Home1

Alphonso J Saiewane

Q1

For each of parts (a) through (d), indicate whether we would generally expect the performan of a flexible statistical learning method to be better or worse than an inflexible method. Justify your answer.

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

Answer: A flexible statistical learning method is generally expected to perform **better** in this scenario. When the sample size n is extremely large, the model has sufficient data to accurately estimate the parameters, reducing the risk of overfitting.

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

Answer: inflexible methods are generally expected to perform better. When p is large, and n is small, an **inflexible model** performs better because it reduces the risk of overfitting. Inflexible models, with fewer parameters, offer lower variance and more reliable predictions, even if they introduce some bias.

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

Answer: a **flexible statistical learning method** is better because it allows you to use enough degrees of freedom to accurately model the complex, non-linear relationship.

(d) The variance of the error terms, i.e. ^{2} = Var( ), is extremely high.

Answer: When the variance of the error terms is extremely high, an **inflexible statistical learning method** is generally better because it produces lower variance in the predictions. Inflexible models are less likely to overfit the noisy data, resulting in more stable and reliable predictions, even if they might introduce some bias.

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, where n is the sample size and p is the number of independent variables.

(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.

Answer:

- **Regression** we're analyzing how continuous variables (like CEO salary) relate to other variables (profit, number of employees, and industry).
- **Inference** the goal is to understand the relationships between the predictors and CEO salary, rather than just predicting salary.
- **Sample Size (n)**: 500 (since data is collected on 500 firms).
- **Number of Independent Variables (p)**: 3 (profit, number of employees, and 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

Answer:

- **Classification**: we're determining a categorical outcome (success or failure).
- **Prediction: Prediction** because the goal is to predict whether the new product will succeed based on the available data.

- **Sample Size (n)**: 20 (as data is collected on 20 similar products).
- **Number of Independent Variables (p)**: 13 (which includes the price charged for the product, marketing budget, competition price, and ten other variables).

(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.

Answer:

- **Type of Problem: Regression** because we're predicting a continuous outcome (the % change in the US dollar).
- **Inference or Prediction: Prediction** since the goal is to forecast the percentage change in the US dollar based on changes in the world stock markets.
- **Sample Size (n)**: 52 (weekly data collected for the year 2012).
- **Number of Independent Variables (**p): 3 (the % change in the US market, the % change in the British market, and the % change in the German market).

Q3

You will now think of some real-life applications for statistical learning. In each example, describe the response and the predictors and state the goal- inference or prediction.

(a) Describe two real-life applications in which classification might be useful.

Answer:

1. **Loan Default Prediction:**

- **Response:** Loan default status (default or no default).**Predictors:** Credit score, income, loan amount, employment status, debt-to-income ratio.**Goal:** Predict default risk to help manage lending decisions and mitigate financial risk.

2. **Image Recognition:**

- **Response:** The object identified in an image (categorical outcome: e.g., cat, dog, car, etc.). **Predictors:** Pixel values, color patterns, shapes, and textures within the image. **Goal: Prediction** – To classify images into categories based on their content, used in applications such as facial recognition and autonomous driving.

(b) Describe two real-life applications in which regression might be useful.

Answer:

1. **Stock Price Prediction:**
2. **Response:** Future stock price. **Predictors:** Historical prices, trading volume, economic indicators, earnings reports, interest rates. **Goal:** Predict future stock prices to guide investment decisions.

2. **Health Care Cost Estimation:**

- **Response:** Total medical costs. **Predictors:** Age, medical history, doctor visits, insurance type, treatment received.**Goal:** Predict healthcare costs to aid in budgeting and pricing.

Q4

This exercise relates to the College data set from our textbook. 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 20% 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.
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

(a) Read the data into R, for example, directly from the package that comes with our book:

install.packages("ISLR2"); library(ISLR2); attach(College);

```
library(ISLR2)
```

```
Warning: package 'ISLR2' was built under R version 4.3.3
```

```
attach(College)
```

(b) Use the summary function to produce a summary of all the variables in the data set.

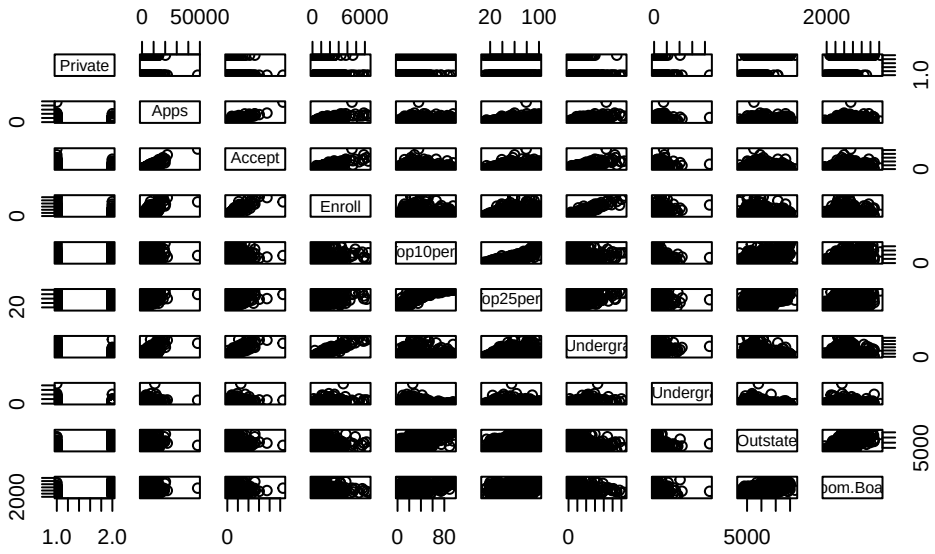
```
summary(College)
```

```
 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
```

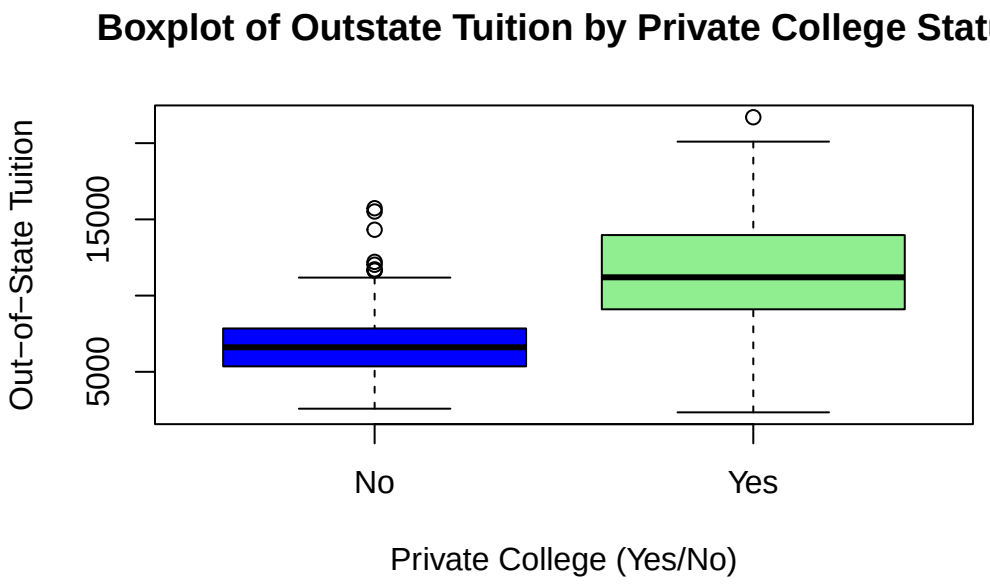
(c)Use the pairs() function to produce a scatterplot matrix of the first ten columns or variables of the data. Recall that you can reference the first ten columns of a matrix A using A[,1:10]

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



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

```
plot(Private, Outstate,
     xlab = "Private College (Yes/No)",
     ylab = "Out-of-State Tuition",
     col=c('blue', 'lightgreen'),
     main = "Boxplot of Outstate Tuition by Private College Status")
```



(e) 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)

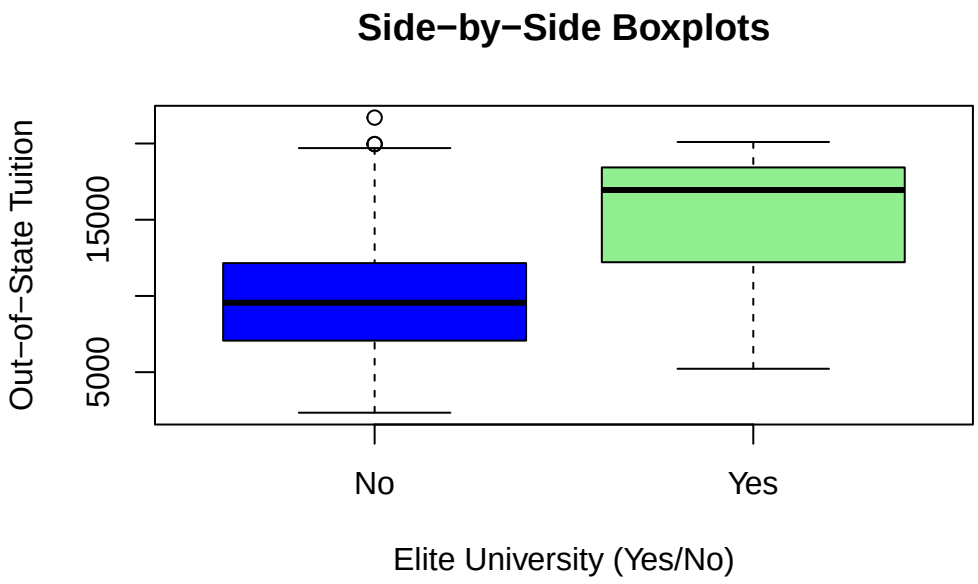
Use the summary() function to see how many elite universities there are. Now use the plot() function to produce side-by-side boxplots of Outstate versus Elite.

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

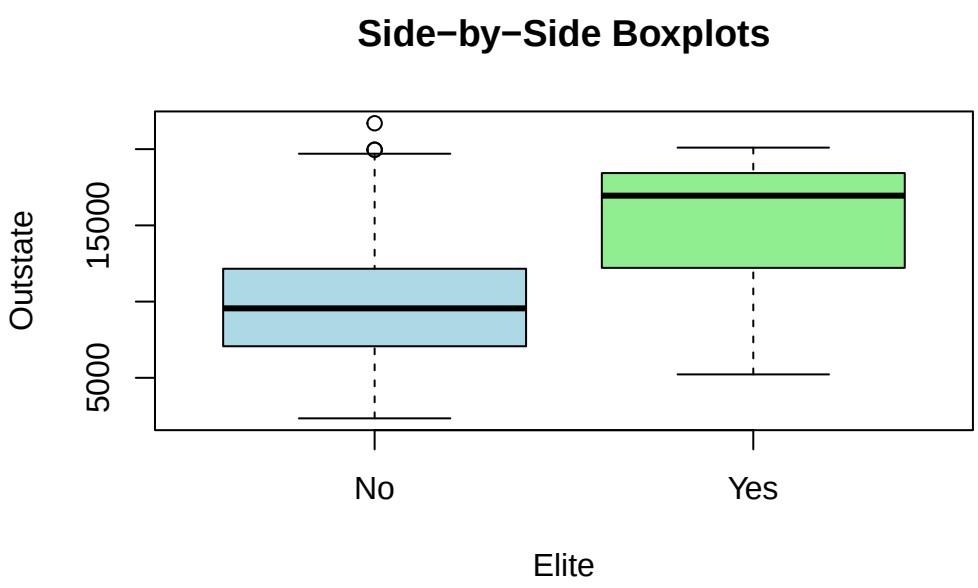
```
#Use the summary() function to see how many elite universities there are
summary(College$Elite)
```

```
 No Yes
699  78
```

```r
#Now use the plot() function to produce side-by-side boxplots of Outstate versus Elite.
plot(College$Elite, College$Outstate,
     xlab = "Elite University (Yes/No)",
     ylab = "Out-of-State Tuition",
     main = "Side-by-Side Boxplots",
     col = c('blue', 'lightgreen'))
```




```r
boxplot(Outstate ~ Elite, data = College, col=c('lightblue', 'lightgreen'), main = "Side-by-S
```



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(3, 2),mar = c(3, 2, 2, 1))

# six quantitative variables with different numbers of bins

# Histogram for Apps (Number of applications received) with 20 bins
hist(College$Apps, main = "Histogram of Applications (Apps)",
     xlab = "Number of Applications",
     col = "lightblue",
     breaks = 20)

# Histogram for Enroll (Number of students enrolled) with 15 bins
hist(College$Enroll,
     main = "Histogram of Enrollments (Enroll)",
     xlab = "Number of Students Enrolled",
```

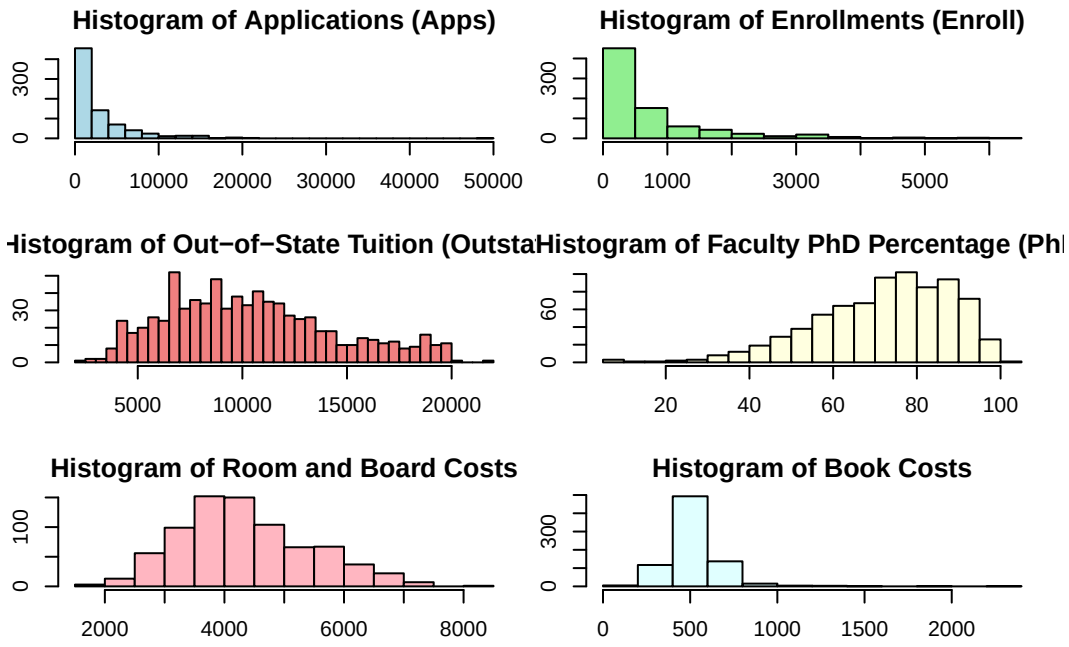
```
     col = "lightgreen",
     breaks = 15)

# Histogram for Outstate (Out-of-state tuition) with 30 bins
hist(College$Outstate,
     main = "Histogram of Out-of-State Tuition (Outstate)",
     xlab = "Out-of-State Tuition",
     col = "lightcoral",
     breaks = 30)

# Histogram for PhD (Percentage of faculty with PhDs) with 25 bins
hist(College$PhD,
     main = "Histogram of Faculty PhD Percentage (PhD)",
     xlab = "Percentage of Faculty with PhDs",
     col = "lightyellow",
     breaks = 25)

# Histogram for Room.Board (Room and board costs) with 20 bins
hist(College$Room.Board,
     main = "Histogram of Room and Board Costs",
     xlab = "Room and Board Costs",
     col = "lightpink",
     breaks = 20)

# Histogram for Books (Book costs) with 15 bins
hist(College$Books,
     main = "Histogram of Book Costs",
     xlab = "Book Costs",
     col = "lightcyan",
     breaks = 15)
```




```
par(mfrow = c(1, 1))
```

(g) Use the lm function to find a regression equation predicting the number of new students

based on the graduation rate, qualifications of the faculty, and various expenses.

```
#   linear regression model
model <- lm(Enroll ~ Grad.Rate + PhD + Terminal + Outstate + Room.Board + Books + Personal +

# summary of the regression model
summary(model)
```

```
Call:
lm(formula = Enroll ~ Grad.Rate + PhD + Terminal + Outstate +
    Room.Board + Books + Personal + Expend, data = College)

Residuals:
    Min      1Q  Median      3Q     Max
-1761.9  -454.9  -122.8   222.1  5127.3
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.236e+03  2.155e+02  -5.738 1.38e-08 ***
Grad.Rate    4.282e+00  2.074e+00   2.064  0.03932 *
PhD          1.476e+01  3.407e+00   4.332 1.67e-05 ***
Terminal     1.084e+01  3.809e+00   2.847  0.00453 **
Outstate    -9.386e-02  1.227e-02  -7.652 5.95e-14 ***
Room.Board   8.804e-03  3.549e-02   0.248  0.80418
Books        3.038e-01  1.813e-01   1.676  0.09422 .
Personal     2.657e-01  4.634e-02   5.733 1.41e-08 ***
Expend       2.254e-02  7.799e-03   2.890  0.00396 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 798.8 on 768 degrees of freedom
Multiple R-squared:  0.2685,    Adjusted R-squared:  0.2609
F-statistic: 35.25 on 8 and 768 DF,  p-value: < 2.2e-16
```