ANLY512HW0

Hongyang Zheng

2019/1/28

Question 2.4.2

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.

This is a regression problem and we are most interested in inference since we want to know the influential factors of salary rather than predicting the salary.

n is 500, since the data set is about top 500 firms in the U.S.

p is 3, they are profit, number of employees, industry.

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.

This is a classification problem and we are most interested in prediction since we want to know whether a new product will be a success.

n is 20, since there are 20 similar products in the data.

p is 13, they are price charged for the product, marketing budget, competition price, and ten other variables.

c)

We are interested in predicting the % change in the USD/Euro exchange rate 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 USD/Euro, the % change in the US market, the % change in the British market, and the % change in the German market.

This is a regression problem and we are most interested in prediction since we want to get the quantitative output of the % change in the USD/Euro exchange rate based on the input variables.

n is 52, since there are 52 weeks in a year.

p is 3, they are the % change in the US market, the % change in the British market, and the % change in the German market.

Question 2.4.5

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?

Advantages: a very flexible model can have a better fit of the data, capture the non-linear features in the data and have a smaller bias.

Disadvantages: a very flexible model will have a large number of parameters, which makes it harder to interpret, and it will overfit the data sometimes, as well as have a higher variance.

A more flexible approach will be preferred if there are few parameters in the data compared to the number of observations or the relationship between predictors and response looks non-linear.

A less flexible approach will be preferred if there are enough parameters in the data compared to the number of observations or we are more interested in interpreting the results rather than prediction.

Question Xtra 4

a)

```
# Load data
load("~/Desktop/other/Data/mnist_all.RData")

# New function
myclosest = function(mydigit,k)
{ digit.dist = function(j)
    {
      return(sqrt(mean((test$x[mydigit,] - train$x[j,])^2) ) )
    }
  mnist.distances = sapply(1:60000,FUN = digit.dist)
  myclosest = head(order(mnist.distances),k)
  mytable <- table(train$y[myclosest])
  myindex = which.max(mytable)
  return(as.numeric(names(mytable[myindex])))
}
```

b)

```
# Demonstrate the modified function
# test digit 1221
c(test$y[1221], myclosest(1221, 50))
```

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

```
# Find k
print(c(test$y[9], myclosest(9, 20), myclosest(9, 900)))
```

```
## [1] 5 5 6
```

When k=20, we can get the correct classification that the true number is 5, while when k=900, for the same test digit, we got a wrong classification which said the number is 6.

Question 2.4.8

a)

```
college=read.csv('~/Desktop/other/Data/College.csv')
```

b)

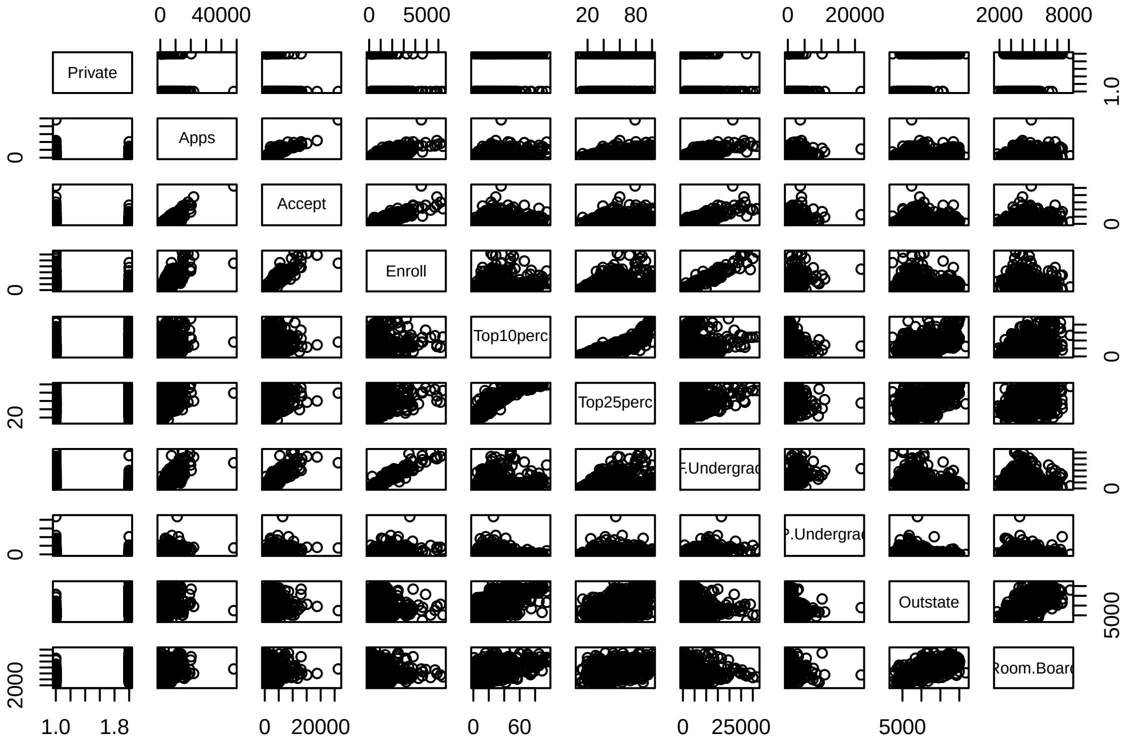
```
rownames(college)=college[,1]
#fix(college)
college=college[,-1]
#fix(college)
```

c)

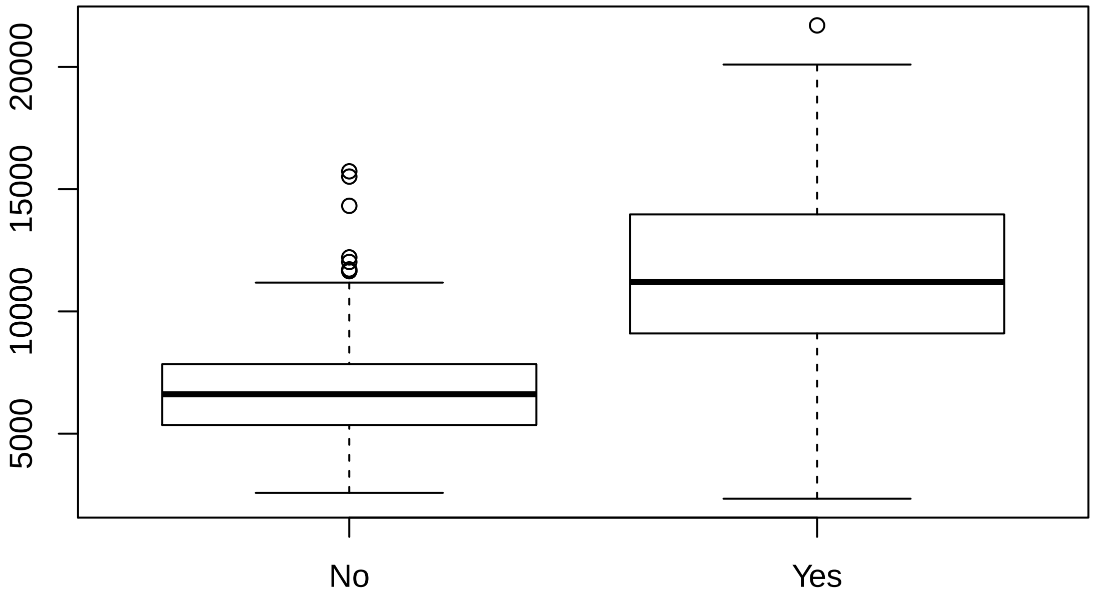
```
# i
summary(college)
```

```
##  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   :27.56
##            3rd Qu.: 3624   3rd Qu.: 2424   3rd Qu.: 902   3rd Qu.:35.00
##            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
```

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



```
# iii
plot(college$Private,college$Outstate)
```

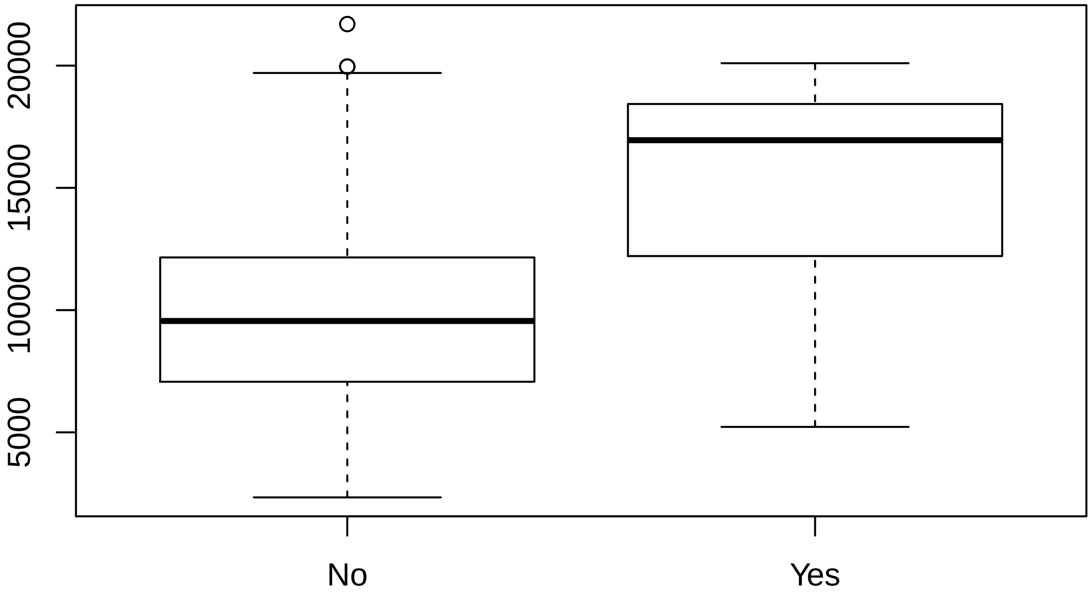


```
# iv
# Create Elite column
Elite=rep("No",nrow(college))
Elite[college$Top10perc >50]="Yes"
Elite=as.factor(Elite)
college=data.frame(college ,Elite)
```

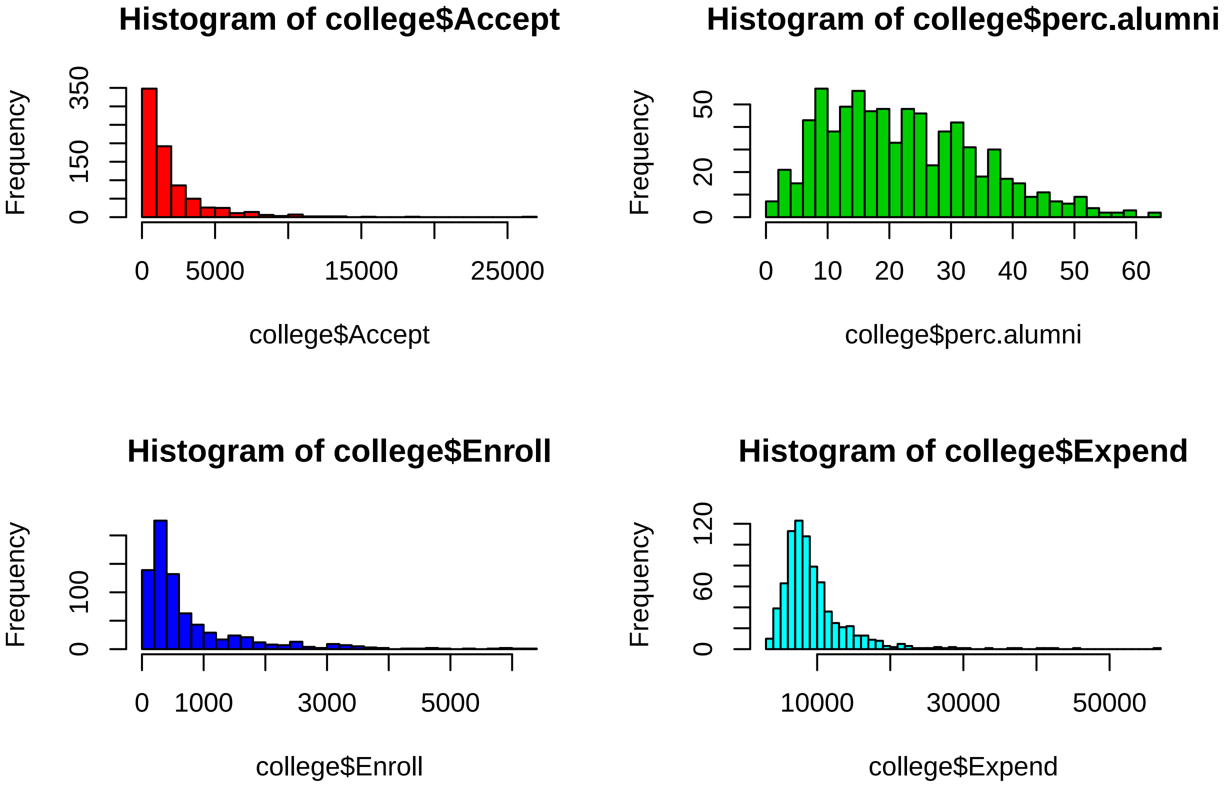
```r
summary(college$Elite)
```

```
##  No Yes
## 699  78
```

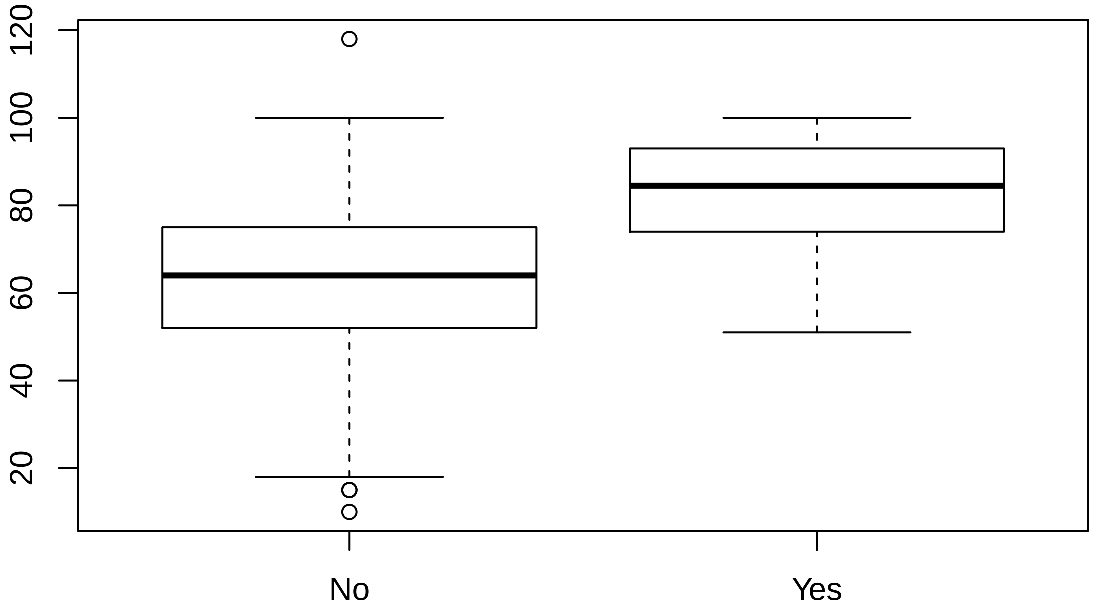
```r
plot(college$Elite, college$Outstate)
```



```r
# v
par(mfrow=c(2,2))
hist(college$Accept, breaks = 20, col=2)
hist(college$perc.alumni, breaks = 30, col=3)
hist(college$Enroll, breaks=40, col=4)
hist(college$Expend, breaks=50, col=5)
```

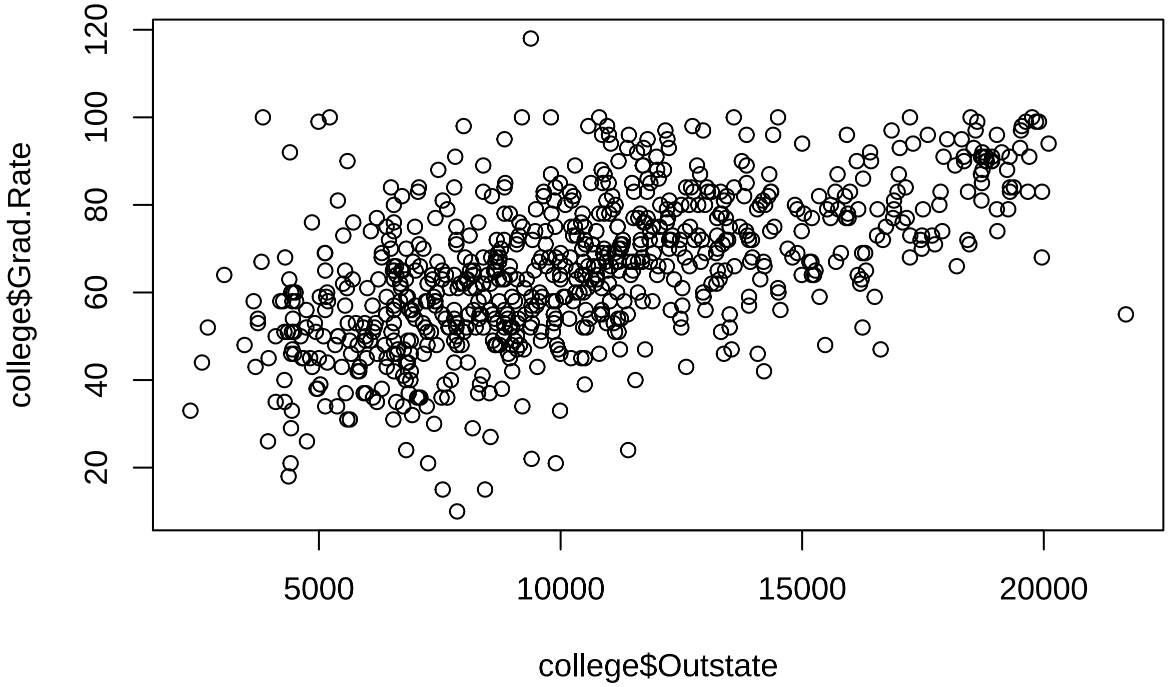


```
# vi
par(mfrow=c(1,1))
plot(college$Elite, college$Grad.Rate)
```



Elite college has a higher graduation rate on average.

```
plot(college$Outstate, college$Grad.Rate)
```



There is a positive relationship between out-of-state tuition and graduation rate. When the tuition is higher, the graduation rate is also higher.

Question Xtra 6

a)

```
AVG_RSSE1=(104.51+101.73+47.21+73.13+67.49
          +117.89+66.37+74.26+90.72+99.53)/10
range1=c(-5,-3)
```

The average residual SSE is 84.284 and the approximate range of the highest order coefficient is -5, -3.

b)

```
AVG_RSSE10=(36.29+19.94+26.56+27.49+3.17
           +29.07+17.52+12.91+9.01+18.36)/10
range10=c(-6100,1300)
```

The average residual SSE is 20.032 and the approximate range of 5th order coefficient is -6100, 1300.

c)

For a low-complexity model, it has a higher bias on average and narrower range of variance, while for a high-complexity model, it has a lower bias on average and wider range of variance. The trade-off is when we increase the complexity of model to reduce bias, the variance of the model will increase; when we employ a simple model to reduce the variance, this model will have bigger bias.

d)

I firstly observed that when complexity=2 or 3, the curve is similar. Then I tried lots of times for each case.

When complexity=3, 7 out of 10 times the curve will be close to the true model.

When complexity=2, 9 out of 10 times the curve will be close to the true model.

Since the residual SSE is similar, I would choose complexity=2 to include less variance and for easy interpretation.