Problem Set 2

Pete Cuppernull

1/30/2020

Load Packages and Clean

```
library(tidyverse)
library(modelr)
library(broom)
library(rsample)
library(patchwork)
library(corrplot)
library(ISLR)
library(caret)
library(rcfss)
library(yardstick)
library(stargazer)
biden <- read_csv("/Users/petecuppernull/Dropbox/UChicago/2019-20/Winter/Machine Learning/Repos/Problem

biden <- biden %>%
  select(biden, female, age, educ, dem, rep) %>%
  na.omit()
```

1. Estimate MSE

```
biden_traditional <- glm(biden ~ .,
                         data = biden)
biden_mse <- modelr::mse(biden_traditional, biden)

stargazer(biden_traditional, type = "text")
```

```
##
## =============================================
##                       Dependent variable:
##                   ---------------------------
##                              biden
## ---------------------------------------------
## female                     4.103***
##                             (0.948)
##
## age                         0.048*
##                             (0.028)
##
## educ                        -0.345*
##                             (0.195)
##
## dem                        15.424***
##                             (1.068)
```

```
##
## rep                         -15.850***
##                               (1.311)
##
## Constant                     58.811***
##                               (3.124)
##
## ---------------------------------------------
## Observations                    1,807
## Log Likelihood               -7,967.563
## Akaike Inf. Crit.            15,947.130
## =============================================
## Note:               *p<0.1; **p<0.05; ***p<0.01
```

The MSE of the traditional model is 395.2701693. The simple linear model shows that partisanship explains much of an individual's feelings towards Biden, as evidenced by the relatively large and statistically significant coefficients on the democrat and republican variables. An individual's gender also can help explain feelings towards Biden. Somewhat surprisingly, education level and age seem to have the least bearing on one's feelings towards Biden — this might indicate that despite age, an individual's party identification explains much of their political preferences.

2. Simple Holdout Validation

```
##Split data
set.seed(1414)
split <- initial_split(biden, prop = .5)
biden_train <- training(split)
biden_test <- testing(split)

#Fit model w/ training data
biden_train_ho <- lm(biden ~ .,
                          data = biden_train)

summary(biden_train_ho)
```

```
##
## Call:
## lm(formula = biden ~ ., data = biden_train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -76.244 -11.285   1.024  13.463  46.399
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 59.94623    4.66070  12.862  < 2e-16 ***
## female       4.91439    1.38557   3.547  0.00041 ***
## age          0.02769    0.04077   0.679  0.49718
## educ        -0.44396    0.29528  -1.504  0.13305
## dem         15.93531    1.54846  10.291  < 2e-16 ***
## rep        -15.73808    1.94635  -8.086 1.99e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

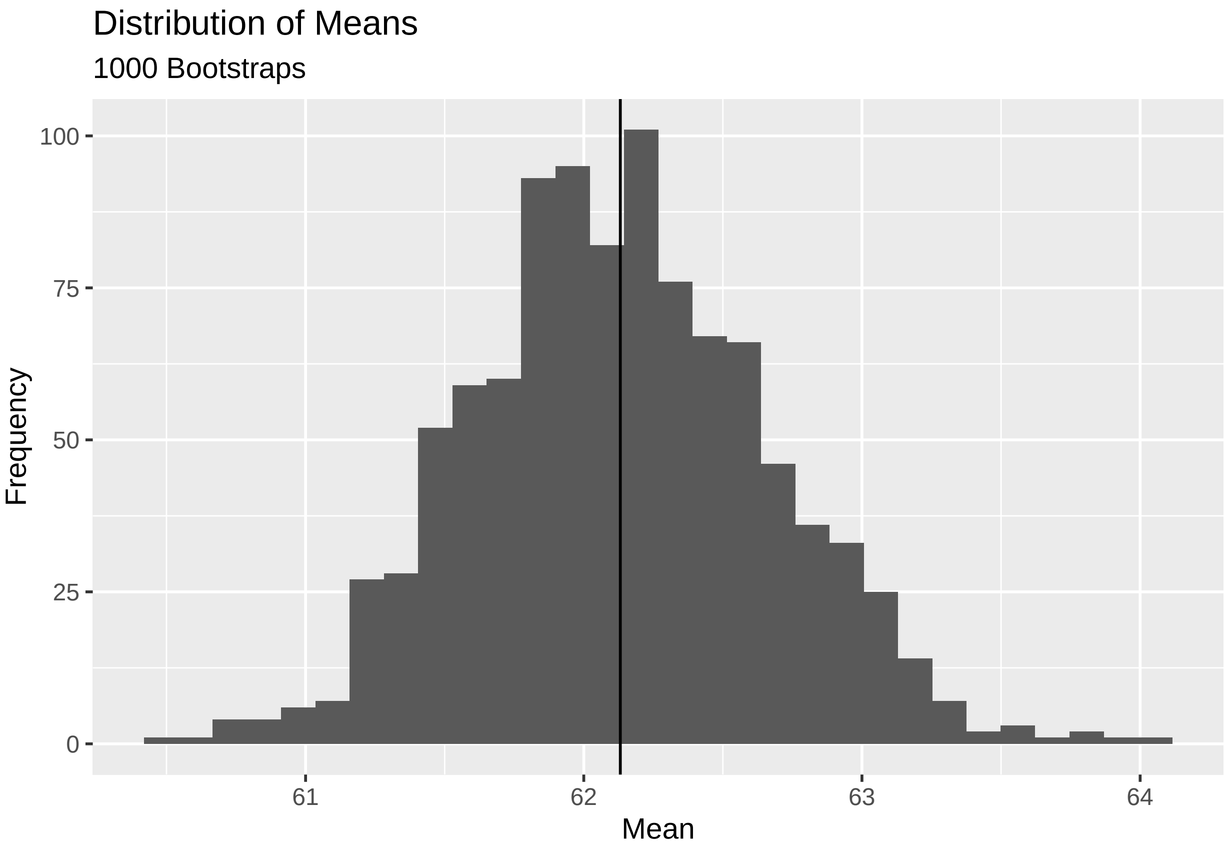
```
##
## Residual standard error: 20.44 on 898 degrees of freedom
## Multiple R-squared:  0.2791, Adjusted R-squared:  0.2751
## F-statistic: 69.53 on 5 and 898 DF,  p-value: < 2.2e-16
```

```
#Calculate MSE
biden_mse_test <- modelr::mse(biden_train_ho, biden_test)
```

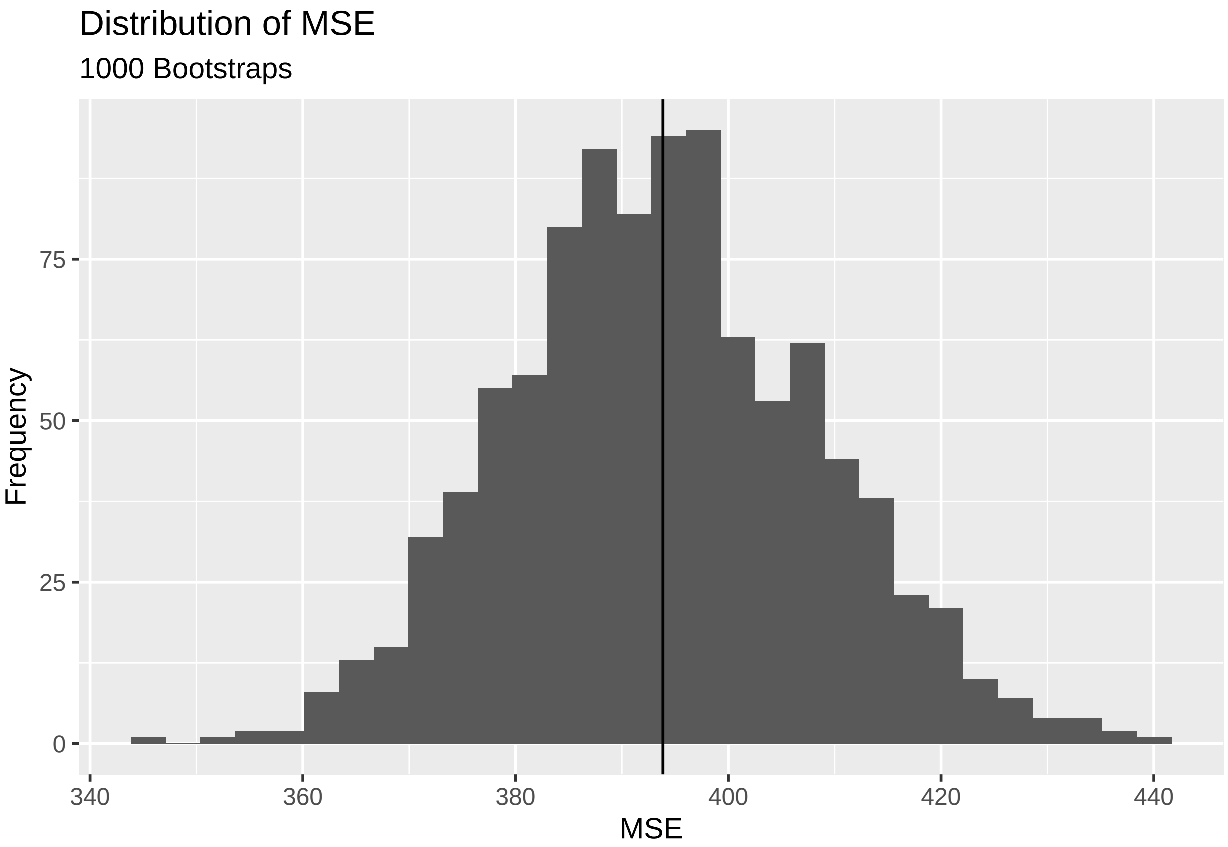
The MSE of the test set of the holdout model is 376.7664128. Considering that the MSE of the traditional model is 395.2701693, the simple holdout appraoch provides a more accurate learner. This suggests that we may have been overfitting the data in the traditional model, and actually building a model on fewer observations in the holdout approach results in a more accurate learner.

3. Bootstrap

```
mean_biden <- function(splits) {
  x <- analysis(splits)
  mean(x$biden)
}

##function for mse
biden_mse_boot <- function(split_data){

biden_train_ho <- lm(biden ~ .,
                          data = split_data)
biden_mse_train <- modelr::mse(biden_train_ho, split_data)
biden_mse_train
}


biden_bootstrap <- biden %>%
  bootstraps(1000) %>%
  mutate(mean = map_dbl(splits, mean_biden)) %>%
  mutate(mse_boot = map_dbl(splits, biden_mse_boot))

#Means
ggplot(biden_bootstrap) +
  geom_histogram(aes(mean)) +
  geom_vline(xintercept = mean(biden_bootstrap$mean)) +
  labs(x = "Mean",
       y = "Frequency",
       title = "Distribution of Means",
       subtitle = "1000 Bootstraps")
```



```
#MSE
ggplot(biden_bootstrap) +
  geom_histogram(aes(mse_boot)) +
  geom_vline(xintercept = mean(biden_bootstrap$mse_boot)) +
  labs(x = "MSE",
       y = "Frequency",
       title = "Distribution of MSE",
       subtitle = "1000 Bootstraps")
```



The mean of the sample means is 62.131378 and the mean of the MSEs is 393.851926. We can observe that the individual values of both appear to be drawn from a normal distribution, which we would expect. The mean of the bootstrapped means and MSEs also approximately converge on the true mean (62.1638074) and MSE (395.2701693) of the original sample.

Question 4. Comparison

```
stargazer(biden_traditional, type = "text")
```

```
##
## =============================================
##                       Dependent variable:
##                    ---------------------------
##                               biden
## ---------------------------------------------
## female                     4.103***
##                             (0.948)
##
## age                         0.048*
##                             (0.028)
##
## educ                        -0.345*
##                             (0.195)
##
## dem                        15.424***
##                             (1.068)
##
```

```
## rep                       -15.850***
##                             (1.311)
##
## Constant                   58.811***
##                             (3.124)
##
## --------------------------------------------
## Observations                  1,807
## Log Likelihood             -7,967.563
## Akaike Inf. Crit.          15,947.130
## ============================================
## Note:                 *p<0.1; **p<0.05; ***p<0.01
```

```
model_coef <- function(splits, ...) {
  mod <- lm(..., data = analysis(splits))
  tidy(mod)
}

biden_bootstrap4 <- biden %>%
  as_tibble()%>%
  bootstraps(1000)%>%
  mutate(coef = map(splits, model_coef, as.formula(biden ~ .)))

bootstrap_coefs <- biden_bootstrap4 %>%
  unnest(coef) %>%
  group_by(term) %>%
  summarize(.estimate = mean(estimate),
            .se = sd(estimate, na.rm = TRUE))

summary(biden_traditional)$coefficients
```

```
##                 Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)  58.81125899  3.1244366  18.822996 2.694143e-72
## female        4.10323009  0.9482286   4.327258 1.592601e-05
## age           0.04825892  0.0282474   1.708438 8.772744e-02
## educ         -0.34533479  0.1947796  -1.772952 7.640571e-02
## dem          15.42425563  1.0680327  14.441745 8.144928e-45
## rep         -15.84950614  1.3113624 -12.086290 2.157309e-32
```

Considering the outputs of our first model and our bootstrapped model, we can see that partisanship remains the strongest indicator of feelings towards Biden in both models. We also observe similar coeffiecients for the gender, age, and education variables. However, the standard errors in the bootstrapped model are slightly larger than in the original model, as we would expect — the bootstrapped model does not assume a strict functional form, and if we had reserved confidence in this assumption when analyzing this data, the bootstrapped model would be a more appropriate modeling choice.