UNIVERSITI PUTRA MALAYSIA

ROBUST RANDOM REGRESSION IMPUTATION METHOD FOR MISSING DATA IN THE PRESENCE OF OUTLIERS

AHAMEFULE HAPPY JOHN

ROBUST RANDOM REGRESSION IMPUTATION METHOD FOR MISSING DATA IN THE PRESENCE OF OUTLIERS

By

AHAMEFULE HAPPY JOHN

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

December 2013

DEDICATION

- To God who made it all possible
- To my family who wanted to see me reach my goals
- To my beloved supervisor who inspired and guided me all the way
- To my friend who was there for me

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

ROBUST RANDOM REGRESSION IMPUTATION METHOD FOR MISSING DATA IN THE PRESENCE OF OUTLIERS

By

AHAMEFULE HAPPY JOHN

December 2013

Chairman: Md. Sohel Rana, PhD

Faculty: Science

The Ordinary Least Square (OLS) estimator is the best regression estimator if all the assumptions are met. However, the presence of missing data and outliers can distort the Ordinary Least Squares estimation and increase the variability of the parameters estimates. The main focus of this research is to take remedial measure in missing data in regression in the presence of outliers. In regression analysis, the dependent variable (Y) is a function of the independent variable X. Thus, in regression, outliers and missing values can come in both X and Y directions. It is very common to use the OLS base Random Regression Imputation (RRI) when missing values are in Y direction. This RRI seems to be a good method if there are no outliers in the data. Unfortunately, this estimate performs poorly in the presence of outliers. It is because the RRI is OLS base imputation

method and OLS is largely affected by outliers. As such, we modified an OLS base Random Regression Imputation (RRRI) methods by incorporating the robust MM estimate which is less affected by outliers. The proposed method is compared with some well-known methods of estimating missing data. The results of the study signify that the RRRI method outperforms the existing methods in the presence of outliers. Since in regression, outliers and missing data can come in both directions, we also considered a situation in which observations are missing in the X explanatory variable. In this respect, the Dummy Variable (DV) approach is one of the best approaches to predict the missing data model. However, this approach also becomes poor in the presence of outliers. As an alternative, Robust Inverse Regression Technique is proposed to get the better estimate. By examining the real data and Monte Carlo Simulation studies, it revealed that our proposed robust methods perform better than the classical methods.

Abstrak tesis dikemukakan kepada Senat Universiti Putra Malaysia sebagai memenuhi syarat bagi Sarjana Statistik Gunaan

KAEDAH IMPUTASI REGRESI RAWAK TEGUH BAGI DATA HILANG DENGAN KEHADIRAN TITIK TERPENCIL

Oleh

AHAMEFULE HAPPY JOHN

Disember 2013

Pengerusi: Md. Sohel Rana, PhD

Fakulti: Sains

Penganggar Kuasa Dua Terkecil Biasa (KDTB) adalah penganggar regresi yang paling baik sekiranya semua andaian dipenuhi. Walau bagaimanapun kehadiran data hilang dan titik terpencil boleh memesongkan anggaran Kuasa Dua Terkecil Biasa dan meningkatkan kebolehubahan penganggaran parameter. Fokus utama kajian ini ialah untuk mengambil langkah pembetulan terhadap data hilang dalam regresi dengan kehadiran titik terpencil. Dalam analisis regresi, pemboleh ubah bersandar (Y) adalah fungsi kepada pembolehubah tak bersandar X. Oleh itu, dalam data regresi, titik terpencil dan nilai hilang boleh wujud pada arah X dan Y. Penggunaan Imputasi Regresi Rawak (IRR) yang berasaskan Kaedah Kuasadua Terkecil Biasa (KKTB) sering digunakan apabila nilai hilang pada arah Y. Kaedah IRR merupakan kaedah yang baik jika titik terpencil tiada dalam data. Malangnya, penganggar ini lemah sekiranya wujud titik terpencil . Ini kerana IRR adalah kaedah imputas yang berasaskan KDTB dan KDTB sangat dipengaruhi oleh titik terpencil. Oleh itu kami telah mengubahsuai kaedah Imputasi Regresi Rawak (IRR) yang berasaskan KDTB dengan menggabungkan penganggar MM teguh yang kurang dipengaruhi oleh titik terpencil. Kaedah yang

dicadangkan ini dibandingkan dengan kaedah penganggar data hilang yang terkenal. Hasil kajian menunjukkan kaedah IRR mengatasi kaedah sedia ada bagi data yang mengandungi titik terpencil. Oleh keruna dalam regresi, titik terpencil dan data hilang boleh berada di kedua-dua arah, kami turut mengambil kira situasi dengan cerapan hilang dalam pemboleh ubah tak bersandar X. Dalam hal ini, pendekatan Pemboleh ubah Dami (PD) merupakan suatu pendekatan yang paling baik bagi meramal model data hilang. Walau bagaimanapun, pendekatan ini menjadi lemah dengan kehadiran titik terperncil. Sebagai alternatif, Teknik Regresi Teguh Songsang dicadangkan bagi mendapatkan anggaran yang lebih baik. Dengan mengkaji data sebenar dan kajian Simulasi Monte Carlo, kaedah teguh yang kami cadangkan didapati lebih baik berbanding kaedah-kaedah yang lama.

ACKNOWLEDGEMENTS

I wish to take this opportunity to express my sincere gratitude to all who have helped me in diverse ways to complete this thesis successfully.

First and foremost, I want to thank God for making it all possible. Special thanks to my supervisor, Dr. Md. Sohel Rana for his excellent guidance, patience and constructive ideas throughout the course of this thesis. I am grateful to have a patient and motivational supervisor to support me at all times during this study. Many thanks also go to my co-supervisor, Prof. Dr. Habshah Midi for her valuable guidance. Without her great support, it was impossible to complete this thesis.

I would also like to extend my deepest appreciation to Dr. John Nwabueze Igabari for his role as a mentor and role model in my studies and also all my other lecturers who taught and shared their valuable knowledge with me.

Last but not least, I would like convey my gratitude to my parents for their inspiration and to my beloved family, my wife and friends too for their support and spiritual encouragement throughout my studies in Universiti Putra Malaysia.

I certify that a thesis Examination Committee has met on 13/ 12/13 to conduct the final examination of Ahamefule Happy John on his thesis entitled "Robust Random Regression Imputation Method for Missing Data in the Presence of Outliers" 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 Master of Science of Applied Statistics.

Members of the Examination Committee are as follows:

Y. Bhg. Prof. Dr. Noor Akma bt Ibrahim
Mathematics Department
Science Faculty
Universiti Putra Malaysia
43400 UPM Serdang
(Chairman)

Prof. Madya Dr. Mohd Rizam b Abu Bakar
Mathematics Department
Science Faculty
Universiti Putra Malaysia
43400 UPM Serdang
(Internal Examiner)

Dr. Jayathi a/p Arasan
Mathematics Department
Science Faculty
Universiti Putra Malaysia
43400 UPM Serdang
(Internal Examiner)

Prof. Madya Dr. Yong Zulina Zubairi
Pusat Asasi Sains
University of Malaya
50603 Kuala Lumpur
(External Examiner)

SEOW HENG FONG, 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 Master of Science. The members of the Supervisory Committee were as follows:

Md. Sohel Rana, PhD

Senior Lecturer
Faculty of Science
Universiti Putra Malaysia
43400 UPM Serdang
(Chairman)

Habshah Midi, PhD

Professor
Faculty of Science
Universiti Putra Malaysia
43400 UPM Serdang
(Member)

BUJANG BIN KIM HUAT, PhD

Professor and Deputy 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 is not concurrently, submitted for any other degree at Universiti Putra Malaysia or any other institutions.

AHAMEFULE HAPPY JOHN

Date: 13/12/2013

TABLE OF CONTENTS

LIST OF TABLES

Table		Page

LIST OF FIGURES

Figure		Page

LIST OF APPENDICES

LIST OF ABBREVIATIONS

DV – Dummy Variable Approach
INV OLS – Inverse Ordinary Least Square
INV LTS – Inverse Least Trimmed Square
LTS Estimator – Least Trimmed Square Estimator
LMS Estimator – Least Median Square Estimator
MSE – Mean Square Error
OLS- Ordinary Least Squares
RSE – Residual Standard Error
SE – Standard Error

CHAPTER 1

INTRODUCTION

1.1 Research Background

Missing data occurs frequently in many fields; for example, surveys often contain missing observation caused by the person being interviewed not understanding the question correctly. The presence of outliers and missing data make a huge interpretative problem in data analysis. In the content of the regression model, missing variables occur in both dependent and independent variables. Let us suppose a model

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad i = 1, \dots, n \text{ with } \varepsilon_i \sim iidN(0, \sigma^2) \qquad (1.1)$$

If missing observation is present in the independent variables (X) or dependent variable (Y), we will get the biased estimate of the parameters β_0 *and* β_1 and also the biased standard error of the parameters. However, the problem becomes more serious if there exists some unusual observations that we often call outlier. It makes a huge interpretative problem in the model as the parameters estimate of the model is affected by outliers.

For dealing with the outliers in the context of regression model, we have some robust estimators such as M-Estimator, MM-Estimator, L_1 - norm Estimator, LTS-Estimator, LMS-Estimator, etc. Unfortunately, to the best of our knowledge, no efficient method is available when both outliers and missing data come together in a data set. In regression analysis, the missing value is estimated based on random regression or regression imputation. However, these imputation methods are not reliable in the presence of outliers since these imputation methods are based on Ordinary Least Squares (OLS).

In this research, the major question that comes to our mind is, 'Are the OLS based imputation estimator robust in the presence of outliers'? The answer is probably, 'No'. The reason is that the classical OLS is greatly affected by outliers (Marrona, 2006; Rousseuw, 2003). We proposed to estimate the missing data by using robust regression approaches. The performance of our proposed methods would be evaluated by Monte Carlo simulation approach and real data analysis. We would be inclined to propose our robust methods based on the existing robust estimator such as MM-Estimator. Our proposed methods would estimate the missing data in the presence of outliers.

1.2 Importance and Motivation of the Study

Missing data is a common problem in many researches. Information is usually missing in some variables for some cases in a typical data set and most data analysis procedures are designed for corrected samples. However, standard methods are not directly applicable if there are missing data.

It is important to know why missing data is a problem. The most serious concern is that missing data can bring bias into estimates derived from a statistical model (Schafer,1997; Alison, 2002). In addition, missing data can lead to a loss of information and statistical power (Little and Rubin, 2002). Also, another problem is that missing data makes common statistical methods difficult to apply. Finally, missing data causes a waste in valuable resources. Due to all these reasons, missing data gives problems in various fields of research.

Many times, difficulties arise when practitioners try to apply Ordinary Least Squares regression estimation to real world data which have some missing observation and outliers. Model adequacy diagnostics show poor Ordinary Least Squares fit because of missing observation and outliers. The need to determine the performance of OLS is demanding due to the quick development of alternative robust estimators. Opportunities are also available for the development of improved robust estimators. Even then, progress continues to be made in making algorithms for robust methods ready for implementation when Ordinary Least Squares fail. It is ideal that improved methods be computationally practical and available in many software such as R, S-Plus, etc.

In regression analysis, when missing is in the response variable Y, the Random Regression Imputation (RRI) is usually used (Little and Rubin, 1987). Unfortunately, this RRI is affected by outliers. This problem has inspired us to develop a new robust method which is resistant to outliers. We incorporate the robust estimators in the missing value imputation methods. We proposed to use the Robust Random Regression Imputation (RRRI) instead of RRI.

In this thesis, we also consider the situation when data missing is on X direction and there exists some outliers in the data. It is very common to use the dummy variable regression approach when data is missing in X direction. However, this dummy variable regression approach also gives biased estimate of parameters and biased standard error of parameters in the presence of outliers. Thus, we proposed a new alternative robust method. Instead of dummy variable approach, in this situation, we proposed robust inverse regression to get more reliable estimates of parameters.

1.3 Research Objectives

In regression analysis, missing data comes in both X and Y directions together with outliers. It is worth mentioning that these types of problems are very common in the regression data. Hence, in this thesis, we have two main aims which are as follows.

1. To propose a robust imputation approach when missing values and outliers present in the Y direction of the regression data. Since in the presence of

outliers, the classical random Regression Imputation (RRI) is affected by outliers, we attempt to robustify RRI by using robust MM-estimator in order to get the robust estimate of missing data in the presence of outliers.

2. To propose a robust imputation approach when missing values are present in X direction and outliers are present in Y direction. It is seen that the classical Dummy variable Regression approach is affected by outliers. Thus, based on the robust inverse regression methods, we aim to estimate the missing values in X direction in the presence of outliers.

3. To get the benefits from our proposed robust imputation methods compared to the classical methods, the real data example will be used. We also conduct simulation studies and applied both classical and robust methods to compare their performance.

1.4 Plan of the Study

This research is organized into five chapters.

Chapter One: This chapter serves as an introduction to this research. It gives the background of the research, the problem statement, the research objectives and plan of the study.

Chapter Two: This chapter provides a brief review of the literature of missing data, the consequences of missing data, and the techniques for its estimation. We have also discussed some robust estimators which can help to estimate the missing data in the presence of outliers.

Chapter Three: In this chapter, a modification of the Random Regression Imputation (RRI) approach is proposed which we call Robust Random Regression Imputation (RRRI). A set of numerical results proved that the Robust Random Regression Imputation approach performs well compared to the Random Regression Imputation (RRI) in the presence of outliers when missing data is present in the dependent variable.

Chapter Four: In this chapter, an alternative method of dummy variable regression approach to estimate the missing data in the independent variables is proposed. As an alternative, the inverse regression is used. Also, we proposed the robust version of inverse regression which performs well in the presence of outliers.

Chapter Five: This last chapter provides the conclusion and significance of the research findings. I thereby also recommend a number of areas for further research which could have essential roles in future studies.

REFERENCES

Allison, P.D. (2002). Missing Data, Sage *Series Quantitative Applications in the social Science*. Unpublished PhD thesis, University of Pennsylvania, Philadelphia.

Barnett, V. and Lewis, T. (1994). *Outliers in Statistical Data,* 3rd edition. New York: Willy.

Buck ,S.F (1960).' A method of estimation of missing values of multivariate data suitable for use with an electronic computer. *Journal of the Royal Statistical Society Series*. 22: 302-306.

Chatterjee, S. and Hadi, A.S. (2006). *Regression Analysis* by Examples, 4th edition. New York: Wiley.

Cohen, J. and Cohen, P. (1985). *Applied Multiple Regression and Correlation Analysis for the Behavioral Sciences,* 2nd edition. Mahwah. N.J : Lawrence Erlbaum Associates.

Daniel, C., and Wood, F. F. (1971). *Fitting Equations to Data*, John & Sons. New York: Willy.

Dempster, A.P, Laird, N.M, Rubin, D.B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series.* B (Methodological). 39: 1-38.

Ellenberg, J.H. (1976). Testing for Single outlier from a general regression. *Biometrics*. 32: 637-645.

Gujarati, D. (2003). Basic Econometrics, 4th edition. New York: McGraw-Hill.

Habshah, M. and Lau, U. H. (2009). The Performance of Latent Root-M based Regression. *Journal of Mathematics and Statistics*. 5: 1-9.

Hampel, F.R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association*. 69 : 383-393.

Hampel, F.R. (1979). *Discussion of the meeting on robustness*. Proceedings of the 42nd Session of the ISI. Manila. 100-102.

Hawkins, D.M. Bradu, D. and Kass, G.V. (1984). Location of Several Outliers in multiple regression data using elemental sets. *Technometrics*. 26: 197-208.

Heckman, J. J. (1976). The common structure of statistical models of the truncated, sample selection and limited dependent variables, and a simple estimator of such models. *Annals of Economic and Social Measurement.* 5, 475-492.

Huber, P.J. (1981). *Robust Statistics*. New York: Wiley.

Imon, A.H.M.R.(2005). Identifying multiple influential observations in linear regression. *Journal of Applied Statistics*. 32: 929-946

Imon, AHMR. (2009). Deletion residuals in the detection of heterogeneity of variances in linear regression. *Journal of Applied Statistics*. 36: 347-358.

Jones, M.P. (1996). Indicator and Stratification methods for missing explanatory variables in Multiple linear regression. *Journal of the American Statistical Association.* 91: 222-230

Kutner, M. H., Nachtsheim, C. J. and Neter, J. (2008). *Applied Linear Regression Models*, 4th edition. New York: McGraw-Hill.

Little, R.J.A. and Rubin, D.B ,(1987). *Statistical Analysis with Missing Data*. New York: Wiley.

Little, R.J.A. and Rubin, D.B. (2002). *Statistical Analysis with missing Data*, 2nd edition, New York: Wiley.

Maronna ,R.A, Martin, R.D and Yohai V.J, (2006). *Robust Statistics-Theory and Methods.* New York: Wiley.

McKnight, Patrick E., McKnight. Katherine M; Sidani , Souraya, and Aurelio Jose Figueredo. (2007) Missing Data: *A Gentle Introduction*. Guilford Press. New York: Willy.

Midi, H,. Rana,.S. and A.H.M.R. Imon. (2009). The Performance of robust weighted least squares in the presence of outliers and heteroscedastic errors. WSEAS, *Transactions on Mathematics*. 8: 351-361.

Montgomery, D. C., Peck, E. A. and Vining, G. G. (2001). *Introduction to Linear Regression Analysis*. New York: Willy.

Rana S..Midi, H and A.H.M.R Imon.(2012). Robust with wild bootstrap for stabilizing the variance of parameter estimates in Heteroscedastic regression models in the presence of outliers. *Mathematical Problems in Engineering*, vol. 2012, Article ID 730328, 14 Pages.

Rousseeuw P.J. (1984). Least median of squares regression. *Journal of American Statistics Association.* 79: 871–880.

Rousseeuw, P. J. and Yohai, V. J. (1984). Robust regression by means of S estimators. *Lecture Notes in Statistics.* 26: 256-272.

Rousseuw, P. J. and Leroy, A. M. (1987). *Robust Regression and Outlier Detection*. New York: Willy.

Rousseuw, P. J. and Leroy, A. M. (2003). *Robust Regression and Outlier Detection*. John & Sons. New York: Willy.

Rubin, D.B..(1976). Inference and Missing Data. *Biometrika*. 63: 581-592.

Ryan, T.P. (1997). *Modern Regression Methods*, New York: Wiley.

Schafer, J. (1997). *Analysis of Incomplete Multivariate Data*. New York: Chapman & Hall.

Schafer, J.L. (2002). Missing Data; Our view of the state of the Art. *Psychological Method*. 2: 147-177.

Simpson, J.R. (1995). *New Methods and Comparative Evaluations for Robust and Biased-Robust Regression Estimate*. Unpublished Phd Thesis. Arizona State University.

Snedecor, G.W and Cochran ,W.G (1967). *Statistical Methods*, 6th edition, Ames: Iowa State University Press.

Srikantan K.S. (1961). Testing for the single outlier in a regression model. *Shankhya*. series A. 23: 251-260

Stromberg, A.J. Hossjer, O. and Hawkins, D.M. (2000). The least trimmed and difference regression estimator and alternatives. *Journal of American Statistical Association*. 95: 853-864.

Tukey, J.W. (1977). *Exploratory Data Analysis*. Addison –Wesley Publisher, Cy.

Yohai, V. J. (1987). High breakdown point and high efficiency robust estimates for regression. *The Annals of Statistics*, 15: 642-656.