L7. Regression models(week4)

Name : Taesoon Kim

Date : Jul-05-2017

Executive summary

Cars can be evaluated by mpg(miles per gallon). If mpg is higher, it is possible to operate more efficiently with less money. I will check which variables are going to affect mpg values using linear regression model. First, using simple linear regression model, I compare mpg level according to transmission ways. I use lm() function in R, and check the relationship between variables, mpg and transmission methods. In conclusion, they have a close relationship, and there is a difference 7.24 mpg whether using automatic transmission and manual transmission. Second, I want to confirm which variables are affected to mpg levels besides the transmission method. As I exclude the variable sequentially, I check the adjusted R squared value. As a result, when I use various variables, I can make more sophisticated results.

Analysis

Load data and check MPG column

```r
# Load dataset
data(mtcars)

#check the "mtcars" data
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 ...
```

```r
# Change from num to factor
mtcars$am[mtcars$am==1]<-"Manual"
mtcars$am[mtcars$am==0]<-"Automatic"
mtcars$am<-as.factor(mtcars$am)
```

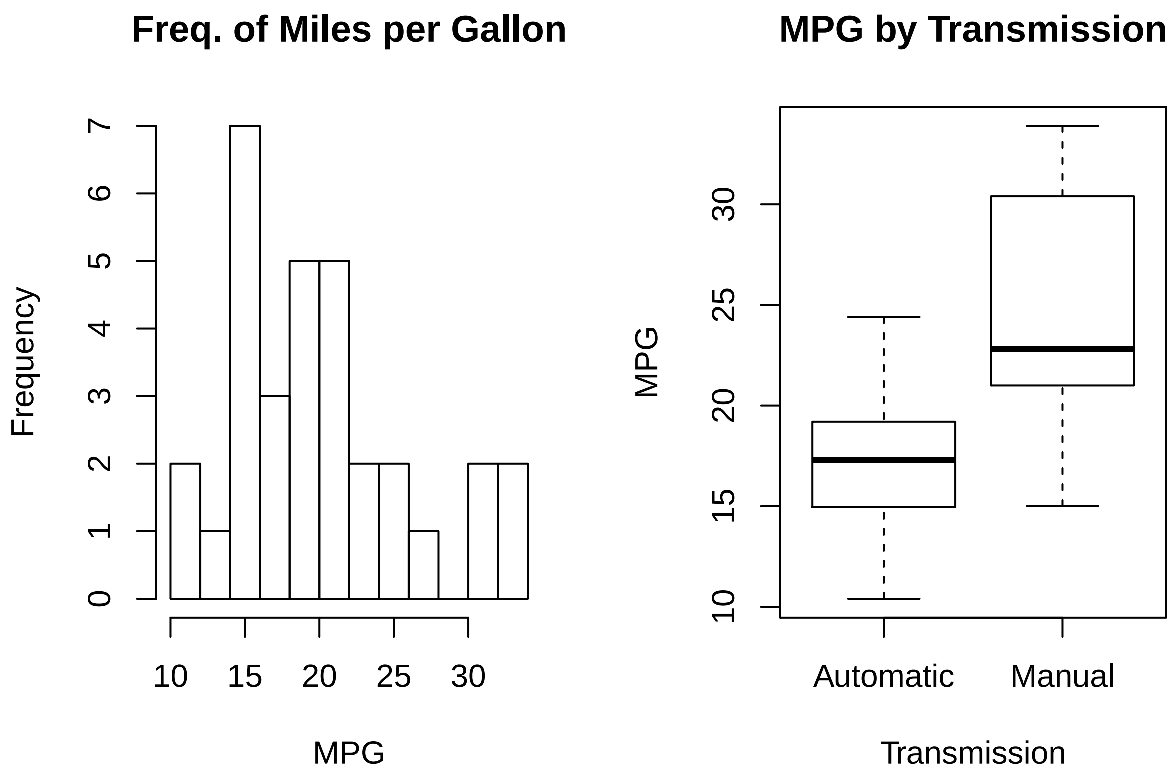
There are 32 rows and 11 columns in “mtcars” data set. Of the 11 variables, mpg means miles per gallon, and am means an automatic and manual transmission. Mpg consists of numbers which mean how many miles can go in one gallon. Am consists of 2 numbers, 0 and 1. 0 means automatic transmission, 1 means manual transmission. I changed the am variable as a factor.

Exploratory data analysis

```
# Focus on MPG and draw a histogram
summary(mtcars$mpg)  # mpg consists of numbers
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   10.40   15.42   19.20   20.09   22.80   33.90
```

```
# Draw a histogram and check the relationship between mpg & transmission
par(mfrow=c(1,2))

hist(mtcars$mpg,breaks=10,main="Freq. of Miles per Gallon",xlab="MPG")
boxplot(mpg~am,mtcars,main="MPG by Transmission",xlab="Transmission",ylab="MPG")
```



The largest number of MPG data is located in the 14~16 range. And also I can check that manual transmission uses more gas than automatic transmission.

Hypothesis testing(Statistical Inference)

```
# Define new variable and store new data frame
mtcars_auto<-mtcars[mtcars$am=="Automatic",]
mtcars_manu<-mtcars[mtcars$am=="Manual",]

t.test(mtcars_auto$mpg,mtcars_manu$mpg)
```

```
## 
```

```
##  Welch Two Sample t-test
##
## data:  mtcars_auto$mpg and mtcars_manu$mpg
## t = -3.7671, df = 18.332, p-value = 0.001374
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.280194  -3.209684
## sample estimates:
## mean of x mean of y
##  17.14737  24.39231
```

P-value is less than 0.05, so it indicates strong evidence against the null hypothesis. The mean difference between automatic and manual is 7.24494.

Simple linear regression

```r
# Using "lm()" function, consider linear regression
fit_mtcars<-lm(mpg~am, mtcars)
summary(fit_mtcars)
```

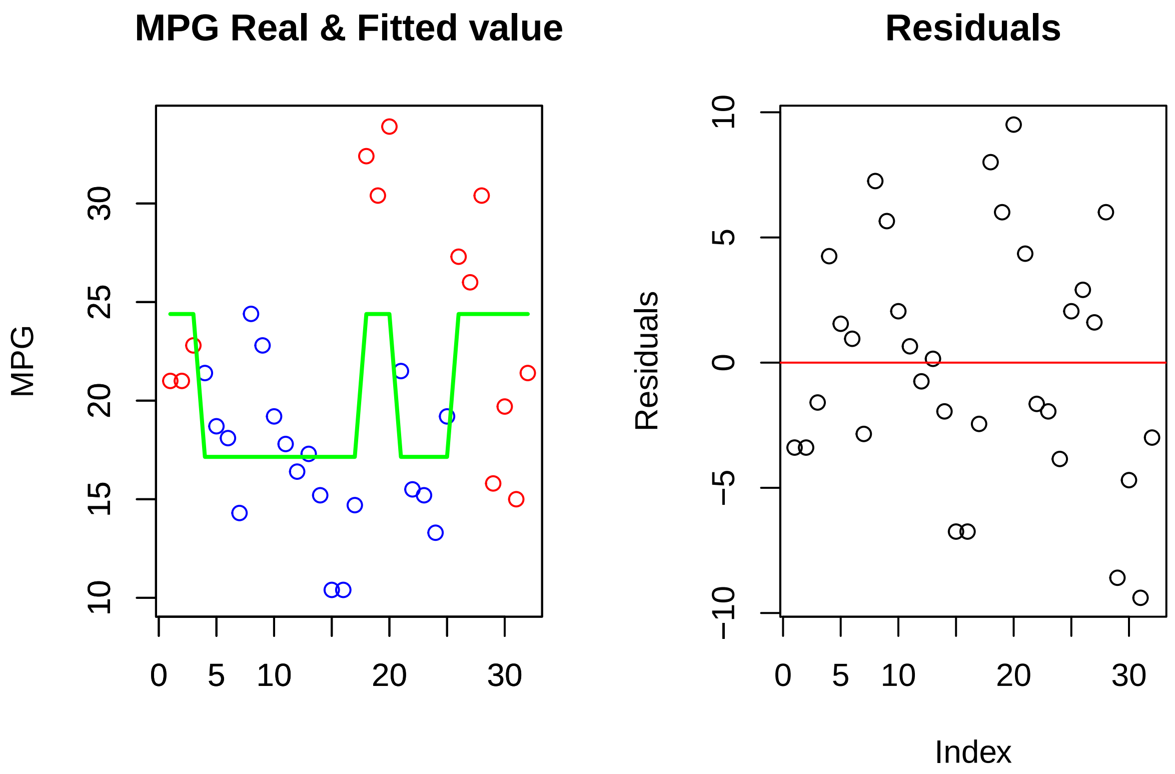
```
##
## Call:
## lm(formula = mpg ~ am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.3923 -3.0923 -0.2974  3.2439  9.5077
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   17.147      1.125  15.247 1.13e-15 ***
## amManual       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
```

I can confirm mean mpg of automatic transmission is 17.147, and mean mpg of manual transmission is 7.245 more.

```r
par(mfrow=c(1,2))

# plot the linear graph
mtcars_fitted<-fitted(fit_mtcars)

plot(mtcars$mpg,main="MPG Real & Fitted value",xlab="",ylab="MPG",ylim=c(10,34),col=c("blue","red")[mtca
par(new=T)
plot(mtcars_fitted,type="l",xlab="",ylab="",ylim=c(10,34),col="green",lwd=2)

plot(residuals(fit_mtcars),main="Residuals",Xlab="",ylab="Residuals")
```

```
## Warning in plot.window(...): "Xlab" is not a graphical parameter
## Warning in plot.xy(xy, type, ...): "Xlab" is not a graphical parameter

## Warning in axis(side = side, at = at, labels = labels, ...): "Xlab" is not
## a graphical parameter

## Warning in axis(side = side, at = at, labels = labels, ...): "Xlab" is not
## a graphical parameter

## Warning in box(...): "Xlab" is not a graphical parameter

## Warning in title(...): "Xlab" is not a graphical parameter
```

```
abline(h=0,col="red")
```



The blue point is the mpg value of automatic transmission, and the red point is the mpg value of manual transmission. The green line is fitted line, and it means the predicted value using linear regression. The right graph is residuals of linear regression model.

Multiple linear regression

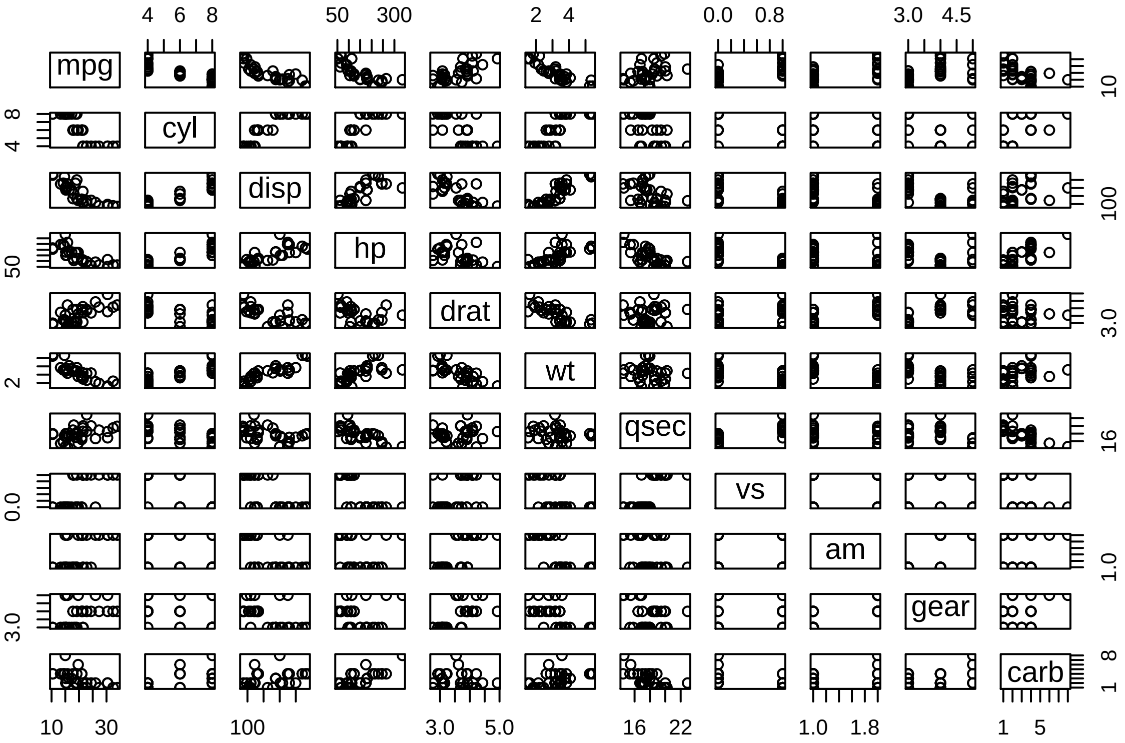
In order to model selection, I am going to use backward elimination method, which removes variables sequentially. First, I starts with the model that includes all variables. Next, I exclude a variable in order, as comparing with adjusted R squared value.

```
# Check the overall parameters
fit_overall<-lm(mpg~.,mtcars)
summary(fit_overall)
```

```
##
```

```
## Call:
## lm(formula = mpg ~ ., data = mtcars)
##
## 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
## amManual     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
```

```r
plot(mtcars)
```



```r
# Exclude the variables sequentially, and compare the R squared value
fit_ex_cyl<-lm(mpg~disp+hp+drat+wt+qsec+vs+am+gear+carb,mtcars)
summary(fit_ex_cyl)
```

```
##
## Call:
## lm(formula = mpg ~ disp + hp + drat + wt + qsec + vs + am + gear +
##     carb, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4286 -1.5908 -0.0412  1.2120  4.5961
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 10.96007   13.53030   0.810   0.4266
## disp         0.01283    0.01682   0.763   0.4538
## hp          -0.02191    0.02091  -1.048   0.3062
## drat         0.83520    1.53625   0.544   0.5921
## wt          -3.69251    1.83954  -2.007   0.0572 .
## qsec         0.84244    0.68678   1.227   0.2329
## vs           0.38975    1.94800   0.200   0.8433
## amManual     2.57743    1.94035   1.328   0.1977
## gear         0.71155    1.36562   0.521   0.6075
## carb        -0.21958    0.78856  -0.278   0.7833
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.59 on 22 degrees of freedom
## Multiple R-squared:  0.8689, Adjusted R-squared:  0.8153
## F-statistic: 16.21 on 9 and 22 DF,  p-value: 9.031e-08
```

```r
fit_ex_vs<-lm(mpg~disp+hp+drat+wt+qsec+am+gear+carb,mtcars)
summary(fit_ex_vs)
```

```
##
## Call:
## lm(formula = mpg ~ disp + hp + drat + wt + qsec + am + gear +
##     carb, data = mtcars)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.356 -1.576 -0.149  1.218  4.604
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.76828   11.89230   0.821   0.4199
## disp         0.01214    0.01612   0.753   0.4590
## hp          -0.02095    0.01993  -1.051   0.3040
## drat         0.87510    1.49113   0.587   0.5630
## wt          -3.71151    1.79834  -2.064   0.0505 .
## qsec         0.91083    0.58312   1.562   0.1319
## amManual     2.52390    1.88128   1.342   0.1928
## gear         0.75984    1.31577   0.577   0.5692
## carb        -0.24796    0.75933  -0.327   0.7470
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.535 on 23 degrees of freedom
## Multiple R-squared:  0.8687, Adjusted R-squared:  0.823
## F-statistic: 19.02 on 8 and 23 DF,  p-value: 2.008e-08
```

```r
fit_ex_carb<-lm(mpg~disp+hp+drat+wt+qsec+am+gear,mtcars)
summary(fit_ex_carb)
```

```
##
## Call:
## lm(formula = mpg ~ disp + hp + drat + wt + qsec + am + gear,
##     data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.1200 -1.7753 -0.1446  1.0903  4.7172
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.19763   11.54220   0.797  0.43334
## disp         0.01552    0.01214   1.278  0.21342
## hp          -0.02471    0.01596  -1.548  0.13476
## drat         0.81023    1.45007   0.559  0.58151
## wt          -4.13065    1.23593  -3.342  0.00272 **
## qsec         1.00979    0.48883   2.066  0.04981 *
## amManual     2.58980    1.83528   1.411  0.17104
## gear         0.60644    1.20596   0.503  0.61964
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.488 on 24 degrees of freedom
## Multiple R-squared:  0.8681, Adjusted R-squared:  0.8296
## F-statistic: 22.56 on 7 and 24 DF,  p-value: 4.218e-09
```

```
fit_ex_gear<-lm(mpg~disp+hp+drat+wt+qsec+am,mtcars)
summary(fit_ex_gear)
```

```
##
## Call:
## lm(formula = mpg ~ disp + hp + drat + wt + qsec + am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2669 -1.6148 -0.2585  1.1220  4.5564
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 10.71062   10.97539   0.976  0.33848
## disp         0.01310    0.01098   1.193  0.24405
## hp          -0.02180    0.01465  -1.488  0.14938
## drat         1.02065    1.36748   0.746  0.46240
## wt          -4.04454    1.20558  -3.355  0.00254 **
## qsec         0.99073    0.48002   2.064  0.04955 *
## amManual     2.98469    1.63382   1.827  0.07969 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.45 on 25 degrees of freedom
## Multiple R-squared:  0.8667, Adjusted R-squared:  0.8347
## F-statistic: 27.09 on 6 and 25 DF,  p-value: 8.637e-10
```

```
fit_ex_drat<-lm(mpg~disp+hp+wt+qsec+am,mtcars)
summary(fit_ex_drat)
```

```
##
## Call:
## lm(formula = mpg ~ disp + hp + wt + qsec + am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5399 -1.7398 -0.3196  1.1676  4.5534
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 14.36190    9.74079   1.474  0.15238
## disp         0.01124    0.01060   1.060  0.29897
## hp          -0.02117    0.01450  -1.460  0.15639
## wt          -4.08433    1.19410  -3.420  0.00208 **
## qsec         1.00690    0.47543   2.118  0.04391 *
## amManual     3.47045    1.48578   2.336  0.02749 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.429 on 26 degrees of freedom
## Multiple R-squared:  0.8637, Adjusted R-squared:  0.8375
## F-statistic: 32.96 on 5 and 26 DF,  p-value: 1.844e-10
```

```
fit_ex_disp<-lm(mpg~hp+wt+qsec+am,mtcars)
summary(fit_ex_disp)
```

```
##
## Call:
## lm(formula = mpg ~ hp + wt + qsec + am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4975 -1.5902 -0.1122  1.1795  4.5404
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 17.44019    9.31887   1.871  0.07215 .
## hp          -0.01765    0.01415  -1.247  0.22309
## wt          -3.23810    0.88990  -3.639  0.00114 **
## qsec         0.81060    0.43887   1.847  0.07573 .
## amManual     2.92550    1.39715   2.094  0.04579 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.435 on 27 degrees of freedom
## Multiple R-squared:  0.8579, Adjusted R-squared:  0.8368
## F-statistic: 40.74 on 4 and 27 DF,  p-value: 4.589e-11
```

```
fit_ex_hp<-lm(mpg~wt+qsec+am,mtcars)
summary(fit_ex_hp)
```

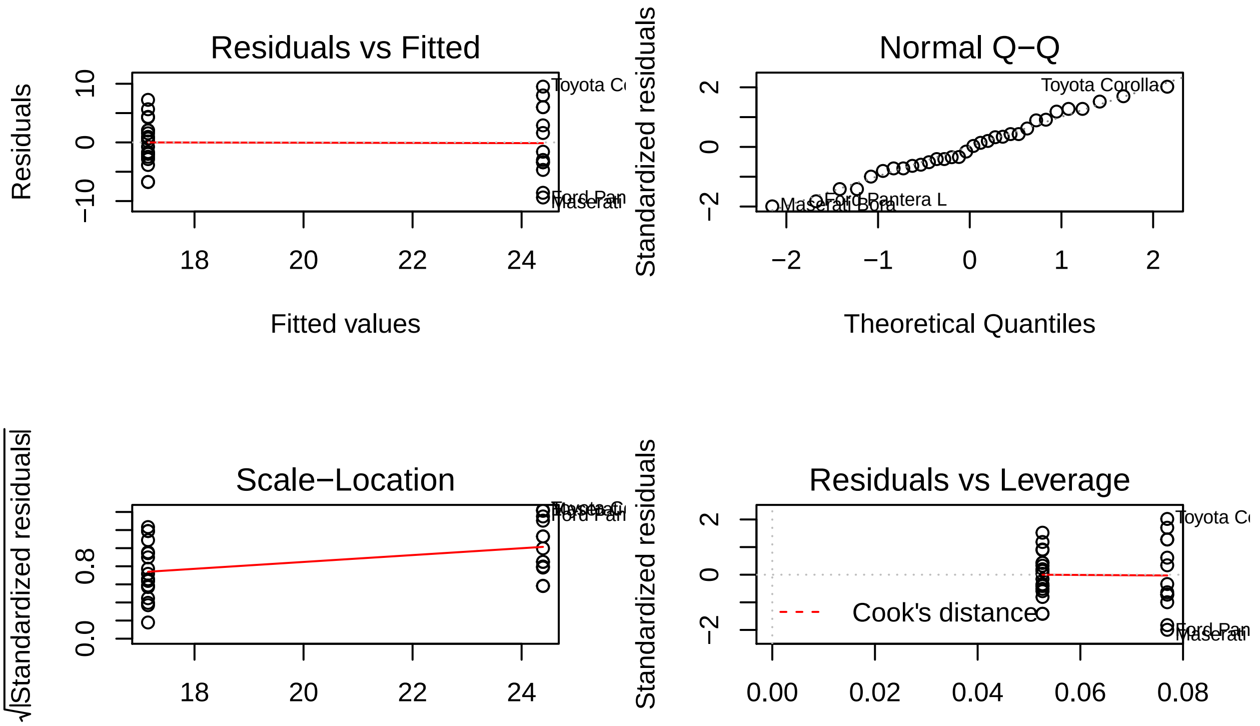
```
##
## Call:
## lm(formula = mpg ~ wt + qsec + am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4811 -1.5555 -0.7257  1.4110  4.6610
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.6178     6.9596   1.382 0.177915
## wt           -3.9165     0.7112  -5.507 6.95e-06 ***
## qsec          1.2259     0.2887   4.247 0.000216 ***
## amManual      2.9358     1.4109   2.081 0.046716 *
## ---
## 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
```

When I use all varibles, the adjusted R squared value is 0.8066. As I exclude variables, “cyl”, “vs”, “carb”, “gear”, “drat”, the adjusted R squared value goes up to 0.8375. However, when I exclude more variables, the adjusted R squared value goes down. So if “disp”, “hp”, “wt”, “qsec”, “am” are used, regression model is the best.

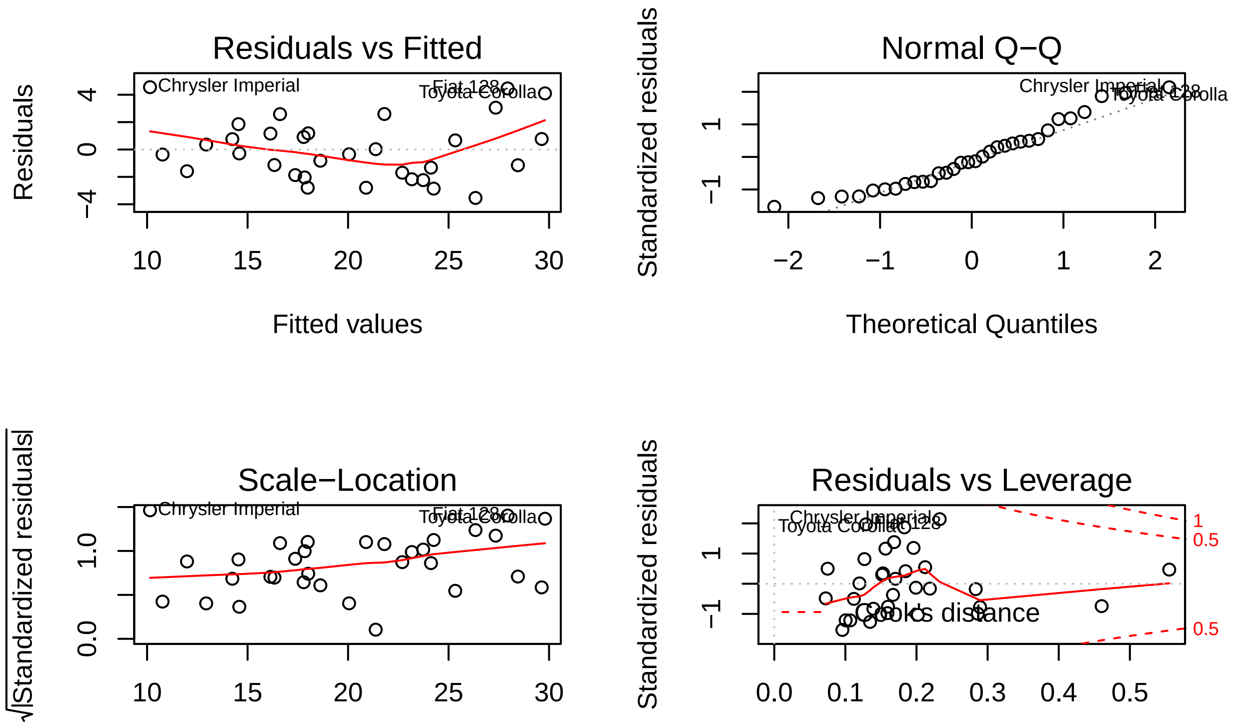
```
# Use anova function
anova(fit_mtcars,fit_ex_drat)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ disp + hp + wt + qsec + am
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     30 720.90
## 2     26 153.44  4    567.46 24.039 2.067e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# plot
par(mfrow=c(2,2))
plot(fit_mtcars)
```



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



P-value is less than 0.05, so I reject the null hypothesis and claim that multivariate model is different from simple linear regression model.

Result

In conclusion, manual transmission cars are better mpg than automatic transmission cars. The average mpg is 7.24 higher, and there is relationship between mpg and transmission ways. Besides, if we consider additional variables, such as “disp”, “hp” and so on, we can make more sophisticated regression model.