Intro to ML Pset 2

William Ansehl

2/3/2020

Problem 1

```r
nes_data <- read.csv("nes2008.csv")
biden_model <- lm(biden~female+age+educ+dem+rep, data = nes_data)
sm <- summary(biden_model)
mse <- mean(sm$residuals^2)
mse
```

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

```r
sm
```

```
##
## Call:
## lm(formula = biden ~ female + age + educ + dem + rep, data = nes_data)
##
## 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 ***
## female        4.10323    0.94823   4.327 1.59e-05 ***
## age           0.04826    0.02825   1.708   0.0877 .
## educ         -0.34533    0.19478  -1.773   0.0764 .
## dem          15.42426    1.06803  14.442  < 2e-16 ***
## 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
# From the given model, the significant variables (at the alpha = 0.05 confidence level)
#appears to be limited to female, dem and rep. At the alpha = 0.1 confidence level, all
#the regressors are significant. We note that education status as republican appear to
#have an inverse relationship with sentiment towards Biden while democrat and age appear
#to have a direct relationship with sentiment towards biden. That said, little variation
#in the data is accounted for, as described by an r-squared value equal to approximately 0.28.
#Furthermore, the MSE, an estimator of the fit of the regression line to the data,
#is approx 395.27. Considering the low r-squared and high mse, the model might be underfitting
```

```
#the data considerably.
```

Problem 2

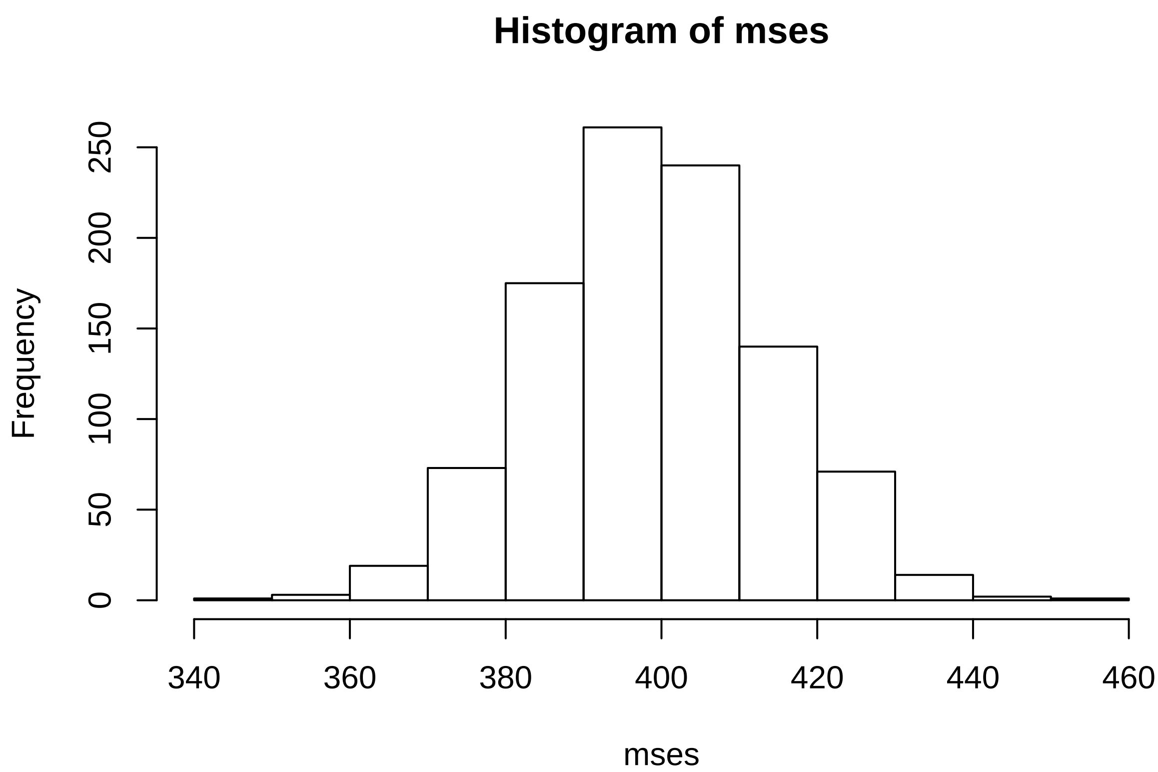
```
set.seed(5)
samples <- sample(1:nrow(nes_data),
                  nrow(nes_data)*0.5,
                  replace = FALSE)
train <- nes_data[samples, ]
test <- nes_data[-samples, ]

train_model <- lm(biden~female+age+educ+dem+rep, data = train)
predictions <- predict(train_model, newdata = test)
new_mse <- mean((test$biden - predictions)^2)
new_mse
```

```
## [1] 408.9851
```

```
# the new mse is approx. 408.99
# the new mse (408.99) is higher than the prior mse (395.27).
# This is to be expected since the first model was trained using the entire data set and thus will
#be more acurate than the second model which was trained using only half the data set and then
#evaluated on its predictions vs actual values for the test set.
```

Problem 3

```
mses <- c()
for (i in 1:1000) {
  set.seed(i)
  samples <- sample(1:nrow(nes_data),
                    nrow(nes_data)*0.5,
                    replace = FALSE)
  train <- nes_data[samples, ]
  test <- nes_data[-samples, ]
  train_model <- lm(biden~female+age+educ+dem+rep, data = train)
  predictions <- predict(train_model, newdata = test)
  new_mse <- mean((test$biden - predictions)^2)
  mses <- append(mses, new_mse, after=length(mses))
}
hist(mses)
```




```r
sd(mses)
```

```
## [1] 14.87322
```

```r
mean(mses)
```

```
## [1] 399.1602
```

```r
# the 1000 simulations seem to represent a standard distribution when they are represented
#as a histogram.
# the 1000 simulations have mean mse 399.1602 and standard deviation 14.87322.
# Therefore, we are 95% confidant that the true population mse under the holdout validation
#approach is in the range [384.287, 414.0334]
```

Problem 4

```r
library('dplyr')
```

```
##
## Attaching package: 'dplyr'
## The following objects are masked from 'package:stats':
##
##     filter, lag
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library('rsample')
```

```
## Loading required package: tidyr
```

```r
library(purrr)
library(tidyverse)
```

```
## -- Attaching packages ----------------------------------------------------------------------------------
## v ggplot2 3.2.1     v stringr 1.4.0
## v tibble  2.1.3     v forcats 0.4.0
## v readr   1.3.1
## -- Conflicts -------------------------------------------------------------------------------------- tid
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```r
library(tidyr)
lm_coefs <- function(splits, ...) {
  mod <- lm(..., data = analysis(splits))
  tidy(mod)
}
my_boot <- nes_data %>%
  bootstraps(1000) %>%
  mutate(coef = map(splits, lm_coefs, as.formula(biden~female+age+educ+dem+rep)))
my_boot %>%
  unnest(coef) %>%
  group_by(term) %>% summarize(.estimate = mean(estimate),
                              .se = sd(estimate, na.rm = TRUE))
```

```
## # A tibble: 6 x 3
##   term        .estimate    .se
##   <chr>           <dbl>  <dbl>
## 1 (Intercept)   58.7    3.11
## 2 age            0.0479 0.0291
## 3 dem           15.4    1.05
## 4 educ          -0.340  0.197
## 5 female         4.10   1.01
## 6 rep          -15.8    1.35
```

```r
# The produced results from the bootstrap methodology produce regression coefficients
#similar to the coefficients given in the full model, espcecially when compared to the
#full model's standard errors. That is to say, the difference between the regression
#coefficients for the bootstrap method and the full model's regression coefficients is
#not statistically significant. The standard errors between the two models are also similar.
#However, since the bootstrap method does not rely on assumptions on the distribution as
#seen in question 1, it's estimate for the population mse is more powerful.

# the holdout validation approach with 1000 simulations produced a mean mse = 399.1602 with standard
#deviation 14.87322.
# the original model had mse = 395.27.
# Under the 95% CI, the bootstrapped method contains the population mean and the difference between
#the two models MSE's are not statistically significant.

# Conceptual Motivation for Bootstrap Method:
# bootstrapping is a great methodology to utilize in order to better understand a population
#paramater(s) without needing to collect more data.
# It operates by randomly sampling from a sample in order to estimate these more complex
#paramaters and of the distribution. It also lets the statistician avoid costs associated with
#collecting more additional data.
```