Homework 1

Xiao-Qun Wang, Zili Ma

Q1. 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. $\sigma^2 = \mathrm{Var}(\varepsilon)$, is extremely high.

Answer 1:

(a) Flexible method is better. For an extremely large sample size, the data set provides enough freedom to figure out a more complex flexible model, which give us a better estimation.

(b) Flexible method is worse than inflexible method. Flexible method can lead to the overfitting when the sample only has a limited observation number.

(c) Flexible method is better. For highly non-linear correlation, flexible method can give a better estimation, if the sample size is large enough.

(d) Flexible method performs worse in this case. The closely fitting result derived from the flexible method will overfit with a large variance the sample whose variance is very large.

Q2. 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.

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

Answer 2:

(a) See the hand-written page in the last part.

(b)

1. The value of **bias** continuously decreases when the flexibility increases. Because the assumed function used in the learning method is typically different from the function. Therefore, a method with less flexibility would introduce more error between the true function and our assumption than a more flexible method, in other words, during the estimation process, the inflexible method can result in more bias.
2. **Variance** value increases monotonically when the flexibility increases. Because for a more flexible method, the regression curve/plane in p dimension will approach to all the sample point more closely.
3. **Bayes error** or **irreducible** error keeps constant below training and test MSE curves for different flexibility. Due to the trade-off relationship between bias and variance, MSE is always larger than Bayer error.
4. The value of **training MSE** continuously decreases when the flexibility increase. Because increasing flexibility of the model will lead to the function of result approach to the training sample point closely or even pass it. Then, parameters will be very sensitive to those new data points. It results in an increasingly larger variance of the model.
5. **Test MSE** value has a U-shape when the flexibility increase. At the beginning, increased flexibility will decrease test value until the function reaches a reasonable fitting to the training sample. However, when the flexibility continues to increase, the function trained by the learning method will start to fit those points whose “patterns” in fact are caused by random noise. The overfitting to training samples increases the test MSE.

Q3. Describe the differences between a parametric and a non-parametric statistical learning approach. What are the advantages of a parametric approach to regression or classification (as opposed to a nonparametric approach)? What are its disadvantages?

Answer 3:

The **parametric statistical learning approach** uses some explicit assumptions to solve the problem. It reduces the problem of estimating f down to one of estimating a set of parameters. .The model using for estimation sometimes is given empirically, such as SLR or MLR. Typically. they involve a two-step model based approach

Non-parametric methods do not make explicit assumptions about the functional form of *f*. The fitting or learning process is done numerically to avoid the danger of using wrong model in the study.

Its advantages and disadvantages are as follows:

Methods	Advantages	Disadvantages
Parametric	1. Simplifies the problem by the given explicit model. 2. Typically, a smaller number of observations is needed in parametric methods than that of non-parametric approaches required	1. The pre-specific model could be not true, or not accurate enough to explain the data set in details. 2. Overfitting occurs when the model is too flexible.
Non-parametric	1. By avoiding the assumption of a particular functional form for *f*, they have the potential to accurately fit a wider range of possible shapes for *f*.	1. Typically, without the simplification of the problem, they require far more number of observation to fit datasets. 2. Overfitting occurs if a too low level of smoothness is used.

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)
```

	row.names	Private	Apps	Accept	Enroll	Top10perc	Top25perc
1	Abilene Christian University	Yes	1660	1232	721	23	52
2	Adelphi University	Yes	2186	1924	512	16	29
3	Adrian College	Yes	1428	1097	336	22	50
4	Agnes Scott College	Yes	417	349	137	60	89
5	Alaska Pacific University	Yes	193	146	55	16	44
6	Albertson College	Yes	587	479	158	38	62
7	Albertus Magnus College	Yes	353	340	103	17	45
8	Albion College	Yes	1899	1720	489	37	68
9	Albright College	Yes	1038	839	227	30	63
10	Alderson-Broaddus College	Yes	582	498	172	21	44
11	Alfred University	Yes	1732	1425	472	37	75
12	Allegheny College	Yes	2652	1900	484	44	77
13	Allentown Coll. of St. Francis de Sales	Yes	1179	780	290	38	64
14	Alma College	Yes	1267	1080	385	44	73
15	Alverno College	Yes	494	313	157	23	46
16	American International College	Yes	1420	1093	220	9	22

(c.)

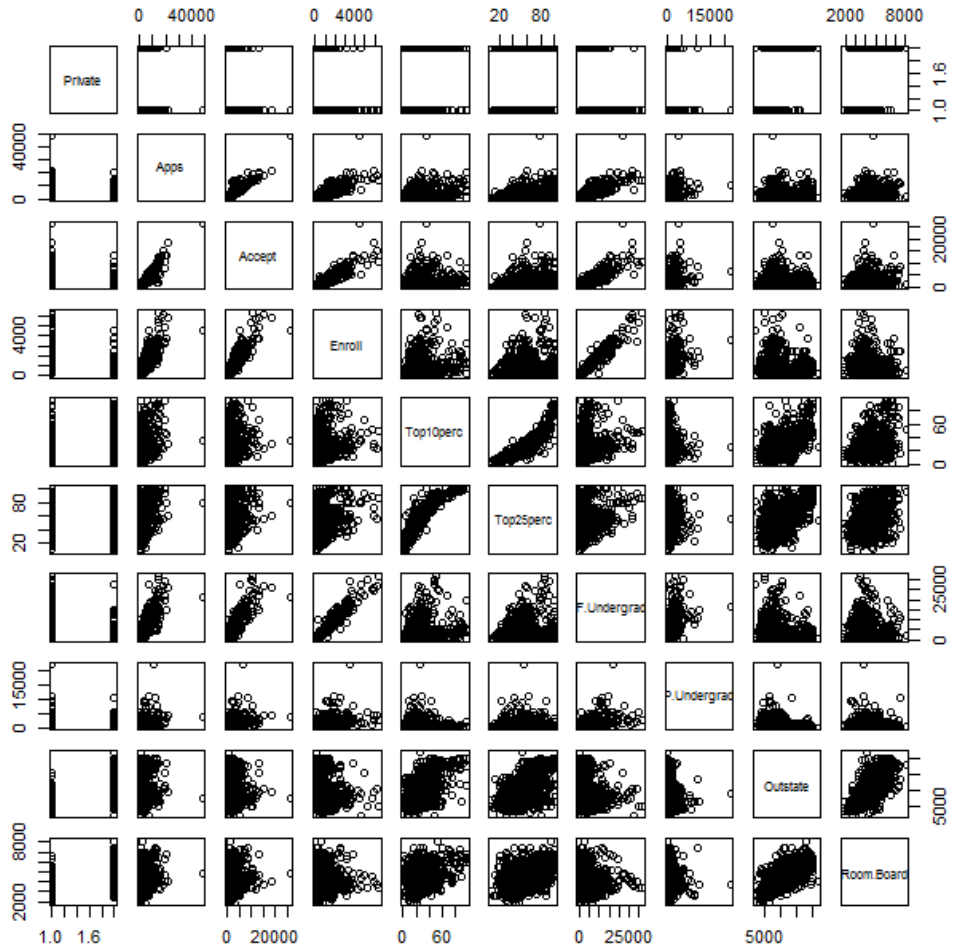
(i.)*Use the summary() function to produce a numerical summary of the variables in the data set.*

```
> 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
```

```
   Grad.Rate
 Min.   : 10.00
 1st Qu.: 53.00
 Median : 65.00
 Mean   : 65.46
 3rd Qu.: 78.00
 Max.   :118.00
```

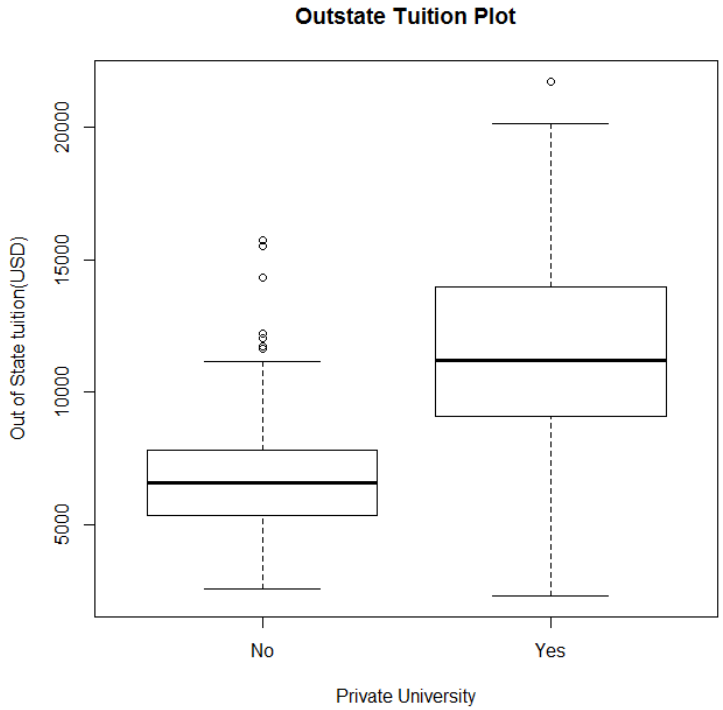
(ii.) ***Use the pairs() function to produce a scatterplot matrix of the first ten columns or variables of the data.***

- `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")
```

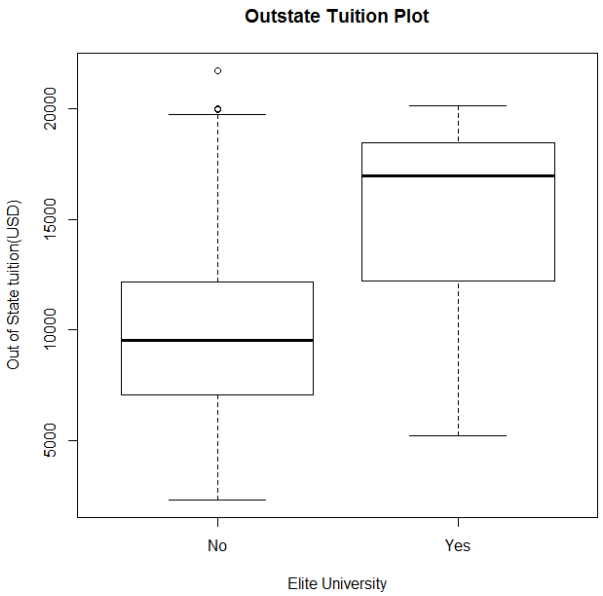


(iv.) ***Create a new qualitative variable, called "Elite", by binning the "Top10perc" variable. We are going to divide universities into two groups based on whether or not the proportion of students coming from the top 10% of their high school classes exceeds 50%. Use the summary() function to see how 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)

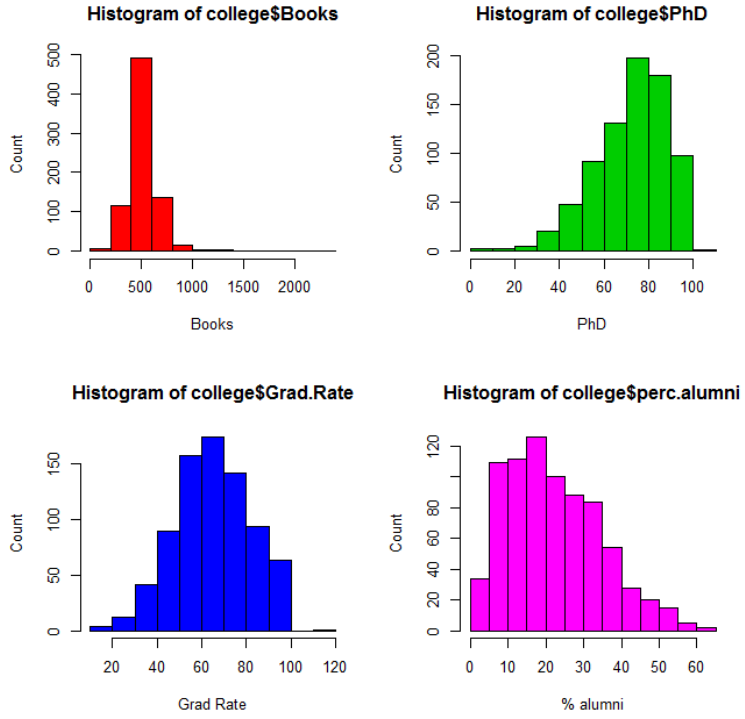
```
Elite
No :699
Yes: 78
```

- 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.

```
> summary(college$PhD)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   8.00   62.00   75.00   72.66   85.00  103.00
```

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)

```
[1] "Texas A&M University at Galveston"
```

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 ?*

```
> auto <- read.csv("Auto (1).csv", na.strings = "?")
> auto <- na.omit(auto)
> str(auto)
'data.frame':   392 obs. of  9 variables:
 $ mpg         : num  18 15 18 16 17 15 14 14 14 15 ...
 $ cylinders   : int  8 8 8 8 8 8 8 8 8 8 ...
 $ displacement: num  307 350 318 304 302 429 454 440 455 390 ...
 $ horsepower  : int  130 165 150 150 140 198 220 215 225 190 ...
 $ weight      : int  3504 3693 3436 3433 3449 4341 4354 4312 4425 3850 ...
 $ acceleration: num  12 11.5 11 12 10.5 10 9 8.5 10 8.5 ...
 $ year        : int  70 70 70 70 70 70 70 70 70 70 ...
 $ origin      : int  1 1 1 1 1 1 1 1 1 1 ...
 $ name        : Factor w/ 304 levels "amc ambassador brougham",..: 49 36 2
 - attr(*, "na.action")=Class 'omit'  Named int [1:5] 33 127 331 337 355
  .. ..- attr(*, "names")= chr [1:5] "33" "127" "331" "337" ...
```

All variables except "horsepower" and "name" are quantitative.

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

```
> summary(auto[, -c(4, 9)])
      mpg          cylinders      displacement       weight      acceleration
 Min.   : 9.00   Min.   :3.000   Min.   : 68.0   Min.   :1613   Min.   : 8.00
 1st Qu.:17.00   1st Qu.:4.000   1st Qu.:105.0   1st Qu.:2225   1st Qu.:13.78
 Median :22.75   Median :4.000   Median :151.0   Median :2804   Median :15.50
 Mean   :23.45   Mean   :5.472   Mean   :194.4   Mean   :2978   Mean   :15.54
 3rd Qu.:29.00   3rd Qu.:8.000   3rd Qu.:275.8   3rd Qu.:3615   3rd Qu.:17.02
 Max.   :46.60   Max.   :8.000   Max.   :455.0   Max.   :5140   Max.   :24.80
      year           origin
 Min.   :70.00   Min.   :1.000
 1st Qu.:73.00   1st Qu.:1.000
 Median :76.00   Median :1.000
 Mean   :75.98   Mean   :1.577
 3rd Qu.:79.00   3rd Qu.:2.000
 Max.   :82.00   Max.   :3.000
```

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:

```
> sapply(auto[, -c(4, 9)], mean)
         mpg    cylinders displacement       weight acceleration         year
   23.445918     5.471939   194.411990  2977.584184    15.541327    75.979592
      origin
    1.576531
```

Standard deviation:

```
> sapply(auto[, -c(4, 9)], sd)
         mpg    cylinders displacement       weight acceleration         year
   7.8050075    1.7057832  104.6440039  849.4025600    2.7588641    3.6837365
      origin
   0.8055182
```

(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:

```
> sapply(subset, range)
      mpg cylinders displacement weight acceleration year origin
[1,] 11.0         3           68   1649          8.5   70      1
[2,] 46.6         8          455   4997         24.8   82      3
```

Mean:

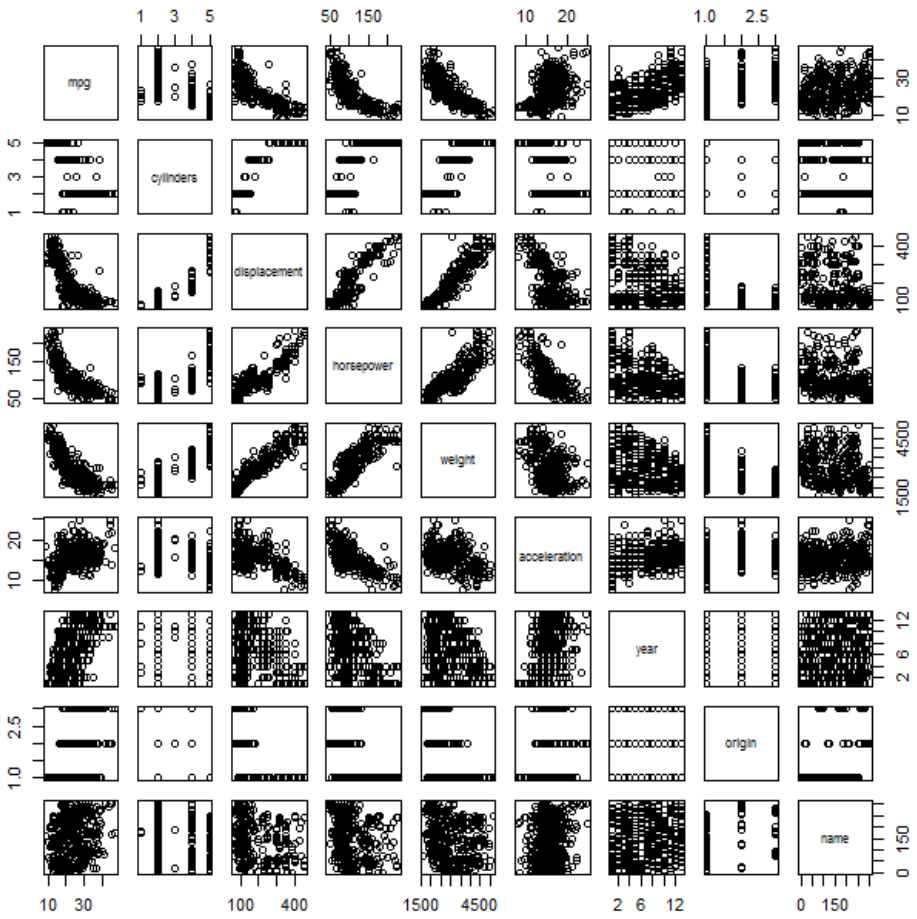
```
> sapply(subset, mean)
         mpg    cylinders displacement       weight acceleration         year
   24.404430     5.373418   187.240506  2935.971519    15.726899    77.145570
      origin
    1.601266
```

Standard deviation:

```
> sapply(subset, sd)
         mpg    cylinders displacement       weight acceleration         year
    7.867283     1.654179    99.678367   811.300208     2.693721     3.106217
      origin
    0.819910
```

(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.

```
> cor(auto$weight, auto$horsepower)
[1] 0.8645377
> cor(auto$weight, auto$displacement)
[1] 0.9329944
> cor(auto$displacement, auto$horsepower)
[1] 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)

```
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
```

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 ?***

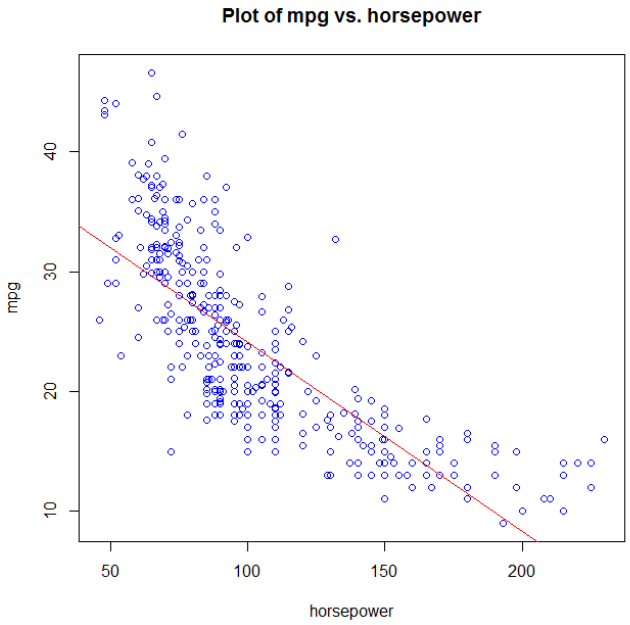
The relationship is negative. As shown above, the coefficient of "horsepower" is negative, therefore the relationship is also negative. Which means, if the automobile equipped with more horsepower, then the less mpg fuel efficiency it will have.

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

```
> predict(fit, data.frame(horsepower = 98), interval = "confidence")
       fit      lwr      upr
1 24.46708 23.97308 24.96108
> predict(fit, data.frame(horsepower = 98), interval = "prediction")
       fit     lwr      upr
1 24.46708 14.8094 34.12476
```

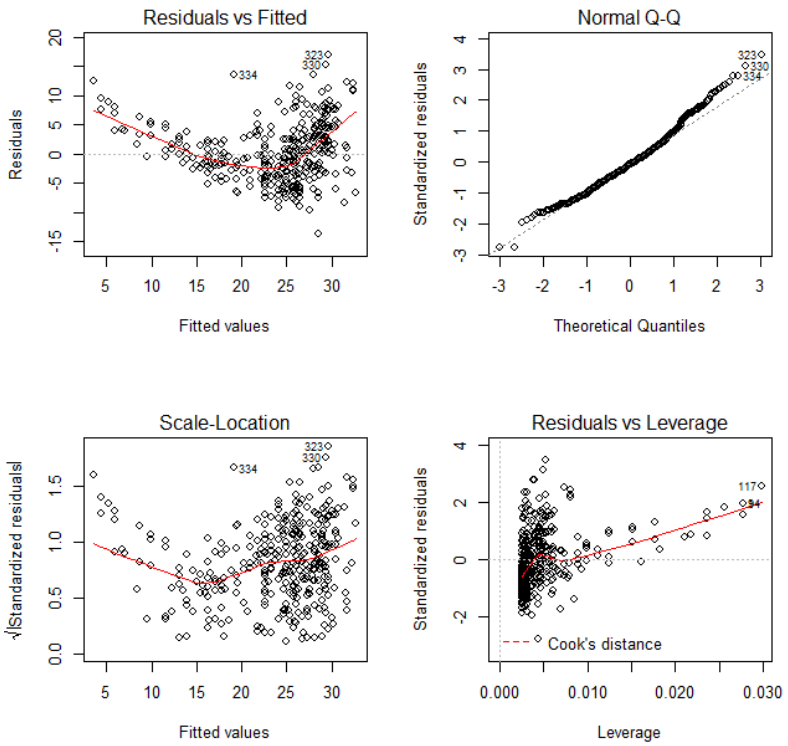
(b.) ***Plot the response and the predictor. Use the abline() function to display the least squares regression line.***

- 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.

Q6. Describe the null hypotheses to which the p-values given in Table 3.4 correspond. Explain what conclusions you can draw based on these p-values. Your explanation should be phrased in terms of sales, *TV*, *radio*, and *newspaper*, rather than in terms of the coefficients of the linear model.

	Coefficient	Std. error	t-statistic	p-value
Intercept	2.939	0.3119	9.42	< 0.0001
TV	0.046	0.0014	32.81	< 0.0001
radio	0.189	0.0086	21.89	< 0.0001
newspaper	-0.001	0.0059	-0.18	0.8599

Answer 6:

In Table 3.4, with the corresponding p-value (smaller than 0.05) is significant, the coefficient for TV represents the average effect of increasing newspaper spending by $1,000 while holding radio and newspaper fixed. For a given amount of radio and newspaper advertising, spending an additional $1,000 on TV advertising leads to an increase in sales by approximately 46 units.

With the corresponding p-value (smaller than 0.05) is significant, the coefficient for radio represents the average effect of increasing newspaper spending by $1,000 while holding TV and newspaper fixed. For a given amount of TV and newspaper advertising, spending an additional $1,000 on radio advertising leads to an increase in sales by approximately 189 units.

The coefficient estimate for newspaper in this multiple regression model is close to zero, and the corresponding p-value (larger than 0.05) is no longer significant.

Q7. Consider the fitted values that result from performing linear regression without an intercept. In this setting, the *i*th fitted value takes the form:

$$\hat{y}_i = x_i\hat{\beta},$$

Where

$$\hat{\beta} = \left(\sum_{i=1}^{n} x_i y_i\right) / \left(\sum_{i'=1}^{n} x_{i'}^2\right). \qquad (3.38)$$

Show that we can write

$$\hat{y}_i = \sum_{i'=1}^{n} a_{i'} y_{i'}.$$

What is $a_{i'}$?

Answer 7:

See the attached paper written by hand.

Q8. Using (3.4), argue that in the case of simple linear regression, the least squares line always passes through the point $(\bar{x}, \bar{y})$.

Answer 8:

The equation of (3.4) is as follows:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \qquad (3.4)$$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x},$$

From the 2nd equation, if we substitute $\bar{\mathbf{x}}$ into this rearrange equation as follows,

$$\mathbf{y} = \hat{\boldsymbol{\beta}}_0 + \hat{\boldsymbol{\beta}}_1\bar{\mathbf{x}} = \bar{\mathbf{y}}$$

Therefore, we can conclude that the linear regression passes through the point **$(\bar{x}, \bar{y})$.**

.

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)

```
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
```

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.

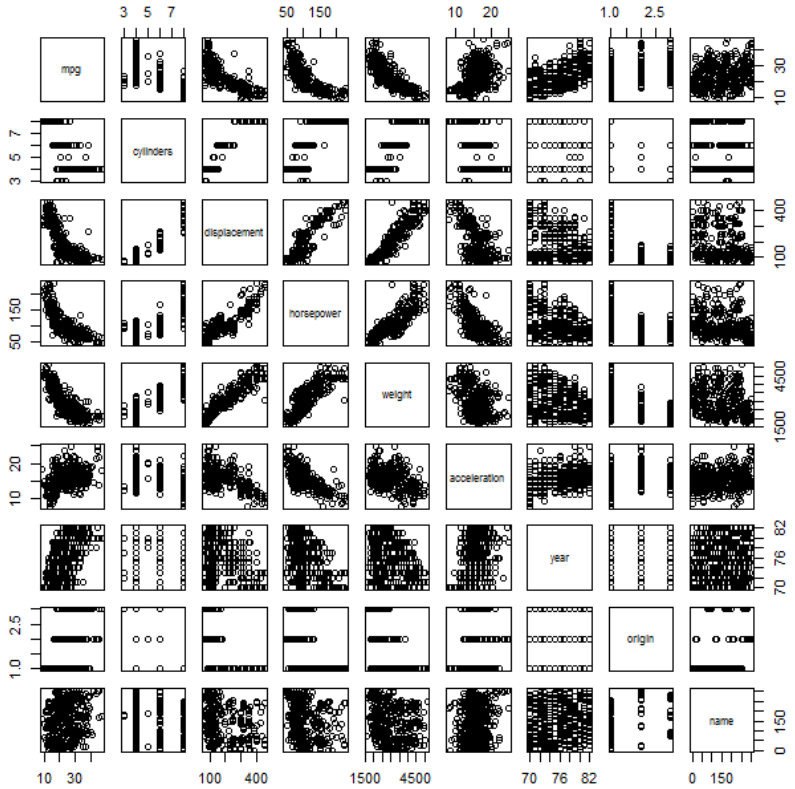
(iv.) ***What is the predicted mpg associated with a "horsepower" of 98 ? What are the associated 95% confidence and prediction intervals ?***

```
> predict(fit, data.frame(horsepower = 98), interval = "confidence")
       fit      lwr      upr
1 24.46708 23.97308 24.96108
> predict(fit, data.frame(horsepower = 98), interval = "prediction")
       fit     lwr      upr
1 24.46708 14.8094 34.12476
```

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.

```
> cor(auto[1:8])
                     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
origin        0.5652088 -0.5689316   -0.6145351 -0.4551715 -0.5850054
              acceleration       year     origin
mpg              0.4233285  0.5805410  0.5652088
cylinders       -0.5046834 -0.3456474 -0.5689316
displacement    -0.5438005 -0.3698552 -0.6145351
horsepower      -0.6891955 -0.4163615 -0.4551715
weight          -0.4168392 -0.3091199 -0.5850054
acceleration     1.0000000  0.2903161  0.2127458
year             0.2903161  1.0000000  0.1815277
origin           0.2127458  0.1815277  1.0000000
```

(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 :

i. Is there a relationship between the predictors and the response ?

```
> fit2 <- lm(mpg ~ . - name, data = auto)
> summary(fit2)

Call:
lm(formula = mpg ~ . - name, data = auto)

Residuals:
    Min      1Q  Median      3Q     Max
-9.5903 -2.1565 -0.1169  1.8690 13.0604

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -17.218435   4.644294  -3.707  0.00024 ***
cylinders     -0.493376   0.323282  -1.526  0.12780
displacement   0.019896   0.007515   2.647  0.00844 **
horsepower    -0.016951   0.013787  -1.230  0.21963
weight        -0.006474   0.000652  -9.929  < 2e-16 ***
acceleration   0.080576   0.098845   0.815  0.41548
year           0.750773   0.050973  14.729  < 2e-16 ***
origin         1.426141   0.278136   5.127 4.67e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.328 on 384 degrees of freedom
Multiple R-squared:  0.8215,    Adjusted R-squared:  0.8182
F-statistic: 252.4 on 7 and 384 DF,  p-value: < 2.2e-16
```

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.

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

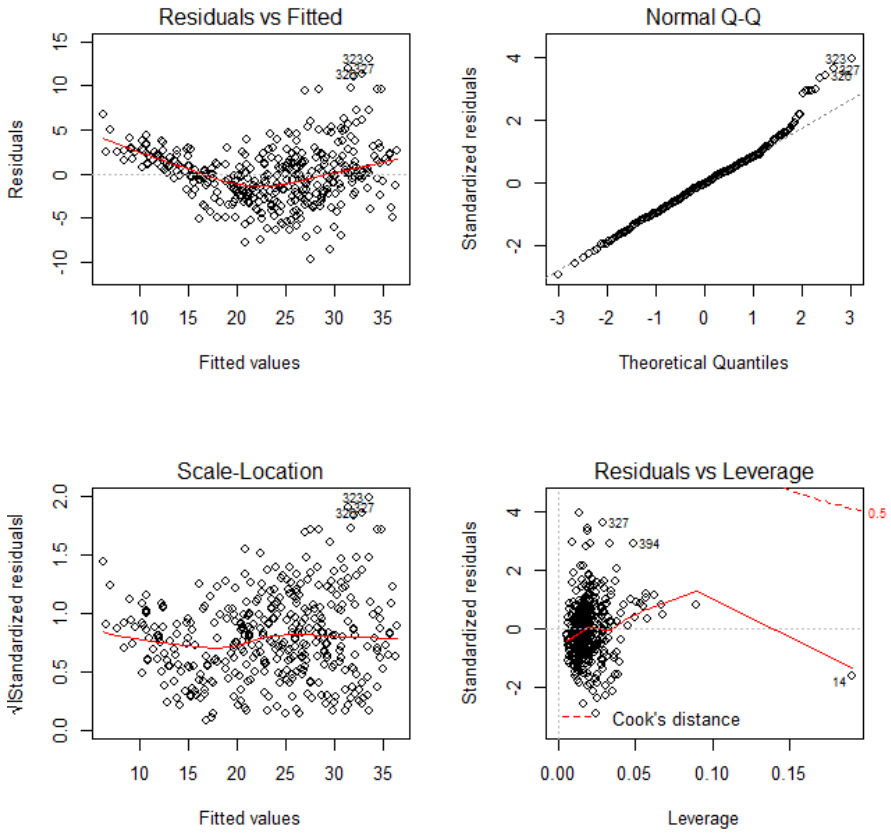
By checking the p-values associated with each predictor's t-statistic. We may conclude that all predictors are statistically significant except "cylinders", "horsepower" and "acceleration".

iii. 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.

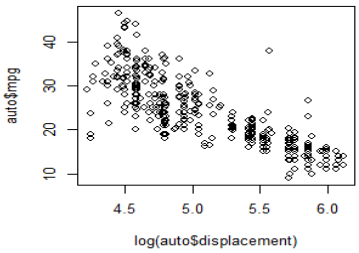
```
> fit3 <- lm(mpg ~ cylinders * displacement+displacement * weight+ cylinders* weight, data = auto[, 1:8])
> summary(fit3)

Call:
lm(formula = mpg ~ cylinders * displacement + displacement *
    weight + cylinders * weight, data = auto[, 1:8])

Residuals:
     Min       1Q   Median       3Q      Max
-13.1599  -2.5204  -0.3546   1.7851  17.8829

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)            4.903e+01  6.743e+00   7.271 2.01e-12 ***
cylinders              1.851e+00  2.075e+00   0.892  0.37289
displacement          -9.357e-02  3.919e-02  -2.387  0.01746 *
weight                -8.351e-03  3.026e-03  -2.759  0.00607 **
cylinders:displacement -2.026e-03  3.826e-03  -0.529  0.59682
displacement:weight    2.499e-05  8.250e-06   3.029  0.00262 **
cylinders:weight      -3.801e-04  6.720e-04  -0.566  0.57197
---
```

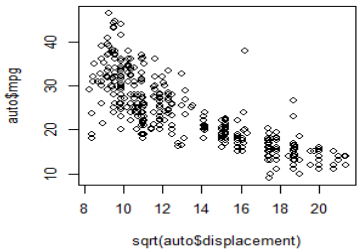
As shown above, from the p-values, we can see that the interaction between displacement and weight is statistically significant, while the interaction between cylinders and displacement, cylinders and weight are not.

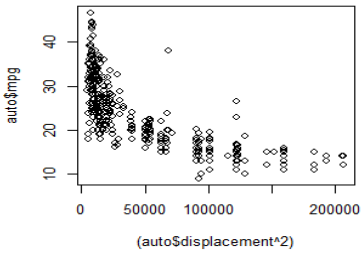
(f.) Try a few different transformations of the variables, such as $\log(X)$, $\sqrt{X}$, X^2, Comment on your findings.

We let "displacement" and "horsepower" be the only predictor respectively.

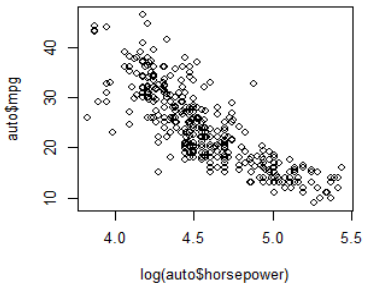
- par(mfrow = c(2, 2))
- plot(log(auto$displacement), auto$mpg)
- plot(sqrt(auto$displacement), auto$mpg)
- plot((auto$displacement^2), auto$mpg)

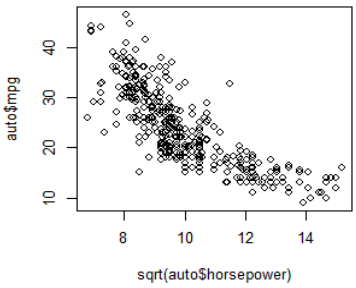


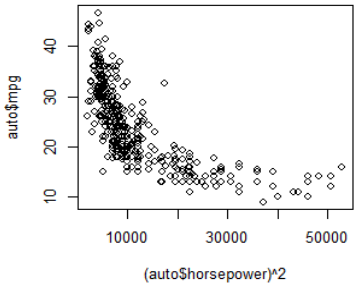




- par(mfrow = c(2, 2))
- plot(log(auto$horsepower), auto$mpg)
- plot(sqrt(auto$horsepower), auto$mpg)
- plot((auto$horsepower)^2, auto$mpg)







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