# Regression models - Course project - Motor Trend Analysis

*01/16/2015*

In this report we try to addres the the following two questions:

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

We have found that manual transmission overall, seems to give a better mpg than automatic transmission. In this report we show though, that other variables such as number of cylinders, weight and horsepower, play an important role as well.

## Sanity check

First we look at the data to have an idea what we're looking at. We see that, next to mpg, we have 10 numeric columns.

```r
data(mtcars)
str(mtcars)  # See appendix for results
```

However, some of them should be treated as factor:

```r
mtcars$cyl <- factor(mtcars$cyl)
mtcars$vs <- factor(mtcars$vs, labels=c("V-engine", "Straight engine"))
mtcars$gear <- factor(mtcars$gear)
mtcars$carb <- factor(mtcars$carb)
mtcars$am <- factor(mtcars$am,labels=c('Automatic','Manual'))
```

## Exploratory analysis

We draw a pairs plot to have a first look at the relation ship between mpg and the other variables (see appendix). There we already see some trends, such as

- An increase in number of cylinders tends to lead to a decrease in mpg.
- An increase in hp tends to lead to a decrease in mpg.
- An increase in rear axle ratio tends (though more varition) to lead to an increase in mpg.
- An increase in weight, seems to lead to a decrease in mpg

## Model fitting

We now fit a linear model, initially using all variables and see if we get results which are usable. When looking at the summary, we see that we get high p-values for all variables, suggesting that this model is not such a good fit. We choose a better model by using a step wise algorithm, in backward direction (i.e. remove candidates from the fit).

```r
vars.all <- lm(mpg ~ ., data = mtcars)
vars.best.fit <- step(vars.all, direction = "backward")
```

```r
summary(vars.best.fit)
```

```
##
## Call:
## lm(formula = mpg ~ cyl + hp + wt + am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9387 -1.2560 -0.4013  1.1253  5.0513
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 33.70832    2.60489  12.940 7.73e-13 ***
## cyl6        -3.03134    1.40728  -2.154  0.04068 *
## cyl8        -2.16368    2.28425  -0.947  0.35225
## hp          -0.03211    0.01369  -2.345  0.02693 *
## wt          -2.49683    0.88559  -2.819  0.00908 **
## amManual     1.80921    1.39630   1.296  0.20646
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.41 on 26 degrees of freedom
## Multiple R-squared:  0.8659, Adjusted R-squared:  0.8401
## F-statistic: 33.57 on 5 and 26 DF,  p-value: 1.506e-10
```

**Interpretation:** From this model's summary we see that the structure in the data is reasonably well explained by the variables (or variable values in case of categorical variables): cylinder 16, cylinder 18, horsepower, weight and manual transmision. Most of those variables have pretty low p-values ($< 0.05$), which increases our confidence in the model. Additionally, we see that the R-squared values are pretty high, that is around 84 percent of the in- or decrease in milage per gallon is explained by our model. Finally, we see that the 1st and 3rd quantiles of the residuals are about the same as `1.5 +/- std error` given a strong hint that the residuals are normally distributed. This observation is further strengthened by the normal q-q plot (see appendix), though there some outliers become apparent.

From this model, we see that we can expect an `1.8092` increase in mpg when changing from automatic to manual, keeping all other variables fixed. This suggests that having a manual car is better for mpg. To further test this hypothesis, we perform a t-test:

```r
t.test(mpg ~ am, data = mtcars) # See appendix
```

Finally, we construct a 95% confidence interval to find the values below. This still includes a possible decrease, which shows that just switching from automatic to manual is not to be sure (with earlier mentioned confidence) that it will actually increase milage.

```r
confint(vars.best.fit)["amManual",]
```

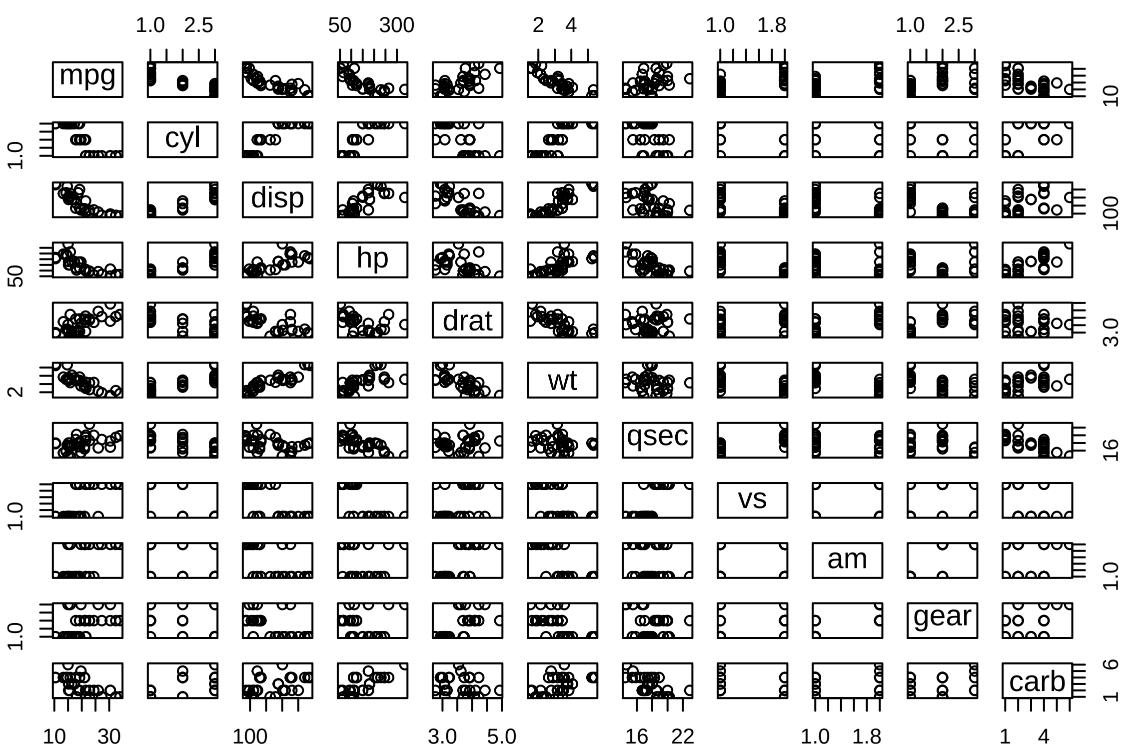
```
##     2.5 %    97.5 %
## -1.060934  4.679356
```

# Appendix

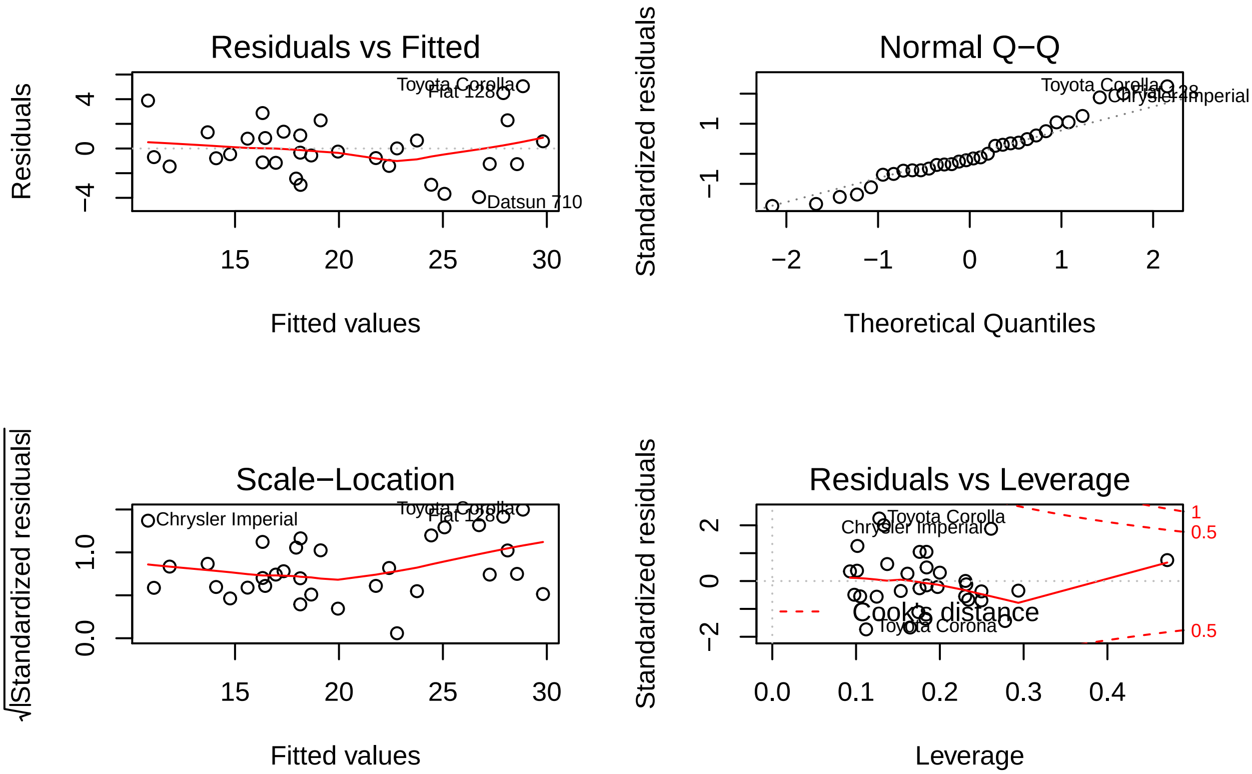
```r
str(mtcars)
```

```
## 'data.frame':    32 obs. of  11 variables:
##  $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ cyl : Factor w/ 3 levels "4","6","8": 2 2 1 2 3 2 3 1 1 2 ...
##  $ disp: num  160 160 108 258 360 ...
##  $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
##  $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
##  $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
##  $ qsec: num  16.5 17 18.6 19.4 17 ...
##  $ vs  : Factor w/ 2 levels "V-engine","Straight engine": 1 1 2 2 1 2 1 2 2 2 ...
##  $ am  : Factor w/ 2 levels "Automatic","Manual": 2 2 2 1 1 1 1 1 1 1 ...
##  $ gear: Factor w/ 3 levels "3","4","5": 2 2 2 1 1 1 1 2 2 2 ...
##  $ carb: Factor w/ 6 levels "1","2","3","4",..: 4 4 1 1 2 1 4 2 2 4 ...
```

```r
pairs(mpg ~ ., data=mtcars)
```




```r
par(mfrow=c(2,2))
plot(vars.best.fit)
```




```r
t.test(mpg ~ am, data = mtcars) # See appendix
```

```
##
##  Welch Two Sample t-test
##
## data:  mpg by am
## t = -3.7671, df = 18.332, p-value = 0.001374
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.280194  -3.209684
## sample estimates:
## mean in group Automatic    mean in group Manual
##                17.14737                24.39231
```