ASSESSING THE PERFORMANCE OF PENALIZED REGRESSION METHODS AND THE CLASSICAL LEAST SQUARES METHOD

BY

Pascalis Kadaro, MATTHEW, B.TECH (FUTY) 2010
MSc/SCI/40557/2012-2013

A THESIS SUBMITTED TO THE SCHOOL OF POSTGRADUATE STUDIES, AHMADU BELLO UNIVERSITY, ZARIA

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE AWARD OF A MASTER DEGREE IN STATISTICS.

DEPARTMENT OF MATHEMATICS,
FACULTY OF SCIENCE
AHMADU BELLO UNIVERSITY, ZARIA
NIGERIA

JULY, 2015

DECLARATION

I hereby declare that this thesis entitled **"ON THE PERFORMANCE OF PENALIZED REGRESSION METHODS AND THE CLASSICAL LEAST SQUARE METHODS"** is a record of my own research work under the supervision of Dr. A.Yahaya and Prof. O.E Asiribo. All literatures cited in this work have been duly acknowledged in the text and a list of references provided. No part of this work has been submitted for the award of degree in any other University.

MATTHEW, Pascalis KadaroDate

CERTIFICATION

This thesis entitled "ON THE PERFORMANCE OF PENALIZED REGRESSION METHODS AND THE CLASSICAL LEAST SQUARE METHODS" by MATTHEW, Pascalis Kadaro (M.Sc/Sci/40557/2012-2013) meets the regulations governing the award of the degree of Master of Science of the Ahmadu Bello University, Zaria and is approved for its contribution to knowledge and literary presentation.

Dr. A. Yahaya _____ _____

Chairman, Supervisory Committee SignatureDate

Prof. O.E. Asiribo _____ _____
Member, Supervisory Committee Signature Date

Name _____ _____
External Examiner Signature Date

Prof. B. Sani _____ _____

Head of Department Signature Date

Prof. A.H. Zoaka _____ _____

Dean, School of Postgraduate Studies Signature Date

DEDICATION

This research work is dedicated to God Almighty for His love and protection over my life and also to my late father DSPMatthew Eliseus Kadaroof blessed memory.

ACKNOWLEDGEMENTS

I am grateful to God Almighty, who has been helpful to me in all areas of my endeavor. Thank you Lord for the wisdom and knowledge you have blessed me with.

I am very grateful to my supervisors, Dr. A.Yahaya and Prof. O.E. Asiribo for their relentless support and time to ensure that this work became a successful one.

My deep appreciation goes to all academic and non-academic staff members of the Department of Mathematics, Ahmadu Bello University Zaria, particularly to the Head of Department, Prof. Babangida Sani, the Postgraduate Coordinator in person of Dr. Abubakar Yahaya, the seminar Coordinator in person of Dr. A. Ibrahim for their immense contributions.

A special thanks to my lovely mother, Mrs. Martina M. Kadaro for her relentless support, fervent prayers and encouragement throughout this programme. Mum, may the Almighty God continue to bless you. My gratitude goes to Mr. Nchekwake Eliseus, Mrs. Eucharia Sunday, Mr. Enock Aloysius,Mrs. Ercharu Manaram, and Miss. Kuti Pejeli for their financial and moral support. My gratitude also goes to my fiancée Pamela John Pam,thank you for your patience and understanding sweetheart.

I am very grateful to my cousins; Panam Joshua, John Fimber, Samson Joshua and Usoko Nzong and all my M.Sc classmates, thank you for your wonderful support during this research work. And for those who have contributed in one way or the other, space and time will not permit me to mention your names, thank you very much for your assistance.

ABSTRACT

Regression is one of the most useful statistical methods for data analysis. Multicollinearity is a problem that, pose a challenge to regression analysis by increasing the standard error of the estimators, making the model to be less predictive and difficult for interpretation. Penalized regression which is a variable selection techniquehave been developed specifically to eliminate the problem of multicollinearity and also reduce the flaws inherent in the prediction accuracy of the ordinary least squares (OLS) regression technique. In this thesis, the focus is on the numerical study of these three penalized methods, namely: least absolute shrinkage selection operator (LASSO), elastic net and the newly introduced correlation adjusted elastic net (CAEN). A diabetes dataset which was shown to possess the qualities of multicollinearitywas obtained from previous literature to compare these well-known techniques. 10-fold cross validation (CV) within *glmnet* package was used to entirely search for the optimal λ.The whole path of results (in λ) for the LASSO, Elastic Net and CAEN models were calculated using the path wise Cyclic Coordinate Descent (CCD) algorithms– in *glmnet* package in R,a computationally effective technique for finding out these convex optimization solutions. A regularized profile plot of the coefficient paths for the three methods, were also shown. Predictive accuracy was also assessed using the mean squared error (MSE) and the penalized regression models were able to produce feasible and efficient models capable of capturing the linearity in the data than the ordinary least squares model.It was observed that correlation adjusted elastic net generates a less complex model with a minimum mean square error (MSE).

TABLE OF CONTENTS

List of Tables

List of Figures

Appendix

CHAPTER ONE

GENERAL INTRODUCTION

1.1 Background of the study

In Multiple linear regression analysis, when a large number of predictor variables are introduced in a model to reduce possible modeling biases or there is serious concern of multicollinearity among the predictor variables, variable selection is an important issue.

Regression is one of the most useful statistical methods for data analysis. However, there are many practical problems and computational issues, such as multicollinearity and high dimensionality that pose a challenge to regression analysis.To deal with these challenges, variable selection and shrinkage estimation are becoming important and popular. The traditional approach of automatic selection (such as forward selection, backward elimination and stepwise selection) and best subset selection are often computationally expensive and may not necessarily produce the best model.

The method of penalized least squares (PLS), which is equivalent to penalized maximum likelihood, helps to deal with the issue of multicollinearity by putting constraints on the values of the estimated parameters. A wonderful consequence is that the entries of the variance-covariance matrices are reduced significantly.

Suppose multicollinearity is detected and the predictor variables that cause multicollinearity are identified. As discussed by (Ryan 2009) multicollinearity may not be a problem if the goal is to use the linear regression model for prediction. However multicollinearity is a problem if we use the linear regression model for description or control.Multicollinearity implies that predictor variables form some groups. Within each group, predictor variables are highly correlated. One solution to multicollinearity is to remove one or more of the predictor variables within the same group, but deciding which ones to eliminate

tends to be a difficulttechnical task. A major consequence of multicollinearity is that the parameter estimators and their variances tend to be large. Therefore the inference on the response is highly variable.

To deal with the challenges mentioned above, penalized regression approaches, also called shrinkage or regularization methods, have been developed. Although shrinking some of the regression coefficients toward zero may result in biased estimates, these regression coefficient estimates will have smaller variance. This can result in enhanced prediction accuracy because of a smaller mean squared error (Hastie *et al*., 2009). Regression coefficients are shrunk by imposing a penalty on their size, which is done by adding a penalty function to the least-squares model. Moreover, some of these procedures e.g. the Least Absolute Shrinkage Selection Operator (LASSO) enable variable selection such that only the important predictor variables stay in the model(Szymczak, et al. 2009).

1.2 Statement of the Problem

When perfect multicollinearity or near-perfect multicollinearity exists in a model, parameter estimates of the multiple linear regression models are not unique. In practice, perfect collinearity occurs rarely, what we often have is nearly-perfect collinearity. However quite often we face the issue of multicollinearity when there are strong linear relationships among two or more predictor variables. This happens when two or more predictor variables contribute more or less to a same characteristic of the subjects.In recent years, alternative methods have been introduced to deal with multicollinearity. In particular, methods of penalization become popular and useful. This is also known as simultaneous shrinkage and variable selection. The purpose of this study is to assess thestatistical performances of LASSO, Elastic Net and the newly introduced Correlation Adjusted Elastic-Net (CAEN) regression methods.

1.3 Research Motivation

The motivation for using penalized regression is that in the presence of nearly-perfect multicollinearity, the ordinary least squares estimates are not unique. However, with penalized least squares, these estimates become unique especially when appropriate tuning parameters are chosen. Similarly, without penalization, the ordinary least squares estimators are subject to high variability when multicollinearity exists. With penalization, the variances of the estimators are controlled. Most of the comparisons done by other researchers werebetween LASSO and elastic net.This research attempts to compare LASSO, elastic net and the newly introducedcorrelation adjusted elastic net.And also assess the advantages of using these methods over the classical least squarestechnique. This research attempts to accentuate some of these differences by using numerical results.

1.4 Aim and objectives of the study

The main aim of this research is to assess the performance and advantages of using LASSO, Elastic Net and CAEN methods over the classical regression methods. We hope to achieve this aim through the following objectives:

i. Application ofpenalized regression methods of eliminating multicollinearity.
ii. Identifying the variables that possess the characteristics of multicollinearity using the Variance Inflation Factor, and
iii. Identifying the number of variables selected by each of the penalized regression methodand the classical least squares method.

1.5 Significance of the study

The significance of this study is geared toward detecting variables with the qualities of multicollinearity in a regression model. Also to show why penalized methods are preferred,

over classical least squares technique when faced with the problem of multicollinearity. In achieving this, we explored and compared three penalized methods used in eliminating multicollinearity. This work is also aimed at providing assistance to researchers to ease their decision making as to which technique to be used when encountered with the problem of multicollinearity.

1.6 Scope and limitations of the study

This research is circumscribed by the use of Leave One-Out Cross Validation (LOOCV) criterion to determine the number of variables selected by each of these methods under study, also by the used of mean square error, to assess the predictive accuracy of the methods. The research also gives an overview of each of the procedures in an attempt to highlight the similarities as well as the differences existing among these three penalized methods with respect to variable selection.

1.7 Multicollinearity

Multicollinearity is another important issue in multiple regression. Collinearity means a linear relationship exists between two or more predictor variables, while multicollinearity refers to a situation in which two or more predictor variables are highly linearly correlated. The most extreme case is perfect collinearity (or multicollinearity) where the linear correlation between two predictor variables is either -1 or 1. This happens, for example, when two predictor variables X_1 and X_2satisfy

$$X_2 = a + bX_1 \qquad (1.1)$$

for two real numbers a and b.

In the presence of perfect multicollinearity, parameter estimates of the population multiple linear regression model are not unique. In practice, perfect collinearity occurs rarely. However quite

often we face the issue of multicollinearity when there are strong linear relationships among two or more predictor variables. This happens when two or more predictor variables contribute more or less to the same characteristic of the subjects. For a matrix **A**, let $\boldsymbol{A}^T$ be its transpose and $\boldsymbol{A}^{-1}$ be its inverse matrix, if it exists. When predictor variables are highly linearly correlated, the most significant consequence is that entries of $\left(\mathbb{X}^T\,\mathbb{X}\;\right)^{-1}$ tend to be large, so the predictor variables contribute overlapping and redundant information. Other consequences of multicollinearity are that some predictor variables may not be statistically significant but the model may overall be significant, and that the usual interpretation of coefficient estimates fails in the presence of multicollinearity.Furthermore there is high variability of parameter estimators, because the estimated variance-covariance matrix has large diagonal entries.

Several methods for detecting multicollinearity exist. These include checking for significant change in the parameter estimate when its corresponding predictor variable is added to or removed from the model, checking for insignificance of individual estimators while the model is overall significant, calculating the Variance Inflation Factor (VIF) and carrying out formal multicollinearity tests.

There are several remedies for dealing with multicollinearity. One method is to select a collection of predictor variables that are minimally correlated with each other. This avoids over fitting the regression model and can be normally done with statistical software. However information from other predictor variables is often lost. Furthermore, there is no clear way of selecting a collection of predictor variables that forms the best subset.

Since omitting predictor variables may result in potential loss of information, another method is to include interaction terms into the model to account for high linear correlation among the predictor variables. There are several problems with this approach. One of such is that the form

of interaction is not unique and must be carefully determined. Another problem is that ,the model is much more complex and has too many terms which reduce the degrees of freedom of the inference of the response, and hence reduces the power for predicting and estimating the response.In recent years, alternative methods have been introduced to deal with multicollinearity. In particular, some methods of penalization become popular and useful. This is also known as simultaneous shrinkage and variable selection.

CHAPTER TWO

LITERATURE REVIEW

2.1 Introduction

This chapter seeks to review related literatures on the works of different scholars regarding the classical regression, penalized regression and their applications. Many authors have proposed several penalized regression methods to beat the defects of ordinary least squares with regards to prediction accuracy.

2.2 Classical Regression Analysis

Regression analysis is a statistical technique used to relate variables. Its basic aim is to build a mathematical model to relate dependent variables to independent variables. The method of least squares was firstdiscovered around 1805 (Stigler, 1986). There was a disagreement about who first discoveredthe method of least squares. It appears that, it was discovered independently by CarlFriedrich Gauss (1777-1855) and Adrien Marie Legendre (1752-1833), that Gauss startedusing it before 1803 (he claimed in about 1795, but there is no corroboration of this earlierdate), and that the first account was published by Legendre in 1805, as indicated in (Draper and Smith, 1981). Stigler (1986) noted that, Sir Francis Galton discovered regression around 1885 in his studies related to heredity. Any contemporary course in regression analysis today starts with themethods of least squares and its variations.

Multiple Linear Regression (MLR) is one of the most commonly used data miningtechniques, and can provide insight information in cases where the rigid assumptionsassociated with MLR are met.MLR is a very versatile tool and can be applied to almost any process,system, or area of study. Much has been published regarding this subject, and we refer an interested reader to

Kutner *et al*. (2004), as well as Myers (1990) which provided thorough accounts of MLR and will be indispensable for most readers.

A key step in developing an appropriate MLR model is selecting a method of modelbuilding and a set of best model criteria. Efromyson (1960) introduced Stepwise regression which iscommonly used for model building. Stepwise regressionwas intended to be an automated procedure that selects the most statistically significant variables from a finite pool of independent variables. There are three separate stepwiseregression procedures: forward selection, backward selection and mixedselection. Mixed selection is the most statistically defendable type of stepwise regression,and is a mixture of the forward and backward procedures as indicated in Kutner *et al*. (2004), Neter *et al*. (1996) as well as Draper and Smith (1981).

As noted by Kutner *et al*. (2004); model validation is the final step in the regressionmodeling-building process. Furthermore it was highlighted therein that, there are three main methods associated withmodel validation, as follows:

1. Collection of new data to validate the current model and its predictability.
2. Comparison of current results with other theoretical values, empirical and simulation results.
3. Use of a cross-validation sample to validate and assess the predictive power of the current model.

The cross-validation approach is used to assess the validity andpredictability of the regression models constructed, i.e., a certain amount of the data are removed from the model-building process say twentyrecords, and then use the constructed model to estimatetheir computed values. A general rule of thumb in regression model building is to use 80percent of the data set for the development of the training model and the remaining 20percent for validation of

the model as noted by (Kutner *et al*. 2004). Validation records can be selectedat random from the entire data set, or in the case of time series data, the validationset can be the most current 20 percent (Kutner *et al*. 2004). Adequate regression models areexpected to yield estimates reasonably close to the actual data values. There are lots of statistics available to aid in assessing the predictive power ofregression models. A popular statistic for assessing this predictability is the Root MeanSquared Error of the Prediction (RMSEP) statistic (André *et al*. 2006). This statistic iscomputed by calculating the square root of the Sum Squared Errors (SSE) for the withheldrecords divided by the corresponding degrees of freedom. Lower RMSEP values indicatebetter model predictability. Another common model validation statistic is the classical coefficient of determination, orR^2 , statistic. This value is also computed for the withheldsample, and provides some insight into the predictability of the model. By definition, higherR^2values are preferred, i.e., theR^2 statistic indicates the amount of variation explained bythe regressors in the regression model.

Breiman and Friedman (1997) observed that, when considering multiple regression models, it is of great importance for the predictors to share strength among different models.(Turlach *et al*. 2005)also observed that, it is of particular interest, when there are large number of covariates,to find a common set of variables that can be used for all models under investigation. In the context ofmean regression, Turlach *et al*. (2005) considered the problem of selecting a subset of 770 wavelengths that are suitable aspredictors for 14 different but correlated infra-red spectrometry measurements, and they proposed a novel regularizationmethod to perform simultaneous variable selection.Because classical regression approaches require thenumber of samples to exceed the number of variables,they are not applicable in case of genome wide association (GWA) data. Additionally,least-squares estimates of regression coefficients may behighly

unstable, especially in cases of correlated predictorvariables, which lead to low prediction accuracy.

In genomic settings, where collinear predictors typically outnumbered available samples ($p>n$), an examplebeing the prediction of cancer patient survival from tumor geneexpression data (Beer *et al.*, 2002; Shedden *et al.*, 2008; Sørlie*et al.*, 2001; van de Vijver *et al.*, 2002; Wigle *et al.*, 2002). In thissetting, ordinary regression is subject to overfitting and instabilityof coefficients (Harrell *et al.*, 1996), and stepwise variable selectionmethods do not scale well as observed by (Yuan and Lin, 2006). Regression has beensuccessfully adapted to high-dimensional situations by penalization methods (see for instance, Hesterberg 2008), and penalized regressionhas been shown to outperform univariate and other multivariateregression methods in multiple genomic datasets (Bøvelstad *et al.*, 2007).

Many simulation studieshave been carried out and it was suggested that, least-squares estimates can be quite poor, see for instance Roecker (1991), Adams (1990) as well as Hurvich and Tsai (1990). Thesestudies show that often prediction errors using OLS are too small and that the usual 95% confidence intervals will only include the true value of the parameterin roughly 50% of cases. When predictor variables are strongly correlated, theprediction errors were shown to become too large.

2.3 Penalized Regression

It is a well-knownfact that, OLS often does poorly in both prediction and interpretation especially when some of the predictor variables are collinear. Penalization techniques have been proposed to improve on the prediction flaws inherent in ordinary least squares. Hoerl and Kennard (1970) introduced the ridge regression which estimates the regression coefficients through an l_2-norm penalized least-squares criterion. Friedman *et al.*,(2007) observed that ridge

regression shrinks the coefficients of correlated predictor variables toward each other, allowing them to borrow strength from each other. However, this behavior is not without its problems. For example, in the case of the *k* identical predictor variables mentioned above, they each get identical coefficients with size 1/*k* that,which any single one would get if, fit alone. The ridge penalty is ideal if there are many predictor variables, and all have non-zero coefficients (from a Bayesian perspective only if these are drawn from a Gaussian prior distribution).Best subset selectionon the other hand produces a sparse model, but it is extremely variable because of itsinherent discreteness, as addressed by Breiman (1996).Frank andFriedman (1993) introduced bridge regression which minimizes the Residual Sum of Squares (RSS). The estimator from bridge regression is not explicit; however Frankand Friedman (1993) argued that the optimal choice of the parameterγyields reasonable predictors. This is because, it controls the degree of preference for the true coefficienttoalign with the original variable axis directions in the predictor space.Also Tibshirani (1996)introduce LASSO regression,which minimizes the RSS subject to a bound on the L_1-normof the coefficients. The Elastic net was proposed by Zou and Hastie (2005a) which is the combination of both $L_1$and L_2 norms. The Dantzig selector which was introduced by (Candes and Tao, 2007), a slightly modified versionof the LASSO.Li and Lin (2010) proposed a related Bayesian Elastic Net method with a slightly different specification of the prior where the two penalty parameters were chosen by the empirical Bayes method. Tan (2012) introduced what is known as Correlation Adjusted Elastic Net regression which is an extension of Elastic Net regression. Tan (2012) also introduced correlation adjusted regression which is an extension of ridge regression.Bondell and Reich (2008) introduced the OSCAR (Octagonal Shrinkage and Clustering Algorithm for Regression). There are so many penalized

regression methods proposed in recent years but this thesis focuses on LASSO, Elastic Net and Correlation Adjusted Elastic Net methods because of their performance.

2.31 LASSO Regression

Tibshirani (1996) proposed the LASSO estimator which estimates the regression coefficients through an l_1-norm penalized least-squares criterion. This is equivalent to minimizing the sums of squares of residuals plus an l_1 penalty on the regression coefficients. Due to the nature of the l_1 penalty, LASSO performscontinuous shrinkage and variable selection simultaneously. In addition LASSO possesses the properties of both the l_2 (ridge) penalization and best-subset selection. It was argued that, the automatic feature selection property makes the LASSO a better choicethan the l_2 penalization in high dimensional problems, especially when there are lots of redundant noise features (Friedman *et al.*,2007). Although the $l_2$regularization has been widely used in various learning problems such as smoothing splines(Wahba, 1990), the support vector machine (Vapnik, 1995) and neural networks where the $l_2$regularization is called weight decay (Hastie *et.al*, 2009). An l_1 methodcalled *basis pursuit* was also used in signal processing (Chen*et al.*, 2007).There are many theoretical works to prove the superiority of the l_1 penalization in sparsesettings. The LASSO estimator has two desirable properties. Firstly, the nature of regularization used in the LASSO leads to sparse solutions. Secondly, it is also computationally feasible as it was seen in the works of (Efron *et al.,* 2004) and (Friedman *et al.,* 2007). The sparse solutions obtained by using LASSO automatically leads to model selection. In the finite dimensional case, many authors have studied the model-consistency properties of the LASSO and investigated conditions under which the Lasso can recover the true sparsity pattern as in the case of Zhao and Yu (2006). Yuan and Lin (2006) have studied the neighborhood selection properties of the LASSO in graphical models.

Donohoand Johnson (1994) prove the near minimax optimalityof soft-thresholding (l_1 shrinkage with orthogonal predictors). Donoho *et.al.*(2004) also shown thatthe l_1 approach is able to discover the "right" sparse representation of the model undercertain conditions.

Knight and Fu (2000) have shown consistency for LASSO type estimators (generally bridge estimators) with fixed *p* under some regularity conditions on the design. They obtainedthe asymptotic normal distribution with a fixed true parameter$\boldsymbol{\beta}$ and local asymptotic,that is, when the true parameter is small but nonzero in finite samples. Also, theyderived asymptotic properties of LASSO type estimators under nearly singular designmatrices.

Osborne *et al.*,(2000a) proposed two algorithms for the computation of the lasso, a*compact descent algorithm* was derived to solve the selection problem for a particularvalue of the tuning parameter, and then a *homotopy method* for the tuning parameter was developed to completely describe the possible selection.

Efron *et al.*,(2004)later proposed *Least Angle Regression Selection* (LARS) for a model selection algorithm.They showed that with a simple modification, the LARS algorithm implements the LASSO. Efron *et al.*,(2004) also studied an efficient way of selecting the optimal fit and the effective degrees of freedom of the LASSO, where it was discovered that, the size of the active set(the indices corresponding to covariates to be chosen) can be used as a measure of thedegrees of freedom, which changes, not necessarily monotonically, along the solutionpaths of LARS.Zou *et al.*,(2007) improved on the work of (Efron *et al.*,2004) and showed that the number of nonzero coefficients is an unbiased estimate for degrees of freedom of the LASSO. In addition,Zou *et al.*,(2007) showed that the unbiased estimator is asymptotically consistent, thusvarious model selection criteria can be used with the LARS algorithm for the optimal LASSO fit.

However, as variable selection becomesincreasingly important in modern data analysis, the LASSO is much more appealing due toits sparse representation.Although the LASSO has shown success in many situations, it has some limitations as shown in Zou and Hastie (2005b). For instance considerthe following three scenarios:

1. In the unsual $p > n$ case, the LASSO selects at most n variables before it becomes saturated, because of the nature of the convex optimization problem. This seems to be a limiting feature for a variable selection method. Moreover, the LASSO is not well defined unless the bound on the l_1 norm of the coefficients is smaller than a certain value.
2. If there is a group of variables among which the pairwise correlations are very high, then the LASSO tends to select only one variable from the group and does not care which one is selected.
3. For usual $n > p$ situations, if there exist high correlations among predictors, it has been empirically established that the prediction performance of the LASSO is dominated by ridge regression (Tibshirani, 1996). Scenarios (1) and (2) make the LASSO an inappropriate variable selection method in some situations. Zou and Hastie (2005a) illustrate their points by considering the gene-selection problem in microarray data analysis. A typical microarray data set has several thousand predictors (genes) and often less than (in most cases) 100 samples. For those genes sharing the same biological pathway, the correlations among them can be high. They think of those genes as forming a group. The ideal gene selection method should be able to do two things: eliminate the trivial genes, and automatically include whole groups into the model once one gene among them is selected (grouped selection). For this kind of $p > n$ and grouped variables situation, the LASSO is not the ideal method, because it can only select at most

n variables out of p candidates (Efron *et al.*, 2004), and it lacks the ability to reveal the grouping information. As for prediction performance, scenario (3) is not rare in regression problems. So it is possible to further strengthen the prediction power of the LASSO.

Kyung *et al*., (2010) observed from a Bayesian point of view, that the LASSO penalty corresponds to a Laplace (double exponential) prior over the regression coefficients, which expects many coefficients to be close to zero, and a small subset to be larger and non-zero. Kyung *et al*., (2010) also showed that the two tuning parameters could be estimated within the Gibbs sampler by assigning hyper priors to them. Hans(2010) argued that, the Bayesian LASSO doesn't set any variables to exactly zero and therefore needs to be combined with some other form of variable selection.To compensate the ordering limitations of the LASSO, Tibshirani *et al.*,(2005) introduced the fused LASSO. The *fused LASSO* penalizes the L_1-norm of both the coefficients and their differences:

One important difference between the LASSO and ridge regression occurs for the predictor variables with the highest regression coefficients. Whereas the l_2 penalty pushes the regression coefficients towards zero with a force proportional to the value of the coefficient, the l_1 penalty exerts the same force on all non-zero coefficients. Hence for the variables that are most valuable (i.e., that clearly should be in the model and where shrinkage toward zero is less desirable) an l_1 penalty shrinks less (Hesterberg *et al*., 2008). The extent of shrinkage in ridge regression is also dependent on the sample size (Gianola, 2013).

2.32 Elastic Net Regression

Zou and Hastie (2005) proposed the *Elastic Net* penalty which is based on combined penalties of LASSO and ridge regression. The penalty parameter α determines how much weight should be

given to either the LASSO or ridge regression. The Elastic Net with α set to 0 is equivalent to ridge regression. The Elastic Net with α close to 1 performs much like the LASSO, but removes any degeneracies and odd behavior caused by high correlations.

Zou and Hastie (2005) highlighted two aspects that are important when evaluating the quality of a model:

(a) Accuracy of prediction on future data—it is difficult to defend a model that predicts poorly;

(b) Interpretation of the model—scientists prefer a simpler model because it puts more light on the relationship between the response and covariates. Parsimony is especially an important issue when the number of predictors is large.

Bühlmann and vandeGeer (2011)have shown that analysis with the *Elastic Net* can result in lower mean squared errors than the LASSO and ridge regression when predictor variables are correlated.(Tutz and Ulbricht, 2009) also shown that, the*Elastic Net* produces higher number of correctly identified influential variables than the LASSO, and has much lower false positive rate than ridge regression. Later Zou and Zhang (2009) proposed the*Adaptive Elastic Net*. Li and Lin (2010) introduced the *Bayesian Elastic Net.*

Fan and Li (2001)proposed the *Smoothly Clipped Absolute Deviation* (SCAD) penalty for penalized leastsquares to reduce bias and satisfy certain conditions to yield continuous solutions. Also, they derived the fixed tuning parameter asymptotic distribution of the estimator and showed that the estimator satisfies the oracle property (consistent model selection).Large sample properties of SCAD estimators are studied in Kwon and Yongdai(2012).

2.33 Correlation Adjusted Elastic Net (CAEN) Regression

Tan(2012) introduces CAEN (correlation adjusted elastic net) which is an extension of elastic net regression. The behavior of the CAEN regression is similar to that of *Elastic Net*

regression. The sample correlation is also included in the penalty term. After applying argumentation to the data set, the CAEN can be reduced to the LASSO regression.

2.4 Applications of Penalized Regression

To achieve better prediction ability in the face of multicollinearity; penalized regression methods have been proposed for variable selection in high dimensional studies which focuses on human genetic data as in the case of Sung *et al.*(2009) and as well as Cho *et al*., (2010). Fu (1998) compared LASSO, Ridge regression and Bridge regression methods using a prediction performance criteria. He argued that, because of the nonlinearity of the Bridge operator, the Bridge model does not always perform well in estimationand prediction compared to the other shrinkage methods: the LASSO and Ridge regression.

Efron *et al.*,(2004) compared LARS with Ridge and LASSO using a diabetes dataset where it wasfound out that, one of the advantages of LARS is the short computation time compared to the other two methods.Usai *et al*.,(2009) tested the least angle regression version of the LASSO on the Quantitative Trait Loci Marker Assisted Selection (QTLMAS 2008) data and found that 169 Single Nucleotide Polymorphism (SNPs) were needed to explain the variation of the 48 simulated Quantitative Trait Loci (QTLs). Yet, they used a rather adhoc cross-validation approach where the highest correlation between genomic breeding values (sum of the regression coefficients of the SNPs) and true simulated breeding values was used as stopping criterion. This approach is difficult to generalize to real data because it relies on the fact that the breeding values are known or estimated without error.

Kooperberg *et al.* (2010) compared the performance of Elastic Net and LASSO using uncorrelated predictor variables.Ayers and Cordell (2010) compared the statistical properties of the LASSO, Ridge regression and Elastic Net regression methods on simulated data.Ayers and

Cordell (2010) considered the effects from groups of highly correlated variables as a single signal to prevent inflated false positive rates, which is more appropriate for prediction of future phenotypes. They also used a permutation approach aimed at controlling Type I error.

Motyer *et al.* (2011) considered a penalized regression using the LASSO procedure for a genome-wide association study (GWAS17) data set and show that post-processing of the penalized-regression results with subsequent stepwise selection may lead to improved identification of causal single-nucleotide polymorphisms.The GWA17 data set contains 24,487 SNPs from 697 individuals. After applying LASSO to the dataset only 6,321 SNPs were left to be considered in the model.

Waldronn *et al.*(2011) compared the LASSO, Ridge and Elastic Net penalties for prediction and variable selection through simulation of contrasting scenarios of correlated high-dimensional survival data. The study found that, a 2D tuning of the Elastic Net penalties was necessary to avoid mimicking the performance of LASSO or Ridge regression. Furthermore, it was also found out that, in a simulated scenario favoring the LASSO penalty, a univariate pre-filter made the Elastic Net behave more like Ridge regression, which was detrimental to prediction performance. This study once more demonstrated the real-life application of these methods to predicting the survival of cancer patients from microarray data, and to classification of obese and lean individuals from meta-genomic data. They used a crossvalidationstrategy for assessment of model prediction (Molinaro*et al.*, 2005; Simon *et al.*, 2011), with an additional inner level ofcross-validation for model tuning in training data (Goeman, 2011).Both examples proved favorable to the L_2 penalty, and modelstrained by Ridge regression and Elastic Net showed independentpredictive ability, whereas models trained by the LASSO did not. They emphasize evidence of overfitting in both simulatedand real experimental data, and summarize

methods for realisticassessment of prediction accuracy with limited sample size. Basedon results from this study and best practices for high-dimensional model validation, they concluded with an end-to-end methodology foreffective application of penalized regression to diverse genomic datafor prediction and variable selection.

Doreswamyand Chanabasayya(2013) considered the performance analysis of regularized linear regression models for oxazolines and oxazoles derivatives descriptor dataset where it was found that, regularized regression models were able to produce feasible and efficient models capable of capturing the linearity in the data than the ordinary least squares model. It was shown that, the *Elastic Net* and *LARS* had similar accuracies as well as LASSO and relaxed LASSO had similar accuracies but outperformed ridge regression in terms of the Root Mean Square Error (RMSE) and R square metrics.Waldmann *et al.*(2013) compared the statistical performance of LASSO and Elastic Net methods on a real data set from a 50K genome-wide Single Nucleotide Polymorphism (SNP) panel of 5570 Fleckvieh bulls. It was concluded that, it is important to analyze GWAS data with both LASSO and the Elastic Net where by an alternative tuning criterion for minimizing MSE is needed for variable selection.

CHAPTER THREE

METHODOLOGY

3.1 Regression Analysis

Regression analysis is one of the most important tools for analyzing relationships between one response variable and one or more explanatory variables. It is widely used in our day-to-day endeavors, ranging in diverse areas of human life including social and biological sciences, economics and so on. Regression analysis has become one of the most important tool in data analysis.

The use of regression analysis has significant applications in medical and countless other research areas, and is an important component of modern data analysis. The central objective is to understand the relationship between a response (or dependent) variable and a set of predictor variables (also known as explanatory variables, regressors, covariates, or independent variables) and to apply the relationship for the purpose of estimating and/or predicting future responses. There are many important theoretical, practical and computational issues related to regression modeling and inference, including specification of the link function that relates the response variable and predictor variables, estimation of regression parameters in the link function, measure of model performance, diagnostic statistics to assess the modeling assumptions and goodness-of-fit, as well as remedial methods in the cases of violation of assumptions.

The response variable can be continuous or categorical. Although some philosophical ideas may be similar for different types of response variables, methodologies are different, in particular on the choice of the link function and assessment of the model's goodness-of-fit.Effective model building is a significant issue. Essentially, we search for the best fitting and most parsimonious model that is practically meaningful and reasonable to describe the relationship between the

response and the set of predictor variables. The fit of the model to the data set is determined by measures of goodness-of-fit, and being most parsimonious requires effective methods of model selection.

For instance, in the multiple linear regression models, let Y denote the response variable (also called the endogenous variable or the dependent variable) and X_1,X_2 , … , $X_p$denote the explanatory variables (also called exogenous variables or independent variables). The relationship between Y andX_1,X_2 , … , X_p can be expressed as

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \ + \cdots + \beta_p X_p + \varepsilon \qquad (3.1)$$

The parameters $\beta_0, \beta_1, \beta_2, \dots, \beta_p$ are called regression coefficients and ε is the random error term.

Given a data set $\{y_i, x_{i1}, x_{i2}, \dots, x_{ip}\}_{i=1}^{n}$ of n statistical units, each statistical unit can be expressed as

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i i = 1,2, \dots, n \qquad (3.2)$$

where $y_i$is the i^{th} response observation, $\beta_0, \beta_1, \beta_2, \dots, \beta_p$ are the unknown parameters and $\varepsilon_i \sim N(0, \sigma_i^2)$. Often those n equations can be rewritten invector form as

$$\mathbb{Y} = \mathbb{X}\boldsymbol{\beta} + \ \boldsymbol{\varepsilon} \qquad (3.3)$$

or

$$E(\mathbb{Y}) = \mathbb{X}\boldsymbol{\beta} \qquad (3.4)$$

Where

$$\mathbb{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \mathbb{X} = \begin{pmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \qquad and \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

- $\mathbb{X}$is called design matrix.
- $\mathbb{Y}$is called response vector.
- $\boldsymbol{\beta}$ is the parameter vector.
- $\boldsymbol{\varepsilon}$ is the error vector.

we plug each individual statistical unit in equation 3.1 to obtain the matrix form as

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i = x_i^T \beta + \varepsilon_i \,, \qquad (3.5)$$

$$where x_i^T = \begin{pmatrix} 1 & x_{i1} \dots x_{1p} \end{pmatrix} and i = 1,2, \dots, n$$

3.21 Assumptions of Multiple Linear Regression

1. Linearity: The relationship between the explanatory variables and the response variable is linear. This is the only restriction on the parameters (not explanatory variables), since the explanatory variables are regarded as fixed values. That is,
   - $E\left(y_i \middle| x_{i1}, x_{2p}, \dots x_{1p}\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i = x_i^T \beta$
   - $\frac{\partial E(y_i | x_i)}{\partial x_i} = \beta$
2. Independence: There are two types of independence.
   - Each combination of explanatory variable and error is independent.
   - The error terms are independent. Therefore, $Cor\left(\varepsilon_i, \varepsilon_j\right) = 0$ or equivalently$Cov\left(y_i, y_j\right) = 0 \; for all \; i \neq j.$
3. Normality: The error terms follow normal distribution.
   - $\varepsilon_i \sim N(0, \sigma_i^2)$
   - $\mathbb{Y} \sim N(\mathbb{X}\beta, \sigma^2)$

where

$$\sigma^2 = \begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \ddots & & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{pmatrix}$$

4. Equal Variance: error terms are assume to have equal variances.

- $Var(\varepsilon_i) = Var(\varepsilon_j) = \sigma^2 \, foralli \neq j$
- $Var(y_i) = Var(y_j) = \sigma^2 \, foralli \neq j$

3.22 Ordinary Least Squares

The ordinary least squares (OLS) is a classic technique to estimate the parameters of the multiple linear regression model. It should be noted at this point that, there are two principles establishing the OLS regression model:

$$\hat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_{i1} + \widehat{\beta}_2 x_{i2} + \cdots + \widehat{\beta}_p x_{ip} = x_i^T \widehat{\beta}, \quad i = 1,2, \ldots, n$$

Firstly, $\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} (y_i - \hat{y}_i) = 0$,

where

$$e_i = y_j - \hat{y}_i, \qquad iscalledthe\ i^{th} residual\ .$$

$$\hat{\varepsilon}^T = (e_1, e_2, \ldots, e_n) and \boldsymbol{j} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}_{\boldsymbol{n \times 1}}$$

Secondly, $\hat{\varepsilon}^T \hat{\varepsilon} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 isminimized.$

Here, $\hat{y}_i$ is an estimator of $E(y_i)$ and there is no distribution assumption required for OLS.Now, in vector form we have,

$$\hat{\varepsilon}^T \hat{\varepsilon} = \sum_{i=1}^{n} e_i^2$$

$$= \sum_{i=1}^{n} \left(y_i - x_i^T \hat{\beta}\right)^2$$

$= \sum_{i=1}^{n} \left(y_i - x_i^T \hat{\beta}\right)^2$; which we can write in matrix form as;

$$= \left(\mathbb{Y} - \mathbb{X}\hat{\beta}\right)^T \left(\mathbb{Y} - \mathbb{X}\hat{\beta}\right)$$

$$= \mathbb{Y}^T \mathbb{Y} - 2\mathbb{Y}^T \mathbb{X}\hat{\beta} + \hat{\beta}^T \mathbb{X}^T \mathbb{X}\hat{\beta}$$

Note that $\mathbb{Y}^T \mathbb{X}\hat{\beta} = \hat{\beta}^T \mathbb{X}^T \mathbb{Y} since the y are both scalars.$

To find $\hat{\beta} = \left(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p\right)^T$ that minimizes $(\hat{\varepsilon}^T \hat{\varepsilon})$ we take the derivatives of $(\hat{\varepsilon}^T \hat{\varepsilon})$ with respect to $\hat{\beta}$ and let the derivative equal to zero to obtain $\hat{\beta} = \left(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p\right)^T$

Hence,

$$\frac{\partial \hat{\varepsilon}^T \hat{\varepsilon}}{\partial \hat{\beta}} = \mathbf{0} - 2\mathbb{X}^T \mathbb{Y} + 2\mathbb{X}^T \mathbb{X}\widehat{\boldsymbol{\beta}}$$

Finally, the OLS estimator is

$$\hat{\beta}_{OLS} = \left(\mathbb{X}^T \mathbb{X}\ \right)^{-1} \mathbb{X}^T \mathbb{Y} \qquad (3.6)$$

where $\hat{\beta}_{OLS}$ is the best linear unbiased estimator$(BLUE)$. Specifically,

- $Best$ Means $Var\left(\hat{\beta}_{OLS}\right) = \sigma^2 \left(\mathbb{X}^T \mathbb{X}\ \right)^{-1}$ has the minimum variance among all linear unbiased estimators.
- The $\hat{\beta}_{OLS}$ is a linear function of $\mathbb{Y}$. That is $\left[\left(\mathbb{X}^T \mathbb{X}\ \right)^{-1} \mathbb{X}^T\right] \mathbb{Y}$.
- The $\hat{\beta}_{OLS}$ is an unbiased estimator for $\boldsymbol{\beta}$ .That is,

$$E\left(\hat{\beta}_{OLS}\right) = \left(\mathbb{X}^T \mathbb{X}\ \right)^{-1} \mathbb{X}^T E(\mathbb{Y}) = \left(\mathbb{X}^T \mathbb{X}\ \right)^{-1} \mathbb{X}^T \mathbb{X}\beta = \beta.$$

Based on $\hat{\beta}_{OLS}$, we can derive an unbiased estimator $\hat{\sigma}^2 for \sigma^2$, given by

$$\hat{\sigma}^2 = \frac{1}{n - (p+1)} \sum_{i=1}^{n} \left(y_i - x_i^T \hat{\beta}\right)^2$$

$$= \frac{1}{n-(p+1)} \left(\mathbb{Y} - \mathbb{X}\hat{\beta}\right)^T \left(\mathbb{Y} - \mathbb{X}\hat{\beta}\right)$$

$$= \frac{SSE}{n-(p+1)} \quad (3.7)$$

Where $(p+1)$ is equal to the number of $\boldsymbol{\beta}^{,s}$

3.23 Maximum Likelihood Estimation

Maximum likelihood estimator (MLE) is another classical method to estimate the multiple, linear regression model. Under the assumptions of multiple regression, the likelihood function is given by

$$\mathcal{L}\left(\beta, \sigma^2\right) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \, exp\left(-\frac{\left(y_i - x_i^T\hat{\beta}\right)^2}{2\sigma^2}\right)$$

$$= \frac{1}{\left(2\pi\sigma^2\right)^{\frac{n}{2}}} \, exp\left(-\frac{\left(\mathbb{Y} - \mathbb{X}\hat{\beta}\right)^T \left(\mathbb{Y} - \mathbb{X}\hat{\beta}\right)}{2\sigma^2}\right)$$

And the log likelihood function is given by

$$\ell\left(\beta, \sigma^2\right) = ln\, \mathcal{L}\left(\beta, \sigma^2\right)$$

$$= -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\left(\sigma^2\right) - \frac{1}{2\sigma^2}\left(\mathbb{Y} - \mathbb{X}\hat{\beta}\right)^T \left(\mathbb{Y} - \mathbb{X}\hat{\beta}\right)$$

To find $\hat{\beta} and \hat{\sigma}^2$ which maximizes the log-likelihood function, let

$$\begin{cases} \dfrac{\partial \ell\left(\beta, \sigma^2\right)}{\partial \beta} = 0 \\ \dfrac{\partial \ell\left(\beta, \sigma^2\right)}{\partial \sigma^2} = 0 \end{cases} \Rightarrow \begin{cases} \hat{\beta}_{ML} = \left(\mathbb{X}^T\, \mathbb{X} \quad\right)^{-1} \mathbb{X}^T\, \mathbb{Y} \\ \hat{\sigma}^2{}_{ML} = \dfrac{1}{n}\left(\mathbb{Y} - \mathbb{X}\hat{\beta}\right)^T \left(\mathbb{Y} - \mathbb{X}\hat{\beta}\right) \end{cases}$$

As we see,$\hat{\beta}_{ML}$ is an unbiased estimator which equals to $\hat{\beta}_{OLS}$ but $\hat{\sigma}^2{}_{ML}$ is a biased estimator of σ^2.

3.3 Penalized Regression

Substantialamount of efforts hasbeen put in the development of penalized regression methods for simultaneous variable selection and coefficient estimation (Hoerl and Kennard, 1970; Tibshirani, 1996; Fred and Friedman, 1993; Zou and Hastie 2005; Tan, 2012). In practice, even if the sample size is small, a large number of predictors are typically included to mitigate modeling biases. With such a large number of predictors, there might exist some problems among explanatory variables, in particular, there could be a problem with multicollinearity. Also, with a large number of predictors there is often a desire to select a smaller subset that not only fits as well as the full set of variables, but also contains more important predictors. Such concerns have led to prominent development of least squares (LS) regression methods with various penalties to discover relevant explanatory factors and to get higher prediction accuracy in linear regression.

We consider a linear regression model with *n* observations on a dependent variable *Y* having *p* predictors:

$$y = \mathbb{X}\beta + \varepsilon \qquad (3.8)$$

Penalized regression approaches havebeen used in cases where $p < n$, and in the ever-more-common case with$p \geq n$. Inthe former case, penalized regression, and its accompanying variable selection features, can lead to finding smaller groups of variables with good prediction accuracy. If$p \geq n$,ordinary least-squares regression (OLS), which minimizes the residual sum of squares

$$RSS = \left(\mathbb{Y} - \mathbb{X}\hat{\beta}\right)^T \left(\mathbb{Y} - \mathbb{X}\hat{\beta}\right) \qquad (3.9)$$

yields an estimator that is not unique since $\mathbb{X}$is not of full rank. Moreover, the variances will be artificially large. Here, again, penalized regression approaches can guide us to good subsets of predictors.

In general, the Penalized Least Squares (PLS) is aimed at minimizing Sum of Squares due to Error (SSE),

$\sum_{i=1}^{n} e_i^2 = \hat{\varepsilon}^T \hat{\varepsilon} = (\mathbb{Y} - \mathbb{X}\beta)^T (\mathbb{Y} - \mathbb{X}\beta)$ subject to $Pen(\beta) \leq t$,

Where $Pen(\beta)$(specific penalty) is a function of $\hat{\beta} = \left(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p\right)^Tandt$is a tuning parameter. This constrained optimization problem can be solved with the equivalent Lagrangian formulation which minimizes

$$\text{PLS} = \text{OLS} + \text{Penalty} = (\mathbb{Y} - \mathbb{X}\beta)^T (\mathbb{Y} - \mathbb{X}\beta) + \lambda Pen(\beta) \qquad (3.10)$$

where λ is a tuning parameter and controls the strength of shrinkage. For example, when λ = 0, no penalty is applied and we have the ordinary least squares regression. When λ gets larger, more weight is given to the penalty term.

Desirable properties of penalization include variable selection and grouping effect. That is, by penalization, it is hoped that the variables that are truly statistically significant are selected into the model, and highly correlated predictor variables should be selected all together or excluded all together.

3.4 LASSO Regression

The Least Absolute Shrinkage and Selection Operator (LASSO) regression method was introduced by Tibshirani (1996)as an estimation and variable selection method. It is also called L_1 penalized regression. The LASSO is a penalized least squares procedure that minimizes RSS subject to the non-differentiable constraint expressed in terms of the L_1 norm of the coefficients.

The penalty function is given by

$$Pen(\beta) = \lambda \sum_{i=1}^{p} |\beta_i| \quad (3.11)$$

The objectives is to minimize

$$\hat{\beta}_{LASSO} = \arg\min_{\beta \in R^p} (\mathbb{Y} - \mathbb{X}\beta)^T (\mathbb{Y} - \mathbb{X}\beta) + \lambda \sum_{i=1}^{p} |\beta_i| \quad (3.12)$$

where λ is a non-negative regularization parameter.

Since the LASSO penalty term is no longer quadratic, there is no explicit formula for the mean squared error of the LASSO estimator. Generally,the $Bias(\hat{\beta}_{LASSO})$also increases as the tuning parameter λ increases. While the variance, $Var(\hat{\beta}_{LASSO})$ decreases.

For instance,

when $\lambda = 0$

$$\begin{aligned} MSE(\hat{\beta}_{LASSO}) &= trace(Var(\hat{\beta}_{LASSO})) + Bias^T(\hat{\beta}_{LASSO})\, Bias(\hat{\beta}_{LASSO}) \\ &= trace(Var(\hat{\beta}_{LASSO})) + 0 \\ &= MSE(\hat{\beta}_{OLS}), \quad (3.13) \end{aligned}$$

And when $\lambda \to \infty$

$$\begin{aligned} MSE(\hat{\beta}_{LASSO}) &= trace(Var(\hat{\beta}_{LASSO})) + Bias^T(\hat{\beta}_{LASSO})\, Bias(\hat{\beta}_{LASSO}) \\ &\to 0 + (-\beta)^T(\beta) = \beta^T\beta \quad (3.14) \end{aligned}$$

Since $Bias^T(\hat{\beta}_{LASSO})\, Bias(\hat{\beta}_{LASSO})$ and $trace(Var(\hat{\beta}_{LASSO}))$move to opposite directions as the tuning parameter λincreases, thus, we can choose an optimal value of the parameter λthat minimizes $MSE(\hat{\beta}_{LASSO})$.

3.5 Elastic Net Regression

Zou and Hastie (2005) proposed the *Elastic Net* penalty which is based on a combined penalties of LASSO and Ridge regression penalties to improve the prediction performance of the naive elasticnet by correcting this double-shrinkage. Suppose that,*a* data set has *n* observations with *p* predictors. Let $y = (y_1, y_2, \dots, y_n)^T$be theresponse and $\mathbb{X} = [X_1 \mid, \dots, \mid X_p]$ be the model matrix, where $X_j = (x_{11}, \dots, x_{np})^T$, for $j = 1, \dots, p$ are predictors. After a location and scale transformation, it can be assumedthat the responseis centered and the predictors are standardized, such that:

$$\sum_{i=1}^{n} y_i = 0, \sum_{i=1}^{n} x_{ij} = 0\,, \sum_{i=1}^{n} {x_{ij}}^2 = 1, for\ j = 1, \dots, p$$

For any fixed non-negative λ_1 and λ_2 , we define the Elastic Net criterion as

$$L(\lambda_1, \lambda_2, \beta) = \mid \boldsymbol{y} - \mathbb{X}\beta \mid^2 + \lambda_2 \mid \beta \mid^2 + \lambda_1 \mid \beta \mid_1 \qquad (3.15)$$

where

$$\mid \beta \mid^2 = \sum_{j=1}^{p} {\beta_j}^2 \ and \ \mid \beta \mid_1 = \sum_{j=1}^{p} \mid \beta_j \mid$$

The naïve elastic net estimator is the minimizer of (3.14)

$$\hat{\beta} = \arg\min_{\beta} L(\lambda_1, \lambda_2, \beta) \qquad (3.16)$$

The above procedure can be viewed as a penalized least-squares method. Let$\alpha = \frac{\lambda_2}{\lambda_1 + \lambda_2}$, then solving for β in (3.14) is equivalent to solving the optimization problem:

$$\hat{\beta} = \arg min_{\beta} \mid \boldsymbol{y} - \mathbb{X}\beta \mid^2, subject\ to\ (1 - \alpha) \mid \beta \mid_1 + \alpha \mid \beta \mid^2$$

$$\leq t\ for\ some\ constant, t \qquad (3.17)$$

where the function $(1-\alpha)\,|\beta|_1+\alpha\,|\beta|^2$ is called the elastic net penalty.Which is a convex combinationof the LASSO and Ridge penalty. When α = 1, the naive elastic net becomes simpleridge regression.The elastic net estimator be defined in matrix form as

$$\hat{\beta}_{EN}=\arg\min_{\beta\in R^P}(\mathbb{Y}-\mathbb{X}\beta)^T(\mathbb{Y}-\mathbb{X}\beta)+\lambda_1\sum_{i=1}^{p}|\beta_i|+\lambda_2\sum_{i=1}^{p}\beta_i^2 \qquad (3.18)$$

$$\hat{\beta}_{EN}=\arg\min_{\beta\in R^P}(\mathbb{Y}-\mathbb{X}\beta)^T(\mathbb{Y}-\mathbb{X}\beta)+\lambda_1\sum_{i=1}^{p}|\beta_i|+\lambda_2\beta^T\beta \qquad (3.19)$$

where λ_1 and λ_2 are non-negative regularization parameters.

According to the Ridge and LASSO regression procedures, the amount of shrinkage increases as λ increases. This implies that when either $\lambda_1\to 0$or$\lambda_2\to 0$, we have $\hat{\beta}_{EN}\to 0$. Since the LASSO penalty term is included in $\hat{\beta}_{EN}$there is no explicit formula of the mean squared error for the Elastic Net estimator except when$\lambda_1=0$.

3.6 Correlation Adjusted Elastic Net Regression

Tan (2012) introduced the Correlation Adjusted Elastic Net $(CAEN)$regression procedure which is a combination of L_1penalized regression and Correlation Adjusted Regression(CAR). It's also an extension of ElasticNet regression. The Correlation Adjusted Elastic Net Regression can be defined as:

$$\hat{\beta}_{CAEN}=\arg\min_{\beta\in R^P}(\mathbb{Y}-\mathbb{X}\beta)^T(\mathbb{Y}-\mathbb{X}\beta)+\lambda_1\sum_{i=1}^{p}|\beta_i|+\lambda_2\beta^T W\ \beta \qquad (3.20)$$

$$=\arg\min_{\beta\in R^P}(\mathbb{Y}-\mathbb{X}\beta)^T(\mathbb{Y}-\mathbb{X}\beta)+\lambda_1\sum_{i=1}^{p}|\beta_i|+\lambda_2\beta^T D^T D\ \beta \qquad (3.21)$$

where λ_1 and λ_2 are non-negative regularization parameters.The$\boldsymbol{W}$iseither$\boldsymbol{W_1}$ or $\boldsymbol{W_2}$and $\boldsymbol{W_k} = D_K^T D_K$ for $K = 1,2.$

And

$$D_1 = \begin{pmatrix} 1 & -r_{1,2} & 0 & \cdots & 0 & 0 \\ 0 & 1 & -r_{2,3} & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & -r_{p-1,p} \\ 0 & 0 & 0 & \cdots & 0 & 1 \end{pmatrix}$$

The $\boldsymbol{r}_{i,j}$ is the sample correlation between the predictor $\boldsymbol{x_i}$ and $\boldsymbol{x_j}$.

$$D_2 = \begin{pmatrix} 1 & -r_{1,2} & 0 & 0 & \cdots & 0 & 0 \\ 1 & 0 & -r_{1,3} & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 0 & 0 & \cdots & 0 & -r_{1,p} \\ 0 & 1 & -r_{2,3} & 0 & \cdots & 0 & 0 \\ 0 & 1 & 0 & -r_{2,4} & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 1 & 0 & 0 & \cdots & 0 & -r_{2,p} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 & -r_{p-1,p} \\ 0 & 0 & 0 & 0 & \cdots & 0 & 1 \end{pmatrix}$$

The $\boldsymbol{r}_{i,j}$ is the sample correlation between the predictor variables $\boldsymbol{x_i}$ and $\boldsymbol{x_j}$

The procedure of correlation adjusted elastic net regression when $\lambda_1 \neq 0$ and$\lambda_2 \neq 0$is, we fixed $\lambda_2$first, and then do the $CAEN$ regression to determine the optimal λ_1. Finally, we choose the optimal combination of $\lambda_1$and $\lambda_2$based on the smallest $MSE(\hat{\beta}_{CAEN}$). Due to quadratic regularization, the solution paths of $CAEN$ aremore stable than the solution paths of LASSO regression. So $CAEN$ can also beregarded as a stabilized case of the LASSO regression.

Given the Cholesky's decomposition $\boldsymbol{W} = CC^T$ and for any λ_1,$\lambda_2 > 0$, define

$$\mathbb{X}^* = \frac{1}{\sqrt{1+\lambda_2}}\begin{pmatrix}\mathbb{X}\\ \sqrt{\lambda_2}C^T\end{pmatrix}, \qquad \mathbb{Y}^* = \begin{pmatrix}\mathbb{Y}\\ 0\end{pmatrix}, \;\; \beta^* = \sqrt{1+\lambda_2}\beta \;, \;\; \gamma = \frac{\lambda_1}{\sqrt{1+\lambda_2}}$$

Tan (2012) proved that minimizing

$$LASSO^* = (\mathbb{Y}^* - \mathbb{X}^*\beta^*)^T(\mathbb{Y}^* - \mathbb{X}^*\beta^*) + \gamma\sum_{i=1}^{p}\left|\beta^*{}_i\right|$$

Is equivalent to minimizing

$$(\mathbb{Y} - \mathbb{X}\beta)^T(\mathbb{Y} - \mathbb{X}\beta) + \lambda_1\sum_{i=1}^{p}|\beta_i| + \lambda_2\beta^T W \;\; \beta = CAEN \qquad (3.22)$$

Proof:

$$OLS = (\mathbb{Y}^* - \mathbb{X}^*\beta^*)^T(\mathbb{Y}^* - \mathbb{X}^*\beta^*)$$

$$= [(\mathbb{Y}^*)^T - (\beta^*)^T(\mathbb{X}^*)^T](\mathbb{Y}^* - \mathbb{X}^*\beta^*)$$

$$= (\mathbb{Y}^*)^T\mathbb{Y}^* - (\mathbb{Y}^*)^T\mathbb{X}^*\beta^* - (\beta^*)^T(\mathbb{X}^*)^T\mathbb{Y}^* + (\mathbb{X}^*\beta^*)^T\mathbb{X}^*\beta^*$$

Now,

$$(\beta^*)^T(\mathbb{X}^*)^T\mathbb{Y}^* = (\beta^T\mathbb{X}^T \quad \sqrt{\lambda_2}\beta^T C)\begin{pmatrix}\mathbb{Y}\\ 0\end{pmatrix} = \beta^T\mathbb{X}^T\mathbb{Y}\mathrm{g}$$

$$(\mathbb{Y}^*)^T\mathbb{X}^*\beta^* = (\mathbb{Y}^T \quad 0)\frac{1}{\sqrt{1+\lambda_2}}\begin{pmatrix}\mathbb{X}\\ \sqrt{\lambda_2}C^T\end{pmatrix}\sqrt{1+\lambda_2}\beta = \mathbb{Y}^T\mathbb{X}\beta$$

$$(\mathbb{X}^*\beta^*)^T\mathbb{X}^*\beta^* = \left(\beta^T\mathbb{X}^T \quad \sqrt{\lambda_2}\beta^T C\right)\begin{pmatrix}\mathbb{X}\beta\\ \sqrt{\lambda_2}C^T\beta\end{pmatrix}$$

$$= \beta^T\mathbb{X}^T\mathbb{X}\beta + \lambda_2\beta^T CC^T\beta$$

$$= \beta^T\mathbb{X}^T\mathbb{X}\beta + \lambda_2\beta^T\boldsymbol{W} \;\; \beta$$

Finally,

$$\gamma \sum_{i=1}^{p} |\beta_i| = \frac{\lambda_1}{\sqrt{1+\lambda_2}} \sum_{i=1}^{p} \left|\sqrt{1+\lambda_2}\beta_i\right| = \lambda_1 \sum_{i=1}^{p} |\beta_i|$$

Therefore,

$$OLS = \mathbb{Y}^T\mathbb{Y} - \mathbb{Y}^T\mathbb{X}\beta - \beta^T\mathbb{X}^T\mathbb{Y} + \beta^T\mathbb{X}^T\mathbb{X}\beta + \lambda_2\beta^T \boldsymbol{W} \ \beta + \lambda_1 \sum_{i=1}^{p} |\beta_i|$$

$= (\mathbb{Y} - \mathbb{X}\beta)^T(\mathbb{Y} - \mathbb{X}\beta) + \lambda_2\beta^T \boldsymbol{W} \ \beta + \lambda_1 \sum_{i=1}^{p} |\beta_i|$(3.23)

The CAEN regression is an extension of Elastic Netregression. The behavior of the CAEN regression is similar to Elastic Net regression. The sample correlation is also included in the penaltyterm. After applying argumentation to the data set, the CAEN regression can be reduced to the LASSO regression.

3.7 Mean square error of an estimator

The mean square error (MSE) of an estimator $\hat{\beta}$of a parameter β is defined as

$$MSE(\hat{\beta}) = E[(\hat{\beta} - \beta)^2]$$

$$= Var(\hat{\beta}) + [bias(\hat{\beta})^2] \qquad (3.24)$$

where the $bias(\hat{\beta})$ is given as $E(\hat{\beta} - \beta)^2$.

As we see, $Var(\hat{\beta})$ measures the variability of the estimator and $bias(\hat{\beta})$measures the bias. Therefore, to find a good estimator we need to find the estimator with the smallest mean square error. There is of course a trade-off between $Var(\hat{\beta})$ and $[bias(\hat{\beta})^2]$.This implies that, we can increase a little bias of the estimator in exchange of a large decrease in the variance. After adjustment, the model may be biased a little bit but is more stable, having less variability.

3.8 Choice of tuning parameters

We now discuss how to choose the type and value of the tuning parameter in the Elastic Net. Although we defined the Elastic Net by using (λ_1,λ_2), it is not the only choice as the tuning parameter. In the LASSO, the conventional tuning parameter is the L_1-norm of the coefficients (t) or the fraction of the L_1-norm. We can also use α to parameterize the Elastic Net. The advantage of usingα is that it is always valued within [0, 1]. In algorithm LARS, the LASSO is described as a forward stagewise additive fitting procedure and shown to be (almost) identical to "-L_2 boosting (Efron *et al.*, 2004). This new view adopts the number of steps k of algorithm LARS as a tuning parameter for the LASSO. For each fixedλ_2, the Elastic Net is solved by the algorithm LARS-EN; hence similarly we can use the number of the LARS-EN steps (k) as the second tuning parameter besidesλ_2.

There are well-established methods for choosing such tuning parameters as explained by Hastie *et al.* (2001). If only training data are available, ten-fold cross-validation (CV) is a popular method for estimating the prediction error and comparing different models, and we use it here. Note that there are two tuning parameters in the Elastic Net and CAEN, so we need to cross-validate on a two-dimensional surface. Typically we first pick a (relatively small) grid of values forλ_2, say (0, 0.01, 0.1, 1, 10, and 100). Then, for eachλ_2, algorithm LARS-EN produces the entire solution path of the Elastic Net and CAEN. The other tuning parameter (λ_1) is selected by ten-fold cross validation. The chosen $\lambda_2$is the one giving the smallest cross validation error. For eachλ_2, the computational cost of tenfold cross validation is the same as 10 OLS fits.

3.9 Source of data

The data used in this research comes from a study conducted by (Efron *et al.,* 2004) “Least Angle Regression.". 442 diabetic patients were measured on 10 baseline variables. A full description of the dataset is given in appendix A. Prediction model was desired for the response variable-a measure of disease progression one year after baseline. The data on ten baseline variables including age, sex, Body Mass Index (BMI), Blood Pressure (BP), and six blood serum measurements were obtained for each of the n = 442 diabetic patients, as well as the response of interest, a quantitative measure of disease progression one year after baseline.The statisticians were asked to construct a model that will predict the response y from the covariatesX_1, X_2, . . . , X_{10}. Two hopes were evident here, that the model would produce accurate baseline predictions of response for future patients, and also that the form of the model would suggest which covariates were important factors in disease progression.

3.10 Methods of data analysis

We have discussed that there is no explicit formula for the Mean Square Error whenever the LASSO penalty term is included in $Pen(\beta)$. In this thesis, we used the leave one out cross validation criterion.Cross-validation can be used to assess the predictive quality of the penalizedprediction model or to compare the predictive ability of different values of thetuning parameter.The cross-validation approach is used to assess the validity andpredictability of the regression models constructed, i.e., a certain amount of the data are removed from the model-building process say twentyrecords, and then use the constructed model to estimatetheir computed values. A general rule of thumb in regression model building is to use 80percent of the data set for the development of the training model and the remaining 20percent for validation of the model as noted by (Kutner *et al*. 2004). Validation records can be selectedat random from the

entire data set, or in the case of time series data, the validationset can be the most current 20 percent (Kutner *et al*. 2004). Adequate regression models areexpected to yield estimates reasonably close to the actual data values.R software was used for all the analysis with the aids of *glmnet*, *lars,glmpath,glmnet,survival,splines,penalized,covTest,MASS,Matrix,psych,DAAG,rgl,scatterplot 3d,lasso2,elasticnet,pls,hydroGOF,xtable*packages which are extremely efficient procedures for fitting the entire lasso or elastic-net regularization path for linear regression.

In order to reduce the magnitude of the error by the ordinary least squares (OLS), we standardized the original data. For each observation we subtract its column mean and divide by its column standard deviation.

CHAPTER FOUR

RESULTS AND DISCUSSION

4.1 Introduction

This chapter seeks to compare the performances of LASSO, Elastic Net and CAEN penalized methods using numerical results.

4.2 Ordinary least squares regression

Table 4.1: Results of Ordinary Least Squares

$Variables$	$Estimate$	$Std.Error$	$t-Value$	Pr $(>\|t\|)$	VIF
$INTERCEPT$	-0.0000	0.0334	-0.0000	1.0000	0.0000
AGE	-0.0062	0.0369	-0.1700	0.8670	1.2170
SEX	-0.1481	0.0378	-3.9200	0.0001	1.2780
BMI	0.3211	0.0411	7.8100	0.0000	1.5090
BP	0.2004	0.0404	4.9600	0.0000	1.4590
TC	-0.4893	0.2574	-1.9000	0.0579	59.2030
LDL	0.2945	0.2094	1.4100	0.1604	39.1930
HDL	0.0624	0.1313	0.4800	0.6347	15.4020
TCH	0.1094	0.0990	1.1000	0.2735	8.8910
LTG	0.4641	0.1062	4.3700	0.0000	10.0760
GLU	0.0418	0.0408	1.0200	0.3060	1.4850

Residual standard error: 0.7025 on 431 degrees of freedom

Multiple R-squared: 0.5177, Adjusted R-squared: 0.5066

F-statistic: 46.27 on 10 and 431 DF, p-value :< 2.2e-16.

From table 4.2 above, variables TC,LDL, $HDLTCH$ and LTG all have Variance Inflation Factors (VIF) greater than 5. This tells us that there is a problem of multicollinearity in the data.

We therefore, do the penalized regression on the data which is one of the best methods for remedying multicollinearity problem.

We calculate the Correlation among independent variables and obtain D_1 and D_2which will be used for the calculation of Correlation Adjusted Elastic Net.

$$\begin{pmatrix} 1.000 & 0.174 & 0.185 & 0.335 & 0.260 & 0.219 & -0.080 & 0.200 & 0.271 & 0.302 \\ 0.174 & 1.000 & 0.088 & 0.241 & 0.035 & 0.143 & -0.380 & 0.332 & 0.150 & 0.208 \\ 0.185 & 0.088 & 1.000 & 0.395 & 0.250 & 0.260 & -0.370 & 0.414 & 0.446 & 0.389 \\ 0.335 & 0.241 & 0.395 & 1.000 & 0.242 & 0.186 & -0.180 & 0.258 & 0.393 & 0.390 \\ 0.260 & 0.035 & 0.250 & 0.242 & 1.000 & 0.896 & 0.052 & 0.542 & 0.516 & 0.326 \\ 0.219 & 0.143 & 0.260 & 0.186 & 0.896 & 1.000 & -0.190 & 0.660 & 0.318 & 0.291 \\ -0.080 & -0.380 & -0.370 & -0.180 & 0.052 & -0.190 & 1.000 & -0.740 & -0.390 & -0.270 \\ 0.200 & 0.332 & 0.414 & 0.258 & 0.542 & 0.660 & -0.740 & 1.000 & 0.618 & 0.417 \\ 0.271 & 0.150 & 0.446 & 0.393 & 0.516 & 0.318 & -0.390 & 0.618 & 1.000 & 0.464 \\ 0.302 & 0.208 & 0.389 & 0.390 & 0.326 & 0.291 & -0.270 & 0.417 & 0.464 & 1.000 \end{pmatrix}$$

$$D_1 = \begin{pmatrix} 1.00 & -0.17 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ -0.17 & 1.00 & -0.09 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & -0.09 & 1.00 & -0.39 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & -0.39 & 1.00 & -0.24 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & -0.24 & 1.00 & -0.89 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & -0.89 & 1.00 & 0.19 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.19 & 1.00 & 0.74 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.74 & 1.00 & -0.62 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & -0.62 & 1.00 & -0.46 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & -0.46 & 1.00 \end{pmatrix}$$

$$D_2 = \begin{pmatrix} 1.00 & -0.17 & 0.00 & 0.00 & \cdots & 0.00 & 0.00 \\ 1.00 & 0.00 & -0.19 & 0.00 & \cdots & 0.00 & 0.00 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1.00 & 0.00 & 0.00 & 0.00 & \cdots & 0.00 & -0.30 \\ 0.00 & 1.00 & -0.09 & 0.00 & \cdots & 0.00 & 0.00 \\ 0.00 & 1.00 & 0.00 & -0.24 & \cdots & 0.00 & 0.00 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0.00 & 1.00 & 0.00 & 0.00 & \cdots & 0.00 & -0.21 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0.00 & 0.00 & 0.00 & 0.00 & \cdots & 1.00 & -0.46 \\ 0.00 & 0.00 & 0.00 & 0.00 & \cdots & 0.00 & 1.00 \end{pmatrix}$$

The R package defines the penalized term as

$$P_\lambda(\beta) = \lambda P_\alpha(\beta)$$

$$= \lambda \sum_{i=1}^{p} \left[\frac{1}{2}(1-\alpha)\beta_i^2 + |\beta_i| \right]$$

- $\alpha = 1 \rightarrow lasso\ method$
- $0 < \alpha < 1 \rightarrow elastic\ net\ method$

4.3 LASSO regression

The figure below gives the relationship between$log(lambda)\lambda\ and\ MSE$. The integer numbers at the top of the graph shows the number of non-zero estimators in the model. The left line gives the smallest MSE with 7 variables in the model and the right line gives the smallest standard deviation with only 4 variables in the model. We can therefore choose the model with 7 variables and minimum *MSE*.

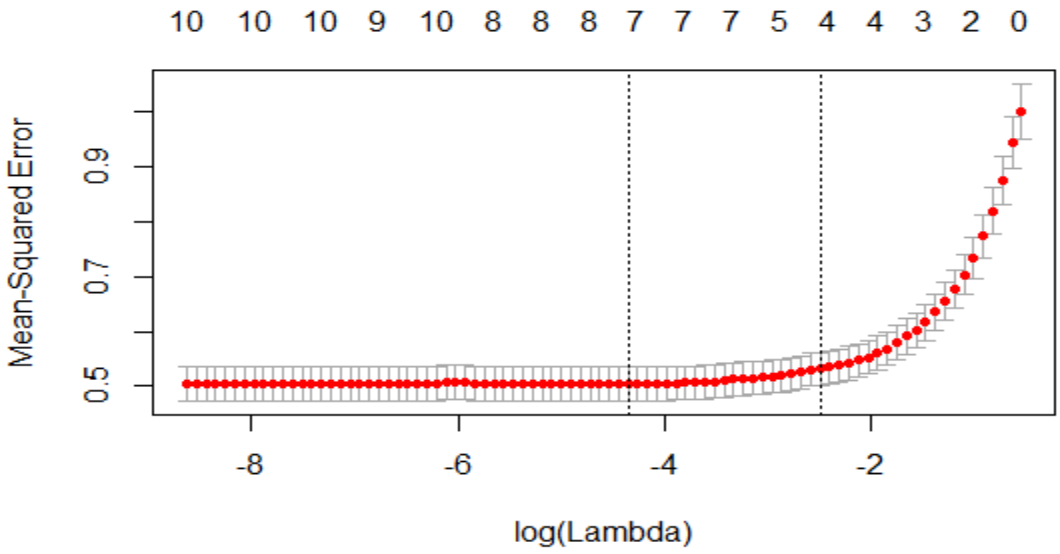

Fig. 4.1: MSE plot and the number of *Variables* in the model as a function of *log* (λ) for the 10-fold cross validation for the LASSO Regression.

Table 4.2: LASSO numerical results

λ	MSE	$Standard\ Error$	$variables\ selected$
0.05857	1.003	0.0504	0
0.10000	0.538	0.0303	4
0.01292	**0.504**	**0.0309**	**7**
0.00614	0.504	0.0312	8
0.00096	0.504	0.0314	9
0.00038	0.505	0.0315	10
0.00018	0.505	0.0315	10

The table above gives the different values of MSE and their respective standard errors at different values of $lambda(\lambda)$. When $lambda(\lambda) = 0.58579$ the value of MSE is 1.003 ,its standard error is 0.0504 and the number of non-zero variables in the model is 0.When $lambda(\lambda) = 0.1000$ the value of MSE is 0.538, its standard error is 0.0303 and the number of

non-zero variables in the model is 4.When $lambda(\lambda) = 0.01292$the value of MSE is 0.504, its standard error is 0.0309 and the number of non-zero variables in the model is 7.When $lambda(\lambda) = 0.00614$the value of MSE is 0.504, its standard error is 0.0312 and the number of non-zero variables in the model is 8.When $lambda(\lambda) = 0.00096$the value of MSE is 0.504, its standard error is 0.0314 and the number of non-zero variables in the model is 9.And When $lambda(\lambda) = 0.00018$the value of MSE is 0.505, its standard error is 0.0315 and the number of non-zero variables in the model is 9. We could see that it is only at When $lambda(\lambda) = 0.01292$ that we have both $MSE\ and$ Standard error are at minimum with 7 non-zero variables included in the model. We shall now use the value of $lambda(\lambda) = 0.01292$ to calculate the penalized regression for $LASSO$. The $LASSO$ penalized regression based on $lambda(\lambda) = 0.01292$ is given below.

Table 4.3: Coefficient Estimates of $LASSO$ regression

$Variable$	$Coefficients$
AGE	0.0000
SEX	−0.1211
BMI	0.3225
BP	0.1830
TC	−0.0630
LDL	0.0000
HDL	−0.1379
TCH	0.0000
LTG	0.3173
GLU	0.0333
MSE	0.5040
$Standard\ Error$	0.0309
$variables\ selected$	7

4.4 Elastic net regression

The elastic net is define by,

$$P_{\alpha}(\beta) = \sum_{i=1}^{p} \left[\frac{1}{2}(1-\alpha)\beta_i^2 + \alpha|\beta_i| \right]$$

Let $\alpha_i = \frac{i}{100} - 0.01, for\ i = 1,2,\dots,101.$ That is $\alpha_1 = 0,\ \alpha_2 = 0.01,\ \alpha_3 0.02,\ \alpha_4 = 0.03,\ \alpha_5 = 0.04, \alpha_6 = 0.05,\ \alpha_7 = 0.06, \alpha_8 = 0.07, \alpha_9 = 0.08, \alpha_{10} = 0.09, \alpha_{11} = 0.1, \alpha_{12} = 0.11, \alpha_{13} = 0.12, \alpha_{14} = 0.13, \alpha_{15} = 0.14, \alpha_{16} = 0.15, \alpha_{17} = 0.16, \dots, \alpha_{101} = 1.$ For each value of the α_i we calculate the elastic net regression and keep the smallest MSE.And finally, we use the value of the minimum MSE among the 101 $MSE's$ to determine the value of α.

Table 4.4: shows the values of $MSE's$ using different values of $alpha(\alpha)$

$\alpha_i\ for\ i = 1,2,\dots,101$	$MSE's$
$\alpha_1 = 0$	0.505135
$\alpha_2 = 0.01$	0.504859
$\alpha_3 = 0.02$	0.504707
$\alpha_4 = 0.03$	0.504863
$\alpha_6 = 0.05$	0.504918
$\alpha_7 = 0.06$	0.504132
$\alpha_9 = 0.08$	0.503872
$\alpha_{10} = 0.09$	0.503756
$\alpha_{11} = 0.10$	0.503636
$\alpha_{12} = 0.11$	0.503473
$\alpha_{13} = 0.12$	0.503501
$\alpha_{14} = 0.13$	0.503469
$\alpha_{15} = 0.14$	0.503446

$\alpha_{16} = 0.15$	0.503410
$\boldsymbol{\alpha_{17} = 0.16}$	**0.503400**
$\alpha_{18} = 0.17$	0.503420
$\alpha_{19} = 0.18$	0.504322
.	.
.	.
.	.
$\alpha_{101} = 1$	0.503750

From the above table it could be observed that the minimum value of MSE is at $alpha(\alpha_{17})$.We will now use the minimum value of $MSE = 0.503400$ at $alpha(\alpha = 0.16)$ to obtain the numerical results for elastic net regression.`

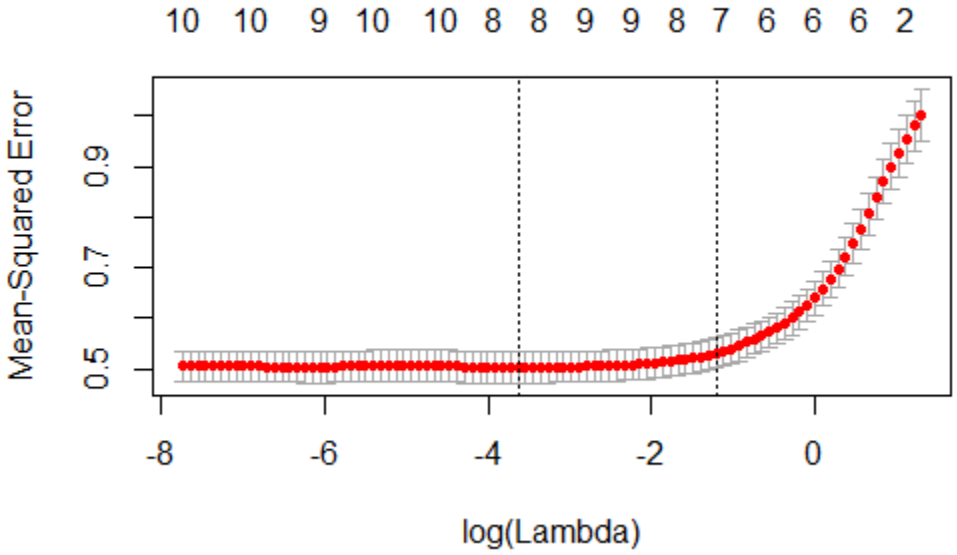

Fig. 4.2: MSE plot and the number of *Variables* in the model as a function of *log* (λ) for the 10-fold cross validation for the Elastic Net Regression.

The figure above gives the relationship between$log\lambda\ and\ MSE$. The integer numbers at the top of the graph shows the number of non-zero estimators in the model. The left line gives the smallest MSE with 8 variables in the model and the right line gives the smallest standard deviation with only 7 variables in the model. We can therefore choose the model with 8 variables and minimum *MSE*.

Table 4.5: Numerical results of elastic net

λ	MSE	$Standard\ Error$	$variables\ selected$
3.6611	1.0024	0.0504	0
1.5848	10.7482	0.0383	6
0.6251	0.5803	0.0313	6
0.0264	**0.5034**	**0.0310**	**8**
0.0037	0.5051	0.0314	10
0.0004	0.5050	0.0315	10

The table above gives the different values of MSE and their respective standard error at different values of $lambda(\lambda)$. When $lambda(\lambda) = 3.6611$ the value of MSE is 1.0024 ,its standard error is 0.0504 and the number of non-zero variables in the model is 0.When $lambda(\lambda) = 1.5848$ the value of $MSE$is 10.7482, its standard error is 0.0383 and the number of non-zero variables in the model is 6.When $lambda(\lambda) = 0.6251$ the value of MSE is 0.5803, its standard error is 0.0313 and the number of non-zero variables in the model is 6.When $lambda(\lambda) = 0.0264$ the value of MSE is 0.5034, its standard error is 0.0310 and the number of non-zero variables in the model is 8.When $lambda(\lambda) = 0.0037$ the value of MSE is 0.5051, its standard error is 0.0314 and the number of non-zero variables in the model is 10.And When $lambda(\lambda) = 0.0004$ the value of MSE is 0.5050, its standard error is 0.0315 and the number

of non-zero variables in the model is 10. We could see that its only at When $lambda(\lambda) = 0.0264$ that we have both $MSE\ and\ Standard\ error$ are at minimum with 8 non-zero variables included in the model.

We shall now use the value of $lambda(\lambda) = 0.0264$ to calculate the penalized regression for $elastic\ net$. The $elastic\ net$ penalized regression based on $lambda(\lambda) = 0.0264$ is given below.

Table 4.6: Numerical results for$ELASTIC\ NET$ regression

$Variable$	$Coefficients$
AGE	0.0000
SEX	−0.1342
BMI	0.3189
BP	0.1907
TC	−0.1010
LDL	0.0000
HDL	−0.1078
TCH	0.0513
LTG	0.3151
GLU	0.0423
MSE	0.5034
$Standard\ Error$	0.0310
$variables\ selected$	8

4.5 Correlation adjusted elastic net regression

Since minimizing

$$LASSO^* = (\mathbb{Y}^* - \mathbb{X}^*\beta^*)^T(\mathbb{Y}^* - \mathbb{X}^*\beta^*) + \gamma \sum_{i=1}^{p} \left|\beta^*_i\right|$$

Is equivalent to minimizing

$$(\mathbb{Y}-\mathbb{X}\beta)^T(\mathbb{Y}-\mathbb{X}\beta)+\lambda_1\sum_{i=1}^{p}|\beta_i|+\lambda_2\beta^T W\ \ \beta=CAEN$$

We can now apply $LASSO$ regression to obtain the numerical results of $CAEN$ using the updated data set.

Let $\lambda_{2,i}=\frac{i}{100}-0.01\ for\ i=1,2,\dots,101$.That is$\lambda_{2,1}=0$,$\lambda_{2,2}=0.01$,$\lambda_{2,3}=0.02\dots\lambda_{2,101}=1$.For each $\lambda_{2,i}$, we update the data set and do the lasso regression to find the optimal MSE and corresponding standard error. Since$CAEN$ does the variable selection.

Table 4.7: shows the values of $MSE's$ using different values of $\lambda_1$andλ_2

$\lambda_{2,i}\ for\ i=1,2\dots,101$	λ_1	MSE
$\lambda_{2,1}=0$	0.0129177	0.50372
$\lambda_{2,2}=0.01$	0.0141542	0.533673
$\boldsymbol{\lambda_{2,3}=0.02}$	**0.014174**	**0.50340**
.	.	.
.	.	.
.	.	.
$\lambda_{2,101}=1$	0.0140633	0.50684

From the above table we could see that when $\lambda_2=0.02$ gives the minimum value of MSEas 0.50340. We will now use the value of $\lambda_1=0.014174$ and $\lambda_2=0.02$ to plot the MSE and obtain the numerical results for correlation adjusted elastic net regression.

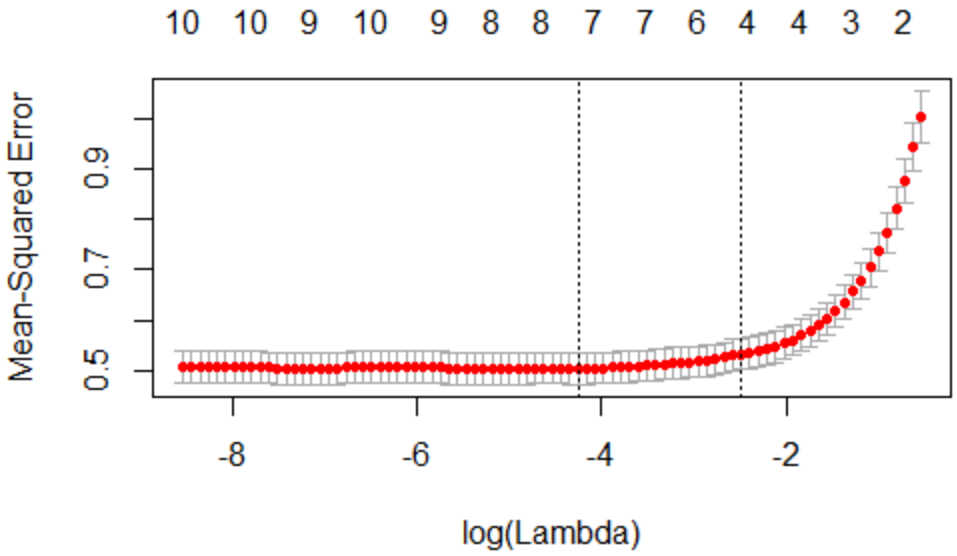

Fig. 4.3: MSE plot and the number of *Variables* in the model as a function of *log* (λ) for the 10-fold cross validation for CAEN Regression.

The figure above gives the relationship between $log\lambda\ and\ MSE$. The integer numbers at the top of the graph shows the number of non-zero estimators in the model. The left line gives the smallest MSE with 7 variables in the model and the right line gives the smallest standard deviation with only 4 variables in the model. We can therefore choose any value of $lambda(\lambda)$between the left line and the right line to obtain the numerical results for correlation adjusted elastic net regression.

Table 4.8:$CAEN$ numerical results

$\lambda_{2,i}\ for\ i = 1,2\ldots,101$	λ_1	MSE	$Std.Error$	$variables\ selected$
$\lambda_{2,1} = 0$	0.0129177	0.50372	0.031090	8
$\lambda_{2,2} = 0.01$	0.0141542	0.533673	0.031046	7
$\boldsymbol{\lambda_{2,3} = 0.02}$	**0.014174**	**0.50340**	**0.03104**	**7**
.	.	.		
.	.	.		
.	.	.		
$\lambda_{2,101} = 1$	0.0140633	0.50684	0.03090	7

Table 4.9: Numerical results for $CORRELATIONADJUSTEDELASTICNET$ regression

$Variable$	$Coefficients$
AGE	0.0000
SEX	−0.8404
BMI	2.2779
BP	1.2857
TC	−0.4265
LDL	0.0000
HDL	−0.9701
TCH	0.0000
LTG	2.2322
GLU	0.2282
MSE	0.50340
$Standard\ Error$	0.03104
$variables\ selected$	7

Table 4.10: Coefficient Comparison of OLS, LASSO, Elastic Net and CAEN Regression

Variable	*OLS*	*LASSO*	*ELASTIC NET*	*CAEN*
AGE	−0.0062	0.0000	0.0000	0.0000
SEX	−0.1481	−0.1211	−0.1342	−0.8404
BMI	0.3211	0.3225	0.3189	2.2779
BP	0.2004	0.1830	0.1907	1.2857
TC	−0.4893	−0.0630	−0.1010	−0.4265
LDL	0.2945	0.0000	0.0000	0.0000
HDL	0.0624	−0.1379	−0.1078	−0.9701
TCH	0.1094	0.0000	0.0513	0.0000
LTG	0.4641	0.3173	0.3151	2.2322
GLU	0.0418	0.0333	0.0423	0.2282
MSE	0.5050	0.5040	0.5034	0.5034
Standard Error	0.03152	0.0309	0.0310	0.03104
variable selected	10	7	8	7

CHAPTER FIVE

SUMMARY, CONCLUSION AND RECOMMENDATION

5.1 Introduction

In this chapter, we present the summary, conclusion and recommendation based on the results obtained from this research work.

5.2 Summary

This research work was aimed at comparing the performances of LASSO, Elastic Net and CAEN regression methods using numerical results. We have applied LASSO, Elastic Net and CAEN methods to the diabetes dataset.From table 4.10 in chapter four it can be seen that the characteristics of each of the methods were observed carefully. The LASSO regression does both the shrinkage and variable selection and there are 7non-zero variables in the final model. The Elastic Net regression also does both the shrinkage and variable selection and there are 8 nonzero variables in the final model. CAEN selects 7 nonzero variables in the final model. According to numerical results, the Elastic Net regression gives a smaller MSE but a larger $Std.Error$, the LASSO regression gives a larger MSE and a smaller $Std.Error$. Also the CAEN gives a smaller MSE but a larger $Std.Error$. According to the dataset CAEN outperforms LASSO and Elastic Net regressions in terms of mean square error and it produced a less complex model than the other two methods.

5.3 Conclusion

To establish an accurate model, one needs to collect numerous variables.Unfortunately, those variables are often highly correlated. As we have discussedin this thesis, those variables that are correlated makes the model less predictive and difficult to interpret. Therefore a penalized

regression method provides a better way of selecting the appropriate variables to establish an effective model as observed in this thesis.

5.4 Recommendations and suggestion for further study

Penalized regression techniques for linear regression have been created in the last few decades to reduce the flaws of ordinary least squares regression with regard to prediction accuracy. Multicollinearity is a problem seldom considered in elementary statistics texts, because it is not really a mathematical-statistical problem, but it is rather a problem in the interpretation of the coefficients. While not extensively considered, however, it is a problem that confronts researchers in actual data analytic situations. Therefore researchers should always be sensitive to the possibility of the problem. And since the multicollinearity diagnostics are so easily obtained, no one should ever report results of regressions with multicollinearity problems. Based on the diabetes dataset The CAEN performs better compared to the other two methods. CAEN can also be applied to survival data, since there are lots of variables in many survival data analysis problems.

5.5 Contribution to Knowledge

1. This research work was able to compare the newly introduced Correlation Adjusted Elastic Net with the existing LASSO and Elastic net regressions. Where CAEN regression outperforms the other two methods in terms of mean square error (MSE) based on thediabetes dataset used.
2. This research has further deepened the discrepancies among the penalized regression methods considered, thereby providing assistance to researchers to ease their decision making as to which technique to be usedwhen encountered with the problem of multicollinearity.

REFERENCES

Adams, J.(1990). A computer experiment to evaluate regression strategies. Proceedings of the Statistical Computing Section. *American Statistical Association*, 3(4): 55-62.

André, N., Young,T.M., and Rials,T.G. (2006).Online monitoring of the buffer capacity of Particleboard furnish by near-infrared spectroscopy,*Applied Spectroscopy*, 60(13):1204-1209.

Ayers,K.L.,andCordell,H.J.(2010).SNPselectioningenome-wideandcandidate gene studies viapenalizedlogisticregression. *GeneticEpidemiology*,34(6):879–891.

Beer, D.G., Kardia, S.L., Huang, C.C., Giordano, T.J., Levin, A.M. (2002). Gene-expression profiles predict survival of patients with lungadenocarcinoma. *National Medical*, 8(2):816–824.

Bondell, H.D. and Reich, B.J. (2008). Simultaneous regression shrinkage,variable selection and clustering of predictors with OSCAR.*Biometrics*,64(5):115-123.

Bøvelstad, H.M., Nygard, S., Storvold, H.L., Aldrin, M., Borgan, O., Frigessi, A.(2007) Predicting survival from microarray data a comparativestudy.*Bioinformatics*, 23(12): 2080–2087.

Breiman, L.(1996). Heuristics of instability and stabilization in model selection. *Annals of Statistics*,24(3):2350-2383.

Breiman, L., Friedman, J., 1997. Predicting multiple responses in multiple linear regression (with discussion). *Journal of the Royal Statistical Society*, SeriesB (59): 3–54.

Buhlmann, P. and VandeGeer, S. (2011). *Statistics for High Dimensional Data*, Springer-verlag.New York, pp. 565-596.

Candes, E and Tao, T (2007). The Dantzig Selector: Statistical Estimation When *p* is much Larger than *n*. *The Annals of Statistics*, 35(6): 2313-2351.

Chen, H.Y., Yu, S.L., Chen, C.H.(2007). A five gene signature and clinical outcome in non–small cell lung cancer.*New England Journal of Medicine*, 356(6): 11–20.

Cho,S.,Kim,K.,Kim,Y.J.,Lee,J.K.(2010).Jointidentificationofmultiplegeneticvariantsvia elastic-netvariable selection inagenome-wideassociationanalysis.*Annals ofHuman Genetics*,74(5): 416–428.

Donoho, D.L., Johnstone, I.M.(1994). Ideal Spatial Adaptation by Wavelet Shrinkage. *Biometrika*,8(3): 90-100.

Donoho. D., Elad, M., and Temlyakov, M. (2004). Stable recovery of sparse overcomplete representations inthe presence of noise.*IEEE transactions on information theory*, 52(1): 200-231.

Doreswamy, V., and Chanabasayya, M.V.(2013):performance analysis of regularized linear regression models.*International Journal of Computational Science and Information Technology*,1(4): 20-33.

Draper, N.R. and H. Smith. (1981). *Applied Regression Analysis*, 2nd Ed. John Wiley and Sons, Inc. New York, pp. 60-80.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R.(2004). Least Angle Regression. The*Annals of Statistics*,32(6)*:* 407-499.

Efroymson, M.A. (1960). *Multiple Regression Analysis*.John Wiley and Sons, Inc. New-York, NY, pp. 30-45.

Fan, J. and Li, R.(2001). Variable selection via non-concave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(10): 1348-1360.

Frank, I. and Friedman, J. (1993). A statistical View of some Chemometrics Regression Tools. *Technometrics*, 35(12): 109-148.

Friedman, J., Hastie, T., and Tibshirani, R. (2007). Pathwise Coordinate Optimization, *Annals of*

Applied Statistics.2(5):302-322.

Fu, W. J. (1998). Penalized Regressions:The Bridge versus the Lasso. *Journal ofComputational and Graphical Statistics*, 7(3):397-416.

Gianola, G.(2013).Priorsinwhole-genomeregression:theBayesianalphabet returns. *Genetics* 194, 573–596.

Goeman, J.J(2011) Multiple testing for exploratory research. *Statistical Science*,26(4): 584-597.

Hans,C.(2010).ModeluncertaintyandvariableselectioninBayesianlasso regression. *Statistical Computing.* 20(3): 221–229.

Harell, J.R, F.E, Lee K.L, and Mark, D.B., (1996).Multivariate Prognostic models, Issues in developing models,evaluating assumptions and adequacy and measuring and reducing errors. *Statistical medical*,15(10): 361-387.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer Series in Statistics. NewYork, pp.400-412.

Hesterberg,T. (2008) Least angle and L_1 penalized regression: a review. *Statistical Survey*,2(3): 61–93.

Hoerl, A.E. and Kennard, R. (1970).Ridge regression: Biased estimation fornonorthogonal problems. *Technometrics*,12(3): 55-67.

Hurvich, C and Tsai, C., (1990). The impact of model selection on inference in linearregression. *American Statistician.* 44: 214-217.

Knight, K. and Fu, W. (2000), Asymptotics for lasso-type estimators, *Annals of Statistics*, 28: 1356-1378.

Kooperberg, C., LeBlanc, M., Obenchain, V., (2010). Risk prediction using genome-wide association studies,*Genetics Epidemiology*, 34: 643 – 652.

Kutner, M. H., Nachtsheim, C. J., Neter, J. and Li, W. (2004). *Applied linear statistical models* (Fifth edition). McGraw-Hill/Irwin, New York. pp.300-321.

Kwon, S., and Yongdai, K.(2012). Large sample properties of the scad-penalizedmaximum likelihood estimationon high dimensions.*Statistica Sinica*,22: 629-653.

Kyung,M.,Gill,J.,Ghosh,M.,andCasella,G.(2010). Penalized regression, standarderrors,and Bayesianlassos. *Bayesian Annals.* 5: 369–412.

Li, Q.,andLin,N (2010).TheBayesianelasticnet. *BayesianAnnals.* 5: 151–170.

Molinaro,A.M.,Simon, R.,and Pfeiffer, R.M.,(2005). Prediction error estimation: a comparison of resamplingmethods.*Bioinformatics*, 21: 3301–3307.

Motyer, J.,Allan**,** C., McKendry**,** S., Galbraith,G.,andSusan, R. W. (2011). LASSO model selection with postprocessingfor a genome-wide association Study data set. *BMC Proceedings,* 5(Suppl 9):S24.

Myers, R.H. (1990). *Classical and Modern Regression with Applications*. PWS-Kent Publishing Company. Boston, pp. 100-120.

Neter, J., Kutner,M.H., Nachtsheim,C.J.,and Wasserman,W. (1996). *Applied Linear Regression Models*. 3rd Ed. Irwin, Inc. Chicago, Illinois,pp.90-98.

Osborne, M.R., Presnell, B. and Turlach, B.A. (2000a). On the LASSO and itsdual. *Journal of Computational and Graphical Statistics.* 9(4): 319-337.

Roecker, E., (1991). Prediction error and its estimation for subset-selection models. *Technometrics.* 33: 459-468.

Ryan, T. (2009). *Modern Regression Methods (Second Edition).*John Wiley & Sons.Hoboken, New Jersey, pp. 50-80.

Shedden,K.,Taylor, J.M, and Enkemann, S.A (2008). Gene expression-based survival prediction in lungadenocarcinoma: a Multi-site, blinded validation study. *National Medical*,14: 822–827.

Simon,R.M.,Subramanian J., Li M.C., Menezes, S.(2011). Using cross-validation to evaluate predictive accuracy ofsurvival risk Classifiers based on high-dimensional data. *Brief. Bioinformatics,* 12,203–214.

Sørlie,T., Perou, C.M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, T.,(2001). Gene expression patterns of breast carcinomas distinguishtumor Subclasses with clinical implications. *Proclamation of National Academic Science. USA*, 98,10869–10874.

Stigler, S.M. (1986). *The History of Statistics: The Measurement of Uncertainty before 1900.* The Belknap Press of Harvard University Press: Cambridge, pp. 140-160.

Sung,Y.J.,Rice,T.K.,Shi,G.,Gu,C.C.,andRao,D.C.(2009).Comparison betweensingle-MarkeranalysisusingMerlinandmulti-markeranalysisusing LASSO forframingham Simulateddata. *BMC Proceedings.* 3(Suppl.7):S27.doi: 10.1186/1753-6561-3-s7-s27

Szymczak, S., Biernacka,J.M., Cordell,H,J., González-Recio,and König,I.R. (2009). "Machine Learning in Genome-Wide Studies." *Epidemiology*,33: 51-56.

Tan, Q. (2012).Correlation Adjusted Penalization In Regression Analysis.PhDThesis, Department of statistics, University of Manitoba.

Tibshirani,R.(1996).Regression shrinkage and selection via the lasso. *Journal of Royal Statistical Society Series* B, 58, 267-288.

Tutz, G. and Ulbricht, J.(2009).Penalized regression with correlation-basedpenalty.*Statistics andComputing,* 19, 239-253.

Turlach, B., Venables, W., Wright, S.(2005). Simultaneous variable selection. *Technometrics* 47, 349–363.

Usai,M.G.,Goddard,M.E.,andHayes,B.J., (2009).LASSOwith cross-validationforgenomic selection. *GeneticResearch.* 91, 427–436.

van de Vijver, M.J., Bergh, J., Piccart, M., Doleronzi, M.(2002). A gene-expression signature as a predictor of survivalin breastcancer. *New England Journal of Medicine*.347,1999–2009.

Vapnik, V.N.(1995).*The nature of statistical learning theory*.Springer N.Y,pp,231-245.

Wahba. G. (1990). Splines models for observational data. *SIAM, CBMS-NFS regional conference in applied mathematics* , v.59

Waldron,L.,Pintilie,M.,Tsao,M.S.,Shepherd, F.A., Huttenhower,C., and Jurisica,I.(2011). Optimizedapplicationofpenalizedregressionmethodsto diversegenomicdata. *Bioinformatics,*27: 3399–3406.

WaldmannP,MészárosG,GredlerB,FuerstCandSölknerJ(2013) Evaluationofthelassoand theelasticnetingenome-wideassociationstudies.*Frontiersin Genetics*. 4:270.

Wigle,D.A., Jurisica, I., and Radulovich, N.,(2002). Molecular profiling of non-small cell lung cancer andcorrelation with disease-free survival. *Cancer Research*, 62, 3005–3008.

Yuan,M. and Lin,Y. (2006) Model selection and estimation in regression with grouped Variables. *Journal of Royal Statistical Society Series B*, 68, 49–67.

Zhao, P. and Yu, B. (2006). On model selection consistency of Lasso. *Journal of Machine Learning Research*(JMLR), 7, 2541-2563.

Zou, H. and Hastie, T. (2005a). The adaptive lasso and its oracle properties.*Journal of the American Statistical Association, 101, 1418-1429.*

Zou, H. and Hastie, T. (2005b). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B,* 67, 301-320.

Zou, H., Hastie, T. and Tibshirani, R. (2007). On the Degrees of Freedom of the Lasso.*The Annals of Statistics,*35, 2173-2192.

Zou, H. and Zhang, H.(2009).On the adaptive elastic-net with a divergingnumber of Parameters.*The Annals of Statistics*, 38, 1149-1173.

APPENDIX A

Diabetes data set of 442 diabetic patients was measured on 10 baseline variables. The response variable Y is a quantitative measure of disease progression one year after baseline.

S/NO	AGE	SEX	BMI	BP	TC	LDL	HDL	TCH	LTG	GLU	Y
1	59	2	32.1	101.00	157	93.2	38.0	4.00	4.8598	87	151
2	48	1	21.6	87.00	183	103.2	70.0	3.00	3.8918	69	75
3	72	2	30.5	93.00	156	93.6	41.0	4.00	4.6728	85	141
4	24	1	25.3	84.00	198	131.4	40.0	5.00	4.8903	89	206
5	50	1	23.0	101.00	192	125.4	52.0	4.00	4.2905	80	135
6	23	1	22.6	89.00	139	64.8	61.0	2.00	4.1897	68	97
7	36	2	22.0	90.00	160	99.6	50.0	3.00	3.9512	82	138
8	66	2	26.2	114.00	255	185.0	56.0	4.55	4.2485	92	63
9	60	2	32.1	83.00	179	119.4	42.0	4.00	4.4773	94	110
10	29	1	30.0	85.00	180	93.4	43.0	4.00	5.3845	88	310
11	22	1	18.6	97.00	114	57.6	46.0	2.00	3.9512	83	101
12	56	2	28.0	85.00	184	144.8	32.0	6.00	3.5835	77	69
13	53	1	23.7	92.00	186	109.2	62.0	3.00	4.3041	81	179
14	50	2	26.2	97.00	186	105.4	49.0	4.00	5.0626	88	185
15	61	1	24.0	91.00	202	115.4	72.0	3.00	4.2905	73	118
16	34	2	24.7	118.00	254	184.2	39.0	7.00	5.0370	81	171
17	47	1	30.3	109.00	207	100.2	70.0	3.00	5.2149	98	166
18	68	2	27.5	111.00	214	147.0	39.0	5.00	4.9416	91	144
19	38	1	25.4	84.00	162	103.0	42.0	4.00	4.4427	87	97
20	41	1	24.7	83.00	187	108.2	60.0	3.00	4.5433	78	168
21	35	1	21.1	82.00	156	87.8	50.0	3.00	4.5109	95	68
22	25	2	24.3	95.00	162	98.6	54.0	3.00	3.8501	87	49
23	25	1	26.0	92.00	187	120.4	56.0	3.00	3.9703	88	68
24	61	2	32.0	103.67	210	85.2	35.0	6.00	6.1070	124	245
25	31	1	29.7	88.00	167	103.4	48.0	4.00	4.3567	78	184
26	30	2	25.2	83.00	178	118.4	34.0	5.00	4.8520	83	202
27	19	1	19.2	87.00	124	54.0	57.0	2.00	4.1744	90	137
28	42	1	31.9	83.00	158	87.6	53.0	3.00	4.4659	101	85
29	63	1	24.4	73.00	160	91.4	48.0	3.00	4.6347	78	131
30	67	2	25.8	113.00	158	54.2	64.0	2.00	5.2933	104	283
31	32	1	30.5	89.00	182	110.6	56.0	3.00	4.3438	89	129
32	42	1	20.3	71.00	161	81.2	66.0	2.00	4.2341	81	59
33	58	2	38.0	103.00	150	107.2	22.0	7.00	4.6444	98	341
34	57	1	21.7	94.00	157	58.0	82.0	2.00	4.4427	92	87
35	53	1	20.5	78.00	147	84.2	52.0	3.00	3.9890	75	65
36	62	2	23.5	80.33	225	112.8	86.0	2.62	4.8752	96	102
37	52	1	28.5	110.00	195	97.2	60.0	3.00	5.2417	85	265
38	46	1	27.4	78.00	171	88.0	58.0	3.00	4.8283	90	276
39	48	2	33.0	123.00	253	163.6	44.0	6.00	5.4250	97	252
40	48	2	27.7	73.00	191	119.4	46.0	4.00	4.8520	92	90

41	50	2	25.6	101.00	229	162.2	43.0	5.00	4.7791	114	100
42	21	1	20.1	63.00	135	69.0	54.0	3.00	4.0943	89	55
43	32	2	25.4	90.33	153	100.4	34.0	4.50	4.5326	83	61
44	54	1	24.2	74.00	204	109.0	82.0	2.00	4.1744	109	92
45	61	2	32.7	97.00	177	118.4	29.0	6.00	4.9972	87	259
46	56	2	23.1	104.00	181	116.4	47.0	4.00	4.4773	79	53
47	33	1	25.3	85.00	155	85.0	51.0	3.00	4.5539	70	190
48	27	1	19.6	78.00	128	68.0	43.0	3.00	4.4427	71	142
49	67	2	22.5	98.00	191	119.2	61.0	3.00	3.9890	86	75
50	37	2	27.7	93.00	180	119.4	30.0	6.00	5.0304	88	142
51	58	1	25.7	99.00	157	91.6	49.0	3.00	4.4067	93	155
52	65	2	27.9	103.00	159	96.8	42.0	4.00	4.6151	86	225
53	34	1	25.5	93.00	218	144.0	57.0	4.00	4.4427	88	59
54	46	1	24.9	115.00	198	129.6	54.0	4.00	4.2767	103	104
55	35	1	28.7	97.00	204	126.8	64.0	3.00	4.1897	93	182
56	37	1	21.8	84.00	184	101.0	73.0	3.00	3.9120	93	128
57	37	1	30.2	87.00	166	96.0	40.0	4.15	5.0106	87	52
58	41	1	20.5	80.00	124	48.8	64.0	2.00	4.0254	75	37
59	60	1	20.4	105.00	198	78.4	99.0	2.00	4.6347	79	170
60	66	2	24.0	98.00	236	146.4	58.0	4.00	5.0626	96	170
61	29	1	26.0	83.00	141	65.2	64.0	2.00	4.0775	83	61
62	37	2	26.8	79.00	157	98.0	28.0	6.00	5.0434	96	144
63	41	2	25.7	83.00	181	106.6	66.0	3.00	3.7377	85	52
64	39	1	22.9	77.00	204	143.2	46.0	4.00	4.3041	74	128
65	67	2	24.0	83.00	143	77.2	49.0	3.00	4.4308	94	71
66	36	2	24.1	112.00	193	125.0	35.0	6.00	5.1059	95	163
67	46	2	24.7	85.00	174	123.2	30.0	6.00	4.6444	96	150
68	60	2	25.0	89.67	185	120.8	46.0	4.02	4.5109	92	97
69	59	2	23.6	83.00	165	100.0	47.0	4.00	4.4998	92	160
70	53	1	22.1	93.00	134	76.2	46.0	3.00	4.0775	96	178
71	48	1	19.9	91.00	189	109.6	69.0	3.00	3.9512	101	48
72	48	1	29.5	131.00	207	132.2	47.0	4.00	4.9345	106	270
73	66	2	26.0	91.00	264	146.6	65.0	4.00	5.5683	87	202
74	52	2	24.5	94.00	217	149.4	48.0	5.00	4.5850	89	111
75	52	2	26.6	111.00	209	126.4	61.0	3.00	4.6821	109	85
76	46	2	23.5	87.00	181	114.8	44.0	4.00	4.7095	98	42
77	40	2	29.0	115.00	97	47.2	35.0	2.77	4.3041	95	170
78	22	1	23.0	73.00	161	97.8	54.0	3.00	3.8286	91	200
79	50	1	21.0	88.00	140	71.8	35.0	4.00	5.1120	71	252
80	20	1	22.9	87.00	191	128.2	53.0	4.00	3.8918	85	113
81	68	1	27.5	107.00	241	149.6	64.0	4.00	4.9200	90	143
82	52	2	24.3	86.00	197	133.6	44.0	5.00	4.5747	91	51
83	44	1	23.1	87.00	213	126.4	77.0	3.00	3.8712	72	52
84	38	1	27.3	81.00	146	81.6	47.0	3.00	4.4659	81	210
85	49	1	22.7	65.33	168	96.2	62.0	2.71	3.8918	60	65
86	61	1	33.0	95.00	182	114.8	54.0	3.00	4.1897	74	141

87	29	2	19.4	83.00	152	105.8	39.0	4.00	3.5835	83	55
88	61	1	25.8	98.00	235	125.8	76.0	3.00	5.1120	82	134
89	34	2	22.6	75.00	166	91.8	60.0	3.00	4.2627	108	42
90	36	1	21.9	89.00	189	105.2	68.0	3.00	4.3694	96	111
91	52	1	24.0	83.00	167	86.6	71.0	2.00	3.8501	94	98
92	61	1	31.2	79.00	235	156.8	47.0	5.00	5.0499	96	164
93	43	1	26.8	123.00	193	102.2	67.0	3.00	4.7791	94	48
94	35	1	20.4	65.00	187	105.6	67.0	2.79	4.2767	78	96
95	27	1	24.8	91.00	189	106.8	69.0	3.00	4.1897	69	90
96	29	1	21.0	71.00	156	97.0	38.0	4.00	4.6540	90	162
97	64	2	27.3	109.00	186	107.6	38.0	5.00	5.3083	99	150
98	41	1	34.6	87.33	205	142.6	41.0	5.00	4.6728	110	279
99	49	2	25.9	91.00	178	106.6	52.0	3.00	4.5747	75	92
100	48	1	20.4	98.00	209	139.4	46.0	5.00	4.7707	78	83
101	53	1	28.0	88.00	233	143.8	58.0	4.00	5.0499	91	128
102	53	2	22.2	113.00	197	115.2	67.0	3.00	4.3041	100	102
103	23	1	29.0	90.00	216	131.4	65.0	3.00	4.5850	91	302
104	65	2	30.2	98.00	219	160.6	40.0	5.00	4.5218	84	198
105	41	1	32.4	94.00	171	104.4	56.0	3.00	3.9703	76	95
106	55	2	23.4	83.00	166	101.6	46.0	4.00	4.5218	96	53
107	22	1	19.3	82.00	156	93.2	52.0	3.00	3.9890	71	134
108	56	1	31.0	78.67	187	141.4	34.0	5.50	4.0604	90	144
109	54	2	30.6	103.33	144	79.8	30.0	4.80	5.1417	101	232
110	59	2	25.5	95.33	190	139.4	35.0	5.43	4.3567	117	81
111	60	2	23.4	88.00	153	89.8	58.0	3.00	3.2581	95	104
112	54	1	26.8	87.00	206	122.0	68.0	3.00	4.3820	80	59
113	25	1	28.3	87.00	193	128.0	49.0	4.00	4.3820	92	246
114	54	2	27.7	113.00	200	128.4	37.0	5.00	5.1533	113	297
115	55	1	36.6	113.00	199	94.4	43.0	4.63	5.7301	97	258
116	40	2	26.5	93.00	236	147.0	37.0	7.00	5.5607	92	229
117	62	2	31.8	115.00	199	128.6	44.0	5.00	4.8828	98	275
118	65	1	24.4	120.00	222	135.6	37.0	6.00	5.5094	124	281
119	33	2	25.4	102.00	206	141.0	39.0	5.00	4.8675	105	179
120	53	1	22.0	94.00	175	88.0	59.0	3.00	4.9416	98	200
121	35	1	26.8	98.00	162	103.6	45.0	4.00	4.2047	86	200
122	66	1	28.0	101.00	195	129.2	40.0	5.00	4.8598	94	173
123	62	2	33.9	101.00	221	156.4	35.0	6.00	4.9972	103	180
124	50	2	29.6	94.33	300	242.4	33.0	9.09	4.8122	109	84
125	47	1	28.6	97.00	164	90.6	56.0	3.00	4.4659	88	121
126	47	2	25.6	94.00	165	74.8	40.0	4.00	5.5255	93	161
127	24	1	20.7	87.00	149	80.6	61.0	2.00	3.6109	78	99
128	58	2	26.2	91.00	217	124.2	71.0	3.00	4.6913	68	109
129	34	1	20.6	87.00	185	112.2	58.0	3.00	4.3041	74	115
130	51	1	27.9	96.00	196	122.2	42.0	5.00	5.0689	120	268
131	31	2	35.3	125.00	187	112.4	48.0	4.00	4.8903	109	274
132	22	1	19.9	75.00	175	108.6	54.0	3.00	4.1271	72	158

133	53	2	24.4	92.00	214	146.0	50.0	4.00	4.4998	97	107
134	37	2	21.4	83.00	128	69.6	49.0	3.00	3.8501	84	83
135	28	1	30.4	85.00	198	115.6	67.0	3.00	4.3438	80	103
136	47	1	31.6	84.00	154	88.0	30.0	5.10	5.1985	105	272
137	23	1	18.8	78.00	145	72.0	63.0	2.00	3.9120	86	85
138	50	1	31.0	123.00	178	105.0	48.0	4.00	4.8283	88	280
139	58	2	36.7	117.00	166	93.8	44.0	4.00	4.9488	109	336
140	55	1	32.1	110.00	164	84.2	42.0	4.00	5.2417	90	281
141	60	2	27.7	107.00	167	114.6	38.0	4.00	4.2767	95	118
142	41	1	30.8	81.00	214	152.0	28.0	7.60	5.1358	123	317
143	60	2	27.5	106.00	229	143.8	51.0	4.00	5.1417	91	235
144	40	1	26.9	92.00	203	119.8	70.0	3.00	4.1897	81	60
145	57	2	30.7	90.00	204	147.8	34.0	6.00	4.7095	93	174
146	37	1	38.3	113.00	165	94.6	53.0	3.00	4.4659	79	259
147	40	2	31.9	95.00	198	135.6	38.0	5.00	4.8040	93	178
148	33	1	35.0	89.00	200	130.4	42.0	4.76	4.9273	101	128
149	32	2	27.8	89.00	216	146.2	55.0	4.00	4.3041	91	96
150	35	2	25.9	81.00	174	102.4	31.0	6.00	5.3132	82	126
151	55	1	32.9	102.00	164	106.2	41.0	4.00	4.4308	89	288
152	49	1	26.0	93.00	183	100.2	64.0	3.00	4.5433	88	88
153	39	2	26.3	115.00	218	158.2	32.0	7.00	4.9345	109	292
154	60	2	22.3	113.00	186	125.8	46.0	4.00	4.2627	94	71
155	67	2	28.3	93.00	204	132.2	49.0	4.00	4.7362	92	197
156	41	2	32.0	109.00	251	170.6	49.0	5.00	5.0562	103	186
157	44	1	25.4	95.00	162	92.6	53.0	3.00	4.4067	83	25
158	48	2	23.3	89.33	212	142.8	46.0	4.61	4.7536	98	84
159	45	1	20.3	74.33	190	126.2	49.0	3.88	4.3041	79	96
160	47	1	30.4	120.00	199	120.0	46.0	4.00	5.1059	87	195
161	46	1	20.6	73.00	172	107.0	51.0	3.00	4.2485	80	53
162	36	2	32.3	115.00	286	199.4	39.0	7.00	5.4723	112	217
163	34	1	29.2	73.00	172	108.2	49.0	4.00	4.3041	91	172
164	53	2	33.1	117.00	183	119.0	48.0	4.00	4.3820	106	131
165	61	1	24.6	101.00	209	106.8	77.0	3.00	4.8363	88	214
166	37	1	20.2	81.00	162	87.8	63.0	3.00	4.0254	88	59
167	33	2	20.8	84.00	125	70.2	46.0	3.00	3.7842	66	70
168	68	1	32.8	105.67	205	116.4	40.0	5.13	5.4931	117	220
169	49	2	31.9	94.00	234	155.8	34.0	7.00	5.3982	122	268
170	48	1	23.9	109.00	232	105.2	37.0	6.00	6.1070	96	152
171	55	2	24.5	84.00	179	105.8	66.0	3.00	3.5835	87	47
172	43	1	22.1	66.00	134	77.2	45.0	3.00	4.0775	80	74
173	60	2	33.0	97.00	217	125.6	45.0	5.00	5.4467	112	295
174	31	2	19.0	93.00	137	73.0	47.0	3.00	4.4427	78	101
175	53	2	27.3	82.00	119	55.0	39.0	3.00	4.8283	93	151
176	67	1	22.8	87.00	166	98.6	52.0	3.00	4.3438	92	127
177	61	2	28.2	106.00	204	132.0	52.0	4.00	4.6052	96	237
178	62	1	28.9	87.33	206	127.2	33.0	6.24	5.4337	99	225

179	60	1	25.6	87.00	207	125.8	69.0	3.00	4.1109	84	81
180	42	1	24.9	91.00	204	141.8	38.0	5.00	4.7958	89	151
181	38	2	26.8	105.00	181	119.2	37.0	5.00	4.8203	91	107
182	62	1	22.4	79.00	222	147.4	59.0	4.00	4.3567	76	64
183	61	2	26.9	111.00	236	172.4	39.0	6.00	4.8122	89	138
184	61	2	23.1	113.00	186	114.4	47.0	4.00	4.8122	105	185
185	53	1	28.6	88.00	171	98.8	41.0	4.00	5.0499	99	265
186	28	2	24.7	97.00	175	99.6	32.0	5.00	5.3799	87	101
187	26	2	30.3	89.00	218	152.2	31.0	7.00	5.1591	82	137
188	30	1	21.3	87.00	134	63.0	63.0	2.00	3.6889	66	143
189	50	1	26.1	109.00	243	160.6	62.0	4.00	4.6250	89	141
190	48	1	20.2	95.00	187	117.4	53.0	4.00	4.4188	85	79
191	51	1	25.2	103.00	176	112.2	37.0	5.00	4.8978	90	292
192	47	2	22.5	82.00	131	66.8	41.0	3.00	4.7536	89	178
193	64	2	23.5	97.00	203	129.0	59.0	3.00	4.3175	77	91
194	51	2	25.9	76.00	240	169.0	39.0	6.00	5.0752	96	116
195	30	1	20.9	104.00	152	83.8	47.0	3.00	4.6634	97	86
196	56	2	28.7	99.00	208	146.4	39.0	5.00	4.7274	97	122
197	42	1	22.1	85.00	213	138.6	60.0	4.00	4.2767	94	72
198	62	2	26.7	115.00	183	124.0	35.0	5.00	4.7875	100	129
199	34	1	31.4	87.00	149	93.8	46.0	3.00	3.8286	77	142
200	60	1	22.2	104.67	221	105.4	60.0	3.68	5.6276	93	90
201	64	1	21.0	92.33	227	146.8	65.0	3.49	4.3307	102	158
202	39	2	21.2	90.00	182	110.4	60.0	3.00	4.0604	98	39
203	71	2	26.5	105.00	281	173.6	55.0	5.00	5.5683	84	196
204	48	2	29.2	110.00	218	151.6	39.0	6.00	4.9200	98	222
205	79	2	27.0	103.00	169	110.8	37.0	5.00	4.6634	110	277
206	40	1	30.7	99.00	177	85.4	50.0	4.00	5.3375	85	99
207	49	2	28.8	92.00	207	140.0	44.0	5.00	4.7449	92	196
208	51	1	30.6	103.00	198	106.6	57.0	3.00	5.1475	100	202
209	57	1	30.1	117.00	202	139.6	42.0	5.00	4.6250	120	155
210	59	2	24.7	114.00	152	104.8	29.0	5.00	4.5109	88	77
211	51	1	27.7	99.00	229	145.6	69.0	3.00	4.2767	77	191
212	74	1	29.8	101.00	171	104.8	50.0	3.00	4.3944	86	70
213	67	1	26.7	105.00	225	135.4	69.0	3.00	4.6347	96	73
214	49	1	19.8	88.00	188	114.8	57.0	3.00	4.3944	93	49
215	57	1	23.3	88.00	155	63.6	78.0	2.00	4.2047	78	65
216	56	2	35.1	123.00	164	95.0	38.0	4.00	5.0434	117	263
217	52	2	29.7	109.00	228	162.8	31.0	8.00	5.1417	103	248
218	69	1	29.3	124.00	223	139.0	54.0	4.00	5.0106	102	296
219	37	1	20.3	83.00	185	124.6	38.0	5.00	4.7185	88	214
220	24	1	22.5	89.00	141	68.0	52.0	3.00	4.6540	84	185
221	55	2	22.7	93.00	154	94.2	53.0	3.00	3.5264	75	78
222	36	1	22.8	87.00	178	116.0	41.0	4.00	4.6540	82	93
223	42	2	24.0	107.00	150	85.0	44.0	3.00	4.6540	96	252
224	21	1	24.2	76.00	147	77.0	53.0	3.00	4.4427	79	150

225	41	1	20.2	62.00	153	89.0	50.0	3.00	4.2485	89	77
226	57	2	29.4	109.00	160	87.6	31.0	5.00	5.3327	92	208
227	20	2	22.1	87.00	171	99.6	58.0	3.00	4.2047	78	77
228	67	2	23.6	111.33	189	105.4	70.0	2.70	4.2195	93	108
229	34	1	25.2	77.00	189	120.6	53.0	4.00	4.3438	79	160
230	41	2	24.9	86.00	192	115.0	61.0	3.00	4.3820	94	53
231	38	2	33.0	78.00	301	215.0	50.0	6.02	5.1930	108	220
232	51	1	23.5	101.00	195	121.0	51.0	4.00	4.7449	94	154
233	52	2	26.4	91.33	218	152.0	39.0	5.59	4.9053	99	259
234	67	1	29.8	80.00	172	93.4	63.0	3.00	4.3567	82	90
235	61	1	30.0	108.00	194	100.0	52.0	3.73	5.3471	105	246
236	67	2	25.0	111.67	146	93.4	33.0	4.42	4.5850	103	124
237	56	1	27.0	105.00	247	160.6	54.0	5.00	5.0876	94	67
238	64	1	20.0	74.67	189	114.8	62.0	3.05	4.1109	91	72
239	58	2	25.5	112.00	163	110.6	29.0	6.00	4.7622	86	257
240	55	1	28.2	91.00	250	140.2	67.0	4.00	5.3660	103	262
241	62	2	33.3	114.00	182	114.0	38.0	5.00	5.0106	96	275
242	57	2	25.6	96.00	200	133.0	52.0	3.85	4.3175	105	177
243	20	2	24.2	88.00	126	72.2	45.0	3.00	3.7842	74	71
244	53	2	22.1	98.00	165	105.2	47.0	4.00	4.1589	81	47
245	32	2	31.4	89.00	153	84.2	56.0	3.00	4.1589	90	187
246	41	1	23.1	86.00	148	78.0	58.0	3.00	4.0943	60	125
247	60	1	23.4	76.67	247	148.0	65.0	3.80	5.1358	77	78
248	26	1	18.8	83.00	191	103.6	69.0	3.00	4.5218	69	51
249	37	1	30.8	112.00	282	197.2	43.0	7.00	5.3423	101	258
250	45	1	32.0	110.00	224	134.2	45.0	5.00	5.4116	93	215
251	67	1	31.6	116.00	179	90.4	41.0	4.00	5.4723	100	303
252	34	2	35.5	120.00	233	146.6	34.0	7.00	5.5683	101	243
253	50	1	31.9	78.33	207	149.2	38.0	5.45	4.5951	84	91
254	71	1	29.5	97.00	227	151.6	45.0	5.00	5.0239	108	150
255	57	2	31.6	117.00	225	107.6	40.0	6.00	5.9584	113	310
256	49	1	20.3	93.00	184	103.0	61.0	3.00	4.6052	93	153
257	35	1	41.3	81.00	168	102.8	37.0	5.00	4.9488	94	346
258	41	2	21.2	102.00	184	100.4	64.0	3.00	4.5850	79	63
259	70	2	24.1	82.33	194	149.2	31.0	6.26	4.2341	105	89
260	52	1	23.0	107.00	179	123.7	42.5	4.21	4.1589	93	50
261	60	1	25.6	78.00	195	95.4	91.0	2.00	3.7612	87	39
262	62	1	22.5	125.00	215	99.0	98.0	2.00	4.4998	95	103
263	44	2	38.2	123.00	201	126.6	44.0	5.00	5.0239	92	308
264	28	2	19.2	81.00	155	94.6	51.0	3.00	3.8501	87	116
265	58	2	29.0	85.00	156	109.2	36.0	4.00	3.9890	86	145
266	39	2	24.0	89.67	190	113.6	52.0	3.65	4.8040	101	74
267	34	2	20.6	98.00	183	92.0	83.0	2.00	3.6889	92	45
268	65	1	26.3	70.00	244	166.2	51.0	5.00	4.8978	98	115
269	66	2	34.6	115.00	204	139.4	36.0	6.00	4.9628	109	264
270	51	1	23.4	87.00	220	108.8	93.0	2.00	4.5109	82	87

271	50	2	29.2	119.00	162	85.2	54.0	3.00	4.7362	95	202
272	59	2	27.2	107.00	158	102.0	39.0	4.00	4.4427	93	127
273	52	1	27.0	78.33	134	73.0	44.0	3.05	4.4427	69	182
274	69	2	24.5	108.00	243	136.4	40.0	6.00	5.8081	100	241
275	53	1	24.1	105.00	184	113.4	46.0	4.00	4.8122	95	66
276	47	2	25.3	98.00	173	105.6	44.0	4.00	4.7622	108	94
277	52	1	28.8	113.00	280	174.0	67.0	4.00	5.2730	86	283
278	39	1	20.9	95.00	150	65.6	68.0	2.00	4.4067	95	64
279	67	2	23.0	70.00	184	128.0	35.0	5.00	4.6540	99	102
280	59	2	24.1	96.00	170	98.6	54.0	3.00	4.4659	85	200
281	51	2	28.1	106.00	202	122.2	55.0	4.00	4.8203	87	265
282	23	2	18.0	78.00	171	96.0	48.0	4.00	4.9053	92	94
283	68	1	25.9	93.00	253	181.2	53.0	5.00	4.5433	98	230
284	44	1	21.5	85.00	157	92.2	55.0	3.00	3.8918	84	181
285	60	2	24.3	103.00	141	86.6	33.0	4.00	4.6728	78	156
286	52	1	24.5	90.00	198	129.0	29.0	7.00	5.2983	86	233
287	38	1	21.3	72.00	165	60.2	88.0	2.00	4.4308	90	60
288	61	1	25.8	90.00	280	195.4	55.0	5.00	4.9972	90	219
289	68	2	24.8	101.00	221	151.4	60.0	4.00	3.8712	87	80
290	28	2	31.5	83.00	228	149.4	38.0	6.00	5.3132	83	68
291	65	2	33.5	102.00	190	126.2	35.0	5.00	4.9698	102	332
292	69	1	28.1	113.00	234	142.8	52.0	4.00	5.2781	77	248
293	51	1	24.3	85.33	153	71.6	71.0	2.15	3.9512	82	84
294	29	1	35.0	98.33	204	142.6	50.0	4.08	4.0431	91	200
295	55	2	23.5	93.00	177	126.8	41.0	4.00	3.8286	83	55
296	34	2	30.0	83.00	185	107.2	53.0	3.00	4.8203	92	85
297	67	1	20.7	83.00	170	99.8	59.0	3.00	4.0254	77	89
298	49	1	25.6	76.00	161	99.8	51.0	3.00	3.9318	78	31
299	55	2	22.9	81.00	123	67.2	41.0	3.00	4.3041	88	129
300	59	2	25.1	90.00	163	101.4	46.0	4.00	4.3567	91	83
301	53	1	33.2	82.67	186	106.8	46.0	4.04	5.1120	102	275
302	48	2	24.1	110.00	209	134.6	58.0	4.00	4.4067	100	65
303	52	1	29.5	104.33	211	132.8	49.0	4.31	4.9836	98	198
304	69	1	29.6	122.00	231	128.4	56.0	4.00	5.4510	86	236
305	60	2	22.8	110.00	245	189.8	39.0	6.00	4.3944	88	253
306	46	2	22.7	83.00	183	125.8	32.0	6.00	4.8363	75	124
307	51	2	26.2	101.00	161	99.6	48.0	3.00	4.2047	88	44
308	67	2	23.5	96.00	207	138.2	42.0	5.00	4.8978	111	172
309	49	1	22.1	85.00	136	63.4	62.0	2.19	3.9703	72	114
310	46	2	26.5	94.00	247	160.2	59.0	4.00	4.9345	111	142
311	47	1	32.4	105.00	188	125.0	46.0	4.09	4.4427	99	109
312	75	1	30.1	78.00	222	154.2	44.0	5.05	4.7791	97	180
313	28	1	24.2	93.00	174	106.4	54.0	3.00	4.2195	84	144
314	65	2	31.3	110.00	213	128.0	47.0	5.00	5.2470	91	163
315	42	1	30.1	91.00	182	114.8	49.0	4.00	4.5109	82	147
316	51	1	24.5	79.00	212	128.6	65.0	3.00	4.5218	91	97

317	53	2	27.7	95.00	190	101.8	41.0	5.00	5.4638	101	220
318	54	1	23.2	110.67	238	162.8	48.0	4.96	4.9127	108	190
319	73	1	27.0	102.00	211	121.0	67.0	3.00	4.7449	99	109
320	54	1	26.8	108.00	176	80.6	67.0	3.00	4.9558	106	191
321	42	1	29.2	93.00	249	174.2	45.0	6.00	5.0039	92	122
322	75	1	31.2	117.67	229	138.8	29.0	7.90	5.7236	106	230
323	55	2	32.1	112.67	207	92.4	25.0	8.28	6.1048	111	242
324	68	2	25.7	109.00	233	112.6	35.0	7.00	6.0568	105	248
325	57	1	26.9	98.00	246	165.2	38.0	7.00	5.3660	96	249
326	48	1	31.4	75.33	242	151.6	38.0	6.37	5.5683	103	192
327	61	2	25.6	85.00	184	116.2	39.0	5.00	4.9698	98	131
328	69	1	37.0	103.00	207	131.4	55.0	4.00	4.6347	90	237
329	38	1	32.6	77.00	168	100.6	47.0	4.00	4.6250	96	78
330	45	2	21.2	94.00	169	96.8	55.0	3.00	4.4543	102	135
331	51	2	29.2	107.00	187	139.0	32.0	6.00	4.3820	95	244
332	71	2	24.0	84.00	138	85.8	39.0	4.00	4.1897	90	199
333	57	1	36.1	117.00	181	108.2	34.0	5.00	5.2679	100	270
334	56	2	25.8	103.00	177	114.4	34.0	5.00	4.9628	99	164
335	32	2	22.0	88.00	137	78.6	48.0	3.00	3.9512	78	72
336	50	1	21.9	91.00	190	111.2	67.0	3.00	4.0775	77	96
337	43	1	34.3	84.00	256	172.6	33.0	8.00	5.5294	104	306
338	54	2	25.2	115.00	181	120.0	39.0	5.00	4.7005	92	91
339	31	1	23.3	85.00	190	130.8	43.0	4.00	4.3944	77	214
340	56	1	25.7	80.00	244	151.6	59.0	4.00	5.1180	95	95
341	44	1	25.1	133.00	182	113.0	55.0	3.00	4.2485	84	216
342	57	2	31.9	111.00	173	116.2	41.0	4.00	4.3694	87	263
343	64	2	28.4	111.00	184	127.0	41.0	4.00	4.3820	97	178
344	43	1	28.1	121.00	192	121.0	60.0	3.00	4.0073	93	113
345	19	1	25.3	83.00	225	156.6	46.0	5.00	4.7185	84	200
346	71	2	26.1	85.00	220	152.4	47.0	5.00	4.6347	91	139
347	50	2	28.0	104.00	282	196.8	44.0	6.00	5.3279	95	139
348	59	2	23.6	73.00	180	107.4	51.0	4.00	4.6821	84	88
349	57	1	24.5	93.00	186	96.6	71.0	3.00	4.5218	91	148
350	49	2	21.0	82.00	119	85.4	23.0	5.00	3.9703	74	88
351	41	2	32.0	126.00	198	104.2	49.0	4.00	5.4116	124	243
352	25	2	22.6	85.00	130	71.0	48.0	3.00	4.0073	81	71
353	52	2	19.7	81.00	152	53.4	82.0	2.00	4.4188	82	77
354	34	1	21.2	84.00	254	113.4	52.0	5.00	6.0936	92	109
355	42	2	30.6	101.00	269	172.2	50.0	5.00	5.4553	106	272
356	28	2	25.5	99.00	162	101.6	46.0	4.00	4.2767	94	60
357	47	2	23.3	90.00	195	125.8	54.0	4.00	4.3307	73	54
358	32	2	31.0	100.00	177	96.2	45.0	4.00	5.1874	77	221
359	43	1	18.5	87.00	163	93.6	61.0	2.67	3.7377	80	90
360	59	2	26.9	104.00	194	126.6	43.0	5.00	4.8040	106	311
361	53	1	28.3	101.00	179	107.0	48.0	4.00	4.7875	101	281
362	60	1	25.7	103.00	158	84.6	64.0	2.00	3.8501	97	182

363	54	2	36.1	115.00	163	98.4	43.0	4.00	4.6821	101	321
364	35	2	24.1	94.67	155	97.4	32.0	4.84	4.8520	94	58
365	49	2	25.8	89.00	182	118.6	39.0	5.00	4.8040	115	262
366	58	1	22.8	91.00	196	118.8	48.0	4.00	4.9836	115	206
367	36	2	39.1	90.00	219	135.8	38.0	6.00	5.4205	103	233
368	46	2	42.2	99.00	211	137.0	44.0	5.00	5.0106	99	242
369	44	2	26.6	99.00	205	109.0	43.0	5.00	5.5797	111	123
370	46	1	29.9	83.00	171	113.0	38.0	4.50	4.5850	98	167
371	54	1	21.0	78.00	188	107.4	70.0	3.00	3.9703	73	63
372	63	2	25.5	109.00	226	103.2	46.0	5.00	5.9506	87	197
373	41	2	24.2	90.00	199	123.6	57.0	4.00	4.5218	86	71
374	28	1	25.4	93.00	141	79.0	49.0	3.00	4.1744	91	168
375	19	1	23.2	75.00	143	70.4	52.0	3.00	4.6347	72	140
376	61	2	26.1	126.00	215	129.8	57.0	4.00	4.9488	96	217
377	48	1	32.7	93.00	276	198.6	43.0	6.42	5.1475	91	121
378	54	2	27.3	100.00	200	144.0	33.0	6.00	4.7449	76	235
379	53	2	26.6	93.00	185	122.4	36.0	5.00	4.8903	82	245
380	48	1	22.8	101.00	110	41.6	56.0	2.00	4.1271	97	40
381	53	1	28.8	111.67	145	87.2	46.0	3.15	4.0775	85	52
382	29	2	18.1	73.00	158	99.0	41.0	4.00	4.4998	78	104
383	62	1	32.0	88.00	172	69.0	38.0	4.00	5.7838	100	132
384	50	2	23.7	92.00	166	97.0	52.0	3.00	4.4427	93	88
385	58	2	23.6	96.00	257	171.0	59.0	4.00	4.9053	82	69
386	55	2	24.6	109.00	143	76.4	51.0	3.00	4.3567	88	219
387	54	1	22.6	90.00	183	104.2	64.0	3.00	4.3041	92	72
388	36	1	27.8	73.00	153	104.4	42.0	4.00	3.4965	73	201
389	63	2	24.1	111.00	184	112.2	44.0	4.00	4.9345	82	110
390	47	2	26.5	70.00	181	104.8	63.0	3.00	4.1897	70	51
391	51	2	32.8	112.00	202	100.6	37.0	5.00	5.7746	109	277
392	42	1	19.9	76.00	146	83.2	55.0	3.00	3.6636	79	63
393	37	2	23.6	94.00	205	138.8	53.0	4.00	4.1897	107	118
394	28	1	22.1	82.00	168	100.6	54.0	3.00	4.2047	86	69
395	58	1	28.1	111.00	198	80.6	31.0	6.00	6.0684	93	273
396	32	1	26.5	86.00	184	101.6	53.0	4.00	4.9904	78	258
397	25	2	23.5	88.00	143	80.8	55.0	3.00	3.5835	83	43
398	63	1	26.0	85.67	155	78.2	46.0	3.37	5.0370	97	198
399	52	1	27.8	85.00	219	136.0	49.0	4.00	5.1358	75	242
400	65	2	28.5	109.00	201	123.0	46.0	4.00	5.0752	96	232
401	42	1	30.6	121.00	176	92.8	69.0	3.00	4.2627	89	175
402	53	1	22.2	78.00	164	81.0	70.0	2.00	4.1744	101	93
403	79	2	23.3	88.00	186	128.4	33.0	6.00	4.8122	102	168
404	43	1	35.4	93.00	185	100.2	44.0	4.00	5.3181	101	275
405	44	1	31.4	115.00	165	97.6	52.0	3.00	4.3438	89	293
406	62	2	37.8	119.00	113	51.0	31.0	4.00	5.0434	84	281
407	33	1	18.9	70.00	162	91.8	59.0	3.00	4.0254	58	72
408	56	1	35.0	79.33	195	140.8	42.0	4.64	4.1109	96	140

409	66	1	21.7	126.00	212	127.8	45.0	4.71	5.2781	101	189
410	34	2	25.3	111.00	230	162.0	39.0	6.00	4.9767	90	181
411	46	2	23.8	97.00	224	139.2	42.0	5.00	5.3660	81	209
412	50	1	31.8	82.00	136	69.2	55.0	2.00	4.0775	85	136
413	69	1	34.3	113.00	200	123.8	54.0	4.00	4.7095	112	261
414	34	1	26.3	87.00	197	120.0	63.0	3.00	4.2485	96	113
415	71	2	27.0	93.33	269	190.2	41.0	6.56	5.2417	93	131
416	47	1	27.2	80.00	208	145.6	38.0	6.00	4.8040	92	174
417	41	1	33.8	123.33	187	127.0	45.0	4.16	4.3175	100	257
418	34	1	33.0	73.00	178	114.6	51.0	3.49	4.1271	92	55
419	51	1	24.1	87.00	261	175.6	69.0	4.00	4.4067	93	84
420	43	1	21.3	79.00	141	78.8	53.0	3.00	3.8286	90	42
421	55	1	23.0	94.67	190	137.6	38.0	5.00	4.2767	106	146
422	59	2	27.9	101.00	218	144.2	38.0	6.00	5.1874	95	212
423	27	2	33.6	110.00	246	156.6	57.0	4.00	5.0876	89	233
424	51	2	22.7	103.00	217	162.4	30.0	7.00	4.8122	80	91
425	49	2	27.4	89.00	177	113.0	37.0	5.00	4.9053	97	111
426	27	1	22.6	71.00	116	43.4	56.0	2.00	4.4188	79	152
427	57	2	23.2	107.33	231	159.4	41.0	5.63	5.0304	112	120
428	39	2	26.9	93.00	136	75.4	48.0	3.00	4.1431	99	67
429	62	2	34.6	120.00	215	129.2	43.0	5.00	5.3660	123	310
430	37	1	23.3	88.00	223	142.0	65.0	3.40	4.3567	82	94
431	46	1	21.1	80.00	205	144.4	42.0	5.00	4.5326	87	183
432	68	2	23.5	101.00	162	85.4	59.0	3.00	4.4773	91	66
433	51	1	31.5	93.00	231	144.0	49.0	4.70	5.2523	117	173
434	41	1	20.8	86.00	223	128.2	83.0	3.00	4.0775	89	72
435	53	1	26.5	97.00	193	122.4	58.0	3.00	4.1431	99	49
436	45	1	24.2	83.00	177	118.4	45.0	4.00	4.2195	82	64
437	33	1	19.5	80.00	171	85.4	75.0	2.00	3.9703	80	48
438	60	2	28.2	112.00	185	113.8	42.0	4.00	4.9836	93	178
439	47	2	24.9	75.00	225	166.0	42.0	5.00	4.4427	102	104
440	60	2	24.9	99.67	162	106.6	43.0	3.77	4.1271	95	132
441	36	1	30.0	95.00	201	125.2	42.0	4.79	5.1299	85	220
442	36	1	19.6	71.00	250	133.2	97.0	3.00	4.5951	92	57