MACS33002: Homework #2

Yujiao Song

Due Monday, Feb 3 by 5pm

Problem Set 2: Uncertainty, Holdouts, and Bootstrapping

Fork the repository

Fork the repository for the problem set 2, `problem-set-2` (https://github.com/macss-ml20/problem-set-2). *Remember, all final submissions should be a single rendered PDF with code produced in-line.* Also, don't forget to **open the pull request** once you've committed your final submission to your forked repository. It needs to be merged back into the course master branch to be considered "submitted". See the syllabus for details.

Joe Biden and Validation

Joe Biden was the 47th Vice President of the United States. He was the subject of many memes, attracted the attention of Leslie Knope, and experienced a brief surge in attention due to photos from his youth.

The goal here is to fit a regression model predicting feelings toward Biden, and then implement a couple validation techniques to evaluate the original findings. The validation techniques include the simple holdout approach and the bootstrap. **Note**: we are *not* covering cross validation (LOOCV or k-fold) in this problem set, as these topics are covered in the following week.

The 2008 NES Data

The `nes2008.csv` data contains a paired down selection of features from the full 2008 American National Election Studies survey. These data will allow you to test competing factors that may influence attitudes towards Joe Biden. The variables are coded as follows:

- `biden` - feeling thermometer ranging from 0-100. Feeling thermometers are a common metric in survey research used to gauge attitudes or feelings of "warmth" towards individuals and institutions. They range from 0-100, with 0 indicating extreme "coldness" and 100 indicating extreme "warmth."
- `female` - 1 if respondent is female, 0 if respondent is male
- `age` - age of respondent in years
- `educ` - number of years of formal education completed by respondent

- `dem` - 1 if respondent is a Democrat, 0 otherwise
- `rep` - 1 if respondent is a Republican, 0 otherwise

For this exercise we consider the following functional form,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon,$$

where Y is the Joe Biden feeling thermometer, and $[X_1 \ldots X_p]$ are the predictive features, including age, gender, education, Democrat, and Republican. The reason for including both `dem` and `rep` party affiliation features is to allow for capturing the preferences of Independents, which must be left out to serve as the baseline category, otherwise we would encounter perfect multicollinearity.

The Questions

```
nes2008 <- read.csv('nes2008.csv', header=T,sep=',')
```

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.

```
auto_lm <- lm(biden ~ female+age+educ+dem+rep, data =nes2008)
summary(auto_lm)
```

```
##
## Call:
## lm(formula = biden ~ female + age + educ + dem + rep, data = nes2008)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -75.55 -11.29   1.02  12.78  53.98
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  58.8113     3.1244   18.82  < 2e-16 ***
## female        4.1032     0.9482    4.33 0.000016 ***
## age           0.0483     0.0282    1.71    0.088 .
## educ         -0.3453     0.1948   -1.77    0.076 .
## dem          15.4243     1.0680   14.44  < 2e-16 ***
## rep         -15.8495     1.3114  -12.09  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.9 on 1801 degrees of freedom
## Multiple R-squared:  0.282,  Adjusted R-squared:  0.28
## F-statistic:  141 on 5 and 1801 DF,  p-value: <2e-16
```

```
(train_mse <- augment(auto_lm, newdata =nes2008 ) %>%
  mse(truth = biden, estimate = .fitted))
```

```
## # A tibble: 1 x 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 mse     standard        395.
```

The p-value of age and educ are lareger than 0.1 which means these two variables does not have a significant effect on votors' feelings toward Biden, whereas other vaiables like female dem and rep has p-value less tahn 0.001 which means they are significant to the output. MSE in this question is 395.3.Mean square error measures the error rate.

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%). **Be sure to set your seed prior to this part of your code to guarantee reproducibility of results.**

```
Auto <- as_tibble(Auto)

set.seed(1234)

nes2008_split <- initial_split(data =nes2008 ,
                               prop = 0.5)
```

```
* (5 points) _Fit_ the linear regression model using _only_ the _training_ observations.
```

```
train_model <- lm(biden ~ female+age+educ+dem+rep,
                  data = training(nes2008_split))

summary(train_model)
```

```
##
## Call:
## lm(formula = biden ~ female + age + educ + dem + rep, data = training(nes2008_split))
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -75.88 -11.95   1.93  11.90  46.12
##
## Coefficients:
```

```
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  58.6894     4.3032   13.64  < 2e-16 ***
## female        4.4134     1.2889    3.42  0.00064 ***
## age           0.0446     0.0386    1.16  0.24798    
## educ         -0.1826     0.2683   -0.68  0.49625    
## dem          13.6387     1.4535    9.38  < 2e-16 ***
## rep         -18.7684     1.7835  -10.52  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 19.1 on 898 degrees of freedom
## Multiple R-squared:  0.308,  Adjusted R-squared:  0.305 
## F-statistic: 80.1 on 5 and 898 DF,  p-value: <2e-16
```

```
* (10 points) Calculate the _MSE_ using _only_ the _test_ set observations.
```

```
(test_mse <- augment(auto_lm, newdata = testing(nes2008_split)) %>%
  mse(truth = biden, estimate = .fitted))
```

```
## # A tibble: 1 x 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 mse     standard        426.
```

```
* (10 points) How does this value compare to the training MSE from question 1? Present nume
```

q2's mse is 426.1 which is larger than 395.3 in q1. MSE in q2 is larger because it applies to the training set model on aset od completely new and unseen data points which is the testing set. MSE in Q1 estimate the accuracy and quality of the fit on the data we used to build this model. MSE in Q2 it estimate the predictive qulity.

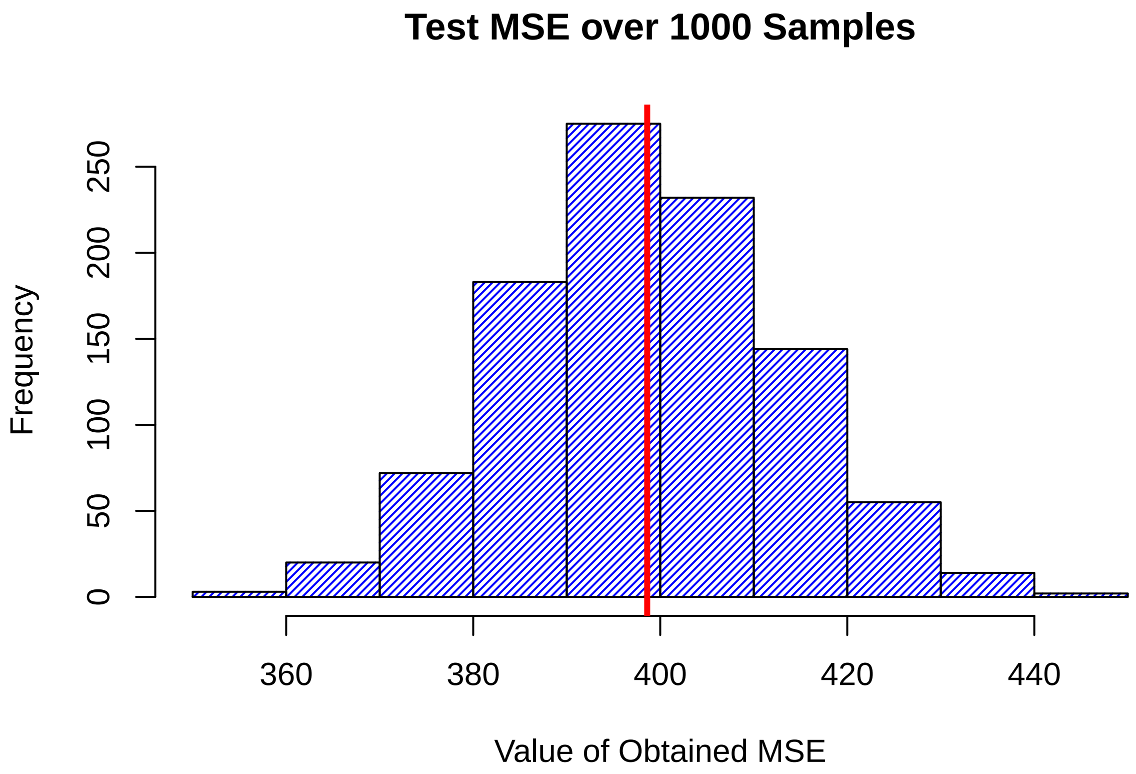
3. (30 points) Repeat the simple validation set approach from the previous question 1000 times, using 1000 different splits of the observations into a training set and a test/validation set. Visualize your results as a sampling distribution (hint: think histogram or density plots). Comment on the results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

```
set.seed(5)

test_mse=c()

for(i in 1:1000){
  train = sample(1:nrow(nes2008),0.5*nrow(nes2008))
  test = setdiff(1:nrow(nes2008),train)
```

```r
  biden_tr <- lm(biden ~ female + age + educ + dem + rep, data = nes2008[train, ])
  preds = predict(biden_tr,newx = nes2008[test, ])
  test_mse[i] = mean((nes2008$biden-predict(biden_tr, nes2008))[test]^2)
}
mean(test_mse)
```

```
## [1] 398.6
```

```r
hist(test_mse, density=35, main = "Test MSE over 1000 Samples", xlab = "Value of Obtained M
abline(v=mean(test_mse), lwd=3, col="red")
```



The bootstrapping method generate a set of standard error(398.4) that is pretty close to the original model(395.3).This tells us thatthe parameters in our original model are close to the true values of the population. The MSE did not change so much, which means the original model does not suffer from overfitting.

4. (30 points) Compare the estimated parameters and standard errors from the original model in question 1 (the model estimated using *all of the available data*) to parameters and standard errors estimated using the bootstrap ($B = 1000$). Comparison should include, at a minimum, both numeric output as well as discussion on differences, similarities, etc. Talk also about the conceptual use and impact of bootstrapping.

```
#Estimated parameters and SE from Original Model
tidy(auto_lm)
```

```
## # A tibble: 6 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)  58.8       3.12       18.8  2.69e-72
## 2 female        4.10      0.948       4.33 1.59e- 5
## 3 age           0.0483    0.0282      1.71 8.77e- 2
## 4 educ         -0.345     0.195      -1.77 7.64e- 2
## 5 dem          15.4       1.07       14.4  8.14e-45
## 6 rep         -15.8       1.31      -12.1  2.16e-32
```

```
# bootstrapped estimates of the parameter estimates and standard errors
lm_coefs <- function(splits, ...) {
  mod <- lm(..., data = analysis(splits))
  tidy(mod)
}

biden_boot <- nes2008 %>%
  bootstraps(1000) %>%
  mutate(coef = map(splits, lm_coefs, as.formula(biden ~ female + age + educ + dem + rep))

biden_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)   58.8    2.92
## 2 age            0.0480 0.0289
## 3 dem           15.4    1.10
## 4 educ          -0.342  0.191
## 5 female         4.11   0.963
## 6 rep          -15.8    1.37
```

The bootstrapping method generate a set of standard error that is pretty close to the original model.This tells us thatthe parameters in our original model are close to the true values of the population.The bootstrapped estimates of parameters are virtually identical, however the standard errors on the bootstrap estimates are slightly larger. This is because they do not rely on any distributional assumptions, whereas the traditional estimates do.The bootstrap method is not biased by distributional assumptions and gives us a more robust estimate.