# *Regression Shrinkage and Selection via the Lasso*

*Report of the project on the Lasso Regression*

*Team Members*

*Ankush Agrekar (A20382887)* *Divya Vasireddy (A20370052)*

*Zhiwei Zhang (A20379914)* *Jingyuan He (A20386460)*

## 1. Introduction

This report is based on the project work done on the basis of the article 'Regression Shrinkage and Selection via the Lasso' by Robert Tibshirani(1996). The author proposes a new method for estimation of the beta coefficients in Linear models. The research mainly focus on measures for improving the prediction accuracy in Linear models. The motivation for the Lasso came from interpreting proposal of non-negative Garotte minimizers. There are other significant works in Lasso which are done by 'Prof. Trevor Hastie' at Stanford. The research done by this author plays a significant role for addressing the convex problem. He used Karush-Kuhn-Tucker approach to minimize the residual sum of squares under the constraint on the sum of the absolute values of regression coefficient estimates.

The Ordinary Least Squares estimates are obtained by minimizing the residual squared error. There are reasons why the results are not satisfactory using the OLS estimates. First, is Prediction Accuracy: the OLS estimates often has low bias but large variance; the prediction accuracy can sometimes be improved by shrinking some coefficients. By doing so we sacrifice a little bit bias to reduce the variance overall. The second reason is interpretation: with large number of predictors we would often like to determine smaller subset that exhibits the strongest effects.

The two standard techniques for improving OLS estimates are Subset selection and ridge Regression. But both have drawbacks. The subset selection is iterative process, it will either drop or add the parameters. This is a discrete process. Ridge Regression is a continuous process that shrinks coefficients but it does not set the coefficients to '0' and hence does not give smaller and interpretable model.

So, the main idea of is 'Lasso' is to minimize the residual sum of squares subject to the sum of absolute values of the coefficients being less than a constant (constraint). Because of this nature of constraint, it tends to produce some coefficients exactly to Zero thus gives the interpretable models. The simulation studies by the author proved that lasso inherits some of the favorable properties of both subset selection and ridge regression. It produces interpretable models like subset selection and exhibits stability of ridge regression.

## 2. The Lasso : Definition

Let $(x_i, y_i), i = 1, 2, \dots, N,$ be the dataset, where $x_i = (x_{i1}, \dots, x_{ip})^T$ are the 'p' predictor variables and $y_i$ are the 'N' responses. We assume similar assumptions as for the linear regression set up. Hence, all the $y_i$s are conditionally independent given the $x_{ij}$s or all the observations are independent. We require the $x_{ij}$s to be standardized so that the mean and variance of each predictor is 0 and 1 respectively.

We use the representation $(\alpha, \beta)$ to represent the Lasso estimates where, $\beta = (\widehat{\beta_1}, \dots, \widehat{\beta_p})^T$

Therefore,

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

Here, $t \geq 0$ is the constraint parameter. Now, for all t, the solution for $\alpha$ is $\alpha = \bar{y}$ , since the mean of all the independent variables is 0. We can assume without loss of generality that $\bar{y}$ = 0 and hence omit $\alpha$. Here, Equation (1) is a quadratic programming problem with linear inequality constraints. The algorithm to solve this quadratic complex problem is given below in Section 5. The parameter $t \geq 0$ controls the amount of shrinkage that is applied to the ordinary least square estimates. Let $\widehat{\beta_j^o}$ be the full least square estimates and let $t_o = \sum_j \left|\widehat{\beta_j^o}\right|$. Values of $t < t_o$ will cause shrinkage towards 0, and some coefficients may be exactly equal to Zero. The motivation for the lasso came from the proposal of Breiman (1993). Breiman's non-negative garotte minimizes

$$\sum_{i=1}^{N}\left(y_i - \alpha - \sum_j c_j \widehat{\beta_j^o} x_{ij}\right)^2 \qquad \text{subject to } c_j \geq 0 \;\; and \; \sum c_j \leq t$$

The Garotte solution shrinks the ols coefficients by non-negative factors whose sum is constrained. The drawback of Garotte estimates is that is dependent on both the sign and the magnitude of the linear estimates. As a result, if the data is highly correlated or in case of overfitting, the ols estimates perform poorly and hence will the Garotte estimates.

## 3. Approach

For implementing Lasso, we considered the Matrix equation and converted the Lasso quadratic problem to match the structure of the complex quadratic problem required by the quadprog package of the form $\min(-d^T b + \frac{1}{2}\, b^T D b)$ with the constraints $A^T b \,\geq\, b_0$. For this we have referred to package 'QuadProg' provided by R. 'solve.QP' function R helps to solve any quadratic problem.

$$(\alpha, \beta) = \arg\min\left\{\sum_{i=1}^{N}\left(y_i - \alpha - \sum_j \beta_j\, x_{ij}\right)^2\right\}$$

can be written in the Matrix form as

$$(\alpha, \beta) = \arg\min(Y - X\beta)^T(Y - X\beta)$$

$$= \arg\min[Y^T Y - Y^T X\beta - \beta^T X^T Y + \beta^T (X^T X)\, \beta]$$

Here, $\beta^T X^T Y$ is a scalar and hence, can be written as $(\beta^T X^T Y\,)^T = Y^T X\beta$. Therefore,

$$(\alpha, \beta) = \arg\min[Y^T Y - 2Y^T X\beta + \beta^T (X^T X)\, \beta]$$

In the above equation, $Y^T Y$ is a scalar/constant, so the problem is equivalent to writing as

$$(\alpha, \beta) = \arg\min[-\, 2Y^T X\beta + \tfrac{1}{2}\beta^T (2X^T X)\, \beta]$$

Comparing this with the equation for the solve.QP function, we get –

$$d = 2X^T Y$$

$$D = 2X^T X$$

$$A = -G_e^T$$

$$b = \beta$$

Where, $G_e$ is the matrix of p-tuples of the form $(\pm 1, \pm 1, \ldots, \pm 1)$ corresponding to all possible signs for the 'p' beta estimates.

The Lasso solution can be considered to be unique when the following Karush-Kuhn-Tucker conditions are met

For any y, X and λ≥0, the lasso problem has the following properties:

(i) There is either a unique solution or infinite number of solutions
(ii) Every lasso solution $\hat{\beta}$ gives the same fitted value $X\widehat{\beta}$
(iii) If λ>0, then every lasso solution $\hat{\beta}$ has the same L1 norm, $\|\widehat{\beta}\|_1$

## 4. PREDICTION ERROR AND ESTIMATION OF t

The computation of solution for the Lasso equation is a quadratic programming problem with linear inequality constraints. After fitting the linear model, we found the Linear estimates of least squares. Since this is non-linear and non-differentiable function of the response values even for the fixed value t, it is difficult to obtain the accurate estimate of its standard error.

There are three methods for estimation of the lasso problem t:

### *4.1 Cross-validation*

The prediction error for the Lasso procedure can be calculated by k-fold cross validation. The lasso is indexed in terms of the normalized parameter $s = t/\sum_j |\widehat{\beta_j^o}|$, and the prediction error is estimated over a grid of values of s from 0 to 1 inclusive. The value s yielding the lowest estimated PE is selected.

### *4.2 Generalized Cross-Validation*

The constraint $\sum_j |\beta_j| \leq t$ can be written as $\sum_j \beta_j^2 / |\beta_j| \leq t$. This constraint is equivalent to adding a Lagrangian penalty X E fl/l,ij to the residual sum of squares, with X depending on t.

The number of effective parameters for any given 't', can be found using

$$p(t) = \mathrm{tr}\left\{\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda \mathbf{W}^{-})^{-1}\mathbf{X}^{\mathrm{T}}\right\}.$$

Where $W = diag\left(\left|\beta_j\right|\right)$ and $W^{-}$ denotes generalized inverse, then the Generalised Cross Validation error can be calculated as

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

The two methods are applicable in the 'X-random' case, where it is assumed that the observations (X, Y) are drawn from the unknown distribution and the third method applies to the X-fixed case. However, in real problems there is often no clear distinction between the two scenarios and we might choose the most convenient method.

## 5. Algorithm for finding the Lasso Solution:

We fix $t \geq 0$. We can express the least squares problem with $2^p$ inequality constraints, corresponding to the $2^p$ different possible signs for the $\beta_j$s.

Lawson and Hansen (1974) provided the ingredients for a procedure which solves the linear least squares problem subject to a general linear inequality constraint $G\beta \leq h$. Here $G$ is an m x p matrix, corresponding to m linear inequality constraints on the p-vector $\beta$. For the lasso problem however, m=$2^p$ may be very large so that direct application of Lawson and Hansen procedure is not practical. However, this problem can be solved by introducing the inequality constraints one by one, seeking a feasible solution satisfying the Kuhn-Tucker conditions. The procedure is outlined below.

Let $g(\beta) = \sum_{i=1}^{N} \left(y_i - \alpha - \sum_j \beta_j x_{ij}\right)^2$ , and let $\delta_i,\ i = 1, 2,\ \ldots, 2^p$ be the p-tuples of the form $(\pm 1, \pm 1,\ \ldots,\ \pm 1)$. Then the condition $\sum_j |\beta_j| \leq t$ is equivalent to $\delta_i^T \beta \leq t$ for all $i$. For a given $\beta$, let $E = \{i : \delta_i^T \beta = t\}$ and $S = \{i :\ \delta_i^T \beta < t\}$. The set $E$ is the equality set, corresponding to those constraints which are exactly met, whereas $S$ is the slack set, corresponding to those constraints for which equality does not hold. Denote by $G_E$ the matrix whose rows are $\delta_i$ for$i \in E$. Let **1** be a vector of 1s of length equal to number of rows of $G_E$.

The following algorithm starts with $E = \{i_0\}$ where $\delta_{io} = sign(\hat{\beta})$, $\hat{\beta}$ being the overall least squares estimate. It solves the least squares problem subject to $\delta_{io}^T \beta \leq t$ and then checks whether $\sum_j |\beta_j| \leq t$. If so, the computation is complete, if not, the violated constraint is added to $E$ and the process is continued until $\sum_j |\beta_j| \leq t$.

Here is the outline of the algorithm:

*Algorithm 1: The Lasso*

a) Start with $E = \{i_0\}$ where $\delta_{io} = sign(\hat{\beta})$, $\hat{\beta}$ being the overall least squares estimate.

b) Find $\hat{\beta}$ to minimize $g(\beta)$ subject to $G_{\mathrm{E}}^T \beta \leq t\mathbf{1}$

c) While $\{\sum_j |\beta_j| \leq t\}$,

d) Add $i$ to set $E$ where $\delta_i = sign(\hat{\beta})$. Find $\hat{\beta}$ to minimize the $g(\beta)$ subject to $G_{\mathrm{E}}^T \beta \leq t\mathbf{1}$

This procedure must always converge in a finite number of steps since one element is added to the set E at each step, and there is a total of $2^p$ elements. The final iterate is a solution to the original problem since the Kuhn-Tucker conditions are satisfied for the sets E and S at convergence.

## 6. Implementation and Results

Here, apply the lasso with methods and algorithms as described in the previous sections. The implementation was carried out in R.

### *6.1 Data*

For solving the Lasso problem, we have used the Prostate Data set in lasso2 package in R. The prostate cancer dataset comes from a study by Stamey that examined the correlation between the level of prostate specific antigen and number of clinical measures. The factors were log (cancer volume - lcavol), log(prostate weight – lweight), log(benign prostatic hyperplasia amount – lbph), seminal vesicle invasion(svi), log(capsular penetration – lcp), gleason score and percentage of Gleason scores(pgg45). We fit a linear model to log(prostate specific antigen) (lpsa) after first standardizing the predictors. We can find the lasso estimates as a function of standardized bound $s=t/|\hat{\beta}j0|$. Notice that the absolute value of each coefficient tends to 0 as s goes to 0. The paper states using s=0.44 as the optimal value obtained from the Generalised Cross Validation and the below results are based on the experiments performed using the same s value. The t constraint is calculated accordingly.

### *6.2 Results*

Figure 1 is the plot for lasso shrinkage of coefficients in the prostate cancer example. Each curve represents the coefficients and as a function of labeled parameter ‘s’.

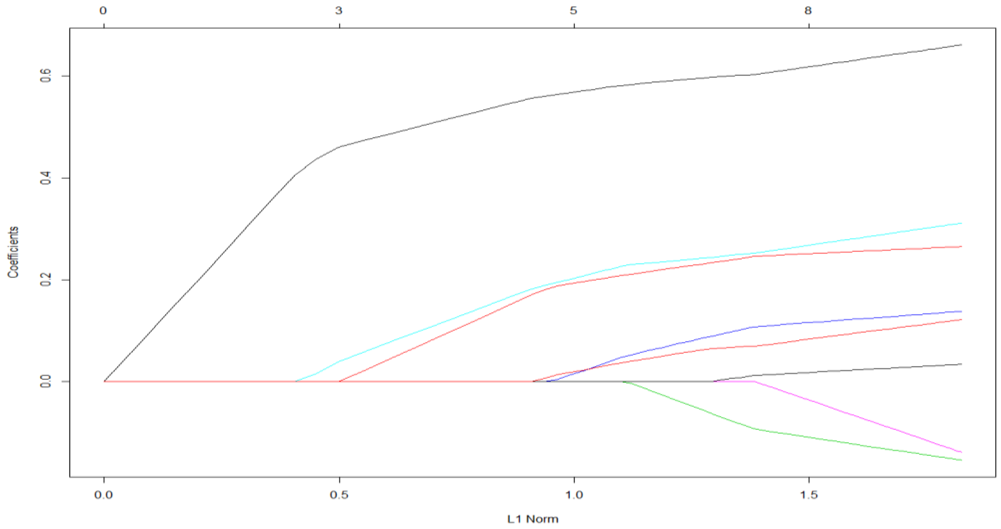


Figure 1. the lasso arc length graph based on s = 0.44.

In this example, the curves decrease in a monotone fashion to 0, but this does not always happen in general.

Later we have compared the Lasso estimate results with the Subset selection and full least squares. We have used the inbuilt functions of R for the full least squares and Subset selection.

Below table shows the results for the full least squares, Subset Selection and Lasso estimates

| **Predictor** | **Least Square Results** | | | **Subset selection Results** | | | **Lasso Results** | | |
|---|---|---|---|---|---|---|---|---|---|
| | **Coeff** | **Std Error** | **Z-score** | **Coeff** | **Std Error** | **Z-score** | **Coeff** | **Std Error** | **Z-score** |
| intercept | 2.48 | 0.07102 | 34.8 | 2.48 | 0.07 | 34.9 | 2.4784 | 0.0798 | 31.058 |
| lcavol | 0.67 | 0.10352 | 6.4 | 0.64 | 0.09 | 7.2 | 0.5588 | 0.11496 | 4.86 |
| lweight | 0.26 | 0.08 | 3.09 | 0.25 | 0.08 | 3.4 | 0.097 | 0.09367 | 1.036 |
| age | -0.15 | 0.08 | -1.917 | 0 | 0 | 0 | 0 | 0.09228 | 0 |
| lbph | 0.14 | 0.08 | 1.67 | 0 | 0 | 0 | 0 | 0.09408 | 0 |
| svi | 0.31 | 0.1 | 3.15 | 0.29 | 0.09 | 3.1 | 0.1556 | 0.11221 | 1.387 |
| lcp | -0.14 | 0.12 | -1.18 | 0 | 0 | 0 | 0 | 0.14118 | 0 |
| gleason | 0.035 | 0.11 | 0.317 | 0 | 0 | 0 | 0 | 0.12615 | 0 |
| pgg45 | 0.125 | 0.12 | 1.021 | 0 | 0 | 0 | 0 | 0.13834 | 0 |

Figure. 2 shows the plots for coefficients for Subset Selection, Lasso and Ridge. For Subset Selection and Ridge we have used the existing functions in R and for Lasso we used the function which we have implemented.

The lasso gave non-zero coefficients to lcavol, lweight and svi; subset selection chose the same three predictors. Notice that the Coefficients and Z-scores for the selected predictors from the subset selection tend to be larger than the full model values: this is common with positively correlated predictors. However, the lasso shows the opposite effect, as it shrinks the coefficients and Z-scores from their full model values.

The standard errors have been calculated by fixing the ŝ at its optimal value 0.44 for the original dataset. The standard error is calculated by the formula

$se(\beta_j) = \hat{\sigma}\sqrt{v_j}$ where $v_j$ is the jth diagonal element ofn $(X^T X)^{-1}$ and $\hat{\sigma}$ is the RMSE.

We can estimate the standard error by Bootstrap sampling. In this method, we will re-estimate the t for each sample. The Figure 3 shows the 200 bootstrap replications of the Lasso estimates, with ŝ fixed at the estimated value 0.44. The predictors whose estimated coefficients is 0 exhibit skewed bootstrap distributions. The central 90% percentile intervals (fifth and 95th percentiles of the bootstrap distributions) all contained the value 0.

The results for lasso are different from the results in the article but our results are matching with 'l1ce' function in 'lasso2' package which is created by Tibshirani.

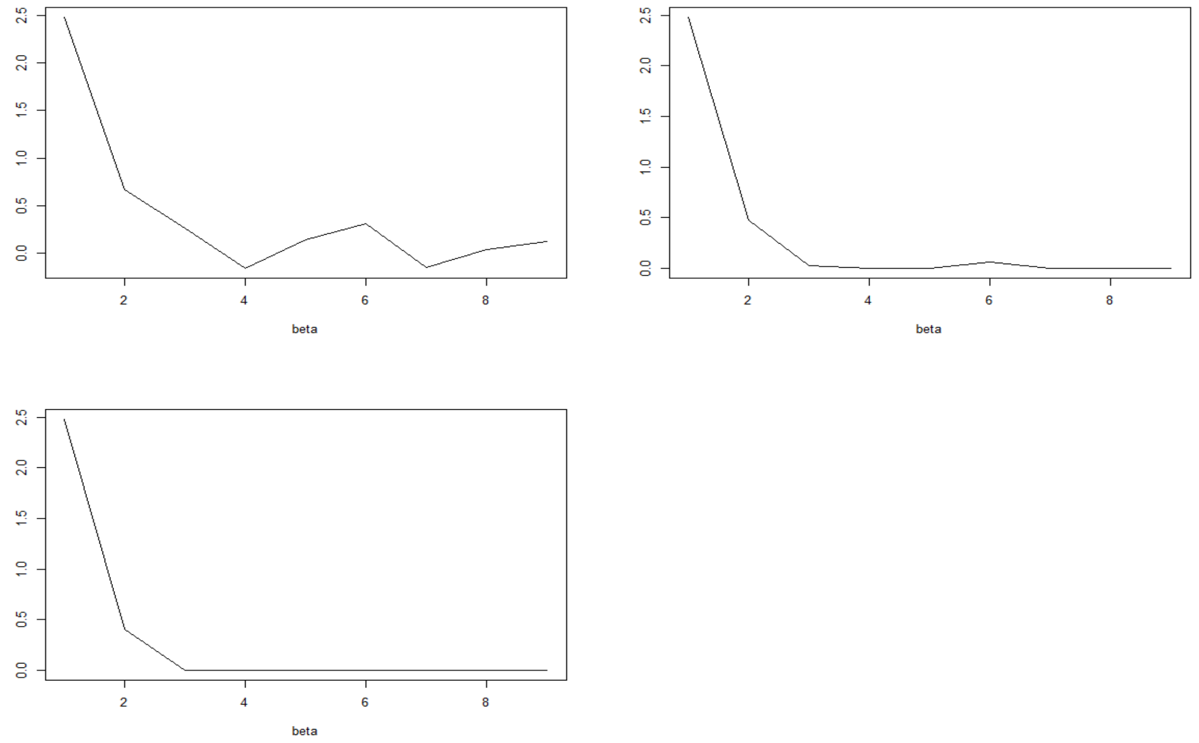


Figure. 2 coefficients of subset selection, lasso and ridge from top left to bottom.

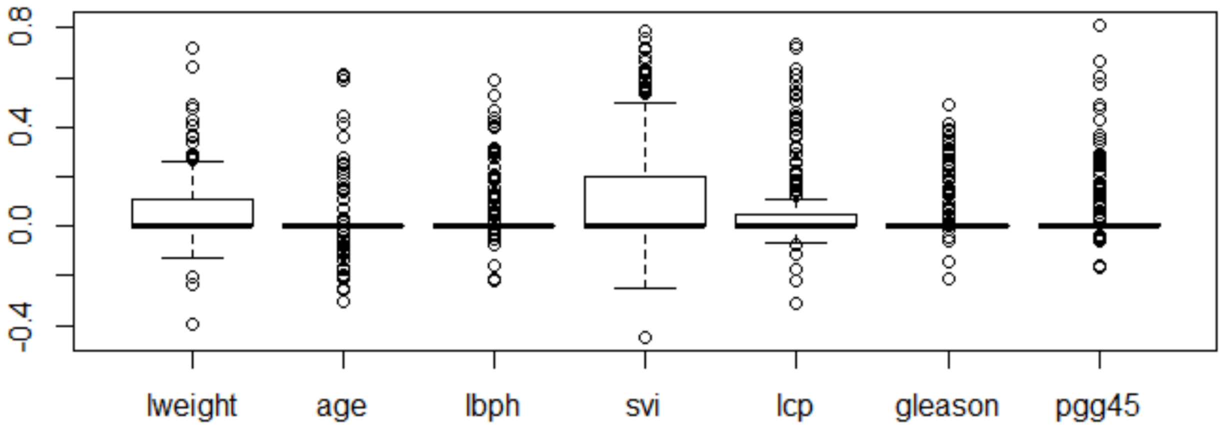
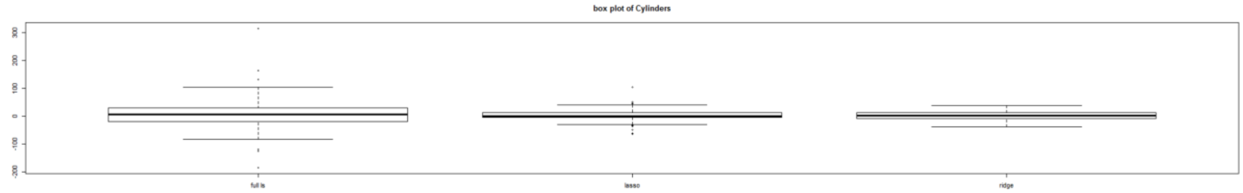
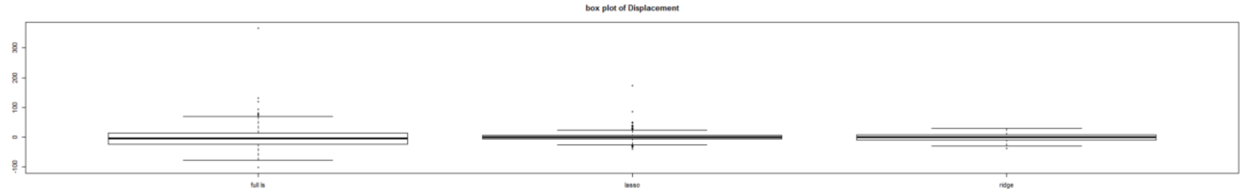


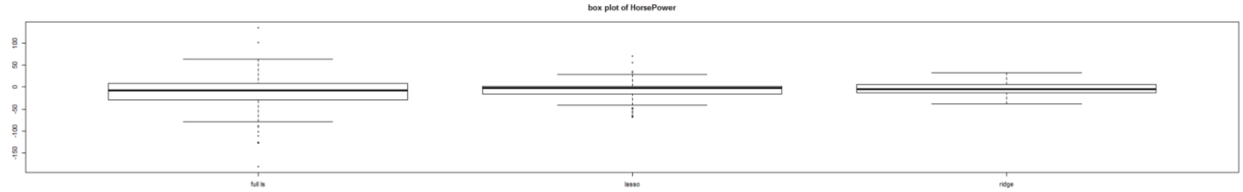
Figure. 3 boxplots for the coefficients obtained after 200 subsets of bootstrapping

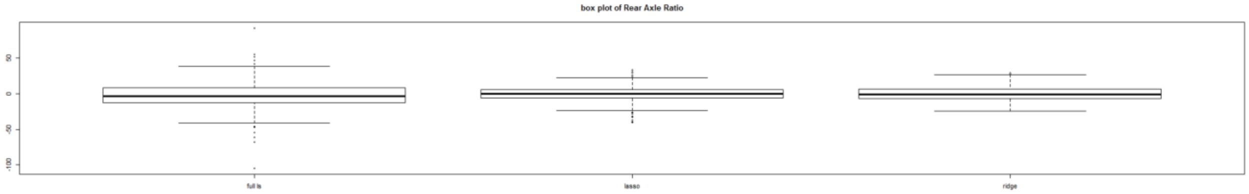
### 6.3 Simulation

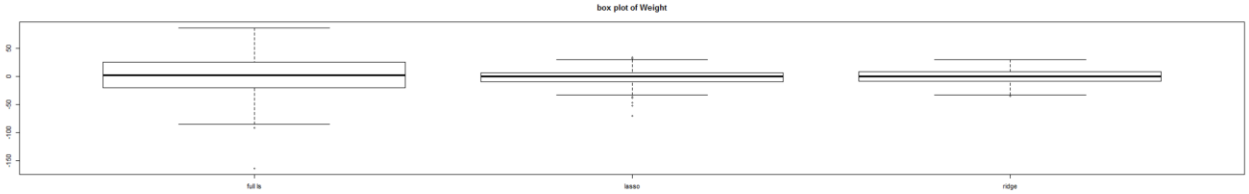
In the following examples, we used mtcars data to compare the full least squares estimates with the lasso, and ridge regression. For full least squares estimates we used the lm(), for the ridge estimates we used the lm.ridge() procedure in the R language.

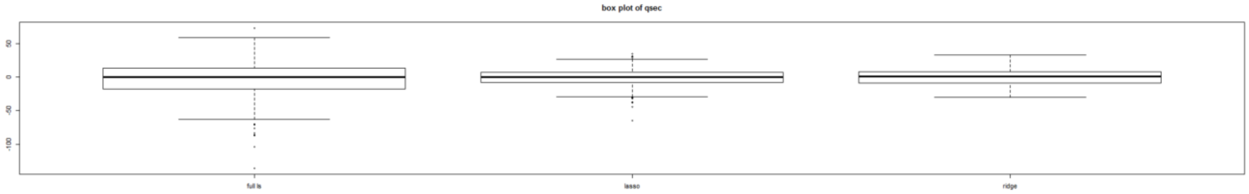

box plot of Cylinders


box plot of Displacement


box plot of HorsePower


box plot of Rear Axle Ratio


box plot of Weight


box plot of qsec

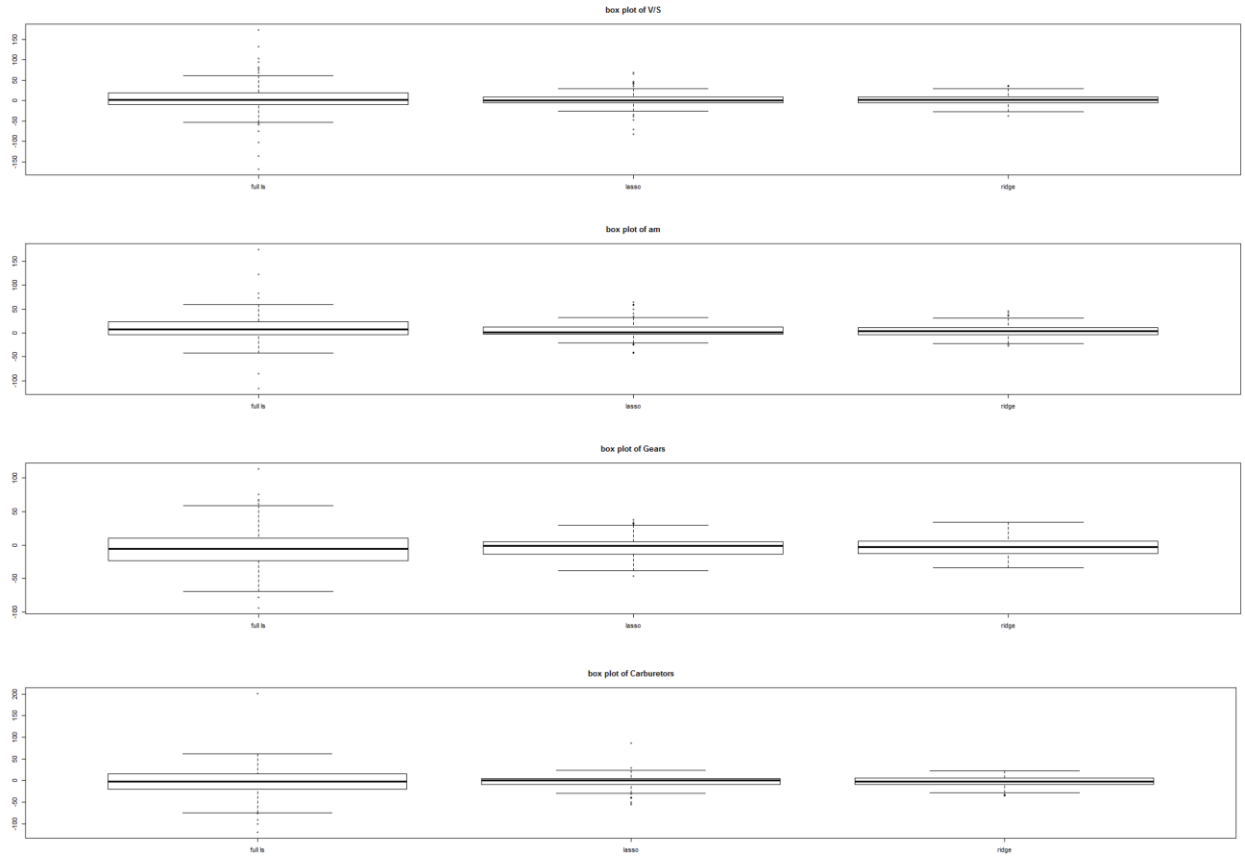


Figure 4. box plot to compare ols, Lasso and ridge estimate for each coefficients of mtcars.

The box plot in Figure 8, shows that the lasso coefficients are shrinked as compared to the OLS estimates and for this particular example are as good as the ridge coefficients.

## 7. Summary and Future Work

In previous sections, we introduced basic ideas about lasso. We can see that lasso actually has a similar result comparing to backwards selection on the selection of subset with the coefficients and the Z-score shrinked. We have discussed about the solution to the lasso quadratic problem in algorithm section using KKT approach. We also implemented a lasso function by transforming lasso problem to the form that can be solved by solve.QP function in R. We then used our program to compare the coefficients from three methods (OLS, Lasso and Ridge) with standardized “mtcars” data.

The potential future work might be to convert this into a generalized q-degree penalty.

## 8. References

[1] Robert Tibshirani 1996, Regression Shrinkage and Selection via the Lasso.

[2] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning.

[3] https://stats.stackexchange.com/questions/44838/how-are-the-standard-errors-of-coefficients-calculated-in-a-regression