

P8106 Homework 1

Pengyuan Su (ps3195)

2/9/2021

```
# Import data

sol_train =
  read_csv("./data/solubility_train.csv") %>%
  janitor::clean_names()

sol_test =
  read_csv("./data/solubility_test.csv") %>%
  janitor::clean_names()

x_train = model.matrix(solubility ~., sol_train)[, -1]
y_train = sol_train$solubility

x_test = model.matrix(solubility ~., sol_test)[,-1]
y_test = sol_test$solubility

ctr <- trainControl(method = "repeatedcv", number = 10, repeats = 5)
```

Question 1

```
set.seed(5)
fit.lm =
  train(
    solubility ~.,
    data = sol_train,
    method = "lm",
    trControl = ctr
  )

RMSE(predict(fit.lm, newdata = sol_test), sol_test$solubility)^2
```

```
## [1] 0.556
```

Question 2

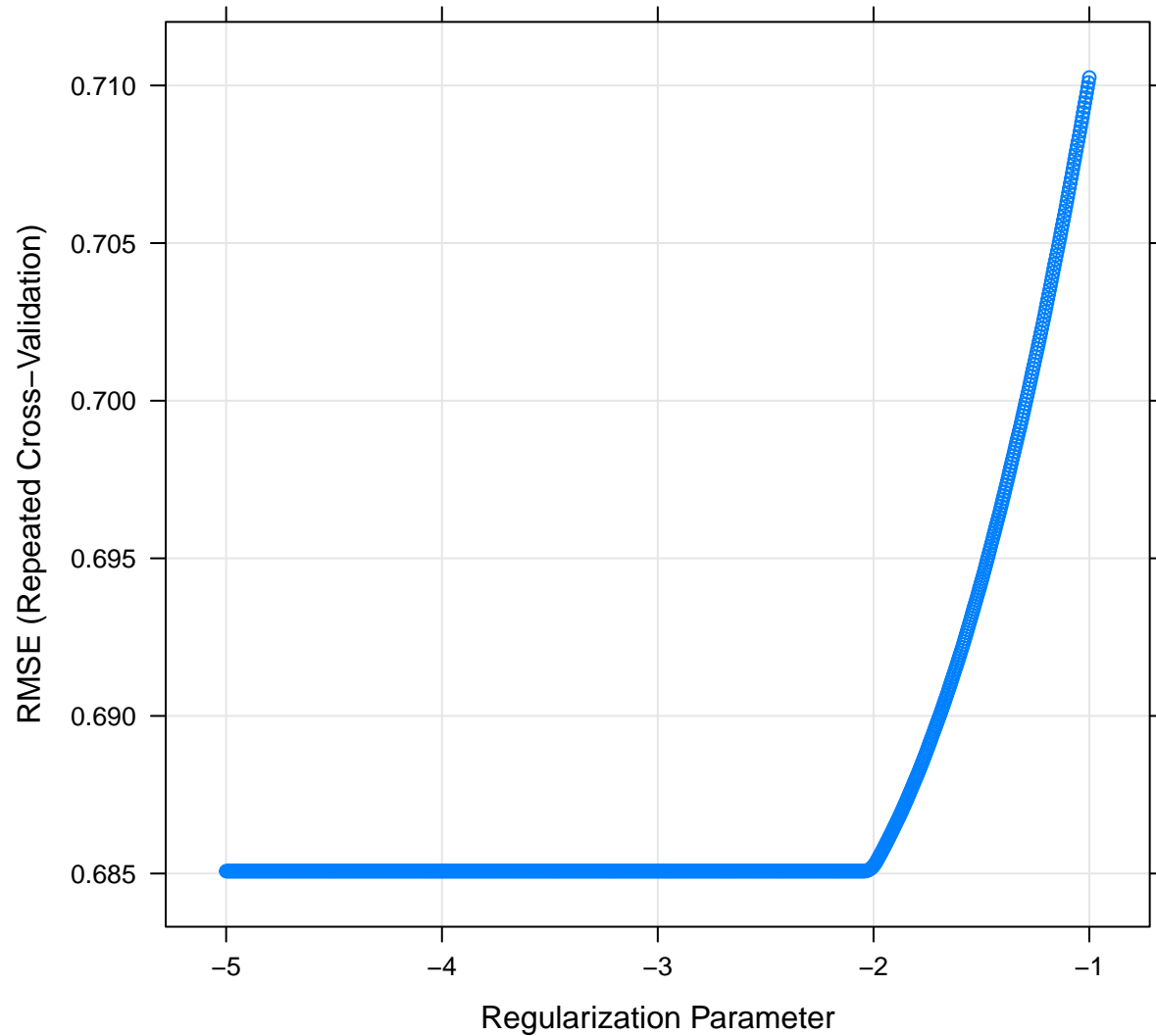
```

set.seed(5)

fit.ridge =
  train(
    solubility ~.,
    data = sol_train,
    method = "glmnet",
    tuneGrid =
      expand.grid(
        alpha = 0,
        lambda = exp(seq(from = -1, to = -5, length = 1000))
      ),
    trControl = ctr,
    preProcess = c("center", "scale")
  )

plot(fit.ridge, xTrans = log)

```



```
fit.ridge$bestTune
```

```
##      alpha lambda
## 740      0    0.13
```

```
RMSE(predict(fit.ridge, s = "lamda.min", newdata = sol_test), sol_test$solubility)^2
```

```
## [1] 0.513
```

Question 3

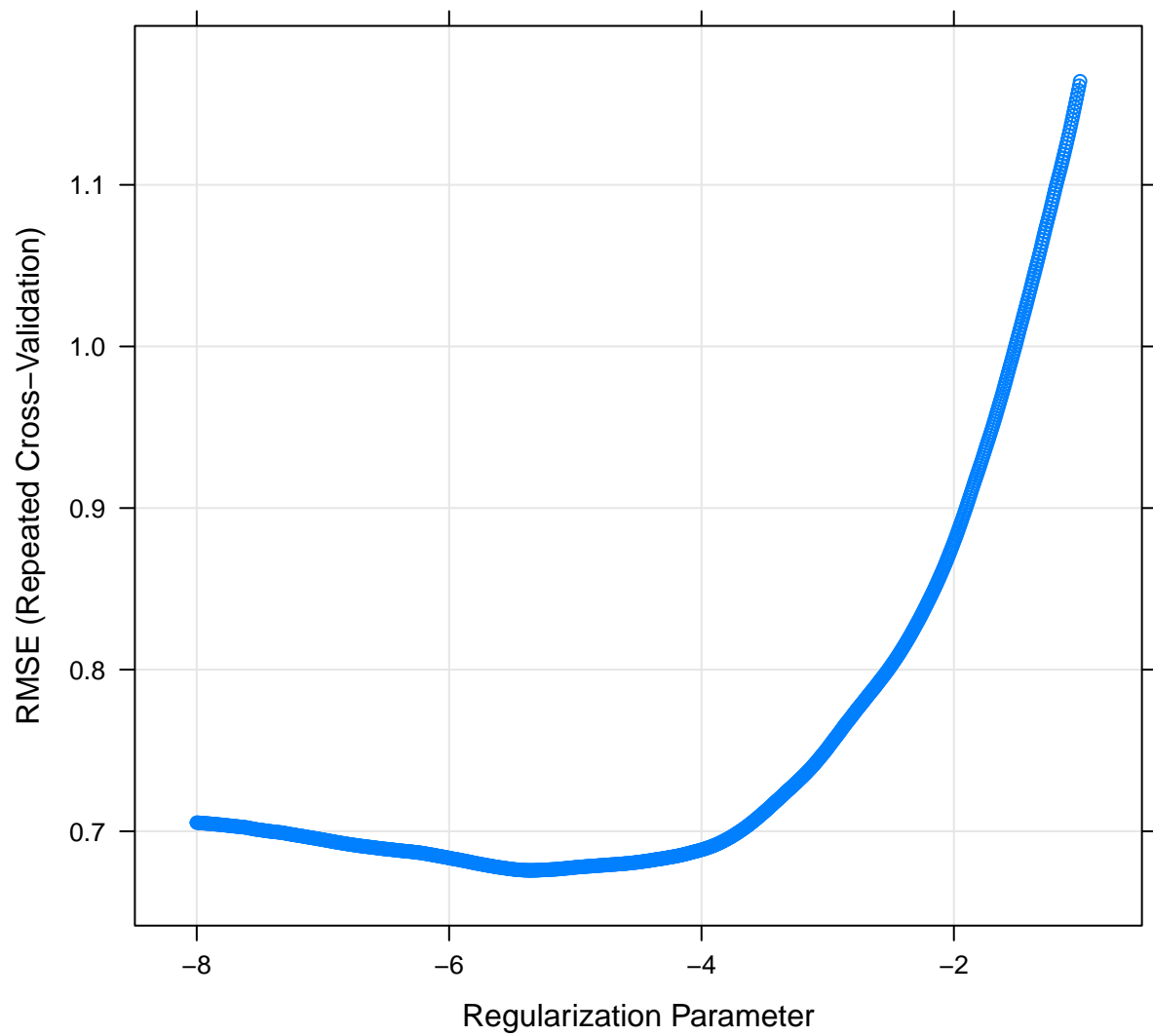
```
set.seed(5)
fit.lasso =
```

```

train(
  solubility ~.,
  data = sol_train,
  method = "glmnet",
  tuneGrid =
    expand.grid(
      alpha = 1,
      lambda = exp(seq(from = -1, to = -8, length = 1000))
    ),
  trControl = ctr,
  preProcess = c("center", "scale")
)

plot(fit.lasso, xTrans = log)

```



```
fit.lasso$bestTune
```

```
##      alpha  lambda  
## 378      1 0.00471
```

```
RMSE(predict(fit.lasso, s = "lambda.min", newdata = sol_test), sol_test$solubility)^2
```

```
## [1] 0.498
```

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

```
## [1] 144
```

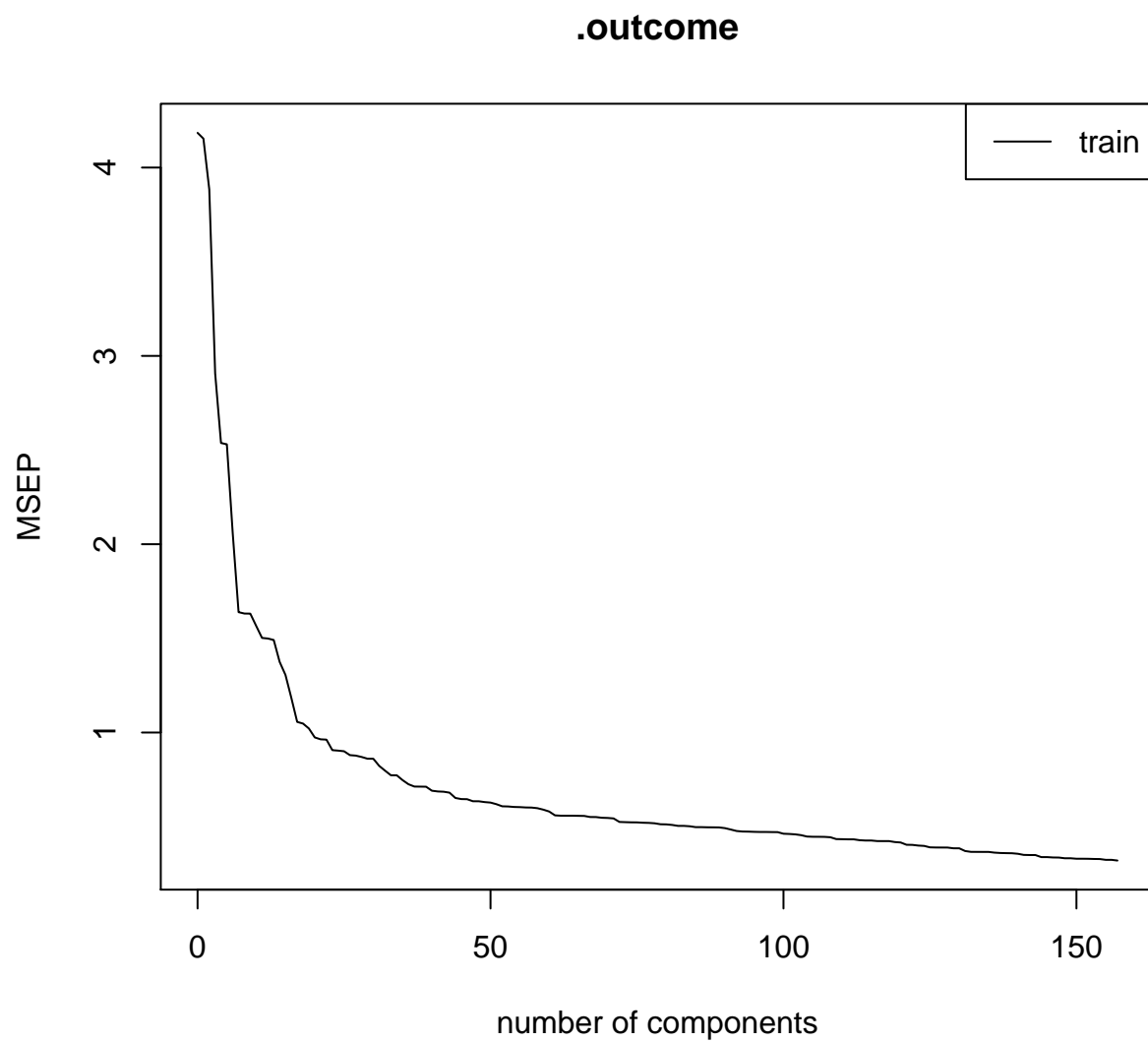
Question 4

```
set.seed(5)  
fit.pcr =  
  train(  
    solubility~.,  
    data = sol_train,  
    method = "pcr",  
    tuneGrid =  
      expand.grid(ncomp = seq(1,ncol(sol_train))),  
    preProcess = c("center","scale"),  
    trControl = ctr  
  )
```

```
fit.pcr$bestTune #Find the best M
```

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

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



```
RMSE(predict(fit.pcr, x_test), y_test)^2 # Calculate MSE
```

```
## [1] 0.55
```

```
mean((predict(fit.pcr, x_test) - y_test)^2) # Check
```

```
## [1] 0.55
```

Question 5

```
resample =  
  resamples(list(  
    lm = fit.lm,
```

```

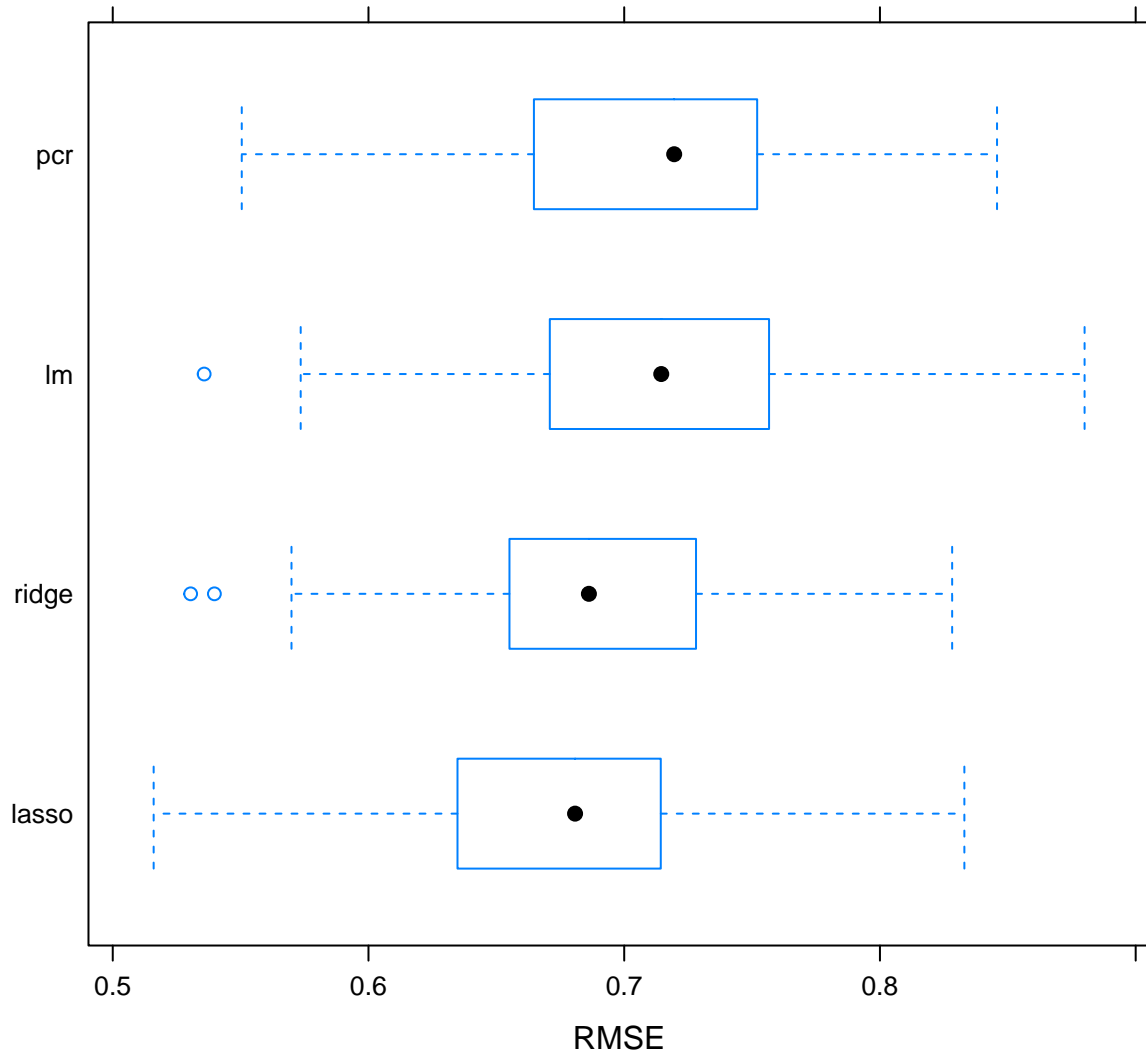
    ridge = fit.ridge,
    lasso = fit.lasso,
    pcr = fit.pcr
  ))

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      0.401   0.499   0.536 0.530   0.564 0.607    0
## ridge   0.412   0.503   0.530 0.522   0.551 0.607    0
## lasso   0.416   0.498   0.518 0.518   0.544 0.605    0
## pcr     0.420   0.519   0.549 0.545   0.583 0.645    0
##
## RMSE
##      Min. 1st Qu. Median  Mean 3rd Qu.  Max. NA's
## lm      0.536   0.671   0.714 0.708   0.754 0.880    0
## ridge   0.530   0.656   0.686 0.685   0.727 0.828    0
## lasso   0.516   0.636   0.681 0.676   0.714 0.833    0
## pcr     0.550   0.666   0.720 0.709   0.752 0.846    0
##
## Rsquared
##      Min. 1st Qu. Median  Mean 3rd Qu.  Max. NA's
## lm      0.839   0.865   0.876 0.882   0.897 0.939    0
## ridge   0.821   0.871   0.885 0.887   0.905 0.943    0
## lasso   0.818   0.873   0.889 0.890   0.908 0.944    0
## pcr     0.821   0.858   0.880 0.880   0.903 0.937    0

bwplot(resample, metric = "RMSE")

```



From numeric result and boxplot, Lasso model has the smallest RMSE, MAE, and the highest R^2 , hence we will choose it for predicting solubility.