HW1

Hanyu Lu hl3439

```
train <- na.omit(train)
test2 <- model.matrix(Solubility ~ ., test)[ ,-1]
train2 <- model.matrix(Solubility ~ ., train)[ ,-1]
set.seed(1)
y_test <- test$Solubility
y_train <- train$Solubility
```

a

```
linear.fit = lm(Solubility ~., data = train)

pred_linear = predict(linear.fit, test)
mean((pred_linear - y_test)^2)
```

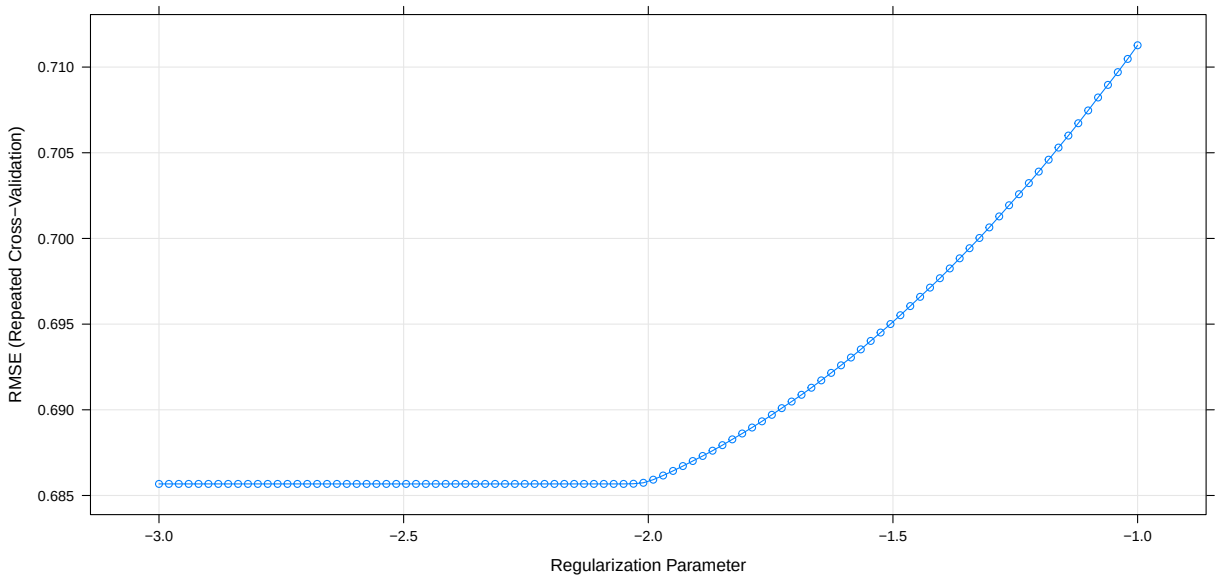
```
## [1] 0.5558898
```

The mean squared error is 0.5558898.

b

```
ctrl1 <- trainControl(method = "repeatedcv", number = 10, repeats = 5) # you can try other options
set.seed(2)
ridge.fit <- train(train2, y_train,
method = "glmnet",
tuneGrid = expand.grid(alpha = 0,
lambda = exp(seq(-1, -3, length=100))), # preProc = c("center", "scale"),
trControl = ctrl1)

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




```
ridge.fit$bestTune
```

```
##    alpha    lambda
## 48     0 0.1286699
```

```
pred_ridge <- predict(ridge.fit, newdata = test)
# test error
mean((pred_ridge - y_test)^2)
```

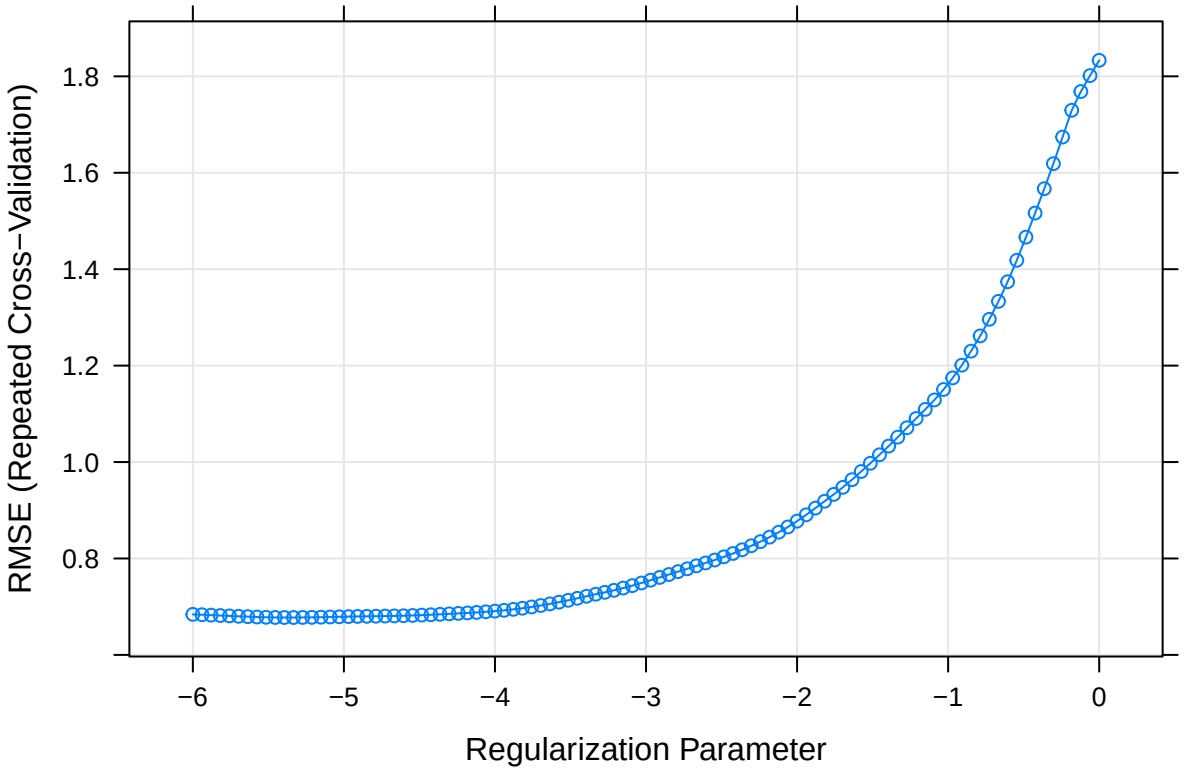
```
## [1] 0.5134603
```

The test error is 0.5134603.

c

```
set.seed(2)
lasso.fit <- train(train2, y_train,
method = "glmnet",
tuneGrid = expand.grid(alpha = 1,lambda = exp(seq(-6, 0, length=100))),
trControl = ctrl1)

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




```
lasso.fit$bestTune
```

```
##    alpha     lambda
## 12     1 0.00482795
```

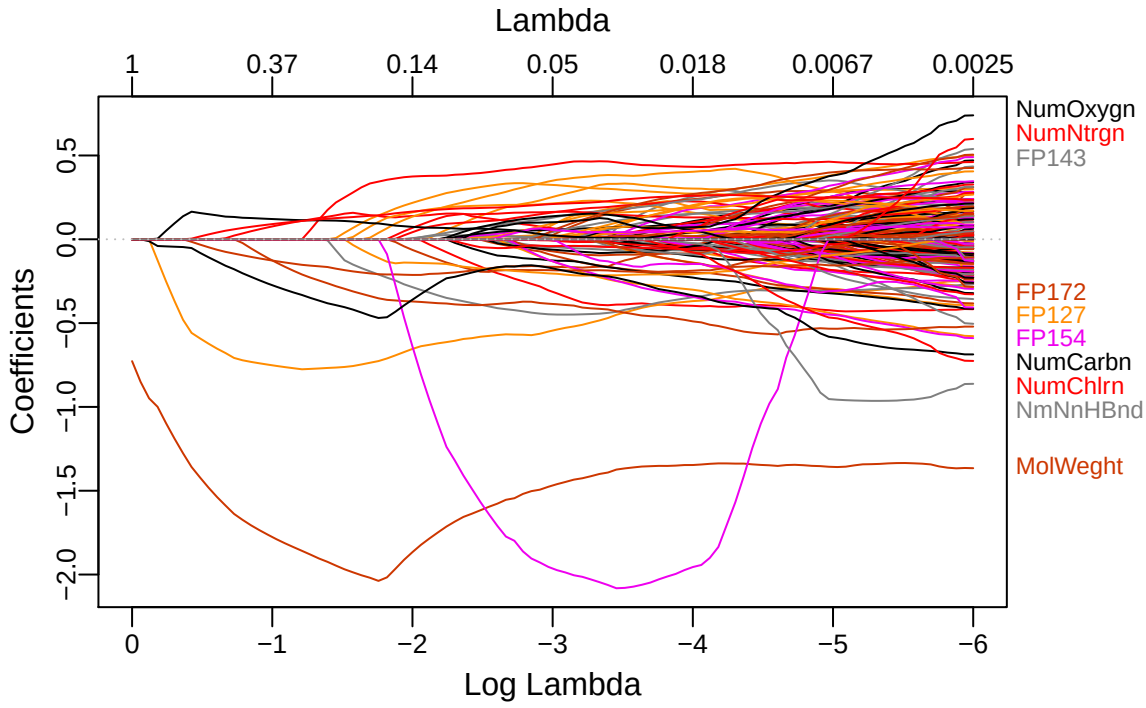
```
pred_lasso <- predict(lasso.fit, newdata = test)
# test error
mean((pred_lasso - y_test)^2)
```

```
## [1] 0.4976488
```

The test error is 0.4976488.

```
set.seed(2)
cv.lasso <- cv.glmnet(train2, y_train, alpha = 1,
lambda = exp(seq(-6, 0, length=100)))

plot_glmnet(cv.lasso$glmnet.fit)
```




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

There are 140 non-zero coefficients.

d

```
set.seed(2)
pcr.mod <- pcr(Solubility ~ .,
data = train,
scale = TRUE, # scale = FALSE by default
validation = "CV")
```

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

```
## 157 comps
##       157
```

```r
predy2.pcr <- predict(pcr.mod, newdata = test, ncomp = ncomp.cv)
# test MSE
mean((y_test - predy2.pcr)^2)
```

```
## [1] 0.549917
```

The test error is 0.549917 and the value of M selected by cross-validation is 157.

e

Since Lasso has the least RMSE and the model is regularized, we choose Lasso as our model for estimation.