# VARIABLE SELECTION USING
# LEAST ANGLE REGRESSION

WAN NUR SHAZIAYANI BT WAN MOHD ROSLY

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 you to all my family members, friends and colleagues that helped me to complete my thesis.

# ABSTRACT

The least-angle regression (LARS) (Efrron, Hastie, Johnstone, and Tibshirani, 2004) is a technique used with the absence of data that consist of many independent variables. Suppose we expect a response variable to be determined by a linear combination of a subset of potential covariates. Then the LARS algorithm provides a means of producing an estimate of which variables to include, as well as their coefficients. The MATLAB programming codes are developed in order to solve the algorithms systematically and effortlessly.

# ABSTRAK

"Least Angle Regression (LARS)" (Efrron, Hastie, Johnstone, and Tibshirani, 2004) adalah satu teknik yang digunakan dengan kehadiran data yang mempunyai banyak pemboleh ubah tidak bersandar. Seperti yang dijangkakan pemboleh ubah bersandar boleh di tentukan dengan gabungan pemboleh ubah tidak bersandar. Oleh itu LARS algoritma bermaksud memberikan anggaran pemboleh ubah yang manakah dapat disertakan begitu juga dengan "coefficient". Kod pengaturcaraan MATLAB dihasilkan bagi menyelesaikan masalah dengan lebih mudah dan sistematik.

## TABLE OF CONTENTS

# LIST OF TABLES

## LIST OF FIGURES

## LIST OF APPENDIX

# CHAPTER 1

# INTRODUCTION

## 1.1 Background of The Problem

In statistics, regression analysis includes any techniques for modeling and analyzing several variables, when the focus is on the relationship between a dependent variable and one or more independent variables. More specifically, regression analysis helps us understand how much dependent variable changes with changes in each of the independent variable, while the other independent variables are held fixed.

Regression analysis is widely used for prediction and forecasting. Regression analysis is also used to understand which among the independent variables are related to the dependent variable, and to explore the forms of these relationships.

There are simple regression and multiple regression. Simple regression is a model with only one independent variable ($X$) while multiple regression is a model with more than one independent variables which are $Y = \beta_0 + \beta_1 X_1 + \beta_2 x_2 + \beta_3 X_3 + \ldots + \beta_p X_p$. In theory, the more of independent variables use the more accurate explanation on the dependent variable. But in most practical situations, however, only a relatively small number of independent variable is to be considered because it will help to reduce the cost and time.

For the variables selection, common methods are being used are stepwise, forward and backward selection method. Stepwise selection has been proposed as a technique that combines advantages of forward and backward selection. At any point in the search, a single predictor variable may be added or deleted based solely on the *t*-statistics of their estimated coefficients. Commonly, the starting subset is the empty set. Some of the problems with stepwise variable selection are it yields R-squared values that are badly biased to be high, the method yields confidence intervals for effects and predicted values that are falsely narrow (See Altman and Anderson, 1989, Statistics in Medicine) and  it has severe problems in the presence of collinearity.

Forward selection, which involves starting with no variables in the model, trying out the variables one by one and including them if they are statistically significant while Backward elimination, which involves starting with all candidate variables and testing them one by one for statistical significance, deleting any that are not significant on the basis of an F-distribution, calculate the p-value associated with restoring the term into the model.

The purpose of model selection algorithms such as all subsets, Forward Selection and Backward Elimination is to choose a linear model on the basis of the same set data to which the model will be applied. Typically we have available a large collection of possible covariates from which we hope to select a parsimonious set for the efficient prediction of a response variable. Least Angle Regression (LARS), a new model selection algorithm, is a useful and less greedy version of traditional forward selection methods.

## 1.2 Problem Statement

The main problem in the multiple regression model is to select the independent variable. The idea is to choose a simpler model where the $X$ is selected from the p variables.

## 1.3 Objectives of the Study

The main objectives of this research are:

1. To apply forward selection and LARS method in variable selection for a regression model
2. To identify the similarities and the differences between forward selection and LARS method

## 1.4 Scope of the Study

The study and proposed efficient algorithms for the extensions of common methods for factor selection and show that these extensions (LARS) give superior performance to the traditional forward selection method in factor selection problems. We study the similarities and the differences between these methods. The body fat data are used to illustrate the methods. This data were used to produce predictive equations for lean body weight, a measure of health.

## 1.5 Significance of the Study

Variable selection is very important to ensure that the result from data analysis will be more accurate. Therefore this research will focus on the least angle regression (LARS) to select the variables.

# REFERENCES

1. George, E. I. and McCulloch, R. E. (1993), Variable selection via Gibbs sampling, *J. Amer. Statist. Assoc.,* Volume 88, 881-889.
2. Ming Yuan, V. Roshan Joseph, and Yi Lin (November 23, 2005), *An Efficient Variable Selection Approach for Analyzing Designed Experiments.* Assoc. Prof Thesis. Department of Statistics, University of Wisconsin.
3. Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004), Least Angle Regression, *Ann. Statist.*, Volume 32, 407-499.
4. Meier L., Sara Van De Geer, and Buhlmann P. (2008), *Statistical Methodology, Journal of royal Statistical Society,* Volume 24, 130-137.
5. DerwinA.Turlach (2004), *Discussion of Least Angle Regression, Institute of Mathematical Statistics*, University of Western Australia, Vol. 32, No 2. 481-490.
6. Breiman, L., Friedman, J., Olshen, R. and Stone, C. (1984). *Classification and Regression Trees*. Wadsworth, Belmont, CA.
7. Friedman, J., Hastie, T. and Tibshirani, R. (2000). *Additive logistic regression: A statistical view of boosting (with discussion).* Ann. Statist. issues 28, pages 337-407.
8. Chris Fraley and Tim Hesterberg (2009) *Least Angle Regression and LASSO for Large Datasets* volume 1, issues 4, pages 251-259.