MGSC 310 Problem Set #1

Ananya Vittal

Ch 2

1.

a) A flexible statistical learning method would perform better since there is a large sample size.

b) Performance would be worse. Since the number of observations is small and the number of predictors is large, a flexible method might overfit the data.

c) A flexible method would fit better since the relationship between the predictors and response is non-linear.

d) A flexible method would perform worse since there is a high degree of variance.

2.

a) This is an inference because we are trying to understand which factors affect salary using regression. n = 500 and p = 3

b) This is a prediction since there are 2 different outcomes: "success" and "failure." This is a classification. n = 20 and p = 13

c) This is a prediction of the % of change in the exchange rate using regression. n = 52 and p = 3

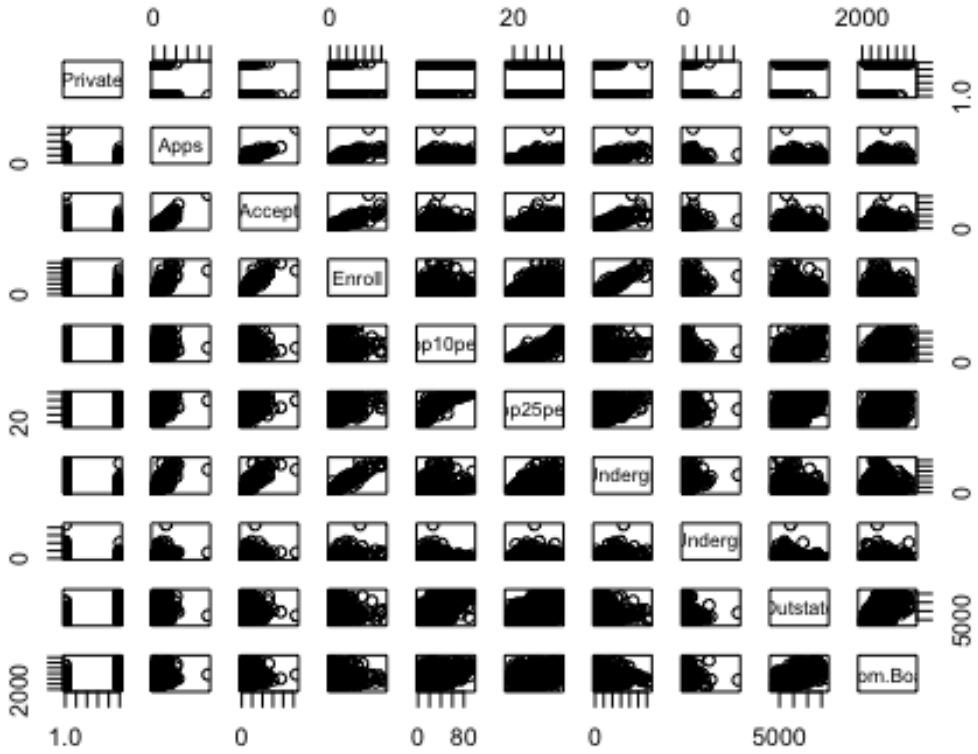
8.

```
#install.packages("ggplot2")
library("ggplot2")

## Warning: package 'ggplot2' was built under R version 3.4.4

setwd("/Users/ananyavittal/Documents/MGSC310")
college<-read.csv("College.csv")

head(college)

##                              X Private Apps Accept Enroll Top10perc
## 1 Abilene Christian University     Yes 1660   1232    721        23
## 2           Adelphi University     Yes 2186   1924    512        16
## 3               Adrian College     Yes 1428   1097    336        22
## 4          Agnes Scott College     Yes  417    349    137        60
```

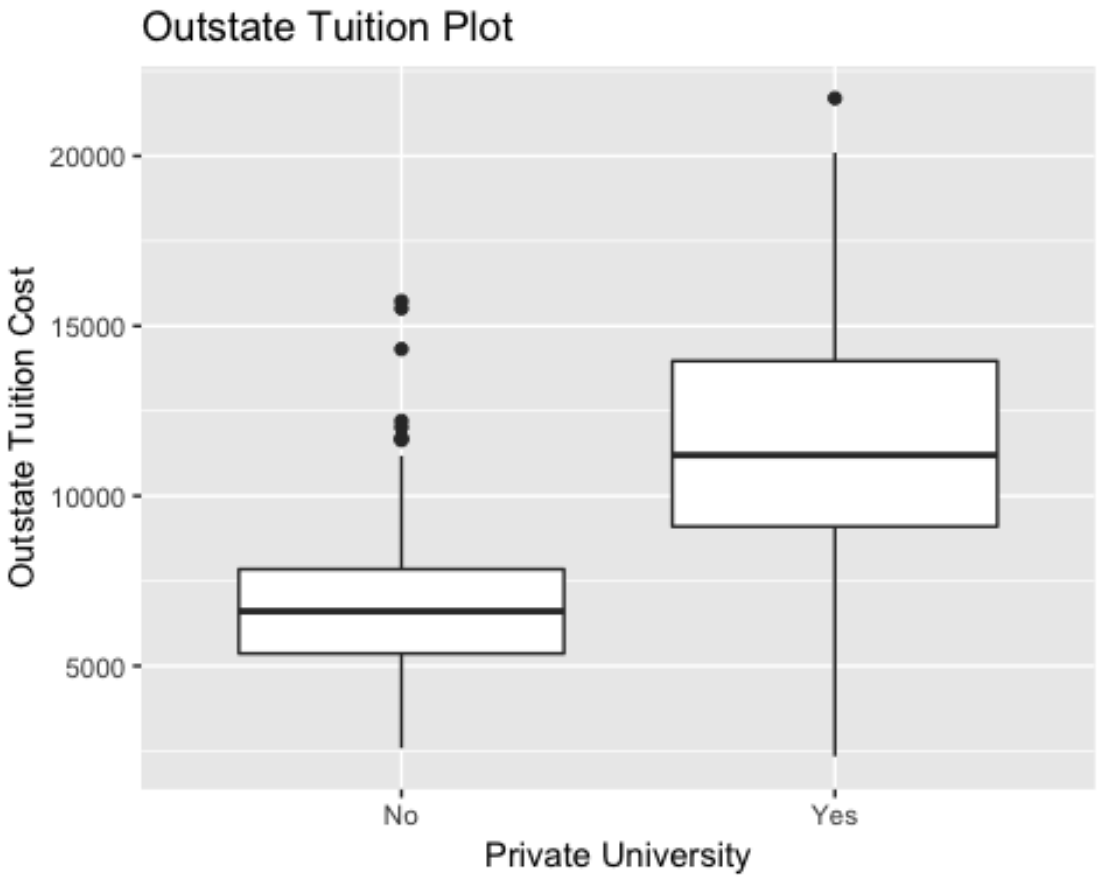
```
## 5    Alaska Pacific University     Yes  193    146     55        16
## 6            Albertson College     Yes  587    479    158        38
##   Top25perc F.Undergrad P.Undergrad Outstate Room.Board Books Personal PhD
## 1        52        2885         537     7440       3300   450     2200  70
## 2        29        2683        1227    12280       6450   750     1500  29
## 3        50        1036          99    11250       3750   400     1165  53
## 4        89         510          63    12960       5450   450      875  92
## 5        44         249         869     7560       4120   800     1500  76
## 6        62         678          41    13500       3335   500      675  67
##   Terminal S.F.Ratio perc.alumni Expend Grad.Rate
## 1       78      18.1          12   7041        60
## 2       30      12.2          16  10527        56
## 3       66      12.9          30   8735        54
## 4       97       7.7          37  19016        59
## 5       72      11.9           2  10922        15
## 6       73       9.4          11   9727        55

rownames<-college[,1]
head(college)

##                              X Private Apps Accept Enroll Top10perc
## 1 Abilene Christian University     Yes 1660   1232    721        23
## 2           Adelphi University     Yes 2186   1924    512        16
## 3               Adrian College     Yes 1428   1097    336        22
## 4          Agnes Scott College     Yes  417    349    137        60
## 5    Alaska Pacific University     Yes  193    146     55        16
## 6            Albertson College     Yes  587    479    158        38
##   Top25perc F.Undergrad P.Undergrad Outstate Room.Board Books Personal PhD
## 1        52        2885         537     7440       3300   450     2200  70
## 2        29        2683        1227    12280       6450   750     1500  29
## 3        50        1036          99    11250       3750   400     1165  53
## 4        89         510          63    12960       5450   450      875  92
## 5        44         249         869     7560       4120   800     1500  76
## 6        62         678          41    13500       3335   500      675  67
##   Terminal S.F.Ratio perc.alumni Expend Grad.Rate
## 1       78      18.1          12   7041        60
## 2       30      12.2          16  10527        56
## 3       66      12.9          30   8735        54
## 4       97       7.7          37  19016        59
## 5       72      11.9           2  10922        15
## 6       73       9.4          11   9727        55

college<-college[,-1]
head(college)

##   Private Apps Accept Enroll Top10perc Top25perc F.Undergrad P.Undergrad
## 1     Yes 1660   1232    721        23        52        2885         537
## 2     Yes 2186   1924    512        16        29        2683        1227
## 3     Yes 1428   1097    336        22        50        1036          99
## 4     Yes  417    349    137        60        89         510          63
## 5     Yes  193    146     55        16        44         249         869
```

```
## 6     Yes  587    479    158        38        62         678          41
##   Outstate Room.Board Books Personal PhD Terminal S.F.Ratio perc.alumni
## 1     7440       3300   450     2200  70       78      18.1          12
## 2    12280       6450   750     1500  29       30      12.2          16
## 3    11250       3750   400     1165  53       66      12.9          30
## 4    12960       5450   450      875  92       97       7.7          37
## 5     7560       4120   800     1500  76       72      11.9           2
## 6    13500       3335   500      675  67       73       9.4          11
##   Expend Grad.Rate
## 1   7041        60
## 2  10527        56
## 3   8735        54
## 4  19016        59
## 5  10922        15
## 6   9727        55

summary(college)

##  Private        Apps           Accept          Enroll       Top10perc
##  No :212   Min.   :   81   Min.   :   72   Min.   :  35   Min.   : 1.00
##  Yes:565   1st Qu.:  776   1st Qu.:  604   1st Qu.: 242   1st Qu.:15.00
##            Median : 1558   Median : 1110   Median : 434   Median :23.00
##            Mean   : 3002   Mean   : 2019   Mean   : 780   Mean   :27.64
##            3rd Qu.: 3624   3rd Qu.: 2424   3rd Qu.: 902   3rd Qu.:36.00
##            Max.   :48094   Max.   :26330   Max.   :6392   Max.   :96.00
##    Top25perc      F.Undergrad     P.Undergrad         Outstate
##  Min.   :  9.0   Min.   :  139   Min.   :    1.0   Min.   : 2340
##  1st Qu.: 41.0   1st Qu.:  992   1st Qu.:   95.0   1st Qu.: 7320
##  Median : 54.0   Median : 1707   Median :  353.0   Median : 9990
##  Mean   : 55.8   Mean   : 3700   Mean   :  855.3   Mean   :10441
##  3rd Qu.: 69.0   3rd Qu.: 4005   3rd Qu.:  967.0   3rd Qu.:12925
##  Max.   :100.0   Max.   :31643   Max.   :21836.0   Max.   :21700
##    Room.Board       Books           Personal         PhD
##  Min.   :1780   Min.   :  96.0   Min.   : 250   Min.   :  8.00
##  1st Qu.:3597   1st Qu.: 470.0   1st Qu.: 850   1st Qu.: 62.00
##  Median :4200   Median : 500.0   Median :1200   Median : 75.00
##  Mean   :4358   Mean   : 549.4   Mean   :1341   Mean   : 72.66
##  3rd Qu.:5050   3rd Qu.: 600.0   3rd Qu.:1700   3rd Qu.: 85.00
##  Max.   :8124   Max.   :2340.0   Max.   :6800   Max.   :103.00
##     Terminal       S.F.Ratio      perc.alumni        Expend
##  Min.   : 24.0   Min.   : 2.50   Min.   : 0.00   Min.   : 3186
##  1st Qu.: 71.0   1st Qu.:11.50   1st Qu.:13.00   1st Qu.: 6751
##  Median : 82.0   Median :13.60   Median :21.00   Median : 8377
##  Mean   : 79.7   Mean   :14.09   Mean   :22.74   Mean   : 9660
##  3rd Qu.: 92.0   3rd Qu.:16.50   3rd Qu.:31.00   3rd Qu.:10830
##  Max.   :100.0   Max.   :39.80   Max.   :64.00   Max.   :56233
##    Grad.Rate
##  Min.   : 10.00
##  1st Qu.: 53.00
##  Median : 65.00
```

```
##  Mean   : 65.46
##  3rd Qu.: 78.00
##  Max.   :118.00

pairs(college[,1:10])
```



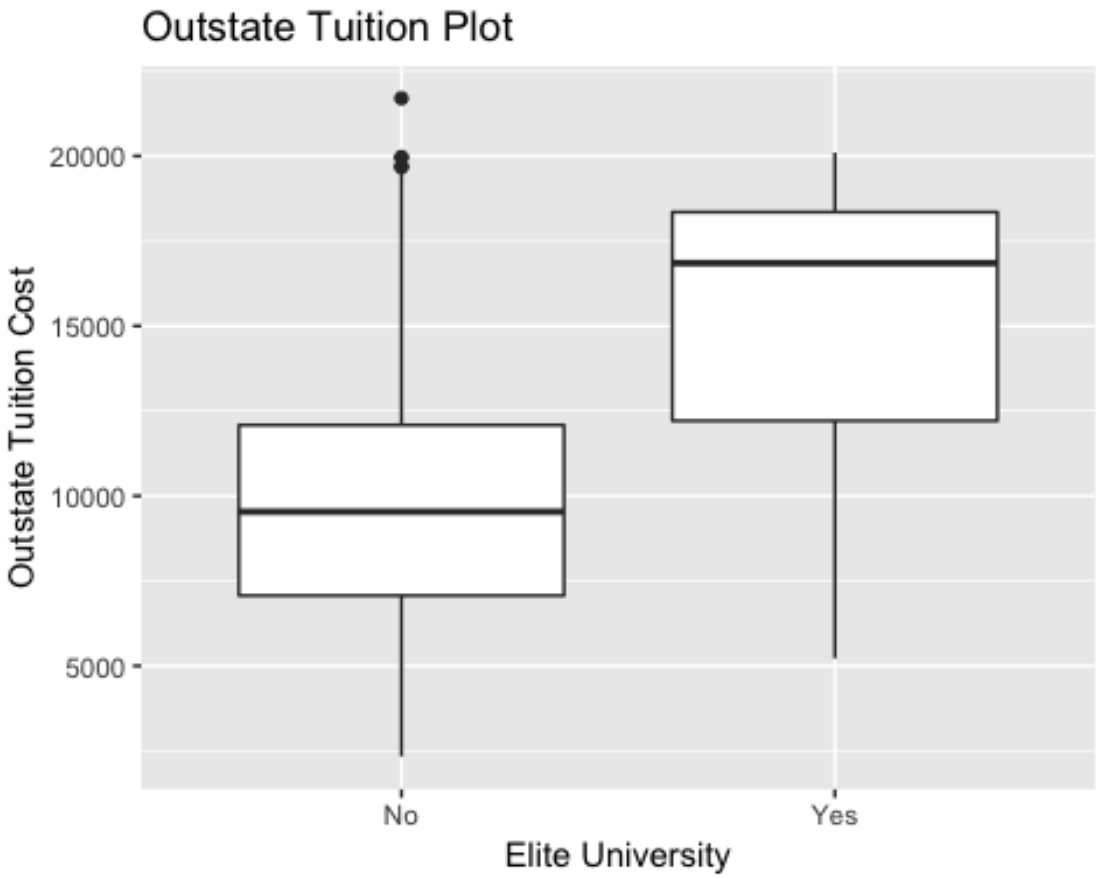

```
ggplot(college,aes(Private, Outstate)) + geom_boxplot() +
labs(title="Outstate Tuition Plot",x="Private University", y = "Outstate
Tuition Cost")
```



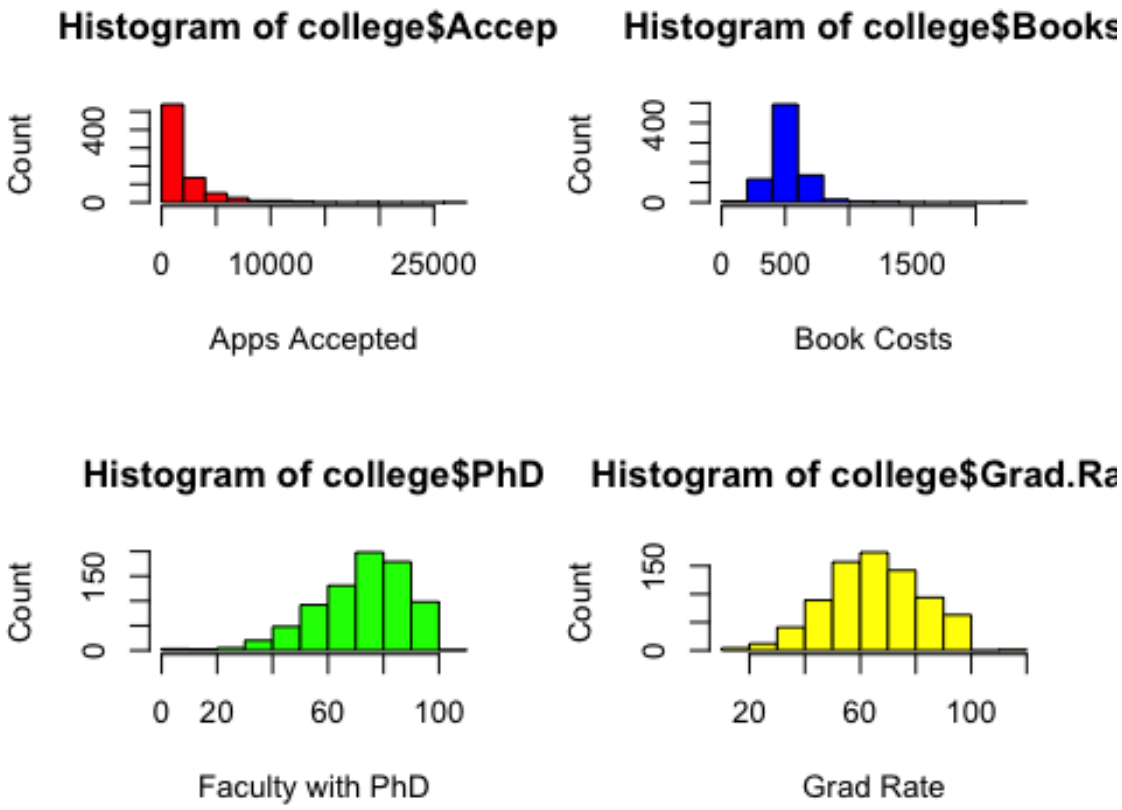

```
Elite<-rep("No",nrow(college))
Elite[college$Top10perc > 50]<-"Yes"
Elite<-as.factor(Elite)
college<-data.frame(college, Elite)
summary(Elite)
```

```
##  No Yes
## 696  81
```

```
ggplot(college,aes(Elite, Outstate)) + geom_boxplot() + labs(title="Outstate
Tuition Plot",x="Elite University", y = "Outstate Tuition Cost")
```




```r
par(mfrow=c(2,2))
hist(college$Accept, col = "red", xlab = "Apps Accepted", ylab = "Count")
hist(college$Books, col = "blue", xlab = "Book Costs", ylab = "Count")
hist(college$PhD, col = "green", xlab = "Faculty with PhD", ylab = "Count")
hist(college$Grad.Rate, col = "yellow", xlab = "Grad Rate", ylab = "Count")
```




```
sum(is.na(college))
```

```
## [1] 0
```

There are 0 missing cases in the dataset.

2) Calculate E[5x+2] and Var(3x)

E[x] = (-1)(0.5) + (1)(0.5) = 0

So: E[x] = 0

E[x2] = (1)(0.5) + (1)(0.5) = 1

So: E[x2] = 1

Now, we use the expected value and variance properties to calculate E[5x+2] and var(3x).

E[5x+2] = 5*E[x] + E[2] = 0 + E[2] = 2

Var(3x) = 3^2 * Var(x) = 9[E(X2) – [E(X)]2] = 9[1 - 0] = 9

Therefore E[5x+2] = 2 and Var(3x) = 9

3) What is the minimum value of the test MSE?

The goal is to minimize the value of the test mean squared error. When there is a perfect model, the variance and bias can be reduced to 0. Therefore, the minimum value or lower bound of the test MSE is the irreducible error or var(ϵ).