4255 Homework 1

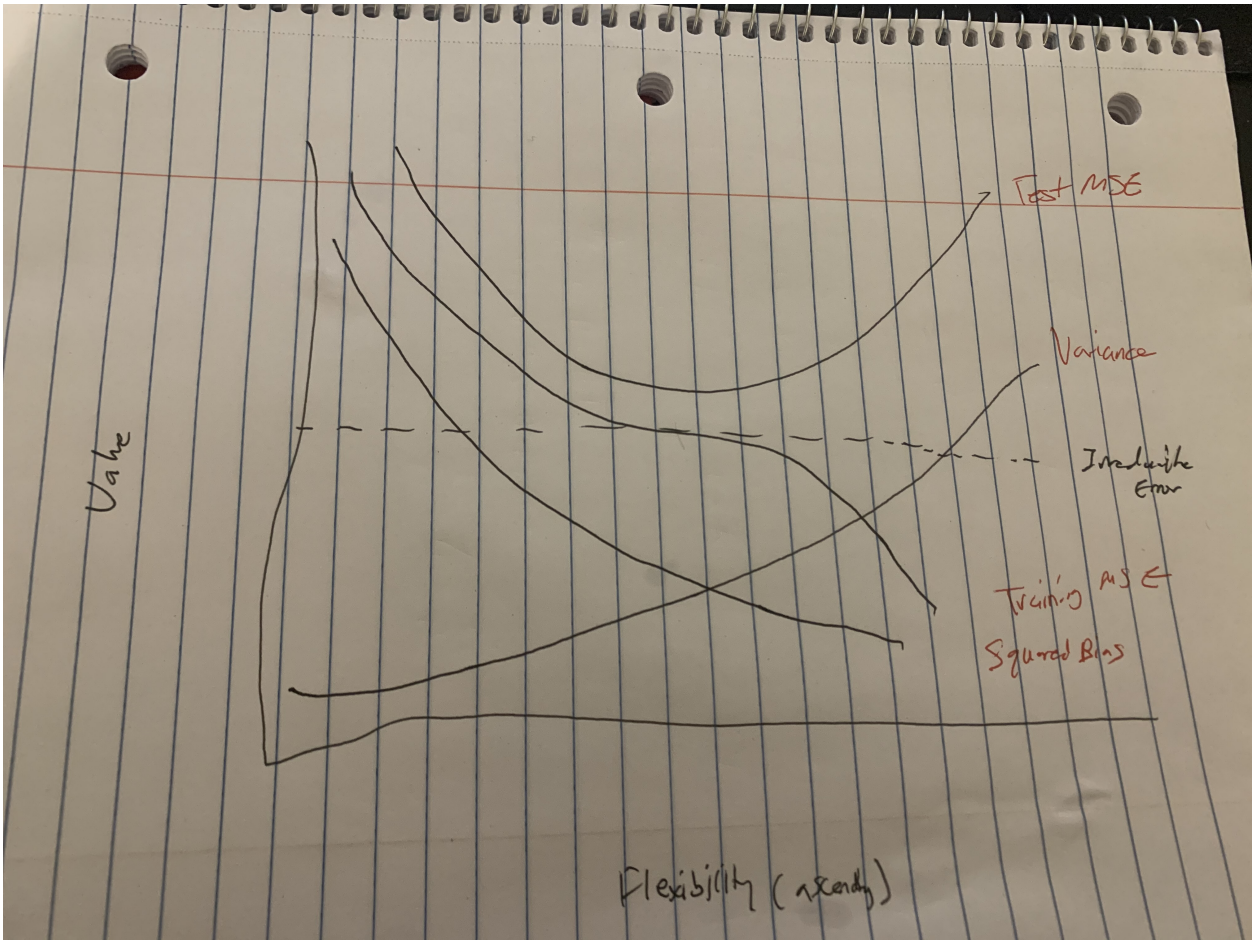
Hyeonjin Lee

2/17/2021

Chapter 2 Exercise 2

a) Regression. n = 500, p = 3
b) Classification. n = 20, p = 13
c) Regression. n = 52,p = 3

Chapter 2 Exercise 3 a)



foo
bar

b)

The test MSE is a U shape because it initially decreases as flexibility increases but at a certain point it will increase again. This is because there may be overfitting leading to the increase.

The training MSE declines as flexibility increases. When flexibility increases, the curve better fits the data.

The squared bias decreases as the flexibility increases. Using simpler models, it is unlikely that the real life data fits perfectly and thereby there will be bias. This will decrease as the flexibility increases.

The variance does the opposite and increases as flexibility increases. If the curve fits the data very closely and we change it to a different training data set, it will cause the curve to also change greatly.

The Bayes error curve stays constant. However, it is below the test MSE because the expected test MSE will be greater than the irreducible error.

Chapter 2 Exercise 7

a) Using the formula, the Euclidean distance between each observation and test point (1-6) is 3, 2, 3.16, 2.23, 1.41, 1.73 respectively.

b)

$$P(Y = Red|X = x_0) = \frac{1}{1}\sum_{i \in N_0} I(y_i = Red) = I(y_5 = Red) = 0$$

$$P(Y = Green|X = x_0) = \frac{1}{1}\sum_{i \in N_0} I(y_i = Green) = I(y_5 = Green) = 1$$

therefore our prediction is Green.

c)

$$P(Y = Red|X = x_0) = \frac{1}{3}\sum_{i \in N_0} I(y_i = Red) = \frac{1}{3}(1 + 0 + 1) = \frac{2}{3}$$

$$P(Y = Green|X = x_0) = \frac{1}{3}\sum_{i \in N_0} I(y_i = Green) = \frac{1}{3}(0 + 1 + 0) = \frac{1}{3}$$

Thus the prediction is Red.

d) The best value is for K to be small. When K becomes bigger, the boundary becomes inflexible.

Chapter 2 Exercise 10

a)

```
library(MASS)
nrow(Boston)
```

```
## [1] 506
```

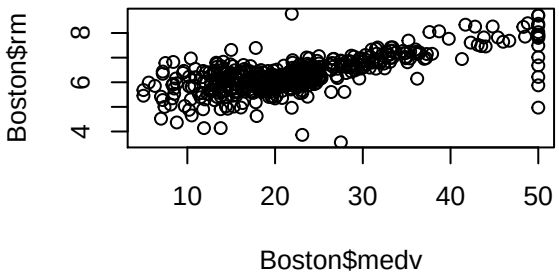
```
ncol(Boston)
```

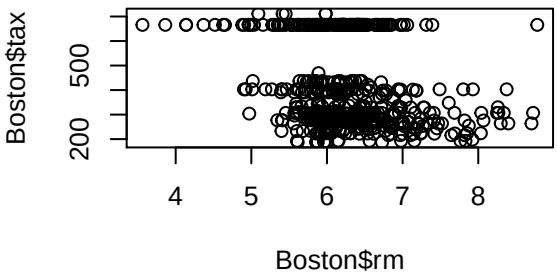
```
## [1] 14
```

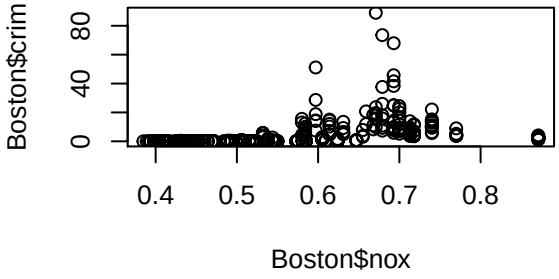
There are 506 rows which represent the data. The 14 columns represent the 14 predictors such as crime rate, average number of rooms per dwelling, etc.

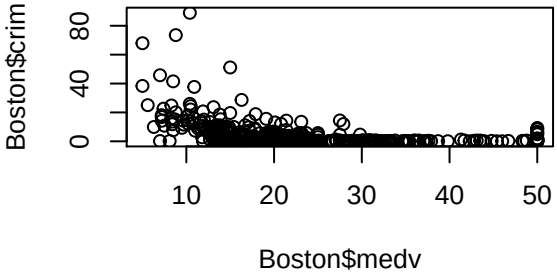
b)

```
par(mfrow = c(2,2))
plot(Boston$medv , Boston$rm)
plot(Boston$rm , Boston$tax)
plot(Boston$nox , Boston$crim)
plot(Boston$medv , Boston$crim)
```







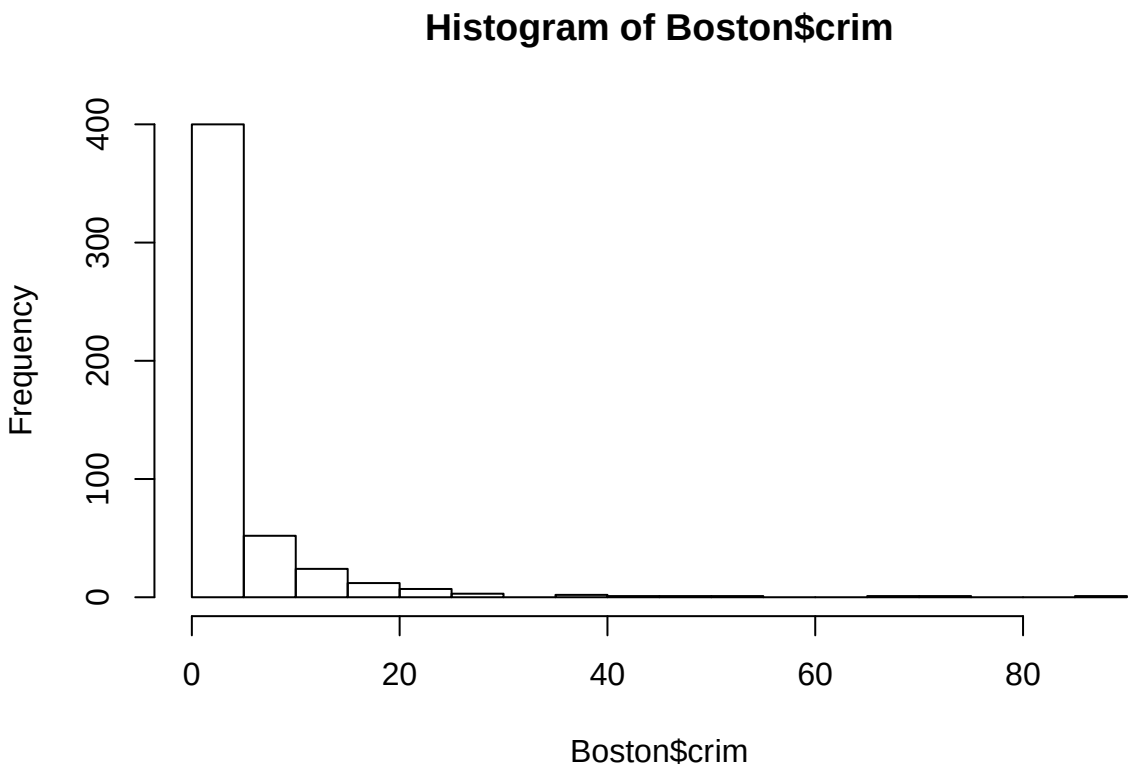


The first chart shows that as the median value of the home increases, we can expect to see more rooms per dwelling. The second chart looks at the property tax rate per $10,000 and the average rooms per dwelling. There is no apparent pattern. The third chart looks at nitrogen oxides concentration compared to the crime rate. It seems to follow a normal distribution with the highest crime rates between .6 and .7 nox. The fourth chart looks at the median value of homes and compares it with the crime rates. We can notice that the highest crime rates tend to occur when the median value of the homes are lower.

c) From the fourth chart above that looked at median home value and crime rate, we can note an association. Crime tends to occur in areas where the home value is low.

d)

```
hist(Boston$crim, breaks = 25)
```

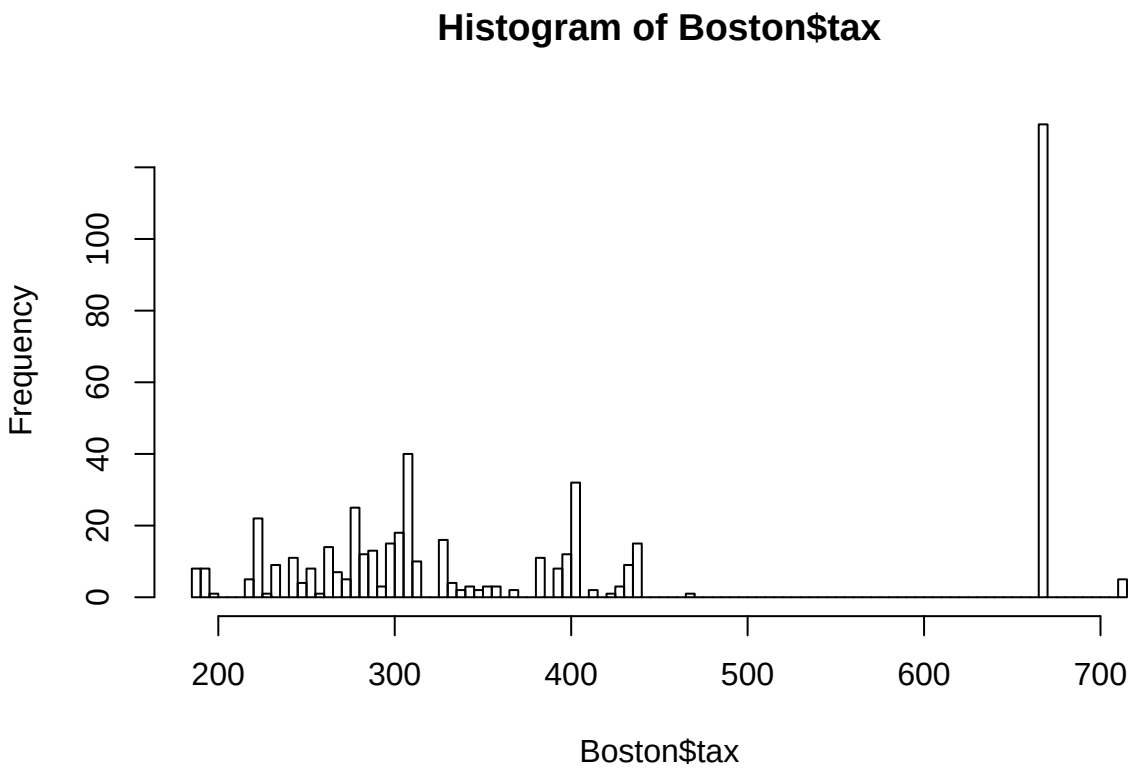



```
nrow(Boston[Boston$crim > 30, ])
```

```
## [1] 8
```

We can see that there are 8 data points where the crime rate was higher than 30% which is relatively high.

```
hist(Boston$tax, breaks = 100)
```

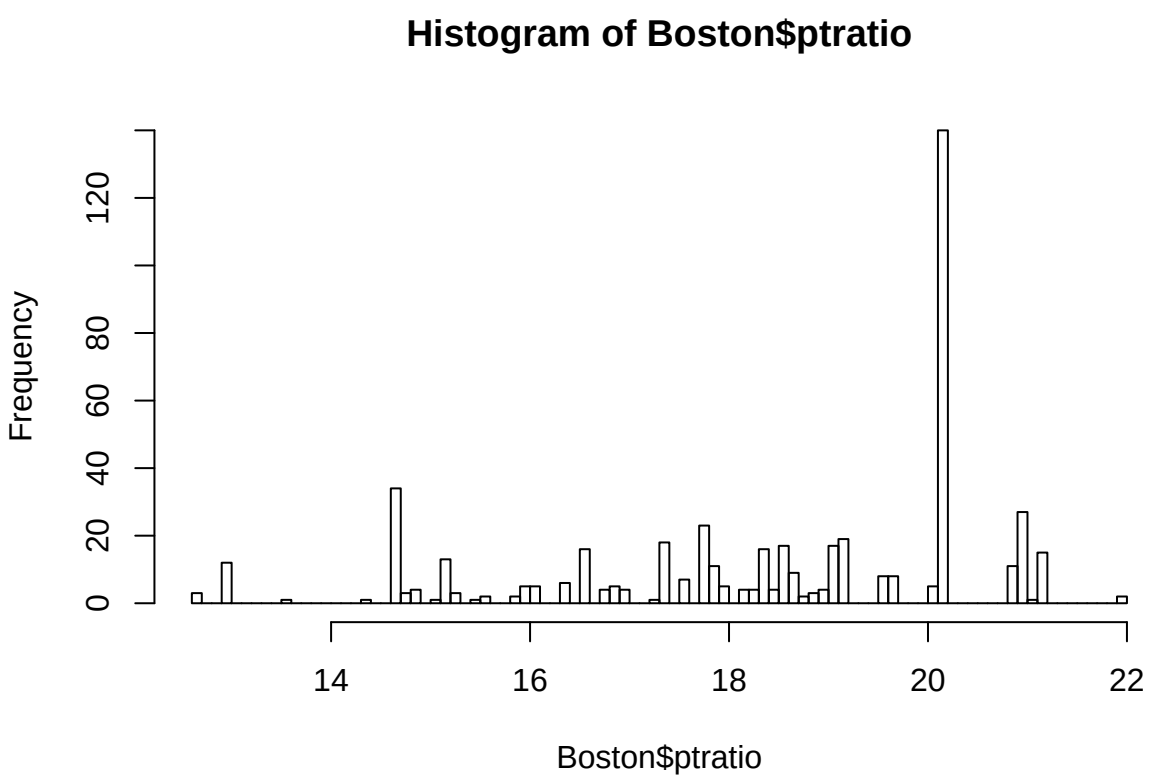



```
nrow(Boston[Boston$tax > 600, ])
```

```
## [1] 137
```

We can see that there are 132 data points with particularly high tax rates (> 600).

```
hist(Boston$ptratio, breaks = 100)
```




```
nrow(Boston[Boston$ptratio > 20, ])
```

```
## [1] 201
```

There were 201 points where there were particularly high pupil-teacher ratios (> 20).

e)

```
nrow(Boston[Boston$chas == 1, ])
```

```
## [1] 35
```

```
# 35 suburbs.
```

f)

```
median(Boston$ptratio)
```

```
## [1] 19.05
```

g)

```r
# Suburb of boston with lowest medv
row.names(Boston[min(Boston$medv), ])
```

```
## [1] "5"
```

```r
# Value of other predictors for that suburb row 5
Boston[5, ]
```

```
##      crim zn indus chas   nox    rm  age    dis rad tax ptratio black
## 5 0.06905  0  2.18    0 0.458 7.147 54.2 6.0622   3 222    18.7 396.9
##   lstat medv
## 5  5.33 36.2
```

```r
# How do these values compare to the overall range of these predictors
data.frame(min = sapply(Boston, min), max=sapply(Boston,max))
```

```
##               min      max
## crim      0.00632  88.9762
## zn        0.00000 100.0000
## indus     0.46000  27.7400
## chas      0.00000   1.0000
## nox       0.38500   0.8710
## rm        3.56100   8.7800
## age       2.90000 100.0000
## dis       1.12960  12.1265
## rad       1.00000  24.0000
## tax     187.00000 711.0000
## ptratio  12.60000  22.0000
## black     0.32000 396.9000
## lstat     1.73000  37.9700
## medv      5.00000  50.0000
```

One interesting thing to note is that the suburb that has the lowest median value of owner-occupied homes has a relatively high rm (average number of rooms per dwelling). It is close to the max of the range when compared to the range of the overall rm values.

h)

```r
# How many suburbs average more than 7 rooms per dwelling
nrow(Boston[Boston$rm > 7, ])
```

```
## [1] 64
```

```r
# How many suburbs average more than 8 rooms per dwelling
nrow(Boston[Boston$rm > 8, ])
```

```
## [1] 13
```

```r
Boston[row.names(Boston[Boston$rm > 8, ]), ]
```

```
##        crim zn indus chas    nox    rm  age    dis rad tax ptratio  black
## 98  0.12083  0  2.89    0 0.4450 8.069 76.0 3.4952   2 276    18.0 396.90
## 164 1.51902  0 19.58    1 0.6050 8.375 93.9 2.1620   5 403    14.7 388.45
## 205 0.02009 95  2.68    0 0.4161 8.034 31.9 5.1180   4 224    14.7 390.55
## 225 0.31533  0  6.20    0 0.5040 8.266 78.3 2.8944   8 307    17.4 385.05
## 226 0.52693  0  6.20    0 0.5040 8.725 83.0 2.8944   8 307    17.4 382.00
## 227 0.38214  0  6.20    0 0.5040 8.040 86.5 3.2157   8 307    17.4 387.38
## 233 0.57529  0  6.20    0 0.5070 8.337 73.3 3.8384   8 307    17.4 385.91
## 234 0.33147  0  6.20    0 0.5070 8.247 70.4 3.6519   8 307    17.4 378.95
## 254 0.36894 22  5.86    0 0.4310 8.259  8.4 8.9067   7 330    19.1 396.90
## 258 0.61154 20  3.97    0 0.6470 8.704 86.9 1.8010   5 264    13.0 389.70
## 263 0.52014 20  3.97    0 0.6470 8.398 91.5 2.2885   5 264    13.0 386.86
## 268 0.57834 20  3.97    0 0.5750 8.297 67.0 2.4216   5 264    13.0 384.54
## 365 3.47428  0 18.10    1 0.7180 8.780 82.9 1.9047  24 666    20.2 354.55
##     lstat medv
## 98   4.21 38.7
## 164  3.32 50.0
## 205  2.88 50.0
## 225  4.14 44.8
## 226  4.63 50.0
## 227  3.13 37.6
## 233  2.47 41.7
## 234  3.95 48.3
## 254  3.54 42.8
## 258  5.12 50.0
## 263  5.91 48.8
## 268  7.44 50.0
## 365  5.29 21.9
```

We can see that out of the 64 suburbs that have more than 7 rooms, only 13 have more than 8. And when looking further, the highest average number of rooms is 8.780. Surprisingly, the median home value of this suburb of 21.9 falls close to the median home value of the entire dataset of 21.2 even though it has the highest rm.