**Homework 1**

Xiao-Qun Wang, Zili Ma

***Q1. For each of parts (a) through (d), indicate whether i. or ii. is correct, and explain your answer. In general, do we expect the performance of a flexible statistical learning method to perform better or worse than an inflexible method when :***

**Answer 1:**

**(a) *The sample size is extremely large, and the number of predictors is small ?***

It will be better. Because for a large sample size, a flexible approach will fit the data more precise and gain less bias than an inflexible approach. Although there may exists some overfitting problems for more flexible approach, but it will less likely to happen for a large sample size.

**(b) *The number of predictors is extremely large, and the number of observations is small ?***

It will be worse. A more flexible method will cause overfitting problem because of the small number of observations. It will usually cause larger increase in variance and smaller reduction on bias.

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

It will be better. For a highly non-linear model, it may have more degrees of freedom compare to linear model, so it is necessary to use a more flexible model in order to find non-linear effect.

**(d) *The variance of the error terms is extremely high ?***

It will be worse. If the variance of the error terms is large, it may have too much noise in the data. For a flexible approach, it may fit to more noise and therefore leads to more error terms. Which will results in the increasing of variance.

**2. 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)***

**Irreducible error:** The irreducible error is a constant so it will not change with flexibility. This curve will lies below the test MSE curve, because the expected test MSE will always greater. (bias-variance-trade-off relation)  
**Training MSE:** When flexibility increases, the curve will fits the observed data more closely. Therefore the training error will declines monotonically as flexibility increases.

**Test MSE:** The test error will declines as flexibility increases at the beginning, but when model yields a small training MSE but a large test MSE, it may actually overfitting the data. So the test error will starts to increase. This is because the flexibility model tries too hard to find as may properties as possible, and fit them together. But some features may merely be noise or irrelevant to the patterns in the training dataset.

**The variance and squared bias:** Generally, the squared bias will decreases monotonically while the variance increases monotonically. This is because variance refers to the amount of curve would change if we using a different training data set to estimate. So if we use more flexible model to fit the observed data, it will fits to them very closely, but changing any point may cause the corresponding curve to change considerably, and therefore will result in larger variance. Meanwhile, bias refers to the error that is introduced by approximating with a much simpler model, which means the model with less flexibility (more close to linear regression). And most of time, real-life problem is unlikely has just a simple linear relationship. So with less flexibility model, it may result in more bias in the estimation.

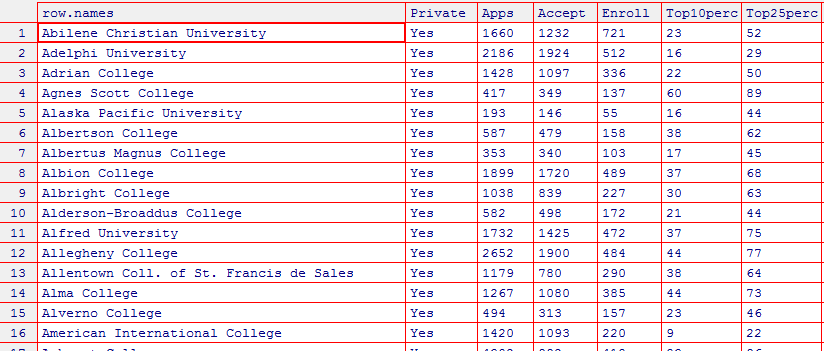
**4. This exercise relates to the “College” data set, which can be found in the file “College.csv”. It contains a number of variables for 777 different universities and colleges in the US.**

**(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.***

* college = read.csv("College(1).csv")

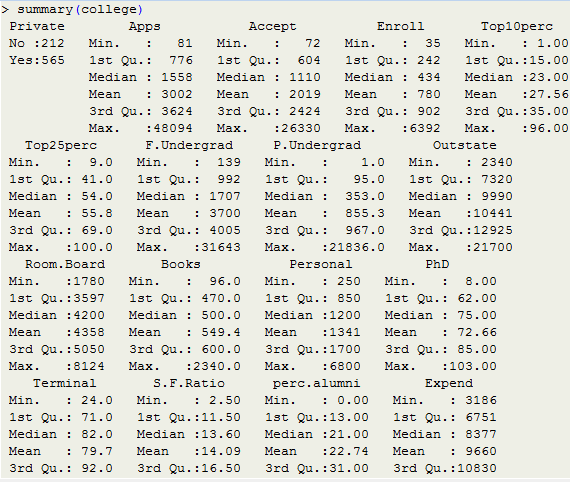
**(b.) *Look at the data using the fix() function. You should notice that the first column is just the name of each university. We don’t really want R to treat this as data. However, it may be handy to have these names for later.***

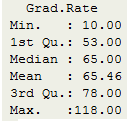
* fix(college)
* rownames(college) = college[,1]
* fix(college)
* college = college[,-1]
* fix(college)

******

**(c.)**

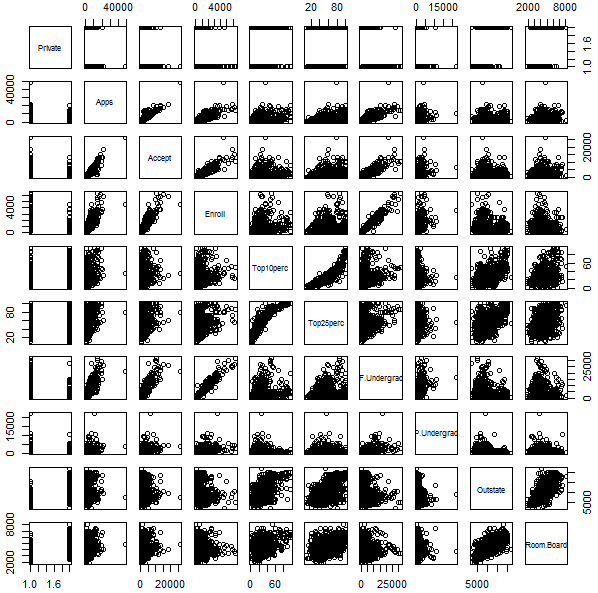
**(i.)*Use the summary() function to produce a numerical summary of the variables in the data set.***

****

****

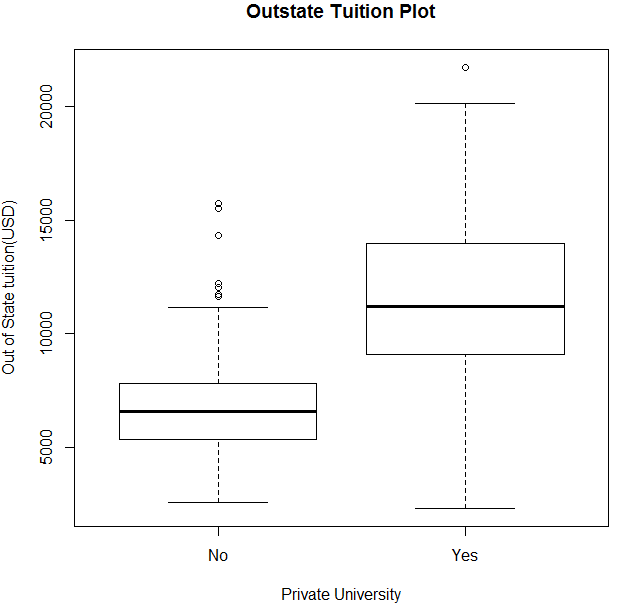
**(ii.) *Use the pairs() function to produce a scatterplot matrix of the first ten columns or variables of the data.***

* pairs(college[,1:10])

******

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

* plot(college$Outstate~college$Private, xlab = "Private University", ylab ="Out of State tuition(USD)", main = "Outstate Tuition Plot")

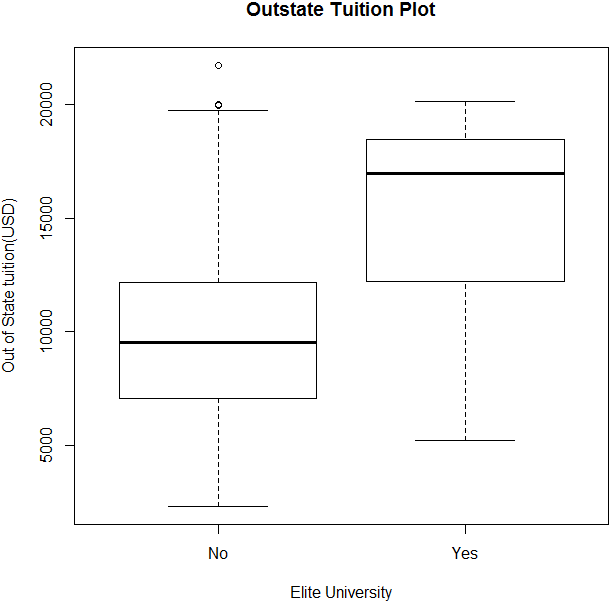


**(iv.) *Create a new qualitative variable, called “Elite”, by binning the “Top10perc” variable. We are going to divide universities into two groups based on whether or not the proportion of students coming from the top 10% of their high school classes exceeds 50%. Use the summary() function to see how many elite universities there are. Now use the plot() function to produce side-by-side boxplots of “Outstate” versus “Elite”.***

* Elite = rep("No",nrow(college))
* Elite[college$Top10perc>50] = "Yes"
* Elite = as.factor(Elite)
* college = data.frame(college,Elite)
* summary(college)

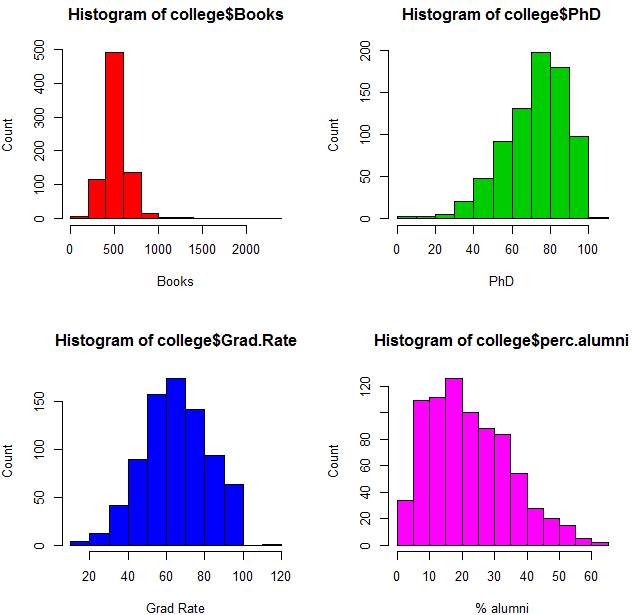
******

* plot(college$Elite, college$Outstate, xlab = "Elite University", ylab ="Out of State tuition(USD)", main = "Outstate Tuition Plot")

******

**(v.) *Use the hist() function to produce some histograms with numbers of bins for a few of the quantitative variables.***

* par(mfrow = c(2,2))
* hist(college$Books, col = 2, xlab = "Books", ylab = "Count")
* hist(college$PhD, col = 3, xlab = "PhD", ylab = "Count")
* hist(college$Grad.Rate, col = 4, xlab = "Grad Rate", ylab = "Count")
* hist(college$perc.alumni, col = 6, xlab = "% alumni", ylab = "Count")

****

**(vi.) Continue exploring the data, and provide a brief summary of what you discover.**

******

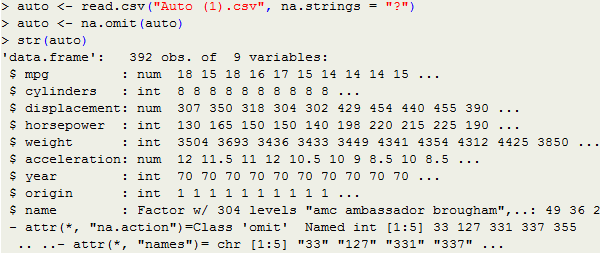
It is weird that some universities with *103%* of faculty with Phd’s. Then I continue to see the name of these universities.

* find.phd <- college[college$PhD == 103, ]
* nrow(find.phd)
* rownames(find.phd)



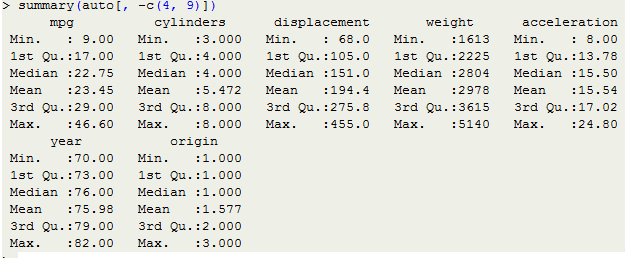
**5. This exercise involves the “Auto” data set studied in the lab. Make sure the missing values have been removed from the data.**

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

****

All variables except “horsepower” and “name” are quantitative.

**(b.) What is the range of each quantitative predictor ?**



As shown above, the range of these predictors are:  
mpg: Min: 9.00 Max: 46.60  
cylinders: Min: 3.00 Max: 8.00  
displacement: Min: 68.00 Max: 455.00

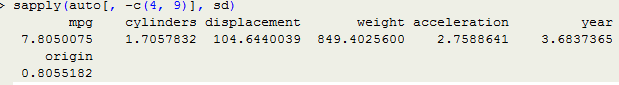
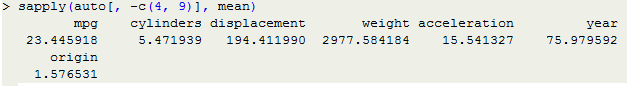
weight: Min: 1613.00 Max: 5140.00

acceleration: Min: 8.00 Max: 24.80

year: Min: 70 Max: 82

origin: Min: 1.00 Max: 3.00

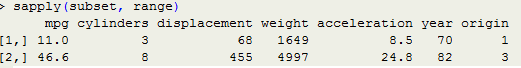
**(c.) *What is the mean and standard deviation of each quantitative predictor ?***  
Mean:

Standard deviation:

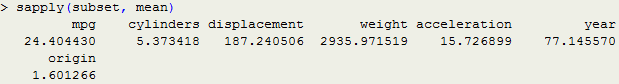
**(d.)** ***Now remove the 10th through 85th observations. What is the range, mean, and standard deviation of each predictor in the subset of the data that remains ?***

* subset <- auto[-c(10:85), -c(4,9)]

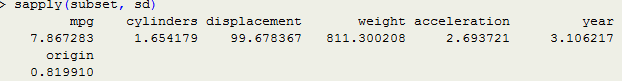
Range:



Mean:

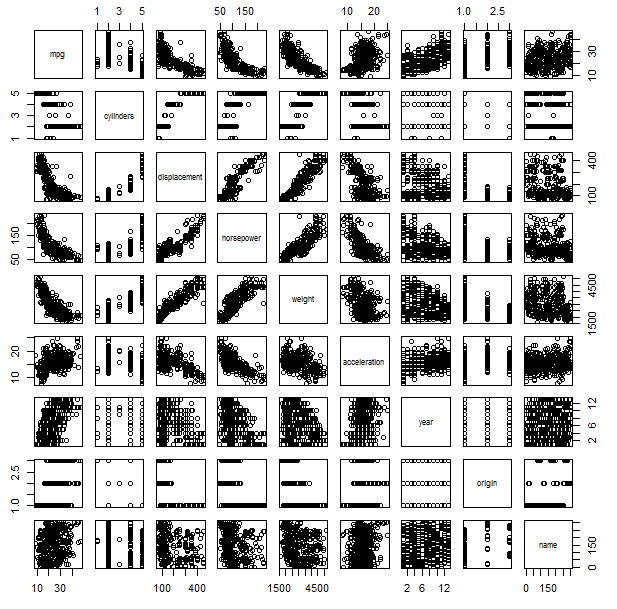


Standard deviation:



**(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.***

* auto$cylinders <- as.factor(auto$cylinders)
* auto$year <- as.factor(auto$year)
* auto$origin <- as.factor(auto$origin)
* pairs(auto)

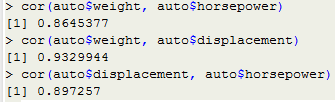


As shown above, Weight, displacement and horsepower have an inverse relation with mpg. There also has overall increase in mpg over the years.

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

Useful: As shown above, the cylinders, horsepower, year and origin can be used as predictors.

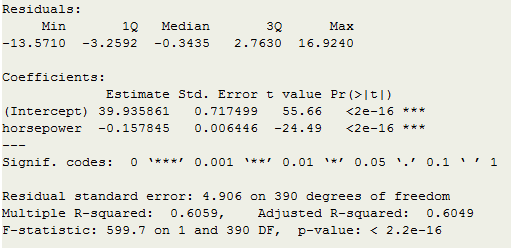
Unuseful: Displacement and weight are not useful. Because they are highly correlated with horespower and with each other, therefore they are redundant factor. The correlation between weight and horsepower is 0.8645377, between weight and displacement is 0.9329944, between displacement and horsepower is 0.897257.



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

***(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 :*(i.) *Is there a relationship between the predictor and the response ?***

* auto <- read.csv("Auto (1).csv", na.strings = "?")
* auto <- na.omit(auto)
* fit <- lm(mpg ~ horsepower, data = auto)
* summary(fit)



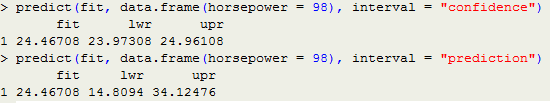
The p-value of F-statistic is smaller than 2.2e-16. So we can reject the hypothesis which states the coefficient between “mpg” and “horsepower” are zero. There has a clear evidence of a relationship between “mpg” and “horsepower”.

**(ii.) *How strong is the relationship between the predictor and the response ?***

We can note that as the R-squared is equal to 0.6059, almost 60.59% of the variability in “mpg” can be explained using “horsepower”.

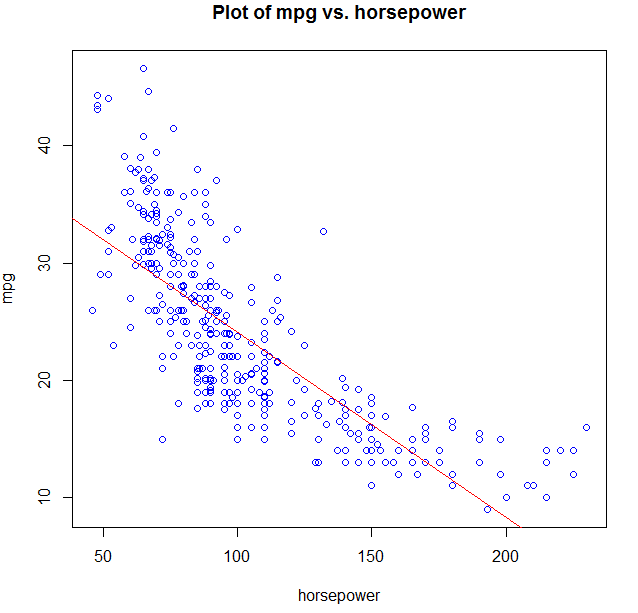
**(iii.) *Is the relationship between the predictor and the response positive or negative ?***

The relationship is negative.As shown above, the coefficient of “horsepower” is negative, therefore the relationship is also negative. Which means, if the automobile equipped with more horsepower, then the less mpg fuel efficiency it will have.

**(iv.) *What is the predicted mpg associated with a “horsepower” of 98 ? What are the associated 95% confidence and prediction intervals ?*****

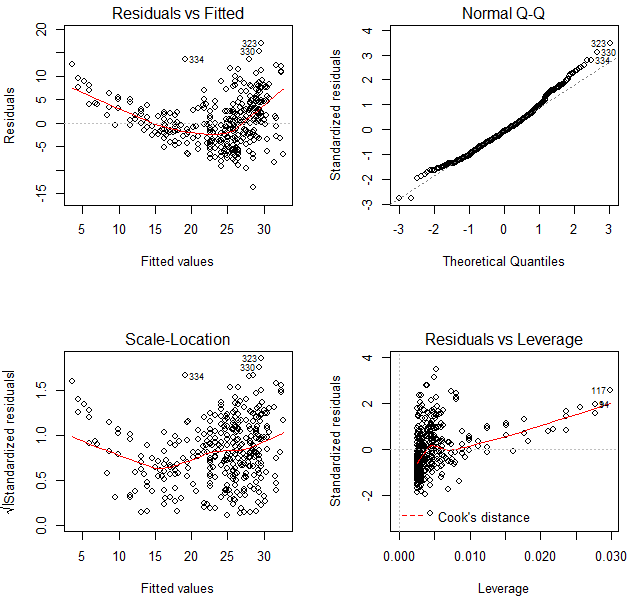
**(b.) *Plot the response and the predictor. Use the abline() function to display the least squares regression line.***

* plot(auto$horsepower, auto$mpg, main = "Plot of mpg vs. horsepower", xlab = "horsepower", ylab = "mpg", col = "blue")
* abline(fit, col = "red")

**

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

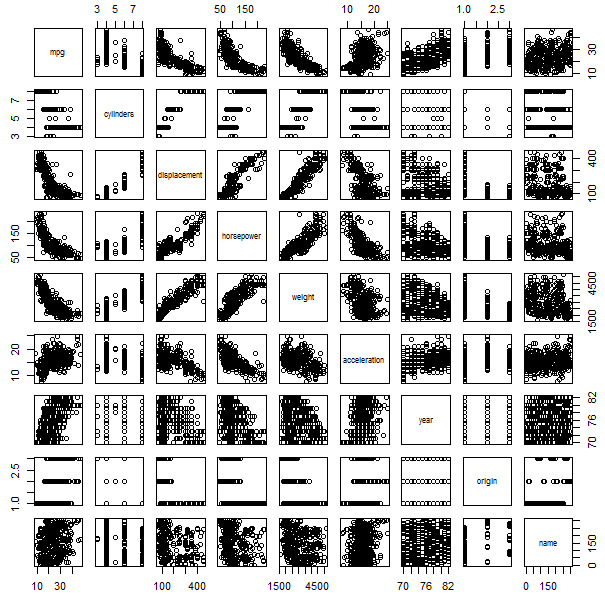
* par(mfrow = c(2, 2))
* plot(fit)



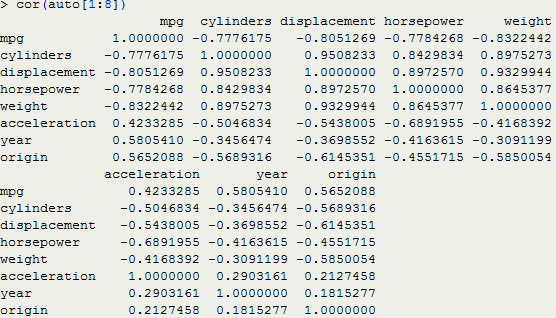
As shown above, first, the plot of residuals versus fitted values indicates there has non-linearity in the data. Second, the plot of residuals versus leverage indicates there has several outliers (higher than 2 or lower than -2) and some high leverage points.

**10. *This question involves the use of multiple linear regression on the “Auto” data set.  
(a.) Produce a scatterplot matrix which include all the variables in the data set.***

* auto <- read.csv("Auto (1).csv", na.strings = "?")
* auto <- na.omit(auto)
* pairs(auto)

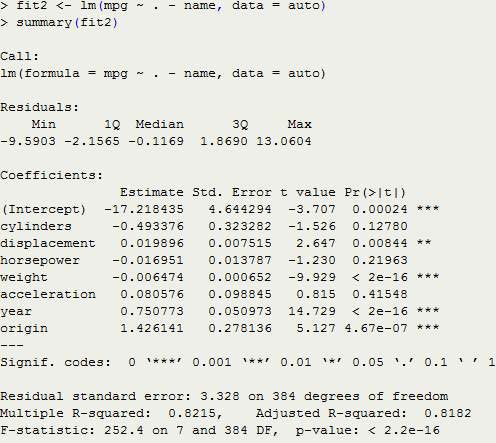


***(b.) Compute the matrix of correlations between the variables using the function cor(). You will need to exclude the “name” variable, which is qualitative.***

******

***(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 :***

1. ***Is there a relationship between the predictors and the response ?***

******

The p-value of F-statistic is smaller than 2.2e-16. So we can reject the hypothesis which states the coefficient between “mpg” and other predictors are zero. There has a clear evidence of a relationship between “mpg” and other predictors.

1. ***Which predictors appear to have a statistically significant relationship to the response ?***

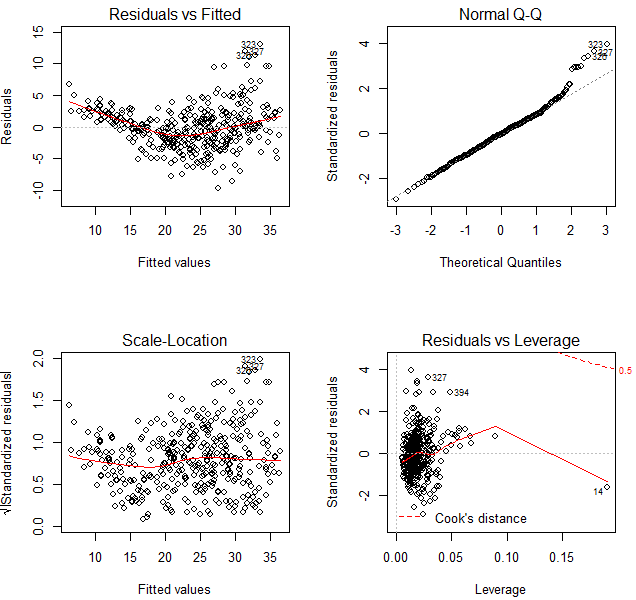
By checking the p-values associated with each predictor’s t-statistic. We may conclude that all predictors are statistically significant except “cylinders”, “horsepower” and “acceleration”.

1. ***What does the coefficient for the “year” variable suggest ?***

While all other predictors remaining constant, increasing the “year” variable by 1 will leads to an increase of 0.750773 in “mpg”. So the cars become more fuel efficient every year by 0.750773 mpg / year

***(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 plots identify any observations with unusually high leverages ?***

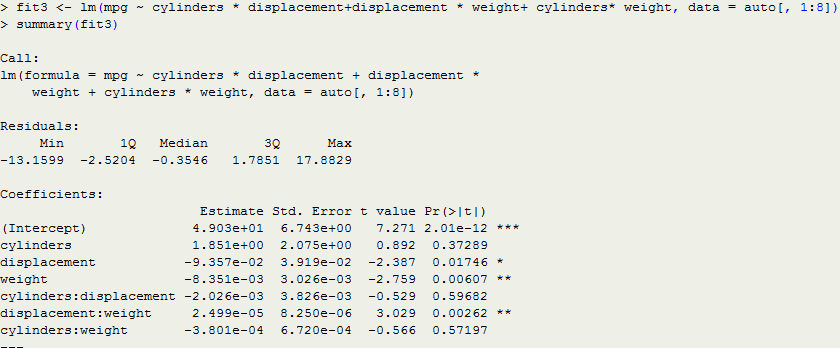
* par(mfrow = c(2, 2))
* plot(fit2)

******

As shown above, first, the plot of residuals versus fitted values indicates the presence of some non linearity in the data. Second, the plot of residuals versus leverage indicates that there are few outliers (higher than 2 or lower than -2) and one high leverage outlier point (point 14).

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

According to the result in **(b.),** we can find that the three highest correlated pairs are (cylinders, displacement), (displacement, weight), (cylinders, weight). So we use them to observe interaction effects.

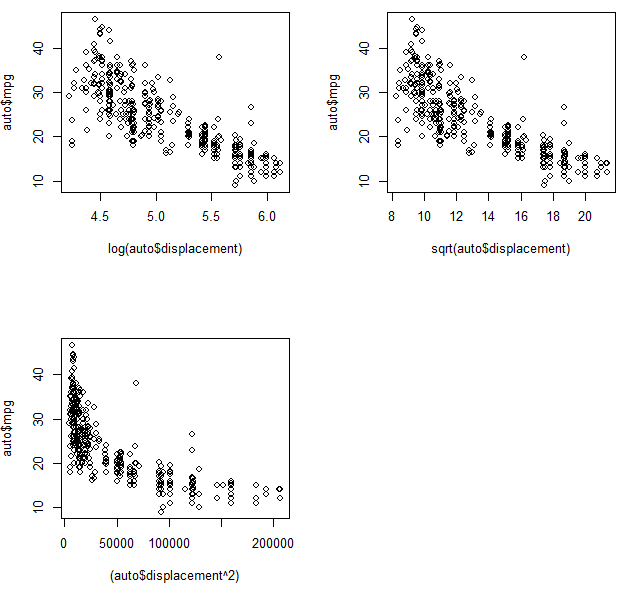
****

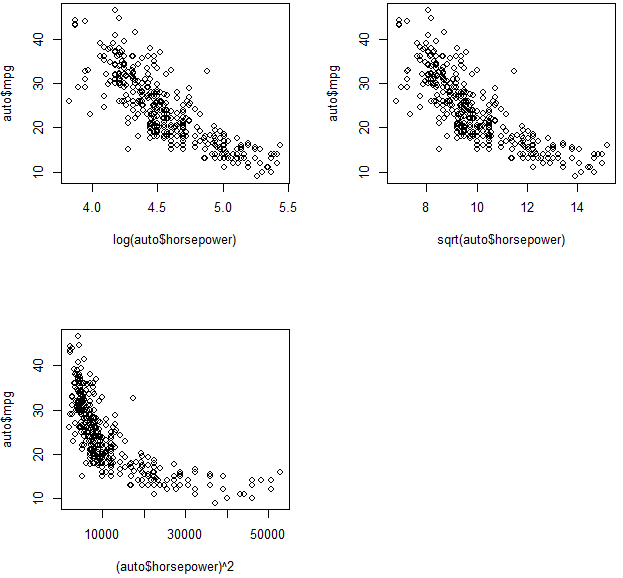
As shown above,from the p-values, we can see that the interaction between displacement and weight is statistically significant, while the interaction between cylinders and displacement, cylinders and weight are not.

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

We let “displacement” and “horsepower” be the only predictor respectively.

* par(mfrow = c(2, 2))
* plot(log(auto$displacement), auto$mpg)
* plot(sqrt(auto$displacement), auto$mpg)
* plot((auto$displacement^2), auto$mpg)



* par(mfrow = c(2, 2))
* plot(log(auto$horsepower), auto$mpg)
* plot(sqrt(auto$horsepower), auto$mpg)
* plot((auto$horsepower)^2, auto$mpg)

For both of them, the most linear looking plot appear in the log transformation.