

Homework_1

Jeffrey LIANG

2/4/2021

```
set.seed(123123)
sl_tr =
  read_csv(here::here("solubility_train.csv")) %>%
  janitor::clean_names()

sl_ts =
  read_csv(here::here("solubility_test.csv")) %>%
  janitor::clean_names()

x_ts = model.matrix(solubility ~ ., sl_ts)[, -1]
y_ts = sl_ts$solubility
```

Q1

```
sl_lm =
  train(solubility~.,
        data = sl_tr,
        method = "lm",
        trControl =
          trainControl(
            method = "repeatedcv",
            number = 10,
            repeats = 5
          ))

RMSE(predict(sl_lm,newdata = sl_ts),sl_ts$solubility)
```

```
## [1] 0.746
```

Q2

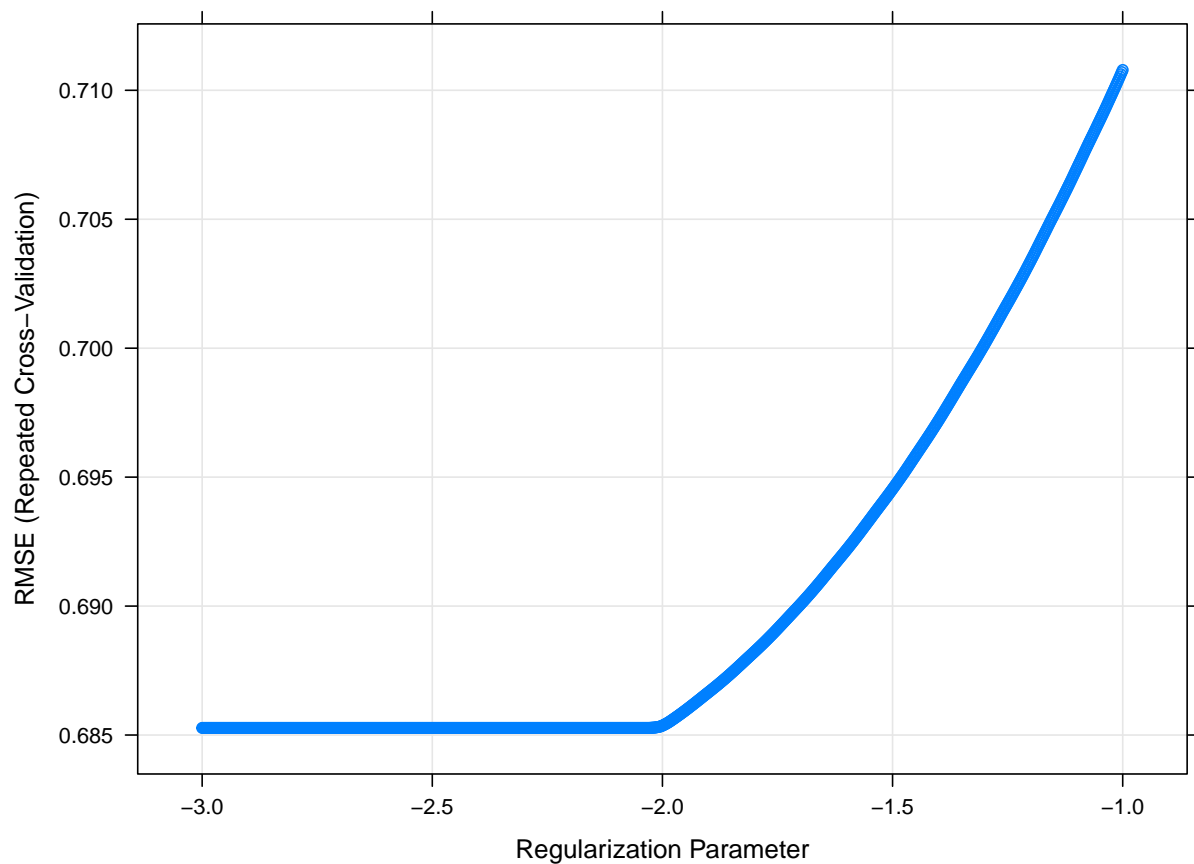
```
sl_ridge =
  train(
    solubility ~ .,
```

```

data = sl_tr,
method = "glmnet",
tuneGrid =
  expand.grid(alpha = 0,
             lambda = exp(seq(from = -1,to = -3, length = 1000))),
trControl =
  trainControl(method = "repeatedcv",
              number = 10,
              repeats = 5),
preProcess = c("center","scale")
)

plot(sl_ridge, xTrans = log)

```



```
sl_ridge$bestTune
```

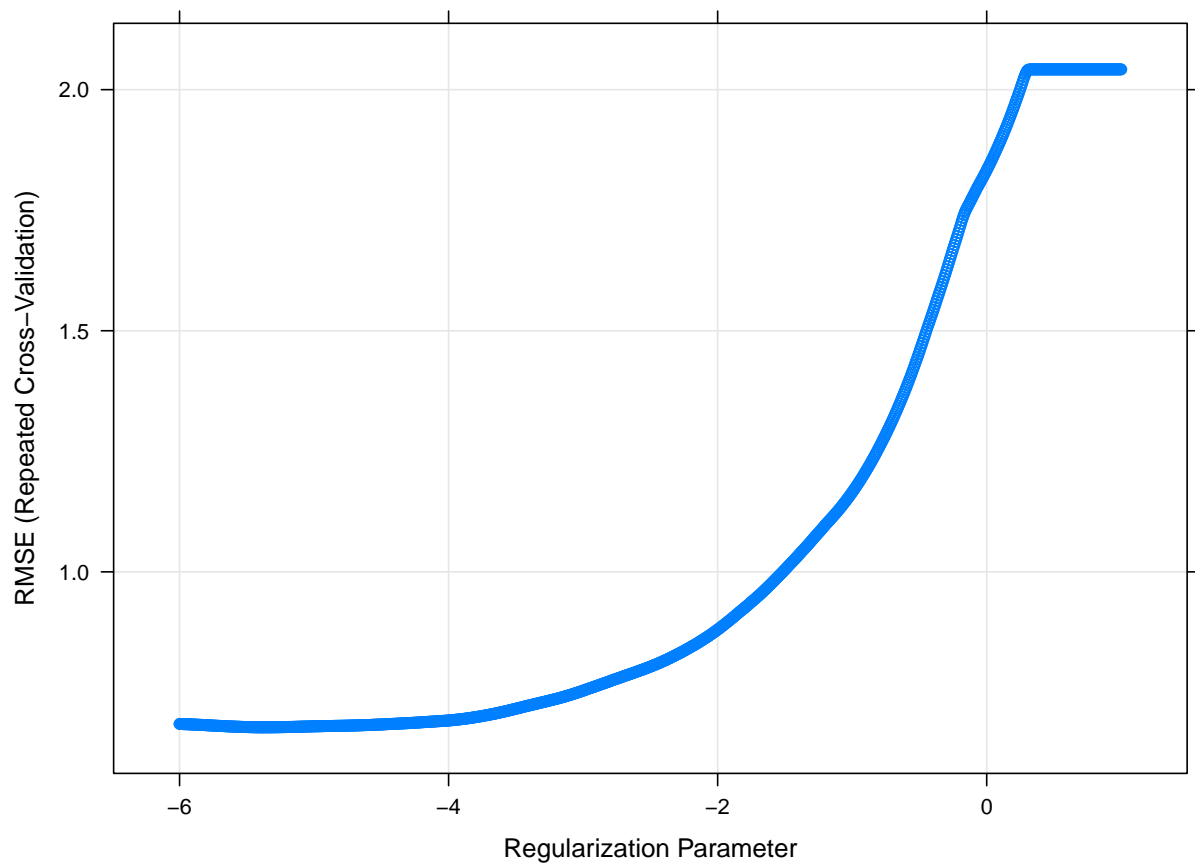
```
##      alpha lambda
## 485      0  0.131
```

```
RMSE(predict(sl_ridge,s="lambda.min",newx = sl_ts),sl_ts$solubility)
```

```
## [1] 2.93
```

Q3

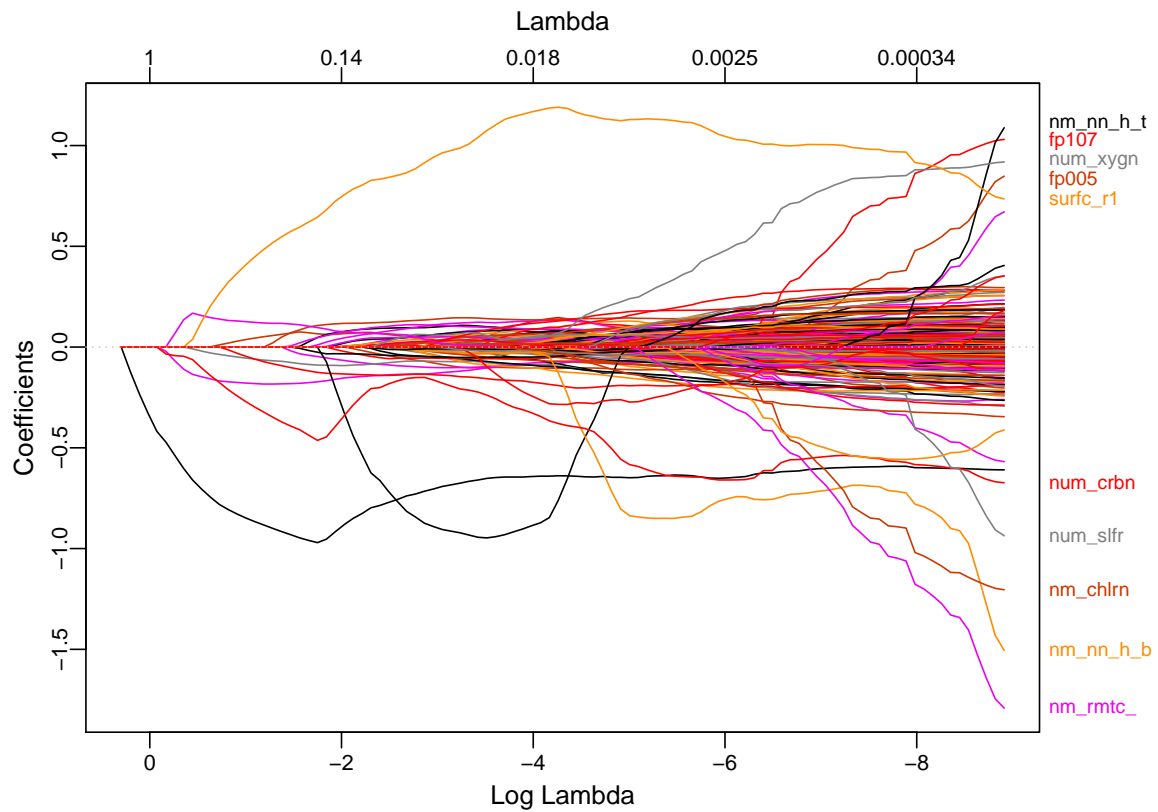
```
sl_lasso =  
  train(  
    solubility~.,  
    data = sl_tr,  
    method = "glmnet",  
    tuneGrid =  
      expand.grid(alpha = 1,  
                  lambda = exp(seq(-6,to=1,length = 1000))),  
    trControl =  
      trainControl(  
        method = "repeatedcv",  
        number = 10,  
        repeats = 5  
      ),  
    preProcess =  
      c("center","scale")  
  )  
  
plot(sl_lasso,xTrans = log)
```



```
sl_lasso$bestTune
```

```
##      alpha  lambda  
## 89      1 0.00459
```

```
plotmo::plot_glmnet(  
  sl_lasso$finalModel  
)
```



```
sum(coef(sl_lasso$finalModel, s = sl_lasso$bestTune$lambda) != 0)
```

```
## [1] 148
```

```
RMSE(predict(sl_lasso,s="lambda.min",newx = sl_ts),sl_ts$solubility)
```

```
## [1] 2.95
```

Q4

```

sl_pcr =
  train(
    solubility~.,
    data = sl_tr,
    method = "pcr",
    tuneGrid =
      expand.grid(ncomp = seq(1,ncol(sl_tr))),
    preProcess = c("center","scale"),
    trControl =
      trainControl(
        method = "repeatedcv",
        number = 10,
        repeats = 5
      )
  )

sl_pcr$bestTune

```

```

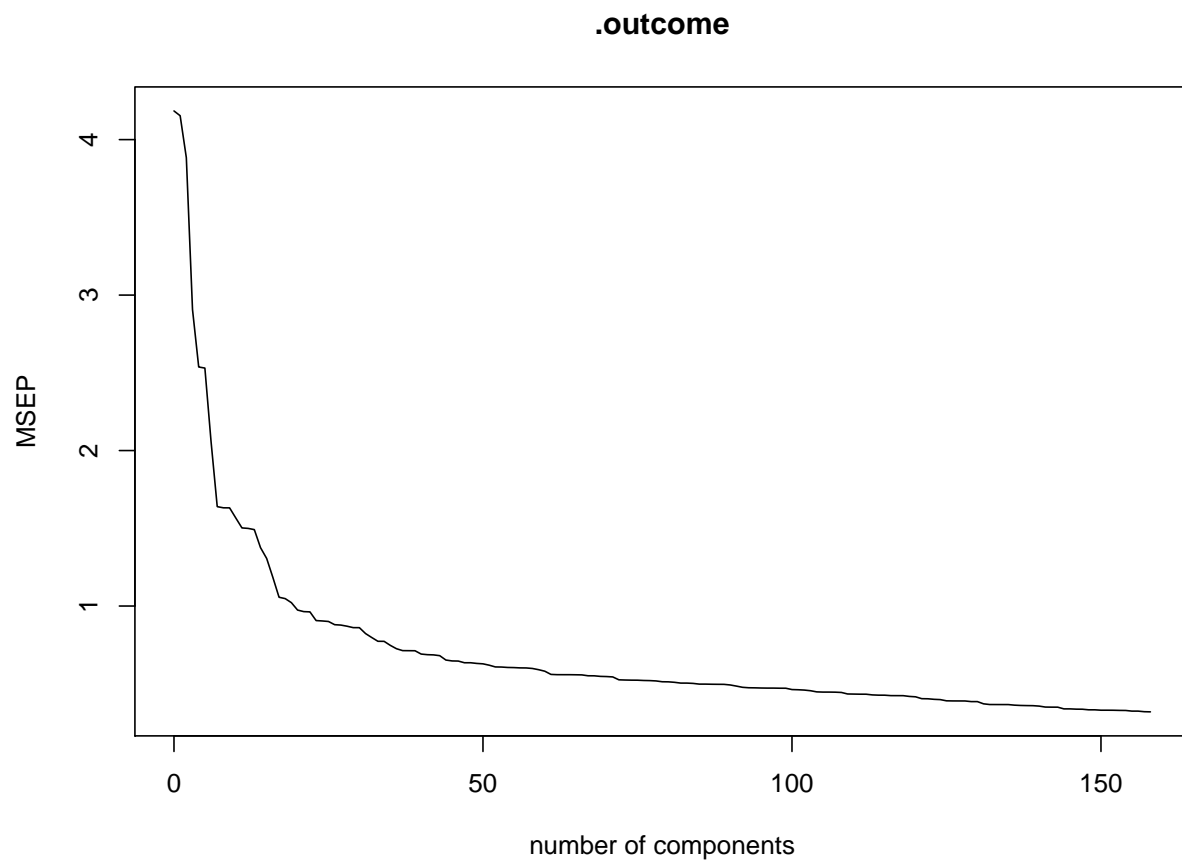
##      ncomp
## 158    158

```

```

validationplot(sl_pcr$finalModel, val.type = "MSEP")

```



```
RMSE(predict(sl_pcr,x_ts),y_ts)
```

```
## [1] 0.741
```

```
mean((predict(sl_pcr,x_ts)-y_ts)^2)
```

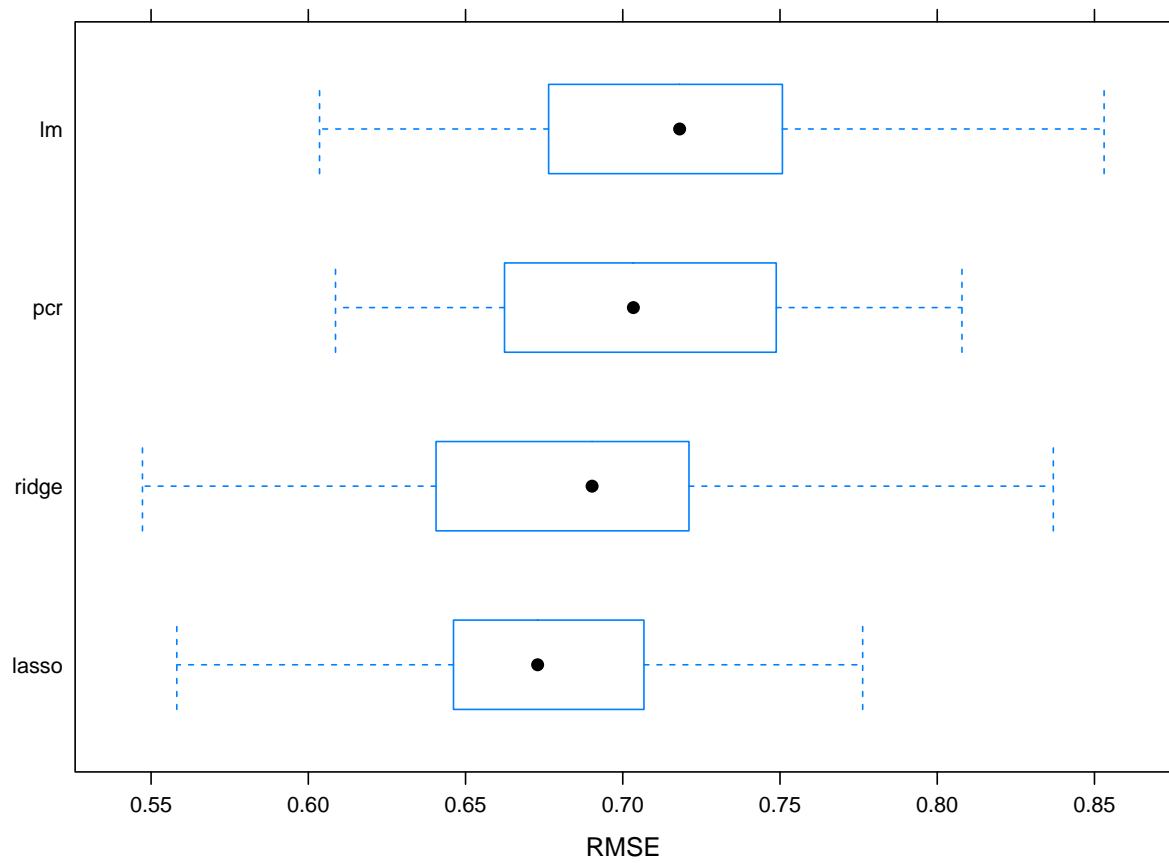
```
## [1] 0.549
```

```
resam =  
  resamples(list(  
    lm = sl_lm,  
    ridge = sl_ridge,  
    lasso = sl_lasso,  
    pcr = sl_pcr  
  ))
```

```
summary(resam)
```

```
##  
## Call:  
## summary.resamples(object = resam)  
##  
## Models: lm, ridge, lasso, pcr  
## Number of resamples: 50  
##  
## MAE  
##      Min. 1st Qu. Median  Mean 3rd Qu.  Max. NA's  
## lm      0.446   0.508  0.538 0.533   0.556 0.635    0  
## ridge   0.426   0.491  0.522 0.522   0.557 0.640    0  
## lasso   0.444   0.498  0.512 0.519   0.543 0.606    0  
## pcr     0.445   0.514  0.541 0.544   0.574 0.643    0  
##  
## RMSE  
##      Min. 1st Qu. Median  Mean 3rd Qu.  Max. NA's  
## lm      0.604   0.678  0.718 0.716   0.750 0.853    0  
## ridge   0.547   0.641  0.690 0.685   0.721 0.837    0  
## lasso   0.558   0.647  0.673 0.678   0.706 0.776    0  
## pcr     0.609   0.663  0.703 0.707   0.748 0.808    0  
##  
## Rsquared  
##      Min. 1st Qu. Median  Mean 3rd Qu.  Max. NA's  
## lm      0.798   0.859  0.880 0.878   0.898 0.924    0  
## ridge   0.822   0.869  0.886 0.887   0.912 0.932    0  
## lasso   0.850   0.876  0.893 0.891   0.903 0.929    0  
## pcr     0.816   0.863  0.886 0.881   0.900 0.921    0
```

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



By the resampling result, we can see that Lasso method has the best/lowest mean RMSE and MAE, as well as the highest R^2 . Lasso will be chosen to use out of the four for its predictivity and goodness of fit.