Motor Trend Data - Regression Models Project Report

Paulo Cardoso

February 28, 2016

Executive Summary

This project is part of Regression Models Course of the Johns Hopkins University in partnership with Coursera. And aims to demonstrate the relationship between Miles Per Gallon with other variables, with the intent to answer the following questions:

- Is an automatic or manual transmission better for MPG?
- Quantify the MPG difference between automatic and manual transmissions?

The data used in this study are from the mtcars dataset and is available in the RStudio. These data were extracted from Motor Trend Magazine US in 1974. It is composed of data such as fuel consumption and 10 aspects that comprise information about design and performace of 32 car models (model 1973-74).

Data Processing

Loading the external libraries, the data and transforming variables. As showed in the Appendix 1.

Exploratory Data Analyses

In order to understand the data set is necessary that a few metrics be presented. As presented on the Appendix 2. I)The comparison of the standard deviations of MPG by transmission type, and Levene's test indicate that the assumption of homogeneity of variance is questionable. II) Shows that on average there is a difference between the fuel efficiency depending on the transmission type. III)Here is presented the Correlation between the variables mpg, wt, qsec and am,which respectively represent: Miles per Gallon, Weight (lb/1000), 1/4 mile time and Transmission (0 = automatic, 1 = manual). III) is the graphic representation of this correlation.

Inference

In order to confirm the significance, a t-test is performed with the H0 Null Hypothesis being the case where there is NO difference between an Automatic and Manual Transmissions and the alternative Hypotesis HA being the cade where there is significant difference between Automatic and Manual Transmissions. The p-value of 0,001374 rejects the N0 Null Hypothesis.

Regression Analysis

I)The first model is compose of all variables of mtcars, as is presented in the Appendix 4. This first model has an Adjusted R-squared of 0,779 and a Residual Standard Error of 2.833 on 15 degress of freedom. However none of the coefficients are significant at 0,05 significant level. II)On this second model that is composed of

multiple regression models “foward selection and backward elimination”, as presented on the Appendix 4. This model has an Adjusted R-squared of 0,833 and a Residual Standard Error of 2.459 on 28 degrees of freedom. III)This last model it is a simple one, is composed only of “mpg” and “am” variables, as is presented by the Appendix 4. This model has an Adjusted R-squared of 0,338 and a Residual Standard Error of 4.902 on 30 degrees of freedom. IV) Here occours a comparison between the two models using anova, as presented on the Appendix 4.

Residuals Analysis, Diagnostics and Conclusion

The Graphical representation of the Residual Analysis can be observed on the Appendix 5. And what can be concluded or diagnosed is that: I)The Residuals vs. Fitted plot presents no consistent pattern, supporting the accuracy of the independence assumption. II) The Normal Q-Q plot indicates that the residuals are normally distributed because the points lie closely to the line. III)The Scale-Location plot confirms the constant variance assumption, as the points are randomly distributed. IV)The Residuals vs. Leverage argues that no outliers are present, as all values fall well within the 0.5 bands.

So what can be concluded is that taking into account the comments of the models produced above it can be concluded that the best model is the “model 2” and moreover can also be concluded that: I)When comparing the fuel consumption per mile “mpg” transmission with manual cars have better performance. To be more economic, spending less fuel. II)That “mpg”" will decrease by 2.5 (adjusted by hp, cyl, and am) for every 1000 lb increase in “wt”. III)That “mpg”" decreases negligibly with increase of “hp”. IV)If number of cylinders, “cyl”" increases from 4 to 6 and 8, “mpg”" will decrease by a factor of 3 and 2.2 respectively (adjusted by hp, wt, and am).

Appendix 1

```
# Loading libraries
library(car)
# Loading the data
data(mtcars)

mtcars$cyl <- factor(mtcars$cyl)
mtcars$vs <- factor(mtcars$vs)
mtcars$gear <- factor(mtcars$gear)
mtcars$carb <- factor(mtcars$carb)
mtcars$am <- factor(mtcars$am,labels=c('Automatic','Manual'))
```

Appendix 2

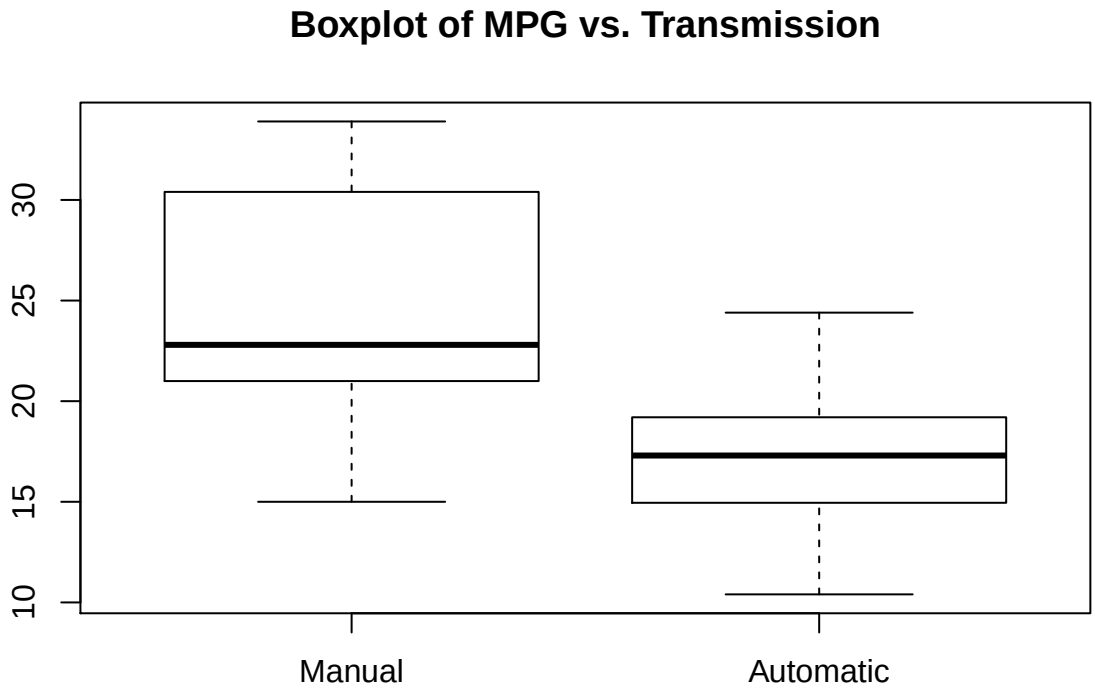
```
# I)Standard Deviation of MPG by Transmission Type
by(mtcars$mpg, mtcars$am, sd)
```

```
## mtcars$am: Automatic
## [1] 3.833966
## ------------------------------------------------------------
## mtcars$am: Manual
## [1] 6.166504
```

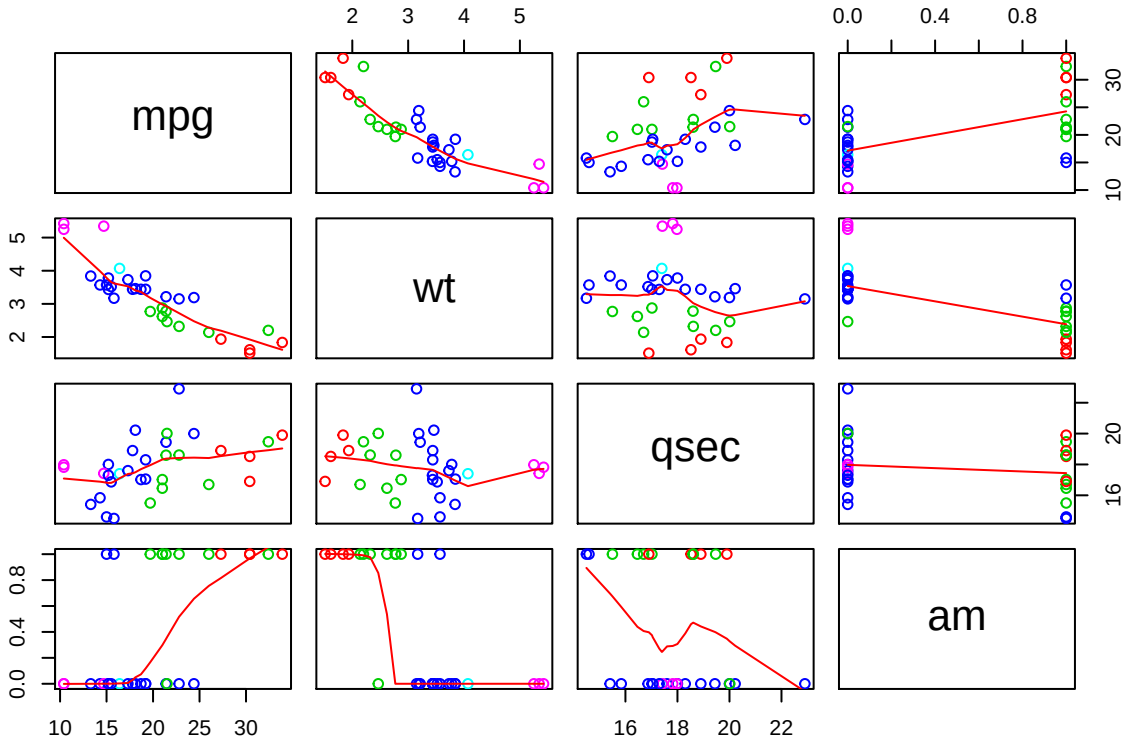
```r
# I)Levene's Test for Homogeneity of Variance
leveneTest(mpg ~ factor(am), data = mtcars)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value  Pr(>F)
## group  1  4.1876 0.04957 *
##       30
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# II)Boxplot MPG vs Transmission
boxplot(mtcars[mtcars$am == 1, ]$mpg, mtcars[mtcars$am == 0, ]$mpg, names = c("Manual", "Automatic"), ma
```




```r
# III)Comparssion plot
pairs(mtcars[, c(1, 6, 7, 9)], panel = panel.smooth, col = 9 + mtcars$wt)
```



Appendix 3

```
# T-Test
tt <- t.test(mpg ~ am, data = mtcars)
tt$p.value
```

Appendix 4

```
# I)Model 1
model1 <- lm(mpg ~ ., data=mtcars)

# II)Model 2
model2 <- step(model1, direction = "both")

# III)Model 3
model3 <- lm(mpg ~ am, data = mtcars)
```

```
# IV)Anova
anova(model3, model2)
```

```
## 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
```

Appendix 5

```
# Residual Analysis
par(mfrow=c(2, 2))
plot(model2)
```