Assignment\_2 Lab1

Assignment 2 Lab1 by Madhvi Jha # Applied Exercise 2.4.8 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 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

## 2.4.8.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.**

# install.packages("ISLR")  
library(ISLR)  
library(MASS)  
# college = read.csv("file name")  
# Instead of running above command we are getting from package  
data("College")  
college = na.omit(College)

## 2.4.8.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 to have these names for later. Try the following commands:**

**It is not needed since we are not loading from csv** However, we could have run below code if done from csv rownames(college) = college[, 1] fix(college)

head(college)

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

## 2.4.8.c

### Part 2.4.8.c.i

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

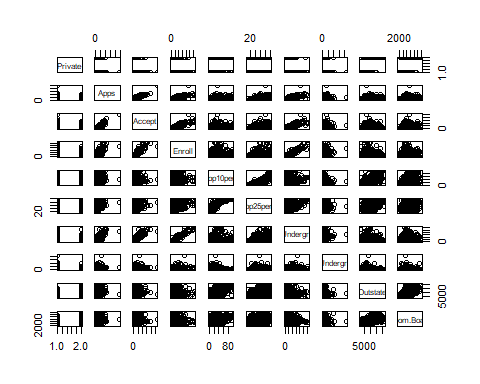
summary(college)

## Private Apps Accept Enroll Top10perc   
## 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

### 2.4.8.c.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].**

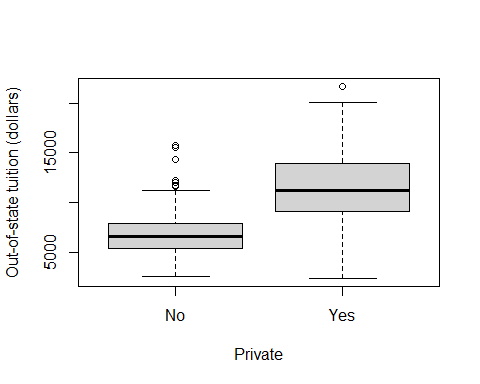
pairs(college[,1:10])



### 2.4.8.c.iii

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

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

 ### 2.4.8.c.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%.**

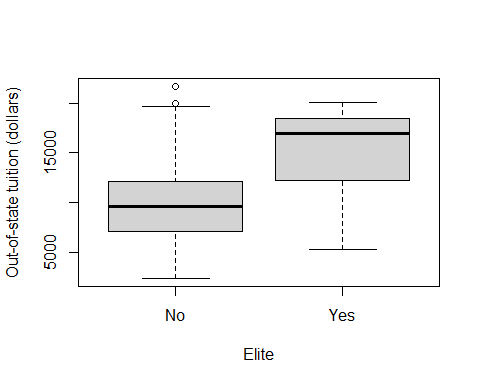
Elite = rep("No", nrow(college))  
Elite[college$Top10per > 50] = "Yes"  
Elite = as.factor(Elite)  
college = data.frame(college, Elite)

**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.**

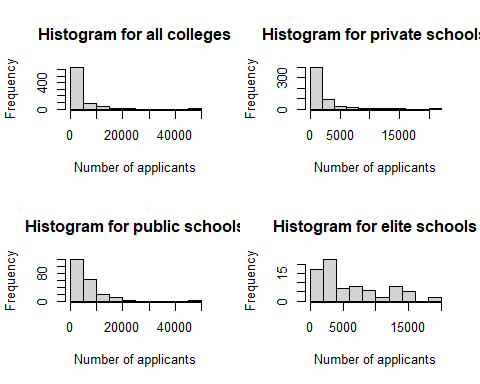
summary(college$Elite)

## No Yes   
## 699 78

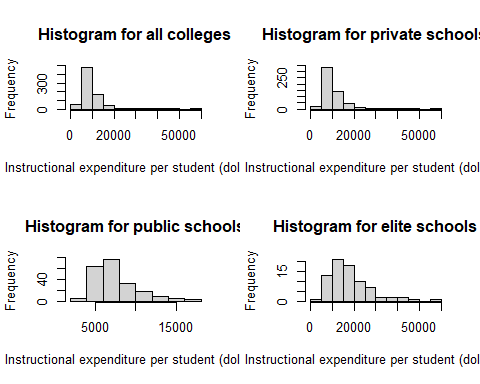
plot(college$Elite, college$Outstate, xlab = "Elite", ylab = "Out-of-state tuition (dollars)")

 ### 2.4.8.c.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.**

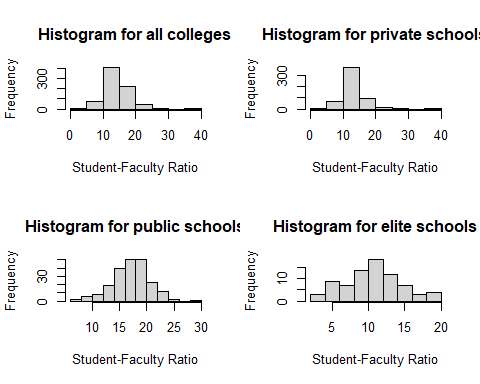
par(mfrow = c(2, 2))  
hist(college$Apps, xlab = "Number of applicants", main = "Histogram for all colleges")  
hist(college$Apps[college$Private == "Yes"], xlab = "Number of applicants", main = "Histogram for private schools")  
hist(college$Apps[college$Private == "No"], xlab = "Number of applicants", main = "Histogram for public schools")  
hist(college$Apps[college$Elite == "Yes"], xlab = "Number of applicants", main = "Histogram for elite schools")



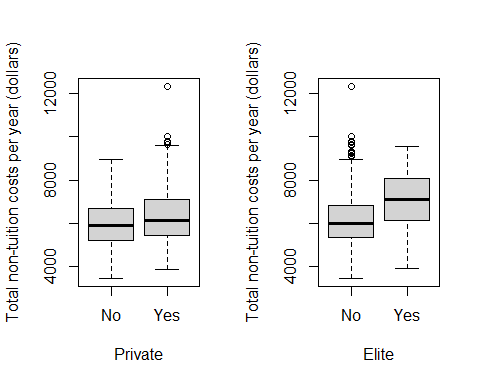
par(mfrow = c(2, 2))  
hist(college$Expend, xlab = "Instructional expenditure per student (dollars)", main = "Histogram for all colleges")  
hist(college$Expend[college$Private == "Yes"], xlab = "Instructional expenditure per student (dollars)", main = "Histogram for private schools")  
hist(college$Expend[college$Private == "No"], xlab = "Instructional expenditure per student (dollars)", main = "Histogram for public schools")  
hist(college$Expend[college$Elite == "Yes"], xlab = "Instructional expenditure per student (dollars)", main = "Histogram for elite schools")



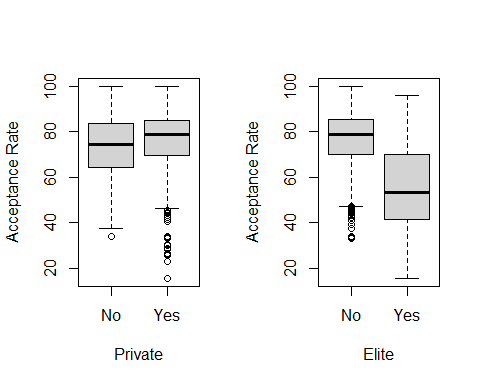
par(mfrow = c(2, 2))  
hist(college$S.F.Ratio, xlab = "Student-Faculty Ratio", main = "Histogram for all colleges")  
hist(college$S.F.Ratio[college$Private == "Yes"], xlab = "Student-Faculty Ratio", main = "Histogram for private schools")  
hist(college$S.F.Ratio[college$Private == "No"], xlab = "Student-Faculty Ratio", main = "Histogram for public schools")  
hist(college$S.F.Ratio[college$Elite == "Yes"], xlab = "Student-Faculty Ratio", main = "Histogram for elite schools")

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

NonTuitionCosts = college$Room.Board + college$Books + college$Personal  
college = data.frame(college, NonTuitionCosts)  
par(mfrow = c(1, 2))  
plot(college$Private, college$NonTuitionCosts, xlab = "Private", ylab = "Total non-tuition costs per year (dollars)")  
plot(college$Elite, college$NonTuitionCosts, xlab = "Elite", ylab = "Total non-tuition costs per year (dollars)")

 Based on the above box plots, it looks like that, aside from some outlier schools with very high costs, there isn’t a wide gap for the median non-tution costs between private schools and public schools. The box plots do show, though, that there is a distinct difference in median non-tuition costs between elite and non-elite schools, with elite schools having higher costs.

AcceptPerc = college$Accept / college$Apps \* 100  
college = data.frame(college, AcceptPerc)  
par(mfrow = c(1, 2))  
plot(college$Private, college$AcceptPerc, xlab = "Private", ylab = "Acceptance Rate")  
plot(college$Elite, college$AcceptPerc, xlab = "Elite", ylab = "Acceptance Rate")



summary(college$AcceptPerc[college$Private == "Yes"])

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 15.45 69.49 78.86 75.46 85.10 100.00

summary(college$AcceptPerc[college$Private == "No"])

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 33.97 64.12 74.43 72.65 83.42 100.00

summary(college$AcceptPerc[college$Elite == "Yes"])

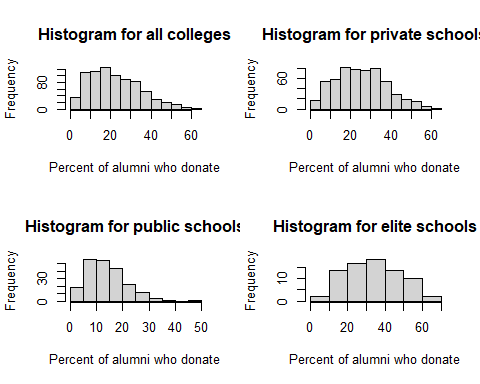
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 15.45 41.53 53.30 54.34 69.59 96.05

summary(college$AcceptPerc[college$Elite == "No"])

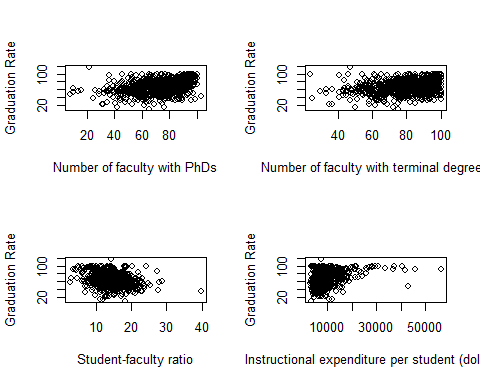
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 32.83 70.13 78.81 76.96 85.48 100.00

The boxplots show that while the median acceptance rates for both private and public schools are pretty close at around 75-80%, private schools have a much wider range of acceptance rates (going down to a minimum of 15.45%). When we distinguish between elite and non-elite schools, elite schools have a much lower median acceptance rate compared to non-elite ones.

par(mfrow = c(2, 2))  
hist(college$perc.alumni, xlab = "Percent of alumni who donate", main = "Histogram for all colleges")  
hist(college$perc.alumni[college$Private == "Yes"], xlab = "Percent of alumni who donate", main = "Histogram for private schools")  
hist(college$perc.alumni[college$Private == "No"], xlab = "Percent of alumni who donate", main = "Histogram for public schools")  
hist(college$perc.alumni[college$Elite == "Yes"], xlab = "Percent of alumni who donate", main = "Histogram for elite schools")

 Based on the above histograms, private schools and elite schools tend to have a higher percent of alumni who donate.

par(mfrow = c(2, 2))  
plot(college$PhD, college$Grad.Rate, xlab = "Number of faculty with PhDs", ylab = "Graduation Rate")  
plot(college$Terminal, college$Grad.Rate, xlab = "Number of faculty with terminal degrees", ylab = "Graduation Rate")  
plot(college$S.F.Ratio, college$Grad.Rate, xlab = "Student-faculty ratio", ylab = "Graduation Rate")  
plot(college$Expend, college$Grad.Rate, xlab = "Instructional expenditure per student (dollars)", ylab = "Graduation Rate")



The above scatterplots explore some of the factors which might be related to student graduation rates. From the upper-left plot, it appears there is a weak positive relationship between the number of faculty with PhDs and graduation rates. The upper-right plot appears to indicate that there isn’t relationship between the number of faculty with terminal degrees and graduation rates. The bottom-left plot indicates that as student-faculty ratios increase, graduation rates generally tend to decrease. Lastly, the bottom-right plot seems to show that there is a definite positive relationship between instructional expenditure per student and graduation rates, with higher expenditures corresponding to higher graduation rates.

# Applied Exercise 2.4.9

**This exercise involves the Auto data set studied in the lab. Make sure that the missing values have been removed from the data.**

Auto = na.omit(Auto)  
dim(Auto)

## [1] 392 9

## 2.4.9.a

**Which of the predictors are quantitative, and which are qualitative?**

head(Auto)

## mpg cylinders displacement horsepower weight acceleration year origin  
## 1 18 8 307 130 3504 12.0 70 1  
## 2 15 8 350 165 3693 11.5 70 1  
## 3 18 8 318 150 3436 11.0 70 1  
## 4 16 8 304 150 3433 12.0 70 1  
## 5 17 8 302 140 3449 10.5 70 1  
## 6 15 8 429 198 4341 10.0 70 1  
## name  
## 1 chevrolet chevelle malibu  
## 2 buick skylark 320  
## 3 plymouth satellite  
## 4 amc rebel sst  
## 5 ford torino  
## 6 ford galaxie 500

The quantitative variables are mpg, displacement, horsepower, weight, and acceleration. Depending on the context, we may want to treat cylinders and year as quantitative predictors or qualitative ones. Lastly, origin and name are qualitative predictors. origin is a quantitative encoding of a car’s country of origin, where 1 being American, 2 being European, and 3 being Japanese.

## 2.4.9.b

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

?range

## starting httpd help server ... done

range(Auto$mpg)

## [1] 9.0 46.6

range(Auto$cylinders)

## [1] 3 8

range(Auto$displacement)

## [1] 68 455

range(Auto$horsepower)

## [1] 46 230

range(Auto$weight)

## [1] 1613 5140

range(Auto$acceleration)

## [1] 8.0 24.8

range(Auto$year)

## [1] 70 82

We have the following ranges for each quantitative predictor:

* mpg = 37.6
* cylinders = 5
* displacement = 387
* horsepower = 184
* weight = 3527
* acceleration = 16.8
* year = 12

## 2.4.9.c

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

colMeans(Auto[, 1:7])

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

apply(Auto[, 1:7], MARGIN = 2, FUN = "sd")

## mpg cylinders displacement horsepower weight acceleration   
## 7.805007 1.705783 104.644004 38.491160 849.402560 2.758864   
## year   
## 3.683737

We have the following mean and standard deviation for each quantitative predictor:

* mpg: mean = 23.45, standard deviation = 7.81
* cylinders: mean = 5.47, standard deviation = 1.71
* displacement: mean = 194.41, standard deviation = 104.64
* horsepower: mean = 104.47, standard deviation = 38.49
* weight: mean = 2977.58, standard deviation = 849.40
* acceleration: mean = 15.54, standard deviation = 2.76
* year: mean = 75.98, standard deviation = 3.68

## 2.4.9.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?**

apply(Auto[-(10:85), 1:7], MARGIN = 2, FUN = "range")

## mpg cylinders displacement horsepower weight acceleration year  
## [1,] 11.0 3 68 46 1649 8.5 70  
## [2,] 46.6 8 455 230 4997 24.8 82

apply(Auto[-(10:85), 1:7], MARGIN = 2, FUN = "mean")

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

apply(Auto[-(10:85), 1:7], MARGIN = 2, FUN = "sd")

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

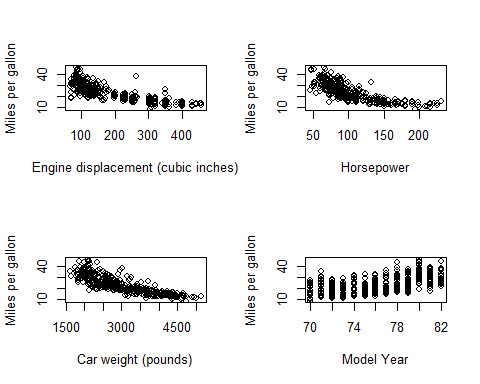
We have the following range, mean,standard deviation for each quantitative predictor after the 10th through 85th rows have been removed:

* mpg: range = 35.6, mean = 24.40, standard deviation = 7.87
* cylinders: range = 5, mean = 5.37, standard deviation = 1.65
* displacement: range = 387, mean = 187.24, standard deviation = 99.68
* horsepower: range = 184, mean = 100.72, standard deviation = 35.71
* weight: range = 3348, mean = 2935.97, standard deviation = 811.30
* acceleration: range = 16.3, mean = 15.73, standard deviation = 2.69
* year: mean = 77.15, standard deviation = 3.11

## 2.4.9.e

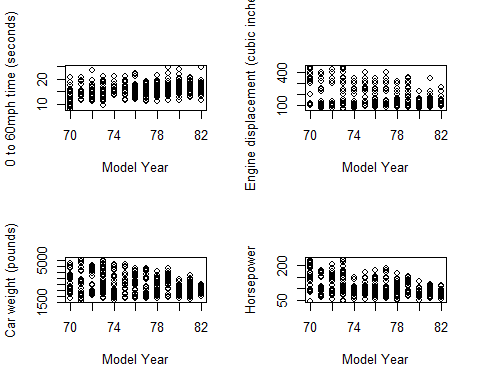
**Using the full data set, investigate the predictors graphically, using scatterplots or other tools of your choice. Create some plots highlighting the relationships among the predictors. Comment on your findings.**

par(mfrow = c(2, 2))  
plot(Auto$displacement, Auto$mpg, xlab = "Engine displacement (cubic inches)", ylab = "Miles per gallon")  
plot(Auto$horsepower, Auto$mpg, xlab = "Horsepower", ylab = "Miles per gallon")  
plot(Auto$weight, Auto$mpg, xlab = "Car weight (pounds)", ylab = "Miles per gallon")  
plot(Auto$year, Auto$mpg, xlab = "Model Year", ylab = "Miles per gallon")



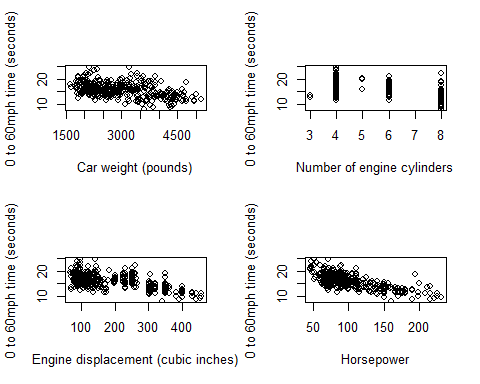
See discussion in Part 6 below.

par(mfrow = c(2, 2))  
plot(Auto$year, Auto$acceleration, xlab = "Model Year", ylab = "0 to 60mph time (seconds)")  
plot(Auto$year, Auto$displacement, xlab = "Model Year", ylab = "Engine displacement (cubic inches)")  
plot(Auto$year, Auto$weight, xlab = "Model Year", ylab = "Car weight (pounds)")  
plot(Auto$year, Auto$horsepower, xlab = "Model Year", ylab = "Horsepower")



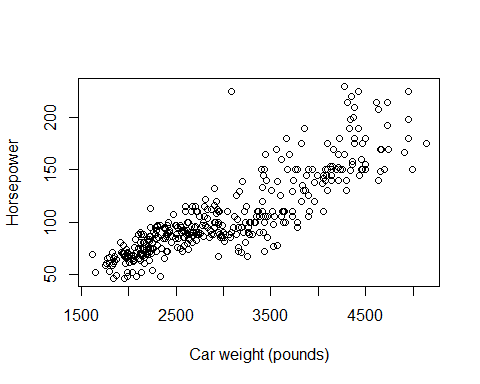
Looking at how various car characteristics change with model year, we see that there aren’t any strong relationships. There are still some weak relationships, such as max engine displacement, car weight, and horsepower generally decreasing from 1970 to 1982. From a historical perspective, these changes could be in response to the 1973 and 1979 oil crises, in which spikes in oil prices pushed auto manufacturers to take measures to improve the efficiency of their cars.

par(mfrow = c(2, 2))  
plot(Auto$weight, Auto$acceleration, xlab = "Car weight (pounds)", ylab = "0 to 60mph time (seconds)")  
plot(Auto$cylinders, Auto$acceleration, xlab = "Number of engine cylinders", ylab = "0 to 60mph time (seconds)")  
plot(Auto$displacement, Auto$acceleration, xlab = "Engine displacement (cubic inches)", ylab = "0 to 60mph time (seconds)")  
plot(Auto$horsepower, Auto$acceleration, xlab = "Horsepower", ylab = "0 to 60mph time (seconds)")

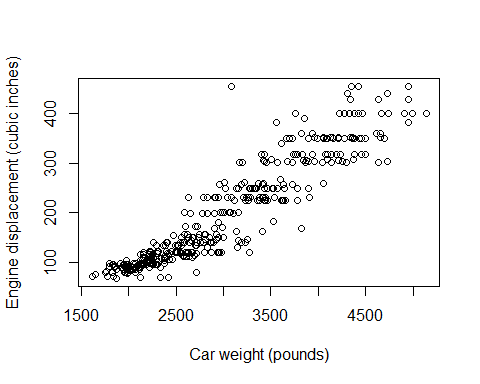


Next, I explored the relationship between the number of seconds it takes a car to accelerate from 0 to 60 miles per hour and a number of different factors. As expected, the 0-to-60 time clearly decreases with increased engine displacement and increased horsepower. There is also a weak relationship that as the number of engine cylinders increases the 0-to-60 time tends to decrease. While it may seem counter-intuitive at first, the 0-to-60 time also tends to decrease with car weight. This makes more sense in the context of the two scatterplots below, which shows that the higher weight is correlated with higher horsepower and higher engine displacement.

par(mfrow = c(1, 1))  
plot(Auto$weight, Auto$horsepower, xlab = "Car weight (pounds)", ylab = "Horsepower")



par(mfrow = c(1,1))  
plot(Auto$weight, Auto$displacement, xlab = "Car weight (pounds)", ylab = "Engine displacement (cubic inches)")



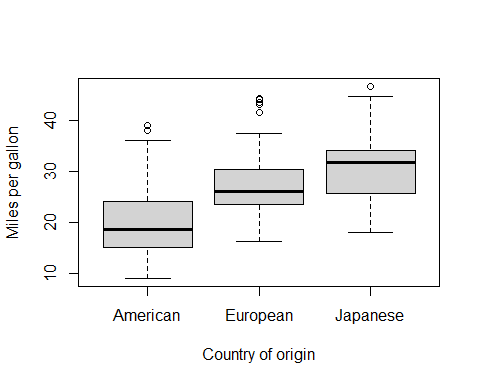
## 2.4.9.f

**Suppose 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.**

Based on the scatter plots I made in part 5 which relate miles per gallon to the predictors engine displacement, horsepower, car weight, and model year, it seems as if the first three factors would be most helpful in predicting mpg, with model year still potentially being helpful but less so. There are clear relationships that increasing engine displacement/horsepower/car weight results in decreased fuel efficiency. There is also a weak relationship that fuel efficiency generally increased going from 1970 to 1982.

Auto$origin[Auto$origin == 1] = "American"  
Auto$origin[Auto$origin == 2] = "European"  
Auto$origin[Auto$origin == 3] = "Japanese"  
Auto$origin = as.factor(Auto$origin)

plot(Auto$origin, Auto$mpg, xlab = "Country of origin", ylab = "Miles per gallon")



Looking at the above box plot, we can also see that there is a relationship between a car’s country of origin and fuel efficiency, where on average Japanese cars are the most efficient, followed by European cars and then by American cars.

# Applied Exercise 2.4.10

**This exercise involves the Boston housing data set.**

## 2.4.10.a

**To begin, load the Boston data set. The Boston data set is part of the MASS *library* in R.**

library(MASS)

library(MASS)

head(Boston)

## crim zn indus chas nox rm age dis rad tax ptratio black lstat  
## 1 0.00632 18 2.31 0 0.538 6.575 65.2 4.0900 1 296 15.3 396.90 4.98  
## 2 0.02731 0 7.07 0 0.469 6.421 78.9 4.9671 2 242 17.8 396.90 9.14  
## 3 0.02729 0 7.07 0 0.469 7.185 61.1 4.9671 2 242 17.8 392.83 4.03  
## 4 0.03237 0 2.18 0 0.458 6.998 45.8 6.0622 3 222 18.7 394.63 2.94  
## 5 0.06905 0 2.18 0 0.458 7.147 54.2 6.0622 3 222 18.7 396.90 5.33  
## 6 0.02985 0 2.18 0 0.458 6.430 58.7 6.0622 3 222 18.7 394.12 5.21  
## medv  
## 1 24.0  
## 2 21.6  
## 3 34.7  
## 4 33.4  
## 5 36.2  
## 6 28.7

**Read about the data set:**

?Boston

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

dim(Boston)

## [1] 506 14

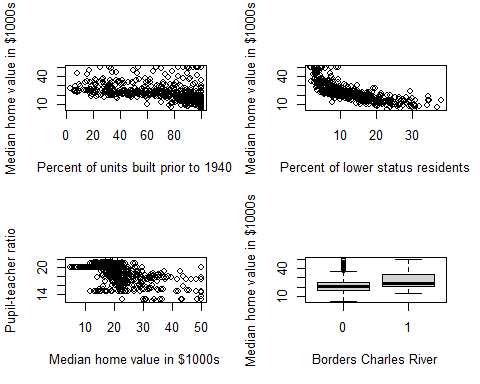
There are 506 observations and 14 features.

* crim - per capita crime rate by town.
* zn - proportion of residential land zoned for lots over 25,000 sq.ft.
* indus - proportion of non-retail business acres per town.
* chas - Charles River dummy variable (= 1 if tract bounds river; 0 otherwise).
* nox - nitrogen oxides concentration (parts per 10 million).
* rm - average number of rooms per dwelling.
* age - proportion of owner-occupied units built prior to 1940.
* dis - weighted mean of distances to five Boston employment centres.
* rad - index of accessibility to radial highways.
* tax - full-value property-tax rate per $10,000.
* ptratio - pupil-teacher ratio by town.
* black - 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town.
* lstat - lower status of the population (percent).
* medv - median value of owner-occupied homes in $1000s.

## 2.4.10.b

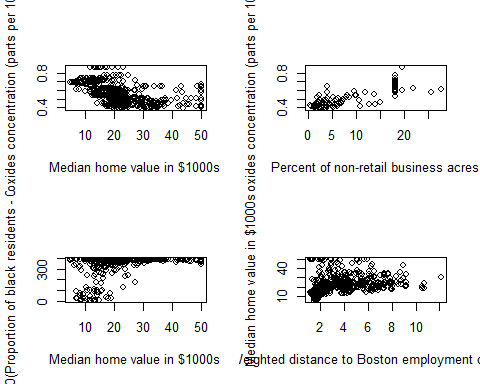
**Make some pairwise scatterplots of the predictors (columns) in the data set. Describe your findings.**

par(mfrow = c(2, 2))  
plot(Boston$age, Boston$medv, xlab = "Percent of units built prior to 1940", ylab = "Median home value in $1000s")  
plot(Boston$lstat, Boston$medv, xlab = "Percent of lower status residents", ylab = "Median home value in $1000s")  
plot(Boston$medv, Boston$ptratio, xlab = "Median home value in $1000s", ylab = "Pupil-teacher ratio")  
plot(as.factor(Boston$chas), Boston$medv, xlab = "Borders Charles River", ylab = "Median home value in $1000s")



First, I generated some plots to explore the relationship between median home value and a number of non-crime factors. There aren’t any especially clear patterns I can discern from thes plots aside from the expected result that as a tracts with higher median home values have a greater proportion of lower-status residence. Also, it appears as if tracts that border the Charles river are a high a slightly higher median home value on average.

par(mfrow = c(2, 2))  
plot(Boston$medv, Boston$nox, xlab = "Median home value in $1000s", ylab = "Nitric oxides concentration (parts per 10 million)")  
plot(Boston$indus, Boston$nox, xlab = "Percent of non-retail business acres", ylab = "Nitric oxides concentration (parts per 10 million)")  
plot(Boston$medv, Boston$black, xlab = "Median home value in $1000s", ylab = "1000(Proportion of black residents - 0.63)^2")  
plot(Boston$dis, Boston$medv, xlab = "Weighted distance to Boston employment centers", ylab = "Median home value in $1000s")



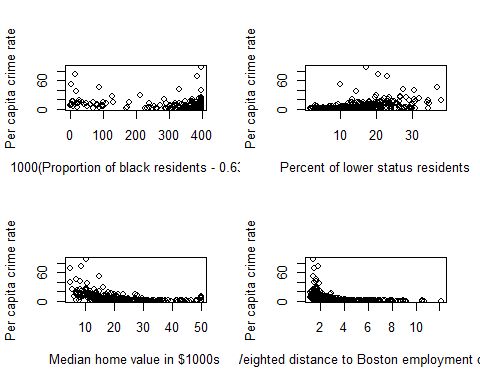
The first two scatter plots in this next group explore factors that might relate to the concentration of nitric oxides. While there isn’t a strong relationship, it appears that tracts with higher median home value also weakly tend to have lower concentrations of nitric oxides. There is a much clearer relationship with the percentage of non-retail business acres – tracts with a higher proportion of non-retail business acres tend to have higher concentrations of nitric oxides. The bottom two plots look at some more factors which might be related to the median home value of a tract.

The bottom-left plot seems to indicate that there is a relationship between the value of black and medv, where black increases as medv increases. If I am interpreting this correctly, this means that tracts with high median home values have a very low (close to 0%) proportion of Black residents, while tracts with low median home values have a much higher proportion (close to 63%). The bottom-right plot appears to indicate that there is also a relationship between proximity to Boston employment centers and median home value, with home values generally increasing as one gets further away from the employment centers.

## 2.4.10.c

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

par(mfrow = c(2, 2))  
plot(Boston$black, Boston$crim, xlab = "1000(Proportion of black residents - 0.63)^2", ylab = "Per capita crime rate")  
plot(Boston$lstat, Boston$crim, xlab = "Percent of lower status residents", ylab = "Per capita crime rate")  
plot(Boston$medv, Boston$crim, xlab = "Median home value in $1000s", ylab = "Per capita crime rate")  
plot(Boston$dis, Boston$crim, xlab = "Weighted distance to Boston employment centers", ylab = "Per capita crime rate")

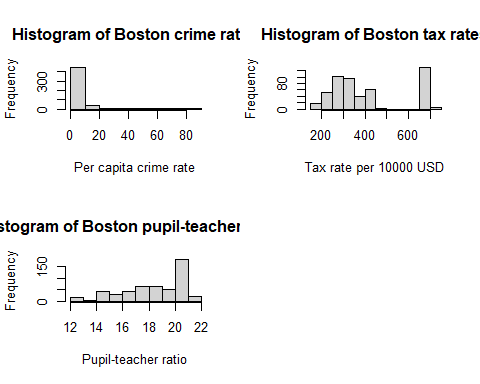


Based on the above four scatter plots, it appears that there are pretty clear relationships between crime rate and median home value, percent of lower status residents, and proximity to Boston employment centers. Tracts with lower home values tend to have higher crime rates, as do tracts which are closer to Boston employment centers. In addiion, tracts with higher proportion of lower status residents tend to have higher crime rates. I was also curious if there would be a relationship between crime rate and black, which serves as some kind of measurement for the proportion of Black residents. Based on the scatter plot between those two variables, there doesn’t appear to be a clear relationship.

## 2.4.10.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**

par(mfrow = c(2, 2))  
hist(Boston$crim, xlab = "Per capita crime rate", main = "Histogram of Boston crime rates")  
hist(Boston$tax, xlab = "Tax rate per 10000 USD", main = "Histogram of Boston tax rates")  
hist(Boston$ptratio, xlab = "Pupil-teacher ratio", main = "Histogram of Boston pupil-teacher ratios")



summary(Boston[, c(1, 10, 11)])

## crim tax ptratio   
## Min. : 0.00632 Min. :187.0 Min. :12.60   
## 1st Qu.: 0.08205 1st Qu.:279.0 1st Qu.:17.40   
## Median : 0.25651 Median :330.0 Median :19.05   
## Mean : 3.61352 Mean :408.2 Mean :18.46   
## 3rd Qu.: 3.67708 3rd Qu.:666.0 3rd Qu.:20.20   
## Max. :88.97620 Max. :711.0 Max. :22.00

Based on the histograms and the numerical summary, there do appear to be tracts within Boston which have particularly high crime rates, tax rates, or pupil-teacher ratios. The minimum crime rate is 0.00632, while the maximum is 88.97620, with a median of 0.25651. The minimum tax rate is \$187 per \$10000, while the maximum is \$711, with a median of \$330. The minimum pupil-teacher ratio is 12.60 pupils per teacher, while the maximum is 22, with a median of 19.05. Given the median value, the maximum pupil-teacher ratio in the data set isn’t outrageously high, since about half of the tracts have a ratio of 19 or more.

## 2.4.10.e

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

sum(Boston$chas)

## [1] 35

In this data set, 35 tracts neighbor the Charles river.

## 2.4.10.f

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

summary(Boston$ptratio)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 12.60 17.40 19.05 18.46 20.20 22.00

The median pupil-teacher ratio among towns in this data set is 19.05 pupils per teacher.

## 2.4.10.g

**Which census tract of Boston has lowest median value of owner-occupied 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.**

t(subset(Boston, medv == min(Boston$medv)))

## 399 406  
## crim 38.3518 67.9208  
## zn 0.0000 0.0000  
## indus 18.1000 18.1000  
## chas 0.0000 0.0000  
## nox 0.6930 0.6930  
## rm 5.4530 5.6830  
## age 100.0000 100.0000  
## dis 1.4896 1.4254  
## rad 24.0000 24.0000  
## tax 666.0000 666.0000  
## ptratio 20.2000 20.2000  
## black 396.9000 384.9700  
## lstat 30.5900 22.9800  
## medv 5.0000 5.0000

Two of the tracts of South Boston have the lowest median value of owner-occupied homes, at $5000. Both of these tracts have very high crime rates compared to the overall range for that variable, with values 38.3518 and 67.9208 putting them far into the upper quartile and into the range of being outliers. These tracts have no land zoned for residential lots of 25000 sq. ft., though this is in line with at least half of the tracts in the overall set given the median for ZN is 0. The two tracts do have a relatively high proportion of non-retail business acres, with values of 18.1 being right at the third quartile. Similarly, the tracts also have concentrations of nitric oxides in the upper quartile of the overall set with a value of 0.693 parts per ten million. The average number of rooms per dwelling for these two tracts is at the low end, with values of 5.453 and 5.683 putting them at the bottom quartile. Next, these two tracts are among those with the highest proportion of owner-occupied homes built prior to 1940, with a value of 100. The tracts are also quite close Boston employment centers with DIS values of 1.4896 and 1.4254 putting them at the bottom quartile. The tracts also are very close to radial highways with the maximum value of RAD at 24. Next, the tracts have above average property tax rates, with a value of \$666 per \$10000, putting them at the third quartile. The pupil-teacher ratio of 20.2 also puts these tracts at the third quartile. The tracts have relatively high values for B, though one tract has a maximum value while the other, with a value of 384.97, is in between the first and second quartiles. Lastly, the tracts have a high proportion of lower status residents (values of 30.59 and 22.98), putting them in the top quartile of the data.

In summary, these two tracts with the lowest median value of owner-occupied homes have predictors generally at the extreme ends of their respective ranges.

## 2.4.10.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 suburbs that average more than eight rooms per dwelling.**

dim(subset(Boston, rm > 7))

## [1] 64 14

dim(subset(Boston, rm > 8))

## [1] 13 14

summary(subset(Boston, rm > 8))

## crim zn indus chas   
## Min. :0.02009 Min. : 0.00 Min. : 2.680 Min. :0.0000   
## 1st Qu.:0.33147 1st Qu.: 0.00 1st Qu.: 3.970 1st Qu.:0.0000   
## Median :0.52014 Median : 0.00 Median : 6.200 Median :0.0000   
## Mean :0.71879 Mean :13.62 Mean : 7.078 Mean :0.1538   
## 3rd Qu.:0.57834 3rd Qu.:20.00 3rd Qu.: 6.200 3rd Qu.:0.0000   
## Max. :3.47428 Max. :95.00 Max. :19.580 Max. :1.0000   
## nox rm age dis   
## Min. :0.4161 Min. :8.034 Min. : 8.40 Min. :1.801   
## 1st Qu.:0.5040 1st Qu.:8.247 1st Qu.:70.40 1st Qu.:2.288   
## Median :0.5070 Median :8.297 Median :78.30 Median :2.894   
## Mean :0.5392 Mean :8.349 Mean :71.54 Mean :3.430   
## 3rd Qu.:0.6050 3rd Qu.:8.398 3rd Qu.:86.50 3rd Qu.:3.652   
## Max. :0.7180 Max. :8.780 Max. :93.90 Max. :8.907   
## rad tax ptratio black   
## Min. : 2.000 Min. :224.0 Min. :13.00 Min. :354.6   
## 1st Qu.: 5.000 1st Qu.:264.0 1st Qu.:14.70 1st Qu.:384.5   
## Median : 7.000 Median :307.0 Median :17.40 Median :386.9   
## Mean : 7.462 Mean :325.1 Mean :16.36 Mean :385.2   
## 3rd Qu.: 8.000 3rd Qu.:307.0 3rd Qu.:17.40 3rd Qu.:389.7   
## Max. :24.000 Max. :666.0 Max. :20.20 Max. :396.9   
## lstat medv   
## Min. :2.47 Min. :21.9   
## 1st Qu.:3.32 1st Qu.:41.7   
## Median :4.14 Median :48.3   
## Mean :4.31 Mean :44.2   
## 3rd Qu.:5.12 3rd Qu.:50.0   
## Max. :7.44 Max. :50.0

From the numerical summary, one thing that stands out is that the tracts which average at least eight rooms per dwelling have low crime rates, low concentrations of nitric oxides, low proportions of Black residents (high values of black), and low proportions of lower status residents compared to the overall data set.

# Applied Exercise 3.7.8

**This question involves the use of simple linear regression on the Auto data set.**

library(ISLR)  
library(MASS)  
Auto = na.omit(Auto)

## 3.7.8.a

**Use the lm() function to perform a simple linear regression with mpg as the response and horsepower as the predictor. Use the summary() function to print the results. Comment on the output. For example:**

### 3.7.8.a.i

**Is there a relationship between the predictor and the response.**

### 3.7.8.a.ii

**How strong is the relationship between the predictor and the response?** ### 3.7.8.a.iii **Is the relationship between the predictor and the response positive or negative.** ### 3.7.8.a.iV **What is the predicted mpg associated with a horsepower of 98? What are the associated 95% confidence and prediction intervals?**

auto.lin.fit = lm(mpg ~ horsepower, data = Auto)  
summary(auto.lin.fit)

##   
## Call:  
## lm(formula = mpg ~ horsepower, data = Auto)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -13.5710 -3.2592 -0.3435 2.7630 16.9240   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 39.935861 0.717499 55.66 <2e-16 \*\*\*  
## horsepower -0.157845 0.006446 -24.49 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.906 on 390 degrees of freedom  
## Multiple R-squared: 0.6059, Adjusted R-squared: 0.6049   
## F-statistic: 599.7 on 1 and 390 DF, p-value: < 2.2e-16

Simple linear regression gives a model between the predictor horsepower and the response mpg. A p-value of essentially zero for gives very strong evidence that there is a relationship between mpg and horsepowerSince , approximately 60.6% of the variability in mpg is explained by a linear regression onto horsepower. This is a modest relationship between the predictor and the response, since as discussed in the chapter we can improve our value to 0.688 by including a quadratic term. The value of itself indicates that in the model each increase of 1 horsepower results on average in a decrease of 0.157845 miles per gallon. In other words, in this model there is a negative relationship between the predictor and the response.

predict(auto.lin.fit, data.frame(horsepower = 98), interval = "confidence")

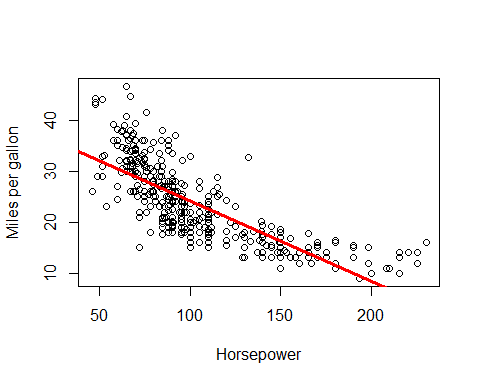
## fit lwr upr  
## 1 24.46708 23.97308 24.96108

predict(auto.lin.fit, data.frame(horsepower = 98), interval = "prediction")

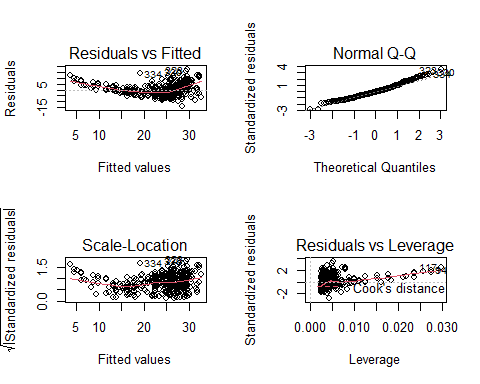
## fit lwr upr  
## 1 24.46708 14.8094 34.12476

Plugging in a horsepower value of 98 gives a predicted mpg of 24.46708. The 95% confidence interval for this prediction is (23.97308, 24.96108) and the 95% prediction interval is (14.8094, 34.12467) ## 3.7.8.b **Plot the response and the predictor. Use the abline() function to display the least squares regression line.**

plot(Auto$horsepower, Auto$mpg, xlab = "Horsepower", ylab = "Miles per gallon")  
abline(auto.lin.fit, lwd = 3, col = "red")

 ## 3.7.8.c **Use the plot() function to produce diagnostic plots of the least squares regression fit. Comment on any problems you see with the fit.**

par(mfrow = c(2, 2))  
plot(auto.lin.fit)



Looking at the Residuals vs. Fitted plot, there is a clear U-shape to the residuals, which is a strong indicator of non-linearity in the data. This, when combined with an inspection of the plot in Part 2, tells us that the simple linear regression model is not a good fit. In addition, when looking at the Residuals vs. Leverage plot, there are some high leverage points (remember that after dropping the rows with null values, there are 392 observations in the data set, giving an average leverage value of ) which also have high standardized residual values (greater than 2), which is also of concern for the simple linear regression model. There are also a number of observations with a standardized residual value of 3 or more, which is evidence to suggest that they would be possibile outliers if we didn’t already have the suspicion that the data is non-linear.

# Applied Exercise 3.7.9

**This question involves the use of multiple linear regression on the Auto data set.**

Auto = na.omit(Auto)  
head(Auto)

## mpg cylinders displacement horsepower weight acceleration year origin  
## 1 18 8 307 130 3504 12.0 70 American  
## 2 15 8 350 165 3693 11.5 70 American  
## 3 18 8 318 150 3436 11.0 70 American  
## 4 16 8 304 150 3433 12.0 70 American  
## 5 17 8 302 140 3449 10.5 70 American  
## 6 15 8 429 198 4341 10.0 70 American  
## name  
## 1 chevrolet chevelle malibu  
## 2 buick skylark 320  
## 3 plymouth satellite  
## 4 amc rebel sst  
## 5 ford torino  
## 6 ford galaxie 500

Note that the origin column actually contains categorical data, even though it is coded using integers. In order to make my life a little easier for performing regression, I’m going replace the values in that column with their meanings and convert it to a factor column. There are also [other options for coding categorical variables](https://stats.idre.ucla.edu/r/modules/coding-for-categorical-variables-in-regression-models/), such as using the factor() function directly within lm(), or using the C() function to have more control over the contrast coding.

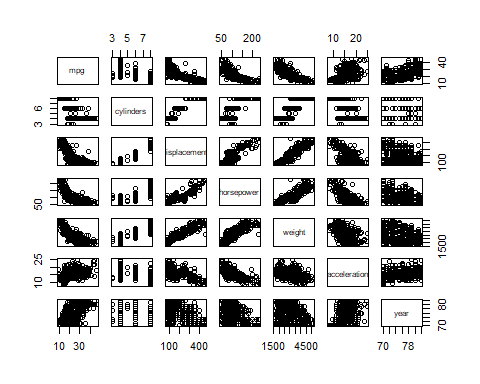
Auto$origin[Auto$origin == 1] = "American"  
Auto$origin[Auto$origin == 2] = "European"  
Auto$origin[Auto$origin == 3] = "Japanese"  
Auto$origin = as.factor(Auto$origin)  
head(Auto)

## mpg cylinders displacement horsepower weight acceleration year origin  
## 1 18 8 307 130 3504 12.0 70 American  
## 2 15 8 350 165 3693 11.5 70 American  
## 3 18 8 318 150 3436 11.0 70 American  
## 4 16 8 304 150 3433 12.0 70 American  
## 5 17 8 302 140 3449 10.5 70 American  
## 6 15 8 429 198 4341 10.0 70 American  
## name  
## 1 chevrolet chevelle malibu  
## 2 buick skylark 320  
## 3 plymouth satellite  
## 4 amc rebel sst  
## 5 ford torino  
## 6 ford galaxie 500

## 3.7.9.a

**Produce a scatterplot matrix which includes all of the variables in the data set.**

pairs(~mpg + cylinders + displacement + horsepower + weight + acceleration + year, Auto)



Since origin and name are categorical columns, I’m excluding them from the scatterplot matrix.

## 3.7.9.b

**Compute the matrix of correlations between the variables using the function cor(). You will need to exclude the name variable, which is qualitative.**

cor(Auto[,-c(8, 9)])

## mpg cylinders displacement horsepower weight  
## mpg 1.0000000 -0.7776175 -0.8051269 -0.7784268 -0.8322442  
## cylinders -0.7776175 1.0000000 0.9508233 0.8429834 0.8975273  
## displacement -0.8051269 0.9508233 1.0000000 0.8972570 0.9329944  
## horsepower -0.7784268 0.8429834 0.8972570 1.0000000 0.8645377  
## weight -0.8322442 0.8975273 0.9329944 0.8645377 1.0000000  
## acceleration 0.4233285 -0.5046834 -0.5438005 -0.6891955 -0.4168392  
## year 0.5805410 -0.3456474 -0.3698552 -0.4163615 -0.3091199  
## acceleration year  
## mpg 0.4233285 0.5805410  
## cylinders -0.5046834 -0.3456474  
## displacement -0.5438005 -0.3698552  
## horsepower -0.6891955 -0.4163615  
## weight -0.4168392 -0.3091199  
## acceleration 1.0000000 0.2903161  
## year 0.2903161 1.0000000

Since the origin column is also qualitative, I excluded it along with the name column when computing the matrix of correlations.

## 3.7.9.c

**Use the lm() function to perform a multiple linear regression with mpg as the response and all other variables except name as the predictors. Use the summary() function to print the results. Comment on the output. For instance:** ### 3.7.9.c.i **Is there a relationship between the predictors and the response?** ### 3.7.9.c.ii **Which predictors appear to have a statistically significant relationship to the response?** ### 3.7.9.c.iii **What does the coefficient for the year variable suggest?**

mpg.fit = lm(mpg ~ . - name, data = Auto)  
summary(mpg.fit)

##   
## Call:  
## lm(formula = mpg ~ . - name, data = Auto)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -9.0095 -2.0785 -0.0982 1.9856 13.3608   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.795e+01 4.677e+00 -3.839 0.000145 \*\*\*  
## cylinders -4.897e-01 3.212e-01 -1.524 0.128215   
## displacement 2.398e-02 7.653e-03 3.133 0.001863 \*\*   
## horsepower -1.818e-02 1.371e-02 -1.326 0.185488   
## weight -6.710e-03 6.551e-04 -10.243 < 2e-16 \*\*\*  
## acceleration 7.910e-02 9.822e-02 0.805 0.421101   
## year 7.770e-01 5.178e-02 15.005 < 2e-16 \*\*\*  
## originEuropean 2.630e+00 5.664e-01 4.643 4.72e-06 \*\*\*  
## originJapanese 2.853e+00 5.527e-01 5.162 3.93e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.307 on 383 degrees of freedom  
## Multiple R-squared: 0.8242, Adjusted R-squared: 0.8205   
## F-statistic: 224.5 on 8 and 383 DF, p-value: < 2.2e-16

contrasts(Auto$origin)

## European Japanese  
## American 0 0  
## European 1 0  
## Japanese 0 1

Since the F-statistic is 224.5, giving a p-value of essentially zero for the null hypothesis , there is strong evidence to believe that there is a relationship between the predictors and the response. The predictors that appear to have a statistically significant relationship to the response mpg are displacement with a p-value of 0.001863, and weight, year, originEuropean, and originJapanese with p-values of essentially zero. The coefficients for cylinders, horsepower, and acceleration have p-values which are not small enough to provide evidence of a statistically significant relationship to the response mpg. The coefficient of 0.777 for the year variable suggests that when we fix the number of engine cylinders, engine displacement, horsepower, weight, acceleration, and country of origin, fuel efficiency increases on average by about 0.777 miles per gallon each year. In other words, the model suggests that we would expect cars from 1971 to be more fuel efficient by 0.777 miles per gallon on average compared to equivalent cars from 1970. Also of interest are the coefficients for originEuropean and originJapanese, which suggest that compared to equivalent cars from the United States, we would expect European cars to be more fuel efficient by 2.630 miles per gallon on average, and Japanese cars to be more fuel efficient by 2.853 miles per gallon on average. Lastly, the value of 0.8242 indicates that about 82% of the variation in mpg is explained by this least squares regression model.

#### Answer 3.7.9.c.i

Yes, there is a relationship between the predictors and the response by testing the null hypothesis of whether all the regression coefficients are zero. The F-statistic is far from 1 (with a small p-value), indicating evidence against the null hypothesis

#### Answer 3.7.9.c.ii

Looking at the p-values associated with each preditor’s t-statistic, we see that displacement, weight, year, and origin have a statistically significant relationship, while cylinders, horsepower, and acceleration do not.

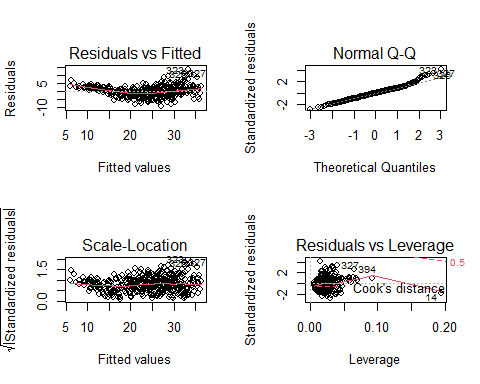
#### Answer 3.7.9.c.iii

The regression coefficient for year, 0.7508 suggests that for every one year, mpg increases by the coefficient. In other words, cars become more fuel efficient every year by almost 1 mpg/year.

## 3.7.9.d

**Use the plot() function to produce diagnostic plots of the linear regression fit. Comment on any problems you see with the fit. Do the residual plots suggest any unusually large outliers? Does the leverage plot identify any observations with unusually high leverage?**

par(mfrow = c(2, 2))  
plot(mpg.fit)



The fit does not appear to be accurate because there is a discernible curve pattern to the residual plots. Form the leverage plot, point 14 appears to have high leverage, although not a high magnitude residual.

There are possible outliers as seen in the plot of studentized residuals becuase there are data with a value greater than 3.

## 3.7.9.e

**Use the \* and : symbols to fit linear regression models with interaction effects. Do any interactions appear to be statistically significant?**

lm.fit2 = lm(Auto$mpg~Auto$cylinders\*Auto$displacement+Auto$displacement\*Auto$weight)  
summary(lm.fit2)

##   
## Call:  
## lm(formula = Auto$mpg ~ Auto$cylinders \* Auto$displacement +   
## Auto$displacement \* Auto$weight)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -13.2934 -2.5184 -0.3476 1.8399 17.7723   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5.262e+01 2.237e+00 23.519 < 2e-16 \*\*\*  
## Auto$cylinders 7.606e-01 7.669e-01 0.992 0.322   
## Auto$displacement -7.351e-02 1.669e-02 -4.403 1.38e-05 \*\*\*  
## Auto$weight -9.888e-03 1.329e-03 -7.438 6.69e-13 \*\*\*  
## Auto$cylinders:Auto$displacement -2.986e-03 3.426e-03 -0.872 0.384   
## Auto$displacement:Auto$weight 2.128e-05 5.002e-06 4.254 2.64e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.103 on 386 degrees of freedom  
## Multiple R-squared: 0.7272, Adjusted R-squared: 0.7237   
## F-statistic: 205.8 on 5 and 386 DF, p-value: < 2.2e-16

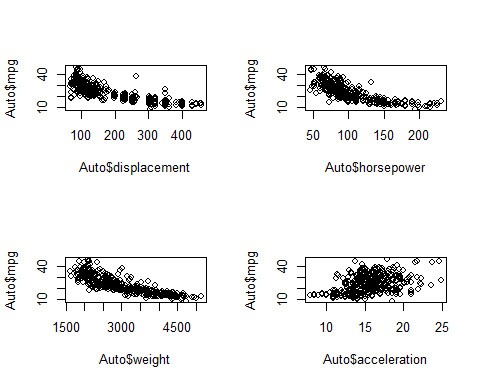
From the correlation matrix, I obtained the two highest correlated pairs and used them in picking my interaction effects. From the p-values, we can see that the interaction between displacement and weight is statistically signifcant, while the interactiion between cylinders and displacement is not.

## 3.7.9.f

**Try a few different transformations of the variables, such as . Comment on your findings.**

To get a sense of which transformations I want to try out for each quantitative variable, focusing on displacement, horsepower and weight, I’ll look at the scatterplots of each one versus mpg.

par(mfrow = c(2, 2))  
plot(Auto$displacement, Auto$mpg)  
plot(Auto$horsepower, Auto$mpg)  
plot(Auto$weight, Auto$mpg)  
plot(Auto$acceleration, Auto$mpg)

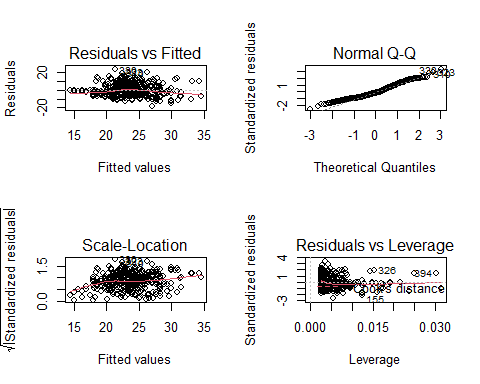


The book already explored nonlinear transformations of horsepower to predict mpg, so I will first look at transforms of acceleration.

summary(lm(mpg ~ acceleration, data = Auto))

##   
## Call:  
## lm(formula = mpg ~ acceleration, data = Auto)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -17.989 -5.616 -1.199 4.801 23.239   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4.8332 2.0485 2.359 0.0188 \*   
## acceleration 1.1976 0.1298 9.228 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 7.08 on 390 degrees of freedom  
## Multiple R-squared: 0.1792, Adjusted R-squared: 0.1771   
## F-statistic: 85.15 on 1 and 390 DF, p-value: < 2.2e-16

par(mfrow = c(2, 2))  
plot(lm(mpg ~ acceleration, data = Auto))

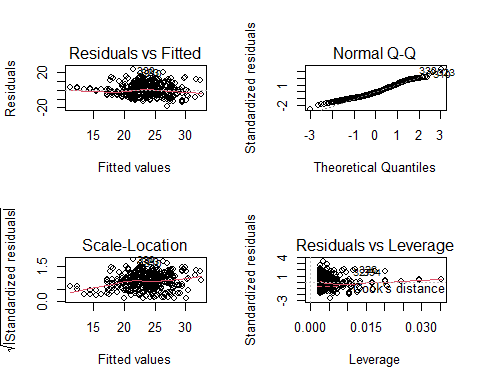


It appears that there might be heteroscedasticity, or non-constant variances in the error terms, so let’s first see how applying a logarithmic transportation affects the model.

summary(lm(mpg ~ log(acceleration), data = Auto))

##   
## Call:  
## lm(formula = mpg ~ log(acceleration), data = Auto)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -18.0234 -5.6231 -0.9787 4.5943 23.0872   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -27.834 5.373 -5.180 3.56e-07 \*\*\*  
## log(acceleration) 18.801 1.966 9.565 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 7.033 on 390 degrees of freedom  
## Multiple R-squared: 0.19, Adjusted R-squared: 0.1879   
## F-statistic: 91.49 on 1 and 390 DF, p-value: < 2.2e-16

par(mfrow = c(2, 2))  
plot(lm(mpg ~ log(acceleration), data = Auto))



While the transformation did bump up the value very slightly, it didn’t really do anything to help with the residuals. This is probably due to the fact that two cars with the same 0 to 60 mile per hour time could be quite different in other ways that would affect fuel economy, such has differences in engine efficiency. For the remainder of the problem, let’s turn our attention the the relationship between engine displacement and fuel efficiency. From the scatterplot, it is pretty clear that there is a nonlinear relationship between the two quantities. Let’s start off by comparing a linear model to one that also includes the quadratic term.

displacement.linear = lm(mpg ~ displacement, data = Auto)  
summary(displacement.linear)

##   
## Call:  
## lm(formula = mpg ~ displacement, data = Auto)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -12.9170 -3.0243 -0.5021 2.3512 18.6128   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 35.12064 0.49443 71.03 <2e-16 \*\*\*  
## displacement -0.06005 0.00224 -26.81 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.635 on 390 degrees of freedom  
## Multiple R-squared: 0.6482, Adjusted R-squared: 0.6473   
## F-statistic: 718.7 on 1 and 390 DF, p-value: < 2.2e-16

displacement.quadratic = lm(mpg ~ poly(displacement, 2), data = Auto)  
summary(displacement.quadratic)

##   
## Call:  
## lm(formula = mpg ~ poly(displacement, 2), data = Auto)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -15.2165 -2.2404 -0.2508 2.1094 20.5158   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 23.4459 0.2205 106.343 < 2e-16 \*\*\*  
## poly(displacement, 2)1 -124.2585 4.3652 -28.466 < 2e-16 \*\*\*  
## poly(displacement, 2)2 31.0895 4.3652 7.122 5.17e-12 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.365 on 389 degrees of freedom  
## Multiple R-squared: 0.6888, Adjusted R-squared: 0.6872   
## F-statistic: 430.5 on 2 and 389 DF, p-value: < 2.2e-16

anova(displacement.linear, displacement.quadratic)

## Analysis of Variance Table  
##   
## Model 1: mpg ~ displacement  
## Model 2: mpg ~ poly(displacement, 2)  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 390 8378.8   
## 2 389 7412.3 1 966.56 50.726 5.175e-12 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

As we can see, the quadratic term has a p-value of essentially zero, which is quite statistically significant. Moreover, the inclusion of the quadratic term improves the value from 0.6482 in the linear model to 0.6888. This, along with the above results of using the anova() function to compare the two models, strongly suggests that the model which includes the quadtratic term is a better fit than the model which does not include it. To finish, lets now compare the quadratic model to a quintic one.

displacement.quintic = lm(mpg ~ poly(displacement, 5), data = Auto)  
summary(displacement.quintic)

##   
## Call:  
## lm(formula = mpg ~ poly(displacement, 5), data = Auto)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -15.3360 -2.3445 -0.2895 2.1635 20.3439   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 23.4459 0.2209 106.158 < 2e-16 \*\*\*  
## poly(displacement, 5)1 -124.2585 4.3728 -28.416 < 2e-16 \*\*\*  
## poly(displacement, 5)2 31.0895 4.3728 7.110 5.67e-12 \*\*\*  
## poly(displacement, 5)3 -4.4655 4.3728 -1.021 0.308   
## poly(displacement, 5)4 0.7747 4.3728 0.177 0.859   
## poly(displacement, 5)5 3.2991 4.3728 0.754 0.451   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.373 on 386 degrees of freedom  
## Multiple R-squared: 0.6901, Adjusted R-squared: 0.6861   
## F-statistic: 171.9 on 5 and 386 DF, p-value: < 2.2e-16

anova(displacement.quadratic, displacement.quintic)

## Analysis of Variance Table  
##   
## Model 1: mpg ~ poly(displacement, 2)  
## Model 2: mpg ~ poly(displacement, 5)  
## Res.Df RSS Df Sum of Sq F Pr(>F)  
## 1 389 7412.3   
## 2 386 7380.8 3 31.425 0.5478 0.6499

First, we notice that none of the terms above order 2 (i.e. the cubic, quartic, and quintic terms) have statistically significant p-values. In addition, the adjusted value has dropped slightly from 0.6872 in the quadratic model to 0.6861. Lastly, p-value from the anova() function is 0.65, which means that there is not sufficient evidence to reject the null hypothesis that the quintic model is a better fit than the quadratic one. These three pieces of evidence suggest that including terms beyond order 2 does not improve the model.

# Applied Exercise 3.7.10

**This question should be answered using the Carseats data set.**

library(ISLR)  
head(Carseats)

## Sales CompPrice Income Advertising Population Price ShelveLoc Age Education  
## 1 9.50 138 73 11 276 120 Bad 42 17  
## 2 11.22 111 48 16 260 83 Good 65 10  
## 3 10.06 113 35 10 269 80 Medium 59 12  
## 4 7.40 117 100 4 466 97 Medium 55 14  
## 5 4.15 141 64 3 340 128 Bad 38 13  
## 6 10.81 124 113 13 501 72 Bad 78 16  
## Urban US  
## 1 Yes Yes  
## 2 Yes Yes  
## 3 Yes Yes  
## 4 Yes Yes  
## 5 Yes No  
## 6 No Yes

## 3.7.10.a

**Fit a multiple regression model to predict Sales using Price, Urban, and US.**

carseats.fit.1 = lm(Sales ~ Price + Urban + US, data = Carseats)  
summary(carseats.fit.1)

##   
## Call:  
## lm(formula = Sales ~ Price + Urban + US, data = Carseats)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -6.9206 -1.6220 -0.0564 1.5786 7.0581   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 13.043469 0.651012 20.036 < 2e-16 \*\*\*  
## Price -0.054459 0.005242 -10.389 < 2e-16 \*\*\*  
## UrbanYes -0.021916 0.271650 -0.081 0.936   
## USYes 1.200573 0.259042 4.635 4.86e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.472 on 396 degrees of freedom  
## Multiple R-squared: 0.2393, Adjusted R-squared: 0.2335   
## F-statistic: 41.52 on 3 and 396 DF, p-value: < 2.2e-16

contrasts(Carseats$Urban)

## Yes  
## No 0  
## Yes 1

contrasts(Carseats$US)

## Yes  
## No 0  
## Yes 1

## 3.7.10.b

**Provide an interpretation of each coefficient in the model. Be careful – some of the variables in the model are qualitative!**

The coefficient of -0.054459 for Price means that, for a given location (i.e. fixed values of Urban and US), increasing the price of a car seat by \$1 results in a decrease of sales by approximately 54.46 units, on average, in the model. The coefficient of -0.021916 for UrbanYes means that, for a given carseat price point and value of US, the model predicts urban areas to have approximately 22 fewer carseat sales on average compared to non-urban areas. The coefficient of 1.200573 for USYes means that, for a given carseat price point and value of Urban, the model predicts that stores in the United States have 1201 more carseat sales on average than stores outside the United States.

## 3.7.10.c

**Write out the model in equation form, being careful to handle the qualitative variables properly.**

The model has the following equation.

Here, is the estimated carseat sales, in thousands of car seats; is the price of the carseat at the jth store, in dollars; and and are dummy variables to represent whether or not the th store at is located in an urban area and in the United States, respectively. More concretely, and use the following coding scheme.

## 3.7.10.d

**For which of the predictors can you reject the null hypothesis ?**

The p-values for the intercept, Price, and USYes are all essentially zero, which provides strong evidence to reject the null hypothesis for those predictors. The p-value for UrbanYes, however, is 0.936, so there is no evidence to reject the null hypothesis that it has a non-zero coefficient in the true relationship between the predictors and Sales.

## 3.7.10.e

**On the basis of your response to the previous question, fit a smaller model that only uses the predictors for which there is evidence of association with the outcome.**

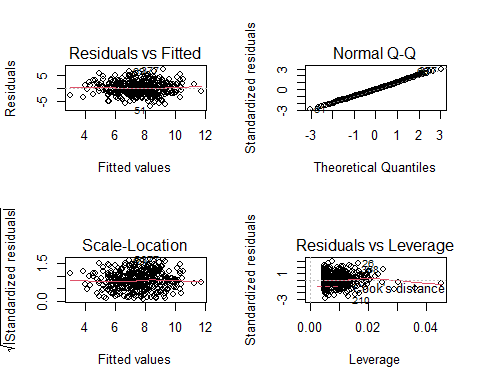
carseats.fit.2 = lm(Sales ~ Price + US, data = Carseats)  
summary(carseats.fit.2)

##   
## Call:  
## lm(formula = Sales ~ Price + US, data = Carseats)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -6.9269 -1.6286 -0.0574 1.5766 7.0515   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 13.03079 0.63098 20.652 < 2e-16 \*\*\*  
## Price -0.05448 0.00523 -10.416 < 2e-16 \*\*\*  
## USYes 1.19964 0.25846 4.641 4.71e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.469 on 397 degrees of freedom  
## Multiple R-squared: 0.2393, Adjusted R-squared: 0.2354   
## F-statistic: 62.43 on 2 and 397 DF, p-value: < 2.2e-16

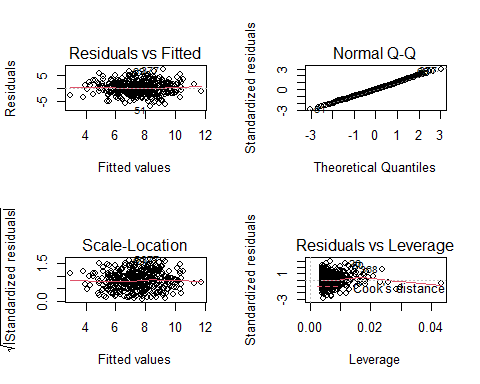
## 3.7.10.f

**How well do the models in (a) and (e) fit the data?**

par(mfrow = c(2, 2))  
plot(carseats.fit.1)



par(mfrow = c(2, 2))  
plot(carseats.fit.2)



The models in Part 1 and Part 5 both fit the data about equally well, with identical values of 0.2393. In addition, when comparing the diagnostic plots between the two models, there isn’t any discernable visual differences that would strongly indicate that one model is a better fit than the other.

## 3.7.10.g

**Using the model from (e), obtain 95% confidence intervals for the coefficient(s).**

confint(carseats.fit.2)

## 2.5 % 97.5 %  
## (Intercept) 11.79032020 14.27126531  
## Price -0.06475984 -0.04419543  
## USYes 0.69151957 1.70776632

## 3.7.10.h

**Is there evidence of outliers or high leverage observations in the model from Part 5?**

When we look at the residuals vs. leverage plot for the model from Part 5 that I generated in Part 6, we see that there are a number of observations with standardized residuals close to 3 in absolute value. Those observations are possible outliers. We can also see in the same plot that there are number of high leverage points with leverage values greatly exceeding the average leverage of , though those high leverage observations are not likely outliers, as they have studentized residual values with absolute value less than 2.