VARIABLE SELECTION USING LEAST ABSOLUTE SHRINKAGE AND SELECTION OPERATOR

RAHAINI BINTI MOHD SAID

A thesis submitted in partial fulfillment of
the requirements for the award of the degree of
Master of Science (Mathematics)

Faculty of Science
Universiti Teknologi Malaysia

MAY 2011

Special dedicated to

My beloved family and friends

ACKNOWLEDGEMENT

First of all, I would like to express my sincere appreciation to my supervisor, Assoc. Prof. Dr Ismail Mohamad for his encouragement and guidance as well as suggestions.

Furthermore, I am grateful to Universiti Teknologi Malaysia for funding my master study, the librarians at Sultanah Zanariah Library (PSZ), UTM plays an important role in supplying the relevant literatures and resources used in this research.

I would also like to say thank all my friends and colleagues that helps me to complete my thesis.

ABSTRACT

Least Absolute Shrinkage and Selection Operator (LASSO) and Forward Selection are variable selection method that implement in this study. The objectives of this study are to apply forward selection method in variable selection for a regression model, to apply LASSO method in variable selection for a regression model using quadratic programming and leave one out cross validation and choosing the better model obtained from forward selection and LASSO method using least mean square error. The forward selection method implemented in the statistical package for social sciences (SPSS). Quadratic programming technique and leave one out cross validation from MATLAB software is applied to solve LASSO. The analyzed result showed forward selection and LASSO are chosen the same variable that should be included in the model. However the coefficient of the regression for both model differ. To choose between the two models, MSE is used as the criteria where the model with the smallest MSE is taken as the best model. The MSE for forward selection and LASSO are 0.4959 and 0.4765 respectively. Thus, LASSO is the better model compared to forward selection model.

ABSTRAK

"Least absolute shrinkage and selection operator" (LASSO) dan "forward selection" merupakan pilihan kaedah pembolehubah dalam kajian ini. Tujuan kajian ini adalah untuk melaksanakan kaedah "forward selection" dalam pemilihan pembolehubah untuk model regresi, untuk melaksanakan kaedah LASSO dalam pemilihan pembolehubah untuk model regresi menggunakan pengaturcaraan kuadratik dan "leave one-ou cross validation", memilih model yang lebih baik daripada "forward selection" dan LASSO dengan menggunakan memilih purata ralat kuasa dua. Kaedah regresi "forward selection" dilaksanakan dalam pakej statistik untuk ilmu sosial (SPSS) dan teknik pengaturcaraan kuadrat dan "leave one out cross validation" daripada perisian MATLAB diterapkan untuk menyelesaikan LASSO. Keputusan analisis menunjukkan "forward selection" dan LASSO adalah memilih pembolehubah yang sama yang harus dimasukkan ke dalam model. Namun pekali regresi untuk kedua model adalah berbeza. Untuk memilih antara dua model, purata ralat kuasa dua digunakan sebagai kriteria yang mana model dengan purata ralat kuasa dua terkecil diambil sebagai model terbaik. Purata ralat kuasa dua untuk "forward selection"dan LASSO adalah 0.4959 dan 0.4765 masing-masing. Dengan demikian, LASSO adalah model lebih baik berbanding dengan "forward selection".

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDIX

CHAPTER 1

INTRODUCTION

1.0 Background of the problem

Regression analysis is a statistical technique for modeling and investigating the relationship between independent (predictor or explanatory) variables and a dependent (response or outcome) variable. In the theory, we would like the model to include as many independent variables as possible so that the information content in these factors can influence or can explain the predicted value of *Y*. But we want the simplest model as good as a full model because we want to reduce the cost of data collection and model maintenance.

Multiple linear regression is one of the regression model, where involved equating response variable Y and many independent variables (X_1, X_2, X_3 . . . X_n), in form $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ ... + \ \beta_p X_p$. However not all of the p variable are useful in explaining the model. Thus a lesser number of the independent variables such as k are chosen to explain Y.

An important problem in statistical analysis is chosen of an optimal model or the best model from a set of a full model. In additional, selection of variables in regression problems has occupied the minds of many statisticians, because variable selection process is very difficult process. Then, these processes require consideration of the factors that maybe can be influence and the precision when choosing the variable.

Variable selection is the general method in selecting the k independent variable p to form a simpler model to explain Y such as $Y = \ \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ ... + \ \beta_k X_k$. There are several common variable selection methods in statistic, such as forward selection (FS), backward elimination (BE), stepwise (SW), R-squared (R-sq), and all-possible method.

In statistics, forward selection is the process that choose the variable, which is involves starting with no variables in the model, trying out the variables one by one and including them if

they are 'statistically significant'. Usually, the variable or predictor are select consider the sequence of *F*-tests, but other techniques are possible, such as *t*-tests, adjusted R-square, akaike information criterion (AIC), bayesian information criterion (BIC), or mallows' Cp. Although, the statistical package for social sciences (SPSS) is a type of sophisticated software that can produced or analysis easier and quickly. The results will produce similar with the manual calculation.

In the practice of statistical modeling, it is often desirable to have an accurate predictive model with a sparse representation. But, nowadays, modern data sets usually have a large number of predictors; hence the simplest model is an important issue. Then, shrinkage methods are the other method of selection variable, where have been widely studied in many branches of statistics such as least absolute shrinkage and selection variable (LASSO). Thus the LASSO possesses the nice properties for best-subset selection. It is forcefully argued that the automatic feature selection property makes the LASSO a better in high dimensional problems, especially when there are lots of redundant noise features.

LASSO is one of variable selection method proposed by Robert Tibshirani, 1996. The LASSO minimizes the residual sum of square and puts a constraint on the sum of absolute values of the parameters. LASSO also promising automatic model building technique, simultaneously producing accurate and parsimonious models. The LASSO is a worthy competitor to subset selection and ridge regression. The LASSO idea is quite general and can be applied in a variety of statistical model.

1.1 Problem statement

Nowadays, as data collection technology and data storage devices become more powerful, scientists have become able to collect many variables and large numbers of observations in their studies. In the absence of prior knowledge, data analysts may include very many variables in their models at the initial stage of modeling in order to reduce possible model bias (approximation error). In general, a complex model that includes many insignificant variables may result in less predictive power, and it may often be difficult to interpret the results. In these cases, a simpler model becomes desirable in practice.

Therefore, variable selection is fundamental to statistical modeling. Variable selections have become the focus of much research in areas of application for which datasets with tens or hundreds of thousands of variables are available. The objectives of variable selection are improving the prediction performance of the predictors, more cost-effective predictors, and providing a better understanding of the underlying process that generated the data. The purpose of variable selection in regression is identifying the best subset among many variables to be included in a model. The idea is to choose a simplest model which is as good as a full model in capability to response value by the dependent variables. This simpler model is known as parsimonious model.

1.2 Objectives of the study

The objectives of this research are:

1. To apply forward selection method in variable selection for a regression model.
2. To apply LASSO method in variable selection for a regression model using quadratic programming and leave one out cross validation.
3. Choosing the better model obtained from forward selection and LASSO method using least mean square error.

1.3 Scope of the study

This study will look into the variable selection method namely forward selection and least absolute shrinkage and selection operator (LASSO). These two methods are applied to heart diseases data comprising of single response and eight independent variables. This study will choose the best parsimonious model obtained from the two methods.

1.4 Significance of the study

This study focuses on two variable selection methods namely the common forward selection method and LASSO method. The study will enrich the finding on these two methods in selecting variables.

1.5 Research organization

This research is divided into five chapters. The first chapter presents the introduction of this research which comprises of the background of the study, statement of the study, objectives of the study and the significance of the study.

Chapter two is the literature review on the studies done on forward selection and least absolute shrinkage and selection operator (LASSO).

Chapter three will be focusing on description, formulation, theoretical and principle of on LASSO. It also shows the step to conduct a forward Selection analysis by using statistical package for social Sciences (SPSS). The quadratic programming and cross validation are a technique from MATLAB software we will use to solve the LASSO problem.

Chapter four consists of case study on the use of two methods, forward selection and LASSO on heart disease data. Discussion and comparison are presented in this chapter.

Finally, chapter five concludes the study. This chapter summarizes the study and conclusion based on analysis and the results of the study. Suggestions for further research are also recommended in this chapter.

REFERENCES

Allen, D. M. (1974), The relationship between variable selection and data augmentation and a method for prediction. Technometrics 16, page: 125–127.

Bruce Ratner,(2003), Statistical modeling and analysis for database marketing: Effective techniques for mining big data, page: 45-61

C. Leng, Y. Lin, and G. Wahba, (2004), A note on the lasso and related procedures in model selection, Journal of Statistica Sinica, page :134-145

Craven, P. and Wahba, G, (1979), Smoothing noisy data with spline functions. Numer. Math. 31, page :377–403.

Dash, M., and Liu, H., 1997, Feature Selection for Classification , Intelligent Data Analysis ,Elsevier Science Inc, page: 455-466

Douglass.C.M, (2001), Introduction to linear regression analysis, Third edition, page:67-76

Efroymson, M. A, (1960), Multiple Regression Analysis, in Mcathlenlcati-cal Methodsf or Digital Conmputters, eds. A. Ralston and H. S. Wilf, New York: John Wiley, page: 191-203.

Elizabeth A.peck, (2001), Introduction to linear regression analysis, third edition, page:291-318

F Guillaume Blanchet, Pierre Legendre,Daniel Borcard, (2008), Forward selection of explanatory variables, Issue: 9, Publisher: Eco Soc America, Page: 2623-2632

Frank, I.E., and Friedman, J.H, (1993), A statistical view of some chemometrics regression tools, technometrics, 35, page 109-148.

Fu, W, (1998), Penalized regressions: the bridge vs. the lasso, Journal of Computational and Graphical Statistics 7(3), page: 397–416.

Geisser, S, (1975), The predictive sample reuse method with applications. J. Amer. Statist. Assoc. 70, page: 320–328

Hastie, T., Botha, J. and Schnitzler, C, (1989), Regression with an ordered categorical response, Statistics in Medicine 43, page: 884–889.

Hastie, T., Tibshirani, R. and Friedman, J, (2003), A note on comparison of model selection for regression, Neural computation 15(7), page :1477–1480.

Hastie.t and Tibshirani.R, (2009), Linear method for regression, the elements of statistical learning.Second edition, page: 43- 56.

Jason D. M. Rennie, (2003), The value of leave-one-out cross-validation bounds, page: 1-6

Lawson, C., and Hansen, R, (1974), Solving least squares problems, Englewood Cliffs, NJ: Prentice-Hall, page : 345-356

Kaspar Fischer, Bernd Gärtner, Sven Schönherr, and Frans Wessendorp, (2007). **Linear and quadratic programming solver**, page: 61-64

M. Frank and P. Wolfe, (1956) An algorithm for quadratic programming, naval research logistics quarterly, **3**, page : 95–110,.

Osborne, M., Presnell, B. and Turlach, B, (2000). A new approach to variable selection in least squares problems, IMA Journal of Numerical Analysis 20, page :389–404.

Stone, M, (1974), Cross-validatory choice and assessment of statistical predictions (with discussion). J. Roy. Statist. Soc. Ser. B 36, page: 111–147.

Seber, G.A.F, (1977), Linear regression analysis, New York: Wiley, page: 4-7

Scott W. Menard, (2002), Applied logistic regression analysis, quantitative applications in the social sciences, page : 4-14

Tibshirani, R, (1996), Regression shrinkage and selection via the lasso, Journal of royal

statistical society, page 267-288.

Yuan, M. and Lin, Y, (2007), Model selection and estimation in regression with grouped variables, Journal of the Royal Statistical Society, Series B 68(1), page: 49–67.

Weisberg, S, (1980), Applied Linear Regression, Wiley, New York, page : 234-245

Zhang, P, (1993), Model selection via multifold cross-validation, Annals of Statistics 21, page:299–311.