Regression Models Course Project

Mario Previdi

26 ottobre 2018

Executive summary

In this project is requested to investigate, considering a dataset of a collection of cars (mtcars), the relationship between MPG variable and the automatic / manual tranmission. Then, at the end, triyng to quantify whether an automatic or manual transmission is better for fuel consumption (the less obiovusly).

The main questions addressed in this project are:

1. "Is an automatic or manual transmission better for MPG?"
2. "Quantify the MPG difference between automatic and manual transmissions"

Note to the reader: This file will be made available on GitHub repo at my personal page address: https://github.com/MarioPre.

First of all it could be easy to compare the average MPG for either an automatic and a manual transmission car; it is ok but also wold like to verify whther or not other variables included in the dataset may have a correlation with the MPG in order to indlude them into the analysis. The analysis has benn conducted, as stated before, first evaluating the average MPG for both manual and automatic transmission, as you can see from the paragraph "Simple approach"; it results a value of average MPG with manual transmission higher than you can obtain using an automatic transmission, as we can intuitively expect if we are experienced drivers.

In the next paragraph, "Regression analysis" there is the strategy, actually as simple as possible, to understand which variables to include in our model based on both empirical considerations and ANOVA function applied to the eligible models.

Once identified the model there is the paragraph related to residual evaluation, "Residuals", in order to make some considerations about the model residual values and test its values to take the model as appropriate to our investigation-

In the end, in the paragraph called "Conclusions", in which are reported the hints to the questions reported in point one and two.

Simple approach

First of all let's show how the dataset is organized; it is composed by 32 rows and 11 columns (cars, attributes). The main attribute we are interested has been categorized indicating with the label "Automatic" the "0" as automatic transmission and "Manual" the "1" as manual transmission. Head function will display the first entries of the dataset showing values contained into.

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.4.4
```

```
require(datasets)
data(mtcars)
knitr::opts_chunk$set(echo = TRUE)
mtcars$am   <- factor(mtcars$am,labels=c("Automatic","Manual"))
dim(mtcars)
```

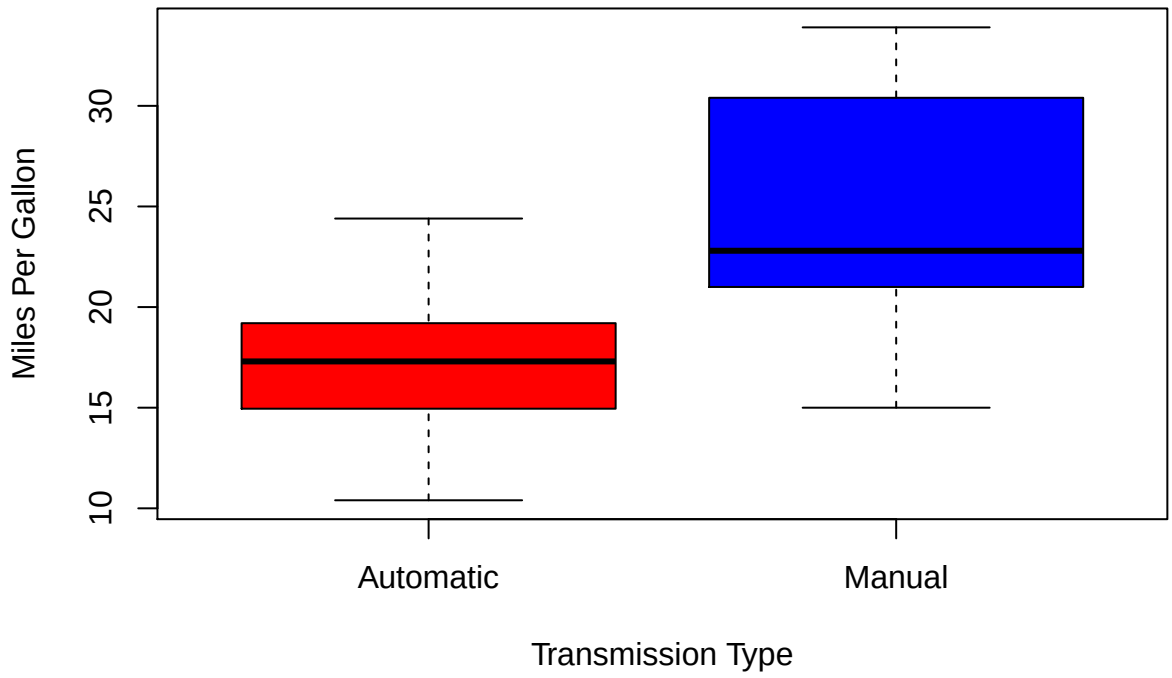
```
## [1] 32 11
```

```
head(mtcars)
```

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

The following figure shows immediately that, as experience may suggests, manual transmission allows to cover an average amount of miles more than the automatic transmission.

```
boxplot(mpg ~ am, data = mtcars, col = (c("red","blue")), ylab = "Miles Per Gallon", xlab = "Transmissi
```



Regression analysis

Now it is time to identify the 'best' model which describes our system. First of all we start with the complete model and will derive some considerations

```r
model0 <- lm(mpg ~ ., data = mtcars)
summary(model0)
```

```
##
## 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
```

At a first glance results showed above suggests that the variables which have a negative impact over the mpg variable are: cyl (number of cylinders), hp (horsepower), wt (weight) and carb (numbers of carburators). Those variables also find some accordance to common experience: miles per gallon actually may decrease as cylynders, horse power, weight or numbers of carburators increases. So, in the next section we gradually increase the number of variables other the am (automatic) one and then with the ANOVA function we investigate which variables take into account in the model.

```r
model1 <- lm(mpg ~ am, data = mtcars)
model2 <- update(model1, mpg ~ am + cyl)
model3 <- update(model2, mpg ~ am + cyl + hp)
model4 <- update(model3, mpg ~ am + cyl + hp + wt)
model5 <- update(model4, mpg ~ am + cyl + hp + wt + carb)
anova(model1,model2,model3,model4,model5)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + cyl
## Model 3: mpg ~ am + cyl + hp
## Model 4: mpg ~ am + cyl + hp + wt
## Model 5: mpg ~ am + cyl + hp + wt + carb
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 720.90
## 2     29 271.36  1    449.53 71.8295 5.895e-09 ***
```

```
## 3     28 220.55  1     50.81  8.1186  0.008458 **
## 4     27 170.00  1     50.56  8.0780  0.008602 **
## 5     26 162.72  1      7.28  1.1633  0.290678
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The P-values are for a test of whether all of the new variables are all zero or not (i.e. whether or not they're necessary). So this model would conclude that all of the added Model4 terms are necessary over the other models. The only one we can discard is the final variable in model5: carb. So, unless there were some other compelling reasons, we'd select Model4.

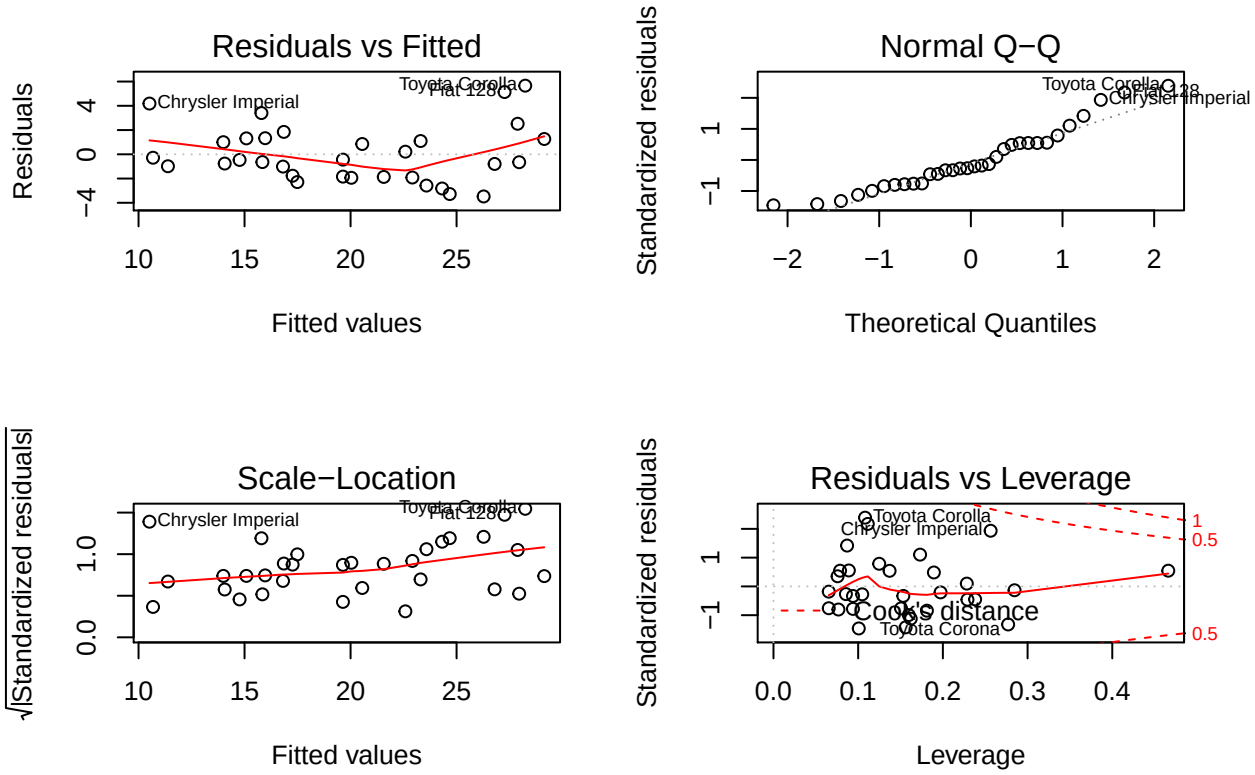
```
summary(model4)
```

```
##
## Call:
## lm(formula = mpg ~ am + cyl + hp + wt, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4765 -1.8471 -0.5544  1.2758  5.6608
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 36.14654    3.10478  11.642 4.94e-12 ***
## amManual     1.47805    1.44115   1.026   0.3142
## cyl         -0.74516    0.58279  -1.279   0.2119
## hp          -0.02495    0.01365  -1.828   0.0786 .
## wt          -2.60648    0.91984  -2.834   0.0086 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.509 on 27 degrees of freedom
## Multiple R-squared:  0.849,  Adjusted R-squared:  0.8267
## F-statistic: 37.96 on 4 and 27 DF,  p-value: 1.025e-10
```

Residuals

In this paragraph some consideration are conducted for what concerns model residuals.

```
par(mfrow = c(2, 2))
plot(model4)
```



From the above plots following observations derives: The "Residuals vs Fitted" graph shows residuals values randomy scattered indicating an indipendence behaviour. Then Normal Q-Q plot is composed by the points that fall on the line indicating the positive correlation between variables identified.The Scale-Location plot indicates a quasi constant variance; in the end the Residual vs Leverage graph shows the absence of points that may introduce high bias in our analysis.

Conclusion

Based on the observations we can conclude the following:

- Cars with Manual transmission can cover more MPG compared to cars with Automatic transmission. 1.4 miles adjusted by cyl, hp and wt.
- MPG will decrease by 2.6 (adjusted by cyl, hp and am) for every 1000 lb increase in wt.
- MPG will decrease by 0.74 (adjusted by wt, hp and am) increasing cylinders.
- MPG will decrease by 0.02 (adjusted by cyl, wt and am) for every hp increase.