ROBUST REGRESSION WITH CONTINUOUS AND CATEGORICAL VARIABLES HAVING HETEROSCEDASTIC NON-NORMAL ERRORS

By

BASHAR ABDUL AZIZ MAJEED AL-TALIB

Thesis Submitted to the School of Graduate Studies, Universiti Putra Malaysia, in Fulfilment of the Requirement for the Degree of Doctor of Philosophy

August 2006

Dedicated to

The memory of my father
my Dear mother
my wife and
my beloved twin children,
Harith & Bara'a

Abstract of the thesis presented to the Senate of Universiti Putra Malaysia in fulfilment of requirement for the degree of Doctor of Philosophy

ROBUST REGRESSION WITH CONTINUOUS AND CATEGORICAL VARIABLES HAVING HETEROSCEDASTIC NON-NORMAL ERRORS

By

BASHAR ABDUL AZIZ AL-TALIB

August 2006

Chairman: Associate Professor Habshah Midi, PhD

Faculty : Science

The performance of the classical Ordinary Least Squares (OLS) method can be very poor when the data set for which one often makes a normal assumption, has a heavy-tailed distribution which may arise as a result of outliers. The problem is further complicated when the variances of the error terms are not constant.

In this thesis, a Reweighted Least Squares based on Robust Distance and a combination between the S and M-estimates (called S/M Estimates). The resulted estimates will be called (RLSRDSM) estimates, which proposed to overcome the problem of outlier. The RLSRDSM estimates are proposed to estimates the parameters of a regression model with both continuous and categorical variables. The Robust Distance S/M Estimates are computed in three stages where on the first stage the Minimum Volume Ellipsoid (MVE) estimator is computed to identify leverage points, then a weighted S/M weights and scale is calculated in the second and third steps respectively.

In many applications, one may encounter errors which are heteroscedastic and not normally distributed. Therefore, in this thesis, a weighted RLSRDSM (WRLSRDSM) is proposed to remedy these two problems simultaneously. This method first computes the residuals scale estimates for each level of the categorical variables based on RDSM residuals. A weighted scheme is then developed and incorporates in the model.

In addition to RLSRDSM and WRLSRDSM, another estimator which is referred as 2D-RDLS procedure that use two-dimensional weighting scheme is also proposed. However, the performance of the 2D-RDLS estimates is not as good as the RLSRDSM and therefore seldom referred in the discussion.

A number of numerical examples and simulation studies have been performed to compare the robustness of the RLSRDSM, and WRLSRDSM with some existing methods in the regression model with both continuous and categorical regressors. Data with various outlier contamination were simulated and analyzed. Design parameters were varied, include sample size (n=20, 50, 100, 300, and 500), number of continuous regressors (p=1,3, and 5) and categorical data (q=1,4), outliers density (0%, 5%, 10%, 20%, 30%, 40%, and 50%), and different error distribution scenarios (*N(0,0.25), N(0,0.5), N(0,1), N(0,2), N(0,3), N(0,4), t(3), and EXP(1))*.

Criteria used to measure the performance of the regression methods are p-values, residual scale, R^2%, and $100\overline{R}^2$%for the real data analysis and The Root Mean Square Error (RMSE) of the overall simulation replications which summarized the variance and bias for the simulated data.

The results in this thesis indicate that the Ordinary Least Squares (OLS) estimators are very sensitive to the presence of outliers and heteroscedastic errors. In the presence of outliers, the RLSRDSM and $RLSRDL_1$ are better than OLS, by producing robust estimates to such kind of data points. The RLSRDSM is slightly better than $RLSRDL_1$ and sometimes their performances are indistinguishable in the presence of outliers. Nonetheless, the $RLSRDL_1$ posed certain computational problems such as producing degenerate solution or singular matrices. The advantage of RLSRDSM is that it has no computational problems. The performance of WRLSRDSM is better than the $WLSRDL_1$ when both outliers and heteroscedastics occurs together.

In order to support the numerical findings, Bootstrap simulation procedures and visual analysis are also been carried out to justify that the RLSRDSM is the most robust estimator compared to the OLS and $RLSRDL_1$, on a ground that this estimator result with robust and stable in the presence of outliers, combined models with continuous and categorical variables, and even heteroscedasticity problem. The results indeed show that they are in close agreement with the earlier conclusion.

Abstrak tesis yang dikemukakan kepada Senat Universiti Putra Malaysia
sebagai memenuhi keperluan untuk ijazah Doktor Falsafah

REGRESI TEGUH DENGAN PEMBOLEHUBAH SELANJAR DAN BER KATEGORI DENGAN RALAT BER HETEROSKEDASTIK

Oleh

BASHAR ABDUL AZIZ AL-TALIB

Ogos 2006

Pengerusi: **Profesor Madya Habshah Midi, PhD**

Fakulti: **Sains**

Prestasi kaedah Kuasadua Terkecil Biasa Klasik boleh menjadi lemah bila set data yang sepatutnya dianggap normal, mempunyai taburan berhujung tebal yang disebabkan oleh titik terpencil. Masalah ini menjadi semakin rumit, apabila varians bagi ralatnya tidak konstan.

Dalam tesis ini, Kaedah Kuasadua Terkecil-Berpemberat berasaskan Jarak Teguh Angaran S-estimator dan M-estimator S/M (RLSRDSM) dicadangkan untuk mengatasi masalah titik terpencil. Anggaran RLSRDSM dicadangkan untuk menganggar parameter suatu model regresi yang mempunyai pembolehubah selanjar dan berkategori. Anggaran Jarak Teguh S/M dikira dalam tiga tahap, di mana tahap pertama penganggar `Minimum Volume Ellipsoid` (MVE) dikira untuk mengenalpasti titik `leverage`.

Dalam banyak penggunaan, seseorang boleh menjumpai ralat yang ber heteroskedastik dan tidak tertabur secara normal. Oleh yang demikian, dalam tesis ini, RLSRDSM berpemberat (WRLSRDSM) dicadangkan untuk mengatasi dua

masalah ini secara serentak. Kaedah ini dimulakan dangan mengira anggaran skala reja bagi setiap paras pembolehubah berkategori berasaskan reja RDSM. Skim berpemberat kemudiannya dibina dan dimasukkan ke dalam model.

Selain daripada RLSRDSM dan WRLSRDSM, penganggar lain yang dirujuk sebagi prosedur 2D-RDLS yang menggunakan skim pemberat dua-dimensi juga dicadangkan. Bagaimanapun, prestasi penganggar 2D-RDLS tak sebaik RLSRDSM dan $RLSRDL_1$ dan oleh yang demikian ia jarang dirujuk dalam perbincangan. Beberapa contoh numerik dan kajian simulasi telah dijalankan untuk membandingkan keteguhan RLSRDSM, dan WRLSRDSM dengan beberapa kaedah yang sedia ada dalam model regresi dengan pembolehuban selanjar dan berkategori.

Data dengan pelbagai pencemaran titik terpencil dijana dan dianalisis. Rekabentuk parameter dipelbagaikan termasuk saiz sampel (n=20,50,100, 300, dan 500), bilangan pembolehubah selanjar (p=1,3,dan 5) dan bilangan pembolehubah berkategori (q=1,4), ketumpatan titik terpencil (0%, 10%, 20%,30%, 40%, dan 50%), dan senario taburan ralat berbeza ($N(0,25)$, $N(0,0.5)$,$N(0,1)$,$N(0,2)$,$N(0,3)$,$N(0,4)$, $t(3)$ dan $Exp(1)$).

Kriteria yang digunakan untuk mengukur kaedah regresi adalah nilai-p, skala-reja, R^2, dan $100\overline{R}^2$ untuk analisis data nyata, dan Punca Min Kuasa Dua Ralat (RMSE) bagi semua replikasi simulasi yang merengkaskan varians dan `bias` data simulasi.

Keputusan yang diperolehi dalam tesis ini menunjukkan bahawa penganggar Kuasadua Terkecil Biasa (OLS) sangat sensitif kepada kehadiran titik terpencil dan

ralat berheteroskedastik.

Dalam kehadiran titik terpencil, RLSRDSM dan $RLSRDL_1$ adalah lebih baik daripada OLS. RLSRDSM lebih baik sedikit daripada $RLSRDL_1$, dan kadang-kala prestasinya lebih kurang sama dengan kehadiran titik terpencil.

Walaubagaimanapun, $RLSRDL_1$ memiliki masalah pengiraan tertentu seperti menghasilkan jawapan yang buruk atau matriksnya `singular`. Kelebihan yang dimiliki oleh RLSRDSM ialah ianya tidak memiliki masalah pengiraan.

Prestasi WRLSRDSM adalah lebih baik daripada $WLSRDL_1$ apabila kedua-dua masalah titik terpencil dan ralat berheteroskedastik wujud bersama.

Tatacara Simulasi `Bootstrap` dan analisis visual juga dijalankan untuk menjustifikasikan behawa RLSRDSM adalah pengangger paling lasak berbanding dengan OLS dan $RLSRDL_1$. Keputusan yang didapati menunjukkan pencapaian mereka hampir sama dengan kesimpulan yang terdahulu.

ACKNOWLEDGEMENTS

First of all I would like to thank All Mighty Allah for everything. I am extremely grateful to my supervisor Assoc. Prof. Dr. Habshah Binti Midi for her invaluable guidance, enthusiastic encourage and support in every stage of my thesis research. This thesis represents a great deal of time and effort not only on my part, but also on the part of my supervisor. She introduced me to the field of Robust Regression and provided me many opportunities for growth: from reading papers, writing a survey, turning ideas to implementation, getting through the inevitable research setbacks, and finishing the thesis. She opens the door for me to enjoy a research work. What I learn from her will provide me with lifetime benefits. I consider myself lucky to access her supervision.

My thanks also goes to the members of my supervisory committee, Assoc. Prof. Dr. Adam Kilicman and Assoc. Prof. Dr. Kassim Bin Haron for their invaluable discussions, comments, and help.

Also I would like to extend my thanks to all members of Dept. of Mathematics, Faculty of Science, Universiti Putra Malaysia, for their kind assistance during my studies. This particularly goes to Assoc. Prof. Dr. Mohd. Rizam Abu Bakar, Head of Dept. of Mathematics, Dr. Mahendran Shitan, Assoc. Prof. Dr. Noor Aisha (Universiti Malaya), Assoc. Prof. Dr. Kassim Bin Haron for giving me permission to attend some of the courses given in the department for the MSc. students, to improve my knowledge, not forgetting forget Dr. Marwan Abdul Malik Thanoon for his kind help me in data collection stage.

I certify that an Examination Committee has met on 10th August 2006 to conduct the final examination of Bashar Abdul Aziz Majeed Al-Talib on his Doctor of Philosophy thesis entitled "Robust Regression with Continuous and Categorical Variables Having Heteroscedastic Non-Normal Errors" in accordance with Universiti Pertanian Malaysia (Higher Degree) Act 1980 and Universiti Pertanian Malaysia (Higher Degree) Regulations 1981. The Committee recommends that the candidate be awarded the relevant degree. Members of the Examination Committee are 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)

Noor Akma Ibrahim, PhD
Associate Professor
Faculty of Science
Universiti Putra Malaysia
(Internal Examiner)

Zainodin Hj. Jubok, PhD
Professor
Faculty of Science
Universiti Malaysia Sabah
(External Examiner)

_____
HASANAH MOHD. GHAZALI, PhD
Professor / Deputy Dean
School of Graduate Studies
Universiti Putra Malaysia

Date:

This thesis submitted to the Senate of Universiti Putra Malaysia and has been accepted as fulfilment of the requirement for the degree of Doctor of Philosophy. The members of the Supervisory Committee are as follows:

Habshah Midi, PhD
Associate Professor
Faculty of Science
Universiti Putra Malaysia
(Chairman)

Adam Kilicman, PhD
Associate Professor
Faculty of Science
Universiti Putra Malaysia
(Member)

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

_____
AINI IDERIS, PhD
Professor / Dean
School of Graduate Studies
Universiti Putra Malaysia

Date:

DECLARATION

I hereby declare that the thesis is based on my original work except for quotations and citations which have been duly acknowledged. I also declare that it has not been previously or concurrently submitted for any other degree at UPM or other institutions.

_____

BASHAR ABDUL AZIZ MAJEED AL-TALIB

Date:

TABLE OF CONTENTS