# Regression Models Final Assignment : mtcars analysis

*Alnour Ribault*

*22 février 2018*

## Introduction

This document is my submission for the final assignment of the Regression Models course from the Coursera Data Science specialization by the John Hopkins University.

## Instructions

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"

Take the `mtcars` data set and write up an analysis to answer their question using regression models and exploratory data analyses.

## Author's note

Since we're doing many significance tests in this study, we have to correct for this. in order to avoid getting p-values by pure lucK. Since we estimate the number of significance tests in the study to be of the order of 10, the Bonferroni correction tells us to look for p-values under 0.005.

## Exploratory analysis

We first get a grip of the data by using basic R commands.

It appears that some of the variables are naturally discrete. We thus convert them to factor variables.

```
mtcars_fac <- mutate(mtcars, cyl = factor(cyl),
                     vs = factor(vs, labels = c("V engine", "Straigth engine")),
                     am = factor(am, labels = c("Automatic", "Manual")),
                     gear = factor(gear),
                     carb = factor(carb))
```

We then plot a pair graph of the original `mtcars` data to get a grip of the correlation between the variables. `mpg` seems to decrease when `cyl`, `disp`, `hp`, `wt` increase, and seems to be higher among V engines than among Straigth engines, and higher among automatic transmission cars than among manual transmission cars as well.
Some of those relations make sense : a heavier car will naturally use more gas, and a car designer will have to sacrifice some efficiency in order to achieve higher horsepower. The others, however, are more obscure.

Since we are especially interested in the relationship between variables `mpg` and `am`, we plot a boxplot of the value of `mpg` for automatic and manual transmission. It appears that cars with manual transmission have a notably higher mpg than those with automatic transmission.

```
fit <- lm(mpg ~ am, mtcars_fac)
```

Fitting a first model we find that the average value of `mpg` is 17.15 for automatic cars and 7.24 for manual cars. Both p-values are low enough for us to reject the null hypothesis that the actual coefficients are zero. However, the model's R-squared is only 0.36 which is not satisfying. We thus have to try adding other variables in order to explain `mpg`'s variance.

# Model Selection

As our first model is not satisfying, we will search for other significant variables in the modelling of `mpg`. We start with a model including all variables and use the AIC (Akaike information criterion) to eliminate variables down to a better model.

This new model is not fully satisfying, however. The p-values for the `cyl8` coefficient is 0.35, which is far from enough to attest for its significance. The p-value for the `am` coefficient is 0.21 which is not good enough either.

Since these coeffcients are not significative, let us fit two models, each with one of these variable dropped. We also fit a model where we drop both `cyl` and `am` variable and compare them to our previous model.

```
fit_no_am <- lm(mpg ~ wt + hp + cyl, mtcars_fac)
fit_no_cyl <- lm(mpg ~ wt + hp + am, mtcars_fac)
fit_no_am_no_cyl <- lm(mpg ~ wt + hp, mtcars_fac)
anova(fit_step, fit_no_am)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ cyl + hp + wt + am
## Model 2: mpg ~ wt + hp + cyl
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     26 151.03
## 2     27 160.78 -1    -9.752 1.6789 0.2065
```

```
anova(fit_step, fit_no_cyl)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ cyl + hp + wt + am
## Model 2: mpg ~ wt + hp + am
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     26 151.03
## 2     28 180.29 -2   -29.265 2.5191    0.1 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(fit_step, fit_no_am_no_cyl)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ cyl + hp + wt + am
## Model 2: mpg ~ wt + hp
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
```

```
## 1     26 151.03
## 2     29 195.05 -3   -44.022 2.5262 0.07947 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The model without either `am` or `cyl` seems to be the only one improving on the previous one. It seems to be a pretty satisfying model since the p-values are better than previously, being under our Bonferroni threshold of 0.005!
So the two remaining variables are horsepower and weight. But those two seem intuitively related : cars that have higher horsepower will be heavier. We thus try correcting for the interaction between the two variables.

```
fit_final <- lm(mpg ~ wt + hp + wt*hp, mtcars_fac)
anova(fit_no_am_no_cyl, fit_final)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ wt + hp
## Model 2: mpg ~ wt + hp + wt * hp
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     29 195.05
## 2     28 129.76  1    65.286 14.088 0.0008108 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The improvement obtained by adding the interaction term seems like a good idea. This model actually explains *almost* as much variance (R-squared = 0.88) than the model containing all the variables (R-squared = 0.89) which many less variables, *and* this time the coefficients of all the terms are highly significant.

# Residuals

We observe no particular pattern in the various plots of the residuals, and they seem to be normally distributed : our model presents no obvious weakness.

# Conclusion

Our study showed that, if the mtcars sample is representative, the transmission mode does not have a significant influence on the MPG, which is mainly explained by weight and horsepower. Therefore, the interaction between the transmission mode and the MPG is not quantifiable : the difference in MPG between the automatic cars and the manual cars is explained by their weight and horsepower.

# Appendices

## mtcars data

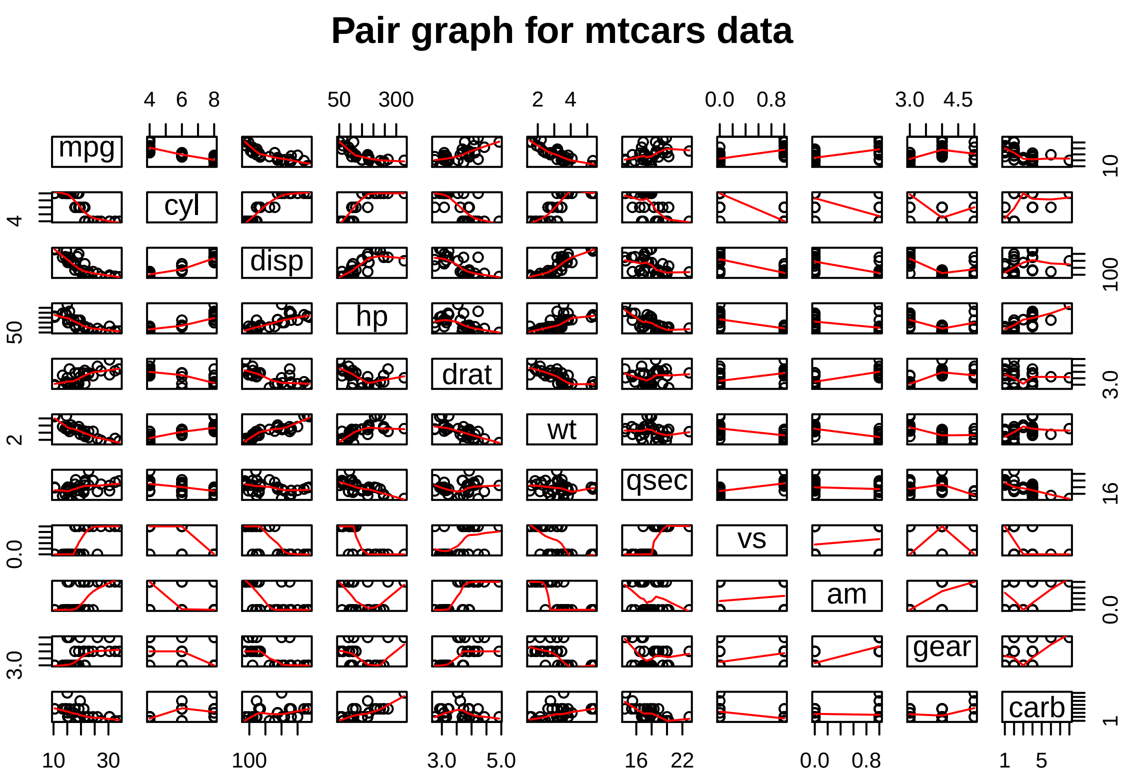
```
##                    mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## Hornet 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
```

```
## '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 : num  6 6 4 6 8 6 8 4 4 6 ...
##  $ 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  : num  0 0 1 1 0 1 0 1 1 1 ...
##  $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
##  $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
##  $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

```
##       mpg             cyl             disp             hp       
##  Min.   :10.40   Min.   :4.000   Min.   : 71.1   Min.   : 52.0  
##  1st Qu.:15.43   1st Qu.:4.000   1st Qu.:120.8   1st Qu.: 96.5  
##  Median :19.20   Median :6.000   Median :196.3   Median :123.0  
##  Mean   :20.09   Mean   :6.188   Mean   :230.7   Mean   :146.7  
##  3rd Qu.:22.80   3rd Qu.:8.000   3rd Qu.:326.0   3rd Qu.:180.0  
##  Max.   :33.90   Max.   :8.000   Max.   :472.0   Max.   :335.0  
##       drat             wt             qsec             vs        
##  Min.   :2.760   Min.   :1.513   Min.   :14.50   Min.   :0.0000  
##  1st Qu.:3.080   1st Qu.:2.581   1st Qu.:16.89   1st Qu.:0.0000  
##  Median :3.695   Median :3.325   Median :17.71   Median :0.0000  
##  Mean   :3.597   Mean   :3.217   Mean   :17.85   Mean   :0.4375  
##  3rd Qu.:3.920   3rd Qu.:3.610   3rd Qu.:18.90   3rd Qu.:1.0000  
##  Max.   :4.930   Max.   :5.424   Max.   :22.90   Max.   :1.0000  
##        am              gear            carb      
##  Min.   :0.0000   Min.   :3.000   Min.   :1.000  
##  1st Qu.:0.0000   1st Qu.:3.000   1st Qu.:2.000  
##  Median :0.0000   Median :4.000   Median :2.000  
##  Mean   :0.4062   Mean   :3.688   Mean   :2.812  
##  3rd Qu.:1.0000   3rd Qu.:4.000   3rd Qu.:4.000  
##  Max.   :1.0000   Max.   :5.000   Max.   :8.000  
```

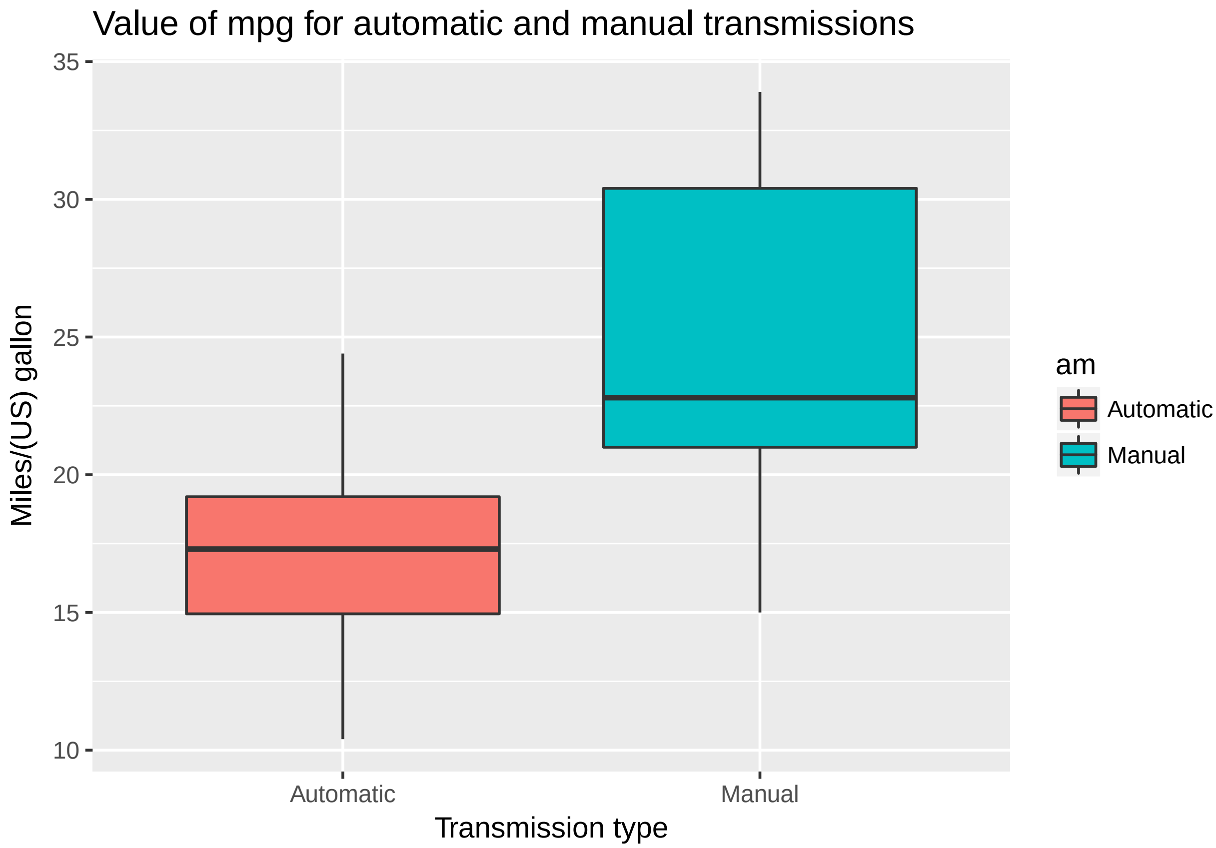
## Pair graph

```r
pairs(mtcars, panel=panel.smooth, main="Pair graph for mtcars data")
```




## Boxplot

```r
g <- ggplot(mtcars_fac, aes(am, mpg)) +
    geom_boxplot(aes(fill = am)) +
    labs(title = "Value of mpg for automatic and manual transmissions",
         x = "Transmission type",
         y = "Miles/(US) gallon")
print(g)
```




# Models

## Simple Model

```r
summary(fit)
```

```
##
## Call:
## lm(formula = mpg ~ am, data = mtcars_fac)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.3923 -3.0923 -0.2974  3.2439  9.5077
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   17.147      1.125  15.247 1.13e-15 ***
## amManual       7.245      1.764   4.106 0.000285 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.902 on 30 degrees of freedom
## Multiple R-squared:  0.3598, Adjusted R-squared:  0.3385
## F-statistic: 16.86 on 1 and 30 DF,  p-value: 0.000285
```

## Model with all variables

```r
summary(fit_all)
```

```
##
## Call:
## lm(formula = mpg ~ ., data = mtcars_fac)
##
## 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
## vsStraigth engine  1.93085    2.87126   0.672   0.5115
## amManual           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
```

## Model obtained with AIC

```r
summary(fit_step)
```

```
##
## Call:
## lm(formula = mpg ~ cyl + hp + wt + am, data = mtcars_fac)
##
## 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
```

### Model with only `wt` and `hp`

```r
summary(fit_no_am_no_cyl)
```

```
##
## Call:
## lm(formula = mpg ~ wt + hp, data = mtcars_fac)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.941 -1.600 -0.182  1.050  5.854
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 37.22727    1.59879  23.285  < 2e-16 ***
## wt          -3.87783    0.63273  -6.129 1.12e-06 ***
## hp          -0.03177    0.00903  -3.519  0.00145 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.593 on 29 degrees of freedom
## Multiple R-squared:  0.8268, Adjusted R-squared:  0.8148
## F-statistic: 69.21 on 2 and 29 DF,  p-value: 9.109e-12
```

### Final Model with `wt`, `hp` and their interaction

```r
summary(fit_final)
```

```
##
## Call:
## lm(formula = mpg ~ wt + hp + wt * hp, data = mtcars_fac)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.0632 -1.6491 -0.7362  1.4211  4.5513
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 49.80842    3.60516  13.816 5.01e-14 ***
## wt          -8.21662    1.26971  -6.471 5.20e-07 ***
```

```
## hp           -0.12010    0.02470  -4.863 4.04e-05 ***
## wt:hp         0.02785    0.00742   3.753 0.000811 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.153 on 28 degrees of freedom
## Multiple R-squared:  0.8848, Adjusted R-squared:  0.8724
## F-statistic: 71.66 on 3 and 28 DF,  p-value: 2.981e-13
```

# Residuals

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