Quantile Regression on COVID-19

Andrew Kenig and Mei Ling Huang*

Department of Mathematics & Statistics, Brock University, Canada

September 1, 2021

Abstract

The COVID-19 pandemic is an extreme disaster for the world during 2020-2021. Predicting the future number of daily deaths is an important task. Quantile regression is a statistical tool to estimate conditional quantiles. Our goal is obtaining accurate prediction on extreme conditional quantiles of the daily deaths. The regular quantile regression method often sets a linear model with estimating the coefficients to obtain the estimated conditional quantile. That approach may be restricted by the model setting, and has computational difficulties. To overcome these difficulties, this paper proposes a nonparametric quantile regression method with a five-step algorithm. Monte Carlo simulations show good efficiency for the proposed nonparametric quantile regression relative to the regular linear quantile regression. This paper studies Ontario, Canada COVID-19 pandemic example by using the proposed method. Comparisons of the proposed method with existing methods are given.

Keywords: *Conditional quantile, COVID-19, extreme value distribution, Burr distribution, generalized Pareto distribution, nonparametric regression.*

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

1. Introduction

The COVID-19 pandemic in Ontario, Canada is an ongoing viral pandemic of coronavirus disease 2019 (COVID-19). The first confirmed case of COVID-19 in Canada was announced on January 25, 2020, with increasing transmission province-wide, a state of emergency was declared by Premier Doug Ford on March 17, 2020. From late spring to early summer, the majority of the deaths were residents of long-term care homes. From May through August 2020, the province instituted a three-stage plan to lift economic restrictions, subject to the employment of social distancing and other guidelines, and continued restrictions on the sizes of gatherings. The state of emergency was lifted on July 24, 2020. A plan was implemented for the return-to-class of public schools, involving more than 2 million children. In early September 2020, the province showed

*Corresponding author. E-mail: mhuang@brocku.ca.

a significant increase in new cases, along with similar spikes in provinces across the country. Nation-wide cases, hospitalizations and deaths spiked preceding and following the Christmas and holiday season in December 2020 and January 2021. Alarmed by hospital capacity issues, fatalities and new cases, heavy restrictions (such as lockdowns and curfews) were put in place in affected areas and across the country. These lockdowns resulted in active cases beginning to steadily decline, reaching a plateau in active cases in mid-February 2021. During a third wave of the virus, cases began rising across Canada in mid-March. However, in late April, the third wave had spread throughout the province. A province-wide shutdown beginning Boxing Day was lifted February 10, 2021. In mid-March 2021, the Ontario Hospital Association, Ontario's COVID-19 scientific advisory table, and Ontario's Chief Medical Officer of Health declared the province was experiencing a third wave of the virus.

The COVID-19 pandemic is an extreme disaster for the world during 2020-2021. Predicting the future number of daily deaths is an important task. The scientists search for the best mathematical model which applies to this event. In the literature, extreme events occur in many fields such as financial markets, natural disasters, industrial engineering and others. The extreme values often follow a heavy tailed distribution. When statisticians are interested in estimating high quantiles of heavy-tailed distributions of extreme events, they often face theoretical difficulties in doing so. It is important to estimate extreme conditional quantiles of a random variable y given a variable vector $\mathbf{x} = (x_1, x_2, ..., x_d)^T$, and we let $\mathbf{x}_p = (1, x_1, x_2, ..., x_d)^T \in R^p$, $p = d+1$. First, we will review the mean regression and linear quantile regression models:

The mean linear regression model assumes

$$\mu_{y|\mathbf{x}} = E(y|x_1, x_2, ..., x_d) = \mathbf{x}_p^T \boldsymbol{\beta} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_d x_d. \quad (1)$$

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

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

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

The mean linear regression provides the mean relationship between a response variable and explanatory variables. We are interested in estimating the conditional quantiles of y given $\mathbf{x}$.

$$Q_Y(\tau|\mathbf{x}) = \inf(t : F_Y(t|\mathbf{x}) \geq \tau\} = F_Y^{-1}(\tau|\mathbf{x}),\ 0 < \tau < 1. \quad (3)$$

Koenker and Bassett (1978) proposed a linear quantile regression model for estimating true conditional quantiles in *(3)*. It is defined as

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

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

In model *(4)*, we estimate the coefficient $\boldsymbol{\beta}(\tau) = (\beta_0(\tau), \beta_1(\tau), \beta_2(\tau), ..., \beta_d(\tau))^T \in R^p$ from a random sample $\{(y_i, \mathbf{x}_i),\ i = 1, ..., n\}$ by using an L_1- loss function to obtain estimator $\widehat{\boldsymbol{\beta}}(\tau)$,

$$\widehat{Q}_L(\tau|\mathbf{x}) = \mathbf{x}_p^T \widehat{\boldsymbol{\beta}}(\tau) = \widehat{\beta}_0(\tau) + \widehat{\beta}_1(\tau) x_1 + \cdots + \widehat{\beta}_d(\tau) x_d,\ 0 < \tau < 1, \quad (5)$$

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

where ρ_τ is a loss function, namely

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

We are motivated to study the Ontario, Canada COVID-19 event. In recent years, many studies have focussed on improvements of estimators $\mu_{y|\mathbf{x}}$ *(1)* and $\widehat{Q}_L(\tau|\mathbf{x})$ *(5)* for estimating extreme conditional quantiles. Let us apply them to the Ontario COVID-19 data.

Example. Ontario, Canada COVID-19 (March 17, 2020 - May 31, 2021)

During March 17, 2020 - May 31, 2021, the provine of Ontario, Canada was highly affected by the COVID-19 pandemic with 3 dangerous waves (see Figure 1) based on the Government of Ontario Website data (https://data.ontario.ca/dataset/status-of-covid-19 cases-in-ontario). Daily deaths is a major indicator for the damage of the COVID-19 pandemic, which relates to hospital capacity, Health Agents management and Government policy for controling disease spread and saving people. We observed that the Ontario COVID-19 pandemics first wave's highest death day is May 4, 2020 with 84 deaths; the second wave's highest death day is January 14, 2021 with 100 deaths; the third wave's highest death day is May 9, 2021 with 47 deaths.

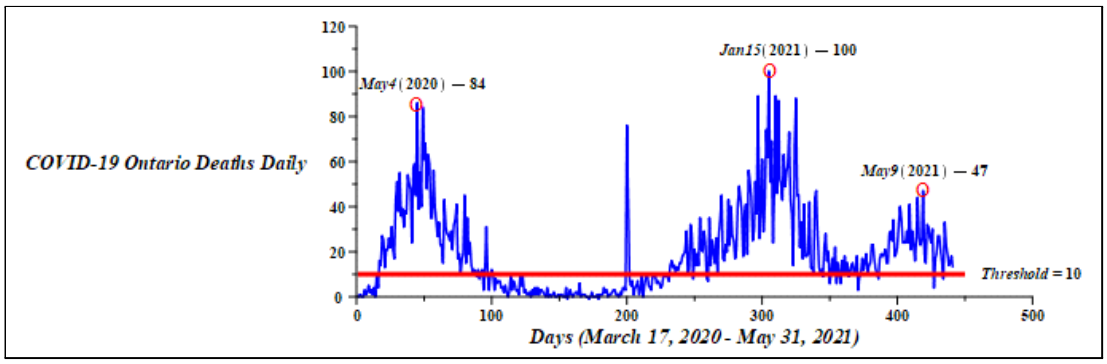


Figure 1. The Ontario COVID-19 daily deaths during March 17, 2020 - May 31st, 2021 ordered by date, $n^* = 441$ days, and a threshold of 10 daily deaths (red).

We focus on the main variable $y-$ the daily COVID-19 deaths which is related to x_1- the daily COVID-19 tests and x_2- the daily new COVID-19 cases. We observed that if daily deaths is less or equal to 10 people, then the pressure for hospitals or public virus spread is controllable on that day. So we chose a threshold of 10 daily deaths (as a red line in Figure 1) for this data set. We only analyze daily deaths over 10. Thus the new data set size is $n = 267$ days. We are interested if y-daily deaths are related to x_1- COVID-19 tests and x_2-daily new COVID-19 cases. We apply the mean regression model in *(1)* and the linear quantile regression model in *(4)* to this data set, the results are as follows.

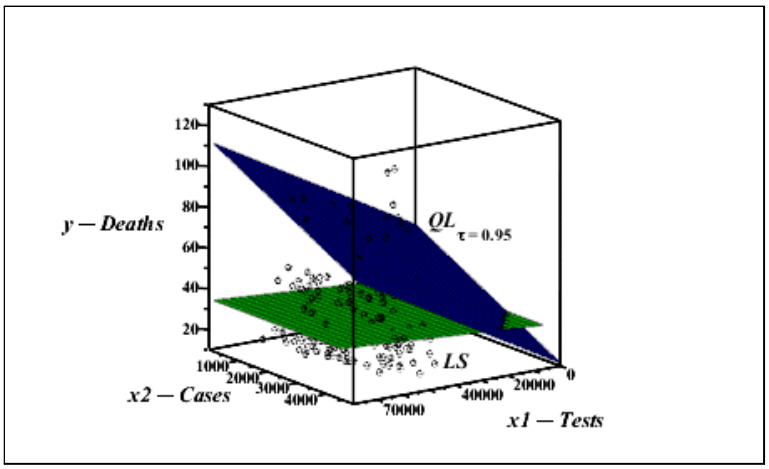
Based on the $n = 267$ days deaths greater than 10, we perform a daily deaths relative to daily tests and daily new cases mean regression model *(1)*, the least squares estimate is

$$\mu_{LSx_1,x_2} = \widehat{\mu}_{y|x_1,x_2} = 27.0494 + 0.00008x_1 - 0.00005x_2,$$

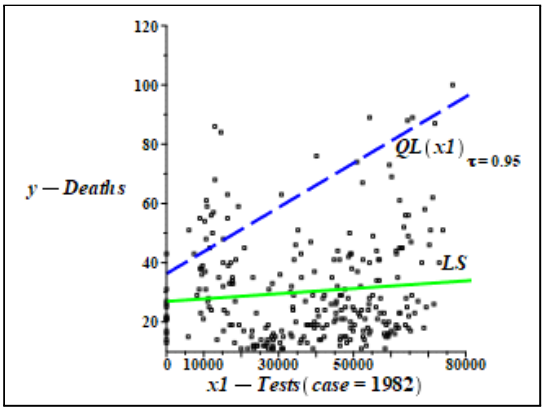
where y represents the response variable daily deaths, and x_1 represents the daily tests and x_2 represents the daily new cases.

The linear quantile regression estimate $\widehat{Q}_L(0.95|x_1, x_2)$ for $\tau = 0.95$ in *(5)* is

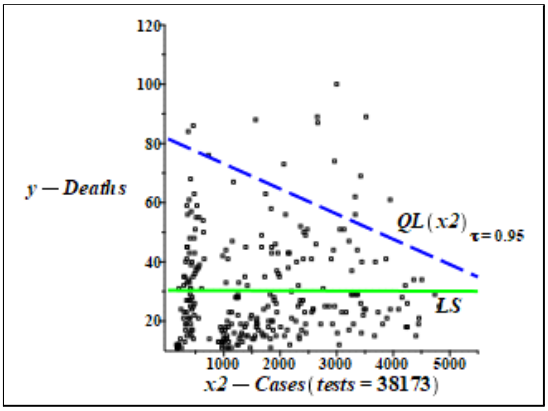
$$\begin{aligned} \widehat{Q}_L(0.95|x_1, x_2) &= 53.1492 + 0.00075x_1 - 0.0085x_2. \\ \widehat{Q}_L(0.95|x_1,\ when\ x_2 &= \overline{x_2} = 1982) = 36.3025 + 0.00075x_1. \\ \widehat{Q}_L(0.95|x_2,\ when\ x_1 &= \overline{x_1} = 38173) = 81.6522 - 0.0085x_2. \end{aligned}$$



(a) 3-D μ_{LSx_1,x_2} and $\widehat{Q}_L(0.95|x_1, x_2)$.



(b) 2-D μ_{LSx_1} and $\widehat{Q}_L(0.95|x_1)$, $n = 267$.



(c) 2-D μ_{LSx_2} and $\widehat{Q}_L(0.95|x_2)$, $n = 267$.

Figure 2. For Ontario COVID-19 data, after threshold 10, $n = 267$ days, scatter plot (black dots), (a) 3-D $\mu_{LS-x1,x2}$ (green) and $\widehat{Q}_L(0.95|x_1, x_2)$ (blue) regression surfaces (b) 2-D μ_{LS-x1} (green) line and $\widehat{Q}_L(0.95|x_1$, when $x_2 = \overline{x_2} = 1982$ is the cases avrage) regression line (blue dash). (c) 2-D μ_{LS-x2} (green) line and $\widehat{Q}_L(0.95|x_2$, when $x_1 = \overline{x_1} = 38173$ is the tests avrage) regression line (blue dash).

Figure 2 shows the 3-D and 2-D mean regression μ_{LS} and the $\tau = 0.95$ quantile regression $\widehat{Q}_L$ surfaces or curves. Note that both μ_{LS} and $\widehat{Q}_L$ planes and lines do not catch the extreme daily deaths data well. Since the mean LS regression only estimates the mean of the COVID-19 daily deaths in Ontario, which does not represent the extreme values of the COVID-19 daily deaths. Also even the linear quantile regression $\widehat{Q}_L$ in *(5)* has improvement over μ_{LS}, but it is also restricted by the linear model setting and does not catch the data pattern.

In recognition of these complications for estimating extreme conditional quantiles, we will study this example by implementing the new proposed nonparametric quantile regression method which is more flexible then regular linear quantile regression. The nonparametric quantile regression has been used (Yu, et al., 2003; Huang and Nguyen, 2018). The nonparametric quantile regression is based on the nonparametric kernel method which avoids the quantile curve crossing problem. This paper's contributions are:

a) Improve the existing kernel estimation method by using the proposed 5-step algorithm to estimate extreme conditional quantile curves;

b) Perform Monte Carlo simulation to confirm the proposed estimator is more efficient relative to the linear quantile regression method;

c) Using the proposed method to the Ontario COVID example to get more reasonable results.

In Section 2, we propose a nonparametric quantile regression estimator with a 5-step algorithm. In Section 3, the results of Monte Carlo simulations generated from Burr distribution of Type XII (Burr, 1942) show that the proposed nonparametric quantile regression produces high efficiencies relative to existing linear quantile regression. A relative measure of comparing goodness of fit for these two models is given in Section 4. In Section 5, we study Canada Ontario COVID-19 example by using three methods: μ_{LS}, linear quantile regression and proposed nonparametric quantile regression. The simulations and the example illustrate that the proposed nonparametric quantile regression model fits the data set better than the linear quantile regression method.

2. Proposed Nonparametric Quantile Regression

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

$$\widehat{Q}_Y(\tau|\mathbf{x}) = \inf(t : \widehat{F}_Y(t|\mathbf{x}) \geq \tau\} = \widehat{F}_Y^{-1}(\tau|\mathbf{x}),\ 0 < \tau < 1,$$

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

2.1. Extreme Value Distribution

Let $X_1, X_2, ..., X_n$ be independent and identically distributed (i.i.d.) continuous random variables with common cumulative distribution function (c.d.f.) $F(x)$. It is important to study the limiting behavior of the sample maxima and minima, defined as $\max(X_1, X_2, ..., X_n)$ and $\min(X_1, X_2, ..., X_n)$, respectively. The main interest of extreme value theory (EVT) is in finding possible limiting distributions of the sample maxima of i.i.d. random variables. Any nondegenerate distribution that can be derived as such a limit is called an *extreme value distribution* (Haan and Ferreira, 2006).

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

$$G_\gamma(x) = \begin{cases} 1-\exp(-(1+\gamma x)^{-1/\gamma}), & 1+\gamma x > 0 \text{ and } \gamma \neq 0; \\ 1-\exp(-e^{-x}), & \gamma = 0, \end{cases} \tag{6}$$

where the parameter γ is called the extreme value index (EVI).

Note that when $\gamma > 0$, the corresponding densities for both $G_\gamma(x)$ and $G_\gamma(ax+b)$ have heavier tails than the exponential distribution, which are referred to as *heavy tailed distributions*. In many applications, it is important to include observations that take extremely high or low values in the statistical analysis.

A limit conditional extreme value distribution for exceeding a threshold has a *generalized Pareto distribution* (GPD).

Definition 2. *(Pickands, 1975) The c.d.f. $H_\gamma(x)$ and its corresponding probability density function (p.d.f.) $f(x)$ of the two-parameter GPD(γ,σ) with shape parameter $\gamma \neq 0$ and scale parameter σ of a random variable X are given by*

$$H_\gamma(x) = 1-\left(1+\gamma\frac{x}{\sigma}\right)^{-1/\gamma}, \quad \gamma \neq 0, \quad \sigma > 0, \quad x > 0. \tag{7}$$

2.2. Propose Nonparametric Quantile Regression

We build the following steps to construct a direct nonparametric quantile regression estimator:

Step 1: Kernel Estimation for conditional c.d.f.: Estimate the conditional c.d.f. $F(y|\mathbf{x})$ of y for given $\mathbf{x} = (x_1, x_2, ..., x_d)$ using kernel estimation method (Scott, 2015)

$$\widehat{F}_Y(y|\mathbf{x}) = \frac{\sum_{i=1}^{n} I(Y_i \leq y\} K_h(\mathbf{x}-\mathbf{X}_i)}{\sum_{i=1}^{n} K_h(\mathbf{x}-\mathbf{X}_i)},$$

where $I(Y_i \leq y\}$ is an indicator function, and

$$K_h(\mathbf{x}-\mathbf{X}_i) = \frac{1}{h^d} K\left(\frac{\mathbf{x}-\mathbf{X}_i}{h}\right),$$

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

Step 2. Normalization: Normalize the estimated conditional c.d.f. $\widehat{F}_Y(y|\mathbf{x}) \in [0,1]$ to a true c.d.f. function.

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

$$\widehat{\xi}(\tau|\mathbf{x}) = \widehat{Q}_Y(\tau|\mathbf{x}) = \inf\{y : \widehat{F}_Y(y|\mathbf{x}) \geq \tau\} = \widehat{F}_Y^{-1}(\tau|\mathbf{x}). \tag{8}$$

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

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

Step 4: Nonparametric Regression based on local estimates in Step 3. We obtain a nonparametric quantile regression estimator for the τth conditional quantile curve of $\mathbf{x}$ by using Nadaraya-Watson (NW) nonparametric regression estimator on $\left(\mathbf{x}_i, \widehat{\xi}_i(\tau|\mathbf{x}_i)\right)$, $i = 1, 2, ..., n$,

$$\widehat{Q}_N(\tau|\mathbf{x}) = \widehat{\xi}(\tau|\mathbf{x}) = \frac{\sum_{i=1}^{n} K\left\{\frac{\mathbf{x}-\mathbf{X}_i}{\mathbf{h}}\right\} \widehat{\xi}_i(\tau|\mathbf{x}_i)}{\sum_{j=1}^{n} K\left\{\frac{\mathbf{x}-\mathbf{X}_j}{\mathbf{h}}\right\}} = \sum_{i=1}^{n} W_h(\mathbf{x}, \mathbf{X}_i)\widehat{\xi}_i(\tau|\mathbf{x}_i), \quad 0 < \tau < 1, \qquad (9)$$

where $W_i(\mathbf{x})$ is called an equivalent kernel,

$$W_h(\mathbf{x}, \mathbf{X}_i) = \frac{K\left\{\frac{\mathbf{x}-\mathbf{X}_i}{\mathbf{h}}\right\}}{\sum_{j=1}^{n} K\left\{\frac{\mathbf{x}-\mathbf{X}_j}{\mathbf{h}}\right\}}, \quad i = 1, 2, ..., n,$$

where

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

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

Step 5: Goodness of fit test: Check that the response variable y is heavy tailed distributed.

3. Monte Carlo Burr Simulation

Similarly. we generate $m = 1,000$ random samples with size $n = 500$ each from one-dimensional random variables X ($d = 1$) uniformly distributed on $E = [0, 1]$. And Y given $X = x$ is Burr distribution of Type XII (Burr, 1942); the conditional c.d.f. is

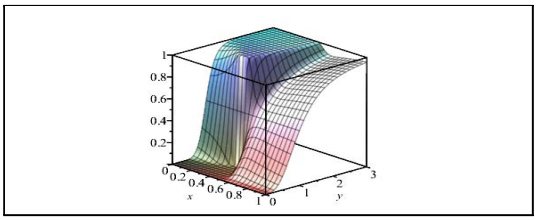
$$F_{Burr}(y|x) = 1 - \frac{1}{1 + y^{1/\gamma(x)}}, \quad 0 \leq x \leq 1, \quad \gamma(x) > 0, \quad y > 0. \qquad (10)$$

with the conditional tail index given by

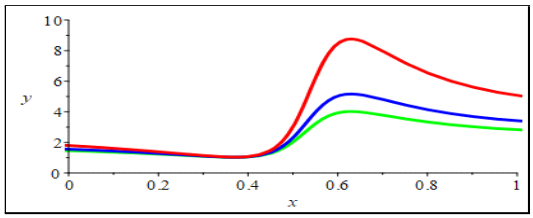
$$\gamma_{Burr}(x) = \frac{3}{100}\left(\frac{120x^2 - 90x + 17}{15x^2 - 15x + 4}\right), \quad 0 \leq x \leq 1.$$

The true conditional quantile of *(10)* is

$$Q_{Y_{Burr}}(\tau|x) = \left(\frac{\tau}{1-\tau}\right)^{\gamma(x)}, \quad 0 \leq x \leq 1, \quad 0 < \tau < 1, \quad \gamma(x) > 0. \qquad (11)$$



(a) The joint Burr and uniform distribution.



(b) The true quantile function.

Figure 3. (a) The joint c.d.f. of Burr distribution of y and uniform distribution of x. (b) The true conditional quantiles at $\tau = 0.99$ (red), $\tau = 0.97$ (blue), $\tau = 0.95$ (green).

We use two conditional quantile estimators $\widehat{Q}_N(\tau|x)$ in *(9)* and $\widehat{Q}_L(\tau|x)$ in *(5)* to estimate the true conditional quantile $Q_{Y_{Burr}}(\tau|x)$ in *(11)*. For each method, we generate size $n = 500$, $m = 1,000$ samples. $\widehat{Q}_{N,i}(\tau|x)$ and $\widehat{Q}_{L,i}(\tau|x)$, $i = 1, 2, ..., m$, are estimated in the ith sample. The simulation mean squared errors (SMSE) and bias (SBIAS) of the two estimators *(9)* and *(5)* for $0 < \tau < 1$ are:

$$SMSE\left(\widehat{Q}_N(\tau)\right) = \frac{1}{m}\sum_{i=1}^{m}\int_0^1 \left(\widehat{Q}_{N,i}(\tau|x) - Q_Y(\tau|x)\right)^2 dx; \quad (12)$$

$$SMSE\left(\widehat{Q}_L(\tau)\right) = \frac{1}{m}\sum_{i=1}^{m}\int_0^1 \left(\widehat{Q}_{L,i}(\tau|x) - Q_Y(\tau|x)\right)^2 dx; \quad (13)$$

$$SBIAS\left(\widehat{Q}_N(\tau)\right) = \frac{1}{m}\sum_{i=1}^{m}\int_0^1 \left|\widehat{Q}_{N,i}(\tau|x) - Q_Y(\tau|x)\right| dx; \quad (14)$$

$$SBIAS\left(\widehat{Q}_L(\tau)\right) = \frac{1}{m}\sum_{i=1}^{m}\int_0^1 \left|\widehat{Q}_{L,i}(\tau|x) - Q_Y(\tau|x)\right| dx; \quad (15)$$

where the true τth conditional quantile $Q_Y(\tau|x)$ is defined in *(11)*. The simulation efficiencies (SEFF) are given by

$$\mathbf{SEFF}_{MSE}|_{\widehat{Q}_L(\tau)}\left(\widehat{Q}_N(\tau)\right) = \frac{SMSE\left(\widehat{Q}_L(\tau)\right)}{SMSE\left(\widehat{Q}_N(\tau)\right)}; \quad (16)$$

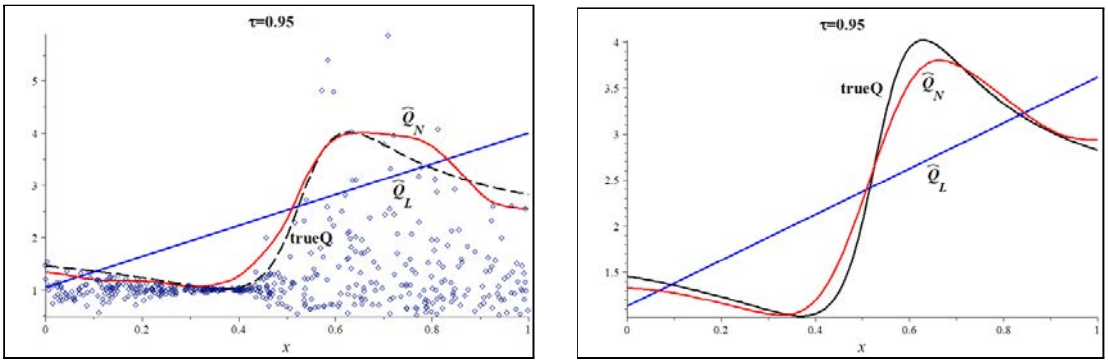
$$\mathbf{SEFF}_{bias}|_{\widehat{Q}_L(\tau)}\left(\widehat{Q}_{NE}(\tau)\right) = \frac{SBIAS\left(\widehat{Q}_L(\tau)\right)}{SBIAS\left(\widehat{Q}_N(\tau)\right)}, \quad (17)$$

where $SMSE\left(\widehat{Q}_N(\tau)\right)$, $SMSE\left(\widehat{Q}_L(\tau)\right)$, $SBIAS\left(\widehat{Q}_N(\tau)\right)$ and $SBIAS\left(\widehat{Q}_L(\tau)\right)$ are defined in *(12)-(15)*.

We have the simulation results in Table 1.

Table 1. Burr simulation mean squared errors (SMSEs), bias and efficiencies (SEFFs) of $\widehat{Q}_N(\tau|x)$ and $\widehat{Q}_L(\tau|x)$ estimating $Q_{Y_{Burr}}(\tau|x), m = 1,000, n = 500$.

τ	0.93	0.94	0.95	0.96	0.97
$SMSE\left(\widehat{Q}_L(\tau)\right)$	0.3511	0.4412	0.5746	0.7933	1.1686
$SMSE\left(\widehat{Q}_N(\tau)\right)$	0.1496	0.1963	0.2839	0.4348	0.9407
$\mathbf{SEFF}_{MSE\|_{\widehat{Q}_L(\tau)}}\left(\widehat{Q}_N(\tau)\right)$	**2.3472**	**2.2479**	**2.0240**	**1.8248**	**1.2423**
$BIAS\left(\widehat{Q}_L(\tau)\right)$	0.4888	0.5485	0.6257	0.7336	0.8890
$BIAS\left(\widehat{Q}_N(\tau)\right)$	0.2561	0.2886	0.3383	0.4080	0.5393
$\mathbf{SEFF}_{bias\|_{\widehat{Q}_L(\tau)}}\left(\widehat{Q}_N(\tau)\right)$	**1.9082**	**1.9007**	**1.8493**	**1.7979**	**1.6484**



(a) $m = 1$, $\widehat{Q}_N$, $\widehat{Q}_L$ vs true Q_{YBurr}. (b) $m = 1000$, Average $\widehat{Q}_N$, $\widehat{Q}_L$ vs true Q_{YBurr}.

Figure 4. Burr Simulation, $\tau = 0.95$, (a) for $m = 1$, $n = 500$, data (black dots), the true quantile $Q_{YBurr}(\tau|x)$ (black dash), $\widehat{Q}_L(\tau|x)$ (blue) and $\widehat{Q}_N(\tau|x)$ (red). (b) for $m = 1,000$, $n = 500$, average $\widehat{Q}_L(\tau|x)$ (blue), $\widehat{Q}_N(\tau|x)$ (red) and true quantile $Q_{YBurr}(\tau|x)$ (black).

From the Burr simulation results above, we conclude that: Table 1 and Figure 4 show that all of the $SEFF(\widehat{Q}_N(\tau)) > 1$ when $\tau = 0.95$,..., 0.99. Thus using the proposed nonparametric estimator $\widehat{Q}_N(\tau|x)$ in *(9)* is more efficient relative to the linear estimator $\widehat{Q}_L(\tau|x)$ in *(5)*.

4. Comparison of Goodness-of Fit on Quantile Regression Models

To compare the nonparametric estimator $\widehat{Q}_N(\tau|\mathbf{x})$ in *(9)* and the linear estimator $\widehat{Q}_L(\tau|x)$ in *(5)*, we use a Relative $R_N(\tau)$ of $\widehat{Q}_N(\tau|\mathbf{x})$ to $\widehat{Q}_L(\tau|\mathbf{x})$, $0 < \tau < 1$, which is defined as

$$Relative\ R_N(\tau) = 1 - \frac{V_N(\tau)}{V_L(\tau)}, \quad -1 \le R_N(\tau) \le 1, \quad \text{where} \tag{18}$$

$$V_N(\tau) = \sum_{y_i \ge \widehat{Q}_N(\tau|\mathbf{x}_i)} \frac{\tau}{n} \left| y_i - \widehat{Q}_N(\tau|\mathbf{x}_i) \right| + \sum_{y_i < \widehat{Q}_N(\tau|\mathbf{x}_i)} \frac{(1-\tau)}{n} \left| y_i - \widehat{Q}_N(\tau|\mathbf{x}_i) \right|,$$

where $\widehat{Q}_N(\tau|\mathbf{x}_i)$ is obtained by *(9)*, and

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

where $\widehat{Q}_L(\tau|\mathbf{x}_i)$ is given by *(5)*.

5. Ontario, Canada COVID-19, 2020-2021 Example

Recall the Ontario, Canada COVID-19 example in Section 1, we use the mean regression model in *(1)* and the linear quantile regression model in *(4)*, assume that

$$\begin{aligned} E(y|x_1, x_2) &= \beta_0 + \beta_1 x_1 + \beta_2 x_2, \\ Q_Y(\tau|x_1, x_2) &= \beta_0(\tau) + \beta_1(\tau) x_1 + \beta_2(\tau) x_2, \quad 0 < \tau < 1. \end{aligned}$$

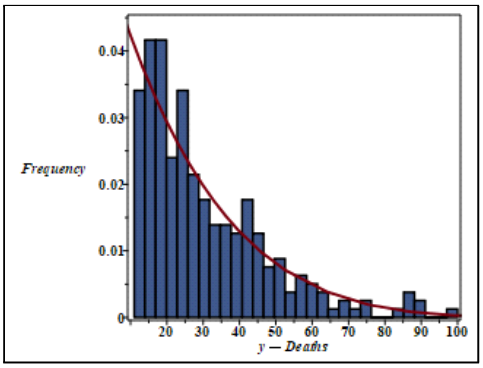
where y represents the COVID-19 daily deaths, and x_1 represents the daily COVID-19 tests, x_2 reprecents the COVID-19 daily new cases.

In this Section, we will compare the proposed nonparametric quantile regression in *(9)* with the linear quantile regression in *(5)*,

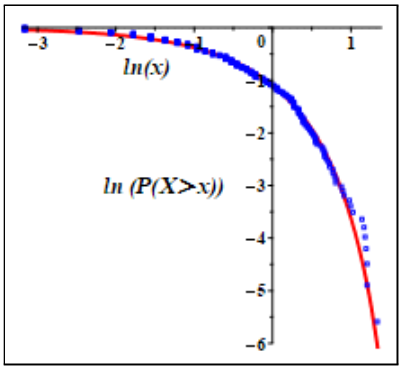
1. The linear estimator $\widehat{Q}_L(\tau|x_1, x_2)$ in *(5)* Section 1.

2. The nonparametric estimator $\widehat{Q}_N(\tau|x_1, x_2)$ in *(9)* obtained by the 5-Step algorithm in Section 2.

Figure 5(a), (b) show the histogram and log-log plot of the Ontario COVID-19 data with GPD model in *(7)* with MLEs of the parameters. The theoretical GDP curve follows the shape of the Ontarion CONVID-19 data very well.



(a) Histogram of COVID-19 data

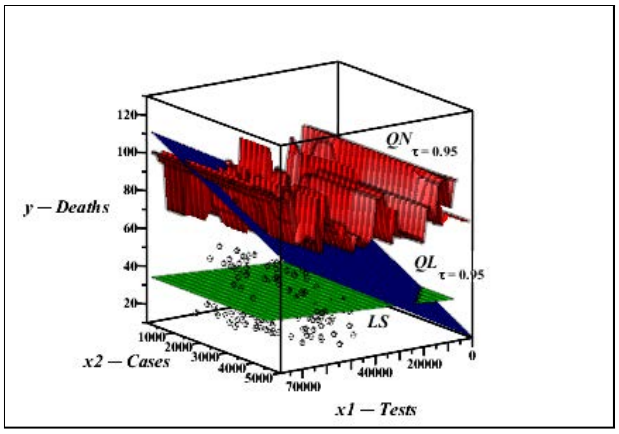


(b) Log-log Plot of COVID-19 data

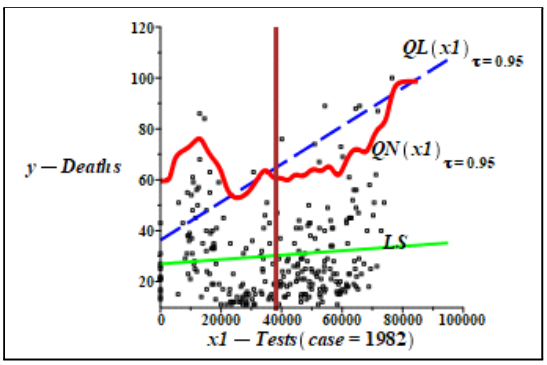
Figure 5. (a) Histogram of the Ontario COVID-19 daily death data, $n = 267$, greater than a threshold of 10 with the GPD (red). (b) a Log-log plot of Ontario COVID-19 daily death cover area. The blue dots are the data and the solid line is the GPD curve (red).

Figure 6 shows that the proposed nonparametric $\widehat{Q}_N(\tau|x_1, x_2)$ predicts that the daily deaths which is slide positive related to the daily tests and fits extreme data pattern well. It has

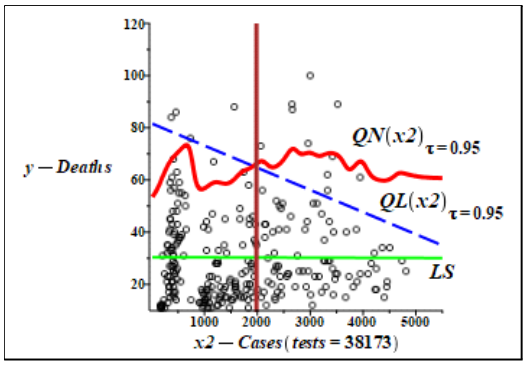
flat relationships to the daily new cases, it fits the data pattern well. We note that the linear regression $\widehat{Q}_L(\tau|x_1, x_2)$ lines are straight positive and negative relating to the daily tests and new cases, respectively. This shows that the proposed $\widehat{Q}_N(\tau|x_1, x_2)$ curves are more reasonable and flexible. The results also suggest that predicting extreme daily deaths is complicated, other regressors, e.g. vaccinating, hospital service may be considered in further studies.



(a) 3-D $\mu_{LSx_1,x_2}, \widehat{Q}_L(0.95|x_1, x_2)$ and $\widehat{Q}_N(0.95|x_1, x_2)$



(b) 2-D $\mu_{LSx_1}, \widehat{Q}_L(0.95|x_1)$ and $\widehat{Q}_N(0.95|x_1)$



(c) 2-D $\mu_{LS}x_2, \widehat{Q}_L(0.95|x_2)$ and $\widehat{Q}_N(0.95|x_2)$

Figure 6. For Ontario COVID-19 daily deaths data above 10 people, $n = 267$ days, scatter plot (black dots), (a) 3-D $\mu_{LS-x1,x2}$ plan (green), $\widehat{Q}_L(\tau|x_1, x_2)$ regression plan (blue) and $\widehat{Q}_N(\tau|x_1, x_2)$ regression surface (red) at $\tau = 0.95$. (b) When $x_2 = \overline{x_2} = 1982$, 2-D μ_{LS-x1} line (green), $\widehat{Q}_L(0.95|x_1)$ (blue dash) and $\widehat{Q}_N(0.95|x_1)$ (red curve), $x_1 = \overline{x_1} = 38173$ is the average vertical red line. (c) When $x_1 = \overline{x_1} = 38173$, 2-D μ_{LS-x2} (green) line, $\widehat{Q}_L(0.95|x_2)$ (Blue dash) and $\widehat{Q}_N(0.95|x_2)$ regression (red curve), $x_2 = \overline{x_2} = 1982$ is the average vertical red line.

Table 2 shows the values of the Relative $R_N(\tau)$ of the nonparametric $\widehat{Q}_N(\tau|x_1, x_2)$ relative to linear $\widehat{Q}_L(\tau|x_1, x_2)$ for given $\tau = 0.93$,..., 0.97. We note that $R_N(\tau) > 0$ means that $V_N(\tau) < V_L(\tau)$, thus $\widehat{Q}_N(\tau|x_1, x_2)$ is better fit to the data than $\widehat{Q}_L(\tau|x_1, x_2)$.

Table 2. Relative $R_N(\tau)$ of $\widehat{Q}_N(\tau|x_1, x_2)$ relative to $\widehat{Q}_L(\tau|x_1, x_2)$ for the Ontario, Canada COVID-19 example.

	$\tau = 0.93$	$\tau = 0.94$	$\tau = 0.95$	$\tau = 0.96$	$\tau = 0.97$
Relative $R_N(\tau)$	0.2586	0.2501	0.2447	0.2418	0.1941

6. Conclusions

After the studies above, we can conclude:

1. Traditional mean regression estimates the conditional mean by using the L_2- loss function. The linear quantile regression uses a L_1- loss function. But both models have limitations for estimating extreme conditional quantiles for the analysis of extreme events. In a heavy tailed population, the proposed five-step nonparametric quantile regression method has advantages to predicting extreme conditional quantiles compared to other existing methods.

2. The Monte Carlo computational simulation results show that the proposed nonparametric quantile regression is more efficient relative to the linear method using linear quantile regression.

3. The proposed nonparametric quantile regression can be used to predict extreme values of the Canada Ontario COVID-19 example. The results are more reasonable compared with the linear quantile regression method. The research work suggests that further studies are expected.

References

[1] Burr, I. W. (1942). Cumulative Frequency Function. *The Annals of Mathematical Statistics, 13, 215-232.*

[2] de Haan, L. and Ferreira, A. (2006). *Extreme Value Theory: An Introduction.* Springer, New York.

[3] Government of Ontario (Canada), https://data.ontario.ca/dataset/status-of-covid-19 cases-in-ontario, 2020-2021, accessed June 2021

[4] Huang, M. L. and Nguyen, C. (2018). A nonparametric approach for quantile regression, *Journal of Statistical Distributions and Applications,* Springer, Vol. 5(1), 1-14.

[5] Koenker, R. (2005). *Quantile Regression.* Cambridge University Press, New York.

[6] National Resouces of Canada (2017). http://www.nrcan.gc.ca.

[7] Picklands, J. (1975). Statistical inference using extreme order statistics. *Ann. Stat.* 3, 119-131.

[8] Scott, D. W. (2015). *Multivariate Density Estimation, Theory, Practice and Visualization.* 2nd edition, John Wiley & Sons, New York.

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

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