Motor Trend - An analysis of the miles per gallon

A.N.

Saturday, August 16, 2014

Executive summary

This analysis is based on the mtcars dataset. The objective is to analyse the relationship between a set of variables and the outcome miles per gallon(mpg) and try to answer to the following two questions:

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

Data analysis

Which variables are correlated?

On the appendix 1, it's possible to see on the graphic above that ***mpg*** is highly correlated with ***vs***, ***am***, ***gear*** and ***qsec***; and negatively correlated with ***cyl***, ***disp*** and ***hp***.

Let visualize the relation between mpg and am ?

As we want to compare the impact of an automatic or manual car on the mile per gallon so we set the variable ***am*** as a factor. We can clear see on the appendix 2 that the automatic car tend to have a lower mile per gallon than the manual cars. The relation between manual car and mile per gallon tend to be quite spread and right skewed.

Estimate the model

```
fit<-lm(mpg~ am-1,data=mtcars)
summary(fit)$coef
```

```
##     Estimate Std. Error t value  Pr(>|t|)
## amA    17.15      1.125   15.25 1.134e-15
## amM    24.39      1.360   17.94 1.376e-17
```

Looking at the coefficients, We estimate an expected increase of 17 miles for every mile increase of mile per gallon, for an automatic cars. For manual cars, we estimate an expected increase of 24 miles for every mile increase of mile per gallon.

Let's include extra variables and see if it helps to explain the mpg. the others variables that will be included in the model are:

- qsec,
- the *10/wt* : as wt is negatively correlated with mpg so we take the inverse.

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am - 1
## Model 2: mpg ~ am + qsec - 1
## Model 3: mpg ~ am + qsec + I(10/wt) - 1
##   Res.Df RSS Df Sum of Sq    F  Pr(>F)
## 1     30 721
## 2     29 353  1       368 76.0 1.8e-09 ***
## 3     28 136  1       217 44.7 2.9e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##          Estimate Std. Error t value  Pr(>|t|)
## amA       -13.596     4.2405  -3.206 3.352e-03
## amM       -11.508     4.0012  -2.876 7.611e-03
## qsec        1.138     0.2601   4.373 1.535e-04
## I(10/wt)    3.658     0.5470   6.687 2.940e-07
```

Looking at the coefficient, all the variables from the third model have meaningful p-values. Also there is an important drop in the sum of squared residuals (RSS). Also the variance of the residual dropped from 4.9 in the initial model to 2.2 in the model 3.

The residuals

Cannot see any pattern on the Residuals Vs Fitted values on the appendix 3, which is a good thing. On the normal Q-Q plot appendix 4, we can clearly see that few points at the tails (left and right) are not inline with the others. It's clear that the distribution of the residual is not normal.

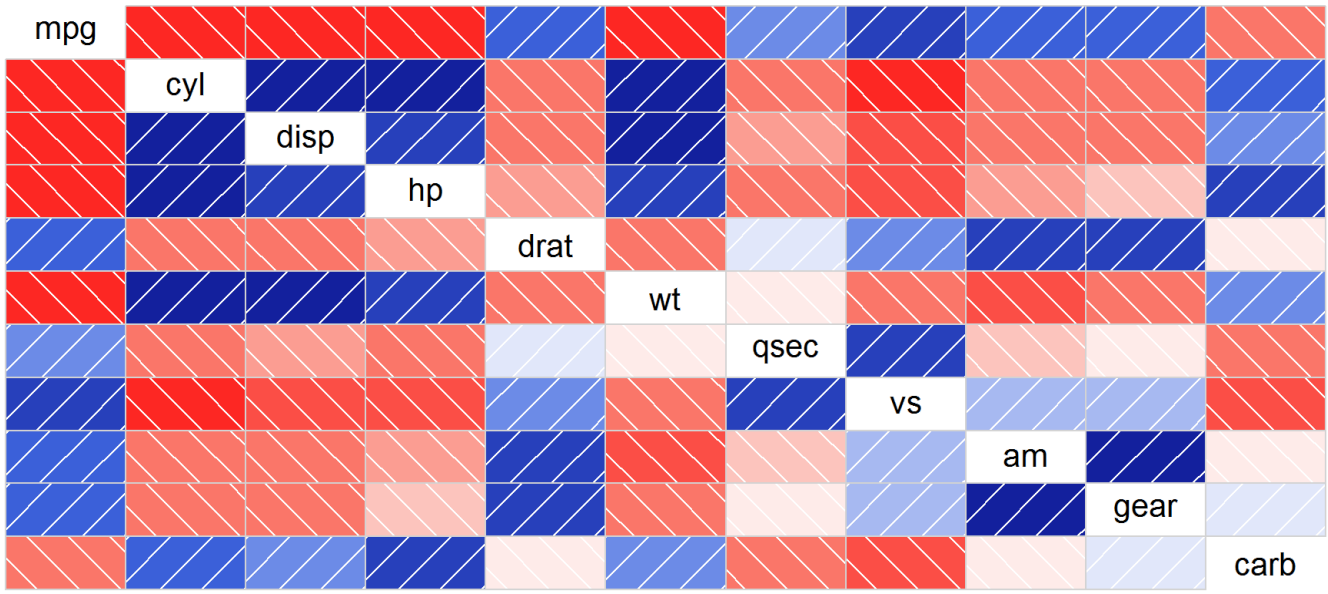
The influential residuals ?

```
## Potentially influential observations of
##   lm(formula = mpg ~ am + qsec + I(10/wt) - 1, data = mtcars) :
##
##               dfb.amA dfb.amM dfb.qsec dfb.I(10 dffit cov.r   cook.d hat
## Merc 230       0.40    0.39   -0.42     0.14    -0.47  1.58_*  0.06   0.32
## Fiat 128      -0.69   -0.50    0.70    -0.27     1.09  0.48_*  0.24   0.13
## Lotus Europa  -0.17   -0.07    0.36    -0.65    -0.72  1.74_*  0.13   0.41_*
## Maserati Bora -0.05   -0.07    0.03     0.05    -0.09  1.47_*  0.00   0.22
```

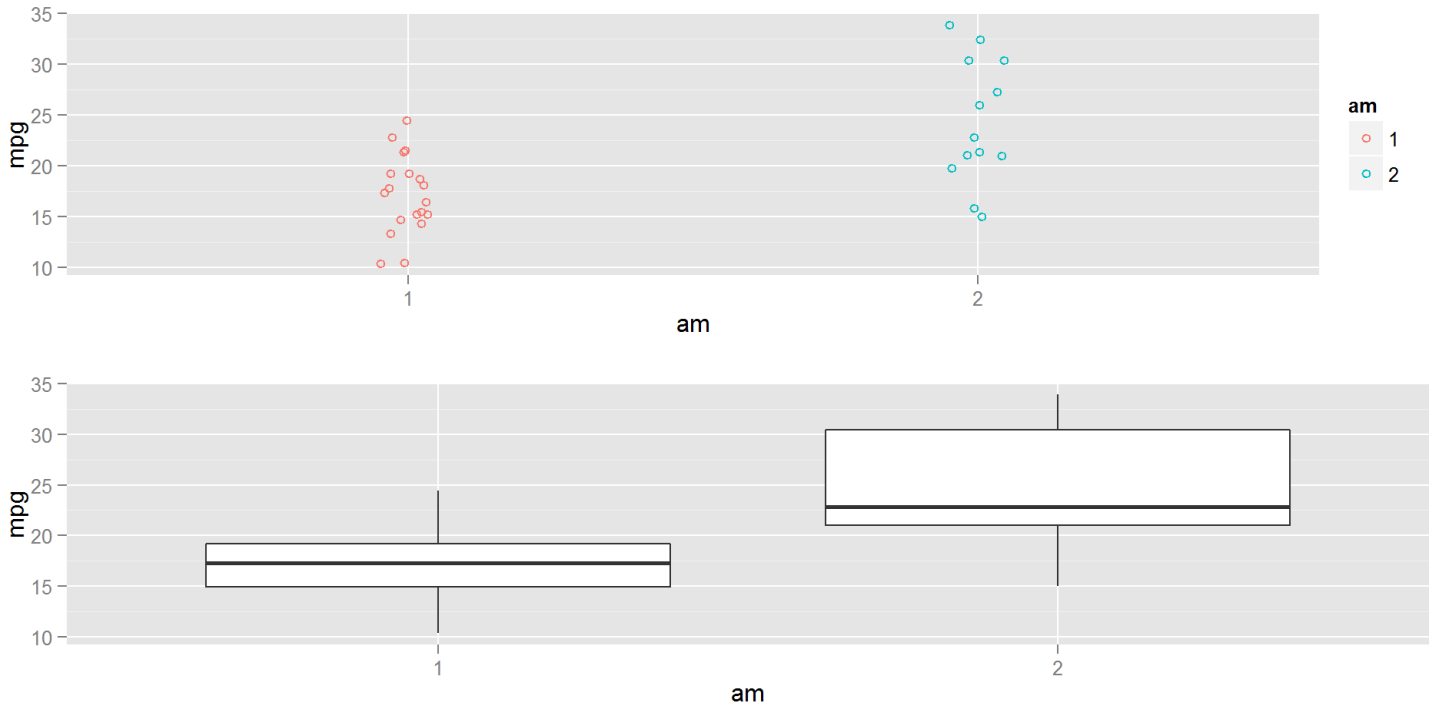
Those four influential points can be seen on the graphic in appendix 5 and are impaction the regression relationship. ***Maserati Bora***,***Fiat 128***, ***Merc 230*** have a high leverage and are impacting the fit. ***Lotus Europa*** has a high leverage too but is not impacting the fit.

Appendix

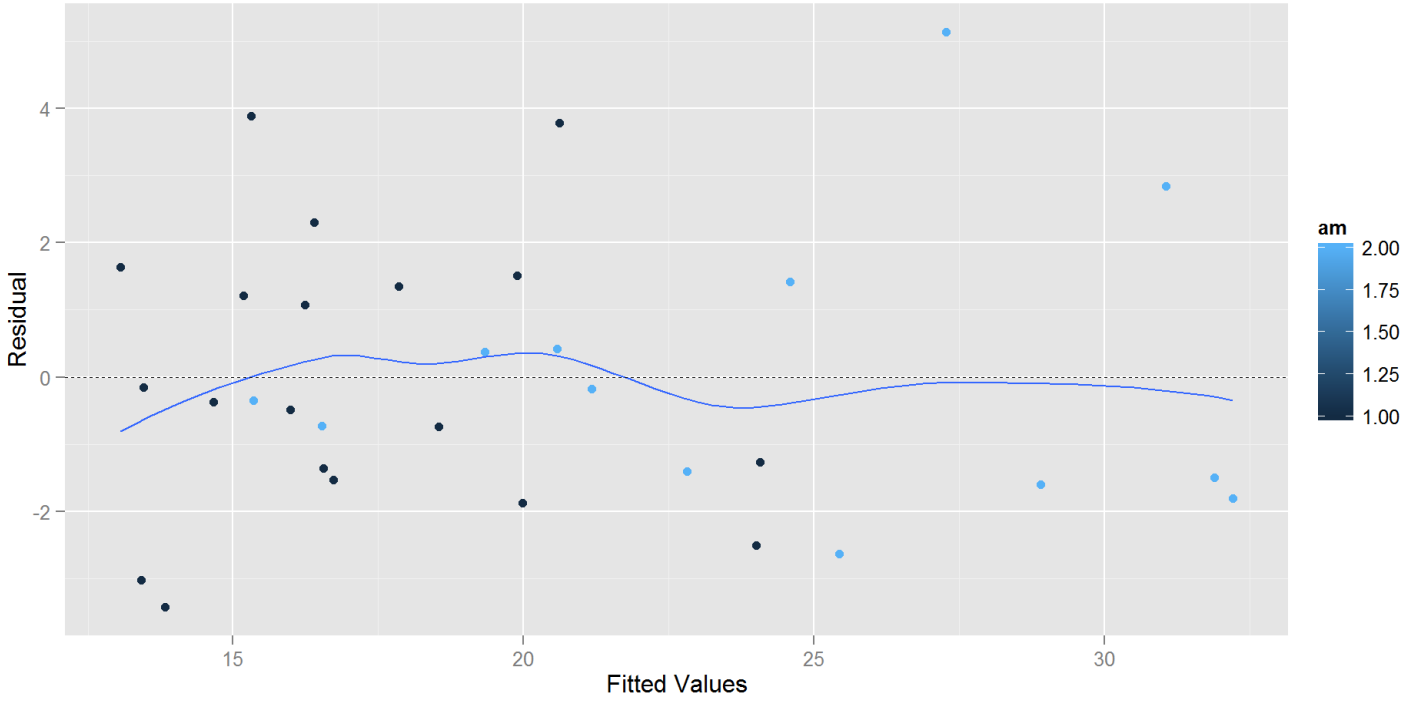
Appendix 1 Kendall correlation map



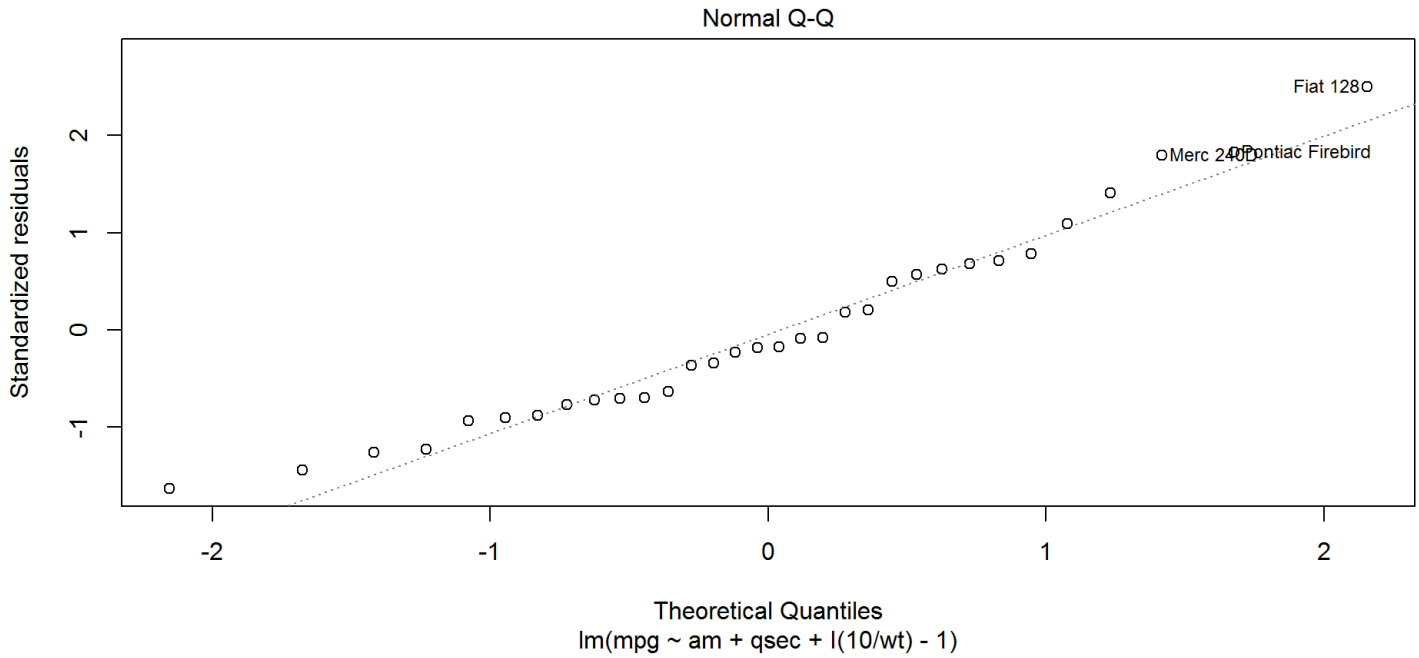
Appendix 2 Relation between mpg and am



Appendix 3 Residuals vs Fitted



Appendix 4 Normal Q-Q plot for model 3



Appendix 5 Influential residuals