
**Electronic Journal of Applied Statistical Analysis**
**EJASA, Electron. J. App. Stat. Anal.**
http://siba-ese.unisalento.it/index.php/ejasa/index
e-ISSN: 2070-5948
DOI: 10.1285/i20705948v9n1p17

# Testing the equality of two parametric quantile regression curves: the application for comparing two data sets

By Tonggumnead

Published: 26 April 2016

*Electronic Journal of Applied Statistical Analysis*
Vol. 09, Issue 01, April 2016, 17-39
DOI: 10.1285/i20705948v9n1p17

# Testing the equality of two parametric quantile regression curves: the application for comparing two data sets

Unchalee Tonggumnead*

*Division of Applied Statistics, Department of Mathematics and Computer Science, Faculty of Science and Technology*
*Rajamangala University of Technology Thanyaburi (RMUTT)*
*39 Moo 1, Rangsit-Nakhonnayok Rd., Klong 6, Thanyaburi, Pathumthani, 12110, Thailand*

Published: 26 April 2016

This study aims to compare the different between two data sets that having the relationship between the dependent and independent variables at each quantile using testing the equality of two parametric quantile regression (QR), the conditional quantile regression and the conditional mean regression function are considered. The influence of the distribution of errors that heavy tailed is also examined through a test statistic that is in the form of the empirical distribution function (EDF), applying the bootstrapping principle in the estimation of the critical value of the test statistic. In addition, comparing the equality of two quantile regression functions at the extream quantile are appied with the actual data. The results show that the type I error and power of the test properties becomes better as the sample size increases. However, with variables that heavy-tailed distribution of errors, the conditional median regression function is more robust. An analysis of the actual data indicates consistent findings.

**keywords:** quantile regression function, conditional mean, conditional median, empirical distribution function, robust regression.

*Corresponding author:unchalee_t@rmutt.ac.th


ISSN: 2070-5948
http://siba-ese.unisalento.it/index.php/ejasa/index

# 1 Introduction

Regression analysis is a statistical tool commonly used to explain the relationship between a dependent or explained variable Y and an independent or explanatory variable X using the mathematical function Y = f(x), where Y represents a conditional expected value, and the mean of Y-values is represented by $E(Y|X)$. The function can be employed in comparing the conditional means of two independent sets of data. However, this approach provides only information on the conditional mean but not other types of important information such as the distribution of data. Another problem with the conditional mean regression function is that the distribution of a dependent variable Y may be heavy-tailed, the distribution of errors may be heavy-tailed, asymmetric, or not unimodel, outlier etc. To solve such problems, the conditional mean regression analysis is extended to the conditional quantile of the response variable (Koenker and Bassett, 1978) and (Koenker and Hallock, 2001). There are several other approaches to comparing regression functions, including testing the equality of conditional means based on the parametric and the nonparametric regression models as well as testing the equality of conditional quantile regression functions. In this research we focus on comparing the quantile regression function when the distribution of errors are heavy tailed. The idea about testing the equality of regression curve such as: Kutner et al. (2005) tested the equality of two linear regression models using the equation $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2} + \varepsilon_i$, i=1,...,n: $n = n_1 + n_2$; $n_1, n_2$=the number of data in the first and the second sets, respectively; $Y_i$=dependent variable; $X_{i1}$= independent variable; $Y_i$ and $X_{i1}$=quantitative data; $X_{i2}$=dummy variable,the value of which equals to 0 for the first set of data and 1 for the second set of data. They also specified three conditions for their model. First, the relationships between two sets of data have to be linear. Second, the errors variance of the two sets of data must not differ. Third, F-test is used to determine the equality of the regression curves corresponding to the two sets of data. King et al. (1991) used a kernel smoother and fixed points with equal intervals in their analysis of a test statistic involving two nonparametric regression curves. Kulasekera (1995) examined a test statistic based on the Quasi-residuals technique and the estimators of variance of the errors distribution. Pardo et al. (2007) evaluated a test statistic by comparing the difference between two empirical distribution functions of errors, estimating the errors of each regression function with the equation $\hat{\varepsilon_{ij}} = \frac{Y_{ij} - \hat{m}_j(X_{ij})}{\hat{\sigma}_j X_{ij}}$ , and estimating the regression curves with the Nadaraya-Watson estimator. Sun (2006) studied the application of consistent nonparametric tests for analyzing the equality of unknown conditional quantile curves and employed the wild bootstrap in estimating the critical value of the test statistic. Kuruwita et al. (2014) examined the equality of nonparametric quantile regression functions based on the marked empirical process. Bera et al. (2014) tested the equality of regression functions by comparing the mean regression and the quantile regression coefficient, applying the asymptotic joint distribution in formulating the test statistic. Tonggumnead and Seangngam (2015) compared slops testing in simple linear regression between using F testing and $Z_{KS}$ test derived from the empirical distribution function (EDF) of errors.This study considered where Y represents a conditional ex-

pected value, and the mean of Y is represented by $E(Y|X)$. Despite their efforts to deal with the problems inherent in the conditional regression function, researchers have not arrived at an effective solution. The present study is aimed to comparing two data sets in each quantile through testing the equality of two quantile regression functions. Because, sometime we would like to compare two data sets that having the relationship between dependent variable and independent variables only the hight values or only low values, the problem about the distribution of errors and the distribution of dependent variable Y etc. So, the regression function at each quantile is important. In this research, the test statistic based on empirical distribution of errors are considered. The estimator of the errors in each quantile regression is: $\varepsilon_{ij\tau} = (Y_{ij} - f_j(X_{ij}, \hat{\beta(\tau)})$. The empirical distribution of errors of each regression function is determined using the equation $\hat{F}_{\varepsilon j\tau}(y) = \frac{1}{n_j}\sum_{i=1}^{n_j} I(Y_{ij} - f_j(X_{ij}, \hat{\beta(\tau)}) \leq y), j = 1, 2, i = 1, ..., n_j, -\infty < y < \infty, \tau\epsilon(0, 1)$, where $f_j(X_{ij}, \hat{\beta}(\tau))$ represents the conditional quantile regression function of Y given X for the set of data j, with the quantile regression function being estimated using a parametric quantile regression function. The idea of the test statistic is that: If the empirical distribution of errors in each quantile regression function between two sets of data is similar, the quantile regression function in each group will be equal, then two data sets have similar relationship between dependent and independent variables. On the other hand, if the empirical distribution of errors in each quantile regression function between two sets of data be different, the quantile regression function in each group will not be equal, then two data sets be different relationship between dependent and independent variables. In addition, the impact of distribution of errors that heavy tailed is analyzed, and the application of the data are also included in the next section.

## 2 Materials and Method

Quantile regression is a type of regression analysis that is very useful when the rate of change in the conditional quantile can be explained using the regression coefficient depend on the quantile. For a random variable Y with the probability distribution function $F(y) = P(Y \leq y)$, the $\tau^{th}$ quantile of Y can be determined using the inverse function $Q(\tau) = inf[y : F(y) \geq \tau]$. Let the observations for two independent sets of data be in the form of $(X_{ij},\ Y_{ij}), i = 1, ..., n_j j = 1, 2$, For $\tau\epsilon(0, 1)$, the conditional quantile function $Q(\tau|X = x) = f(X, \beta(\tau))$ can be estimated using the equation $\hat{\beta}(\tau) = argmin_{\beta\epsilon R}\sum_{i=1}^{n} \rho_\tau(Y_{ij} - f_j(X_{ij}, \beta(\tau)))$ (Chen, 2005). In this research, the quantile regression function model is represented by:

$$\hat{\beta(\tau)} = argmin_{\beta\epsilon R}\sum_{i=1}^{n} \rho_\tau(Y_{ij} - f_j(X_{ij}, \beta(\tau))), i = 1, ..., n_j, j = 1, 2. \quad (1)$$

Where $f_j(X_{ij}, \beta(\tau))$ represents the $\tau^{th}$ conditional quantile function of Y given X for the $j^{th}$ population. For the present study, the comparison of the difference between two independent sets of data is conducted by testing the equality of two quantile regression functions, applying the following hypothesis $H_0 : f_1(X_{ij}, \beta(\tau)) = f_2(X_{ij}, \beta(\tau))$

versus $H_1 : f_1(X_{ij}, \beta(\tau)) \neq f_2(X_{ij}, \beta(\tau))$. If the empirical distribution of errors in each quantile regression function between two sets of data is similar, the quantile regression function in each group will be equal (accept $H_0$). Namely, two data sets have similar relationship between dependent and independent variables. On the other hand, if the empirical distribution of errors in each quantile regression function between two sets of data be different, the quantile regression function in each group will not be equal (accept $H_1$). Namely, two data sets be different relationship between dependent and independent variables. The idea of the test statistic we apply follow Tonggumnead and Seangngam (2015). The principles for testing the hypothesis are that $f_1(X_{ij}, \beta(\tau))$,$f_2(X_{ij}, \beta(\tau))$, $j = 1, 2; i = 1, ..., n_j$, represents the quantile regression function for the first and the second sets of data, and $f(X_{ij}, \beta(\tau))$ represents the common quantile regression function under the null hypothesis, and estimated from the two sets of data combined. When the different between the empirical distribution function of errors of the common quantile regression ($\hat{F}^0_{\varepsilon\tau}(y)$) and the empirical distribution function of errors in each quantile regression ($\hat{F}_{\varepsilon j\tau}(y)$), $j = 1, 2$) not different, the null hypothesis is confirmed. Namely, $f_1(X_{ij}, \beta(\tau)) = f_2(X_{ij}, \beta(\tau) = f(X_{ij}, \beta(\tau))$. A comparison of the errors of each population is performed using the two-dimensional process $\hat{G(y)} = (\hat{G_{1\tau}}(y), \hat{G_{2\tau}}(y))$: where $\hat{G_{j\tau}}(y) = n_j^{1/2}(\hat{F_{\varepsilon\tau}}^0(y) - \hat{F_{\varepsilon j\tau}}(y))$, $j = 1, 2; i = 1, ..., n_j$,and the estimation of empirical distribution function of errors in each quantile regression is: $\hat{F_{\varepsilon j\tau}}(y) = \frac{1}{n_j}\sum_{i=1}^{n_j} I(Y_{ij} - f_j(X_{ij}, \hat{\beta}(\tau)) \leq y)$,$j = 1, 2; i = 1, ..., n_j$, $-\infty < y < \infty$. The estimation of empirical distribution function of errors when the null hypothesis is confirmed is: $\hat{F_{\varepsilon\tau}}^0(y) = \frac{1}{n_j}\sum_{i=1}^{n_j} I(Y_{ij} - f(X_{ij}, \hat{\beta}(\tau)) \leq y)$,$j = 1, 2; i = 1, ..., n_j$, $-\infty < y < \infty$.The Kolmogorov-Smirnov type statistic $A_{ks\tau} = \sum_{j=1}^{2} sup_y|\hat{G}_{j\tau}|(y)$ is applied in testing the equality of the two quantile regression functions, following Pardo et al. (2007), if the empirical distributiom functions (EDF) of the errors of the two quantile regression functions do not differ, the null hypothesis is confirmed. On the other hand, if the EDF different, the alternative hypothesis is confirmed. In this study, assume the estimator of the error in each quantile regression function: $\hat{\varepsilon_{ij\tau}} = Y_{ij} - f_j(X_{ij}, \hat{\beta}(\tau))$, $j = 1, 2; i = 1, ..., n_j$; the estimator of the errors under the null hypothesis $\hat{\varepsilon^0_{ij\tau}} = Y_{ij} - f(X_{ij}, \hat{\beta}(\tau))$, where $f(X_{ij}, \hat{\beta}(\tau))$ represents the common predicted value under the null hypothesis. The common quantile regression function under the null hypothesis is estimated from the two sets of data combined. $f_1(X_{ij}, \hat{\beta}(\tau))$ and $f_2(X_{ij}, \hat{\beta}(\tau))$ are the estimators of the first and the second quantile regression functions respectively.

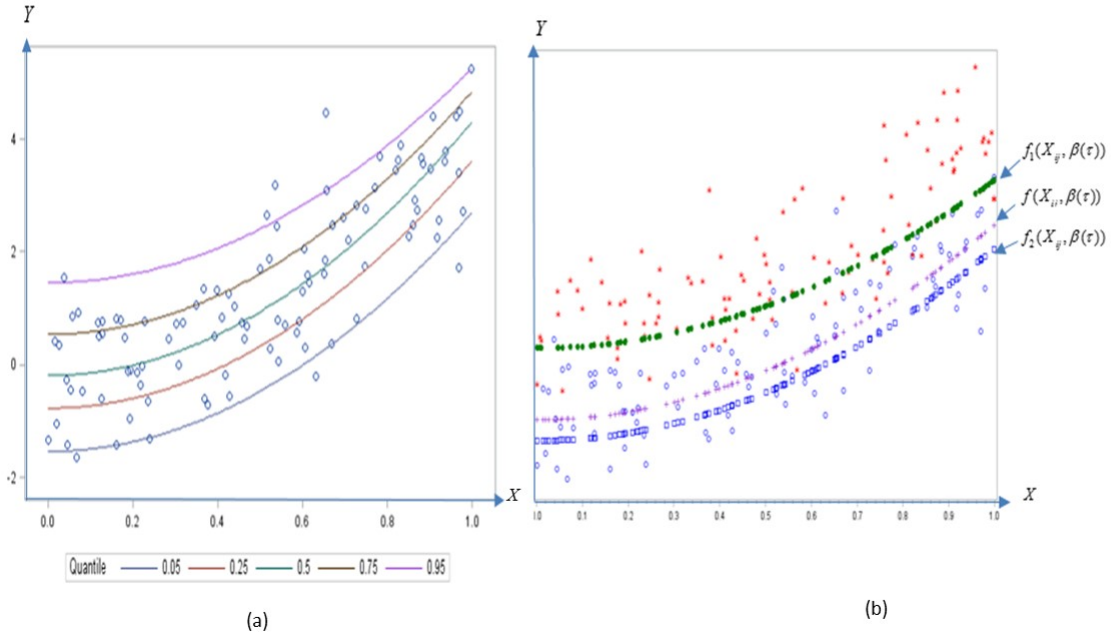


Figure 1: (a)quantile regression curves correspond with model $f(x) = 4x^2$ and the lines from down to upper are $\tau = 0.05, 0.25, 0.50, 0.75, 0.95$.(b) plot of two data sets, stars present the model $f_1(x) = 4x^2 + 2$, circles present the model $f_2(x) = 4x^2$, $f_1(X_{ij}, \beta(\tau))$, $f_2(X_{ij}, \beta(\tau))$ and $f(X_{ij}, \beta(\tau))$ are the first quantile regression function, the second quantile regression function, and the common quantile regression function with $\tau = 0.25$ respectively.

## 3 Bootstrap and Simulation Studies

The present study follows Pardo et al. (2007), Freedman (1981), Silverman and Young (1987), Akritas and Van Keilegom (2001), Lloyd (2014), Efron and Tibshirani (1994) applying bootstrapping for estimating the critical value of the test statistic $A_{ks\tau}$, because bootstrap can be use to estimate the critical value when the distribution of the test statistics are unknown, and the p-value from bootstrap more accurate than standard p - value. In addition the estimators from bootstrap are consistent . The procedures are as follows:

Assume the bootstrap replication b = 1,..., B (B=300), for j =1, 2, i = 1,..., $n_j$, and construct the new response under the null hypothesis form the equation:

$$Y^*_{ij,b} = f_j(\underline{X}, \underline{\beta}) + \varepsilon^*_{ij,b}, j = 1, 2, i = 1, ..., n_j. \tag{2}$$

From equation (2), the quantile regression function $Y^*_{ij,b} = f_j(\underline{X}, \underline{\beta}(\tau)) + \varepsilon^*_{ij\tau,b(\tau)}$ can be determined by assigning $\tau = 0.05, 0.25, 0.50, 0.75$ and 0.95. In this study, we determine the six regression function $(f_j(\underline{X}, \underline{\beta}))$ for testing the equality of the quantile regression functions. The first three are applied when the null hypothesis is confirmed, whereas the other three are applied when the alternative hypothesis is confirmed.

(1) $f_1(x) = f_2(x) = 4x$.
(2) $f_1(x) = f_2(x) = 4x^2$.
(3) $f_1(x) = f_2(x) = 4e^x$.
(4) $f_1(x) = 4x, f_2(x) = 4x + 2$.
(5) $f_1(x) = 4x^2, f_2(x) = 4x^2 + 2$.
(6) $f_1(x) = 4e^x, f_2(x) = 4e^x + 2$.

The distribution of errors $\varepsilon_{ij,b}$, $i = 1, ..., n_j, j = 1, 2$ are as follows: 1)Let 100% of the distribution of errors of each set of data are the standard normal distribution, $\varepsilon_{i1} \sim N(0, 1)$ and $\varepsilon_{i2} \sim N(0, 1)$, $i = 1, ..., n_j, j = 1, 2$. 2) The errors distribution have heavy tailed, let 95% of the distribution of errors of each set of data are the standard normal distribution, and let the remaining 5% of the distribution of errors are the Cauchy distribution, and 3)The errors distribution have heavy tailed, let 90% of the distribution of errors of each set of data are the standard normal distribution, and let the remaining 10% of the distribution of errors are the Cauchy distribution. The probability density function of the Cauchy distribution is in the form of $f(x) = \frac{1}{\pi b[1+(\frac{X-a}{b})^2]}$, a=0 and b=1. For j =1, 2, i = 1,.., $n_j$, calculate the value of the test statistics $A_{ks\tau}$ from the bootstrappling samples $X_{ij}$, $Y^*_{ij,b}$ in each $\tau^{th}$ conditional quantile function. Let $A^*_{ks\tau,b}$ be the order statistics of $A^*_{ks\tau(1)}$,...,$A^*_{ks\tau(b)}$ and $A^*_{ks\tau(1-\alpha)B}$ approximate the $(1-\alpha)$-quantile of the distribution of $A_{ks\tau}$ under the null hypothesis. Calculate the value of the test statistic when the $\tau^{th}$ conditional quantile: $\tau = 0.05$, 0.25, 0.50, 0.75 and 0.95. Iterate the test statistic $A_{ks\tau}$ 1,000 times. Display the rejection proportion when the null hypothesis is confirmed and when the alternative hypothesis is confirmed.

## 4 Results

As for the ratio of rejection proportion when the null hypothesis is confirmed under different conditions, it is found that the type I error is closer to 0.05 when the sample size is larger. In addition, the figure is higher when the $\tau$ conditional quantile equals 0.25, 0.50, and 0.75 than when it is equal to 0.05 and 0.95. However, when the distribution of errors are heavy-tailed the ratio of hypothesis rejection slightly declines from (95% standard normal errors+ 5% Cauchy errors) to (90% standard normal errors + 10% Cauchy errors) for model (1), model (2), and model (3). So, the impact of the heavy tailed distribution of errors, the condition quatile regression function is more robust. The result of rejection proportion are displayed in Table 1, Table 2, and Table 3.

Table 1: Rejection proportions under the null hypothesis (type I error) of model ( 1) : $f_1(x) = f_2(x) = 4x, \alpha = 0.05$ .

| sample size | $\tau$ | 100% N(0,1) | 95% N(0,1)+5%Cauchy | 90% N(0,1)+10%Cauchy |
|---|---|---|---|---|
| (20,20) | 0.05 | 0.034* | 0.034* | 0.032* |
| | 0.25 | 0.042 | 0.041 | 0.041 |
| | 0.50 | 0.041 | 0.041 | 0.041 |
| | 0.75 | 0.043 | 0.040 | 0.040 |
| | 0.95 | 0.033* | 0.035* | 0.034* |
| (60,60) | 0.05 | 0.037 | 0.035* | 0.034* |
| | 0.25 | 0.048 | 0.045 | 0.044 |
| | 0.50 | 0.045 | 0.043 | 0.042 |
| | 0.75 | 0.043 | 0.040 | 0.040 |
| | 0.95 | 0.035* | 0.034* | 0.034* |
| (100,100) | 0.05 | 0.039 | 0.037 | 0.036 |
| | 0.25 | 0.050 | 0.048 | 0.048 |
| | 0.50 | 0.048 | 0.047 | 0.047 |
| | 0.75 | 0.048 | 0.046 | 0.045 |
| | 0.95 | 0.039 | 0.039 | 0.038 |

*The type I error out of control interval.

Table 2: Rejection proportions under the null hypothesis (type I error) of model ( 2) : $f_1(x) = f_2(x) = 4x^2, \alpha = 0.05$ .

| sample size | $\tau$ | 100% N(0,1) | 95% N(0,1)+5%Cauchy | 90% N(0,1)+10%Cauchy |
|---|---|---|---|---|
| (20,20) | 0.05 | 0.037 | 0.037 | 0.035* |
| | 0.25 | 0.037 | 0.038 | 0.037 |
| | 0.50 | 0.041 | 0.040 | 0.040 |
| | 0.75 | 0.041 | 0.040 | 0.039 |
| | 0.95 | 0.032* | 0.030* | 0.030* |
| (60,60) | 0.05 | 0.039 | 0.037 | 0.037 |
| | 0.25 | 0.042 | 0.040 | 0.040 |
| | 0.50 | 0.045 | 0.043 | 0.042 |
| | 0.75 | 0.046 | 0.044 | 0.043 |
| | 0.95 | 0.037 | 0.034* | 0.035* |
| (100,100) | 0.05 | 0.042 | 0.041 | 0.041 |
| | 0.25 | 0.046 | 0.044 | 0.044 |
| | 0.50 | 0.048 | 0.047 | 0.047 |
| | 0.75 | 0.040 | 0.040 | 0.040 |
| | 0.95 | 0.039 | 0.036 | 0.035* |

*The type I error out of control interval.

Table 3: Rejection proportions under the null hypothesis (type I error) of model (3) : $f_1(x) = f_2(x) = 4e^x, \alpha = 0.05$ .

| sample size | $\tau$ | 100% N(0,1) | 95% N(0,1)+5%Cauchy | 90% N(0,1)+10%Cauchy |
|---|---|---|---|---|
| (20,20) | 0.05 | 0.032* | 0.030* | 0.030* |
| | 0.25 | 0.033* | 0.029* | 0.029* |
| | 0.50 | 0.035* | 0.034* | 0.033* |
| | 0.75 | 0.037 | 0.035* | 0.035* |
| | 0.95 | 0.031* | 0.028* | 0.028* |
| (60,60) | 0.05 | 0.035* | 0.034* | 0.033* |
| | 0.25 | 0.037 | 0.035* | 0.035* |
| | 0.50 | 0.042 | 0.040 | 0.041 |
| | 0.75 | 0.046 | 0.044 | 0.042 |
| | 0.95 | 0.037 | 0.035* | 0.035* |
| (100,100) | 0.05 | 0.042 | 0.041 | 0.041 |
| | 0.25 | 0.046 | 0.045 | 0.045 |
| | 0.50 | 0.048 | 0.047 | 0.047 |
| | 0.75 | 0.040 | 0.039 | 0.039 |
| | 0.95 | 0.039 | 0.032* | 0.033* |

*The type I error out of control interval.

With regards to the rejection proportion when the alternative hypothesis is confirmed under the different conditions, the findings reveal that the power of the test is closer to 1.00 when the sample size is larger. Additionally, the figure is higher when the $\tau^{th}$ conditional quantile is equal to 0.25, 0.50, and 0.75 than when it equals 0.05 and 0.95. However, when the distribution of errors are heavy-tailed the ratio of hypothesis rejection slightly declines from (95% standard normal errors+ 5% Cauchy errors) to (90% standard normal errors + 10% Cauchy errors), for model (4), model (5), and model (6). Namely, quatile regression function more robust as the distribution of errors are heavy tailed. The result of rejection proportion are displayed in Table 4, Table 5, and Table 6.

Table 4: Rejection proportions under the alternative hypothesis (power of the test) of model (4) : $f_1(x) = 4x, f_2(x) = 4x + 2, \alpha = 0.05$ .

| sample size | $\tau$ | 100% N(0,1) | 95% N(0,1)+5%Cauchy | 90% N(0,1)+10%Cauchy |
|---|---|---|---|---|
| (20,20) | 0.05 | 0.700 | 0.675 | 0.673 |
| | 0.25 | 0.750 | 0.730 | 0.730 |
| | 0.50 | 0.780 | 0.772 | 0.770 |
| | 0.75 | 0.690 | 0.700 | 0.690 |
| | 0.95 | 0.600 | 0.590 | 0.590 |
| (60,60) | 0.05 | 0.730 | 0.710 | 0.710 |
| | 0.25 | 0.900 | 0.880 | 0.878 |
| | 0.50 | 0.800 | 0.800 | 0.798 |
| | 0.75 | 0.820 | 0.800 | 0.800 |
| | 0.95 | 0.740 | 0.725 | 0.723 |
| (100,100) | 0.05 | 0.790 | 0.784 | 0.784 |
| | 0.25 | 0.930 | 0.920 | 0.918 |
| | 0.50 | 0.890 | 0.877 | 0.877 |
| | 0.75 | 0.900 | 0.889 | 0.885 |
| | 0.95 | 0.790 | 0.770 | 0.770 |

Table 5: Rejection proportions under the alternative hypothesis (power of the test) of model (5) : $f_1(x) = 4x^2, f_2(x) = 4x^2 + 2, \alpha = 0.05$ .

| sample size | $\tau$ | 100% N(0,1) | 95% N(0,1)+5%Cauchy | 90% N(0,1)+10%Cauchy |
|---|---|---|---|---|
| (20,20) | 0.05 | 0.650 | 0.640 | 0.640 |
| | 0.25 | 0.780 | 0.760 | 0.757 |
| | 0.50 | 0.800 | 0.790 | 0.788 |
| | 0.75 | 0.780 | 0.770 | 0.770 |
| | 0.95 | 0.670 | 0.660 | 0.658 |
| (60,60) | 0.05 | 0.741 | 0.730 | 0.730 |
| | 0.25 | 0.900 | 0.888 | 0.885 |
| | 0.50 | 0.870 | 0.870 | 0.868 |
| | 0.75 | 0.850 | 0.840 | 0.838 |
| | 0.95 | 0.752 | 0.746 | 0.746 |
| (100,100) | 0.05 | 0.800 | 0.788 | 0.785 |
| | 0.25 | 0.920 | 0.909 | 0.905 |
| | 0.50 | 0.890 | 0.870 | 0.870 |
| | 0.75 | 0.910 | 0.900 | 0.890 |
| | 0.95 | 0.770 | 0.760 | 0.760 |

Table 6: Rejection proportions under the alternative hypothesis (power of the test) of model (6) : $f_1(x) = 4e^x, f_2(x) = 4e^x + 2, \alpha = 0.05$ .

| sample size | $\tau$ | 100% N(0,1) | 95% N(0,1)+5%Cauchy | 90% N(0,1)+10%Cauchy |
|---|---|---|---|---|
| (20,20) | 0.05 | 0.652 | 0.644 | 0.640 |
| | 0.25 | 0.780 | 0.770 | 0.770 |
| | 0.50 | 0.700 | 0.780 | 0.768 |
| | 0.75 | 0.765 | 0.770 | 0.770 |
| | 0.95 | 0.672 | 0.650 | 0.648 |
| (60,60) | 0.05 | 0.742 | 0.739 | 0.743 |
| | 0.25 | 0.895 | 0.890 | 0.888 |
| | 0.50 | 0.870 | 0.870 | 0.868 |
| | 0.75 | 0.848 | 0.834 | 0.845 |
| | 0.95 | 0.747 | 0.740 | 0.745 |
| (100,100) | 0.05 | 0.791 | 0.788 | 0.786 |
| | 0.25 | 0.918 | 0.913 | 0.910 |
| | 0.50 | 0.888 | 0.880 | 0.878 |
| | 0.75 | 0.909 | 0.898 | 0.896 |
| | 0.95 | 0.768 | 0.765 | 0.763 |

As for the difference in the rejection proportion between the conditional median and the conditional mean regression functions (bracketed numbers) under the null hypothesis, it is found that when the distribution of errors form a standard normal distribution, the test statistic derived from the conditional mean regression function yields a relatively low type I error value close to 0.05. Additionally, the test statistic performs better as the sample size becomes larger. However, when the distribution of errors are heavy-tailed : 1) 95% standard normal errors+ 5% Cauchy errors, and 2) 90% standard normal errors + 10% Cauchy errors, the ratio of hypothesis rejection slightly declines from (95% standard normal errors+ 5% Cauchy errors) to (90% standard normal errors + 10% Cauchy errors) for model (1), model (2), and model (3). Moreover, this impact is found to be greater for the test statistic derived from the conditional mean regression function than for that obtained from the conditional median regression function. The results are displayed in Table 7.

Table 7: Rejection proportion under the null hypothesis (type I error) compare with the test statistic from the conditional median and conditional mean,$\alpha = 0.05$.

| sample size | model | 100% N(0,1) | 95% N(0,1)+5%Cauchy | 90% N(0,1)+10%Cauchy |
|---|---|---|---|---|
| (20,20) | (1) | 0.041 | 0.041 | 0.041 |
| | | (0.042) | (0.035*) | (0.030*) |
| (60,60) | | 0.045 | 0.043 | 0.042 |
| | | (0.049) | (0.040) | (0.039) |
| (100,100) | | 0.048 | 0.047 | 0.047 |
| | | (0.049) | (0.044) | (0.045) |
| (20,20) | (2) | 0.041 | 0.040 | 0.040 |
| | | (0.043) | (0.038) | (0.035*) |
| (60,60) | | 0.045 | 0.043 | 0.042 |
| | | (0.047) | (0.041) | (0.040) |
| (100,100) | | 0.048 | 0.047 | 0.047 |
| | | (0.049) | (0.041) | (0.040) |
| (20,20) | (3) | 0.035* | 0.034* | 0.033* |
| | | (0.040) | (0.032*) | (0.028*) |
| (60,60) | | 0.042 | 0.040 | 0.041 |
| | | (0.047) | (0.039) | (0.035*) |
| (100,100) | | 0.048 | 0.047 | 0.047 |
| | | (0.049) | (0.044) | (0.040) |

*The type I error out of control interval.

As regards the difference in the rejection proportion between the conditional median regression function and the conditional mean regression function (bracketed numbers) under the alternative hypothesis, it is found that when the distribution of errors form a standard normal distribution, the test statistic derived from the conditional mean regression function yields a high power of the test value close to 1.00. Additionally, the test statistic performs better with a larger sample size. However, when the distribution of errors are heavy-tailed : 1) 95% standard normal errors+ 5% Cauchy errors and 2) 90% standard normal errors + 10% Cauchy errors, the ratio of hypothesis rejection slightly declines from (95% standard normal errors+ 5% Cauchy errors) to (90% standard normal errors + 10% Cauchy errors) for model(4), model(5), and model(6). Furthermore, the impact of heavy-tailed distribution of errors are greater for the test statistic derived from

the conditional mean regression function than for that obtained from the conditional median regression function. The results are displayed in Table 8.

Table 8: Rejection proportion under the alternative hypothesis (power of the test) compare with the test statistic from the conditional median and conditional mean,$\alpha = 0.05$.

| sample size | model | 100% N(0,1) | 95% N(0,1)+5%Cauchy | 90% N(0,1)+10%Cauchy |
|---|---|---|---|---|
| (20,20) | (4) | 0.780 | 0.772 | 0770 |
| | | (0.800) | (0.770) | (0.765) |
| (60,60) | | 0.800 | 0.800 | 0.798 |
| | | (0.830) | (0.794) | (0.790) |
| (100,100) | | 0.890 | 0.877 | 0.877 |
| | | (0.920) | (0.870) | (0.865) |
| (20,20) | (5) | 0.800 | 0.790 | 0.788 |
| | | (0.800) | (0.780) | (0.778) |
| (60,60) | | 0.870 | 0.870 | 0.868 |
| | | (0.900) | (0.862) | (0.860) |
| (100,100) | | 0.890 | 0.870 | 0.870 |
| | | (0.910) | (0.868) | (0.865) |
| (20,20) | (6) | 0.700 | 0.780 | 0.768 |
| | | (0.800) | (0.775) | (0.770) |
| (60,60) | | 0.870 | 0.870 | 0.868 |
| | | (0.900) | (0.868) | (0.865) |
| (100,100) | | 0.888 | 0.880 | 0.878 |
| | | (0.915) | (0.878) | (0.873) |

## 5 Application of the Data

The actual data used for testing the test statistic $A_{ks\tau}$ are comprised of two data sets: the first, the independent variable X representing the number of households in each of the 76 provinces of Thailand for 2002 and 2004, the dependent variable Y representing the number of population in each of 76 provinces of Thailand for 2002 and 2004. The first data set is retrieved from the census conducted by the National Statistical Office of the Prime Minister of Thailand in 2002 and 2004 (National Statistical Officer Thailand, 2002) and (National Statistical Officer Thailand, 2004).The second, the independent variable X representing the relative humidity of 46 provinces of Thailand for 2009 and 2010, and the dependent variable Y representing the rainfall in each 46 provinces of Thailand in 2009 and 2010. The second data set is retrieved from Thai Meteorological Department (Thai Meteorological Department, 2010). To calculate the p-value of the test statistic $A_{ks\tau}$, the conditional median regression function and the conditional mean regression function are estimated from 1,000 replications of bootstrapping.

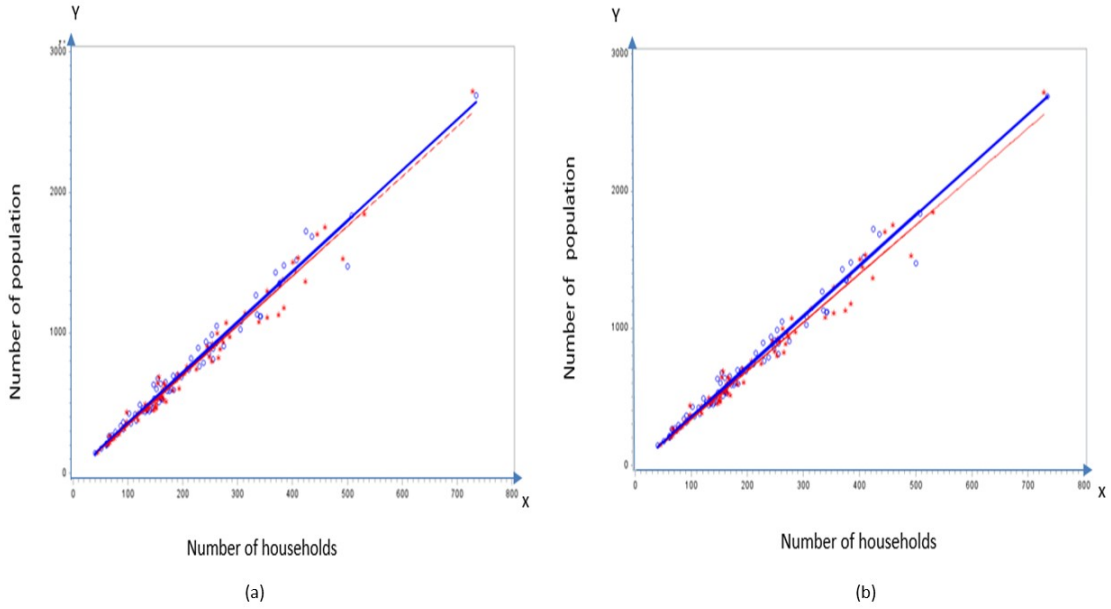


Figure 2: (a) illustrates the scatter plot and the conditional mean regression function of the number of household and the number of population in each province. The data for 2004 are represented by circles and the solid line, whereas those for 2002 are represented by stars and the dash line (b) Illustrates the scatter plot and the conditional median regression function of the number of household and the number of population in each province. The data for 2004 are represented by circles and the solid line, whereas those for 2002 are represented by stars and the dash line.

From Figure 2(a). It can be seen that the relationship between the two variables in both years are almost the same and nearly form a single line. Also, the p-value from 1,000 replications of bootstrapping equals 0.915. This implies that, conditional mean regression function between two sets of data be equal, two data sets have similar relationship between dependent and independent variables. From Figure 2(b). It can be seen that in comparison with the conditional median regression function, the conditional median regression function leads to more discrepancies in the regression lines and the lower p-value of 0.890 from 1,000 replications of bootstrapping. This implies that, conditional median regression function between two sets of data be equal, two data sets have similar relationship between dependent and independent variables. In spite of this difference, it can be said that the conditional mean and the conditional median regression functions yield similar results, namely, accept the null hypothesis, the relationship between two regression functions not different, when we consider about the distribution of the errors, from conditional mean regression function, the result are displayed in Table 9 and Figure 3.

Table 9: The distribution of error of the first data set : The test statistic for testing the normality of the distribution of errors that estimate from conditional mean.

| data | $\epsilon_{ij\tau}$ | K-S | p-value | Kurtosis |
|---|---|---|---|---|
| No. of population (Y), No. of households (X) in 2002 | $\epsilon_{i1\tau}$ | 0.083 | 0.200 | 0.636 |
| No. of population (Y), No. of households (X) in 2004 | $\epsilon_{i2\tau}$ | 0.093 | 0.178 | 0.920 |
| over all data | $\epsilon^0_{ij\tau}$ | 0.099 | 0.069 | 0.740 |

From Table 9. According to Kolmogorov-Smirnov test statistic, the distribution of errors that estimate from conditional mean regression function of the first regression function $\epsilon_{i1\tau}$, the second regression function $\epsilon_{i2\tau}$, and the common regression function $\epsilon^0_{ij\tau}$ are normal distribution. When we consider about the kurtosis, the result was found that the kurtosis value $< 3$, this implies that the distribution of errors is normal distribution that not heavy tailed. The distribution of errors (histogram and normal probability) are displayed in Figure 3.

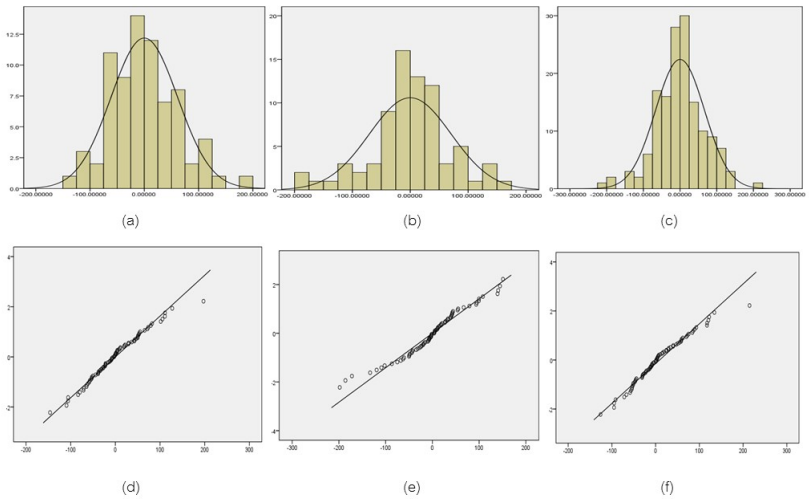


Figure 3: (a) and (d) illustrates the histogram and normal probability plot of the distribution of errors from the conditional mean of the relationship between the number of households (X) and number of population (Y) in 2002. (b) and (e) illustrates the histogram and normal probability plot of the distribution of errors from the conditional mean of the relationship between the number of households (X) and number of population (Y) in 2004. (c) and (f) illustrates the histogram and normal probability plot of the distribution of errors from the conditional mean of the relationship between the number of households (X) and number of population (Y) of common regression function.

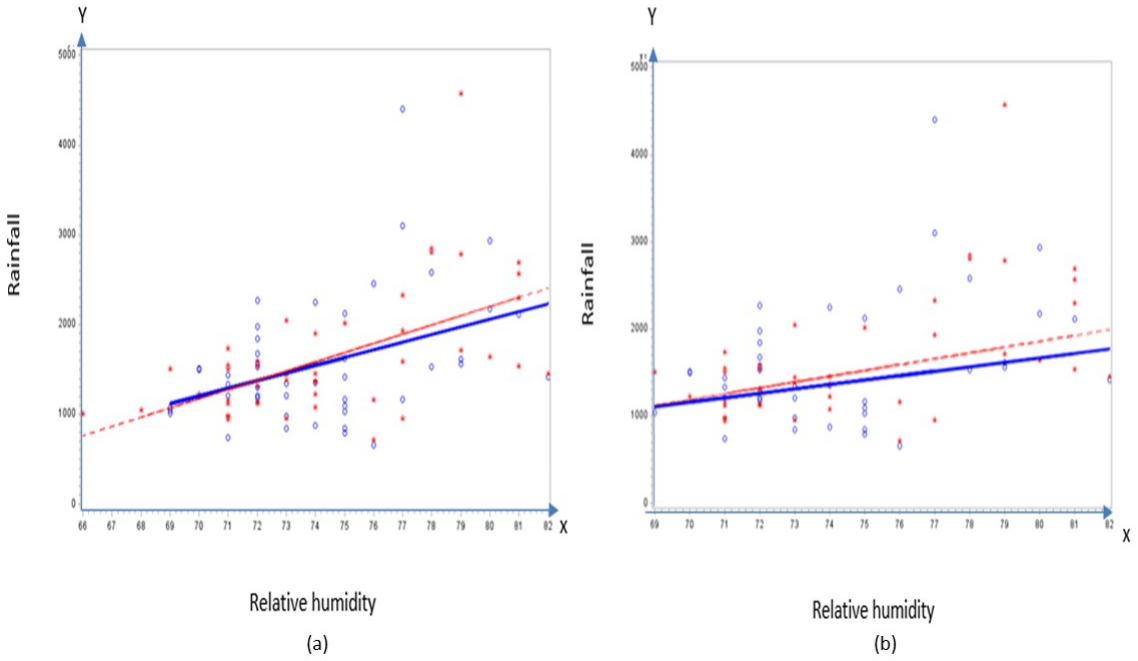


Figure 4: (a) illustrates the scatter plot and the conditional mean regression function of the relative humidity and the rainfall in each province. The data of 2009 are represented by circles and the solid line, whereas those for 2010 are represented by stars and the dash line. (b) illustrates the scatter plot and the conditional median regression function of the relative humidity and the rainfall in each province. The data of 2009 are represented by circles and the solid line, whereas those for 2010 are represented by stars and the dash line.

Figure 4 (a) presents the conditional mean regression functions of the relative humidity and the rainfall in each province, compare with 2009 and 2010. It can be seen that the relationship between two groups in 2009 and 2010 are different. Also, the p-value from 1,000 replications of bootstrapping equals 0.0472. Figure 3(b). Shows the conditional median regression function of the relative humidity and the rainfall in each province, compare with 2009 and 2010, It can be seen that the relationship between two groups in 2009 and 2010 are different. Also, the p-value from 1,000 replications of bootstrapping equals 0.0465, it can be said that the conditional mean and the conditional median regression functions yield similar results, namely, accept the alternative hypothesis: two regression functions have different at 0.05 significant level, when we consider about the distribution of the errors, from conditional mean regression function, the result are displayed in Figure 5 and Table 10.

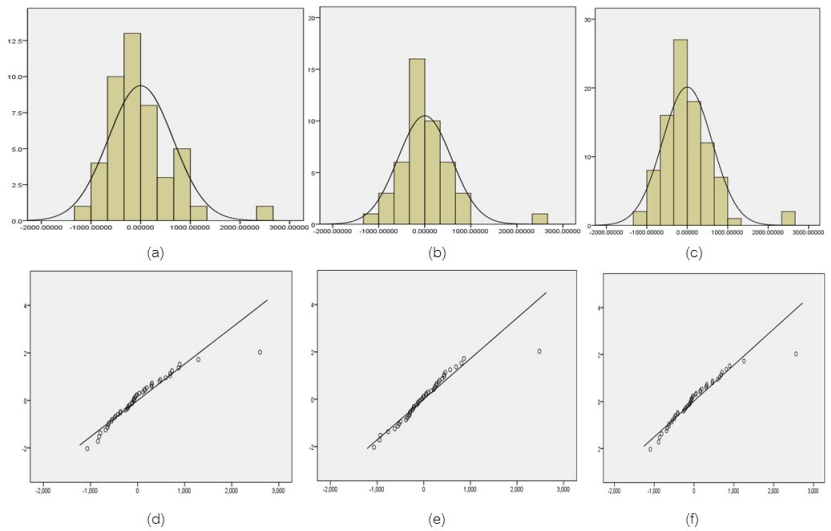


Figure 5: (a) and (d) illustrates the histogram and normal probability plot of the distribution of errors from the conditional mean of the relationship between the relative humidity (X) and rainfall (Y) in 2009. (b) and (e) illustrates the histogram and normal probability plot of the distribution of errors from the conditional mean of the relationship between the relative humidity (X) and rainfall (Y) in 2010. (c) and (f) illustrates the histogram and normal probability plot of the distribution of errors from the conditional mean of the relationship between relative humidity (X) and rainfall (Y) of common regression function.

Table 10: The distribution of error of the second data set : The test statistic for testing the normality of the distribution of errors that estimate from conditional mean.

| data | $\epsilon_{ij\tau}$ | K-S | p-value | Kurtosis |
|---|---|---|---|---|
| rainfall (Y), relative humidity (X) in 2009 | $\epsilon_{i1\tau}$ | 0.115 | 0.154 | 4.374 |
| rainfall (Y), relative humidity (X) in 2010 | $\epsilon_{i2\tau}$ | 0.105 | 0.200 | 6.255 |
| over all data | $\epsilon^0_{ij\tau}$ | 0.116 | 0.143 | 4.147 |

Figure 5 illustrates the histogram and normal probability plot of the distribution of errors from the conditional mean of the relationship between the relative humidity (X) and rainfall (Y) in 2009 and 2010. It can be seen that the distribution of errors are normal distribution that heavy tailed, and the Kolmogorov-Smirnov test statistic from Table 10 show that: the distribution of errors that estimate from conditional mean regression function of the first regression function $\epsilon_{i1\tau}$, the second regression function $\epsilon_{i2\tau}$, and the common regression function $\epsilon^0_{ij\tau}$ are normal distribution at 0.05 significant level. When we consider about the kurtosis, the result was found that the kurtosis value $> 3$, this implie that the distribution of errors is normal distribution that heavy tailed. It can be said that comparing the regression function with conditional mean (p - value=0.0472) and the conditional median regression functions (p - value = 0.465) yield very similar results. This implies that, with the heavy-tailed errors distribution, the conditional median regression function is more robust and give the performance similar to conditional mean.

From the second data sets, we simply compares the equality of two quantile regression functions that having the relationship between the relative humidity and rainfall in 2009 and 2010 as the maximum and minimum value : extream quantile regression function such as $\tau = 0.95$ and 0.25. Testing the equality of two quantile regression function as $\tau = 0.95$ and 0.25 are displayed in Figure 6.

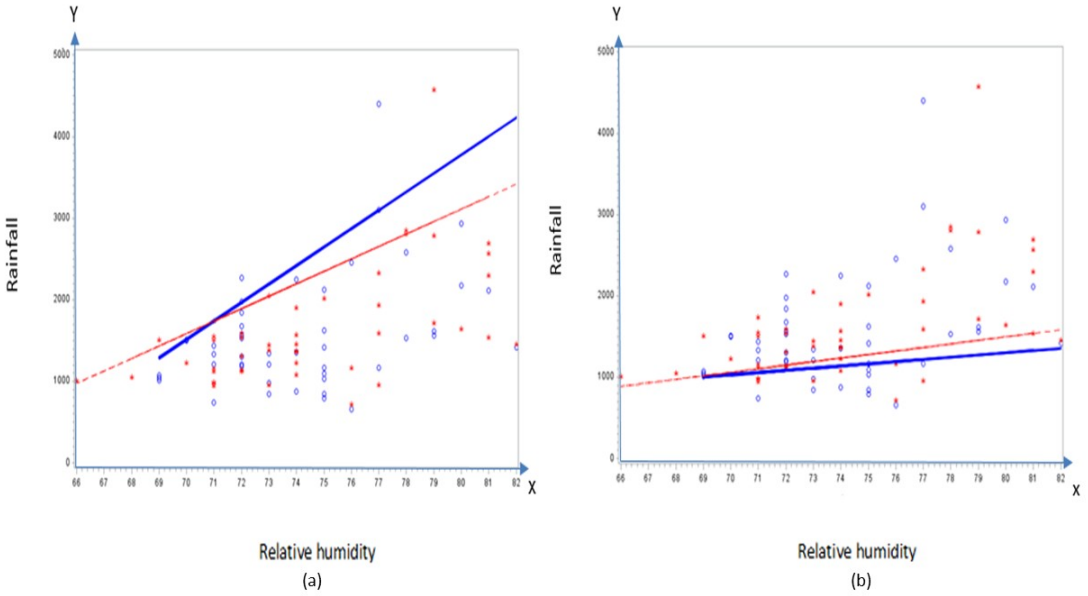


Figure 6: (a) illustrates the scatter plot and the conditional quatile regression function ($\tau = 0.95$) of the relative humidity and the rainfall in each province. The data of 2009 are represented by circles and the solid line, whereas those for 2010 are represented by stars and the dash line. (b) illustrates the scatter plot and the conditional quatile regression function ($\tau = 0.25$) of the relative humidity and the rainfall in each province. The data of 2009 are represented by circles and the solid line, whereas those for 2010 are represented by stars and the dash line.

Figure 6(a) pesents for testing the equality of two quantile regression functions as $\tau = 0.95$ of the relative humidity and the rainfall in each province, compare with 2009 and 2010. It can be seen that the relationship between two groups in 2009 and 2010 are different. Also, the p-value from 1,000 replications of bootstrapping equals 0.028, and as the $\tau = 0.95$ the quantile regression function in 2009 is higher than 2010. Figure 3(b). pesents for testing the equality of two quantile regression functions as $\tau = 0.25$ of the relative humidity and the rainfall in each province, compare with 2009 and 2010. It can be seen that the relationship between two groups in 2009 and 2010 are different. Also, the p-value from 1,000 replications of bootstrapping equals 0.046, and as the $\tau = 0.25$ the quantile regression function in 2010 is higher than 2009. It can be said that as the extreme conditional quatile: $\tau = 0.95$ and $\tau = 0.25$, accept the alternative hypothesis: two regression functions have different at 0.05 significant level.

## 6 Discussion Conclusions

The present study compares the conditional mean and the conditional median regression functions as well as examines the impact of the distribution of errors by determining the difference between the empirical distribution function of errors of each set of data and the common regression function. The results indicate that power of the test increases as the sample size becomes larger. Additionally, when the distribution of errors are heavy-tailed the ratio of hypothesis rejection slightly declines from (95% standard normal errors+ 5% Cauchy errors) to (90% standard normal errors + 10% Cauchy errors), and under normal conditions, the test statistic estimated from the conditional mean regression function performs better than that obtained from the conditional median regression function. In contrast, with heavy-tailed distribution of errors, the conditional median regression function is more robust, namely the impact of heavy-tailed distribution of errors are greater for the test statistic derived from the conditional mean regression function than for that obtained from the conditional median regression function. The approach presented in this research can provide guidelines for other studies aiming to compare two regression functions, especially conditional quantile regression functions subject to such conditions as heavy-tailed distribution of errors. It can also be applied for other special purposes, such as analyzing only the high values or only the low values in a set of data, and in the next research, the regression model that flexible assumption and good fitted such as nonparametric regression should be considered, and applying the quantile regression function for detection the outlier is more interesting.

## References

Akritas, M. G., and Van Keilegom, I. (2001). Non parametric Estimation of the Residual Distribution. *Scandinavian Journal of Statistics*, 28(3), 549-567.

Bera, A. K., Galvao, A. F., and Wang, L. (2014). On Testing the Equality of Mean and Quantile Effects. *Journal of Econometric Methods*, 3(1), 47-62.

Chen, C. (2005). An introduction to quantile regression and the QUANTREG procedure. *In Proceedings of the Thirtieth Annual SAS Users Group International Conference*, SAS Institute Inc.

Efron, B., and Tibshirani, R.J. (1994). An Introduction to the Bootstrap. *CRC press.*

Freedman, D. A. (1981). Bootstrapping regression models. *The Annals of Statistics*, 9(6), 1218-1228.

King, E., Hart, J. D., and Wehrly, T. E. (1991). Testing the equality of two regression curves using linear smoothers. *Statistics and Probability Letters*, 12(3), 239-247.

Koenker, R., and Bassett Jr, G. (1978). Regression quantiles. *Econometrica: journal of the Econometric Society*, 33-50.

Koenker, R., and Hallock, K. (2001). Quantile regression: An introduction. *Journal of Economic Perspectives*, 15(4), 43-56.

Kulasekera, K. B. (1995). Comparison of regression curves using quasi-residualss. *Journal of the American Statistical Association*, 90(431), 1085-1093.

Kuruwita, C., Gallagher, C., and Kulasekera, K. (2014). A consistent nonparametric equality test of conditional quantile functions. *International Journal of Statistics and Probability*, 3(1), 55.

Kutner, M. H., Nachtsheim, C. J., Neter, J., and Li, W. (2005). Applied linear statistical models. *McGraw-Hill Irwin New York*.

Lloyd, C.J. (2014). How Colse are Alternative Bootstrap p-value?. *Statistics and Probability Letters*, 80(23), 1972-1976.

National Statistical Officer Thailand (2002). Ministry of Information and Communication Technology, The Analytical report of income distribution in the province level. Available online at: `http://service.nso.go.th/nso/nso_center/project/search_center/23project-th.htm`.

National Statistical Officer Thailand (2004). Ministry ofInformation and Communication Technology, The Analytical report of income distribution in the province level. Available online at: `http://service.nso.go.th/nso/nso_center/project/search_center/23project-th.htm`.

Pardo-Fernndez, J. C., Van Keilegom, I., and Gonzlez-Manteiga, W. (2007). Testing for the equality of k regression curves. *Statistica Sinica*, 17(3), 1115,

Silverman, B. W., and Young, G. A. (1987). The bootstrap: To smooth or not to smooth?. *Biometrika*, 74(3), 469-479.

Sun, Y.(2006). A consistent nonparametric equality test of conditional quantile functions. *Econometric Theory*, 22(04), 614-632.

Thai Meteorological Department (2010). Annual Rainfall, Rain-day and Relative Humidity : Selected Location by Region 2009 - 2010. Available online at: `http://www.dnp.go.th/statistics/2556/table`.

Tonggumnead, U., and Seangngam, N. (2007). A comparison of Simple Linear Regression Slopes Testing Using F-test and $Z_{KS}$ test Based on Empirical Distribution Function of Errors. *Journal of Research Methodology*, 27(2), 135 - 155.