homework1

Na Yun Cho

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

```
train_data = read.csv("./data/solubility_train.csv")
train_data <- na.omit(train_data)

test_data = read.csv("./data/solubility_test.csv")
test_data <- na.omit(test_data)

train2 <- model.matrix(Solubility ~ .,train_data)[ ,-1]
test2 <- model.matrix(Solubility ~ ., test_data)[ ,-1]

#matrix of predictors
x <- train2
y1 <- test2
#vector of response
y <- train_data$Solubility
y2 <- test_data$Solubility
```

Part(a)

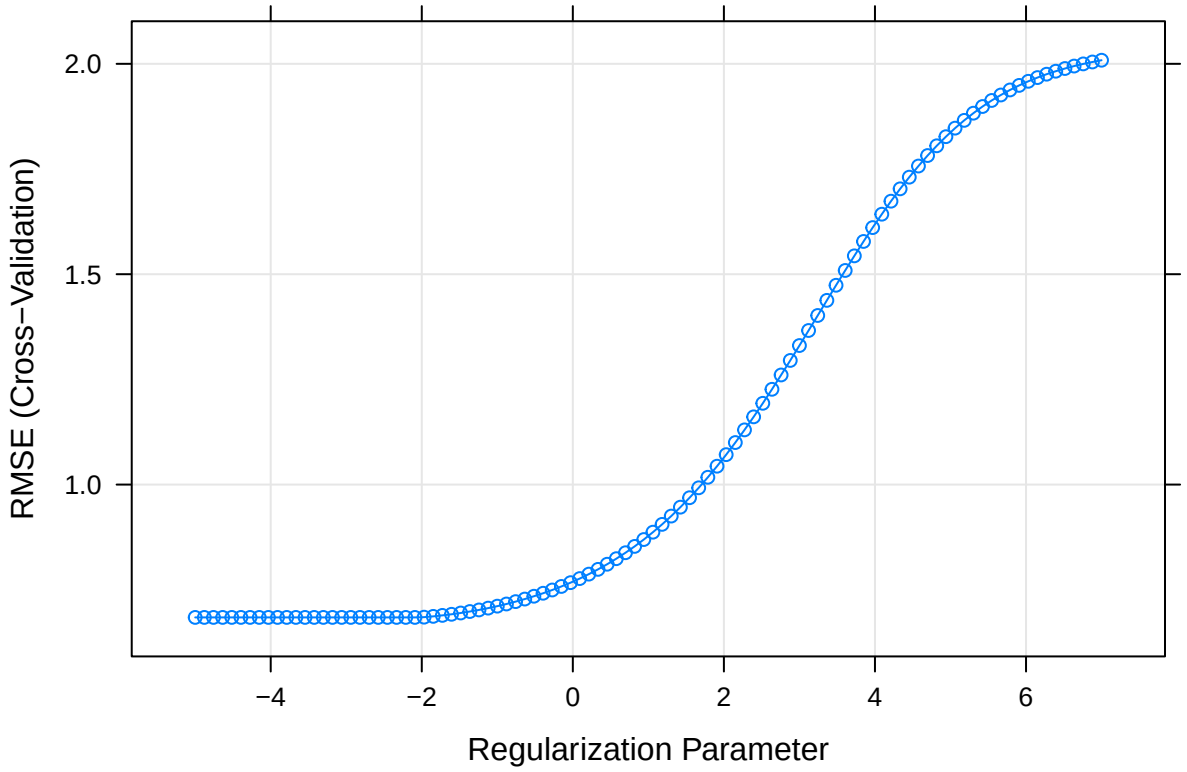
```
#fit a linear model using least squares
ctrl1 <- trainControl(method = "cv", number = 10)
set.seed(1)
lm.fit <- train(x, y,
                method = "lm",
                trControl = ctrl1)

#calculate mean squared error using test data
linear.pred <- predict(lm.fit, newdata = test2)
mean((linear.pred - test_data$Solubility)^2)
```

```
## [1] 0.5558898
```

The mean squared error using the test data is 0.5558898

Part(b)

```
#fit a ridge regression model
ctrl1 <- trainControl(method = "cv", number = 10)
set.seed(1)
ridge.fit <- train(x, y,
                   method = "glmnet",
                   tuneGrid = expand.grid(alpha = 0,
                   lambda = exp(seq(7, -5, length=100))),
                   preProc = c("center", "scale"),
                   trControl = ctrl1)
#choose lambda
plot(ridge.fit, xTrans = log)
```



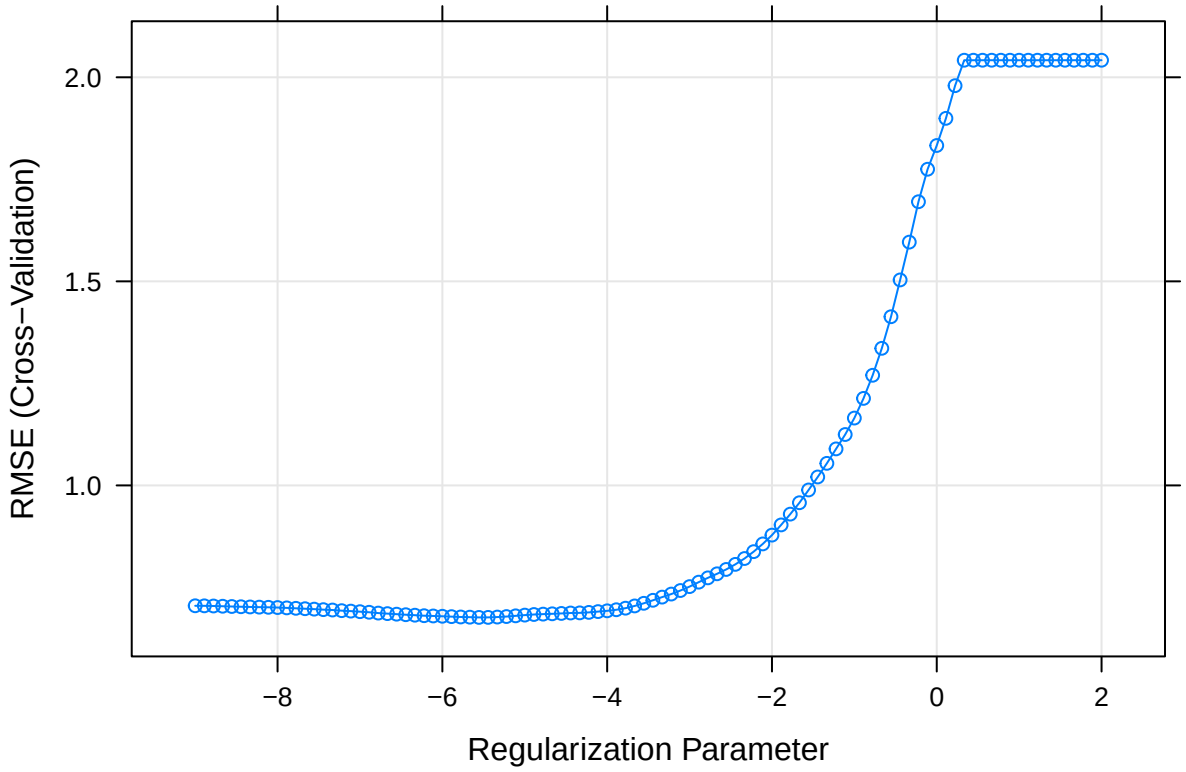

```
best_ridge = ridge.fit$bestTune

#calculate test error
ridge.pred <- predict(ridge.fit, newdata = test2)
mean((ridge.pred - test_data$Solubility)^2)
```

```
## [1] 0.5134603
```

The chosen lambda is 0.1235747 and the test error is 0.5134603

Part(c)

```
#fit a lasso model
set.seed(1)
lasso.fit <- train(x, y,
             method = "glmnet",
             tuneGrid = expand.grid(alpha = 1,
             lambda = exp(seq(2, -9, length=100))),
             preProc = c("center", "scale"),
             trControl = ctrl1)

#choose lambda
plot(lasso.fit, xTrans = log)
```

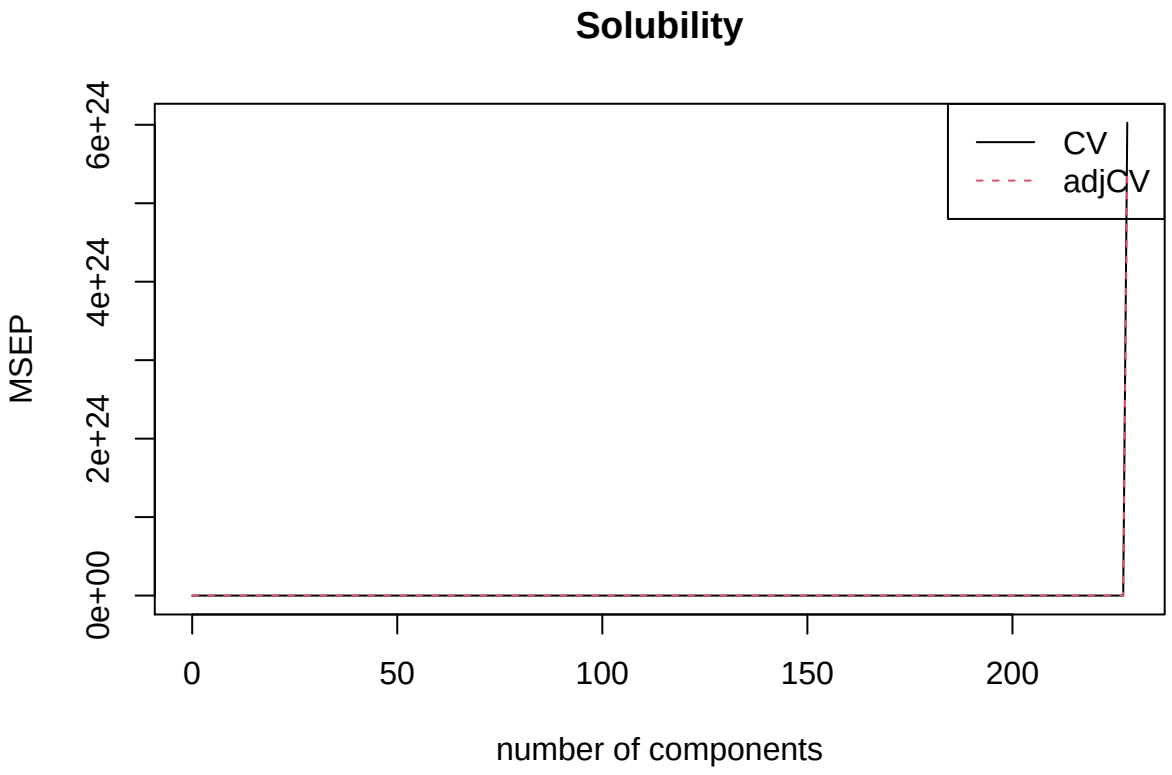



```
best_lasso = lasso.fit$bestTune

#calculate test error
lasso.pred <- predict(lasso.fit, newdata = test2)
mean((lasso.pred - test_data$Solubility)^2)

#find number of coefficient estimates
coef(lasso.fit$finalModel, lasso.fit$bestTune$lambda)
```

The chosen lambda is 0.00386592, the test error is 0.50333, and there are 149 non-zero coefficient estimates in the model.

Part(d)

```
#fit model using function pcr()
set.seed(1)
pcr.mod <- pcr(Solubility ~ .,
data = train_data,
scale = TRUE,
validation= "CV")
summary(pcr.mod)

validationplot(pcr.mod, val.type="MSEP", legendpos = "topright")
```




```
cv.mse <- RMSEP(pcr.mod)
ncomp.cv <- which.min(cv.mse$val[1,,])-1
ncomp.cv

predy2.pcr2 <- predict(pcr.mod, newdata = y1, ncomp = ncomp.cv)
mean((y2 - predy2.pcr2)^2)


#fit model using caret
ctrl1 <- trainControl(method = "cv", selectionFunction = "best")
modelLookup("pcr")
set.seed(1)
pcr.fit <- train(x, y,
           method = "pcr",
           tuneGrid = data.frame(ncomp = 1:190),
```

```
            trControl = ctrl1,
            preProcess = c("center", "scale"))
```

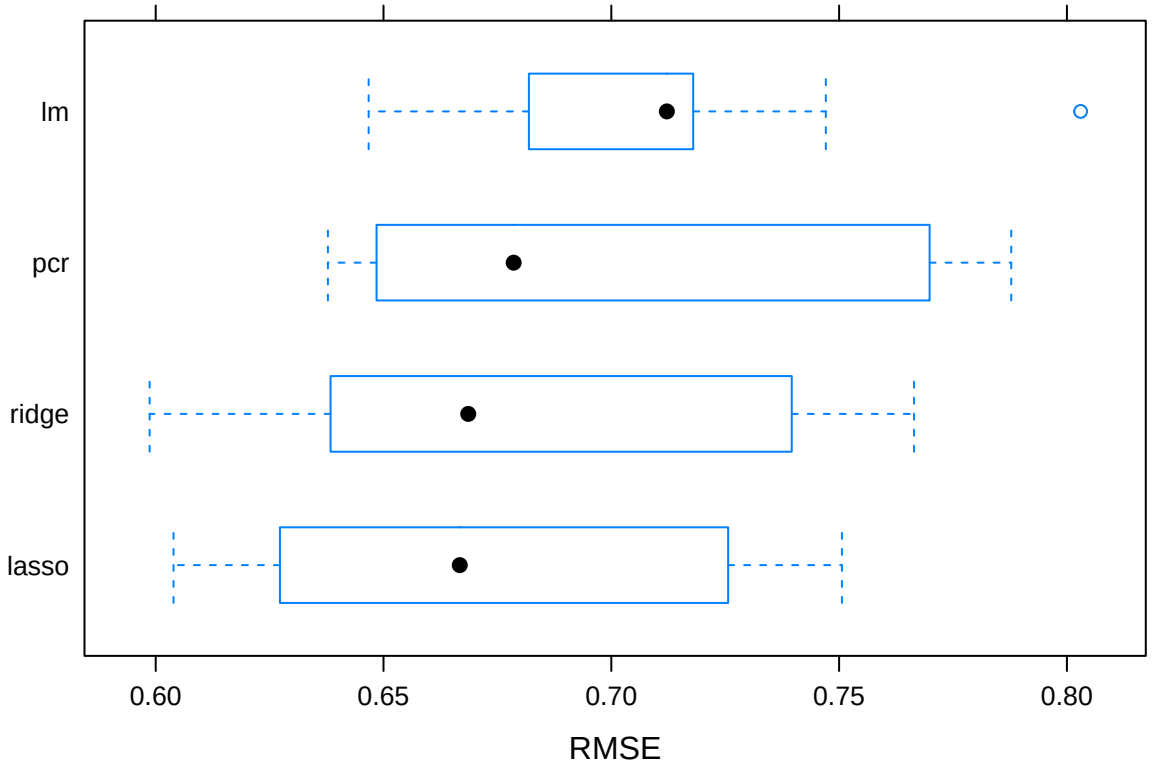
The chosen M is 152 and the test error is 0.5477905

Part(e)

```
#choose a model to predict solubility
resamp <- resamples(list(lasso = lasso.fit, ridge = ridge.fit, lm = lm.fit, pcr = pcr.fit))
summary(resamp)
```

```
##
## Call:
## summary.resamples(object = resamp)
##
## Models: lasso, ridge, lm, pcr
## Number of resamples: 10
##
## MAE
##            Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
## lasso 0.4745132 0.4912323 0.5080393 0.5185207 0.5510879 0.5687773    0
## ridge 0.4660475 0.4953502 0.5167305 0.5225447 0.5556487 0.5830117    0
## lm    0.4787720 0.5028170 0.5332078 0.5281167 0.5509856 0.5859704    0
## pcr   0.4867494 0.5017483 0.5271328 0.5398784 0.5736772 0.6125581    0
##
## RMSE
##            Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
## lasso 0.6039065 0.6338959 0.6667391 0.6765875 0.7231725 0.7506134    0
## ridge 0.5986715 0.6425918 0.6686019 0.6843122 0.7365268 0.7664399    0
## lm    0.6467371 0.6850769 0.7121902 0.7080065 0.7178712 0.8030063    0
## pcr   0.6377952 0.6528636 0.6785692 0.7034588 0.7636394 0.7877797    0
##
## Rsquared
##            Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
## lasso 0.8697192 0.8811758 0.8927381 0.8918467 0.8996490 0.9215221    0
## ridge 0.8632437 0.8783195 0.8901298 0.8891165 0.9006366 0.9187341    0
## lm    0.8600259 0.8770213 0.8871223 0.8841123 0.8893032 0.9052887    0
## pcr   0.8500621 0.8753012 0.8806902 0.8833410 0.8982732 0.9152954    0
```

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



I will choose the Lasso model to predict solubility because it has the lowest mean and median RMSE among all the above models.