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/i20705948v7n2p279

Testing the difference between two sets of data using comparison two linear regression functions
By Tonggumnead U.

Published: 14 October 2014

Electronic Journal of Applied Statistical Analysis
Vol. 07, Issue 02, 2014, 279-291
DOI: 10.1285/i20705948v7n2p279

Testing the difference between two sets of data using comparison two linear regression functions

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: 14 October 2014

This study aims to compare two sets of data with each having a linear relationship between the independent and dependent variables. The problem is solved by testing the equality of two regression functions. The test statistics based on empirical distribution function: the Kolmogorov-Smirnov and Kuiper type statistics are considered, under the alternative hypotheses comprised of a constant shift and an affine shift. Additionally, the rejection proportion is calculated using the bootstrap method. The test statistics are also applied to the analysis of two sets of data, the characteristics of which are found to be consistent with the p-value after 1,000 trials of bootstrapping..

keywords: regression function, linear relationship, empirical distribution function, bootstrap procedure, error distribution.

1 Introduction

A large number of studies comparing two independent sets of data using the t-test have been carried out for a variety of purposes, such as examining the effects of two fertilizers on the difference in corn yields per acre with controlled cultivation fields. This is considered quantitative data with the dependent variable Y and qualitative (categorical) with an independent variable X, and no assumption is made about the nature of the

*Corresponding author: unchalee-t@hotmail.com

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

relationship. A different problem is one in which two sets of data with each having a linear relationship between the independent variable X and the dependent variable Y, such as comparing the expenditure per household of thai citizens in the first year $(Y_{i1}, i = 1, ..., n_1)$ with that in the second year $(Y_{i2}, i = 1, ..., n_2)$,$X_{i1}, X_{i2}$ represent the income per household of thai citizens from the first year and the second respectively. Generally, the linear relationship between the independent and dependent variables can be investigated with regression analysis using the following regression equation (1):

$$\underline{Y} = \underline{X}\underline{\beta} + \underline{\varepsilon} = f(\underline{X}, \underline{\beta}) + \underline{\varepsilon}. \tag{1}$$

Where $\underline{X}\underline{\beta} = f(\underline{X}, \underline{\beta})$ is the linear regression function, $\underline{\mathrm{Y}}$ is an $(n \times 1)$ vector of the observations, $\underline{\mathrm{X}}$ is an $(n \times p)$ matrix of the level of the independent variables, $\underline{\beta}$ is a $(p \times 1)$ vector of the regression coefficients, and $\underline{\varepsilon}$ is an $(n \times 1)$ vector of random errors. The error in equation (1) are normally and independently distributed with mean zero and constant variance $\sigma^2[NID(0, \sigma^2)]$. The vector of fitted valued $\underline{\widehat{Y}}$ correspondent to the observed value Y_{ij} is :

$$\underline{\widehat{Y}} = \underline{X}\underline{b} = f(\underline{X}, \underline{b}). \tag{2}$$

This research focuses on the difference between two sets of data with the independent variable X and the dependent variable Y having the simple linear regression relationship, $\underline{Y} = \underline{X}\underline{\beta} + \underline{\varepsilon} = f(\underline{X}, \underline{\beta}) + \underline{\epsilon}$,where $\underline{X}\underline{\beta} = f(\underline{X}, \underline{\beta})$ is the linear regression function, $\underline{\mathrm{Y}}$ is an $(n \times 1)$ vector of the observations, $\underline{\mathrm{X}}$ is an $(n \times 2)$ matrix of the level of the independent variables, $\underline{\beta}$ is a (2×1) vector of the regression coefficients, and $\underline{\varepsilon}$ is an $(n \times 1)$ vector of random errors, that is, if $\underline{Y}_1$ represent the vector of the expenditure per household in the first year and $\underline{Y}_2$ represent the vector of the expenditure per household in the second year, $\underline{X}_1$, $\underline{X}_2$ represent $(n \times 2)$ matrix of the income per household, $f_1(\underline{X}, \underline{\beta})$ and $f_2(\underline{X}, \underline{\beta})$ will represent the expected values of $\underline{Y}$ in the first and the second years respectively, if $f_1(\underline{X}, \underline{\beta}) = f_2(\underline{X}, \underline{\beta})$ it can be concluded that the expenditure per household from two years are not different.

This problem has been examined in some research, Clogg et al. (1995) discussed for comparison of the regression coefficients of two models in the situation where one is nested in the other. Comparison of this are of interest when ever two explanations of given phenomenon are specified as linear model.Brame et al. (1998) discussed a test for the equality of two independent equations with respect to the maximum-likelihood regression coefficient where the maximum-likelihood estimate of regression coefficients is derived for the respective populations from which the two largest independent samples are drawn. Moreno et al. (2005) presented a test for testing the equality of regression coefficients in heteroscedastic normal regression models. This research addresses the problem of testing whether the vector of regression coefficients are equal for two independent normal regression models when the error variance are unknown.This study assume two normal regression: $\underline{Y}_1 = \underline{X}_1\underline{\beta}_1 + \varepsilon_1$, $\varepsilon_1 \sim N(0, \sigma_1^2\underline{I})$, and $Y_2 = \underline{X}_2\underline{\beta}_2 + \varepsilon_2$, $\varepsilon_2 \sim N(0, \sigma_2^2\underline{I})$, the hypothesis is $H_0 : \underline{\beta}_1 = \underline{\beta}_2$ versus $H_1 : \underline{\beta}_1 \neq \underline{\beta}_2$.The Bayesian approach are applied in this problem. Nevertheless, such studies do not involve a real comparison of two sets of data having a relationship between the independent variable X

and the dependent variable Y. To solve the problem in question. Pardo-Fernndez (2007) presented the test base on Kolmogorov-Smirnov and Cramer-von Mises type statistics for comparing the equality of k-error distributions,and formulate the test hypotheses $H_0 : F_{\varepsilon 1} = F_{\varepsilon 2} =, ..., = F_{\varepsilon k}, H_1 : F_{\varepsilon i} \neq F_{\varepsilon j}$ for some i,j, $j \in 1, ..., k$, while Pardo-Fernndez et al. (2007) presented the test statistics for testing the difference between k-regression curves using the principles of the Kolmogorov-Smirnov and Cramer-von Mises type statistics in the framework of non-parametric curve estimation. Mohdeb et al. (2010) presented a new methodology for comparing regression function f_1 and f_2 in the case of homoscedastic error structure and fixed design. The test statistics based on the empirical fourier coefficients of the regression function f_1 and f_2 are considered. Feng et al. (2014) discussed a test concern about robust comparison of two regression curves, and a robust testing procedure is recommended under a framework of the generalized likelihood(GLR). These papers are mainly devoted to testing for the equality of two or more regression curves. Therefore, testing for the equality of two regression function is the best way to compare the difference between two sets of data of which the relationship between the independent and dependent variables. The objective of the present research is to compare the difference between two sets of data of which the relationship between the independent and dependent variables is in the form of a simple linear regression. This is examined by testing the difference between two simple linear regression functions using the following hypotheses:

$$H_0 : f_1(\underline{X}, \underline{\beta}) = f_2(\underline{X}, \underline{\beta}) \; versus \; H_1 : f_1(\underline{X}, \underline{\beta}) \neq f_2(\underline{X}, \underline{\beta}). \tag{3}$$

According to Pardo-Fernndez et al. (2007), if the empirical distribution function of the residuals of each regression function is similar, the null hypothesis $H_0 : f_1(\underline{X}, \underline{\beta}) = f_2(\underline{X}, \underline{\beta})$ will be confirmed. On the other hand, should the empirical distribution function of the residuals of each regression function be different, the alternative hypothesis $H_1 : f_1(\underline{X}, \underline{\beta}) \neq f_2(\underline{X}, \underline{\beta})$ will be confirmed. In this study, the estimator of the error $\hat{\varepsilon}_{ij} = Y_{ij} - \hat{Y}_{ij}, j = 1, 2, i = 1, ..., n_j$ the estimator of the error under the null hypothesis $\hat{\varepsilon}_{ij} = Y_{ij} - \hat{Y}_{ij}^0$ where $\hat{Y}_{ij}^0 = f(\underline{X}, b)$ is the common predicted value under the null hypothesis. The common linear regression function under the null hypothesis is estimated from the overall of two sets of data. $f_1(\underline{X}, \underline{\beta})$ and $f_2(\underline{X}, \underline{\beta})$ are the estimator of the first linear regression function and the second respectively, The test statistics based on empirical distribution: Kolmogorov-Smirnov and Kuiper type statistics are applied in comparing two sets of data based on a test of equality between two linear regression functions.

In the following section, the details of testing procedure is given, then the bootstrap and simulation studies are presented. The application of the data and conclusions are also included in the next section.

2 Materials and Method

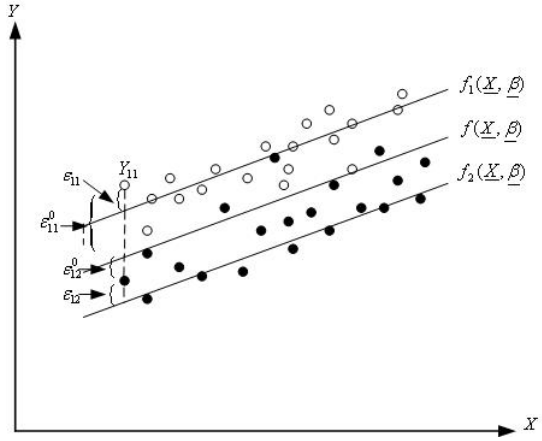

Figure 1: Illustrate the expectation of the error when varies the regression function

From Figure1. For $j = 1, 2, i = 1, ..., n_1$. $f_1(\underline{X}, \underline{\beta})$ and $f_2(\underline{X}, \underline{\beta})$ are the first and the second linear regression function respectively, $f(\underline{X}, \underline{\beta})$ is the common linear regression function when the null hypothesis is confirmed. This function is estimated form overall of two sets of data, $E(\epsilon_{ij})$ is the expectation of the error in each regression function $f_j(\underline{X}, \underline{\beta})$, and $E(\epsilon_{ij}) = 0$, at the same time $E(\epsilon^0_{ij})$ is the expectation of the error from common regression function $f(\underline{X}, \underline{\beta})$, normally $E(\varepsilon^0_{ij}) > 0$, $E(\varepsilon^0_{ij}) = 0$ where$f(\underline{X}, \underline{\beta}) = f_1(\underline{X}, \underline{\beta}) = f_2(\underline{X}, \underline{\beta})$.The idea of the testing procedure is to construct the estimator of the error in each population and the estimator of the error under the null hypothesis, and defined the empirical distribution functions of these estimated residual. In this research, $\hat{\varepsilon}_{ij} = Y_{ij} - \hat{Y}_{ij}$ is an estimator of the error in each linear regression function, $\hat{\varepsilon}_{ij} = Y_{ij} - \hat{Y}^0_{ij}$ is an estimator of the error under the null hypothesis, the estimator of the distribution of errors in each regression function is:

$$\hat{F}_{\varepsilon j}(y) = \frac{1}{n_j} \sum_{i=1}^{n_j} I(Y_{ij} - \hat{Y}_{ij} \leq y), j = 1, 2, i = 1, ..., n_j, -\infty < y < \infty. \tag{4}$$

The distribution of errors when the null hypothesis is confirmed is:

$$\hat{F}^0_{\varepsilon}(y) = \frac{1}{n_j} \sum_{i=1}^{n_j} I(Y_{ij} - \hat{Y}^0_{ij} \leq y), j = 1, 2, i = 1, ..., n_j, -\infty < y < \infty. \tag{5}$$

According to Akritas and Keilegom (2001), Pardo-Fernndez (2007), and Pardo-Fernndez et al. (2007), if the null hypothesis is confirmed, both $\hat{F}_{\varepsilon j}(y)$ and $\hat{F}^0_{\varepsilon}(y)$ are the estimators of $F_{\varepsilon j}(y)$. In contrast, if the alternative hypothesis is confirmed, the distribution of the error will be estimated from a different linear regression function, In this study,

the distribution of errors for each population is compared using the two-dimensions process $\hat{V}(y) = (\hat{V}_1(y), \hat{V}_2(y))$ when $\hat{V}_j(y) = n_j^{1/2}(\hat{F}_\varepsilon^0(y) - \hat{F}_{\varepsilon j}(y)), j = 1, 2, i = 1, ..., n_j$the Kolmogorov-Smirnov type statistic $Z_{ks} = \sum_{i=j}^{2} sup_y|n_j^{1/2}(\hat{F}_\varepsilon^0(y) - \hat{F}_{\varepsilon j}(y))|$ and Kuiper type statistic $Z_{ku} = \sum_{i=j}^{2}[sup_y|n_j^{1/2}(\hat{F}_\varepsilon^0(y) - \hat{F}_{\varepsilon j}(y))| - inf_y|n_j^{1/2}(\hat{F}_\varepsilon^0(y) - \hat{F}_{\varepsilon j}(y))|]$ are applied. According to Pardo-Fernndez et al. (2007), let $f_j(\underline{X}, \underline{\beta})$ be a continuous function. For $j = 1, 2, i = 1, ..., n_j$, $F_\varepsilon^0 y = F_{\varepsilon j} y$, $-\infty < y < \infty$, if and only if $f(\underline{X}, \underline{\beta}) = f_1(\underline{X}, \underline{\beta}) = f_2(\underline{X}, \underline{\beta})$. Namely, assume $F_\varepsilon^0(y) = F_{\varepsilon j}(y)$. This implies that two empirical distributions of the errors are equal, there are evidence for the equality of the 1^{st} moment, $E[Y_{ij} - f(\underline{X}, \underline{\beta})] = E[Y_{ij} - f_j(\underline{X}, \underline{\beta})] = 0$, then $f(\underline{X}, \underline{\beta}) = f_j(\underline{X}, \underline{\beta})$. In the same way, the 2^{nd} moment have originate from the 1^{st} moment, then $f(\underline{X}, \underline{\beta}) = f_j(\underline{X}, \underline{\beta})$. Namely, $f(\underline{X}, \underline{\beta}) = f_1(\underline{X}, \underline{\beta}) = f_2(\underline{X}, \underline{\beta})$. Conversely, assume $f(\underline{X}, \underline{\beta}) = f_1(\underline{X}, \underline{\beta}) = f_2(\underline{X}, \underline{\beta})$. Claim that $F_\varepsilon^0(y) = F_{\varepsilon j}(y)$. Consider the 1^{st} moment: From $f(\underline{X}, \underline{\beta}) = f_j(\underline{X}, \underline{\beta})$ then, $E[Y_{ij} - f(\underline{X}, \underline{\beta})] = E[Y_{ij} - f_j(\underline{X}, \underline{\beta})] = 0$. Consider the 2^{nd} moment: From $f(\underline{X}, \underline{\beta}) = f_j(\underline{X}, \underline{\beta})$ then, $E[Y_{ij} - f(\underline{X}, \underline{\beta})]^2 = E[Y_{ij} - f_j(\underline{X}, \underline{\beta})]^2$. From the 1^{st} moment and the 2^{nd} moment, if $f(\underline{X}, \underline{\beta}) = f_1(\underline{X}, \underline{\beta}) = f_2(\underline{X}, \underline{\beta})$. then $F_\varepsilon^0 y = F_{\varepsilon j} y$. Therefore, a comparison of two sets of data with each having a linear relationship between the independent and dependent variables can be examined through the equality of two linear regression functions by considering the equality of the distribution of errors. As regards the distribution of $\hat{V}_j(y)$ when the null hypothesis is confirmed, according to Donsker (1952), Donskers theorem,$\hat{V}_j(y) = n_j^{1/2}(\hat{F}_\varepsilon^0(y) - \hat{F}_{\varepsilon j}(y))$ are the random elements of the Skorokhod space $D(-\infty, \infty)$, and converge in distribution to Gaussian process with zero mean and covariance $F_\varepsilon^0(y)(1 - F_\varepsilon^0(y))$.

3 Bootstrap and Simulation Studies

3.1 Bootstrap

According to Freedman (1981), Silverman and Young (1987), Akritas and Keilegom (2001), and Pardo-Fernndez et al. (2007) , bootstrap method bring many benefit for estimate the critical value of the test statistics. In this section, bootstrap procedure are applied for estimating the critical value of the test statistics Z_{ks} and Z_{ku}. The procedures are as follows:

3.1.1 Assume the bootstrap replication b = 1,..., B (B=300), for j =1, 2, i = 1,.., n_j, the new response under the null hypothesis $Y_{ij,b}^*$, b=1,..., B defined as:

$$Y_{ij,b}^* = f_j(\underline{X}, \underline{\beta}) + \varepsilon_{ij,b}^*, j = 1, 2, i = 1, ..., n_j. \tag{6}$$

3.1.2 For j =1, 2, i = 1,.., n_j, calculate the test statistics Z_{ks} and Z_{ku} from the bootstrap samples X_{ij}, $Y_{ij,b}^*$.

3.1.3 Let $Z_{ks,b}^*$ and $Z_{ku,b}^*$ be the order statistics of $Z_{ks(1)}^*$,...,$Z_{ks(b)}^*$ and $Z_{ku(1)}^*$,...,$Z_{ku(b)}^*$ from 300 bootstrap replications respectively, $Z_{ks(1-\alpha)B}^*$ and $Z_{ku(1-\alpha)B}^*$ approximate the $(1 - \alpha)$-quantile of the distribution of Z_{ks} and Z_{ku} under the null hypothesis.

3.1.4 The test statistics Z_{ks} and Z_{ku} are iterated for 1,000 trials, and the proportion of rejections are displayed.

3.2 Simulation

3.2.1 Assume four regression functions in a linear form, and the shifts under the alternative hypothesis are made up of a constant shift and an affine shift. The model is as follows:

(1) $f_1(x) = f_2(x) = 2x$.
(2) $f_1(x) = f_2(x) = 2x + 2$.
(3) $f_1(x) = 2x, f_2(x) = 2x + 2$. (constant shift)
(4) $f_1(x) = 2x, f_2(x) = 2x + x$. (affine shift)

3.2.2 The distribution of error $\varepsilon_{i1} \sim N(0, 1)$ and $\varepsilon_{i2} \sim N(0, 1)$, i=1,...,n, j=1,2.

3.2.3 In all cases, the covariates X_{i1} and X_{i2}, i = 1,...,n, j =1, 2 are uniformly distributed on the interval [0,1].

3.2.4 The sample size is determined in two way: equal in which case $(n_1, n_2) = (20, 20)$,$(n_1, n_2) = (50, 50)$ $(n_1, n_2) = (100, 100)$and unequal in which case $(n_1, n_2) = (20, 50)$,$(n_1, n_2) = (20, 100)$ $(n_1, n_2) = (50, 100)$.

4 Results

As shown in Table 1. and Figure 2. Under the null hypothesis, the type I error is well-approximated when the sample size becomes larger, and model (2), in the form of Y = bx + a, will have a slightly higher rejection proportion than model (1), which does not have a y-intercept. Next, under the null hypothesis we examine the type I error of two methods: Z_{ks} and Z_{ku}, the value in Table1. and Figure 2. clearly shows that the type I error of Z_{ku} is slightly lower than Z_{ks} for model (1)and model(2)for all situation. When we consider about the type I error controlling, the performance of Z_{ks} is better than Z_{ku}. As shown in Table 2. and Figure 3. Under the alternative hypothesis, the power of the test statistic will get higher with a larger sample size. Additionally, the affine shift (model (4)) results in a higher rejection percentage than the constant shift (model (3)). Next, under the alternative hypothesis we examine the power of the test of two methods:Z_{ks} and Z_{ku} , the value in Table2. and Figure 3. clearly shows that the power of the test of Z_{ku} is slightly lower than Z_{ks} for all situation, and the sample size: equal in which case or unequal in which case not effect to the performance of the test statistics. Finally, the performance of the test statistics based on Kolmogorov-Smirnov is better than Kuiper for all situation. Therefore, Kolmogorov-Smirnovtest statistic based on comparing empirical distribution of error is the one choice for comparing two sets of data with each having a linear relationship between the independent and dependent variables.However, this research focus under the assumption of linear regression model is true, in practice the assumption about the data may not hold, such as: error distribution, heavy tailed of error distribution, outlier etc., the test statistic for solve these problem should be considered.

Table 1: Rejection proportions under the null hypothesis; models (1) and (2) of the test statistics Z_{ks} and Z_{ku} .

sample size	model	$z_{ku} : \alpha = 0.05$	$z_{ku} : \alpha = 0.10$	$z_{ks} : \alpha = 0.05$	$z_{ks} : \alpha = 0.10$
(20,20)	1	0.040	0.068*	0.041	0.074*
(20,20)	2	0.042	0.071*	0.043	0.080*
(20,50)	1	0.042	0.072*	0.045	0.084
(20,50)	2	0.043	0.075*	0.049	0.085
(20,100)	1	0.045	0.075*	0.048	0.085
(20,100)	2	0.045	0.075*	0.048	0.086
(50,50)	1	0.047	0.081*	0.048	0.088
(50,50)	2	0.048	0.083*	0.052	0.089
(50,100)	1	0.050	0.089	0.051	0.101
(50,100)	2	0.050	0.092	0.051	0.102
(100,100)	1	0.050	0.098	0.052	0.102
(100,100)	2	0.051	0.101	0.052	0.102

*The type I error out of control interval.

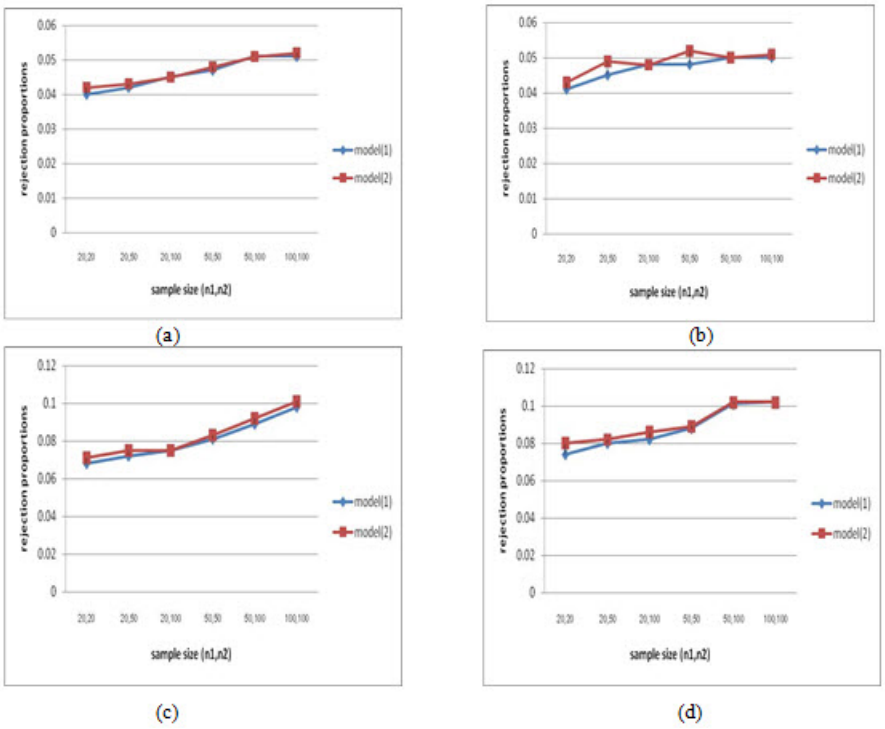

Figure 2: (a) Presents the rejection proportions under the null hypothesis of the test statistic Z_{ku}, $\alpha = 0.05$, (b) Presents the rejection proportions under the null hypothesis of the test statistic Z_{ks}, $\alpha = 0.05$, (c) Presents the rejection proportions under the null hypothesis of the test statistic Z_{ku}, $\alpha = 0.10$,(d) Presents the rejection proportions under the null hypothesis of the test statistic Z_{ks}, $\alpha = 0.10$.

Table 2: Rejection proportions under the alternative hypothesis; models (3) and (4) of the test statistics Z_{ks} and Z_{ku}.

sample size	model	$z_{ku} : \alpha = 0.05$	$z_{ku} : \alpha = 0.10$	$z_{ks} : \alpha = 0.05$	$z_{ks} : \alpha = 0.10$
(20,20)	3	0.885	0.900	0.886	0.912
(20,20)	4	0.887	0.910	0.900	0.918
(20,50)	3	0.890	0.911	0.899	0.919
(20,50)	4	0.888	0.911	0.905	0.920
(20,100)	3	0.898	0.914	0.920	0.935
(20,100)	4	0.898	0.921	0.942	0.945
(50,50)	3	0.905	0.927	0.925	0.938
(50,50)	4	0.908	0.928	0.927	0.948
(50,100)	3	0.911	0.928	0.925	0.948
(50,100)	4	0.913	0.930	0.930	0.966
(100,100)	3	0.913	0.932	0.928	0.964
(100,100)	4	0.914	0.932	0.930	0.975

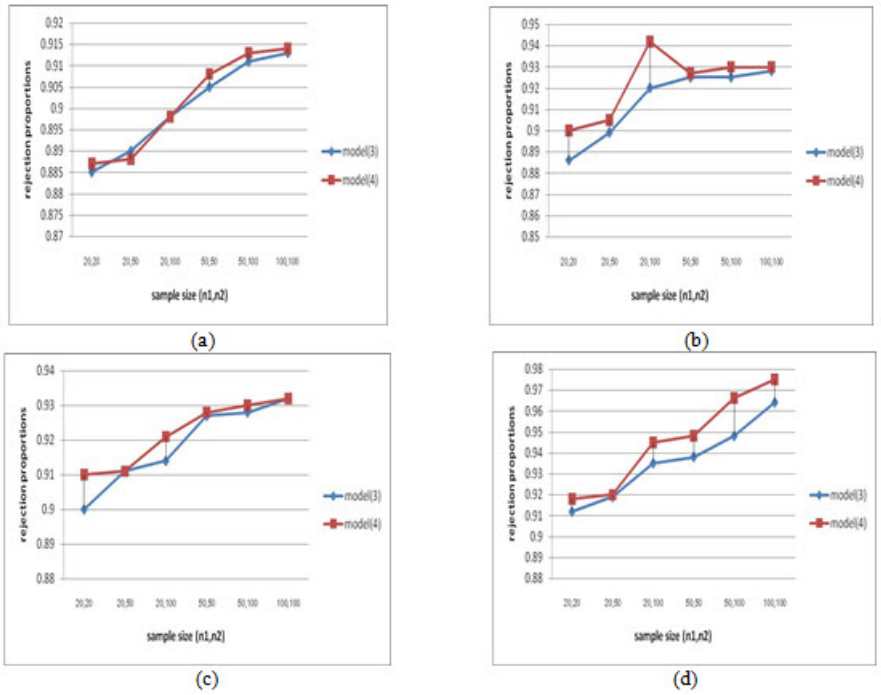

Figure 3: (a) Presents the rejection proportions under the alternative hypothesis of the test statistic Z_{ku}, $\alpha = 0.05$, (b) Presents the rejection proportions under the alternative hypothesis of the test statistic Z_{ks}, $\alpha = 0.05$, (c) Presents the rejection proportions under the alternative hypothesis of the test statistic Z_{ku}, $\alpha = 0.10$,(d) Presents the rejection proportions under the alternative hypothesis of the test statistic Z_{ks}, $\alpha = 0.10$.

5 Application of the Data

This section applies the test statistic Z_{ks} and Z_{ku} to the comparison of two sets of data from the National Statistical Officer Thailand that having a linear relationship between the independent variable X and the dependent variable Y. The first set of data is comprised of the average income per household (X) and the average expenditure per household (Y) of Thai citizens in each province for 2006 and 2007, while the second is made up of the average income per household (X) and the average expenditure per household (Y) of Thai citizens in each province for 2004 and 2007. (National Statistical Officer Thailand, 2009) The linear regression function is estimated using the maximum-likelihood estimator before the p-value is calculated from the test statistics Z_{ks} and $Z_{ku}$by carrying out 1,000 times of bootstrapping. The results are shown in Figure 4. (a) and (b).

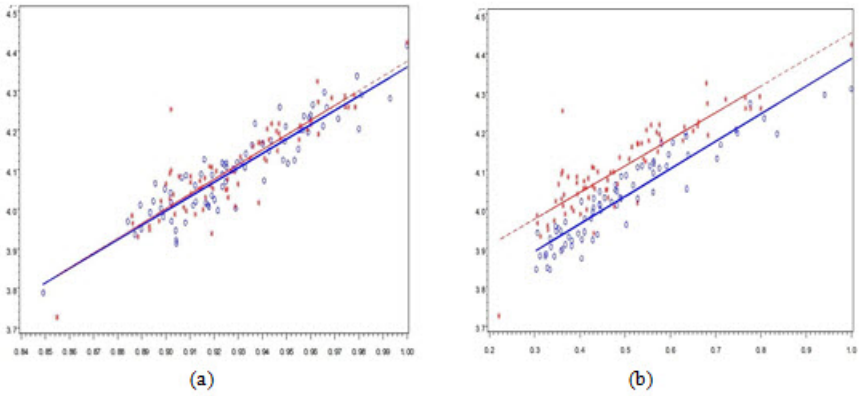

Figure 4: (a) Illustrates the scatter plot and the linear regression function of the average income per household with the data being transformed in the range of [0,1] and log of the average expenditure per household . The data for 2006 are represented by circles and the solid line, whereas those for 2007 are represented by stars and the dash line. Figure 4. (b) Illustrates the scatter plot and the linear regression function of the average income per household with the data being transformed in the range of [0,1] and log of the average expenditure per household. The data for 2004 are represented by circles and the solid line, while those for 2007 are represented by stars and the dash line.

As shown in Figure 4. (a) when the distribution of data and the two linear regression functions are considered, the data relating to the average income per household and the average expenditure per household in 2006 and 2007 are nearly similar with the p-value from 1,000 trials of bootstrapping, the p-value are 0.213 and 0.204 for the test statistics based on Z_{ks} and Z_{ku} respectively . In contrast, an analysis of Figure 4. (b) reveals that the corresponding data for 2004 and 2007 are remarkably different with the p-value of 0.015 and 0.011 for the test statistics based on Z_{ks} and Z_{ku} respectively.

6 Conclusions

The present study differs from previous research in that it does not involve only the dependent variable Y and qualitative (categorical) with an independent variable X that no assumption about the nature of the relationship, but takes into account the relationship between independent and dependent variables having a simple linear regression relationship. The problem is solved by testing the equality of two linear regression functions, based on the principle that two similar linear regression functions represent two similar sets of data . From the simulation results using bootstrapping under the null hypothesis, the type I error is found to be well-approximated with a increasing sample size, consistent with Pardo-Fernndez et al. (2007). In comparison, under the alternative hypothesis, the power of the test will become stronger when the sample size gets larger, and the affine shift will yield a higher rejection percentage than the constant shift. For type I error controlling, the performance of Z_{ks} better than Z_{ku}. As for the case of the test statistics based on Kolmogorov-Smirnov type statistic and Kuiper type statistic for testing the equality between two sets of data, the performance of the test statistics based on Kolmogorov-Smirnov is better than Kuiper for all situation. When the test statistics are applied to the actual data, the findings are consistent with the p-value after 1,000 trials of bootstrapping. This research is the one choice for comparing two sets of data with each having a linear relationship between the independent and dependent variables. Recommendations for further research are provided as follows. First, the relationship between dependent variable Y and more than one independent variable X should be investigated. Second, in practice the assumption about the data may not hold, especially: error distribution, heavy tailed of error distribution, outlier etc., the test statistic for solve these problem should be considered. Third, linear models may not be practical in actual settings mostly involving nonlinear relationships. Thus, studies along this line should analyze the phenomenon using more than one independent variable, develops a flexible and robust testing procedure to compare two sets of data, and consider about a nonlinear function in order that problems involving a comparison of two sets of data can be truly solved.

References

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

Brame, R., Paternoster, R., Mazerolle, P., and Piquero, A. (1998). Testing for the equality of maximum-likelihood regression coefficients between two independent equations. *Journal of Quantitative Criminology*, 14(3), 245-261.

Clogg, C. C., Petkova, E., and Haritou, A. (1995). Statistical methods for comparing regression coefficients between models. *American Journal of Sociology*, 1261-1293.

Donsker, M. D. (1952). Justification and extension of Doob's heuristic approach to the Kolmogorov-Smirnov theorems. *The Annals of mathematical statistics*, 277-281.

Feng, L., Zou, C., Wang, Z., and Zhu, L. (2014). Robust comparison of regression curves. *TEST*, 1-20.

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

Mohdeb, Z., Mezhoud, K. A., and Boudaa, D. (2010). Testing the equality of nonparametric regression curves based on Fourier coefficients. *Afrika Statistika*, 5(1).

Moreno, E., Torres, F., and Casella, G. (2005). Testing equality of regression coefficients in heteroscedastic normal regression models. *Journal of statistical planning and inference*,131(1), 117-134.

National Statistical Officer Thailand (2009). 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. (2007). Comparison of error distributions in nonparametric regression. *Statistics and probability letters*, 77(3), 350-356.

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.