Homework 1

Lanston

1. What is the main objective of Lasso regression?

a) To shrink the coefficients of the model towards zero
b) To perform feature selection by forcing some coefficients to be exactly equal to zero
c) To minimize the residual sum of squares
d) To maximize R-squared

Answer 1: b)

2. What is the main difference between Ridge and Lasso regression?

a) Ridge regression shrinks the coefficients of the model towards zero, while Lasso regression forces some coefficients to be exactly equal to zero
b) Lasso regression shrinks the coefficients of the model towards zero, while Ridge regression forces some coefficients to be exactly equal to zero
c) Both Ridge and Lasso regression shrink the coefficients of the model towards zero
d) Both Ridge and Lasso regression force some coefficients to be exactly equal to zero

Answer 2: a)

3. What is the main advantage of splitting the data into training and testing sets when building a linear regression model?

a) It allows us to evaluate the model's performance on new, unseen data
b) It allows us to estimate the model's true error, which is the error that will be made when the model is used to make predictions on new data
c) It prevents overfitting such that the fitted model that has a high accuracy on the given data but performs poorly on new, unseen data
d) All of the above

Answer 3:d)

4. Boston housing data

```r
Boston <- read.csv('https://zhang-datasets.s3.us-east-2.amazonaws.com/Boston.csv')
```

(a) Fit a simple linear regression that predicts *medv* (median house value) using *lstat* (percent households with low socioeconomic status). Use this model to answer the following questions.

```r
lm1 <-lm(medv~lstat,data=Boston)
summary(lm1)
```

```
##
## Call:
## lm(formula = medv ~ lstat, data = Boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.168  -3.990  -1.318   2.034  24.500
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 34.55384    0.56263   61.41   <2e-16 ***
## lstat       -0.95005    0.03873  -24.53   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.216 on 504 degrees of freedom
## Multiple R-squared:  0.5441, Adjusted R-squared:  0.5432
## F-statistic: 601.6 on 1 and 504 DF,  p-value: < 2.2e-16
```

From the p_values, is *lstat* a significant predictor for *medv*? What is the R-squared of this linear regression?

Answer: yes, the p_value is significantly less than 5%

What happens to *medv* with one precent increase in lstat?

Answer: one percent increase in lstat, the the medv will descrise by 1%*-0.95005%

(b) Predict *medv* for *lstat* = 5, 10, 15, respectively.

```r
new_lstat <-data.frame(lstat=c(5,10,15))
predict(lm1,new_lstat)
```

```
##        1        2        3
## 29.80359 25.05335 20.30310
```

(c) Fit a multiple linear regression that predicts *medv* using all the covariates in the data set. Use this model to answer the following questions.

```r
lm2 <- lm(medv~.,data = Boston)
summary(lm2)
```

```
##
## Call:
## lm(formula = medv ~ ., data = Boston)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.1304  -2.7673  -0.5814   1.9414  26.2526
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  41.617270   4.936039   8.431 3.79e-16 ***
## crim         -0.121389   0.033000  -3.678 0.000261 ***
## zn            0.046963   0.013879   3.384 0.000772 ***
## indus         0.013468   0.062145   0.217 0.828520
## chas          2.839993   0.870007   3.264 0.001173 **
## nox         -18.758022   3.851355  -4.870 1.50e-06 ***
## rm            3.658119   0.420246   8.705  < 2e-16 ***
## age           0.003611   0.013329   0.271 0.786595
## dis          -1.490754   0.201623  -7.394 6.17e-13 ***
## rad           0.289405   0.066908   4.325 1.84e-05 ***
## tax          -0.012682   0.003801  -3.337 0.000912 ***
## ptratio      -0.937533   0.132206  -7.091 4.63e-12 ***
## lstat        -0.552019   0.050659 -10.897  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.798 on 493 degrees of freedom
## Multiple R-squared:  0.7343, Adjusted R-squared:  0.7278
## F-statistic: 113.5 on 12 and 493 DF,  p-value: < 2.2e-16
```

Which covariates are significant?

Answer: instead of indus and age, all the covariates are significant

What is the R-squared now?

Answer: the multiple R-squared are 0.0734 and adjusted R-squared are 0.7278

(d) We will now try to predict per capita crime rate by town (the column *crim*) in the Boston data set. Considering all predictors in this data set, try out the standard linear regression, ridge regression and lasso regression. Propose a model that performs the best on this data set, and justify your answer using evidence from model fitting. (Note: make sure that you evaluate model performance using testing error, as opposed to training error.)

```r
library(glmnet)
```

```
## Warning: package 'glmnet' was built under R version 4.3.2
```

```
## Loading required package: Matrix
## Warning: package 'Matrix' was built under R version 4.3.2
## Loaded glmnet 4.1-8
```

```r
library(car)
```

```
## Warning: package 'car' was built under R version 4.3.2
## Loading required package: carData
## Warning: package 'carData' was built under R version 4.3.2
```

```r
grid = 10^seq(10,-2,length=100)
x = model.matrix(crim~.,Boston)
y = Boston$crim

set.seed(1)

dim(Boston)
```

```
## [1] 506  13
```

```r
train <- sample(1:nrow(x),nrow(x)/2)
test <- setdiff(1:nrow(x), train)

lm3 <- lm(crim~.,data = Boston, subset = train)
lm_lasso <- glmnet(x[train,],y[train], alpha=1,lambda=grid)
lm_ridge <- glmnet(x[train,],y[train], alpha=0,lambda=grid)

#calculate the min lambda
cv.out1 <- cv.glmnet(x[train,],y[train],alpha=1)
best_lambda_lasso <- cv.out1$lambda.min
best_lambda_lasso
```

```
## [1] 0.01161955
```

```r
cv.out2 <- cv.glmnet(x[train,],y[train],alpha=0)
best_lambda_rige <- cv.out2$lambda.min
best_lambda_rige
```

```
## [1] 0.5919159
```

```r
#make the prediction
lm3.pred <- predict(lm3,newdata = Boston[test,])
lm_lasso.pred <- predict(lm_lasso,s=0.003893966,newx = x[test,])
lm_ridge.pred <- predict(lm_ridge,s=0.502925,newx = x[test,])

#mse calculation
mse_lm <- mean((y[test]-lm3.pred)^2)
mse_lasso <- mean((y[test]-lm_lasso.pred)^2)
mse_ridge <- mean((y[test]-lm_ridge.pred)^2)
```

```r
#compare the 3 kinds of models
print(list(lm = mse_lm, lasso = mse_lasso, ridge = mse_ridge))
```

```
## $lm
## [1] 41.19923
##
## $lasso
## [1] 41.03398
##
## $ridge
## [1] 40.24592
```

```r
summary(lm3)
```

```
##
## Call:
## lm(formula = crim ~ ., data = Boston, subset = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -10.574  -2.723  -0.566   1.351  57.279
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  18.665277  10.693890   1.745  0.08219 .
## zn            0.046105   0.028198   1.635  0.10336
## indus        -0.127032   0.116665  -1.089  0.27730
## chas         -0.916885   1.769171  -0.518  0.60476
## nox         -11.606805   7.924234  -1.465  0.14431
## rm            0.738859   0.913675   0.809  0.41951
## age          -0.010585   0.026291  -0.403  0.68761
## dis          -1.184115   0.427288  -2.771  0.00602 **
## rad           0.671788   0.130702   5.140  5.7e-07 ***
## tax          -0.004607   0.007552  -0.610  0.54237
## ptratio      -0.515160   0.284824  -1.809  0.07175 .
## lstat         0.296310   0.114591   2.586  0.01031 *
## medv         -0.249594   0.097588  -2.558  0.01115 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.693 on 240 degrees of freedom
## Multiple R-squared:  0.5021, Adjusted R-squared:  0.4772
## F-statistic: 20.17 on 12 and 240 DF,  p-value: < 2.2e-16
```

```r
vif(lm3)
```

```
##       zn    indus     chas      nox       rm      age      dis      rad
## 1.989247 3.532785 1.108023 4.667127 2.272078 3.178961 4.237537 7.524784
##      tax  ptratio    lstat     medv
```

```
## 9.381149 1.883494 3.640571 4.134604
```

answers for comparing 3 kinds of model

1.from the mse, the ridge regression model is better.

2.as we can see the summary of standard linear regression, there are several variables are significant, so ridge is better which make sense

3.use VIF value we can see that tax and rad which vif value is 7 and 9 which bigger than 5, means there are collinearity in standard linear regression model, so ridge regression is better.