Regression Project _ Motor Trend

Luoning

5/15/2020

1.Synopsis

The work is based on data of Motor Trend, a magazine about the automobile industry. The data set of a collection of cars was reviewed. And the relationship between a set of variables and miles per gallon (MPG) (outcome) was analyzed. Two particular questions were answered: 1) whether an automatic or manual transmission is better for MPG; 2) whether the the MPG difference between automatic and manual transmissions can be quantified.

2.Data download and process

```
# Datasets
library(datasets)
data(mtcars)

# View of the data
head(mtcars)
```

```
##                    mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

```
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 ...
```

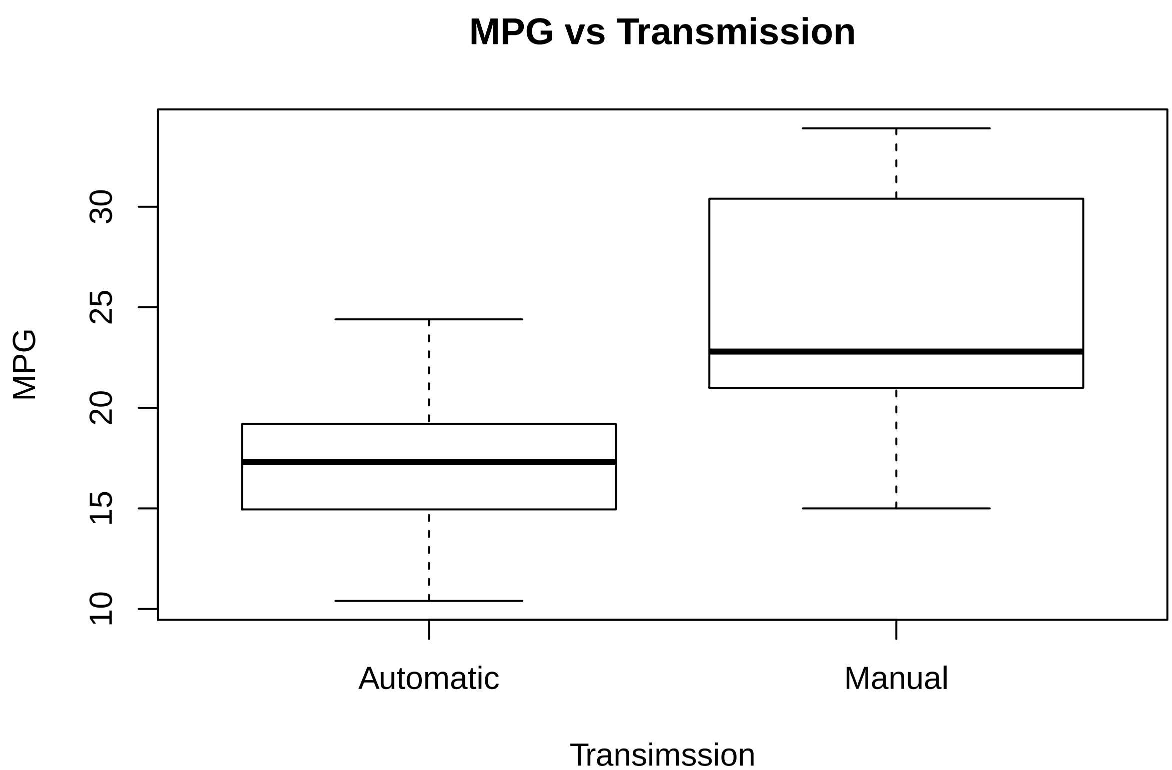
The data frame has 32 observatoins on 11 variables.
[, 1] mpg Miles/(US) gallon
[, 2] cyl Number of cylinders
[, 3] disp Displacement (cu.in.)

[, 4] hp Gross horsepower
[, 5] drat Rear axle ratio
[, 6] wt Weight (1000 lbs)
[, 7] qsec 1/4 mile time
[, 8] vs Engine (0 = V-shaped, 1 = straight)
[, 9] am Transmission (0 = automatic, 1 = manual)
[,10] gear Number of forward gears
[,11] carb Number of carburetors

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
mtcars<-mutate(mtcars,Transmission=am)
mtcars$Transmission[mtcars$Transmission==1]<-'Manual'
mtcars$Transmission[mtcars$Transmission==0]<-'Automatic'
mtcars$Transmission<-factor(paste(mtcars$Transmission))
# Plot the mpg of manual vs automatic
boxplot(mtcars$mpg ~ mtcars$Transmission, data=mtcars, outpch = 19, ylab="MPG", xlab="Transimssion", mai
```



From visuallization, the mpg of manual transmission is higher than that of automatic transmission. Now let's evaluate the significance.

```
auto<-mtcars[mtcars$Transmission=='Automatic',]
manual<-mtcars[mtcars$Transmission=='Manual',]
t.test(auto$mpg,manual$mpg)
```

```
##
##  Welch Two Sample t-test
##
## data:  auto$mpg and manual$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
```

The t value is negative and the confidence interval is absolutely below zero, which means the hypothesis that the automatic and manual transmission are the same (t=0) is rejected. The difference in the mpg of automatic and manual transmission is significant, and the mpg of automatic is lower than manual.

Regression Model

To quantify the difference between the automatic and manual transmission. The regression model is applied to evaluate the correlation of all other variables vs. transmission and mpg relationship.

```
fit1<-lm(mpg~am,mtcars)
summary(fit1)
```

```
##
## 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 ***
## 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
```

The p-value is low but the R value is only 0.3385, which means other variables may influence the mpg. Now let's consider other variables and perform the multivariable linear regression.

```
fit2<-lm(mpg~am+cyl,mtcars)
anova(fit1,fit2)
```

```
## Analysis of Variance Table
```

```
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + cyl
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     30 720.90
## 2     29 271.36  1    449.53 48.041 1.285e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value is low and the cyl is significant, which means the addition of cyl as a variable is necessary. We can add more variables to check the significance.

```r
fit2.1<-lm(mpg~am+cyl+disp+hp+drat,mtcars)
anova(fit2,fit2.1)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am + cyl
## Model 2: mpg ~ am + cyl + disp + hp + drat
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     29 271.36
## 2     26 214.50  3    56.865 2.2976  0.101
```

The addition of displacement (disp), Gross horsepower (hp) and Rear axle ratio (drat) is not necessary.

```r
fit2.2<-lm(mpg~am+cyl+wt+qsec+vs,mtcars)
anova(fit2,fit2.2)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am + cyl
## Model 2: mpg ~ am + cyl + wt + qsec + vs
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1     29 271.36
## 2     26 167.52  3    103.84 5.3725 0.005163 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
fit2.3<-lm(mpg~am+cyl+gear+carb,mtcars)
anova(fit2,fit2.3)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am + cyl
## Model 2: mpg ~ am + cyl + gear + carb
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     29 271.36
## 2     27 206.58  2    64.782 4.2335 0.02516 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The new variables weight (wt), 1/4 mile time (qsec) and Engine (vs) has at least one variable necessary. And at least one in number of forward gears (gear) and number of carburetors (carb) is necessary.

```r
fit2.4<-lm(mpg~am+cyl+wt,mtcars)
fit2.5<-lm(mpg~am+cyl+wt+qsec,mtcars)
fit2.6<-lm(mpg~am+cyl+wt+qsec+vs,mtcars)
anova(fit2,fit2.4,fit2.5,fit2.6)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am + cyl
## Model 2: mpg ~ am + cyl + wt
## Model 3: mpg ~ am + cyl + wt + qsec
## Model 4: mpg ~ am + cyl + wt + qsec + vs
##   Res.Df    RSS Df Sum of Sq       F   Pr(>F)
## 1     29 271.36
## 2     28 191.05  1    80.315 12.4655 0.001569 **
## 3     27 167.78  1    23.262  3.6105 0.068561 .
## 4     26 167.52  1     0.268  0.0416 0.840041
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Only weight is significant. carburetors (carb) is necessary.

```r
fit3<-lm(mpg~am+cyl+wt,mtcars)
fit3.1<-lm(mpg~am+cyl+wt+gear,mtcars)
fit3.2<-lm(mpg~am+cyl+wt+gear+carb,mtcars)
anova(fit3,fit3.1,fit3.2)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am + cyl + wt
## Model 2: mpg ~ am + cyl + wt + gear
## Model 3: mpg ~ am + cyl + wt + gear + carb
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     28 191.05
## 2     27 183.93  1    7.1156 1.0994 0.3040
## 3     26 168.27  1   15.6588 2.4195 0.1319
```

Neither of forward gears (gear) and number of carburetors (carb) is necessary.
Therefore, transmission (am), number of cylinders (cyl) and weight (wt) play a significant role on mpg. This can also be seen from the plots of mpg vs. number of cylinders and weight.

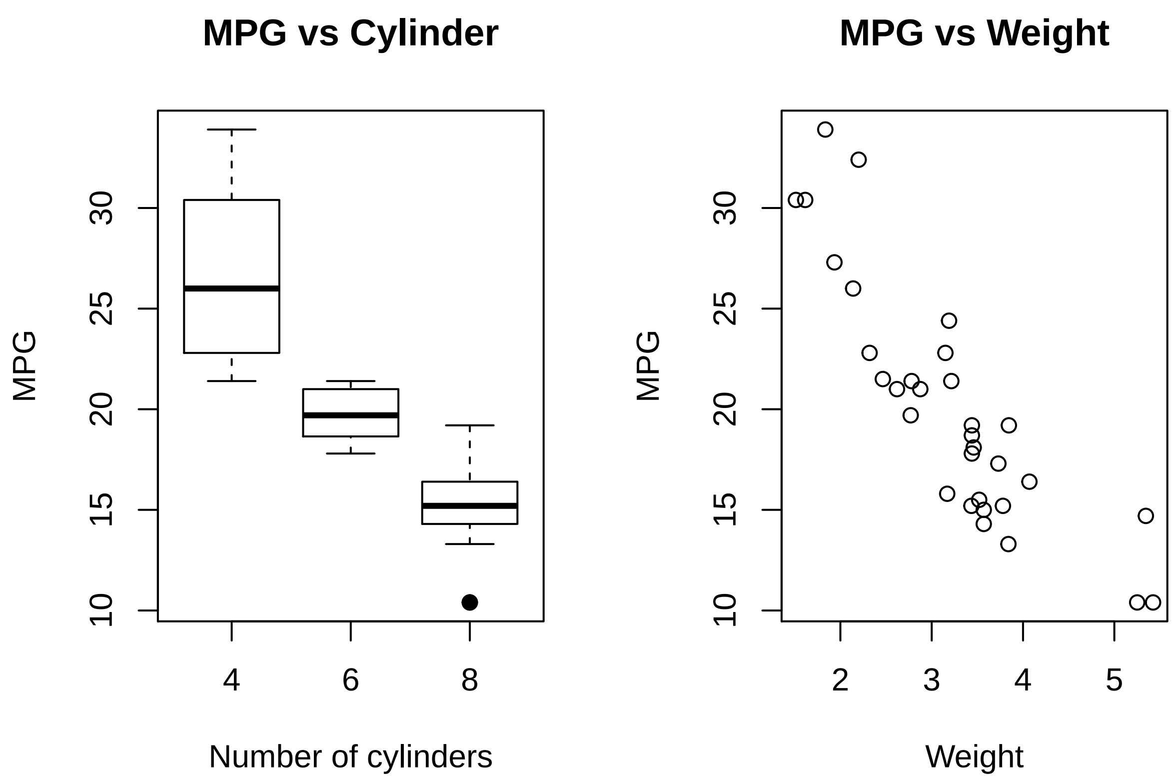
```r
summary(fit3)
```

```
##
## Call:
## lm(formula = mpg ~ am + cyl + wt, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.1735 -1.5340 -0.5386  1.5864  6.0812
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  39.4179     2.6415  14.923 7.42e-15 ***
## am            0.1765     1.3045   0.135  0.89334
## cyl          -1.5102     0.4223  -3.576  0.00129 **
## wt           -3.1251     0.9109  -3.431  0.00189 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.612 on 28 degrees of freedom
## Multiple R-squared:  0.8303, Adjusted R-squared:  0.8122
```

```
## F-statistic: 45.68 on 3 and 28 DF,  p-value: 6.51e-11
```

```
par(mfrow=c(1,2))
boxplot(mtcars$mpg ~ mtcars$cyl, data=mtcars, outpch = 19, ylab="MPG", xlab="Number of cylinders", main=
plot(mtcars$mpg ~ mtcars$wt, data=mtcars, outpch = 19, ylab="MPG", xlab="Weight", main="MPG vs Weight")
```

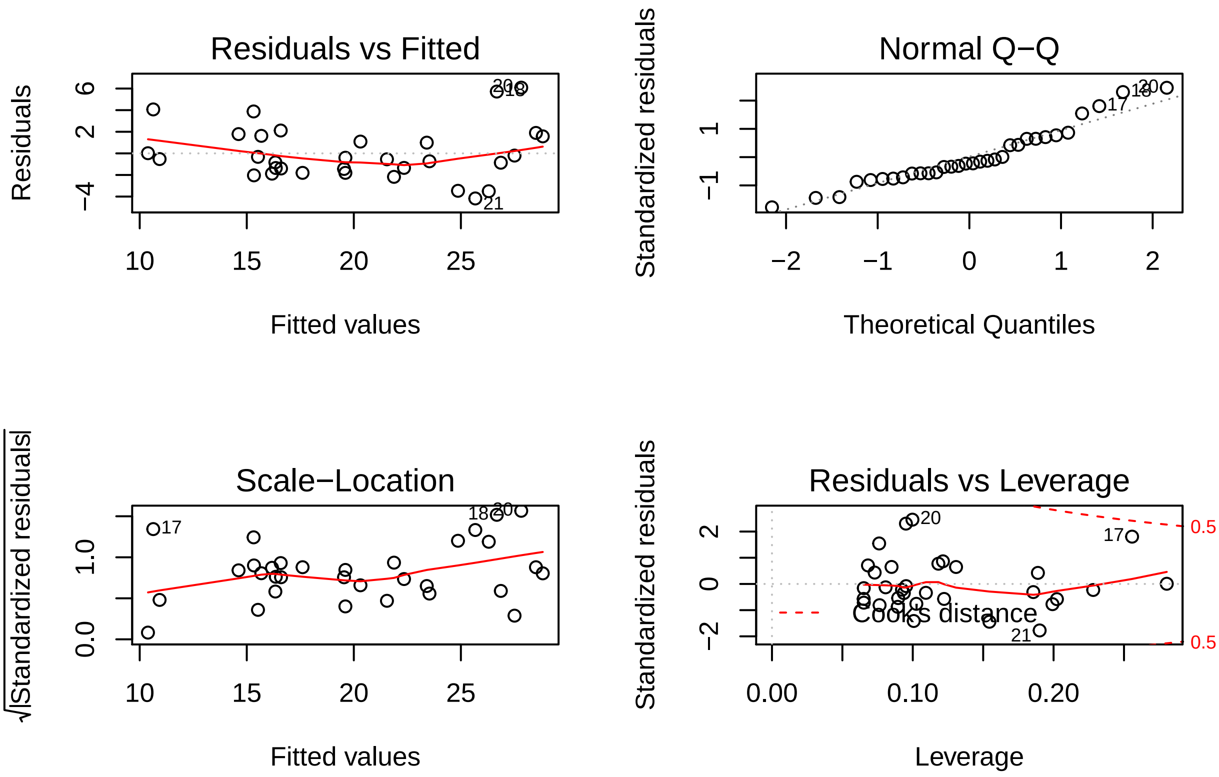
```
## Warning in plot.window(...): "outpch" is not a graphical parameter

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

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

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

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

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



Residual and Diagnostics

Multivariable regression model residuals

```
par(mfrow=c(2,2))
plot(fit3)
```




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

From the above plots, we can make the following observations:
The residuals appear to be randomly scattered on the plot and verify the independence condition.
The points in Q-Q plot mostly fall on the line which indicates the normally distributed residuals.
The Scale-Location plot consists of points scattered in a constant band pattern, indicating constant variance.
The outliers or leverage points are limited and acceptible.