Spring 2019

Ridge Regression and Lasso Estimators for Data Analysis

Dalip Kumar
Missouri State University, Dalip488@live.missouristate.edu

RIDGE REGRESSION AND LASSO ESTIMATORS FOR DATA ANALYSIS

A Master's Thesis

Presented to

The Graduate College of

Missouri State University

In Partial Fulfillment

Of the Requirements for the Degree

Master of Science, Mathematics

By

Dalip Kumar

May 2019

RIDGE REGRESSION AND LASSO ESTIMATORS FOR DATA ANALYSIS

Mathematics

Missouri State University, May 2019

Master of Science

Dalip Kumar

ABSTRACT

An important problem in data science and statistical learning is to predict an outcome based on data collected on several predictor variables. This is generally known as a regression problem. In the field of big data studies, the regression model often depends on a large number of predictor variables. The data scientist is often dealing with the difficult task of determining the most appropriate set of predictor variables to be employed in the regression model. In this thesis we adopt a technique that constraints the coefficient estimates which in effect shrinks the coefficient estimates towards zero. Ridge regression and lasso are two well-known methods for shrinking the coefficients towards zero. These two methods are investigated in this thesis. Ridge regression and lasso techniques are compared by analyzing a real data set for a regression model with a large collection of predictor variables.

KEYWORDS: ridge regression, lasso, cross validation, mean square error, Akaike information criterion, Bayesian information criterion

RIDGE REGRESSION AND LASSO ESTIMATORS FOR DATA ANALYSIS

By

Dalip Kumar

A Master's Thesis
Submitted to the Graduate College
Of Missouri State University
In Partial Fulfillment of the Requirements
For the Degree of Master of Science, Mathematics

May 2019

Approved:

George Mathew, Ph.D., Thesis Committee Chair

Songfeng Zheng, Ph.D., Committee Member

Yingcai Su, Ph.D., Committee Member

Julie Masterson, Ph.D., Dean of the Graduate College

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

The linear regression model consists of a response variable which depends on several predictors or explanatory variables. Linear regression models are widely used today in business administration, economics, engineering, and the social, health, and biological sciences. Successful applications of these models require a sound understanding of both the underlying theory and the practical problems that are encountered in using the models in real-life situations. Many methods can be applied to linear regression such as least square approach, maximum likelihood function and bayesian approach. We will use ridge regression and lasso techniques to analyze a data set for a regression model with a large collection of predictor variables.

Beyond this introduction to analyzing a data using ridge regression and lasso techniques, the layout of this expository work is as follows:

In Chapter 2, we will discuss some methods such as best subset selection, stepwise selection and ridge regression. We are always concerned about choosing the optimal model. We will discuss some techniques such as cross-validated prediction error, Bayesian Information Criterion (BIC), Mallow's C_P, Akaike Information Criterion (AIC), or adjusted R^2 to select the best model.

In Chapter 3, we will discuss about the lasso estimator, variable selection property of lasso, orthonormal design case, standard errors, prediction errors and lasso parameter, and Bayesian interpretation for ridge regression and the lasso.

In Chapter 4, we will compare ridge regression and lasso techniques to analyze a data set for a regression model with a large collection of predictor variables.

RIDGE REGRESSION

2.1 Introduction

Regression is a statistical procedure that attempts to determine the strength of the relationship between one response variable and a series of other variables known as independent or explanatory variables. Linear regression models are widely used in diverse fields. Many methods can be employed to obtain a linear regression model. These methods include least square approach, maximum likelihood estimation and Bayesian approach. All these methods estimates the parameters in linear regression model so that we can apply these estimates for getting predictions of outcomes. Generally, we use least squares method for determining the set of coefficients which minimizes the residual sum of squares. Model selection is an important part of any statistical analysis with a wide range of applications where the number of variables is much larger than the sample size. When the number of predictors is more than the sample size, we use alternative fitting procedures which can yield better prediction accuracy and model interpretability. If the relationship between the response and predictors is linear then the least squares estimates will have low bias. If the number of observations is much larger than number of variables then the least squares estimates will perform on test observations and it will have low variance. If it is not much larger than number of predictors then it will have lot of variability. Otherwise, if number of predictors is much larger than number of observations then it will have variance infinite. So we cannot use least squares estimate. However, we can reduce the variance by shrinking the estimated coefficients. Therefore, we can predict the response for observations with significant improvements in the accuracy. Moreover, if we set these estimated coefficients to zero then we can acquire a model that is more easily interpreted (see James et al., 2013).

2.2 Model selection methods

2.2.1 Discrete process. In discrete process, variables are either retained or dropped from the model. Methods such as best subset selection and stepwise selection fall under the category of discrete process. We discuss these methods one by one briefly.

a) Best Subset Selection: Generally, it fits 2^p possible models involving p predictors. We select the best subset which has the smallest Residual Sum of Squares (RSS) or largest value of R^2 where R^2 is the coefficient of determination. Our objective is to select a model that has a low test error. Therefore, we use Cross-validated prediction error, Bayesian Information Criterion (BIC), Mallow's C_p, Akaike Information Criterion (AIC), or adjusted R^2 to select the best subset. If the values of p is large then this method becomes computationally infeasible (see James et al., 2013).

b) Stepwise Selection: We know that best subset selection approach suffers from statistical problems when p is large. We use two approaches- forward stepwise selection and backward stepwise selection that are best alternatives to best subset selection. The forward stepwise selection starts with a null model containing zero predictors. It then adds predictors one at a time until a model with the smallest RSS or largest R^2 value is obtained. On the other hand, the backward stepwise selection starts with a full model containing all p predictors. It then removes the predictor which is less useful one at a time. This process goes on until we get the best model having smallest RSS or largest value of R^2. We can use these approaches when the value of p is large. Moreover, both these approaches are computationally efficient (see James et al., 2013).

However, we are always concerned about choosing the optimal model. As discussed above, we select a model which has the smallest RSS or highest value of R^2. We wish to have a model

that has a low test error. There are many techniques such as Cross-validated prediction error, Bayesian Information Criterion (BIC), Mallow's C_P, Akaike Information Criterion (AIC), or adjusted R^2 to estimate the test error. We discuss these techniques one by one.

i) Cross-validated error: This approach takes fewer assumptions into account and provides a direct estimate of the test error. We can use this approach in model selection tasks where it is hard to determine the number of predictors or to estimate the error variance σ^2 . Cross-validation is a very general approach that can be administered to any statistical learning method in order to identify the method which has the lowest test error. It has low bias and is computationally cheap (see James et al., 2013).

Cross- validated method mainly addresses two issues- k- fold cross validation and Leave-one-out cross validation (LOOCV). In k-fold cross validation, we split the data into K-equally sized folds. The first fold is treated as a validation set, and the method is fit on the remaining k-1 folds. We then compute the mean squared error, MSE_1 , on the observations in the validation set. We repeat this procedure k times and each time, a different group of observations is treated as a validation set. This process provides us k estimates of the test error, $MSE_1,....,MSE_k$. This k-fold cross validation is computed by averaging these values,

$$CV_{(k)} = \frac{1}{k}\sum_{i=1}^{k} MSE_i$$

The common value of k we choose is 5 or 10 for performing k-fold cross-validation.

On the other hand, the LOOCV divides the n data points into two subsets. We consider the first observation (x_1, y_1) as validation set and the remaining (n-1) as training set. The first training set consists of all observations but not (x_1, y_1) and second one contains all but not (x_2, y_2) and so on. Every time we leave out one observation from the data set. The first training

set does not use (x_1, y_1) in the fitting process. By this fitting process we predict y_1 by $\hat{y}_1$. Then $MSE_1 = (y_1 - \hat{y}_1)^2$ provides an approximately unbiased estimate for the test error and in second one, we compute $MSE_2 = (y_2 - \hat{y}_2)^2$. Repeating this approach n times produces n squared errors, $MSE_1, ..., MSE_n$. The LOOCV estimate for the test MSE is the average of these n test error estimates:

$$CV_{(n)} = \frac{1}{n}\sum_{i=1}^{n} MSE_i$$

Moreover, LOOCV may pose computational problems when the value of n is too large. However, we can perform calculations easily with fast computers. We can observe that LOOCV is a special case of k- fold cross validation when k is equal to n. So we can use this method directly to estimate the test error and it helps us to better assess a model on its predictive performance (see James et al., 2013).

ii) Akaike Information Criterion (AIC): When we include additional variables in a model, it penalizes those variables which leads to increased error. Using this technique, we can fit a large class of models by maximum likelihood estimation procedure. It is given by

$$AIC = \frac{1}{n\hat{\sigma}^2}(RSS + 2p\hat{\sigma}^2)$$

where RSS is the residual sum of squares on a training set of data, p is the number of predictors and $\hat{\sigma}^2$ is an estimate of the variance associated with each response in the linear model. Then we choose a model that has the lowest AIC.

i) Bayesian Information Criterion (BIC): This technique is derived from a Bayesian point of view. BIC applies strong penalty on models with many variables. It also uses maximum likelihood estimation to fit the model. It is determined by the relation

$$BIC = \frac{1}{n}(RSS + \log(n) p\hat{\sigma}^2)$$

where RSS is the residual sum of squares on a training set of data, p is the number of predictors and $\hat{\sigma}^2$ is an estimate of the variance associated with each response in the linear model. Then we select a model that has the lowest BIC.

ii) Adjusted R^2: The coefficient of determination is a statistical measure of how well the real data points are approximated by the regression predictions. The value of R^2 lies between 0 and 1. If the value of R^2 is 0 then it shows that there is no linear association between response and predictor variable. If the value of R^2 closer to 1 then it indicates that all the observations fall on the fitted regression line. The model would be best when the value of R^2 is higher. For a least squares model with p variables, the adjusted R^2 statistic is calculated as

$$\text{Adjusted } R^2 = 1 - \frac{RSS/(n-p-1)}{TSS/(n-1)}$$

where RSS is the residual sum of squares on a training set of data, TSS is the total sum of squares, p is number of predictors, and n is the number of observations.

Using this technique, we choose a model by maximizing R^2 or equivalently, selecting a model by minimizing $\frac{RSS}{n-d-1}$. The higher value of adjusted R^2, the better the model is (see James et al., 2013).

iii) Mallows C_P : It uses ordinary least squares to assess the fit of a regression model that has been estimated. It is applied in the context of model selection, where a number of predictor variables are available for predicting some outcome. Our goal is to find the best model involving a subset of these predictors. It is given by

$$C_P = \frac{1}{n}(RSS + 2p\hat{\sigma}^2)$$

where RSS is the residual sum of squares on a training set of data, p is the number of predictors and $\hat{\sigma}^2$ is an estimate of the variance associated with each response in the linear model. We choose a model that has lowest C_P (see Mallows, C.L., 1973, Steven, G, 1996).

To summarize, we can say that all these techniques have the same objective that is finding a model that predicts well.

2.2.2 Continuous process. In continuous process, we can fit a model by including all the predictors but constrains or shrinks the estimated coefficients towards zero. Using this technique, we can remarkably reduce the variance. In continuous process, Ridge regression and Least Absolute Shrinkage and Selection Operator (LASSO) are the two main methods for shrinking the regression coefficients towards zero. We discuss these methods one by one.

a) Ridge regression: The ridge regression coefficients estimates $\hat{\beta}^R$ are the values that minimize

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}\beta_j{}^2 = RSS + \lambda\sum_{j=1}^{p}\beta_j{}^2 \qquad (1)$$

where $\lambda \geq 0$ is a tuning parameter, to be determined separately. Ridge regressions seeks coefficient estimates that fit the data well, by making the RSS small. However, the second term, $\lambda\sum_j \beta_j{}^2$, called a shrinkage penalty, is small when $\beta_1,...\beta_p$ are close to zero, and it has the effect of shrinking the estimates of β_j towards zero. When $\lambda = 0,$ the penalty term has no effect, and ridge regression will produce the least squares estimates. However, as $\lambda \to \infty,$ the impact of the shrinkage penalty grows, and the ridge regression coefficient

estimates will approach zero. It is a technique that analyzes multiple regression data that suffer from multi-collinearity. Least squares estimates are unbiased and variances are large when this type of situation arises. It then reduces the standard errors by adding a degree of bias to the regression estimates (see James et al., 2013).

We briefly discuss the term of Multi-collinearity, its effects, sources and modifications.

Multicollinearity is the existence of near-linear relationships among the response variables. It means some of the variables are linear combinations of the others. It can create faulty estimates of the regression coefficients and degrade the predictability of the model by inflating the standard errors of the regression coefficients as well as by deflating the partial t-tests for the regression coefficients.

We need to identify the sources of multicollienarity. The analysis, the corrections, and the interpretation of the model are impacted by the source of multicollinearity. Some of the sources are the following:

a) It leads to problem of multicollinearity when we gather data from narrow subspace of the response variables.
b) It creates multicollinearity when manufacturing or service processes have constraints on response variables, either physically, politically, or legally.
c) When there are more variables than observations, it leads to over-defined model and thus multicollinearity occurs.
d) It increases the problem of multicollinearity when we use response variables that are powers or interactions of an original set of variables (see Montgomery, D.C., 1982).

The modifications to be done depend on the source of collinearity. If it is due to data collection then we should gather the data over a wider subspace of the explanatory variables. We could

simplify the model by using variable selection techniques if the choice of the linear model when multicollinearity is present. If an observation or two induced the multicollinearity, then remove those observations. Also, we need to be careful while selecting the variables (see Montgomery, D.C., 1982).

In the following chapter, we discuss Least Absolute Shrinkage and Selection Operator (Lasso) which falls under continuous process. We will see the models generated by Lasso are generally much easier to interpret than those by ridge regression.

LEAST ABSOLUTE SHRINKAGE AND SELECTION OPERATOR

3.1 Introduction

We have seen in the previous that methods in discrete processes either include or exclude variables. When the goal is to select a subset of variables that best represent the model, then this technique of including or excluding variables may be beneficial. On the other hand, the methods such as ridge regression under continuous process deals with tuning parameter that shrinks the estimated coefficients towards zero. Although it gives better accuracy in predicting the model but it creates challenge in interpreting the model when the number of variables is quite large. We use Lasso to overcome this drawback of ridge regression. We discuss this method in detail.

The term 'LASSO' stands for Least Absolute Shrinkage and Selection Operator. The main function of this method is to shrink some coefficients and set others to zero. Lasso uses a penalty named l_1 which makes some of the estimated coefficients exactly equal to zero. It minimizes the residual sum of squares subject to the sum of the absolute value of the coefficients being less than a constant. This method enjoys some of the good features of both subset selection and ridge regression. We can easily interpret the models generated by Lasso than those produced by subset selection and ridge regression.

The motivation for the lasso came from an interesting proposal of Breiman (1993). Breiman's non-negative garotte minimizes

$$\sum_{i=1}^{N}\left(y_i-\alpha-\sum_j c_j\hat{\beta}_j^0 x_{ij}\right)^2 \text{ subject to } c_j\geq 0, \sum c_j \leq t \ .$$

The garotte starts with the Ordinary Least Squares (OLS) estimates and shrinks them by non-negative factors whose sum is constrained. Breiman showed that the garotte has consistently lower prediction error than subset selection and is competitive with ridge regression except when the true model has many small non-zero coefficients. The garotte has a drawback that its solution depends on both the sign and the magnitude of the OLS estimates. Lasso evades the explicit use of the OLS estimates. Lasso is even better than OLS. Lasso also deals with overfitting by deriving a biased estimator but with possibly lower variance than the variance of OLS parameter. Moreover, models created by Lasso are sparse models meaning that they involve only a subset of the variables. The idea behind Lasso is quite general and we can apply Lasso method to a variety of statistical models ranging from extensions to generalized regression models (see Breiman, 1993).

Assume that we have data $(x^i, y_i), i = 1, 2, ...N$ where $x^i = (x_{i1}, x_{i2}, ..., x_{ip})^T$ are the predictor variables and y_i are the responses. We assume either that the observations are independent or that $y_i's$ are conditionally independent given the $x_{ij}'s$. Let us suppose that the x_{ij} are standardized so that

$$\sum_i x_{ij} / N = 0 \ , \quad \sum_I x_{ij}^{\ 2} / N = 1 .$$

Assuming $\hat{\beta} = (\hat{\beta}_1, ..., \hat{\beta}_p)^T$, the lasso estimate $(\hat{\alpha}, \hat{\beta})$ is defined by

$$(\hat{\alpha}, \hat{\beta}) = \quad \arg\min \left\{ \sum_{i=1}^{N} \left(y_i - \alpha - \sum_j \beta_j x_{ij} \right)^2 \right\} \text{ subject to } \sum_j |\beta_j| \le t$$

where $t \ge 0$ is a tuning parameter. The solution for α is $\hat{\alpha} = \bar{y}$ for all t (see Tibshirani, 1996). Without loss of generality, we can assume that $\bar{y} = 0$ and hence omit α . The solution of equation (1) is a quadratic programming problem with linear inequality constraints. Here $t \ge 0$ is a tuning parameter which controls the amount of shrinkage is applied to the estimates.

We assume that $\overline{\beta}_j^0$ be the full least squares estimates and let $t_0 = \sum | \beta^0{}_j |$. Some coefficients may be exactly equal to zero and the solution approaches towards zero when the values of t is less than t_0. The main objective of the Lasso is to minimize the quantity

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} | \beta_j | = RSS + \lambda \sum_{j=1}^{p} | \beta_j | \qquad (2)$$

where $\lambda \geq 0$ is a tuning parameter. The penalty l_1 norm is used by Lasso in statistics. The l_1 norm of a coefficient vector β is given by $\| \beta \|_1 = \sum | \beta_j |$. When the tuning parameter λ is large, the penalty l_1 has the effect of forcing some of the coefficients estimates to be exactly equal to zero. Lasso gives the least squares fit when the value of $\lambda = 0$. Lasso can generate a model involving any number of variables depending on the value of λ (see James et al., 2013).

Using another formulation for Ridge regression and the Lasso, we can prove that the Lasso and Ridge regression coefficient estimates solve the problems

$$\min_{\beta} imize \left\{ \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij} \right)^2 \right\} \text{ subject to } \sum_{j=1}^{p} | \beta_j | \leq s \qquad (3)$$

and

$$\min_{\beta} imize \left\{ \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij} \right)^2 \right\} \text{ subject to } \sum_{j=1}^{p} \beta_j{}^2 \leq s \qquad (4)$$

respectively. For some value of s, the equations (2) and (3) will give the same lasso coefficient estimates for every value of λ. Similarly, for every value of λ, there is a corresponding s such that equations (1) and (4) will give the same ridge regression coefficient estimates (see James et al., 2013).

In particular, when p=2, the equation (3) indicates that the lasso coefficient estimates have the smallest RSS out of all points that lie within the diamond defined by $|\beta_1| + |\beta_2| \leq s$. Similarly, the ridge regression estimates have the smallest RSS out of all points that lie within the circle defined by $\beta_1^2 + \beta_2^2 \leq s$.

We can think of lasso as follows. Our objective is to find the set of estimates of coefficients such that the RSS is as small as possible, subject to the constraint that there is a budget s for how large $\sum_{j=1}^{p} |\beta_j|$ can be. The equation (2) will yield the least squares solution when s is large enough that the least squares falls within the budget. If s is small, then $\sum_{j=1}^{p} |\beta_j|$ must be small in order to avoid the budget. Similarly, we try to find the set of coefficient estimates that lead to the smallest RSS, subject to the condition that $\sum_{j=1}^{p} \beta_j^2$ is less than the budget s (see James et al., 2013).

The equations (3) and (4) reveal a close relationship between the lasso, ridge regression, and best subset selection. Consider the problem

$$\min_{\beta} imize \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\} \text{ subject to } \sum_{j=1}^{p} I(\beta_J \neq 0) \leq s \qquad (5)$$

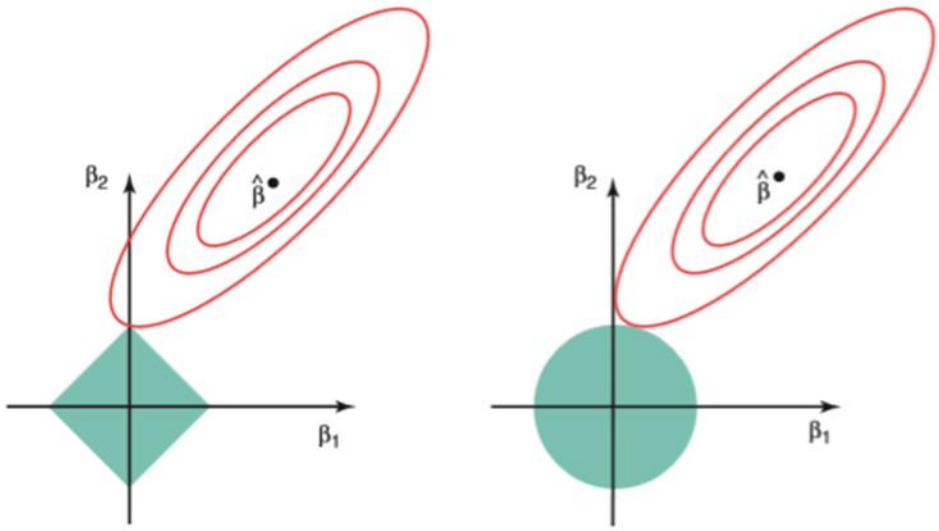
where $I(\beta_J \neq 0) = \begin{Bmatrix} 1, if\, \beta_j \neq 0 \\ 0, otherwise \end{Bmatrix}$ is an indicator variable.

Given the requirement that no more than s coefficients can be nonzero, we try to find the set of coefficient estimates such that RSS is as small as possible. The above equation is equivalent to best subset selection. When p is large, then solving (5) is computationally infeasible because it requires all $\begin{pmatrix} p \\ s \end{pmatrix}$ models containing s predictors. Therefore, we can observe that ridge regression

and the lasso are best alternatives to best subset selection as far as computation is concerned. Moreover, when s is sufficiently small, the lasso performs feature selection (see James et al., 2013).

3.2 Variable selection property of lasso

We know that lasso produces coefficient estimates that are exactly equal to zero. Moreover, it does not occur with ridge regression, which has the requirement $\sum \beta_j^{\,2} \le s$ rather than $\sum |\beta_j| \le s$. Figure 1 represents estimation picture for lasso and ridge regression.



a) b)

Figure 1. Estimation picture for (a) the lasso and (b) ridge regression.

It shows that contours of the error and constraint functions for the lasso and ridge regression. The solid blue areas are the constraint regions, $|\beta_1| + |\beta_2| \le s$ and $\beta_1^{\,2} + \beta_2^{\,2} \le s$, while the red ellipses are the contours of the RSS (see James et al., 2013).

The formulations (3) and (4) can be used to shed light on the issue. Here we consider only two parameters β_1 and β_2 . The estimates by the lasso and ridge regression are illuminated in above Figure 1. The least squares solution is marked as $\hat{\beta}$, while the blue diamond and circle represent the lasso and ridge regression constraints in (3) and (4), respectively. The ridge regression and lasso estimates will be the same as the least squares estimates when the constraint regions will constrain $\hat{\beta}$ for large value of s. However, in above figure, the least squares estimates are not the same as the lasso and ridge regression estimates because least squares estimates lie outside of the diamond and the circle (see James et al., 2013).

The ellipses that are centered around $\hat{\beta}$ represent regions of constant RSS. It means that all the points on a given ellipse share a common value of the RSS. The RSS increases when the ellipses expand away from the least squares coefficient estimates. Equations (3) and (4) indicate that the lasso and ridge regression coefficient estimates are given by the first point at which an ellipse contacts the constraint region. Since ridge regression has a circular constraint with no sharp points, this intersection will not generally occur on an axis, and so the ridge regression coefficient estimates will be exclusively non-zero. On the other hand, the lasso constraint has corners at each of the axes, and so the ellipse will often intersect the constraint region at an axis. When this situation occurs, then one of the coefficients will be equal to zero. In higher dimensions, many of the coefficient estimates may be equal to zero simultaneously. In figure 1, when the contours touch the vertex then $\beta_1 = 0$, and so β_2 is chosen as the only relevant parameter in the model (as can be seen from James et al., 2013).

Here, we considered the simple case of $p = 2$. When $p = 3$, then the constraint region for the ridge regression becomes a sphere, and the constraint region for the lasso becomes a

polyhedron. When $p > 3$, the constraint for ridge regression becomes a hypersphere, and the constraint for the lasso becomes a polytope. However, the key ideas represented in figure 1 still hold. In particular, the lasso leads to feature selection when $p > 2$ due to the sharp corners of the polyhedron or polytope (more details can be found in James et al., 2013).

3.3 Orthonormal Design Case

Now we discuss the nature of the shrinkage which can be gathered from the orthonormal design case. Suppose that **X** be the $n \times p$ design matrix with x_{ij} as the $(i, j)^{th}$ entry, and letting $X^T X = I$, the identity matrix.

The solutions to equation (2) are easily shown to be

$$\hat{\beta}_j = sign(\hat{\beta}_j^{\,0})(|\,\hat{\beta}_j^{\,0}\,| - \gamma)^+ \qquad (6)$$

where γ is determined by the condition $\sum |\,\hat{\beta}_j\,| = s$ (see Tibshirani, 1996).

In the orthonormal design case, best subset selection of size k reduces to choosing the k largest coefficients in absolute value and setting the rest of the coefficients equal to zero (more details can be found in James et al., 2013). For some choice of λ, this is equivalent to the following setting

$$\hat{\beta}_j = \begin{Bmatrix} \hat{\beta}_j^0, if\ |\,\hat{\beta}_j^{\,0}\,| > \lambda \\ 0, otherwise \end{Bmatrix}.$$

Ridge regression minimizes

$$\sum_{i=1}^{N} \left(y_i - \sum_j \beta_j x_{ij} \right)^2 + \lambda \sum_j \beta_j^{\,2}\ .$$

Or, equivalently, minimizes

$$\sum_{i=1}^{N}\left(y_i-\sum_j \beta_j x_{ij}\right)^2 \text{ subject to } \sum \beta_j^{\,2} \le s. \qquad (7)$$

The ridge solutions are $\frac{1}{1+\lambda}\hat{\beta}_j^0$ where γ depends on λ or s. The garotte estimates are $\left(1-\frac{\gamma}{\hat{\beta}_j^0}\right)^+ \hat{\beta}_j^0$. Figure 2 shows the coefficient shrinkage in orthonormal design case.

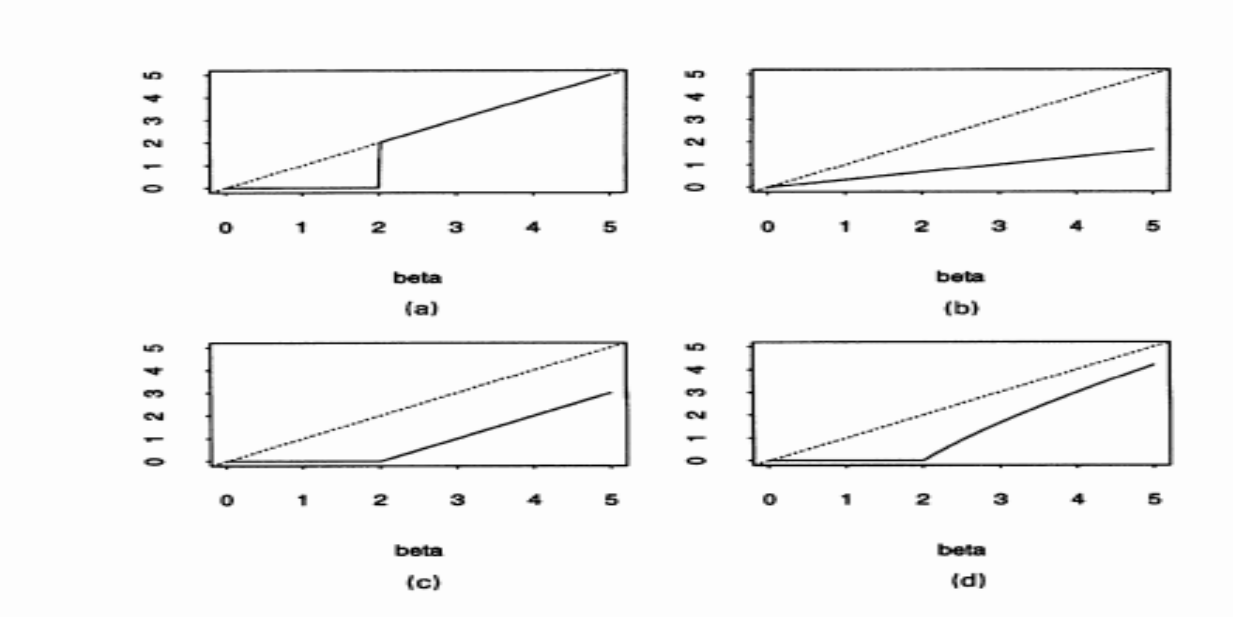


Figure 2. (a) Subset regression (b) Ridge regression (c) Lasso (d) the Garotte

It shows the form of these functions. In the figure, the solid line forms of coefficient shrinkage in the orthonormal design case and dotted line form 45° line for reference. Ridge regression scales the coefficients by a constant factor. On the other hand, lasso translates by a constant factor, truncating at 0. The garotte function is very similar to the lasso, with less shrinkage for larger coefficients (see Tibshirani, 1996).

3.4 More on two-predictor case

Without loss of generality, we assume that the least squares estimates $\hat{\beta}_j^0$ are both positive for p=2. Then we can prove that the lasso estimates are $\hat{\beta} = \left(\hat{\beta}_j^0 - \gamma\right)^+$ where γ is chosen so that $\hat{\beta}_1 + \hat{\beta}_2 = s$. This formula holds for $s \le \hat{\beta}_1^0 + \hat{\beta}_2^0$ and is valid even if the predictors are correlated. Solving for γ yields

$$\hat{\beta}_1 = \left(\frac{s}{2} + \frac{\hat{\beta}_1^0 - \hat{\beta}_2^0}{2}\right)^+, \hat{\beta}_2 = \left(\frac{s}{2} - \frac{\hat{\beta}_1^0 - \hat{\beta}_2^0}{2}\right)^+.$$

On the other hand, the form of ridge regression shrinkage depends on the correlation of the predictors (see Tibshirani, 1996).

3.5 Special case for Ridge regression and the Lasso

In order to get a better intuition about the behavior of ridge regression and the lasso, we consider a special case with $n = p$, and X a diagonal matrix with 1's on the diagonal and 0's in all off-diagonal elements. We assume that we are performing regression without an intercept in order to simplify the problem further. With these assumptions, the usual least squares problem simplifies to finding $\beta_1, \beta_2, ..., \beta_p$ that minimize $\sum_{j=1}^{p} (y_j - \beta_j)^2$ (more details can be found in James et al., 2013).

In this case, the least squares solution is given by $\hat{\beta}_j = y_j$ and in this setting, ridge regression amounts to finding $\beta_1, ..., \beta_p$ such that

$$\sum_{j=1}^{p} \left(y_j - \beta_j\right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

is minimized, and the lasso amounts to finding the coefficients such that

$$\sum_{j=1}^{p}\left(y_j-\beta_j\right)^2+\lambda\sum_{j=1}^{p}|\beta_j|$$

is minimized. One can prove that in this setting, the ridge regression estimates take the form

$$\hat{\beta}_j^{\,R}=\frac{y_j}{1+\lambda}$$

and the lasso estimates take the form

$$\hat{\beta}_j^{\,L}=\begin{cases}y_j-\lambda/2, y_j>\lambda/2;\\ y_j+\lambda/2, y_j<-\lambda/2;\\ 0, if\ |y_j|\le\lambda/2.\end{cases} \qquad (8)$$

More details can be found from James et al, 2013.

3.6 Standard Errors

It is difficult to obtain an accurate estimate of its standard error because lasso estimate is a non-linear and non-differentiable function of the response values even for a fixed value of s. By using the approach of bootstrap, we can fix s or may optimize over s for each bootstrap sample. By fixing s is analogous to selecting a best subset, and then using the least squares standard error for that subset (see Tibshirani, 1996).

We may derive an approximate closed form estimate by writing the penalty $\sum|\beta_j|$ as $\sum\beta_j^{\,2}/|\beta_j|$. Hence, at the lasso estimate $\tilde{\beta}$ we approximate the solution by a ridge regression of the form $\beta^*=\left(X^TX+\lambda W^-\right)^{-1}X^Ty$ where W is a diagonal matrix with diagonal elements $|\tilde{\beta}_j|$ and W^- denotes the generalized inverse of W and λ is chosen so that $\sum|\beta_j|^*=t$. The covariance matrix of the estimates may then be approximated by

$$\left(X^T X + \lambda W^-\right)^- X^T X (X^T X + \lambda X^-)^{-1} \hat{\sigma}^2$$

where $\hat{\sigma}^2$ is an estimate of the error variance. One difficulty lies with above formula is that it gives an estimated variance of zero for predictors with $\hat{\beta}_j = 0$ (more details about the derivation can be found in Tibshirani, 1996).

3.7 Prediction error and Estimation of lasso parameter t

Here we provide estimate of the lasso parameter t using three methods: cross-validation, generalized cross-validation and an analytical unbiased estimate of risk. The first two methods are applicable in the 'X-random' case, where it is assumed that the observations (X, Y) are drawn from some unknown distribution, and the third method applies to the X-fixed case. We might simply choose the most convenient method because in real problems there is often no clear distinction between the two scenarios (see Tibshirani, 1996).

Suppose that

$$Y = \eta(X) + \varepsilon$$

where $\mathrm{E}(\varepsilon) = 0$ and $Var(\varepsilon) = \sigma^2$. The mean-squared error (MSE) of an estimate $\hat{\eta}(X)$ is defined by

$$\mathrm{M}SE = \mathrm{E}\{\hat{\eta}(X) - \eta(X)\}^2 .$$

Here, the expected value taken over the joint distribution of X and Y, with $\hat{\eta}(X)$ fixed. A similar measure is the prediction error (PE) of $\hat{\eta}(X)$ given by

$$\mathrm{PE} = E\{Y - \hat{\eta}(X)\}^2 = MSE + \sigma^2 .$$

We estimate the prediction error for the lasso procedure by five fold cross-validation as described in Chapter 17 of Efron and Tibshirani (1993). The lasso is indexed in terms of the normalized

parameter $s = t / \sum \hat{\beta}_j^{\,0}$, and the prediction error is estimated over a grid of values of s from 0 to 1 inclusive. The value $\hat{s}$ yielding the lowest estimated PE is selected. Simulation results are reported in terms of MSE rather than PE. In this paper, we consider the linear models $\eta(X) = X\hat{\beta}$, the mean-squared error has the simple form

$$MSE = (\hat{\beta} - \beta)^T V (\hat{\beta} - \beta)$$

where V is the population covariance matrix of X (see Tibshirani, 1996).

We may derive a second method to estimate t from a linear approximation to the lasso estimate. The method is adopted from Tibshirani (1996). We write the constraint $\sum |\beta_j| \le t$ as $\sum \beta_j^{\,2} / |\beta_j| \le t$. This constraint is equivalent to adding a Lagrangian penalty $\lambda \sum \beta_j^{\,2} / |\beta_j|$ to the residual sum of squares, with λ depending on t. Thus we may write the constrained solution $\tilde{\beta}$ as the ridge regression estimator

$$\tilde{\beta} + \lambda W^{-})^{-1} X^T y$$

where W= $diag(|\tilde{\beta}$ W^{-} denotes a generalized inverse. Therefore the number of effective parameters in the constrained fit $\tilde{\beta}$ approximated by

$$p(t) = tr\{X(X^T X + \lambda W^{-})^{-1} X^T\}$$

(more details can be found in Tibshirani, 1996).

We assume that RSS(t) is the residual sum of squares for the constrained fit with constraint t, we construct the generalized cross-validation style statistic

$$GCV(t) = \frac{1}{N} \frac{RSS(t)}{\{1 - p(t)/N\}^2}.$$

Now, we describe a third method based on Stein's unbiased estimate of risk. Suppose that z is a multivariate normal random vector with mean vector μ and variance the identity matrix. Let

$\hat{\mu}$ be an estimator of μ, and write $\hat{\mu} = z + g(z)$ where g is an almost differential function from R^p to R^p. Then, Stein (1981) proved that

$$E_\mu \| \hat{\mu} - \mu \|^2 = p + E_\mu \left(\| g(z) \|^2 + 2 \sum_1^p dg_i / dz_i \right)$$

We may apply the above result to the lasso estimator (6). Denote the estimated standard error of $\hat{\beta}_j^0$ by $\hat{\tau} = \hat{\sigma} / \sqrt{N}$, where $\hat{\sigma}^2 = \sum (y_i - \hat{y}_i)^2 / (N - p)$. Then, as in Tibshirani (1996), the $\hat{\beta}_j{}^0 / \hat{\tau}$ are approximately independent standard normal variates, and from the above equation, we may derive the formula

$$R\{\hat{\beta}(\gamma)\} \approx \hat{\tau}^2 \left\{ p - 2\#(j; | \hat{\beta}_j{}^0 / \hat{\tau} | < \gamma) + \sum_{j=1}^p \max(| \hat{\beta}_j{}^0 / \hat{\tau} |, \gamma)^2 \right\}$$

as an approximately unbiased estimate of the risk or mean-square error $E\{\hat{\beta}(\gamma) - \beta\}^2$, where $\hat{\beta}_j(\gamma) = sign(\hat{\beta}_j{}^0)(| \hat{\beta}_j{}^0 / \hat{\tau} | - \gamma)^+$. Donoho and Johnstone (1994) gave a similar formula in the function estimation setting. Hence an estimate of γ can be obtained as the minimizer of

$$R\{\hat{\beta}(\gamma)\} : \gamma = \arg\min_{\gamma \geq 0} [R\{\hat{\beta}(\gamma)\}].$$

We can obtain an estimate of the lasso parameter t from the above equation and it is given by

$$\hat{t} = \sum (| \hat{\beta}_j^0 | - \hat{\gamma})^+.$$

We may use the above derivation in non-orthogonal setting, although the derivation of $\hat{t}$ assumes an orthogonal design.

3.8 Bayesian Interpretation for ridge regression and the lasso

One can view ridge regression and the lasso through Bayesian point of view. A Bayesian viewpoint for regression assumes that the coefficient vector β has some prior distribution, say

$p(\beta)$, where $\beta = (\beta_0, \beta_1, ...\beta_p)^T$. The likelihood of the data can be written as $f(Y \mid X, \beta)$, where $X = (X_1, ...X_p)$. Multiplying the prior distribution by the likelihood gives us the posterior distribution, which takes the form

$$p(\beta \mid X, Y) \propto f(Y \mid X, \beta) p(\beta \mid X) = f(Y \mid X, \beta) p(\beta),$$

where the proportionality above follows from Bayes' theorem, and the equality above follows from the assumption that X is fixed (see James et al., 2013). We assume that usual linear model,

$$Y = \beta_0 + X_1\beta_1 + ... + X_p\beta_p + \in,$$

and suppose that the errors are independent and drawn from a normal distribution. Furthermore, assume that $p(\beta) = \prod_{j=1}^{p} g(\beta_j)$, for some density function g. It turns out that ridge regression and the lasso follow naturally from two special cases of g.

If g is a Gaussian distribution with mean zero and standard deviation a function of λ, then it follows that the posterior mode for β is the ridge regression solution. In fact, the ridge regression solution is also the posterior mean.

If g is a double-exponential distribution with mean zero and scale parameter a function of λ, then it follows that the posterior mode for β is the lasso solution. However, the lasso solution is not the posterior mean, and in fact, the posterior mean does not yield a sparse coefficient vector. The Gaussian and double-exponential priors are displayed in the figure below reproduced from James et al (2013). The left diagram displays ridge regression as the posterior mode for β under a Gaussian prior. The right one diagram displays lasso as the posterior mode for β under a double-exponential prior. Therefore, from a Bayesian point of view, ridge regression and the lasso follow directly from assuming the usual linear model with normal errors, together with a simple prior

distribution for β (see James et al., 2013). Figure 3 shows that lasso uses double-exponential prior and ridge regression uses Gaussian prior.

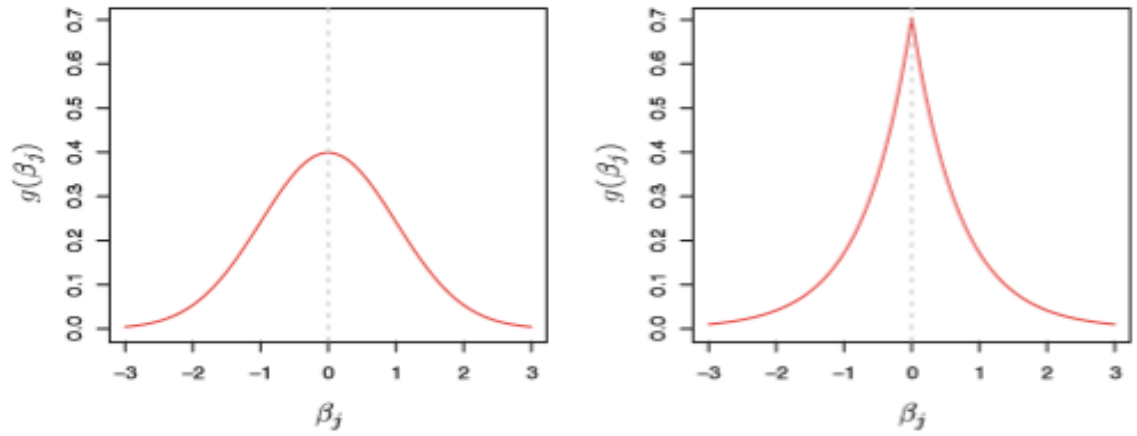


Figure 3**.** Left : Ridge regression is the posterior mode β under a Gaussian prior. Right: The lasso is the posterior mode β under a double-exponential prior.

It shows that ridge regression and the lasso follow directly from assuming the usual linear model with normal errors, together with a simple prior distribution. Lasso prior is steeply peaked at zero, while the Gaussian is flatter and flatter at zero. Therefore, the lasso expects a priori that many of the coefficients are exactly zero, while ridge assumes the coefficients are randomly distributed about zero (the figure is taken from James et al., 2013).

DATA ANALYSIS

In this chapter, we apply ridge regression and lasso procedures to analyze a data set, and compare the results. We will perform calculations using R software for statistical computing. The data set is known as Ischemic Heart Disease data which can found as APPENCO9 in the data set of Kutner, Nachtsheim and Neter (2004). In Table 1, the data is collected by a health insurance company on 788 of its subscribers who made claims resulting from ischemic heart disease. Data were obtained on total costs of services provided for these 788 subscribers and the nature of the various services for the period of January,1998 through December,1999.

Table 1. Ischemic heart disease data for the period of January,1998 through December,1999. (Partial)

y	x_1	x_2	x_3	x_4	x_5	x_6	x_7
179.1	63	2	1	4	0	3	300
319.0	59	2	0	6	0	0	120
9310.7	62	17	0	2	0	5	353
280.9	60	9	0	7	0	2	332
18727.1	55	5	2	7	0	0	18
453.4	66	1	0	3	0	4	296
323.1	64	2	0	3	0	1	247
..	..	..	..	..	..	..	..
586.0	56	4	4	6	0	3	336

In Table 1, y represents the total cost (in dollars) of claims by subscriber, x_1 represents age (in years) of subscriber, x_2 is the total number of interventions or procedures carried out, x_3 is the number of tracked drugs prescribed, x_4 is the number of emergency room visits, x_5 is the number of other complications arose during heart disease treatment, x_6 is the number of other diseases that the subscriber had during period, x_7 is the number of days of duration of treatment condition.

In the regression model we introduce two way interactions, three way interactions and one four way interactions. The model consists of the variables x_1, x_2, x_3, x_4, x_5, x_6, x_7 and the interaction variables are x_{10}, x_{11},.., x_{34}.

We will perform ridge regression and lasso in order to predict the total cost of claims by subscriber on the Ischemic heart disease data. To perform ridge regression and the lasso, we will use glmnet package. This package has a function named glmnet ( ), which can be used to fit ridge regression models and lasso models. This function has an alpha argument that determines what type of model is fit. If alpha = 0, then a ridge regression model is fit, and if alpha = 1, then a lasso model is fit. We first fit a ridge regression model.

The glmnet ( ) function performs ridge regression over a range of λ values and we chose alpha = 0. However, we chose λ ranging from 10^{10} to 10^{-2} to implement this function. There is a vector ridge regression coefficients associated to each value of λ. In this case, it is a 33×100 matrix containing 33 rows and 100 columns. For a large value of λ, we expect small values of coefficient estimates or a small value of λ, we expect largest values of coefficients estimates. In Table 2, estimated coefficients are evaluated using $\lambda = 11498$ and in Table 3, we calculated coefficient estimates using $\lambda = 705$.

Table 2. Coefficient estimates are calculated when $\lambda = 11498$.

Coefficient Estimates	**Value**
x_1	-2.311515e+01
x_2	1.633078e+02
x_3	3.714766e+01
x_4	1.136144e+02
x_5	1.623361e+02
x_6	-4.848064e+00
x_7	-2.727258e-01
x_{10}	9.173380e+00
x_{11}	1.840240e+01
x_{12}	5.915916e+01
x_{13}	8.472665e+00
x_{14}	5.257641e-01
x_{15}	9.104796e-01
x_{16}	2.187727e-02
x_{17}	2.640743e-01
x_{18}	9.794958e+00
x_{19}	3.764478e+00
x_{20}	3.263379e-01
x_{21}	2.154675e+01
x_{22}	9.925574e-01
x_{23}	2.171972e+01
x_{24}	9.963126e-01
x_{25}	1.799363e-01
x_{26}	-5.279760e-02
x_{27}	-9.481987e-01
x_{28}	-2.198440e+01
x_{29}	-2.206871e+00
x_{30}	3.102326e-02
x_{31}	-5.110960e+00
x_{32}	5.539768e-02
x_{33}	-3.464758e-02
x_{34}	9.621908e-03

Table 3. Coefficient estimates are calculated when $\lambda = 705$.

Coefficient Estimates	**Value**
x_1	-3.972794e+01
x_2	2.822483e+02
x_3	1.293302e+02
x_4	5.516112e+02
x_5	1.900810e+02
x_6	2.242231e+01
x_7	-2.394098e+00
x_{10}	-2.910851e+01
x_{11}	5.025507e+01
x_{12}	1.57787e+01
x_{13}	1.338450e+01
x_{14}	7.254279e-01
x_{15}	-1.414356e+00
x_{16}	4.66191e-03
x_{17}	3.802481e-01
x_{18}	5.020874e+01
x_{19}	3.551409e+00
x_{20}	-1.222256e-01
x_{21}	-9.795041e+01
x_{22}	1.050626e+00
x_{23}	-9.190754e+01
x_{24}	1.234670e+00
x_{25}	5.551288e-01
x_{26}	-1.677179e-01
x_{27}	-4.592036e+01
x_{28}	-7.480861e+01
x_{29}	-9.072001e+00
x_{30}	4.719415e-01
x_{31}	-4.187675e+00
x_{32}	-6.748773e-02
x_{33}	-3.464758e-02
x_{34}	9.621908e-03

In order to estimate the test error of ridge regression and the lasso, we split the samples into a training set and a test set. Then we fit a ridge regression on the training set, and using $\lambda = 4$, we get the value of test MSE = 34987825.

If we simply predict a model with just an intercept and predict each test observation using the mean of the training observations, then we get the test MSE = 42082752.

We generalize the value of tuning parameter λ, instead of choosing $\lambda = 4$, using the method of cross-validation. This can be done using the built-in cross validation function, cv.glmnet ( ) of R. This function generally performs ten-fold cross-validation. The cross-validation result can be visualized with plot command. Figure 4 gives the cross-validation result for ridge regression.

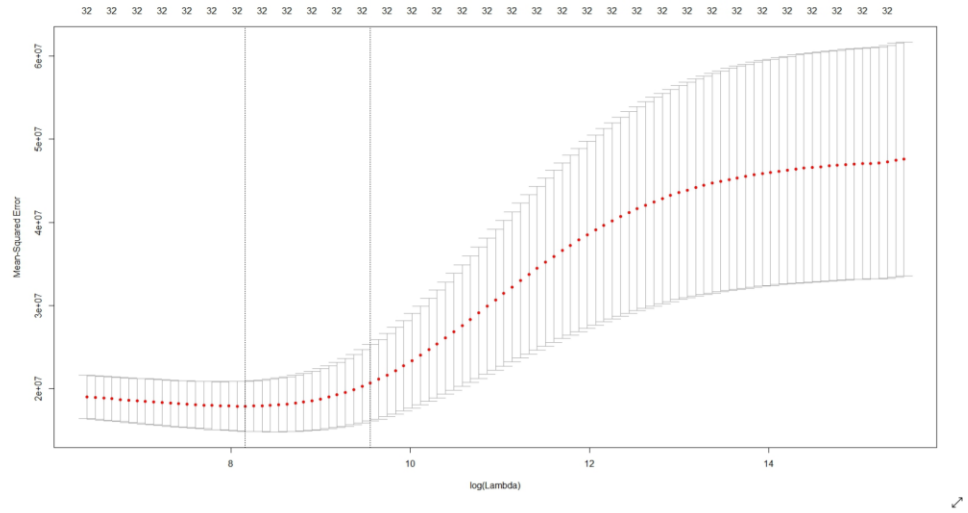


Figure 4. Cross-validated estimate of the mean squared prediction error for ridge regression.

It shows that cross-validated estimate of the mean squared prediction error for ridge regression is a function of $\log\lambda$. The upper part of the plot shows that the number of non-zero coefficients in the regression model for a given value of λ. The dashed line shows the location

of the function minimum and the 'one-standard-error' location. The first vertical dashed line shows the location of the minimum of MSE and other one shows the point selected by the 'one-standard-error' rule. We see that coefficients estimates get more shrunk towards zero. In this case, we get the smallest value of $\lambda = 3491.48$ and MSE = 29847591. This value of MSE is smaller than that test MSE which we got using Now we fit our ridge regression model on the full data, using the value of λ chosen by cross-validation, and we examine the coefficient estimates.

In Table 4, we observe that none of the coefficients estimates are zero. There are some coefficients close to zero because ridge regression does not perform variable selection.

Now we apply lasso to yield a more accurate and interpretable model than the ridge regression. We use the same function glmnet ( ) to fit a lasso model, and alpha parameter in this case takes the value of 1. Figure 5 depicts the plot of lasso model.

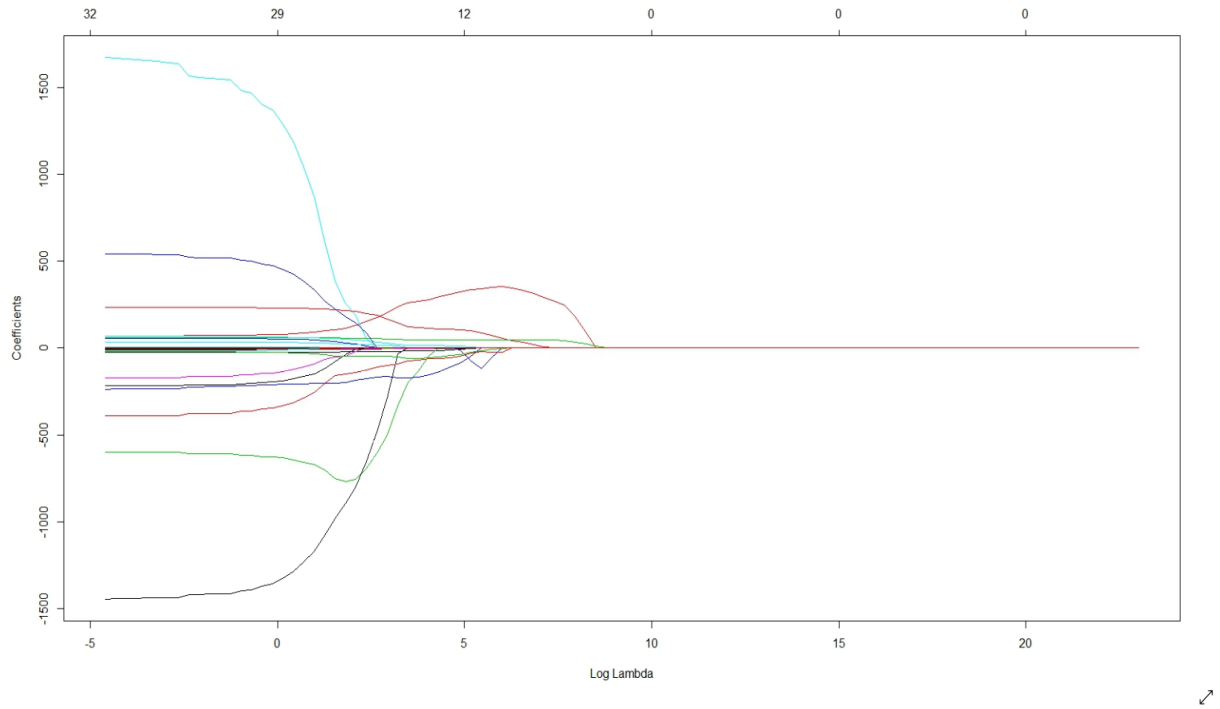


Figure 5. Coefficients estimates for lasso.

Table 4. Coefficient estimates are calculated using the value of $\lambda = 3491.48$.

Coefficient Estimates	Value
x_1	-3.601058e+01
x_2	2.379662e+02
x_3	2.001988e+01
x_4	1.181190e+02
x_5	194451e+02
x_6	-7.384343e+00
x_7	-1.524470e+00
x_{10}	-1.851799e+00
x_{11}	2.928069e+01
x_{12}	5.759298e+01
x_{13}	1.108550e+01
x_{14}	7.002426e-01
x_{15}	1.954041e-01
x_{16}	2.143584e-02
x_{17}	3.084597e-01
x_{18}	4.999844e-01
x_{19}	2.550092e+00
x_{20}	2.534284e-01
x_{21}	-1.411609e+01
x_{22}	8.229143e-01
x_{23}	-1.374466e+01
x_{24}	8.415797e-01
x_{25}	1.884453e-01
x_{26}	-1.151413e-01
x_{27}	-1.0875e+01
x_{28}	-3.975029e+01
x_{29}	-5.851659e+00
x_{30}	1.111922e-01
x_{31}	-3.841174e+00
x_{32}	-6.520050e-02
x_{33}	-1.795449e-02
x_{34}	1.105041e-04

It shows coefficients estimates for lasso for the data plotted versus $\log\lambda$. The upper part of the plot shows the number of non-zero coefficients in the regression model for a given $\log\lambda$. We see that the lasso regression tends to shrink the regression coefficients to zero as λ increases. When $\log\lambda = 5$ there are 12 non-zero coefficients and when $\log\lambda = 10$ we get the model where all the coefficients are zero. That is, the lasso performs variable selection when λ is large enough. We now perform cross-validation to select the λ parameter. The cross-validation result can be visualized with plot command, and is given below. Figure 6 provides cross-validation result for lasso model.

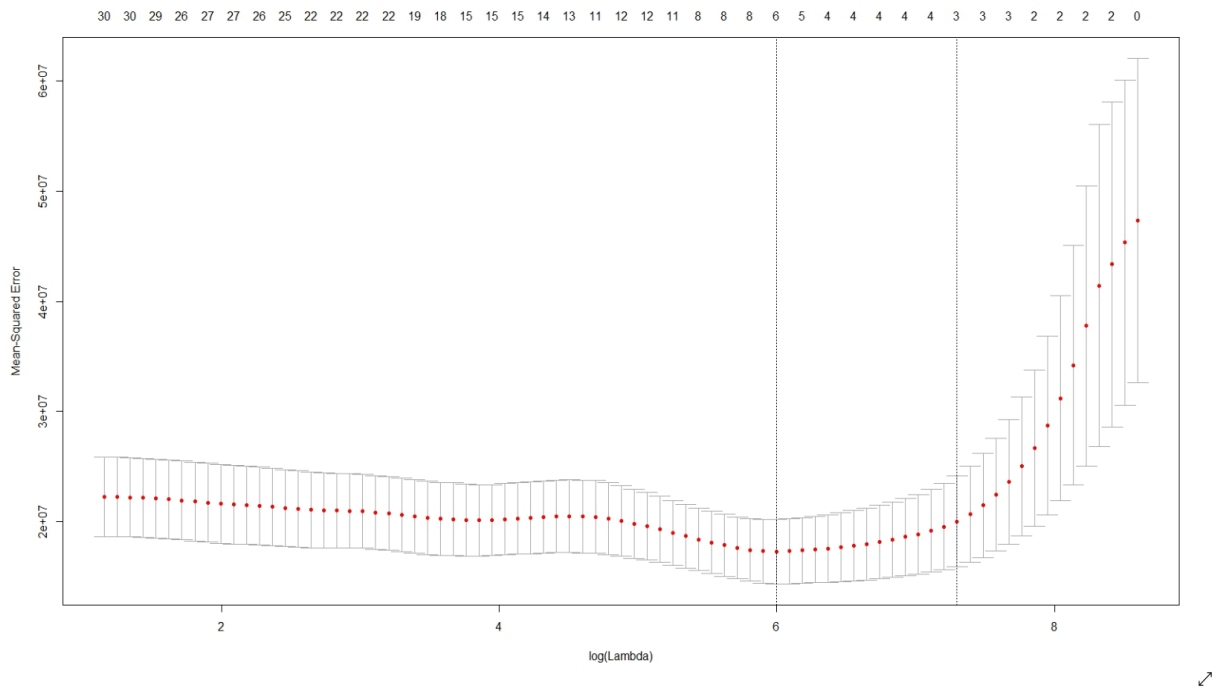


Figure 6. Cross-validated estimate of the mean squared prediction error for lasso.

It shows that cross-validated estimate of the mean squared prediction error for lasso as a function of $\log\lambda$. The upper part of the plot shows that the number of non-zero coefficients in the regression model for a given value of $\log\lambda$. The dashed line shows the location of the function minimum and the 'one-standard-error' location. The first vertical dashed line shows the location of the minimum of MSE and other one shows the point selected by the 'one-standard-error' rule. In this case, we get the smallest value of λ is 483.5302 and MSE = 28169646. This value MSE is similar to the test MSE of ridge regression with λ chosen by cross-validation.

In Table 5, we observe that 25 of the 32 coefficients estimates are exactly zero by the lasso method because it also performs variable selection.

We have seen that the lasso has a major advantage over the ridge regression, in that it produces simpler and more interpretable models that involve only a subset of the predictors. As noted in James et al (2013), ridge regression performs better when the response is a function of many predictors, all with coefficients of roughly the same size. On the other hand, lasso performs better in a setting where a relatively small number of predictors have substantial coefficients, and the remaining predictors have coefficients that are very small or that equal to zero. In the case of ridge regression, we see that none of the coefficients are zero because it does not perform variable selection. On the other hand, as in James et al (2013), the lasso performs variable selection, and hence results in models that are easier to interpret. Lasso solution can yield a reduction in variance at the expense of a small increase in bias, and hence can generate more accurate predictions.

Table 5. Coefficient estimates are calculated using the value of $\lambda = 483.5302$.

Coefficient Estimates	**Value**
x_1	-3.26486
x_2	347.03378
x_3	0
x_4	0
x_5	0
x_6	0
x_7	0
x_{10}	-12.94
x_{11}	46.4037
x_{12}	0
x_{13}	9.5689105
x_{14}	0
x_{15}	0.4030609
x_{16}	-12.94
x_{17}	46.4037
x_{18}	0
x_{19}	9.5689105
x_{20}	0
x_{21}	0
x_{22}	0
x_{23}	0
x_{24}	0
x_{25}	0
x_{26}	0
x_{27}	0
x_{28}	0
x_{29}	0
x_{30}	0
x_{31}	0
x_{32}	0
x_{33}	0
x_{34}	0

REFERENCES

Breiman, L. (1993). Better subset selection using the non-negative garotte. *Technometrics*, Department of Statistics, University of California, Berkeley.

Efron, B. & Tibshirani, R. (1993). *An Introduction to the bootstrap*, Chapman and Hall.

Gilmour, Steven G. (1996*).* The interpretation of Mallow's C_p-statistic. *Journal of the Royal Statistical Society Series D.* 45(1): 49-56.

James, G., Witten, D., Hastie, T., Tibshirani, R., (2013). *An introduction to Statistical Learning*. New York: Springer.

Mallows, C.L. (1973). Some comments on C_p. *Technometrics*, 15(4): 661-675.

Montgomery, D.C. & Peck, E.A. (1982*). Introduction to linear regression analysis*. New York: John Wiley and sons.

Stein, C. (1981). Estimation of the mean of a multiplicative normal distribution. *Ann. Statist*. 9, 1135-1151.

Tibshirani, R (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B*. 58(1): 267-288.