Motor Trend Car Road Tests - Effects of transmission on MPG

Tanguy Levent

8 mars 2017

Github account for the scripts : https://github.com/TanguyLevent

Synopsis

The data was extracted from the 1974 Motor Trend US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973–74 models). We are going to analyze and select the predictors of our data set by AIC method (Akaike Information Criterion) to regressing the fuel consumption - unit *miles per gallon (MPG)*.

The questions of our researches are :

- Is an automatic or manual transmission better for MPG?
- Quantifying how different is the MPG between automatic and manual transmissions?

Data Processing

Please install the R package **PerformanceAnalytics** if you don't have it : `install.packages("PerformanceAnalytics")`. Then call `mtcars` data set, verify there are no missing values and compute basic summaries to have an idea of the variables inside.

```
library(dplyr)
library(PerformanceAnalytics)

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

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

```
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)
```

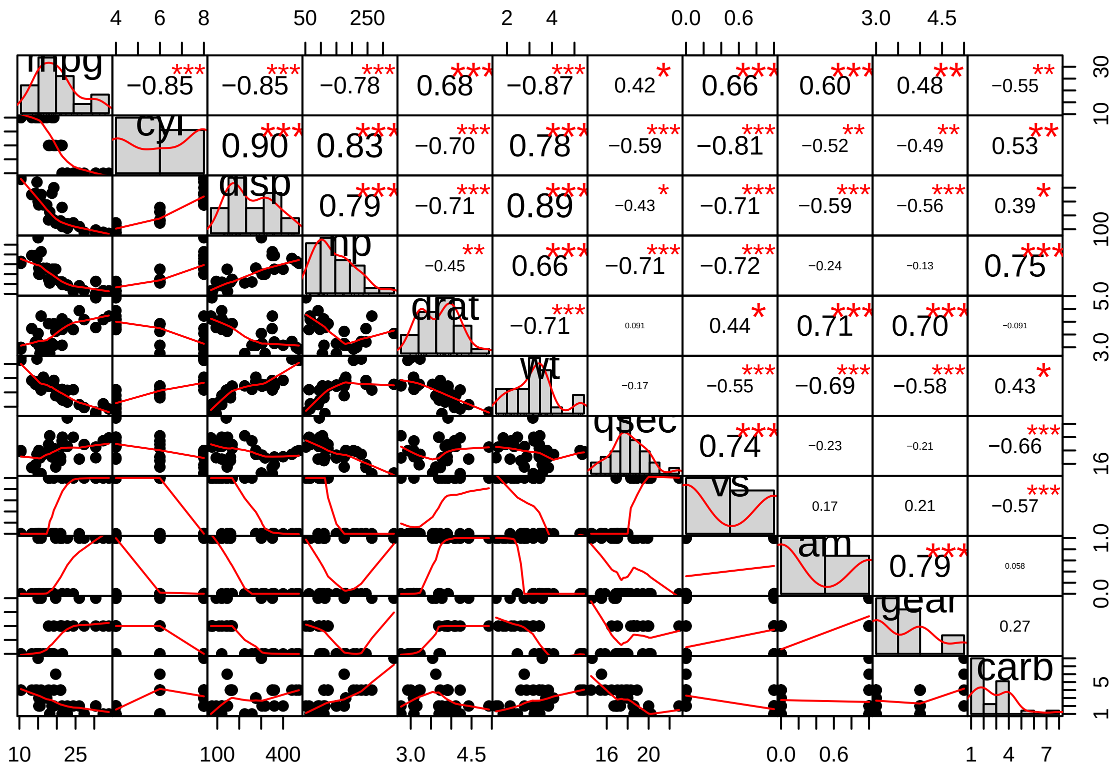
```
## 'data.frame':    32 obs. of  11 variables:
##  $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
##  $ disp: num  160 160 108 258 360 ...
##  $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
```

```
##  $ 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 ...
```

Looking at the data, we find our 10 automobile performances and 32 different automobiles.
Please find the explanation of the variables on the original documentation website: https://stat.ethz.ch/R-manual/R-devel/library/datasets/html/mtcars.html or use the command **?mtcars**.

Exploratory analysis

First I used to plot a chart correlation in order to have a general view of our variables and the link they have with each other.



Conspicuously there are variables which are related with the response `mpg` :

- `cyl` which is in fact a factor variable
- `disp`
- `hp`
- `drat`
- `wt`
- `vs` which is in fact a factor variable
- `am` which is in fact a factor variable

To finish this part, we arrange our dataset with all the informations that we have processing before.

```
mdata <- mtcars %>% select(mpg,cyl,disp,hp,drat,wt,vs,am)
mdata$cyl <- factor(mdata$cyl)
mdata$vs <- factor(mdata$vs)
mdata$am = factor(mdata$am, labels = c("Automatic","Manual"))
```

Statistical inference

The purpose of our analyze is to focus on the relationship between mpg and am. The first way to look at it is to perform a hypothesis test by test the *null hypothesis* :

1. H_0 : There is no relationship between MPG and the transmission (am)

versus the *alternative hypothesis* :

2. H_α : There is some relationship between MPG and the transmission (am)

Below the computation of the *t-statistic* and his associate result :

```
test$p.value
```

```
## [1] 0.001373638
```

```
test$estimate
```

```
## mean in group Automatic    mean in group Manual
##                17.14737                24.39231
```

The *t table* with $n = 32$ observations give me a cutoffs around 2.5%. Here we have a p-value equals to 0.1 %, largely small enough to *reject the null hypothesis* and assure there is an association between the predictor `am` and the response `mpg`. The result of the mean tells us that automatic is better for the fuel consumption than manual transmission.

Regression linear

Once we have rejected the null hypothesis in favor of the alternative hypothesis, it is natural to want to quantify *the extent to which the model fits the data* with the Residual Standard Error (RSE ou ϵ) even if it not always clear what constitutes a good RSE. Hence we also use R^2 statistics provides an alternative measure of fit. It takes the form of a *proportion* - the proportion of variability that can be explained using X. A number near 0 indicates that the regression did not explain much of the variability in the response.

Mathematically the linear relationship can be write as :

$$MPG \approx \beta_0 + \beta_1 \times am + \epsilon$$

where ϵ is the error term. Let's write the linear regression r code :

```
mpg.lm.am <- lm(mpg ~ am, mdata)
```

that gives us the RSE and R^2 :

```
round(sigma(mpg.lm.am),2)
```

```
## [1] 4.9
```

```
round(summary(mpg.lm.am)$adj.r.squared,3)
```

```
## [1] 0.338
```

In other words, actual mpg deviate from the true regression line by approximately $4,9$ on average and 34% variation of the fuel consumption is explained by the transmission - that is not enough.

Previously we have list the variables related to `mpg`. Let's see if we can draw a better model by adding one or more of these variables. In order to not testing the 127 combinaisons possible one by one, I use the `step` function which is based on AIC method and make the work for us.

```
mpg.lm.global <- lm(mpg ~ ., mdata)
mpg.lm.selection <- step(mpg.lm.global, direction = "both", trace=FALSE)
mpg.lm.selection$call
```

```
## lm(formula = mpg ~ cyl + hp + wt + am, data = mdata)
```

The powerful of this function gives us instantly the best model to fit the regression. Now we write the RSE and R^2 of our new multiple linear regression model fit by AIC method.

```
round(sigma(mpg.lm.selection),2)
```

```
## [1] 2.41
```

```
round(summary(mpg.lm.selection)$adj.r.squared,3)
```

```
## [1] 0.84
```

It's bear no comparison with the previous simple linear regression. The RSE is halved and the R^2 imply that now 84% of our predictors explained the variation of our response `mpg`. It's quite well! To finish this part of the data set analysis we are going to use the `anova` function which performs a hypothesis test comparing the two models. The null hypothesis is that the two models fit the data equally well, and the alternative hypothesis is that the full model is superior.

```
anova(mpg.lm.am,mpg.lm.selection)
```

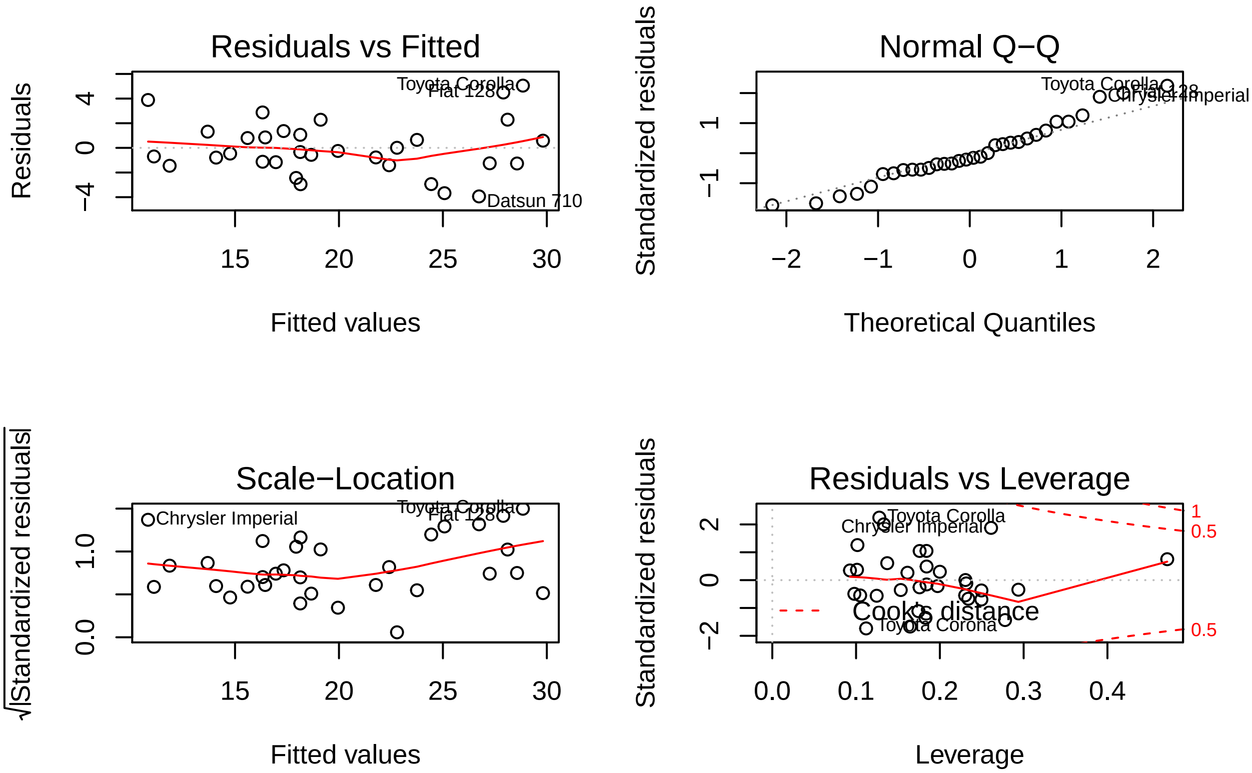
```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ cyl + hp + wt + am
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     30 720.90
## 2     26 151.03  4    569.87 24.527 1.688e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Here we have a F-statistic equals to $24,5$ and a p-value near to zero that confirm an improvement in the model fit by adding more variables. Now the Mathematical linear relationship can be write as :

$$MPG \approx \beta_0 + \beta_1 \times am + \beta_2 \times cyl + \beta_3 \times hp + \beta_4 \times wt + \epsilon$$

Potential Problems : Residuals

Residuals are leftover of the outcome variable after fitting a model (predictors) to data and they could reveal unexplained patterns in the data by the fitted model. Using this information, not only could you check if linear regression assumptions are met, but you could improve your model in an exploratory way. To realize the diagnostic we need to use the `plot(model)` function which shows us four different ways to understand our residuals and/or fitting.



1. Non linearity of the Data : Residuals vs Fitted

At the top left in the figures above, the plot shows no discernible pattern and the residuals no exhibit a clear U-shape which provide a strong indication of well linearity in the data and our model improve the fit.

2. Non-constant Variance of Error Terms : Residuals vs Fitted

At the top left in the figures above, the plot shows no heteroscedasticity from no presence of a *funnel shope* in the residual plot. If we have found this kind of shape then we would have tried to solve the problem by using a log or a square root on the response.

3. Non Normal Distribution of the Residual : Normal Q-Q

At the top right in the figures above, the plot shows no residuals deviate so much of the straight dashed line. Hence the residuals follow a normal distribution.

3. Outliers : Scale-Location

At the bottom left in the figures above, the plot shows the *studentized residuals* to check the assumption of homoscedasticity computed by dividing each residual by its estimated standard error. Observations whose

studentized residuals are greater than 3 in absolute value are possible outliers. However, care should be taken, since an outlier may instead indicate a deficiency with the model, such as a missing predictor.
Here the scale-location plot confirms the constant variance assumption, as the points are randomly distributed and no outliers in the observations.

4. High Leverage Points : Residuals vs Leverage

We are just looking at the plot to confirm or not outliers can influence our regression. As all the values fall under the Cook's distance and the leverage value are low we can assume that no obervations influence the fit regression.

Conclusion

Related to our previous analyzes we have conclude that the transmission influence the fuel consumption and an automatic is better than a manual transmission. But tansmission is not the only one predictor which permit to fit a linear regression. Hence we have fit a multiple linear regression and have found with the AIC method that the number of cylinders, the gross horsepower, the weight and the transmission are the best predictors to fitting linear regression with a very small p-value and an important R^2 of 84%. Finally our last researches don't gives us enough evidences that outliers are included inside the observations.