Quantile Regression Models and Their Applications: A Review

Qi Huang1, Hanze Zhang2, Jiaqing Chen3* and Mengying He4

1School of Law, Wuhan University, Wuhan, Hubei, 430072, PR China
2Department of Epidemiology and Biostatistics, College of Public Health, University of South Florida, Tampa, FL 33612, USA
3College of Science, Wuhan University of Technology, Wuhan, Hubei, 430070, PR China
4Department of Management, College of Business, Texas A&M University San Antonio, San Antonio, TX 78224, USA

Abstract

Quantile regression (QR) has received increasing attention in recent years and applied to wide areas such as investment, finance, economics, and engineering. Compared with conventional mean regression, QR can characterize the entire conditional distribution of the outcome variable, may be more robust to outliers and misspecification of error distribution, and provides more comprehensive statistical modeling than traditional mean regression. QR models could not only be used to detect heterogeneous effects of covariates at different quantiles of the outcome, but also offer more robust and complete estimates compared to the mean regression, when the normality assumption violated or outliers and long tails exist. These advantages make QR attractive and are extended to apply for different types of data, including independent data, time-to-event data and longitudinal data. Consequently, we present a brief review of QR and its related models and methods for different types of data in various application areas.

Keywords: Quantile regression; Check function; Asymmetric laplace distribution; Time-to-event; Longitudinal data

Introduction

In statistical modeling, regression has been developed to quantify the relationship between dependent variable (outcome) and independent variables (covariates) for over 200 years. The classic regression has been one of the most widely used statistical methods to capture the effects at the mean. These conventional regressions assume that the regression coefficients/covariates effects are constant across the population. However, such average effects are not always of interest in many areas, and sometimes quite heterogeneous. For example, Quantile regression (QR) with applications by exploring the relation of the foreign direct investment and economic growth [1,2] and in “precision health/medicine” [3,4] have been widely adopted in related fields currently. A lot of researchers, economists, financial investors, clinicians and policymakers have showed increasing attention on group differences across the entire population rather than that solely on the average. Mean regression cannot satisfy with all of these needs or requirements.

Developed by Koenker and Bassett in 1978 [5], QR complements and improves the traditional mean regression models. In this situation of homogeneity assumption violated, QR quantifies the heterogeneous effects of covariates through conditional quantiles of the outcome variable, and provides a comprehensive scan of the whole distribution of the outcome. Additionally, it is well known that when asymmetries and heavy tails exist, the sample median (the 50th percentile), one of the best-known examples of quantiles, provides a better summary of centrality than the mean. As a consequence, compared to the standard mean regression models, QR is more robust to outliers and more flexible, because the distribution of the outcome does not need to be strictly specified as certain parametric assumptions. Although mean regression-based methods still dominate the statistical modeling field, QR can be viewed as a critical extension and complement when assumptions are violated. Thus, QR has become a subject of intense investigation and application in the past decades.

QR has attracted considerable research interest in decades, and has been widely applied to independent data and time-to-event data. Recently, the use of QR for longitudinal data has also received increasing attention. This review article is organized to provide a brief overview of QR models and associated statistical methods for these three types of data with applications in different areas.

QR Models for Independent Data

In analogy with traditional linear regression, QR model for independent data was formally formulated by Koenker and Bassett [5] in 1978 as an extension from the notion of ordinary percentiles. The different QR approaches can be roughly classified into two groups: (i) minimization of weighted absolute deviations, which is a typical inferential method used in QR; and (ii) the maximization of a Laplace likelihood.

The former is based on Koenker and Bassett’s work [5], which estimated the conditional median and a full range of other quantile functions by minimizing asymmetrically weighted absolute residuals. Generally, let y and x denote the outcome of interest and the corresponding covariate vector for subject i (i=1, . . . , n), where y is independent scalar observations of a continuous random variable with common cumulative distribution function (cdf) F(y). The QR model with rth quantile for the response y given x takes the form of

\[ Q_{r}(y | x) = g(x, \beta), \]  

where \( Q_{r}(\cdot) = F_{r}^{-1}(\cdot) \) is the inverse of cdf of y given x evaluated at \( \tau \) with \( 0 < \tau < 1 \), g( ) is a known function. The regression coefficient vector \( \beta \) is estimated by minimizing

\[ \sum_{i=1}^{n} Q_{r}(y_{i} - g(x_{i}, \beta)). \]  

*Corresponding author: Chen J, College of Science, Wuhan University of Technology, Wuhan, Hubei, 430070, PR China, Tel: 86-27-87163146; E-mail: hyx123@hotmail.com

Received May 30, 2017; Accepted June 16, 2017; Published June 20, 2017


Copyright: © 2017 Huang Q, et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.
Where $\rho_\tau(\cdot)$ is the check function defined by $\rho_\tau(u)=u(\tau-1(u<0))$ and $I(\cdot)$ denotes the indicator function. A full discussion of this class of methods could be found from many related publications [5-8]. Traditional QR makes minimal assumptions on the form of the error term, which is flexible, but inference for these models is challenging, particularly when the data features are complicated.

The latter is built on the asymmetric Laplace distribution (ALD) [9-11], and other parametric distributions, like an infinite mixture of Gaussian densities [12]. ALD, which is closely related to the check function for QR, has been discussed in the literature [6,9,11,13]. A random variable $Y$ is said to follow ALD if its probability density function (pdf) with parameters $\mu$, $\sigma$ and $\tau$ is given by

$$f(y|\mu,\sigma,\tau) = \frac{\tau(1-\tau)}{\sigma} \exp\left\{ -\frac{\tau(y-\mu)}{\sigma} \right\}.$$  \hspace{1cm} (3)

Where $\rho_\tau(u)=u(\tau-1(u<0))$ is the check function, $I(\cdot)$ is the indicator function, $0<\tau<1$ is the skewness parameter, $\sigma>0$ is the scale parameter and $-\infty<y<\infty$ is the location parameter. The range of $y$ is $(-\infty, \infty)$. We denote the above distribution by ALD $(\mu, \sigma, \tau)$. Briefly, if $Y \sim$ ALD $(\mu, \sigma, \tau)$, then $P(Y \leq \mu) = \tau$ and $P(Y > \mu) = 1-\tau$, which shows that the parameters $\mu$ and $\tau$ in ALD satisfy $\mu$ to be the rth quantile of the distribution. However, the ALD is not smooth and thus difficult to maximize its likelihood function. Fortunately, as shown in these studies [14,15], the ALD has various mixture representations. A hierarchical mixture of exponential and normal distributions is utilized to develop algorithms for the QR models [14,15]. These important features of ALD have been generally adopted for likelihood based quantile inference, as well as the Bayesian inference. See Yu and Zhang’s work [11] for further properties and generalizations of this distribution as well as its close relationship with QR. By utilizing this property, under independent data setting, a large number of QR-based statistical models and various associated analysis methods have been investigated in the literature. For example, a likelihood-based goodness-of-fit test has been proposed for QR [6]; Bayesian QR [9], and the Bayesian estimation procedure for the Tobit QR model with censored data [16,15], have also been developed.

Importantly, these two classes of QR inferential methods are not mutually exclusive. The relationship between the check function and ALD can be used to reformulate the QR method in the likelihood framework. Considering $\sigma$ a nuisance parameter, it can be easily shown that the minimization of equation (2) in the former method with respect to the parameter $\beta$ is exactly equivalent to the maximization of an ALD-based likelihood function in the latter.

It has been demonstrated that QR is widely used to analyze independent data in many important application areas. First, due to the importance of modeling extreme values accurately, the foreign direct investment (FDI), finance and economics are the most important area where QR is utilized. Girma and Gorg [1] and Zhou [2] used QR modeling to explore relationship between the foreign direct investment and economic growth. Several economists have examined wage structure and wealth distribution using QR [17-19]. Specifically, research has been conducted to explore the gap in wage and wealth distribution [20], including the effect of gender on wage [21,22], wage differences between public and private entities [23-25], wealth inequality between urban and rural areas [26], and the impact of education on wage [27-29] and intergenerational earnings [30]. Focusing on the women group, Moshe Buchinsky studied female wage distribution in the USA [22] and their return to education [31] using QR as an analytic tool. In addition, QR has also been applied in economic-based discipline. In the area of economics and education, QR has been applied to examine the impact of school choice [32] and quality [33,34] on student performance and achievement [33]. With the application of economics on management, QR has been used to study the effect of innovation on firm growth [35] and relationship between companies foreign ownership and production efficiency [36] as well as association of FDI and economic growth [1,2]. In the sub-area of economics and policy, existing corruption levels have been explored [37] and the relationship between FDI and corruption level [38] has been examined using QR models. In the finance field, QR has been adopted to study housing price [39], capital structure [40], FDI and investment treaties [1] and stock market returns [41] based on cross-sectional data.

In addition, QR is becoming more popular in clinical, biomedical, and other health related research. For example, Austin et al. [42] examined the varying gender differences in the delivery of thrombolyis in patients with an acute myocardial infarction by QR. Brolliais and Durrieu [43] provided a review of recent applications of QR to the area of genetics. Azagha and Sharaf [44] identified that increasing the intake of fruits and vegetables may be an effective dietary strategy to control weight and mitigate the risk of obesity, which is more effective at the higher quantiles of the body mass index (BMI) distribution. More applications of QR to independent data could be found in various fields, such as public health [45,46], bioinformatics [47], healthcare [48,49], environmental science [50], and ecology [51,52].

**QR Models for Time-to-Event Data**

Time-to-event data arise when interest is focused on the time elapsing before an event is experienced. Application to analysis of this kind of data, called survival analysis or duration models, is objective to investigate the effects of covariates on the survival/duration time. These effects can be heterogeneous on low, medium, and high risk subjects. In other words, covariates may have greater effects at an early period of survival, and weaker effects or even no effect later, or vice versa. QR has been considered to apply to measure the differences of covariates effects at different quantiles of survival/duration time [53]. Furthermore, the survival/duration time often exists non-normality and long tails, and thus QR-based survival models provide more robust estimation than traditional mean regression-based ones.

Although, Coxs proportional hazard model is the most often used for survival analysis, it is rarely generalized to QR-based models. Alternatively, the accelerated failure time (AFT) model with the transformed survival time can be employed to QR field, in which logarithm transformation is the most commonly used one [53-56]. Due to the complexity of the time-to-event data, large number of studies has contributed to the QR-based AFT model under different scenarios. Ying et al. [57] studied a semiparametric procedure for median regression. Yang [54] extended the median regression with weighted empirical survival and hazard functions based estimation. Portnoy [58] generalized the principle of the Kaplan-Meier estimate under QR framework. Yin et al. [59] investigated the quantile regression model for correlated failure time data. Peng and Huang [60] developed an estimator which is very close to Nelson-Aalen estimator. Most recently, great work is still expanding this area to recurrent events [61-63], various censoring types [64-66], competing risks [65,67-69].

There are many applications of QR to survival analysis or duration models. For instance, in finance and economics, Schaeck [70] assessed the association among bank liability structure and time to failure by a
QR Models for Longitudinal Data

Longitudinal data, sometimes called panel data, show great complexity in statistical analysis and application due to the correlation between and within repeatedly measured observations. In statistics, mixed-effects models are becoming increasingly popular in longitudinal data analysis. However, the majority of longitudinal modeling methods are based on mean regression to concentrate only on the average effect of covariate and the mean trajectory of longitudinal outcome. Thus, mimic to independent data, QR has been also extended and applied to longitudinal data. Longitudinal QR has the capability, at both of the population and individual level, to identify heterogeneous covariates effects, and describe differences in longitudinal changes at different quantiles of the outcome, and provides more robust estimates when heavy tails and outliers exist.

Similar as QR for independent data, longitudinal QR models, specifically QR-based mixed-effects models have been proposed via different statistical approaches, which could also be classified into two categories: distribution-free and likelihood-based. In details, for example, Jung [79] firstly developed a quasi-likelihood method for median regression considering correlations between repeated measures for dependent data. He et al. [80] proposed a median regression based linear mixed-effects model for longitudinal data. Koenerk [81] generalized his previous work on QR to longitudinal data via penalized least squares method. Other methods or algorithms used to QR includes Barrodale-Roberts algorithm [82], Expectation-Maximization (EM) algorithm [83], Monte Carlo Expectation-Maximization (MC) algorithm [13,84,85], and Bayesian approach by Markov chain Monte Carlo (MCMC) procedure [86-93]. Longitudinal QR has been rapidly expanded in many areas, including investment and finance [94,95], economics [96], environmental science [97,98], geography [99], public health [100,101] and biomedical research [102-105]. In investment and finance areas, Bassett and Chen [94] utilized longitudinal QR to provide additional information from the time series data of portfolio returns based on the way style that affects returns at places other than the expected value of return. In economics, Buchinsky [96] studied US wage structure from 1963 to 1987 with the application of longitudinal QR. It provided a full scan of information among time effects, education level, and years of experience in different wage quintile. In public health, Smith et al. [100] revealed that the association between high blood pressure and living in an urban area has evolved from positive to negative, with the strongest changes occurring in the upper tail. In meteorology, Timofeev and Sterin [97] utilized longitudinal QR to analyze various changes in climate characteristics. In biomedical studies, Rezvian et al. [104] investigated the effect of a naturally derived biological peptide, P28, and found that it produced slower rates of growth in the upper quantiles of melanoma tumor volumes in mice.

Data collected in many longitudinal studies record much information, not only repeated measures, but also time-to-event information. For example, in HIV/AIDS studies, viral load (the number of copies of HIV-1 RNA) and CD4 cell counts are important biomarkers of the severity of a viral infection, disease progression, and treatment evaluation, and their time trends of longitudinal measures may also be predictive of the risk of a terminal event. Thus, joint models are an active area of statistics, because of its capability on the bias reduction and improvement of estimates' efficiency. More recently, QR has been extended to more complicated joint models in AIDS research. Farcomeni and Viviani [85] developed QR-based survival joint models in the presence of informative dropout. Huang et al. proposed QR-based mixed-effects joint models by considering many longitudinal data features simultaneously, including covariate measurement errors [90,93,89], missing [90,91], non-normality [90-92], left-censoring [89,92], and time-to-event outcomes [91].

Summary and Conclusion

This review provided a general overview of QR-based models and methods targeting different types of data and application areas. We have illustrated that QR is a powerful tool to detect heterogeneous effects of covariates at different quantiles of the outcome, and complements excellently the mean regression when data are in presence of outliers and long tails. Recent developments and extensions in QR-based models offer increasing ability and flexibility in capturing independent, time-to-event, and longitudinal data with different data features, which can benefit applications in various scientific and finance areas.

We believe that QR, a comprehensive strategy, has a bright future. In financial/investment market, QR is more powerful for investors to predict investment strategies; in medicine, according to the idea of "precision medicine", QR is more precise for physicians to evaluate treatment and make clinical decisions, compared to mean regression models. In statistics, especially in the "big data" era, data sources get richer, data structures become more complicated, extreme values and heterogeneity increase. Instead of the mean regression, which hardly meets our expectation, QR methods dig deeper into the data, grab more information, and become more relevant. Last but not the least, as the power of the computer has advanced, the computational load for QR-based models and methods has decreased substantially. Thus, more complicated QR-based models could be considered under a Bayesian framework [89-93] and applied to more diverse areas in near future.

A final note that we would like to make is possible software to implement QR modeling methods. The most widely used software for QR models is R with "quantreg" package [106]. It covers linear, nonlinear parametric and non-parametric (total variation penalized) models for conditional quantiles of a univariate response, and several methods for handling censored time-to-event data. Other R packages are also available for specific QR topics. For example, R package "cmprskQR" [69] is developed for analysis of competing risks using QR; package "lqm" [107], and "qrLMM" [108], deal mainly with longitudinal data via QR-based linear or non-linear mixed-effects models. SAS currently also includes a "quantreg" procedure, which is similar as the R "quantreg" package. Stata software has "qreg" function to fit QR models, but the capabilities are limited. QR also has been added to SPSS (version 22.0.0 or later), just simply estimate one or more conditional quantiles for a linear model. When the model components are very complicated, especially for survival and longitudinal data with multiple data features, which bring extremely heavy computational load, the Bayesian method shows its advantages. The WinBUGS software [109] interacted with the package "R2WinBUGS" in R and "Rstan" package [110] in R are good choices with a lot of flexibility for Bayesian inference.
Acknowledgements
This research is partially supported by the National Natural Science Foundation of China (61671833) to J. Chen.

References
34. Levin J (2002) For whom the reductions count A quantile regression analysis of class size and peer effects on scholastic achievement. Economic Applications of Quantile Regression Springer 221-246.


