UNIVERSITI PUTRA MALAYSIA

ROBUST ESTIMATION METHODS AND ROBUST MULTICOLLINEARITY DIAGNOSTICS FOR MULTIPLE REGRESSION MODEL IN THE PRESENCE OF HIGH LEVERAGE COLLINEARITY-INFLUENTIAL OBSERVATIONS

AREZOO BAGHERI

IPM 2011 1

ROBUST ESTIMATION METHODS AND ROBUST MULTICOLLINEARITY DIAGNOSTICS FOR MULTIPLE REGRESSION MODEL IN THE PRESENCE OF HIGH LEVERAGE COLLINEARITY-INFLUENTIAL OBSERVATIONS

By

AREZOO BAGHERI

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

February 2011

DEDICATION

- *To my respectful father and lovely mother who thought me the meaning of courage and always had confident in me*

- *To my husband for all his contribution, patience and understanding throughout my doctorial studies. He incredibly supported me and made it all possible for me*

- *To my son, Kiarash, who was accompanying me in all different parts of my study and his love has always been my greatest inspiration*

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

ROBUST ESTIMATION METHODS AND ROBUST MULTICOLLINEARITY DIAGNOSTICS FOR MULTIPLE REGRESSION MODEL IN THE PRESENCE OF HIGH LEVERAGE COLLINEARITY-INFLUENTIAL OBSERVATIONS

By

AREZOO BAGHERI

February 2011

Chairman: Associate Professor Habshah Midi, PhD

Faculty: Institute for Mathematical Research

The presence of outliers and multicollinearity are inevitable in real data sets and they have an unduly effect on the parameter estimation of multiple linear regression models. It is now evident that outliers in the **X**-direction or high leverage points are another source of multicollinearity. These leverage points may induce or hide near-linear dependency of explanatory variables in a data set. We call these leverages, high leverage collinearity-influential observations either enhancing or reducing multicollinearity. By proposing High Leverage Collinearity-Influential Measure, denoted as HLCIM, we study several criteria such as sample size and magnitude, percentage, and position of high leverage

points which cause these leverages to change the multicollinearity pattern of collinear and non-collinear data sets.

The Ordinary Least Squares (OLS) estimates are heavily influenced by the presence of high leverage collinearity-influential observations. To rectify this problem, two new groups of robust regression methods are proposed. The Diagnostic Robust Generalized Potentials (DRGP) based on Minimum Volume Ellipsoid (MVE) is incorporated with different types of robust methods such as L_1, LTS, M, and MM in the establishment of the first proposed group of robust methods. The new proposed methods are called GM-DRGP-L_1, GM-DRGP-LTS (or Modified GM-estimator1(MGM1)), M-DRGP, MM-DRGP, and DRGP-MM. The second group of the proposed robust methods is formulated by modifying the existing Generalized M-estimator which is called as GM6. Two new GM-estimators which we call the Modified GM-estimator 2 and the Modified GM-estimator 3, denoted as MGM2 and MGM3, respectively are developed. Some indicators are employed to assess the performance of several existing robust methods and the new proposed methods. The results for real data set and Monte Carlo simulation study reveal that our proposed MGM3 outperforms the OLS and some of the existing robust methods.

The classical multicollinearity diagnostic methods may not be suitable to diagnose correctly the existence of multicollinearity in the presence high leverage collinearity-influential observations. To remedy this problem, two different approaches are proposed in the establishment of robust multicollinearity

diagnostic methods. In the first approach, we propose robust variance inflation factors, namely the RVIF(MM) and the RVIF(MGM3). The later is based on the proposed robust coefficient determination of MGM3. In the second approach, the diagnostic robust methods are proposed, specifically the Robust Condition Number (RCN), Robust Variance Inflation Factors (RVIF) and Robust Variance Decomposition Properties (RVDP) which are based on Minimum Covariance Determinant (MCD). The findings of this study suggest that the developed robust multicollnearity diagnostic methods are able to identify the source of multicollinearity in non-collinear data sets in the presence of high leverage collinearity-enhancing observations. On the other hand, for collinear data sets, in the presence of high leverage collinearity-reducing observations, the developed robust multicollinearity diagnostic methods are able to diagnose the multicollinearity pattern of the data set, correctly.

This thesis also addresses the problems of identifying multiple high leverage collinearity- influential observations in a data set. Since, the existing collinearity-influential measures fail to identify multiple collinearity-influential observations in a data set, a new High Leverage Collinearity-Influential Measure based on DRGP, denoted as HLCIM(DRGP) is proposed. The results of the study signify that this new diagnostic measure surpasses the existing measures. Furthermore, some non-parametric cutoff points for the proposed and some existing collinearity-influential measures are suggested in this thesis.

High leverage points may be considered as good or bad leverage point which depend on their residuals values. Unfortunately, researchers do not consider good leverage points to be problematic. However, these points may be collinearity-influential observations and need more attention. Regression diagnostic plots are one of the easiest and efficient tools for virtualizing the influential observations in a data set. Unfortunately, there is no existing plot in the literatures that identifies high leverage collinearity-influential observations. Finally, in this regard, we proposed three diagnostic plots specifically the SR(LMS)-DRGP, the DRGP-HLCIM, and the SR(LMS)-HLCIM. These new proposed diagnostic plots serve as powerful tools in separating outliers in the **y**-direction and the **X**-direction and able to identify any high leverage point which is collinearity-influential observation.

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

KAEDAH PENGANGGARAN TEGUH DAN MULTIKOLINEARAN TEGUH BERDIAGNOSTIK BAGI MODEL LINEAR BERGANDA DENGAN KEHADIRAN CERAPAN TUASAN TINGGI BERPENGARUH KOLINEARAN

Oleh

AREZOO BAGHERI

Februari 2011

Pengerusi: Profesor Madya Habshah Midi, PhD

Fakulti: Institute Penyelidikan Matematik

Kehadiran titik terpencil dan multikolinearan di dalam suatu set data tidak boleh dielakkan dan mempunyai kesan buruk ke atas penganggaran parameter bagi model linear regresi berganda. Bukti terkini menunjukkan bahawa titik terpencil pada arah **X** atau titik tuasan tinggi adalah satu lagi punca multikolinearan. Titik tuasan ini mungkin menampakkan atau menyembunyikan kebersandaran linear hampir bagi pembolehubah tak bersandaran dalam suatu set data. Kita namakan titik tuasan ini cerapan kolinearan berpengaruh sama ada menambah atau mengurang multikolinearan. Dengan mencadangkan Ukuran Tuasan Tinggi Kolinearan Berpengaruh, ditandakan sebagai HLCIM, kami mengkaji beberapa kriteria seperti saiz sampel dan magnitude, peratusan dan posisi titik tuasan tinggi

yang menyebabkan titik tuasan ini menukar corak multikolinearan bagi set data kolinear dan tak linear.

Pengangagr Kuasadua Terkecil Biasa (OLS) banyak dipengaruhi oleh kehadiran cerapan tuasan tinggi berpengaruh kolinearan. Untuk mengatasi masalah ini, dua kumpulan baharu kaedah regresi teguh, dicadangkan. Kaedah Teguh Berdiagnostik Potensi Teritlak (DRGP) berasaskan Isipadu Minimum Ellipsoid (MVE) digabungkan dengan beberapa kaedah teguh yang berbeza seperti L_1, LTS, M, dan MM bagi membangunkan kaedah teguh bagi kumpulan pertama yang dicadangkan. Kaedah baharu yang disarankan ini dinamakan GM-DRGP-L_1, GM-DRGP-LTS (atau Pengubahsuaian penganggar1 GM (MGM1)), M-DRGP, MM-DRGP, dan DRGP-MM. Kumpulan kedua kaedah teguh yang dicadangkan diformulasikan dengan mengubahsuai penganggar GM6 yang sedia ada. Dua penganggar baharu GM yang kami bangunkan, dinamakan pengubahsuaian penganggar 2-GM dan pengubahsuaian penganggar 3-GM, masing-masing ditandakan dengan MGM2 dan MGM3. Beberapa petunjuk diguna bagi menilai pencapaian beberapa kaedah teguh sedia ada dan kaedah baharu yang dicadangkan. Keputusan dari set data sebenar dan kajian simulasi Monte Carlo menunjkkan bahawa kaedah MGM3 yang dibangunkan lebih baik daripada kaedah OLS dan beberapa kaedah teguh yang sedia ada.

Kaedah multikolinearan berdiagnostik klasik mungkin tidak sesuai untuk mengenalpasti dengan betul, kehadiran multikolinearan dan cerapan titik tuasan tinggi berpengaruh kolinearan. Untuk mengatasi masalah ini, dua pendekatan

yang berbeza dicadangkan bagi membangunkan kaedah multikolinearan berdiagnostik. Bagi pendekatan pertama, kami mencadangkan inflasi varians teguh yang dinamakan RVIF(MM) dan RVIF(MGM3). Kaedah yang terkemudian ini berasaskan pekali penetapan teguh *MGM3*. Bagi pendekatan kedua, kaedah teguh berdiagnostik berasaskan Penentu Kovarians Minimum (MCD) dicadangkan yang dinamakan *Robust Condition Number (RCN), Robust Variance Inflation Factors (RVIF) dan Robust Variance Decomposition Properties (RVDP)*. Keputusan kajian menunjukkan bahawa kaedah multikolinearan teguh yang dibangunkan berupaya untuk menentukan punca multikolinearan bagi set data tak kolinear dengan kehadiran cerapan tuasan tinggi penambahan kolinearan. Manakala bagi set data kolinear, kaedah multikolinearan teguh berdiagnostik berjaya untuk menunjukkan corak multikolinearan pada set data dengan betul.

Tesis ini juga menyebut masalah yang dihadapi untuk mengenalpasti cerapan dengan titik tuasan tinggi berganda berpengaruh kolinearan, di dalam suatu set data. Olehkerana ukuran berpengaruh kolinearan yang sedia ada gagal untuk mengenalpasti cerapan berpengaruh kolinearan berganda, suatu ukuran baharu titik tuasan tinggi berpengaruh kolinearan berganda berasaskan DRGP dicadangkan, ditandakan dengan HLCIM(DRGP). Keputusan kajian menunjukkan bahawa ukuran baharu berdiagnostik ini lebih baik daripada ukuran yang sedia ada. Selain daripada itu, beberapa titik genting tak berparameter disarankan bagi kaedah yang sedia ada dan kaedah yang dicadangkan di dalam tesis ini.

Terdapat kemungkinan bahawa titik tuasan tinggi adalah titik tuasan baik atau buruk bergantung kepada nilai reja masing-masing. Malangnya, penyelidik tidak mengambil kira titik tuasan baik sebagai suatu masalah. Bagaimanapun, titik sebegini mungkin cerapan berpengaruh kolinearan dan memerlukan perhatian yang lebih. Plot berdiagnostik regresi adalah salah satu alat yang mudah dan efisien bagi menunjukkan cerapan berpengaruh pada suatu set data. Malangnya, tidak terdapat plot yang sedia ada pada literatur bagi mengenalpasti cerapan tuasan tinggi berpengaruh kolinearan. Dalam hal ini akhir sekali, kami mencadangkan tiga plot berdiagnostik yang dinamakan SR(LMS)-DRGP, DRGP-HLCIM, dan SR(LMS)-HLCIM. Plot baharu yang dicadangkan ini berperanan sebagai alat berkuasa untuk memisahkan titik terpencil dalam arah **y** dan arah **X** dan berupaya untuk mengenalpasti sebarang cerapan titik tuasan tinggi yang berpengaruh kolinearan.

ACKNOWLEDGEMENTS

Now in the name of God whose power controls
Wisdom, and has created human souls,
Exulted beyond all that thought or speech,
Is able to encompass or to reach,
The lord of Saturn and Stars at night,
Who gives the sun and moon and Venus light,
Above all name and thoughts, exceeding all
Of his creation, an unknowable,…

Shahnameh (Abolqasem Ferdowsi)

First of all, I wish to thank God who always supported me in all difficulties of my study life.

To have successful children has been one of my parent's dreams. I tried as much as I could afford to fulfil their dreams in order to thank them sincerely for scarifying their life to grow me up.

I am gratefully indebted to my committee chairman, Dr Habshah Midi. I deeply appreciate to have the opportunity to complete my degree under her supervision. She has provided the constant inspiration, efficient guidance, instruction and constructive criticisms and most importantly her friendly manner that I value greatly. I have benefited enormously from her continued support and confidence in my abilities.

My special heartfelt thanks go to my internal co-supervisors Dr. Isthrinayagy Krishnarajah and Dr. Basher Abdul Aziz Majeed Al-Talib, senior lecturers of my institute serving in my committee supervisory. A special word of thanks are

reserved to Professor Dr Rahmatullah Imon, associate professor of statistics, Department of Mathematical Sciences, Ball State University, U.S.A, my external committee member whose immeasurable assistance guided me a lot. Moreover, I am deeply thankful to him for dedicating his valuable time to give me new ideas through sending me electronic messages.

I could not possibly forget all the wonderful people that have offered me their friendship and have enriched my life during these doctorial studies duration. My acknowledgement would be incomplete without mentioning of my friends Sohel, Hassan, Munther, Dr. Saroje and all the others who made wonderful memories for me. Thanks you all.

And lastly but importantly, my special thanks to my husband whose patience is admirable for me. Without his undoubting faith, my thesis would never have been completed. My sincerely regards to my sisters, specially my elder one, my brother, and my niece, Negar, who encouraged me not to miss my hope in doing my research and supported me a lot mentally.

My doctorial studies wouldn't be possible without the scholarship granted to me by School of Graduate Studies of Universiti Putra Malaysia. Much gratitude is also due to all of the INSPEM members who created an environment in which PhD students can flourish. I was lucky to have the chance to be graduated from this institute.

I certify that a Thesis Examination Committee has met on 16/02/2011 to conduct the final examination of Arezoo Bagheri on her thesis entitled "Robust Estimation Methods and Robust Multicollinearity Diagnostics for Multiple Regression Model in the Presence of High Leverage Collinearity-Influential Observations" in accordance with 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 Doctor of Philosophy of Statistics.

Member of the Examination committee were as follows

Mohd Rizam Abu Bakar
Associate Professor
Faculty of Science
Universiti Putra Malaysia
(Chairman)

Mahendran a/I S. Shitan
Associate Professor
Faculty of Science
Universiti Putra Malaysia
(Internal Examiner)

Isa Daud
Associate Professor
Faculty of Science
Universiti Putra Malaysia
(Internal Examiner)

Olimjon Shukurovich Sharipov
Professor
Department of Probability Theory and Mathematical Statistics
Institute of Mathematics and Informational Technologies,
Uzbek Academy of Sciences, Uzbekistan
(External Examiner)

SHAMSUDDIN SULAIMAN, PhD
Professor and Deputy Dean
School of Graduate Studies
Universiti Putra Malaysia

Date:

This thesis was 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 Supervisory committee were as follows:

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

Isthrinayagy Krishnarajah, PhD
Senior Lecturer
Faculty of science
Universiti Putra Malaysia
(Member)

Bashar Abdul Aziz Majeed Al-Talib, PhD
Senior Lecturer
Faculty of science
Universiti Putra Malaysia
(Member)

A. H. M. Rahmatullah Imon
Associate Professor
Ball State University
Muncie, IN 47306, U.S.A.
(Member)

_____

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

Date:

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 it is not concurrently, submitted for any other degree at University Putra Malaysia or at any institutions.

_____

AREZOO BAGHERI

Date: 16 February 2011

TABLE OF CONTENETS