Regression Models Project

Carlos Alberto Guevara Díez

18 de septiembre de 2015

Executive Summary

For the purpose of this analisys I´m using the mtcars dataset available in R, for further information about the dataset please refer to its help file. The goal of this analysis is to answer the questions: "Is an automatic or manual transmission better for MPG?" and "Quantify the MPG difference between automatic and manual transmissions". As a result the analysis will prove that manual transmission is better than automatic, and that that in a car with average horse power, and a 4 cylinder, straight engine, a manual transmission that expected difference is about 5.2 miles per gallon.

Exploratory Data Analysis

The first thing to do is pre-process the dataset and add some descriptive an case transformations, this will help to standardize the data and make the exploratory analisys easier.

Whit the pre procesed data now I can make a basic exploratory analysis and a summary exploration explained with the **Fig. 1 of the appendix "Exploratory BoxPlot"**, with this I can easily answer to the first question that manual transmission is better than automatic. In addition a t-test have been made comparing the mean between two transmission data groups (manual and auto), the confidence interval (95%) does not contain zero (-11.28,-3.21) and p-value is greater than 0.005. Then, it can be concluded that the average consumption, in miles per gallon, with automatic transmission is higher than the manual transmission. It is possible to quantify the MPG difference between automatic and manual transmissions: 7.24 mpg greater subtracting means. Additionally it is concluded that there are other variables correlated with mpg according to the graph analysis in the **Fig. 2 of the appendix "Pair analisys"**.

```r
datAuto <- mtcars$mpg[mtcars$am == "automatic transmission"]
datManual <- mtcars$mpg[mtcars$am == "manual transmission"];
t.test(datAuto, datManual, paired = FALSE, alternative="two.sided", var.equal=FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  datAuto and datManual
## 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
```

Prediction Models

To obtain different approaches of the solution, several models will be analyzed in this section. To evaluate the accuraccy of each model I will use: The adjusted R-squared, the interpretablility of the model (e.g. how much sense does it make to include the included variables), and the signifigance of the term of interest, the trasmission type variable.

For the first approach I´ve taken only fuel eficciency and transmission type:

```
## $coefficients
##                        Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)           17.147368   1.124603 15.247492 1.133983e-15
## ammanual transmission  7.244939   1.764422  4.106127 2.850207e-04
##
## $adj.r.squared
## [1] 0.3384589
```

This is the easiest model and the results are simple to explain, the Transmission variable is significant, nevertheless as we dod not take into account other variables the adjustes R-squared is low.

The second approach involves all the variables in the mtcars dataset, for space limitations I'm only showing the R-squared result that is bigger than the first model, nevertheless, the most of the time its not a good strategy to include all the variables because some of them may be representing the same characteristics or may add noise to the model.

```
## $adj.r.squared
## [1] 0.7790215
```

To add accuraccy to the results I need to explore a third model, this one is using the variables fuel economy, transmission type, horse power, number of cylinders and engine shape as variables.

```
## $coefficients
##                          Estimate Std. Error     t value     Pr(>|t|)
## (Intercept)           27.01624146 1.42041453 19.01996976 8.863125e-17
## ammanual transmission  5.16287130 1.45386237  3.55114170 1.489216e-03
## hp                    -0.04687855 0.01451486 -3.22969437 3.346750e-03
## cyl6                  -2.65245486 1.79590505 -1.47694604 1.517011e-01
## cyl8                  -0.27710473 3.48664077 -0.07947613 9.372625e-01
## vsv engine            -2.56902830 1.94243080 -1.32258421 1.974901e-01
##
## $adj.r.squared
## [1] 0.8043595
```

In this model I'm showing that the selected variables have an obvious effect on fuel economy (Refer to Fig. 3 of the appendix "Effects of included variables in fuel economy"). This model also shows a higher adjusted R-squared than the other two models.

Conclusions

The final model (and, in fact, all of the models included) show that manual transmissions had better fuel efficiency in 1974. The model has a positive coefficient for the "Manual Transmission" term. This matches what we expected from the initial exploratory analysis.

The third and final model show us that if a car changes from the base case - automatic transmission, average horse power, 4 cylinders, and a straight engine - and switches to a manual transmission, it can expect to gain 5.16 miles per gallon of fuel efficiency. The 95% confidence interval for this value is [2.1744144, 8.1513282]:, Fig. 4 of the appendix "Residual Plot in Final Model" shows that the behavior of the third model is adequate considering normal residuals and constant variability. The leverage is within reasonable upper limit.

Appendix

Fig. 1 .- Exploratory BoxPlot

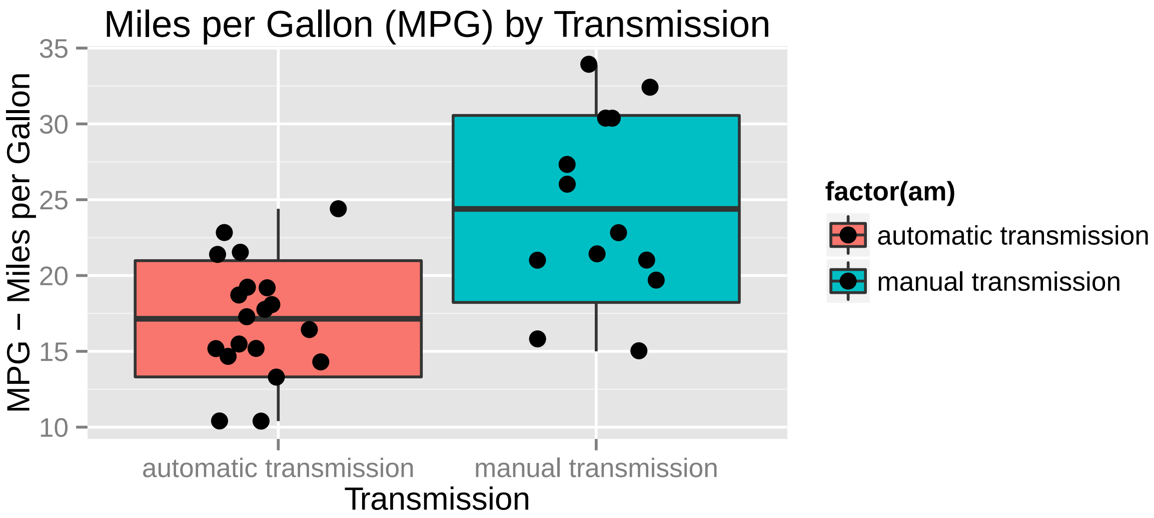


Fig. 2 .- Pair Analisys

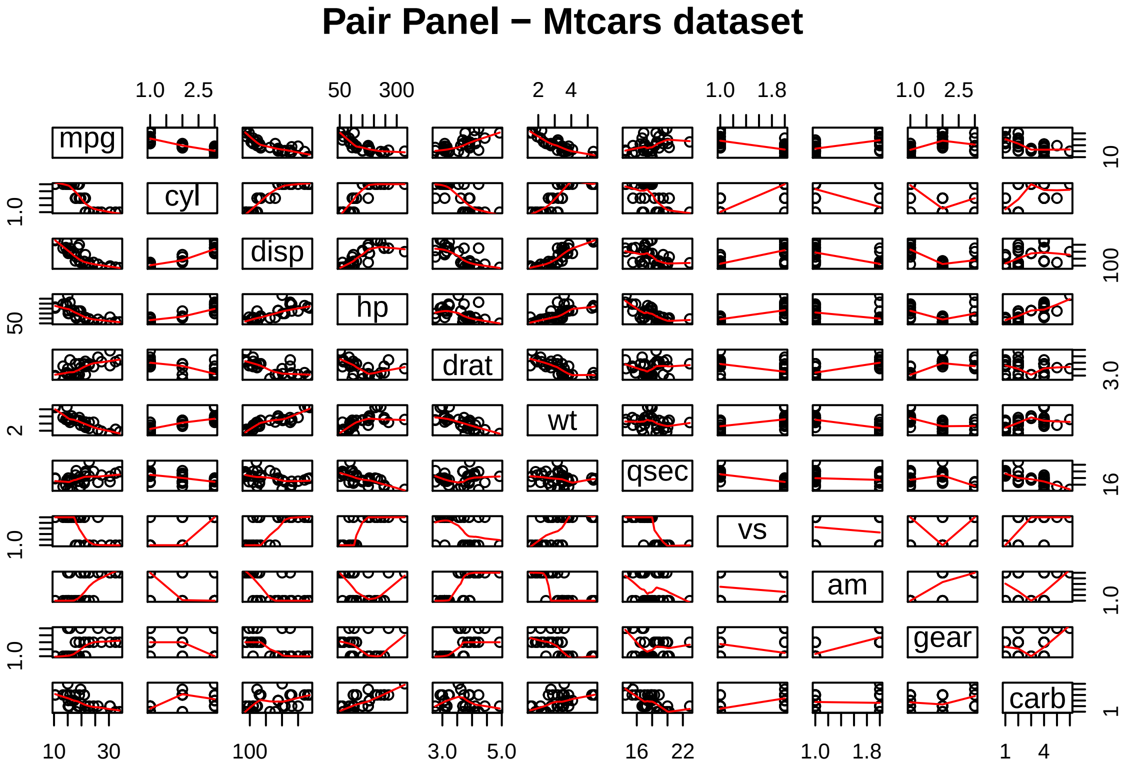
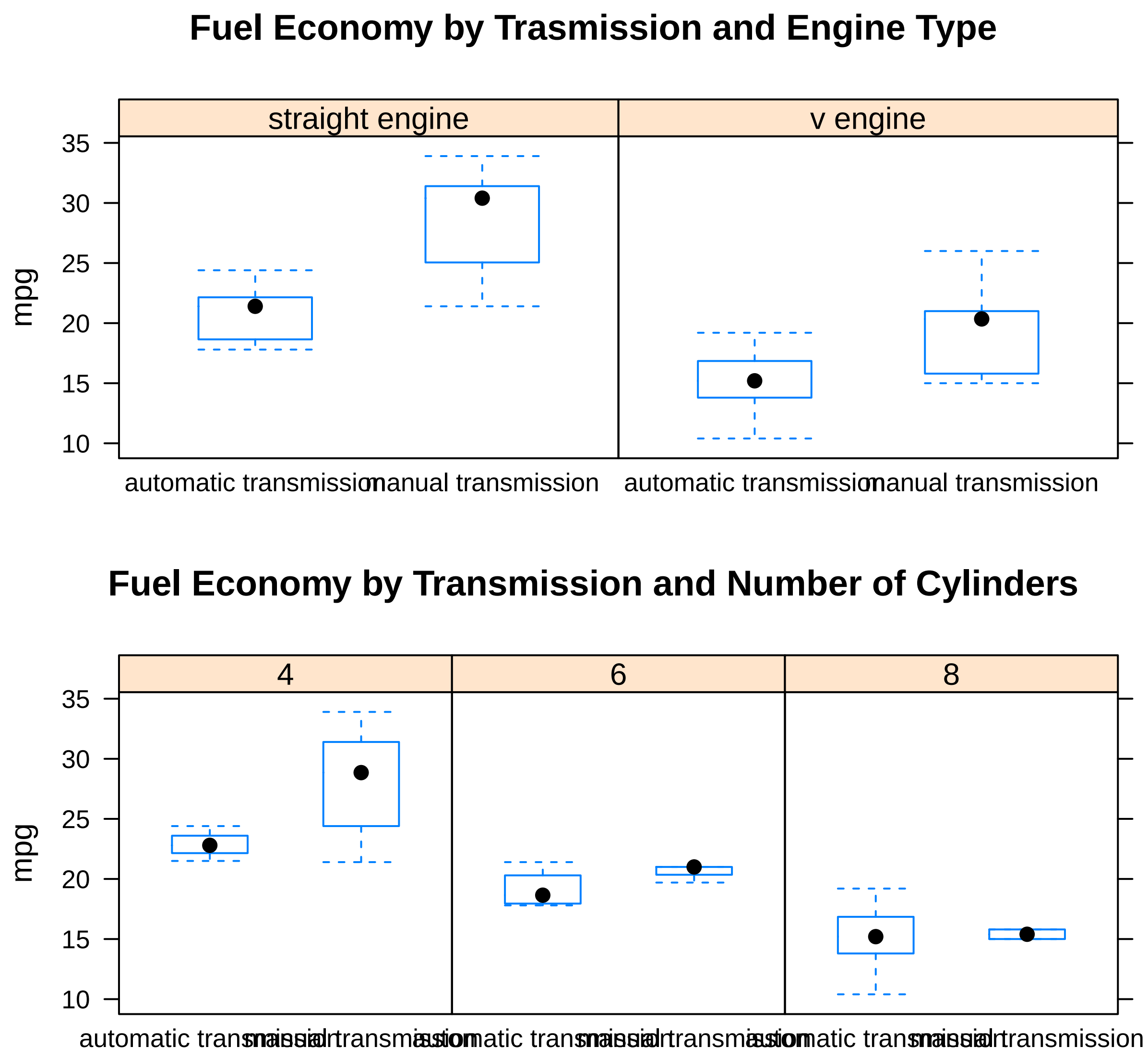


Fig. 3 .- Effects of included variables in fuel economy



Fuel Economy by Trasmission and Engine Type

Fuel Economy by Transmission and Number of Cylinders

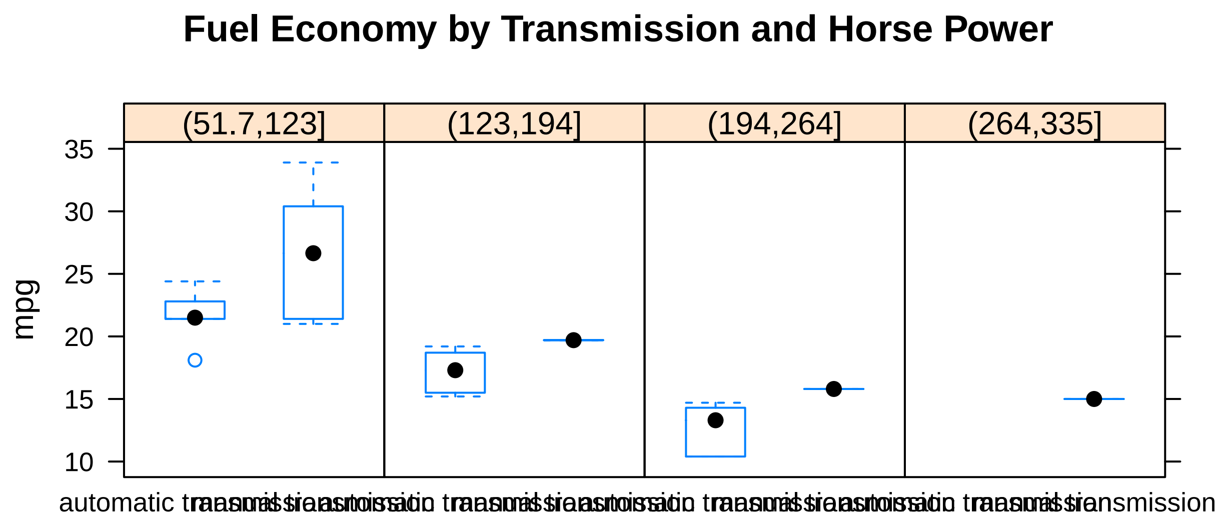
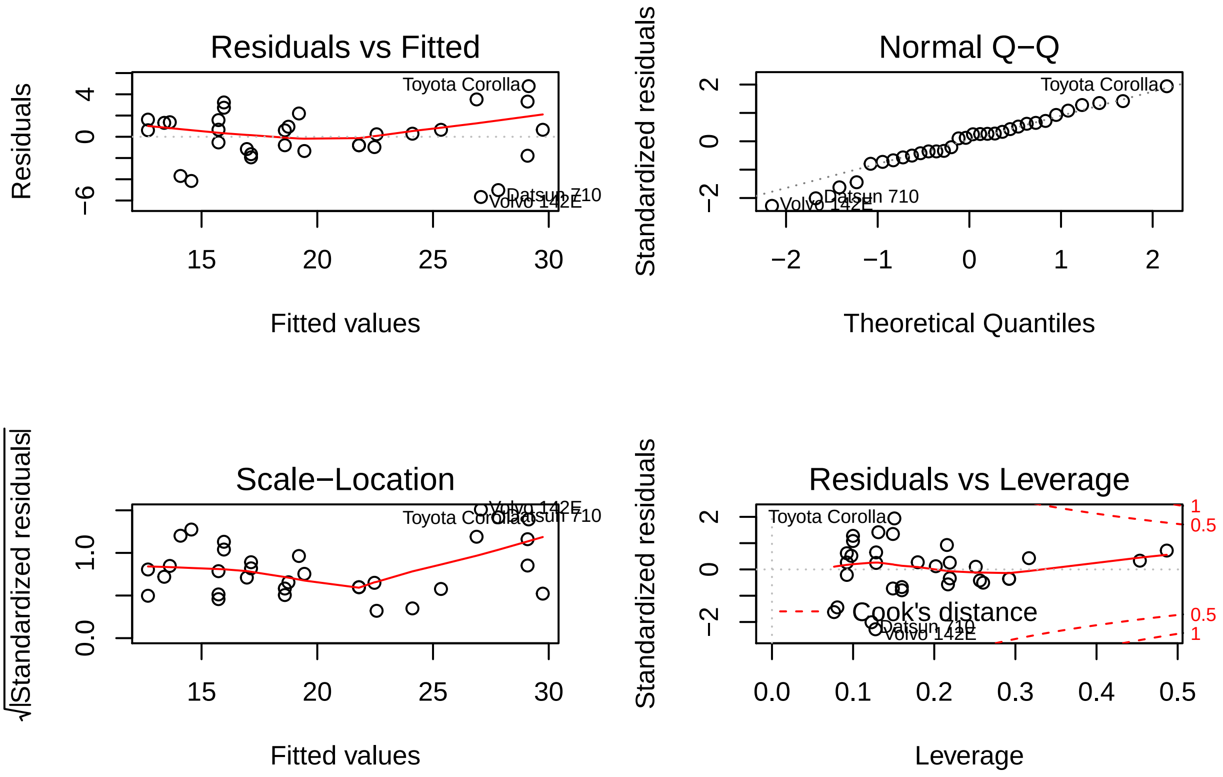


Fig. 4 .- Residual Plot in Final Model


Residuals vs Fitted
Residuals
Fitted values
Toyota Corolla
Datsun 710
Volvo 142E
Normal Q−Q
Standardized residuals
Theoretical Quantiles
Toyota Corolla
Datsun 710
Volvo 142E
Scale−Location
√|Standardized residuals|
Fitted values
Volvo 142E
Toyota Corolla
Datsun 710
Residuals vs Leverage
Standardized residuals
Leverage
Toyota Corolla
Cook's distance
Datsun 710
Volvo 142E
1
0.5
0.5
1