Motor Trend Data Analysis Report

Alessandro Camillò

October 19, 2015

Executive Summary

We were tasked by the **Motor Trend US** magazine to evaluate the performance of interior combustion engine vehicles in terms of miles per gallon (MPG) with respect to other characteristics of the vehicles; e.g transmission type, number of cylinders, horse power, etc. To provide an answer we have analyzed the `mtcars` dataset extracted from the 1974 Motor Trend US magazine, which comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973–74 models).

Using linear regression techniques we were able to assess that the **MPG** is affected by the vehicle *weight* and by the number of *cylinders* as well as the type of transmission (whether being it *automatic* instead of *manual*). [1]

Exploratory Data Analysis

First, we load the data set `mtcars` and produce a new working data set *cars* appropriately modified. We change the categorical variables (e.g.: transmission, gears, cylynder ecc) from `numeric` class to `factor` class.

```
require(ggplot2)
require(gridExtra)
library(ggplot2)
library(gridExtra)

data(mtcars)
```

We perform a basic exploratory data analyses using the ANOVA analisys for hints of highly correlated variables.

```
analysis <- aov(mpg ~ ., data = cars)
# Shows the ANOVA outcome
summary(analysis)  ## Results omitted
```

The ANOVA's result suggests that the *cyl*, *disp* and *wt* are significant variables. We use these variables in our first linear regression model.

Model definition

Using the information provided by ANOVA we build a first linear regression model with the variables: *cyl*, *disp*, *wt*, and *am*.

```
fit1 <- lm(mpg ~ cyl + disp + wt + am, data = cars)
fit2 <- lm(mpg ~ cyl + wt + am, data = cars)
summary(fit1); summary(fit2) ## Results omitted
```

[1] All the project code is available online at https://github.com/acamillo/Coursera-Regression-Models

As the p.value of the *disp* variable is quite high **0.9064** and after a quick verification with another ANOVA analisys we immediately decide to drop the *disp* variable being it related to the number of cylinders *cyl*. This model accounts for up to **81.34%** (adjusted R-squared) of the variance of *mpg*. All the coefficients are significant at **0.05** level. However the analysis shows a very high value p-value (**0.908**) for the *am* variable inducing us to review again the model. However we rule out this first impression as explained in the following section.

In Figure 2 we see that in general manual gear provides an higher amount of MPG w.r.t. automatic gear. However, in Figure 1 we see that the type of transmission makes some differences on MPG only when choosing a 4 cylinders or 6 cylinders cars. Despite the MPG has a negative trend in both cases, for automatic transmission cars the drawback is greater. For all the rest of the cases the kind of tranmission is largely uninfluencial and other factors are more relevant.

To enforce this conclusion we make a NULL hypotesis H_0 that MPG and *transmission* data are from the same population and reject it by means of an hypothesis test:

```
result <- t.test(mpg ~ am, data = cars) ## Results omitted
```

With a p-value of 0.0013736 we indeed reject our H_0 hypothesis. Moreover in the box-plot the two boxes do not overlaps confirming that the automatic and manual transmissions are from two distinct populations, so that we can assert that the milage for automatic cars is inferior to the mileage achievable by cars with manual gear.

After these results we have selected the following model: "mpg ~ am + wt:am + hp"

```
fit3 <- lm(mpg ~ am + wt:am + hp, data = cars)
summary(fit3) ## Results omitted
confint(fit3) ## Results omitted
```

The residuals plot, visible in Figure 3 shows no consistent pattern, supporting the accuracy of the independence assumption.

Conclusions

The variables *am*, *wt* and *cyl* are all relevant to explain the variability in *mpg*. However AM by its own cannot explain MPG but is an important variable that explains (partially) car's performance. This drove us so finally select the model $fit3$ "mpg ~ am + wt:am + hp". It has the residual standard error as 2.332 on 27 degrees of freedom and an adjusted R-squared value is 0.8503, meaning that this model can explain about 85% of the variance of the MPG variable. All of the coefficients are significant at 0.05 significant level.

The 95% confidence interval indicates that when "wt" remain constant, cars with manual transmission yeld from 3.2 up to 20 MPG more w.r.t. to an automatic gear cars. Moreover the model indicates an average increse of 11.5 MPG with manual transmssion.

As for the Dfbetas, our model obtain a value of 0 convalidating all the basic assumpetion of linre regression technique.

```
sum( (abs(dfbetas(fit3))) > 1 )
```

Appendix: Figures

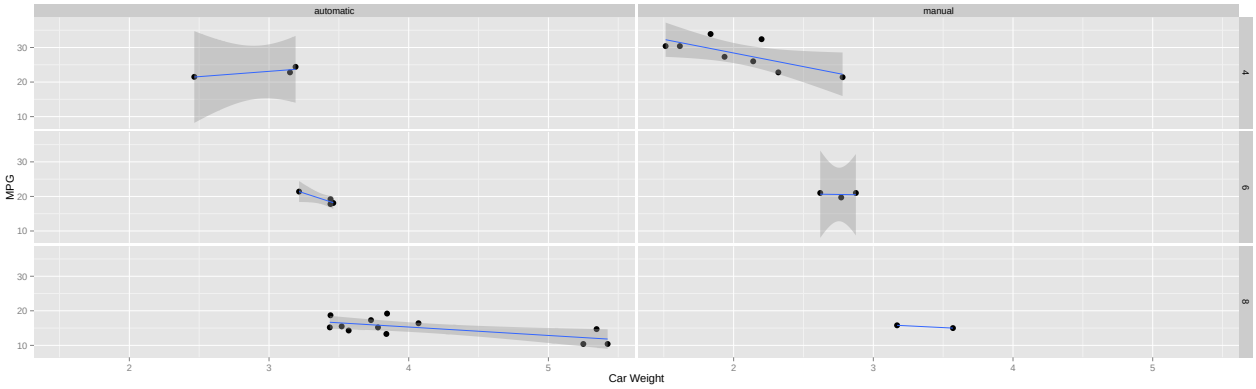

Figure 1: MPG vs Weight by Transmission and Cylinder

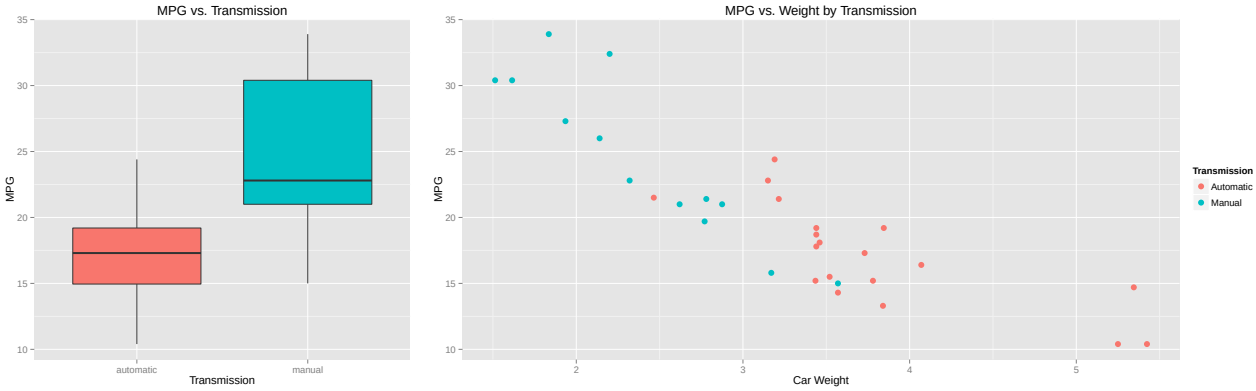

Figure 2: MPG vs Weight by Transmission

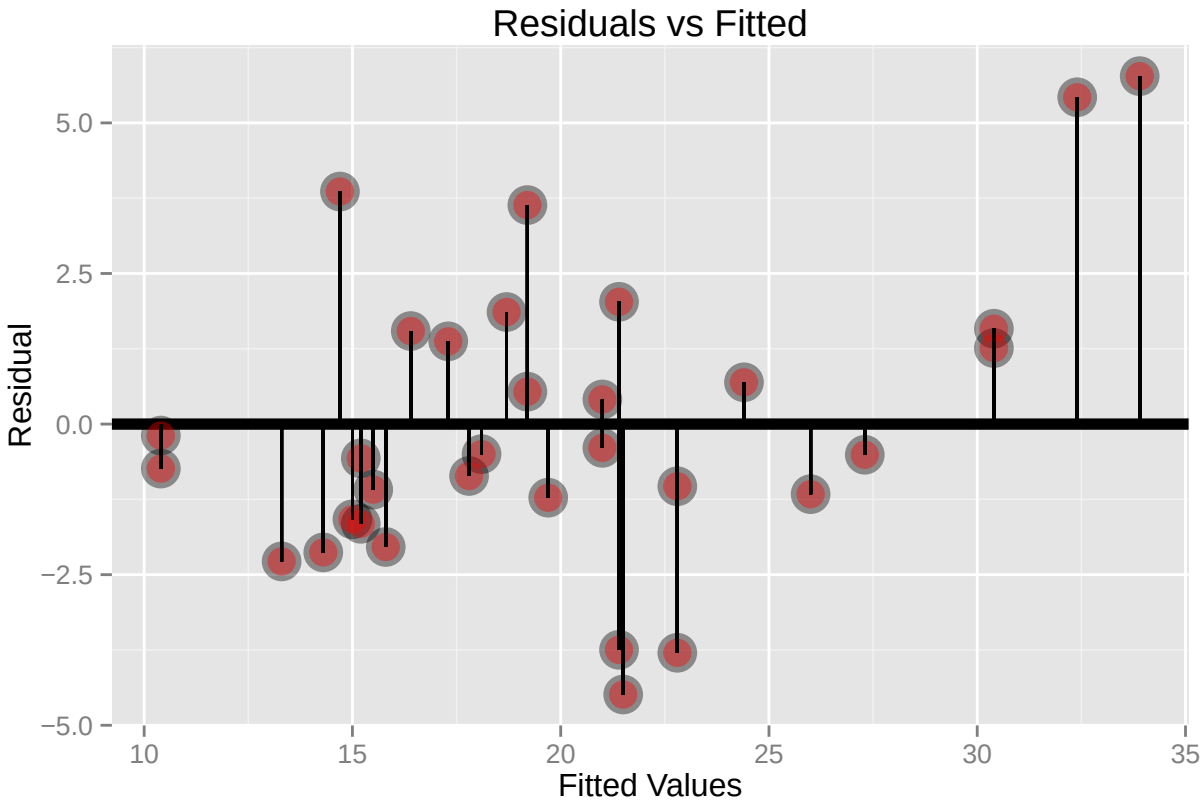

Figure 3: Residuals vs X axis