International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

# Eliminate the Heterogeneous Variances Effect Using Quantile Regression

**Wafa Omar El_khafifi[1], Abdelbaset Abdalla[2], Nasir Elmesmari [3], Ahmed M. Mami[4]**

[1]Department of Statistics, Libyan Academy, Benghazi Branch, Benghazi, Libya
[2,4]Department of Statistics, Faculty of Science, University of Benghazi, Benghazi, Libya
[3]Department of Statistics, Faculty of arts and Science/ Al Maraj, University of Benghazi, Benghazi, Libya

**Abstract**
Quantile regression (QR) is a statistical method that addresses the issue of inconsistent data errors. QR utilizes minimum absolute deviation to reduce the absolute deviation by employing percentile estimators such as the median, 1st and 3rd quartiles, and 10th and 90th percentiles. This study focuses on the qth percentile estimators of QR. QR models not only detect varying effects of explanatory variables at different quantiles of the response variable but also provide more robust and accurate estimates compared to mean regression when normality assumptions are violated or when outliers and long tails are present. This study conducts several simulation studies to compare the suggested qth percentile estimators of QR under different sample sizes, explanatory variables, and qth percentiles. Additionally, the study examines the impact of error terms dependency on normal/non-normal distribution. The asymptotic properties of these estimators are also investigated. The study concludes with a discussion of the advantages and disadvantages of using these qth percentile estimators of QR.

**Keywords:** Heterogeneous Variances Effect; Variance Inflation Factor (VIF); Quantile Regression

## 1.introduction

Regression models have three important conditions to be valid for estimation and inference: linearity between response variable and predictor variables, normal distribution of error terms, and homogeneous variances. However, these conditions may not always hold, requiring alternative approaches. In this paper, we focus on heterogeneous variances, which mean that some observations contain more information than others. Ordinary least squares (OLS) assume equal weighting, which can lead to loss of precision in estimation if the variances are not equal. Two approaches to handle heterogeneous variances are transformation and quantile regression (QR) [2].

QR has gained popularity in various fields and is used to address inconsistency in data errors, which OLS cannot handle. QR estimates the model parameters by minimizing absolute deviation using median estimator and other percentiles estimators such as the 1st and 3rd quartiles and the 10th and 90th percentiles.

It is expected that certain types of data will have heterogeneous variances, just like non-normal distributions. This is because in most non-normal distributions, the variance is related to the mean of the

distribution. Even if the underlying distributions are normal within groups, the variances may differ between groups. Typically, groups with larger means will have larger variances. To identify heterogeneous variances, several residual plots can be used, as suggested by [7].

The use of median regression for larger datasets has historically been less popular among statisticians compared to the least squares method because it can be tedious to calculate. However, with the widespread availability of computers in the late 20th century, quantile regression has gained renewed interest from both a theoretical and practical perspective. Quantile regression is a statistical technique that estimates conditional quantile functions by minimizing asymmetrically weighted absolute residuals. This method is analogous to classical linear regression, which minimizes sums of squared residuals to estimate conditional mean functions [9].

Quantile regression can estimate conditional median functions and a full range of other conditional quantile functions. This was proposed by [7]

According to the discussion in [6] paper, using conventional least squares estimators in linear models that are based on non-Gaussian settings can lead to poor results. In such cases, quantile regression provides more robust and efficient estimates. Although there may be a loss of efficiency compared to least squares estimators for data that follows normal distribution, the gain in accuracy for non-Gaussian data outweighs this. As a result, using quantile regression in addition to traditional mean regression models can enhance the overall performance of the models, as stated in [7] paper.

## 2. Quantile Regression Analysis

This section will provide the concept of quantile regression, which will be an important tool for the analysis in this paper. The aim is to establish a basic understanding of the methodology and its relevance to the research question. Following this, the proposed methodology for the analysis will be described in detail, including the statistical techniques, and analytical approaches [6].

**Definition1**(**Quantile**): *Let* Y *be a real valued random variable with cumulative distribution function* $F_Y(y) = P(Y \leq y)$. *For all* $0 \leq \tau \leq 1$ *the* $\tau$*-quantile of Y is given by*

$$Q_Y(\tau) = \mathcal{F}_Y^{-1}(y) = \inf\{y : \mathcal{F}_Y(y) \geqslant \tau\}.$$

From the definition, τ-quantile of a continuous random variable Y is the point at which the area under the PDF curve, from the left to that point, is equal to 1, see Figure (1) for examples.

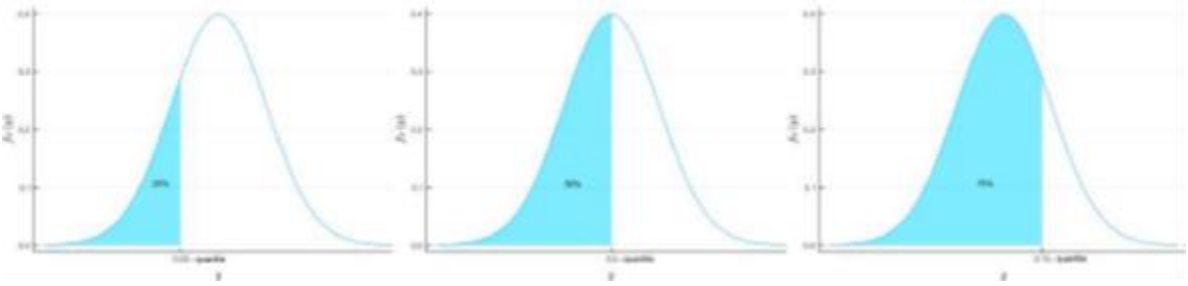


Figure (1): Examples of the quantiles of a Normal distribution.

It is well-known that, in linear regression, the most commonly used loss function is the mean squared error (**MSE**) function. Similar to the linear regression framework, in order to penalize and infer the parameters, we need a loss function for quantile regression:

**Definition 2** (**Quantile loss function**): *Given* $0 \leq \tau \leq 1$, *the quantile loss function is defined as*

$$\ell_\tau(u) = u(\tau - \mathbb{I}_{\{u<0\}}), \quad \forall u \in \mathbb{R},$$

*where* $\mathbb{I}$ *is the indicator function*

Note that the quantile loss function could also be rewritten as

$$\ell_\tau(u) = \begin{cases} \tau u & \text{if } u \geqslant 0 \\ (\tau - 1)u & \text{if } u < 0 \end{cases}.$$

Now, given $\tau \in (0, 1)$, let Y be a real-valued random variable with cumulative distribution function $F_Y(y) = P(Y \leq y)$, the problem under consideration is the minimization of a convex stochastic [8].

## 2. The Basic Quantile Regression Model

The classic model of quantile regression was originally proposed by [6] as a natural extension of the concept of ordinary quantiles in a location model, to a more general class of linear models where the conditional quantiles have a linear form. To briefly explain the concept of ordinary quantiles, consider a real-valued random variable Y that is characterized by the following distribution function.

$$F(y) = Pr(Y \leq y) \qquad (1)$$

Then, for any $\tau \in (0, 1)$, the $\tau$ -th quantile of Y is defined as follows:

$$Q(\tau) = \inf \{y : F(y) \geq \tau\} \qquad (2)$$

The median is then Q(1/2), the first quartile Q(1/4) and the first decile Q(1/10).

The quantiles may be formulated as the solution to a simple optimization problem. For any $0 < \tau < 1$, by defining the piecewise linear function demonstrate by $\rho_\tau(u)$

$$\rho_\tau(u) = u(\tau - I(u < 0)) \qquad (3)$$

where I(.) is the common indicator function. The solution to the minimization problem is then

$$\hat{\alpha}(\tau) = \arg\min_{\xi \in \mathbb{R}} E[\rho_\tau(Y - \xi)] \qquad (4)$$

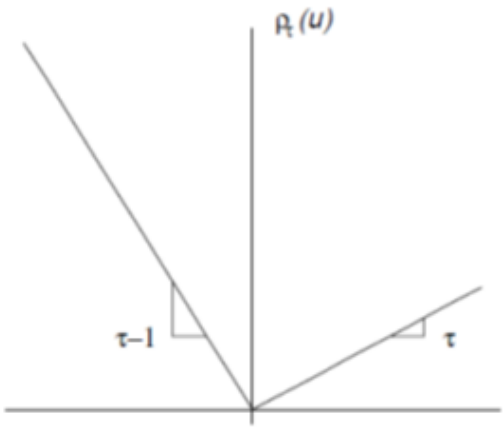



Figure (2) *Quantile Regression _ Function Minimizing the expectation of*

$\rho\tau$

*( Y- ξ)* *with respect to ξ yields solutions*

$\xi\hat{}$

*(τ) the smallest of which is Q(T) defined above.*

The sample analogue of Q(τ), based on a random sample, $\{y_1,\ldots\ldots,y_n\}$, of Y 's, is called the $Q^{th}$ sample quantile. The τ-th quantile can then be identified, and in the spirit of equation (4) above, may be found the solution by solving,

$$\min_{\xi \in \mathbf{R}} \sum_{i=1}^{n} \rho_\tau (y_i - \xi), \tag{5}$$

Let $x_i$, i = 1...n, a (K × 1) vector of repressors. We can then rewrite the equivalent of expression (2.1) as:

$$F_{u_\tau}(\tau - x_i'\beta_\tau | x_i) = \Pr(y_i \leq \tau | x_i) \tag{6}$$

which is basically a different form derived from the more familiar

$$y_i = x_i'\beta_\tau + u_{\tau i} \tag{7}$$

where the distribution of the error term $u^\tau_i$ is left undetermined, the only constraint being the quantile restriction $Q^\tau (u^\tau_i | x_i) = 0$.

In a similar manner of the estimation of conditional mean functions as in

$$\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^K} \sum_{i=1}^{n} (y_i - x_i'\beta)^2 \tag{8}$$

Thus, the linear conditional quantile function

$$Q_Y(\tau | X = x) = x_i'\beta_\tau \tag{9}$$

can be estimated by solving the equivalent of expression (8) for this case:

$$\widehat{\beta}_\tau = \arg\min_{\beta \in \mathbb{R}^K} \sum_{i=1}^{n} \rho_\tau (y_i - x_i' \beta) \qquad (10)$$

**2.1 The Quantile Regression Interpretation**

the least squares estimator of the mean regression model focuses on the dependency of the conditional mean of Y on the explanatory variables X. In contrast, the quantile regression estimator addresses this problem at each quantile of the conditional distribution, providing a more complete description of how the conditional distribution of Y given X = x depends on x. Quantile regression examines the possible effects on the shape of the distribution, rather than assuming that explanatory variables only shift the location or scale of the distribution
[7].

[1] had provided a practical answer to the interpretation of quantile regression coefficients. The coefficients represent the partial derivative of the conditional quantile of Y with respect to one of the explanatory variables, say the jth one. This derivative can be interpreted as the marginal change in the τ-th quantile due to the marginal change in the jth element of X [1].

Furthermore, as stated above in this section, x has *K* distinct variables, then this derivative would simply be the coefficient on the $j^{th}$ variable, βj. One must be careful when interpreting these results. They do not necessarily imply that a subject who happens to be in the $\tau^{th}$ quantile of one conditional distribution would still find itself there, had the corresponding value of x changed [7].

Although, the quantile regression estimates are inherently robust to contamination of the response observations, they could be rather sensitive to contamination of the design (**X**) observations [7].

**2.2 Quantile regression using a linear model**

A simple but very effective way to estimate $Q_Y(\tau)$ by using the linear model, in which τ -quantile of the response variable depends linearly on the features of the explanatory variables. Specifically, in linear model, Q (τ) is estimated by

XTβτ

where

βτ ∈ Rn

the unknown parameter (sometimes called coefficient of weight) is. Therefore, the Optimization problem becomes

$$\underset{\beta_\tau \in \mathbb{R}^n}{\text{minimize}} \sum_{i=1}^{m} \ell_\tau (y_i - x_i^{\mathsf{T}} \beta_\tau). \qquad (11)$$

Now, recall that, we used a single explanatory variable $x_i$ to refer to an observation of the stochastic variable X, that is the $i^{th}$ row of the matrix. So, if

β∗τ

s the solution of equation (11) then the τ -quantile of the response variable Y at an unseen input

x ∈ Rn

can be simply estimate by

$$\mathcal{Q}_Y(\tau) = x^{\mathsf{T}} \beta_\tau^*.$$

Starting from here, the quantile regression can be considered as a linear model [3]

### 2.3 From Quantile Regression to Linear Programming

In this section we will demonstrate that the Optimization problem in equation (11) is basically equivalent to a standard linear programming. By rewriting a quantile regression problem as a Linear Programming, we simultaneously throw out the ambiguous in constrained quantile regression

By definition, the quantile loss functions are able to solve the problem efficiently using simplex method, following the steps:

First: we have by definition of quantile regression the following loss function,

$$\sum_{i=1}^{m} \ell_\tau(y_i - x_i^\intercal \beta_\tau) = \sum_{i=1}^{m} (y_i - x_i^\intercal \beta_\tau)\left(\tau - \mathbb{I}_{\{y_i - x_i^\intercal \beta_\tau < 0\}}\right)$$
$$= \sum_{i=1}^{m} u_i \left(\tau - \mathbb{I}_{\{u_i < 0\}}\right),$$

where $u_i := y_i - x_i^\intercal \beta_\tau$ for all i = 1, 2,…….., n. Therefore, the Optimization problem in equation (11) can be rewritten as

$$\begin{cases} \text{minimize } \sum_{i=1}^{m} \left(\tau u_i^+ + (1-\tau) u_i^-\right) \\ \text{subject to: } \quad y_i - x_i^\intercal \beta_\tau = u_i^+ - u_i^-, \\ \qquad\qquad\quad \beta_\tau \in \mathbb{R}^n, \\ \qquad\qquad\quad u_i^+, u_i^- \in \mathbb{R}, \quad u_i^+, u_i^- \geqslant 0 \quad \forall i = 1, 2, \ldots, n, \end{cases}$$

here, we expressed $u_i$ as the difference of two non-negative variables

$$u_i^+ = \max\{u_i, 0\} \text{ and } u_i^- = -\min\{u_i, 0\}.$$

In order to formulate this Optimization problem to a standard Linear Programming, we again split $\mathbf{B}^T$ into two non-negative variables as we did with $u_i$'s:

$$\beta_\tau = \beta_\tau^+ - \beta_\tau^-, \quad \text{with } \beta_\tau^+ = \max\{\beta_\tau, 0\} \text{ and } \beta_\tau^- = -\min\{\beta_\tau, 0\}.$$

Therefore, the standard Linear Programming for a quantile regression Optimization problem becomes

$$\begin{cases} \text{minimize } \sum_{i=1}^{m} \left(\tau u_i^+ + (1-\tau) u_i^-\right) \\ \text{subject to: } \quad x_i^\intercal \beta_\tau^+ - x_i^\intercal \beta_\tau^- + u_i^+ - u_i^- = y_i, \\ \qquad\qquad\quad \beta_\tau^+, \beta_\tau^- \in \mathbb{R}^n, \quad \beta_\tau^+, \beta_\tau^- \geqslant 0, \\ \qquad\qquad\quad u_i^+, u_i^- \in \mathbb{R}, \quad u_i^+, u_i^- \geqslant 0 \quad \forall i = 1, 2, \ldots, n. \end{cases}$$

This Linear Programming can be rewritten in matrix notation form as

$$(QR) \quad \begin{cases} \text{minimize} & \tau e^{\mathsf{T}} u^{+} + (1-\tau) e^{\mathsf{T}} u^{-} \\ \text{subject to:} & \boldsymbol{X}\beta_{\tau}^{+} - \boldsymbol{X}\beta_{\tau}^{-} + u^{+} - u^{-} = Y, \\ & \beta_{\tau}^{+}, \beta_{\tau}^{-} \in \mathbb{R}^{n} \quad \beta_{\tau}^{+}, \beta_{\tau}^{-} \geqslant 0, \\ & u^{+}, u^{-} \in \mathbb{R}^{m}, \quad u^{+}, u^{-} \geqslant 0, \end{cases}$$

where $e = (1, \ldots, 1)^{\mathsf{T}} \in \mathbb{R}^{m}$.

In this paper, occasionally, the QR is used to refer to the non-standard Linear Programming form

$$(QR) \quad \begin{cases} \text{minimize} & \tau e^{\mathsf{T}} u^{+} + (1-\tau) e^{\mathsf{T}} u^{-} \\ \text{subject to:} & \boldsymbol{X}\beta_{\tau} + u^{+} - u^{-} = Y, \\ & \beta_{\tau} \in \mathbb{R}^{n}, \\ & u^{+}, u^{-} \in \mathbb{R}^{m}, \quad u^{+}, u^{-} \geqslant 0, \end{cases}$$

in which, the unknown parameters are kept as unspecified variables. As an extension to this, one might consider multiple kinds of training sets, and depending on the circumstances, the notations in (QR) could be changed [8].

As a final remark, as interesting and favorable properties of the quantile regression:

Quantile regression offers a complete strategy for regression analysis by going beyond the primary goal of determining only the conditional mean and allowing one to examine the relationship between the response variable and explanatory variables at any quantile of the conditional distribution function. Since the loss function is piecewise linear, solving linear quantile regression is a linear programming problem that uses the standard doubling trick to substitute the absolute values by positivity constraints. Local estimation of conditional regression quantiles can be approximated at a point X = x by the quantile of the training observations in the neighborhood. By overcoming problems with heteroscedasticity and information loss in the tails of a distribution that are often encountered with OLS, quantile regression is a valuable tool for regression analysis. Finally, the trade-offs between quantities of bias and variance in quantile regression are essentially similar to those in the conditional mean least square [5].

## 3. The Simulation Study

We conducted two sets of simulations to compare the performance of two estimators, namely the Ordinary Least Squares (OLS) and Quantile Regression estimators. These estimators were employed to address the issue of heterogeneity in variances. To gain a better understanding of the properties of the Quantile estimator and its alternatives, we performed various computations and generated graphics using the R software package, which is based on the statistical language S (Statistical Science, Inc 2015). We are confident that the results obtained from these simulations provide valuable insights into assessing the practical performance of the two proposed estimators in addressing the problem of collinearity.

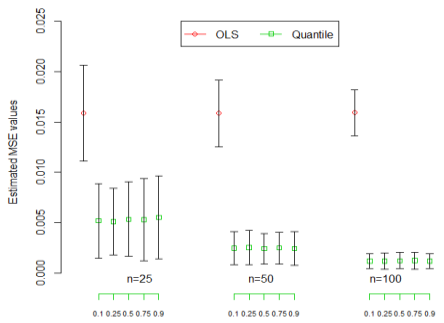
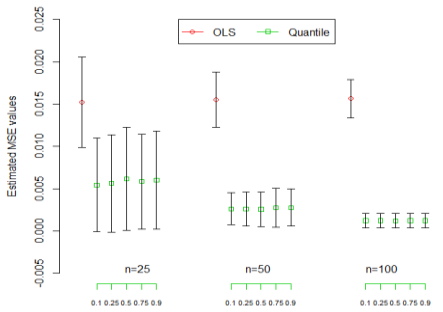
### 3.1 Description of The Experiment

The primary aim of conducting 24 simulation studies was to compare the performance of the Quantile estimator with the OLS estimator using different percentiles (q = 0.10, 0.25, 0.50, 0.75, and 0.90). This comparison aimed to better understand the advantages often associated with Quantile regression estimators, particularly their effectiveness when there are fewer explanatory variables. The simulations were designed to explore the impact of various scenarios of heterogeneity in variances on the regression

model, and they were divided into two distinct settings: Setting 1: The error terms follow a normal distribution and Setting 2: The error terms do not follow a normal distribution.
Within each setting, we explored assorted options for the number of explanatory variables (p = 1, 5, 10, and 20). The sample sizes considered were n = 25, 50, and 100. The regression model followed the form Y=X
β
+
ε
. Additionally, we incorporated two distinct types of marginal distributional errors in the simulations. The first was the normal distribution, while the second involved non-normal distributions. By considering these varied factors, we aimed to examine the performance of the estimators under different conditions and gain a comprehensive understanding of their behavior.

### 3.2 The Results of Setting 1: The error terms follow a normal distribution

The experiment involved investigating various percentiles (q = 0.10, 0.25, 0.50, 0.75, and 0.90) by considering different numbers of explanatory variables (p = 1, 5, 15, and 20) and three different sample sizes (n = 25, 50, and 100). The simulations were performed for 1000 independent runs, and in each run, the marginal errors were distributed as normal. The results obtained from these simulations are presented in the following graphs, supplying insights into the performance of the estimators under different combinations of percentiles, explanatory variables, and sample sizes.

<table>
<tr><td>(a) p=1</td><td>(b) p=5</td></tr>
<tr><td> 
</td><td> 
</td></tr>
<tr><td>(c) p=10</td><td>(d) p=20</td></tr>
</table>

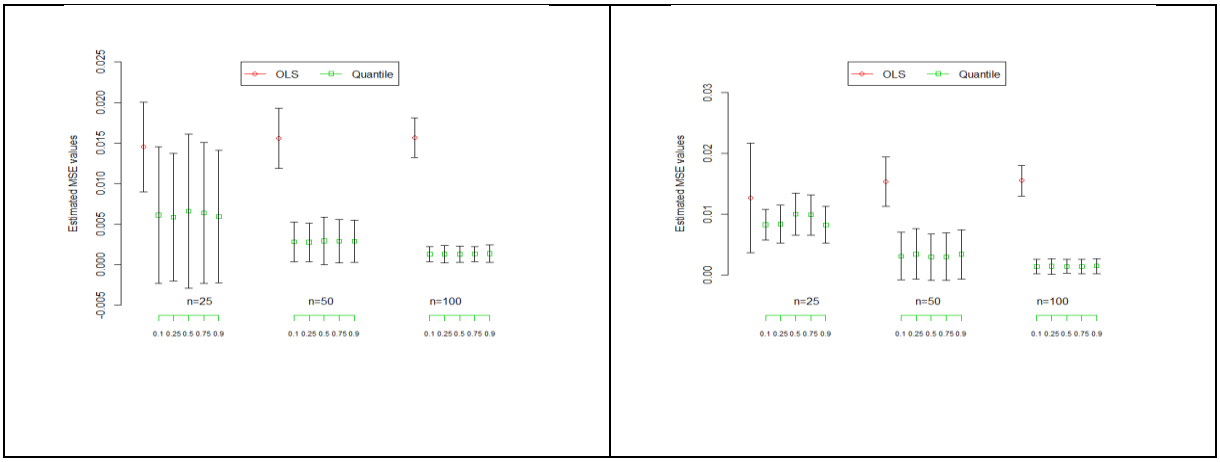



Figure (3): *The Box plots show how various choices of quantiles affect the mean (standard deviations) of the MSE values for the proposed quantile estimators along with the OLS estimator when for p=1,5,10,and 20 assuming the proposed marginal errors belongs to normal distribution.*

Based on the observations from Figure (3), which examines the performance of estimators under normal distribution errors with p = 1, 5, 10, and 20 we can draw the following important conclusion:
Firstly, The OLS estimator consistently produces significantly higher Mean Squared Error (MSE) values compared to its corresponding Quantile estimators. This writes down that the OLS estimator is not practically useful for handling heteroscedastic errors. Additionally, in case P=1, the best estimators and their corresponding MSE values vary depending on the sample size. When the sample size is 25, the 1st quarter estimator achieves the minimum MSE. For a sample size of 50, the Median estimator performs the best, while the 10% quantile estimator shows the highest performance for a sample size of 100. Furthermore, it is noteworthy that as the sample size increases, both the MSE values and their corresponding standard errors consistently decrease. This pattern holds true regardless of the specific sample size or percentile used.

Secondly, the Quantile estimators consistently demonstrate the best performance regardless of the number of independent variables (P), whether it is 5, 10, or 20. The minimum MSE values and their corresponding standard error values vary based on the sample size (n) and the specific estimator used. For example, when the sample size is 25, the 10% quantile estimator, 1st quarter estimator, Median estimator, and 3rd quarter estimator each achieve their minimum MSE values at P = 5. Similarly, for a sample size of 50, these estimators reach their minimum MSE values at P = 5. Meanwhile, the 90% quantile estimator attains its lowest MSE value at P = 10. Furthermore, when the sample size increases to 100, the 10% quantile estimator, 1st quarter estimator, Median estimator, and 3rd quarter estimator all show their minimum MSE values at P = 5.

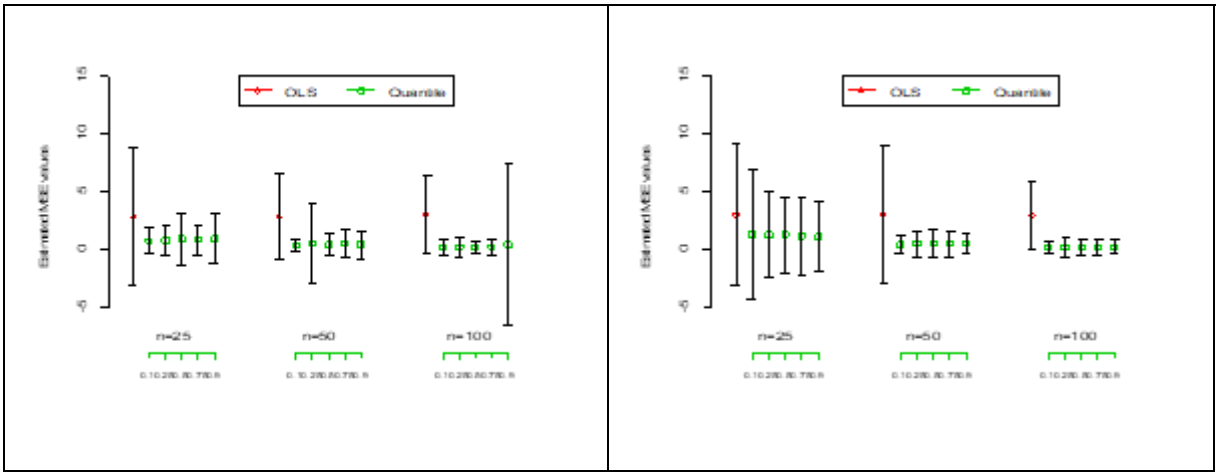
Thirdly, regardless of the specific percentile chosen (10th quantile, 1st quarter, Median, 3rd quarter, or 90th quantile) or the sample size (n), the best performance of the Quantile estimators is consistently achieved at P = 5. The minimum MSE values and their corresponding standard error values vary depending on the percentile and sample size. Regardless of the chosen percentile or the number of independent variables (P), increasing the sample size (n) leads to a consistent decrease in both the MSE values and their standard error values.

Conversely, increasing the number of independent variables consistently results in an increase in both the MSE values and their standard error values, regardless of the chosen percentile or the sample size (n). These findings highlight the impact of sample size and the number of independent variables on the accuracy and precision of the estimators.

### 3.3 The Results of Setting 2: The error terms do not follow a normal distribution.

The experiment involved testing different percentiles (0.10, 0.25, 0.50, 0.75, and 0.90) with varying numbers of explanatory variables (1, 5, 10, and 20) and three different sample sizes (25, 50, and 100). The errors in the experiment were assumed to follow a non-normal distribution, and the experiment was repeated 1000 times independently. The results of this experiment are summarized in graphical form in Figures 3.8, 3.11, 3.12, and 3.14, which display the outcomes for different choices of correlation and sample sizes for each of the four cases (p = 1, 5, 10, and 20).

(a) p=1 (b) p=5




(c) p=10 (d) p=20

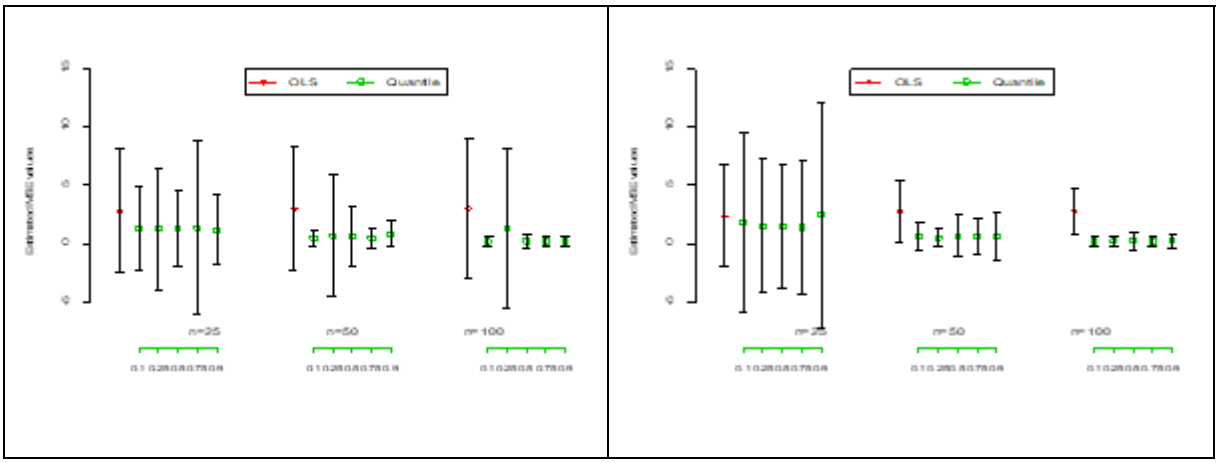



Figure (4): *The Box plots show how various choices of quantiles affect the mean (standard deviations) of the MSE values for the proposed quantile estimators along with the OLS estimator when for p=1,5,10,and 20 assuming the proposed marginal errors belongs to non-normal distribution.*

Figures (4), focusing on the scenario where errors follow a non-normal distribution, the following observations were made: Firstly, regardless of the number of independent variables (P) or the sample size (n), it is evident that the OLS estimator consistently yields higher Mean Squared Error (MSE) values compared to all its corresponding Quantile estimators. This shows that the OLS estimator is not suitable for use in cases of heteroscedastic errors.

Therefore, it is strongly recommended to avoid its usage under such conditions. Secondly, in the context of the second simulation study, which aims to investigate the impact of non-normality on the regression model using the proposed Quantile estimators, direct comparisons were made regarding the performance of the various Quantile estimators based on the selection of P and n. These comparisons provide insights into the effectiveness of the Quantile estimators in the presence of non-normal errors, considering different combinations of independent variables and sample sizes.

Table (1): The values for the minimum Mean Squared Error corresponding to different number of independent variables (**P**), sample size(n), and "Quantile" for the normally and non-normally distributed errors.

*Table (1): The values for the minimum Mean Squared Error corresponding to different number of independent variables (**P**), sample size(n), and "Quantile" for the normally and non-normally distributed errors.*

| | | ***normally distributed error*** | | ***non-normally distributed error*** | |
|---|---|---|---|---|---|
| P | n | Quantile | MSE | Quantile | MSE |
| **1** | 25 | 1st Quarter | 0.0051 | 1st Quarter | 0.7640 |
| | 50 | Median | 0.0024 | Median | 0.3750 |
| | 100 | 10% Quantile | 0.0018 | 10% Quantile | 0.2023 |
| **5** | 25 | 10% Quantile | 0.0054 | 10% Quantile | 1.1225 |
| | 50 | Median | 0.0026 | Median | 0.4350 |
| | 100 | Median | 0.0012 | Median | 0.2192 |
| **10** | 25 | 1st Quarter | 0.0059 | 1st Quarter | 1.1662 |
| | 50 | 1st Quarter | 0.0027 | 1st Quarter | 0.4920 |
| | 100 | 10% Quantile | 0.0013 | 10% Quantile | 0.2255 |
| **20** | 25 | 90% Quantile | 0.0082 | 90% Quantile | 1.4105 |
| | 50 | Median | 0.0029 | Median | 0.5344 |
| | 100 | 3rd Quarter | 0.0013 | 3rd Quarter | 0.2308 |

Table 3.1 provides comparing the Mean Squared Error (MSE) values for normally distributed error and non-normally distributed error, we can draw the following conclusions:

**1. Effect of number of independent variables (P):**

- As the number of independent variables (P) increases from 1 to 20, the MSE values generally increase for both the normally distributed and non-normally distributed errors. This indicates that larger parameter values result in higher prediction errors.

**2. Effect of Sample Size (n):**

- Increasing the sample size from 25 to 100 generally leads to lower MSE values for both the normally distributed and non-normally distributed errors. This suggests that larger sample sizes tend to improve the accuracy of predictions.

**3. Comparison between Normally Distributed and Non-normally Distributed Errors:**

- Overall, the MSE values for the non-normally distributed errors tend to be higher than those for the normally distributed errors across different number of independent variables of (P) and sample sizes (n). This suggests that when errors deviate from a normal distribution, they introduce more variability and uncertainty into the predictions, leading to elevated levels of prediction errors.

**4. Comparison of Quantiles:**

- Within each number of independent variables (P) and sample size (n), different quantiles of the error distribution exhibit varying levels of prediction accuracy. For example, the 10% quantile generally has higher MSE values compared to the median or other quantiles, indicating that the lower end of the error distribution contributes more to the prediction errors.

In summary, the analysis of the Mean Squared Error (MSE) values suggests that non-normally distributed errors can significantly impact the accuracy of predictions compared to normally distributed errors. Additionally, larger parameter values and smaller sample sizes tend to increase prediction errors. Therefore, it is important to consider the distribution of errors and the sample size when evaluating and improving the performance of regression models.

## 4. Summary and Conclusion

In conclusion, the study conducted simulation studies to compare the performance of quantile regression (QR) estimators at different quantiles of the response variable. The study focused on the qth percentile estimators of QR and examined their performance under various conditions. The results of the study indicate that QR models offer advantages over mean regression when normality assumptions are violated or when outliers and long tails are present in the data. The qth percentile estimators of QR not only detect varying effects of explanatory variables at different quantiles but also provide more robust and accurate estimates.

**The study identified several key findings:**

1. **Effect of number of independent variables (P):** Increasing the number of independent variables generally resulted in higher mean squared error (MSE) values for both normally distributed and non-normally distributed errors. This suggests that larger parameter values lead to higher prediction errors. Therefore, it is important to carefully consider the number of independent variables when building regression models.
2. **Effect of sample size (n):** Increasing the sample size generally led to lower MSE values for both normally distributed and non-normally distributed errors. This implies that larger sample sizes improve the accuracy of predictions. Researchers and practitioners should strive to obtain an adequate sample size to enhance the reliability of QR models.
3. **Comparison between normally distributed and non-normally distributed errors:** The study found that non-normally distributed errors generally resulted in higher MSE values compared to normally distributed errors. This indicates that errors deviating from a normal distribution introduce more variability and uncertainty into the predictions, leading to elevated levels of prediction errors. It

is crucial to assess the distributional properties of errors and consider appropriate modeling techniques when dealing with non-normal data.

4. **Comparison of quantiles:** Different quantiles of the error distribution exhibited varying levels of prediction accuracy. The 10% quantile generally had higher MSE values compared to the median or other quantiles, suggesting that the lower end of the error distribution contributes more to the prediction errors. Researchers should pay attention to the specific quantiles they are interested in and evaluate their performance accordingly.

Overall, the study highlights the importance of considering the distribution of errors and the sample size when using QR models. It emphasizes that non-normally distributed errors can significantly impact the accuracy of predictions compared to normally distributed errors. Additionally, larger parameter values and smaller sample sizes tend to increase prediction errors. By understanding these factors and making appropriate adjustments, researchers and practitioners can enhance the performance of regression models in real-world applications.

**References**

1. Buchinsky, M. (1998) Recent advances in quantile regression models: A practical guide for empirical research, Journal of Human Resources 33, pp. 88-126
2. Draper, N.R and Smith, H. (1998). *Applied Regression Analysis*, 3rd edition. John Wiley and Sons, Inc.
3. Furno, M., and Vistocco, D. (2018) Quantile Regression: Estimation and Simulation. Hoboken, NJ: John Wiley and Sons.
4. Koenker, R. (2005). Quantile regression. New York : Cambridge University Press.
5. Koenker, R., and Bassett, G. (1978) Regression quantiles, Econometrica 46, pp. 33-50.
6. Koenker, R., and Bassett, G. (1982) Robust Tests for Heteroscedasticity Based on Regression Quantiles, Econometrica 50, issue 1, pp. 43-61.
7. Koenker, R. and Hallock, K. (2001), 'Quantile regression', Journal of Economic Perspectives 15, pp. 143—156.
8. Pham, Nhat-Thien (2018) Quantile Regression in Large Energy Datasets, Unpublished M.Sc. Thesis, Universite Paris-Sud, France
9. Rawlings, John O., Pantula, Sastry G., and Dickey, David A. (2001) Applied regression analysis: a research tool. Springer Science & Business Media.