Motor Trend - Effects of transmission on MPG

Pedro Magalhães Bernardo

June 5, 2016

Executive Summary

This report is part of a course project within the **Regression Models** course on the **Data Science Specialization** by **Johns Hopkins University** on **Coursera**. On this report we will analyze the **mtcars** data set and explore the relationship between the type of transmission (manual or automatic), among other variables, and miles per gallon (MPG), which will be our 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.

Data Pre-Processing

In this section we load the data set and perform the necessary data transformations that are needed for a better analysis. Specifically for this data set we will transform the necessary variables into factors.

```
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’))
```

Exploratory Data Analysis

In this section we explore some relationships between variables of the data set and our outcome. First we plot the relationship between all variables of the data set (see Figure 1 in the appendix). From this plot we can see a strong correlation between variables such as: **disp**, **hp**, **drat**, **wt**, and **am** with our outcome **mpg**.

Since we are actually interested in quantifying the MPG difference between automatic and manual transmissions we also plot a boxplot between **mpg** and **am** (see Figure 2 in the appendix). We can see that there is an increase in **mpg** when the transmission is **manual**.

Regression

In this section we build a linear regression using different variables as predictors and **mpg** as outcome. To find the best model we perform a stepwise selection using both forward selection and backward elimination. The code below takes care of this.

```
first_model <- lm(mpg ~., mtcars)
best_model <- step(first_model, direction = "both")
```

We can see that the best model uses the variables **cyl**, **hp** and **wt** as confounders and **am** as the independent variable.

```
best_model
```

```
##
## Call:
## lm(formula = mpg ~ cyl + hp + wt + am, data = mtcars)
##
## Coefficients:
## (Intercept)         cyl6         cyl8           hp           wt
##    33.70832     -3.03134     -2.16368     -0.03211     -2.49683
##    amManual
##     1.80921
```

We can also use **anova** to compare a base model, that only uses **am** as a predictor, and the best model that was found performing stepwise selection.

```
base_model <- lm(mpg~am, mtcars)
anova(base_model, best_model)
```

```
## 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
```

Since the **p-value** is highly significant we reject the null hypothesis that the variables **cyl**, **hp** and **wt** do not contribute to the model.

Appendix

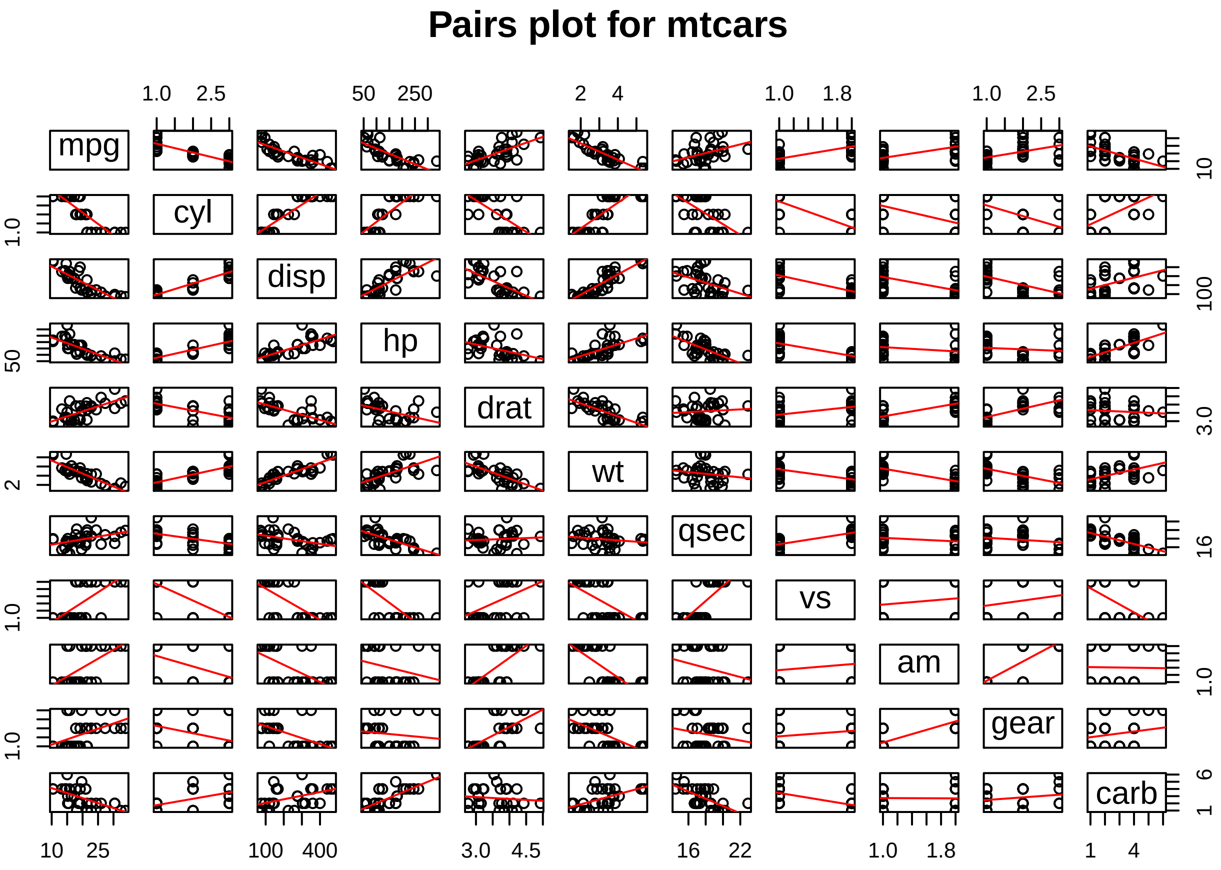

Figure 1: Pairs plot for mtcars

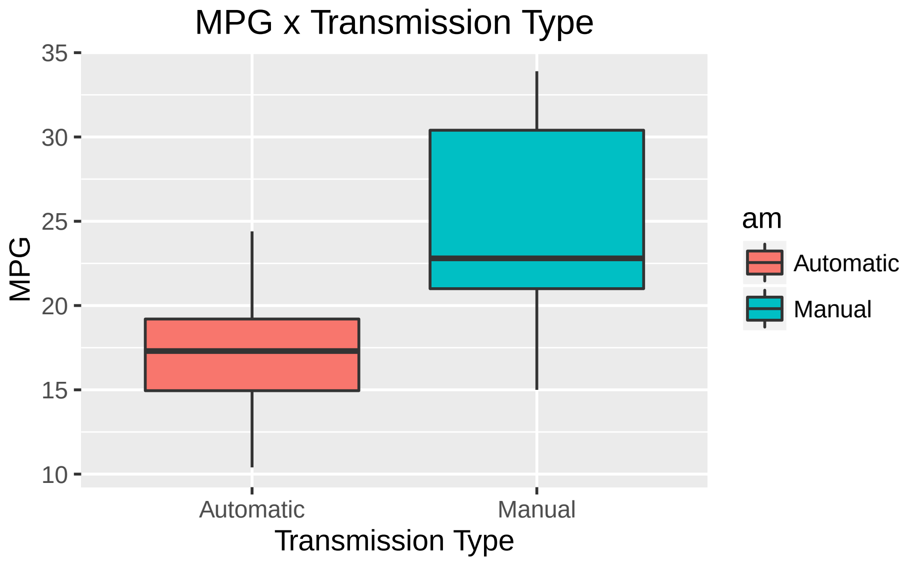

Figure 2: Boxplot between mpg and am