P8106 Assignment 1

Yihan Feng

2/14/2021

Set up libraries

```
library(tidyverse)
library(caret)
library(glmnet)
library(pls)
```

Data import and set up

```
set.seed(888)

setwd("C:/Users/irene/OneDrive - cumc.columbia.edu/2021 M1 Spring/Data Science 2/HW/hw1/p8106_hw1_yf255
train = read_csv("./solubility_train.csv")
test = read_csv("./solubility_test.csv")

train_x = model.matrix(Solubility ~ ., train)[,-1]
train_y = train$Solubility

test_x = model.matrix(Solubility ~ ., test)[,-1]
test_y = test$Solubility

ctrl = trainControl(method = "repeatedcv", number = 10, repeats = 5)
```

a. Least Squares

```
set.seed(888)
lm.fit = train(train_x, train_y,
               method = "lm",
               trControl = ctrl)

pred.lm = predict(lm.fit, newdata = test_x)
rmse.lm = RMSE(test_y, pred.lm)
```

The mean squared error is 0.746.

b. Ridge Regression

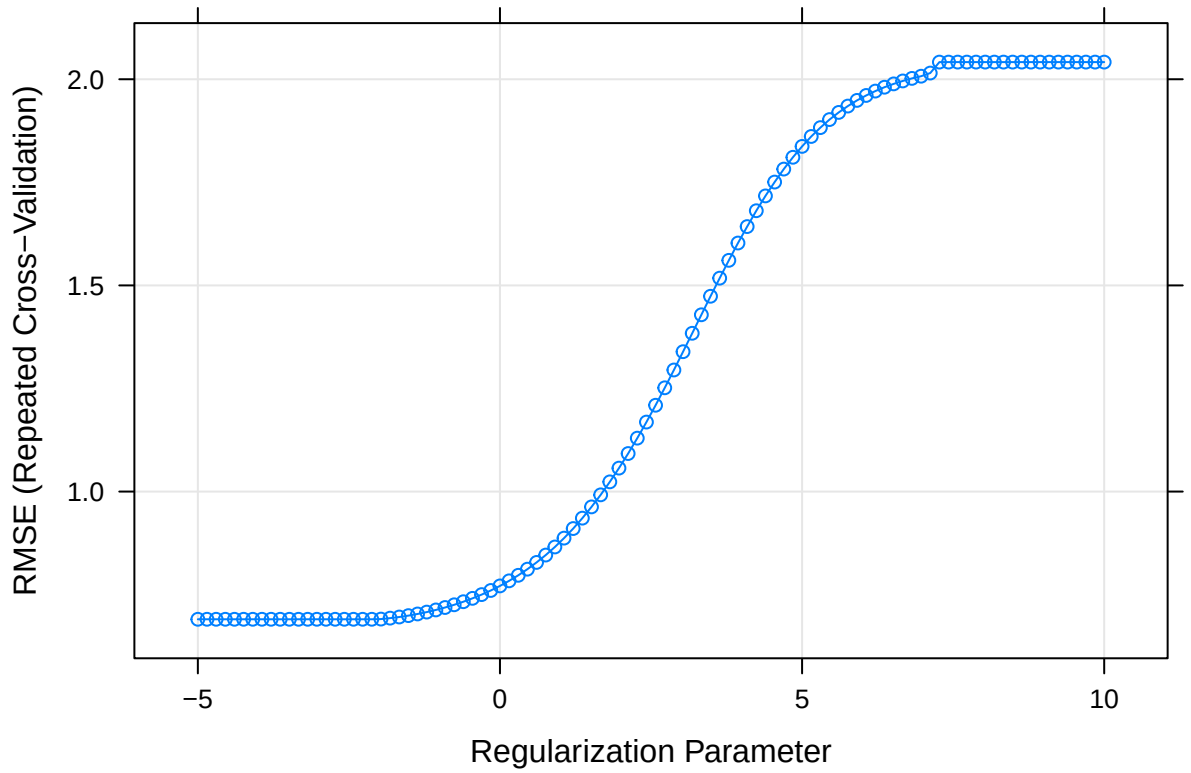
```
set.seed(888)
ridge.fit = train(train_x, train_y,
                  method = "glmnet",
                  tuneGrid = expand.grid(alpha = 0,
                                         lambda = exp(seq(-5, 10, length = 100))),
                  preProc = c("center", "scale"),
                  trControl = ctrl)

ridge.fit$bestTune
```

```
##    alpha    lambda
## 20     0 0.1198862
```

```
pred.ridge = predict(ridge.fit, newdata = test_x)
rmse.ridge = RMSE(test_y, pred.ridge)

plot(ridge.fit, xTrans = log)
```



The λ chosen by cross-validation is 0.126. The test error is 0.717.

c. Lasso Regression

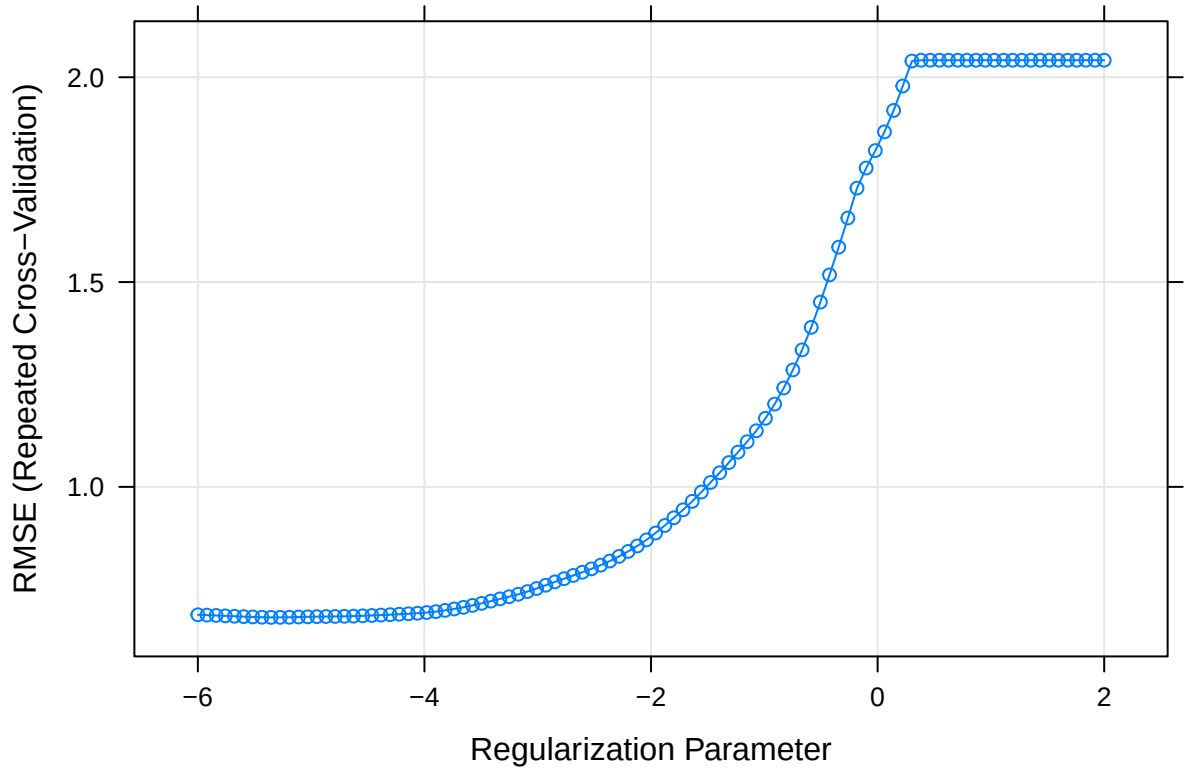
```
set.seed(888)
lasso.fit = train(train_x, train_y,
                  method = "glmnet",
                  tuneGrid = expand.grid(alpha = 1,
                                         lambda = exp(seq(-6, 2, length = 100))),
                  preProc = c("center", "scale"),
                  trControl = ctrl)
lasso.fit$bestTune
```

```
##   alpha      lambda
## 9     1 0.004731394
```

```
pred.lasso = predict(lasso.fit, newdata = test_x)
rmse.lasso = RMSE(test_y, pred.lasso)

non_zero = coef(lasso.fit$finalModel, s = lasso.fit$bestTune$lambda) != 0

plot(lasso.fit, xTrans = log)
```



The λ chosen by cross-validation is 0.0047. The test error is 0.706. The number of non-zero coefficient estimates in the model is 144.

d. PCR

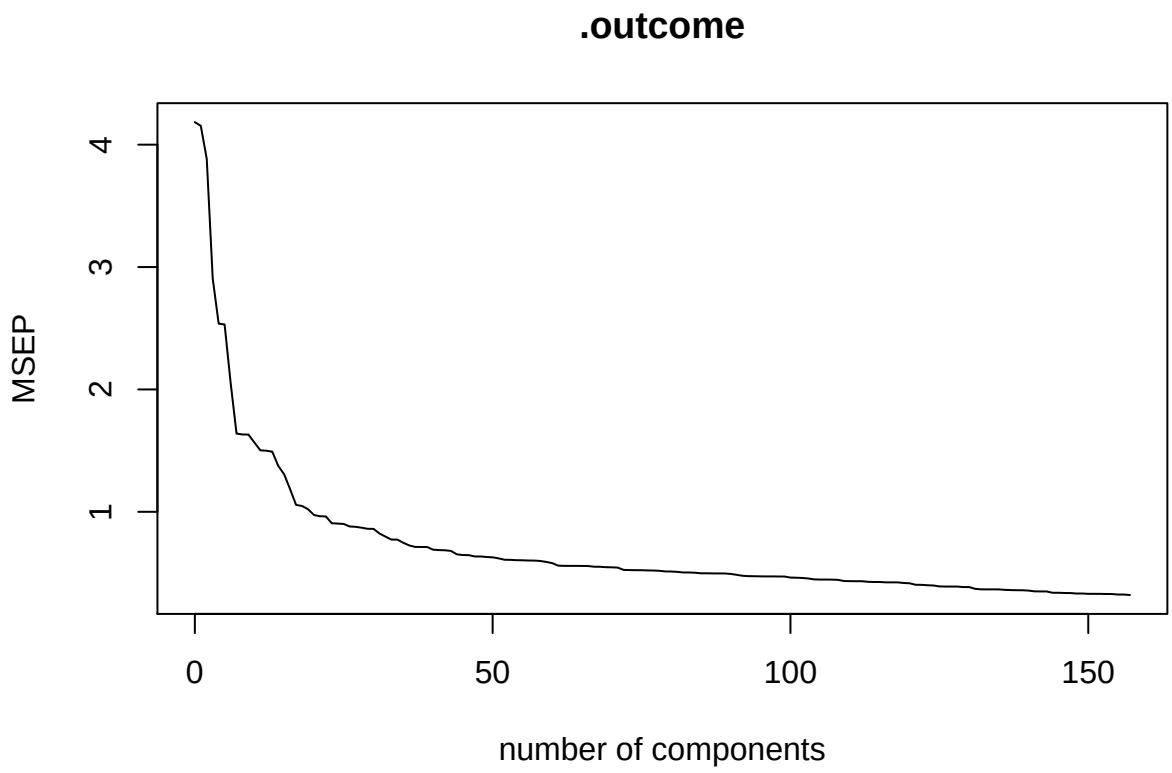
```
set.seed(888)
pcr.fit = train(train_x, train_y,
                method = "pcr",
                tuneGrid = data.frame(ncomp = 1:226),
                tuneLength = length(train),
                preProc = c("center", "scale"),
                trControl = ctrl)

pcr.fit$bestTune
```

```
##     ncomp
## 157   157
```

```
pred.pcr = predict(pcr.fit, newdata = test_x)
rmse.pcr = RMSE(test_y, pred.pcr)

validationplot(pcr.fit$finalModel, val.type = "MSEP")
```



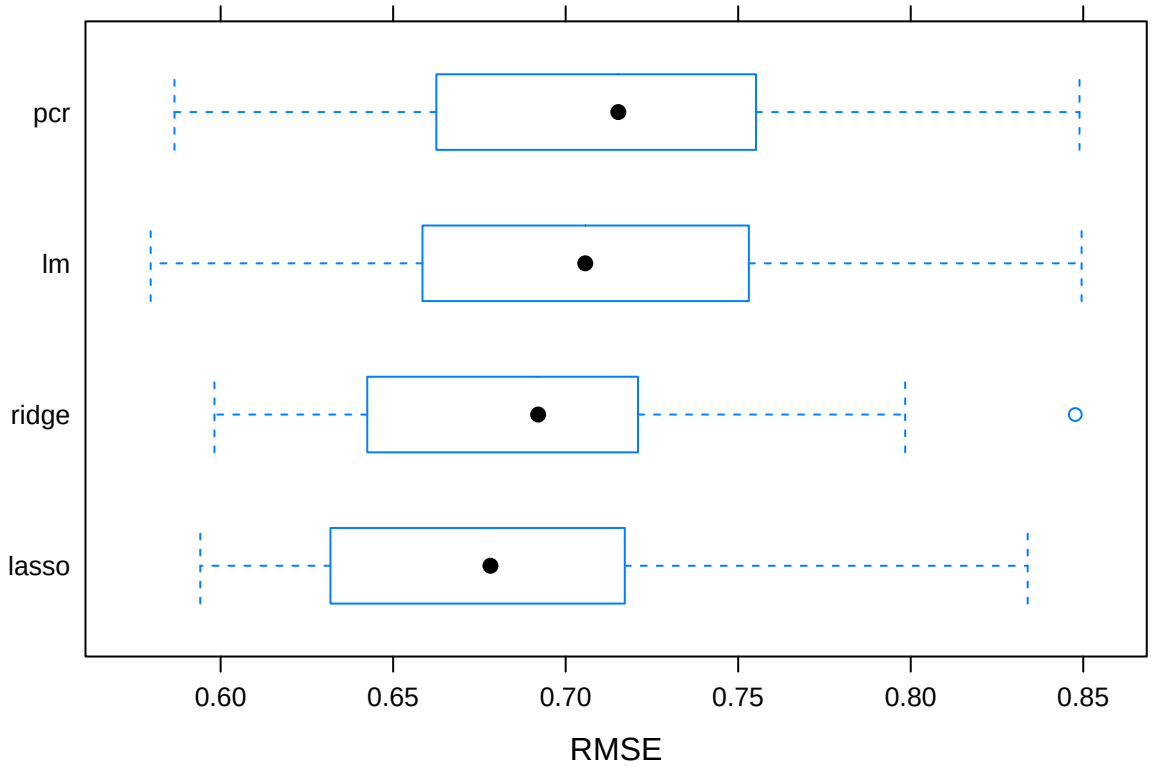
The M chosen by cross-validation is 157. The test error is 0.742.

e. Model selection

```
set.seed(888)
resample = resamples(list(lm = lm.fit,
                          ridge = ridge.fit,
                          lasso = lasso.fit,
                          pcr = pcr.fit))
summary(resample)
```

```
##
## Call:
## summary.resamples(object = resample)
##
## Models: lm, ridge, lasso, pcr
## Number of resamples: 50
##
## MAE
##            Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
## lm    0.4468867 0.4881411 0.5305533 0.5279005 0.5552502 0.6389394    0
## ridge 0.4502546 0.4978506 0.5197123 0.5247712 0.5513622 0.6112160    0
## lasso 0.4430475 0.4947076 0.5169765 0.5192260 0.5445943 0.6250290    0
## pcr   0.4375922 0.5115133 0.5448737 0.5437747 0.5745087 0.6488673    0
##
## RMSE
##            Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
## lm    0.5797165 0.6589665 0.7056781 0.7075712 0.7524940 0.8495215    0
## ridge 0.5982036 0.6439133 0.6920174 0.6903004 0.7203360 0.8476839    0
## lasso 0.5941131 0.6352401 0.6781913 0.6812847 0.7171065 0.8338925    0
## pcr   0.5866279 0.6626439 0.7152601 0.7125931 0.7549252 0.8489115    0
##
## Rsquared
##            Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
## lm    0.8271904 0.8636814 0.8835413 0.8811753 0.9005140 0.9271165    0
## ridge 0.8370724 0.8712622 0.8892398 0.8862398 0.9015232 0.9187072    0
## lasso 0.8510410 0.8737466 0.8901338 0.8890886 0.9022358 0.9285259    0
## pcr   0.8396785 0.8632298 0.8828183 0.8796094 0.8920293 0.9174622    0
```

```
bwplot(resample, metric = "RMSE")
```



Based on the resample summary, I would choose Lasso regression, because it has the lowest mean MAE and RMSE, as well as the highest Rsquared.