Monte Carlo Simulation in Quantile Regression Parameter for Sparsity Estimate

A Hapsery\textsuperscript{1}\textsuperscript{*} and A B Tribhuwaneswari\textsuperscript{2}
\textsuperscript{1}Statistic Department, PGRI Adi Buana University, Surabaya-60234, Indonesia
\textsuperscript{2}Urban and Regional Planning Department, PGRI Adi Buana University, Surabaya-60234, Indonesia

*Email : alfisyahrina@unipasby.ac.id

Abstract. Monte Carlo is a method used to generate data according to the distribution and resampling until the parameters of the method used became converge. The purpose of this simulation is first to prove that quantile regression with the estimated sparsity function parameter can model the data according to the non-uniform distribution of the data. Secondly, it’s to prove that the quantile regression is a developed method from the linear regression. The pattern of data which is not uniform is generally referred to as heterogeneous data, while the pattern of uniform data distribution is called homogeneous data. Data in this study will be generated for small and large samples on homogeneous and heterogeneous data. Uniformity of variance will be carried out on both heterogeneous and homogeneous data types, namely 0.25, 1 and 4. The parameter estimation process and data generation will be resampled 1000 times. Thus, in conclusion of the simulation studies was the parameter estimates in the classical regression will be the same as the parameter estimates in the quantile regression at quantile 0.5. In the simulation, it is decided that the quantile regression method can be used on heterogeneous and homogeneous data to the unconstrained number of samples and variances.

Keywords: Monte carlo simulation, quantile regression, simulation studies, sparsity

1. Introduction
Monte Carlo simulation is an algorithm for generating data from a distribution. Monte Carlo simulations contain randomized iterative simulations associated with modelling a data \cite{1}. A method can be justified theoretically based on data driven facts that are carried out. The algorithm to do this is called the Monte Carlo simulation. In the Monte Carlo simulation there is also a resampling process, a procedure of modelling data and generating parameters where the practice of modelling to produce these parameters is repeated as many times as desired by the researcher \cite{2}. Prior to the resampling stage, the most important and necessary algorithm in Monte Carlo is to generate data according to the distribution of a data driven method. In statistical science, statisticians will generate data from errors in the model and generate data for independent variables. Methods in the world of statistics are not only quantile regression and classical regression that can be applied to Monte Carlo, all groups of science in the field of statistics can use the Monte Carlo method to find and prove a driven theory. In this study, the focus is on simulating heterogeneous data using the quantile regression method.

Quantile regression is a parameter estimation modeling method that could overcome non-uniform distribution of data. Quantile regression was first discovered \cite{3} which is an extension of the
median quantile regression method, where the conditional quantile distribution of the response variables is expressed as a function of the observed covariates. Quantile regression is excellent to utilise when one desire to understand the level of change in certain conditional quantiles expressed by their coefficient values [4]. In quantile regression the principle used is to minimize the absolute number of residuals known as Least Absolute Deviation (LAD). Meanwhile, parameter estimation can be done with three approaches, namely direct, rank-score, and resampling methods. In the direct method of interval estimation in this study, the sparsity function is used which is different from the rank-score approach which uses the rank-score test. While the resampling method uses interval estimation with bootstrap technique. Quantile regression is a robust nonparametric method, it could be utilized in large quantity of samples [5].

The nature of the quantile regression with three approaches through simulation studies has been studied by [6]. The results of this study indicate that direct produces a robust estimator against non-uniform data distribution and provides the most efficient results for computationally narrower interval estimates. Various interval estimation methods have been developed, such as research by [1] and research conducted by [6]. Based on previous research, there has been no simulation based on the type of data, the amount of data and the data variance.

Based on this description, this study focuses on examining the advantages of the quantile regression method carried out by simulation on heterogeneous and homogeneous data types, the amount of data are low, medium and large. The variance of the data is expected to affect the quantile regression method in terms of modeling the data. In addition, we also want to prove, theoretically quantile regression is formed from median regression, therefore classical regression is believed to be part of quantile regression at the 0.5 or 50% quantile. The estimate of the interval used is the direct sparsity function which is believed to be an interval estimate with a good range. With a variety of theoretical information about quantitative regression, this research is expected to provide results of the advantages of the theoretical method that has been described.

2. Research method
The method used in this study consists of a combination of several methods, namely multiple linear regression and quantile regression with the interval estimate used is the direct interval estimate of the sparsity function.

2.1 Linier regression
Regression analysis discusses the relationship between one variable and another. Multiple regression analysis is a statistical method that can be used to analyse the relationship between one dependent variable and several independent variables [7]. By generalizing the two and three variable linear regression models, the dependent Y variable and p variable regression model with independent X1, X2, ..., Xp variables can be written

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_p X_{pi} + \epsilon_i,$$

with i = 1, 2, ..., n and \(\beta_0\) as intercept, \(\beta_i\) to \(\beta_p\) as slope, \(\epsilon_i\) as residual.

2.2 Quantile regression
Quantile regression was first introduced by [3]. This method is an extension of the conditional quantile regression model in which the conditional quantile distribution of the dependent variable is expressed as a function of the observed covariates [8]. With this approach, it is possible to estimate the quantile function of the conditional distribution of the dependent variable on each quantile value according to the desired quantile [4]. Because of its robustness to outlier data, quantile regression is highly recommended to analyses a number of data that are not symmetrical and have an inhomogeneous distribution. Estimated intervals in quantile regression can be done by several methods, namely the direct approach, rank-score, and resampling [9]. The general linear quantile regression equation specifically for the conditional quantile \(Q_{\tau} (\tau \mid X_{1i}, X_{2i}, \ldots, X_{pi})\) from dependent \(Y_i\) variable, such as:
with \( i = 1,2,\ldots,n \). The optimal solution for quantile regression is as follows:

\[
\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} \rho(\tau) \left( y_i - X_i^T \hat{\beta}(\tau) \right), \quad i = 1,2,\ldots,n \quad \tau \in (0,1)
\]

where \( y_i = \{ y_1, y_2, \ldots, y_n \} \) is a random sample with the dependent variable \( Y \) and \( x_i \in \mathbb{R}^p \) is a covariance vector, while \( \rho(\tau) = \left( \tau - I(\tau < 0) \right) \) holds for \( 0 < \tau < 1 \) which is an asymmetric loss function of \( \varepsilon \) which is the residual of the parameter estimate [9].

Regression is estimated by minimizing the number of squares of residuals, while quantile regression will minimize the absolute number of residuals which is better known as Least Absolute Deviation (LAD) [10]. The product of the residual with the given weight forms a loss function \( \rho(\tau) \) such as:

\[
\rho(\tau)(\varepsilon) = \sum_{i=1,\varepsilon_i \neq 0}^{n} \tau | \varepsilon_i | + \sum_{i=1,\varepsilon_i = 0}^{n} (1-\tau) | \varepsilon_i |.
\]

Thus, in quantile regression there is a conditional quantile function \( \tau \) which considers the estimator \( \hat{\beta}(\tau) \), so that the solution to the problem is obtained which is stated as follows:

\[
\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} \rho(\tau) \left( y_i - Q(\tau)(y \mid X) \right); \tau \in (0,1)
\]

with :

\begin{align*}
\rho(\tau) & : \text{asymmetric loss function} \\
\tau & : \text{index quantile with} \ \tau \in (0,1) \\
Q(\tau)(y \mid X) & : \text{quantile function} \ \tau \text{ of variable} \ Y \ \text{with the provision of} \ X.
\end{align*}

The simplex method is a way to determine the optimal combination of three or more variables, where this algorithm can provide solutions to linear programming problems with computations and involve many variables [11].

\[
\max_{z} \left\{ y^T z \mid X^T z = 0, z \in [-1,1]^p \right\},
\]

which is transformed, we get the equation (7)

\[
\min_{\beta} \sum_{i=1}^{n} \rho(\tau) \left( y_i - x_i^T \beta(\tau) \right).
\]

2.3 Estimating the direct interval of the sparsity function

The method in quantile regression used to calculate the confidence interval \( \beta \) one of them is the sparsity function. On each \( \tau \) that has been determined, estimate the interval with the following formula [11] :

\[
P\left( \hat{\beta}_j(\tau) - t_{(\alpha/2,df_j)} \text{se}(\hat{\beta}_j(\tau)) \leq \beta_j(\tau) \leq (\hat{\beta}_j(\tau) + t_{(\alpha/2,df_j)} \text{se}(\hat{\beta}_j(\tau))) = 1 - \alpha.
\]

\( \text{se}(\hat{\beta}_j(\tau)) \) value is a diagonal value achieved by estimating the matrix diagonal of \( \hat{\sigma}^2(\tau)D^1 \), where \( D^1 = (X^T X)^{-1} \). For \( F \) is the cumulative distribution function and \( f = F^{-1} \) is a density function. The square form of the matrix \( X \) assumed to converge to a positive definite matrix \( D \) as follows:

\[
D = \lim_{n \to \infty} \frac{1}{n} (X^T X)^{-1}.
\]
As for the matrix \( D \) in equation (9) is obtained from the following steps:

\[
D = \lim_{n \to \infty} \frac{1}{n}(X^T X)^{-1}.
\]

\[
= \lim_{n \to \infty} \frac{1}{n} \begin{bmatrix}
1 & 1 & \cdots & 1 \\
X_{11} & X_{12} & \cdots & X_{1n} \\
X_{21} & X_{22} & \cdots & X_{2n} \\
\vdots & \vdots & \ddots & \vdots \\
X_{p1} & X_{p2} & \cdots & X_{pn}
\end{bmatrix} \begin{bmatrix}
1 & X_{21} & \cdots & X_{p1} \\
1 & X_{12} & \cdots & X_{p2} \\
\vdots & \vdots & \ddots & \vdots \\
1 & X_{1n} & \cdots & X_{pn}
\end{bmatrix}
\]

The next step is to get the regression quantile estimator from the coefficient vector 
\( \hat{\beta}(\tau) = [\beta_0(\tau), \beta_1(\tau), \ldots, \beta_p(\tau)]^T \). Thus:

\[
\sqrt{n} \left[ \hat{\beta}(\tau) - \beta(\tau) \right] \to N\left(0, \omega^2(\tau)D^{-1}\right). \tag{10}
\]

Where \( \omega^2(\tau) \) for each quantile \( \tau \) defined as follows [12]:

\[
\omega^2(\tau) = \frac{(1-\tau)}{f\left(F^{-1}(\tau)\right)}. \tag{11}
\]

For \( F \) is the cumulative distribution function and is a density function, then the sparsity function is denoted as follows:

\[
S(\tau) = f\left(F^{-1}(\tau)\right)^{-1}. \tag{12}
\]

On each \( \tau \) that has been determined, the confidence interval with the formula as in equation (13) below:

\[
P\left( \hat{\beta}_j(\tau) - t_{(a/2,d)},se(\hat{\beta}_j(\tau)) \leq \beta_j(\tau) \leq (\hat{\beta}_j(\tau) + t_{(a/2,d)},se(\hat{\beta}_j(\tau)) \right) = 1 - \alpha. \tag{13}
\]

\( se(\hat{\beta}_j(\tau)) \) value is the diagonal value obtained from the calculation of the matrix \( \omega^2(\tau)D^{-1} \).

3. Results and analysis

This research simulation study was conducted in the case of univariate quantile regression with the aim of proving theoretically the advantages of the quantile regression method. According to Koenker, quantile regression is used to model data that has a non-uniform distribution of data. With regard to the types of heterogeneous and homogeneous data, the distribution of data that is not uniform can be categorized into heterogeneous data types. This simulation study is not only to prove the type of data, but also to prove whether quantile regression can also model data that has various data variance values, the data variances used are 0.25, 1 and 4. The third simulation study is aimed at proving whether quantile regression also can well model the data based on the difference in the amount of data that is 50, 150 and 500 data. Each of these three types of simulations will be modeled based on classical regression and quantile regression with Monte Carlo simulation where the \( k \) replication.

Monte Carlo is a computational algorithm where this simulation method can help in proving the advantages of a method by sampling repeatedly on a data from the results of a particular distribution. The first idea was discovered by [13] in his research on neutron radiation, which later got a solution by modeling experiments with computer aids. Monte Carlo is the basis for all algorithms of the simulation method whose basic principle is to provide as much value as possible from the generated data and be able to evaluate it repeatedly. The procedure for performing simulations with Monte Carlo are:

1. Generating data as independent variables with normal and uniform distribution 
   \( X_i \sim U(a, b), \) and \( X_j \sim N(\mu, \sigma^2) \)
2. Generating data for errors with distribution $\varepsilon_i \sim U(a, b)$, and $\varepsilon_i \sim N(\mu, \sigma^2)$

3. Determine the parameter values in the model in the method used $\beta_0 = 10$ and $\beta_1 = 5$.

4. Modeling according to the homogeneous model design $Y_i = 10 + 5X_{1i} + \varepsilon_i$, and heterogeneous $Y_i = 10 + 5X_{1i} + \exp^{0.5X_i} \varepsilon_i$.

5. Calculating the parameters of the modeling results

6. Repeating procedures 1 to 4 as many as the 1000 replication specified in this study.

7. Calculating the accuracy of the average parameter values

### 3.1 Simulation study based on data

The first step in simulating data based on the type of data is first generating data from the distribution. The distribution of data for classical regression is a normal distribution, while for quantile regression the appropriate data distribution is a Uniform distribution. The amount of data used in this sub-chapter is 500 data. The model used in this study is:

#### Homogeneous data

$$Y_i = 10 + 5X_{1i} + \varepsilon_i,$$

#### Heterogeneous data

$$Y_i = 10 + 5X_{1i} + \exp^{0.5X_i} \varepsilon_i.$$

Where $X$ is the independent variable, $\varepsilon_i$ is the random error of the data distribution, $\beta$ is a parameter of the regression model both quantile regression and classical regression.

In heterogeneous data, the error is multiplied by the exponential of the variable value[14]. These two types of data will be modelled using two methods, namely classical regression and quantile regression. The expected conclusion is that the classical regression centred on the mean of the data distribution will be equal in value to the quantile regression on the 0.5 or 50% quantile.

![Figure 1. Scatter plot data replication to – 1000 from residual with homogeneous data simulation study.](image1)

![Figure 2. Scatter plot data replication to 1000 from residual with heterogeneous data simulation study.](image2)

Figure 2 shows that the residual variance is not constant. This means that the distribution of the data from the simulation results is heterogeneous or the distribution of the data is not uniform. Figure 1 above shows that the residual variance is constant, which means...
that the distribution of the data from the simulation results is homogeneous. The next modelling uses classical and quantile regression whose estimation results are as in Table 1.

Table 1. Homogeneous data modeling with classical and quantile regression on 500 samples

|                | Classical Regression |               | Quantile Regression |               |
|----------------|----------------------|---------------|---------------------|---------------|
|                | $\hat{\beta}_p$ Value | $\tau$      | $\hat{\beta}_p(\tau)$ | Value | Accuration |
| 10%            | $b_{0.01}$           | 3.59         | 100%                | $b_{0.01}$ | 4.99        |
|                | $b_{0.1}$            | 4.99         | 100%                | $b_{0.1}$ | 4.99        |
| 25%            | $b_{0.25}$           | 6.62         | 100%                | $b_{0.25}$ | 4.99        |
|                | $b_{0.25}$           | 4.99         | 100%                | $b_{0.25}$ | 4.99        |
| 50%            | $b_{0.5}$            | 9.99         | 100%                | $b_{0.5}$ | 9.99        |
|                | $b_{0.5}$            | 9.99         | 100%                | $b_{0.5}$ | 9.99        |
| 75%            | $b_{0.75}$           | 13.3         | 100%                | $b_{0.75}$ | 4.99        |
|                | $b_{0.75}$           | 4.99         | 100%                | $b_{0.75}$ | 4.99        |
| 90%            | $b_{0.9}$            | 16.37        | 100%                | $b_{0.9}$ | 16.37       |

Replication in the homogeneous error simulation study was carried out as much as 1000, then the average value of all parameter values was $\bar{\beta}_0 = 9.98$ and $\bar{\beta}_1 = 4.99$. This value is between the lower and upper limits, indicating the correctness of the simulation study results. It can be seen that all the estimated values are between the upper and lower limits with an accuracy rate of 100%. This indicates that the simulation results are appropriate. In addition, it can also be seen that the intercept value is getting bigger and bigger in line with the increase in the quantile value. At the time of quantile 50%, the mean value of the intercept in the quantile regression is identical to the intercept value in the linear regression. Other information is that the 50% quantile is identical to the slope of the linear regression. As shown in Figures 3 and 4.

The model for the 50%, and 95% quantiles can be written as follows:
Model for 50% quantile: $Q_{0.50}(y \mid X) = 9.992 + 4.998X_i$.
Model for 95% quantile: $Q_{0.95}(y \mid X) = 18.19 + 4.999X_i$. 

![Figure 3. Parameter value boxplot $\hat{\beta}_0$ on quantile regression for homogeneous error simulation studies](image1)

![Figure 4. Parameter value boxplot $\hat{\beta}_1$ on quantile regression for homogeneous error simulation studies](image2)
3.2 Simulation study based on number of samples

The second simulation study was conducted to find out whether quantile regression can model data with a variety of data ranging from small to large data. The data generated is data from residuals with homogeneous errors of 50, 150, and 500 data. The replication in this simulation study is the same as the previous simulation study, namely 1000. The model according to model 20 obtained the following results:

| Sample size | \( \hat{\beta}_0 \) Value | Accuracy |
|-------------|-----------------|----------|
| 50          | 10.04           | 100%     |
|             | 4.97            | 100%     |
| 150         | 9.98            | 100%     |
|             | 5.01            | 100%     |
| 500         | 10.01           | 100%     |
|             | 4.994           | 100%     |

The parameter estimation values in classical regression are already in the lower and upper limits, indicating high accuracy and correctness of the simulation study results. But none \( \beta_1 \) are significant at the 5% level. Visually, parameter estimates at sample sizes of 50, 150 and 300 provide the same visualization, namely, the estimated intercept value of the boxplot image shows an increasing pattern as the quantile increases. It can be concluded that the amount of data does not affect the accuracy of quantile regression in modeling data whose distribution is not uniform.

![Figure 5. Boxplot intercept of heterogeneous error with number of samples 50](image)

![Figure 6. Boxplot intercept of heterogeneous error with number of samples 150](image)

![Figure 7. Boxplot intercept of heterogeneous error with number of samples 500](image)
3.3 Simulation study based on data variance
The third simulation study was conducted to find out whether quantile regression can model data with data variances ranging from small to large data. The amount of data used is 500 with differences in variance of 0.25, 1 and 4. From Figure 8, it is known that quantile regression can still produce intercept values according to theory, increasing with time with a small variance of 0.5. The difference in variance will only visually indicate that the resulting image is getting wider or narrower. The three variances produce the same pattern with different boxplot shapes, what is meant by different is the width of the boxplot image which is determined by the magnitude of the variance.

![Boxplot of heterogeneous error intercept with variance 0.25](image1)

**Figure 8.** Boxplot of heterogeneous error intercept with variance 0.25

![Boxplot of heterogeneous error intercept value with variance 1](image2)

**Figure 9.** Boxplot of heterogeneous error intercept value with variance 1

![Boxplot of heterogeneous error intercept with variance 4](image3)

**Figure 10.** Boxplot of heterogeneous error intercept with variance 4

4. Conclusion
Quantile regression can model data in a variety of data states. With the simplex algorithm and Monte carlo simulation, it is known that quantile regression can not only model accurately on homogeneous data whose data distribution is around the average and model accurately on heterogeneous data whose data distribution does not spread around the data average. In addition, the difference in the amount of data and variance does not affect the accuracy of the quantitative regression. Further study about the theory of quantile regression could refer to nonstandard quantile regression for it was an estimate method which parameter will be more robust [15].

5. References
[1] Chen D 2017 Monte Carlo Simulation based Statistical modeling University of Pretoria
[2] Binder K and Heermann D W 2002 Monte Carlo Simulation in Statistical Physics An Introduction Springer.
[3] Koenker R and Bassett G 1978 Regression quantiles Econometrica p 33
[4] Chen C and Wei Y 2005 Computational Issues for Quantile Regression The Indian Journal of Statistics 67 No 2 p 399
[5] Noor N and Muhammad A 2013 Model of Robust Regression with Parametric and Nonparametric Methods Mathematical Theory and Modeling 3 No 5 p 27
[6] Zhou K Q and Portnoy S L 1996 Direct Use of Regression Quantile to Construct Confidence Sets in Linear Models The Annals of Statistics 24 No 1 p 287

[7] Gujarati D N 2004 Basic Econometrics 4th edition United States of America: The McGraw-Hill Companies

[8] John O O and Nduka E C 2009 Quantile Regression Analysis as a Robust Alternative to Ordinary Least Square The Journal of Scientia Africana 8 No 2 p 61

[9] Koenker R 2005 Quantile Regression First published Cambridge University Press

[10] Wardani D K 2014 Penerapan Regresi Kuantil pada Data dengan Pelanggaran Asumsi Kenormalan Sisaan Statistic Department Brawijaya University

[11] Davino C and Furno M 1998 Quantile Regression Theory and Application SPi Publishers Services, Pondicherry, India.

[12] Hapsery A 2017 Regresi Kuantil Berbasis Model Rekursif dan Estimasi Sparsity untuk Publikasi Dosen ITS di Scopus Statistic Department ITS

[13] Fermi E 1966 Neutron Physics, A Course of Lecture University of ChicagoPress

[14] Navianti D R 2014 Regresi Kuantil untuk Pemodelan Tingkat Pengangguran Terbuka di Indonesia Statistic Department ITS

[15] Goh S C and Knight K 2008 Nonstandard Quantile Regression Inference Econometric Theory 25 p 1415