Problem Set 2

Siva Bhavani

1/29/2020

The Questions

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.

```
linear <- lm(biden ~ ., data=nes2008)
summary(linear)
```

```
##
## Call:
## lm(formula = biden ~ ., data = nes2008)
##
## 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
```

```
nes2008$prediction = linear%>%predict(nes2008)
nes2008$diff = (nes2008$biden - nes2008$prediction)^2
RSS = sum(nes2008$diff)
MSE = RSS/nrow(nes2008)
MSE
```

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

The MSE is 395.3. The MSE is the mean squared error of the predictions compared to the actual values of feelings about Biden. MSE values are in the original units squared (so Biden thermometer squared). Since it is a sum of squared values, it will always be positive, and the lower it is the better the model fits the data. It represents the variance, bias and the random error of the model.

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. • (5 points) Fit the linear regression model using only the training observations. • (10 points) Calculate the MSE using only the test set observations. • (10 points) How does this value compare to the training MSE from question 1? Present numeric comparison and discuss a bit.

```
drop = c(7,8)
nes2008 = nes2008[-(drop)]

sample_size = floor(0.5*nrow(nes2008))
set.seed(122)

# randomly split data in r
picked = sample(seq_len(nrow(nes2008)),size = sample_size)
training =nes2008[picked,]
holdout =nes2008[-picked,]

linear <- lm(biden ~ ., data=training)

holdout$prediction = linear%>%predict(holdout)
holdout$diff = (holdout$biden - holdout$prediction)^2
RSS = sum(holdout$diff)
MSE = RSS/nrow(holdout)
MSE
```

```
## [1] 420.9873
```

The MSE using the test set observations is much higher (421 vs 395). The mean squared error is a lot higher because the model was not trained on the hold-out dataset. It did not generalize well. This suggests that there may be some overfitting.

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.

```
x <- c(1:10000)
rm(MSE)
MSE=0
sample_size = floor(0.5*nrow(nes2008))

for (val in x) {
set.seed(val)
picked = sample(seq_len(nrow(nes2008)),size = sample_size)
```

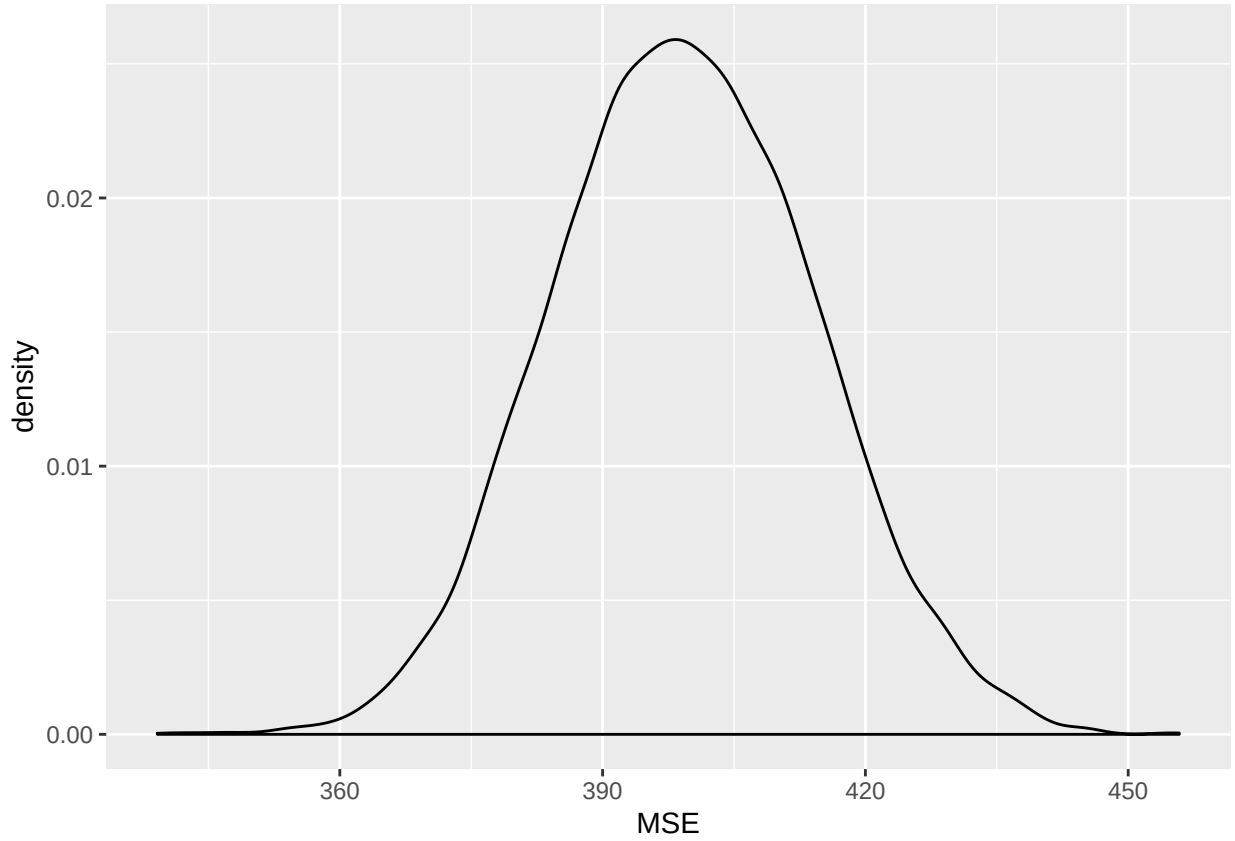
```
training =nes2008[picked,]
holdout =nes2008[-picked,]
linear <- lm(biden ~ ., data=training)

prediction = linear%>%predict(holdout)
diff = (holdout$biden - prediction)^2
RSS = sum(diff)
MSE[val] = RSS/nrow(holdout)
}

mean(MSE)
```

```
## [1] 399.3154
```

```
sd(MSE)
```

```
## [1] 14.83257
```

```
MSE = as.data.frame(MSE)
ggplot(data=MSE, aes(MSE)) +
  geom_density()
```



```
dev.off()
```

```
## null device
##           1
```

FIGURE 1

The MSE over the 1000 training/testing splits appears normally distributed around a mean of 399.26 and standard deviation of 14.8 The MSE is a composite of bias, variance and random error.

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.

```
Coef<-data.frame(intercept=numeric(0), female=numeric(0), age=numeric(0), educ=numeric(0), dem=numeric(
sample_size = floor(0.5*nrow(nes2008))


x <- c(1:1000)
for (val in x) {
set.seed(val)
picked = sample(seq_len(nrow(nes2008)),size = sample_size)
training =nes2008[picked,]
holdout =nes2008[-picked,]
linear <- lm(biden ~ ., data=training)
data = summary(linear)$coefficients[,1:2]
Coef[val,1] = data[1,1]
Coef[val,2] = data[2,1]
Coef[val,3] = data[3,1]
Coef[val,4] = data[4,1]
Coef[val,5] = data[5,1]
Coef[val,6] = data[6,1]
Coef[val,7] = data[1,2]
Coef[val,8] = data[2,2]
Coef[val,9] = data[3,2]
Coef[val,10] = data[4,2]
Coef[val,11] = data[5,2]
Coef[val,12] = data[6,2]
}

summary<-data.frame(intercept=numeric(0), female=numeric(0), age=numeric(0), educ=numeric(0), dem=numeri
x = c(1:12)
for (val in x){
summary[1,val]=mean(Coef[,val])
}

linear <- lm(biden ~ ., data=nes2008)
data = summary(linear)$coefficients[,1:2]
summary[2,1] = data[1,1]
summary[2,2] = data[2,1]
```

```
summary[2,3] = data[3,1]
summary[2,4] = data[4,1]
summary[2,5] = data[5,1]
summary[2,6] = data[6,1]
summary[2,7] = data[1,2]
summary[2,8] = data[2,2]
summary[2,9] = data[3,2]
summary[2,10] = data[4,2]
summary[2,11] = data[5,2]
summary[2,12] = data[6,2]

rownames(summary) = c("bootstrap", "overall")
rownames(summary) <- summary$X
summary$X <- NULL

summary_transpose <- as.data.frame(t(as.matrix(summary)))
colnames(summary_transpose) = c("bootstrap", "overall")

kable(summary_transpose, caption="Bootstrap results compared to Initial model")
```

Table 1: Bootstrap results compared to Initial model

	bootstrap	overall
intercept	58.7021597	58.8112590
female	4.0840895	4.1032301
age	0.0481390	0.0482589
educ	-0.3325402	-0.3453348
dem	15.3824445	15.4242556
rep	-15.9294047	-15.8495061
interceptse	4.4287023	3.1244366
femalese	1.3432697	0.9482286
agese	0.0400361	0.0282474
educse	0.2760625	0.1947796
demse	1.5126202	1.0680327
repse	1.8583713	1.3113624

The coefficients are very similar between the original model and the bootstrap averaged model. However, the standard errors of the coefficient are higher in the bootstrap model. The standard errors capture the uncertainty around the parameter, and the bootstrap model reflects this more accurately. The bootstrap model better captures the generalizability, since it is trained on portion of the data set and then tested on the other.