Problem Set 2: Uncertainty, Holdouts, and Bootstrapping

Jiewen Luo

1/30/2020

```
library(tidyverse)

## Warning: package 'tidyverse' was built under R version 3.6.2

## -- Attaching packages ----------------------------------------- tidyverse
1.3.0 --

## v ggplot2 3.2.1     v purrr   0.3.3
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0

## Warning: package 'ggplot2' was built under R version 3.6.2

## Warning: package 'tibble' was built under R version 3.6.2

## Warning: package 'tidyr' was built under R version 3.6.2

## Warning: package 'readr' was built under R version 3.6.2

## Warning: package 'purrr' was built under R version 3.6.2

## Warning: package 'dplyr' was built under R version 3.6.2

## Warning: package 'stringr' was built under R version 3.6.2

## Warning: package 'forcats' was built under R version 3.6.2

## -- Conflicts -------------------------------------------- tidyverse_confli
cts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()

library(rsample)

## Warning: package 'rsample' was built under R version 3.6.2

library(broom)

## Warning: package 'broom' was built under R version 3.6.2

library(rcfss)
library(yardstick)

## Warning: package 'yardstick' was built under R version 3.6.2
```

```
## For binary classification, the first factor level is assumed to be the eve
nt.
## Set the global option `yardstick.event_first` to `FALSE` to change this.

##
## Attaching package: 'yardstick'

## The following object is masked from 'package:readr':
##
##     spec

da<-read.csv(file.choose())
da<-as_tibble(da)
```

1.

```
da_lm<-glm(biden~female+age+educ+dem+rep,data=da)
summary(da_lm)

##
## Call:
## glm(formula = biden ~ female + age + educ + dem + rep, data = da)
##
## Deviance 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
##
## (Dispersion parameter for gaussian family taken to be 396.587)
##
##     Null deviance: 994144  on 1806  degrees of freedom
## Residual deviance: 714253  on 1801  degrees of freedom
## AIC: 15947
##
## Number of Fisher Scoring iterations: 2

print(da_mse<-augment(da_lm,newdata=da)%>%mse(truth=biden, estimate=.fitted))

## # A tibble: 1 x 3
##   .metric .estimator .estimate
```

```
##   <chr>   <chr>          <dbl>
## 1 mse     standard        395.
```

When we fit the linear regression model using the entire dataset, the MSE is 395.2702. Since MSE is a measure of accuracy, one of our major goals is to pick a model with MSE as small as possible.

2.

```
set.seed(12345)
da_split<-initial_split(data=da,prop=.5)
da_train<-training(da_split)
da_test<-testing(da_split)
train_lm<-glm(biden ~ female + age + educ + dem + rep,data=da_train)
test_mse<-augment(train_lm,newdata=da_test)%>%mse(truth=biden,estimate=.fitte
d)
tibble(da_mse$.estimate,test_mse$.estimate)
```

```
## # A tibble: 1 x 2
##   `da_mse$.estimate` `test_mse$.estimate`
##                <dbl>                <dbl>
## 1               395.                 407.
```

The test MSE from the simple holdout validation approach was 407.333, whereas the MSE from Q1 is 395.2702.

The traditional model utilizes the entire sample set, and the MSE was calculated using the data that trained the model. On the other hand, the simple holdout model utilizes only half of the sample set, and the MSE was calculated using a new subset of data. Therefore, it makes sense that the traditional MSE is lower than the trained MSE from the simple holdout model.

However, this does not necessarily mean the model in Q1 is a better representative of the population because the traditional model introduces an optimistic bias due to overfitting.

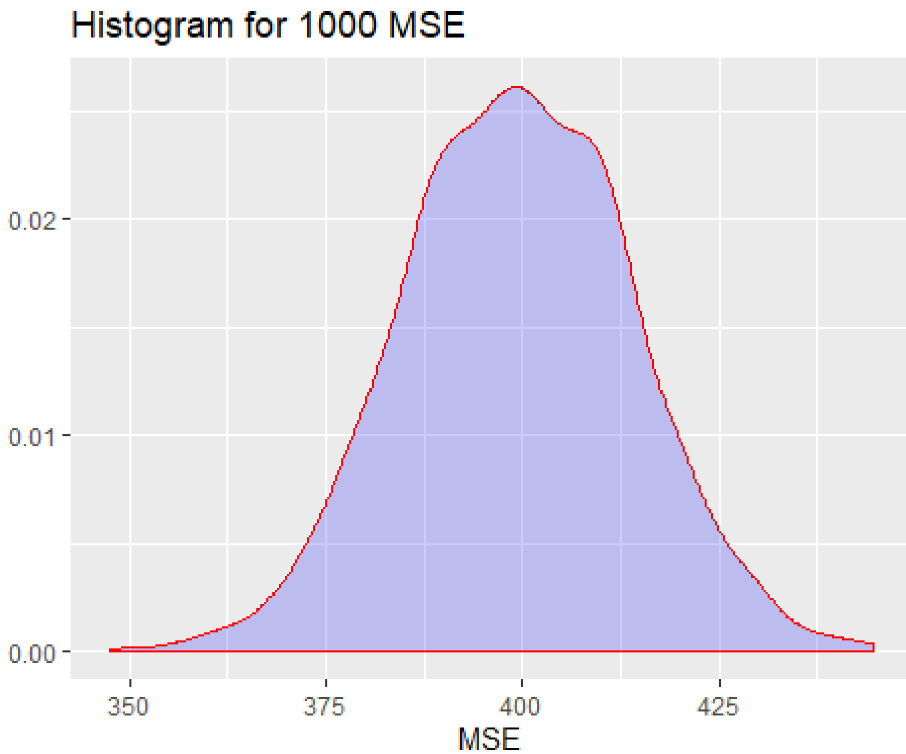
3.

```
mse1000<-c()
for (i in 1:1000) {
da_split<-initial_split(data=da,prop=.5)
da_train<-training(da_split)
da_test<-testing(da_split)
train_lm<-glm(biden ~ female + age + educ + dem + rep,data=da_train)
test_mse<-augment(train_lm,newdata=da_test)%>%mse(truth=biden,estimate=.fitte
d)
mse1000[[i]]=test_mse$.estimate
}

qplot(mse1000,
      geom="density",
      binwidth = 0.5,
```

```
      main = "Histogram for 1000 MSE",
      xlab = "MSE",
      fill=I("blue"),
      col=I("red"),
      alpha=I(.2))

## Warning: Ignoring unknown parameters: binwidth
```



The density has roughly a bell shape, this means the sample MSE has approximately a normal distribution with a mean around 400 and range from 350 to 450. This implies wide variance in results when using simple holdout approach and error predictions dependent on composition of the split.

4.

```
da_coefs<-function(splits, ...){
  mod<-glm(biden~female+age+educ+dem+rep,data=analysis(splits))
  tidy(mod)
}

da_boot<-da %>%
  bootstraps(1000) %>%
  mutate(coef=map(splits, da_coefs, as.formula(biden~female+age+educ+dem+rep)
))
s2<-da_boot %>%
```

```
  unnest(coef) %>%
  group_by(term) %>%
  summarize(boot_estimate=mean(estimate), boot_se=sd(estimate, na.rm = TRUE))
```

```
s1<-tibble(term=tidy(da_lm)$term, da_estimate=tidy(da_lm)$estimate, da_se=tid
y(da_lm)$std.error)
merge(s2,s1, by='term')
```

```
##          term boot_estimate    boot_se  da_estimate     da_se
## 1 (Intercept)   58.66873832 3.03866789  58.81125899 3.1244366
## 2         age    0.04863959 0.02921596   0.04825892 0.0282474
## 3         dem   15.38038587 1.11778078  15.42425563 1.0680327
## 4        educ   -0.33535238 0.19290975  -0.34533479 0.1947796
## 5      female    4.13245763 0.94392789   4.10323009 0.9482286
## 6         rep  -15.95208304 1.43134988 -15.84950614 1.3113624
```

From the output result we can see both models have very similar estimates. The differences are that Bootstrap method has slightly smaller estimates for Intercept, Democrat, and Republican, but slightly larger estimates for age, education and female.

Both models have very similar standard errors. The differences are that Bootstrap method has slightly smaller standard error for intercept, education and female, but slightly larger standard errors for age, Democrat and Republican

The above differences between these two models lie in the fact that the bootstrap estimates do not rely on any distributional assumptions, but the traditional estimates do.

Bootstrapping is a resampling technic used for estimating the sampling distribution. The idea of the bootstrap method is to generate multiple sets of new data with the same original sample size by sampling the original sample pool with replacement. Therefore, the bootstrap method is not biased by distributional assumptions, and it gives us a more robust estimate. Bootstrapping is commonly used to estimate the uncertainty of a performance estimate. It is very useful when we have limited access to data or when we are uncertain about the distribution assumption. However, sampling with replacement can cause problems like biased (non-generalizable) samples for fitting, testing, and evaluation.