, New York
12. Gilks WR, Richardson EM, Spiegelhalter DJ (1996) Markov chain monte carlo in practice. Chapman & Hall, Cambridge, pp 1–19
13. Government of Ontario (2021) Status of COVID-19 cases in Ontario. Retrieved from Ontrio.ca: https://data.ontario.ca/dataset/status-of-covid-19-cases-in-ontario
14. Government of Canada (2021) Regulatory Decision Summary - AstraZeneca COVID-19 Vaccine - Health Canada. Retrieved from Canada.ca: https://covid-vaccine.canada.ca/info/regulatory-decision-summary-detailTwo.html?linkID=RDS00772&pType=rds&lang=en
15. High Blood Pressure Symptoms, Causes, and Problems, cdc.gov (2020) Retrieved from https://www.cdc.gov/bloodpressure/about.htm#:~:text=High%20blood%20pressure%2C%20also%20called,blood%20pressure%20(or%20hypertension).
16. Harvard Health Publishing (2021) Treatments for COVID-19. Retrieved from https://www.health.harvard.edu/diseases-and-conditions/treatments-for-covid-19
17. Hosking JRM, Wallis JR (1987) Parameter and quantile estimation for generalized Pareto distribution. Technometrics 29(3):339–349
18. Huang M, Xu X, Tashnev D (2015) A weighted linear quantile regression. J Stat Comput Simul 85(13):2596–2618
19. Islam T, Das K (2021) Predicting Bitocoin return using extreme value theory. Am J Math Manag Sci 40(2):177–180
20. Koenker R (2005) Quantile regression. Cambridge University Press, Cambridge
21. Koenker R, Chernozhukov V, He X, Limin Peng (2018) Handbook of quantile regression. Chapman & Hall/CRC, Cambridge
22. Kolmogorov AN (1933) Sulla determinations empirica di una legge di distribuzione. Giornale dell'Istituto Italiano degli Attuari 4:83–91
23. National High Blood Pressure Education Program (2003) The seventh report of the joint national committee on prevention, detection, evaluation, and treatment of high blood pressure. National Heart, Lung, and Blood Institute, Bethesda, MD
24. Pickands J (1975) Statistical inference using extreme order statistics. Ann Stat 3(1):119–131
25. Phua J, Weng L, Ling L, Egi M, Lim C-M, Divatia JV, Shrestha BR, Arabi YM, Ng J, Gomersall CD, Nishimura M, Koh Y, Du B (2020) Intensive care management of coronavirus disease 2019 (covid-19): challenges and recommendations. Lancet Respir Med 8(5):506–517
26. Scott DW (2015) Multivariate density estimation: theory, practice, and visualization. Wiley, Hoboken
27. Silverman BW (1986) Density estimation for statistics and data analysis. Chapman & Hall, London
28. Takezawa K (2006) Introduction to nonparametric regression. Wiley-Interscience, Hoboken, NJ
29. Whelton PK, Carey RM, Aronow WS, Casey DE, Collins KJ, Dennison Himmelfarb CD, Wright JT (2018) 2017 ACC/AHA/AAPA/ABC/ACPM/AGS/apha/ash/ASPC/NMA/PCNA guideline for the prevention, detection, evaluation, and management of high blood pressure in adults. J Am College of Cardiol 71(19):e127–e248. https://doi.org/10.1016/j.jacc.2017.11.006
30. Yan Y, Wang KH, He X (2016) Posterior inference in bayesian quantile regression with asymmetric laplace likelihood. Int Statist Rev 84(3):327–344

31. Yu K, Moyeed RA (2001) Bayesian quantile regression. Statist Probab Lett 54:437–447