Homework 1

Jingyi Zhang

2/6/21

Load dataset

```
sol_test <-
  read_csv("./data/solubility_test.csv") %>%
  janitor::clean_names()

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

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

(a) Fit a linear model using least squares on the training data and calculate the mean squared error using the test data.

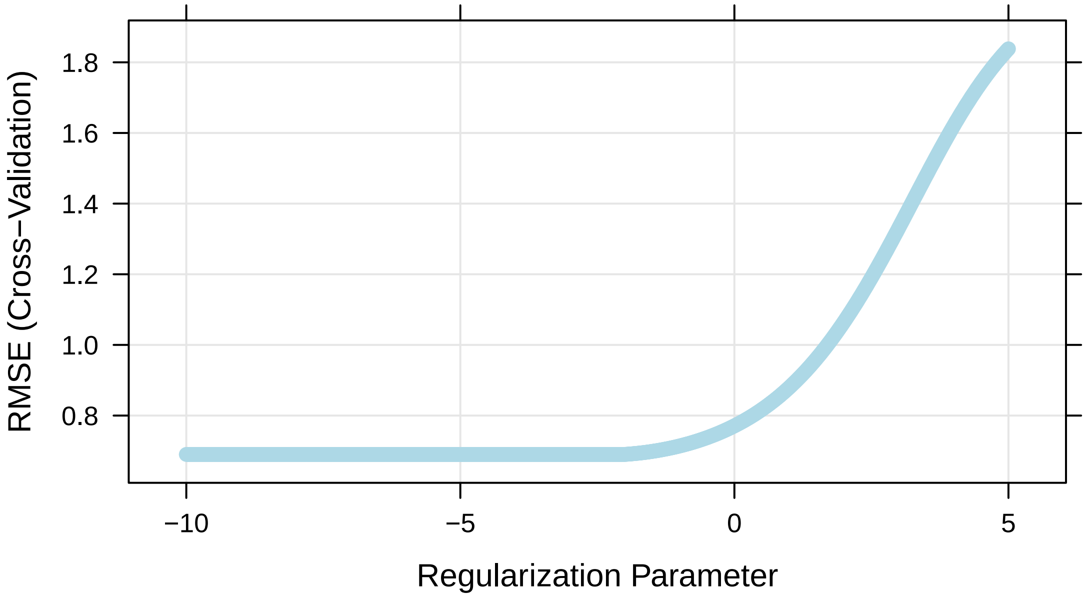
```
set.seed(36)
ctrl1 <- trainControl(method = "cv", number = 10)
fit_lm <- train(solubility~.,
                data = sol_train,
                method = "lm",
                trControl = ctrl1)
RMSE(predict(fit_lm, newdata = sol_test), sol_test$solubility)
```

```
## [1] 0.7455802
```

(b) Fit a ridge regression model on the training data, with lambda chosen by cross-validation. Report the test error.

```
set.seed(7)
ridge.fit <-
  train(solubility~.,
        data = sol_train,
        method = "glmnet",
        tuneGrid =
          expand.grid(alpha = 0,
                      lambda = exp(seq(5, -10, length = 1000))),
        trControl = ctrl1,
        preProc = c("center", "scale"))

plot(ridge.fit, xTrans = log, pch = 1, col = "light blue")
```



```
ridge.fit$bestTune
```

```
##     alpha    lambda
## 532     0 0.1317266
```

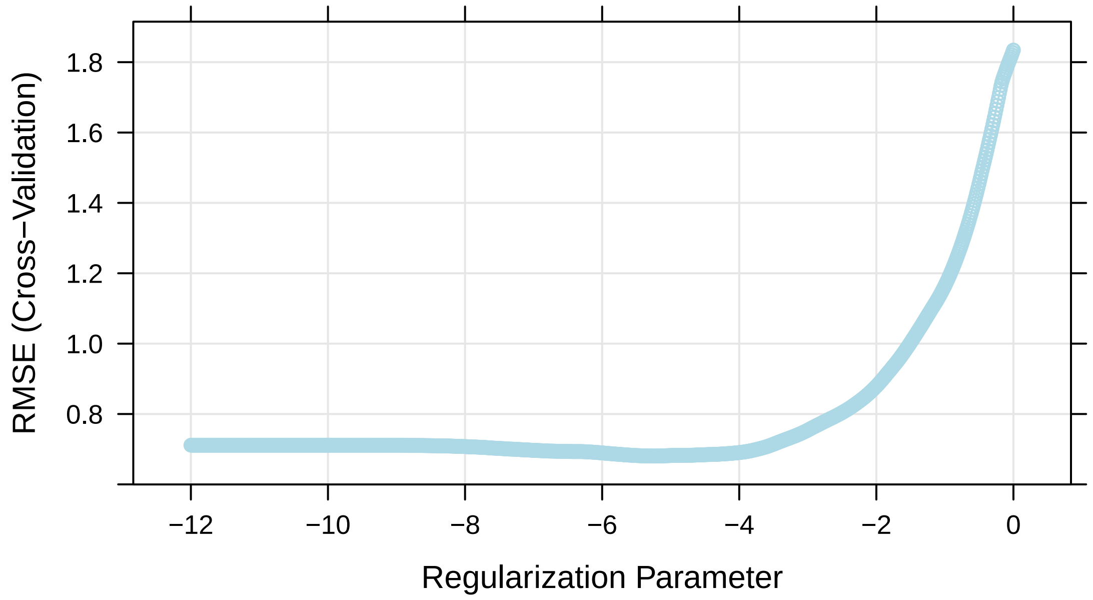
```
RMSE(predict(ridge.fit, s = "lambda.min", newx = sol_test), sol_test$solubility)
```

```
## [1] 2.929737
```

(c) Fit a lasso model on the training data, with lambda chosen by cross-validation. Report the test error and the number of non-zero coefficient estimates in your model.

```
set.seed(7)
lasso.fit <-
  train(solubility~.,
        data = sol_train,
        method = "glmnet",
        tuneGrid =
          expand.grid(alpha = 1,
                      lambda = exp(seq(0, -12, length = 1000))),
        trControl = ctrl1,
        preProc = c("center", "scale"))

plot(lasso.fit, xTrans = log, col = "light blue", pch = 1)
```



```
lasso.fit$bestTune
```

```
##     alpha      lambda
## 565     1 0.005379148
```

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

```
## [1] 2.945769
```

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

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

(d) Fit a principle component regression model on the training data, with M chosen by cross-validation. Report the test error and the value of M selected by cross-validation.

(e) Which model will you choose for predicting solubility?