Regression Models

Course Project

Delermando Branquinho Filho

20 de julho de 2016

Fuel Consumption Analysis : Difference Between Automatic and Manual Transmissions

1. Executive Summary

Scope

You work for Motor Trend, a magazine about the automobile industry. Looking at a data set of a collection of cars, they are interested in exploring the relationship between a set of variables and miles per gallon (MPG) (outcome). They 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"

In order to answer these questions, we performed a very quick exploratory data analysis, and then used hypothesis testing and linear regression as methodologies to make the necessary inferences. Both simple and multivariate linear regression analysis (supported by an ANOVA of the variables to be included into the final model) have been used. Using model selection strategy, it has been found out that :

For higher MPG, manual transmission is better than automatic. In a simple linear regression model between MPG and transmission, it is observed that cars with manual transmission would travel 7.245 more miles per gallon on average than cars with automatic transmission. When using the multivariable regression analysis that includes other impacting variables (weight - wt - and quarter mile time - qsec), the adjusted model shows that manual transmission cars allowed in reality 2.936 miles per gallon more than automatic transmission (when keeping the other variables constant).

Exploratory Analysis

First, we load the data set mtcars and change some variables from numeric class to factor class.

```r
library(knitr)
library(ggplot2)
library(GGally)
```

```
## Warning: package 'GGally' was built under R version 3.3.1
```

```r
library(datasets)
library(MASS)

# Data Sample
data(mtcars)
mtcars[1:3, ]
```

```
##                 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
```

```r
dim(mtcars)
```

```
## [1] 32 11
```

```r
mtcars$cyl <- as.factor(mtcars$cyl)
mtcars$vs <- as.factor(mtcars$vs)
mtcars$am <- factor(mtcars$am)
mtcars$gear <- factor(mtcars$gear)
mtcars$carb <- factor(mtcars$carb)
attach(mtcars)
```

```
## The following object is masked from package:ggplot2:
##
##     mpg
```

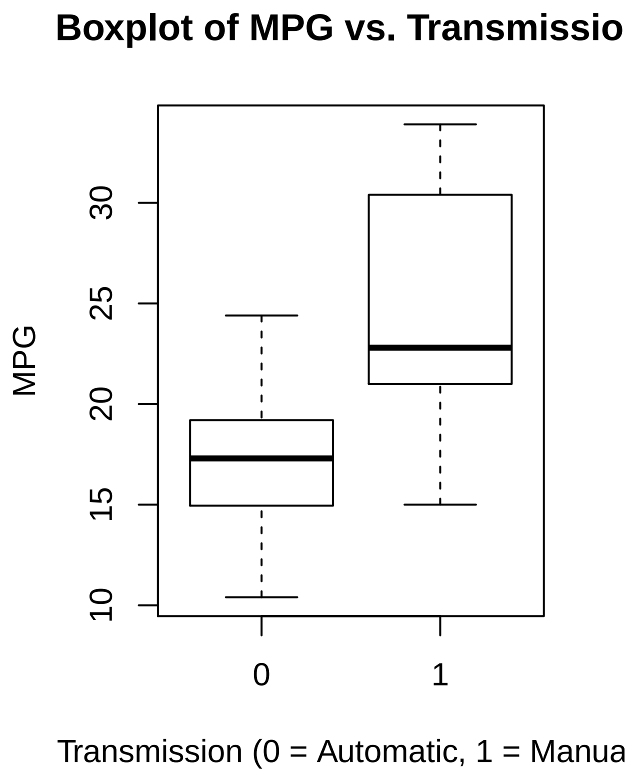
For the purpose of this analysis we use mtcars dataset which is a dataset extracted from the 1974 Motor Trend US magazine, and comprises fuel autonomy and 10 more aspects of automobile design and performance for 32 automobiles (1973-74 models). The table below shows a brief description of the variables in the dataset:

```r
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
```

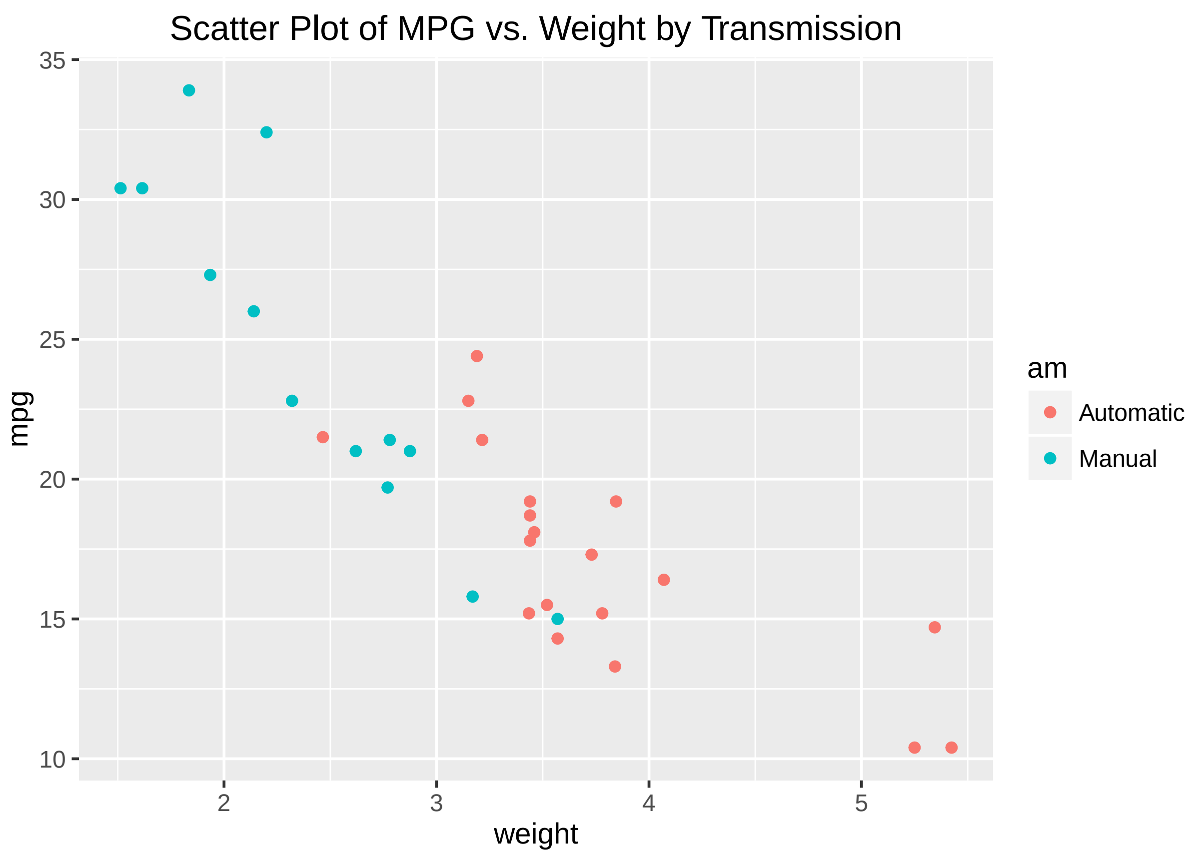
As an initial test, dependence of mpg on wt is analysed. box-scat A: Boxplot, B: Scatterplot of mpg against wt.

```r
par(mfrow=c(1,2))
boxplot(mpg ~ am, xlab="Transmission (0 = Automatic, 1 = Manual)", ylab="MPG",
        main="Boxplot of MPG vs. Transmission")
```

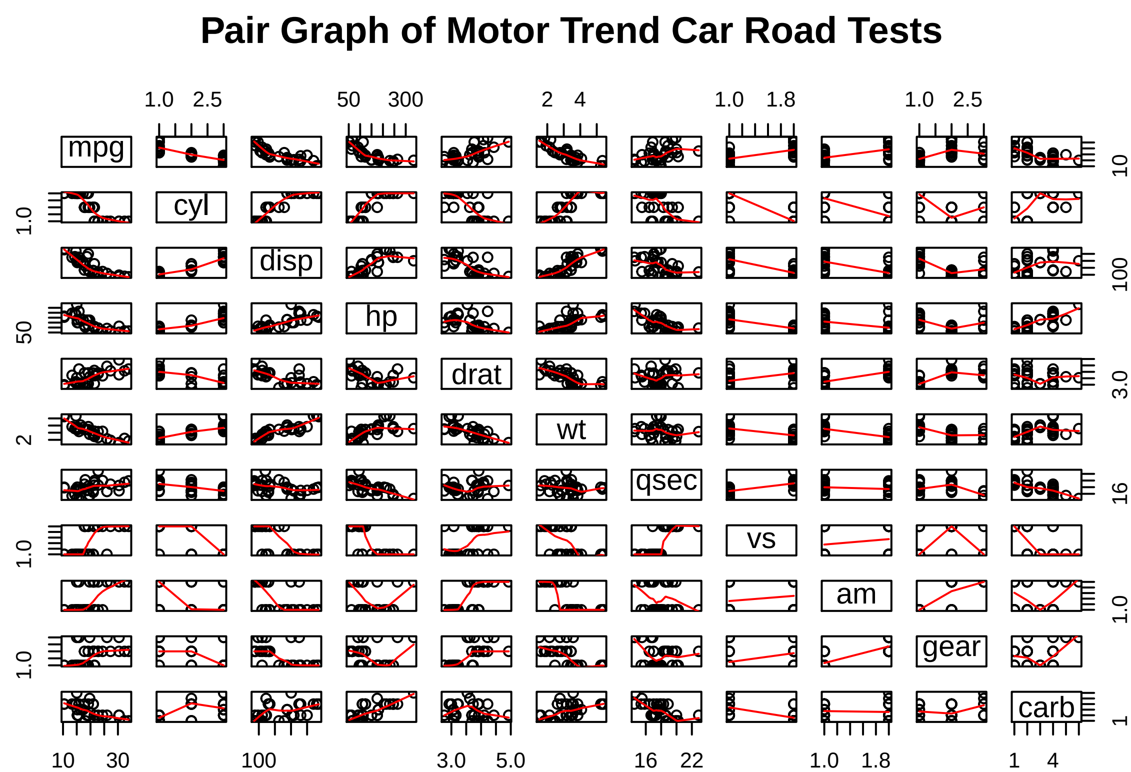


According to the boxplot, automatic cars have lower MPG (and possibly lower variance in the data). Importantly, the relationships appear linear and no outliers which could affect correlation values are identified. The only aspect which is slightly problematic is the limited dataset size (n=32). It is noted that apparently most cars with automatic transmission also are heavier which possibly confounds the observation, this would be subject to further research.

```
ggplot(mtcars, aes(x=wt, y=mpg, group=am, color=am, height=3, width=3)) + geom_point() +
scale_colour_discrete(labels=c("Automatic", "Manual")) +
xlab("weight") + ggtitle("Scatter Plot of MPG vs. Weight by Transmission")
```




```
pairs(mtcars, panel=panel.smooth, main="Pair Graph of Motor Trend Car Road Tests")
```



The Tests - t.test

In order to check for significant difference on MPG between automatic and manual transmissions (to justify further analyses) it has been performed a t Test with the data.

F Test for equal variances:

```
trAutom  <- mtcars$mpg[mtcars$am == 0]
trManual <- mtcars$mpg[mtcars$am == 1]
var.test(trAutom,trManual)
```

```
##
##  F test to compare two variances
##
## data:  trAutom and trManual
## F = 0.38656, num df = 18, denom df = 12, p-value = 0.06691
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.1243721 1.0703429
## sample estimates:
## ratio of variances
##          0.3865615
```

At this step, we make the null hypothesis as the MPG of the automatic and manual transmissions are from the same population (assuming the MPG has a normal distribution). We use the two sample T-test to show it.

```r
result <- t.test(mpg ~ am)
result$p.value
```

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

```r
result$estimate
```

```
## mean in group 0 mean in group 1
##        17.14737        24.39231
```

Since the p-value is 0.00137, we reject our null hypothesis. So, the automatic and manual transmissions are from different populations. And the mean for MPG of manual transmitted cars is about 7 more than that of automatic transmitte

Regression Analysis

First, we fit the full model as the following

```r
fullModel <- lm(mpg ~ ., data=mtcars)
summary(fullModel) # results hidden
```

```
##
## Call:
## lm(formula = mpg ~ ., data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5087 -1.3584 -0.0948  0.7745  4.6251
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 23.87913   20.06582   1.190   0.2525
## cyl6        -2.64870    3.04089  -0.871   0.3975
## cyl8        -0.33616    7.15954  -0.047   0.9632
## disp         0.03555    0.03190   1.114   0.2827
## hp          -0.07051    0.03943  -1.788   0.0939 .
## drat         1.18283    2.48348   0.476   0.6407
## wt          -4.52978    2.53875  -1.784   0.0946 .
## qsec         0.36784    0.93540   0.393   0.6997
## vs1          1.93085    2.87126   0.672   0.5115
## am1          1.21212    3.21355   0.377   0.7113
## gear4        1.11435    3.79952   0.293   0.7733
## gear5        2.52840    3.73636   0.677   0.5089
## carb2       -0.97935    2.31797  -0.423   0.6787
## carb3        2.99964    4.29355   0.699   0.4955
## carb4        1.09142    4.44962   0.245   0.8096
## carb6        4.47757    6.38406   0.701   0.4938
## carb8        7.25041    8.36057   0.867   0.3995
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 2.833 on 15 degrees of freedom
## Multiple R-squared:  0.8931, Adjusted R-squared:  0.779
## F-statistic:  7.83 on 16 and 15 DF,  p-value: 0.000124
```

This model has the Residual standard error as 2.833 on 15 degrees of freedom. And the Adjusted R-squared value is 0.779, which means that the model can explain about 78% of the variance of the MPG variable. However, none of the coefficients are significant at 0.05 significant level.

Then, we use backward selection to select some statistically significant variables.

This model is "mpg ~ wt + qsec + am". It has the Residual standard error as 2.459 on 28 degrees of freedom. And the Adjusted R-squared value is 0.8336, which means that the model can explain about 83% of the variance of the MPG variable. All of the coefficients are significant at 0.05 significant level.

The coefficients are interpreted as follows. The wt dependence of automatic cars (am = 0) is such that for every unit in wt, the MPG decreases by 5.7 units. However, given that the p-value for the dummy variable I(am*wt) is around 0.34, it is not plausible to believe that the wt dependence of manual cars is different from automatic cars. This is illustrated in a figure in the appendix where a single linear model can explain the MPG dependence of both manual and automatic cars.

Data Correlations

A first glimpse on the correlations of all the variables with MPG is shown in the table below.

Regression Analysis

a) Linear Regression

A first Linear Regression Analysis, using only MPG and transmission type (am) as variables was made to show the impact of transmission on MPG witout taking into account the other variables.

As said before, it shows a big difference in MPG favorable to manual transmission (+ 7.245 miles per gallon) when the other variables are not considered. By looking at the correlations table, it is easy to see that there are other variables also impacting on MPG and a Multivariable Regression Analysis is then performed below.

b) Multivariable Regression

Including all variables we have:

```
trMVAR <- lm(mpg ~ . , data = mtcars)
summary(trMVAR)$coefficients
```

```
##                Estimate  Std. Error     t value   Pr(>|t|)
## (Intercept) 23.87913244 20.06582026  1.19004018 0.25252548
## cyl6        -2.64869528  3.04089041 -0.87102622 0.39746642
## cyl8        -0.33616298  7.15953951 -0.04695316 0.96317000
## disp         0.03554632  0.03189920  1.11433290 0.28267339
## hp          -0.07050683  0.03942556 -1.78835344 0.09393155
## drat         1.18283018  2.48348458  0.47627845 0.64073922
## wt          -4.52977584  2.53874584 -1.78425732 0.09461859
## qsec         0.36784482  0.93539569  0.39325050 0.69966720
## vs1          1.93085054  2.87125777  0.67247551 0.51150791
## am1          1.21211570  3.21354514  0.37718957 0.71131573
## gear4        1.11435494  3.79951726  0.29328856 0.77332027
```

```
## gear5        2.52839599  3.73635801  0.67670068 0.50889747
## carb2       -0.97935432  2.31797446 -0.42250436 0.67865093
## carb3        2.99963875  4.29354611  0.69863900 0.49546781
## carb4        1.09142288  4.44961992  0.24528452 0.80956031
## carb6        4.47756921  6.38406242  0.70136677 0.49381268
## carb8        7.25041126  8.36056638  0.86721532 0.39948495
```

We may observe that all variables have p-values higher than 0.05, which shows that all of them have some sort of impact on MPG. To separate the ones that are really impacting, an ANOVA (using MASS package stepAIC function) is performed.

c) Model Fitting

```
fitModel <- stepAIC(lm(mpg ~ . ,data=mtcars), direction = 'both', trace = FALSE)
fitModel
```

```
##
## 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
##         am1
##     1.80921
```

According to the analysis above, the most impacting variables on MPG, besides transmission type (am), are the weight of the car (wt) and quarter mile time (qsec). This means that other variables are less significant than those two or that the correlation among variables allows us to choose only those, minimizing the deviations (variances) in the final model.

d) Final Model The final model, including the relationship among MPG and transmission (am), weight (wt) and quarter mile time (qsec) is:

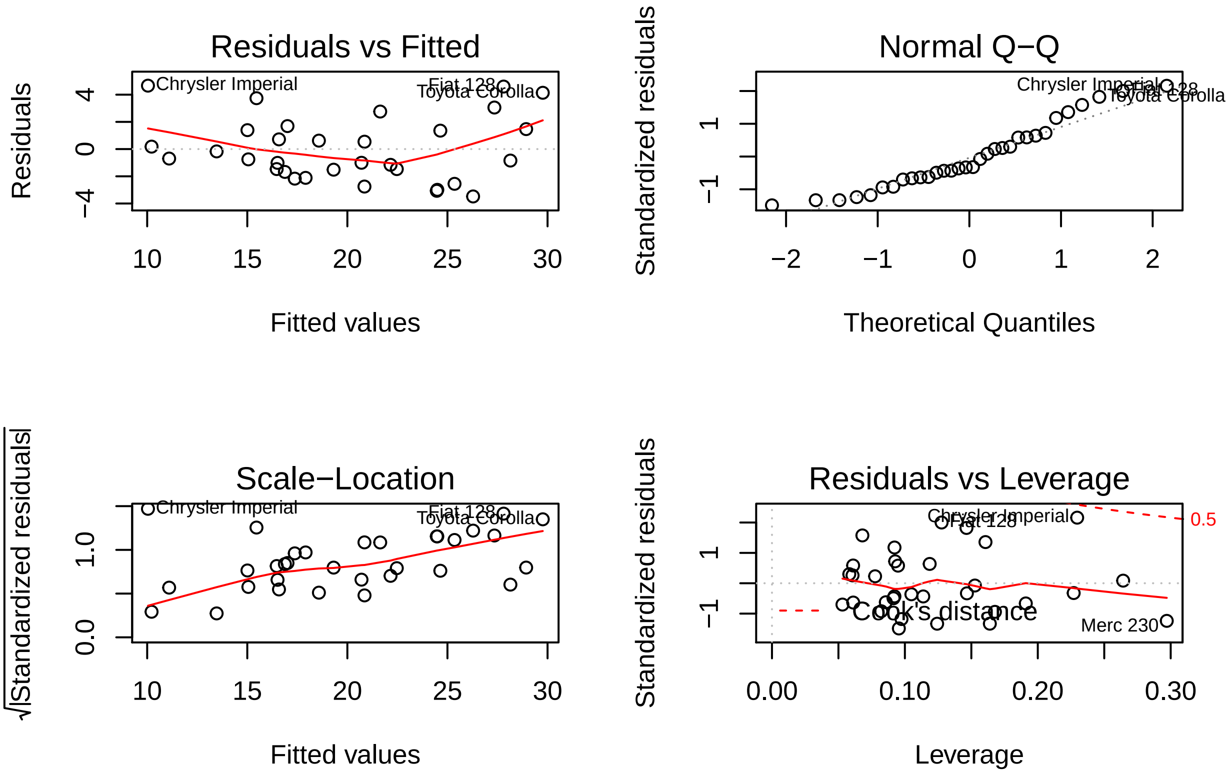
```
finalModel <- lm(mpg ~ factor(am) + qsec + wt, data = mtcars)
summary(finalModel)$coefficients
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)  9.617781  6.9595930  1.381946 1.779152e-01
## factor(am)1  2.935837  1.4109045  2.080819 4.671551e-02
## qsec         1.225886  0.2886696  4.246676 2.161737e-04
## wt          -3.916504  0.7112016 -5.506882 6.952711e-06
```

In this model, we see a reduced impact of transmission on MPG, closer to reality. If the other variables are kept constant, the new impact of transmission on MPG would be only 2.936 miles per gallon (in average), favorable to the manual transmission.

d) Residuals Analysis

```r
par(mfrow = c(2, 2))
plot(finalModel)
```



There are no significant visual trends on the residuals of the final model, and it can be observed good normality pattern. These allow us to conclude that the model could be validated.

Conclusion

MPG is higher for manual cars and the wt dependence of MPG is not different between automatic and manual cars.

As conclusions of the analysis above, we reinforce that:

Manual transmission is better fuel autonomy MPG than the automatic (+2.936 miles per gallon favorable to manual). The final model for MPG considering the most impacting variables is: mpg = 9.618 - 3.917 wt + 1.226 qsec + 1.4109 am in this sense, for the same weight (wt) and quarter mile time (qsec), manual transmission cars get 2.936 miles per gallon more than automatic transmission cars.