Problem Set 2

Seth Harrison

02/03/2020

```
# Load Packages/Data

library(tidyverse)
library(readr)
library(rsample)
library(broom)
library(rcfss)
library(ISLR)
library(yardstick)
library(caret)
library(randomForest)
library(pls)

nes <- read_csv("./nes2008.csv")
```

1. (10 points) Estimate the MSE of the model using the traditional approach. That is, fit the linear regression model using the *entire* dataset and calculate the mean squared error for the *entire* dataset. Present and discuss your results at a simple, high level.

```
# Fit Linear Model

Model1 <- lm(nes$biden~nes$female+nes$age+nes$educ+nes$dem+nes$rep)
summary(Model1)
```

```
##
## Call:
## lm(formula = nes$biden ~ nes$female + nes$age + nes$educ + nes$dem +
##     nes$rep)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -75.546 -11.295   1.018  12.776  53.977
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  58.81126    3.12444  18.823  < 2e-16 ***
## nes$female    4.10323    0.94823   4.327 1.59e-05 ***
## nes$age       0.04826    0.02825   1.708   0.0877 .
## nes$educ     -0.34533    0.19478  -1.773   0.0764 .
## nes$dem      15.42426    1.06803  14.442  < 2e-16 ***
## nes$rep     -15.84951    1.31136 -12.086  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.91 on 1801 degrees of freedom
```

```
## Multiple R-squared:  0.2815, Adjusted R-squared:  0.2795
## F-statistic: 141.1 on 5 and 1801 DF,  p-value: < 2.2e-16
```

```r
# Calculate MSE

mse1 <- mean(Model1$residuals^2)
mse1
```

```
## [1] 395.2702
```

The mean squared error (MSE) first eliminates negative directionality by squaring the residuals, or the distance betweened the observed observation and observation expected by the regression line. It then takes arithmatic mean of those products. One easy way to intepret MSE is to convert it back to a mean error by taking the square root of the MSE. In this case, the root mean squared error (RMSE) is about 20, meaning that, on average, the expected thermometer rating was about 20 units off (either higher or lower) from the observed rating.

2. (30 points) Calculate the test MSE of the model using the simple holdout validation approach.

```
- (5 points) Split the sample set into a training set (50%) and a holdout set (50%).
```

```r
# Split nes into test and train

set.seed(1)

nes_split <- initial_split(data = nes,
                           prop = 0.5)

nes_train <- training(nes_split)
nes_test <- testing(nes_split)
```

```
- (5 points) Fit the linear regression model using only the training observations.
```

```r
# Fit Linear Model Using nes_train

Model2 <- lm(biden~female+age+educ+dem+rep, data = nes_train)
      mse2  <- mean(Model2$residuals^2)
summary(Model2)
```

```
##
## Call:
## lm(formula = biden ~ female + age + educ + dem + rep, data = nes_train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -75.875 -10.974   0.638  13.968  45.989
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 61.94663    4.52928  13.677  < 2e-16 ***
## female       5.14561    1.38493   3.715 0.000215 ***
## age         -0.02402    0.04197  -0.572 0.567281
## educ        -0.46983    0.28126  -1.670 0.095179 .
## dem         16.27265    1.55652  10.454  < 2e-16 ***
```

```
## rep         -16.41671    1.96592  -8.351 2.55e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.74 on 898 degrees of freedom
## Multiple R-squared:  0.2799, Adjusted R-squared:  0.2759
## F-statistic:  69.8 on 5 and 898 DF,  p-value: < 2.2e-16
```

```
- (10 points) Calculate the MSE using only the test set observations.
```

```
# Calculate MSE on nse_test

mse3 <- augment(Model2, newdata = nes_test) %>%
                  mse(truth = biden, estimate = .fitted)

mse3$.estimate
```

```
## [1] 370.1792
```

```
- (10 points) How does this value compare to the training MSE from question 1?
Present numeric comparison and discuss a bit.
```

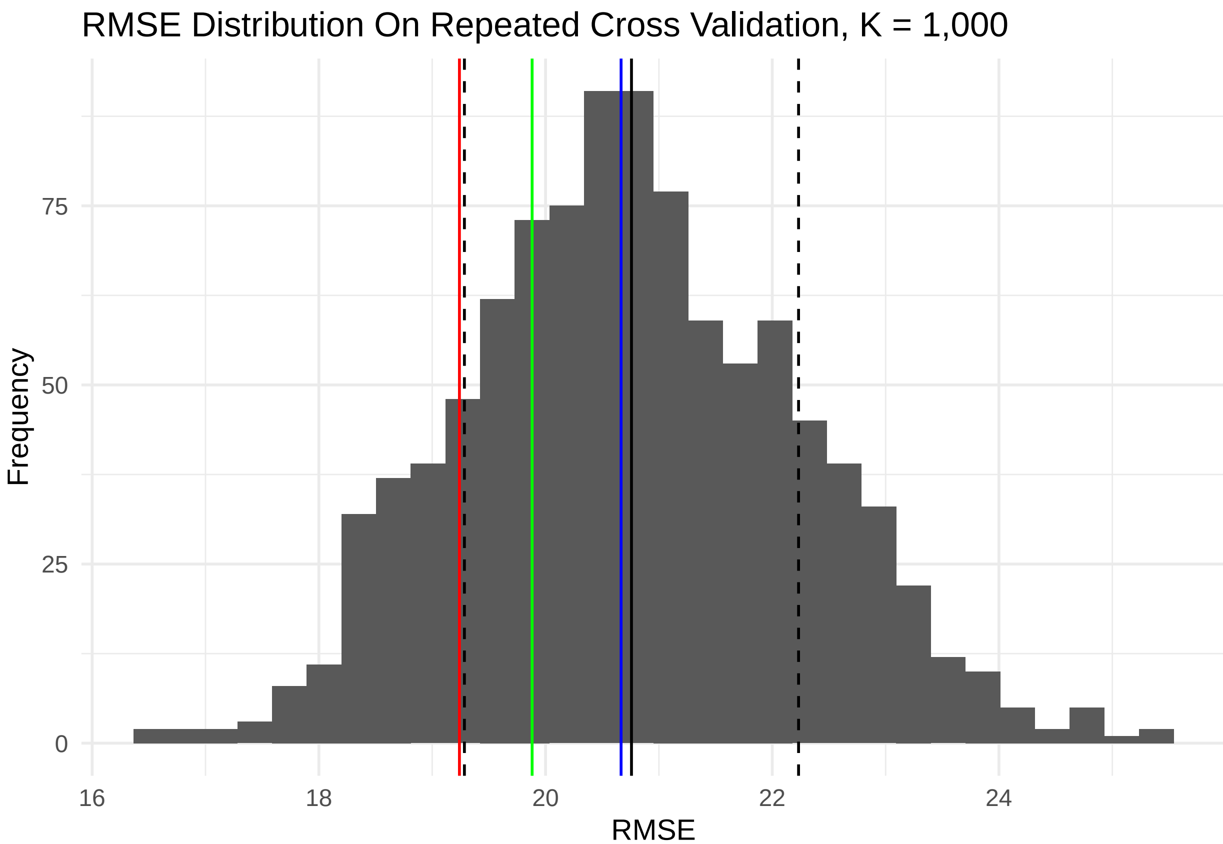
The first MSE was 395.27, while the second from the split data set was lower at 370.18. We avoided an optimistically biased, overfit evaluation by estimating the model from the training data and evaluating it on the test data but since the training and test sets were generated through random processes, we can expect high variance between MSE predictions dependent on the composition of the split.

3. (30 points) Repeat the simple validation set approach from the previous question 1000 times, using 1000 different splits of the observations into a training set and a test/validation set. Visualize your results as a sampling distribution (hint: think histogram or density plots). Comment on the results obtained.

```
#Use caret to estimate model 1000 times

ctrl <- trainControl(method = "repeatedcv", repeats = 100)

# Note repeats = 100 because default number is 10

Model3 <- train(biden~female+age+educ+dem+rep,
                data = nes_train,
                method = "pls",
                trControl = ctrl
                )
```

```
# Visualize distribution of RMSEs

ggplot(Model3$resample, aes(x = Model3$resample$RMSE)) +
      geom_histogram(bins = 30) +
      labs(title = "RMSE Distribution On Repeated Cross Validation, K = 1,000",
           x = "RMSE",
           y = "Frequency") +
      theme_minimal() +
      geom_vline(xintercept =  sqrt(mse1),color = "green") +
      geom_vline(xintercept = sqrt(mse2), color = "blue") +
```

```r
    geom_vline(xintercept = sqrt(mse3$.estimate), color = "red") +
    geom_vline(xintercept = mean(Model3$resample$RMSE), color = "black") +
    geom_vline(xintercept = mean(Model3$resample$RMSE) + sd(Model3$resample$RMSE),
               color = "black",
               linetype = "dashed") +
    geom_vline(xintercept = mean(Model3$resample$RMSE) - sd(Model3$resample$RMSE),
               color = "black",
               linetype = "dashed")
```



The RMSE distributions suggest some instability in our model. Only two out of the three RMSE estimated by our models (the green and blue lines) fell within one standard deviation of the mean RMSE. In the histogram this is represented by the two dashed black lines about the solid black line (the mean). The RMSEs appear to be (roughly) normally distributed, so we would expect that about 68% of the RMSEs would fall within that range. The RMSE generated from our second model on the test data (the red line) fell outside of that range, however, highlighting how unrepresentative statistics can be when estimated from single, randomly generated holdouts.