Applied Mathematical Sciences, Vol. 12, 2018, no. 16, 759 - 772
HIKARI Ltd www.m-hikari.com
https://doi.org/10.12988/ams.2018.8465

Robust Nonparametric Regression for Testing the Equality of Nonparametric Regression Curves

Unchalee Tonggumnead

Department of Mathematics and Computer Science
Faculty of Science and Technology, Rajamangala University of Technology
Thanyaburi, Phatum Thani 12110, Thailand

Abstract

This research aims to examine the equality of two nonparametric regression functions using two test statistics: the robust kernel regression (A_{rks}), which estimates regression functions from the robust kernel regression, and the kernel regression (A_{ks}), which estimates regression functions with the Nadaraya-Watson Estimator. Influenced by the Kolmogorov-Smirnov test statistic, the test statistic A_{rks} in the present study is created from the empirical distribution function (EDF) of errors. The efficiency of each of the statistics is also compared when the distribution of errors is heavy-tailed and outliers are present in the data. It is found that in case of normal distribution of errors with no outliers, the test statistics A_{rks} and A_{ks} are almost equally efficient. In contrast, in case of heavy-tailed distribution of errors or presence of outliers, the test statistic A_{rks} is much more efficient than the test statistic $A_{ks.}$ Additionally, as the size of n is larger, both statistics become more efficient. In addition, in case the regression function is linear, both test statistics are highly efficient. Finally, the application of the two test statistics to actual data yields consistent results.

Keywords: robust nonparametric regression, kernel regression, empirical distribution function, bootstrap, nonparametric regression

1 Introduction

Nonparametric regression is a category of regression analysis in which the regression function is unknown and hence must be estimated from independent

random samples that have been smoothed. In this type of analysis, the regression model is in the form of $Y = g(x) + \varepsilon$, where Y represents a conditional expected value and the mean of Y-values is represented by E(Y|X). Some studies compare the differences between two sets of data that involve the relationship between the independent variable X and the dependent variable Y through nonparametric regression using an analysis of the differences between regression curves. For instance, [1] examined the differences between two EDFs of errors based on the principle of kernel regression by estimating errors of the regression curves using $\hat{\varepsilon}_{ij} = \frac{Y_{ij} - \hat{g}_j(X_{ij})}{\hat{\sigma}_j(X_{ij})}$ and $\hat{g}_j(X_{ij})$ using the Nadaraya-Watson Estimator. In their bootstrap test, [2] tested the equality of the nonparametric regression curves of the test statistics based on the functional distances between nonparametric estimators of the regression functions as well as estimated the critical values of the test statistics using the bootstrap resampling method. [3] employed Fourier coefficients to investigate the equality of the regression curves $g_1(x)$ and $g_2(x)$ in case the data involved fixed-design, homoscedastic error. [4] conducted a nonparametric analysis of regression functions by applying a new method for comparing the regression curves $g_1(x)$ and $g_2(x)$, that is, making estimation based on the chosen points. They also examined the distribution of the test statistic $\hat{T}$ under the null hypothesis $H_0 : g_1(x) = g_2(x)$ and the alternative hypothesis $H_1 : g_1(x) \neq g_2(x)$ as well as explored the weight functions that would equip the test statistic with the highest power of the test. [5] investigated the test statistics for comparing nonparametric regression curves to one-sided curves under the null hypothesis $H_0 : g_1(x) = g_2(x)$ and the alternative hypothesis $H_1 : g_1(x) \leq g_2(x)$ with the sample average of errors being estimated from each of the nonparametric regression curves. [6] employed nonparametric tests to compare the regression curves in case the data involved more than two sets of population, the test statistics were dependent on local linear estimates, and a data-driven approach was used to select the bandwidth. [7] compared nonparametric regression curves based on a scale-space visualization tool for statistical inferences referred to as significant ZERo crossing of the differences (SiZer) analysis. This method does not require any specification of smoothing parameters but involves a comparison of a wide range of resolutions to determine the differences between two regression curves at each resolution level and a comparison of k regression curves through error analysis.

However, little research has focused on the robust kernel regression, which can provide an effective solution in case random errors $\varepsilon_1, \varepsilon_2, ..., \varepsilon_n$ are heavy-tailed. Another problem with previous studies is that outliers heavily influence the estimation of regression functions involving a kernel estimator, leading to inefficiency in testing the equality of regression functions. Such issues have been addressed by few researchers. [8] employed a marked empirical process to test the equality of nonparametric regression curves and compared the efficiency of the test

statistics in case of heavy-tailed errors. [9] tested the equality of two parametric quantile regression curves and compared the conditional quantile regression and the conditional mean regression in case errors were heavy-tailed and outliers were present in the data.

Therefore, the main objective of the present study is to examine the efficiency of robust nonparametric regression in testing the equality of two regression curves where errors are represented by $\varepsilon_{ij} = Y_{ij} - g_j(X_{ij})$ and $g_j(X_{ij})$ represents the nonparametric regression function derived from the robust estimator $\hat{g}_j(x_0, h, k)$, which will be detailed in the next section. This research also addresses issues pertaining to the impact of outliers, the distribution of heavy-tailed errors, model construction, and the application of the test statistics to actual data. The assumption behind the test statistics is that two sets of data have a similar relationship between the independent variable and the dependent variable provided that the EDF of each of the regression functions involved is equal.

2. Methods

2.1 Robust nonparametric kernel regression

This study performs a nonparametric regression analysis using the nonparametric kernel regression in the form of $Y_{ij} = g_j(X_{ij}) + \varepsilon_{ij}, i = 1, ..., n_j$, where $j = 1, 2$ and the expected value of random errors is derived from $E(\varepsilon_{ij} \mid X_{ij}) = 0$. The main objective of the research is to estimate the regression mean function $g_j(X_{ij})$. However, as the data used to calculate (X_{ij},Y_{ij}) may be comprised of outliers, a robust $g_j(X_{ij})$ is necessary to account for the impact of outliers on the traditional nonparametric kernel regression. Specifically, outliers will have more impact in case the regression curve are estimated using the nonparametric kernel regression, pulling the regression function toward them, details are shown in figure 1.

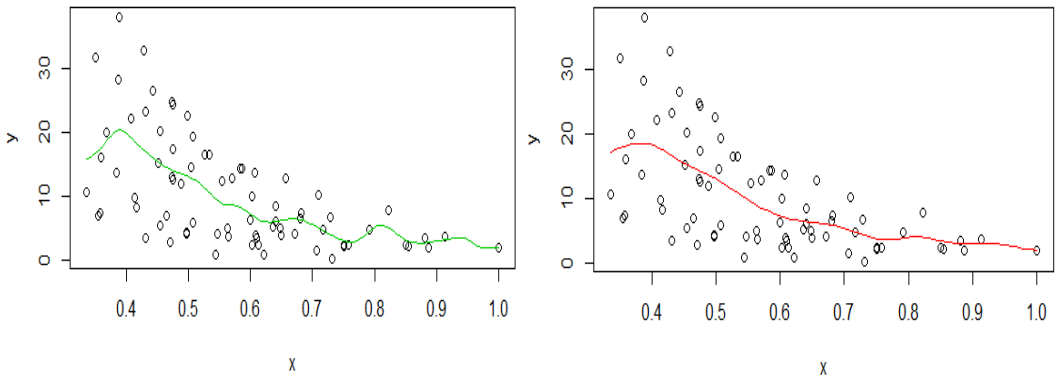


Figure 1: (a) Illustrates the scatter plot and estimate regression curve with classical nonparametric kernel regression (Nadaraya- Watson estimator) (b) Illustrates the scatter plot and estimate regression curve with robust nonparametric kernel regression.

Therefore, to ensure the power of the test of the test statistic, the regression curve is estimated based on the robust nonparametric kernel regression with the robust kernel regression. According [10], the estimator of regression function from kernel regression $\hat{g}_j(x_0,h,k)$ being calculated from the following equation:

$$\sum_{i=1}^{n_j} w_j(x_i,x_0,h)\psi_k\{y_i-\hat{g}_j(x_0,h,k)\}=0,\ \text{j=1,2} \tag{1}$$

Where w represents the weight function and

$$w_j(x_i,x_0,h)=h^{-1}\int_{x_{i-1}+x_1/2}^{x_i+x_{i+1}/2} K\{(x_0-u)/h\}du$$

for the kernel function K(.); and ψ represents the truncation function that can be calculated from $\psi_k(s)=sI(|s|<k)+kI(s\geq k)-kI(s\leq -k)$, where k represents the trimmer and I(.) represents the indicator function.
The parameter estimation procedure is as follows:

1. The beginning trimmer is determined assuming k > k_0.
2. For each k, the bandwidth value is selected using the leave-one-out cross-validation method.

2.1)The equation $\sum_{l\neq i} w_j(x_l,x_i,h)\psi_k(y_i-\hat{y}_{-1,i})=0$ is solved, which yields the value of $\hat{y}_{-1,i}$.

2.2) The equation $\arg\min \phi_{1-q}\left\{\left|\hat{y}_{-1,1}-y_1\right|,...,\left|\hat{y}_{-1,n_j}-y_{n_j}\right|\right\}$, j=1,2, $i=1,...,n_j$ is solved to obtain the value of h, where $\phi_p(s_1...,s_{n_j})=\sum_{i=1}^{|n_j p|} s_{(i)}$ represents the lowest value of $s_1...,s_{n_j}$.

3. A set of points between $x_{(1)}$ and $x_{(n_j)}$, $x_t^{'},t=1,...,r$ is selected before the value of k is updated.

3.1)The equation $\sum_{i=1}^{n_j} w(x_i,x_t^{'},h)\psi_k\left\{y_i-\hat{g}(x_t^{'},h)\right\}=0$ is solved to obtain the value of $\hat{g}(x_t^{'},h)$, where $x_t^{'}$ represents the points between $x_{(1)}$ and $x_{(n)}$.

3.2)The equation $\sum_{t=1}^{r}\sum_{i=1}^{n_j} I\left\{\left|y_i-\hat{g}(x_t^{'},h)\right|<k\right\}=r[n(1-q)]$ is solved to obtain the value of k.

4. The steps in 2 and 3 are repeated until the values of k and h approach to k* and h*. In the last step, the estimator of $\hat{g}_t(x,h^*,k^*)$ is obtained from the solution to the equation $\sum_{i=1}^{n} w(x_i,x_t^{'},h^*)\psi_k\left\{y_i-\hat{g}(x_t^{'},h^*)\right\}=0$.

2.2 Test statistics

The robust kernel regression (A_{rks}) is created from the EDF of errors. The test statistic $A_{rks}=\sum_{i=1}^{2}\sup_{y}\left|\hat{G}_j(y)\right|$, where $\hat{G}_j(y)=n_j^{1/2}\left(\hat{F}_{\varepsilon}^{0}(y)-\hat{F}_j(y)\right)$, j=1,2, and $i=1,...,n_j$, where $\hat{F}_{\varepsilon}^{0}(y)=\frac{1}{n_j}\sum_{i=1}^{n_j}I(Y_{ij}-(\hat{g}(x,h^*,k^*)))$, is the estimator of the EDF of errors when the null hypothesis is true. In addition, $\hat{F}_j(y)$ is the estimator of the EDF of errors of each regression curve. Additionally, $\hat{F}_j(y)=\frac{1}{n_j}\sum_{i=1}^{n_j}I(Y_{ij}-(\hat{g}_j(x,h^*,k^*)))$ estimates $\hat{g}_j(x,h^*,k^*)$ and $\hat{g}(x,h^*,k^*)=\sum_{j=1}^{2}\hat{g}_j(x,h^*,k^*)$ based on the robust nonparametric kernel regression principle. By contrast, the test statistic $A_{ks}=\sum_{i=1}^{2}\sup_{y}\left|\hat{G}_j(y)\right|$ estimates the regression function using the Nadaraya-Watson Estimator. Influenced by the Kolmogorov-Smirnov test statistic, the test statistics A_{rks} and A_{ks} will convergence to a normal distribution with the average value equaling 0 and the covariance equaling $F_{\varepsilon}^{0}(y)(1-\hat{F}_{\varepsilon}^{0}(y))$.

3. Simulation study

The construction of the test statistics A_{rks} and A_{ks} applies the bootstrap resampling method for convenience in critical value estimation, following [1], [11], [12] and [13]. The procedure is as follows.

1.The bootstrap replication is set at b = 1,…,B(B=1,000) for j=1,2 and i=1,…,n_j before the transformation of the bootstrap $Y_{ij,b}^{*}$, b=1,…,B into the following equation.

$$Y_{ij,b}^{*}=g_j(X_{ij})+\varepsilon_{ij,b}^{*},\ \text{j=1,2, i=1,…, } n_j$$

The test statistic A_{rks} estimates the regression function $\hat{g}_j(X_{ij})$ using $\hat{g}_j(x,h^*,k^*)$ based on the robust nonparametric regression principle, whereas the test statistic A_{ks} estimates the regression function using the Nadaraya-Watson Estimator.

2. For j=1,2, i=1,…, n_j, the test statistics A_{rks} and A_{ks} are calculated from the bootstrap sample $X_{ij}, Y_{ij,b}^{*}$ by determining the regression functions in six forms with the first three representing the forms under true null hypothesis and the last three representing the forms under true alternative hypothesis as follows.

a. $g_1(x) = g_2(x) = 3x$

b. $g_1(x) = g_2(x) = 3x^2$

c. $g_1(x) = g_2(x) = \sin(2x)$

d. $g_1(x) = 3x$, $g_2(x) = 3x + 2$

e. $g_1(x) = 3x^2$, $g_2(x) = 3x^2 + 2$

f. $g_1(x) = \sin(2x)$, $g_2(x) = \sin(2x) + 2$

3. The distribution of errors is determined as follows.

3.1)100% of the errors in each set of data are determined to have a standard normal distribution according to $\varepsilon_{i1} \sim N(0,1)$ and $\varepsilon_{i2} \sim N(0,1)$, j=1,2., i=1,…,n,

3.2) 95% and 90% of the errors in each set of data are determined to have a standard normal distribution and the respective remaining 5% and 10% of the errors in each set of data are determined to have the Cauchy distribution, whose probability function is in the form of $f(x) = \dfrac{1}{\pi b\left[1+\left(\dfrac{x-a}{b}\right)^2\right]}$, $-\infty < x < \infty, -\infty < a < \infty$ and b>0, with a=0 and b=1 in the present study.

4. The distribution of X_{i1} and X_{i2}, j=1,2, i=1,…,n_j are determined as follows.

4.1) X_{i1} and X_{i2}, j=1,2, i=1,…,n_j are determined to have a uniform distribution in the range [0,1].

4.2) X_{ij} and j=1,2, i=1,…,n_j are determined to have a uniform distribution in the range [0,1] with mild outliers and independent variable values in the range $[Q_1 - 3(IQR), Q_1 - 1.5(IQR)]$ or $[Q_1 + 1.5(IQR), Q_1 + 3(IQR)]$ with the outlier rates of 5% and 10% of the sample size $(n_1, n_2) = (20,20),(50,50)$, $(100,100)$

4. Results

A comparison of the proportions of type I errors for the first three forms (1)-(3) under true null hypothesis when the errors are normally distributed and when the errors are heavy-tailed is conducted at the significance level of 0.05. The results show that the test statistics A_{rks} and A_{ks} produce a very similar degree of type I errors in case of normal distribution of errors. By contrast, in case of 90%N(0,1)+10% Cauchy, the type I errors associated with the test statistic A_{rks} more closely approach 0.05 than do those associated with the test statistic A_{ks}, demonstrating greater robustness of the former than the latter. Additionally, with a larger sample size, the type I errors associated with both test statistics better approximate 0.05. Finally, the two test statistics are relatively highly efficient when the regression function is linear, as shown in Table 1.

Table 1 Rejection proportions under the null hypothesis (Type I error) of model (1)-(3) $\alpha = 0.05$, when the distribution of error are 100% N(0,1) , 95% N(0,1) + 5% Cauchy, and 90% N(0,1) + 10% Cauchy

model	sample size	100% N(0,1)		95% N(0,1) + 5% Cauchy		90% N(0,1) + 10% Cauchy	
		A_{rks}	A_{ks}	A_{rks}	A_{ks}	A_{rks}	A_{ks}
(1)	(20,20)	0.035	0.036	0.035	0.033	0.035	0.030
	(50,50)	0.040	0.039	0.041	0.039	0.040	0.037
	(100,100)	0.042	0.042	0.042	0.039	0.042	0.038
(2)	(20,20)	0.033	0.035	0.033	0.032	0.032	0.030
	(50,50)	0.038	0.036	0.038	0.035	0.038	0.031
	(100,100)	0.041	0.041	0.040	0.038	0.040	0.034
(3)	(20,20)	0.033	0.032	0.032	0.030	0.031	0.028
	(50,50)	0.037	0.035	0.037	0.035	0.036	0.032
	(100,100)	0.043	0.039	0.042	0.040	0.041	0.035

A comparison of the proportions of type I errors for the first three forms (1)-(3) under true null hypothesis when the rate of outliers stands at 5% and 10% is also conducted at the significance level of 0.05. It is found that at such outlier rates, the type I errors associated with the test statistic A_{rks} are closer to 0.05 than are those associated with the test statistic A_{ks}, demonstrating that the former is more robust to outliers than the latter. Again, when the sample size is larger, the type I errors associated with the two test statistics more closely approach 0.05. Finally, both test statistics are relatively highly efficient in case of linear regression function, as shown in Tables 2.

Table 2 Rejection proportions under the null hypothesis (Type I error) of model (1)-(3) $\alpha = 0.05$, when the rate of outliers stands at 5% and 10%.

model	sample size	No outlier		5% outlier		10% outlier	
		A_{rks}	A_{ks}	A_{rks}	A_{ks}	A_{rks}	A_{ks}
(1)	(20,20)	0.035	0.036	0.033	0.028	0.033	0.025
	(50,50)	0.040	0.039	0.038	0.033	0.036	0.029
	(100,100)	0.042	0.042	0.040	0.035	0.038	0.030
(2)	(20,20)	0.033	0.035	0.031	0.028	0.033	0.028
	(50,50)	0.038	0.036	0.035	0.030	0.035	0.030
	(100,100)	0.041	0.041	0.038	0.032	0.038	0.030
(3)	(20,20)	0.033	0.032	0.030	0.027	0.028	0.022
	(50,50)	0.037	0.035	0.036	0.032	0.035	0.025
	(100,100)	0.043	0.039	0.036	0.030	0.035	0.027

A comparison of the proportions of power of the test for the first three forms (4)-(6) under alternative hypothesis when the errors are normally distributed and when the errors are heavy-tailed is conducted at the significance level of 0.05. The results show that the test statistics A_{rks} and A_{ks} produce a very similar degree of power of the test in case of normal distribution of errors. By contrast, in case of

90%N(0,1)+10% Cauchy, the power of the test associated with the test statistic A_{rks} more closely approach 1.00 than do those associated with the test statistic A_{ks}, demonstrating greater robustness of the former than the latter. Additionally, with a larger sample size, the power of the teat associated with both test statistics better approximate 1.00. Finally, the two test statistics are relatively highly efficient when the regression function is linear, as shown in Table 3.

Table 3 Rejection proportions under the alternative hypothesis (power of the test) of model (4)-(6), $\alpha = 0.05$, when the distribution of error are 100% N(0,1), 95% N(0,1) + 5% Cauchy, and 90% N(0,1) + 10% Cauchy.

model	sample size	100% N(0,1)		95% N(0,1) + 5% Cauchy		90% N(0,1) + 10% Cauchy	
		A_{rks}	A_{ks}	A_{rks}	A_{ks}	A_{rks}	A_{ks}
(4)	(20,20)	0.780	0.770	0.760	0.710	0.750	0.680
	(50,50)	0.850	0.850	0.850	0.800	0.830	0.780
	(100,100)	0.910	0.880	0.900	0.850	0.880	0.820
(5)	(20,20)	0.750	0.740	0.740	0.700	0.720	0.670
	(50,50)	0.820	0.830	0.800	0.750	0.780	0.720
	(100,100)	0.900	0.880	0.880	0.830	0.880	0.820
(6)	(20,20)	0.730	0.720	0.720	0.680	0.700	0.650
	(50,50)	0.820	0.800	0.820	0.780	0.800	0.750
	(100,100)	0.880	0.860	0.880	0.820	0.860	0.800

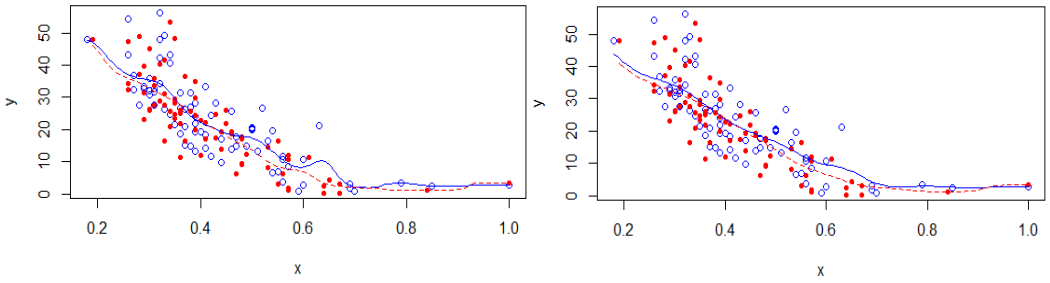
A comparison of the proportions of power of the test for the first three forms (4)-(6) under true alternative hypothesis when the rate of outliers stands at 5% and 10% is also conducted at the significance level of 0.05. It is found that at such outlier rates, the power of the test associated with the test statistic A_{rks} are closer to 1.00 than are those associated with the test statistic A_{ks}, demonstrating that the former is more robust to outliers than the latter. Again, when the sample size is larger, the power of the tes associated with the two test statistics more closely approach 1.00. Finally, both test statistics are relatively highly efficient in case of linear regression function, as shown in Tables 4.

Table 4 Rejection proportions under the alternative hypothesis (power of the test) of model (4)-(6), $\alpha = 0.05$, when the rate of outliers stands at 5% and 10%.

model	sample size	100% N(0,1)		95% N(0,1) + 5% Cauchy		90% N(0,1) + 10% Cauchy	
		A_{rks}	A_{ks}	A_{rks}	A_{ks}	A_{rks}	A_{ks}
(4)	(20,20)	0.780	0.770	0.750	0.720	0.750	0.680
	(50,50)	0.850	0.850	0.820	0.790	0.800	0.750
	(100,100)	0.910	0.880	0.880	0.830	0.850	0.790
(5)	(20,20)	0.750	0.740	0.720	0.680	0.720	0.680
	(50,50)	0.820	0.830	0.820	0.780	0.800	0.720
	(100,100)	0.900	0.880	0.890	0.840	0.880	0.800
(6)	(20,20)	0.730	0.720	0.700	0.650	0.700	0.650
	(50,50)	0.820	0.800	0.800	0.700	0.800	0.730
	(100,100)	0.880	0.860	0.850	0.800	0.840	0.780

5. Application of the data

The application of the actual data used for testing the test statistic A_{rks} and A_{ks} are comprised of three data sets : the first, the independent X representing the average income per person in each of the 76 provinces of Thailand for 2007, 2008 and 2014 , the dependent variable Y representing head Count Index calculated from the population with consumption expenditure below the poverty line divided by the total population multiplied by 100 in each of the 76 provinces of Thailand for 2007, 2008 and 2014 [14]. The first application of the actual data is comparing two data sets between 2007 and 2008, details are shown in figure 7. Figure 7 (a) estimate regression function with Nadaraya-Watson estimator. Figure 7 (b) estimate regression function with robust kernel regression. In this research we calculate the p-value of test statistics A_{rks} and A_{ks} from 1,000 replications of bootstrapping. From Figure 2. It can be seen that the relationship between the independent variable X and dependent variable Y in both years are almost the same, the test statistics A_{ks} and A_{rks} give the p-value from 1,000 replications of bootstrapping equals 0.188 and 0.192 respectively. This implies that, the relationship between independent variable X and dependent variable y of two data have similar, namely, accept the null hypothesis $(g_1(x) = g_2(x))$ at 0.05 significant level. When we consider about the distribution of the error of the test statistic A_{rks}, the result are displayed in Figure 3.



(a) (b)

Figure 2 (a) Illustrates the scatter plot and estimate regression curves with Nadaraya Watson estimator of average income per person in each of the 76 provinces of Thailand (X) and head Count Index in each of the 76 provinces (Y).

The data of 2007 are represented by circles and dash line, whereas those of 2008 are represented by solid circles and the solid line (b) Illustrates the scatter plot and estimate regression curves with robust kernel regression of average income per person in each of the 76 provinces of Thailand (X) and head Count Index of the 76 provinces. The data of 2007 are represented by circles and dash line, whereas those of 2008 are represented by solid circles and the solid line

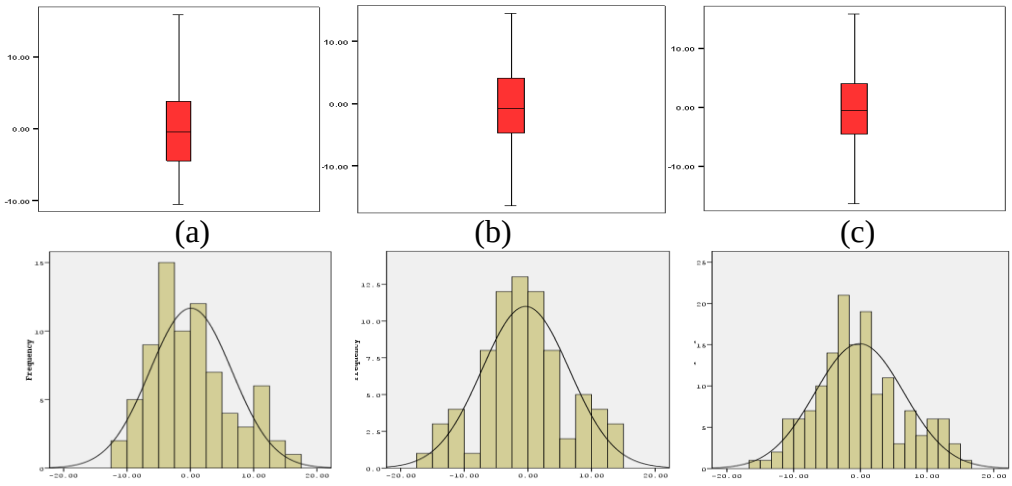


Figure 3 (a) and (d) Illustrates the Boxplot and histogram of the distribution of errors from robust kernel regression of the relationship between average income per person in each of the 76 provinces of Thailand (X) and head Count Index of the 76 provinces (Y) in 2007. (b) and (e) Illustrates the Boxplot and histogram of the distribution of errors from robust kernel regression of the relationship between average income per person in each of the 76 provinces of Thailand (X) and head Count Index of the 76 provinces (Y) in 2008. (c) and (f) Illustrates the Boxplot and histogram of the distribution of errors from robust kernel regression of the relationship between average income per person in each of the 76 provinces of Thailand (X) and head Count Index of the 76 provinces (Y) of common regression function

From Figure 3, the distribution of error that estimate from robust kernel regression of the first regression function ε_{i1}, the second regression function ε_{i2}, and the common regression function ε_{ij}^{0} are normal distribution that not heavy tailed and no outlier. This implies that under the assumption of normal distribution of errors that not heavy tailed and no outlier, the test statistics A_{rks} and A_{ks} are relatively equally efficient .

The second application of the actual data is comparing two data sets between 2008 and 2014, details are shown in figure 4. Figure 4 (a) estimate regression function with Nadaraya-Watson estimator. Figure 4 (b) estimate regression function with robust kernel regression. In this research we calculate the p-value of test statistics A_{rks} and A_{ks} from 1,000 replications of bootstrapping.

From Figure 4. It can be seen that the relationship between the independent variable X and dependent variable Y in both years are almost the different, the test statistics A_{ks} and A_{rks} give the p-value from 1,000 replications of bootstrapping equals 0.012 and 0.025 respectively. This implies that, the relationship between independent variable X and dependent variable Y of two data have different, namely, reject the null hypothesis at 0.05 significant level.
When we consider about the distribution of the error of the test statistic A_{rks}, the result are displayed in Figure 5.

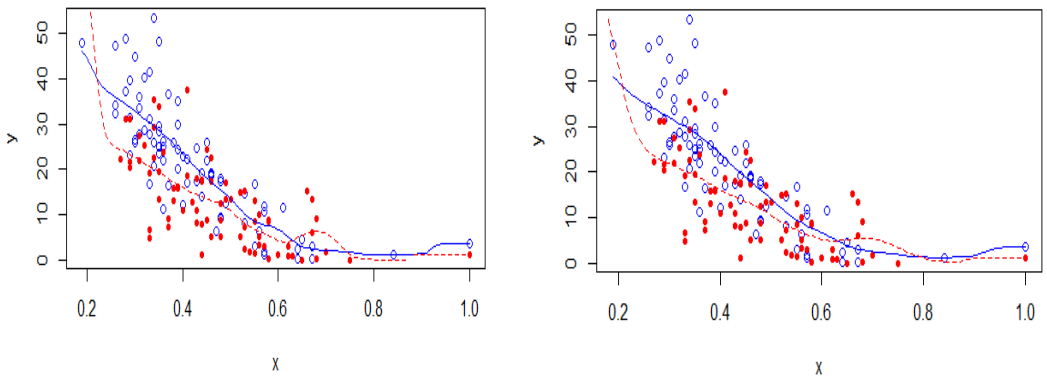


Figure 4 (a) Illustrates the scatter plot and estimate regression curves with Nadaraya Watson estimator of average income per person in each of the 76 provinces of Thailand (X) and head Count Index calculated from the population with consumption expenditure below the poverty line divided by the total population multiplied by 100 in each of the 76 provinces (Y). The data of 2008 are represented by circles and dash line, whereas those of 2008 are represented by solid circles and the solid line (b) Illustrates the scatter plot and estimate regression curves with robust kernel regression of average income per person in each of the 76 provinces of Thailand (X) and head Count Index of the 76 provinces. The data of 2014 are represented by circles and dash line, whereas those of 2008 are represented by solid circles and the solid line

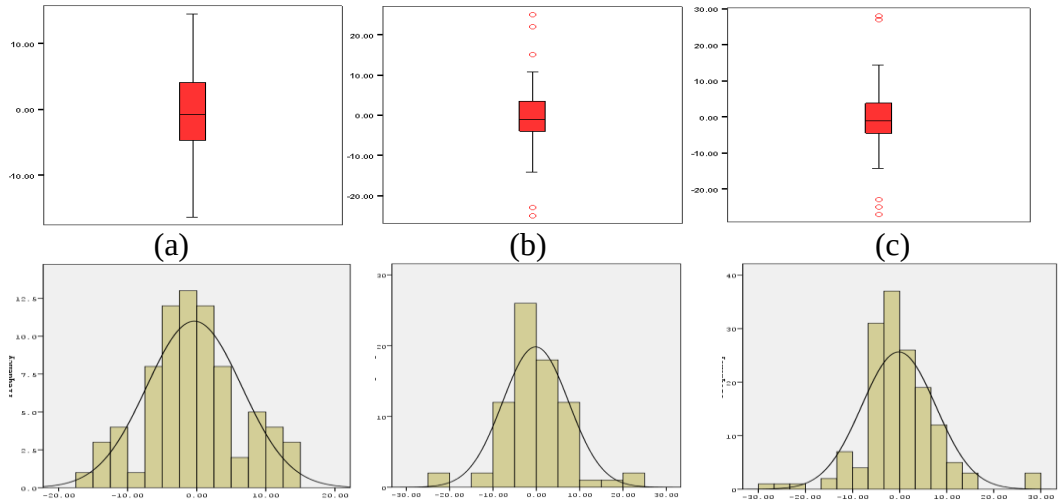


Figure 5 a) and d) Illustrates the Boxplot and histogram of the distribution of errors from robust kernel regression of the relationship between average income per person in each of the 76 provinces of Thailand (X) and head Count Index of the 76 provinces (Y) in 2008. b) and e) Illustrates the Boxplot and histogram of the distribution of errors from robust kernel regression of the relationship between average income per person in each of the 76 provinces of Thailand (X) and head Count Index of the 76 provinces (Y) in 2014. c) and f) Illustrates the Boxplot and histogram of the distribution of errors from robust kernel regression of the relationship between average income per person in each of the 76 provinces of Thailand (X) and head Count Index of the 76 provinces (Y) of common regression function

From Figure 5, the distribution of error that estimate from robust kernel regression of the first regression function ε_{i1}, the second regression function ε_{i2}, and the common regression function ε_{ij}^{0} are normal distribution that have heavy tailed and outlier. This implies, when the distribution of errors is heavy-tailed or outliers are present, the test statistics A_{rks} is more robust than the test statistics A_{ks}

6. Conclusion and Discussion

This research aims to examine the equality of two data sets that having the relationship between the dependent and independent variable though testing the equality two nonparametric regression functions using two test statistics: the robust kernel regression (A_{rks}), which estimates regression functions robust kernel regression, and the test statistic A_{ks}, which estimates regression functions with the Nadaraya-Watson Estimator. The result was found that the impact of outlier and heavy tail distribution of error are greater for the test statistic A_{ks}. When we consider about the proportions of type I errors under true null hypothesis at the significance level of 0.05. The results show that the test statistics A_{rks} and A_{ks} produce a very similar degree of type I errors in case of normal distribution of errors. By contrast, in case of 90%N(0,1)+10% Cauchy, the type I errors associated with the test statistic A_{rks} more closely approach 0.05 than do those associated with the test statistic A_{ks}, demonstrating greater robustness of the former than the latter. Meanwhile, the proportions of type I errors under true null hypothesis when the rate of outliers stands at 5% and 10% is also conducted at the significance level of 0.05. It is found that at such outlier rates, the type I errors associated with the test statistic A_{rks} are closer to 0.05 than are those associated with the test statistic A_{ks}, demonstrating that the former is more robust to outliers than the latter. When we consider about the proportions of power of the test under alternative hypothesis when the errors are normally distributed and at the significance level of 0.05. The results show that the test statistics A_{rks} and A_{ks} produce a very similar degree of power of the test in case of normal distribution of errors. By contrast, in case of 90%N(0,1)+10% Cauchy, the power of the test associated with the test statistic A_{rks} more closely approach 1.00 than do those associated with the test statistic A_{ks}, demonstrating greater robustness of the former than the latter. Meanwhile, the proportions of type I errors under true null hypothesis at the significance level of 0.05. It is found that at such outlier rates, the power of the test associated with the test statistic A_{rks} are closer to 1.00 than are those associated with the test statistic A_{ks}, demonstrating that the former is more robust to outliers than the latter. From this research the findings indicate that when the errors are normally distributed with no outliers, the test statistics A_{rks} and A_{ks} are relatively equally efficient. In contrast, when the distribution of errors is heavy-tailed or outliers are present, the former is more robust than the latter. Thus, a test statistic like A_{rks} provides an efficient alternative in case the data under investigation does not follow the predetermined assumptions in terms of distribution of errors or outliers. In addition, in case the regression function is

linear, both test statistics are highly efficient. As for further research, it is recommended that the test statistics be used to compare more than two sets of data that involve the relationship between the independent variable X and the dependent variable Y by testing the equality of k regression curves, especially when there is more than one independent variable.

Acknowledgements. The author wishes to gratefully acknowledge the referee of this paper who helped to clarify and improve its presentation.

References

[1] J.C. Pardo-Fernández, I. Van Keilegom and W. González-Manteiga, Testing for the equality of k regression curves, *Statistica Sinica*, **17** (2007) 1115-1137.

[2] J.M. Vilar and J.C. Vilar, A bootstrap test for the equality of nonparametric regression curves under dependence, *Communications in Statistics-Theory and Methods*, **41** (2012), 1069-1088. https://doi.org/10.1080/03610926.2010.535634

[3] Z. Mohdeb, K.A. Mezhoud and D. Boudaa, Testing the equality of nonparametric regression curves based on Fourier coefficients. *Afrika Statistika*, **5** (2010). https://doi.org/10.4314/afst.v5i1.71036

[4] R. Srihera and W. Stute, Nonparametric comparison of regression functions, *Journal of Multivariate Analy*sis, **101** (2010), 2039-2059. https://doi.org/10.1016/j.jmva.2010.05.001

[5] N. Neumeyer and J.C. Pardo-Fernández, A simple test for comparing regression curves versus one-sided alternatives, *Journal of Statistical Planning and Inference*, **139** (2009), 4006-4016. https://doi.org/10.1016/j.jspi.2009.05.005

[6] T. Gørgens, Nonparametric comparison of regression curves by local linear fitting, *Statistics & Probability Letters*, **60** (2002), 81-89. https://doi.org/10.1016/s0167-7152(02)00283-3

[7] C. Park and K.H. Kang, SiZer analysis for the comparison of regression curves, *Computational Statistics & Data Analysis*, **52** (2008), 3954-3970. https://doi.org/10.1016/j.csda.2008.01.006

[8] C. Kuruwita, C. Gallagher and K.B. Kulasekera, Testing equality of nonparametric quantile regression functions, *International Journal of Statistics and Probability*, **3** (2014), 55. https://doi.org/10.5539/ijsp.v3n1p55

[9] U. Tonggumnead, Testing the equality of two parametric quantile regression curves: the application for comparing two data sets, *Electronic Journal of Applied Statistical Analysis*, **9** (2016), 17-39.

[10] G. Zhao and Y. Ma, Robust nonparametric kernel regression estimator, *Statistics & Probability Letters*, **116** (2016), 72-79. https://doi.org/10.1016/j.spl.2016.04.010

[11] D.A. Freedman, Bootstrapping regression models, *The Annals of Statistics*, **9** (1981), 1218-1228. https://doi.org/10.1214/aos/1176345638

[12] B.W. Silverman and G.A. Young, The bootstrap: To smooth or not to smooth?. *Biometrika*, **74** (1987), 469-479. https://doi.org/10.1093/biomet/74.3.469

[13] M.G. Akritas and I. Van Keilegom, Non-parametric Estimation of the Residual Distribution, *Scandinavian Journal of Statistics*, **28** (2001), 549-567. https://doi.org/10.1111/1467-9469.00254

[14] National Statistical Office, Processing by the Development Indicators database and social NESDB, Office of The National Economic and Social Development Board, (2015).

Received: May 11, 2018; Published: June 30, 2018