Johns Hopkins - Regression Models Course Project

Benjamin Berhault

September 26, 2015

Mission

Looking at a data set of a collection of cars, 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

URL

- Coursera - Johns Hopkins : Regression Models (https://class.coursera.org/regmods-032/)
- Code source can be found here : github.com/benjamin-berhault/regression-models (https://github.com/benjamin-berhault/regression-models)

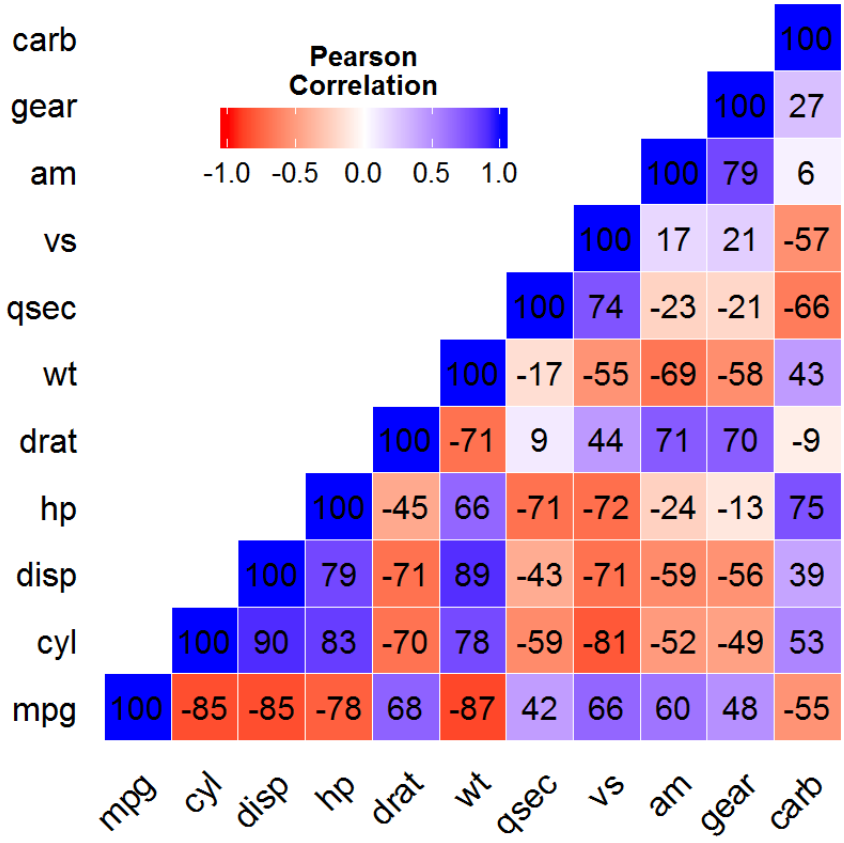
Parameters :

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

Take a look at what the datasets consists of

```
##                   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
```

We are interested by exploring the relationship between variables. For this purpose, we compute the correlation matrix.



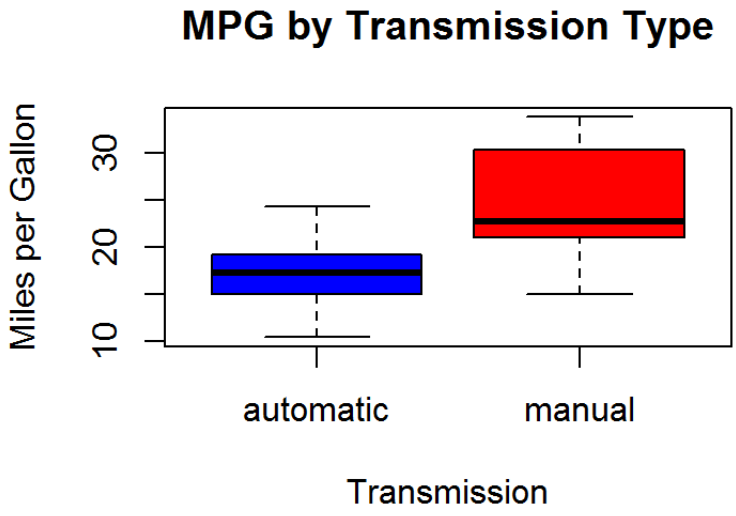
We are especially interested in exploring the relationship between miles per gallon (MPG) and the other parameters. For this purpose we compute the absolute correlation to highlight parameters that best explain the MPG parameter.

```
##   wt  cyl disp   hp drat   vs   am carb gear qsec
## 0.87 0.85 0.85 0.78 0.68 0.66 0.60 0.55 0.48 0.42
```

wt, **cyl**, **disp** parameters best explain the **MPG** parameter. Those variables are quite anticorrelated with **MPG** meaning that the miles driven for a given quantity of fuel, tend to decrease when the weight, the number of cylinders, the displacement size of an engine increase.

Question 1: Is an automatic or manual transmission better for MPG?

Let’s take a look at the consumption distribution (or MPG distribution) by transmission system :



We clearly see a difference between automatic and manual distribution systems. Manual distribution systems are mostly less fuel consuming.

But we want assert that statistically. For this purpose we implement a t-test in the second part of our analysis.

Question 2 : Quantify the MPG difference between automatic and manual transmissions.

Two samples t-test on the MPG parameter distinguishing auto vs manual systems.

Null hypothesis: There is no difference between MPG means for automatic and manual transmissions

```
##
##  Welch Two Sample t-test
##
## data:  mtcars$mpg by mtcars$am
## 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 in group 0 mean in group 1
##        17.14737        24.39231
```

We get a p-value of 0.00137, so we reject the null hypothesis. We are 99.86% confident that the mean of both transmissions are significantly different.

And the average miles per gallon for Manual Transmission is 24.39 which is 7.24 higher than the average miles of Automatic Transmission.

The boxplot with respect to the Transmission show significant overlap, indicating that it may not be the best predictor of the MPG. To determine the best predictor, further analysis needs to be done.

Appendix : Uncertainty in the analysis

First of all, lets take a look to what extent a model only based on the **am** parameter can explain in regard to MPG consumption.

R-squared (mpg~am)	R-squared adjusted (mpg~am)
35.98%	**33.85**%

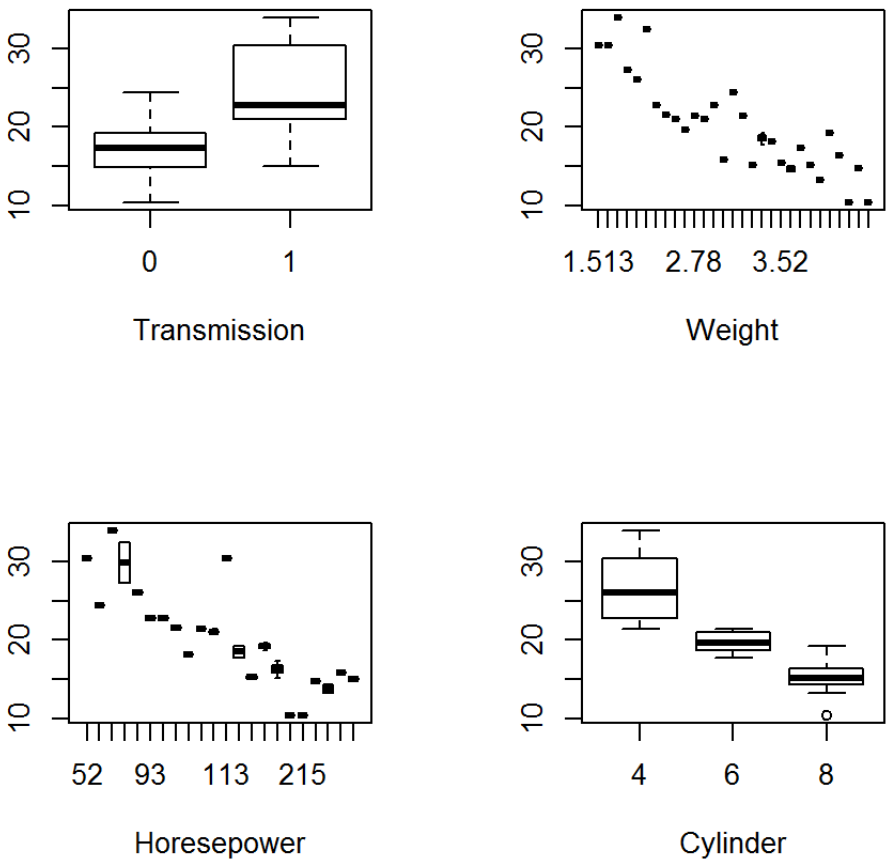
Looking for the best predictive model :

To determine which variables to include in our model and to avoid multi-colinearity issue, we used an R stepwise regression function. This function adds and removes independent variables to the model until it finds the combination of independent variables minimizing the Akaike Information Criterion (AIK : measures the relative quality of statistical models).

```
##
## Call:
## lm(formula = mpg ~ cyl + hp + wt + am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9387 -1.2560 -0.4013  1.1253  5.0513
##
## Coefficients:
##             Estimate Std. Error t value          Pr(>|t|)
## (Intercept) 33.70832    2.60489  12.940 0.000000000000773 ***
## cyl6        -3.03134    1.40728  -2.154           0.04068 *
## cyl8        -2.16368    2.28425  -0.947           0.35225
## hp          -0.03211    0.01369  -2.345           0.02693 *
## wt          -2.49683    0.88559  -2.819           0.00908 **
## am           1.80921    1.39630   1.296           0.20646
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.41 on 26 degrees of freedom
## Multiple R-squared:  0.8659, Adjusted R-squared:  0.8401
## F-statistic: 33.57 on 5 and 26 DF,  p-value: 0.0000000001506
```

After computation our best model get an R-squared adjusted of **84.01%**. **wt**, **hp** and **cyl** are the variables that best explain miles per gallon consumption if we look at the asterix marks.

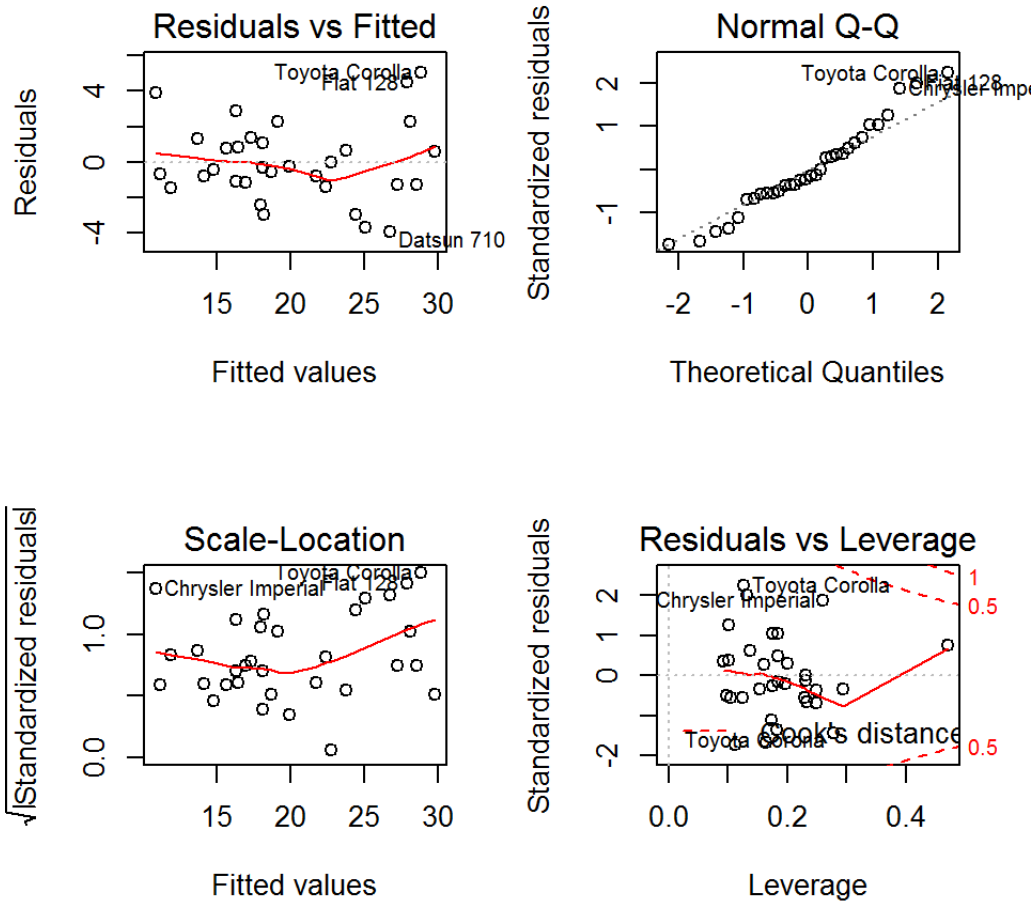
Lets boxplot those parameters :



Here, we observe that the **wt**, **hp** and **cyl** predictors are more significant than **am**. And with a p-value of 0.2, we conclude that **am** seems to have some kind of influence over **"MPG"** but is not so significant that it could be in the first part of our analysis.

According to our best predictive model the impact of having a manual distribution system only enhance by **1.80921 Miles per Gallon** the efficiency of a car in comparison to automatic distribution system.

Residual analysis



In the residual plot, we don't see any pattern that causes us to believe that the Fuel consumption could be explained more by any other predictors available in the dataset.