1-1-2018

An Application of Ridge Regression and LASSO Methods for Model Selection

Katie Lynn Phillips

An application of ridge regression and LASSO methods for model selection

By

Katie Lynn Phillips

A Thesis
Submitted to the Faculty of
Mississippi State University
in Partial Fulfillment of the Requirements
for the Degree of Master of Science
in Statistics
in the Department of Mathematics and Statistics

Mississippi State, Mississippi

August 2018

An application of ridge regression and LASSO methods for model selection

By

Katie Lynn Phillips

Approved:

_____

Jonathan Woody
(Major Professor)

_____

Janice DuBien
(Committee Member)

_____

Prakash Patil
(Committee Member)

_____

Mohammad Sepehrifar
(Graduate Coordinator)

_____

Rick Travis
Dean
College of Arts & Sciences

Name: Katie Lynn Phillips

Date of Degree: August 10, 2018

Institution: Mississippi State University

Major Field: Statistics

Major Professor: Dr. Jonathan Woody

Title of Study: An application of ridge regression and LASSO methods for model selection

Pages of Study: 41

Candidate for the Degree of Master of Science

Ordinary Least Squares (OLS) models are popular tools among field scientists, because they are easy to understand and use. Although OLS estimators are unbiased, it is often advantageous to introduce some bias in order to lower the overall variance in a model. This study focuses on comparing ridge regression and the LASSO methods which both introduce bias to the regression problem. Both approaches are modeled after the OLS but also implement a tuning parameter. Additionally, this study will compare the use of two different functions in R, one of which will be used for ridge regression and the LASSO while the other will be used strictly for the LASSO. The techniques discussed are applied to a real set of data involving some physiochemical properties of wine and how they affect the overall quality of the wine.

Key words: ridge regression, LASSO, shrinkage

DEDICATION

To my husband, Spencer, for all of his support throughout this process.

ACKNOWLEDGEMENTS

I would like to begin by expressing my gratitude to my advisor, Dr. Jonathan Woody, for his guidance and mentorship throughout my time in graduate school. His eagerness to learn and teach new things and his enthusiasm about the subject material is what motivated me to pursue a degree in Statistics. For that, I am forever grateful.

I would also like to thank my committee members, Dr. Janice DuBien and Dr. Prakash Patil, as well as both the past and present graduate coordinators, Dr. Corlis Johnson and Dr. Mohammed Sepehrifar, for their continued support throughout my academic career. Each of them has played a critical role in my degree, both inside and outside of the classroom.

It is with great pleasure that I thank my supervisors and coworkers at the Institute for Clean Energy Technology for their understanding. There are many people there that have supported me throughout this endeavor, but Dr. Charles Waggoner, Jaime Rickert, and Ronald Unz have been particularly supportive in my pursuit of this degree.

Finally, I would like to give thanks to my parents who have encouraged me to dream big and work hard my whole life. I can think of no better role models.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

When data sets contain a large number of variables, it may be difficult to determine the "best" regression model for the data. Just because information about a covariate is available does not mean that the information is significant and should be included in the regression model. In some cases, an independent variable that was thought helpful in explaining the variation in response may actually offer only a small decrease in the Mean Square Error (MSE). Therefore, variable selection should be considered in the regression setting.

Many regression techniques have been developed over the years, including ridge regression and the Least Absolute Shrinkage and Selection Operator (LASSO). Ridge regression "shrinks" coefficient estimates in a model towards zero via a bounded ℓ_2-norm regression penalty. This continuous process is more reliable than most subset selection methods such as Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC), because dropping and retaining predictors can drastically change the prediction accuracy of a model. However, since ridge regression does not actually set any coefficient estimate to zero, it can be difficult to interpret a model. The LASSO is seen as a sort of hybrid of the subset selection and shrinkage techniques. A small change in the

tuning parameter allows the LASSO to remove some covariates from the model by setting their estimates to zero [23].

As new discoveries about various regression techniques are made (i.e. how they relate to other methods), new software functions are also developed. The *glmnet* and *lars* functions in R, a statistical programming language, are of particular interest when using ridge regression and the LASSO. The *glmnet* function was published by J.H. Friedman, T. Hastie, and R. Tibshirani as a tool for inference on general linear models using the LASSO, ridge regression, and mixtures of the two [11,12]. After connections were made from the LASSO to least angle regression and infinitesimal forward stagewise regression, the *lars* function was created, which fits models using each of the three [7,13]. The ridge regression model can be fit in SAS using the *reg* procedure, and the LASSO model can be fit using procedures such as *quantselect* or *glmselect*. This thesis will focus on the comparison of ridge regression and the LASSO, as well as the outcome of using both the *glmnet* and *lars* functions in R as applied to a real-world data set.

CHAPTER II

THE MODELS

Ordinary Least Squares

Consider the standard linear model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i \tag{2.1}$$

for i=1,…,n and j=1,…,p where y_i is the i[th] response and x_{ij} is the j[th] covariate of the i[th] observation. Let $\{\varepsilon_i\}_{i=1}^{n}$ be an independently and identically distributed (i.i.d.) mean zero sequence of errors with finite variance. In matrix form, equation (2.1) may be denoted

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

where $\boldsymbol{y}$ is the nx1 vector of responses, $\boldsymbol{X}$ is the nxp matrix of covariates, $\boldsymbol{\beta}$ is the px1 parameter vector, and $\boldsymbol{\varepsilon}$ is the nx1 vector of errors. In the Ordinary Least Squares (OLS) setting, $\widehat{\boldsymbol{\beta}}$ solves the normal equations:

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^g\boldsymbol{X}^T\boldsymbol{y} \tag{2.2}$$

where the generalized inverse $(\boldsymbol{X}^T\boldsymbol{X})^g = (\boldsymbol{X}^T\boldsymbol{X})^{-1}$ when $\boldsymbol{X}$ is full column rank. In this manner, $\hat{\beta}$'s minimize the sum of squared errors

$$\left(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\right)^T\left(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\right).$$

Assuming $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ exists, the fitted residuals are

$$\hat{\boldsymbol{\varepsilon}} = \boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}} = (\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T)\boldsymbol{y},$$

and the Residual Sum of Squares (RSS) is found to be

$$RSS = \hat{\boldsymbol{\varepsilon}}^T\hat{\boldsymbol{\varepsilon}} = \boldsymbol{y}^T(\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T)\boldsymbol{y}.$$

When $\boldsymbol{X}$ is not full column rank, $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ is replaced with $(\boldsymbol{X}^T\boldsymbol{X})^g$. While $\widehat{\boldsymbol{\beta}}$ may not be unique in this case, any $\widehat{\boldsymbol{\beta}}$ solving equation (2.2) will produce the same RSS. While some assumption violations may be worked around, generally for OLS to work properly the data must have a few characteristics: strict exogeneity, no linear dependence within errors, and spherical errors (i.e. homoscedasticity and no autocorrelation) [15]. The Gaussian distribution is the only optionally invariant distribution with finite moments [2]. While not absolutely necessary, normality is frequently assumed when using OLS. However, Central Limit Theorem (CLT) results exist for $\widehat{\boldsymbol{\beta}}$ in equation (2.2) under mild design and error assumptions [24].

OLS is frequently used to fit linear models, because it is easy to understand, easy to implement, and maintains nice statistical properties (e.g. $E[\hat{y}] = E[y]$). However, it may sometimes be helpful to introduce a small amount of bias in order to decrease the variance of the model estimates. OLS estimates will not be unique if the design matrix $\boldsymbol{X}$ is not full column rank, i.e. $\text{rank}(\boldsymbol{X}) = \text{k} < \text{p}$. In some cases even if k < p, it may be helpful in the regression setting to look at a model with less than k variables in the model via a model or subset selection method [9,16].

OLS models often do not predict well, especially when there are a large number of predictors. For this reason, it is often a good idea to look at a model's Prediction Error (PE) in addition to its MSE. PE can be expressed as

$$\text{PE}(X_0\hat{\beta}) = \sigma^2 + \text{MSE}(X_0\hat{\beta})$$

where

$$\mathrm{MSE}(X_0\hat{\beta}) = \left[\mathrm{Bias}(X_0\hat{\beta})\right]^2 + \mathrm{Var}(X_0\hat{\beta})$$

for a particular covariate, X_0 [22]. Hence, having a "good" PE implies that a model also has a "good" MSE. Because OLS is unbiased, its PE

$$PE(X\hat{\beta}) = \sigma^2 + \frac{1}{n}\sum_{i=1}^{n}\left[Bias(x_i^T\hat{\beta})\right]^2 + \frac{1}{n}\sum_{i=1}^{n} Var(x_i^T\hat{\beta})$$

$$= \sigma^2 + 0 + \frac{p\sigma^2}{n}$$

is heavily dependent on the number of covariates. Adjusting the flexibility of the OLS model through the addition of a tuning parameter will allow a tradeoff between bias and variance [22]. Ideally, a small amount of bias can be introduced in order to decrease the variance by a larger margin, reducing the overall PE of the model. Shrinkage methods, such as LASSO and ridge regression, restrict the coefficient estimates to some constrained parameter space usually centered about the origin. This helps reduce the variance of prediction, because it keeps estimates close to zero. In this manner, shrinkage methods can sometimes outperform OLS as seen in Figure 2.1 [22].

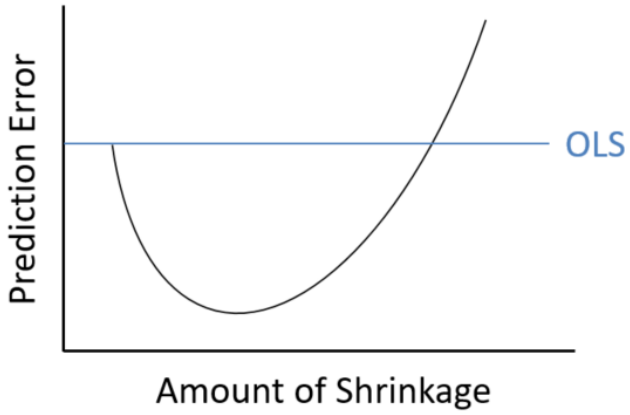

Figure 2.1. Prediction error of shrinkage methods compared to OLS estimates

The LASSO Method

The LASSO technique was originally introduced by Tibshirani [23]. In this method, $\hat{\beta}^{LASSO}$ is chosen to minimize

$$\sum_{i=1}^{n}\left(y_i - \left[\beta_0 + \sum_{j=1}^{p}\beta_j x_{ij}\right]\right)^2 \text{ subject to } \sum_{j=1}^{p}|\beta_j| \leq t(\lambda)$$

where $t(\lambda)$ is a tuning parameter. In other words,

$$\widehat{\boldsymbol{\beta}}^{LASSO} = \underset{\beta \epsilon \mathbb{R}^p}{\operatorname{argmin}}\, RSS + \lambda\|\boldsymbol{\beta}\|_1$$

where $\|\beta\|_1$ is the ℓ_1-norm of the vector ($\beta_1,\ldots,\beta_p$). The ℓ_1-norm of $\boldsymbol{\beta}$ is defined by

$$\|\boldsymbol{\beta}\|_1 = \sum_{j=1}^{p}|\beta_j|.$$

For the case when p=2,

$$|\beta_1| + |\beta_2| \leq t(\lambda).$$

It is easy to see that there are instances when either β_1 or β_2 will be equal to zero, leaving only the other parameter in the model. For large $t(\lambda)$, both β_1 and β_2 will be included. This can be applied to higher dimensions of p, justifying that LASSO is indeed a valid subset selection method [9,16].

Ridge Regression

The LASSO method is in a sense similar to ridge regression, which also regularizes the coefficients of the regression model. However, for ridge regression, $\hat{\beta}^{ridge}$ minimizes

$$\sum_{i=1}^{n}\left(y_i - \left[\beta_0 + \sum_{j=1}^{p}\beta_j x_{ij}\right]\right)^2 \text{ subject to } \sum_{j=1}^{p}{\beta_j}^2 \leq t(\lambda)$$

Alternatively,

$$\widehat{\boldsymbol{\beta}}^{ridge} = \underset{\beta \epsilon \mathbb{R}^p}{\operatorname{argmin}}\, RSS + \lambda \|\boldsymbol{\beta}\|_2^2$$

where $\|\boldsymbol{\beta}\|_2$ is the ℓ_2-norm of the vector ($\beta_1,\ldots,\beta_p$). The ℓ_2-norm of $\boldsymbol{\beta}$ is defined by

$$\|\boldsymbol{\beta}\|_2 = \sqrt{\sum_{j=1}^{p}{\beta_j}^2}.$$

For p=2, this can be written as

$$\beta_1^2 + \beta_2^2 \leq t(\lambda).$$

From Figure 2.1 below, it is shown that β_j will almost surely never be equal to zero for j=1,…,p in ridge regression. However, there are instances in which the LASSO will force some subset of the coefficients to equal zero [9,16]. The LASSO and ridge regression typically both follow the same assumptions as OLS. However, these subset selection methods are better equipped to handle multicollinearity than OLS.

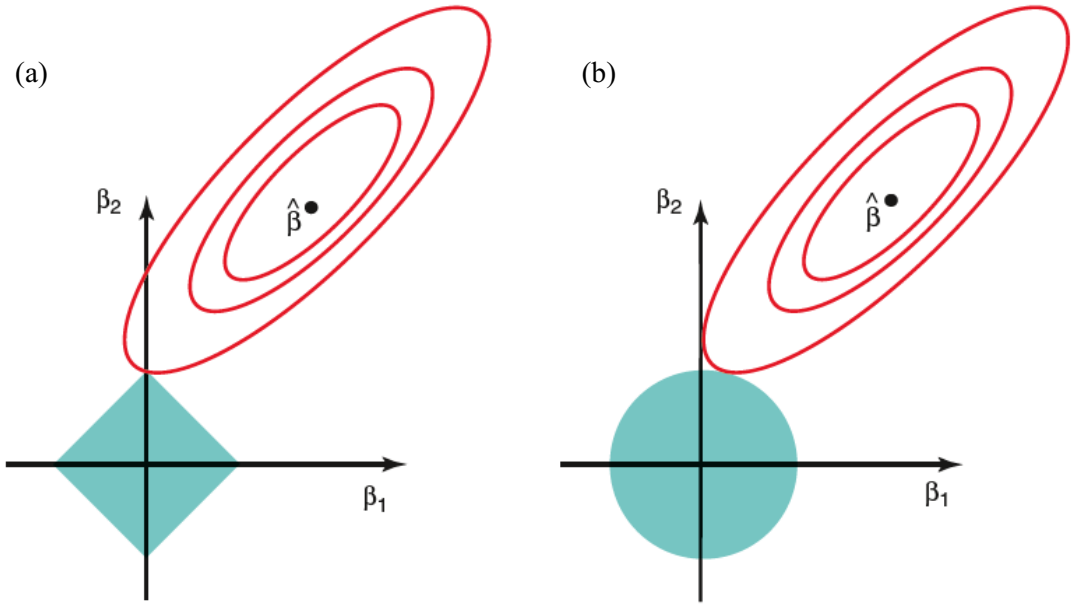

Figure 2.2 Contours of the RSS in comparison to the shaded region of restraint parameter for LASSO (a) and Ridge Regression (b) [16]

Selection of Lambda

The goal is to select a λ such that the error of the model is minimized. There are several ways to do this including multiple iterations of forward stepwise regression or a choice of different cross-validation techniques. This study will focus on k-fold cross-validation, which is detailed in a later section [16].

Least Angle Regression

Another well-known technique for model selection is forward stagewise regression. In this method, coefficients are initially set to zero. Each model produced includes one more variable than the last, choosing the covariate that leads to the largest drop in Residual Sum of Squares (RSS) for the model. Alternatively, some variations of

forward stagewise regression select the covariate which is most highly correlated with the residuals [14]. This continues until either the residuals are zero or all covariates have been added to the model. Similarly, the Least Angle Regression (LAR) method starts with all variables set to zero and introduces the covariates most highly correlation to the residuals one at a time. However, for LAR the coefficient for a variable is only increased until a point in which another covariate has as much correlation with the current residuals [7,14]. The LAR then moves in the direction of the joint least squares coefficient, referred to as the "least angle direction" [7], until the next highly correlated variable is introduced. This process continues until all covariates have been added. The only modification to the LAR method used in *lars* from R necessary to follow LASSO regression is that if a coefficient that has been introduced to the model reaches zero, it is removed from the active joint least angle direction [14]. Further comparisons of these methods can be seen in Efron et. al. (2004) [7] and Hastie et. al. (2007) [14].

Information Criterion-Based Subset Selection Methods

Criterion-based procedures are often used to choose the number of predictors in a statistical model and thus may be seen as competing methods for LASSO and ridge regression. Each information criteria aims to select the model with the "best" penalized log-likelihood. The likelihood function may be written

$$L(\hat{\theta}) = \prod_{i=1}^{n} f_i(y_i|\hat{\theta})$$

where $\hat{\theta}$ is the Maximum Likelihood Estimator (MLE) of the parameters in the function and $f_i(y_i|\hat{\theta})$ is the fitted density of the i[th] observation. Unlike LASSO and ridge regression, note that a statistical distribution for $\boldsymbol{y}$ must be assumed before criterion-based methods may be applied. The log-likelihood function is expressed as

$$\ell(\hat{\theta}) = ln[L(\hat{\theta})] = \sum_{i=1}^{n} ln[f_i(y_i|\hat{\theta})].$$

One method that utilizes this function is the AIC, which is

$$AIC = -2\ell(\hat{\theta}) + 2p.$$

AIC is an estimate of the relative distance between the fitted likelihood function of the model and the unknown true likelihood function of the data [6]. Another criterion-based procedure is BIC, which is written as

$$BIC = -2\ell(\hat{\theta}) + p * \ln(n).$$

In each case, the model with the lowest value is considered the "best" model. A major difference is that BIC penalizes more against models with more complexity [6]. It is worth noting that the criterion-based methods as well as the LASSO and ridge regression methods are of the form of a function plus a penalty term. While models based on AIC and BIC usually agree with each other, AIC is more likely to choose too large of a model, and BIC is more likely to choose too small of a model. This is a very different approach from the LASSO, which does not require an assumed distributional function of $\boldsymbol{y}$, and hence no likelihood needs to be evaluated.

Principle Component Analysis

Before beginning to model a set of data, a preliminary exploratory data analysis is performed on $\boldsymbol{X}$. Principle Component Analysis (PCA) involves taking a set of correlated variables and transforming them into a smaller set of uncorrelated variables without losing much information. This is particularly useful when there are a large number, p, of covariates. PCA is a means of dimensionality reduction. These principle components are linear combinations of the covariates that sequentially maximize the amount of variance accounted for by the principle components (i.e. the first component accounts for the largest amount of variance, the second component accounts for the second largest amount of variance, and so on). Let z_i represent the i[th] principle component. Then

$$z_1 = u_1^T \boldsymbol{X} \text{ maximizes } Var(u_1^T \boldsymbol{X}) \text{ subject to } u_1^T u_1 = 1$$

$$z_2 = u_2^T \boldsymbol{X} \text{ maximizes } Var(u_2^T \boldsymbol{X}) \text{ subject to } Cov(u_1^T \boldsymbol{X}, u_2^T \boldsymbol{X}) = 0 \text{ and } u_2^T u_2 = 1$$

and

$$z_i = u_i^T \boldsymbol{X} \text{ maximizes } Var(u_i^T \boldsymbol{X}) \text{ subject to } Cov(u_k^T \boldsymbol{X}, u_i^T \boldsymbol{X}) = 0 \text{ and } u_i^T u_i = 1$$

$$\forall k < i$$

where u_i is a linear rotation vector [9,17].

Logit Function

When dealing with discrete data, as will be used in this study, logistic regression is generally used to fit a model. Logistic regression is a form of the Generalized Linear Model (GLM) which assumes the response variable follows a Bernoulli distribution as

opposed to a normal distribution. Instead of fitting a linear model, logistic regression fits a probability curve between 0 and 1 using the logit function,

$$logit(\pi_i) = log\left(\frac{\pi_i}{1-\pi_i}\right) = \sum_{j=0}^{p} \beta_j x_{ij}$$

where π_i is the probability of the i[th] response being a success (i.e. $P(y_i = 1) = \pi_i$) [10].

The Data

The data considered in this study contains eleven physiochemical properties on the quality of wine courtesy of the University of California, Irvine, Machine Learning Repository [5]. The quality of each wine was determined by professional wine judges. When determining the quality of a wine, there are many things to consider. A wine expert typically marks on a few specific areas: appearance – the color and clarity, aroma – the smell at the wine and above the glass, body – how the wine feels in the mouth, taste – the actual flavor of the wine, and finish – how the wine lingers on the taste buds after it has been swallowed. These characterizations are molded based on the variety of grape, region and climate that the vineyard is located, the fermentation process, and how the wine is aged. The wine in the data set for this study comes from the northwest region of Minho in Portugal. The wines were tested for several common physiochemical traits and then evaluated by a minimum of three sensory judges for overall quality on a scale of 0 to 10 with 10 being the best. The median score of quality was recorded for each wine. Data was collected from May 2004 to February 2007 [4]. While data on both red and white wines

is available, this study will focus on the analysis of the white wine. However, there is a similar study by L.E. Melkumova et al. who analyzed red wine data [19].

The covariates sampled for this study are now more precisely defined. Fixed acidity refers to the amount of tartaric acid (g/dm^3) in the wine. Tartaric acid controls the acidity of wine and contributes to the overall tartness. The volatile acid measured is acetic acid (g/dm^3), which may lead to a sour taste in high concentrations. Citric acid (g/dm^3) is essential for fermentation and adds a "freshness" to wine. However, a large portion of the citric acid is consumed by bacteria during this process. The overall sweetness of a wine can be quantified by the amount of residual sugar (g/dm^3) left in the wine after the yeast has been absorbed [3,4]. A high amount of sodium chloride (g/dm^3), referenced in this study more broadly as "chlorides," may result in a salty or soapy taste. Total sodium dioxide (mg/dm^3) is broken into two groups: free and bound. Bound SO_2 combines with pigment and sugar, but it does not have much influence on the overall taste or smell of the wine. Free SO_2 is able to react with the oxygen in the wine and affect the flavor as well as the bouquet and aroma. Too much SO_2 can lead to a pungent odor similar to that of a recently struck match. Generally, sweeter, fruitier wines have a higher amount of SO_2 [4,8]. The density (g/cm^3) of a wine is highly correlated to the amount of dry extract in a wine, which helps determine the mouthfeel [4,20]. The pH of a wine is an assessment of the fixed acidity, including tartaric, malic, citric, and succinic acid. Potassium sulphate (g/dm^3), referred to here as "sulphates," is important for the improvement of the aroma of a wine [4]. The alcohol content (percent by volume) is a natural result of fermentation and affects the aroma, taste, mouthfeel, and finish of a wine [3,4]. Each of these properties helps to shape the overall quality of a wine.

Table 2.1

Table of Covariates and Corresponding Variable Names

y	quality
X_1	fixed.acidity
X_2	volatile.acidity
X_3	citric.acid
X_4	residual.sugar
X_5	chlorides
X_6	free.sulfur.dioxide
X_7	total.sulfur.dioxide
X_8	density
X_9	pH
X_{10}	sulphates
X_{11}	alcohol

CHAPTER III

METHODS

Data Preparation

To begin the analysis, summary statistics of the data and the OLS regression model are reported. Diagnostic plots of the residuals are also included. Additionally, the data are checked for multicollinearity using Variance Inflation Factors (VIF). The VIF is calculated using OLS regression for each explanatory variable as a function of all of the other explanatory variables [19]. For example,

$$X_1 = \alpha_0 + \alpha_2 X_2 + \cdots + \alpha_p X_p + \epsilon \tag{3.1}$$

is the model for variable $X_1 = [x_{11}\ x_{12}\ \cdots\ x_{1n}]^T$ where ϵ is the nx1 error vector. Then the VIF for $\widehat{\beta_j}$ from equation (2.1) is

$$VIF_j = \frac{1}{1 - R_j^2}$$

where R_j^2 is the coefficient of determination associated with a regression of X_j onto all of the other predictors as established in equation (3.1). For the purposes of this study, the VIF factors were calculated using the "car" package in R. As a rule of thumb, a VIF value greater than 10 indicates a high level of multicollinearity. Exploratory investigations into the covariates were also made using PCAs of the centered data. The

loadings associated with each principle component tell which covariates make up that component and how much each contributes to the variance.

The quality of the different regression models used in this study will be verified by a comparison of the residual sum of squares (RSS) between a set of data used to determine the coefficient estimates and another set of data in which to fit the model. To begin, the wine data set was randomly and evenly split into a "training" subset and a "testing" subset. The training data set was used to create the regression models using the techniques that follow. Those models were then used to predict the responses of both the training and testing data set.

Statistical Computations

The computations of ridge regression estimators and LASSO estimators are evaluated using two main packages in R, *glmnet* and *lars*. The *glmnet* function can be used to fit a generalized linear model for either LASSO or ridge regression by changing the value of alpha for the parameter

$$\frac{(1-\alpha)}{2}\|\beta\|_2^2 + \alpha\|\beta\|_1.$$

Clearly, a value of 1 will result in the LASSO tuning parameter while a value of 0 will result in the ridge regression parameter [11,12]. Alternatively, the *lars* function fits a LASSO regression sequence as well as least angle regression and forward stagewise regression, which are related to the LASSO. Note that the default of this function is the LASSO [13].

To begin using the *glmnet* function, a range of 100 values for λ was chosen with equally-spaced powers from 10^{-2} to 10^{5}. Using R, the coefficients for both ridge regression and LASSO were plotted against log(λ) on the lower axis. The upper axis represents the number of coefficients in the model for that value of log(λ). The values for coefficients seen in these plots are standardized, but the results are ultimately given on the original scale of the data. Using the *lars* function, a similar yet very distinctive plot is depicted. Instead of plotting against log(λ), the *lars* function plots against fraction of the ℓ_1-norm, i.e. $\sum_{j=1}^{p}|\beta_j| / \max \sum_{j=1}^{p}|\beta_j|$. As opposed to starting with every coefficient and slowly removing them as in the *glmnet* plots, this approach starts with all but one coefficient set to zero and introduces another with each step. The number of steps taken is displayed on the top horizontal axis while the number corresponding to the variable appears on the right vertical axis.

Mean Square Error Computations

The next step of the process is to select the λ tuning parameter. This parameter controls the magnitude of the absolute value of the parameter estimates, $\hat{\beta}_i$. It is important that λ is large enough to give an accurate representation of the data without being so large that the model is overfitted. Cross-validation is then used to determine a "best" λ to use. Both the *glmnet* and *lars* functions have an existing k-fold cross-validation procedure built in. During cross-validation, the data are split into k equal-sized subsamples. Each subsample is then used to validate the model as the other k-1 subsamples are used as training data. This is repeated using each of the k subsamples as

the testing set. Note that this study uses k=10 as this is widely accepted as a value that is large enough for proper validation without causing overfitting of the data. For a discussion on other methods of selecting λ, see Marron, J.S. [18]. The "best" λ is chosen such that the model has the lowest MSE. The MSE for a model based on λ_s, s=1,…,t, can be written as

$$MSE_{\lambda_s} = \frac{1}{K}\sum_{k=1}^{K} RSS_{k,\lambda_s}$$

where there are k=1,…,K subsamples and

$$RSS_{k,\lambda_s} = \sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p} x_{ij}\widehat{\beta}_j(k,\lambda_s)\right)^2$$

is the RSS for a given set of regression coefficients [19]. To avoid overfitting the models, the λ_s with the highest MSE within 1 Standard Error (s.e.) of the minimum MSE was also considered for future analysis. Plots of the MSE for each log(λ_s) were produced for both ridge regression and the LASSO method.

Once the values for λ are established, the estimates for each ridge regression and the LASSO are calculated. Then each set of coefficient estimates are used to predict the values of both the training set and testing set of data. For each of these models, the RSS for the predicted values is calculated. These are the values that will ultimately judge how well each type of regression model worked for this data set.

Binomial Computations

The data being analyzed in this study are discrete and ordinal, not binary. While functions have been created in R to analyze ordinal data using Ridge Regression and LASSO [1], those functions have been removed from the Comprehensive R Archive Network (CRAN) database because errors were found and not addressed. However, studies have been done to show that categorical data with at least 5-7 categories may be treated as continuous [21]. Thus, the data in this study do not require a logistic regression approach. For the remainder of this study, the data are treated as continuous unless otherwise stated.

It may be of interest to model the data from the standpoint of separating "superior" wines from those that are "not superior." For the purposes of this study, a wine with a quality rating of 8 or above is considered "superior." Wines with a superior quality rating are considered a success and given a response value of 1. All other wines are given a response value of 0. The data are then analyzed similarly to previous calculations using the binomial family for the *glmnet* function.

CHAPTER IV

RESULTS

Analysis of Covariates

A summary of the sampled data can be seen in the Table 4.1. The OLS regression for parameter estimates from equation (2.2) are listed in Table 4.2. This model has a coefficient of multiple determination of $R^2 = 0.2818704$. Diagnostic plots of the residuals are included in Figure 4.1. In the plot of Residuals vs Fitted in Figure 4.1 (a), there is some fluctuation in variance, but most values appear to be centered around zero. This near-linear trend is also depicted in the Normal Q-Q plot in Figure 4.1 (b). The Scale-Location plot in Figure 4.1 (c) shows an increase in the standardized residuals as fitted values approach 6, indicating that the data may be heteroscedastic. The plot of Residuals vs Leverage in Figure 4.1 (d) suggests that data point number 2782 may be an outlier. The Variance Inflation Factors, VIF_j, for each of the explanatory variables can be found in Table 4.3. There are multiple factors with values over 10, implying that the multicollinearity is high. This multicollinearity combined with the heteroscedasticity suggests OLS may not provide a good model for the data.

Table 4.1

Summary Statistics of White Wine Data

	y	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}
Min.	3	3.8	0.08	0.00	0.6	0.009	2.0	9.0	0.9871	2.72	0.22	8.0
Q1	5	6.3	0.21	0.27	1.7	0.036	23.0	108.0	0.9917	3.09	0.41	9.5
Med.	6	6.8	0.26	0.32	5.2	0.043	34.0	134.0	0.9937	3.18	0.47	10.4
Mean	5.878	6.855	0.2782	0.3342	6.391	0.04577	35.31	138.4	0.9940	3.188	0.4898	10.51
Q3	6	7.3	0.32	0.39	9.9	0.05	46.0	167.0	0.9961	3.28	0.55	11.4
Max.	9	14.2	1.1	1.66	65.8	0.346	289.0	440.0	1.0390	3.82	1.08	14.2

Table 4.2

OLS Regression Parameter Estimates

β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}	β_{11}
150.2	0.06552	-1.863	0.02209	0.08148	-0.2473	0.003733	-0.0002857	-150.3	0.6863	0.6315	0.1935

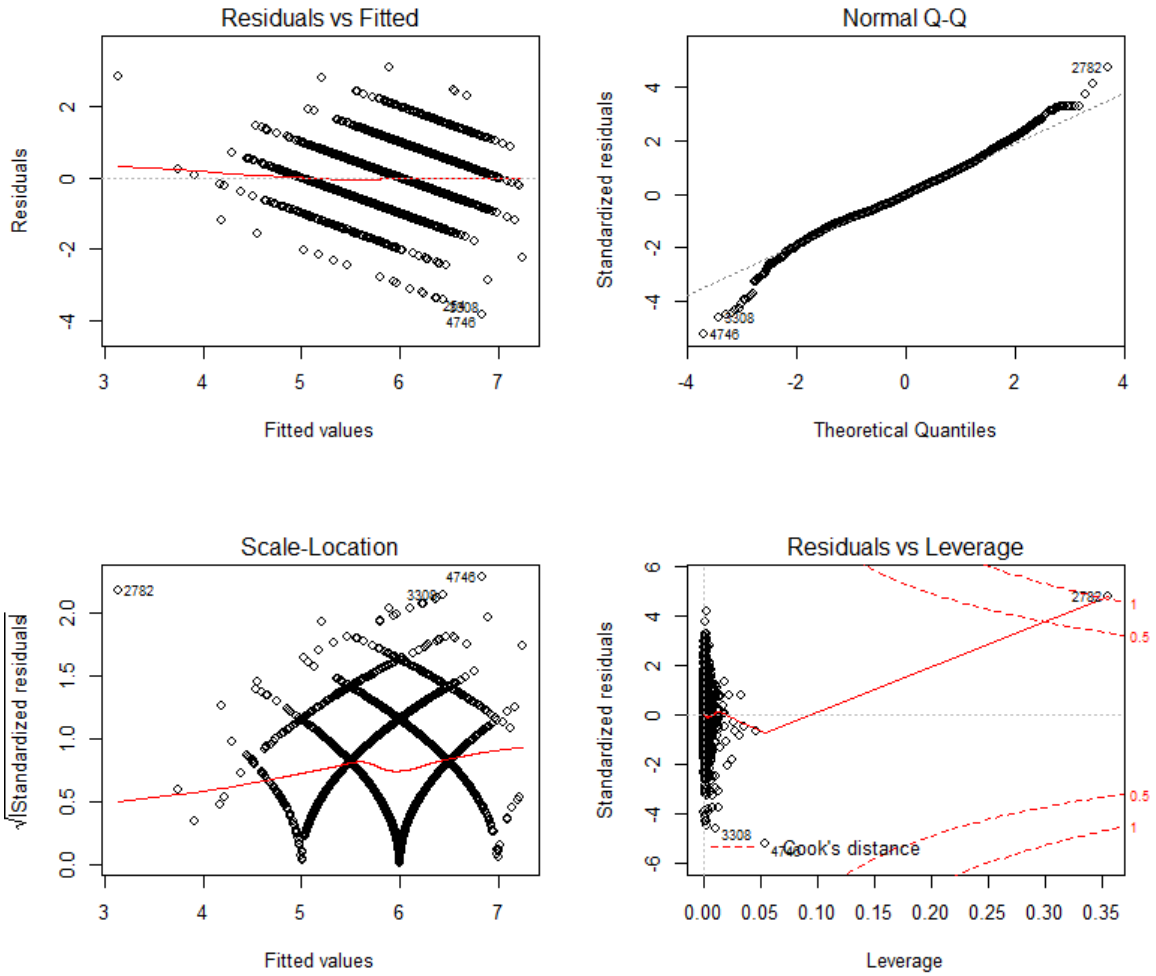

Figure 4.1 Diagnostic plots of residuals

Table 4.3

Variance Inflation Factors for j=1,…,11 Predictors

VIF_1	VIF_2	VIF_3	VIF_4	VIF_5	VIF_6	VIF_7	VIF_8	VIF_9	VIF_{10}	VIF_{11}
3.30	1.14	1.15	15.38	1.26	1.82	2.38	39.29	2.52	1.18	11.19

Further analysis of the covariates involved a look at the PCAs. The data was first centered so that the mean of each covariate was zero. This centered data was used to

ensure that the resulting components are orthogonal. A summary of the first four components and their loadings are depicted in the figures below. The first component accounts for almost 91% of the total variation of the covariates. Combining this with the second component makes up nearly 99% of the total variation. This is further exemplified in the scree plot in Figure 4.4. From the loadings, it is shown that the amount of free sulfur dioxide and total sulfur dioxide make up the first two components. Note that the loadings are orthogonal. This is a product of using the centered data for analysis. This relationship can also be seen in Figure 4.5 where the amount of variation from each covariate is plotted for the first principle component on the x-axis and the second component on the y-axis.

```
Importance of components:
                          Comp.1      Comp.2     Comp.3       Comp.4
Standard deviation     43.9444987 12.97761545 4.64290763 1.0364366647
Proportion of Variance  0.9096573  0.07933386 0.01015427 0.0005060045
Cumulative Proportion   0.9096573  0.98899121 0.99914548 0.9996514857
```

Figure 4.2 Summary of first four principle components

```
                            Comp.1        Comp.2        Comp.3       Comp.4
fixed.acidity         1.544402e-03 -9.163498e-03 -1.290026e-02  0.147657857
volatile.acidity      1.690037e-04 -1.545470e-03 -9.288874e-04 -0.015451710
citric.acid           3.386506e-04  1.403069e-04 -1.258444e-03  0.005004529
residual.sugar        4.732753e-02  1.494318e-02 -9.951917e-01 -0.084200484
chlorides             9.757405e-05 -7.182998e-05 -7.849881e-05  0.006573232
free.sulfur.dioxide   2.618770e-01  9.646854e-01  2.639318e-02  0.006381109
total.sulfur.dioxide  9.638576e-01 -2.627369e-01  4.278881e-02 -0.010613506
density               3.596983e-05 -1.836319e-05 -4.468979e-04  0.001151657
pH                    3.384655e-06 -4.169856e-05  7.017342e-03 -0.017027136
sulphates             3.409028e-04 -3.611112e-04  2.142053e-03 -0.002600913
alcohol              -1.250375e-02  6.455196e-03  8.272268e-02 -0.985062967
```

Figure 4.3 Loadings of first four principle components

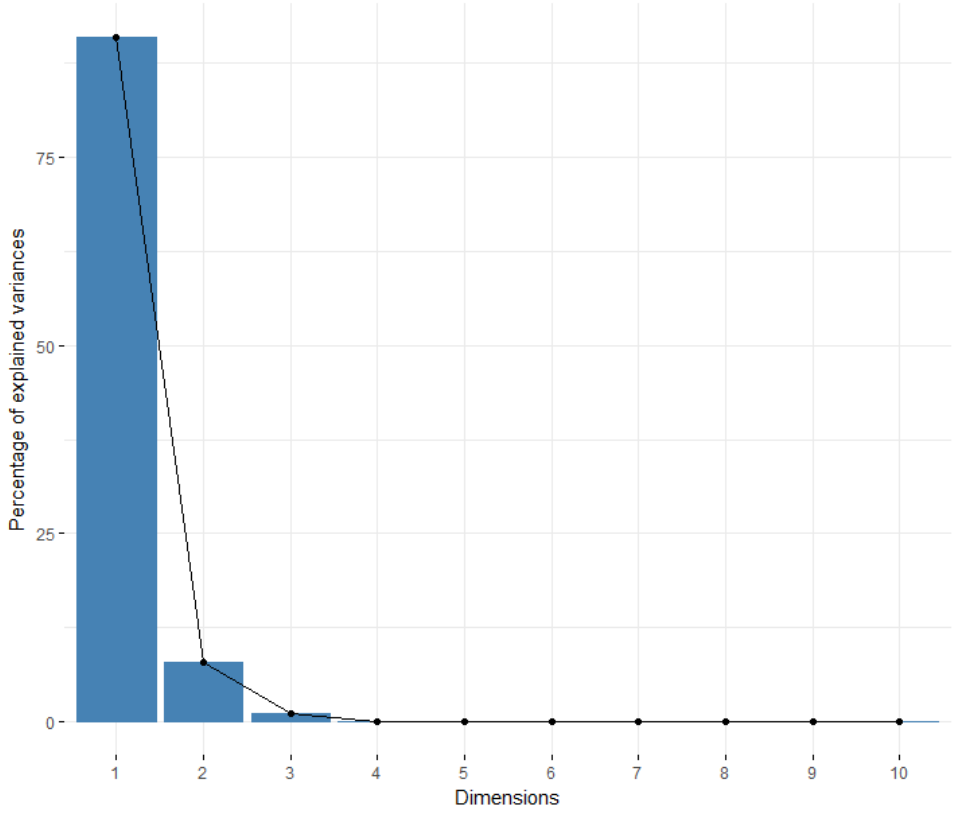

Figure 4.4 Scree plot of principle components for centered data

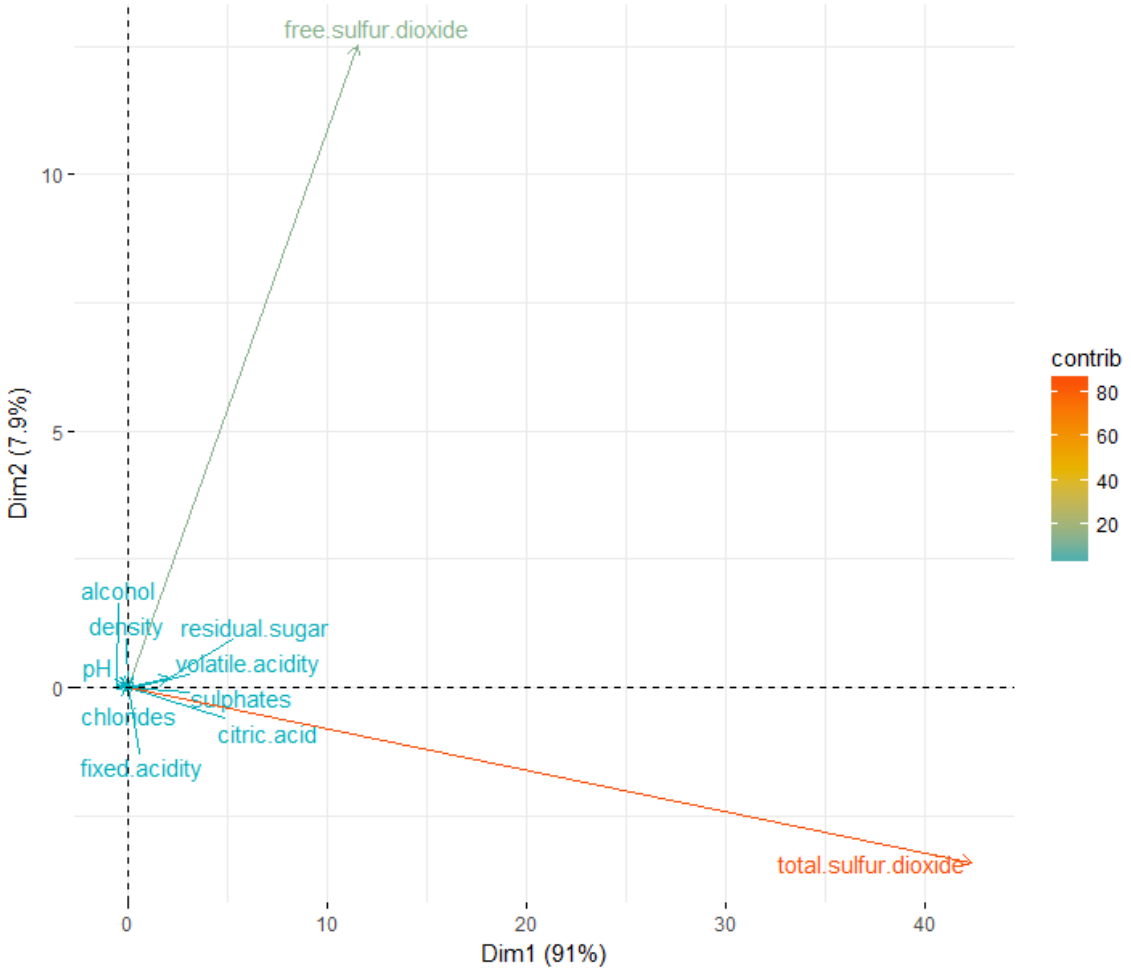

Figure 4.5 Plot displaying the amount each covariate contributes to the first and second principle components

k-Fold Cross Validation

Displayed in Figure 4.6 are the plots of the standardized coefficient estimates for ridge regression and the LASSO method using *glmnet*. The top horizontal axis shows that for ridge regression, the coefficients begin equivalent to those of the OLS model and move closer to zero, but they never actually reach it. Until λ becomes larger than e^5, the density appears to have the most influence on the quality of the wine. On the other hand, the plot of LASSO coefficients shows that the model essentially ends up with no

covariates for large enough λ (e.g. when $\log(\lambda)$ is equal to or greater than 0). The plot of LASSO coefficients gives further insight into the process. Covariates are added into this model in the following order: alcohol, volatile acidity, free sulfur dioxide, residual sugar, fixed acidity, chlorides, sulphates, pH, density, total sulfur dioxide, -fixed acidity, fixed acidity, citric acid. The negative sign here implies that a covariate was removed after it had been introduced to the model. Note that the covariates which contributed most to the variation in the PCA analysis are not necessarily of most importance to these regression models. This is because the parameters which have the most variability do not necessarily have a higher effect on the response than those covariates with less variability. For instance, total sulfur dioxide has a larger variance than volatile acidity, because the units of measurement are several orders of magnitude different. However, volatile acidity appears to have a much more drastic effect on the quality of wine than total sulfur dioxide. For this reason, it is better to use LASSO or ridge regression than PCA for this data set.

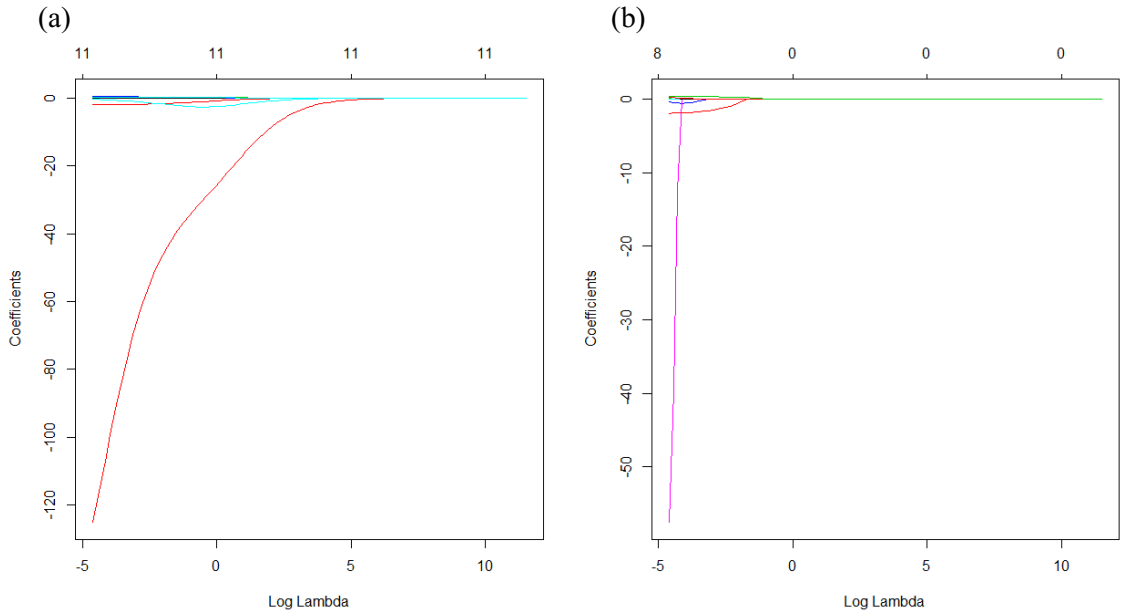

Figure 4.6 Coefficient estimates for ridge regression (a) and LASSO (b) using *glmnet* function

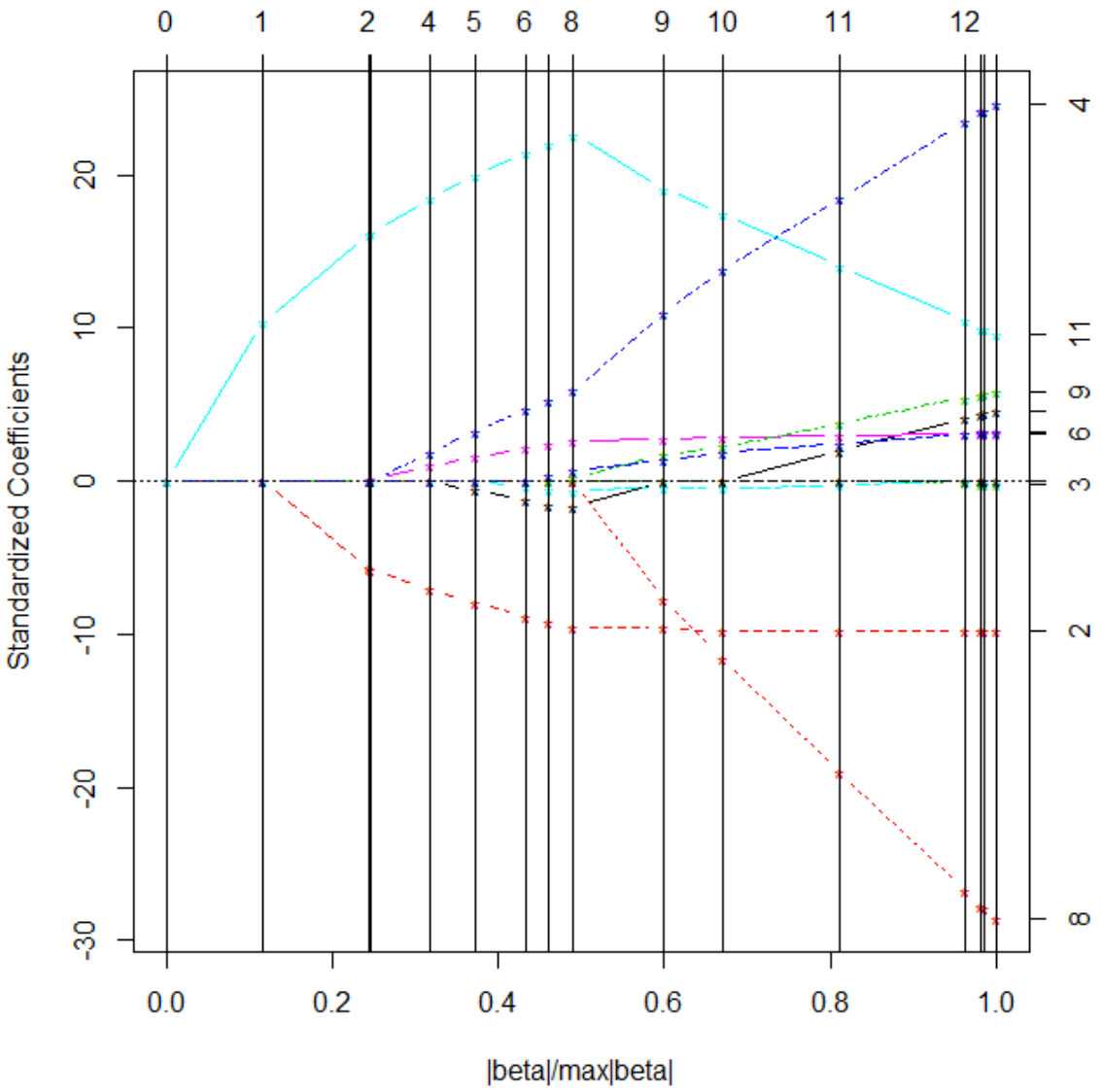

Figure 4.7 Plot of coefficients of LASSO regression using *lars* function

For the k-fold cross validation to determine which λ to use, k was left at a default of 10 for both the *glmnet* and *lars* functions. The plots in Figure 4.8 show the MSE in relation to log(λ) as a result of the 10-fold cross validation from *glmnet*. The numbers on the top x-axis represent the number of non-zero coefficients in the model for that log(λ). The first dotted line in each plot aligns with the minimum MSE_{λ_s}, and the second dotted

line corresponds to the smallest MSE_{λ_s} within 1 s.e. of the minimum. The plot produced by *lars* is not quite as informative as it lacks the upper x-axis and the dotted lines associated with the MSE_{λ_s}. Note that the x-axis for the *lars* LASSO plot is $\sum_{j=1}^{p}|\beta_j| / \max \sum_{j=1}^{p}|\beta_j|$, so it appears as almost a mirror image of the plot of the LASSO produced by *glmnet*.

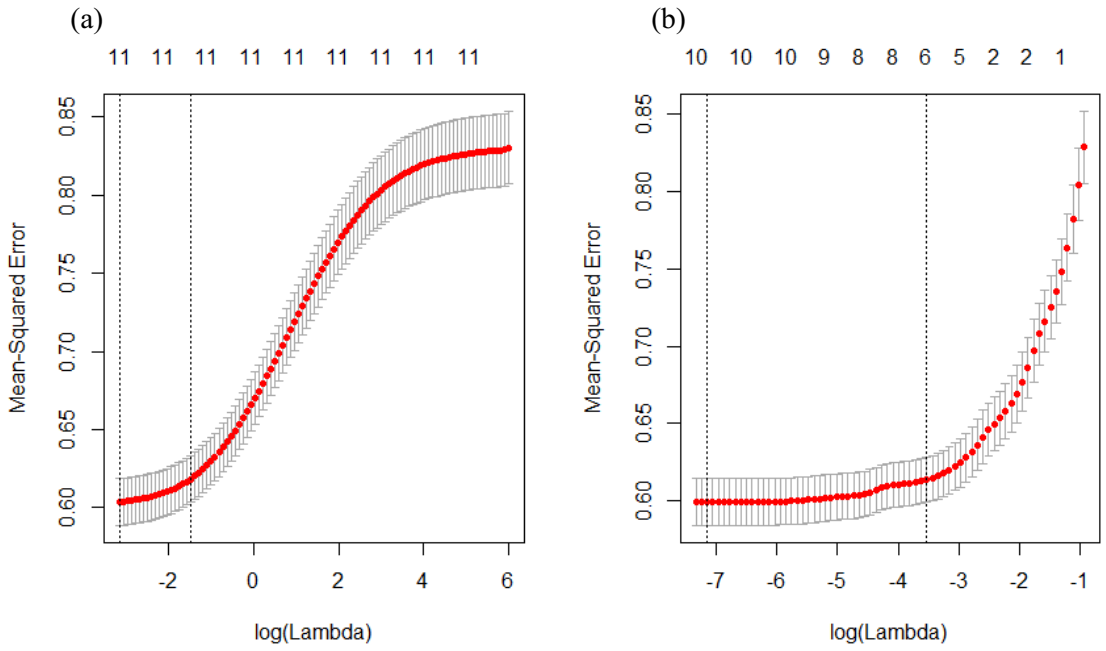

Figure 4.8 Plots of MSE_{λ_s} for each $\log(\lambda)$ and the corresponding number of coefficients in the model produced by *glmnet* for ridge regression (a) and LASSO (b)

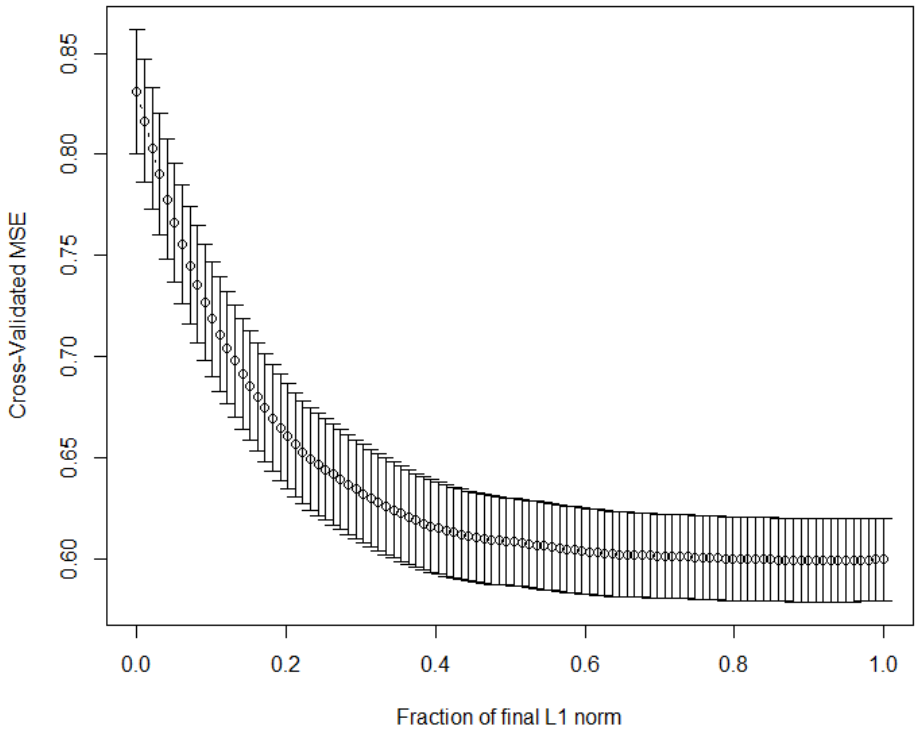

Figure 4.9 Plot of MSE_{λ_s} as a function of a fraction of the ℓ_1-norm produced by *lars* for LASSO

The *glmnet* function allows for easy extraction of the "best" λ and the λ 1 s.e. above the minimum MSE. For ridge regression, λ_{best}=0.04364007 and λ_{1se}=0.2328939 whereas, for the LASSO method, λ_{best}=0.0007805678 and λ_{1se}=0.02938788. These values were then used to find the model coefficients for ridge regression and the LASSO. The λ values are not readily available from *lars*, so additional coding was necessary to find the coefficients for the LASSO model in this format. All of the resulting coefficients can be seen in Table 4.4. It is worth noting that the models for the minimum MSE_{λ_s} using the LASSO have a different number of coefficients for each function used. However, the LASSO models 1 s.e. above the minimum MSE_{λ_s} have nearly identical coefficients. Finally, each model was then used to predict the quality values of both the training set

and testing set of data. The resulting RSS's for each model are listed in Table 4.5. The similarities between the LASSO model found using the *glmnet* function and the *lars* function are evident in these tables as would be expected. The RSS values for 1 s.e. above the minimum are reasonably higher than those calculated for the minimum MSE_{λ_s}. The OLS model is still the "best" for the training set, but the LASSO, particularly when using *lars* funcion, is a better fit than ridge regression for the white wine data. However, ridge regression is the "best" fit, followed by the LASSO and then OLS, for the testing set.

Table 4.4

Coefficient Estimates for Each Regression Model

R function	lm	glmnet				lars	
Regression Model	OLS	Ridge, min	Ridge, 1 s.e.	LASSO, min	LASSO, 1 s.e.	LASSO, min C_p	LASSO, 1 s.e.
(Intercept)	150.2	73.15127	42.19911	177.7704	2.772966324	189.676654	2.762675
fixed.acidity	0.06552	0.006409598	-0.02785866	0.0926979	-0.028790597	0.1038489	-0.02994802
volatile.acidity	-1.863	-1.879111	-1.471504	-1.965073	-1.766327517	-1.961865	-1.77989169
citric.acid	0.02209	-0.02746897	0.03523509	-0.02495405	.	-0.02829795	.
residual.sugar	0.08148	0.05066576	0.02598576	0.09353071	0.018531079	0.09812901	0.01895951
chlorides	-0.2473	-0.9739314	-2.227193	-0.0253627	-0.314432475	.	-0.3399665
free.sulfur.dioxide	0.003733	0.004326843	0.003892894	0.003661909	0.002519617	0.003689699	0.00257406
total.sulfur.dioxide	-0.0002857	-0.000506909	-0.000706571	.	.	.	.
density	-150.3	-71.61539	-39.13527	-178.0809	.	-190.1895	.
pH	0.6863	0.3213479	0.1834351	0.6983401	.	0.7462817	.
sulphates	0.6315	0.3704554	0.2457842	0.5479971	.	0.5724971	.
alcohol	0.1935	0.2710696	0.2262042	0.174793	0.344050765	0.1612205	0.34580294

NOTE: A period represents a covariate not included in the model (i.e. $\widehat{\beta}_j = 0$).

Table 4.5

RSS Values for Each Regression Model

R function	lm	glmnet				lars	
Regression Model	OLS	Ridge, min	Ridge, 1 s.e.	LASSO, min	LASSO, 1 s.e.	LASSO, min C_p	LASSO, 1 s.e.
Training data set	1449.813	1460.769	1500.089	1450.012	1489.077	1449.833	1487.929
Testing data set	1316.359	1311.863	1334.39	1314.046	1339.885	1315.521	1339.312

A scatterplot matrix is provided in Figure 4.10 for each of the six covariates present in the LASSO model for λ_s within 1 s.e. of MSE_{λ_s}. It is easy to see trends in these particular covariates for "good" wines marked by green circles for a quality rating of 8 and brown circles for a quality rating of 9. Similarly, the "bad" wines marked by blue circles for a quality rating of 3 and pink circles for a quality rating of 4 tend to be aggregated together for each of the covariates. It appears that sensory judges preferred wines that were medium to high in alcohol level, had a light aroma due to the lack of free SO_2, were not very sweet or had a small amount of residual sugar, and had moderate to low amounts of both tartaric acid (fixed acidity) and acetic acid (volatile acidity).

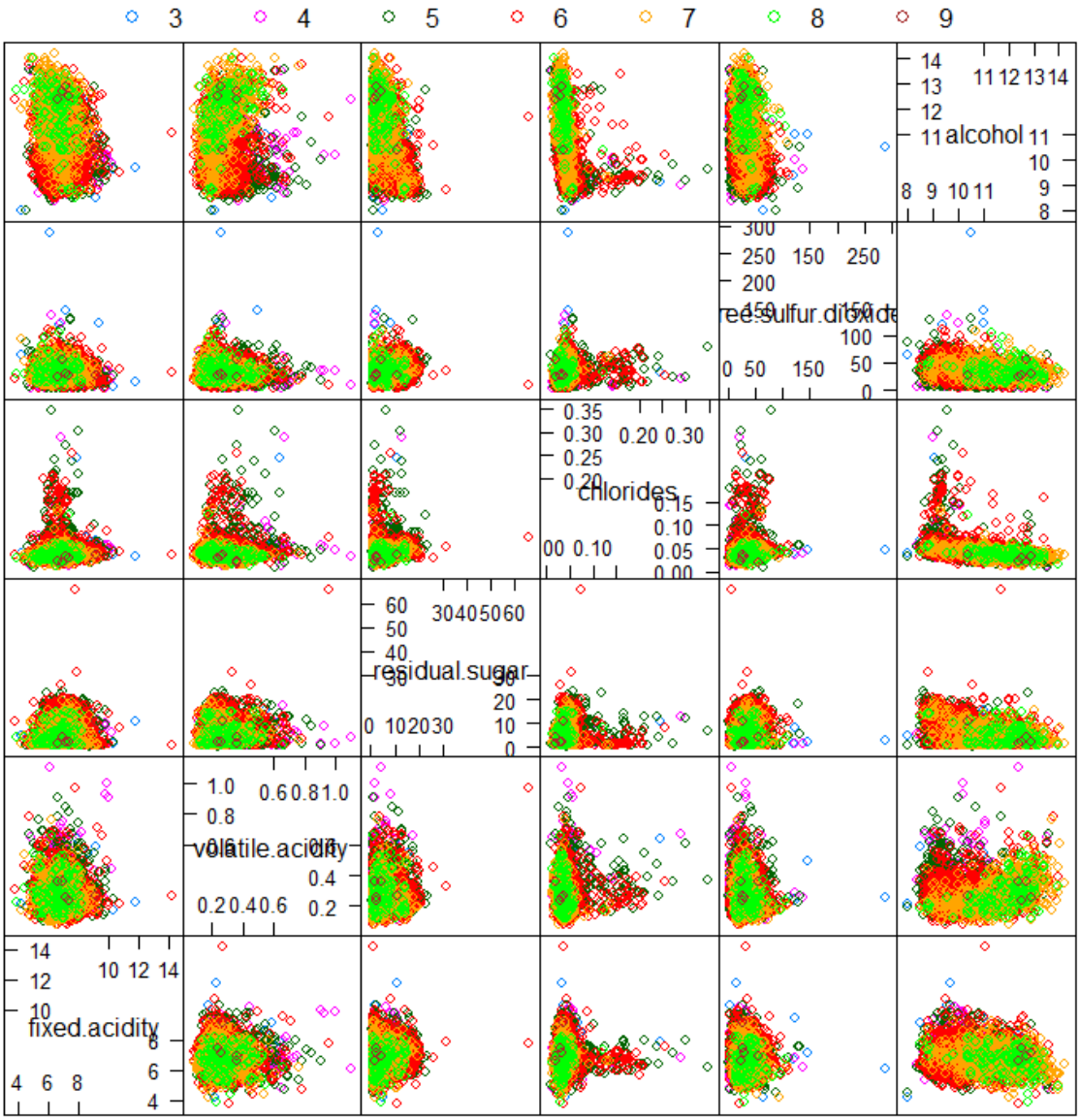

Figure 4.10 Scatterplot matrix of six covariates introduced into LASSO model within 1 s.e. of minimum MSE_{λ_s}

Binomial Analysis

The VIF values in Table 4.6 show that there is still a high amount of multicollinearity in the binary data, suggesting that Ridge Regression and LASSO may be better options for fitting the model. It is obvious from Figure 4.11 that density still plays a large role in the model, but the coefficient estimates are vastly different from

those of the model treated as continuous data. These differences are seen even more clearly by the choice of λ in the plot of MSE_{λ_s} in Figure 4.12.

Table 4.6

Variance Inflation Factors for Binary Data

VIF_1	VIF_2	VIF_3	VIF_4	VIF_5	VIF_6	VIF_7	VIF_8	VIF_9	VIF_{10}	VIF_{11}
4.22	1.33	1.18	19.30	1.30	2.19	2.84	55.20	2.88	1.26	15.05

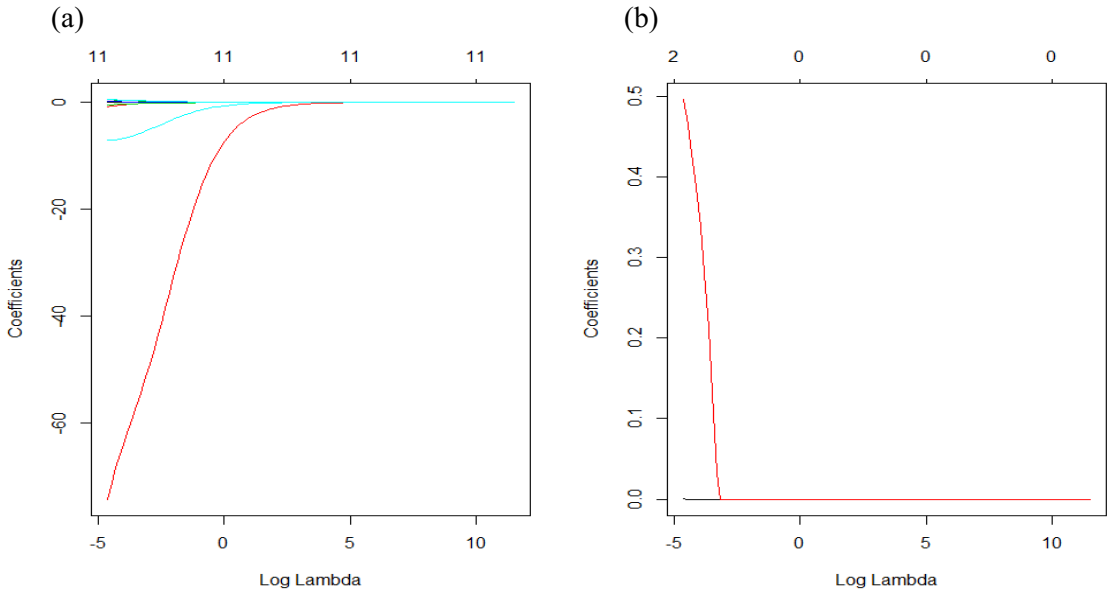

Figure 4.11 Coefficient estimates for binary data for ridge regression (a) and LASSO (b) using *glmnet* function

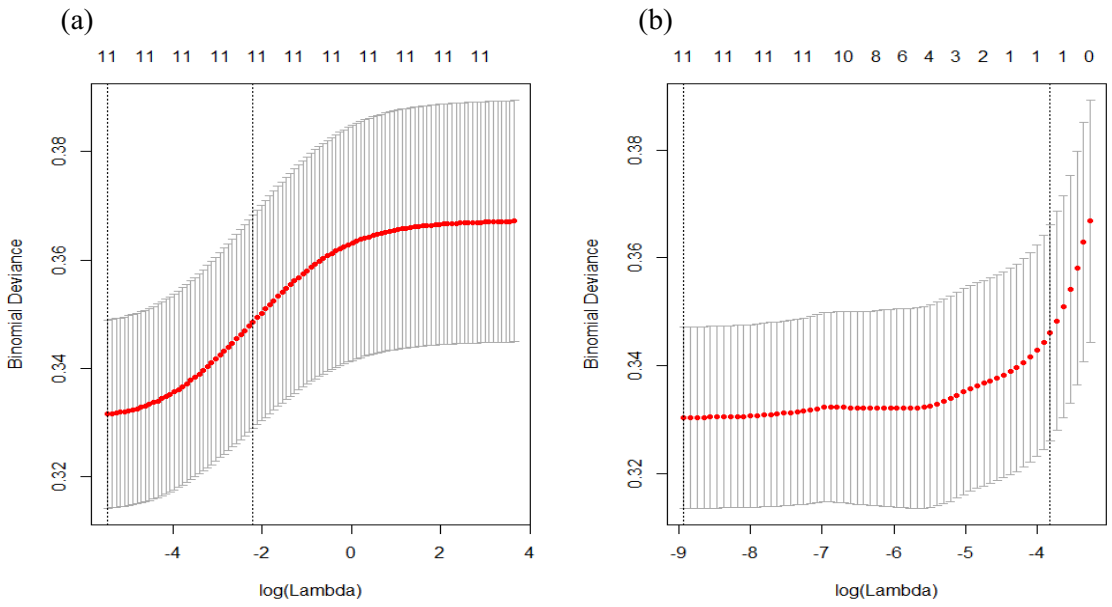

Figure 4.12 Plots of MSE_{λ_s} for each log(λ) produced by *glmnet* for ridge regression (a) and LASSO (b) for binary data

The coefficients for the binary data for each model using the "best" λ and the λ 1 s.e. above the minimum MSE are compared to the logistic regression model in Table 4.7. The "best" λ for ridge regression and the LASSO produce similar coefficient estimates for this data. However, the estimates for λ 1 s.e. above the minimum MSE appear vastly different for ridge regression and the LASSO. The RSS's in Table 4.8 show that ridge regression and LASSO fit the binary data much more efficiently than the standard logistic model. For this particular data set, LASSO using the "best" λ produces the smallest RSS for the training data while ridge regression using the λ 1 s.e. above the minimum MSE produces the smallest RSS for the testing data.

Table 4.7

Coefficient Estimates of Binary Data for Each Regression Model

R function	glm	glmnet			
Regression Model	Logistic	Ridge, min	Ridge, 1 s.e.	LASSO, min	LASSO,1s.e.
(Intercept)	576.7	9.122772763	2.642726	14.67172	-0.0938288
fixed.acidity	0.5985	0.009085571	-0.00032256	0.01356793	.
volatile.acidity	-2.101	-0.01651811	-0.00169704	-0.01401589	.
citric.acid	-0.9714	-0.27006172	-0.01125881	-0.02545494	.
residual.sugar	0.321	0.006689001	0.001782247	0.008870137	.
chlorides	-1.547	0.069864531	-0.1728456	0.1020844	.
free.sulfur.dioxide	0.01698	0.000774085	0.000444592	0.000748073	.
total.sulfur.dioxide	-0.002733	-0.0001312	-0.00007885	-0.00010958	.
density	-600.5	-9.66458017	-2.842664	-15.3032	.
pH	2.679	0.046345164	0.0109989	0.06601116	.
sulphates	1.201	0.020690385	0.00221647	0.02868537	.
alcohol	0.2013	0.025300048	0.0179908	0.01958325	0.01316304

NOTE: A period represents a covariate not included in the model (i.e. $\widehat{\beta_J} = 0$).

Table 4.8

RSS Values of Binary Data for Each Regression Model

R function	glm	glmnet			
Regression Model	Logistic	Ridge, min	Ridge, 1 s.e.	LASSO, min	LASSO, 1 s.e.
Training data set	35639.94	100.2675	101.18	100.2261	102.6142
Testing data set	36691.6	66.41774	66.26058	66.45467	66.96833

CHAPTER V

CONCLUSIONS

With the addition of a tuning parameter to form a penalized OLS for ridge regression and LASSO, there are definite differences in the outcomes from the models as seen by the coefficient estimates and the error values. The OLS model remains unbiased, but it is shown by the RSS values that there may be times when a model with some bias is a better fit. Additionally, there may be instances when it is better to "shrink" the coefficient estimates so small that a variable is removed completely from the model. This is the case in using the LASSO method. The *glmnet* and *lars* functions are both adequate for modeling continuous data using the LASSO. While the available commands and plots are slightly different, the general outcomes are very similar between the two. For continuous data, it is essentially a choice of whether the user is more comfortable speaking in terms of $\log(\lambda)$ or a fraction of the ℓ_1-norm. However, using data sets with binary responses limits the options. Comparing all of the RSS values from this study to those of the study by L.E. Melkumova et al. shows that there is no single "best" regression model to use [19]. Each data set is unique and must be treated as such. More time should be spent on developing ways to use ridge regression and LASSO for ordinal data and other discrete data that is not necessarily binary.

REFERENCES

[1] Archer, K., Hou, J., Zhou, Q., Ferber, K., Layne, J., Gentry, A. (2014), "ordinalgmifs: An R Package for Ordinal Regression in High-dimensional Data Settings," *Cancer Inform*, 13, pp. 187-195.

[2] Billingsley, P. (1995), *Probability and Measure*, 3.

[3] California Wineries, "Wine Chemistry," www.calwineries.com/learn/wine-chemistry.

[4] Cortez, P., Cerdeira, A., Almeida, F., Matos, T., Reis, J. (2009), "Modeling wine preferences by data mining from physiochemical properties," *Decision Support Systems*, 47(4), pp. 547-553.

[5] Dua, D. and Karra Taniskidou, E. (2017), UCI Machine Learning Repository, Irvine, CA: University of California, School of Information and Computer Science, http://archive.ics.uci.edu/ml.

[6] Dziak, J., Coffman, D., Lanza, S., Li, R. (2012), "Sensitivity and specificity of information criteria," *The Pennsylvania State University Technical Report Series #12-119*, The Methodology Center, College of Health and Human Development, The Pennsylvania State University.

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

[8] Eisenman, L., "Sulfur Dioxide in Wine," www.gencowinemakers.com/docs/Sulfur%20Dioxide.pdf.

[9] Faraway, J. (2015), "Shrinkage Methods," *Linear Models with R*, 2, pp. 161-181.

[10] Fonti, V., Belitser, E. (2017), "Feature Selection using LASSO," *Research Paper in Business Analytics*, Vrije Universiteit Amsterdam.

[11] Friedman, J., Hastie, T., Tibshirani, R. (2010), "Regularization Paths for Generalized Linear Models via Coordinate Descent," *Journal of Statistical Software*, 33(1).

[12] Friedman, J., Hastie, T., Tibshirani, R., Simon, N., Narasimhan, B., Qian, J. (2018), "LASSO and Elastic-Net Regularized Generalized Linear Models," https://cran.r-project.org/pub/R/web/packages/glmnet/glmnet.pdf.

[13] Hastie, T., Efron, E. (2013), "Least Angle Regression, LASSO and Forward Stagewise," https://cran.r-project.org/web/packages/lars/lars.pdf.

[14] Hastie, T., Taylor, J., Tibshirani, R., Walther, G. (2007), "Forward stagewise regression and the monotone lasso," *Electronic Journal of Statistics*, 1(1-29).

[15] Hayashi, F. (2000), "Finite-Sample Properties of OLS," *Econometrics*, pp. 3-69.

[16] James, G., Witten, D., Hastie, T., Tibshirani, R. (2017), "Linear Model Selection and Regularization," *An Introduction to Statistical Learning with Applications in R*, pp. 214-229.

[17] Johnson, R., Wichern, D. (2013), "Principle Components," *Applied Multivariate Statistical Analysis*, 6, pp. 430-453.

[18] Marron, J.S. (1987), "A Comparison of Cross-Validation Techniques in Density Estimation," *The Annals of Statistics*, 15(1), pp. 152-162.

[19] Melkumova, L.E., Shatskikh, S.Ya. (2017), "Comparing Ridge and LASSO estimators for data analysis," *Procedia Engineering*, 201, pp. 746-755.

[20] Neto, F.S., de Castilhos, M.B., Telis, V.R., Telis-Romero, J. (2015), "Effect of ethanol, dry extract and reducing sugars on density and viscosity of Brazilian red wines," *Journal of the Science of Food and Agriculture*, 95, pp. 1421-1427.

[21] Rhemtulla, M., Brosseau-Liard, P., Savalei, V. (2012), "When Can Categorical Variables Be Treated as Continuous? A Comparison of Robust Continuous and Categorical SEM Estimation Methods Under Suboptimal Conditions," *Psychological Methods*, 3, pp. 354-373.

[22] Tibshirani, R. (2013), "Regression 2: More perspectives, shortcomings (continued)," *ST 36-462/36-662: Data Mining, lecture 14* [PowerPoint slides], http://www.stat.cmu.edu/~ryantibs/datamining/lectures/14-reg2-marked.pdf.

[23] Tibshirani, R. (1996), "Regression Shrinkage and Selection via the LASSO," *Journal of the Royal Statistical Society: Series B*, 58, pp. 267-288.

[24] van der Vaart, A.W. (2000), "Stochastic Convergence," *Asymptotic Statistics*, pp. 5-24.