UNIVERSITI PUTRA MALAYSIA

GOODNESS-OF-FIT TEST FOR STANDARD LOGISTICS DISTRIBUTION WITH OUTLIERS

LIM FONG PENG

FS 2010 11

R^2_{MAD} GOODNESS-OF-FIT TEST FOR STANDARD LOGISTICS DISTRIBUTION WITH OUTLIERS

LIM FONG PENG

MASTER OF SCIENCE
UNIVERSITI PUTRA MALAYSIA

2010

R^2_{MAD} GOODNESS-OF-FIT TEST
FOR STANDARD LOGISTICS DISTRIBUTION
WITH OUTLIERS

By

LIM FONG PENG

Thesis Submitted to the School of Graduate Studies, Universiti Putra Malaysia, in Fulfilment of the Requirements for the Degree of Master of Science

March 2010

Abstract of thesis presented to the Senate of Universiti Putra Malaysia
in fulfilment of the requirement for the degree of the Master of Science

R^2_{MAD} GOODNESS-OF-FIT TEST

FOR STANDARD LOGISTICS DISTRIBUTION
WITH OUTLIERS

By

LIM FONG PENG

March 2010

Chairman: Noor Akma Ibrahim, PhD

Faculty: Science

Alternative to the least square coefficient of determination (R^2_{OLS}), the coefficient of determination based on median absolute deviation, R^2_{MAD}, is an attractive consideration in the construction of goodness-of-fit test based on regression and correlation, due to its robustness.

This study presents the observations made from the resulting plots and descriptive measures obtained from contaminated standard logistic distribution. Contamination is introduced to investigate perseverance of robustness property of R^2_{MAD} for samples from the standard logistic distribution. The sampling distribution of R^2_{MAD} is simulated for various sample sizes (n = 20, 40, 100), percentage of contamination (5%, 15%, 25%) and distribution of the contaminants (logistic (2, 0.2), logistic (0, 0.2), logistic (2, 1) and normal (3, 0.2) contaminants). The symmetricity of the sampling distribution of R^2_{MAD} is observed and followed by the investigation of the

confidence intervals of R^2_{MAD} in the presence of outliers. The study of confidence interval estimates for the mean and standard deviation of R^2_{MAD} was conducted using the bootstrap (BCa) method.

Tables of critical values for samples from the standard logistic distribution using $Z_{MAD} = 1 - R^2_{MAD}$ and $Z_{OLS} = 1 - R^2_{OLS}$ are constructed. The tables obtained then are used in the power study on the goodness-of-fit tests using test statistic for alternative distributions and contaminated alternative distributions. For lognormal, exponential and standard logistic alternatives, Z^*_{MAD} and Z^*_{OLS} are simulated for various sample sizes (n =10, 20, 30, 50, 100), percentage of contamination (5%, 15%, 25%, 40%) and distribution of the contaminants (logistic (2, 0.2), logistic (0, 0.2), logistic (2, 1) and normal (3, 0.2) contaminants) for different percentiles (=0.01, 0.025, 0.05, 0.1), respectively. The results indicated that the test statistic Z_{MAD} is able to discriminate the sample that comes from alternative distributions as the test statistic Z_{OLS}.

Abstrak tesis yang dikemukakan kepada Senat Universiti Putra Malaysia sebagai memenuhi keperluan untuk ijazah Master Sains

R^2_{MAD} UJIAN KEBAGUSAN PENYUAIAN BAGI TABURAN LOGISTIK PIAWAI DENGAN PENCEMAR

Oleh

LIM FONG PENG

March 2010

Pengerusi: Noor Akma Ibrahim, PhD

Fakulti: Sains

Alternatif bagi pekali penentuan kuasa dua terkecil (R^2_{OLS}), iaitu pekali penentuan yang berdasarkan median lencongan mutlak, R^2_{MAD}, merupakan satu pertimbangan menarik dalam pembinaan ujian kebagusan penyuaian berdasarkan regresi dan korelasi, disebabkan keteguhannya.

Kajian ini memerihalkan pemerhatian yang dibuat daripada plot yang diperolehi dan ukuran pemerihalan daripada taburan logistik piawai tercemar. Pencemaran diperkenalkan untuk menyelidiki sifat keteguhan R^2_{MAD} bagi sampel daripada taburan logistik piawai. Taburan pensampelan R^2_{MAD} disimulasikan bagi pelbagai saiz sampel (n = 20, 40, 100), peratusan pencemaran (5%, 15%, 25%) dan taburan pencemar (pencemar logistik (2, 0.2), logistik (0, 0.2), logistik (2, 1) dan normal (3, 0.2)). Sifat simetri bagi taburan pensampelan R^2_{MAD} diperhatikan dan ini diikuti dengan kajian selang keyakinan bagi R^2_{MAD} dengan kehadiran titik terpencil.

Jadual nilai kritis bagi sampel daripada taburan logistik piawai yang menggunakan $Z_{MAD} = 1 - R^2_{MAD}$ dan $Z_{OLS} = 1 - R^2_{OLS}$ masing-masing dibina. Jadual-jadual yang diperolehi itu kemudian digunakan dalam kajian kuasa pada ujian kebagusan penyuaian yang menggunakan statistik ujian bagi taburan alternatif dan taburan alternatif tercemar. Bagi alternatif lognormal, eksponen dan logistik piawai, Z^*_{MAD} dan Z^*_{OLS} masing-masing disimulasikan bagi pelbagai saiz sampel (n =10, 20, 30, 50, 100), peratusan pencemaran (5%, 15%, 25%, 40%) dan taburan pencemar (pencemar logistik (2, 0.2), logistik (0, 0.2), logistik (2, 1) dan normal (3, 0.2)) bagi persentil yang berbeza (0.01, 0.025, 0.05, 0.1). Keputusan yang diperoleh menunjukkan bahawa statistik ujian Z_{MAD} berupaya untuk mendiskriminasi sampel yang datang daripada taburan alternatif seperti statistik ujian Z_{OLS} .

ACKNOWLEDGEMENTS

First of all, I sincerely express my infinite gratitude to my supervisor, Associate Professor Dr. Noor Akma Ibrahim for her encouragement, patience and critical reviews towards the completion of this project. Without her guidance and help, I could never have accomplished this difficult task. I would also like to extend my gratitude to Puan Fauziah bt Maarof and Prof. Madya Dr. Kassim bin Harun for their guidance, providing related information, and advice. I feel thankful to all lecturers who had taught me before.

I am especially indebted to those who had given me advice and providing some suggestions for my study. I would also like to thank all my fellow friends who had given me support and encouragement when I encountered problems and on the brink of giving up.

Finally, my appreciation goes to my dear parents and my family for their encouragement and continuous support. I would like to share my joy with them, they will always be in my heart.

I certify that a Thesis Examination Committee has met on 3 March 2010 to conduct the final examination of Lim Fong Peng on her thesis entitled "R^2_{MAD} Goodness-of-fit Test for Standard Logistic Distribution with Outliers" in accordance with the Universities and University Colleges Act 1971 and the Constitution of the Universiti Putra Malaysia [P.U.(A) 106] 15 March 1998. The Committee recommends that the student be awarded the Master of Science.

Members of the Examination Committee were as follows:

Isa Daud, PhD
Associate Professor
Faculty of Science
Universiti Putra Malaysia
(Chairman)

Mohd Rizam Abu Bakar, PhD
Associate Professor
Faculty of Science
Universiti Putra Malaysia
(Internal Examiner)

Jayanthi a/p Arasan, PhD
Faculty of Science
Universiti Putra Malaysia
(Internal Examiner)

Kamarulzaman Ibrahim, PhD
Associate Professor
Faculty of Science and Technology
Universiti Kebangsaan Malaysia
Malaysia
(External Examiner)

BUJANG KIM HUAT, PhD
Professor and Deputy Dean
School of Graduate Studies
Universiti Putra Malaysia

Date: 20 May 2010

This thesis was submitted to the Senate of Universiti Putra Malaysia and has been accepted as fulfilment of the requirement for the degree of Master of Science. The members of the Supervisory Committee were as follows:

Noor Akma Ibrahim, PhD
Associate Professor
Institute for Mathematics Research
Universiti Putra Malaysia
(Chairman)

Kassim Harun, PhD
Associate Professor
Faculty of Science
Universiti Putra Malaysia
(Member)

____

HASANAH MOHD. GHAZALI, PhD
Professor and Dean
School of Graduate Studies
Universiti Putra Malaysia

Date: 10 June 2010

DECLARATION

I declare that the thesis is my original work except for quotations and citations which have been duly acknowledged. I also declare that it has not been previously, and is not concurrently, submitted for any other degree at Universiti Putra Malaysia or at any other institution.

____

Lim Fong Peng

Date: 30 January 2010

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Background

Goodness-of-fit techniques are methods of examining how well a sample of data agrees with a given distribution. In the formal framework of hypothesis testing for goodness-of-fit techniques, the hypothesis set is as below,

$\boldsymbol{H}_0$: the data follow a specified distribution.

$\boldsymbol{H}_1$: the data do not follow the specified distribution.

Then, it is usually hoped to accept the null hypothesis and proceed with other analyses (D' Agostino and Stephens, 1986).

In addition to formal hypothesis testing procedures, goodness-of-fit techniques also include less formal method, in particular, graphical techniques. In these methods, graphs are drawn so that deviation from the hypothesized distribution is reflected in certain features of the graph. Graphical techniques are useful because of their ease and informality.

Regression and correlation based test is a graphical technique, related to probability plots, in which the order statistics $\boldsymbol{X_{(i)}}$ (on the vertical axis) are plotted against a suitable function of $\boldsymbol{i}$, $\boldsymbol{T_i}$, (on the horizontal axis). A straight line is then fitted to the points, and goodness-of-fit tests are based on the correlation coefficient between $\boldsymbol{X}$ and $\boldsymbol{T}$.

In this study, a regression and correlation based test is conducted for samples from standard logistic distribution. The statistics of interest are the coefficient of determination from regression obtained by using ordinary least square (OLS) and median absolute deviation (MAD).

The coefficient of determination from ordinary least square regression, $\boldsymbol{R}^2_{OLS}$, is the ratio of the variability of the modeled values, which is obtained by ordinary least square regression, to the variability of the original data values. Or it can be defined as the square of the correlation coefficient, $\boldsymbol{r}_{OLS}$, between the original and modeled data values.

Meanwhile, the median absolute deviation for the correlation coefficient ($\boldsymbol{r}_{MAD}$) has been defined by Gideon (2007). This definition is discussed in Chapter 3. He discovered that $\boldsymbol{r}_{MAD}$ sometimes may go beyond +1 or –1, that is $|\boldsymbol{r}_{MAD}| \leq 1$ is not always true. $\boldsymbol{R}^2_{MAD}$ is an attractive alternative to the coefficient of determination from ordinary least square regression, $\boldsymbol{R}^2_{OLS}$, due to its robustness. Mean absolute deviation (MAD) about the median is a very robust scale estimator (Hampel, 1974). The MAD is clearly not influenced very much by the presence of a large outlier, and as such provides a good robust alternative to the sample standard deviation (Maronna, 2006).

However, the properties of the sampling distribution of $\boldsymbol{R}^2_{MAD}$ have not been well researched. Hence, a simulation study is conducted in this study to obtain graphical and numerical description on the sampling distribution of $\boldsymbol{R}^2_{MAD}$, focusing on samples from the standard logistic distribution. The sampling distribution of $\boldsymbol{R}^2_{MAD}$ above can be used to construct tables of the critical values and to conduct a power study on the goodness-of-fit tests.

In addition, checking for the bound of $\boldsymbol{R}^2_{MAD}$ for samples from standard logistic distribution is also presented before conducting a power study on the goodness-of-fit tests. It is an advantage if $|\boldsymbol{r}_{MAD}|$ is always less than 1, so that the critical values, $\boldsymbol{Z}_{MAD} = 1 - \boldsymbol{R}_{MAD}$, can be calculated in conducting the power study. Hence, the bootstrapping can be a good "tool" to investigate the bound of $\boldsymbol{R}^2_{MAD}$ whether $|\boldsymbol{r}_{MAD}|$ is always less than 1 or not. The bootstrapping method is often used as an alternative for inference based on parametric assumptions when those assumptions are in doubt, or where parametric inference is impossible or requires very complicated formulas for the calculation of standard errors. The goal of bootstrap confidence interval theory is to construct dependable confidence limits for a parameter of interest from the bootstrap distribution of $\hat{\ }$. For more details, refer to DiCiccio and Efron (1996).

There are several bootstrap techniques for constructing confidence intervals, such as bootstrap-***t***, percentile, BC_a (***bias-corrected and accelerated***) and ABC (***approximate bootstrap confidence intervals***) methods. However, $\boldsymbol{BC_a}$ method is preferred and widely used in statistical study since it is an improved version of the percentile