UNIVERSITI PUTRA MALAYSIA

ROBUST ESTIMATION METHODS AND OUTLIER DETECTION IN MEDIATION MODELS

ANWAR FITRIANTO

FS 2010 24

ROBUST ESTIMATION METHODS AND OUTLIER DETECTION IN MEDIATION MODELS

ANWAR FITRIANTO

DOCTOR OF PHILOSOPHY
UNIVERSITI PUTRA MALAYSIA
2010

ROBUST ESTIMATION METHODS AND OUTLIER DETECTION IN MEDIATION MODELS

By

ANWAR FITRIANTO

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

March 2010

DEDICATION

I dedicate this thesis and its related knowledge to God. I also humbly dedicate my thesis to my wife, Greiche Dian Kusumawardhani, and my sweet daughter, Khazbiika Shahrinaz Anwar. My parent (Maksum and Sunifah), my parent-in-law (Roelche Chairul Syahfri and Hermien Sulianthy), who have always believed in me, and my three elder sisters (Mutmainnah, Sulismiati, Sri Kusrini), my two elder brothers (Dr. Imam Hanafi and Nurahmad Fauzi), and my brother-in-law (Syahfreal Dion Kusumawardhana).

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

ROBUST ESTIMATION METHODS AND OUTLIER DETECTION IN MEDIATION MODELS

By

ANWAR FITRIANTO

March 2010

Chair : Associate Professor Habshah Midi, PhD

Faculty : Science

Mediation models refer to the relationships among three variables: an independent variables (IV), a potential mediating variable (*M*), and a dependent variable (DV). When the relationship between the dependent variable (DV) and an independent variables (IV) can be accounted for by an intermediate variable *M*, mediation is said to occur. Simple mediation model consists of three regression equations. The Ordinary Least Squares (OLS) method is often use to estimate the parameters of the mediation model. However, due to the fact that outliers have an unduly effect on the OLS estimates, we propose to incorporate robust M and MM estimator which are not easily affected by outliers, in the estimation of the mediation model which is called RobSim1 and RobSim2, respectively. The numerical example indicates that various types of contamination in the simulated data have arbitrarily large effect on the OLS estimates and the Sobel test. The MM-estimator incorporated in RobSim2 has improved the precision of the indirect effect of mediation model. The overall analysis clearly shows that the

Simple Mediation Model based on RobSim2 is prominently the most excellent result, because it is able to withstand various contamination in the m, x , and y-axes (direction).

There is also concern not only when the data contain observations that are extreme in the response variable but also in the regressor space, namely the leverage points. A new measure for the identification of high-leverage point is called Diagnostic Robust Generalized Potentials (DRGP) which is proposed previously. The DRGP procedures incorporated the Robust Mahanalobis Distance (RMD) based on the minimum volume ellipsoid (MVE) for identifying the set of cases 'remaining' (R) and a set of cases 'deleted'(D), and then diagnostic approach is used to confirm the suspected values. The DRGP procedure uses MAD as its cut-off points. We suggest an alternative method for identification of high leverage points in the mediation model. A modification is made to the DRGP procedure. It was verified that both MAD and Q_n have the same breakdown point that is 50%. Nonetheless, the efficiency of the Q_n is higher (86%) than the MAD (37%). This work inspired us to incorporate the Q_n instead of the MAD in the proposed algorithm. We refer the above new method of identifying potential outliers in mediation analysis as ModDRGP1 where the MAD is incorporated in the second step of the ModDRGP1 algorithm. In this thesis we also propose another DRGP, which has modified step 2 and step 4 for identifying potential outliers in mediation model. We called the second proposed method as ModDRGP2.

In order to strengthen the analysis, we provided a Monte Carlo simulation to evaluate the performance of our proposed ModDRGP1 and ModDRGP2. The simulation results suggested that by applying our newly proposed method has improved the accuracy of the identification of high leverage point when the percentage of high leverage points is medium or high. Even though the method was studied in simple mediation model, it can also be used for the identification of high leverage points in multiple mediation models, as well.

Based on the new proposed DRGP, we proposed a clean-assured data generating procedures (CADGP), a screening algorithm in mediation analysis. The importance of this procedure is that data screening methods provide the researcher with a means to detect potential data problems by identifying data entry errors, missing values, possible outliers, in generating data. Special attention to this research is about simulation study which is very important in model validation. Simulated data are presented because with these data, the underlying structure is known with certainty. Specifically in mediation analysis, the clean data generating process in a simulation is needed before continuing to further analysis so that accurate result of an analysis can be obtained. Our analysis shows that CADGP incorporating ModDRGP2 that we proposed provides a procedure for obtaining clean datasets, especially in mediation analysis and multiple linear regression models. The generated dataset through CADGP will be free of high leverage points. The CADGP is needed especially in simple mediation analysis which is usually used in social sciences. In social sciences, researchers are commonly needs larger sample size.

Finally, we proposed a new bootstrap procedure of indirect effect in mediation model which is resistant to outliers. Bootstrap has been the object of research in statistics. The proposed approach was based on residual bootstrap which incorporated rescaled studentized residuals, namely the Rescaled Studentized Residual Bootstrap using Least Squares (ReSRBLS). From the empirical data, we found that our Rescaled Studentized Residual Bootstrap using Least Squares (ReSRBLS), has produced similar sampling distribution (Normal) compared to some common methods; Raw Residual Bootstrap using Least Squares (RRBLS), Studentized Residual Bootstrap using Least Squares (SRBLS), Jackknifed Residual Bootstrap using Least Squares (JRBLS). The analysis signify that the ReSRBLS has outstanding performances compared to the other methods in the presence of outliers. The ReSRBLS not only has smaller bias and root of mean squares error (RMSE), but also narrower confidence intervals. The results from empirical data have been strengthened by the result from simulation study. For the contaminated data, the better performance of our proposed method with regard to bias and RMSE did not only happen to small or medium percentage of outliers, but also at large percentage of outliers. The advantages of the ReSRBLS over other methods are even more apparent in data sets with medium or large percentage of outliers.

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

KAEDAH-KAEDAH PENGANGGARAN TEGUH DAN PENGESANAN DATA TERPENCIL DALAM MODEL PENGANTARAAN

Oleh

ANWAR FITRIANTO
March 2010

Pengerusi : Professor Madya Habshah Midi, PhD
Faculti : Sains

Model pengantara merujuk pada hubungan di antara tiga pembolehubah: pembolehubah tak bersandar (IV), pembolehubah pengantara potensi, dan pembolehubah bersandar (DV). Apabila hubungan diantara pembolehubah bersandar (DV) dan pembolehubah tak bersandar (IV) boleh diterangkan oleh pembolehubah pengantara *M*, pengantaraan dikatakan berlaku. Model pengantara ringkas terdiri daripada tiga persamaan regresi. Kaedah kuasadua terkecil(OLS) sering digunakan untuk menganggar parameter bagi model pengantara. Bagaimanapun, ramai penyelidik tidak menyedari hakikat bahawa titik terpencil memberi kesan kepada penganggar OLS. Kami mencadang digabungkan penganggar teguh M dan MM yang tidak mudah dipengaruhi oleh titik terpencil, dalam penganggaran model pengantara yang dinamakan, masing-masing RobSim1 dan RobSim2. Contoh berangka menunjukkan bahawa pelbagai jenis pencemaran dalam data simulasi mempunyai pengaruh yang besar secara sewenang-wenangnya ke atas anggaran OLS dan ujian Sobel. Kaedah MM yang digabungkan ke dalam RobSim2 telah menunjukkan peningkatan kejituan kesan

tidak langsung bagi model pengantara. Keseluruhan análisis menunjukkan dengan jelas bahawa model pengantara ringkas berdasarkan RobSim2 memberikan keputusan yang paling baik, kerana mampu menangani pelbagai pencemaran dalam arah m , x , dan y .

Perhatian juga tidak hanya tertumpu apabila data mengandungi titik yang ekstrim pada pembolehubah bersandar tetapi juga di ruang pembolehubah tak bersandar, iaitu titik tuasan. Suatu ukuran baru untuk mengenalpasti titik tuasan yang disebut Diagnosis Potensi Teritlak Teguh (DRGP) telah dicadangkan sebelumnya. Prosedur-DRGP telah menggabungkan jarak teguh Mahalanobis (RMD) berdasarkan kaedah isipadu terkecil ellipsoid (MVE), bagi mengesan set kes 'sisa' (R) dan set kes 'dihapuskan' (D), dan kemudiannya pendekatan berdiagnostik digunakan untuk mengesahkan nilai yang dicurigai. Tatacara DRGP menggunakan MAD sebagai titik genting. Kami menyarankan suatu kaedah alternatif untuk mengenalpasti titik tuasan dalam model pengantara. Telah disahkan bahawa kedua- dua MAD dan Q_n mempunyai titik musnah yang sama iaitu 50%. Walau bagaimanapun, kecekapan Q_n adalah lebih tinggi (86%) daripada MAD (37%). Kerja ini telah mengilhamkan kami untuk menggunakan Q_n dan bukannya MAD didalam tatacara yang dicadangkan. Kita merujuk kaedah baru di atas bagi mengenalpasti data terpencil berpotensi dalam analisis pengantara sebagai ModDRGP1 yang mana MAD digabungkan dalam langkah kedua tatacara ModDRGP1. Dalam tesis ini kita juga mencadangkan DRGP lain, yang telah mengubahsuai langkah 2 dan langkah 4 bagi mengenalpasti titik

terpencil dalam model pengantara. Kami namakan kaedah kedua yang dicadangkan ini, sebagai ModDRGP2.

Untuk pengukuhan analisis, kami menyediakan simulasi Monte Carlo untuk menilai prestasi ModDRGP1 dan ModDRGP2 yang kami cadangkan. Keputusan simulasi menunjukkan bahawa dengan menerapkan kaedah baru yang kami cadangkan, kejituan meningkat bagi pengenalpastian titik tuasan tinggi, apabila peratusan titik tuasan adalah sederhana atau tinggi. Walaupun kaedah itu telah dikaji dalam analisis pengantara ringkas, ianya boleh juga digunakan untuk mengenalpasti titik tuasan dalam model pengantara berganda.

Berdasarkan kepada cadangan baru DRGP, kami mencadangkan prosedur menjana data pasti bersih (CADGP), satu algoritma penyaringan dalam analisis pengantaraan. Kepentingan prosedur ini adalah bahawa kaedah penyaringan data menyediakan penyelidik cara untuk mengesan data yang mempunyai potensi masalah dengan mengenalpasti kesilapan kemasukan data, nilai yang hilang, data terpencil yang mungkin, dalam menjana data. Perhatian khusus untuk bahagian ini adalah mengenai kajian simulasi yang sangat mustahak dalam pengesahan model. Data simulasi disajikan kerana dengan data ini, struktur yang mendasarinya diketahui dengan pasti. Terutamanya dalam analisis pengantaraan, menghasilkan data yang bersih dalam simulasi adalah diperlukan sebelum meneruskan kepada analisis lanjut supaya keputusan analisis yang tepat boleh didapati. Analisis kami menunjukkan bahawa CADGP yang menggabungkan ModDRGP2 yang kami cadangkan memberikan prosedur untuk mendapatkan set data yang bersih,

terutama sekali dalam analisis pengantaraan dan model regresi linear berganda. Set data yang dijana melalui CADGP akan bebas daripada titik tuasan yang tinggi. CADGP diperlukan terutama dalam analisis pengantaraan ringkas yang biasanya digunakan dalam ilmu sains sosial. Dalam sains sosial, penyelidik biasanya memerlukan saiz sampel yang lebih besar.

Akhirnya, kami mencadangkan suatu prosedur bootstrap untuk mengesan secara tidak langsung dalam model pengantara yang tahan terhadap data terpencil. Bootstrap telah menjadi objek kajian dalam statistik. Pendekatan yang kami cadangkan berdasarkan pada “bootstrap” sisa yang menggabungkan reja penskalaan-kembali reja “studentized”, iaitu Bootstrap Penskalaan-kembali Reja Studentized menggunakan Kuasadua Terkecil (ReSRBLS). Daripada data empirik, kami mendapati bahawa ReSRBLS, telah menghasilkan taburan contoh yang sama (Normal) berbanding dengan beberapa kaedah biasa; Bootstrap Reja Mentah menggunakan Kuasadua Terkecil (RRBLS), Bootstrap Reja Studentized menggunakan Kuasadua Terkecil (SRBLS), Bootstrap Reja Pisaulipat menggunakan Kuasadua Terkecil (JRBLS). Analisis juga menunjukkan bahawa ReSRBLS secara nyata mempunyai prestasi cemerlang berbanding dengan kaedah-kaedah lain didalam kewujudan nilai terpencil. ReSRBLS bukan sahaja mempunyai pincangan yang lebih kecil dan punca min kuasadua ralat (RMSE), tetapi juga selang keyakinan yang sempit. Keputusan daripada data empirik telah diperkukuhkan oleh hasil daripada kajian simulasi. Untuk data tercemar, prestasi yang lebih baik daripada kaedah yang kami cadangkan merujuk kepada RMSE dan pincangan bukan sahaja berlaku pada peratusan kecil atau sederhana data-data

terpencil, tetapi juga pada peratusan data terpencil yang tinggi. Kelebihan ReSRBLS keatas kaedah lain adalah lebih nyata dalam set data dengan peratusan data terpencil sederhana atau tinggi.

ACKNOWLEDGEMENTS

In the name of Allah, the most gracious and merciful.

I am deeply indebted to acknowledge and thank a number of people for their contribution in the writing and completion of this dissertation. Primarily, I thank almighty Allah for having granted me a livelihood through UPM Fellowship Program and the Graduate Research Assistantship thus far to complete my studies.

I would like to thank my thesis advisors Assoc. Prof. Dr. Habshah Midi, Assoc. Prof. Kassim Harun, Ph.D, and Assoc. Prof. Mahendran Shitan, Ph.D who underwent with me through many of the labor intense moments in the delivery of this work.

I am extremely grateful to my supervisor Assoc. Prof. Dr. Habshah Midi for her invaluable guidance, enthusiastic encourage and support in every stage of my thesis research. You patiently guided me towards an interesting and workable project. Your experience, your insight, and your openness to my ideas made the pages that follow possible. I will strive with my future students to be the kind of mentor that you have been to me.

I certify that a Thesis Examination Committee has met on 8 March 2010 to conduct the final examination of Anwar Fitrianto on his thesis entitled "Robust Estimation Methods and Outlier Detection in Mediation Models" 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 Doctor of Philosophy.

Members of the Thesis Examination Committee were as follows:

Adem Kilicman, PhD
Professor
Faculty of Science
Universiti Putra Malaysia
(Chairman)

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

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

Anton Abdulbasah Kamil, PhD
Associate Professor
School of Distance Education
Universiti Sains Malaysia
Malaysia
(External Examiner)

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

Date: 12 April 2010

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 were as follows:

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

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

Mahenran Shitan, 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 : 13 May 2010

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.

_____

ANWAR FITRIANTO

Date : April, 5 2010

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES