# RegMod-Project. Research on the effect of automatic transmission on mileage per gallon (MPG)

*Gerrit Versteeg*

*July 2nd, 2017*

## 1. Executive summary

This report investigates the influence of automatic versus manual transmission on the mileage per gallon for cars. The results indicate that for a car of average weight and average 1/4-mile time, a manual transmission adds 2.94 miles per gallon on the average 9.62 miles per gallon of a car with an automatic transmission (95% certainty). So using a manual transmission is better for fuel-efficiency.

## 2. Loading and preparing the data

```
library("GGally", quietly=TRUE, warn.conflicts = FALSE)    ## for ggpairs
library("dplyr", quietly=TRUE, warn.conflicts = FALSE)     ## for data prep
library("ggplot2", quietly=TRUE, warn.conflicts = FALSE)   ## for the graphs
library("datasets", quietly=TRUE, warn.conflicts = FALSE)  ## for the data iself
DF <- tbl_df(mtcars)                ## read data into tibble
DF$am <- as.factor(DF$am)           ## set am from num to factor
attach(DF)                          ## to reference column-names
```

## 3. Exploratory Data Analysis

There are no missing values (NULL, NA's) for any of the variables in mtcars, so nothing needs to be imputed via some imputing strategy. First let's have a look at all variables and their correlations to get an impression of the dataset by looking at the ggpairs plot (see appendix, figure 1). The variables that have the largest correlation with our independent variable "am" are in descending order: gear (0.79), drat (0.71), wt (-0.69), disp (-0.59), cyl (-0.52). The remaining variables appear to be further specifications of the type of engine. Also there is a 0.6 correlation between am and mpg and the plot suggests an increase in mpg for cars with a manual transmission.

## 4. Choosing the model

A normal linear model seems in order, because the outcomes (mpg) are not binary, Bernoulli nor binomial and there are no unbounded count data, rates or proportions. Modelling a first fit with mpg as outcome and am as sole predictor, yields a promising result.

```
fit1 <- lm(mpg~am)                          ## fitting just am
summary(fit1)$coef
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 17.147368   1.124603 15.247492 1.133983e-15
## am1          7.244939   1.764422  4.106127 2.850207e-04
```

The expected mpg for cars with automatic transmission is 17.15 mpg, while the expected mpg for cars with manual transmission is 24.39 (an additional 7.24 miles per US gallon). The relationship appears to be strongly

significant (2.85e-04). BUT the adjusted R-squared is only 0.3385, therefor only 34% of the variation in the dataset is explained through this first simple model.

A lot of lesser correlated (with am) variables have to do with the type of engine (horsepower, no of cylinders, no of carburators, V/S-engine, displacement). Consider the variable qsec (1/4 mile time, acceleration speed) to be a combined 'representation' of the strength of the engine. The remaining variables: rear axle ratio (drat), the no of forward gears (gear) and weight (wt) are less or not related to the engine type and need to be taken into account. To determine which one of those variables are of influence, ANOVA is used to provide a variance table through fitting nested models.

```
fit1 <- lm(mpg~am)                          ## fitting just am
fit2 <- lm(mpg~am+wt)                       ## fitting am and weight
fit3 <- lm(mpg~am+wt+qsec)                  ## fitting am, weight and 1/4-mile time
fit4 <- lm(mpg~am+wt+qsec+drat)             ## fitting am, weight, 1/4-mile time and rear axle ratio
fit5 <- lm(mpg~am+wt+qsec+drat+gear)        ## fitting also no of gears
anova(fit1, fit2, fit3, fit4, fit5)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + wt
## Model 3: mpg ~ am + wt + qsec
## Model 4: mpg ~ am + wt + qsec + drat
## Model 5: mpg ~ am + wt + qsec + drat + gear
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 720.90
## 2     29 278.32  1    442.58 68.8055 8.913e-09 ***
## 3     28 169.29  1    109.03 16.9510 0.0003442 ***
## 4     27 167.89  1      1.40  0.2176 0.6447654
## 5     26 167.24  1      0.65  0.1006 0.7537011
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The resulting table shows that adding weight and qsec has a high significance (using the normal error level ($\alpha = 0.05$) and therefor should be added to the final model. Gear and drat do not add much and are dropped.

## 5. Model fitting

The final model is a linear regression comparing transmission types on mpg (as outcome) while keeping weight en 1/4-mile time fixed on their average.

```
fit <- lm(mpg ~ am + wt + qsec)   ## fitting the linear model
round(summary(fit)$coef,4)       ## looking at the results
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.6178     6.9596  1.3819   0.1779
## am1           2.9358     1.4109  2.0808   0.0467
## wt           -3.9165     0.7112 -5.5069   0.0000
## qsec          1.2259     0.2887  4.2467   0.0002
```

The resulting table shows that the expected value for mpg is 9.62 for automatic transmission cars with an average weigth & 1/4-mile time. Manual transmission generates an increase of 2.94 mpg. The significance of the relationship between am and mpg, while taking wt and qsec into account is just slightly better than the normal error level ($\alpha = 0.05$). So in essence the hypothesis $H_0 = 0$ that claims that there is no relationship at all between am and mpg, is rejected. The adjusted R-squared is 0.8336, so approx. 83% of the total variance is explained by the model used.

```
confint(fit)
```

```
##                    2.5 %    97.5 %
## (Intercept) -4.63829946 23.873860
## am1          0.04573031  5.825944
## wt          -5.37333423 -2.459673
## qsec         0.63457320  1.817199
```

The resulting confidence interval for the increase going from automatic to manual is 0.05 to 5.83 mpg. With slightly more than 95% certainty we can predict that manual transmission will add mileage to the gallon.

## 6. Checking the residuals

Figure 2 in the appendix shows 4 residuals plots, to enable residual analysis. Residuals vs. Fitted is nicely scattered with no clear patterns and no hetero-skedasticity. The Normal Q-Q plot, appears to be pretty straight, showing that the residuals are mainly normally distributed. The Residuals vs. Leverage plot shows no points that go towards Cook's distance, indicating there are no real influential outliers.

## APPENDIX

**Figure 1. The correlation between all variables of mtcars**

```
g = ggpairs(mtcars, lower=list(continuous="smooth"))
g
```

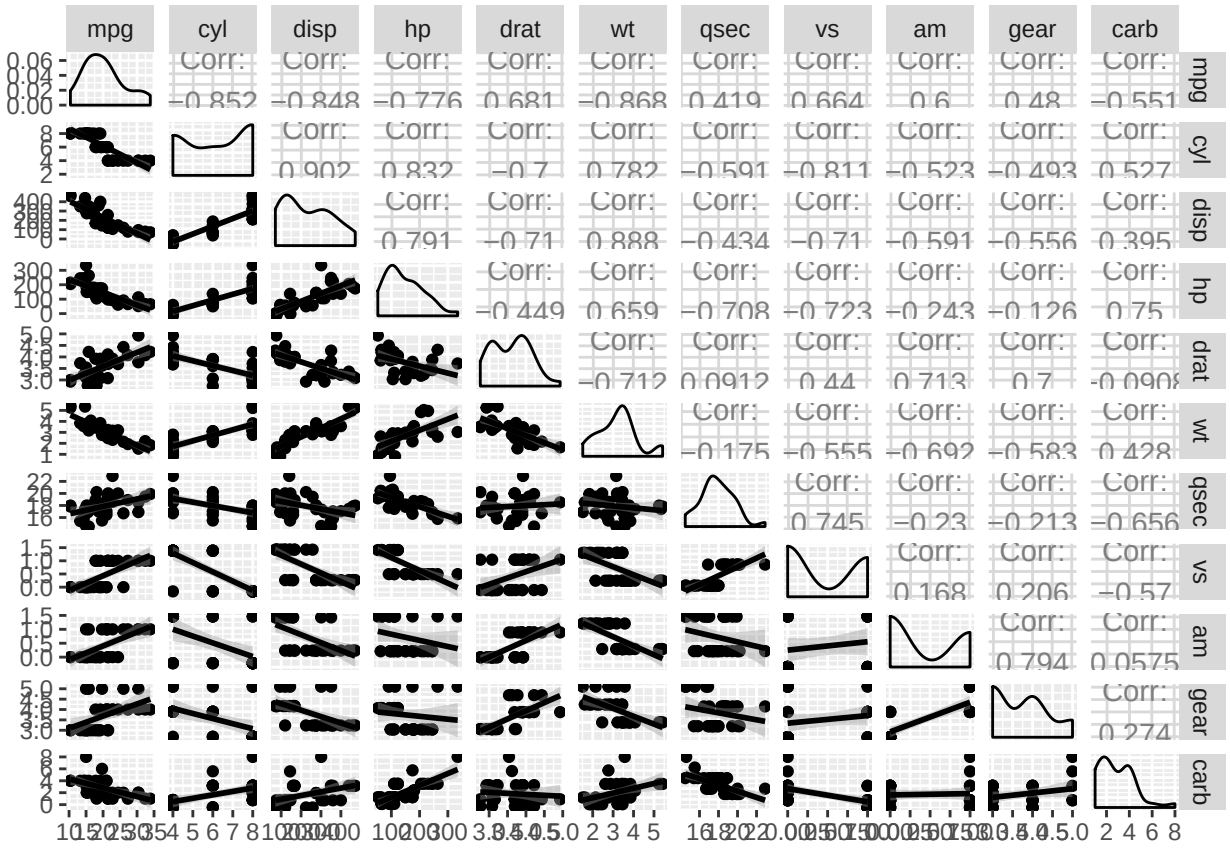

**Figure 2. Residual-plots for fit**

```
par(mfrow=c(2,2))    ## to plot 4 plots on one page
plot(fit)            ## plot the 4 standard residual plots
```