COMPARISM OF THE PENALIZED REGRESSION TECHNIQUES WITH CLASSICAL LEAST SQURES IN MINIMIZING THE EFFECT OF MULTICOLLINEARITY

By

Moses JOHNSON

(P13SCMT8051)

DEPARTMENT OF STATISTICS,
AHMADU BELLO UNIVERSITY, ZARIA
NIGERIA

JUNE, 2018

COMPARISM OF THE PENALIZED REGRESSION TECHNIQUES WITH CLASSICAL LEAST SQURES IN MINIMIZING THE EFFECT OF MULTICOLLINEARITY

By

Moses JOHNSON

B.Sc. Statistics (A.B.U, 2012)
P13SCMT8051

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

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

DEPARTMENT OF STATISTICS,
FACULTY OF PHYSICAL SCIENCE,
AHMADU BELLO UNIVERSITY, ZARIA
NIGERIA

Title Page

JUNE, 2018.

Declaration

I declare that the work in this dissertation titled “COMPARISM OF THE PERFORMANCE OF PENALIZED REGRESSION WITH CLASSICAL LEAST SQUARES IN MINIMIZING THE EFFECT OF MULTICOLLINEARITY.” has been carried out by me in the Department of Statistics under the supervision of Prof. E.O Asiribo and Dr. H.G Dikko. The information derived from the literature has been duly acknowledged in the text and a list of references provided. No part of this dissertation proposal was previously presented for another degree or diploma at this or any other institution.

_____ _____

Johnson Moses Date

Certification

This dissertation titled "COMPARISM OF THE PERFORMANCE OF PENALIZED REGRESSION WITH CLASSICAL LEAST SQUARES IN MINIMIZING THE EFFECT OF MULTICOLLINEARITY" meets the regulations governing the award of the degree of Master of Science of Ahmadu Bello University, and is approved for its contribution to knowledge and literary presentation.

Prof. E.O Asiribo Chairman, Supervisory committee	_____ (Signature)	_____ (Date)
Dr. H. G. Dikko Member, Supervisory committee	_____ (Signature)	_____ (Date)
Dr. H. G. Dikko Head of Department	_____ (Signature)	_____ (Date)
_____ External Examiner	_____ Signature	_____ Date
Prof. S. Z. Abubakar Dean, School of Postgraduate Studies	_____ Signature	_____ Date

Dedication

This study is dedicated to God Almighty for his love and protection over my life and also to my family for their love and encouragement towards the completion of my work.

Acknowledgements

I am grateful to God for seeing me through in my entire endeavor especially for the wisdom and knowledge He blessed me with. I am very grateful to my supervisors, Prof. O.E. Asiribo and Dr. H.G. Dikko for their intense 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 Statistics. Ahmadu Bello University, Zaria, particularly to the Head of Department, Dr. H.G. Dikko, the postgraduate Coordinator in person of Dr. Abubakar Yahaya, the Seminar Coordinator in person Mr. David Reuben for their immense contributions.

A Special thanks to my lovely parents, Mr. Obonyilo Johnson and Comfort Johnson and my brothers, Mr. Sunday Johnson and Linus Johnson for their intense support, fervent prayers and encouragement throughout this programme, may God almighty continue to bless you. My profound gratitude goes to Mr. Emmanuel Oyelowo, My sweetheart Janet Ejembi, Matthew Isaac, for their financial and moral support.

I am very grateful to my friend, Paul John Ogwuche, Mr. Reuben David, Samson Agboola and all my 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

A penalized regression techniques which is a variable selection has been developed specifically to eliminate the problem of multicollinearity and also reduce the flaws inherent in the prediction accuracy of the classical ordinary least squares (OLS) regression technique. In this dissertation, we focus on the numerical study of four penalized regression methods. A diabetes dataset was used to compare four of these well-known techniques, namely: Least Absolute Shrinkage Selection Operator (LASSO), Smoothly Clipped Absolute Deviation (SCAD) and Correlation Adjusted Elastic Net (CAEN) and Elastic Net (EN). The whole paths of results (in λ) for the LASSO, SCAD and CAEN models were calculated using the path wise Cyclic Coordinate Descent (CCD) algorithms– in *glmnet* in R. We used 10-fold cross validation (CV) within *glmnet* to entirely search for the optimal λ. Regularized profile plots 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. Since there are lots of variables in many survival data analysis problems, SCAD can also be applied to survival data. After thorough analysis it was observed that SCAD generates a less complex model with a minimum mean square error (MSE) than the three penalized regression compared namely: Least Absolute Shrinkage Selection Operator (LASSO), Elastic Net (EN) and Correlation Adjusted Elastic Net (CAEN).

Table of Contents

List of Tables

List of Figures

CHAPTER ONE
INTRODUCTION

1.1 Background of the Study

In order to reduce possible biasness, large number of predictor variables was introduced in a model and that lead to a serious concern of multicollinearity among the predictor variables in multiple linear regressions, variable selection is an important issue. (Mathew and Yahaya , 2015)

Multicollinearity and high dimensionality are two problems and computational issue that bring challenges to regression analysis. To deal with these challenges, variables selection and shrinkage estimation are becoming important and useful. The traditional approach of automatic selection (such as forward selection, backward elimination and stepwise selection) and best subset selection are computationally expensive and may not necessarily produce the best model.

Multicollinearity problem is being dealt with by Penalized least square (PLS) method by putting some constraints on the values of the parameters estimated. The aftermath is that the entries of the variance covariance matrix are significantly reduced. When multicollinearity exist that predictor's variables that are highly correlated form some groups. One of the way collinearity problem can be dealt with is to remove one or more of the predictor variables within the same group, by making decision which among the group variables is to be eliminated tend to be difficult and complicated. The aftermath of multicollinearity is th

at the parameter estimator and their variance or standard error tends to be large and prediction may be very inaccurate.

In a situation where there exist correlated data or data where the number of predictors is much larger than the sample size, penalized regression methods have been introduced to deal with this challenge, because they produce more stable results, penalized regression methods do not clearly select the variables; instead they minimize the Regression Sum of Square by using a penalty on

the size of the regression coefficients. This penalty causes the regression coefficients to shrink toward zero and this may result in biased estimates through these regression coefficient estimates will have smaller variance. This can improve the prediction accuracy because of the smaller mean squared error (Hastie *et al.,* 2009). This is why penalized regression methods are also known as shrinkage or regularization methods. Some regression coefficients are set to zero exactly if the shrinkage is large enough, thus, penalized regression methods perform variable selection and coefficient estimation simultaneously. The Least Absolute Shrinkage Selection Operator (LASSO) enables selection such that only the important variable stays in the model (Szymeezak,*et al*., 2009).

1.2 Research Motivation

The motivation for using penalized regression is that the ordinary least square estimation methods is not unique and are subjected to high variability due to the presence of multicollinearity. However, with penalization it becomes unique when appropriate turning parameters are chosen and the variances of the estimators are controlled. Most of the comparisons done by Mathew and Yahaya (2015) were between Least Absolute Shrinkage Selector Operator (LASSO), Elastic Net (EN) and Correlation Adjusted Elastic Net (CAEN). This research attempt to compare LASSO, EN, CAEN and Smoothly Clipped Absolute Deviation (SCAD) regression.

1.3 Statement of the Problem

When multicollinearity exist in a model, Parameter estimates $\left(\widehat{\beta}\right)$ of the multiple linear regression models are not unique. Most often we face the issue of multicollinearity when there are strong linear relationships between two or more predictors. In recent years, alternative methods known as shrinkage and variable selection have been introduced to deal with multicollinearity in

particular, penalized regression methods. This study deal with multicollinearity by considering different penalized regression methods.

1.4 Aim and Objectives of the Study

The aim of this study is to compare the performance of penalized regression techniques with classical regression methods in minimizing the effect of multicollinearity. We intend to achieve this aim through the following objectives:

i. Determine variables that possess multicollinearity using Variance Inflation Factor;
ii. Apply penalized regression techniques such as LASSO, CAEN, EN, and SCAD regression to eliminate multicollinearity; and
iii. Assess the adequacy of the fitted penalized regression models and the classical least squares.

1.5 Significance of the Study.

This study is expected at the end to show the importance of variable selection through Penalized regression as a prior step in removing unimportant factors or variables before model building, also, 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 study revolves around the use of Generalized Cross-Validation (GCV) as a good approximation of the leave-one-out cross-validation (LOOCV) to determine the number of variables selected by each of the methods (LASSO, CAEN, EN and SCAD) under study and also by the use of Mean Square Error and linear fits to determine the predictive accuracy of the methods. The research gives an insight of each of the procedure in an attempt to highlight the similarities and the differences existing between three penalized methods.

CHAPTER TWO
LITERATURE REVIEW

2.1 Introduction

In this section we review some of the earlier model selection criteria. Classical model selection criteria are still useful in modeling that describe adequately on interpretable description of the relationship between variables. Shrinkage methods, also known as regularization methods, were developed in an effort to select models that provide interpretability and accuracy.

2.2 Classical Regression Methods

One of the statistical technique used to relate variables is the regression analysis, its aim is to build a mathematical model that relate dependent variable to independent variables. One of the most commonly used data mining technique is the Multiple Linear Regression (MLR) and it can provide piece of information in cases where the rigid assumptions associated with MLR are met. MLR is a versatile tool that can be applied to almost any process or system. Much work has been done regarding this subject, (Kutner *et al.*, (2004), and Myers (1990)). A key step in developing an appropriate MLR is selection of appropriate Variables. Efromyson (1960) introduced stepwise regression which is commonly used for model building. The steps used to pick the most significant variable from a finite pool of independent variables are the stepwise regression which are three procedures: forward selection, backward selection and mixed selection. The most recommended is the mixed selection, is the combination of forward and backward procedures as stated in Kutner *et al.,* (2004), Neter *et al.,* (1996).Also Draper and Smith (1981),Kutner *et al.,* (2004)stated that the last step in the regression modeling-building process is the model validation. Therefore, it was highlighted therein that, there are three main methods associated with model validation:

i. Determine its predictability by collecting a new data to validate the current model

ii. Assess the current results with empirical, theoretical values and simulation results;

iii. Validate and compare the predictive performance of the current model using cross-validation.

One of the most important tools used to assess the validity and predictability of the regression models constructed is cross-validation approach i.e., from the model building process say twenty records, certain amount of data are removed and then use the constructed model to estimate their computed values. In building regression model a general rule of thumb is to use 80 percent of the data set for the development of the training model and the remaining 20 percent for validation of the model as noted by Kutner *et al.,* (2004). From the entire data set, validation records can be selected at random, or in the case of time series data, the validation set can be the most current 20 percent (Kutner *et al.,* 2004). Suitable regression models are expected to yield estimates to a moderate or acceptable degree close to the actual data values. In determining the predictive power of regression models there are lots of statistics available which can be used. Root Mean Square Error of the Prediction (RMSEP) statistic is the popular statistics for determining this predictability (Andre *et al*., 2006). This statistic (RMSEP) is computed by calculating the square root of the sum square Errors (SSE) for the withheld records divided by the corresponding degrees of freedom. Lower RMSEP values indicate better model predictability. Classical coefficient of determination, or R^2– statistic which is another common model validation statistic indicates the amount of variation explained by the regression model. Breiman and Friedman (1997) stated that, when looking at multiple regression models, it is very importance for the predictors to share strength among different models. Turlach *et al.,* (2005) looked at the problems of selecting a subset of 770 wavelengths that are suitable as predictors for 14 different correlated infra-red spectrometry measurements,

and they also proposed a novel regularization method to perform simultaneous variable selection. Classical regression approaches required that the number of samples exceed the number of variables. The approach may not be applicable in the case of Genome Wide Association (GWA) data. Also, in cases of correlated predictor variables, least-squares estimate of regression coefficient may be highly unstable, which lead to low prediction accuracy. This is common in genomic settings, for instance, where collinear predictors typically outnumbered available samples (p>n), such as the prediction of cancer patient survival from tumor gene expression data (Beer *et al.*, 2002; Shedden *et al*., 2008; Sorlie, 2001; Van de Vijver *et al.,* 2002; and Wigle *et al.,* 2002). In this case, ordinary regression is subject to over fitting and instability of coefficients (Harrel *et al.,* 1996), and stepwise variable selection methods do not perform well (Yuan and Lin, 2006).In penalization methods, regression has been successfully fitted to high-dimensional situation (Hesterberg, 2008), and penalized regression has been shown to outperform univariate and other multivariate regression methods in multiple genomic datasets (Bovelstad *et al.,* 2007).

Roecker (1991), Adams (1990) as well as Hurvich and Tsai (1990) carried out simulation studies and it was suggested that, least-squares estimates can be quite poor. The studies also showed that, often prediction errors using OLS are too small and that the usual 95% confidence interval will only include the true values of parameters in roughly 50% of cases. The prediction errors were shown to become too large, when predictor variables are strongly correlated.

The Means Square Errors (MSE) is the measure of quality of an estimator and also the amount by which the estimated parameter differs from the true value of parameter being estimated.

Although unbiasedness may seem attractive, it does not guarantee the lowest MSE. The lowest MSE is found when a proper tradeoff is made between the bias of the estimator and its variance. It

is commonly observed that introducing a certain amount of bias can lead to substantial reduction of it variance and thus, a reduction in its MSE. This is the key insight to how shrinkage regression can be used to produce better- performing models.

2.3 Penalized Regression

The method of Penalized regression techniques has been introduced to eliminate the problem of multicollinearity and also reduce the flaws inherent in the prediction accuracy of the ordinary least squares (OLS). OLS often does poorly in both prediction and interpretation when some of the predictor variables are collinear.

The ridge regression which estimates the regression coefficients through an L_2 -norm penalized least square criterion was introduced by Hoerl and Kennard (1970). Friedman *et al.,* (2007) stated that the ridge regression shrinks the coefficient of correlated predictor's variables towards each other, allowing them to borrow strength from each other. However, this behavior is not without its problems. For example in case of the *k*-identical predictor variables mentioned above, they each get identical coefficient of size 1/k, which any single one would get if, fitted alone. The ridge penalty is ideal if there are many predictors' variables, and all have non-zero coefficient. Braiman (1996) said that sparse model on the other hand produces best subset selection but it is extremely variable because of its inherent discreteness. Also Frank and Friedman (1993) introduced bridge regression which maximizes the residual sum of squares (RSS) and the estimator from bridge regression is not explicit; however, Frank and Friedman (1993) argued that the optimal choice of the parameter yields reasonable predictors because it controls the degree of preference for the true coefficient to align with the original variable axis direction in the predictor space. Also, Tibshirani (1996) introduced LASSO regressions, which minimizes the RSS subject to a bound on the L_1 -norm of the coefficient. The elastic net was proposed by Zou and Hastie(2005) which is

the combination of both L_1 and L_2 norms. *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. The extension of elastic net regression known as correlation adjusted elastic net regression was introduced by Tan (2012). Bondell and Reich (2008) introduced the OSCAR (Octagonal Shrinkage and Clustering Algorithm for regression). There are lot of penalized regression methods proposed in recent years but this dissertation focuses on Least Absolute Shrinkage Selector Operator (LASSO), Correlation adjusted Elastic Net (CAEN) and Smoothly Clipped Absolute Deviation (SCAD) to ascertain if SCAD outperform the Three Penalized regression.

2.3.1 LASSO Regression

LASSO (least absolute shrinkage and selection operator) is regression analysis methods that perform both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the statistical model it produces.

LASSO estimator which estimates the regression coefficients through an l_1-norm penalized least-squares criterion was introduced by Tibshirani (1996). This is equal to minimizing the sum of squares of the residuals plus an l_1 penalty on the regression coefficients. Due to the nature of the $l_1$penalty, LASSO performs continuous shrinkage and variable selection simultaneously. Also, 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 choice than the l_2 penalization in high dimensional problems, especially when there are lots of redundant noise features (Friedman *et al.,* 2007). The LASSO estimator has two properties which include the nature of the regularization used which leads to sparse solutions and its computational feasibility as shown by Efron *et al.,*(2004) and Friedman *et al.,*(1996). LASSO automatically

leads to model selection when the sparse solutions is obtained. Donoho (1995) proved the near minimax optimality of soft threshold, which is a LASSO shrinkage estimate with orthonormal predictor matrix. Zhao and Yu (2007). proved that a variable selection procedure is consistent if the probability of selecting exactly the set of variables with nonzero coefficients, that is identifying the subset $\{j : \beta_J \neq 0,\ j = 1, ..., p\}$, converges$\rightarrow 1$ and if the probability of point mass at $\beta_j = 0$ is equal to 1 where $j = 1, ..., p$. Knight and Fu (2000) showed that with a fixed p, under certain conditions, LASSO has the model selection consistency. In many situations LASSO has shown success, although, it has some limitations as follows:

1. In cases where p>*T*, LASSO selects at most *T* variables before it saturates, due to the nature of the convex optimization problem. This seems to be a limiting feature for a variable selection method. Moreover, LASSO is not well-defined unless the bound on the l_1-norm of the coefficients is smaller than a certain value.

2. Supposed there exist group of variables among which the pair wise correlations are very high, then LASSO tends to select an arbitrary variable from the group and does not care which one is selected.

3. In situations, where T > p, and if there exist medium or lower correlations among predictors, it has been cleared empirically that the prediction performance of LASSO is dominated by the ridge regression Tibshirani (1996). Limitation (1) and (2) declared by Zou and Hastie (2005) makes LASSO an inappropriate variable selection method in some situations.

Osborn *et al.,* (2000) proposed two algorithms for the computation of the LASSO: a compact descent algorithm was derived to solve the selection problem for a particular value of the tuning parameter, and then a homotopy method for the tuning parameter was develop to completely describe the possible selection. Efron *et al.,* (2004) proposed Least Angle Regression Selection

(LARS) for a model selection algorithm. They showed that, the LARS algorithm implements the LASSO. They also studied an efficient way of selecting the optimal fit and the effective degrees of freedom of the LASSO, where it was discover that, the size of the active set (the indices corresponding to covariates to be chosen) can be used as a measure of the degrees of freedom, which changes, not necessarily monotonically, along the solution paths of LARS.

Kyung *et al.,* (2010) showed from a Bayesian point of view, that the LASSO penalty correspond to a Laplace (double experimental) prior over regression coefficients, which expects many coefficients to be close to zero, and a small subset to be larger and non-zero. They also showed that the two tuning parameters could be estimated within the Gibbs sampler by assigning hyper priors to them. Hans (2010) argued that, 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 limitation of the LASSO, Tibshirani *et al.,* (2005) introduced the fuse LASSO. The fused LASSO penalizes the L_1-norm of both the coefficients and their differences. One important difference existing between LASSO and Ridge regression occurs for the predictor variables with the highest regression coefficients. The L_2 penalty pushes the regression coefficients towards zero with a force proportional to the value of the variables that are most valuable (i.e. that clearly should be in the model where shrinked toward zero but L_1penalty shrinks less (Hesterberg *et. al.,* 2008).

2.3.2 Elastic Net Regression

In the fitting of linear models, the elastic net is a regularization regression method that linearly combines the L_2 and L_2 penalties of the lasso and ridge methods. It was proposed by Zou and Hastie (2005). 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, while with α close to 1 perform much like the LASSO, but removes any degeneracies and odd behaviour caused by high correlations.

Zou and Hastie (2005) highlighted two aspect 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: scientist prefer a simpler model because it put more light on the relationship between the response and covariates. Parsimony is especially an important issue when the number of predictors is large.

Buhlmann and VandeGeer (2011) have showed 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 showed that, the Elastic Net produces higher number of correctly identified influential variables than the LASSO, and has much lower false positive rate than the ridge regression.

2.3.3 Correlation Adjusted Elastic Net (CAEN) Regression

Correlation Adjusted elastic net (CAEN) is another method of penalized regression technique that does variable selection and shrinkage selection. It was introduced by Tan(2012) which is a combination of L_1 penalized regression and Correlation Adjusted Regression (CAR). The behavior of the CAEN regression is similar to that of the 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.

2.3.4 Smoothly Clipped Absolute Deviation (SCAD) Regression

Smoothly Clipped Absolute Deviation (SCAD) penalty for variable selection and efficient estimation was introduced by Fan and Li (2001) $\hat{\beta}$ is said to possess oracle property if there exist a sequence of λ_n such that with $\lambda = \lambda_n$, Here, the oracle property means that the estimator can correctly select the nonzero coefficients with probability converging to one, that is $p_r\left(\hat{\beta} = \hat{\beta}_0\right) \to 1$ and that the estimators of the nonzero coefficients are asymptotically normal with the same means and covariance that they would have if the zero coefficients were known in advance, which works as well as if the correct sub-model was known. We use the SCAD method to achieve simultaneous consistent variable selection and estimation of β . This method uses a specially designed penalty function. Fan and Li (2001) derived the fixed tuning parameter and asymptotic distribution of the estimator, and also showed that the estimator satisfy the oracle property (consistent model selection).

2.4 Application of Penalized Regression

In order to achieve better prediction ability in the face of multicollinearity; penalized regression method 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 Cho *et al*. (2010).

Fu (1998) compared LASSO, Ridge regression and Bridge regression method using prediction performance criteria. He argued that, because of the Bridge operator, the Bridge model does not always perform well in estimation and prediction compared to the other shrinkage methods: the LASSO and Ridge regression.

Efron *et al.,* (2004) used diabetes dataset to compare LARS with Ridge and LASSO where it was found 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) data and found that 169 single Nucleotide polymorphism (SNPS) were needed to explain the variation of the 48 simulated Quantitative Trait Loci (QTLs). They used a rather adhoc cross-validation approach where the highest correlation between genomic breeding values was used as stopping criterion (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) used uncorrelated predictor variables to compare the performance of Elastic Net and LASSO. Ayers and Cordell (2010) compared statistical properties of the LASSO, Ridge and Elastic Net Regression methods on simulated data, and also 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 1 error.

Motyer *et al.,* (2011) considered a penalized regression using the LASSO procedure for a genome-wide associated study (GWAS17) data set and show that post-processing of the penalized-regression results with subsequent stepwise selection may lead to improved identification of casual single-nucleotide polymorphism. 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.

Doreswamy *et al.,* (2013) determined the performance analysis of regularized linear regression models for oxalines and oxazoles derivatives descriptor dataset where it 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 outperform 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 on a real data set from a 50k genome-wide Single Nucleotide polymorphism (SNP) panel of 5570 Fleekvich bulls. It was concluded that, it is important to analyze GWAS data with both LASSO and the Elastic Net whereby an alternative tuning criterion for minimizing MSE is needed for variable selection.

Matthew and Yahaya (2015) used a diabetes data obtained from (Efron *etal* 2004) to compare these well-known penalized regressions namely: Least Absolute Shrinkage Selector Operator (LASSO), Elastic Net and Correlation Adjusted Elastic Net (CAEN), and it was observed that CAEN generated a less complex model. In this Study, we add SCAD penalized regression methods in comparison to existing methods. A better results was obtained that take care of any inherent multicollinearity existing among the variables.

CHAPTER THREE

METHODOLOGY

Introduction

3.1 Penalized Regression Techniques

The following equation shows the general form of the shrinkage and regularization methods for linear models:

$$\hat{\beta} = \arg\min_{\beta}\left(\sum_{i=1}^{N}\left(y_i - \left(\mathrm{X}\beta\right)_i\right)^2\right) \tag{3.1}$$

Subject to $Pen(\beta) \le t$

where $Pen(\boldsymbol{\beta})$ is a specific penalty and a function of $\widehat{\boldsymbol{\beta}} = \left(\widehat{\beta}_0, \widehat{\beta}_1, ..., \widehat{\beta}_p\right)^T$, while t is a tuning parameter. These methods are formulated in the constrained minimization form, where the solution for the vector of regression coefficients, $\hat{\beta}$ is obtained by minimizing the Residual Sum of Square subject to a penalty on the regression coefficient $Pen(\beta)$. The shrinkage (tuning) parameter t determines the amount of shrinkage on the regression coefficients. Note that you do not place a penalty on the size of the regression coefficients if you choose t to be very large, and thus the optimum is the OLS solution. The regression coefficients shrink from the OLS solution toward zero as t decreases,

Much effort had been made 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; Fan and Li, 2001; and Tan, 2012). Large numbers of predictors are included to mitigate modeling biases if the sample size is small. With a large

number of predictors there could be a multicollinearity problem so there is a need to select a smaller subset that only fits the set of full variables, but also contains more important predictors. This led us to the development of least squares (LS) regression methods with various penalties to discover relevant explanatory factors, also to get the higher prediction accuracy in linear regression. According to Friedman *et al* (2010) looked at a linear regression model with n observation on a dependent variable y having p predictors

$$\mathrm{Y}=\mathrm{X}\beta+\varepsilon \qquad (3.2)$$

$$E\left(Y\right)=X\beta \qquad (3.3)$$

where

$\varepsilon_i=\begin{pmatrix}\varepsilon_1\\ \varepsilon_2\\ \vdots\\ \varepsilon_n\end{pmatrix}$ and $\varepsilon_{ii}=\begin{pmatrix}\varepsilon_1\\ \varepsilon_2\\ \vdots\\ \varepsilon_n\end{pmatrix}$ are the error terms for equation (3.2) and (3.3)

Penalized regression and its accompanying variable selection features, can lead to finding smaller groups of variables with good prediction accuracy. Tibshirani(1996)If $p \geq n$, ordinary least squares which minimizes the residual sum of squares can be written as:

$$RSS\left(\beta\right)=\sum_{i=1}^{N}\left(\beta_0+\beta_1\mathrm{X}_{i1}+\cdots+\beta_p\mathrm{X}_{ip}\right) \qquad (3.4)$$

The aim of penalized least squares (PLS) is minimizing sum of squares due to Error (SSE) ;

$$\sum_{i=1}^{n}e_i^2=\hat{\varepsilon}^T\hat{\varepsilon}=\left(Y-X\beta\right)^T\left(Y-X\beta\right)\text{ Subject to }\ pen\left(\beta\right)\leq t \qquad (3.5)$$

Residual sum of squares yields an estimator that is not unique since X is not of full rank. The variance will be artificially large. Hence, Penalized regression can guide us to good subset of predictors.

$$PLS = OLS + penalty = \left(Y - XB\right)^T + \lambda pen\left(\beta\right) \tag{3.6}$$

Where λ is the turning parameter that controls the strength of shrinkage. For example, $\lambda = 0$ implies no penalty is applied and we have ordinaryLeast squares regression. Assuming λ gets larger, more weight is given to the penalty term. The properties of penalization include variable selection and grouping effect. Using Penalization, it is hoped that the variables that are truly statistically significant are selected into the model, and highly correlated predictor variables would be selected or excluded all together.

3.1.1 LASSO Regression Approach

The Least Absolute Shrinkage and Selector Operator estimator was proposed by Tibshirani (1996) as an estimation and variable selection method. It's called L_1 penalized regression. The LASSO is a procedure that minimizes RSS subject to the constraint expressed in term of L_1-norm of the coefficient

The penalty function is given by

$$\hat{\beta}_{LASSO} = \arg\min_{\beta \varepsilon R^p} \left(Y - X\beta\right)^T \left(Y - X\beta\right) + \lambda \sum_{i=1}^{p} \left|\beta_i\right| \tag{3.7}$$

where λ is a nonnegative regularization parameter. $\lambda \sum_{i=1}^{p} \left|\beta_i\right|$ is called "ℓ_1 penalty," which is crucial for the success of the LASSO. It estimates the regression coefficients through an l_1-norm penalizing the least squares criterion. The $MSE\left(\hat{\beta}_{LASSO}\right)$ also increases as the tuning parameter λ increases, while the $Variance\left(\hat{\beta}_{LASSO}\right)$ decreases. For instance, when $\lambda = 0$

$$MSE\left(\hat{\beta}_{LASSO}\right) = trace\left(\text{var}\left(\hat{\beta}_{LASSO}\right)\right) + Bias^T\left(\hat{\beta}_{LASSO}\right) Bias\left(\hat{\beta}_{LASSO}\right)$$

$$= trace\left(\text{var}\left(\hat{\beta}_{LASSO}\right)\right) + 0$$

$$= MSE\left(\hat{\beta}_{OLS}\right)$$

And when $\lambda \to \infty$

$$MSE\left(\hat{\beta}_{LASSO}\right) = trace\left(Var\left(\hat{\beta}_{LASSO}\right)\right) + Bias^{T}\left(\hat{\beta}_{LASSO}\right) Bias\left(\hat{\beta}_{LASSO}\right)$$

$$\to 0 + \left(-\beta\right)^{T}\left(\beta\right) = \beta^{T}\beta \qquad (3.8)$$

Since $Bias^{T}\left(\hat{\beta}_{LASSO}\right) Bias\left(\hat{\beta}_{LASSO}\right)$ and $trace\left(\text{var}\left(\hat{\beta}_{LASSO}\right)\right)$ move to opposite directions as the tuning parameter λ increases, thus, we can choose an optimal value of the parameter λ that minimize $MSE\left(\hat{\beta}_{LASSO}\right)$

3.1.2 Elastic Net Regression Approach

Elastic Net was proposed by Zou and Hastie (2005) which is based on a combined penalties of LASSO and Ridge regression to improve the prediction performance of the naïve elastic net by correcting this double-shrinkage. Suppose a data set has n observations with p predictors. Let $y = \left(y_1, y_2, \ldots, y_n\right)^{T}$ be the response and $\mathrm{X} = \left|\mathrm{X}_1\right|, \ldots, \left|\mathrm{X}_p\right|$ be the model matrix, where $\mathrm{X}_j = \left(x_{11}, \ldots, x_{np}\right)$, for $j = 1, \ldots, p$ are 22 predictors. After a location and scale transformation, it can be assumed that the response is centered and the predictors are standardized, such that:

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

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

$$L(\lambda_1, \lambda_2, \beta) = |y - \mathrm{X}\beta|^2 + \lambda_2 |\beta|^2 + \lambda_1 |\beta|_1 \tag{3.9}$$

where

$$|\beta|^2 = \sum_{j=1}^{p} \beta_j^2 \text{ and } |\beta|_1 = \sum_{j=1}^{p} |\beta_j|$$

The naïve elastic net estimator is the minimizer of (3.8)

$$\text{i:e } \hat{\beta} = \arg\min_{\beta} L(\lambda_1, \lambda_2, \beta) \tag{3.10}$$

The above procedure can be viewed as a penalized least –squares method. Suppose $\alpha = \frac{\lambda_1}{\lambda_1 + \lambda_2}$, then solving for β in (3.8) is equivalent to solving the optimization problem:

3.1.3 Correlation Adjusted Elastic Net Approach

Correlation Adjusted elastic net (CAEN) was introduced by Tan(2012) which is a combination of L_1 penalized regression and Correlation Adjusted Regression (CAR). It is also an extension of Elastic Net regression. The Correlation Adjusted Elastic Net Regression can be defined as:

$$\hat{\beta}_{CAEN} = \arg\min_{\beta \varepsilon \ p} (Y - X\beta)^T (Y - X\beta) + \lambda_1 \sum_{i=1}^{p} |\beta_i| + \lambda_2 \beta^T W \beta \tag{3.11}$$

$$= \arg\min_{\beta \varepsilon \ p} (Y - X\beta)^T (Y - X\beta) + \lambda_1 \sum_{i=1}^{p} |\beta_i| + \lambda_2 \beta^T D^T D \beta \tag{3.12}$$

where λ_1 and λ_2 are non-negative regularized parameters. The W is either W_1 or W_2 and $W_k = D_k^T D_K$ for K=1,2.

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

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

The $r_{i,j}$ is the sample correlation between the predictor variables x_i and x_j, W_k is a matrix

The procedure of correlation adjusted elastics net regression when $\lambda_1 \neq 0$ and $\lambda_2 \neq 0$, we fixed λ_2 first, then we do the CAEN Regression to determine the optimal λ_1. Finally, we choose the optimal combination of λ_1 and λ_2 based on the smallest $MSE\left(\hat{\beta}_{CAEN}\right)$. Due to quadratic regularization, the solution paths of CAEN are more stable than the solution paths of LASSO regression. Therefore CAEN can also be regarded as a stabilized case of the LASSO Regression.

3.1.4 SCAD Regression Approach

LASSO is not asymptotically consistent and so the LASSO can be biased for large coefficients. Fan and Li (2001) addressed this problem and proposed the SCAD penalty function. They described the conditions of a good penalty function: (a) unbiasedness – the resulting estimator is

nearly unbiased when the true unknown parameter is large; (b) sparsity – the resulting estimator is a threshold rule, which automatically sets small estimated coefficients to be zero; and (c) continuity – the resulting estimator is continuous in the data.Least Squares Estimator (LSE) with the SCAD penalty function minimizes the criterion function.

$$\hat{\beta}_{SCAD} = \sum_{i=1}^{n}\left(y_i - \left(\alpha - x_i^T\beta\right)\right) + \sum_{j=1}^{d} P_\lambda\left|\beta_j\right| \quad (3.12)$$

where $P_\lambda\left|\beta_j\right|$ is the SCAD function defined as

$$P_\lambda(\beta) = \begin{cases} \lambda|\beta|, 0 \le |\beta| < \lambda \\ -\dfrac{\left(a^2-1\right)\lambda^2 - \left(|\beta| - a\lambda\right)^2}{2\left(a-1\right)}, \lambda \le |\beta| < a\lambda \\ \dfrac{1}{2}\left(a+1\right)\lambda^2, |\beta| > a\lambda \end{cases} \quad (3.13)$$

Where a can be chosen using cross-validation or generalized cross-validation.

But different from the LASSO penalty, the SCAD penalizes large coefficients equally while the LASSO penalty increases linearly as the magnitude of the coefficient increases. In this way, the SCAD results can be unbiased penalized estimators for large coefficients. It is a method with good variable selection performance (i.e. it selects the signal variables and few or none of the noise variables),. Fan *et al.,* (2012) defined their SCAD variance estimator as:

$$\hat{\sigma}^2_{SCAD} = \frac{1}{n - \hat{s}_{\hat{\lambda}}}\left\|Y - X\hat{\beta}_{SCAD,\hat{\lambda}}\right\|_2^2 \quad (3.14)$$

where $\hat{\beta}_{SCAD,\hat{\lambda}}$ is the SCAD estimator of β at the regularization parameter $\hat{\lambda}$ selected by cross-validation Consistency and asymptotic normality can be shown for this estimator with an appropriately chosen, deterministic regularization parameter sequence λ_n

3.2 Ordinary Least Squares

Ordinary least squares (*OLS*) is one of classical techniques used to estimate the parameter of the multiple regression model given by Dismuke *etal* (2006), as

$$\hat{\beta}_{ols} = \left(X^T X\right)^{-1} X^T Y \tag{3.15}$$

3.3 Assumptions of Multiple Linear Regression

1. Linearity: when there is a relationship between the explanatory variables and the response variable, it known as linear. This restriction is only to parameters but not explanatory variables, since the explanatory variables are known to be fixed values. That is,

- $\mathrm{E}\left(y_i / x_{i1}, x_{2p}, \ldots x_{1p}\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i = x_i^T \beta$
- $\dfrac{\partial \mathrm{E}\left(y_i / x_i\right)}{\partial x_i} = \beta$

2. Independence: There are two types of independence.
    - Each combination of explanatory variable and error is independent
    - The error terms are independent. for all $i \neq j$

3. Normality: The error terms follow normal distribution.
    - $\varepsilon_i = N\left(0, \sigma_i^2\right)$
    - $Y = N\left(X\beta, \sigma^2\right)$

    where

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

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

- $Var(\varepsilon_i) = Var(\varepsilon_j) = \sigma^2$ for all $i \neq j$
- $Var(y_i) = Var(y_j) = \sigma^2$ for all $i \neq j$

3.4. Variance Inflation Factor

A measure of the amount of multicollinearity in a set of multiple regression variables. The presence of multicollinearity within the set of independent variables can cause a number of problems in the understanding the significance of individual independent variables in the regression model. Using variance inflation factors helps to identify multicollinearity issues so that the model can be adjusted. We can calculate k different VIFs (one for each X_i) in three steps:

Step one

First we run an ordinary least square regression that has X_i as a function of all the other explanatory variables in the first equation.

If $i = 1$, for example, the equation would be where c_0 is a constant and e is the error term

Step two

Then, calculate the VIF factor for $\hat{\beta}_i$ with the following formula:

$$\text{VIF}_i = \frac{1}{1 - R_i^2} \tag{3.16}$$

where R_i^2 is the coefficient of determination of the regression equation

Step three

Analyze the magnitude of multicollinearity by considering the size of the $VIF\left(\hat{\beta}_i\right)$. A rule of thumb is that if $VIF\left(\hat{\beta}_i\right) > 5$ then multicollinearity is high. (O'brien *et al.,* 2007)

3.5 Mean Square Error

The mean square error of an estimator $\hat{\beta}$ of a parameter β assesses the quality of an estimator in terms of its variation and unbiasedness is defined as

$$MSE\left(\hat{\beta}\right) = E\left[\left(\hat{\beta}-\beta\right)^2\right] = Var\left(\hat{\beta}\right) + \left[bias\left(\hat{\beta}\right)^2\right] \quad (3.17)$$

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

$Var\left(\hat{\beta}\right)$ Measures the variability of the estimator and $bia\left(\hat{\beta}\right)$ measures the bias. Therefore, to find a good estimator we need to find the estimator with the smallest mean square error. There is a course of trade-off between $Var\left(\hat{\beta}\right)$ and $\left[bias\left(\hat{\beta}\right)^2\right]$. 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.

3.6 Choice of turning Parameters

In the LASSO, the conventional tuning parameter is the L_1-norm of the coefficients (t) (Tibshirani, 1996). We can also use α to parameterize the Elastic Net. The advantage of using α is that it is always valued within [0, 1]. In algorithm LARS, the LASSO is described as a forward stage wise additive fitting procedure and showed 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 fixed λ_2 , the SCAD is solved by the algorithm LARS-EN; similarly, we can use the number of the LARS-EN steps (k) as the second tuning sides λ_2 .

There are well-established methods for choosing such tuning parameter as explained by Hastie *et al.*, (2009). If only training data are available, ten- fold cross-validation (CV) is a popular method for estimating the prediction error and comparing different models, therefore we are using it here. There are two parameters in CAEN, and EN so we need cross validation on two-dimensional

surface. We first, pick a (relatively small) grid of values for λ_2 , say (0,0.01,0.1, 1, 10 and 100). Then, for each λ_2 , algorithm LARS-EN produces the entire solution path of the CAEN and EN. The other tuning parameter (λ_1) is selected by ten-fold cross validation. The chosen λ_2 is the one giving the smallest cross validation error.

3.7 Source of data

The study used data obtained from a study conducted by (Efron *et al.,* 2004). 442 diabetes patients were measured on 10 baseline variables. 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 $n = 442$ diabetes patients; the interest here is providing a quantitative measure of disease progression one year after baseline.

CHAPTER FOUR

RESULTS AND DISCUSSION

4.1 Introduction

This chapter seeks to compare the performances of LASSO, CAEN, EN and SCAD penalized regression methods using numerical results. R software was used.

In order to reduce the magnitude of the error by the ordinary least squares (OLS), we standardized the original data. First, the mean is subtracted from the value of each case, resulting in a mean of zero. Then the difference of the individual's score and the mean is divided by the standard deviation, which result in standard deviation of one.

4.2 Determining the Ordinary least squares regression.

Table 4.1 Results of OLS.

$Variables$	$Estimate$	$t-Value$	$\Pr(>\|t\|)$	VIF	R^2
$INTERCEPT$	0.0000	−0.0000	1.0000	0.0000	
AGE	−0.0062	−0.1700	0.8670	1.2170	0.035
SEX	−0.1481	−3.9200	0.0001	1.2780	0.002
BMI	0.3211	7.8100	0.0000	1.5090	0.344
BP	0.2004	4.9600	0.0000	1.4590	0.195
TC	−0.4893	−1.9000	0.0579	59.2030	0.045
LDL	0.2945	1.4100	0.1604	39.1930	0.030
HDL	0.0624	0.4800	0.6347	15.4020	0.156
TCH	0.1094	1.1000	0.2735	8.8910	0.185
LTG	0.4641	4.3700	0.0000	10.0760	0.320
GLU	0.0418	1.0200	0.3060	1.4850	0.146

Residual standard error: 0.7025 on 431 degrees of freedom, $n = 442$

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 4.1 above, variables TC, LDL, HDL, TCH 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.

Table 4.1.1 Optimal Result of OLS

$Variables$	$Estimate$	$t-Value$	Pr $(>\|t\|)$
$INTERCEPT$	0.0000	-0.0000	1.0000
SEX	-0.1399	-3.748	0.0001
BMI	0.3321	8.216	0.0000
BP	0.2028	5.221	0.0000
TC	-0.08440	-2.024	0.0436
HDL	-0.1486	-3.454	0.0006
LTG	0.3432	7.234	0.0000

Residual standard error: 0.7025 on 435 degrees of freedom Multiple R-squared: 0.5132, Adjusted R-squared: 0.5065 F-statistic: 76.44 on 6 and 435 DF, p-value: < 2.2e-16

From table 4.1 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.

4.3 Determining the Correlation among independent variables

Table 4.2: Correlation Matrix

$$\begin{pmatrix} 1.00 & 0.174 & 0.185 & 0.335 & 0.260 & 0.219 & -0.080 & 0.200 & 0.271 & 0.302 \\ 0.174 & 1.000 & 0.088 & 0.241 & 0.035 & 0.143 & -0.380 & 0.332 & 0.150 & 0.208 \\ 0.185 & 0.088 & 1.000 & 0.395 & 0.250 & 0.260 & -0.370 & 0.414 & 0.446 & 0.389 \\ 0.335 & 0.241 & 0.395 & 1.000 & 0.242 & 0.186 & -0.180 & 0.258 & 0.393 & 0.390 \\ 0.260 & 0.035 & 0.250 & 0.242 & 1.000 & 0.896 & 0.052 & 0.542 & 0.516 & 0.326 \\ 0.219 & 0.143 & 0.260 & 0.186 & 0.896 & 1.000 & -0.190 & 0.660 & 0.318 & 0.291 \\ -0.080 & -0.380 & -0.370 & -0.180 & 0.052 & -0.190 & 1.000 & -0.740 & -0.390 & -0.270 \\ 0.200 & 0.332 & 0.414 & 0.258 & 0.542 & 0.660 & -0.740 & 1.000 & 0.618 & 0.417 \\ 0.271 & 0.150 & 0.446 & 0.393 & 0.516 & 0.318 & -0.390 & 0.618 & 1.000 & 0.464 \\ 0.302 & 0.208 & 0.389 & 0.390 & 0.226 & 0.291 & -0.270 & 0.417 & 0.464 & 1.000 \end{pmatrix}$$

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

$$D_2 = \begin{pmatrix}
1.00 & -0.17 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\
1.00 & 0.00 & -0.19 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\
1.00 & 0.00 & 0.00 & -0.34 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\
1.00 & 0.00 & 0.00 & 0.00 & -0.26 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\
1.00 & 0.00 & 0.00 & 0.00 & 0.00 & -0.22 & 0.00 & 0.00 & 0.00 & 0.00 \\
1.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.80 & 0.00 & 0.00 & 0.00 \\
1.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & -0.20 & .00 & 0.00 \\
1.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & -0.27 & 0.00 \\
1.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & -0.30 \\
0.00 & 1.00 & -0.09 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\
0.00 & 1.00 & 0.00 & -0.24 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\
0.00 & 1.00 & 0.00 & 0.00 & -0.04 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\
0.00 & 1.00 & 0.00 & 0.00 & 0.00 & -0.14 & 0.00 & 0.00 & 0.00 & 0.00 \\
0.00 & 1.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.38 & 0.00 & 0.00 & 0.00 \\
0.00 & 1.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & -0.33 & 0.00 & 0.00 \\
0.00 & 1.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & -0.15 & 0.00 \\
0.00 & 1.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.21 \\
0.00 & 0.00 & 1.00 & -0.310 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\
0.00 & 0.00 & 1.00 & 0.00 & -0.25 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\
0.00 & 0.00 & 1.00 & 0.00 & 0.00 & -0.26 & 0.00 & 0.00 & 0.00 & 0.00 \\
0.00 & 0.00 & 1.00 & 0.00 & 0.00 & 0.00 & 0.37 & 0.00 & 0.00 & 0.00 \\
0.00 & 0.00 & 1.00 & 0.00 & 0.00 & 0.00 & 0.00 & -0.41 & 0.00 & 0.00 \\
0.00 & 0.00 & 1.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & -0.45 & 0.00 \\
0.00 & 0.00 & 1.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & -0.39 \\
0.00 & 0.00 & 0.00 & 1.00 & -0.24 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\
0.00 & 0.00 & 0.00 & 1.00 & 0.00 & -0.19 & 0.00 & 0.00 & 0.00 & 0.00 \\
0.00 & 0.00 & 0.00 & 1.00 & 0.00 & 0.00 & 0.18 & 0.00 & 0.00 & 0.00 \\
0.00 & 0.00 & 0.00 & 1.00 & 0.00 & 0.00 & 0.00 & -0.26 & 0.00 & 0.00 \\
0.00 & 0.00 & 0.00 & 1.00 & 0.00 & 0.00 & 0.00 & 0.00 & -0.39 & 0.00 \\
0.00 & 0.00 & 0.00 & 1.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & -0.39 \\
0.00 & 0.00 & 0.00 & 0.00 & 1.00 & -0.810 & 0.00 & 0.00 & 0.00 & 0.00 \\
0.00 & 0.00 & 0.00 & 0.00 & 1.00 & 0.00 & -0.05 & 0.00 & 0.00 & 0.00 \\
0.00 & 0.00 & 0.00 & 0.00 & 1.00 & 0.00 & 0.00 & -0.54 & 0.00 & 0.00 \\
0.00 & 0.00 & 0.00 & 0.00 & 1.00 & 0.00 & 0.00 & 0.00 & 0.52 & 0.00 \\
0.00 & 0.00 & 0.00 & 0.00 & 1.00 & 0.00 & 0.00 & 0.00 & 0.00 & -0.33 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 1.00 & 0.19 & 0.00 & 0.00 & 0.00 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 1.00 & 0.00 & -0.66 & 0.00 & 0.00 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 1.00 & 0.00 & 0.00 & -0.32 & 0.00 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 1.00 & 0.00 & 0.00 & 0.00 & -0.29 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 1.00 & 0.74 & 0.00 & 0.00 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 1.00 & 0.00 & 0.39 & 0.00 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 1.00 & 0.00 & 0.00 & 0.27 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 1.00 & -0.62 & 0.00 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 1.00 & 0.00 & -0.42 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 1.00 & -0.46 \\
0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 1.00
\end{pmatrix}$$

D_1 and D_2 are matrices of sample correlation between the predictor variables and the graphs are obtain using R Package. Df is the Number of variables selected

The R package defines the penalized term as

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

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

$\alpha = 1 \rightarrow lasso\ method$

$0 < \alpha < 1 \rightarrow elastic\ net\ method$

Table 4.3: LASSO numerical results

λ	MSE	Df
0.05857	1.003	0
0.10000	0.538	4
0.01292	**0.504**	**7**
0.00614	0.5041	8
0.00096	0.5043	9
0.00038	0.5050	10
0.00018	0.5051	10

Table 4.3 gives the different values of MSE at different values of $lambda(\lambda)$. For example when $(\lambda) = 0.05857$ the value of MSE is 1.003 and the number of non-zero variables in the model is 0.When $(\lambda) = 0.1000$ the value of MSE is 0.538, etc. We could see that it is only at $(\lambda) = 0.01292$ that we have MSE at minimum with 7 non-zero variables included in the model. We

shall now use the value of $(\lambda) = 0.01292$ to calculate the penalized regression for $LASSO$. As given in table 4.3.

Table 4.3.1: 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
Df	7

4.4: Results Based LASSO regression

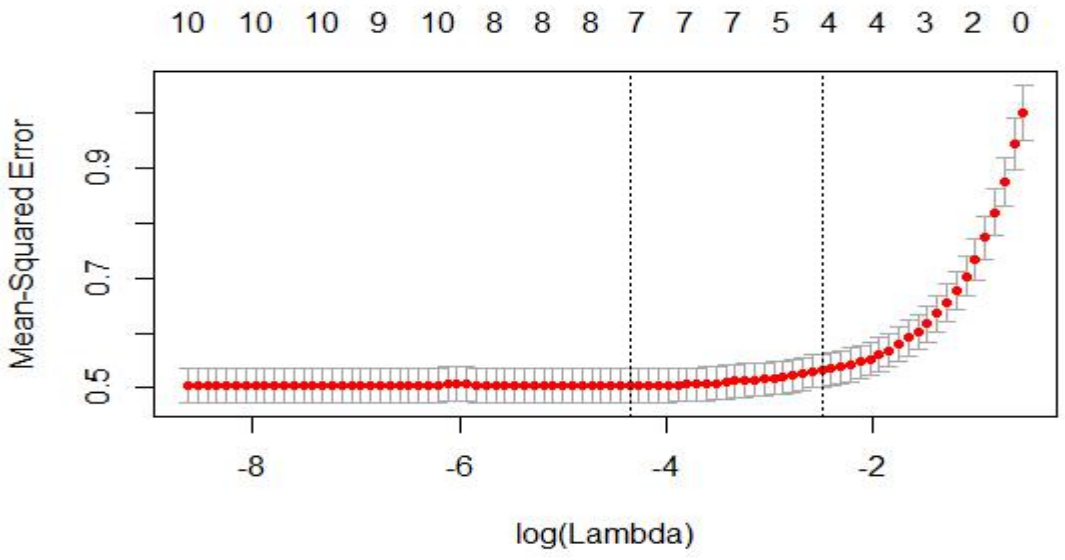

Fig. 4.1: MSE plot and the number of *Variables* in the model as a function of ln (λ) for the 10-fold cross validation for the LASSO Regression.

The figure above gives the relationship between $ln\lambda \; and \; MSE$. 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.

4.5 Results Based 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 $\alpha_i = \frac{i}{100} - 0.01, for\ i = 1,2,...,101.$ That is $\alpha_1 = 0,\ \alpha_2 = 0.01,\ \alpha_3 0.02,\ \alpha_4 = 0.03,\ \alpha_5 = 0.04, \alpha_6 = 0.05,\ \alpha_7 = 0.06, \alpha_8 = 0.07, \alpha_9 = 0.08, \alpha_{10} = 0.09, \alpha_{11} = 0.1, \alpha_{12} = 0.11, \alpha_{13} = 0.12, \alpha_{14} = 0.13, \alpha_{15} = 0.14, \alpha_{16} = 0.15, \alpha_{17} = 0.16, ..., \alpha_{101} = 1$. 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: Numerical results of elastic net

λ	MSE	Df
3.6611	1.0024	0
1.5848	10.7482	6
0.6251	0.5803	6
0.0264	**0.5034**	**8**
0.0037	0.5051	10
0.0004	0.5050	10

Table 4.4 gives the different values of MSE at different values of (λ). For example when $(\lambda) = 3.6611$ the value of MSE is 1.0024 and the number of non-zero variables in the model is 0 and when $(\lambda) = 0.0264$ etc. which is the minimum MSE with 8 non-zero variables included in the model.

Table 4.4.1: Numerical results for***ELASTICNET*** 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
Df	8

Table 4.4.2: shows the values of $\boldsymbol{MSE's}$ using different values of $\boldsymbol{alpha}(\boldsymbol{\alpha})$

$\alpha_i \, for \, i = 1,2, \ldots, 101$	$MSE's$
$\alpha_1 = 0$	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
$\boldsymbol{\alpha_{17} = 0.16}$	$\mathbf{0.503400}$
$\alpha_{18} = 0.17$	0.503420
$\alpha_{19} = 0.18$	0.504322
.	.
.	.
.	.
$\alpha_{101} = 1$	0.503750

From Table 4.3.2, 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. `

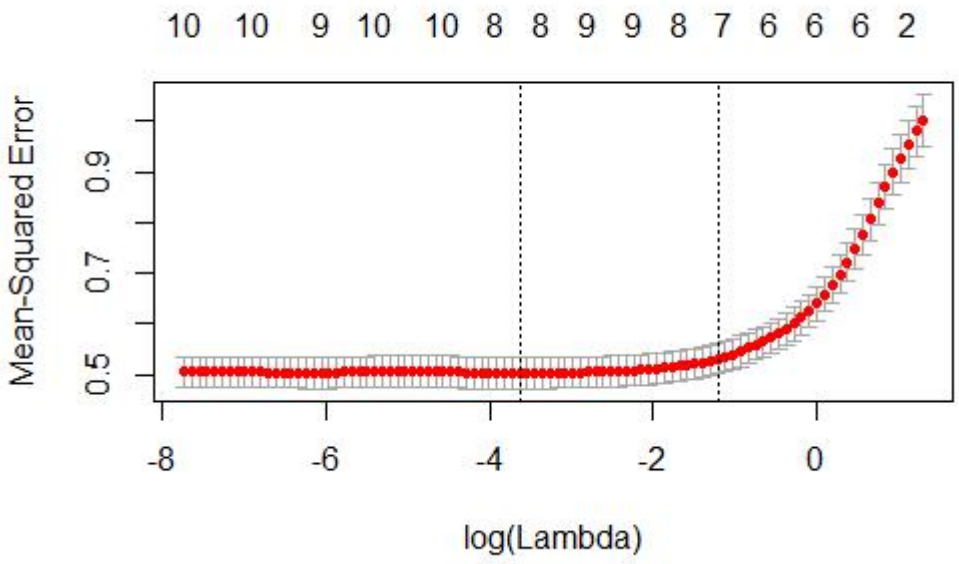

Fig. 4.3: MSE plot and the number of *Variables* in the model as a function of ln (λ) for the 10-fold cross validation for the Elastic Net Regression.

The figure above gives the relationship between $ln\lambda \; and \; MSE$. 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 any value of $lambda(\lambda)$ between the left line and the right line to obtain the numerical results for elastic net.

We shall now use the value of $(\lambda) = 0.0264$ to calculate the penalized regression for $elastic \; net$. As given in Table 4.4.

4.6 Results Based Correlation adjusted elastic net regression

Since minimizing

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

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 $\lambda_{2,i} = \frac{i}{100} - 0.01 \; for \; i = 1,2, \ldots, 101$ .That is$\lambda_{2,1} = 0, \lambda_{2,2} = 0.01, \lambda_{2,3} = 0.02, \ldots, \lambda_{2,101} = 1$.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, we also show the number of non-zero variablesin the model which is called Df.

Table 4.5: ***CAEN*** numerical results

$\lambda_{2,i} \; for \; i = 1,2 \ldots, 101$	λ_1	MSE	Df
$\lambda_{2,1} = 0$	0.0129177	0.50372	8
$\lambda_{2,2} = 0.01$	0.0141542	0.533673	7
$\boldsymbol{\lambda_{2,3} = 0.02}$	**0.014174**	**0.50340**	**7**
.	.	.	
.	.	.	
.	.	0.50684	7
$\lambda_{2,101} = 1$	0.0140633		

Table 4.5.1: Numerical results for ***CORRELATIONADJUSTEDELASTICNET*** regression

$Variable$	$Coefficients$
AGE	0.0000
SEX	−0.8404
BMI	

BP	2.2779
TC	1.2857
LDL	−0.4265
HDL	0.0000
TCH	−0.9701
LTG	0.0000
GLU	2.2322
	0.2282
MSE	0.50340
Df	7

Table 4.6. Shows the values of $\boldsymbol{MSE's}$ using different values of $\boldsymbol{\lambda_1 and \lambda_2}$

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

Table 4.6 it could be observed when $\lambda_2 = 0.02$ gives the minimum value of MSE as 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.

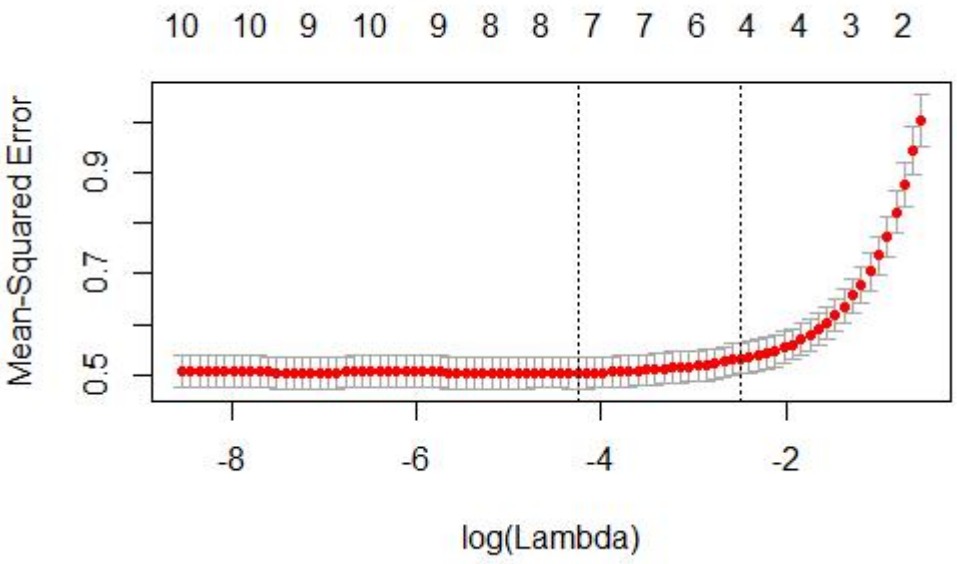

Fig. 4.3: MSE plot and the number of Variables in the model as a function of ln (λ) for the 10-fold cross validation for CAEN Regression.

The figure above gives the relationship between $ln\lambda\, and\, MSE$. 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.

4.7: Smoothly Clipped Absolute Deviation regression

Table 4.7: Numerical results for $\boldsymbol{SCAD}$ regression

$Variable$	$Coefficients$
AGE	0.0000
SEX	−0.5286
BMI	

BP	1.1993
TC	0.7188
LDL	0.0001
HDL	−0.9224
TCH	−0.6577
LTG	0.0000
GLU	1.0924
	1.4708
MSE	0.5000
Df	7

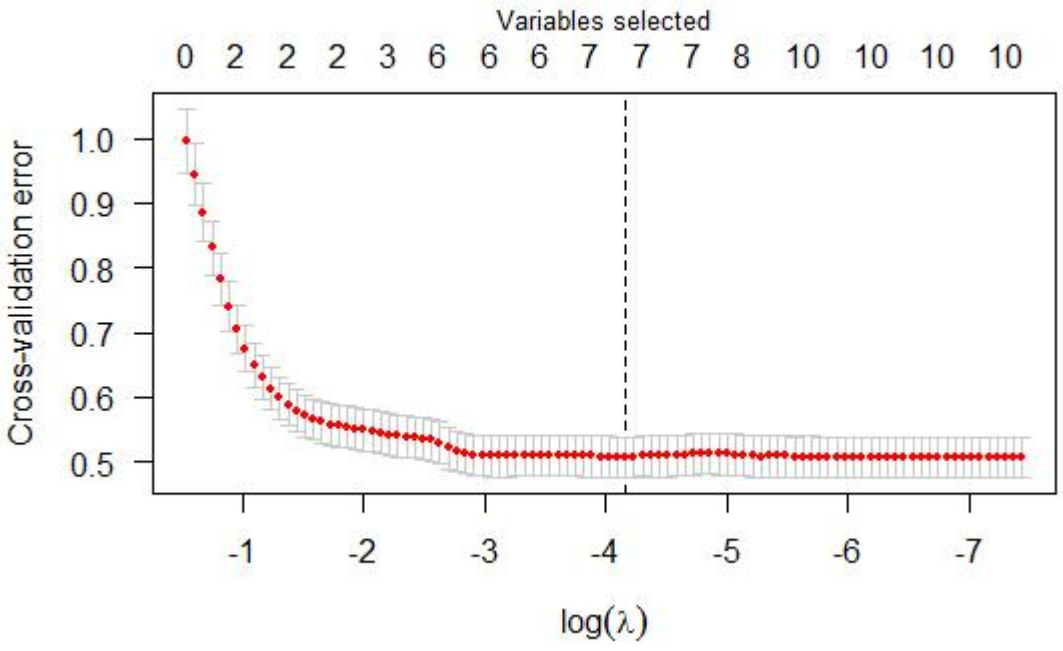

Fig. 4.4: MSE plot and the number of *Variables* in the model as a function of ln (λ) for the 10-fold cross validation for the SCAD Regression.

The figure above gives the relationship between $ln\lambda \; and \; MSE$. 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. We can therefore choose any value of $lambda(\lambda)$ between the left lines. The minimum value of lambda used for the calculation of SCAD coefficients is 0.0413.

Table 4.8: Coefficient Comparison of OLS, LASSO, CAEN, EN and SCAD Regression

Variable	*OLS*	*Optimal OLS*	*LASSO*	*ELASTIC NET*	*CAEN*	*SCAD*

AGE	-0.0062	-	0.0000	0.0000	0.0000	0.0000
SEX	-0.1481	-0.1399	-0.1211	-0.1342	-0.8404	-0.5286
BMI	0.3211	0.3321	0.3225	0.3189	2.2779	1.1993
BP	0.2004	0.2028	0.1830	0.1907	1.2857	0.7188
TC	-0.4893	-0.08440	− 0.0630	-0.1010	-0.4265	0.0000
LDL	0.2945	-	0.0000	0.0000	0.0000	-0.9224
HDL	0.0624	-0.1486	-0.1379	-0.1078	-0.9701	-0.6577
TCH	0.1094	-	0.0000	0.0513	0.0000	0.0001
LTG	0.4641	0.3432	0.3173	0.3151	2.2322	1.0924
GLU	0.0418	-	0.0333	0.0423	0.2282	1.4708
MSE	0.5050	0.493	0.5040	0.5034	0.5034	0.5000
Df	10	6	7	8	7	7

The resultant models for the compared regression models are as follows:

OLS Regression

$$y = -0.0062Age - 0.1481Sex + 0.3211Bmi + 0.2004Bp - 0.4893Tc + 0.2945Ldl + 0.0624Hdl + 0.1094Tch + 0.4641Ltg + 0.0418Glu$$

LASSO Regression

$$y = -0.1211Sex + 0.3225Bmi + 0.1830Bp - 0.0630Tc - 0.1379Hdl + 0.3173Ltg + 0.0333Glu$$

***ELASTIC* Net Regression**

$$y = \text{-}0.1342Sex + 0.3189Bmi - 0.1907Bp - 0.1010\text{Tc} + \text{-}0.1078\text{Hdl} + 0.0513\text{Tch} + 0.3151\text{Ltg} + 0.0423\text{Glu}$$

CAEN Regression

$$y = -0.8404Sex + 2.2779Bmi + 1.2857Bp - 0.4265Tc + 0.9701Hdl + 2.2322Ltg + 0.2282Glu$$

SCAD Regression

$$y = -0.5286Sex + 1.1993Bmi - 0.7188Bp - 0.6577Hdl + 0.0001Tch + 1.0924Ltg + 1.4708Glu$$

COMMENT ON TABLE 4.8

It is observed that Smoothly Clipped Absolute Deviation (SCAD) outperform the three other models namely the LASSO, Elastic Net and CAEN with the lowest Mean square error of 0.5000 with 7 variables in the model.

CHAPTER FIVE:

SUMMARY, CONCLUSION AND RECOMMENDATION

5.1 Summary

In this study the performance of LASSO, EN, CAEN and SCAD regression methods was examined using numerical results. We applied LASSO, EN, CAEN and SCAD methods to

diabetes dataset. The characteristics of each of the penalized regression methods were observed carefully. The LASSO, EN, CAEN and SCAD regression does both the shrinkage and variable selection and there are 7 nonzero variables selected by LASSO, 8 nonzero variables by EN, 7 nonzero variables selected by CAEN and 7 nonzero variables selected by SCAD in the final model and it also gives a smaller *MSE.* According to the dataset, SCAD outperforms LASSO, EN, CAEN regression in terms of mean square error and it produced a less complex model than the other three Penalized Methods.

5.2 Conclusion

To develop an accurate model, one needs to collect numerous variables and those variables are often highly correlated, as discussed earlier in the dissertation, those variables that are correlated make the model less predictive and difficult to interpret. Penalized regression method provides a better way of selecting the appropriate variables to develop an effective model as observed in this dissertation.

5.3 Recommendation

In order to reduce the flaw inherent in the prediction accuracy of the ordinary least squares, Penalized regression techniques have been developed. Multicollinearity is a problem that is rarely considered in elementary statistics texts, because it is really a mathematical-statistical problem, but it is rather a problem in the interpretation of the coefficients. It is a problem that confronts researchers in actual data analytic situations. Therefore, researchers should always be at alert to the possibility of the problem

5.4 Suggestion for further study

Since SCAD perform better compared to the other three methods. SCAD can also be applied to survival data, since there are lots of variables in many survival data analysis problems also there is time constrain

5.5 Contribution to knowledge

The following contributions are made which are vital for academic and other purposes;

- We introduce the SCAD penalized to compare with the three existing methods (LASSO,EN,CAEN) in order to get the penalized regression methods that will best minimize the Multicollinearity
- R was used to find the numerical result for existing penalized regression methods which include **lasso**, **Elastic Net** and **Correlation Adjusted Elastic Net**..
- Suitable data argumentation was used to find and determine the numerical results for Smoothly Clipped Absolute Deviation.
- The numerical results was summarize and compared.

REFERENCES

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

Andre, N., Young, T. M. and Rials, T. (2006). Online monitoring of the buffer capacity of particle board furnish by near-infrared spectroscopy. *Applied Spectroscopy*, 60(10), 1204-1209.

Ayers, K. L., and Cordell, H. J. (2010).SNP selection in genome-wide and candidate gene studies via penalized logistic regression. *Genet. Epidemiol.* 34, 879–89.1

Bovelstad. H. M.,Nygard, S., Storvold, H. I., Aldrin, M., Borgan, O., and Frigessi, A., (2007) predicting survival from microray data a comparative study. *Bioinformatics.* 23(16): 2080-2087

Beer, D.G. Kardia. S. I., Huang. C.C, Giordano, T.J. and Levin. A. M. (2002). Gene-expression Profiles predict survival Patients with lung adenocarcinoma. *Nat. Med.*, 8(8): 816-824.

Breiman, I. (1996). Heuristics of instability and stability in model selection. *Annals of Statistics,* 24(6): 2350-2383

Breiman. I. and Friedman, J., (1997). Predicting multiple responses in multiple linear regression (with discussion). *Journal of the Royal Statistical Society: Series B* 59, 3-54.

Bondell, H. D. and Reich, B. J. (2008). Simultaneous regession shrinkage, variable selection and clustering of predictors with OSCAR. *Biometrics* 64(1), 115-123.

Buhlmann, P., and VandeGeer, S. (2011). Statistics for High Dimensional Data: methods, theory and applications. Springer Science and Business Media.

Cho, S., Kim,K.,Kim,Y.J.,Lee,J.K.,(2010). Joint identification of multiple genetic variants via elastic-net variable selection in a genome-wide association analysis. *Annals of Human Genetics.* 74(5), 416-428

Donoho, D., Elad, M. and Temlyakov, M (2004). Stable recovery of space over complete representatation in the presence of noise. *IEEE transactions on information theory*, 52(1)

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, NY.

Dismuke, C., and Lindrooth, R. (2006). Ordinary least Squares. Methods and designs for outcomes Research, 93, 93-104.

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

Efron, B. Hastie, T., johnstone, I. and Tibshirani, R. (2004). Least Angle Regression. *The Annals of Statistics,* 32(2), 407 – 499.

Efroymson, M.A. (1960). Multiple regression Analysis. In: Ralston, A. and Wilf, HS, editors, Mathematical Methods for Digital Computers. John Wiley and Sons, Inc. NewYork, NY.

Fan, J. and Li, R. (2001). Variable selection via non-concave penalized likelihood and its oracle Properties. *Journal of the American Statistical Association,* 96(456), 1348-1360.

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

Friedman, J. Hastie, T, and Tibshirani, R. (2007). Pathwise Coordinate Optimization, *Annals of Applied Statistics.* 1(2), 302-322.

Fu, W.J. (1998). Penalized Regression: The Bridge Versus the Lasso. *Journal of Computational* and Graphical Statistics, 7(3): 397-416.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization Path for generalized linear models via coordinate decent. *Journal of Statistical Software,* 33(1), 1.

Hans, C. (2010). Model uncertainly and variable selection in Bayesian LASSO regression. *Statistical Computing.* 20(2): 221-229

Harell, J.R, F.E, Lee K.L, and Mark, D.B. (1996). Multivariate Prognostic models, lssues in developing models, evaluating assumptions and adequacy and measuring and reducing errors. *Stat med* 15(4), 361-387

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

Hesterberg, T., Choi, N.H., Meier, L., and Fraley, C. (2008) least angle and L1 penalized regression: a review. *Stat. Surv.*, 2, 61-93

Hoerl, A.E. and Kennard, R. (1970). Ridge regression: Biased estimation for non-orthogonal problems, *Technometrics,* 12(1), 55-67.

Hurvich, C and Tsai, C., (1990). The impact of model selection on inference in linear regression *American Statistician*. 44(3), 214-217.

Knight, K. and Fu, W. (2008), Asymptotic for lasso-type estimators. *The Annals of Statistics,* 28, 1356-1378.

Kooperberg, C., LeBlane, M. and Obenchain, V., (2010) Risk prediction using genome- wide Association studies, *Genetics Epidemiology*., 34(7), 643 – 652.

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

Kyung, M., Gill, J. Ghosh, M., and Casella, G. (2010). Penalized regression, standard errors, and Bayesian lasso. *Bayesian Annals*. 5(2), 369-412.

Li, Q., and Lin, N (2010). The Bayesian elastic net. *Bayesian Annals* 5(1), 151-170.

Matthew, P.K. and Yahaya, A. (2015). Performance analysis on LASSO, Elastic net and Correlation Adjusted Elastic Net regression methods. *International Journal of Advanced Statistics and Probability,* 3(1), 93-99.

Motyer, J., Allan, C., Mc Kendry, S., Galbraith, G., and Susan, R.W. (2011). LASSO model selection with post processing for a genome-wide association Study dataset. *BMC proceedings,* 5(9): S24.

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

Neter, J. Kutner, M.H. Nachtsheim, C.J. and Wasserman, W. (1996). Applied linear Regression Models.4, p. 318. Chicago: Irwin.

Osborne, M.R. Presnell, B. and Turlach, B.A. (2000). On the LASSO and it's dual. Jouney of *Computational and Graphical Statistics.* 9(2), 319-337.

O'brien, Robert, M."A caution regarding rule of thumb for variance inflation factor".Quality and Quantity. 41(5), 673-690.

Roecker, E. (1991). Prediction error and its estimation for subset-selection models *Technometrics.* 33(4), 459-468.

Shedden, K., Taor, J. M, and Enkemann, S. A (2008). Gene expression- based survival prediction in lung adenocarcinoma: a Multi-site, blinded validation study. *Nat. Med.,* 14(8), 822-827.

Sorlie, T., Perou, C.M.,Tibshirani, R.,Aas, T., Geisher, S., Johnsen, H., Hastie, T.,Esien, M.B., Van de Rijin, M., Jefrey, S.S., Thorsen, T., Quist, H., Matese, J.C., Brown, P.O., Bostein, D.,Lonning, P.E.,Borresen-Dale, A.L.,(2001). Gene expression patterns of breast carcinomas distinguish tumor Subclasses with clinical implications. *Proc. Natl Acad. Sci. USA,* 98, 10869-10874.

Szymezak, S., Biernacka, J.M., Cordell, H.J., Gonzalez-Recio, and Konig, I.R., (2009). "Machine Learning in Genome-Wide Studies" *Epidemiology,* 33, pp 51-56

Sung Y.J., Rice, T.K., Shi, G., Gu, C.C., and Rao, D. C (2009). Comparison between single-market analysis using Merlin and Multi-marker Analysis using LASSO for the Framing Ham simulated data. *BMC proceedings*. 3(7):S27.

Tan. Q. (2012). Correlation Adjusted penalization in regression Analysis. PhD Thesis. Department of statistics, University of Manitoba. (Unpublished)

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

Tibshirani, R.,Michael S., Saharon (2005). Sparsity and smoothness Via the fused lasso. *Journal of Royal Statistical Society Series B*, 67(1), 91-108

Tutz, G. and Ulbricht, J. (2009). Penalized regression with correlation-based penalty. Statistics and computing, 19(3), 239-253.

Turlach, B., Venables, W., Wright, S., (2005). Simultaneous variable selection. *Technometrics* 47(3), 349-363.

Usai, M.G. Goddard, M.E., and hayes, B.J., (2009). LASSO with cross-validation for genomic Selection. *Genetic Research*. 91(6), 427-436.

Van de Vijver, M.J., Bergh, J., Piccart, M., Doleronzi, M (2002). A gene-expression signature as a prediction of survival in breast cancer. *New England Journal of Medicine*.347(25), 1999-2009

Wigle, D. A., Jurisica, I., Radulovich, N., Pintilie, M., Rossant, J., Liu, N., Lu, C., Woodgett, J., (2002), "Molecular profiling of non-small cell lung cancer and correlation with disease-free survival", *Cancer Res.*, 62, pp: 3005 – 3008.

Waldmann P, Meszaros G, Gredler B. Fuerst C and Solkner J (2013) Evaluation of the lasso and the elasticnet in genome-wide association studies. *Frontiers in Genetics.* 4:270.

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.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net."journal of the Royal Statistical Society: Series B(Statistical Methodology) 67(2), 301-320.

Zou, H. and Hastie, T., and Tibshirani, R. (2007). On the degree of freedom of the lasso. The Annals of Statistics 35(5)*,* 2173-2192.

Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *Journal of machine learning research,* 7(Nov). 2541-2563

APPENDIX A

Data set of diabetes patient

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

Rcodes

```
> data=as.matrix(read.table("theoriginaldata.txt",header=TRUE))

> datas=stdize(data)

> datasx=data[,-11]

> datasy=data[, 11]

> Xij=datasx
```

```
> dim(Xij)
[1] 442  10
> n=dim(Xij)[1]
> p=dim(Xij)[2]
> one=rep(1,n)
> X.means=t(one)%*%Xij/n
> X.diff=Xij-one%*%X.means
> View(X.diff)
> X.cov=t(X.diff)%*%X.diff/(n-1)
> View(X.cov)
> sdi=diag(1/sqrt(diag(X.cov)))
> View(sdi)
> X.cor=sdi%*% X.cov%*%sdi
> View(X.cor)
> X.cor=round(cor(datasx),7)
> datasx=as.matrix(datasx)
> datasy=as.matrix(datasy)
> zeromatrix=matrix(c(rep(0,p)),p,1)
> Ynew=rbind(datasy,zeromatrix)
> Da=matrix(c(1,-X.cor[1,2],0,0,0,0,0,0,0,0,0,1,-X.cor[2,3],0,0,0,0,0,0,0,0,0,1,-
X.cor[3,4],0,0,0,0,0,0,0,0,0,0,1,-X.cor[4,5],0,0,0,0,0,0,0,0,0,0,1,-X.cor[5,6],0,0,0,0,0,0,0,0,0,0,1,-
X.cor[6,7],0,0,0,0,0,0,0,0,0,0,1,-X.cor[7,8],0,0,0,0,0,0,0,0,0,0,1,-X.cor[8,9],0,0,0,0,0,0,0,0,0,0,1,-
X.cor[9,10],0,0,0,0,0,0,0,0,0,0,1),nrow=10)
> View(Da)
> Wa=Da%*%t(Da)
> View(Wa)
```

```
> Db<-matrix(c(1,-X.cor[1,2],0,0,0,0,0,0,0,0,1,0,-X.cor[1,3],0,0,0,0,0,0,0,1,0,0,-
X.cor[1,4],0,0,0,0,0,0,0,1,0,0,0,-X.cor[1,5],0,0,0,0,0,0,1,0,0,0,0,-X.cor[1,6],0,0,0,0,0,1,0,0,0,0,0,-
X.cor[1,7],0,0,0,0,1,0,0,0,0,0,0,-X.cor[1,8],0,0,1,0,0,0,0,0,0,0,-X.cor[1,9],0,1,0,0,0,0,0,0,0,0,-
X.cor[1,10],0,1,-X.cor[2,3],0,0,0,0,0,0,0,0,0,1,0,-X.cor[2,4],0,0,0,0,0,0,0,1,0,0,-
X.cor[2.5],0,0,0,0,0,0,0,1,0,0,0,-X.cor[2,6],0,0,0,0,0,0,1,0,0,0,0,-X.cor[2,7],0,0,0,0,0,1,0,0,0,0,0,-
X.cor[2,8],0,0,0,0,1,0,0,0,0,0,0,-X.cor[2,9],0,0,0,1,0,0,0,0,0,0,0,-X.cor[2,10],0,0,0,1,-
X.cor[3,4],0,0,0,0,0,0,0,0,0,1,0,-X.cor[3,5],0,0,0,0,0,0,0,0,1,0,0,-X.cor[3,6],0,0,0,0,0,0,0,1,0,0,0,-
X.cor[3,7],0,0,0,0,0,0,1,0,0,0,0,-X.cor[3,8],0,0,0,0,0,1,0,0,0,0,0,-X.cor[3,9],0,0,0,0,1,0,0,0,0,0,0,-
X.cor[3,10],0,0,0,0,1,-X.cor[4,5],0,0,0,0,0,0,0,0,0,1,0,-X.cor[4,6],0,0,0,0,0,0,0,0,1,0,0,-
X.cor[4,7],0,0,0,0,0,0,0,1,0,0,0,-X.cor[4,8],0,0,0,0,0,0,1,0,0,0,0,-X.cor[4,9],0,0,0,0,0,1,0,0,0,0,0,-
X.cor[4,10],0,0,0,0,0,1,-X.cor[5,6],0,0,0,0,0,0,0,0,0,1,0,-X.cor[5,7],0,0,0,0,0,0,0,0,1,0,0,-
X.cor[5,8],0,0,0,0,0,0,0,1,0,0,0,-X.cor[5,9],0,0,0,0,0,0,1,0,0,0,0,-X.cor[5,10],0,0,0,0,0,0,1,-
X.cor[6,7],0,0,0,0,0,0,0,0,0,1,0,-X.cor[6,8],0,0,0,0,0,0,0,0,1,0,0,-X.cor[6,9],0,0,0,0,0,0,0,1,0,0,0,-
X.cor[6,10],0,0,0,0,0,0,0,1,-X.cor[7,8],0,0,0,0,0,0,0,0,0,1,0,-X.cor[7,9],0,0,0,0,0,0,0,0,1,0,0,-
X.cor[7,10],0,0,0,0,0,0,0,0,1,-X.cor[8,9],0,0,0,0,0,0,0,0,0,1,0,-X.cor[8,10],0,0,0,0,0,0,0,0,0,1,-
X.cor[9,10],0,0,0,0,0,0,0,0,0,0,1),nrow=10)

>

> Wb<-Db%*%t(Db)

> Cb<-chol(Wb)

> lambda<-seq(from=0,to=2,length=201)

> min.cv.caen.index<-numeric(201)

> min.cv.caen<-numeric(201)

> min.lambda.caen<-numeric(201)

> min.cvsd.caen.index<-numeric(201)

> for(i in 1:201)

+ X.new<-(1/sqrt(1+lambda[i]))*rbind(Xij,sqrt(lambda[i])*t(Cb))

> Y.new<-rbind(datasy,zeromatrix)

> diabetesdata.new<-cbind(X.new,Y.new)

> View(diabetesdata.new)

> diabetesdata.new<-stdize(diabetesdata.new)

> diabetesdata.new<-as.matrix(diabetesdata.new)
```

```
> x.new<-diabetesdata.new[,-11]
> y.new<-diabetesdata.new[, 11]
> cv.caen.lasso.glm<-cv.glmnet(x.new, y.new, nfold=452, alpha=1)
Warning message:
Option grouped=FALSE enforced in cv.glmnet, since < 3 observations per fold
> X.new<-(1/sqrt(1+lambda[3]))*rbind(Xij,sqrt(lambda[3])*t(Cb))
> Y.new<-rbind(diabetesdatay,zeromatrix)
> diabetesdata.new<-cbind(X.new,Y.new)
> diabetesdata.new<-stdize(diabetesdata.new)
> diabetesdata.new<-as.matrix(diabetesdata.new)
> x.new<-diabetesdata.new[,-11]
> y.new<-diabetesdata.new[, 11]
> cv.caen.lasso.glm<-cv.glmnet(x.new, y.new, nfold=452, alpha=1)
Warning message:
Option grouped=FALSE enforced in cv.glmnet, since < 3 observations per fold
> cv.caen.lasso.glm$cvsd[,3]
> index<-which.min(cv.caen.lasso.glm$cvm)
> cv.caen.lasso.glm$lambda[3]
[1] 0.4862633
> cv.caen.lasso.glm$lambda[,3]
> cv.caen.lasso.glm$lambda[index]
[1] 0.01417483
> cv.caen.lasso.glm$cvm[index]
[1] 0.5037843
> cv.caen.lasso.glm$cvsd[index]
```

```
[1] 0.03104379
> caen1.final.glm$beta[,index]/sqrt(lambda[3])
> caen1.final.glm$beta[,index]/sqrt(lambda[3])
      AGE        SEX        BMI         BP         S1         S2
 0.0000000 -0.8404332  2.2779440  1.2857395 -0.4265534  0.0000000
       S3         S4         S5         S6
-0.9701990  0.0000000  2.2322880  0.2282234
> plot(caen1.final.glm)
> plot(cv.lasso.glm)
> plot(cv.caen.lasso.glm)
> par(mfrow=c(2,2))
> fit <- ncvreg(datasx,datasy)
> plot(fit,main=expression(paste(gamma,"=",3)))
> fit <- ncvreg(datasx,datasy,gamma=10)
> plot(fit,main=expression(paste(gamma,"=",10)))
> cvfit <- cv.ncvreg(datasx,datasy)
> plot(cvfit)
> > summary(cvfit)
MCP-penalized linear regression with n=442, p=10
At minimum cross-validation error (lambda=0.0413):
-------------------------------------------------
  Nonzero coefficients: 6
  Cross-validation error (deviance): 0.50
  R-squared: 0.50
  Signal-to-noise ratio: 1.00
```

```
  Scale estimate (sigma): 0.707

> coef(fit, lambda=0.0413)

  (Intercept)          AGE          SEX          BMI           BP

-2.756927e-16  0.000000e+00 -8.053654e-02  3.481872e-01  1.584260e-01

           S1           S2           S3           S4           S5

-3.686495e-02  0.000000e+00 -1.055784e-01  0.000000e+00  3.287210e-01

           S6

 0.000000e+00
```

```
lasso.glm<-glmnet(datax,datay, alpha= 1)
summary(lasso.glm)
cv.lasso.glm<-cv.glmnet(datax,datay,
summary(cv.lasso.glm)
nfold=442, alpha= 1)
plot(cv.lasso.glm)
plot(lasso.glm)
min.cv.index<-which.min(cv.lasso.glm$cvm)
min.cv.lasso<-cv.lasso$cv [min.cv.index]
lasso.glm$beta[,42]
sum(abs(lasso.glm$beta[,1]))
```