Motor Trends - Regression Analysis

Executive Summary

Analyse cars data to detect if some difference exists between gaz consumption of manual cars, or automatic ones.

For that, we'll use the data from mtcars which containing 11 measures for 32 cars from the years 73-74. In fact, when the car is light enough (~1000lb), manual transmission is initially at 16.8 mpg which is around 10 more than for automatic ones. However, when the weight increases, there is a decrease of the mpg of (around) -7 for the manual cars, and only -3 for the automatic ones. Thus, starting, with cars more than 3500lb, the automatic cars should be the choice of preference. As a side note, if we had only looked at mean within group we would have always preferred manual cars to automatic, but this is influenced by the fact that the dataset has a few number of lighter cars and all are automatic! See (fig-3).

Analysis

First of all we're going to load the data, `mtcars`, avaible in R's package `datasets`. Then we'll head into to look at what it contains (and adapt the types if necessary).

```
##                    mpg cyl disp  hp drat    wt  qsec vs        am gear
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0    Manual    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0    Manual    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1    Manual    4
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1 Automatic    3
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0 Automatic    3
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1 Automatic    3
##                   carb
## Mazda RX4            4
## Mazda RX4 Wag        4
## Datsun 710           1
## Hornet 4 Drive       1
## Hornet Sportabout    2
## Valiant              1
```

So there are 32 observations of cars and 11 measures.

We're interested in the role playing by the transmission (`am`) in the evolution of miles/gallon consumption (`mpg`). To see a visual interpretation of the relation between the two, a boxplot is available in the appendix (fig-1). Still visualy, it looks like the difference between the two groups (manual and automatic transmissions) is true, let's see the mean of each first and then perform a between two-groups t-test (assuming normality and independence).

The difference between both is significant (p-value 0.0013736) and the 95% confidence interval doesn't contains 0, and thus, at this stage, we can say that the manual cars are better than automatic cars from 3.2096842 to 11.2801944 miles per galon. However, using the transmission alone is not enough to quantify the difference for specific cases. To see that, we can look at the prediction of mile per gallon using the single transmission independent variable in a linear regression.

So the relation between the to is rather clear, with p-value of the change being 2.850207410^{-4}. However, the variance explained is quite low, with an R^2 at 0.3597989. So something is missing in the mix.

In order to find another model explaining better the miles per gallon, we'll use the best model selection since our dataset is quite small. For the sake of sanity, models selected by the stepwise method (forward and backward) have been ran, and they were only diverging at the third variable selecting `hp` or `qsec`. To run these selections, we'll use the useful package `leaps`.

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

After playing around, `wt` is the first candidate to try out. A visual representation on how `mpg` is related to both `am` and `wt`, a plot has been provided in the appendix fig-2. The linear regression involving both is increasing R^2 to 0.7528348, however the effect of the change of `am` is not more significant. To solve that, we'll will increase the model by adding the `qsec` (based on the model selection results above). Now the model has a R^2 of .

Nevertheless, the fig-2 is more or less showing an interaction between `wt` and `am`. So, we're going to add this interaction to the moedl and check if it's significant.

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)  9.723053  5.8990407  1.648243 0.1108925394
## wt          -2.936531  0.6660253 -4.409038 0.0001488947
## amManual    14.079428  3.4352512  4.098515 0.0003408693
## qsec         1.016974  0.2520152  4.035366 0.0004030165
## wt:amManual -4.141376  1.1968119 -3.460340 0.0018085763
```

Not only the coefficient are all significant but the new interaction term is explaining more intuitively how the weight of a car is affecting the miles per gallon consumption when it's a manual or an automatic. That's to say, a manual car is worst by a factor of -4.1413764 per 1000lb increase in weigth. Last but not least, this regression is showing a pretty good residual plot and doesn't present evidence of outliers - see figures after fig-4, specially the Cook's distance.

Now we can test how this new variable (interaction) in the model is significant to explain the variance, for this we can run an anova.

It's fair enough to include it, since the p-value is 0.0018086!

For the sake of sanity, we can have a quick look at the VIF of the models, using the `vif` function in the `car` package.Without interaction, we have this very good VIF:

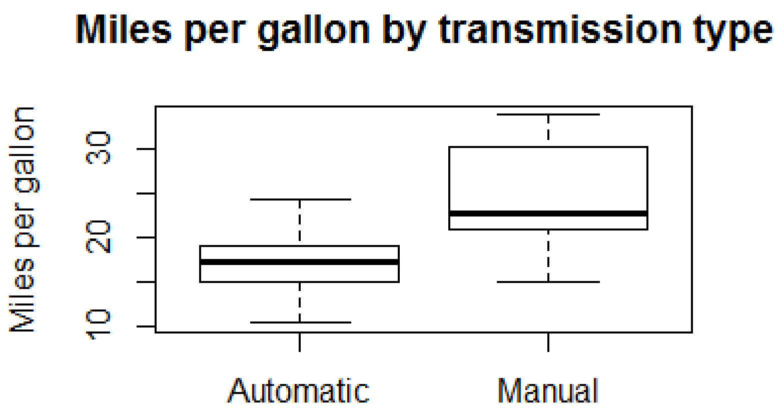
```
## Warning: package 'car' was built under R version 3.3.1

##       wt     qsec       am 
## 2.482952 1.364339 2.541437
```

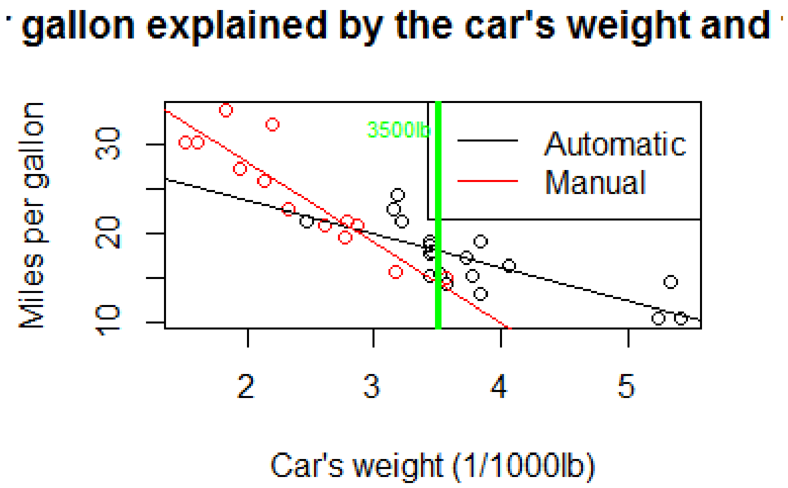
However, with the interaction we have this one:

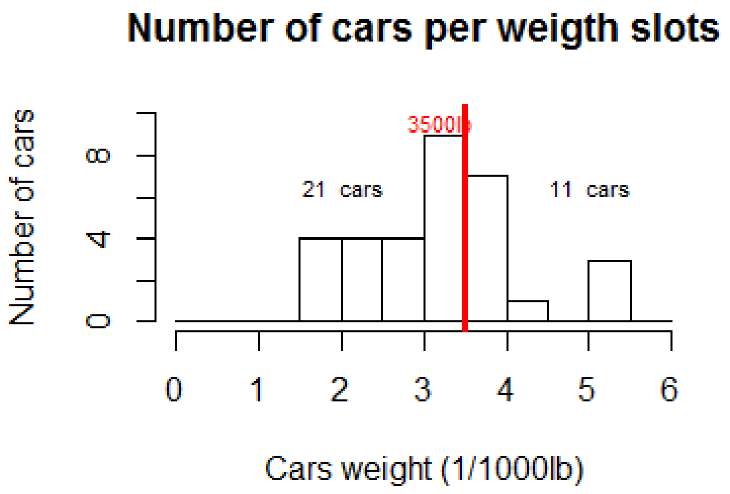
```
##        wt        am      qsec     wt:am 
##  3.030963 20.970925  1.447406 16.302453
```

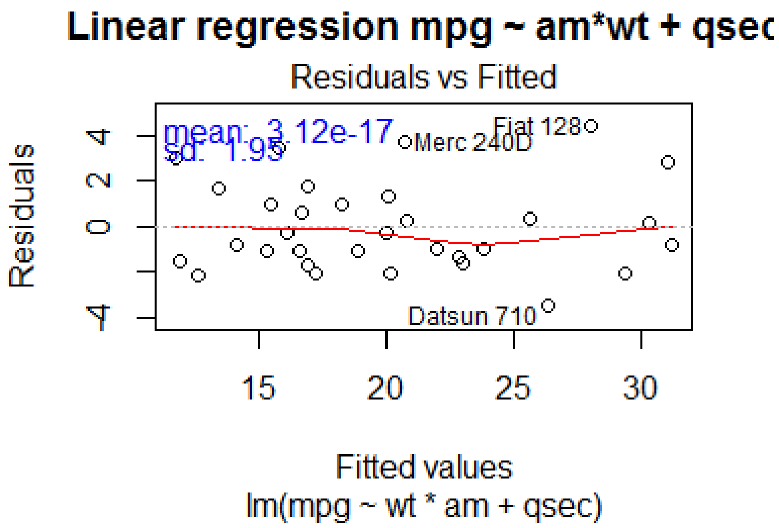
The model including the interaction inflates the variance due to colinearity, but we could have foresee it regarding fig-2. However, still it's inclusion allow better explanation of the difference between the groups.

Appendix

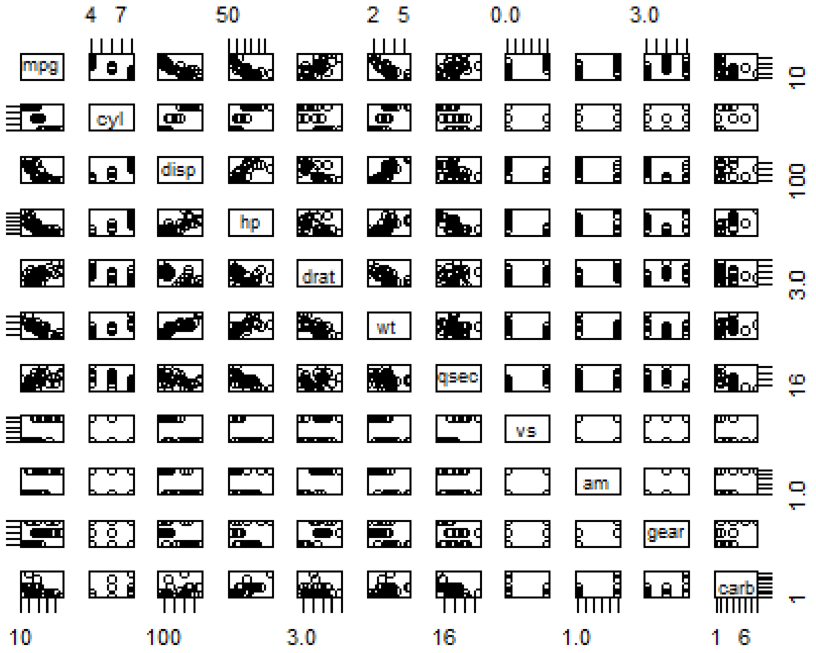


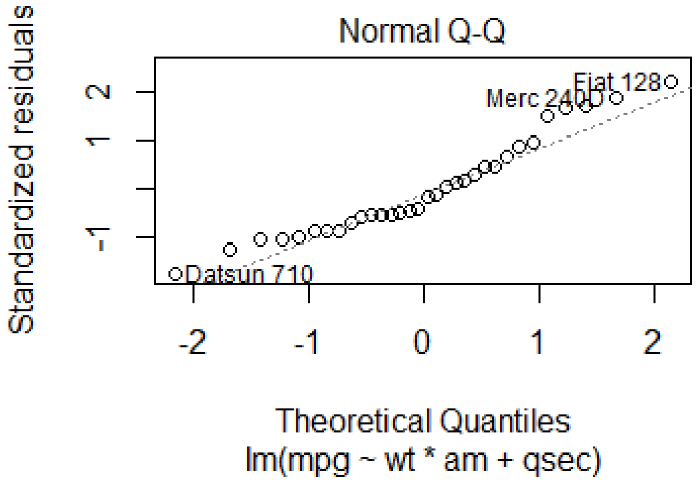



Number of cars per weigth slots
3500lb
21 cars
11 cars
Number of cars
0
4
8
0
1
2
3
4
5
6
Cars weight (1/1000lb)


Linear regression mpg ~ am*wt + qsec
Residuals vs Fitted
mean: 3.12e-17
sd: 1.95
Merc 240D
Fiat 128
Datsun 710
Residuals
-4
0
2
4
15
20
25
30
Fitted values
lm(mpg ~ wt * am + qsec)

Cross-plot of all variables in mtcars


mpg
cyl
disp
hp
drat
wt
qsec
vs
am
gear
carb


Normal Q-Q
Standardized residuals
Fiat 128
Merc 240D
Datsun 710
Theoretical Quantiles
lm(mpg ~ wt * am + qsec)