Tanmay Gupta
tg289

Homework-1

Q1. For each of parts (a) through (d), indicate whether i. or ii. is correct, and explain your answer. In general, do we expect the performance of a flexible statistical learning method to perform better or worse than an inflexible method when :

a. The sample size n is extremely large, and the number of predictors p is small ?

->Better. A flexible method will fit the data closer and with the large sample size, would perform better than an inflexible approach.

b. The number of predictors p is extremely large, and the number of observations n is small ?

->Worse. A flexible method would overfit the small number of observations.

c. The relationship between the predictors and response is highly non-linear ?

->Better. With more degrees of freedom, a flexible method would fit better than an inflexible one.

d. The variance of the error terms, i.e. $\sigma 2=Var(\varepsilon)\sigma 2=Var(\varepsilon)$, is extremely high ?

->Worse. A flexible method would fit to the noise in the error terms and increase variance.

Q2. Explain whether each scenario is a classification or regression problem, and indicate whether we are most interested in inference or prediction. Finally, provide n and p.

a. We collect a set of data on the top 500 firms in the US. For each firm we record profit, number of employees, industry and the CEO salary. We are interested in understanding which factors affect CEO salary.

->Regression and inference with n=500 and p=3

b. We are considering launching a new product and wish to know whether it will be a success or a failure. We collect data on 20 similar products that were previously launched. For each product we have recorded whether it was a success or failure, price charged for the product, marketing budget, competition price, and ten other variables.

->Classification and prediction with n=20 and p=13

c. We are interested in predicting the % change in the US dollar in relation to the weekly changes in the world stock markets. Hence we collect weekly data for all of 2012. For each week we record the % change in the dollar, the % change in the US market, the % change in the British market, and the % change in the German market.

->Regression and prediction with n=52 and p=3

4. You will now think of some real-life applications for statistical learning.
(a) Describe three real-life applications in which *classification* might be useful. Describe the response, as well as the predictors. Is the goal of each application inference or prediction? Explain your answer.

- Classification example 1- Is the email a spam or not

Response- Yes/No
Predictors-Promotions/newsletters/subscriptions/updates/type of mail/sender's address/type of account
Goal- Prediction

- Classification example 2- Should a product be launched in the market or not

Response- Yes/No
Predictors- Features/Price/Market/Buyer/Effectiveness in market/ability to sustain/competitors/type of product
Goal-Prediction

- Classification example 3- Polio Vaccine test

Response- Successful/Not-successful
Predictors- Geography, health condition, age, test group, mental health
Goal-Prediction

(b) Describe three real-life applications in which *regression* might be useful. Describe the response, as well as the predictors. Is the goal of each application inference or prediction? Explain your answer.

- Regression example 1- Gas mileage that a new car will give

Response-The mileage of this car will be XX
Predictors-Model, make, Engine,Company
Goal-Inference

- Regression example 2- GDP growth in an economy

Response- The countries GDp growth for this year is …
Predictors- GDP, Employment, Education,Trade,Relations,
Goal- Inference

- Regression example 3- Increase in house prices over next 3 years

Response- The price of houses have increased by this amount
Predictors- Parks, Schools, Average size of family, Average Income of Family, Crime Rate
Goal- Inference

(c) Describe three real-life applications in which *cluster analysis* might be useful.

- Clustering example 1- Division of income group in USD-5000-10000, 10000-15000, 15000 and above

Response- This % of people fall under this income group
Predictors- Countries, social life, employment, political status, economic status
Goal-Prediction

- Clustering example 2- Ratings given to cars as good/average/bad

Response-This car is rated good
Predictors-Companies, model,make,popularity,Advertisement
Goal-Prediction

- Clustering example 3- Division of countries into developed, developing and others

Response- Asian countries are developing while European are developed
Predictors- GDP, Employment, Education,Trade,Relations,

Goal-Prediction

5. What are the advantages and disadvantages of a very flexible (versus a less flexible) approach for regression or classification? Under what circumstances might a more flexible approach be preferred to a less flexible approach? When might a less flexible approach be preferred?

->The advantages of a very flexible approach are that it may give a better fit for non-linear models and it decreases the bias.The disadvantages of a very flexible approach are that it requires estimating a greater number of parameters, it follows the noise too closely (overfit) and it increases the variance.

A more flexible approach would be preferred to a less flexible approach when we are interested in prediction and not the interpretability of the results.A less flexible approach would be preferred to a more flexible approach when we are interested in inference and the interpretability of the results.

8. This exercise relates to the College data set, which can be found in the file College.csv. It contains a number of variables for 777 different universities and colleges in the US. The variables are

- Private : Public/private indicator
- Apps : Number of applications received
- Accept : Number of applicants accepted
- Enroll : Number of new students enrolled
- Top10perc : New students from top 10% of high school class
- Top25perc : New students from top 25% of high school class
- F.Undergrad : Number of full-time undergraduates
- P.Undergrad : Number of part-time undergraduates
- Outstate : Out-of-state tuition
- Room.Board : Room and board costs
- Books : Estimated book costs
- Personal : Estimated personal spending
- PhD : Percent of faculty with Ph.D.'s
- Terminal : Percent of faculty with terminal degree
- S.F.Ratio : Student/faculty ratio
- perc.alumni : Percent of alumni who donate
- Expend : Instructional expenditure per student
- Grad.Rate : Graduation rate

Before reading the data into R, it can be viewed in Excel or a text editor.

(a) Use the read.csv() function to read the data into R. Call the loaded data college. Make sure that you have the directory set to the correct location for the data.

```
> library(ISLR)
> data("College")
> college<- read.csv("College.csv")
```

(b) Look at the data using the fix() function. You should notice that the first column is just the name of each university. We don't really want R to treat this as data. However, it may be handy tohave these names for later.

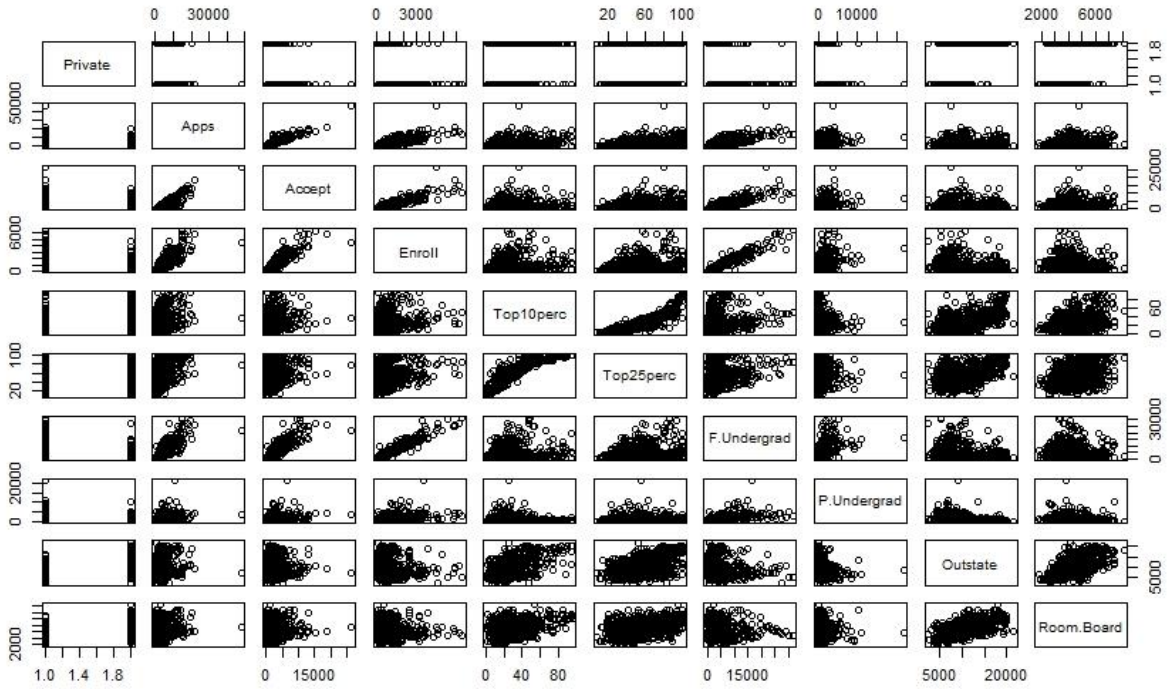
```
> head(college[,1:5])
                            Private Apps Accept Enroll Top10perc
Abilene Christian University     Yes 1660   1232    721       23
Adelphi University               Yes 2186   1924    512       16
Adrian College                   Yes 1428   1097    336       22
Agnes Scott College              Yes  417    349    137       60
Alaska Pacific University        Yes  193    146     55       16
Albertson College                Yes  587    479    158       38
```

(c) i. Use the summary() function to produce a numerical summary of the variables in the data set.

```
> summary(College)
 Private        Apps           Accept          Enroll       Top10perc       Top25perc      F.Undergrad
 No :212   Min.   :   81   Min.   :   72   Min.   :  35   Min.   : 1.00   Min.   :  9.0   Min.   :  139
 Yes:565   1st Qu.:  776   1st Qu.:  604   1st Qu.: 242   1st Qu.:15.00   1st Qu.: 41.0   1st Qu.:  992
           Median : 1558   Median : 1110   Median : 434   Median :23.00   Median : 54.0   Median : 17
07
           Mean   : 3002   Mean   : 2019   Mean   : 780   Mean   :27.56   Mean   : 55.8   Mean   : 3700
           3rd Qu.: 3624   3rd Qu.: 2424   3rd Qu.: 902   3rd Qu.:35.00   3rd Qu.: 69.0   3rd Qu.: 4005
           Max.   :48094   Max.   :26330   Max.   :6392   Max.   :96.00   Max.   :100.0   Max.   :31643
  P.Undergrad       Outstate       Room.Board      Books           Personal         PhD            Terminal
 Min.   :    1.0   Min.   : 2340   Min.   :1780   Min.   :  96.0   Min.   : 250   Min.   :  8.00   Min.   : 24.0
 1st Qu.:   95.0   1st Qu.: 7320   1st Qu.:3597   1st Qu.: 470.0   1st Qu.: 850   1st Qu.: 62.00   1st Qu
.: 71.0
 Median :  353.0   Median : 9990   Median :4200   Median : 500.0   Median :1200   Median : 75.00
Median : 82.0
 Mean   :  855.3   Mean   :10441   Mean   :4358   Mean   : 549.4   Mean   :1341   Mean   : 72.66   Me
an   : 79.7
 3rd Qu.:  967.0   3rd Qu.:12925   3rd Qu.:5050   3rd Qu.: 600.0   3rd Qu.:1700   3rd Qu.: 85.00   3r
d Qu.: 92.0
 Max.   :21836.0   Max.   :21700   Max.   :8124   Max.   :2340.0   Max.   :6800   Max.   :103.00   Max.
:100.0
   S.F.Ratio      perc.alumni        Expend        Grad.Rate
 Min.   : 2.50   Min.   : 0.00   Min.   : 3186   Min.   : 10.00
 1st Qu.:11.50   1st Qu.:13.00   1st Qu.: 6751   1st Qu.: 53.00
 Median :13.60   Median :21.00   Median : 8377   Median : 65.00
 Mean   :14.09   Mean   :22.74   Mean   : 9660   Mean   : 65.46
 3rd Qu.:16.50   3rd Qu.:31.00   3rd Qu.:10830   3rd Qu.: 78.00
 Max.   :39.80   Max.   :64.00   Max.   :56233   Max.   :118.00
```

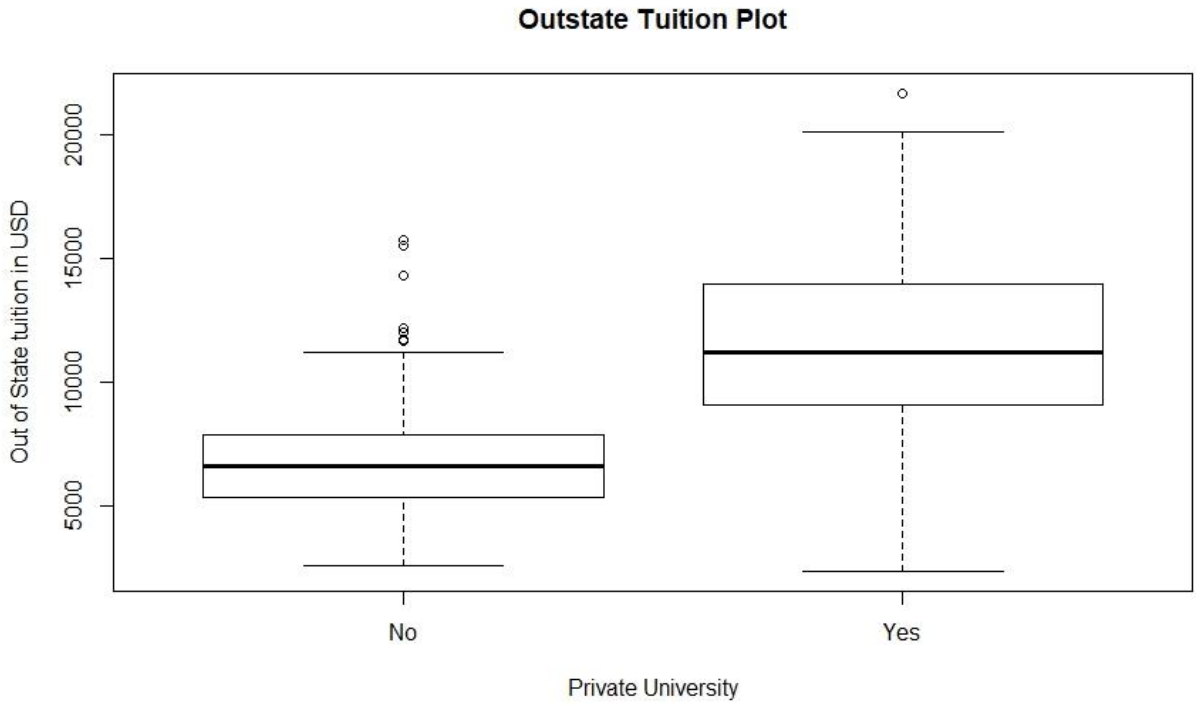
ii. Use the pairs() function to produce a scatterplot matrix of the first ten columns or variables of the data. Recall that you can reference the first ten columns of a matrix A using A[,1:10].

```
> pairs(College[1:10])
```



iii. Use the plot() function to produce side-by-side boxplots of Outstate versus Private.

```
> plot(College$Private, College$Outstate, xlab = "Private University", ylab ="Out of State tuition in USD", main = "Outstate Tuition Plot")
```



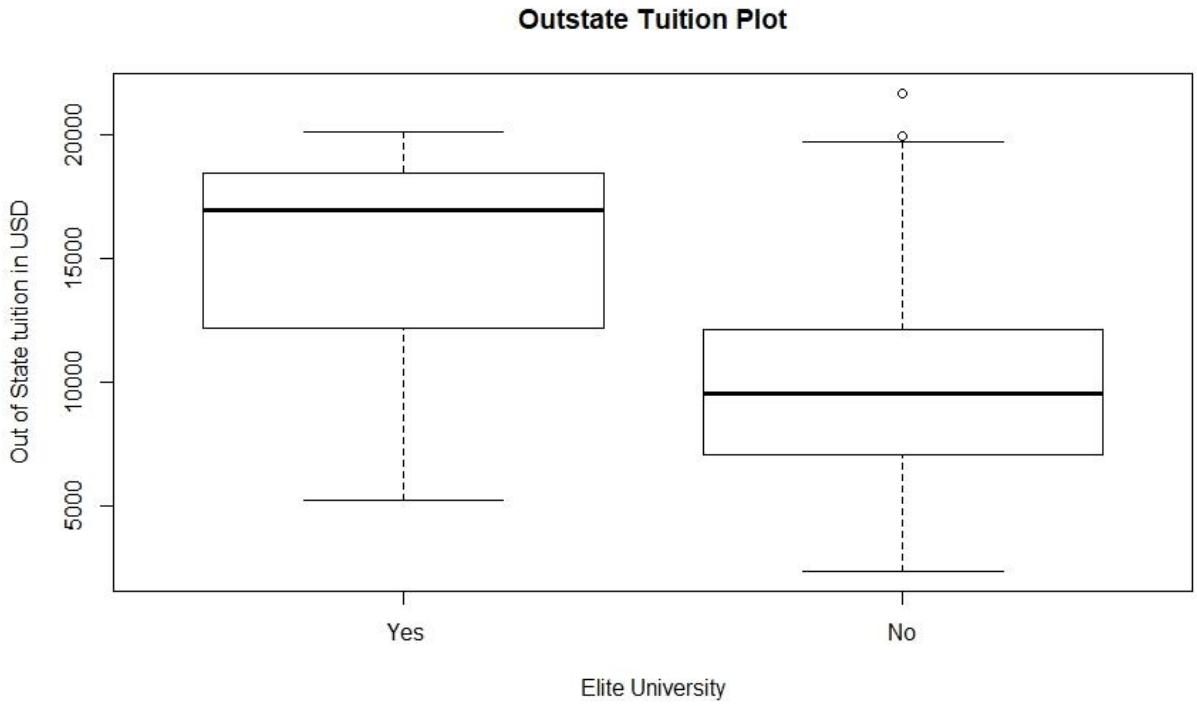
iv. Create a new qualitative variable, called Elite, by *binning* the Top10perc variable. We are going to divide universities into two groups based on whether or not the proportion of students coming from the top 10% of their high school classes exceeds 50%.

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

Use the summary() function to see how many elite universities there are. Now use the plot() function to produce side-by-side boxplots of Outstate versus Elite.

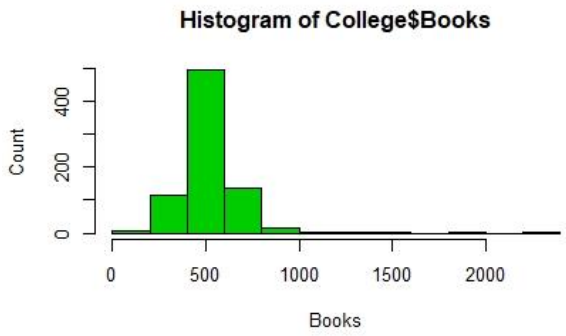
```
> Elite =rep ("No",nrow(college ))
> Elite [college$Top10perc >50]=" Yes"
> Elite =as.factor (Elite)
> college$Elite <- Elite
> summary(college$Elite)
 Yes   No
  78  699
```

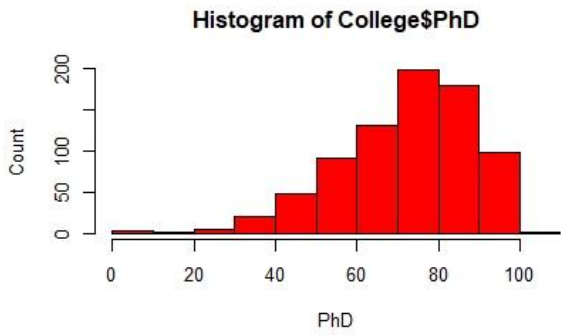
```
> plot(college$Elite, college$Outstate, xlab = "Elite University", ylab ="Out of State tuition in USD", main = "Outstate Tuition Plot")
```

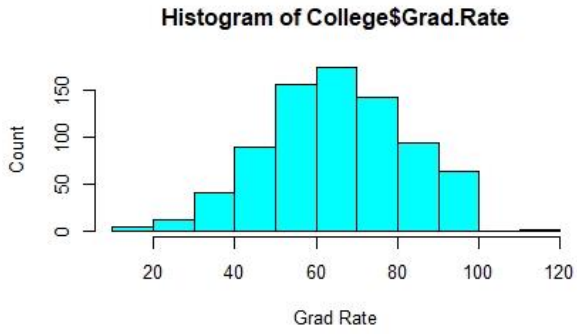


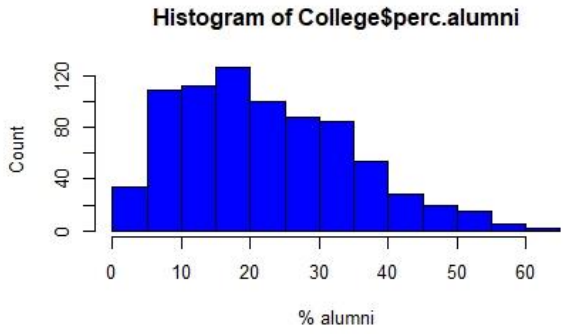
v. Use the hist() function to produce some histograms with differing numbers of bins for a few of the quantitative variables. You may find the command par(mfrow=c(2,2)) useful:
it will divide the print window into four regions so that four plots can be made simultaneously. Modifying the arguments to this function will divide the screen in other ways.

```
> par(mfrow = c(2,2))
> hist(college$Books, col = 3, xlab = "Books", ylab = "Count")
> hist(college$PhD, col = 2, xlab = "PhD", ylab = "Count")
> hist(college$Grad.Rate, col = 5, xlab = "Grad Rate", ylab = "Count")
> hist(college$perc.alumni, col = 4, xlab = "% alumni", ylab = "Count")
```









vi. Continue exploring the data, and provide a brief summary of what you discover.

```
> summary(college$Books)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   96.0   470.0   500.0   549.4   600.0  2340.0
> weird.books <- college[College$Books == 96, ]
> nrow(weird.books)
[1] 1
> row.names(weird.books)
[1] "Appalachian State University"
```

It is weird that the university has a min of only 96 books, Appalachian State University is one of them . A university having only 96 books is weird because there are several streams and the number of books per stream is a big number , so this little number is rare.