Regression Models Project

Mehdi HAMDOUNE

3/24/2021

Introduction

We Look at a data set of a collection of cars (mtcars), we are interested in exploring the relationship between a set of variables and miles per gallon (MPG) (outcome). we 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"

The variables included in this dataset are:

- mpg : Miles/(US) gallon
- cyl : Number of cylinders
- disp : Displacement (cu.in.)
- hp : Gross horsepower
- drat : Rear axle ratio
- wt : Weight (1000 lbs)
- qsec : 1/4 mile time
- vs : Engine (0 = V-shaped, 1 = straight)
- am : Transmission (0 = automatic, 1 = manual)
- gear : Number of forward gears

Loading and preprocessing the data

```
setwd("D:/Auto_formation/Coursera_Regression_Models/Project")

# Importing The Librairies

library(plyr)
library(tidyr)
library(dplyr)
library(ggplot2)
library(ggpubr)


data("mtcars")
```

```
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 ...
```

```
sum(is.na(mtcars))
```

```
## [1] 0
```

Analysis

```
atData<-mtcars[mtcars$am == 0,]
mtData<-mtcars[mtcars$am == 1,]
t.test(atData$mpg, mtData$mpg)
```

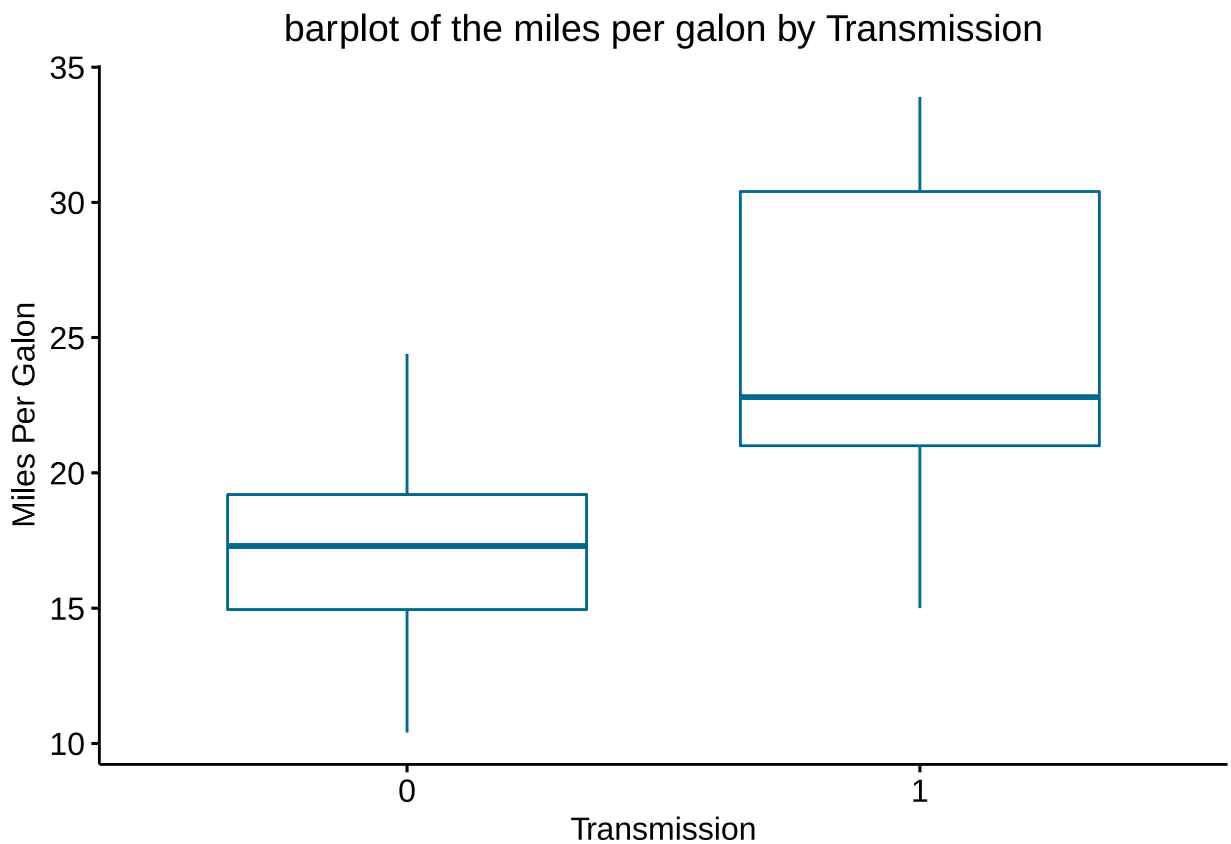
Running t-test

```
##
##  Welch Two Sample t-test
##
## data:  atData$mpg and mtData$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
```

```r
mtcars$am <- as.factor(mtcars$am)
```

```r
# Box plot
g <- ggboxplot(mtcars, x = "am", y = "mpg",
    color = "deepskyblue4", palette = "jco") +
  labs(title="barplot of the miles per galon by Transmission", x= "Transmission",
       y = "Miles Per Galon") +
  theme(plot.title = element_text(hjust = 0.5))

print(g)
```

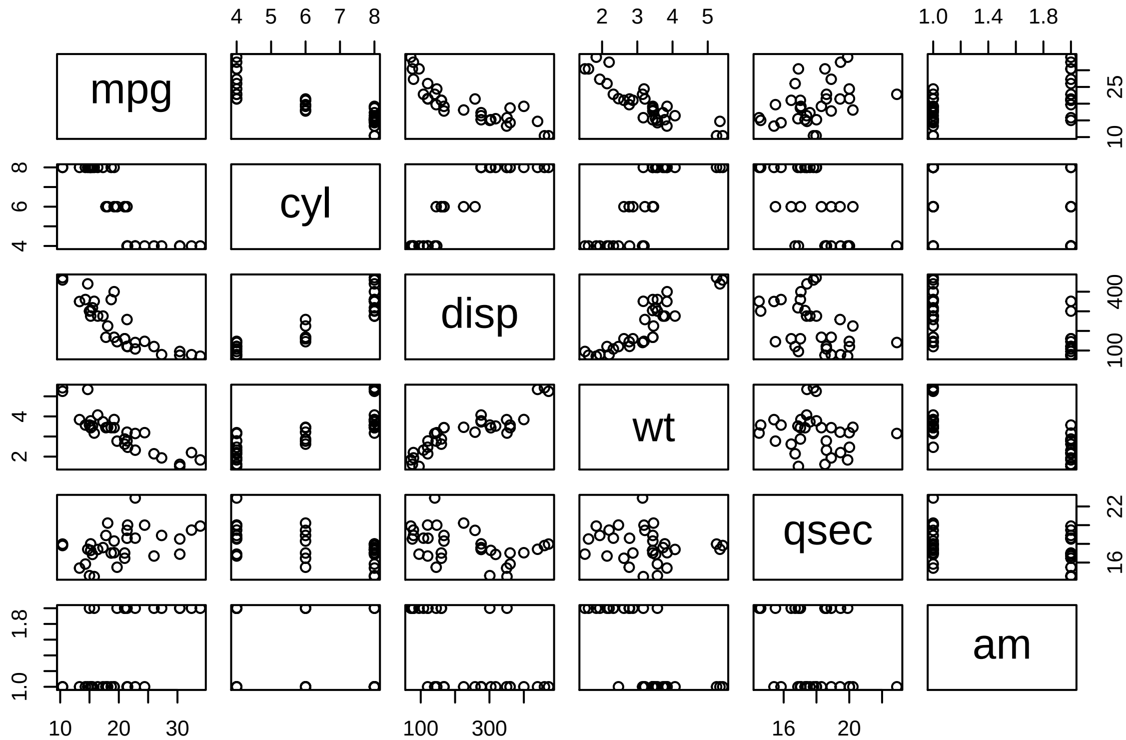
Barplot of the miles per galon by Transmission



From this barplot we can assume that there is a significant association between Miles Per Galon and Transmission.

```r
corr <- select(mtcars, mpg,cyl,disp,wt,qsec, am)
cor <- pairs(corr)
```

Correlations




```
print(cor)
```

```
## NULL
```

Linear Models

```
fit <- step(lm(data = mtcars, mpg ~ .),trace=0,steps=10000)
summary(fit)
```

Choosing The Predictors

```
##
## 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 ***
## am1           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
```

This model uses a step algorithm to pick the variables that affect the mpg of cars the most. From the model, the weight, acceleration as well as the transmission mode affect the mpg of the car the most.

Based on this multivariate regression model, a manual transmission car has a fuel efficiency of 2.94 MPG higher than that of automatic transmission cars. The adjusted R^2 of the model is 0.834, meaning that 83% of the variance in mpg can be explained by the model.

```r
fit2 <- lm(mpg ~ cyl + disp + wt + qsec + am, mtcars)
summary(fit2)
```

2nd Model

```
##
## Call:
## lm(formula = mpg ~ cyl + disp + wt + qsec + am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.6943 -1.2712 -0.4874  1.3427  4.7652
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 14.94065   12.89964   1.158 0.257302
## cyl         -0.67811    0.78813  -0.860 0.397431
## disp         0.01173    0.01158   1.013 0.320386
## wt          -4.41955    1.18871  -3.718 0.000971 ***
## qsec         1.10914    0.53173   2.086 0.046943 *
## am1          2.61118    1.72324   1.515 0.141767
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.492 on 26 degrees of freedom
## Multiple R-squared:  0.8567, Adjusted R-squared:  0.8291
## F-statistic: 31.08 on 5 and 26 DF,  p-value: 3.523e-10
```

Based on this multivariate regression model, a manual transmission car has a fuel efficiency of 2.54 MPG higher than that of automatic transmission cars. The adjusted R^2 of the model is 0.826, meaning that 82% of the variance in mpg can be explained by the model.

```r
fit3 <- lm(mpg ~ ., mtcars)
summary(fit3)
```

3rd Model

```
##
## 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
## am1          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
```

Based on this multivariate regression model, a manual transmission car has a fuel efficiency of 2.94 MPG higher than that of automatic transmission cars. The adjusted R^2 of the model is 0.8165, meaning that 81.5% of the variance in mpg can be explained by the model.

```r
anova(fit, fit2, fit3)
```

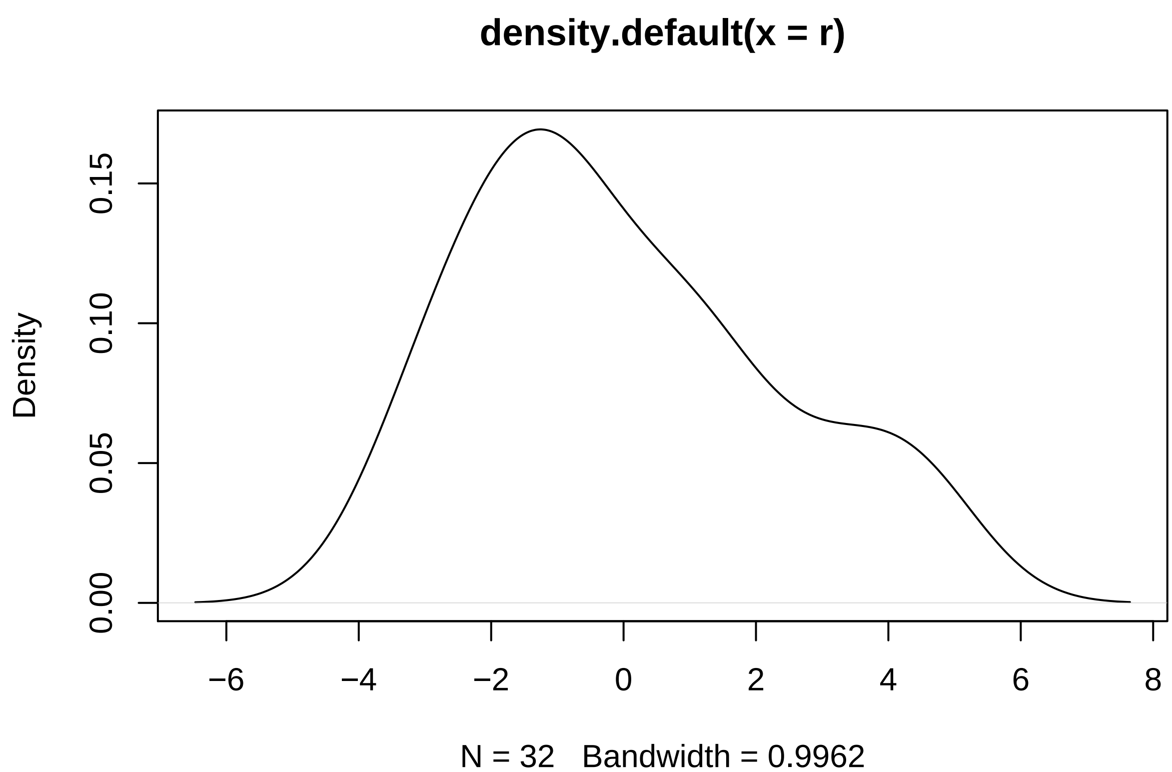
Evaluating The Models

```
## Analysis of Variance Table
##
## Model 1: mpg ~ wt + qsec + am
## Model 2: mpg ~ cyl + disp + wt + qsec + am
## Model 3: mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     28 169.29
## 2     26 161.41  2     7.872 0.5604 0.5793
## 3     21 147.49  5    13.919 0.3964 0.8457
```

The p-values are highs. Which means that the best model is the first one, and also means that the weight, acceleration and transmission of the car have a significant impact on it's MPG.

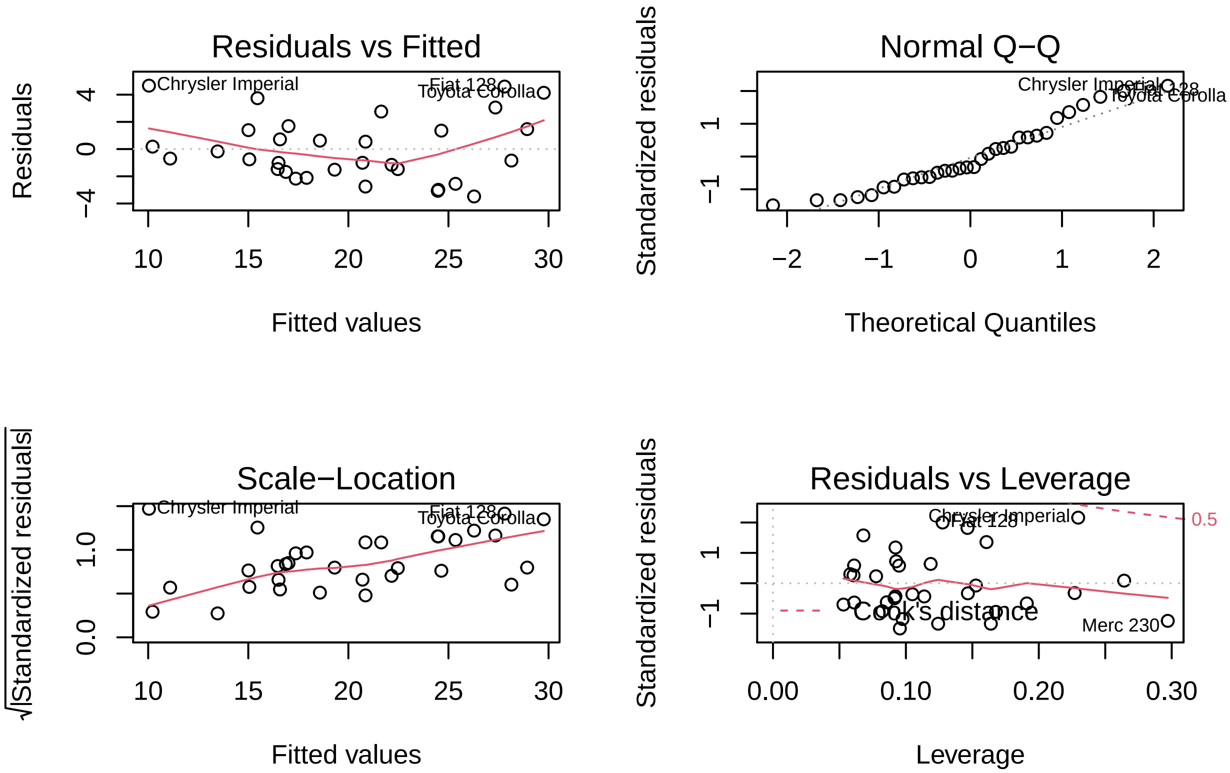
Residuals

```
r <- resid(fit)

d <- density(r)

plot(d)
```



The residuals are normally distributed.

```
par(mfrow=c(2,2))
plot(fit, 1)
plot(fit, 2)
plot(fit, 3)
plot(fit, 5)
```



We have some outliers in the dataset, but they don't have high leverage or influence. So we don't need to omit them.

Conclusion

To conclude, we assume that there is a significant association between Miles Per Galon and Transmission, and a manual transmission car has a fuel efficiency of 2.54 MPG higher than that of automatic transmission cars.