# Analysis Motor Trend

Ben Kooi

11-11-2021

## Synopsis

This report describes the analysis of the reach (miles per gallon) of 32 different type of cars. The dataset that is used is the mtcars-dataset from the library 'datasets'. The dataset seems to be clean and there are no NA's or NaN's. Imputation is not needed. The variables am (transmission) and vs (engine-type) are categorical, so these will be transformed as factors.

Questions that will be answered in this analysis are:
1. Is an automatic or manual transmission better for MPG?
2. Quantify the MPG difference between automatic and manual transmissions?

De most important conclusions are:
1. In average, cars with automatic transmission have more reach (miles per gallon) than with automatic transmission;
2. If we model the dependent variable mpg only with the predictor variable am (mpg ~ am), manual transmission has a advantage of 7.245 miles per gallon with a standard error of 1.764 miles per gallon versus automatic transmission.

The steps taken to find the best model with at least transmission as a predictor for the dependent variable mpg are:
1. determine a statistically significant difference in miles per gallon between automatic- and manual transmission;
2. calculate correlation-coefficients of all variables and plot a correlation-matrix;
3. determine if there is (multi)collineairity in the dataset;
4. fit different models with mpg (miles per gallon) as dependent variable and at least am (transmission) as predictor variable;
5. determine the best fitted and most simple model.

The formula mpg ~ am + wt + qsec seems to give a very useful model with 85% of the variance (R-squared) explained by the model with a p-value of 1.21e-11 in the F-statitics. Collecting ore data could improve normality.

## Configuring the environment for the analysis

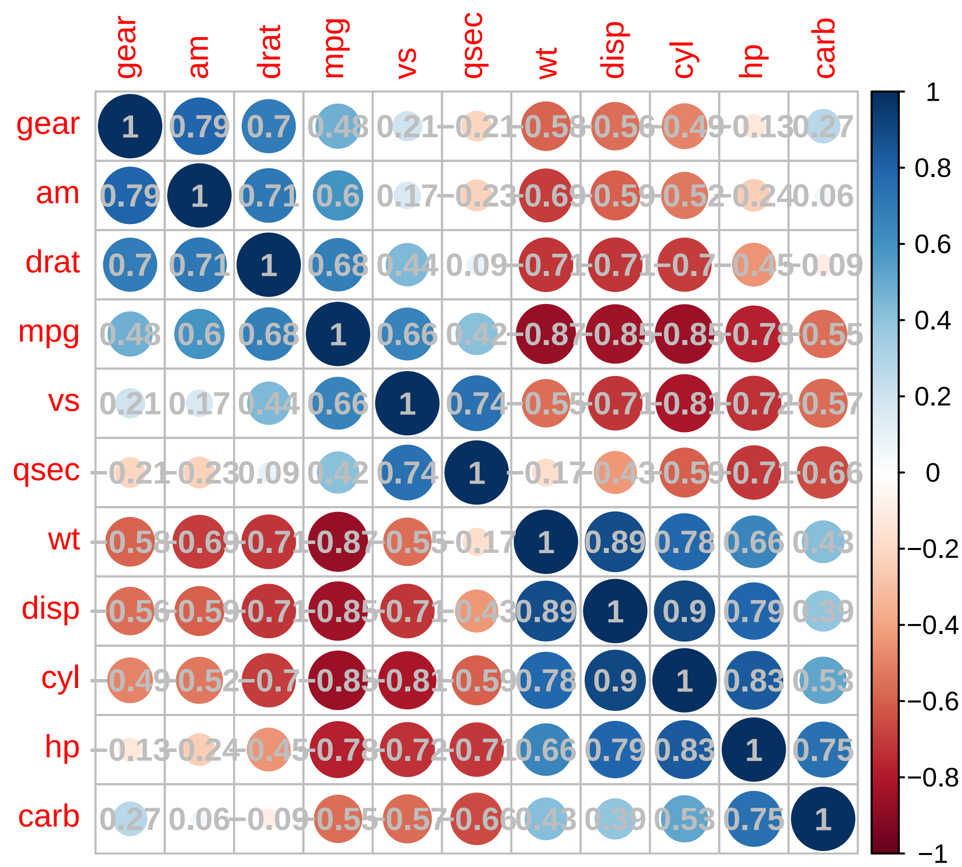
Loading R-packages and configuring the environment..

```
##     dplyr  corrplot   ggplot2 tidyverse     stats    modelr       zoo   tsibble
##      TRUE      TRUE      TRUE      TRUE      TRUE      TRUE      TRUE      TRUE
##       car  EnvStats
##      TRUE      TRUE
```

## Correlation

First we generate a correlation-matrix of all variables in the dataset.

```r
# For using cor() the factors must be transformed to numeric.
mtcars2 <- mtcars
mtcars2$am <- as.numeric(mtcars2$am)
mtcars2$vs <- as.numeric(mtcars2$vs)

# Calculate the correlation-coefficents
cm<-cor(mtcars2)

# Plot a correlation-matrix
corrplot(cm, order = "AOE",method = "circle",addCoef.col = "gray" , insig = "p-value")
```




Inspecting the correlation-matrix tells:
* a significant positive correlation of transmission (am) and miles per gallon (mpg). Overall a car with manual transmission gives you more miles per gallon;
* weight (wt) has the strongest correlation with miles per gallon. The more a car weighs, the less miles per gallon;
* transmission (am) is negative correlated with weight (wt). Overall a car with automatic transmission weighs less than with manual transmission;
* there seems to be a lot of multicollinearity.

Modeling the dependent variable with all other variables as predictors and calculate the variation inflation factors of all the predictors, confirms multicollinearity.

```r
# model dependent mpg with all other variables as predictors
model1 <- lm(mpg ~ ., data=mtcars)

# calculate variation inflation factors
vif(model1)
```

```
##       cyl      disp        hp      drat        wt      qsec        vs        am
## 15.373833 21.620241  9.832037  3.374620 15.164887  7.527958  4.965873  4.648487
##      gear      carb
##  5.357452  7.908747
```

The strategy for selecting the predictors for modelling the dependent variable mpg is:
1. At least the variable of interest am (transmission) is selected;
2. The most significant correlation-coefficient with is selected, which is wt (weight);
3. The correlation-coefficient with no or less collinearity with am and wt is selected, which is qsec (1/4 mile time).

## Linear modelling MPG

For selecting the best model, four different models are created by adding one predictor at the time. Then the ANOVA-test is used to compare the models.

```r
# create different models by adding one variable at the time
model2 <- lm(mpg ~ am, data=mtcars)
model3 <- lm(mpg ~ am + wt, data=mtcars)
model4 <- lm(mpg ~ am + wt + qsec, data=mtcars)
model5 <- lm(mpg ~ am + wt + qsec + disp, data=mtcars)

# determine coefficient and standard error for transmission
summary(model2)$coeff[2]
```

```
## [1] 7.244939
```

```r
summary(model2)$coeff[2,2]
```

```
## [1] 1.764422
```

```r
# compare the performance of the models
anova(model2,model3,model4,model5)
```

```
## 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 + disp
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 720.90
## 2     29 278.32  1    442.58 71.9811  4.26e-09 ***
```

```
## 3     28 169.29  1    109.03 17.7334 0.0002527 ***
## 4     27 166.01  1      3.28  0.5328 0.4717085
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The best model seems to be model4. Adding more variables to model4 will not improve the model. This is checked by model5.

```r
# calculate variation inflation factors of model4
vif(model4)
```

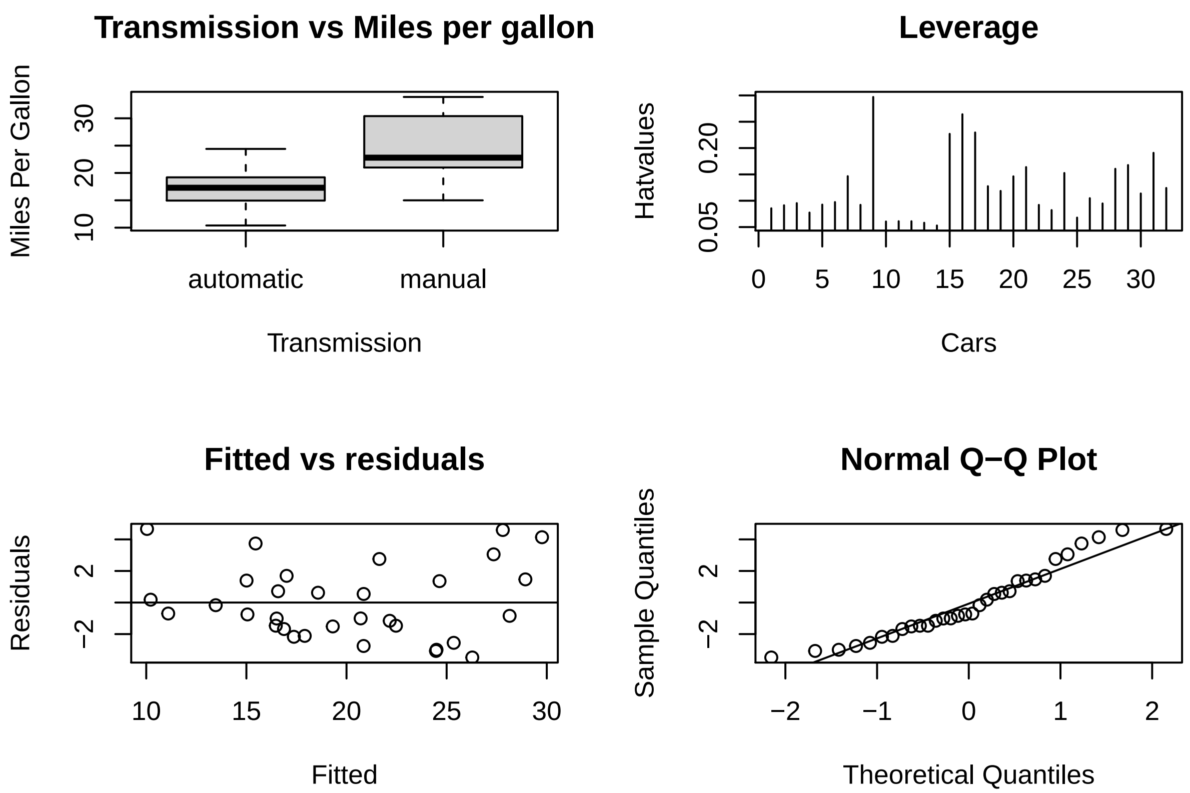
```
##       am       wt     qsec
## 2.541437 2.482952 1.364339
```

```r
# summary of the model
summary(model4)
```

```
##
## Call:
## lm(formula = mpg ~ am + wt + qsec, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4811 -1.5555 -0.7257  1.4110  4.6610
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.6178     6.9596   1.382 0.177915
## ammanual      2.9358     1.4109   2.081 0.046716 *
## wt           -3.9165     0.7112  -5.507 6.95e-06 ***
## qsec          1.2259     0.2887   4.247 0.000216 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.459 on 28 degrees of freedom
## Multiple R-squared:  0.8497, Adjusted R-squared:  0.8336
## F-statistic: 52.75 on 3 and 28 DF,  p-value: 1.21e-11
```

```r
# collect the residuals form the model
res<-resid(model4)

par( mfrow = c(2,2) )

# Plot a boxplot of mpg vs am
boxplot(mpg ~ am, mtcars, main="Transmission vs Miles per gallon",
   xlab="Transmission", ylab="Miles Per Gallon")

# plot leverage values for each observation for checking outliers
plot(hatvalues(model4), type = 'h', main="Leverage", xlab="Cars", ylab="Hatvalues")

#produce residual vs. fitted plot
plot(fitted(model4), res, main="Fitted vs residuals", xlab="Fitted", ylab="Residuals")
```

```r
#add a horizontal line at 0
abline(0,0)

#create Q-Q plot for residuals
qqnorm(res)

#add a straight diagonal line to the plot
qqline(res)
```




```r
# test if the model is a normal distribution
shapiro.test(res)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  res
## W = 0.9411, p-value = 0.08043
```

Checking the variation inflation factors tells us that there is no multicollinearity in this model. The summary of the model says that 85% of the variance is explained by the model with a p-value much less than 0.05. There are no hatvalues with a value>2, so there are no outliers with a big influence on the regression-line. The plot of the residuals versus the fitted model shows no pattern which is a good sign. The QQ-plot shows imperfection in normality of the distribution. The shapiro-test of the mode confirms this with p-value of 0.08, which is just a bit greater than 0.05.

Overall the model seems to be useful, but collecting more data could improve normality.