Chapter2-Exercises

Anthony Chau

12/8/2019

Contents

Conceptual Exercises

Question 1

For each of parts (a) through (d), indicate whether we would generally expect the performance 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.

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

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

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

Solution:

(a) We would expect the performance of a flexible statistical learning method to be better than an inflexible method because the flexible methods can model the data more accurately given the large amount of observations. Also, since the number of predictors is small, the flexible methods would not estimate a large number of parameters - increasing overall interprebility.

(b) The performance of flexible methods is worse than inflexible methods. Given the large number of predictors and low number of observations, flexible methods would most likely result in overfitting as well as have low interprebility.

(c) The performance of flexible methods is better than inflexible methods because the flexible methods can better capture non-linear relationships over inflexible methods (such as linear regression).

(d) Flexible methods would perform better than inflexible methods because the flexible methods could include more potential variables that could be useful in predicting the response. The inclusion of those variables can offset the presence of a high variance for the error terms.

Question 2

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.

(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.

(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.

Solution:

(a) This scenario is a regression problem because we are predicting a quantitative variable, CEO salary. We are mainly interested in inference - understanding the factors that affect CEO salary. For this case, n = 500 and p = 3.

(b) This scenario is a classification problem because we are predicting a qualitative variable - whether the product is a success or a failure. We are mainly interested in predicting - correctly determining if a new product will be a success or a failure. For this case, n = 20 and p = 13.

(c) This scenario is a regression problem because we are predicting a quantitative variable, USD/Euro exchange rate. We are mainly interested in prediction - what the USD/Euro exchange rate is given some predictor variables. For this case, n = 52 (52 weeks in a year) and p = 4.

Question 3

3. We now revisit the bias-variance decomposition.

(a) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The x-axis should represen the amount of flexibility in the method, and the y-axis should represent the values for each curve. There should be five curves. Make sure to label each one.

(b) Explain why each of the five curves has the shape displayed in part (a).

Solution:

(a)

(b)

Question 4

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.

(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.

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

Solution:

(a) One application could be to determine if an applicant will default on a loan. The response is Yes/No, indicating whether the applicant defaulted on a loan. Potential explanatory variables could include credit score, age, income, employment status, number of dependents, etc. The goal of this application is prediction - it is important to predict if an applicant will default on a loan!

Another application could be to determine if a person would move up in socioeconomic status. The response is Yes/No, indicating whether the applicant moved up (Yes) or down (No) from their SES status when they were born. Potential explantatory variables could be parent's education level, parent's employment, household income, born in city or rural area, number of siblings, etc. The main goal of this application is inference. We would like to understand the relationship between the response and the explanatory variables.

Yet another application is to determine if a basketball team will win the championship game. The response is Yes/No, indicating if the team will win the championship game. Potential explanatory variables could include: season record, average number of points scored during season, whether team won a championship in previous year(s), etc. This application could be both for inference and prediction. It is an interesting question to pose: what leads to basketball teams winning championship games? On the other hand, you may be betting with your buddies on the championship game, so your money is on the line!

(b) One application of regression is to predict the yearly sales for a clothing company (response variable). Potential explanatory variables could include: past year sales, advertising expenditure, number of physical stores, number of unique visitors to website, etc.

Another application of regression is to model the relationship between age of death (response) with the following explanatory variables: average minutes spent exercising per week, time spent engaging in mental activities, average number of servings of fruits and vegetables, smoker status, weight, etc. This application focuses on inferences - understanding the relationship between response and predictors.

Another application of regression could be to model the number of delayed flights at an airport in a day(response). Potential predictors could include: average daily passengers total, location of airport, domestic/international airport status, average number of thunderstorms/severe weather incidents occured per year, number of average daily nonstop passengers, number of TSA checkpoints, etc. The main interest of this application is inference: what are the key factors which contribute to long airport delays.

(c) One application of cluster analysis is market segmentation of customers when a company wants to introduce a product to a new market. Cluster analysis can help the company gain insight into which customers would be most likely to be interested in the product or buy more of the product.

Another application of cluster analysis is to identify regions with similar weathers by using info such as latitude/longitude, humidity, temperature, sea level, altitude, etc.

Lastly, another application of cluster analysis could be to group together students within a school based on criterion such as age, ethnicity, family income, GPA, etc.

Question 5

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?

Solution:

The advantage of a very flexible model is that the model will have low bias because it can model the data precisely. Another advantage of flexible models is that they can model more complex relationship between variables. Undoubtedly, any real-world relationships are not constant or linear.

A disadvantage of a very flexible model is the large number of parameters that need to be estimated. This makes the model less interpretable. This problem is exaceberated if we have a low number of observations but a high number of predictor variables. Another disadvantage of a flexible model is that it will most likely have high variance. The model will have a low train MSE, but will have a high test MSE since the model will struggle to accurately predict given brand new observations. The model was taught to adhere too closely to the training data which causes it to struggle with new test data.

A more flexible approach is preferred if the main goal is accurate prediction. Perhaps, accurate prediction would help a company's bottom line or provide a critical diagonosis on a patient.

On the other hand, a less flexible approach is preferred if the main goal is inference. That is, we value interprebility and seek to understand the relationship between variables.

Question 6

6. Describe the differences between a parametric and a non-parametric statistical learning approach. What are the advantages of a parametric approach to regression or classification (as opposed to a nonparametric approach)? What are its disadvantages?

Solution:
Parametric statistical methods make assumptions about the functional form of f and reduce the problem of estimating f to solving for a set of parameters For example, linear regression is a parametric method which assumes that f is linear in its predictors X. To implement linear regression, one can use least squares to solve for β.

Nonparametric statistical methods do not make assumptions about the functional form of f. Instead, they attempt to estimate f through mathematical techniques.

An advantage of a parametric method is that it is easy to estimate a set of parameters compared to estimating an entire function for nonparametric methods.

A disadvantage of a parametric method is that the true form of f may be completely different than what the stated assumptions of the parametic method. This is where nonparametric methods shine as they can estimate a greater variety of functional forms of f. However, because of the high flexibility of nonparametric methods, these methods can result in high variance and overfitting. Nonparametrics methods require a very large number of observations to produce accurate estimate because they cannot reduce problem to estimating a small set of coefficients.

Question 7

7. The table below provides a training data set containing six observations, three predictors, and one qualitative response variable.

Suppose we wish to use this data set to make a prediction for Y when X1 = X2 = X3 = 0 using K-nearest neighbors. (a) Compute the Euclidean distance between each observation and the test point, X1 = X2 = X3 = 0. (b) What is our prediction with K = 1? Why? (c) What is our prediction with K = 3? Why? (d) If the Bayes decision boundary in this problem is highly nonlinear, then would we expect the best value for K to be large or small? Why?

Solution:

The dataset is recreated below.

```r
df <- data.frame(X1 = c(0, 2, 0, 0, -1, 1), X2 = c(3, 0, 1, 1, 0, 1),
                 X3 = c(0, 0, 3, 2, 1, 1),
                 Y = c("Red", "Red", "Red", "Green", "Green", "Red"))

# Test Point
TestPoint <- c(0, 0, 0)

df <- rbind(df, TestPoint)
```

```
## Warning in `[<-.factor`(`*tmp*`, ri, value = 0): invalid factor level, NA
## generated
```

```r
# calculate euclidean distance
# last row is eucliden distance of test point to all observations
# in training set
dist(df[1:7, 1:3], method = "euclidean")
```

```
##          1        2        3        4        5        6
## 2 3.605551
## 3 3.605551 3.741657
## 4 2.828427 3.000000 1.000000
## 5 3.316625 3.162278 2.449490 1.732051
## 6 2.449490 1.732051 2.236068 1.414214 2.236068
## 7 3.000000 2.000000 3.162278 2.236068 1.414214 1.732051
```

(b) When K = 1, our predicition is Green. Because the closest observation to the test point is observation 5 and it is Green.

(c) When K = 3, our prediction is Green. The three closest observations to the test point are observations 4, 5, and 6. Observations 4 and 5 are Green and observation 6 is Red. So the estimated probability for the Green Class is 2/3 and K-Nearest Neighbors predicts that the test point belongs to the Green class.

Applied Exercises

Question 8

Solution:

(a) Load data

```
library(here)
```

```
## here() starts at C:/Users/Anthony Chau/Documents/ISLR-exercises
```

```
college <- read.csv(here("Chapter2-StatisticalLearning", "College.csv"))
```

(b) View data

```
# opens a data editor
rownames(college) <- college[ , 1]
fix(college)
college <- college[ ,-1]
fix(college)
```

(c)

i. Summary of all the variabes

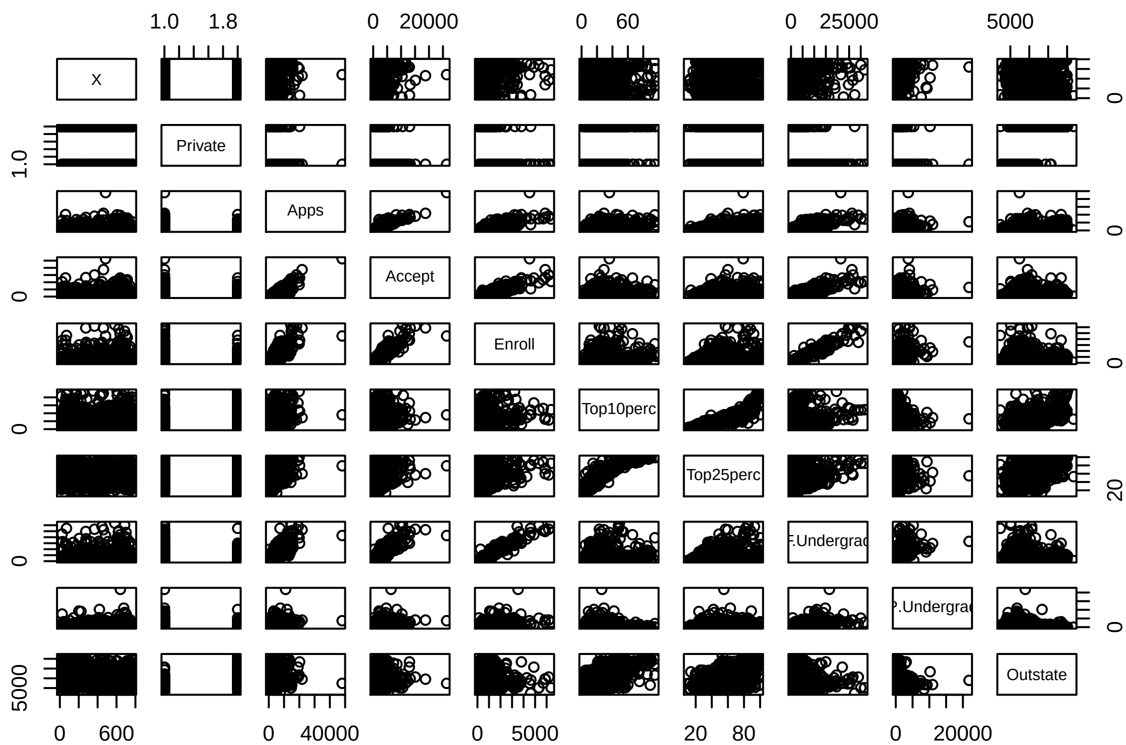
```
summary(college)
```

```
##                              X        Private        Apps           Accept
##  Abilene Christian University:  1   No :212    Min.   :   81   Min.   :   72
##  Adelphi University          :  1   Yes:565    1st Qu.:  776   1st Qu.:  604
##  Adrian College              :  1              Median : 1558   Median : 1110
##  Agnes Scott College         :  1              Mean   : 3002   Mean   : 2019
##  Alaska Pacific University   :  1              3rd Qu.: 3624   3rd Qu.: 2424
##  Albertson College           :  1              Max.   :48094   Max.   :26330
##  (Other)                     :771
##      Enroll        Top10perc       Top25perc      F.Undergrad
##  Min.   :  35   Min.   : 1.00   Min.   :  9.0   Min.   :  139
##  1st Qu.: 242   1st Qu.:15.00   1st Qu.: 41.0   1st Qu.:  992
##  Median : 434   Median :23.00   Median : 54.0   Median : 1707
##  Mean   : 780   Mean   :27.56   Mean   : 55.8   Mean   : 3700
##  3rd Qu.: 902   3rd Qu.:35.00   3rd Qu.: 69.0   3rd Qu.: 4005
##  Max.   :6392   Max.   :96.00   Max.   :100.0   Max.   :31643
##
##   P.Undergrad         Outstate       Room.Board       Books
##  Min.   :    1.0   Min.   : 2340   Min.   :1780   Min.   :  96.0
##  1st Qu.:   95.0   1st Qu.: 7320   1st Qu.:3597   1st Qu.: 470.0
##  Median :  353.0   Median : 9990   Median :4200   Median : 500.0
##  Mean   :  855.3   Mean   :10441   Mean   :4358   Mean   : 549.4
##  3rd Qu.:  967.0   3rd Qu.:12925   3rd Qu.:5050   3rd Qu.: 600.0
##  Max.   :21836.0   Max.   :21700   Max.   :8124   Max.   :2340.0
##
```

```
##     Personal         PhD             Terminal        S.F.Ratio
##  Min.   : 250   Min.   :  8.00   Min.   : 24.0   Min.   : 2.50
##  1st Qu.: 850   1st Qu.: 62.00   1st Qu.: 71.0   1st Qu.:11.50
##  Median :1200   Median : 75.00   Median : 82.0   Median :13.60
##  Mean   :1341   Mean   : 72.66   Mean   : 79.7   Mean   :14.09
##  3rd Qu.:1700   3rd Qu.: 85.00   3rd Qu.: 92.0   3rd Qu.:16.50
##  Max.   :6800   Max.   :103.00   Max.   :100.0   Max.   :39.80
##
##   perc.alumni        Expend         Grad.Rate
##  Min.   : 0.00   Min.   : 3186   Min.   : 10.00
##  1st Qu.:13.00   1st Qu.: 6751   1st Qu.: 53.00
##  Median :21.00   Median : 8377   Median : 65.00
##  Mean   :22.74   Mean   : 9660   Mean   : 65.46
##  3rd Qu.:31.00   3rd Qu.:10830   3rd Qu.: 78.00
##  Max.   :64.00   Max.   :56233   Max.   :118.00
##
```

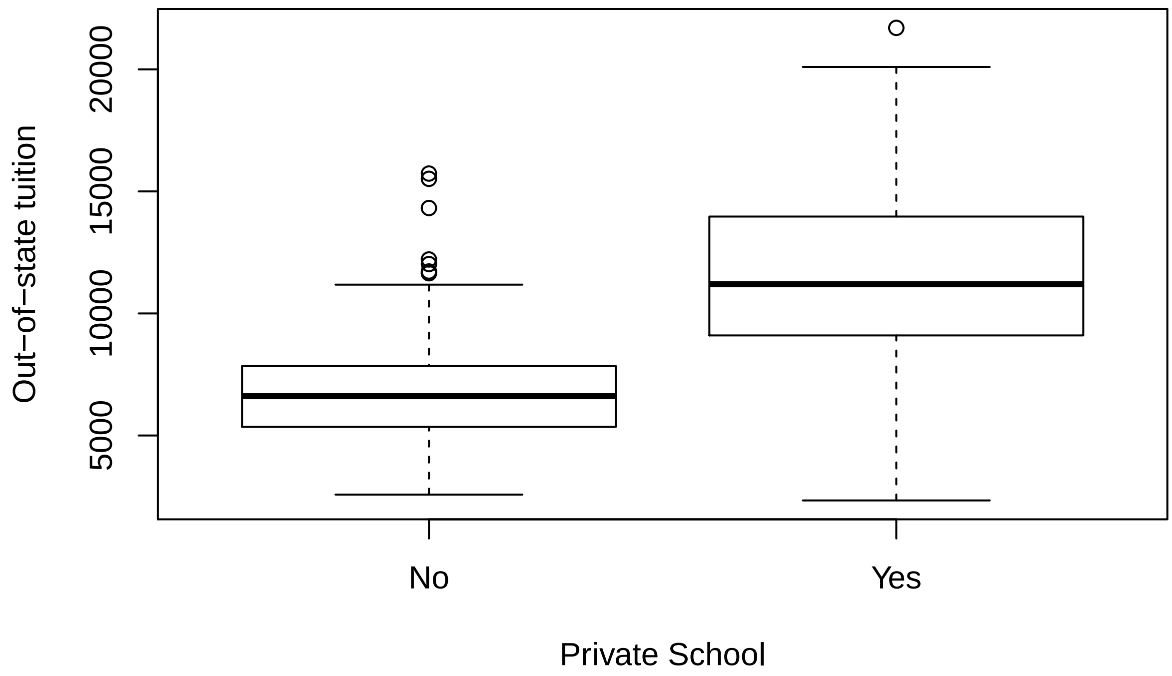
ii. Scatterplot matrix

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



iii. Boxplot of `Outstate` vs `Private`

```r
plot(college$Private, college$Outstate,
     xlab = "Private School",
     ylab = "Out-of-state tuition")
```



iv.

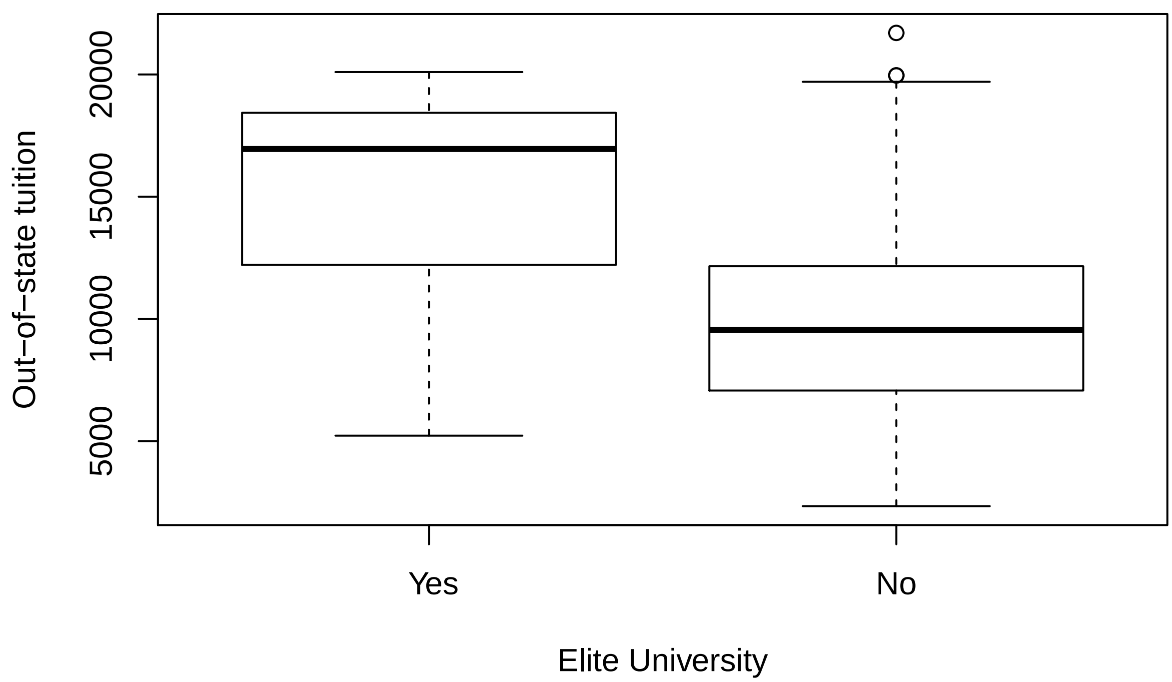
After creating the new `Elite` column, we find that there are 78 elite universites.

```r
# create vector of "No" with length of the dataset
Elite <- rep ("No",nrow(college ))

# University is elite if the top 10% of the high
# school class exceeds 50%
Elite[college$Top10perc > 50] <- " Yes"

# coerce to factor
Elite = as.factor(Elite)

# bind to data frame
college <- data.frame(college, Elite)

summary(college$Elite)
```

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

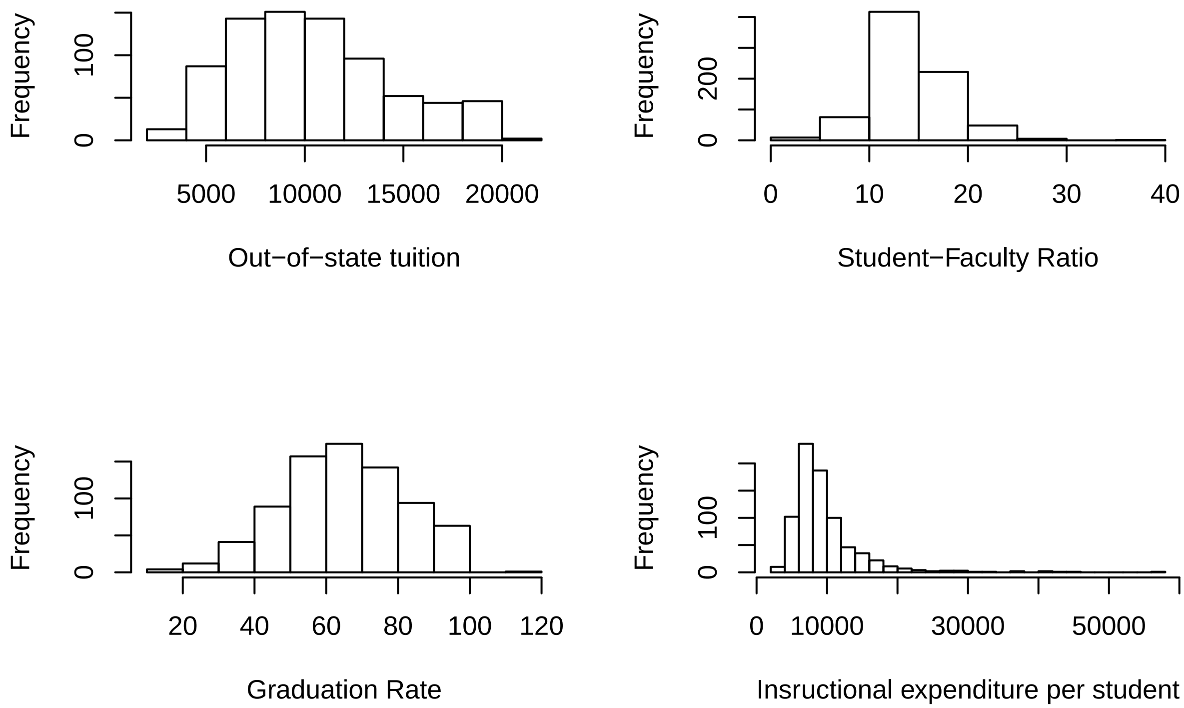
By plotting elite status against out-of state tuition, we find that elite universities tend to have higher out-of-state tuition costs.

```
plot(college$Elite, college$Outstate,
     xlab = "Elite University",
     ylab = "Out-of-state tuition")
```



(v) Some histograms of the quantitative variables

```
par(mfrow=c(2,2))
hist(college$Outstate, xlab = "Out-of-state tuition", main = "")
hist(college$S.F.Ratio, xlab = "Student-Faculty Ratio", main = "")
hist(college$Grad.Rate, xlab = "Graduation Rate", main = "")
hist(college$Expend, xlab = "Insructional expenditure per student", main = "", breaks = 20)
```


Frequency
0
100
5000
10000
15000
20000
Out−of−state tuition
Frequency
0
200
0
10
20
30
40
Student−Faculty Ratio
Frequency
0
100
20
40
60
80
100
120
Graduation Rate
Frequency
0
100
0
10000
30000
50000
Insructional expenditure per student

Question 9

Solution: (a) The quantitative varibles are: `mpg`, `cylinders`, `displacement`, `horsepower`, `weight`, `acceleration`, `year`, and `origin` The sole qualitiative variables is `name`

```
auto <- ISLR::Auto
# determine quantiative variables
sapply(auto, is.numeric)
```

```
##          mpg    cylinders displacement   horsepower       weight acceleration 
##         TRUE         TRUE         TRUE         TRUE         TRUE         TRUE 
##         year       origin         name 
##         TRUE         TRUE        FALSE
```

(b) Range of numerical variables

```
# auto data with only numeric columns
auto_numeric <- auto[, purrr::map_lgl(auto, is.numeric)]
t(sapply(auto_numeric, range))
```

```
##              [,1]   [,2]
## mpg             9   46.6
## cylinders       3    8.0
## displacement   68  455.0
## horsepower     46  230.0
## weight       1613 5140.0
## acceleration    8   24.8
## year           70   82.0
## origin          1    3.0
```

(c) Mean and standard deviation of numerical variables

```
sapply(auto_numeric, mean)
```

```
##          mpg    cylinders displacement   horsepower       weight acceleration 
##    23.445918     5.471939   194.411990   104.469388  2977.584184    15.541327 
##         year       origin 
##    75.979592     1.576531
```

```
sapply(auto_numeric, sd)
```

```
##          mpg    cylinders displacement   horsepower       weight acceleration 
##    7.8050075    1.7057832  104.6440039   38.4911599  849.4025600    2.7588641 
##         year       origin 
##    3.6837365    0.8055182
```

(d) Remove some observations. Recompute mean and sd.

```r
auto_numeric_sparse <- auto_numeric[-c(10:85),]
t(sapply(auto_numeric_sparse, range))
```

```
##                 [,1]   [,2]
## mpg             11.0   46.6
## cylinders        3.0    8.0
## displacement    68.0  455.0
## horsepower      46.0  230.0
## weight        1649.0 4997.0
## acceleration     8.5   24.8
## year            70.0   82.0
## origin           1.0    3.0
```

```r
sapply(auto_numeric_sparse, mean)
```

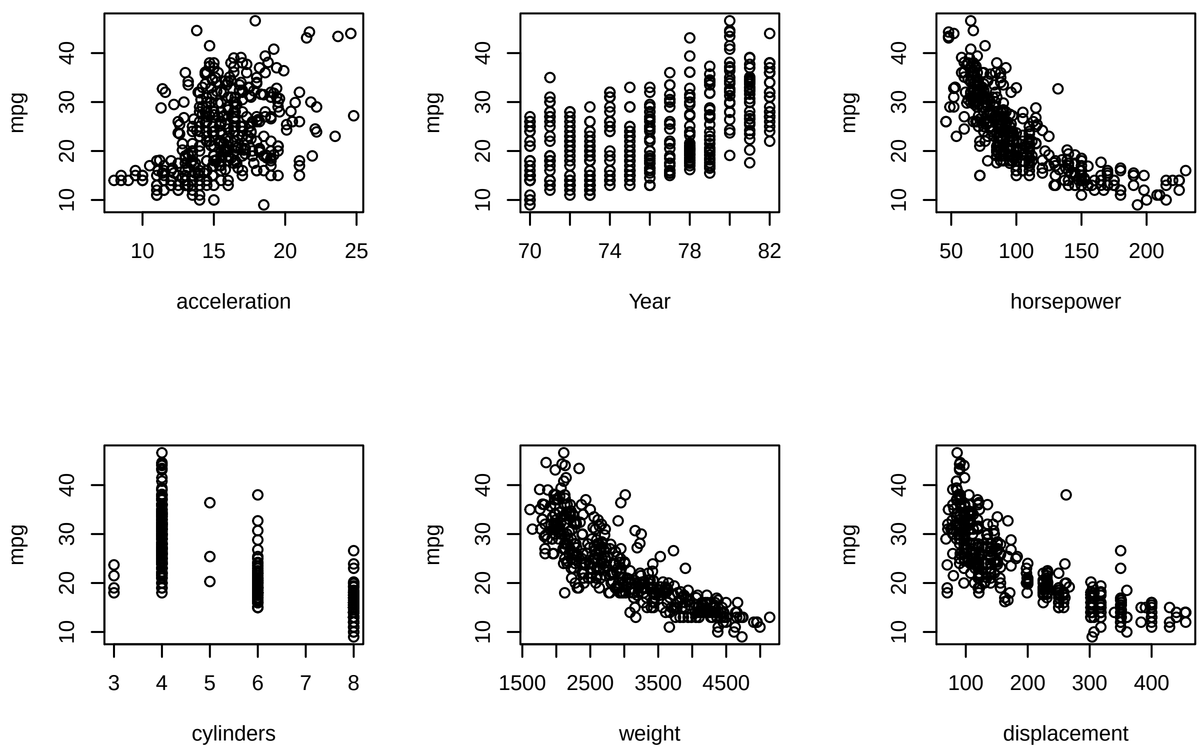
```
##          mpg    cylinders displacement   horsepower       weight acceleration 
##    24.404430     5.373418   187.240506   100.721519  2935.971519    15.726899 
##         year       origin 
##    77.145570     1.601266
```

```r
sapply(auto_numeric_sparse, sd)
```

```
##          mpg    cylinders displacement   horsepower       weight acceleration 
##     7.867283     1.654179    99.678367    35.708853   811.300208     2.693721 
##         year       origin 
##     3.106217     0.819910
```

(e) Exploratory analysis

```r
par(mfrow=c(2,3))
plot(auto$acceleration, auto$mpg,
     xlab = "acceleration",
     ylab = "mpg")
plot(auto$year, auto$mpg,
     xlab = "Year",
     ylab = "mpg")
plot(auto$horsepower, auto$mpg,
     xlab = "horsepower",
     ylab = "mpg")
plot(auto$cylinders, auto$mpg,
     xlab = "cylinders",
     ylab = "mpg")
plot(auto$weight, auto$mpg,
     xlab = "weight",
     ylab = "mpg")
plot(auto$displacement, auto$mpg,
     xlab = "displacement",
     ylab = "mpg")
```


mpg
acceleration
Year
horsepower
cylinders
weight
displacement

Question 10

Solution: