Peer Assignment of Regression Models

author: “winnie”

date: “Sunday, October 25, 2015”

Overview

You work for Motor Trend, a magazine about the automobile industry. Looking at a data set of a collection of cars, they are interested in exploring the relationship between a set of variables and miles per gallon (MPG) (outcome). They are particularly interested in the following two questions:

- Is an automatic or manual transmission better for MPG?
- Quantify the MPG difference between automatic and manual transmissions?

Analyse and Methods

- Do some exploratory data works before getting start.
- Use the automatic transmission , the manual transmission and both as the preditors to make OLS regression of the outcome MPG respectively and make some tests and plots to find out which one or both is a better preditor.
- Understand the coefficients and residuals
- Because MPG might be influenced by many other variables in the dataset, make a multivariable regression is reasonable as well.

Data Processing

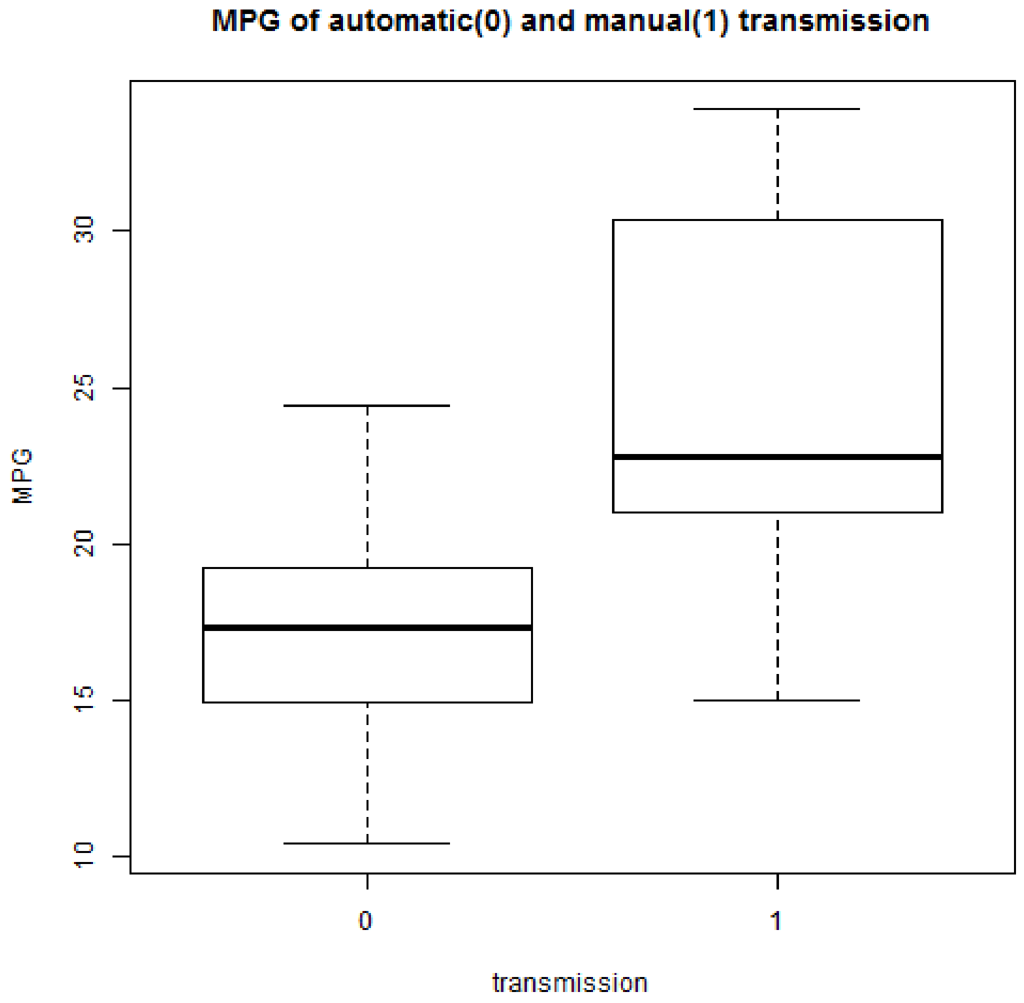
```
library(datasets)
data(mtcars)
str(mtcars)
```

```
## 'data.frame':    32 obs. of  11 variables:
##  $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
##  $ disp: num  160 160 108 258 360 ...
##  $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
##  $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
##  $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
##  $ qsec: num  16.5 17 18.6 19.4 17 ...
##  $ vs  : num  0 0 1 1 0 1 0 1 1 1 ...
##  $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
##  $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
##  $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

- use help(mtcars) to find out what the variables meanIn the help file,we can see that >[, 9]am Transmission (0 = automatic, 1 = manual)
  which indicates the transmission.

Analyse Processing -step1

```
mtcars0 <- transform(mtcars,am=factor(am))
plot(mtcars0$am,mtcars0$mpg,main="MPG of automatic(0) and manual(1) transmission",xlab="transmission", ylab="MPG"
```



- From the boxplot we can see that as the transmission changed from auto to manual, the mean and range of the MPG both increase, so we can assume that manual transimission let the car go more far with manual transmisson.

```
attach(mtcars0)
```

```
## The following objects are masked from mtcars (pos = 6):
## 
##     am, carb, cyl, disp, drat, gear, hp, mpg, qsec, vs, wt
## 
## The following objects are masked from mtcars (pos = 8):
## 
##     am, carb, cyl, disp, drat, gear, hp, mpg, qsec, vs, wt
## 
## The following objects are masked from mtcars0 (pos = 9):
## 
##     am, carb, cyl, disp, drat, gear, hp, mpg, qsec, vs, wt
```

```
table(am)
```

```
## am
##  0  1 
## 19 13
```

```
aggregate(mpg,by=list(am),FUN=mean)
```

```
##   Group.1        x
## 1       0 17.14737
## 2       1 24.39231
```

```
aggregate(mpg,by=list(am),FUN=sd)
```

```
##   Group.1        x
## 1       0 3.833966
## 2       1 6.166504
```

```
fit1 <- aov(mpg~am)
summary(fit1)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)    
## am           1  405.2   405.2   16.86 0.000285 ***
## Residuals   30  720.9    24.0                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- the table tells us that we have 19 observations of auto transmisison and 13 of manual transimission.
- aggregate() function launch a Single factor analysis of variance on the data. we can see that auto transmisison has a mean of 17.14737 MPG while manual transimission has 24.39231, and auto has standard deviation of 3.833966 while manual transimission has 6.166504.
- The p value of F test in aov() function tells us that the transmission has significant influence on the MPG.

Analyse Processing -step2

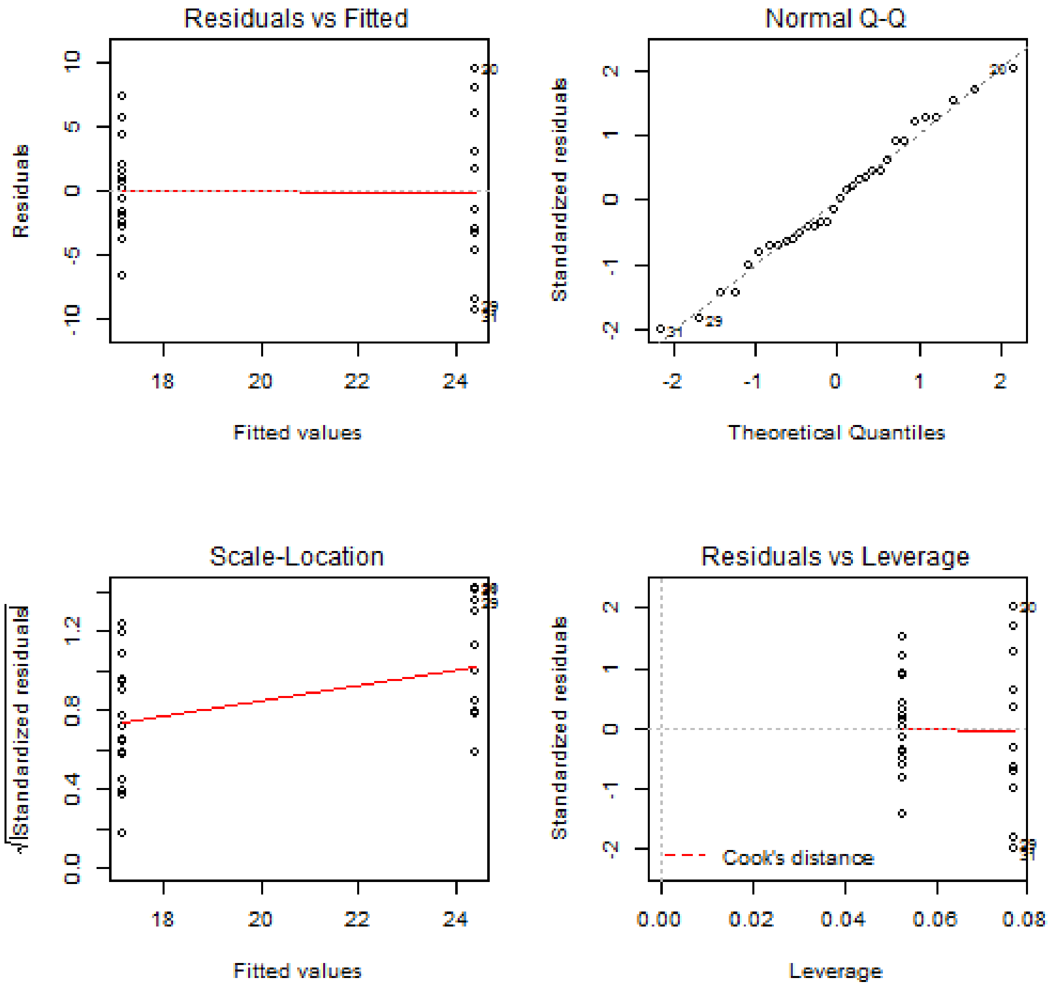
```
attach(mtcars)
```

```
## The following objects are masked from mtcars0 (pos = 3):
## 
##     am, carb, cyl, disp, drat, gear, hp, mpg, qsec, vs, wt
## 
## The following objects are masked from mtcars (pos = 7):
## 
##     am, carb, cyl, disp, drat, gear, hp, mpg, qsec, vs, wt
## 
## The following objects are masked from mtcars (pos = 9):
## 
##     am, carb, cyl, disp, drat, gear, hp, mpg, qsec, vs, wt
## 
## The following objects are masked from mtcars0 (pos = 10):
## 
##     am, carb, cyl, disp, drat, gear, hp, mpg, qsec, vs, wt
```

```
m1 <- lm(mtcars$mpg~mtcars$am)
summary(m1)
```

```
## 
## Call:
## lm(formula = mtcars$mpg ~ mtcars$am)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -9.3923 -3.0923 -0.2974  3.2439  9.5077 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)   17.147      1.125  15.247 1.13e-15 ***
## mtcars$am      7.245      1.764   4.106 0.000285 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 4.902 on 30 degrees of freedom
## Multiple R-squared:  0.3598, Adjusted R-squared:  0.3385 
## F-statistic: 16.86 on 1 and 30 DF,  p-value: 0.000285
```

```
par(mfrow=c(2,2))
plot(m1)
```



```
detach(mtcars)
```

- We can see both the results of t test and dianosis graph indicated that such model of MPG as outcome and transmissions as regressors works well. p-value=0.000285 means that transmissions coefficient is significant.
- And residuals distributed randomly on both sides of line which means residuals meets the normal distribution, and the Q-Q graph with lots of points on the line means MPG are normally distributed, the graph of scale-location also indicates that the MPG meets Homoscedasticity.
- *Interpret the coefficients in the models: when a car change from automatic transmission to manmual transimission(value increase from 0 to 1), the expected miles per gallon increase 7.245. *
- Problem: as the other variables might cause impacts on MPG as well, we can not draw a conclusion on this model because the other variables are not fixed, so we need to make a more reasonable model to interpret the coefficient.

Analyse Processing -step3

```
attach(mtcars)
```

```
## The following objects are masked from mtcars0 (pos = 3):
## 
##     am, carb, cyl, disp, drat, gear, hp, mpg, qsec, vs, wt
## 
## The following objects are masked from mtcars (pos = 7):
## 
##     am, carb, cyl, disp, drat, gear, hp, mpg, qsec, vs, wt
## 
## The following objects are masked from mtcars (pos = 9):
## 
##     am, carb, cyl, disp, drat, gear, hp, mpg, qsec, vs, wt
## 
## The following objects are masked from mtcars0 (pos = 10):
## 
##     am, carb, cyl, disp, drat, gear, hp, mpg, qsec, vs, wt
```

```
m2 <- lm(mpg~cyl+disp+hp+drat+wt+qsec+vs+am+gear+carb)
library(car)
vif(m2)
```

```
##       cyl      disp        hp      drat        wt      qsec        vs 
## 15.373833 21.620241  9.832037  3.374620 15.164887  7.527958  4.965873 
##        am      gear      carb 
##  4.648487  5.357452  7.908747
```

```
sqrt(vif(m2))>2
```

```
##    cyl  disp    hp  drat    wt  qsec    vs    am  gear  carb
##   TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
```

- make a regression model of all of the variables and do some tests to pick them in or out one by one.
- vif() function in the car package, which means Variance Inflation Factor, is useful for validate multicollinearity in the mutivariable regression model.
- sqrt(vif)>2 = there are multicollinearity in the model. From the result above, except for drat, all of the others are related to each other.

```
mtcars1 <- mtcars[,-1]
correlation <- abs(cor(mtcars1))
class(correlation)
```

```
## [1] "matrix"
```

```
diag(correlation) <- 0
which(correlation>0.8,arr.ind=T)
```

```
##      row col
## disp   2   1
## hp     3   1
## vs     7   1
## cyl    1   2
## wt     5   2
## cyl    1   3
## disp   2   5
## cyl    1   7
```

- the originalvariables are seriouly related to each other because many of their R value > 0.8

```
full_model <- lm(mpg~cyl+disp+hp+drat+wt+qsec+vs+am+gear+carb)
summary(full_model)
```

```
## 
## Call:
## lm(formula = mpg ~ cyl + disp + hp + drat + wt + qsec + vs + 
##     am + gear + carb)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -3.4506 -1.6044 -0.1196  1.2193  4.6271 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 12.30337   18.71788   0.657   0.5181  
## cyl         -0.11144    1.04502  -0.107   0.9161  
## disp         0.01334    0.01786   0.747   0.4635  
## hp          -0.02148    0.02177  -0.987   0.3350  
## drat         0.78711    1.63537   0.481   0.6353  
## wt          -3.71530    1.89441  -1.961   0.0633 .
## qsec         0.82104    0.73084   1.123   0.2739  
## vs           0.31776    2.10451   0.151   0.8814  
## am           2.52023    2.05665   1.225   0.2340  
## gear         0.65541    1.49326   0.439   0.6652  
## carb        -0.19942    0.82875  -0.241   0.8122  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 2.65 on 21 degrees of freedom
## Multiple R-squared:  0.869,  Adjusted R-squared:  0.8066 
## F-statistic: 13.93 on 10 and 21 DF,  p-value: 3.793e-07
```

```
reduce_model <- step(full_model,direction="backward")
```

```
## Start:  AIC=70.9
## mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb
## 
##        Df Sum of Sq    RSS    AIC
## - cyl   1    0.0799 147.57 68.915
## - vs    1    0.1601 147.66 68.932
## - carb  1    0.4067 147.90 68.986
## - gear  1    1.3531 148.85 69.190
## - drat  1    1.6270 149.12 69.249
## - disp  1    3.9167 151.41 69.736
## - hp    1    6.8399 154.33 70.348
## - qsec  1    8.8641 156.36 70.765
## <none>              147.49 70.898
## - am    1   10.5467 158.04 71.108
## - wt    1   27.0144 174.51 74.280
## 
## Step:  AIC=68.92
## mpg ~ disp + hp + drat + wt + qsec + vs + am + gear + carb
## 
##        Df Sum of Sq    RSS    AIC
## - vs    1    0.2685 147.84 66.973
## - carb  1    0.5201 148.09 67.028
## - gear  1    1.8211 149.40 67.308
## - drat  1    1.9826 149.56 67.342
## - disp  1    3.9009 151.47 67.750
## - hp    1    7.3632 154.94 68.473
## <none>              147.57 68.915
## - qsec  1   10.0933 157.67 69.032
## - am    1   11.8359 159.41 69.384
## - wt    1   27.0280 174.60 72.297
## 
## Step:  AIC=66.97
## mpg ~ disp + hp + drat + wt + qsec + am + gear + carb
## 
##        Df Sum of Sq    RSS    AIC
## - carb  1    0.6855 148.53 65.121
## - gear  1    2.1437 149.99 65.434
## - drat  1    2.2139 150.06 65.449
## - disp  1    3.6467 151.49 65.753
## - hp    1    7.1060 154.95 66.475
## <none>              147.84 66.973
## - am    1   11.5694 159.41 67.384
## - qsec  1   15.6830 163.53 68.200
## - wt    1   27.3799 175.22 70.410
## 
## Step:  AIC=65.12
## mpg ~ disp + hp + drat + wt + qsec + am + gear
## 
##        Df Sum of Sq    RSS    AIC
## - gear  1     1.565 150.09 63.457
## - drat  1     1.932 150.46 63.535
## <none>              148.53 65.121
## - disp  1    10.110 158.64 65.229
## - am    1    12.323 160.85 65.672
## - hp    1    14.826 163.35 66.166
## - qsec  1    26.408 174.94 68.358
## - wt    1    69.127 217.66 75.350
## 
## Step:  AIC=63.46
## mpg ~ disp + hp + drat + wt + qsec + am
## 
##        Df Sum of Sq    RSS    AIC
## - drat  1     3.345 153.44 62.162
## - disp  1     8.545 158.64 63.229
## <none>              150.09 63.457
## - hp    1    13.285 163.38 64.171
## - am    1    20.036 170.13 65.466
## - qsec  1    25.574 175.67 66.491
## - wt    1    67.572 217.66 73.351
## 
## Step:  AIC=62.16
## mpg ~ disp + hp + wt + qsec + am
```

```
##
##        Df Sum of Sq    RSS    AIC
## - disp  1     6.629 160.07 61.515
## <none>              153.44 62.162
## - hp    1    12.572 166.01 62.682
## - qsec  1    26.470 179.91 65.255
## - am    1    32.198 185.63 66.258
## - wt    1    69.043 222.48 72.051
##
## Step:  AIC=61.52
## mpg ~ hp + wt + qsec + am
##
##        Df Sum of Sq    RSS    AIC
## - hp    1     9.219 169.29 61.307
## <none>              160.07 61.515
## - qsec  1    20.225 180.29 63.323
## - am    1    25.993 186.06 64.331
## - wt    1    78.494 238.56 72.284
##
## Step:  AIC=61.31
## mpg ~ wt + qsec + am
##
##        Df Sum of Sq    RSS    AIC
## <none>              169.29 61.307
## - am    1    26.178 195.46 63.908
## - qsec  1   109.034 278.32 75.217
## - wt    1   183.347 352.63 82.790
```

```
min_model <- lm(mpg~am+qsec+wt)
summary(min_model)
```

```
##
## Call:
## lm(formula = mpg ~ am + qsec + wt)
##
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -3.4811 -1.5555 -0.7257  1.4110  4.6610 
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)   9.6178     6.9596   1.382 0.177915    
## am            2.9358     1.4109   2.081 0.046716 *  
## qsec          1.2259     0.2887   4.247 0.000216 ***
## wt           -3.9165     0.7112  -5.507 6.95e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.459 on 28 degrees of freedom
## Multiple R-squared:  0.8497, Adjusted R-squared:  0.8336 
## F-statistic: 52.75 on 3 and 28 DF,  p-value: 1.21e-11
```

- use step() to make regression model step by step, it will automatically reduce the variables by AIC criteria, so all of the coefficient in min_model are significant.
- Remove the intercept in min_model because it is not significant

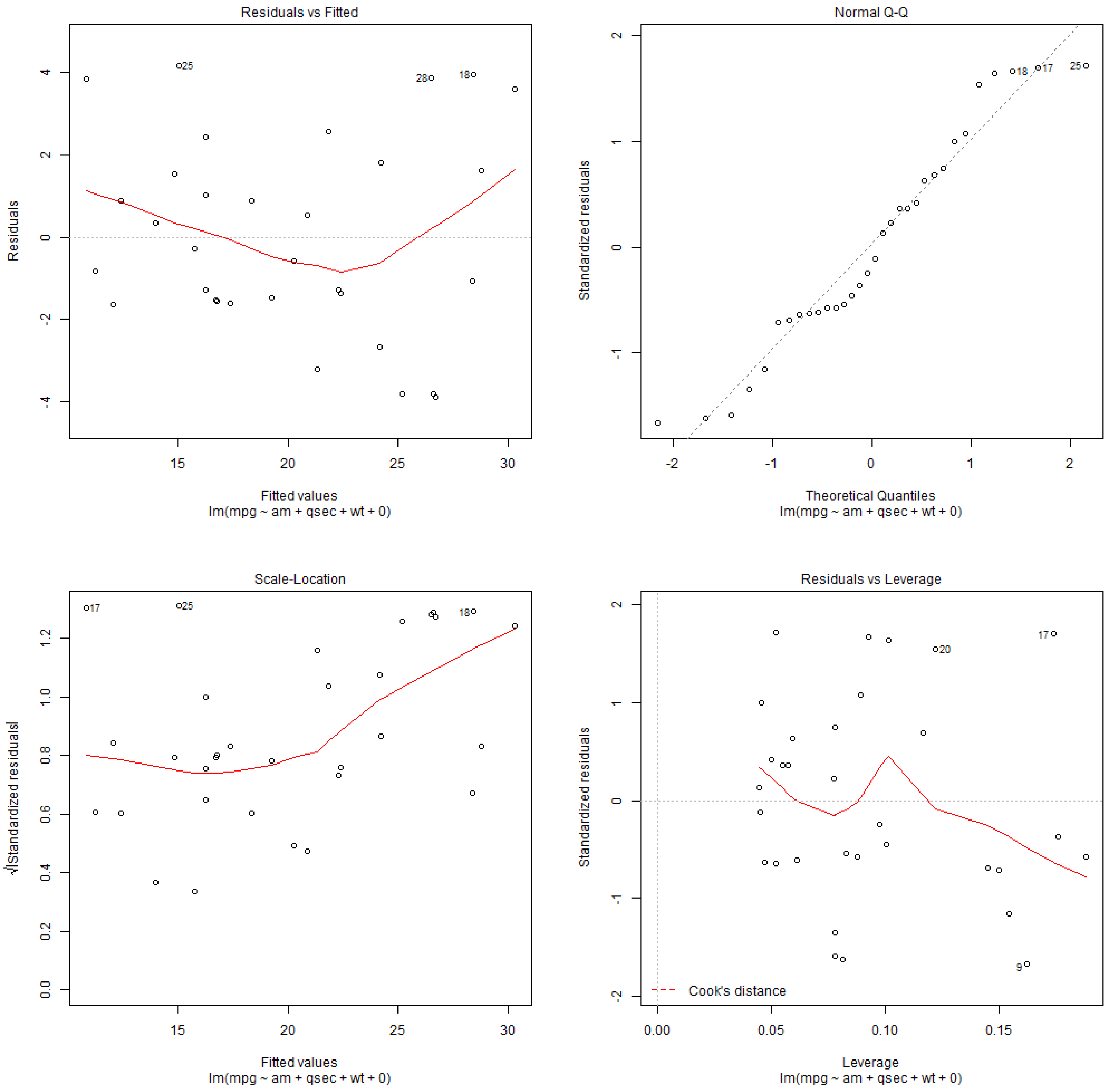
```
min_model2 <- lm(mpg~am+qsec+wt+0)
summary(min_model2)
```

```
##
## Call:
## lm(formula = mpg ~ am + qsec + wt + 0)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8820 -1.5401 -0.4246  1.6623  4.1711
##
## Coefficients:
##      Estimate Std. Error t value Pr(>|t|)
## am     4.2995     1.0241   4.198 0.000233 ***
## qsec   1.5998     0.1021  15.665 1.09e-15 ***
## wt    -3.1855     0.4828  -6.598 3.13e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.497 on 29 degrees of freedom
## Multiple R-squared:  0.9871, Adjusted R-squared:  0.9858
## F-statistic:   741 on 3 and 29 DF,  p-value: < 2.2e-16
```

- interpretation of a coefficient in min_model: When the values of weight and qsec are fixed, if a car change from automatic transmission to manmual transimission(value increase from 0 to 1), the miles per gallon increase 4.2995.

Make the Residuals and Dianogsis plot

```
plot(min_model2)
```



- We can see from the dianosis graph that such model of MPG as outcome and am, qsec and weight as regressors works well. residuals distributed randomly on both sides of line which means residuals meets the normal distribution, and the Q-Q graph with lots of points on the line means MPG are normally distributed, the graph of scale-location also indicates that the MPG meets Homoscedasticity.