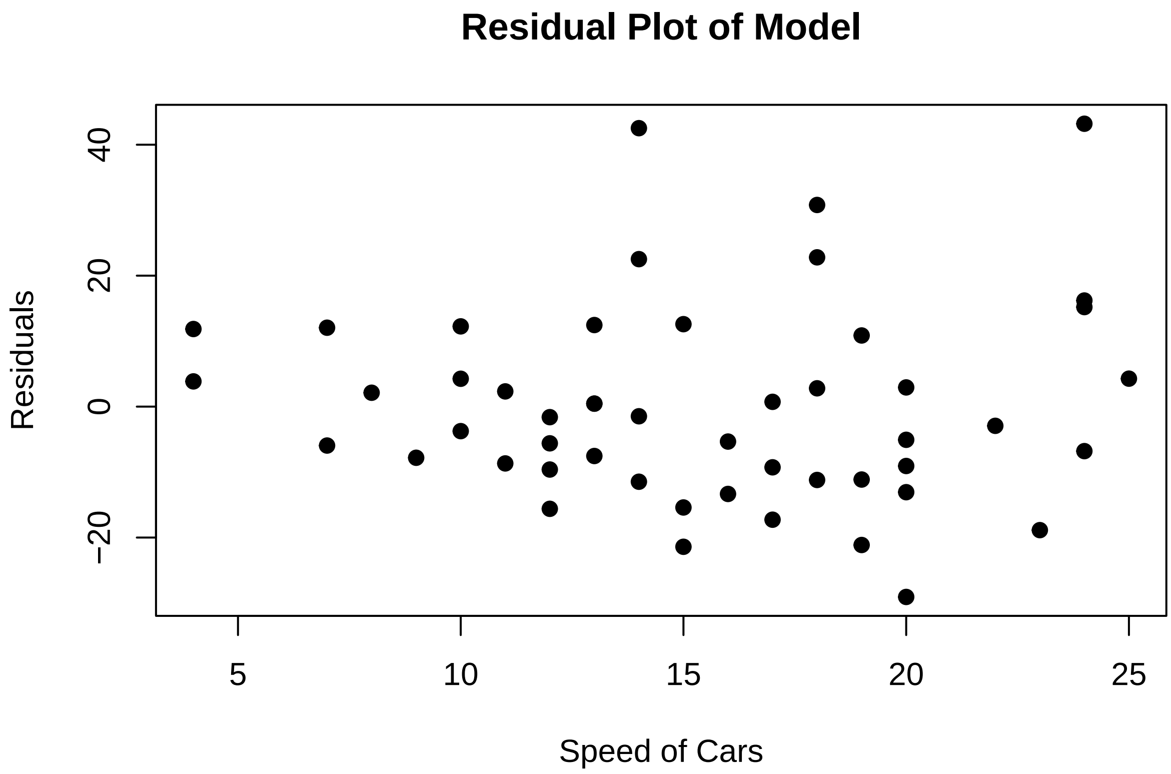
STATS_204_HW3

Qi Wang

Question 1: (4.3)

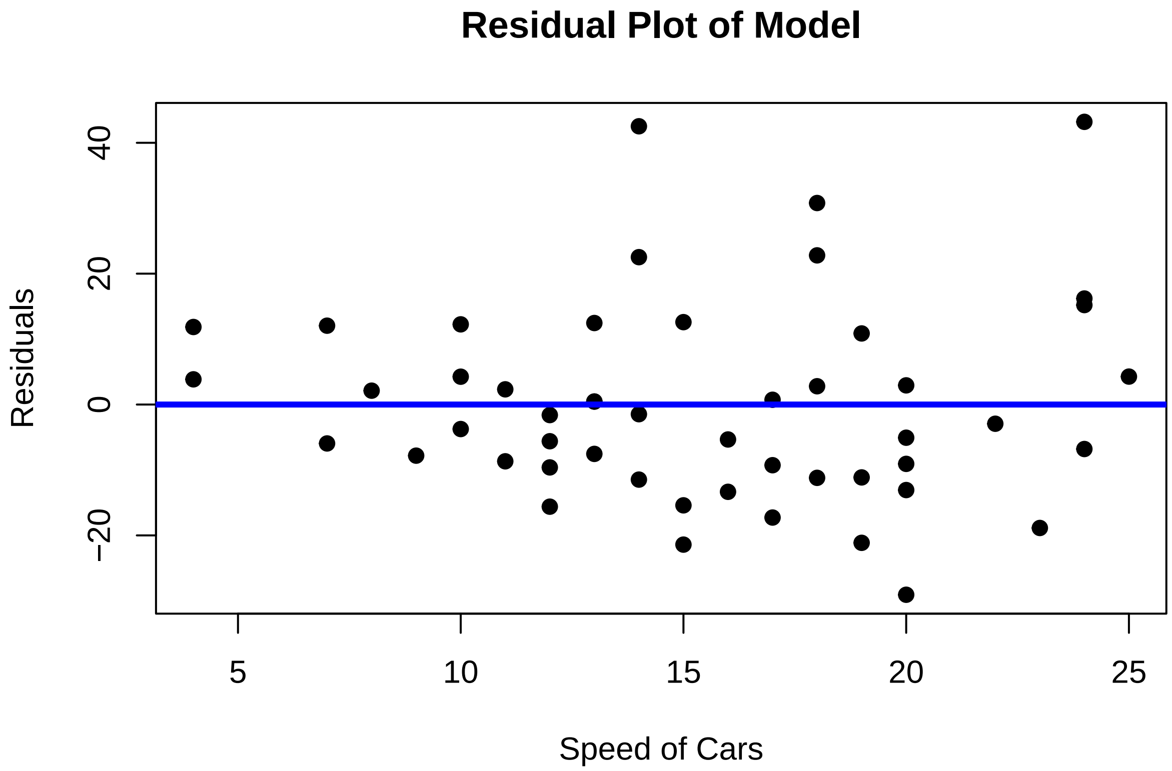
(a)

```
M_cars <- lm(dist ~ speed, data = cars)
plot(cars$speed, M_cars$residual, pch = 19, xlab = "Speed of Cars",
     ylab = "Residuals", main = "Residual Plot of Model")
```



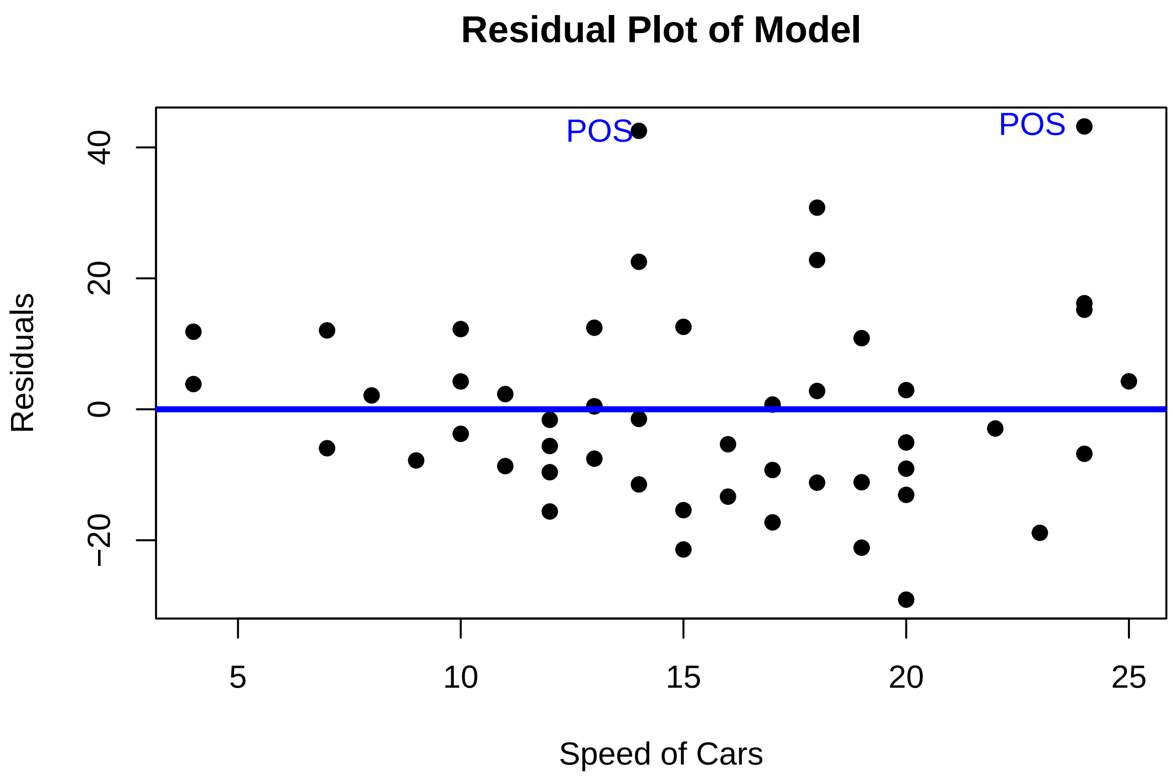
(b)

```
M_cars <- lm(dist ~ speed, data = cars)
plot(cars$speed, M_cars$residual, pch = 19, xlab = "Speed of Cars",
     ylab = "Residuals", main = "Residual Plot of Model")
abline(h = 0, lwd = 3, col = 'blue')
```



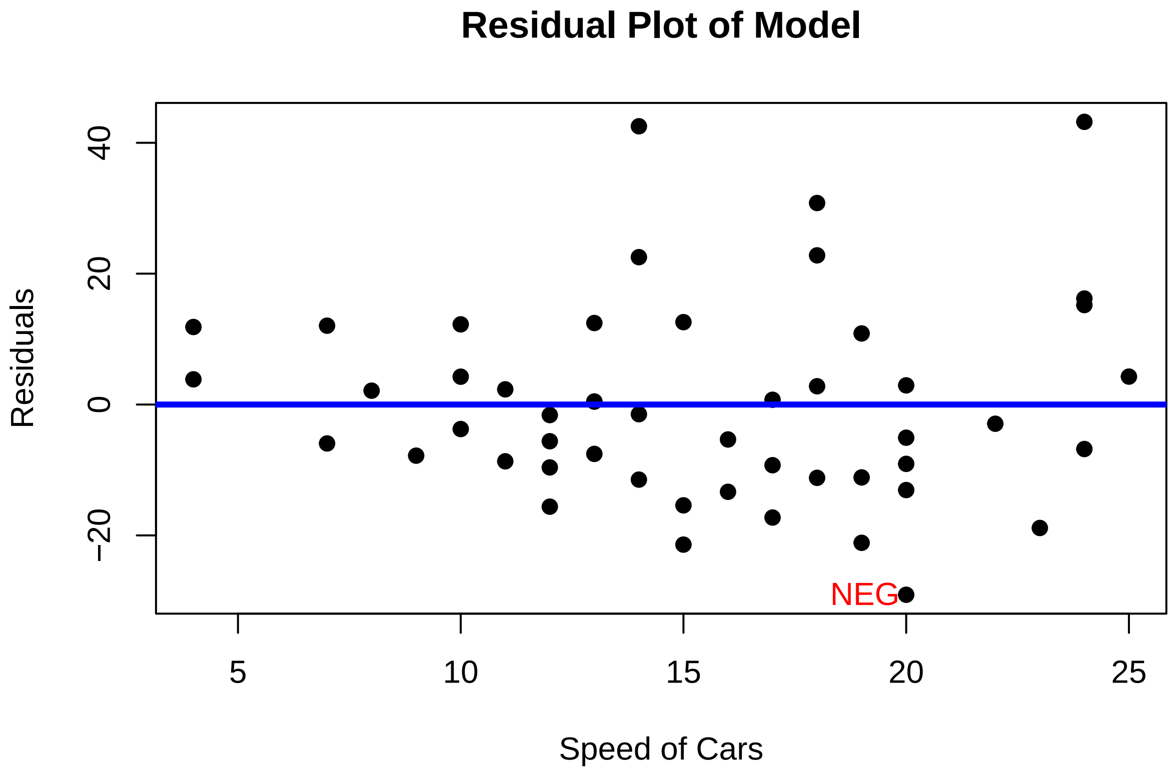
(c)

```
M_cars <- lm(dist ~ speed, data = cars)
plot(cars$speed, M_cars$residual, pch = 19, xlab = "Speed of Cars",
     ylab = "Residuals", main = "Residual Plot of Model")
abline(h = 0, lwd = 3, col = 'blue')
#locator(n = 2)
text(x = c(14.11889, 23.83116)-1, y = c(42.63116, 43.65511), c("POS", "POS"), col = 'blue')
```



(d)

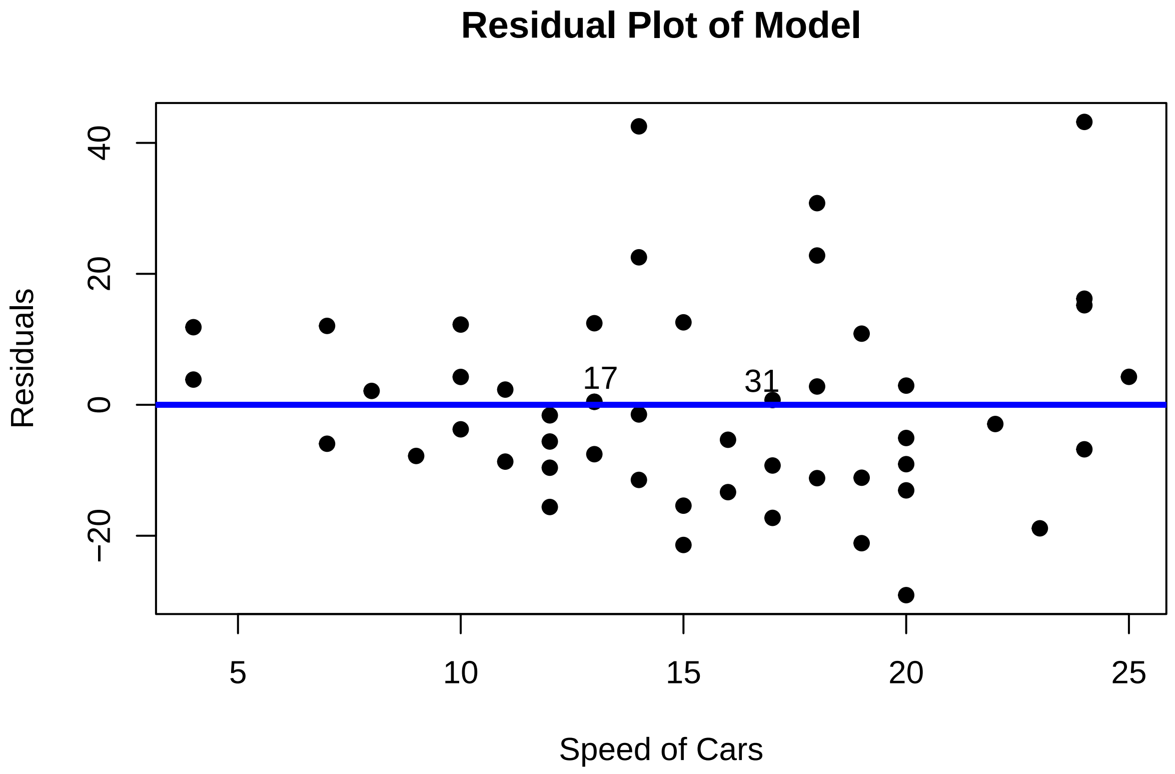
```
M_cars <- lm(dist ~ speed, data = cars)
plot(cars$speed, M_cars$residual, pch = 19, xlab = "Speed of Cars",
     ylab = "Residuals", main = "Residual Plot of Model")
abline(h = 0, lwd = 3, col = 'blue')
#locator(n = 1)
text(x = 20.57324-1.5, y = -28.8744, c("NEG"), col = 'red')
```



(e)

```
trans_cars <- cbind(1:nrow(cars),cars)
colnames(trans_cars) <- c("orders", colnames(cars))
attach(trans_cars)
```

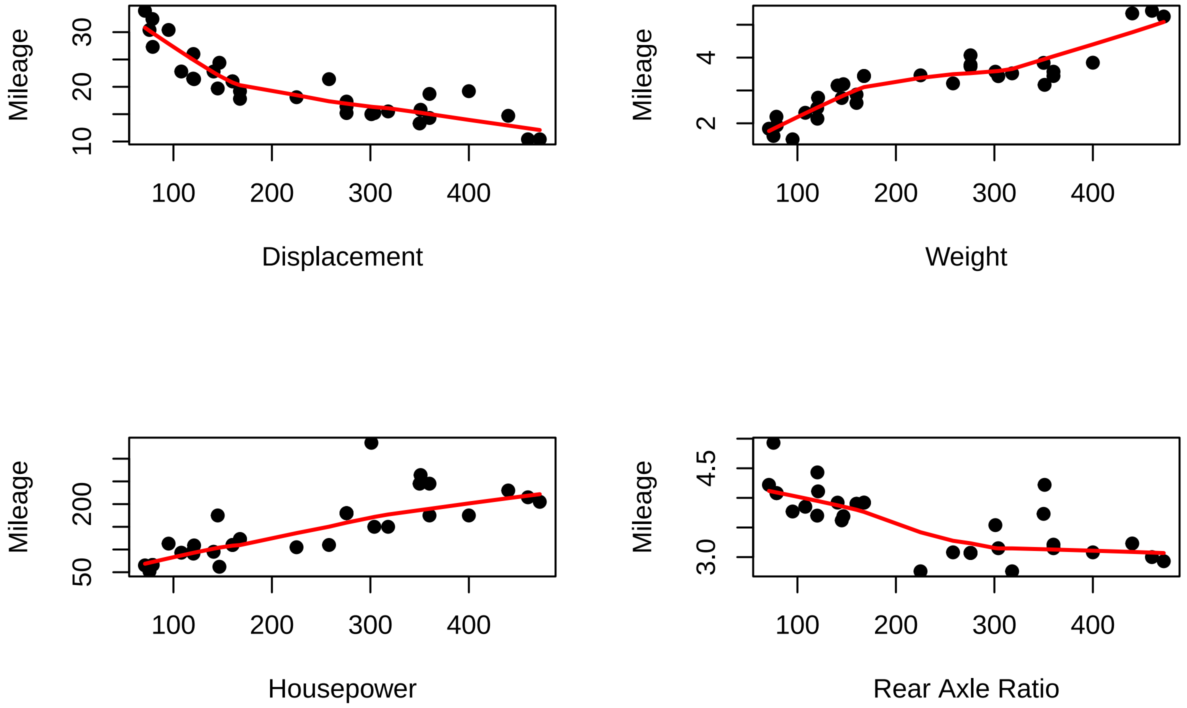
```
M_transcars <- lm(dist ~ speed, data = cars)
plot(cars$speed, M_transcars$residual, pch = 19, xlab = "Speed of Cars", ylab = "Residuals", main = "Res
abline(h = 0, lwd = 3, col = 'blue')
#identify(x=speed, y=M_transcars$residuals, labels = orders, n = 2)
#The row number is 17 and 31.
#locator(2)
text(x = c(13.13536, 16.76210), y = c(1.1613466, 0.6493736)+3, c(17, 31))
```



Question 2: (4.4)

```
attach(mtcars)
```

```
par(mfrow = c(2,2))
plot(x = disp, y = mpg, pch = 19, xlab = "Displacement", ylab = "Mileage", main = "")
lines(lowess(disp, mpg), lwd = 2, col = 'red')
plot(x = disp, y = wt, pch = 19, xlab = "Weight", ylab = "Mileage", main = "")
lines(lowess(disp, wt), lwd = 2, col = 'red')
plot(x = disp, y = hp, pch = 19, xlab = "Housepower", ylab = "Mileage", main = "")
lines(lowess(disp, hp), lwd = 2, col = 'red')
plot(x = disp, y = drat, pch = 19, xlab = "Rear Axle Ratio", ylab = "Mileage", main = "")
lines(lowess(disp, drat), lwd = 2, col = 'red')
```

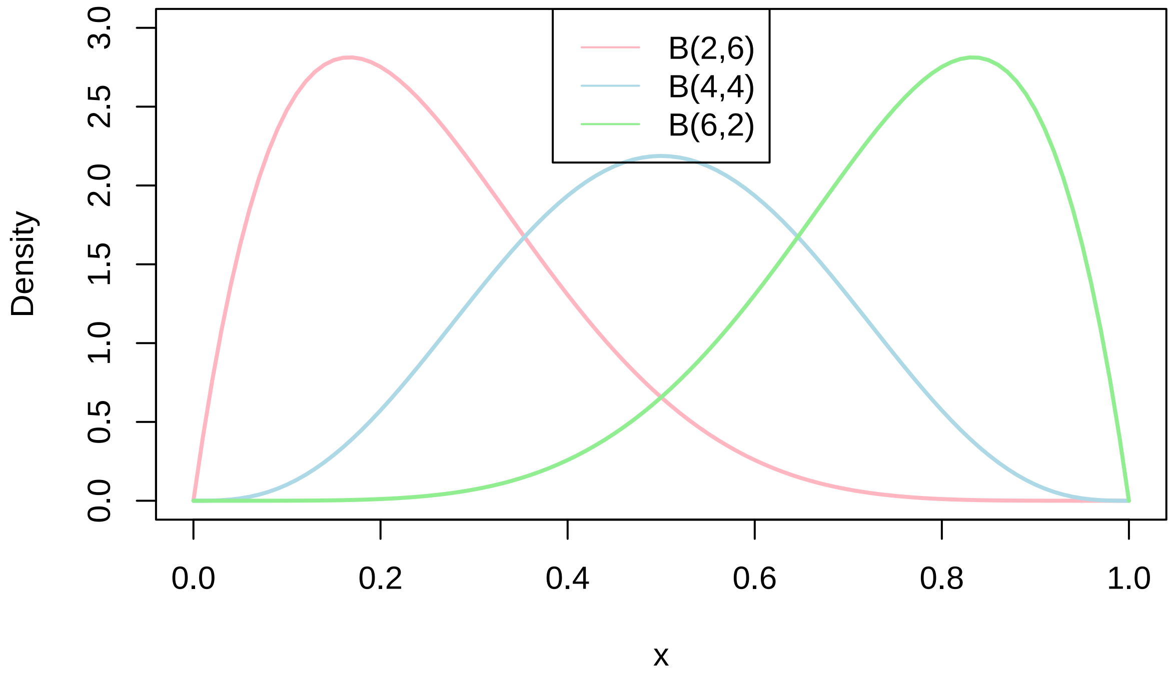


It seems that displacement and weight have stronger relationship than the other two variables. For these two, it is a little hard to distinguish, but intuitively from the graph, displacement have stronger relationship with mileage since the absolute value of slope is greater than the other one.

Question 2: (4.6)

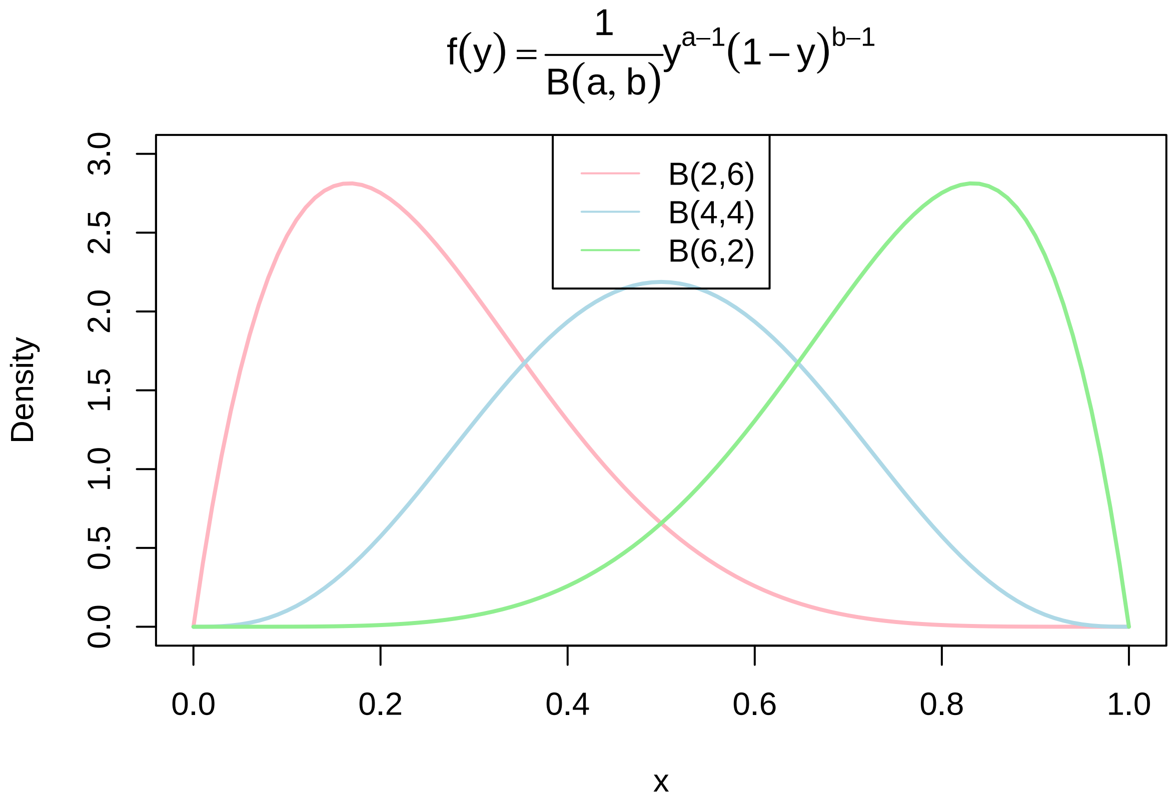
(a)

```
curve(dbeta(x, 2, 6), col = "lightpink", lwd = 2, ylab = "Density", ylim = c(0,3))
curve(dbeta(x, 4, 4), col = "lightblue", lwd = 2, add = TRUE)
curve(dbeta(x, 6, 2), col = "lightgreen", lwd = 2, add = TRUE)
legend("top", col = c("lightpink", "lightblue", "lightgreen"), lty = c(1,1,1), c("B(2,6)","B(4,4)","B(6
```



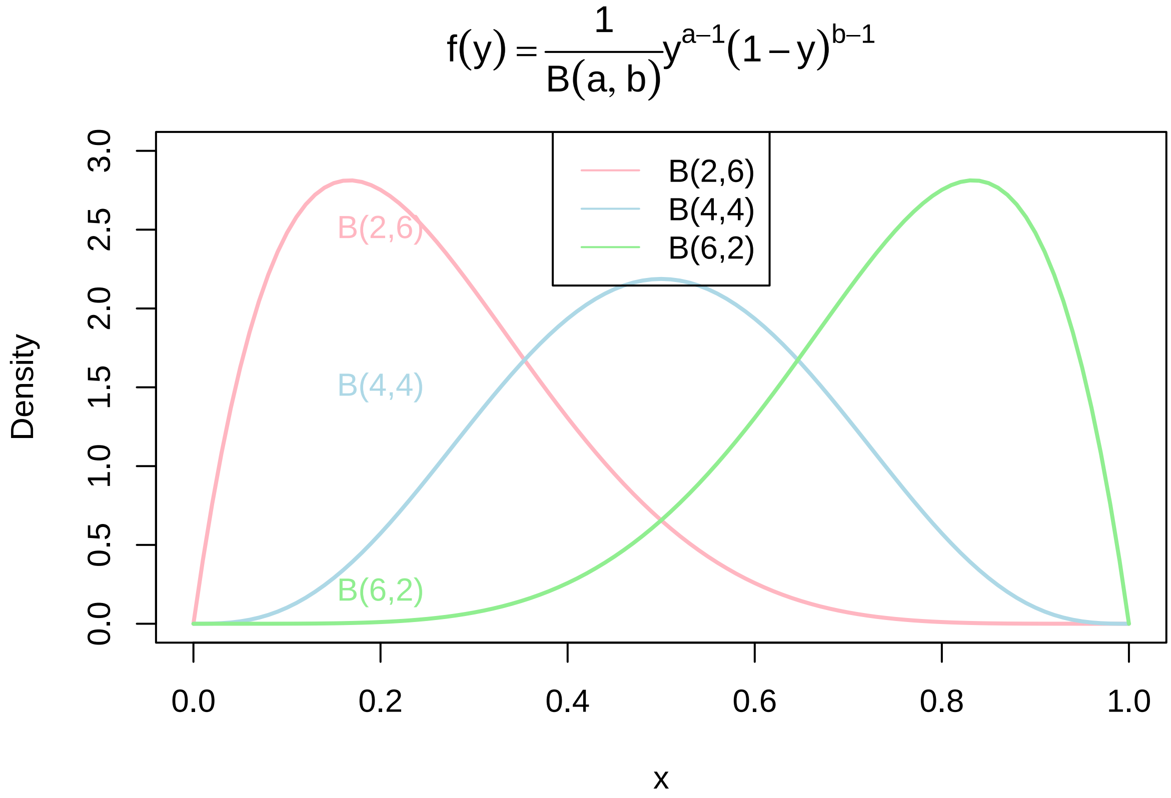
(b)

```
curve(dbeta(x, 2, 6), col = "lightpink", lwd = 2, ylab = "Density", ylim = c(0,3))
curve(dbeta(x, 4, 4), col = "lightblue", lwd = 2, add = TRUE)
curve(dbeta(x, 6, 2), col = "lightgreen", lwd = 2, add = TRUE)
legend("top", col = c("lightpink", "lightblue", "lightgreen"), lty = c(1,1,1), c("B(2,6)","B(4,4)","B(6
title(expression(f(y)==frac(1,B(a,b))*y^{a-1}*(1-y)^{b-1}))
```



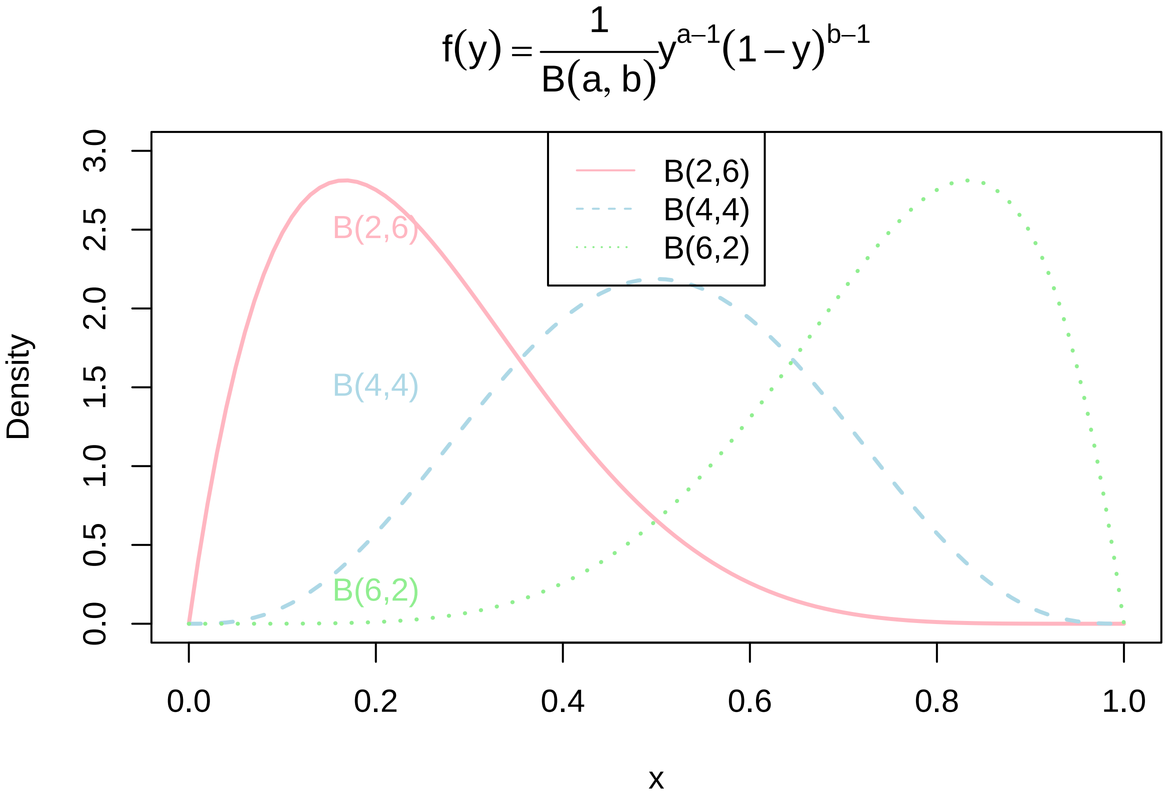
(c)

```
curve(dbeta(x, 2, 6), col = "lightpink", lwd = 2, ylab = "Density", ylim = c(0,3))
curve(dbeta(x, 4, 4), col = "lightblue", lwd = 2, add = TRUE)
curve(dbeta(x, 6, 2), col = "lightgreen", lwd = 2, add = TRUE)
legend("top", col = c("lightpink", "lightblue", "lightgreen"), lty = c(1,1,1), c("B(2,6)","B(4,4)","B(6
title(expression(f(y)==frac(1,B(a,b))*y^{a-1}*(1-y)^{b-1}))
text(c(0.2,0.2,0.2), y = c(0.2, 1.5, 2.5), col = c("lightgreen", "lightblue", "lightpink"), c("B(6,2)",
```



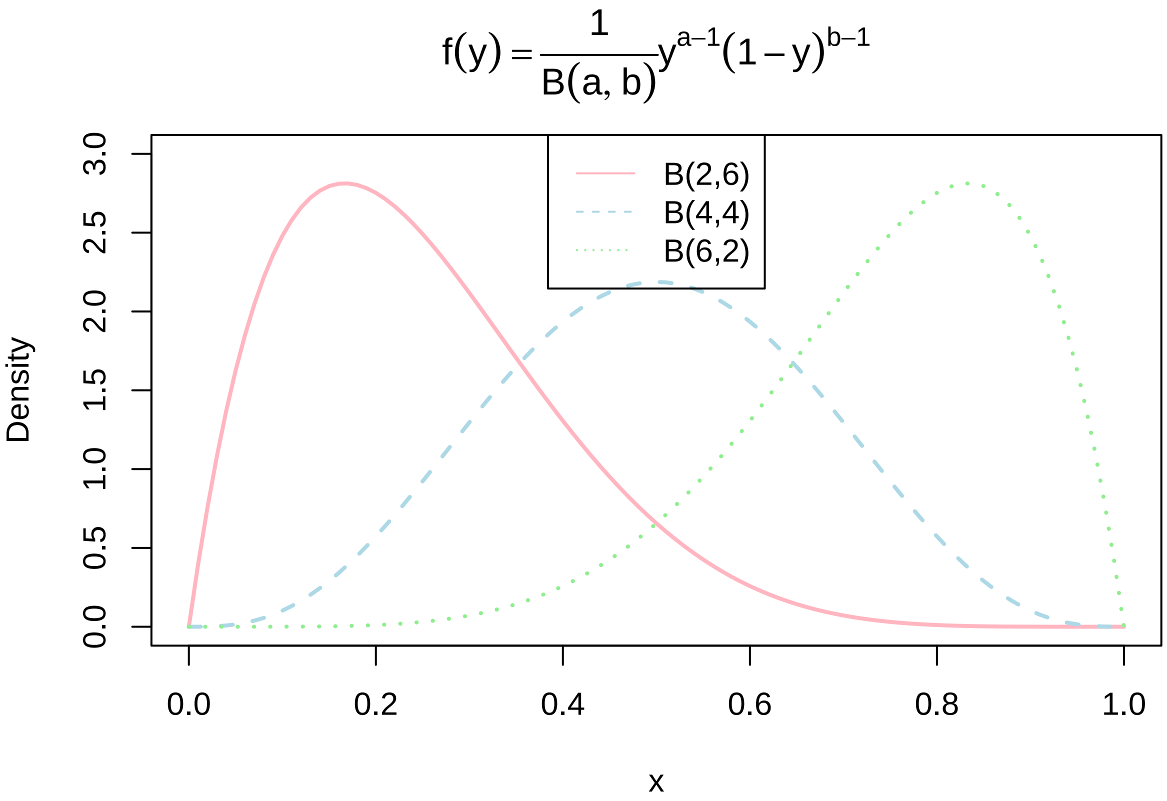
(d)

```
curve(dbeta(x, 2, 6), col = "lightpink", lwd = 2, ylab = "Density", ylim = c(0,3), lty = 1)
curve(dbeta(x, 4, 4), col = "lightblue", lwd = 2, add = TRUE, lty = 2)
curve(dbeta(x, 6, 2), col = "lightgreen", lwd = 2, add = TRUE, lty = 3)
legend("top", col = c("lightpink", "lightblue", "lightgreen"), lty = c(1,2,3), c("B(2,6)","B(4,4)","B(6
title(expression(f(y)==frac(1,B(a,b))*y^{a-1}*(1-y)^{b-1}))
text(c(0.2,0.2,0.2), y = c(0.2, 1.5, 2.5), col = c("lightgreen", "lightblue", "lightpink"), c("B(6,2)",
```



(e)

```
curve(dbeta(x, 2, 6), col = "lightpink", lwd = 2, ylab = "Density", ylim = c(0,3), lty = 1)
curve(dbeta(x, 4, 4), col = "lightblue", lwd = 2, add = TRUE, lty = 2)
curve(dbeta(x, 6, 2), col = "lightgreen", lwd = 2, add = TRUE, lty = 3)
legend("top", col = c("lightpink", "lightblue", "lightgreen"), lty = c(1,2,3), c("B(2,6)","B(4,4)","B(6
title(expression(f(y)==frac(1,B(a,b))*y^{a-1}*(1-y)^{b-1}))
```



Question 4: (6.3)

(a)

```
nyc <- read.table(here::here("nyc-marathon.txt"), sep = ",", header = TRUE)
nycf <- nyc[which(nyc$Gender == "female"),c(1,3)]
colnames(nycf) <- c("time.f", "age.f")
attach(nycf)

nycm <- nyc[which(nyc$Gender == "male"),c(1,3)]
colnames(nycm) <- c("time.m", "age.m")
attach(nycm)
```

```
var.test(age.m, age.f)
```

```
##
##  F test to compare two variances
##
## data:  age.m and age.f
## F = 1.4096, num df = 168, denom df = 106, p-value = 0.05602
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.9911202 1.9775481
## sample estimates:
## ratio of variances
##           1.409591
```

```
t.test(age.m, age.f, alternative = "greater", var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  age.m and age.f
## t = 2.3597, df = 274, p-value = 0.009495
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.8968103       Inf
## sample estimates:
## mean of x mean of y
##  44.54438  41.56075
```

For the variance test, we cannot reject the null hypothesis that the ration of the variance of men and women's age is 1 at 0.95 significance level, so I used the var.equal = TRUE in the t.test. The p-value is so small that we can reject the null hypothesis, and the mean of the age of all men is greater than the mean of all women at significant level 0.05.

(b)

First, I need to calculate the pooled variance, and then calculate the interval.

$$S_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{(n_1-1)+(n_2-1)}$$

In which, s_1 is the sample variance of men's age, s_2 is the sample variance of the women's age. n_1 is the number of men in the survey and n_2 is the number of women in the survey.

```
n1 <- length(age.m)
n2 <- length(age.f)
s1 <- var(age.m)
s2 <- var(age.f)
sp <- ( (n1-1)*s1 + (n2-1)*s2 ) / ( (n1-1) + (n2-1) )
```

And the confidence interval is:

$$\bar{A}ge_m - \bar{A}ge_f \ \pm \ t_{\frac{\alpha}{2}, n_1+n_2-2} \times \sqrt{\frac{S_p^2}{n_1} + \frac{S_p^2}{n_2}}$$

```
a <- qt(0.95, n1+n2-2) * sqrt(sp/n1 + sp/n2)
diff_mu <- mean(age.m) - mean(age.f)
lower <- diff_mu - a
upper <- diff_mu + a
intv <- matrix(c(lower, upper),1,2)
colnames(intv) <- c("Lower", "Upper")
print(intv)
```

```
##          Lower    Upper
## [1,] 0.8968103 5.070452
```

Question 5: (6.4)

(a)

```
dat_length <-  as.integer( scan(text = "22 18 27 23 24 15 26 22 24 25 24 18
18 26 20 24 27 16 30 22 17 18 22 26", what = "integer") )
```

(b)

```
t.test(dat_length, mu = 26)
```

```
##
##  One Sample t-test
##
## data:  dat_length
## t = -4.6148, df = 23, p-value = 0.0001216
## alternative hypothesis: true mean is not equal to 26
## 95 percent confidence interval:
##  20.569 23.931
## sample estimates:
## mean of x
##     22.25
```

According to the result of the one sample t test, the p-value is very small and we can reject the null hypothesis, and the mean of the population is not equal to 26 at significance level 0.05.

(c)

```
t.test(dat_length, mu = 26, conf.level = 0.9)
```

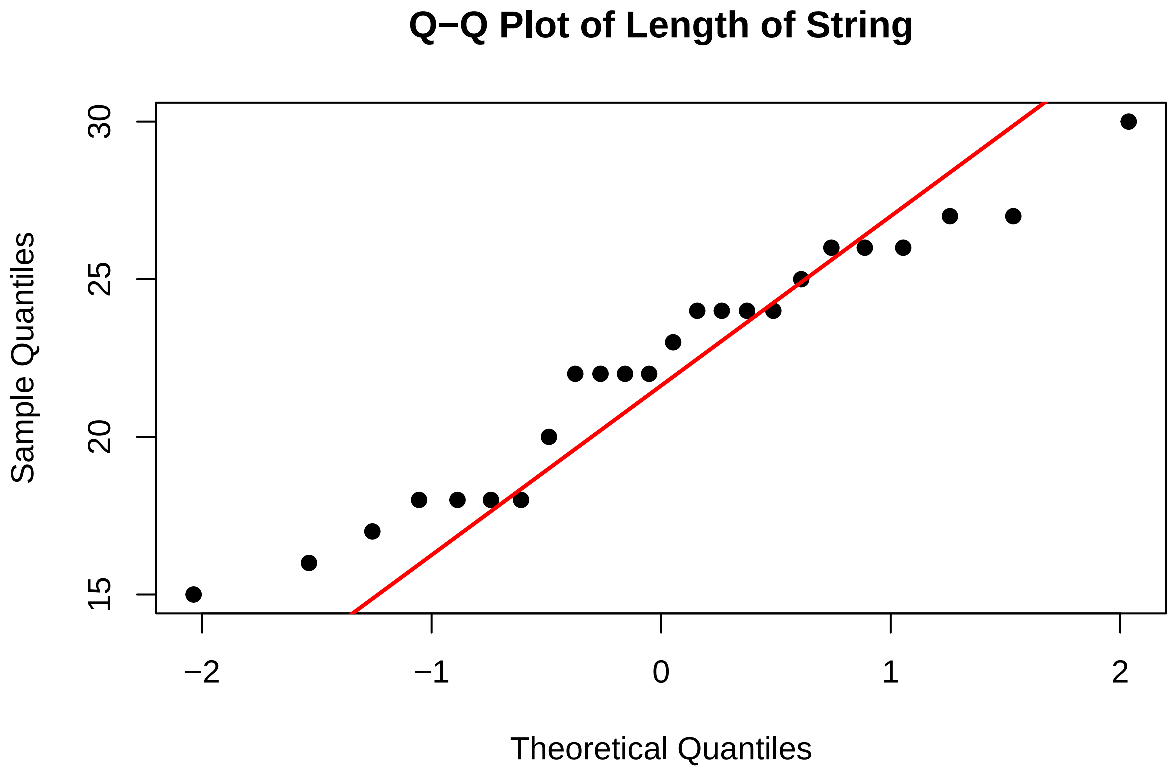
```
##
##  One Sample t-test
##
## data:  dat_length
## t = -4.6148, df = 23, p-value = 0.0001216
## alternative hypothesis: true mean is not equal to 26
## 90 percent confidence interval:
##  20.8573 23.6427
## sample estimates:
## mean of x
##     22.25
```

Here, the same and result is same as b, but we want the 95% CI for the population mu. From the result, the CI should be:

$$[\ 20.858, 23.642\ ]$$

(d)

```
qqnorm(dat_length, pch = 19, main = "Q-Q Plot of Length of String")
qqline(dat_length, col = 'red', lwd = 2)
```



It seems that the length of the string is not that normally distributed since there are many points far from the qqline, especially on the tail of the distribution.

Question 6: (6.6)

(a)

```
dat_it <- read.table(here::here("Etruscan-Italian.txt"))
etr <- dat_it[which(dat_it$group == "Etruscan"),]
ita <- dat_it[which(dat_it$group == "Italian"),]
attach(dat_it)
```

First, I want to carry out one variance test to check whether the variance of them are the same.

```
var.test(etr$x, ita$x)
```

```
##
##  F test to compare two variances
##
## data:  etr$x and ita$x
## F = 1.0782, num df = 83, denom df = 69, p-value = 0.7503
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.6807504 1.6912757
## sample estimates:
## ratio of variances
##            1.07819
```

The p-value is 0.75, so we cannot reject the null hypothesis that the variance of these two populations are the same. Then we will use the t test and the variance are equal.

```
t.test(etr$x, ita$x, var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  etr$x and ita$x
## t = 11.925, df = 152, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   9.45365 13.20825
## sample estimates:
## mean of x mean of y
##  143.7738  132.4429
```

Here, the p-value is very small. So null hypothesis is reject, which means that there are some difference in the skill size of the mean of the population of these two groups.

(b)

Still, from the result above, the 95% CI should be:

$$[\ 9.45365,\ 13.20825\ ]$$

Question 7: (6.7)

Since it is a paired t-test, we are going to first make a pairwise difference and save it as the difference. Then test whether the mean of the difference is 0.

```
winner = c(185, 182, 182, 188, 188, 188, 185, 185, 177, 182, 182, 193, 183, 179, 179, 175)
opponent = c(175, 193, 185, 187, 188, 173, 180, 177, 183,185, 180, 180, 182, 178, 178, 173)
dif <- winner - opponent
t.test(dif, mu = 0)
```

```
##
##  One Sample t-test
##
## data:  dif
## t = 1.3279, df = 15, p-value = 0.2041
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -1.361429  5.861429
## sample estimates:
## mean of x
##      2.25
```

Here the p-value is 0.2041, it is not so significant. We cannot reject the null hypothesis that there is no difference between the mean of election winner and loser.

Question 8: (7.5)

```r
library(gamair)
data("hubble")
M_hub <- lm(hubble$y ~ hubble$x - 1)
summary(M_hub)
```

```
##
## Call:
## lm(formula = hubble$y ~ hubble$x - 1)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -736.5 -132.5  -19.0  172.2  558.0
##
## Coefficients:
##          Estimate Std. Error t value Pr(>|t|)
## hubble$x   76.581      3.965   19.32 1.03e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 258.9 on 23 degrees of freedom
## Multiple R-squared:  0.9419, Adjusted R-squared:  0.9394
## F-statistic: 373.1 on 1 and 23 DF,  p-value: 1.032e-15
```

So the estimated Hubble constant is 76.581.

Question 9: (7.7)

```r
M1 <- lm(cars$dist ~ cars$speed)
M2 <- lm(cars$dist ~ cars$speed - 1 )
summary(M1)
```

```
##
## Call:
## lm(formula = cars$dist ~ cars$speed)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -29.069  -9.525  -2.272   9.215  43.201
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -17.5791     6.7584  -2.601   0.0123 *
## cars$speed    3.9324     0.4155   9.464 1.49e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.38 on 48 degrees of freedom
## Multiple R-squared:  0.6511, Adjusted R-squared:  0.6438
## F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12
```

```r
summary(M2)
```

```
##
## Call:
## lm(formula = cars$dist ~ cars$speed - 1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -26.183 -12.637  -5.455   4.590  50.181
##
## Coefficients:
##            Estimate Std. Error t value Pr(>|t|)
## cars$speed   2.9091     0.1414   20.58   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.26 on 49 degrees of freedom
## Multiple R-squared:  0.8963, Adjusted R-squared:  0.8942
## F-statistic: 423.5 on 1 and 49 DF,  p-value: < 2.2e-16
```

The R^2 of the model with intercept is 0.651 with adjusted R^2 is 0.644. For the model without the intercept, the R^2 is 0.896 and adjusted R^2 0.894. R square means how much of the variance of dependent variable can be explained by the independent variables in our model. It is obvious that the model without the intercept is better since it explains more information.

Question 10: (7.11)

```
twins <-  read.table(here::here("twins.txt"), header = TRUE, sep = ',', na.strings = '.')
```

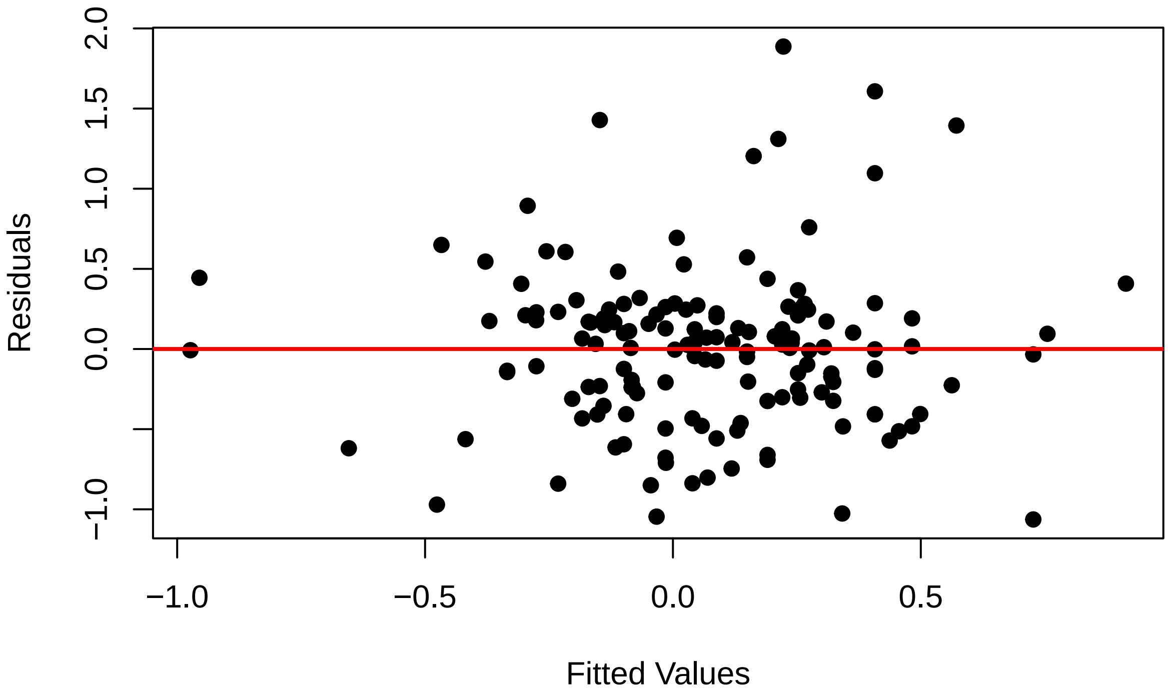
```
twins_new <- cbind(twins$DLHRWAGE, twins$HRWAGEL)
twins_new <- na.omit(twins_new)
colnames(twins_new) <- c("DLHRWAGE", "HRWAGEL")
twins_new <- as.data.frame(twins_new)


twins_new$DLHRWAGE <- as.numeric(twins_new$DLHRWAGE)
twins_new$LHRWAGEL <- log(as.numeric(twins_new$HRWAGEL))
M3 <- lm(twins_new$DLHRWAGE ~  twins_new$LHRWAGEL)
summary(M3)
```

```
##
## Call:
## lm(formula = twins_new$DLHRWAGE ~ twins_new$LHRWAGEL)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.06338 -0.30402  0.01665  0.22892  1.88689
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         1.14922    0.15887   7.234 2.40e-11 ***
## twins_new$LHRWAGEL -0.46090    0.06545  -7.042 6.75e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

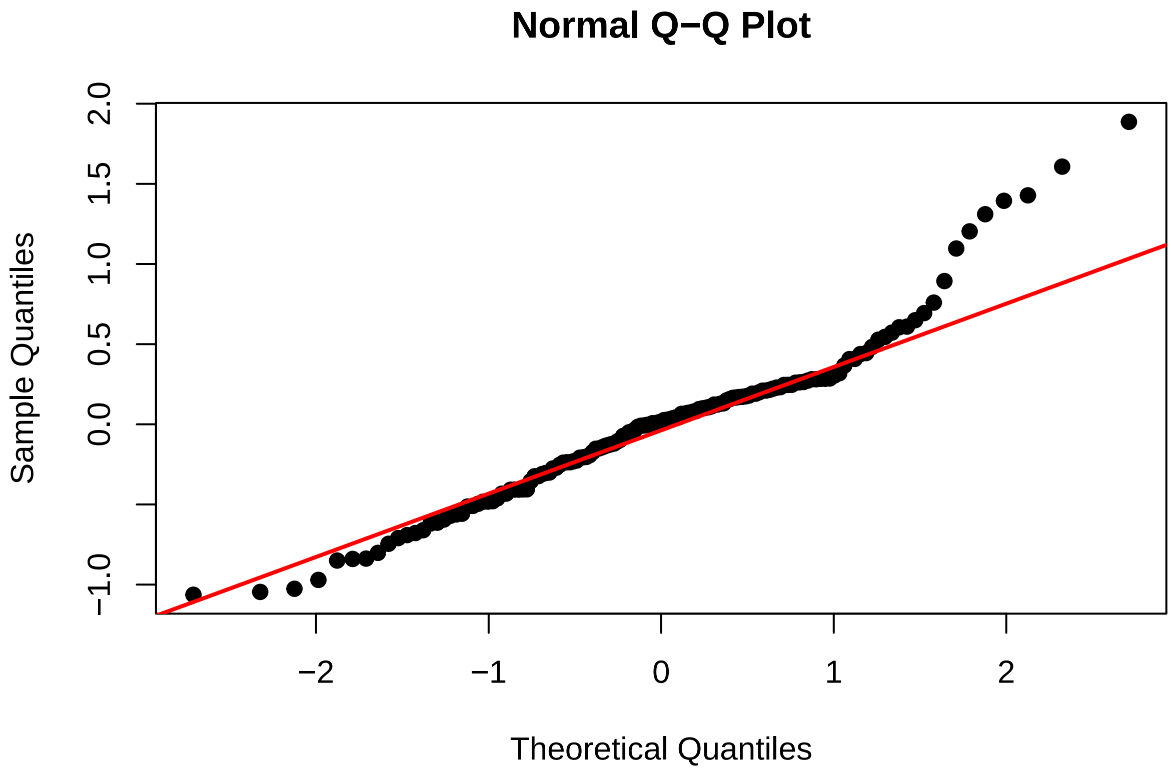
```
## Residual standard error: 0.5029 on 147 degrees of freedom
## Multiple R-squared:  0.2523, Adjusted R-squared:  0.2472
## F-statistic: 49.59 on 1 and 147 DF,  p-value: 6.751e-11
```

Here is the residual plot of the model:

```
plot(M3$fitted.values, M3$residuals, pch = 19, xlab = "Fitted Values", ylab = "Residuals")
abline(h = 0, col = 'red', lwd = 2)
```



```
qqnorm(M3$residuals, pch = 19)
qqline(M3$residuals, lwd = 2, col = 'red')
```



The right of the plot is not that normally distributed. So I think the assumption is violated. And the assumption of constant variance is still okay for this model, but still seems to have lower variance when the fitted values are bigger. And the prediction can be expressed by:

$$\hat{Log}_{DIFF} = 1.149 - 0.461 \times log(HRWAGEL)$$

Question 11: James(3.1) The null hypothesis is the true coefficient is 0, ant the alternative hypothesis is the true coefficient is not 0. From the table, I can see that, the p-value of intercept, TV and radio is very small, but the p-value of newspaper is very large. p-value measures when null hypothesis is true, the probability we observe a more extreme case than the one we observed now. Therefore, TV and radio has a significant impact on the sales, which means that the coefficient is significantly not 0 at significance level 0.05. However, newspaper does not have a significant impact on the sale, which means that we cannot reject the null hypothesis that the coefficient of newspaper is 0.

Question 12: James(3.3)

(a)

The true answer should be iii. Since the $\hat{\beta}_3$ is positive but $\hat{\beta}_5$ is negative, which means that if there is no interactions, and other variables remain the same, mean of the salary of women will be greater than men's. However, here is a interaction between GPA and gender and the coefficient is negative, which means that if IQ and GPA are the same, and the GPA is higher than 3.5, then males on average earn more than females.

(b)

$$\hat{Y} = 50 + 20 \times 4.0 + 0.07 \times 110 + 35 + 0.01 \times 110 \times 4.0 - 10 \times 4 = 137.1$$

(c)

It is not correct. The evidence is provided by the p-value but not the value of coefficient. If the p-value of the coefficient is smaller than a given significance level, there should be evidence that the interaction exist. We cannot judge it only from the coefficient, we should also consider the significance.

```
50 +80+7.7+4.4-40+35
```

```
## [1] 137.1
```

Question 13: James(3.8)

```
library(ISLR)
attach(Auto)
```

```
## The following object is masked from mtcars:
##
##     mpg
```

(a)

```
M_auto <- lm(mpg ~ horsepower, data = Auto)
summary(M_auto)
```

```
##
## 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
```

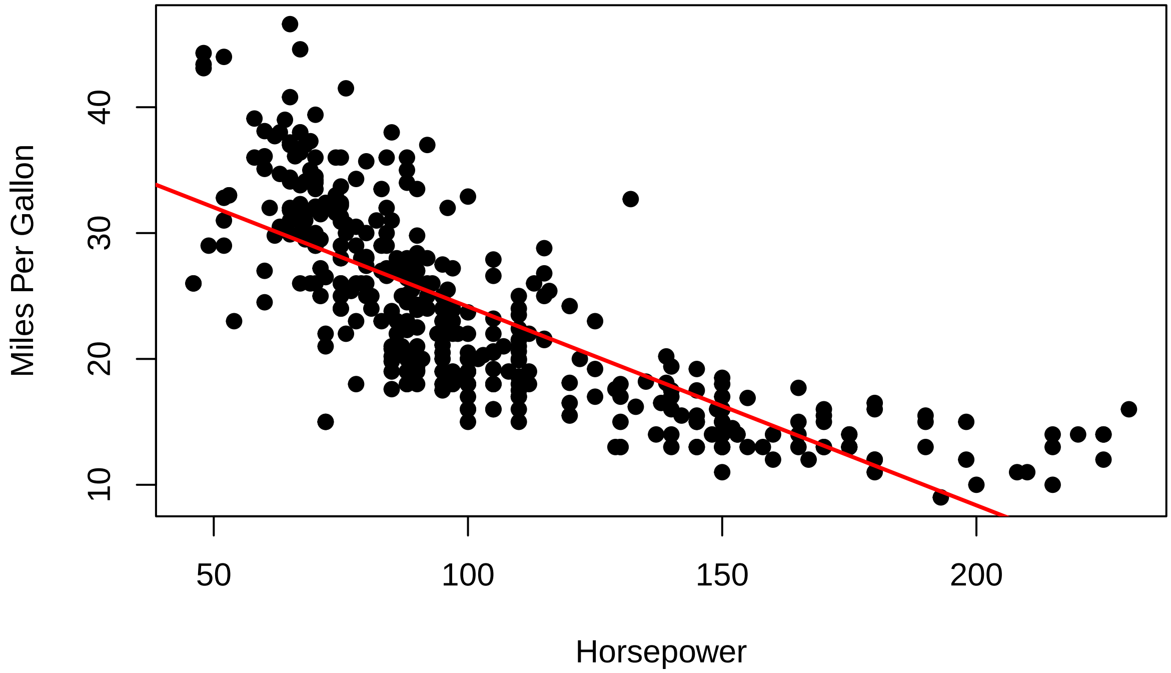
There is significant relationship between horsepower and miles per gallon. And the relationship is negative, the relationship is strong since the p-value is small enough and the coefficient is significant. The 95% CI for the one with horsepower of 98 should be:

```
pre <- data.frame(horsepower = 98)
PI <- predict(M_auto, pre, interval = "prediction")
CI <- predict(M_auto, pre, interval = "confidence")
res <- rbind(CI, PI)
rownames(res) <- c("Confidence Interval", "Prediction Interval")
colnames(res) <- c("Predict", "Lower", "Upper")
print(res)
```

```
##                         Predict    Lower    Upper
## Confidence Interval 24.46708 23.97308 24.96108
## Prediction Interval 24.46708 14.80940 34.12476
```

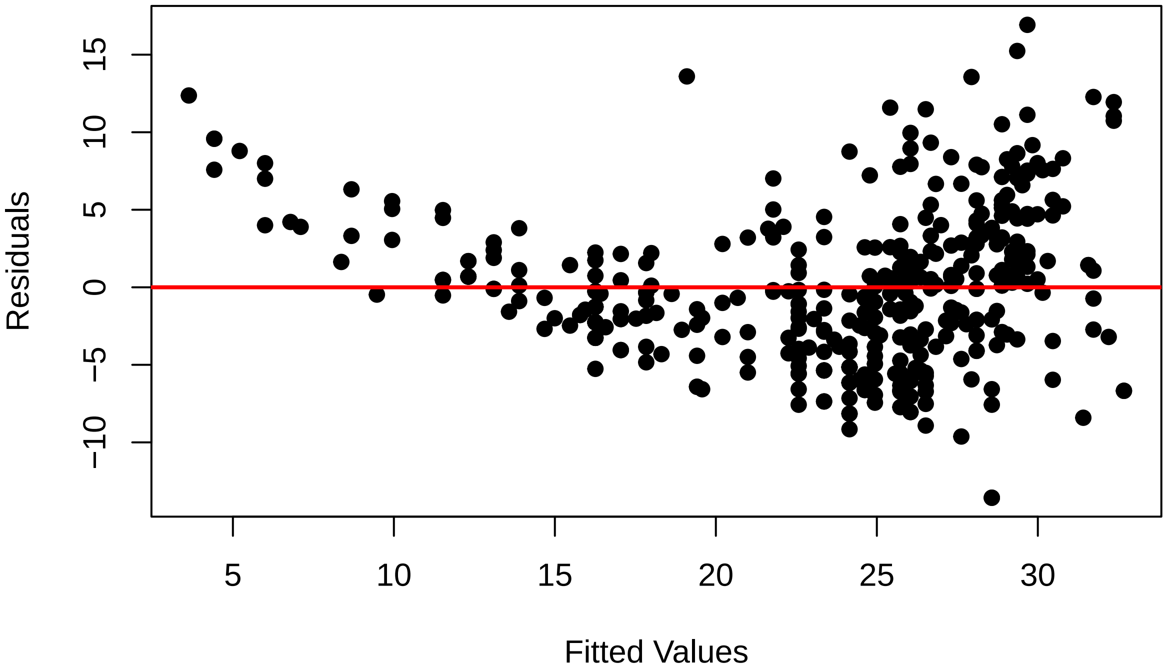
(b)

```
plot(x = horsepower, y = mpg, pch = 19, xlab = "Horsepower", ylab = "Miles Per Gallon", main = "")
abline(M_auto$coefficients, col = 'red', lwd = 2)
```



(c)

```
plot(M_auto$fitted.values, M_auto$residuals, xlab = "Fitted Values", ylab = "Residuals", pch = 19)
abline(h = 0, lwd = 2, col = 'red')
```

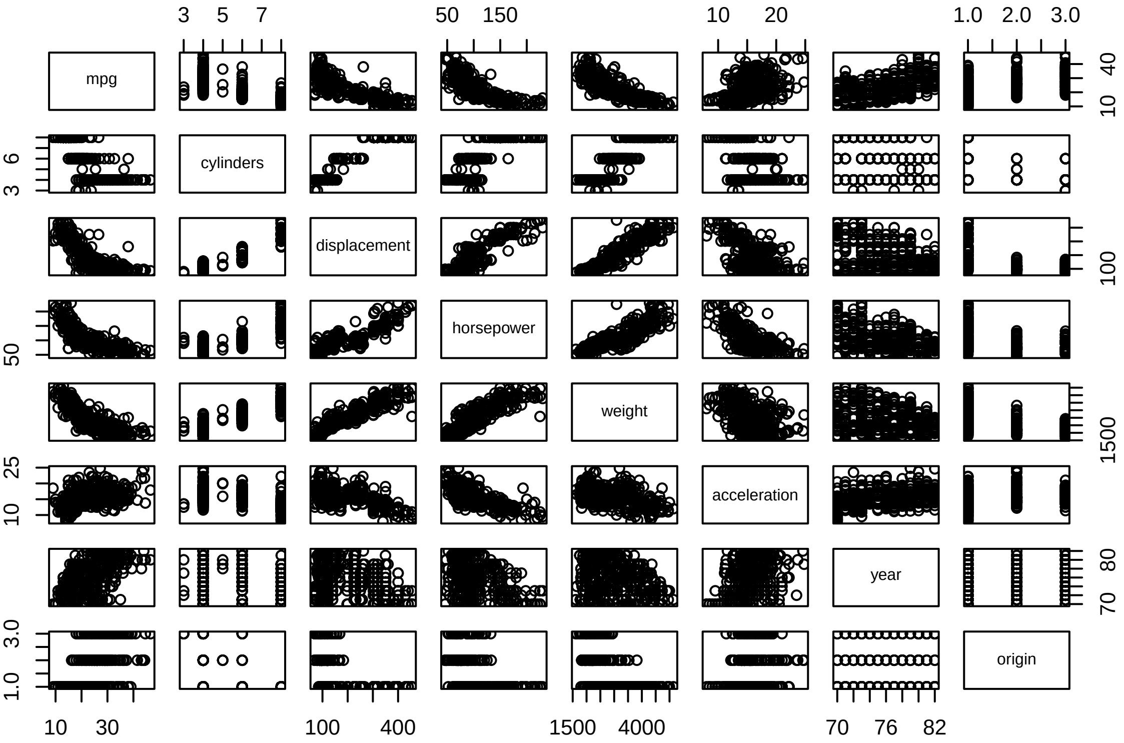


The residuals don't have constant variance, on the contrary, they have a form of quadratic, and also does not seem to be normal. Therefore, to improve this model, we can add one quadratic form of horsepower to this model.

Question 14: James(3.9)

(a)

```
Auto_new <- Auto[,1:8]
pairs(Auto_new)
```



(b)

```
cor(Auto_new)
```

```
##                     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)

```
M_mauto <- lm(mpg ~ ., data = Auto_new)
summary(M_mauto)
```

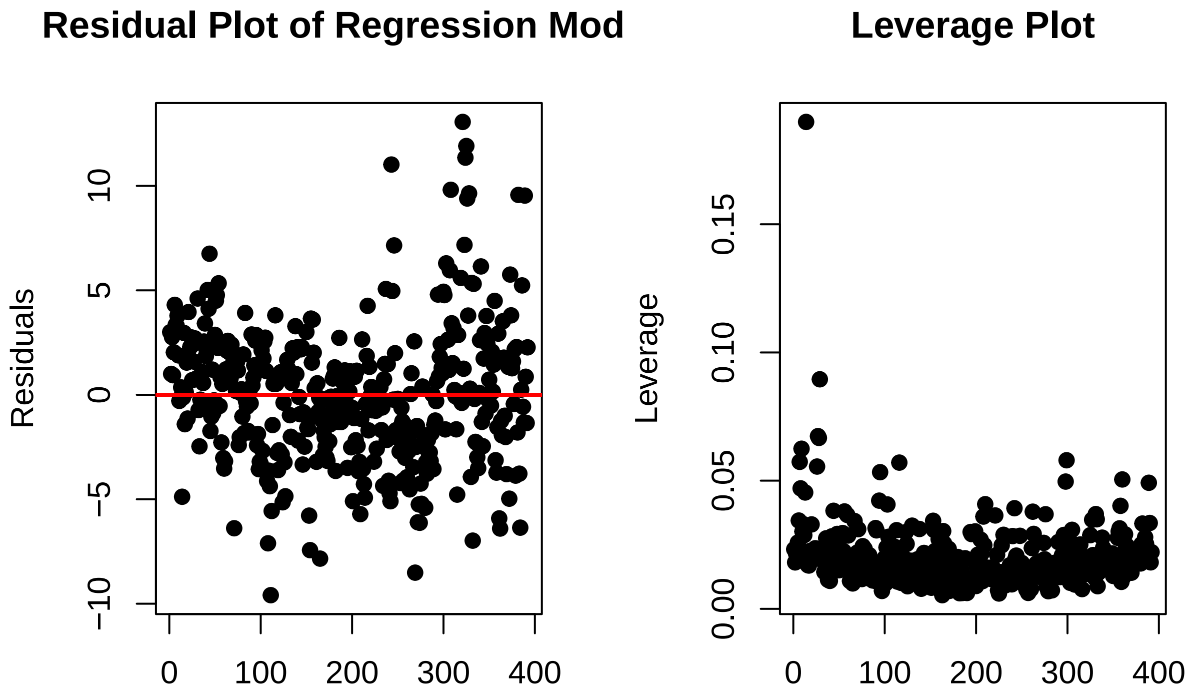
```
##
## Call:
## lm(formula = mpg ~ ., data = Auto_new)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.5903 -2.1565 -0.1169  1.8690 13.0604
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -17.218435   4.644294  -3.707  0.00024 ***
## cylinders     -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
```

After checking the F-statistic and p-value of the F test, we know that there is some relationship between covariates and response. Displacement, weight, year and origin (and intercept) are significant. The coefficient of year means, the expectation of miles per gallon will increase around 0.75 if the year increases one, with all the other variables remain the same.

(d)

I will use leverage plot and residual plot to check whether there are any outliers and see high leverage points.

```
par(mfrow = c(1,2))
plot(M_mauto$residuals, xlab = "", ylab = "Residuals", main = "Residual Plot of Regression Model", pch
abline(h = 0, col = 'red', lwd = 2)
plot(hatvalues(M_mauto), pch = 19, ylab = "Leverage", xlab = "", main = "Leverage Plot" )
```



From the residual plot, I don't think the residual plot is normal since there seems to be a quadratic trend for the residuals. But no outliers for the residuals since I don't think there are some points that are very far from others. However, when it comes to the leverage plot, I think there are some x which has extremely high leverage value since there is one beyond 0.15 and one around 0.1, which is far larger than the others.

(e)

```
M_mauto_int <- lm(mpg ~ .^2, data = Auto_new)
summary(M_mauto_int)
```

```
##
## Call:
## lm(formula = mpg ~ .^2, data = Auto_new)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.6303 -1.4481  0.0596  1.2739 11.1386
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                3.548e+01  5.314e+01   0.668  0.50475
## cylinders                  6.989e+00  8.248e+00   0.847  0.39738
## displacement              -4.785e-01  1.894e-01  -2.527  0.01192 *
## horsepower                 5.034e-01  3.470e-01   1.451  0.14769
## weight                     4.133e-03  1.759e-02   0.235  0.81442
## acceleration              -5.859e+00  2.174e+00  -2.696  0.00735 **
## year                       6.974e-01  6.097e-01   1.144  0.25340
## origin                    -2.090e+01  7.097e+00  -2.944  0.00345 **
## cylinders:displacement    -3.383e-03  6.455e-03  -0.524  0.60051
## cylinders:horsepower       1.161e-02  2.420e-02   0.480  0.63157
## cylinders:weight           3.575e-04  8.955e-04   0.399  0.69000
## cylinders:acceleration     2.779e-01  1.664e-01   1.670  0.09584 .
## cylinders:year            -1.741e-01  9.714e-02  -1.793  0.07389 .
## cylinders:origin           4.022e-01  4.926e-01   0.816  0.41482
## displacement:horsepower   -8.491e-05  2.885e-04  -0.294  0.76867
## displacement:weight        2.472e-05  1.470e-05   1.682  0.09342 .
## displacement:acceleration -3.479e-03  3.342e-03  -1.041  0.29853
## displacement:year          5.934e-03  2.391e-03   2.482  0.01352 *
## displacement:origin        2.398e-02  1.947e-02   1.232  0.21875
## horsepower:weight         -1.968e-05  2.924e-05  -0.673  0.50124
## horsepower:acceleration   -7.213e-03  3.719e-03  -1.939  0.05325 .
## horsepower:year           -5.838e-03  3.938e-03  -1.482  0.13916
## horsepower:origin          2.233e-03  2.930e-02   0.076  0.93931
## weight:acceleration        2.346e-04  2.289e-04   1.025  0.30596
## weight:year               -2.245e-04  2.127e-04  -1.056  0.29182
## weight:origin             -5.789e-04  1.591e-03  -0.364  0.71623
## acceleration:year          5.562e-02  2.558e-02   2.174  0.03033 *
## acceleration:origin        4.583e-01  1.567e-01   2.926  0.00365 **
## year:origin                1.393e-01  7.399e-02   1.882  0.06062 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.695 on 363 degrees of freedom
## Multiple R-squared:  0.8893, Adjusted R-squared:  0.8808
## F-statistic: 104.2 on 28 and 363 DF,  p-value: < 2.2e-16
```

We can see that the interaction between displacement and year, acceleration and year, and acceleration and origin are significant at significance level 0.05.

(f)

I would like to use variable displacement, year, weight and origin for base term, and an interaction term including displacement and year, also with all quadratic form of the base variable since it seems that there are quadratic residual forms in the plot there, then I will re-select variables based on this huge model.

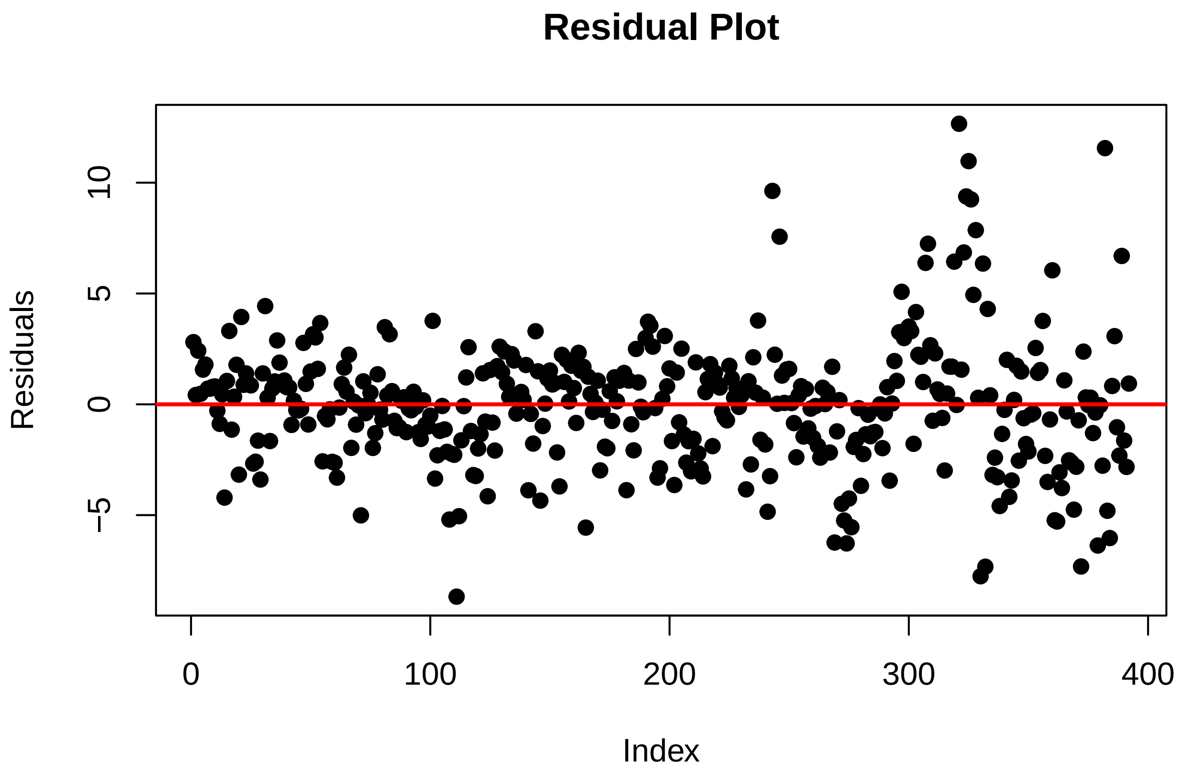
```
M_mauto_trans_1 <- lm(mpg ~ displacement + year + weight + origin + displacement*year + I(displacement^
summary(M_mauto_trans_1)
```

```
##
## Call:
## lm(formula = mpg ~ displacement + year + weight + origin + displacement *
##     year + I(displacement^2) + I(year^2) + I(weight^2) + I(origin^2),
##     data = Auto_new)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.6792 -1.6436  0.0453  1.4188 12.6585
##
## Coefficients:
```

```
##                       Estimate Std. Error t value Pr(>|t|)    
## (Intercept)          2.654e+02  8.052e+01   3.297 0.001071 ** 
## displacement         1.015e-01  4.526e-02   2.242 0.025553 *  
## year                -6.606e+00  2.080e+00  -3.176 0.001615 ** 
## weight              -1.889e-02  2.197e-03  -8.599  < 2e-16 ***
## origin               5.262e+00  1.840e+00   2.860 0.004473 ** 
## I(displacement^2)   -3.456e-06  2.737e-05  -0.126 0.899578    
## I(year^2)            5.054e-02  1.343e-02   3.764 0.000193 ***
## I(weight^2)          2.019e-06  3.130e-07   6.450 3.38e-10 ***
## I(origin^2)         -1.202e+00  4.505e-01  -2.667 0.007974 ** 
## displacement:year   -1.346e-03  5.180e-04  -2.599 0.009714 ** 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 2.881 on 382 degrees of freedom
## Multiple R-squared:  0.8669, Adjusted R-squared:  0.8637 
## F-statistic: 276.4 on 9 and 382 DF,  p-value: < 2.2e-16
```

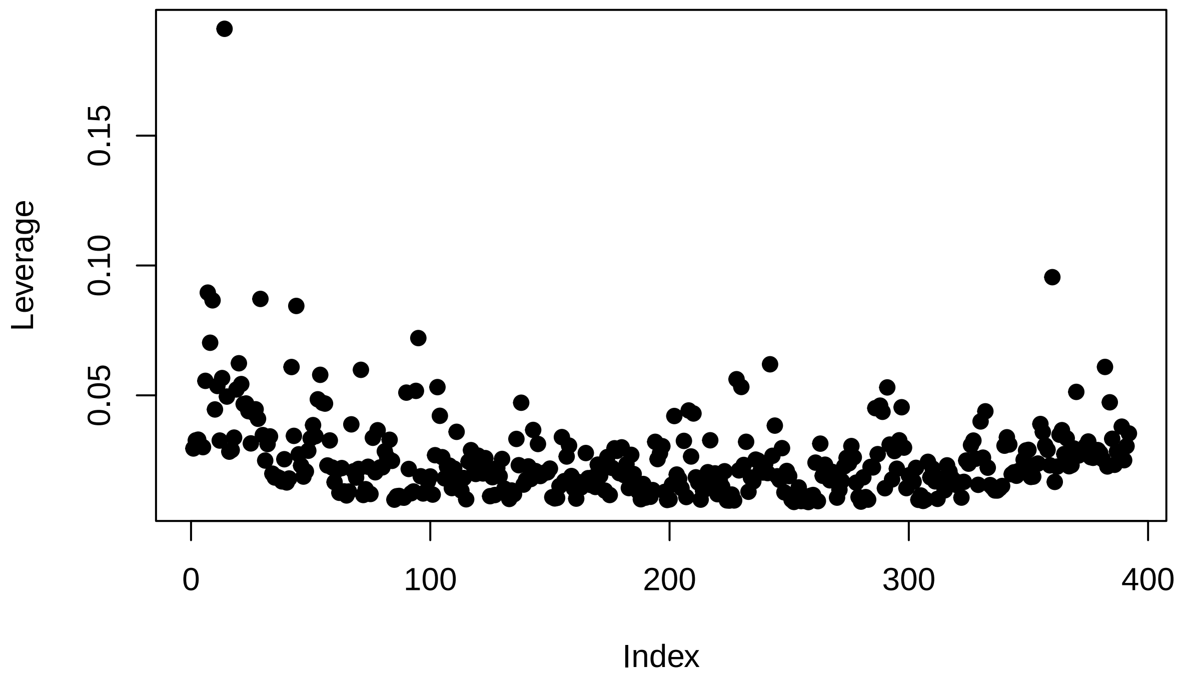
There seems to be quadratic relationship among all of them except variable displacement since all the quadratic form seems to be significant except the displacement, here is the residual plot:

```
plot(M_mauto_trans_1$residuals, ylab = "Residuals", main = "Residual Plot", pch = 19)
abline(h = 0, col = 'red', lwd = 2)
```



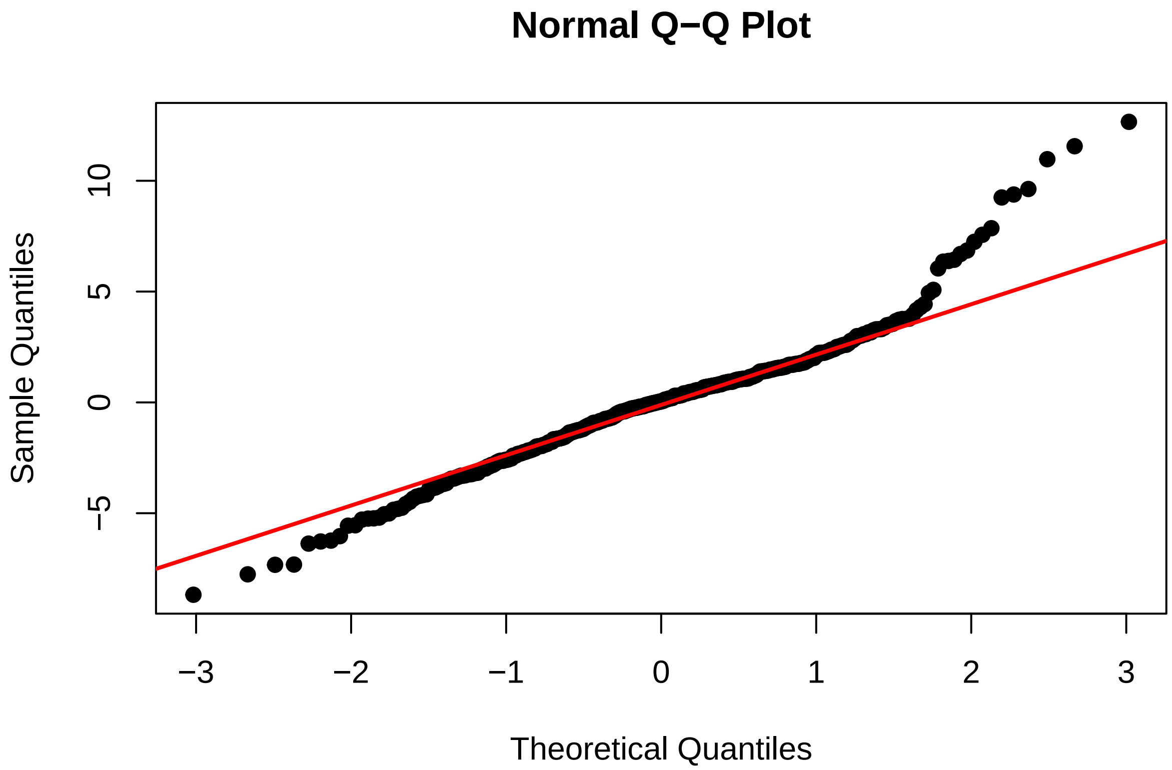
The residual plot seems better but there seems to be a polynomial form but not that obvious, but the adjusted R^2 seems to be greater.

```r
plot(hatvalues(M_mauto_trans_1), ylab = "Leverage", pch = 19)
```



There is still high leverage x existing.

```r
qqnorm(M_mauto_trans_1$residuals, pch = 19)
qqline(M_mauto_trans_1$residuals, lwd = 2, col = 'red')
```



However, here is the problem, the residuals are not so normally distributed, we should try to change a model or do some further transformations.

Question 15: James(3.10)

(a)

```
library(ISLR)
M_seat <- lm(Sales ~ Price + Urban + US, data = Carseats)
summary(M_seat)
```

```
##
## 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
```

(b)

The unit sales (in thousands) will decrease 0.054 on average if the price goes up for 1 unit, with all the other variables remain the same. With the other variables remaining the same, stores not in urban will have 0.022 more unit sales than those in urban. With the other variables remaining the same, stores in US will have 1.2 more unit sales than those in not in US.

(c)

$$Y_i = \beta_0 + X_{i,price}\beta_{price} + X_{Urban}\beta_{Urban} + X_{US}\beta_{US} + \epsilon_i$$

Here, X_{Urban} is an indicator that whether this store locates in urban, if this store is in urban, it is 1, or it will be 0 for this store. X_{US} is also an indicator that whether this store locates in US, if it locates in US, then it is 1, otherwise 0.

(d) The predictors that are significant are the intercept, price and whether the store locates in US.

(e)

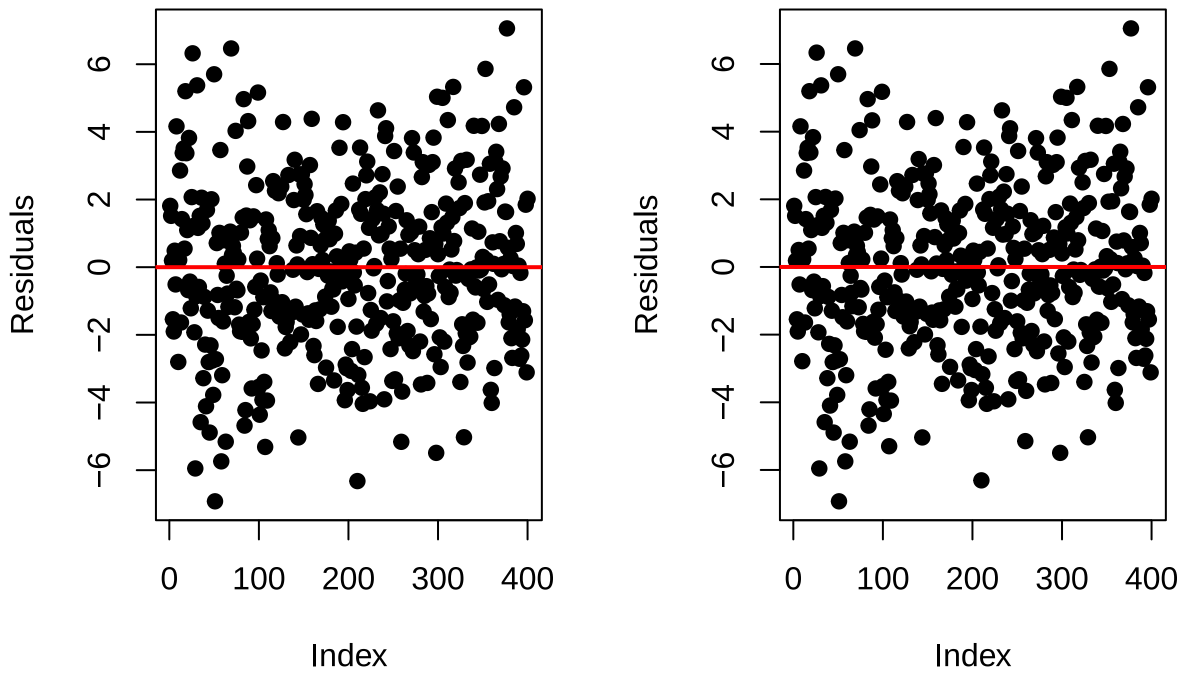
```r
M_seat_small <- lm(Sales ~ Price + US, data = Carseats)
summary(M_seat_small)
```

```
##
## 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
```

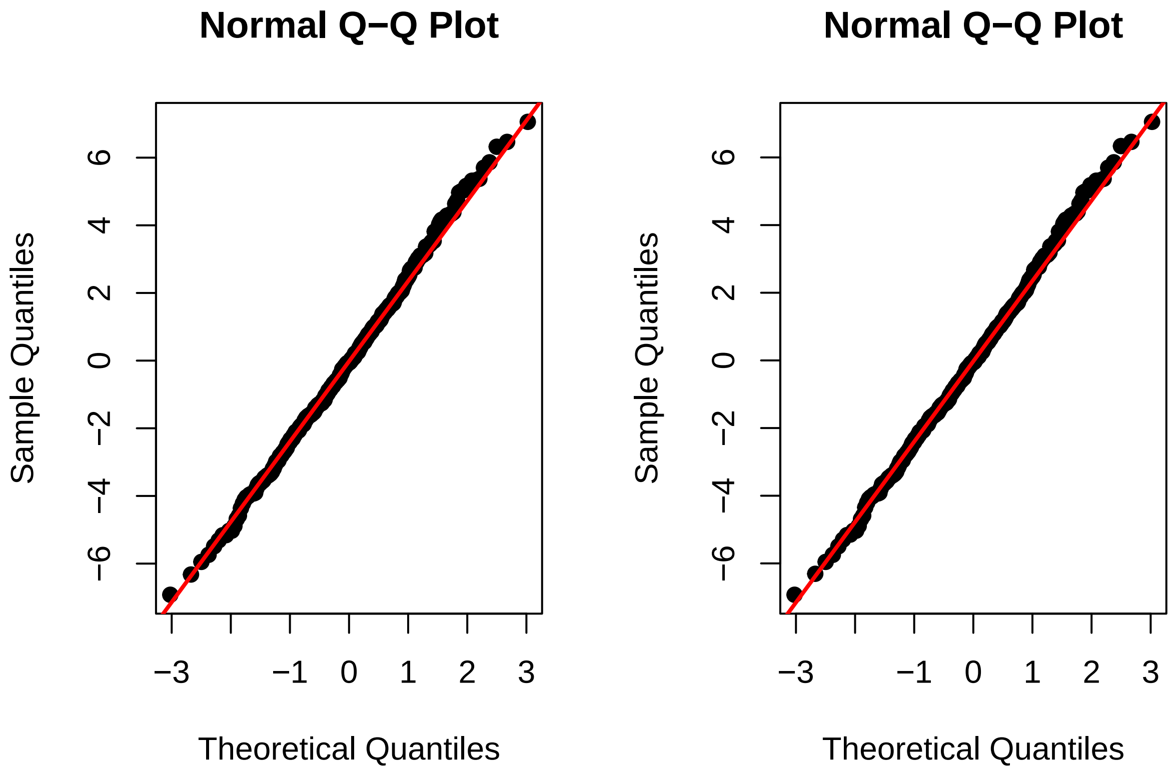
(f)

First, I will do the residual analysis then I will use AIC, BIC and R squared to make judgement and selection between them.

```
par(mfrow = c(1,2))
plot(M_seat$residuals, pch = 19, ylab = "Residuals")
abline(h = 0, col = 'red', lwd = 2)
plot(M_seat_small$residuals, pch = 19, ylab = "Residuals")
abline(h = 0, col = 'red', lwd = 2)
```

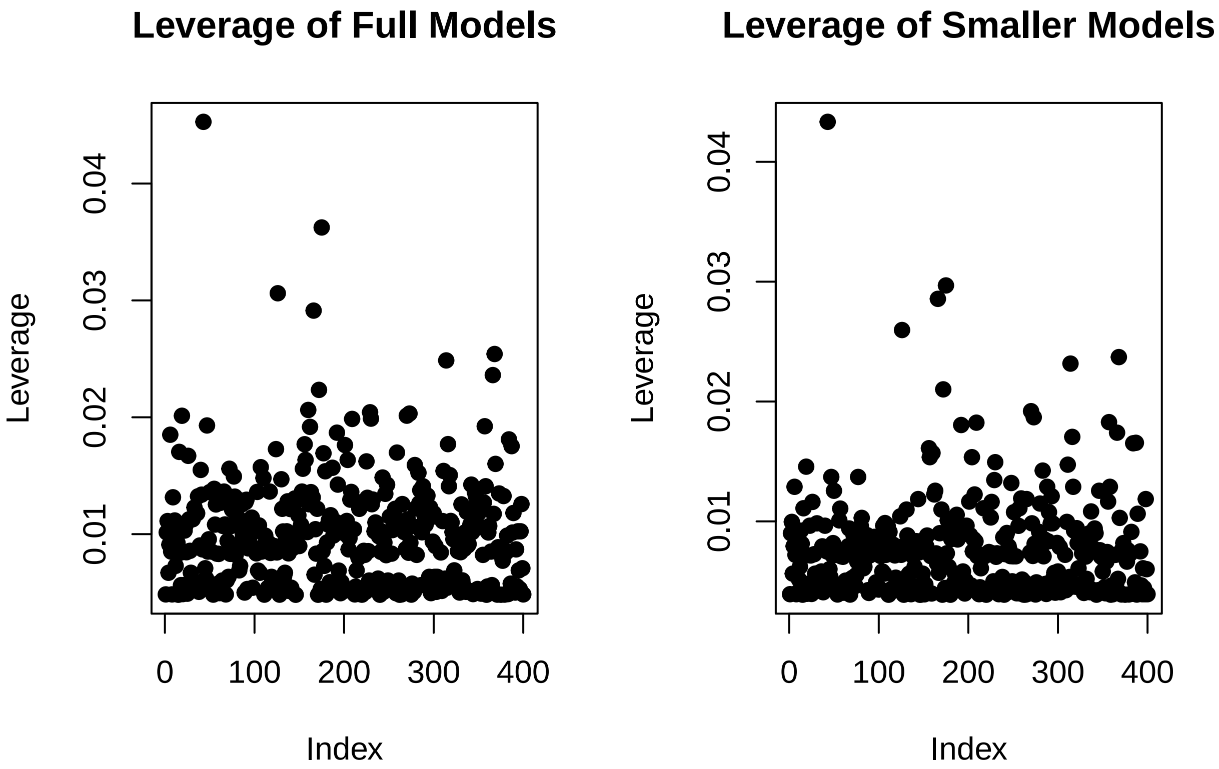


```
par(mfrow = c(1,2))
qqnorm(M_seat$residuals, pch = 19)
qqline(M_seat$residuals, lwd = 2, col = 'red')

qqnorm(M_seat_small$residuals, pch = 19)
qqline(M_seat_small$residuals, lwd = 2, col = 'red')
```



It seems that there are no outliers and the residuals are normally distributed for both of them.

```
par(mfrow = c(1,2))
plot(hatvalues(M_seat), pch = 19, ylab = "Leverage", main = "Leverage of Full Models")
plot(hatvalues(M_seat_small), pch = 19, ylab = "Leverage", main = "Leverage of Smaller Models")
```



Both cases have high leverage values, but smaller model performs better.

I will use AIC, BIC and adjusted R^2 these three criteria to make judgement.

```
aic <- c(AIC(M_seat),AIC((M_seat_small)))
bic <- c(BIC(M_seat),BIC((M_seat_small)))
r <- c(0.2335, 0.2354)
com <- rbind(aic, bic, r)
rownames(com) <- c("AIC", "BIC", paste("Adjusted", expression(R^2)))
colnames(com) <- c("Base Model", "Smaller Model")
knitr::kable(com, format = "latex")
```

	Base Model	Smaller Model
AIC	1865.3121	1863.3186
BIC	1885.2694	1879.2845
Adjusted R^2	0.2335	0.2354

Here, the smaller model has smaller AIC, BIC and greater adjusted R square, which means that smaller model fits the data better.

(g)

```
coe <- M_seat_small$coefficients
std <- c(0.63098, 0.00523, 0.25846)
lower_CI <- coe + qt(0.025, 397)*std
upper_CI <- coe + qt(0.975, 397)*std
Conf.int <- cbind(lower_CI, upper_CI)
knitr::kable(Conf.int, format = "latex")
```

	lower_CI	upper_CI
(Intercept)	11.7903129	14.2712726
Price	-0.0647596	-0.0441957
USYes	0.6915216	1.7077643

(h)

From the picture form question f, it seems that there are not any obvious outliers but there are some points with high leverage values.

```
which.max(hatvalues(M_seat_small))
```

```
## 43
## 43
```

So the highest leverage value point is the 43th data, and it is very far from other leverage values. But there are no outliers.