# Project Report: Project 1 LASSO vs OLS

MAT 4376

Victor Matyiku, Rebecca Ly

2025-03-30

## Contents

## 1.Introduction

### 1.1 Background of OLS and LASSO Method

**OLS** (Ordinary Least Squares) and **LASSO** (Least Absolute Selection and Shrinkage Operator) are both methods that can be used to estimate regression coefficients for a line of best fit on data. These coefficients are typically represented by $\beta$, a $p$-dimensional vector where $p$ is the number of predictors.

The objective of the OLS method is to find a line of best fit that minimizes the sum of squared residuals, which are the differences between the predicted and actual data. This method is effective when the sample size ($n$) of the chosen dataset is greater than the number of predictors ($p$), i.e $n > p$. Otherwise the method is undefined. The OLS estimator derived from the method can be written as follows:

$$\hat{\beta}_{OLS} = argmin_{\beta} \frac{1}{2n} \|Y - X\beta\|_2^2$$

where

- $Y = (Y_1, ..., Y_n)^T$ is a $n \times 1$ response vector
- $X = (X_1, ..., X_n)^T$ is a $n \times p$ design matrix of covariates
- $\beta = (Y_1, ..., Y_n)^T$ is a $p$-dimensional regression coefficient vector

For $p > n$, the LASSO method is more favourable. The LASSO method uses L1 regularization where a penalty is added to the structure of the OLS method. The penalty achieves sparsity: an effect where the coefficients of some features are reduced close to, or exactly to zero, resulting in only certain features being selected in the final model. In other words, predictors with the least predictive power are dropped from the model and the remaining are seen as the most significant predictors. The estimator is typically written in Lagrangian form as follows:

$$\hat{\beta}_{LASSO} = argmin_{\beta}\{L(\beta) + \lambda\|\beta\|_1\}$$

$$\hat{\beta}_{LASSO} = argmin_{\beta}\{\frac{1}{2n}\sum_{i=1}^{n}(Y_i - \sum_{j=1}^{p} X_{ij}\beta_j)^2 + \lambda\sum_{j=1}^{p}|\beta_j|\} = argmin_{\beta}F(\beta)$$

One way of tuning the parameter lambda from a set of candidates is through *Cross-Validation (CV)*. The candidate that produces the LASSO estimator that minimizes the loss function is selected to be the most optimal lambda. A detailed procedure of the CV method is listed below:

1. Divide the data set $n$ into $m$ disjoint set $D_1, ..., D_m$ of size $n/m$ each.
2. For each $\lambda \in \Lambda$, evaluate $\hat{\beta}_{LASSO}^{-(h)}(\lambda)$. the LASSO estimator based on the dataset $D \setminus D_h$, $h = 1, ..., m$. Each $D_h$ is treated as a *test dataset*, while $D \setminus D_h$ as a *training dataset*.
3. Thus, for each $\lambda \in \Lambda$ we get $h$ LASSO estimators, making in total $h \times q$ LASSO estimators.
4. Define the loss function

$$CV(\lambda) = \sum_{h=1}^{m} \sum_{i:(X_i,Y_i)\in D_h} (Y_i - X_i^T \hat{\beta}_{LASSO}^{-(h)}(\lambda))^2$$

5. Choose $\lambda$ that minimized the loss function.

It can be seen that the OLS estimator can be derived from the LASSO estimator as $t \to \infty$. It should also be noted that while LASSO can set coefficients to zero, there is the risk of over-shrinkage.

## 1.2 Hypothesis Statement

- LASSO is expected to show better results [finish]

### 1.3 Goals

1. Write a function for OLS using `lm()`
2. Write a function for LASSO using $\lambda$ chosen from cross-validation
3. Successfully implement one hundred iterations of both functions for each chosen value of $p$ (500 iterations total)
4. Summarize the number of chosen predictors for each $p$ and record the mean squared value

## 2. Data Source and Methodology

### 2.1 Generating Data

The program, designed in R, used five values of $p = \{2, 5, 10, 15, 25\}$. Each iteration on $p$ created a randomly generated, $p$-dimensional dataset based on the following linear model:

$$Y = \beta_0 + X\beta + \epsilon$$

where

- $Y = (Y_1, ..., Y_n)^T$ is a $n \times 1$ response vector
- $X = (X_1, ..., X_n)^T$ is a $n \times p$ design matrix of covariates
- $\beta = (Y_1, ..., Y_n)^T$ is a $p$-dimensional regression coefficient vector
- $\beta_0$ is the intercept
- $\epsilon$ is the $n$-dimensional noise vector

No other manipulation was performed on the dataset.

### 2.2 Code Design

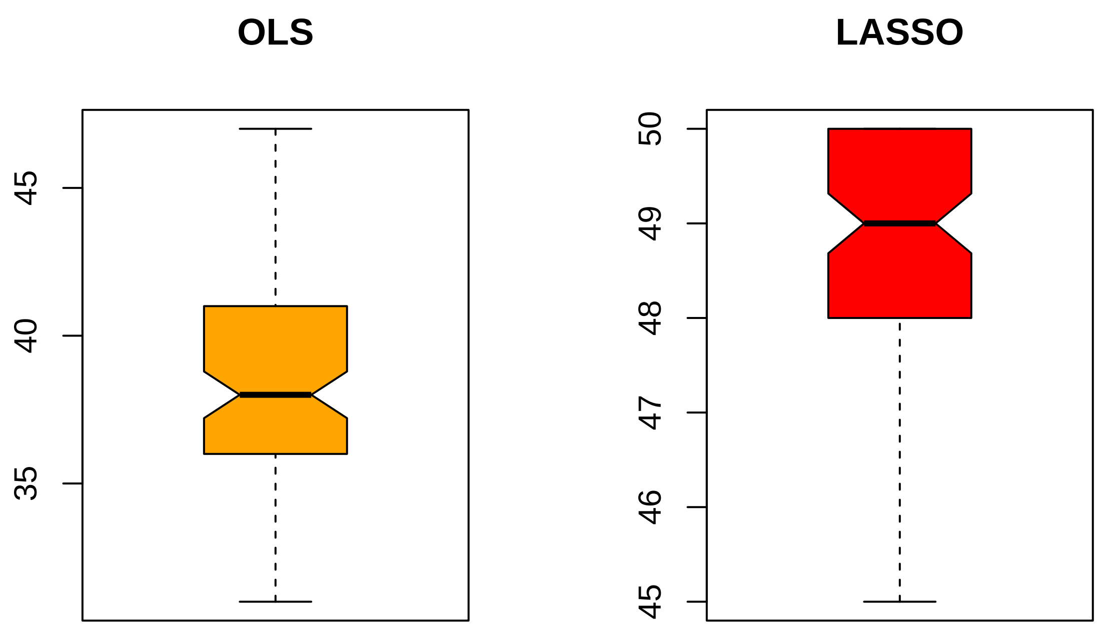
The following functions were written for this project:

- `generate_data(p, n)` , which initially creates the normally distributed data set $X$ using the number of predictors ($p$) and number of observations ($n$);
- `ols(input_data)` calculates the OLS regression using `lm()` then extracts various derived terms (e.g. number of significant predictors, squared error);
- `lasso(input_data)` calculates the LASSO regression using `glmnet::cv.glmnet` then also extracts various derived terms (e.g. number of relevant predictors, squared error);
- `analysis_for(p)` performs 100 OLS and LASSO regressions using the number of predictors ($p$) then generates the summary statistics (e.g. mean squared error) for $p$.

Then, the final code block runs the `analysis_for` function for the five values of $p$ above and generates summary statistics.

## 3. Results

### Figure 1: Boxplot of q's for p = 50



From this figure and the rest (see Appendix for other $p$ values), we can see that LASSO consistently has an at least equal number of relevant predictors ($q$) as OLS, though LASSO often has a much higher number. Interestingly, while OLS tends to distribute its $q$'s near the median as the $p$ increases, LASSO's $q$'s are distributed near the maximum.

### Table 1: MSE for OLS and LASSO

| p | OLS | LASSO |
|---:|---:|---:|
| 2 | 403.7811 | 95.05817 |
| 5 | 695.8608 | 92.60197 |
| 10 | 1206.1890 | 87.32800 |
| 25 | 2583.2156 | 73.99934 |
| 50 | 4979.0414 | 49.84653 |

From the table, we can see that LASSO has a much smaller mean squared error compared to OLS every time. Specifically, we can see that LASSO actually decreases its mean squared error as $p$ increases, whereas OLS's mean squared error increases as $p$ increases.

## 4. Conclusion

## 5. Appendix

**Figure 2: Boxplots of q's for p = 2**

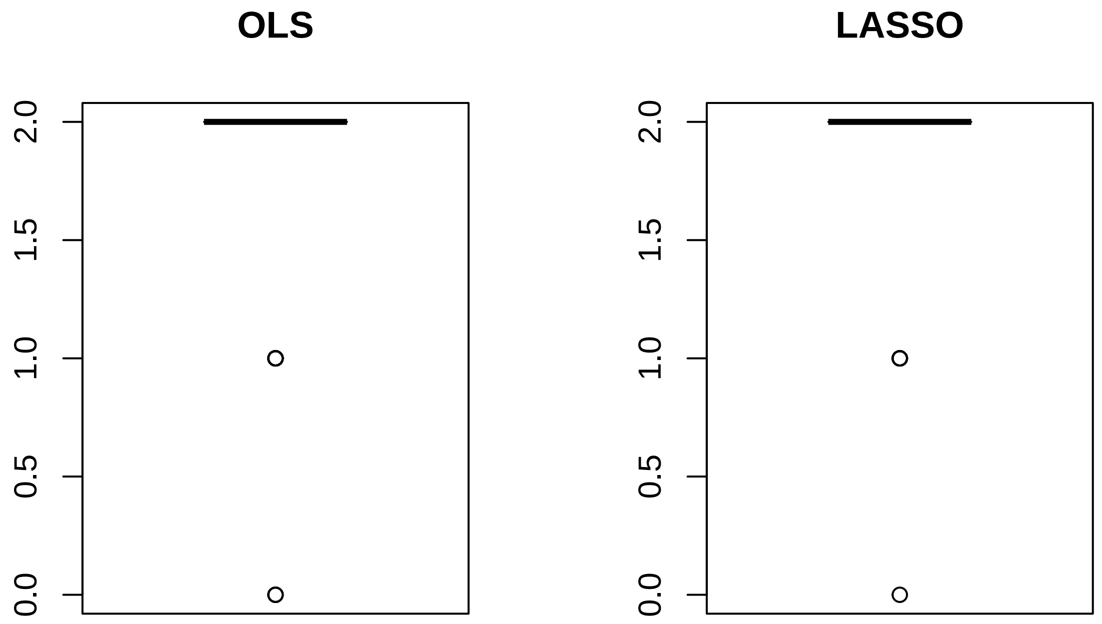


**Figure 3: Boxplots of q's for p = 5**

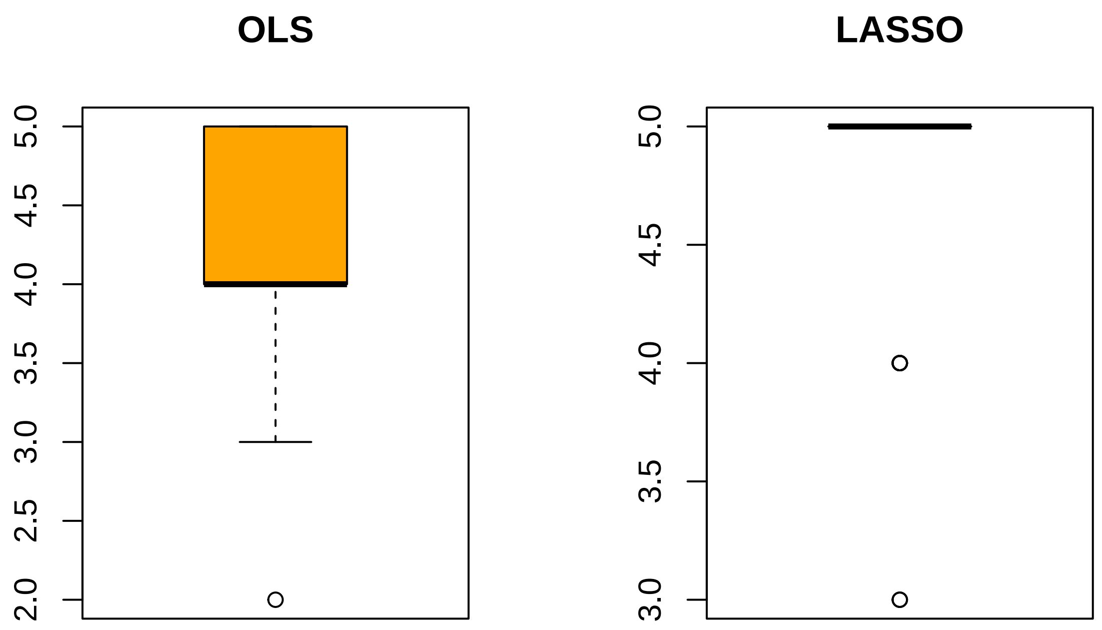


**Figure 4: Boxplots of q's for p = 10**

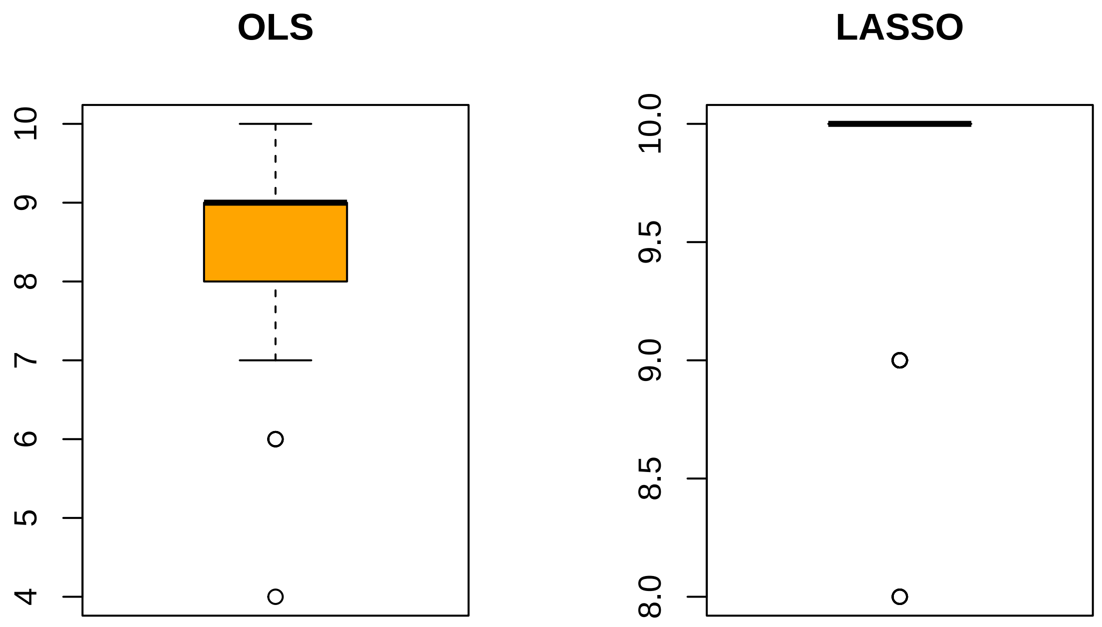
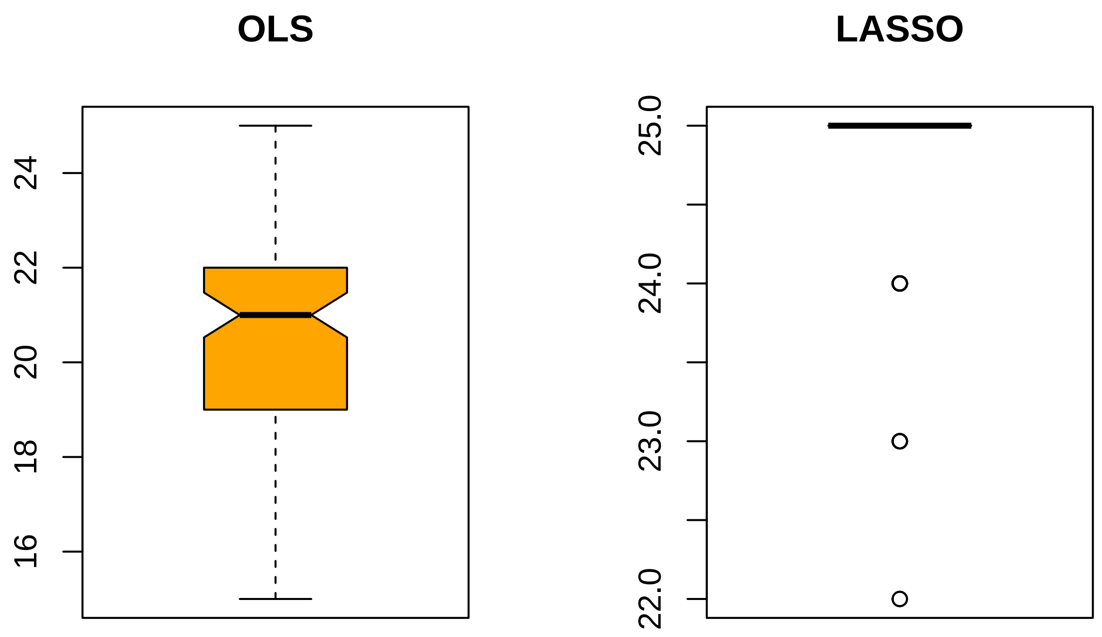


**Figure 5: Boxplots of q's for p = 25**



## Code

```
library(knitr)
knitr::opts_chunk$set(echo = TRUE)
set.seed(123) # so our conclusions don't suddenly change
# Generating data
generate_data <- function(p, n){
  beta0 <- rnorm(1)
  beta <- matrix(rnorm(p), nrow = p, ncol = 1)

  predictors <- matrix(
    rnorm(n*p), nrow = n, ncol = p,
    dimnames = list(rows = 1:n, cols = paste("X", 1:p, sep=""))
  )

  errors <- rnorm(n)
  observations <- data.frame(Y = beta0 + predictors %*% beta + errors)

  output <- as.data.frame(
    cbind(observations, predictors)
  )
  return(output)
}
#generated_data <- generate_data(10, 100)
#head(generated_data)

# OLS
ols <- function(input_data){
  ols_model <- lm(Y ~ ., input_data)

  # extract only coefficients whose p value <= 0.5
  significant_coefs <- data.frame(Coefficients = summary(ols_model)$coef[summary(ols_model)$coef[,4] <=

  # transpose vector so calculations work properly later
  ols_coefs <- t(significant_coefs)
  significant_predictors <- colnames(ols_coefs)[colnames(ols_coefs) != "(Intercept)"]
  ols_coefs <- ols_coefs[, significant_predictors]

  # check if intercept is significant
  if (is.na(significant_coefs["(Intercept)",])) {
    ols_beta_0 <- 0
  } else {
    ols_beta_0 <- significant_coefs["(Intercept)",]
  }

  sig_predictors_matrix <- as.matrix(input_data[, significant_predictors])
  errors <- input_data$Y - ols_beta_0 - sig_predictors_matrix %*% ols_coefs

  # calculating squared error using euclidean norm
  output <- list(sq_error = (norm(errors, "e"))^2,
                 q = length(significant_predictors))

  return(output)
}
#ols(generated_data)

# LASSO
library(glmnet)
lasso <- function(input_data){
  X <- as.matrix(input_data[colnames(input_data) != "Y"])
  Y <- as.matrix(input_data$Y)

  # LASSO with CV
  cv.fit <- cv.glmnet(X, Y, nfolds=10, family="gaussian")

  # coefficients for LASSO model using optimal lambda
  lasso_coefs <- coef(cv.fit, s="lambda.min")

  no_intercept <- lasso_coefs@i[lasso_coefs@i != 0]

  #predict new coefficients using LASSO & optimal lambda
  prediction <- predict(cv.fit, newx = X, s="lambda.min")

  # calculating squared error using euclidean norm
  output <- list(sq_error = (norm(Y-prediction, "e"))^2,
                 q = length(no_intercept))

  return(output)
}
#lasso(generated_data)

# Repetition
analysis_for <- function(p) {
  statistics <- list(
    ols_err = 1:100,
    lasso_err = 1:100,
    ols_q = 1:100,
    lasso_q = 1:100
  )

  for (i in 1:100) {
    generated_data <- generate_data(p, 100)
    ols_stats <- ols(generated_data)
    lasso_stats <- lasso(generated_data)

    statistics$ols_err[i] <- ols_stats$sq_error
    statistics$lasso_err[i] <- lasso_stats$sq_error
    statistics$ols_q[i] <- ols_stats$q
    statistics$lasso_q[i] <- lasso_stats$q
  }

  statistics$ols_mse <- sum(statistics$ols_err)/100
  statistics$lasso_mse <- sum(statistics$lasso_err)/100

  return(statistics)
}

#analysis_for(10)
# p Values
p_values <- c(2, 5, 10, 25, 50)

summary_df <- data.frame(
  p = p_values,
  OLS = p_values,
  LASSO = p_values #just to be consistent in size
)

q_dispersions <- list(ols = list(), lasso = list())

for (i in 1:5) {
  p <- p_values[i]

  #getting statistics
  statistics <- analysis_for(p)
  ols_mse <- statistics$ols_mse
  lasso_mse <- statistics$lasso_mse
```

```r
  #inputting into matrix
  summary_df$OLS[i] <- ols_mse
  summary_df$LASSO[i] <- lasso_mse

  q_dispersions$ols[p] <- list(statistics$ols_q)
  q_dispersions$lasso[p] <- list(statistics$lasso_q)
}

par(mfrow=c(1,2))

boxplot(q_dispersions$ols[50],
  main = "OLS",
  col = "orange",
  notch = TRUE
)

boxplot(q_dispersions$lasso[50],
  main = "LASSO",
  col = "red",
  notch = TRUE
)
kable(summary_df)
par(mfrow=c(1,2))

boxplot(q_dispersions$ols[2],
  main = "OLS",
  col = "orange"
)

boxplot(q_dispersions$lasso[2],
  main = "LASSO",
  col = "red"
)
par(mfrow=c(1,2))

boxplot(q_dispersions$ols[5],
  main = "OLS",
  col = "orange"
)

boxplot(q_dispersions$lasso[5],
  main = "LASSO",
  col = "red"
)
par(mfrow=c(1,2))

boxplot(q_dispersions$ols[10],
  main = "OLS",
  col = "orange"
)

boxplot(q_dispersions$lasso[10],
  main = "LASSO",
  col = "red"
)
par(mfrow=c(1,2))

boxplot(q_dispersions$ols[25],
  main = "OLS",
  col = "orange",
  notch = TRUE
)

boxplot(q_dispersions$lasso[25],
  main = "LASSO",
  col = "red"
)
```