Lab 1: Introduction To Statistical Learning

ACTL3142 and ACTL5110

Questions

Conceptual Questions

1. ⋆ (ISLR2, Q2.1) 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., $\sigma^2 = \mathbb{V}(\epsilon)$, is extremely high.

   Solution

2. (ISLR2, Q2.2) 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

3. ⋆ (ISLR2, Q2.3) 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 represent 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

4. ⋆ (ISLR2, Q2.5) 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

5. ⋆ (ISLR2, Q2.7) The table below provides a training data set containing six observations, three predictors, and one qualitative response variable.

Obs.	X_1	X_2	X_3	Y
1	0	3	0	Red
2	2	0	0	Red
3	0	1	3	Red
4	0	1	2	Green
5	-1	0	1	Green
6	1	1	1	Red

   Suppose we wish to use this data set to make a prediction for Y when $X_1 = X_2 = X_3 = 0$ using K-nearest neighbors.

   a. Compute the Euclidean distance between each observation and the test point, $X_1 = X_2 = X_3 = 0$.

   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 non-linear, then would we expect the `best` value for K to be large or small? Why?

Solution

Applied Questions

1. $\star$ (ISLR2, Q2.8) This exercise relates to the `College` data set, which can be found in the file `College.csv` on the book website. 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.

b. Look at the data using the `View()` 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 to have these names for later. Try the following commands:

```
rownames(college) <- college[, 1]
View(college)
```

   You should see that there is now a `row.names` column with the name of each university recorded. This means that `R` has given each row a name corresponding to the appropriate university. `R` will not try to perform calculations on the row names. However, we still need to eliminate the first column in the data where the names are stored. Try

```
college <- college[, -1]
View(college)
```

   Now you should see that the first data column is `Private`. Note that another column labeled `row.names` now appears before the `Private` column. However, this is not a data column but rather the name that `R` is giving to each row.

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

   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]`.

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

d. 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`.

e. 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.

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

Solution

2. ⋆ (ISLR2, Q2.9) This exercise involves the `Auto` data set studied in the lab. Make sure that the missing values have been removed from the data.

   a. Which of the predictors are quantitative, and which are qualitative?

   b. What is the range of each quantitative predictor? You can answer this using the `range()` function.

   c. What is the mean and standard deviation of each quantitative predictor?

   d. Now remove the 10th through 85th observations. What is the range, mean, and standard deviation of each predictor in the subset of the data that remains?

   e. Using the full data set, investigate the predictors graphically, using scatterplots or other tools of your choice. Create some plots highlighting the relationships among the predictors. Comment on your findings.

   f. Suppose that we wish to predict gas mileage (`mpg`) on the basis of the other variables. Do your plots suggest that any of the other variables might be useful in predicting `mpg`? Justify your answer.

   Solution

3. (ISLR2, Q2.10) This exercise involves the `Boston` housing data set.

   a. To begin, load in the `Boston` data set. The `Boston` data set is part of the `ISLR2` *library*.

   ```r
   library(ISLR2)
   ```

   Now the data set is contained in the object `Boston`.

   ```r
   Boston
   ```

   Read about the data set:

   ```r
   ?Boston
   ```

   How many rows are in this data set? How many columns? What do the rows and columns represent?

b. Make some pairwise scatterplots of the predictors (columns) in this data set. Describe your findings.

c. Are any of the predictors associated with per capita crime rate? If so, explain the relationship.

d. Do any of the census tracts of Boston appear to have particularly high crime rates? Tax rates? Pupil-teacher ratios? Comment on the range of each predictor.

e. How many of the census tracts in this data set bound the Charles river?

f. What is the median pupil-teacher ratio among the towns in this data set?

g. Which census tract of Boston has lowest median value of owneroccupied homes? What are the values of the other predictors for that census tract, and how do those values compare to the overall ranges for those predictors? Comment on your findings.

h. In this data set, how many of the census tracts average more than seven rooms per dwelling? More than eight rooms per dwelling? Comment on the census tracts that average more than eight rooms per dwelling.

Solution

Solutions

Conceptual Questions

1. a. Better: flexible models are better able to capture all the trends in the large amount of data we have.

   b. Worse: flexible models will tend to overfit the small amount of data we have using the large number of predictors.

   c. Better: inflexible models tend to have a hard time fitting non-linear relationships.

   d. Worse: flexible models will tend to fit the noise, which is not desired.

2. a. Regression: the response (CEO salary) is continuous. Inference: We are interested in the factors influencing CEO salary – we don't want to estimate it using a company's information! $n = 500$ (500 companies in the data set) $p = 3$ (predictors: profit, number of employees, industry; response: CEO salary)

   b. Classification: the response (success or failure) is discrete. Prediction: based on various input factors, we want to estimate how well the product will do $n = 20$ (20 similar products in the data set) $p = 13$ (predictors: marketing budget, price charged, competition price, +10 others; response: whether it was a success or failure)

c. Regression: the response (% change in US dollar) is continuous. Prediction: it's written in the question! We are interested in predicting changes in the US dollar. $n \approx 50$ (number of trading weeks in a year) $p = 3$ (predictors: % change in US market, % change in UK market, % change in DE market; response % change in US dollar)

3. a. See the figure

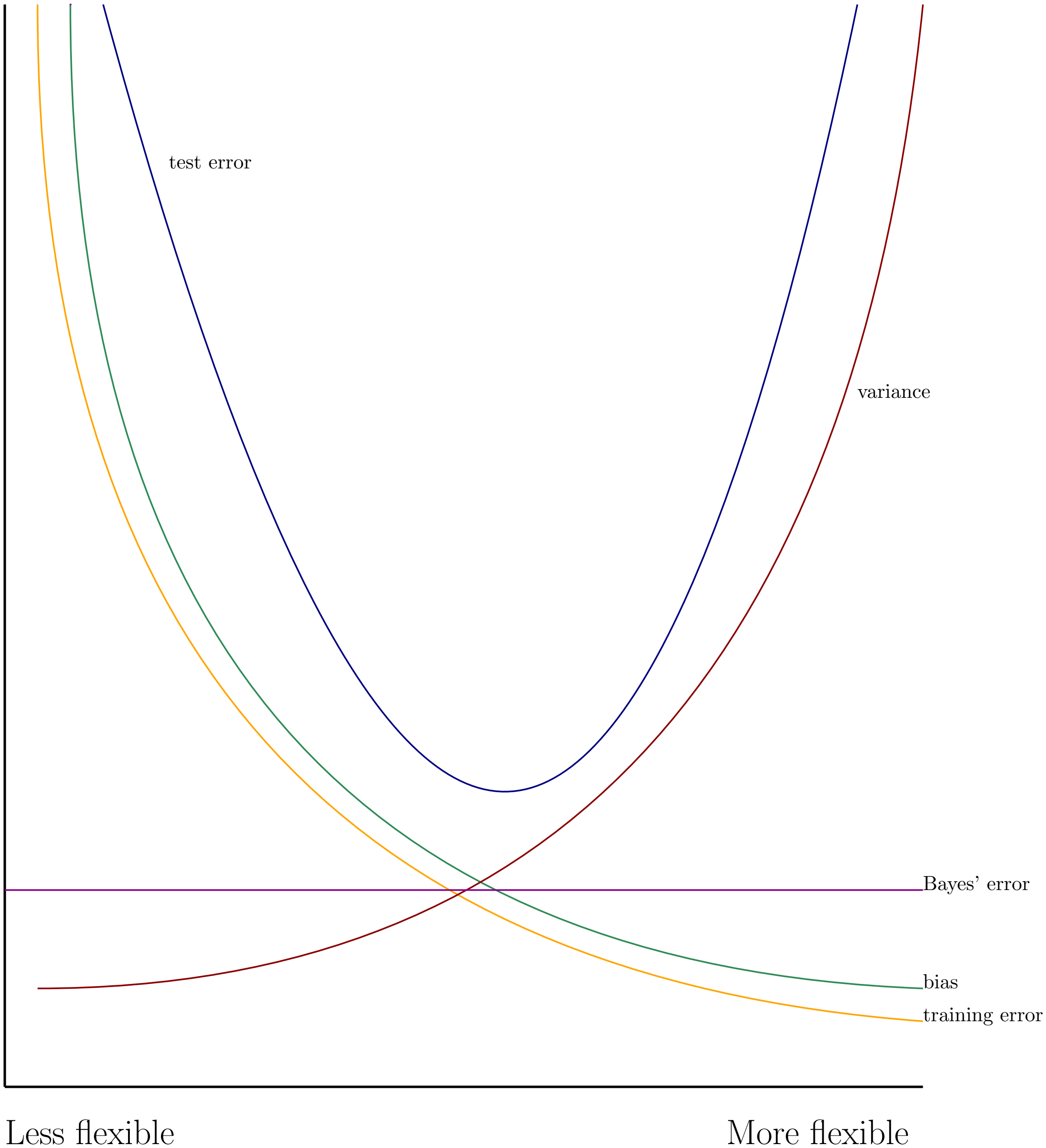


Figure 1: image

b. Bias: increasing flexibility reduces model bias
Training error: increasing flexibility makes the model fit the training data better
Variance: increasing flexibility makes the model incorporate more noise (in other

words, it makes the fit bumpier)
Test error: concave up, since increasing flexibility makes the model fit more of the trend in the data until it starts fitting the noise in the data
Bayes' (irreducible) error: horizontal line, since it's a constant for all models. When the training error dips below the Bayes' error, the model is overfitting, so the test error starts to increase

4. Advantages: can fit a larger variety of (non-linear) models, decreasing bias.
   Disadvantages: can lead to overfitting (hence worse results), requires estimating more parameters, and increasing model variance as it incorporates more noise.
   More flexible models would be preferred if the model is non-linear in nature, or interpretability is not a major issue. Less flexible models are preferred when inference is the goal of the model fitting exercise

5. a. See the table below

Obs.	X_1	X_2	X_3	Distance	Y
1	0	3	0	3	Red
2	2	0	0	2	Red
3	0	1	3	$\sqrt{10} \approx 3.2$	Red
4	0	1	2	$\sqrt{5} \approx 2.2$	Green
5	-1	0	1	$\sqrt{2} \approx 1.4$	Green
6	1	1	1	$\sqrt{3} \approx 1.7$	Red

   b. Green. $K = 1$ so we only take the closest observation (5).

   c. $K = 3$ so we consider the closest 3: 5, 6 and 2. The majority are Red, so this classifies as Red.

   d. A smaller K would lead to a more flexible decision boundary, which would account for the non-linearity better.

Applied Questions

Refer to Section 2.3 of ISLR2 for a primer of applied questions.

1. a. Note that the dataset is available from the course Moodle site

```
college <- read.csv("College.csv")
```

   b. The row names need to be changed to college names as follow

```
rownames(college) <- college[, 1]
college <- college[, -1]
college$Private <- as.factor(college$Private)
```

```r
head(college) # You should instead try `View(college)`
```

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

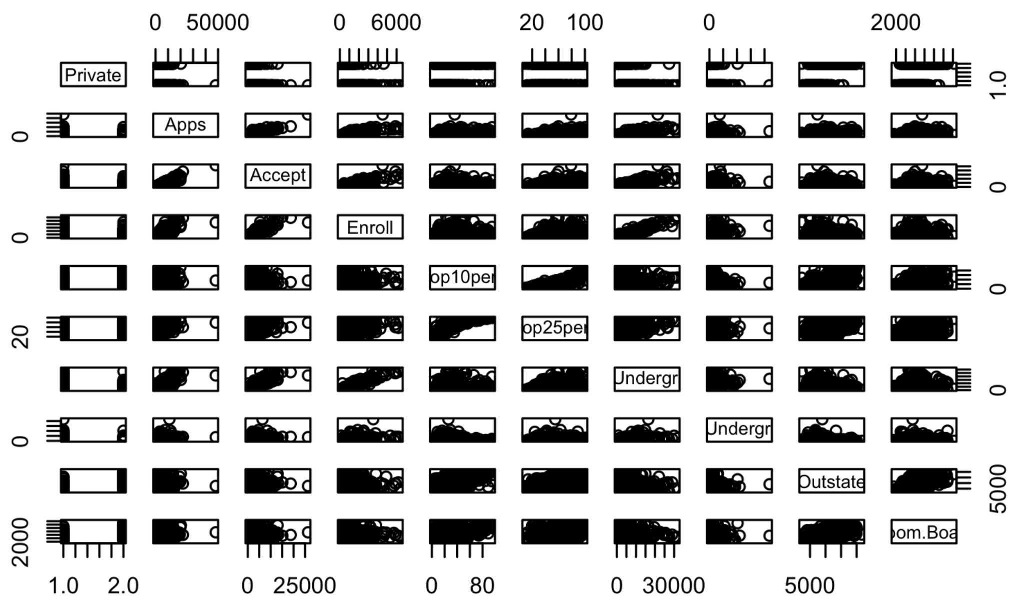
c. i. `summary(college)`

```
 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  
```

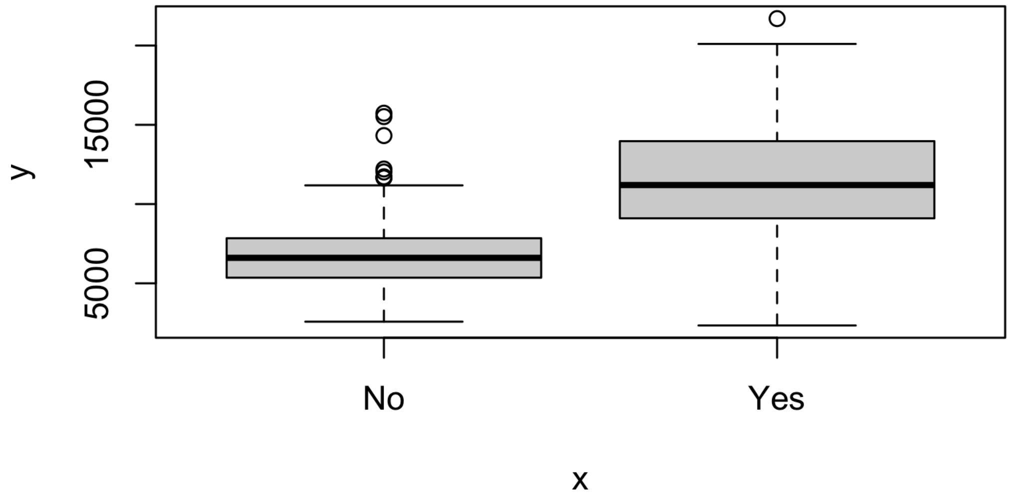
```
 Median :4200   Median : 500.0   Median :1200   Median : 75.00
 Mean   :4358   Mean   : 549.4   Mean   :1341   Mean   : 72.66
 3rd Qu.:5050   3rd Qu.: 600.0   3rd Qu.:1700   3rd Qu.: 85.00
 Max.   :8124   Max.   :2340.0   Max.   :6800   Max.   :103.00
    Terminal       S.F.Ratio      perc.alumni        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. `Private` is not numerical, so cannot be used in pairs so we plot from column 2 onward

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



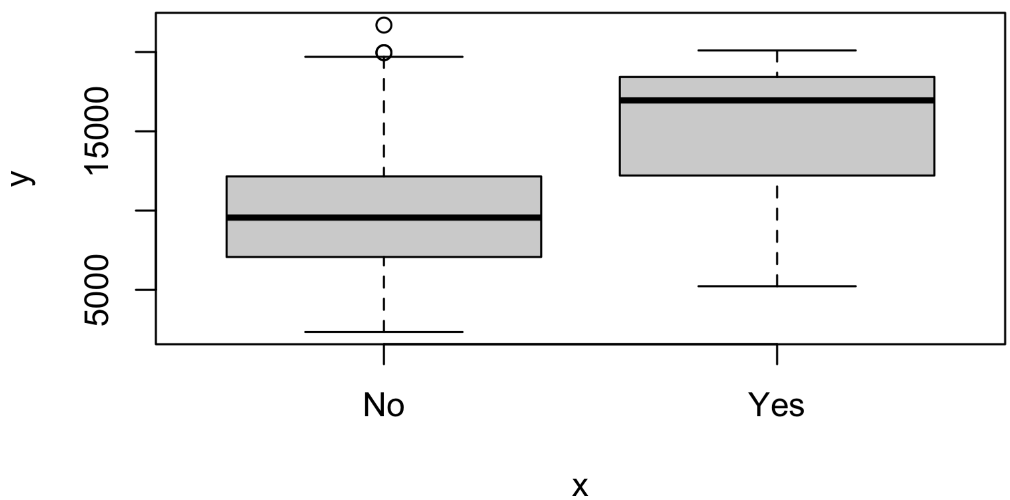
iii. `plot(college$Private, college$Outstate)`



iv.
```r
Elite <- with(college, ifelse(Top10perc > 50, "Yes", "No"))
Elite <- as.factor(Elite)
college <- data.frame(college, Elite)
summary(Elite) # there are 78 elite universities
```

```
 No Yes
699  78
```

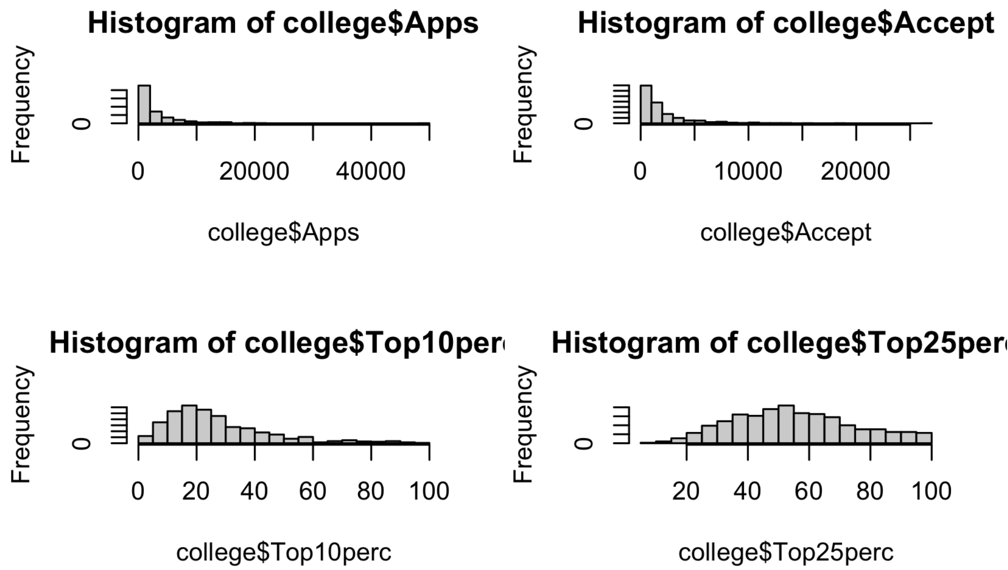
```r
plot(college$Elite, college$Outstate)
```



v. For instance the following code gives histograms for variables `Apps`, `Accept`, `Top10perc` and `Top25perc`.

```r
par(mfrow = c(2, 2))
hist(college$Apps, breaks = 20)
hist(college$Accept, breaks = 20)
```

```r
hist(college$Top10perc, breaks = 20)
hist(college$Top25perc, breaks = 20)
```



2. Note that the dataset is available from the course Moodle site

```r
auto <- read.csv("Auto.csv", na.strings = "?")
auto <- na.omit(auto) # remove missing values
```

   a. Qualitative: `name`, `origin`. Quantitative: `mpg`, `cylinders`, `displacement`, `horsepower`, `weight`, `acceleration`, `year`.

   b. We can use function `apply` combined with function `range`:

```r
quant.var <- c(
  "mpg", "cylinders", "displacement", "horsepower",
  "weight", "acceleration", "year"
)
ranges.df <- apply(auto[, quant.var], 2, range)
rownames(ranges.df) <- c("min", "max")
ranges.df
```

```
     mpg cylinders displacement horsepower weight acceleration year
min  9.0         3           68         46   1613          8.0   70
max 46.6         8          455        230   5140         24.8   82
```

   c. Use a similar application of function `apply`:

```r
means.df <- apply(auto[, quant.var], 2, mean)
std.df <- apply(auto[, quant.var], 2, sd)
distns.df <- rbind(means.df, std.df)
rownames(distns.df) <- c("mean", "sd.")
t(distns.df)
```

```
                    mean        sd.
mpg            23.445918   7.805007
cylinders       5.471939   1.705783
displacement  194.411990 104.644004
horsepower    104.469388  38.491160
weight       2977.584184 849.402560
acceleration   15.541327   2.758864
year           75.979592   3.683737
```

d. Semantic note: the following will remove the 10th to the 85th row, which may not be what we want, since we have already removed some rows to begin with:

```r
subauto <- auto[-(10:85), ]
```

You will find that observation #86 has errantly been removed. That is because the `na.omit` from earlier removed an observation in this range. It is possible to refer to the rows by observation number, which is a character string. In other words, `auto["5",]` will give me observation #5, even if 1-4 are missing. This does complicate the procedure, though.

```r
rid <- rownames(auto)
rid <- rid[as.numeric(rid) < 10 | as.numeric(rid) > 85]
subauto <- auto[rid, ]
```

Use `apply` function:

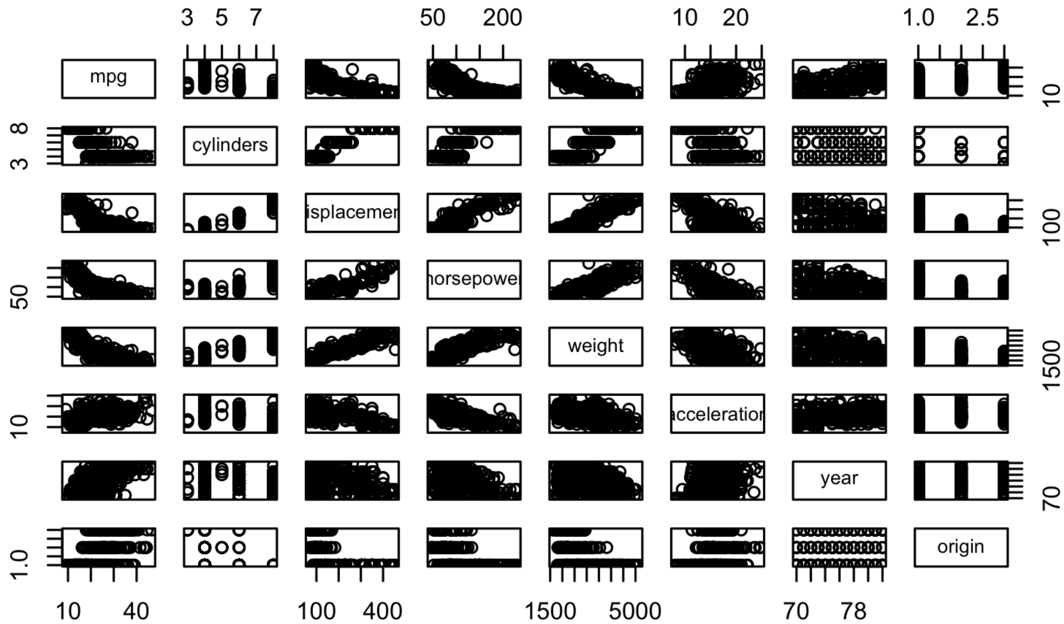
```r
subranges.df <- apply(subauto[, quant.var], 2, range)
submeans.df <- apply(subauto[, quant.var], 2, mean)
substd.df <- apply(subauto[, quant.var], 2, sd)
subdistns.df <- rbind(subranges.df, submeans.df, substd.df)
rownames(subdistns.df) <- c("min", "max", "mean", "sd.")
t(subdistns.df)
```

```
                min    max        mean        sd.
mpg            11.0   46.6   24.368454   7.880898
cylinders       3.0    8.0    5.381703   1.658135
displacement   68.0  455.0  187.753943  99.939488
horsepower     46.0  230.0  100.955836  35.895567
weight       1649.0 4997.0 2939.643533 812.649629
acceleration    8.5   24.8   15.718297   2.693813
```

```
year            70.0    82.0    77.132492    3.110026
```

e. For instance a pairwise plot can be produced using:

```
pairs(auto[, -9])
```



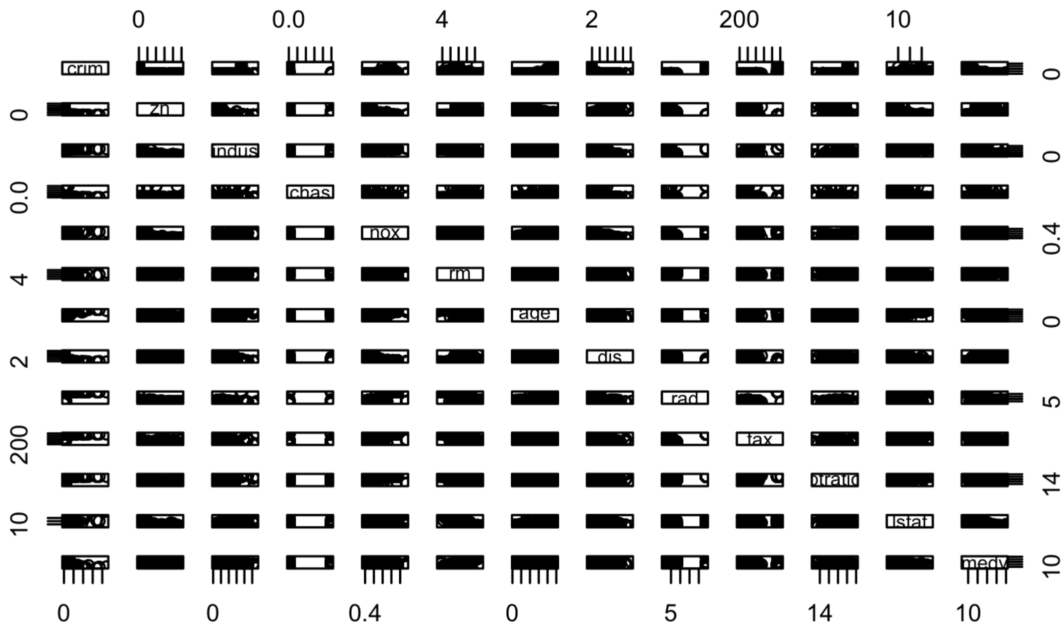
f. Briefly looking at the pairwise plots, the factors `cylinders`, `displacement`, `horsepower`, `weight`, and possibly `year` are worth investigating.

3. a.
```
library(ISLR2)
dim(Boston)
```

```
[1] 506  13
```

506 rows each representing a town, 13 columns each with some data on the towns.

b.
```
pairs(Boston)
```



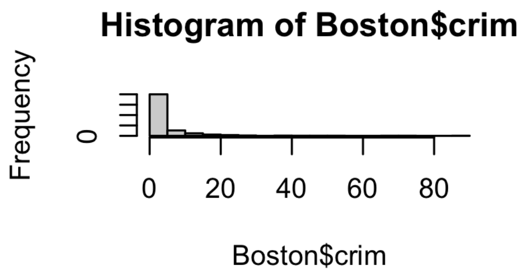
Various answers can exist:

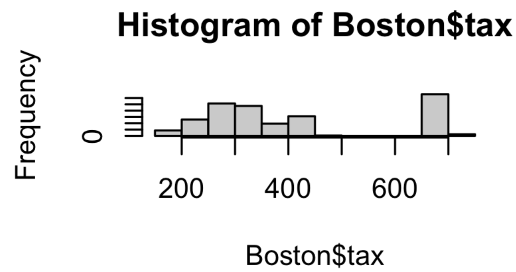
- `crim` relates to `zn`, `indus`, `age`, `dis`, `rad`, `tax`, `ptratio`
- `nox` relates to `age`, `dis`, `rad`
- `age` relates to `lstat`, `medv`
- `lstat` relates to `medv`

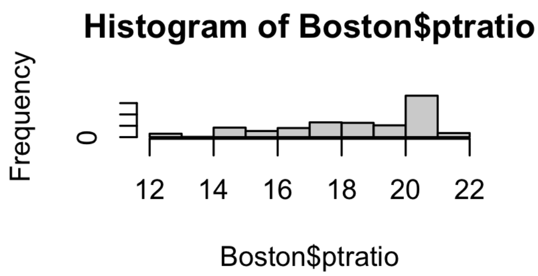
c.
- `zn`: Very low, unless `zn` is very close to 0. Then `crim` can be much higher.
- `indus`: Very low, unless `indus` is close to 18%. Then `crim` can be much higher.
- `age`: `crim` increases as this increases
- `dis`: `crim` decreases as this increases
- `tax`: Very low, unless `tax` is at 666
- `ptratio`: Very low, unless `ptratio` is at 20.2

d.
```r
par(mfrow = c(2, 2))
hist(Boston$crim, breaks = 25)
hist(Boston$tax)
hist(Boston$ptratio)
length(Boston$crim[Boston$crim > 20])
```

```
[1] 18
```







- `crim`: Vast majority of cities have low crime rates, but 18 of them have a crime rate of greater than 20, reaching up to.
- `tax`: Divided into two sections: low < 500, high ≥ 660.
- `ptratio`: Mode at about 20, max at 22, minimum at 12.6.

e. `length(Boston$chas[Boston$chas == 1])`

```
[1] 35
```

f. `median(Boston$ptratio)`

```
[1] 19.05
```

g. `Boston[Boston$medv == min(Boston$medv), ]`

	crim	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	lstat	medv
399	38.3518	0	18.1	0	0.693	5.453	100	1.4896	24	666	20.2	30.59	5
406	67.9208	0	18.1	0	0.693	5.683	100	1.4254	24	666	20.2	22.98	5

Crime rates are quite high, `indus` is on the upper end, all owner-occupied units are built before 1940, both don't bound the Charles river, both are relatively close to employment centres, they're both very close to radial highways, pupil/teacher ratio is at the mode, `lstat` is also on the higher end.

h. `length(Boston$rm[Boston$rm > 7])`

```
[1] 64
```

```r
length(Boston$rm[Boston$rm > 8])
```

```
[1] 13
```

`crim`, `lstat` relatively low.