A Numerical Study Of Penalized Regression

by

Han Yu

A Thesis submitted to the Faculty of Graduate Studies of
The University of Manitoba
in partial fulfillment of the requirements of the degree of

Master of Science

Department of Statistics
University of Manitoba
Winnipeg

Abstract

In this thesis, we review important aspects and issues of multiple linear regression, in particular on the problem of multi-collinearity.

The focus is on a numerical study of different methods of penalized regression, including the ridge regression, lasso regression and elastic net regression, as well as the newly introduced correlation adjusted regression and correlation adjusted elastic net regression. We compare the performance and relative advantages of these methods.

Acknowledgments

Firstly of all, I would like to express my gratitude to my supervisor, Dr. Xikui Wang, whose encouragement, patience and guidance helped me in the research and writing of the thesis. I am grateful to the committee members, Dr. Saumen Mandal of the Department of Statistics and Dr. Jijun Gao of the Department of Business Administration for their useful suggestions. I would like to give my special thanks to my parents and my wife for their love, belief and encouragement throughout my life.

Lastly, I wish to express my appreciation to all those who have supported me in any respect during the completion of this thesis.

This thesis is dedicated to my family
for their love, support
and encouragement.

Contents

Abstract i

Acknowledgments ii

Dedication iii

Contents iv

List of Tables vii

1 Introduction 1

1.1 Regression analysis 1

1.2 The issue of multi-collinearity 8

1.3 Objectives and scope of research 9

1.4 Structure of the thesis 10

2 Methods of penalized regression **11**

2.1 Introduction 11

2.2 Ridge regression 12

2.3 Lasso regression 14

2.4 Elastic net regression 16

2.5 Correlation adjusted regression 17

2.6 Correlation adjusted elastic net regression 21

2.7 Summary 26

3 A numerical comparison of penalization methods **28**

3.1 Introduction 28

3.2 Methods 29

3.3 Numerical results and explanation 31

3.3.1 Preliminary material 31

3.3.2 Ridge regression 37

3.3.3 Lasso regression 39

3.3.4 Elastic net regression 41

3.3.5 Correlation adjusted regression 43

3.3.6 Correlation adjusted elastic net 47

3.4 Summary 51

4 Conclusion **54**

4.1 Summary of results 54

4.2 Future research 55

Appendix **56**

Bibliography **74**

List of Tables

3.1 original data frame 31

3.2 standardized data frame 32

3.3 ordinary least squares regression 32

3.4 correlation matrix among independent variables 33

3.5 ridge numerical results 37

3.6 lasso numerical results 39

3.7 elastic net numerical results 42

3.8 CAR_1 numerical results 44

3.9 CAR_2 numerical results 44

3.10 CAR estimators and the percentage change 45

3.11 $CAEN_1$ numerical results 47

3.12 $CAEN_2$ numerical results 49

3.13 coefficient comparison of classic methods 51

3.14 coefficient comparison of ridge and CAR 52

3.15 coefficient comparison of elastic net and $CAEN$ 53

Chapter 1

Introduction

1.1 Regression analysis

Regression analysis is one of the most important tools for analyzing relationships between one response variable and one or more explanatory variables. It is widely used in our real lives, including the social and biological sciences, economics and so on. Regression analysis has became one of the most important data analysis methods.

The term "regression" was first introduced by Francis Galton (1822-1911). At that time, regression had only the biological meaning since Galton used "regression" to describe a biological phenomenon. Later, Udny Yule and Karl Pearson extended it to a more general statistical context. However, the earliest form of regression which is called the method of least squares was published by Gauss in 1809. In 1821, Gauss published a further development of the theory of the least squares, including a version of the Gauss-Markov theorem. In recent decades, many new regression methods have been developed, including linear regression, logistic regression and penalized regression.

When there is only one explanatory variable corresponding to the response variable, we call it simple regression. For example, if we want to know if there is enough evidence that the father's height affects his child's height, we use the simple regression. As we see, there is only one explanatory variable which is the father's height and one corresponding response variable which is his child's height. If there are at least two explanatory variables corresponding to the response variable, we call it multiple regression. In the previous example, we have to consider not only the father's height, but also the mother's height and the family income. Now, we have three explanatory variables which are the father's height, the mother's height and family income. The one corresponding response variable is their child's height. We see that the simple regression can be regarded as a special case of the multiple regression.

In the multiple linear regression model, let Y denote the response variable (also called the endogenous variable or the dependent variable) and $X_1, X_2, \cdots, X_p$ denote the explanatory variables (also called exogenous variables or independent variables). The relationship between Y and $X_1, X_2, ..., X_p$ can be expressed as

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon.$$

The parameters $\beta_0, \beta_1, \cdots, \beta_p$ are called regression coefficients and ε is a random variable.

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

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i, \quad i = 1, 2, \cdots, n. \tag{1.1}$$

where y_i is the i^{th} response observation, x_{ij} is the i^{th} observation on the j^{th} independent variable, $\beta_0, \beta_1, \cdots, \beta_p$ are the unknown parameters and $\varepsilon_i \sim N(0, \sigma_i^2)$. Often, those above n equations can be rewritten in the matrix form as

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

or

$$E(\boldsymbol{Y}) = \boldsymbol{X}\boldsymbol{\beta},$$

where

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

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

We plug each individual statistical unit in equation 1.1 to obtain the matrix form as

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

where $\boldsymbol{x_i^T} = \begin{pmatrix} 1 & x_{i1} \cdots & x_{ip} \end{pmatrix}$ and i=1,2,$\cdots$, n.

Assumptions of multiple linear regression model include $LINE$. That is,

1. Linearity: the relationship between the explanatory variables and the respond variable is linear. This is the only restriction on the parameters (not explanatory variables), since the explanatory variables are regarded as fixed values. That is,

   - $E\left(y_i \mid x_{i1}, x_{i2}, \cdots x_{ip}\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} = \boldsymbol{x_i^T \beta}$,
   - $\frac{\partial E(y_i|\boldsymbol{x_i})}{\partial \boldsymbol{x_i}} = \boldsymbol{\beta}$.

2. Independence: there are two types of independence.

   - Each combination of explanatory variable and error is independent. $E(\varepsilon_i \mid X_j) = 0$ for all $i = 1, 2, \cdots, n$ and $j = 1, 2, \cdots, n$.
   - The error variables are independent. Therefore, $Cov(\varepsilon_i, \varepsilon_j) = 0$ or equivalently $Cov(y_i, y_j) = 0$ for all $i \neq j$.

3. Normality: the error variables follow normal distributions.

   - $\varepsilon_i \sim N(0, \sigma_i^2)$
   - $\boldsymbol{\varepsilon} \sim N(0, \boldsymbol{\sigma^2})$
   - $\boldsymbol{Y} \sim N(\boldsymbol{X\beta}, \boldsymbol{\sigma^2})$

     where

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

4. Equal Variance: each error variable has the same variance.

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

The ordinary least squares (OLS) is a classic technique to estimate the parameters of the multiple linear regression model. There are two principles to establish the OLS regression model.

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

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

  where

  $e_i = y_i - \hat{y}_i$ is called residual of the i^{th} observation.

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

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

Here $\hat{y}_i$ is an estimator of $E(y_i)$ and there is no distribution assumptions required

for OLS. Now, in vector form, we have

$$\begin{aligned}
\widehat{\boldsymbol{\varepsilon}}^T\widehat{\boldsymbol{\varepsilon}} &= \sum_{i=1}^{n} e_i^2 \\
&= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \\
&= \sum_{i=1}^{n}(y_i - \boldsymbol{x_i^T}\widehat{\boldsymbol{\beta}})^2 \\
&= (\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})^T(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}) \\
&= \boldsymbol{Y}^T\boldsymbol{Y} - 2\boldsymbol{Y}^T\boldsymbol{X}\widehat{\boldsymbol{\beta}} + \widehat{\boldsymbol{\beta}}^T\boldsymbol{X}^T\boldsymbol{X}\widehat{\boldsymbol{\beta}}
\end{aligned}$$

Note that $\boldsymbol{Y}^T\boldsymbol{X}\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}^T\boldsymbol{X}^T\boldsymbol{Y}$, since they are both numbers.

To find $\widehat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_p)^T$ that minimizes $\widehat{\boldsymbol{\varepsilon}}^T\widehat{\boldsymbol{\varepsilon}}$, we take derivative of $\widehat{\boldsymbol{\varepsilon}}^T\widehat{\boldsymbol{\varepsilon}}$ with respect to $\widehat{\boldsymbol{\beta}}$ and let the derivative equal to zero to obtain $\widehat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_p)^T$. Then,

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

Finally, the OLS estimator is

$$\widehat{\boldsymbol{\beta}}_{OLS} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y},$$

where $\widehat{\boldsymbol{\beta}}_{OLS}$ is a best linear unbiased estimator ($BLUE$). Specifically,

- *Best* means $Var(\widehat{\boldsymbol{\beta}}_{OLS}) = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ has the minimum variance among all linear unbiased estimators.

- The $\widehat{\boldsymbol{\beta}}_{OLS}$ is a *linear* function of $\boldsymbol{Y}$. That is, $\widehat{\boldsymbol{\beta}}_{OLS} = [(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T]\boldsymbol{Y}$.

- The $\widehat{\boldsymbol{\beta}}_{OLS}$ is an *unbiased estimator* for $\boldsymbol{\beta}$. That is,

  $E(\widehat{\boldsymbol{\beta}}_{OLS}) = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\mathrm{E}(\boldsymbol{Y}) = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{\beta}$.

Based on the $\widehat{\boldsymbol{\beta}}_{OLS}$, we can devise an unbiased estimator $\hat{\sigma}^2$ for σ^2, given by

$$\begin{aligned}\hat{\sigma}^2 &= \frac{1}{n-(p+1)}\sum_{i=1}^{n}(y_i - \boldsymbol{x_i^T}\widehat{\boldsymbol{\beta}})^2 \\ &= \frac{1}{n-(p+1)}(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})^T(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}) \\ &= \frac{\mathrm{SSE}}{n-(p+1)},\end{aligned}$$

where $(p+1)$ is equal to the number of $\boldsymbol{\beta}^{'s}$. Moreover, $\mathrm{E}(\hat{\sigma}^2) = \sigma^2$.

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

$$\begin{aligned}\mathcal{L}(\boldsymbol{\beta}, \sigma^2) &= \prod_{i=1}^{n}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(y_i - \boldsymbol{x_i^T}\boldsymbol{\beta})^2}{2\sigma^2}\right) \\ &= \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}}\exp\left(-\frac{(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})}{2\sigma^2}\right)\end{aligned}$$

and the log-likelihood function is given by

$$\begin{aligned}\ell(\boldsymbol{\beta}, \sigma^2) &= ln\mathcal{L}(\boldsymbol{\beta}, \sigma^2) \\ &= -\frac{n}{2}ln(2\pi) - -\frac{n}{2}ln(\sigma^2) - \frac{1}{2\sigma^2}\left(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\right)^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}).\end{aligned}$$

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

$$\begin{cases} \frac{\partial \ell(\boldsymbol{\beta},\sigma^2)}{\partial \boldsymbol{\beta}} = 0 \\ \frac{\partial \ell(\boldsymbol{\beta},\sigma^2)}{\partial \sigma^2} = 0 \end{cases} \implies \begin{cases} \widehat{\boldsymbol{\beta}}_{ML} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} \\ \hat{\sigma}^2_{ML} = \frac{1}{n}(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})^T(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}). \end{cases}$$

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

1.2 The issue of multi-collinearity

Let B be the independent variables and E_i be the set of all the independent variables expect variable X_i, Therefore, R_{YB} is the correlation between the dependent variable Y and the independent variables B, $R^2_{X_iE_i}$ is the coefficient of multiple determination of the independent variable X_i on all other independent variables in E_i. The standard error of the variable X_i is

$$S_{b_i} = \frac{S_y}{S_x}\sqrt{\frac{1 - R^2_{YB}}{(1 - R^2_{X_iE_i})(N - P - 1)}} \quad .$$

Where N is the number of observations and P is the number of coefficients. As we see, as $R^2_{X_iE_i} \to \pm 1$, $S_{b_i} \to \infty$. This is called the issue of multi-collinearity.

What are the causes of multi-collinearity?

- Some explanatory variables are computed from other explanatory variables. (e.g. the price of the house and the loan amount are both included in the explanatory variables.)
- Two or more variables measure the same object. (e.g. weight in pounds and weight in kilogram are both included in the explanatory variables.)

- The explanatory variables are truly highly correlated.

What are the consequences of multi-collinearity?

- Increase the estimators' standard errors.
- Produce confusing or even misleading results.

How do we detect and deal with multi-collinearity?

- Check the correlations between explanatory variables and keep only one explanatory variable in the model if some of the explanatory variables are highly correlated.
- Calculate the variance inflation factors (VIF [1]) and carry out formal multi-collinearity tests [2].
- Do penalized regression.

1.3 Objectives and scope of research

The objective of this thesis is to compare the numerical results for the penalized regression methods, which include ridge regression method, lasso regression method, elastic net regression method, CAR (Correlation Adjusted Regression) method and $CAEN$ (Correlation Adjusted Elastic Net) regression method. The CAR method and $CAEN$ method were introduced by Qier Tan (2012).

[1] $VIF = \frac{1}{1-R_i^2}$, where R_i^2 = is the multiple coefficient of determination in a regression of the X_i on all other explanatory variables.

[2] When $VIF > 5$, the multi-collinearity is high.

1.4 Structure of the thesis

In Chapter 1, we review the general background of ordinary least squares regression and explain the issue of multi-collinearity, which is the fundamental purpose of penalized regression.

In Chapter 2, we give the detailed introduction and explanation of penalized regression methods, including ridge regression method, lasso regression method, elastic net regression method, CAR method and $CAEN$ method.

In Chapter 3, we use R to compute the leave one out cross validation ($LOOCV$) and the standard deviation of $LOOCV$ for each method. We also compute the optimal estimator and plot the path of coefficients for each method. Finally, we give the detailed summary of those penalized regression methods.

In Chapter 4, we summarize the results for this thesis and discuss the future research questions.

Chapter 2

Methods of penalized regression

2.1 Introduction

In a multiple linear regression model, $Var(\widehat{\boldsymbol{\beta}}_{OLS}) = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$. When there is severe multi-collinearity, $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ becomes large. This implies that $Var(\widehat{\boldsymbol{\beta}}_{OLS})$ is large. On the other hand, when $p \gg n$, matrix X is no longer of full rank. This implies that $\widehat{\boldsymbol{\beta}}_{OLS}$ is not unique. Based on the above two reasons, penalized regression have received a great deal of attention in recent years. The penalized regression can be defined as

$$\widehat{\boldsymbol{\beta}}_{PENALTY} = \arg\min_{\boldsymbol{\beta}\in R^p}(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}) \quad \text{subject to} \quad P(\boldsymbol{\beta}) \leq t, \tag{2.1}$$

or equivalently

$$\widehat{\boldsymbol{\beta}}_{PENALTY} = \arg\min_{\boldsymbol{\beta}\in R^p}(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}) + \lambda P(\boldsymbol{\beta}), \tag{2.2}$$

where λ is a non-negative regularization parameter. The penalty term $P(\boldsymbol{\beta})$ is a function of the parameters and depends on the method of penalized regression.

There is a one to one correspondence between λ and t making the above two equations equivalent. The intercept is not included in the above models, since we assume that the data are centered.

There are many different methods of penalized regression, such as ridge regression, lasso regression and elastic net regression. In this chapter, we would like to introduce those methods; moreover, there are two new methods called correlation adjusted regression (CAR) and correlation adjusted elastic net regression ($CAEN$) which were introduced by Qier Tan (2012).

2.2 Ridge regression

To detect multi-collinearity issue, Hoerl and Kennard (1970) introduced ridge regression. It's also called L_2 penalized regression. The ridge estimator is defined as

$$\widehat{\boldsymbol{\beta}}_{RIDGE} = \arg\min_{\boldsymbol{\beta}\in R^p}(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}) \quad \text{subject to} \quad \sum_{i=1}^{p}\beta_i^2 \le t,$$

or equivalently

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_{RIDGE} &= \arg\min_{\boldsymbol{\beta}\in R^p}(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}) + \lambda\sum_{i=1}^{p}\beta_i^2 \\ &= \arg\min_{\boldsymbol{\beta}\in R^p}(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}) + \lambda\boldsymbol{\beta}^T\boldsymbol{\beta},\end{aligned}$$

where λ is a non-negative regularization parameter.

We can take derivative of $\widehat{\boldsymbol{\beta}}_{RIDGE}$ with respect to $\boldsymbol{\beta}$, since $\widehat{\boldsymbol{\beta}}_{RIDGE}$ is a quadratic function in $\boldsymbol{\beta}$. After taking the derivative of $\widehat{\boldsymbol{\beta}}_{RIDGE}$ with respect to $\boldsymbol{\beta}$, the ridge

regression estimator can be derived as

$$\widehat{\boldsymbol{\beta}}_{RIDGE} = (\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{Y}.$$

- As $\lambda \to 0$, $\widehat{\boldsymbol{\beta}}_{RIDGE} \to \widehat{\boldsymbol{\beta}}_{OLS}$.
- As $\lambda \to \infty$, $\widehat{\boldsymbol{\beta}}_{RIDGE} \to \mathbf{0}$.

The variance of the ridge regression estimator is

$$\begin{aligned} Var(\widehat{\boldsymbol{\beta}}_{RIDGE}) &= (\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^T Var(\boldsymbol{Y})\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-1} \\ &= (\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\sigma^2. \end{aligned}$$

The bias of the ridge regression estimator is

$$\begin{aligned} Bias(\widehat{\boldsymbol{\beta}}_{RIDGE}) &= E(\widehat{\boldsymbol{\beta}}_{RIDGE}) - \boldsymbol{\beta} \\ &= (\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{\beta} \\ &= (\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}(\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I} - \lambda\boldsymbol{I})\boldsymbol{\beta} - \boldsymbol{\beta} \\ &= -\lambda(\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{\beta}. \end{aligned}$$

The mean squared error of the ridge regression estimator is

$$MSE(\widehat{\boldsymbol{\beta}}_{RIDGE}) = trace(Var(\widehat{\boldsymbol{\beta}}_{RIDGE})) + Bias^T(\widehat{\boldsymbol{\beta}}_{RIDGE})Bias(\widehat{\boldsymbol{\beta}}_{RIDGE}),$$

where the $trace(Var(\widehat{\boldsymbol{\beta}}_{RIDGE}))$ equal to the sum of the main diagonal elements of the $Var(\widehat{\boldsymbol{\beta}}_{RIDGE})$ matrix.

2.3 Lasso regression

Although ridge regression gives us the the prediction performance, it cannot delete any unnecessary coefficients. Tibshirani (1996), introduced the least absolute shrinkage and selection operator ($LASSO$) regression method. It's also called L_1 penalized regression. The name "$LASSO$" stands not only for shrinkage, but also does the variable selection. Since lasso regression does both continuous shrinkage and variable selection, it has received a great deal of attention in recent years. The lasso estimator is defined as

$$\widehat{\boldsymbol{\beta}}_{LASSO} = \arg\min_{\boldsymbol{\beta}\in R^p} (\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}) \quad \text{subject to} \quad \sum_{i=1}^{p}|\beta_i| \leq t,$$

or equivalently

$$\widehat{\boldsymbol{\beta}}_{LASSO} = \arg\min_{\boldsymbol{\beta}\in R^p} (\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}) + \lambda\sum_{i=1}^{p}|\beta_i|,$$

where λ is a non-negative regularization parameter.

- As $\lambda \to 0$, $\widehat{\boldsymbol{\beta}}_{LASSO} \to \widehat{\boldsymbol{\beta}}_{OLS}$.
- As $\lambda \to \infty$, $\widehat{\boldsymbol{\beta}}_{LASSO} \to \mathbf{0}$.

Since the lasso penalty term is no longer quadratic, there is no explicit formula for the mean squared error of the lasso estimator [1]. Generally, $Bias(\widehat{\boldsymbol{\beta}}_{LASSO})$ increases as the tuning parameter λ increases. However, $Var(\widehat{\boldsymbol{\beta}}_{LASSO})$ decreases as the tuning parameter λ increases. What are the $Bias(\widehat{\boldsymbol{\beta}}_{LASSO})$, $Var(\widehat{\boldsymbol{\beta}}_{LASSO})$ and $MSE(\widehat{\boldsymbol{\beta}}_{LASSO})$ at the extreme values?

[1] Efron et al. (2004) introduced $LARS$ to solve lasso regression.

- When $\lambda = 0$, we have $\begin{cases} Bias(\widehat{\boldsymbol{\beta}}_{LASSO}) = Bias(\widehat{\boldsymbol{\beta}}_{OLS}) = \mathbf{0} \\ Var(\widehat{\boldsymbol{\beta}}_{LASSO}) = Var(\widehat{\boldsymbol{\beta}}_{OLS}). \end{cases}$

- When $\lambda \to \infty$, we have $\begin{cases} Bias(\widehat{\boldsymbol{\beta}}_{LASSO}) \to \mathbf{0} - \boldsymbol{\beta} = -\boldsymbol{\beta} \\ Var(\widehat{\boldsymbol{\beta}}_{LASSO}) \to Var(\mathbf{0}) = \mathbf{0}. \end{cases}$

Finally, when $\lambda = 0$,

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

and when $\lambda \to \infty$,

$$\begin{aligned} MSE(\widehat{\boldsymbol{\beta}}_{LASSO}) &= trace(Var(\widehat{\boldsymbol{\beta}}_{LASSO})) + Bias^T(\widehat{\boldsymbol{\beta}}_{LASSO})Bias(\widehat{\boldsymbol{\beta}}_{LASSO}) \\ &\to 0 + (-\boldsymbol{\beta})^T(-\boldsymbol{\beta}) = \boldsymbol{\beta}^T\boldsymbol{\beta}. \end{aligned}$$

Since $Bias^T(\widehat{\boldsymbol{\beta}}_{LASSO})Bias(\widehat{\boldsymbol{\beta}}_{LASSO})$ and $trace(Var(\widehat{\boldsymbol{\beta}}_{LASSO}))$ move to opposite directions as the tuning parameter λ increases, we can choose the optimal parameter λ to make $MSE(\widehat{\boldsymbol{\beta}}_{LASSO})$ minimized theoretically.

The properties of lasso regression can be described as:

- Lasso regression does both continuous shrinkage and variable selection.
- For $p > n$, the lasso regression method selects at most n variables.
- For highly correlated explanatory variables, the lasso regression method only selects one variable among those highly correlated explanatory variables.

2.4 Elastic net regression

Zou and Hastie (2005) introduced elastic net regression which is a combination of L_1 penalized regression and L_2 penalized regression. The elastic net estimator is defined as

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_{EN} &= \arg\min_{\boldsymbol{\beta}\in R^p}(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})+\lambda_1\sum_{i=1}^{p}|\beta_i|+\lambda_2\sum_{i=1}^{p}\beta_i^2 \\ &= \arg\min_{\boldsymbol{\beta}\in R^p}(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})+\lambda_1\sum_{i=1}^{p}|\beta_i|+\lambda_2\boldsymbol{\beta}^T\boldsymbol{\beta},\end{aligned}$$

where λ_1 and λ_2 are non-negative regularization parameters.

- As $\lambda_1 \to 0$, $\widehat{\boldsymbol{\beta}}_{EN} \to \widehat{\boldsymbol{\beta}}_{RIDGE}$.
- As $\lambda_2 \to 0$, $\widehat{\boldsymbol{\beta}}_{EN} \to \widehat{\boldsymbol{\beta}}_{LASSO}$.

What's the behavior of $\widehat{\boldsymbol{\beta}}_{EN}$ at $\lambda_1 \to \infty$ or $\lambda_2 \to \infty$? According to the ridge regression and lasso regression we have proved before, the amount of shrinkage increases as λ increases. This implies that when either $\lambda_1 \to \infty$ or $\lambda_2 \to \infty$, we have $\widehat{\boldsymbol{\beta}}_{EN} \to \mathbf{0}$. Since the lasso penalty term is included in $\widehat{\boldsymbol{\beta}}_{EN}$, there is no explicit formula of the mean squared error for the elastic net estimator except when $\lambda_1 = 0$. What is the $MSE(\widehat{\boldsymbol{\beta}}_{EN})$ at each combination of extreme values?

- When $\begin{cases}\lambda_1 = 0 \\ \lambda_2 = 0\end{cases}$, we have $MSE(\widehat{\boldsymbol{\beta}}_{EN}) = MSE(\widehat{\boldsymbol{\beta}}_{OLS})$.
- When $\begin{cases}\lambda_1 = 0 \\ \lambda_2 \neq 0\end{cases}$, we have $MSE(\widehat{\boldsymbol{\beta}}_{EN}) = MSE(\widehat{\boldsymbol{\beta}}_{RIDGE})$.

- When $\begin{cases} \lambda_1 \neq 0 \\ \lambda_2 = 0 \end{cases}$, we have $MSE(\widehat{\boldsymbol{\beta}}_{EN}) = MSE(\widehat{\boldsymbol{\beta}}_{LASSO})$.

- When $\lambda_i \to \infty$, we have $MSE(\widehat{\boldsymbol{\beta}}_{EN}) \to \boldsymbol{\beta}^T\boldsymbol{\beta}$, for $i = 1$ or 2.

What's the process of elastic net regression when $\lambda_1 \neq 0$ and $\lambda_2 \neq 0$? We fixed λ_2 first, then do the elastic net regression to determine the optimal λ_1. Finally, we choose the optimal combination of λ_1 and λ_2 based on the smallest $MSE(\widehat{\boldsymbol{\beta}}_{EN})$. Due to quadratic regularization, the solution paths of elastic net regression are more stable than the solution paths of lasso regression. So elastic net regression can be regarded as a stabilized version of the lasso regression.

2.5 Correlation adjusted regression

Tan (2012) introduced the correlation adjusted regression (CAR). It's an extension of ridge regression. There are two types of correlation adjusted regression.

The 1^{st} type correlation adjusted estimator can be defined as

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}_{CAR_1} &= \arg\min_{\boldsymbol{\beta}\in R^p} (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda[\sum_{j=1}^{p-1}(\beta_j - r_{j,j+1}\beta_{j+1})^2 + \beta_p^2] \\
&= \arg\min_{\boldsymbol{\beta}\in R^p} (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda\boldsymbol{\beta}^T\boldsymbol{W}_1\boldsymbol{\beta} \\
&= \arg\min_{\boldsymbol{\beta}\in R^p} (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda\boldsymbol{\beta}^T D_1^T D_1\boldsymbol{\beta},
\end{aligned}$$

where λ is a non-negative regularization parameter and

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

The $r_{i,j}$ is the sample correlation between the predictor variables $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$. When $r_{i,i+1} = 0$ for all $i = 1, 2, \cdots, p-1$, we have $CAR_1 = RIDGE$.

The 2^{nd} type correlation adjusted estimator can be defined as

$$\begin{aligned} \widehat{\boldsymbol{\beta}}_{CAR_2} &= \arg\min_{\boldsymbol{\beta} \in R^p} (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda [\sum_{j=1}^{p-1} \sum_{k>j} (\beta_j - r_{j,k}\beta_k)^2 + \beta_p^2] \\ &= \arg\min_{\boldsymbol{\beta} \in R^p} (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda \boldsymbol{\beta}^T \boldsymbol{W}_2 \boldsymbol{\beta} \\ &= \arg\min_{\boldsymbol{\beta} \in R^p} (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda \boldsymbol{\beta}^T D_2^T D_2 \boldsymbol{\beta}, \end{aligned}$$

where λ is a non-negative regularization parameter and

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

The $r_{i,j}$ is the sample correlation between the predictor variables $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$.

Finally, the correlation adjusted estimator can be defined as

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_{CAR} &= \arg\min_{\boldsymbol{\beta}\in R^p}(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})+\lambda\boldsymbol{\beta}^T\boldsymbol{W}\boldsymbol{\beta} \\ &= \arg\min_{\boldsymbol{\beta}\in R^p}(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})+\lambda\boldsymbol{\beta}^T D^T D\boldsymbol{\beta},\end{aligned}$$

where $\boldsymbol{W}$ either is $\boldsymbol{W}_1$ or $\boldsymbol{W}_2$ and $\boldsymbol{W}_K = D_K^T D_K$ for $K = 1, 2$. When $\boldsymbol{W} = \boldsymbol{I}$, $CAR = RIDGE$. So ridge regression can be regarded as a special version of CAR.

We can take derivative of $\widehat{\boldsymbol{\beta}}_{CAR}$ with respect to $\boldsymbol{\beta}$, since $\widehat{\boldsymbol{\beta}}_{CAR}$ is a quadratic function in $\boldsymbol{\beta}$. After taking the derivative of $\widehat{\boldsymbol{\beta}}_{CAR}$ with respect to $\boldsymbol{\beta}$, the correlation adjusted estimator can be derived as

$$\widehat{\boldsymbol{\beta}}_{CAR} = (\boldsymbol{X}^T\boldsymbol{X}+\lambda\boldsymbol{W})^{-1}\boldsymbol{X}^T\boldsymbol{Y}.$$

- As $\lambda \to 0$, $\widehat{\boldsymbol{\beta}}_{CAR} \to \widehat{\boldsymbol{\beta}}_{OLS}$.
- As $\lambda \to \infty$, $\widehat{\boldsymbol{\beta}}_{CAR} \to \mathbf{0}$.

The variance of the correlation adjusted estimator is

$$\begin{aligned}Var(\widehat{\boldsymbol{\beta}}_{CAR}) &= (\boldsymbol{X}^T\boldsymbol{X}+\lambda\boldsymbol{W})^{-1}\boldsymbol{X}^T Var(\boldsymbol{Y})\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X}+\lambda\boldsymbol{W})^{-1} \\ &= (\boldsymbol{X}^T\boldsymbol{X}+\lambda\boldsymbol{W})^{-1}\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X}+\lambda\boldsymbol{W})^{-1}\sigma^2.\end{aligned}$$

The bias of the correlation adjusted estimator is

$$\begin{aligned}Bias(\widehat{\boldsymbol{\beta}}_{CAR}) &= E(\widehat{\boldsymbol{\beta}}_{CAR})-\boldsymbol{\beta} \\ &= (\boldsymbol{X}^T\boldsymbol{X}+\lambda\boldsymbol{W})^{-1}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{\beta} \\ &= (\boldsymbol{X}^T\boldsymbol{X}+\lambda\boldsymbol{W})^{-1}(\boldsymbol{X}^T\boldsymbol{X}+\lambda\boldsymbol{W}-\lambda\boldsymbol{W})\boldsymbol{\beta}-\boldsymbol{\beta} \\ &= -\lambda(\boldsymbol{X}^T\boldsymbol{X}+\lambda\boldsymbol{W})^{-1}\boldsymbol{W}\boldsymbol{\beta}.\end{aligned}$$

The mean squared error of the correlation adjusted estimator is

$$MSE(\widehat{\boldsymbol{\beta}}_{CAR}) = trace(Var(\widehat{\boldsymbol{\beta}}_{CAR})) + Bias^T(\widehat{\boldsymbol{\beta}}_{CAR})Bias(\widehat{\boldsymbol{\beta}}_{CAR}),$$

where the $trace(Var(\widehat{\boldsymbol{\beta}}_{CAR}))$ equal to the sum of the main diagonal elements of the $Var(\widehat{\boldsymbol{\beta}}_{CAR})$ matrix.

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

$$\boldsymbol{X}^* = \frac{1}{\sqrt{1+\lambda}} \begin{pmatrix} \boldsymbol{X} \\ \sqrt{\lambda} C^T \end{pmatrix}, \ \boldsymbol{Y}^* = \begin{pmatrix} \boldsymbol{Y} \\ \boldsymbol{0} \end{pmatrix}, \ \boldsymbol{\beta}^* = \sqrt{1+\lambda}\boldsymbol{\beta}.$$

Tan (2012) proved that minimizing

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

is equivalent to minimizing

$$(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda\boldsymbol{\beta}^T\boldsymbol{W}\boldsymbol{\beta} = CAR.$$

Proof.

$$
\begin{aligned}
OLS^* &= \sum_{i=1}^{n+p}(y_i^* - x_i^*\beta^*)^2 \\
&= (\boldsymbol{Y}^* - \boldsymbol{X}^*\boldsymbol{\beta}^*)^T(\boldsymbol{Y}^* - \boldsymbol{X}^*\boldsymbol{\beta}^*) \\
&= [(\boldsymbol{Y}^*)^T - (\boldsymbol{\beta}^*)^T(\boldsymbol{X}^*)^T][\boldsymbol{Y}^* - \boldsymbol{X}^*\boldsymbol{\beta}^*] \\
&= (\boldsymbol{Y}^*)^T\boldsymbol{Y}^* - (\boldsymbol{Y}^*)^T\boldsymbol{X}^*\boldsymbol{\beta}^* - (\boldsymbol{\beta}^*)^T(\boldsymbol{X}^*)^T\boldsymbol{Y}^* + (\boldsymbol{X}^*\boldsymbol{\beta}^*)^T\boldsymbol{X}^*\boldsymbol{\beta}^* \\
&= \begin{pmatrix}\boldsymbol{Y}^T & \mathbf{0}\end{pmatrix}\begin{pmatrix}\boldsymbol{Y} \\ \mathbf{0}\end{pmatrix} - \begin{pmatrix}\boldsymbol{Y}^T & \mathbf{0}\end{pmatrix}\frac{1}{\sqrt{1+\lambda}}\begin{pmatrix}\boldsymbol{X} \\ \sqrt{\lambda}C^T\end{pmatrix}\sqrt{1+\lambda}\boldsymbol{\beta} \\
&\quad - \sqrt{1+\lambda}\boldsymbol{\beta}^T\frac{1}{\sqrt{1+\lambda}}\begin{pmatrix}\boldsymbol{X}^T & \sqrt{\lambda}C\end{pmatrix}\begin{pmatrix}\boldsymbol{Y} \\ \mathbf{0}\end{pmatrix} \\
&\quad + \sqrt{1+\lambda}\boldsymbol{\beta}^T\frac{1}{\sqrt{1+\lambda}}\begin{pmatrix}\boldsymbol{X}^T & \sqrt{\lambda}C\end{pmatrix}\frac{1}{\sqrt{1+\lambda}}\begin{pmatrix}\boldsymbol{X} \\ \sqrt{\lambda}C^T\end{pmatrix}\sqrt{1+\lambda}\boldsymbol{\beta} \\
&= \boldsymbol{Y}^T\boldsymbol{Y} - \boldsymbol{Y}^T\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{Y} + \boldsymbol{\beta}^T(\boldsymbol{X}^T\boldsymbol{X} + \lambda CC^T)\boldsymbol{\beta} \\
&= (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda\boldsymbol{\beta}^T\boldsymbol{W}\boldsymbol{\beta} \\
&= CAR.
\end{aligned}
$$

2.6 Correlation adjusted elastic net regression

Tan (2012) introduced the correlation adjusted elastic net regression ($CAEN$) which is a combination of L_1 penalized regression and CAR. It's also an extension of elastic net regression. There are two types of correlation adjusted elastic net regression.

The 1^{st} type of correlation adjusted elastic net estimator can be defined as

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_{CAEN_1} &= \arg\min_{\boldsymbol{\beta}\in R^p}(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})+\lambda_1\sum_{i=1}^{p}|\beta_i| \\ &\quad +\lambda_2[\sum_{j=1}^{p-1}(\beta_j - r_{j,j+1}\beta_{j+1})^2+\beta_p^2] \\ &= \arg\min_{\boldsymbol{\beta}\in R^p}(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})+\lambda_1\sum_{i=1}^{p}|\beta_i|+\lambda_2\boldsymbol{\beta}^T\boldsymbol{W}_1\boldsymbol{\beta} \\ &= \arg\min_{\boldsymbol{\beta}\in R^p}(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})+\lambda_1\sum_{i=1}^{p}|\beta_i|+\lambda_2\boldsymbol{\beta}^T D_1^T D_1\boldsymbol{\beta},\end{aligned}$$

where λ_1 and λ_2 are non-negative regularization parameters and

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

The $r_{i,j}$ is the sample correlation between the predictor variables $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$. When $r_{i,i+1}=0$ for all $i=1,2,\cdots,p-1$, we have $CAEN_1 = EN$.

The 2^{nd} type of correlation adjusted elastic net estimator can be defined as

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}_{CAEN_2} &= \arg\min_{\boldsymbol{\beta}\in R^p}(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}) + \lambda_1\sum_{i=1}^{p}|\beta_i| \\
&\quad + \lambda_2[\sum_{j=1}^{p-1}\sum_{k>j}(\beta_j - r_{j,k}\beta_k)^2 + \beta_p^2] \\
&= \arg\min_{\boldsymbol{\beta}\in R^p}(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}) + \lambda_1\sum_{i=1}^{p}|\beta_i| + \lambda_2\boldsymbol{\beta}^T\boldsymbol{W}_2\boldsymbol{\beta} \\
&= \arg\min_{\boldsymbol{\beta}\in R^p}(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}) + \lambda_1\sum_{i=1}^{p}|\beta_i| + \lambda_2\boldsymbol{\beta}^T D_2^T D_2\boldsymbol{\beta},
\end{aligned}$$

where λ_1 and λ_2 are non-negative regularization parameters and

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

The $r_{i,j}$ is the sample correlation between the predictor variables $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$.

Finally, the correlation adjusted elastic net estimator can be defined as

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

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

- As $\lambda_1 \to 0$, $\widehat{\boldsymbol{\beta}}_{CAEN} \to \widehat{\boldsymbol{\beta}}_{CAR}$.
- As $\lambda_2 \to 0$, $\widehat{\boldsymbol{\beta}}_{CAEN} \to \widehat{\boldsymbol{\beta}}_{LASSO}$.

The $\boldsymbol{W}$ either be $\boldsymbol{W}_1$ or $\boldsymbol{W}_2$ and $\boldsymbol{W}_K = D_K^T D_K$ for $K = 1, 2$. When $\boldsymbol{W} = \boldsymbol{I}$, $CAEN = EN$. So elastic net regression can be regarded as a special version of $CAEN$.

The behavior of $\widehat{\boldsymbol{\beta}}_{CAEN}$ at $\lambda_1 \to \infty$ or $\lambda_2 \to \infty$ is similar to that of $\widehat{\boldsymbol{\beta}}_{EN}$. According to the CAR regression and lasso regression we proved before, the amount of shrinkage increases as λ increases. This means that when either $\lambda_1 \to \infty$ or $\lambda_2 \to \infty$, we have $\widehat{\boldsymbol{\beta}}_{CAEN} \to \mathbf{0}$. Since the lasso penalty term is included in $\widehat{\boldsymbol{\beta}}_{CAEN}$, there is no explicit formula of the mean squared error for the correlation adjusted elastic net estimator except when $\lambda_1 = 0$. What is $MSE(\widehat{\boldsymbol{\beta}}_{CAEN})$ at each combination of extreme values?

- When $\begin{cases}\lambda_1 = 0 \\ \lambda_2 = 0\end{cases}$, we have $MSE(\widehat{\boldsymbol{\beta}}_{CAEN}) = MSE(\widehat{\boldsymbol{\beta}}_{OLS})$.
- When $\begin{cases}\lambda_1 = 0 \\ \lambda_2 \neq 0\end{cases}$, we have $MSE(\widehat{\boldsymbol{\beta}}_{CAEN}) = MSE(\widehat{\boldsymbol{\beta}}_{CAR})$.

- When $\begin{cases} \lambda_1 \neq 0 \\ \lambda_2 = 0 \end{cases}$, we have $MSE(\widehat{\boldsymbol{\beta}}_{CAEN}) = MSE(\widehat{\boldsymbol{\beta}}_{LASSO})$.

- When $\lambda_i \to \infty$, we have $MSE(\widehat{\boldsymbol{\beta}}_{CAEN}) \to \boldsymbol{\beta}^T\boldsymbol{\beta}$, for i=1 or 2.

What is the process of correlation adjusted elastic net regression when $\lambda_1 \neq 0$ and $\lambda_2 \neq 0$? We fixed λ_2 first, then 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 $MSE(\widehat{\boldsymbol{\beta}}_{CAEN})$. Due to quadratic regularization, the solution paths of $CAEN$ are more stable than the solution paths of lasso regression. So $CAEN$ can also be regarded as a stabilized version of the lasso regression.

Given the Cholesky's decomposition $\boldsymbol{W} = CC^T$ and for any $\lambda_1, \lambda_2 > 0$, define

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

Tan (2012) proved that minimizing

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

is equivalent to minimizing

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

Proof.

$$
\begin{aligned}
LASSO^* &= (\boldsymbol{Y}^* - \boldsymbol{X}^*\boldsymbol{\beta}^*)^T(\boldsymbol{Y}^* - \boldsymbol{X}^*\boldsymbol{\beta}^*) + \gamma \sum_{i=1}^{p} |\beta_i^*| \\
&= OLS^* + \gamma \sum_{i=1}^{p} |\beta_i^*| \\
&= CAR + \lambda_1 \sum_{i=1}^{p} |\beta_i| \\
&= (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda_2 \boldsymbol{\beta}^T \boldsymbol{W} \boldsymbol{\beta} + \lambda_1 \sum_{i=1}^{p} |\beta_i| \\
&= CAEN.
\end{aligned}
$$

2.7 Summary

When there exists multi-collinearity, the ordinary least squares (OLS) regression may produce large variance based on $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ and contribute to $MSE(\widehat{\boldsymbol{\beta}}_{OLS})$. This causes our model very unstable or highly variable.

By shrinking the coefficients, ridge regression reduces the variability. It releases a little bias in exchange for reduced variability. Ridge regression shrinks the coefficients towards zero simultaneously. If there are many predictors in the model, ridge regression cannot provide a sparse model which can be easily interpreted.

The lasso regression reduces the variability by shrinking the coefficients toward zero and shrinks some coefficients to exactly zero. It makes up the disadvantage of ridge regression but still have some limitations. For $p > n$, the lasso regression

selects at most n variables. When there exists high correlations among explanatory variables, the lasso only selects one explanatory variable among those highly correlated explanatory variables. Lasso does not perform group variable selection.

The elastic net regression makes up the disadvantage of ridge regression and lasso regression. It does shrinkage, variable selection and group variable selection. Elastic net regression is a better method than both ridge regression and lasso regression.

The correlation adjusted regression is an extension of ridge regression. The behavior of the correlation adjusted regression is similar with ridge regression. The sample correlation is included in the penalty term. After applying argumentation to the data set, the correlation adjusted regression can be reduced to the ordinary least squares regression.

The correlation adjusted elastic net regression is an extension of elastic net regression. The behavior of the correlation adjusted elastic net regression is similar with elastic net regression. The sample correlation is also included in the penalty term. After applying argumentation to the data set, the correlation adjusted elastic net regression can be reduced to the lasso regression.

Chapter 3

A numerical comparison of penalization methods

In this chapter, we perform numerical study of the penalized regression methods. We use the diabetes data frame [1] which includes 442 rows and 11 columns. Those 442 rows come from the 442 patients and those 11 columns correspond to 10 independent variables and 1 dependent variable.

3.1 Introduction

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

$$\begin{aligned} MSE(\hat{\beta}) &= E[(\hat{\beta} - \beta)^2] \\ &= Var(\hat{\beta}) + [bias(\hat{\beta})]^2. \end{aligned}$$

As we see, $Var(\hat{\beta})$ measures the variability of the estimator and $bias(\hat{\beta})$ measures the bias. Therefore, to find a good estimator we need to find the estimator with

[1] The data set is available at: http://www.stanford.edu/ hastie/Papers/LARS/diabetes.data

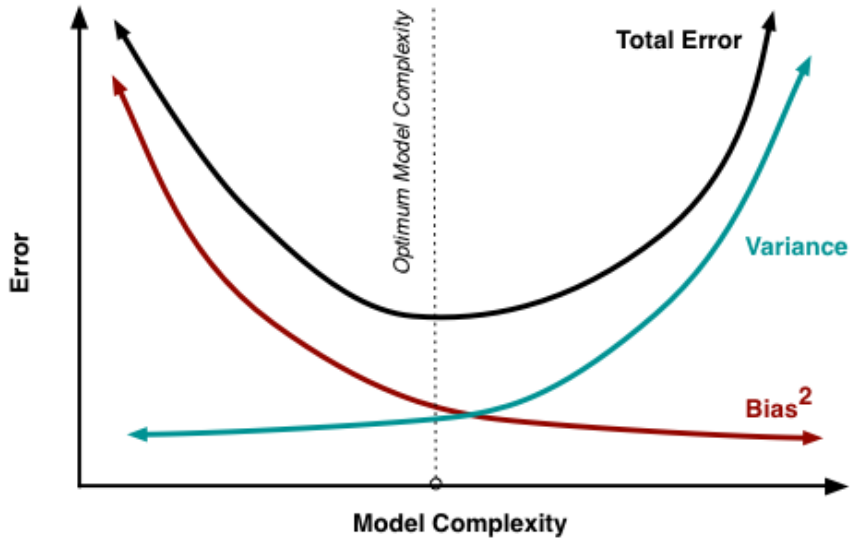


Figure 3.1: trade-off relationships

the smallest mean square error. There is a trade-off between $Var(\hat{\beta})$ and $[bias(\hat{\beta})]^2$. This implies that we can increase a little bias of the estimator in exchange of a large decrease in the variance. After adjustment, the model may bias a little bit but is more stable.

Figure 3.1 [2]shows the trade-off relationship between $Var(\hat{\beta})$ and $[bias(\hat{\beta})]^2$. This implies the simple model with high bias and low variance. However, the complex model has low bias and high variance.

3.2 Methods

In the last chapter, we have discussed that there is no explicit formula for the mean square error whenever the lasso penalty term is included in $P(\boldsymbol{\beta})$. How do we select the best λ? There are several criteria available for selecting the best λ, such as the

[2]This graph is available at: http://scott.fortmann-roe.com/docs/BiasVariance.html

Bayesian Information Criterion (BIC), Alkaike Information Criterion (AIC) and Leave-One-Out Cross Validation ($LOOCV$) criterion. In this thesis, we will use the leave one out cross validation criterion.

Suppose there are n observations. If the i^{th} row observation $(y_i, x_{i1}, x_{i2}, \cdots, x_{ip})$ is removed from the data set, define the X matrix without the i^{th} row as $X_{(i)}$ and the Y matrix without i^{th} element as $Y_{(i)}$. For $i = 1, 2, \cdots, n$, obtain $\widehat{\boldsymbol{\beta}}_{(n-1)}$ using $X_{(i)}$ and $Y_{(i)}$, then the prediction error square for $(y_i, x_{i1}, x_{i2}, \cdots, x_{ip})$ is

$$\begin{aligned} CV_{-i}(\lambda) &= (y_i - X_i\widehat{\boldsymbol{\beta}}_{(n-1)})^2 \\ &= (y_i - \sum_{j=1}^{p} x_{ij}\widehat{\beta}_{(n-1)})^2, \quad i = 1, 2, \cdots, n. \end{aligned}$$

Finally, the $LOOCV$ is

$$CV(\lambda) = \frac{1}{n}\sum_{i=1}^{n} CV_{-i}(\lambda).$$

The standard deviation of the $LOOCV$ is

$$S_{CV(\lambda)} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}[CV_{-i}(\lambda) - CV(\lambda)]^2}.$$

Our target is to find the $\lambda_{CV_{optimal}}$ that makes both $CV(\lambda)$ and $S_{CV(\lambda)}$ optimal which means being as small as well. Then we use the $\lambda_{CV_{optimal}}$ to obtain the penalized estimators.

3.3 Numerical results and explanation

All the simulations are computed by R. There are five types of penalized regression included in this thesis (ridge regression, lasso regression, elastic net regression, correlation adjusted regression and correlation adjusted elastic net regression).

3.3.1 Preliminary material

To reduce the error of the least square estimation and omit β_0, we standardize our data frame. The original data frame are displayed as follows [3]:

Table 3.1: original data frame

	age	sex	bmi	$\cdots$	tch	ltg	glu	y
1	59.00	2.00	32.10	$\cdots$	4.00	4.86	87.00	151.00
2	48.00	1.00	21.60	$\cdots$	3.00	3.89	69.00	75.00
3	72.00	2.00	30.50	$\cdots$	4.00	4.67	85.00	141.00
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
440	60.00	2.00	24.90	$\cdots$	3.77	4.13	95.00	132.00
441	36.00	1.00	30.00	$\cdots$	4.79	5.13	85.00	220.00
442	36.00	1.00	19.60	$\cdots$	3.00	4.60	92.00	57.00
$E(column)$	48.518	1.468	26.376	$\cdots$	4.070	4.641	91.260	152.134
S_{column}	13.109	0.500	4.418	$\cdots$	1.290	0.522	11.496	77.093

where $E(column)$ equal to the mean of the variables and S_{column} equal to the standard deviation of the variables. For each observation, we subtract its column mean and divide by its column standard deviation. The standardized data frame are displayed as follows:

[3] Based on Bradley Efron, Trevor Hastie, Iain Johnstone and Robert Tibshirani (2004) "Least Angle Regression.", the 10 independent variables' name are age, sex, body mass index, average blood pressure, and six blood serum measurements. The response variable is a quantitative measure of disease progression one year after baseline.

Table 3.2: standardized data frame

	age	sex	bmi	map	tc	ldl	hdl	tch	ltg	glu	y
1	0.80	1.06	1.30	0.46	-0.93	-0.73	-0.91	-0.05	0.42	-0.37	-0.01
2	-0.04	-0.94	-1.08	-0.55	-0.18	-0.40	1.56	-0.83	-1.43	-1.94	-1.00
3	1.79	1.06	0.93	-0.12	-0.96	-0.72	-0.68	-0.05	0.06	-0.54	-0.14
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
440	0.88	1.06	-0.33	0.36	-0.78	-0.29	-0.52	-0.23	-0.98	0.33	-0.26
441	-0.95	-0.94	0.82	0.03	0.34	0.32	-0.60	0.56	0.94	-0.54	0.88
442	-0.95	-0.94	-1.53	-1.71	1.76	0.58	3.65	-0.83	-0.09	0.06	-1.23

Since the purpose of investigating the penalized estimator behavior is to compare with the ordinary least squares estimator, we give the detailed numerical results and explanation of $\widehat{\boldsymbol{\beta}}_{OLS}$.

Firstly, we calculate the numerical results by the ordinary least squares regression method using our data set.

Table 3.3: ordinary least squares regression

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.0000	0.0334	-0.00	1.0000
age	-0.0062	0.0369	-0.17	0.8670
sex	-0.1481	0.0378	-3.92	0.0001
bmi	0.3211	0.0411	7.81	0.0000
map	0.2004	0.0404	4.96	0.0000
tc	-0.4893	0.2574	-1.90	0.0579
ldl	0.2945	0.2094	1.41	0.1604
hdl	0.0624	0.1313	0.48	0.6347
tch	0.1094	0.0997	1.10	0.2735
ltg	0.4641	0.1062	4.37	0.0000
glu	0.0418	0.0408	1.02	0.3060

```
Residual standard error: 0.7025 on 431 degrees of freedom
Multiple R-squared: 0.5177,     Adjusted R-squared: 0.5066
F-statistic: 46.27 on 10 and 431 DF,  p-value: < 2.2e-16.
```

Table 3.3 gives the estimated parameters, standard deviations, t-values and corresponding p-values. We also calculate the $\sum_{i=1}^{p} |\beta_i| = 2.137$, $\sum_{i=1}^{p} \beta_i^2 = 0.724$ and $LOOCV = 0.505$, which can be used to compare with other penalized regressions later. Since $\widehat{\boldsymbol{\beta}}_{OLS}$ is an unbiased estimator of $\boldsymbol{\beta}$ and excluding λ, the $LOOCV$ is just a constant. Actually, we can calculate the mean square errors of $\widehat{\boldsymbol{\beta}}_{OLS}$ directly. However, we want to use the same method ($LOOCV$) to compare with the penalized methods.

Secondly, we calculate the correlation matrix and obtain D_1 and D_2 which are used in CAR and $CAEN$.

Table 3.4: correlation matrix among independent variables

	age	sex	bmi	map	tc	ldl	hdl	tch	ltg	glu
age	1.00	0.17	0.19	0.34	0.26	0.22	-0.08	0.20	0.27	0.30
sex	0.17	1.00	0.09	0.24	0.04	0.14	-0.38	0.33	0.15	0.21
bmi	0.19	0.09	1.00	0.40	0.25	0.26	-0.37	0.41	0.45	0.39
map	0.34	0.24	0.40	1.00	0.24	0.19	-0.18	0.26	0.39	0.39
tc	0.26	0.04	0.25	0.24	1.00	0.90	0.05	0.54	0.52	0.33
ldl	0.22	0.14	0.26	0.19	0.90	1.00	-0.20	0.66	0.32	0.29
hdl	-0.08	-0.38	-0.37	-0.18	0.05	-0.20	1.00	-0.74	-0.40	-0.27
tch	0.20	0.33	0.41	0.26	0.54	0.66	-0.74	1.00	0.62	0.42
ltg	0.27	0.15	0.45	0.39	0.52	0.32	-0.40	0.62	1.00	0.46
glu	0.30	0.21	0.39	0.39	0.33	0.29	-0.27	0.42	0.46	1.00

$$
\begin{aligned}
D_1 &= \begin{pmatrix}
1 & -r_{1,2} & 0 & \cdots & 0 & 0 \\
0 & 1 & -r_{2,3} & \cdots & 0 & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
0 & 0 & 0 & \cdots & 1 & -r_{p-1,p} \\
0 & 0 & 0 & \cdots & 0 & 1
\end{pmatrix} \\
&= \begin{pmatrix}
1 & -0.17 & 0 & \cdots & 0 & 0 \\
0 & 1 & -0.09 & \cdots & 0 & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
0 & 0 & 0 & \cdots & 1 & -0.46 \\
0 & 0 & 0 & \cdots & 0 & 1
\end{pmatrix},
\end{aligned}
$$

and

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

$$= \begin{pmatrix} 1 & -0.17 & 0 & 0 & \cdots & 0 & 0 \\ 1 & 0 & -0.19 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 0 & \cdots & 0 & -0.30 \\ 0 & 1 & -0.09 & 0 & \cdots & 0 & 0 \\ 0 & 1 & 0 & -0.24 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 1 & 0 & 0 & \cdots & 0 & -0.21 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 & -0.46 \\ 0 & 0 & 0 & 0 & \cdots & 0 & 1 \end{pmatrix}.$$

Thirdly, for the correlation adjusted regression (CAR) and the correlation adjusted elastic net regression ($CAEN$), we use the OLS and the $LASSO$ to calculate the numerical results. We update the data set and calculate the Cholesky's decomposition $\boldsymbol{W} = CC^T$.

Finally, we introduce the R package which is called $glmnet$ [4](Authors: Jerome Friedman, Trevor Hastie and Rob Tibshirani). Most of those numerical results are

[4]This package is available at: http://cran.r-project.org/web/packages/glmnet/index.html

calculated by this package. This package defines the penalized term as

$$\begin{aligned} P_\lambda(\boldsymbol{\beta}) &= \lambda P_\alpha(\boldsymbol{\beta}) \\ &= \lambda \sum_{i=1}^{p} [\frac{1}{2}(1-\alpha)\beta_i^2 + \alpha\,|\beta_i|], \end{aligned}$$

- $\alpha = 1 \rightarrow$ lasso method,
- $\alpha = 0 \rightarrow$ ridge regression method,
- $0 < \alpha < 1 \rightarrow$ elastic net regression method.

3.3.2 Ridge regression

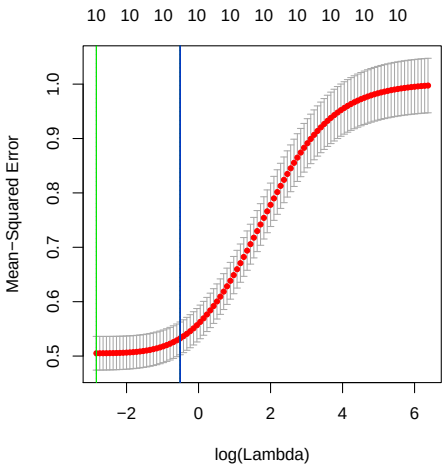


Figure 3.2: ridge $LOOCV$ plot

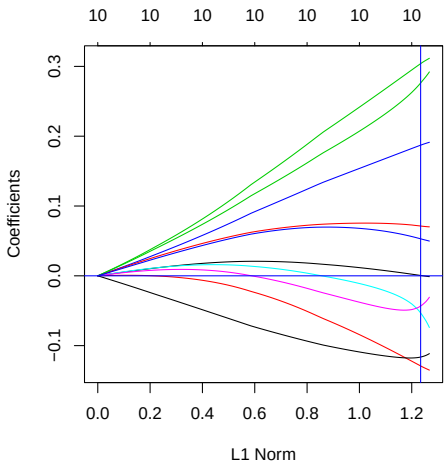


Figure 3.3: coefficients path

Figure 3.2 gives the relationship between $ln\lambda$ and $LOOCV$. The integer numbers at the top of this graphic show the number of non-zero estimators in the model. The left line gives the smallest $LOOCV$ and the right line gives the less complex model. We can pick up any λ between the left line and the right line.

Table 3.5: ridge numerical results

λ	$LOOCV$	Std. Error	$\sum_{i=1}^{p} \beta_i^2$
585.786	0.9973818	0.05012704	0
16	0.8677504	0.04402528	0.00383
2	0.6180950	0.03296649	0.05699633
1	0.5596442	0.03080430	0.09978268
0.5	0.5051091	0.03104667	0.2691374
0.02	**0.5050562**	**0.03129740**	**0.3104395**
0.01	**0.5049250**	**0.03138451**	**0.3653977**
0	0.5050890	0.03152150	0.7239687

Based on the result from Table 3.5, either $\lambda = 0.01$ or $\lambda = 0.02$ is our optimal value. $\lambda = 0.01$ gives the smaller $LOOCV$, but $\lambda = 0.02$ gives the smaller $Std.Error$ [5]. Finally, we plot the path of the coefficient [6]. As we see from Figure 3.3, the optimal ridge estimators are at the intersections of the vertical line and each coefficient path. The ridge regression only does the variables shrinkage but not the variable selection where L_1 norm equal to $\sum_{i=1}^{p} |\beta_i|$.

[5] It's the standard error of $LOOCV$ and the smaller standard error gives the less complex model.

[6] For each coefficient path line, the variables from top to bottom are: ltg, bmi, ldl, map, tch, hdl, glu, age, sex and tc. They are in the same order for other methods of coefficients path.

3.3.3 Lasso regression

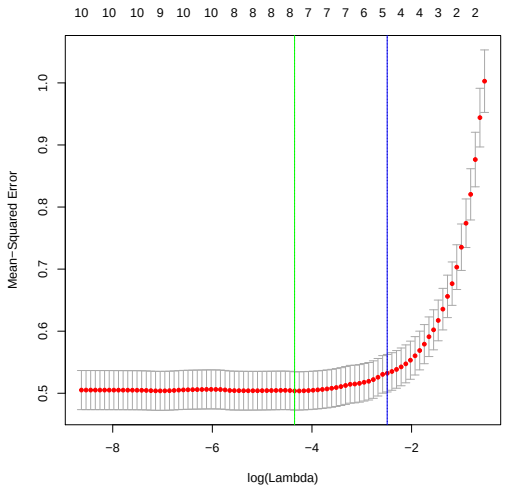


Figure 3.4: lasso $LOOCV$ plot

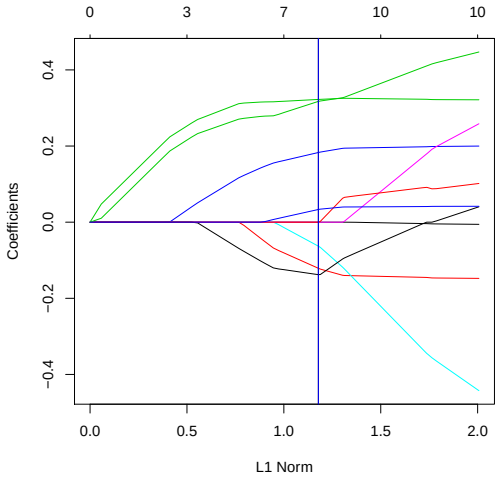


Figure 3.5: coefficients path

Figure 3.4 gives the relationship between $ln\lambda$ and $LOOCV$. The integer numbers at the top of this graphic show the number of non-zero estimators in the model. The left line gives the smallest $LOOCV$ with 7 variables in the model and the right line gives smallest standard deviation with only 4 variables in the model. We can pick any λ between the left line and the right line.

Table 3.6: lasso numerical results

λ	$LOOCV$	Std. Error	$\sum_{i=1}^{p} \lvert\beta_i\rvert$	Df
58.579×10^{-2}	1.003	0.0504	0	0
10.00×10^{-2}	0.538	0.0303	0.734	4
$\mathbf{1.292 \times 10^{-2}}$	**0.504**	**0.0309**	**1.178**	**7**
0.614×10^{-2}	0.504	0.0312	1.267	8
0.096×10^{-2}	0.504	0.0314	1.748	9
0.038×10^{-2}	0.505	0.0315	1.917	10
0.018×10^{-2}	0.505	0.0315	2.007	10

Based on Table 3.6, when $\lambda = 1.292 \times 10^{-2}$, both $LOOCV$ and $Std.Error$ are at

minimum and there are 7 variables included in the model. Since the lasso regression does the variable selection, we also indicate the number of non-zero variables in the model which is called Df. Finally, we plot the path of the coefficients. As we see from Figure 3.5, the optimal lasso estimators are at the intersections of the vertical line and each coefficient path. The lasso regression does not only the estimators' shrinkage but also variable selection.

Figure 3.6: sequence of lasso moves

```
Sequence of LASSO moves:
     bmi ltg map hdl sex glu tc tch ldl age hdl hdl
Var    3   9   4   7   2  10  5   8   6   1  -7   7
Step   1   2   3   4   5   6  7   8   9  10  11  12
```

Figure 3.6 shows the sequence of lasso moves. In each step, either one variable is added (with a positive number) or one variable is deleted (with a negative number).

3.3.4 Elastic net regression

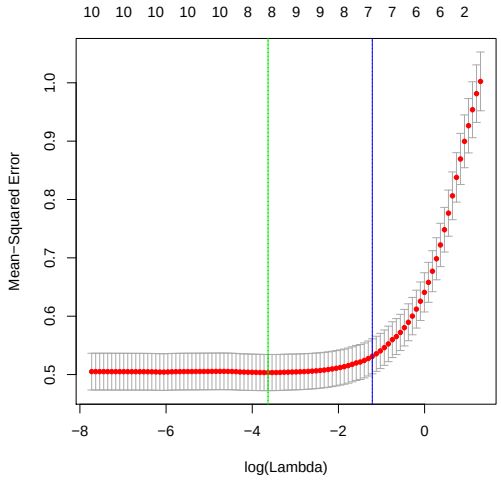


Figure 3.7: elastic net *LOOCV* plot

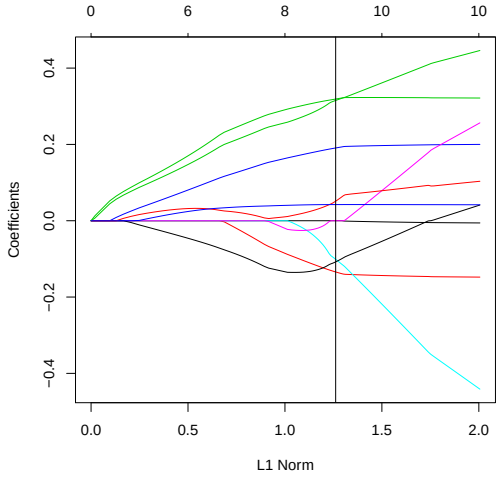


Figure 3.8: coefficients path

Since $P_\alpha(\boldsymbol{\beta}) = \sum_{i=1}^{p}[\frac{1}{2}(1-\alpha)\beta_i^2 + \alpha\,|\beta_i|]$, let $\alpha_i = \frac{i}{100} - 0.01$ for $i = 1, 2, \cdots, 101$. That is, $\alpha_1 = 0$, $\alpha_2 = 0.01$, $\alpha_3 = 0.02$, $\cdots$, $\alpha_{101} = 1$. For each α_i, we do the elastic net regression and keep the smallest $LOOCV_i$. Finally, we use the minimum *LOOCV* among those 101 $LOOCV's$ to determine the value of α.

Figure 3.9: α and corresponding smallest *LOOCV*

```
alpha and corresponding smallest CV:
alpha=          0        0.05     0.16     ...   0.48     ... 1
smallest CV     0.50513  0.50398  0.50341  ...   0.50383  ... 0.50375
```

Based on Figure 3.9, $\alpha_{17} = 0.16$ gives the smallest $LOOCV = 0.50341$. Then we do the elastic net regression with $\alpha = 0.16$ and obtain the detail of numerical results. Since the elastic net regression does the variable selection, we show the number of non-zero variables in the model which is called Df.

Table 3.7: elastic net numerical results

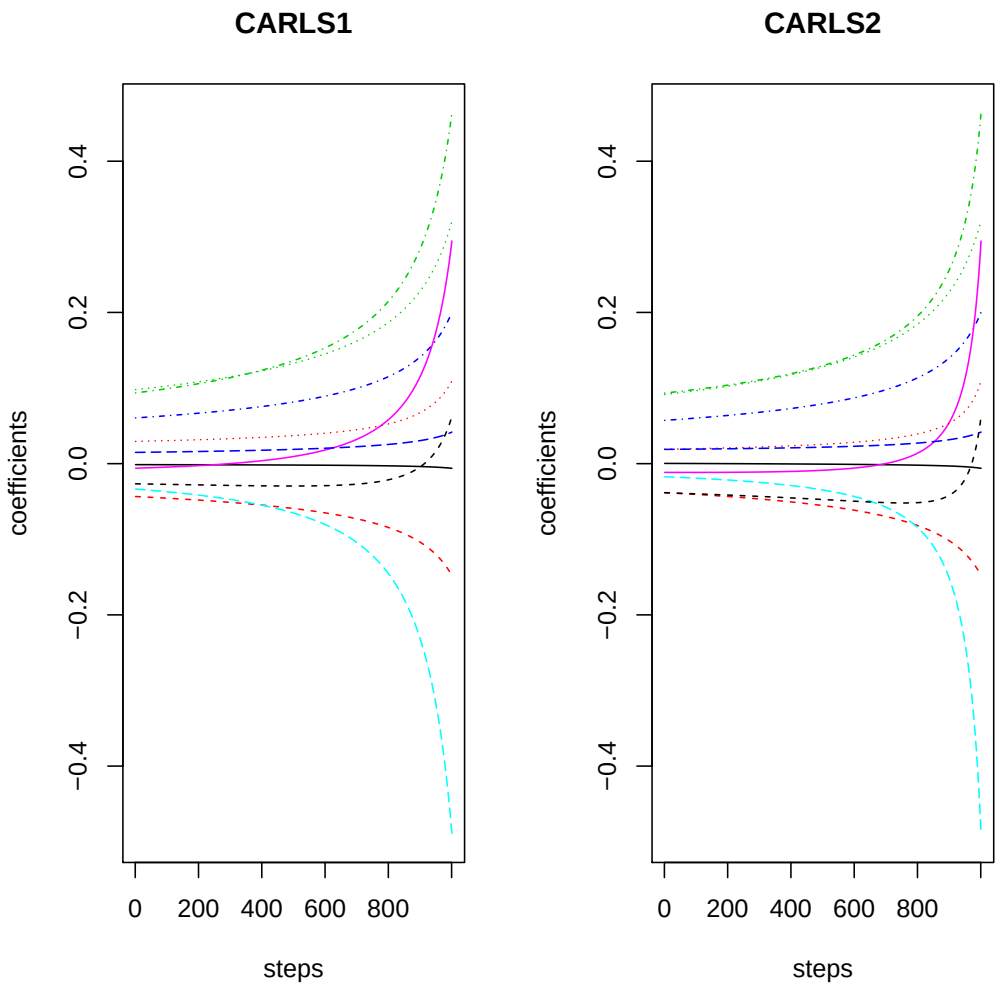
λ	$LOOCV$	$Std.Error$	$P_\alpha(\boldsymbol{\beta})$	Df
3.6611	1.0024	0.0504	0	0
1.5848	0.7482	0.0383	0.0563	6
0.6251	0.5803	0.0313	0.1372	6
0.0264	**0.5034**	**0.0310**	**0.3202**	**8**
0.0037	0.5051	0.0314	0.4505	10
0.0004	0.505	0.0315	0.5367	10

Based on Table 3.7, when $\lambda = 0.0264$, both $LOOCV$ and $Std.Error$ are at minimum and there are 8 variables included in the model.

Figure 3.7 gives the relationship between $ln\lambda$ and $LOOCV$. The integer numbers at the top of this graphic show the number of non-zero estimators in the model. The left line gives the smallest $LOOCV$ with 8 variables in the model and the right line gives the smallest standard deviation with only 7 variables in the model. We can pick any λ between the left line and the right line. Finally, we plot the path of the coefficients. As we see from Figure 3.8, the optimal elastic net estimators are at the intersections of the vertical line and each coefficient path. The elastic net regression does not only the estimators' shrinkage but also variable selection.

3.3.5 Correlation adjusted regression

Figure 3.10: CAR coefficients path



Since minimizing $OLS^* = (\boldsymbol{Y}^* - \boldsymbol{X}^*\boldsymbol{\beta}^*)^T(\boldsymbol{Y}^* - \boldsymbol{X}^*\boldsymbol{\beta}^*)$ is equivalent to minimizing $(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda\boldsymbol{\beta}^T\boldsymbol{W}\boldsymbol{\beta} = CAR$, we calculate the numerical results by ordinary least squares regression method using the updated data set.

Let $\lambda_i = \frac{i}{100} - 0.01$ for $i = 1, 2, \cdots, 1001$. That is, $\lambda_1 = 0$, $\lambda_2 = 0.01$, $\lambda_3 = 0.02$,

$\cdots$, $\lambda_{1001} = 10$. For each λ_i, we update the data set and calculate the $LOOCV$. For each ordinary least squares regression, the $LOOCV$ is a constant. Finally, we pick the optimal value based on the $LOOCV$ and $Std.Error$.

Table 3.8: CAR_1 numerical results

λ_{CARLS_1}	$LOOCV$	$Std.Error$	$\boldsymbol{\beta^T W_1 \beta}$
15	0.5145330	0.03143519	0.226623
11.32	0.5125900	**0.0314135**	0.2351756
0.05	0.5051369	0.03171106	0.9926207
0.03	0.5051079	0.03171519	1.011528
0.02	0.5050914	0.03171711	1.020666
0.01	0.5050766	0.03171921	1.029955
0	0.5050625	0.03172138	1.039399

Based on Table 3.8, as λ_{CAR_1} increases, $LOOCV$ increases and $Std.Error$ decreases until $\lambda_{CAR_1} = 11.32$. Since both $LOOCV$ and $Std.Error$ are not sensitive, we choose the optimal values between $\lambda_{CAR_1} = 0$ and $\lambda_{CAR_1} = 0.05$.

Table 3.9: CAR_2 numerical results

λ_{CARLS2}	$LOOCV$	$Std.Error$	$\boldsymbol{\beta^T W_2 \beta}$
2	0.5129545	0.03145781	1.364795
1.8	0.5122766	**0.03145447**	1.392341
0.05	0.5053233	0.03169246	3.140114
0.03	0.5052181	0.03170346	3.247574
0.02	0.5051672	0.03170925	3.303003
0.01	0.5051135	0.03171523	3.363654
0	0.5050625	0.03172138	3.426745

Based on Table 3.9, as λ_{CAR_2} increases, $LOOCV$ increases and $Std.Error$ decreases until $\lambda_{CAR_2} = 1.8$. Since both $LOOCV$ and $Std.Error$ are not sensitive, we choose the optimal values between $\lambda_{CAR_2} = 0$ and $\lambda_{CAR_2} = 0.05$.

Table 3.10: CAR estimators and the percentage change

	OLS	CAR_1	CAR_2	$\frac{\lvert CAR_2\rvert - \lvert CAR_1\rvert}{\lvert CAR_1\rvert} \times 100\%$
age	-0.006184	-0.006109(-1.21%)	-0.006070(-1.84%)	(-0.64%)
sex	-0.1481	-0.1480(-0.07%)	-0.1479(-0.14%)	(-0.07%)
bmi	0.3211	0.3213(0.06%)	0.3214(0.09%)	(0.03%)
map	0.2004	0.2003(-0.05%)	0.2002(-0.10%)	(-0.05%)
tc	-0.4893	-0.4665(-4.66%)	-0.4590(-6.19%)	(-1.61%)
ldl	0.2945	0.2765(-6.11%)	0.2706(-8.11%)	(-2.13%)
hdl	0.06241	0.05228(-16.23%)	0.04893(-21.60%)	(-6.41%)
tch	0.1094	0.1066(-2.56%)	0.1056(-3.47%)	(-0.94%)
ltg	0.4641	0.4556(-1.83%)	0.4529(-2.41%)	(-0.59%)
glu	0.04177	0.04183(0.14%)	0.04189(0.29%)	(0.14%)

In Table 3.10, we list the estimators of CAR at their optimal values. The percentage values in the brackets are calculated by $\frac{\lvert CAR\rvert - \lvert OLS\rvert}{\lvert OLS\rvert} \times 100\%$ which is the percentage shrinkage at the optimal λ. The negative sign means shrinkage and the positive sign means zooming. For both methods, variables called ldl, hdl, tch and tc obviously have the larger percentage shrinkage. If we go back to Table 3.4, which is the correlation matrix among independent variables, we see that the correlation between variable ldl and variable tc is equal to 0.9, also the correlation between variable tch and variable hdl is equal to -0.74. This implies that if there are multi-collinearity, it is likely produced by those variables. To detect the multi-collinearity issue, we shrinkage the estimators and shrink fast for the highly correlated variables. This is what CAR does. The last column is the percentage change between the CAR_1 and the CAR_2 at the optimal λ. It's calculated by $\frac{\lvert CAR_2\rvert - \lvert CAR_1\rvert}{\lvert CAR_1\rvert} \times 100\%$.

The negative sign means shrinkage and the positive sign means zooming. As we see, eight of the ten signs are negative. This implies the CAR_2 has the more shrinkage than the CAR_1.

According to Figure 3.10, the coefficient path of CAR_1 and CAR_2 are very similar. However, the coefficients of the variables ldl, hdl, tch and tc are converge to 0 more faster at beginning for CAR_2. Based on Table 3.8 and Table 3.9, at the optimal values, CAR_2 prefers shrinkage for highly correlated estimators than CAR_1 at the beginning.

3.3.6 Correlation adjusted elastic net

Since minimizing $LASSO^* = (\boldsymbol{Y}^* - \boldsymbol{X}^*\boldsymbol{\beta}^*)^T(\boldsymbol{Y}^* - \boldsymbol{X}^*\boldsymbol{\beta}^*) + \gamma \sum_{i=1}^{p} |\beta_i^*|$ is equivalent to minimizing $(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda_1 \sum_{i=1}^{p} |\beta_i| + \lambda_2 \boldsymbol{\beta}^T \boldsymbol{W} \boldsymbol{\beta} = CAEN$, we apply the lasso regression to obtain the numerical results by using the update data set.

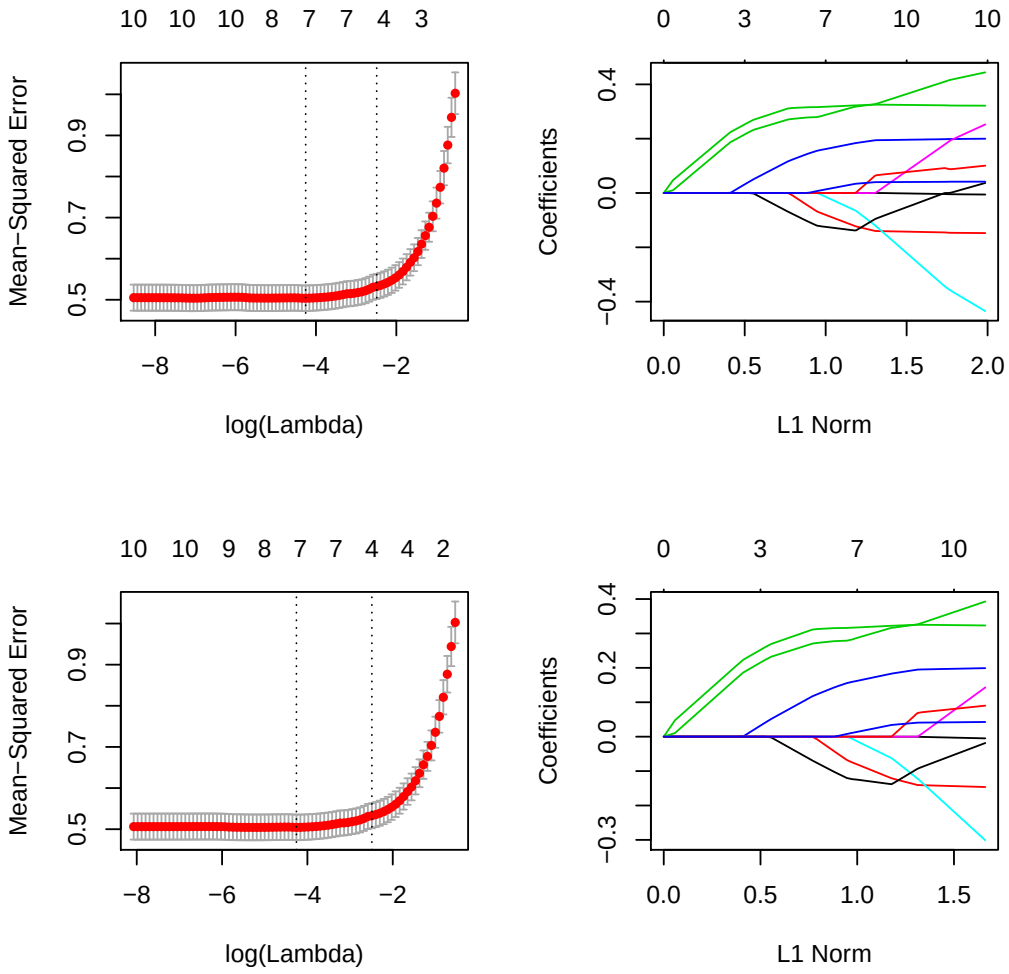
Let $\lambda_{2,i} = \frac{i}{100} - 0.01$ for $i = 1, 2, \cdots, 201$. That is, $\lambda_{2,1} = 0$, $\lambda_{2,2} = 0.01$, $\lambda_{2,3} = 0.02$, $\cdots$, $\lambda_{2,201} = 2$. For each $\lambda_{2,i}$, we update the data set and do the lasso regression to find the optimal $LOOCV$ and corresponding standard error. Since $CAEN$ does the variable selection, we also show the number of non-zero variables in the model which is called Df.

Table 3.11: $CAEN_1$ numerical results

λ_1	λ_2	$LOOCV$	$Std.Error$	$\sum_{i=1}^{p} \lvert\beta_i^*\rvert$	Df
1.29177×10^{-2}	0	0.50372	3.10988×10^{-2}	1.17813	8
1.41768×10^{-2}	0.02	0.50374	3.10465×10^{-2}	1.16839	7
1.41734×10^{-2}	0.2	0.50387	3.10383×10^{-2}	1.16819	7
1.41697×10^{-2}	0.4	0.50403	3.10305×10^{-2}	1.16796	7
1.41660×10^{-2}	0.6	0.50419	3.10227×10^{-2}	1.16773	7
1.41623×10^{-2}	0.8	0.50436	3.10157×10^{-2}	1.16750	7
1.41614×10^{-2}	0.85	0.50441	3.10141×10^{-2}	1.16744	7
0.55855×10^{-2}	0.86	0.50442	3.13507×10^{-2}	1.27530	8

Based on Table 3.11, we choose any λ_2 between 0.02 and 0.85. The smaller value of λ_2 gives the smaller $LOOCV$ but a larger $Std.Error$ and the larger value of λ_2 gives the opposite result.

Figure 3.11: $CAEN_1$ $LOOCV$ plot and path of coefficients



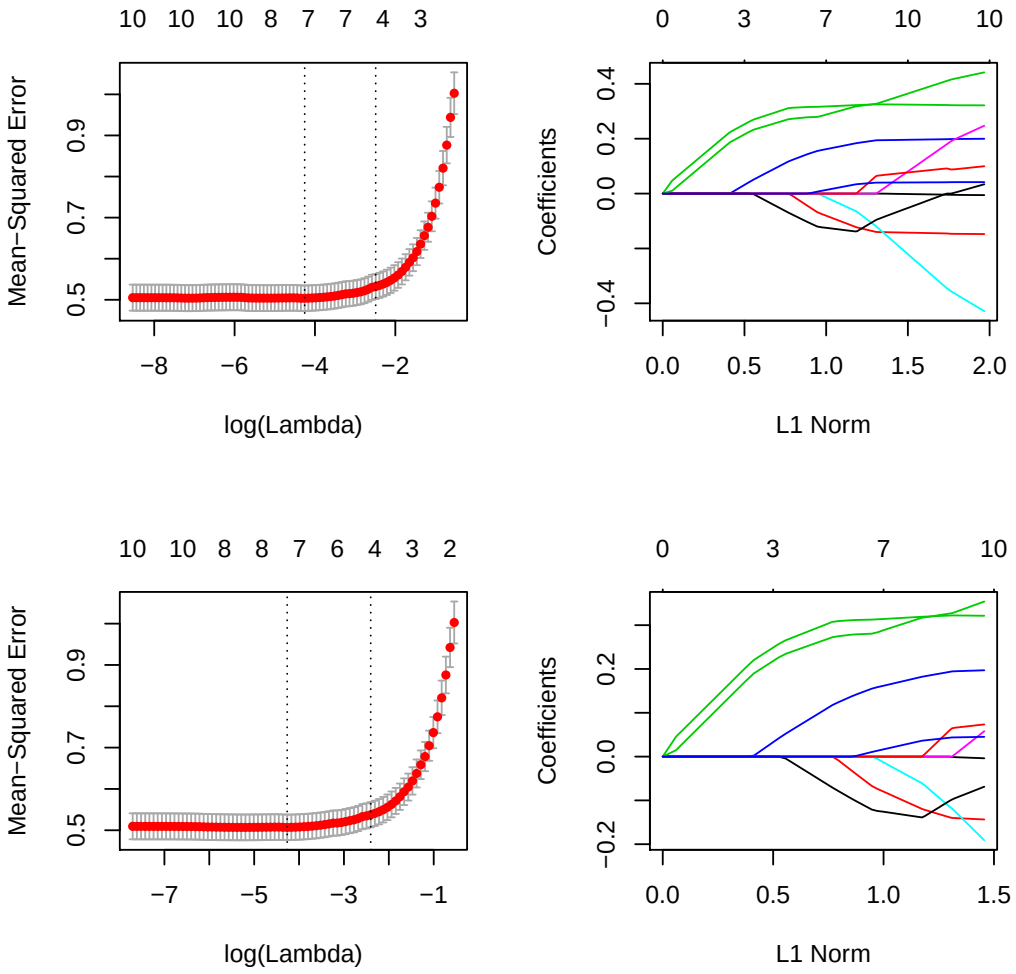
In Figure 3.11, the top two graphics are the $LOOCV$ plot and path of coefficients for $CAEN_1$ with $\lambda_2 = 0.02$. The bottom two graphics are the $LOOCV$ plot and path of coefficients for $CAEN_1$ with $\lambda_2 = 0.85$.

Table 3.12: $CAEN_2$ numerical results

λ_1	λ_2	$LOOCV$	$Std.Error$	$\sum_{i=1}^{p} \lvert\beta_i^*\rvert$	Df
1.29177×10^{-2}	0	0.50372	3.10988×10^{-2}	1.17813	8
1.41748×10^{-2}	0.02	0.50378	3.10438×10^{-2}	1.16834	7
1.41542×10^{-2}	0.2	0.50433	3.10104×10^{-2}	1.16763	7
1.41313×10^{-2}	0.4	0.50494	3.09768×10^{-2}	1.16685	7
1.41085×10^{-2}	0.6	0.50557	3.09488×10^{-2}	1.16608	7
1.40859×10^{-2}	0.8	0.50620	3.09247×10^{-2}	1.16531	7
1.40633×10^{-2}	1	0.50684	3.09053×10^{-2}	1.16454	7
1.40588×10^{-2}	1.04	0.50697	3.09019×10^{-2}	1.16439	7
0.55446×10^{-2}	1.05	0.50700	3.12106×10^{-2}	1.26562	8

According to Table 3.12, we choose any λ_2 between 0.02 and 1.04. The smaller value of λ_2 gives the smaller $LOOCV$ but a larger $Std.Error$ and the larger value of λ_2 gives the opposite result.

Figure 3.12: $CAEN_2$ $LOOCV$ plot and path of coefficients



In Figure 3.12, the top two graphics are the $LOOCV$ plot and path of coefficients for $CAEN_1$ with $\lambda_2 = 0.02$. The bottom two graphics are the $LOOCV$ plot and path of coefficients for $CAEN_1$ with $\lambda_2 = 1.04$.

3.4 Summary

Since different penalized regression methods have their own characteristic, we summarize the numerical results in three aspects.

Firstly, we compare the classical penalized regression methods which include the ridge regression, lasso regression and elastic net regression.

Table 3.13: coefficient comparison of classic methods

	OLS	$RIDGE$[7]	$LASSO$	$ElasticNet$
age	-0.006184	-0.003548	0.0000	0.0000
sex	-0.1481	-0.1426	-0.1211	-0.1342
bmi	0.3211	0.3200	0.3225	0.3189
map	0.2004	0.1964	0.1830	0.1907
tc	-0.4893	-0.1537	-0.063	-0.1011
ldl	0.2945	0.0290	0.0000	0.0000
hdl	0.06241	-0.0826	-0.1379	-0.1078
tch	0.1094	0.0730	0.0000	0.0513
ltg	0.4641	0.3324	0.3173	0.3152
glu	0.04177	0.0454	0.0333	0.0424
$LOOCV$	0.5050891	0.5046807	0.504	0.5034
$Std.Error$	0.03152146	0.0307	0.0309	0.0310
Df	10	10	7	8

Based on Table 3.13, their characteristics are observed obviously. The ridge regression does the shrinkage only (there are 10 nonzero variables in the final model), the lasso regression does both the shrinkage and variable selection (there are 7 nonzero variables in the final model) and the elastic net regression is the combination of the ridge regression and the lasso regression (there are 8 nonzero variables in the final model). According to numerical results, the elastic net regression gives a

[7]The numerical results are calculated based on $\lambda = 0.02$.

smaller $LOOCV$ but a larger $Std.Error$. Ridge regression gives a smaller $Std.Error$ but a larger $LOOCV$. The lasso regression gives the neutral numerical result.

Secondly, we compare the ridge regression and the correlation adjusted regression. In the last chapter, we have summarized that the ridge regression is a special version of CAR. So, we compare the ridge regression and CAR together.

Table 3.14: coefficient comparison of ridge and CAR

	OLS	$RIDGE$	CAR_1	CAR_2
age	-0.006184	-0.003548	-0.006108718	-0.00606960
sex	-0.1481	-0.1426	-0.148013070	-0.14793351
bmi	0.3211	0.3200	0.321322661	0.32135148
map	0.2004	0.1964	0.200256526	0.20019397
tc	-0.4893	-0.1537	-0.466495798	-0.45902860
ldl	0.2945	0.0290	0.276490700	0.27060242
hdl	0.06241	-0.0826	0.052283105	0.04893186
tch	0.1094	0.0730	0.106551737	0.10555651
ltg	0.4641	0.3324	0.455587466	0.45285726
glu	0.04177	0.0454	0.041825159	0.04188734
$LOOCV$	0.5050891	0.5046807	0.5050914	0.5051672
$Std.Error$	0.03152146	0.0307	0.03171711	0.03170925
Df	10	10	10	10

Based on Table 3.14, the ridge regression shows both smaller $LOOCV$ and smaller $Std.Error$. According our data set, the ridge regression performs better results than CAR.

Finally, we compare the elastic net regression and the correlation adjusted elastic net regression. In the last Chapter, we have summarized that the elastic net regression is a special version of $CAEN$. So, we compare the elastic net regression and $CAEN$ together.

Based on Table 3.15, both the $LOOCV$ and $Std.Error$ are really close for the

Table 3.15: coefficient comparison of elastic net and $CAEN$

	OLS	$ElasticNet$	$CAEN_1$	$CAEN_2$
age	-0.006184	0.0000	0.0000	0.0000
sex	-0.1481	-0.1342	-0.11771	-0.11769
bmi	0.3211	0.3189	0.31903	0.31897
map	0.2004	0.1907	0.18006	0.18004
tc	-0.4893	-0.1011	-0.05975	-0.05973
ldl	0.2945	0.0000	0.0000	0.0000
hdl	0.06241	-0.1078	-0.13584	-0.13586
tch	0.1094	0.0513	0.0000	0.0000
ltg	0.4641	0.3152	0.31257	0.31258
glu	0.04177	0.0424	0.03192	0.03196
$LOOCV$	0.5050891	0.5034	0.50374	0.50378
$Std.Error$	0.03152146	0.0310	0.0310465	0.0310438
Df	10	8	7	7

elastic net regression and the correlation adjusted elastic net regression. However, there are only 7 nonzero variables in the $CAEN$ final model. To compare with the elastic net regression, the $CAEN$ generates the less complex model with the same $LOOCV$. According to our data set, the $CAEN$ performs better than the elastic net method.

Chapter 4

Conclusion

With the rapid development of computer information technology, high dimensional data analysis have became an important problem in modern statistics. It has became increasingly common in many fields such as the social sciences, genetics and medical studies. To establish an accurate model, numerous variables are collected. Unfortunately, those variables are often highly correlated. As we have discussed in this thesis, those variables make the model less predictive and difficult to interpret. Therefore penalized regression methods provide a good way to select the appropriate variables and establish an effective model.

4.1 Summary of results

In this thesis, we have given both theoretical and numerical results of penalized regression methods, including ridge regression, lasso regression, elastic net regression, CAR and $CAEN$. According to our numerical results, the CAR method prefers shrinkage for highly correlated estimators. The $CAEN$ method generates the less complex model.

The main results are as follows:

1. We give a detailed introduction to penalized regression and explain the issue of multi-collinearity.

2. We use R to compute the numerical result for existing penalized regression methods which include ridge, lasso and elastic net.

3. We use suitable data argumentation and edit the code in R package to compute and investigate the numerical results for CAR and $CAEN$.

4. We summarize the numerical results and compare them in three aspects.

4.2 Future research

Based on this data set, the CAR method prefers shrinkage for highly correlated variables at the beginning, especially for CAR_2. In the future research, we would like to investigate this characteristic. The $CAEN$ method performs better results compared to other methods. I would like to continue to test and verify the advantages and disadvantages of the $CAEN$ method. Moreover, we will apply CAR and $CAEN$ to survival data, since there are lot's of variables in many survival data analysis problems.

Appendix

R Code

```
####################    Package Installation    ####################

library(lars)

library(glmpath)

library(glmnet)

library(survival)

library(splines)

library(penalized)

library(covTest)

library(MASS)

library(Matrix)

library(psych)

library(DAAG)

library(rgl)

library(scatterplot3d)

library(lasso2)
```

```
library (elasticnet)
library (pls)
library (hydroGOF)
library(xtable)

#######################    Linear regression    ####################

diabetesdata<-as.matrix(read.table("theoriginaldata.txt",header=TRUE))
diabetesdata<-stdize(diabetesdata)
diabetesdatax<-diabetesdata[,-11]
diabetesdatay<-diabetesdata[, 11]
linear<-lm(diabetesdatay ~ diabetesdatax)
summary(linear)
diabetesdata<-as.data.frame(diabetesdata)
is.data.frame(diabetesdata)
set.seed(100)
cv.linear<-CVlm(df=diabetesdata,m=442,
form.lm=formula(y~age+sex+bmi+map+tc+ldl+hdl+tch+ltg+glu))

######################    Ridge regression    ########################

diabetesdata<-as.matrix(read.table("theoriginaldata.txt",header=TRUE))
diabetesdata<-stdize(diabetesdata)
diabetesdatax<-diabetesdata[,-11]
diabetesdatay<-diabetesdata[, 11]
```

```
ridge.glm<-glmnet(diabetesdatax,diabetesdatay, alpha= 0)
summary(ridge.glm)
cv.ridge.glm <-cv.glmnet(diabetesdatax,diabetesdatay,
              nfold=442, alpha= 0)
summary(cv.ridge.glm)
ridge.glm<-glmnet(diabetesdatax,diabetesdatay, alpha= 0,
           lambda= c(585.1,16,2,1,0.05,0.02,0.01,0))
cv.ridge.glm <-cv.glmnet(diabetesdatax,diabetesdatay,nfold=442,
               alpha= 0,lambda= c(585.1,16,2,1,0.05,0.02,0.01,0))
t(ridge.glm$beta[,1])%*%ridge.glm$beta[,1]
t(ridge.glm$beta[,3])%*%ridge.glm$beta[,3]
t(ridge.glm$beta[,8])%*%ridge.glm$beta[,8]
plot(cv.ridge.glm)
plot(ridge.glm)

########################    Lasso regression    ########################

diabetesdata<-as.matrix(read.table("theoriginaldata.txt",header=TRUE))
diabetesdata<-scale(diabetesdata)
diabetesdatax<-diabetesdata[,-11]
diabetesdatay<-diabetesdata[, 11]
lasso.glm<-glmnet(diabetesdatax,diabetesdatay, alpha= 1)
summary(lasso.glm)
cv.lasso.glm<-cv.glmnet(diabetesdatax,diabetesdatay,
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]))
sum(abs(lasso.glm$beta[,10]))
sum(abs(lasso.glm$beta[,42]))
sum(abs(lasso.glm$beta[,88]))

############### Correlation adjusted regression  #################

diabetesdata<-as.matrix(read.table("theoriginaldata.txt",header=TRUE))
diabetesdata<-scale(diabetesdata)
diabetesdatax<-diabetesdata[,-11]
diabetesdatay<-diabetesdata[, 11]
Xij<-diabetesdatax
dim(Xij)
n <- dim(Xij)[1]
p<- dim(Xij)[2]
one <- rep(1, n)
X.means <- t(one) %*% Xij/n
X.diff <- Xij - one %*% X.means
X.cov <- t(X.diff) %*% X.diff/(n - 1)
```

```
sdi <- diag(1/sqrt(diag(X.cov)))
X.cor <- sdi %*% X.cov %*% sdi

######  or use round(cor(diabetesdatax),4)

x.cor<- round(cor(diabetesdatax),7)


diabetesdatax<-as.matrix(diabetesdatax)
diabetesdatay<-as.matrix(diabetesdatay)
zeromatrix<-matrix(c(rep(0, p)),p,1)
Ynew<- rbind(diabetesdatay,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,
            0,0,0,1,-x.cor[4,5],0,0,0,0,0,
            0,0,0,0,1,-x.cor[5,6],0,0,0,0,
            0,0,0,0,0,1,-x.cor[6,7],0,0,0,
            0,0,0,0,0,0,1,-x.cor[7,8],0,0,
            0,0,0,0,0,0,0,1,-x.cor[8,9],0,
            0,0,0,0,0,0,0,0,1,-x.cor[9,10],
```

```
            0,0,0,0,0,0,0,0,0,1),nrow=10)


Wa=Da %*% t(Da)
Ca= chol(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,
            1,0,0,0, -x.cor[1,5],0,0,0,0,0,
            1,0,0,0,0, -x.cor[1,6],0,0,0,0,
            1,0,0,0,0,0, -x.cor[1,7],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,
            0,1,0, -x.cor[2,4],0,0,0,0,0,0,
            0,1,0,0, -x.cor[2,5],0,0,0,0,0,
            0,1,0,0,0, -x.cor[2,6],0,0,0,0,
            0,1,0,0,0,0, -x.cor[2,7],0,0,0,
            0,1,0,0,0,0,0, -x.cor[2,8],0,0,
            0,1,0,0,0,0,0,0, -x.cor[2,9],0,
            0,1,0,0,0,0,0,0,0, -x.cor[2,10],

            0,0,1, -x.cor[3,4],0,0,0,0,0,0,
            0,0,1,0, -x.cor[3,5],0,0,0,0,0,
```

```
0,0,1,0,0, -x.cor[3,6],0,0,0,0,
0,0,1,0,0,0, -x.cor[3,7],0,0,0,
0,0,1,0,0,0,0, -x.cor[3,8],0,0,
0,0,1,0,0,0,0,0, -x.cor[3,9],0,
0,0,1,0,0,0,0,0,0,-x.cor[3,10],

0,0,0,1, -x.cor[4,5],0,0,0,0,0,
0,0,0,1,0, -x.cor[4,6],0,0,0,0,
0,0,0,1,0,0, -x.cor[4,7],0,0,0,
0,0,0,1,0,0,0, -x.cor[4,8],0,0,
0,0,0,1,0,0,0,0, -x.cor[4,9],0,
0,0,0,1,0,0,0,0,0,-x.cor[4,10],

0,0,0,0,1, -x.cor[5,6],0,0,0,0,
0,0,0,0,1,0, -x.cor[5,7],0,0,0,
0,0,0,0,1,0,0, -x.cor[5,8],0,0,
0,0,0,0,1,0,0,0, -x.cor[5,9],0,
0,0,0,0,1,0,0,0,0,-x.cor[5,10],

0,0,0,0,0,1, -x.cor[6,7],0,0,0,
0,0,0,0,0,1,0, -x.cor[6,8],0,0,
0,0,0,0,0,1,0,0, -x.cor[6,9],0,
0,0,0,0,0,1,0,0,0,-x.cor[6,10],

0,0,0,0,0,0,1, -x.cor[7,8],0,0,
0,0,0,0,0,0,1,0, -x.cor[7,9],0,
0,0,0,0,0,0,1,0,0,-x.cor[7,10],
```

```
            0,0,0,0,0,0,0,1, -x.cor[8,9],0,
            0,0,0,0,0,0,0,1,0,-x.cor[8,10],

            0,0,0,0,0,0,0,0,1,-x.cor[9,10],
            0,0,0,0,0,0,0,0,0,1),nrow=10)

Wb= Db%*%t(Db)
Cb= chol(Wb)


lambda=seq(from=0,to=1,length=101)

### Using the different lambda[i] to find numerical result
### Ca for CAR1 and Cb for CAR2
X.new<-(1/sqrt(1+lambda[101]))*rbind(Xij,sqrt(lambda[101])*t(Ca))
Y.new<-rbind(diabetesdatay,zeromatrix)
diabetesdata.new<- cbind(X.new, Y.new)
diabetesdata.new< stdize(diabetesdata.new)
diabetesdata.new<-as.data.frame(diabetesdata.new)
diabetesdata.new<-as.matrix(diabetesdata.new)
cv.car<-CVlm(df=diabetesdata.new, m=452,form.lm=formula
        (V11~age+sex+bmi+map+tc+ldl+hdl+tch+ltg+glu))


### find loocv and standard errors

car.glm<-glmnet(diabetesdata.new[,-11], diabetesdata.new[,11],
         alpha=0 ,lambda=c(0.1,0))
```

```
cv.ridge.glm<-cv.glmnet(diabetesdata.new[,-11],
              diabetesdata.new[,11],
              nfold=452, alpha=0,lambda= c(0.1,0))
cv.ridge.glm

### find the coefficient
diabetesdata.new <-as.matrix(diabetesdata.new)
linear.car<-lm(diabetesdata.new[,11]~-1+diabetesdata.new[,-11])
summary(linear.car)
car.beta<-coef(linear.car)/sqrt(1+lambda[1])
beta_w_beta<-t(car.beta)%*%Wa%*%car.beta


### To find the coefficients path of CAR
### Ca for CAR1 and Cb for CAR2

lambda=seq(from=10,to=0,length=1001)
car.beta.v1<-numeric(1001)
car.beta.v2<-numeric(1001)
car.beta.v3<-numeric(1001)
car.beta.v4<-numeric(1001)
car.beta.v5<-numeric(1001)
car.beta.v6<-numeric(1001)
car.beta.v7<-numeric(1001)
car.beta.v8<-numeric(1001)
car.beta.v9<-numeric(1001)
car.beta.v10<-numeric(1001)
```

```
for(i in 1:1001){

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

     Y.new<-rbind(diabetesdatay,zeromatrix)

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

     diabetesdata.new<- stdize(diabetesdata.new)

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

     linear.car<-lm(diabetesdata.new[,11]~-1+diabetesdata.new[,-11])

     as.matrix(coef(linear.car)/sqrt(1+lambda[i]))

     car.beta.v1[i]<-coef(linear.car)[1]/sqrt(1+lambda[i])

                    as.matrix(car.beta.v1)

     car.beta.v2[i]<-coef(linear.car)[2]/sqrt(1+lambda[i])

                    as.matrix(car.beta.v2)

     car.beta.v3[i]<-coef(linear.car)[3]/sqrt(1+lambda[i])

                    as.matrix(car.beta.v3)

     car.beta.v4[i]<-coef(linear.car)[4]/sqrt(1+lambda[i])

                    as.matrix(car.beta.v4)

     car.beta.v5[i]<-coef(linear.car)[5]/sqrt(1+lambda[i])

                    as.matrix(car.beta.v5)

     car.beta.v6[i]<-coef(linear.car)[6]/sqrt(1+lambda[i])

                    as.matrix(car.beta.v6)

     car.beta.v7[i]<-coef(linear.car)[7]/sqrt(1+lambda[i])

                    as.matrix(car.beta.v7)

     car.beta.v8[i]<-coef(linear.car)[8]/sqrt(1+lambda[i])
```

```
                    as.matrix(car.beta.v8)
     car.beta.v9[i]<-coef(linear.car)[9]/sqrt(1+lambda[i])
                    as.matrix(car.beta.v9)
     car.beta.v10[i]<-coef(linear.car)[10]/sqrt(1+lambda[i])
                    as.matrix(car.beta.v10)

     }

coef.car<-rbind(car.beta.v1,car.beta.v2,car.beta.v3,car.beta.v4,
                car.beta.v5,car.beta.v6, car.beta.v7,car.beta.v8,
                car.beta.v9,car.beta.v10)

matplot(t(coef.car),type = "l")

par(mfrow=c(1,2))

matplot(t(coef.car),xlab="steps",ylab="coefficients",
                            type = "l",main="CARLS1")


############  Correlation Adjusted Elastic Net  ######################

diabetesdata<-as.matrix(read.table("theoriginaldata.txt",header=TRUE))
diabetesdata<-stdize(diabetesdata)
diabetesdatax<-diabetesdata[,-11]
diabetesdatay<-diabetesdata[, 11]
Xij<-diabetesdatax
dim(Xij)
n<-dim(Xij)[1]
```

```
p<-dim(Xij)[2]

one<-rep(1, n)

X.means<-t(one) %*% Xij/n

X.diff<-Xij - one %*% X.means

X.cov<-t(X.diff) %*% X.diff/(n - 1)

sdi<-diag(1/sqrt(diag(X.cov)))

X.cor <- sdi %*% X.cov %*% sdi

######  or use round(cor(diabetesdatax),4)

x.cor<- round(cor(diabetesdatax),7)



diabetesdatax<-as.matrix(diabetesdatax)

diabetesdatay<-as.matrix(diabetesdatay)

zeromatrix<-matrix(c(rep(0, p)),p,1)

Ynew<-rbind(diabetesdatay,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,

            0,0,0,1,-x.cor[4,5],0,0,0,0,0,

            0,0,0,0,1,-x.cor[5,6],0,0,0,0,

            0,0,0,0,0,1,-x.cor[6,7],0,0,0,
```

```
              0,0,0,0,0,0,1,-x.cor[7,8],0,0,

              0,0,0,0,0,0,0,1,-x.cor[8,9],0,

              0,0,0,0,0,0,0,0,1,-x.cor[9,10],

              0,0,0,0,0,0,0,0,0,1),nrow=10)


Wa=Da %*% t(Da)
Ca= chol(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,
              1,0,0,0, -x.cor[1,5],0,0,0,0,0,
              1,0,0,0,0, -x.cor[1,6],0,0,0,0,
              1,0,0,0,0,0, -x.cor[1,7],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,
              0,1,0, -x.cor[2,4],0,0,0,0,0,0,
              0,1,0,0, -x.cor[2,5],0,0,0,0,0,
              0,1,0,0,0, -x.cor[2,6],0,0,0,0,
              0,1,0,0,0,0, -x.cor[2,7],0,0,0,
```

```
0,1,0,0,0,0,0, -x.cor[2,8],0,0,
0,1,0,0,0,0,0,0, -x.cor[2,9],0,
0,1,0,0,0,0,0,0,0, -x.cor[2,10],

0,0,1, -x.cor[3,4],0,0,0,0,0,0,
0,0,1,0, -x.cor[3,5],0,0,0,0,0,
0,0,1,0,0, -x.cor[3,6],0,0,0,0,
0,0,1,0,0,0, -x.cor[3,7],0,0,0,
0,0,1,0,0,0,0, -x.cor[3,8],0,0,
0,0,1,0,0,0,0,0, -x.cor[3,9],0,
0,0,1,0,0,0,0,0,0,-x.cor[3,10],

0,0,0,1, -x.cor[4,5],0,0,0,0,0,
0,0,0,1,0, -x.cor[4,6],0,0,0,0,
0,0,0,1,0,0, -x.cor[4,7],0,0,0,
0,0,0,1,0,0,0, -x.cor[4,8],0,0,
0,0,0,1,0,0,0,0, -x.cor[4,9],0,
0,0,0,1,0,0,0,0,0,-x.cor[4,10],

0,0,0,0,1, -x.cor[5,6],0,0,0,0,
0,0,0,0,1,0, -x.cor[5,7],0,0,0,
0,0,0,0,1,0,0, -x.cor[5,8],0,0,
0,0,0,0,1,0,0,0, -x.cor[5,9],0,
0,0,0,0,1,0,0,0,0,-x.cor[5,10],

0,0,0,0,0,1, -x.cor[6,7],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<-numeric(201)
min.cvsd.caen.index<-numeric(201)

### Ca for CAEN1 and Cb for CAEN2

for( i in 1:201){

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

```
    Y.new<-rbind(diabetesdatay,zeromatrix)
    diabetesdata.new<-cbind(X.new, Y.new)

    diabetesdata.new<-stdize(diabetesdata.new)
    cv.caen.lasso.glm<-cv.glmnet(diabetesdata.new[,-11],
                        diabetesdata.new[,11] ,nfold=452, alpha= 1)

     min.cv.caen.index[i]<-which.min(cv.caen.lasso.glm$cvm)
     min.cvsd.caen.index[i]<-which.min(cv.caen.lasso.glm$cvsd)
     min.cv.caen[i]<-cv.caen.lasso.glm$cvm
                    [which.min(cv.caen.lasso.glm$cvm)]
     min.cvsd.caen[i]<-cv.caen.lasso.glm$cvsd
                    [which.min(cv.caen.lasso.glm$cvm)]
     min.lambda.caen[i]<-cv.caen.lasso.glm$lambda
                       [which.min(cv.caen.lasso.glm$cvm)]

     }

### Using the different lambda[i] to find numerical result
### Ca for CAEN1 and Cb for CAEN2

X.new<-(1/sqrt(1+lambda[i]))*rbind(Xij,sqrt(lambda[i])*t(Ca))
Y.new<-rbind(diabetesdatay,zeromatrix)
diabetesdata.new<-cbind(X.new, Y.new)

diabetesdata.new<-stdize(diabetesdata.new)
cv.caen.lasso.glm<-cv.glmnet(diabetesdata.new[,-11],
```

```
diabetesdata.new[,11], nfold=452, alpha= 1)

index<-which.min(cv.caen.lasso.glm$cvm)
cv.caen.lasso.glm$lambda[index]
cv.caen.lasso.glm$cvm[index]
cv.caen.lasso.glm$cvsd[index]
caen1_final.glm<-glmnet(diabetesdata.new[,-11],
                diabetesdata.new[,11], alpha= 1)
sum(abs(caen1_final.glm$beta[,index]))
index
caen1_final.glm$beta[,index]/sqrt(lambda[i])

###  plot the graphic
###  Ca for CAEN1 and Cb for CAEN2

X.new<-(1/sqrt(1+lambda[3]))*rbind(Xij,sqrt(lambda[3])*t(Ca))
Y.new<-rbind(diabetesdatay,zeromatrix)
diabetesdata.new<-cbind(X.new, Y.new)
diabetesdata.new<-stdize(diabetesdata.new)
cv.caen.lasso.glm<-cv.glmnet(diabetesdata.new[,-11],
                  diabetesdata.new[,11],nfold=452, alpha= 1)
caen1_final.glm<-glmnet(diabetesdata.new[,-11],
                diabetesdata.new[,11],alpha= 1)

par(mfrow=c(2,2))
plot(cv.caen.lasso.glm)
plot(caen1_final.glm)
```

```
abline(v=1.16839)

X.new<-(1/sqrt(1+lambda[86]))*rbind(Xij,sqrt(lambda[86])*t(Ca))
Y.new<-rbind(diabetesdatay,zeromatrix)
diabetesdata.new<-cbind(X.new, Y.new)
diabetesdata.new<-scale(diabetesdata.new)
cv.caen.lasso.glm<-cv.glmnet(diabetesdata.new[,-11],
                  diabetesdata.new[,11],nfold=452, alpha= 1)
caen1_final.glm<-glmnet(diabetesdata.new[,-11],
                  diabetesdata.new[,11],alpha= 1)

plot(cv.caen.lasso.glm)
plot(caen1_final.glm)
```

Bibliography

[1] Alheety, M. and Ramanathan, T. (2009). Confidence interval for shrinkage parameters in ridge regression. *Communications in Statistics-Theory and Methods*, 38 3489-3497.

[2] Anbari, M. and Mkhadri, A. (2008). Penalized regression combining the $L1$ norm and a correlation based penalty. *Research Report, Institue National de Recherche en Informatique et en Automatique*, 6746 1-32.

[3] Brant, R. (2007). Mutltiple linear regression. Available from *http://www.stat.ubc.ca/ rollin/teach/BiostatW07/reading/MLR.pdf*

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

[5] Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96, 1348-1360.

[6] Fan, J. and Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *The Annals of Statistics*, 32, 928-961.

[7] Friedman,J., Hastie, T. and Tibshirani, R. (2013). The glmnet package. Available from *http://cran.r-project.org/web/packages/glmnet/glmnet.pdf*

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

[9] Hastie, T., Tibshirani, R., and Friedman, J. (2001). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. *Springer Series in Statistics*.

[10] Hoerl, A.E. and Kennard, R. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12,55-67.

[11] Kutner, M. H., Nachtsheim, C. J., Neter, J. and Li, W. (2005). *Applied linear statistical models* ($Fifth\ edition$). McGraw-Hill/Irwin, New York.

[12] Osborne, M.R., Presnell, B. and Turlach, B.A. (2000a). On the LASSO and its dual. *Journal of Computational and Graphical Statistics*. 9, 319-337.

[13] Osborne, M.R., Presnell, B. and Turlach, B.A. (2000b). A new approach to variable selection in least squares problems. *IMA Journal of Numerical Analysis*, 20, 389-403.

[14] Park, M. Y. and T. Hastie (2007). $l1$ regularized path algorithm for generailized linear models. *Journal of the Royal Statistical Society, Series B*, 69 (4), 659-677.

[15] Seber, G.A.F. (1977). Linear Regression Analysis. *New York: Wiley*.

[16] Seber, G. and Lee, A. (2003). Linear Regression Analysis, 2*nd* Edition. *Wiley Series in Probability and Statistics*.

[17] Tan, Q.(2012). Correlation Adjusted Penalization In Regression Analysis. *PhD Thesis, Department of statistics, University of Manitoba*.

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

[19] Tutz, G. and Ulbricht, J. (2009). Penalized regression with correlation-based penalty. *Statistics and Computing*, 19, 239-253.

[20] Ulbricht, J. and Tutz, G. (2008). Boosting correlation based penalization in generalized linear models. *Recent Advances in Linear Models and Related Areas Essays in Honour of Helge Toutenburg*, 165-180.

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

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

[23] Zou, H. and Zhang, H. (2009). On the adaptive elastic-net with a diverging number of parameters. *The Annals of Statistics*, 12, 1149-1173.