Thailand Statistician
January 2025; 23(1): 115-128
http://statassoc.or.th
Contributed paper

Performance Comparison of the Quantile Regression Coefficient Estimation with Outliers

Pimpan Ampanthong* [a] and Rungruttikarn Moungmai [b]

[a] Department of Mathematics, Faculty of Science and Technology, Rajamangkala University of Technology, Suvarnabhumi, Suphanburi Campus, Thailand.

[b] Department of Mathematics and Statistics, Faculty of Science and Technology, Nakhon Sawan Rajabhat University, Nakhon Sawan, Thailand.

*Corresponding author; e-mail: pimpan.a@rmutsb.ac.th

Received: 27 July 2023
Revised: 9 July 2024
Accepted: 10 July 2024

Abstract

This research aims to compare the coefficients estimation performance of quantile regression (QR) at different percentiles and simple regression (SR) for dataset with outliers. Five quantile positions were considered comprising QR(10)th, QR(25)th, QR(50)th, QR(75)th and QR(90)th. These were obtained by modifying the probability density functions for the kernel function adjustment under errors. The results were then compared with the simple regression coefficient estimation. After applying variety situations of the simulation, the mean absolute error (MAE) was used as criteria for consideration. The results showed that although the best performance model for large sample size was the SR, it still gave the best performance for small sample size as well as QR(25)th and QR(50)th models. Furthermore, the QR(25)th and the QR(50)th models were the most efficient estimation of the quantile regression coefficients with outliers for moderate sample size. They also indicated that there were changes scattered around zero value when the sample size was small. Comparing between the SR model and the QR(50)th model for every sample size, it was found that their model coefficients are slightly different. Considering kurtosis and skewness for the SR and all QR models, the results revealed that both values increased with small sample size. Then, they decreased with moderate sample size and increased again with large sample size. Therefore, the quantile regression coefficient estimation is effective in relationship analysis and provides estimates that are more accurate than one answer and suitable as an alternative analysis for skewed data.

Keywords: Regression coefficient, simple regression, quantile regression, interquartile range, outliers

1. Introduction

Regression analysis is a statistical process for analysis of relationship between variables. This correlation is then presented in a form of a model or extrapolation issue may occur if the model is used to predict the future. There are many applications of the model, including Manoj and Suresh (2023) showing the application of time-varying coefficient regression models in forecasting financial data. Araveeporn et al. (2010) were interested in the exchange rate forecast of the Thailand index

using a conditional heteroscedastic nonlinear model with autocorrelated errors. Kumari and Tan (2018) study on modeling and forecasting of volatility series with reference to gold prices. From the model for predicting the relationship of variables, it is divided into two groups. One variable is the variable of interest or the variable that is affected by another variable called the response or dependent variable. The other is the variable that influences the dependent variable or the factor(s) that causes the response changes, called the independent or explanatory variable which in regression analysis may have only one or more independent variables. Regression models are widely applied in many areas such as marketing, business, agriculture, and industry. For example, the field of economics shows the need to use analysis of basic financial data that relates to different periods of time, such as annually, quarterly, or monthly, to show the capabilities of the financial system, income statement, balance sheet, and cash flow, etc. In this study, only one independent variable is included in the model, called simple regression. To obtain the best and unbiased estimator based on the initial assumptions of the regression model, the ordinary least squares (OLS) method is widely used to estimate model parameters. The assumptions comprise the errors is independent, the errors' average is close to zero, the errors' variance is stable (homoscedasticity), and the sum of covariance is close to zero. However, if the regression model does not meet these assumptions, it might be from several reasons, for example, there is an outlier(s) associated with the dataset or the dataset is skewed. As a result, predictive results of the regression models created from such data are abnormal or inaccurate. For instance, the analysis does not cover all the values of the data scattered around the average or the predicted value(s) may be higher than the average or lower than the average. The estimation of regression coefficients with outliers was researched by Ampanthong and Suwattee (2010) who interested that for robust estimation with outliers, the comparison of regression methods under high-dimensional sparse data with multicolinearity (Choosawat et al. 2020; Dawoud 2022; Kleinbaum et al. 2008). Therefore, if the regression model according to the above assumptions is adapted to regression analysis from the percentile position of the data, called quantile regression model. The quantile regression for real data was reported by Ninbai (2019), the model will allow the analysis to cover the dataset with outliers or the skewed dataset. It can be said that the quantile regression model can solve for the distribution of the responses outside the average such as the value at the tail (location shift), the unsteady variance (scale shift), and the skewness of the response (shape shift).

In a study by Rosenblatt (1956), quantile regression analysis can explain and is suitable for skewed dataset or dataset with an outlier(s). The analysis mainly considers the data location as well as finding an interval between ranges called interquartile range (IQR) (Devroye and Gyorfi 1985) which is a distribution measurement of the data by considering a difference between the data at the third quartile (Q_3) and the first quartile (Q_1). A dataset whose distribution deviates from the median toward Q_1 and Q_3 are called right skewed and left skewed respectively. Therefore, the regression analysis from the data position will show a better analysis of the data with such anomalies. Although it does not have the mean distribution property of the response like simple regression does. It can be noted that these two regression analysis methods are similar and different. In other words, both regression models were constructed by estimating the regression coefficients of unknown parameters using the dataset. However, the modelling of these two models relies on different analytical principles. Currently, both regression analysis and quantile regression analysis are attractive and widely used for creating predictive models.

In this study, the regression coefficient estimation performance of the simple regression and the quantile regression models for a dataset with outlier(s) are studied and compared. The quantile regression coefficients are estimated by adjusting the probability density function at various percentile positions and choosing a kernel function that determines the optimal bandwidth to optimise

the regression coefficient estimation. The interquartile range (IQR) is then used to detect abnormal data. Also, a boundary range of the outlier(s) of the dataset that is distributed differently from the majority of data is sought. After that the regression models from both methods were analyzed to investigate a proper model by comparing the values of mean absolute errors (MAE) through simulated data and real data. For data simulation, various datasets will be generated under different sizes of skewness and sample sizes that are 20, 30, 50, 100, and 200. Section 2 introduces coefficients estimation of the simple regression model. Section 3 displays an estimate of the quantile regression coefficient, scaling value of the quantile, and the skewness of the quantile of the data. Section 4 presents the comparison results of the regression coefficient estimation performance of the proposed models obtained from the data simulation. Section 5 reveals the results of applying the proposed regression models to real data on government taxation. Section 6 indicates summary and discussion of the data analysis results.

2. Materials and Methods

2.1. Simple regression (SR) coefficient estimation

Simple regression model is to find a relationship between a response (y_i) and one independent variable (x_i) under the assumption that the mean of the conditional response in the independent variable, for n pairs of $(x_1, y_1),(x_2, y_2),...,(x_n, y_n), i = 1,2,...,n.$ Simple regression model can be written as,

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, i = 1,2,...,n. \quad (1)$$

Let $\hat{\beta}_0$ and $\hat{\beta}_1$ be parameter estimators of β_0 and β_1, respectively. Estimation of the regression model parameters in (1) using the ordinary least squares method (OLS) can be obtained by estimating the sum of the least squared errors. These errors can be calculated from the difference between the observations of the response and their predicted values (Hao and Naiman 2007). Therefore, $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$ and $\hat{\beta}_1 = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$, respectively. From simple regression inference, a regression model is obtained with a normal distribution where the average is zero and the variance of the response is the same for any value of the predictor. That is the standard error is calculated as the distribution difference from the parameter estimates, (Pagano and Gauvreau 2000). If the hypothesis is true, this will get $\varepsilon \sim N(0, \sigma^2)$, for all $i = 1,2,...,n.$ Under the OLS method, the same parameters were estimated as the maximum likelihood estimates.

2.2. Quantile regression coefficient estimation

The quantile regression model was developed by Koenker and Bassett (1978) as an alternative method for finding relationship between the response and independent variable(s) of the skewed dataset. It was also extended to include outlier(s) for modelling. The outlier(s) is caused by the independent variable having an abnormally high or low value(s) called shift scale, therefore resulting in this outlier(s) is involved in the values of the response. To estimate the quantile regression coefficients. The distribution of the response is considered due to its dataset is skewed namely shape shift. This skewness may be skewed to the left or to the right depending on the outlier(s), that will appear on which side or deviate from the mean.

The r^{th} quantile is the value of the response located at the position where the number of data is less than the entire population proportionally at represent in quantile position. Let μ_r be a value of y at the r^{th} of quantile position and can be written as,

$$F_y(\mu_r) = P(y \le \mu_r) = r.$$

The quantile regression model can be given by

$$y_i = \beta_0^{(r)} + \beta_1^{(r)} \underline{x}_i + \varepsilon_i^{(r)}, i = 1, 2, ..., n. \quad (2)$$

The estimation of the quantile regression coefficients can display the values of all distributions as the proportion r^{th} of the response shown in (2) (Koenker and Bassett 1978) and is equal to $P(Y > y) = 1 - F(y)$. Moreover, it represents a meaningful estimate of the quantile regression coefficients with a conditional probability distribution $F_y(y)$ when x is given at the r^{th} quantile terms (Hao and Naiman 2007) which is

$$Q_{y/x}(r) = \inf imum\{y : F_y(y) \ge r\},$$

where $E(Y/X) = \beta_0^{(r)} + \beta_1^{(r)} x_i, i = 1, 2, ..., n.$ Let r^{th} be a percentile position of the estimation, $\beta_i^{(r)}$ be a 1×2 vector of the regression coefficients where $\beta_i^{(r)} = \left[\beta_1^{(r)}, \beta_2^{(r)}\right]'_{1\times 2}$ and $\varepsilon_i^{(r)}$ be $n \times 1$ vector of the errors at r^{th} quantile where $\varepsilon_i^{(r)} = \left[\varepsilon_1^{(r)}, \varepsilon_2^{(r)}, ..., \varepsilon_n^{(r)}\right]'_{1\times n}$. The estimation of the r^{th} quantile regression coefficients can be calculated from as,

$$Q_{y/x}(r) = \hat{\beta}_0^{(r)} + \hat{\beta}_1^{(r)} x_i, i = 1, 2, ..., n.$$

Under calculating the weighted sum (r) of the given error, $\beta_i^{(r)}$ can be estimated from as,

$$\hat{\beta}_i^{(r)} = \min_{\hat{\beta}_0^{(r)}, \hat{\beta}_1^{(r)}} \left(\sum_{\substack{i=1 \\ y \ge \beta_0^{(r)} + \hat{\beta}_1^{(r)} x_i}}^{n} \left| y_i - \hat{\beta}_0^{(r)} + \hat{\beta}_1^{(r)} x_i \right| + (1-r) \sum_{\substack{i=1 \\ y \ge \hat{\beta}_0^{(r)} + \hat{\beta}_1^{(r)} x_i}}^{n} \left| y_i - \hat{\beta}_1^{(r)} + \hat{\beta}_2^{(r)} x_i \right| \right).$$

Let $\rho(X, x_i)$ be a probability density function of $x_i = (x_1, x_2, ..., x_n)$ which is a random sample of size n from the population $\underline{X} = (\underline{X}_1, \underline{X}_2, ..., \underline{X}_n)$ whose distribution is unknown and $\hat{\rho}(X, x_i)$ be an estimator of such function. The density function of a random variable can be written as,

$$\rho(X, x_i) = \lim_{h \to 0} \frac{1}{2h} P(x_i - h < X < x_i + h).$$

For window width h corresponding to a symmetrical function with $(x_i - h < X < x_i + h)$ and using $\hat{\rho}(X, x_i)$ as a kernel function. It gives the weight of the nearby data point in solving the estimators, the estimator of $\rho(X, x_i)$ can be approximated from as,

$$\hat{\rho}(X, x_i) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{X - x_i}{h}\right).$$

Proposition 1. *Let x_i be a random sample of size n from the population for which the probability density function is unknown, $K(t)$ be a kernel function, k_2 be a constant, and h be window width*

where $h = h(n) \to 0$ *for* $n \to \infty$. *The kernel function property comprises,* $\int K(t)dt = 1, \int tK(t)dt = 0$ *and* $\int t^2 K(t)dt = k_2 \neq 0$. *The bias of* $\hat{\rho}(X, x_i)$ (Härdle et al. 1993) *can be written as,*

$$bias\,\hat{\rho}(X, x_i) = \frac{1}{2}h^2 f^2(x_i)k_2 + ... + o(h^2).$$

According to the study of Devroye and Gyorfi (1985), it was found that the best window width for a population with a normal distribution and a kernel function used as the window width was $h = \min\left(\sigma_\varepsilon, \frac{IQR}{1.34}\right) \times n^{\left(-\frac{1}{5}\right)}$ where σ_ε stands for standard deviation of the errors which can adjust the probability density function for the proper estimation $\hat{y}_i$.

Lemma 1. T*he properties of quantile regression coefficients* $\hat{\beta}_i^{(r)} = \left(\hat{\beta}_1^{(r)}, \hat{\beta}_2^{(r)}\right)$ *in (6) are as follows,*

1. Scale equivariance for all $c > 0$.

1.1 $\underline{\hat{\beta}}^{(r)}(cy_i, x_i) = c\underline{\hat{\beta}}^{(r)}(y_i, x_i), i = 1, 2, ..., n.$

1.2 $\underline{\hat{\beta}}^{(r)}(-cy_i, x_i) = -c\underline{\hat{\beta}}^{(r)}(y_i, x_i), i = 1, 2, ..., n.$

2. Shift equivariance for all $d \in R^k$.

$$\underline{\hat{\beta}}^{(r)}(y_i + x_i d, x_i) = \underline{\hat{\beta}}^{(r)}(y_i, x_i) + d, i = 1, 2, ..., n.$$

3. Equivariance to reparameterization and A be a $k \times k$ *matrix.*

Lemma 2. *Estimations of the quantile regression estimation at the percentile position QR(10)th, QR(25)th, QR(50)th, QR(75)th and QR(90)th were obtained by adjusting the probability density function for the estimation in the kernel function under the errors. Their edges and skewness are as follows:*

1. Quantile based scale

1.1 Inner edge (Ie) is equal to QR(75)th - QR(25)th.

1.2 Outer edge (Oe) is equal to QR(90)th - QR(10)th.

2. Quantile based skewness

2.1 Inner skewness (Is) is equal to (QR(75)th - QR(50)th / QR(50)th - QR(25)th)-1.

2.2 Outer skewness (Os) is equal to(QR(90)th - QR(50)th / QR(50)th - QR(10)th)-1.

3. Results and Discussion

3.1. Simulation data

3.1.1 Simulation design

In this study, there were 1,000 iterations with each iteration was as follows:

1. Generating a dataset under the data replication requirements detailed below:

(1) The independent variable is under a uniform distribution $x_i \sim U(0,3)$ and relationship condition of $y_i = 10 + 0.5x_i + \varepsilon_i$.

(2) The errors belong a normal distribution with different skewness that is $\varepsilon_i \sim N(0, \sigma^2)$ where variance equal 1 and 6.

(3) Sample sizes comprise 20, 30, 50, 100, and 200.

2. Creating regression models that are simple regression model (SR) and quantile regression models QR(r)th at 5 percentiles position as, QR(10)th, QR(25)th, QR(50)th, QR(75)th and QR(90)th.

3. Determining the regression coefficients efficiency of the proposed models using the mean absolute error (MAE). Let y_i and be an observed value and its predicted value respectively and n be sample size. The value of MAE can be written as

$$MAE = \frac{\sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right|}{n}.$$

3.1.2 Simulation results

The results can be summarized as follows:

1. Performance comparison between simple regression model (SR) and quantile regression model at median position (QR(50)th). After 1,000 iterations were operated, the values of standard error (Std. errors) and mean absolute error (MAE) were calculated and evaluated as shown in Tables 1 and 2 and Figures 1 and 2.

1.1 Model comparison results for generated data where variance 1.

Table 1 Comparison of MAE values and Std. errors with variance 1

Sample sizes	20		30		50		100		200	
Models	SR	QR(50)th	SR	QR(50)th	SR	QR(50)th	SR	QR(50)th	SR	QR(50)th
Std. errors	27.836	28.883	23.964	25.055	27.766	30.592	23.007	26.340	26.721	29.734
MAE (%)	0.277	0.235	0.375	0.324	0.906	0.725	1.450	1.219	3.056	2.470
Average of the estimates of the parameters										
$\hat{\beta}_0$	11.771	0.116	11.725	0.115	11.756	0.115	11.653	0.114	11.711	0.114
	(0.177)	(0.998)	(0.172)	(0.988)	(0.175)	(0.988)	(0.165)	(0.988)	(0.171)	(0.988)
$\hat{\beta}_1$	−1.189	−0.010	−1.173	−0.010	−1.249	−0.010	−1.168	−0.009	−1.201	−0.009
	(3.379)	(1.034)	(3.347)	(1.021)	(3.493)	(1.021)	(3.337)	(1.019)	(3.403)	(1.019)
Change from mean	1.778	1.016	1.759	1.004	1.834	1.004	1.751	1.004	1.787	1.745

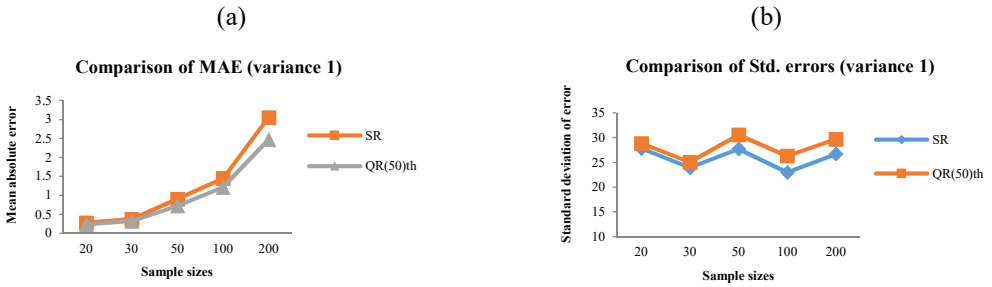

Figure 1 (a) Comparison of MAE at variance 1 and (b) Comparison of Std. errors at variance 1

Table 1 and Figure 1 (a) show that the MAE value of the QR(50)th model was lower than that of the SR model for all sample sizes. It can also be observed from Figure 1(a) that when the number of samples increases, the MAE values of both models also increases. However, the sample size had no

effect on the variation of Std.errors as shown in Figure 1(b). Furthermore, Figure 1 shows that the MAE and Std.errors values of both models are similar for small sample sizes (n = 20 and 30). Considering the regression coefficient estimation from the change in the mean, it was found that the range of interval estimates of the regression coefficients ($\hat{\beta}_0$ and $\hat{\beta}_1$) of the QR(50)th model was smaller than the estimator of the SR model for every sample size. This can be observed from the quantile estimates that there is very little change in the regression coefficient from the mean for every sample size.

1.2 Model comparison results for generated data where variance 6.

Table 2 Comparison of MAE values and Std. errors with variance 6

Sample sizes	20		30		50		100		200	
Models	SR	QR(50)th	SR	QR(50)th	SR	QR(50)th	SR	QR(50)th	SR	QR(50)th
Std. errors	167.016	173.318	143.788	149.591	166.596	182.786	138.047	158.035	160.330	178.404
MAE (%)	1.447	1.230	4.002	4.343	8.480	8.668	15.662	17.057	49.227	64.492
Average of the estimates of the parameters										
$\hat{\beta}_0$	20.631 (1.063)	0.196 (0.980)	20.350 (1.035)	0.196 (0.980)	20.539 (1.053)	0.194 (0.980)	19.920 (0.992)	0.188 (0.981)	20.267 (1.026)	0.189 (0.981)
$\hat{\beta}_1$	–9.639 (20.279)	–0.089 (1.179)	–9.542 (20.084)	–0.089 (1.178)	–9.981 (20.963)	–0.088 (1.176)	–9.513 (20.026)	–0.083 (1.167)	–9.710 (20.421)	–0.084 (1.168)
Change from mean	10.671	1.079	10.559	1.079	11.008	1.078	10.509	1.074	10.723	1.074

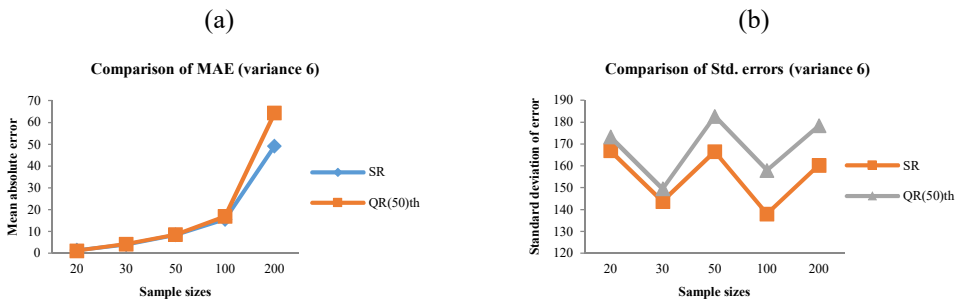

Figure 2 (a) Comparison of MAE at variance 6 and (b) Comparison of Std. errors at variance 6

From the comparison of the performance of two models shown in Table 2 and Figure 2, it revealed that the results were similar to the results of the previous section (variance 1). That is the MAE value of the QR(50)th model is less than the SR model. Also, the MAE values vary with the sample sizes whereas the sample sizes does not affect the change of the Std.errors values for both models. In addition, the interval estimate of the parameter estimator of the QR(50)th model is still smaller than the interval estimate of the SR model. Considering Figure 2, it was found that the MAE and Std.errors values of both models remained similar for small sample size. However, when the sample size increased, i.e. n = 50 and 100, it was observed that the MAE values of the two models remained slightly different from the results of the previous section.

2. Performance comparison of skewness (Sk) and kurtosis (Ku) between the simple regression model (SR) and the quantile regression model (QR) at different percentiles, QR(10)th, QR(25)th, QR(50)th, QR(75)th and QR(90)th. After 1,000 iterations were operated, the values of skewness (Sk),

kurtosis (Ku), inner and outer edges, and inner and outer skewness of each model were calculated and evaluated as shown in Table 3 to 4 and Figure 3 to 4.

2.1 Ku and Sk comparison results for generated data where variance 1.

Table 3 Comparison of Ku, Sk and quantile base scale and skewness with variance 1

Sample sizes		20		30		50		100		200
Models	Ku	Sk	Ku	Sk	Ku	Sk	Ku	Sk	Ku	Sk
QR(10)th	15.416	3.738	19.912	4.836	10.485	0.311	12.161	0.488	19.100	2.775
QR(25)th	15.074	3.699	19.813	4.635	10.373	0.440	13.614	0.353	25.747	3.108
QR(50)th	15.033	3.514	19.415	4.036	10.273	0.428	13.188	0.234	24.293	2.922
QR(75)th	14.890	3.262	19.185	3.982	10.107	0.462	13.101	0.375	25.654	3.167
QR(90)th	11.992	2.027	16.702	3.582	8.484	0.489	11.057	0.564	23.585	3.121
SR	11.993	3.090	16.328	3.526	6.551	0.474	10.830	0.502	22.867	3.073
Quantile based scale and skewness										
Inner edge	–0.183	–0.436	–0.627	–0.653	–0.265	0.022	–0.513	0.021	–0.093	0.058
Outer edge	–3.424	–1.711	–3.210	–1.254	–2.000	0.177	–1.104	0.076	4.485	0.345
Inner Skewness	2.485	0.361	–0.422	–0.909	0.668	–3.964	–0.794	–2.183	–1.935	–2.315
Outer Skewness	6.928	5.632	4.456	–0.432	7.454	–0.484	–3.075	–2.299	–1.136	0.348

From Table 3, it was found that the kurtosis values of both SR model and all QR models had increased for the small sample size alternated with a decreasing kurtosis value at the moderate sample size and increased again for the large sample size. For comparison of skewness, the data were skewed more to the left than average at small sample sizes. Then, the skewness had decreased at moderate sample size and it had skewed more to the left at large sample size. As for the estimation of the inner and outer quantile scales, the range was wider with small sample size and the range decreased with increasing sample size. Considering the diffuse skewness around the median of the quantiles, it was found that the span width was small with small sample size, and the span increased with increasing sample size.

(a) (b)

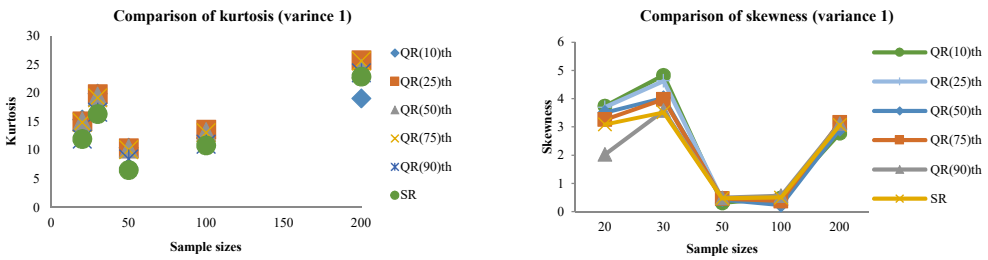

Figure 3 (a) Comparison of Ku at variance 1 and (b) Comparison of Sk at variance 1

It can be seen from Figure 3 that the kurtosis value of the SR model has distributed near zero when the sample size is small and moderate whereas the kurtosis value of the QR(10)th model has distributed near zero when the sample size is large. Considering the skewness comparison, the skewness of the SR, the QR(75)th and the QR(90)th models have skewed around zero when the sample sizes are moderate and large. This means the data moved more towards the middle average.

2.2 Ku and Sk comparison results for generated data where variance 6.

Table 4 Comparison of Ku, Sk and quantile base scale and skewness with variance 6

Sample sizes	20		30		50		100		200	
Models	Ku	Sk	Ku	Sk	Ku	Sk	Ku	Sk	Ku	Sk
QR(10)th	15.416	3.738	19.975	4.036	10.572	0.311	12.161	0.488	19.100	2.775
QR(25)th	15.088	3.635	19.913	4.065	10.475	0.439	13.617	0.352	25.742	3.108
QR(50)th	15.033	3.514	19.808	4.136	10.178	0.432	13.189	0.234	24.293	2.922
QR(75)th	14.919	3.430	19.474	3.980	10.309	0.462	13.101	0.357	25.654	2.823
QR(90)th	11.979	3.033	16.701	3.581	8.473	0.489	11.043	0.565	23.526	3.117
SR	11.993	3.090	16.328	3.526	6.551	0.474	10.830	0.502	22.867	3.073
Quantile based scale and skewness										
Inner edge	–0.168	–0.204	–0.439	–0.085	–0.166	0.023	–0.516	0.004	–0.088	–0.284
Outer edge	–3.437	–0.704	–3.274	–0.454	–2.098	0.178	–1.118	0.077	4.426	0.342
Inner Skewness	1.059	–0.305	2.182	–3.196	–1.440	–5.593	–0.794	–2.037	–1.939	–0.465
Outer Skewness	6.965	1.149	17.582	–6.525	3.320	–0.532	–3.088	–2.302	–1.147	0.323

Table 4 shows that the kurtosis of both the SR and all QR(r)th models will increase when the sample size is small. Then, it decreases for moderate sample size and increases again for larger sample sizes. Considering the skewness, it can be seen that the left skewed is greater than the mean for small sample sizes. Also, it can be noticed that it will decrease if the sample size is moderate and will increase if the sample size is large. For inner and outer edges estimation of the quantile models, it found that their ranges are narrow for small sample sizes and will expand if the sample sizes are larger. Moreover, the detail of the Table 4 displays that the skewness will be around the median of the quantile. However, their ranges will be wide for small sample size and will be decreased if the sample size is increased.

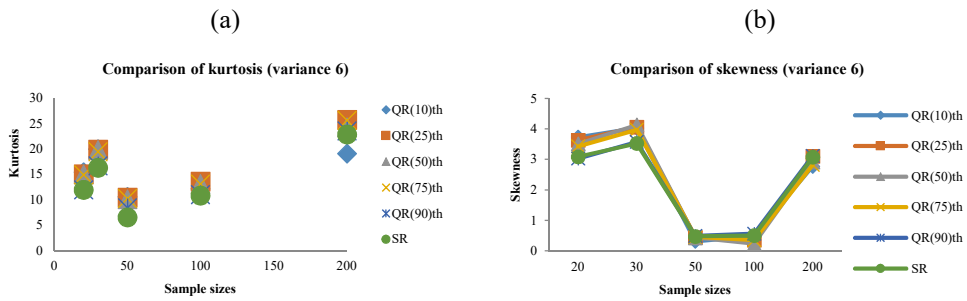

Figure 4 (a) Comparison of Ku at variance 6 and (b) Comparison of Sk at variance 6

Figure 4 states that the Kurtosis of the SR model closely spread to zero when the sample sizes are small and moderate. However, when the sample sizes are increased, the kurtosis will spread around zero for QR(90)th model. The skewness comparison reveals that the skewness of the SR, QR(50)th, and QR(90)th models will spread around zero for the moderate sample size. This means the data is close to the middle mean.

3.2 An application in real data

The Ministry of Finance's authority is to collect the country's revenue by collecting taxes from three important agencies, the Customs Department, the Excise Department, and the Revenue Department. This study only focuses on the country's revenue collection from the taxation of the Revenue department whose primary duty is to collect tax according to the revenue code and related laws consisting of direct taxes and indirect taxes. Tax burden analysis considers the impact of taxation on the fair distribution of goods in society. In such collection, the government may choose to collect various types of taxes. To make taxes a tool that can achieve various objectives and is a source of income to maintain economic stability and distribution of justice in society. The researcher is therefore interested in analyzing data that has changed during the 2019 coronavirus outbreak by studying data with outliers and comparing them with the estimation performance of the model. Therefore, actual data of such was applied to compare the efficiency of coefficient estimation between simple regression and quantile regression models using the data from January 2017 to March 2023, totaling 75 months (Fiscal Policy Office 2023). The data were then divided into 2 subgroups, training and testing datasets. The training dataset was for modelling using the data from January 2017 to December 2021, 60 months in total. The testing dataset was for evaluating the accuracy of the proposed models using the data from January 2022 to March 2023, 15 months in total.

The modeling stage started with scatter diagram to check the data distribution as shown in Figure 5(a). It was found that the data were scattered and some values were outside the criteria, namely outliers. These outliers appeared in 2 periods that are the year 2017 and the year 2020. The first period of the outliers were in February, July and October of year 2017. These might be caused by unusual situations around the world including the global economy slowing down, causing a decrease in purchases, the global oil price fluctuation situation and rising consumption costs. As a result, the price of fresh food products was stable, the manufacturing sector did not expand, and there was no expansion of private investment. Therefore, people were thrifty, turned to save money, and were careful with their spending including a decrease in lending. The overall picture of the supply was declining resulting in slow economic growth. As a result, the collection of income continues to decrease. The second outliers occurred in April, July, October, November, and February 2021. This was a result of the COVID-19 pandemic around the world since the end of 2020 hence such caused severe economic impacts around the world including Thailand. The determination of government policies to help people affected by the COVID-19 epidemic resulted in uncertain spending and product prices. Therefore, the economy contracted and the government revenue collection decreased as well.

Next, the training dataset was used to create the simple regression and the quantile regression models with outliers. After that the mean of regression coefficients, Std.errors, MAE, skewness (Sk), and kurtosis (Ku) of each model were approximated and compared shown in Table 5.

Table 5 Comparison of MAE, Std.errors, Ku and Sk values in each model

Models	QR(10)th		QR(25)th		QR(50)th		QR(75)th		QR(90)th		SR	
Distribution	Ku	Sk	Ku	Sk	Ku	Sk	Ku	Sk	Ku	Sk	Ku	Sk
	1.535	0.958	1.565	0.833	1.572	0.841	1.091	0.571	0.204	0.282	1.231	0.632
Std. errors	195.558		186.350		186.598		183.946		192.132		199.204	
MAE	5.519		3.982		3.557		4.504		6.879		60.367	
$\hat{\beta}_0$	0.063		0.085		0.095		0.120		0.149		10.757	
$\hat{\beta}_1$	0.010		0.009		0.009		0.008		0.007		0.854	

Table 5 states that the least MAE value is from QR(50)th model, 3.557, and the best kurtosis value which is the closest to zero is from QR(75)th model, 1.091. In addition, the values of quantile base scale and skewness were presented in Figure 5(b). The inner and outer scales are "–0.262 and –0.676", respectively and the inner and outer skewness are "–3.475 and –3.777", respectively. It can be seen that the inner and outer values of the quantile base scale are greater than the values of the quantile skewness. This indicates that a change in position has a greater effect on the estimation than a change in shape.

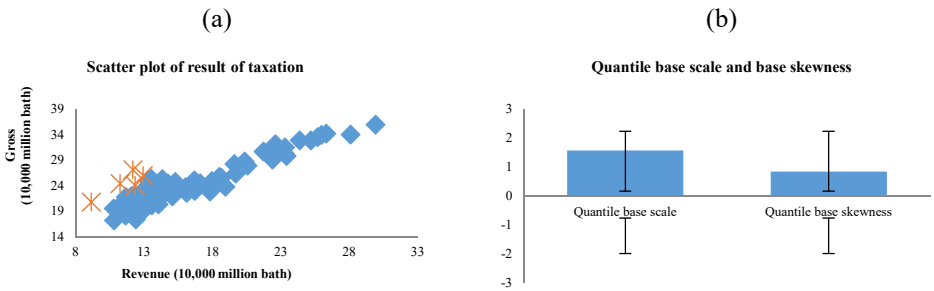

Figure 5 (a) Scatter plot of tax revenue and (b) Quantile base scale and skewness

To verify the accuracy of the proposed models, the testing dataset was applied and then the difference between the actual and the predicted data of each model were calculated and displayed in Table 6.

Table 6 Comparison of percentage of quantile regression model in tests data from January 2022 to March 2023

Methods	QR(10)th	QR(25)th	QR(50)th	QR(75)th	QR(90)th	SR
January	30.467	30.250	31.250	31.334	31.817	31.396
February	26.481	26.662	27.662	28.144	29.026	27.991
March	28.242	28.248	29.248	29.554	30.259	29.496
April	31.031	30.758	31.758	31.784	32.211	31.877
May	39.238	38.144	39.144	38.350	37.957	38.886
June	40.371	39.164	40.164	39.257	38.749	39.853
July	27.862	27.905	28.905	29.249	29.993	29.171
August	32.683	32.244	33.244	33.106	33.368	33.288
September	42.287	40.888	41.888	40.789	40.090	41.489
October	31.667	31.331	32.331	32.294	32.657	32.421
November	29.083	29.005	30.005	30.227	30.848	30.214
December	30.783	30.534	31.534	31.586	32.038	31.665
January	31.382	31.074	32.074	32.066	32.458	32.177
February	26.640	26.806	27.806	28.272	29.138	28.127
March	28.918	28.856	29.856	30.094	30.732	30.073
MAE	0.127	0.124	0.133	0.134	0.138	0.135

Table 6 reveals that the QR(25)th model gave the best predicted values which close to the actual data followed by QR(10)th model and QR(10)th model, respectively. It was also found that the SR model gave the most the difference between the actual and the predicted data. This indicates that if the dataset involve outliers or skewness, the quantile regression model yields the best performance predictive model.

4. Conclusions and Discussions

4.1. Conclusions

Data simulation results under modeling with different sample sizes and vary variances and the application to actual data on government tax collection from January 2017 to March 2023 are summarized as follows:

1. Performance of both SR model and QR model at different percentiles under the generated data yielded similar results as follows:

(1) The comparison results show that the standard deviations of error (Std.errors) and the mean absolute error (MAE) are the same. That is, if the sample size is small, the MAE value of the SR model will close to the value of the QR(50)th model. Also, the SR method will give lower MAE values than the QR(50)th models if the sample size increases. Considering the regression coefficient estimation, it found that the range of the interval estimate of $\hat{\beta}_0$ and $\hat{\beta}_1$ of the QR(50)th model were smaller than the SR model for all sample sizes. This can be observed from the QR(50)th model there was very little change in the regression coefficients from the mean.

(2) The comparison results display that the change in regression coefficients differs from the mean is the same. That is, the regression coefficient estimates of the QR(50)th model were very little difference from the mean compared to the estimates of the SR method. It can be observed from the QR(50)th model that there was very little change in the regression coefficient estimates from the mean for every sample size.

(3) The kurtosis and skewness comparison results state that there was no different. That is, the kurtosis of both SR and QR(r)th models becomes more kurtosis when the small sample size alternates with the moderate sample size where the kurtosis decreases and increases again as the large sample size.

(4) Comparison of the quantile scale values was considered from the position change of the data. It was found that, the inner and outer scale values were wide with small sample size and decreased with increasing sample size. For the skewness of the quantile, it was found that the shape change of the skewness for $n = 50$ yields the most similar inner and outer skewness values.

2. The performance comparison results between the regression model and quantile regression models at different percentiles under the actual data of the taxation of the Revenue Department from January 2017 to March 2023, totaling 75 months, showed that the best model is the QR(25)th model because it gave the best predictive data which close to the actual data. Its MAE value was also the smallest. Therefore, it can be concluded that the quantile regression model is suitable for skewed dataset or dataset with outliers.

4.2. Discussion

This study aims to compare the model efficiency between simple regression and quantile regression models for a dataset with outliers under the simulation of various situation and real data. It was found that the quantile regression models were able to estimate more than one answer whereas the simple regression model focused only on the mean and could not provide suitable estimates for the dataset with outliers. Ali and Nae'l (2024) studies on outliers-detection procedures in binary logistic regression model, which affect the accuracy of model prediction. Thus, the quantile regression model is effective in relationship analysis between the response and independent variables, provides precision estimates more than one answer, and is suitable as an alternative analysis in case of skewed dataset or dataset with outliers in multiple regression. Even with different sample sizes, the quantile regression model in particular low quartile positions such as 25th quantile are more appropriate. If the probability density function is adjusted for the quantile regression coefficient

estimates or there are more options for implementing the kernel functions, the quantile regression coefficient estimation will give more options. For example, parameter estimation was performed using a maximum likelihood estimation method with Fisher's data matrix method. Choosawat et al. (2020) studied the efficiency and compared ridge regression, LASSO, and Adaptive LASSO by using the criterion of median for mean squared error prediction and using two invalid variable selection criteria. In addition, Dawoud (2022) has introduced an improved two-parameter regression estimator called the Dawoud biased regression (DBR) estimator. Moreover, theoretically, the performance of the DBR estimator was compared with the estimator of OLS and some existing estimators on the basis of mean squared error, which could be extended in future studies (Yimnak and Piampholphan 2021).

Acknowledgements

Thank you to Dr.Saravut Sangkawanna for his constant support throughout this research, knowledge, and guidance on the logic for completion of this research. He was a good teacher for me. Thank you very much for making it all possible.

References

Ali HA, Nae'l AA. Comparative study on outliers-detection procedures in binary logistic regression model. Thail Stat. 2024; 22(1): 180-191.

Ampanthong P, Suwattee P. Robust estimation of regression coefficients with outliers. Thail Stat. 2010; 8(2): 183-205.

Araveeporn A, Ghosh SK, Budsaba K. Forecasting the stock exchange rate of Thailand Index by conditional heteroscedastic autoregressive nonlinear model with autocorrelated errors. Thail Stat. 2010; 8(2): 109-122.

Choosawat C, Reangsephet O, Srisuradetchai P, Lisawadi S. Performance comparison of penalized regression methods in poisson regression under high-dimensional sparse data with multicollinearity. Thail Stat. 2020; 18(3): 306-318.

Dawoud I. Modified two parameter regression estimators for solving the multicollinearity. Thail Stat. 2022; 20(4): 842-859.

Devroye L, Gyorfi L. Nonparametric density estimation: the L1 view. New York: John Wiley Sons; 1985.

Fiscal Policy Office, 2023, [cited 2023 Jul 1]; Available from: https://govspending.data.go.th/dashboard/6?language=en.

Härdle W, Hall P, Ichimura H. Optimal smoothing in single-index models. Ann Stat. 1993; 21(1): 157-178.

Hao L, Naiman D. Quantile regression. London: Sage; 2007.

Kleinbaum DG, Kupper LL, Nizam A, Muller KE. Applied regression analysis and other multivariable methods. 4th ed. Belmont, CA: Duxbury; 2008.

Koenker R, Bassett G. Regression quantiles. Econometrica. 1978; 46(1): 33-50.

Kumari SN, Tan A. Modeling and forecasting volatility series: with reference to gold price. Thail Stat. 2018; 16(1): 79-93.

Manoj J, Suresh KK. Application of time-varying coefficient regression model for forecasting financial data. Thail Stat. 2023; 21(1): 180-195.

Ninbai T. Quantile regression. Economics, Ramkhamhaeng University. 2019 [cite 2023 May 8]. Available from: http://www.eco.ru.ac.th/images/gallery/km/KMecon.pdf.

Pagano M, Gauvreau K. Principles of biostatistics. 2nd ed. Duxbury: Pacific Grove; 2000.

Rosenblatt M. Remarks on some nonparametric estimates of a density function. Ann Math Stat. [serial on the internet]. 1956 [cited 2023 May 8]; 27(3): 832-837. Available from https://doi:10.1214/aoms/1177728190.

Yimnak K, Piampholphan K. A Comparison of MM-estimation and fuzzy robust regression for multiple regression model with outliers. Thail Stat. 2021; 19(2): 411-419.