Regression Analysis of the *mtcars* Data Set

Executive Summary

The purpose of this analysis is to gain insight on the relationship between the MPG (miles per gallon) and other variables in the *mtcars* data set.

This analysis will show: 1. The transmission type is not the only factor that can generally explain whether mpg is lower or higher in a car, and 2. we find that - on average - mpg is lower by a factor of 0.7 for automatic transmission cars in comparison to manual ones.

```
data(mtcars)
```

Processing The data set contains records for 32 different types of cars. In total, there are 11 variables to be found. The codebook is included with the help pages for mtcars.

Using visual inspection and small tests we can convince ourselves that no further cleaning is necessary.

In order to improve readability and allow for partitioning of the data set into manual and automatic transmission types, we transform the "am" varaible into a factor variable and re-label the factors:

```
mtcars$am <- as.factor(mtcars$am)
levels(mtcars$am) <- c("Automatic", "Manual")
```

Exploring A first quick idea would be to compare the mean mpg for both transmission types:

```
aggregate(mpg~am, data = mtcars, mean)
```

```
##          am   mpg
## 1 Automatic 17.15
## 2    Manual 24.39
```

It appears that the mpg is lower (i.e. higher fuel consumption) for cars with an automatic transmission. This is further supported by Figure 1 in the appendix.

Additionally, a t-test with a 95% confidence interval can be performed to verify that the difference in means is significant and the null hypothesis - i.e. there is no siginificant difference in the means - can be rejected:

```
t.test(mpg ~ am, data = mtcars)
```

```
##
##  Welch Two Sample t-test
##
## data:  mpg by am
## t = -3.767, df = 18.33, p-value = 0.001374
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.28  -3.21
## sample estimates:
## mean in group Automatic    mean in group Manual
##                   17.15                   24.39
```

Since there are many other variables in the data set, however, we cannot be sure that the transmission type is the only contributing factor for mpg.

Linear Regression Model Building Based on the results above, it might seem useful to try and build a linear regression model:

```
lmfit <- lm(mpg~am, data = mtcars)
summary(lmfit)
```

```
##
## Call:
## lm(formula = mpg ~ am, data = mtcars)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.392 -3.092 -0.297  3.244  9.508
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    17.15       1.12   15.25  1.1e-15 ***
## amManual        7.24       1.76    4.11  0.00029 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.9 on 30 degrees of freedom
## Multiple R-squared:  0.36,   Adjusted R-squared:  0.338
## F-statistic: 16.9 on 1 and 30 DF,  p-value: 0.000285
```

The summary yields that only 33.8% of variability around the mean in the resonse data is explained by this model. This clearly shows that the type of transmission on its own is not enough. In the next steps we determine a better model with a step-wise process:

```
allmodel <- lm(mpg ~ ., data = mtcars)
sink("/dev/null"); best <- step(allmodel, direction = "backward"); sink()
summary(best)
```

```
##
## Call:
## lm(formula = mpg ~ wt + qsec + am, data = mtcars)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.481 -1.556 -0.726  1.411  4.661
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    9.618      6.960    1.38  0.17792
## wt            -3.917      0.711   -5.51    7e-06 ***
## qsec           1.226      0.289    4.25  0.00022 ***
## amManual       2.936      1.411    2.08  0.04672 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.46 on 28 degrees of freedom
## Multiple R-squared:  0.85,   Adjusted R-squared:  0.834
## F-statistic: 52.7 on 3 and 28 DF,  p-value: 1.21e-11
```

This seems to suggest, that weight, 1/4 mile time, and transmission type are actually the main contributing factors to mpg. This model is better than the previous one, as it explains 83.4% of the variability around the mean. Running an analysis of variance proves this to be the case:

```
anova(lmfit, best)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ wt + qsec + am
##   Res.Df RSS Df Sum of Sq    F  Pr(>F)
## 1     30 721
## 2     28 169  2       552 45.6 1.6e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Diagnostics The plottet residuals in Figure 2 in the appendix seem to be randomly scattered, which is what we are expecting. As regards the outliers seen in the plots, we can use hatvalues and dfvalues to further look into their respective contribution to the model parameters:

```
leverage <- hatvalues(best)
tail(leverage, 3)
```

```
##  Ferrari Dino Maserati Bora    Volvo 142E
##        0.1138        0.1910        0.1243
```

```
infp <- dfbetas(best)
tail(infp, 3)
```

```
##                (Intercept)         wt     qsec amManual
## Ferrari Dino     -0.05799  3.523e-06  0.07688 -0.04804
## Maserati Bora    -0.03977 -1.325e-01  0.12037 -0.16843
## Volvo 142E        0.31198 -2.543e-01 -0.28378 -0.41565
```

Results Based on the analysis results, we can answer the questions presented in the first paragraph as follows:

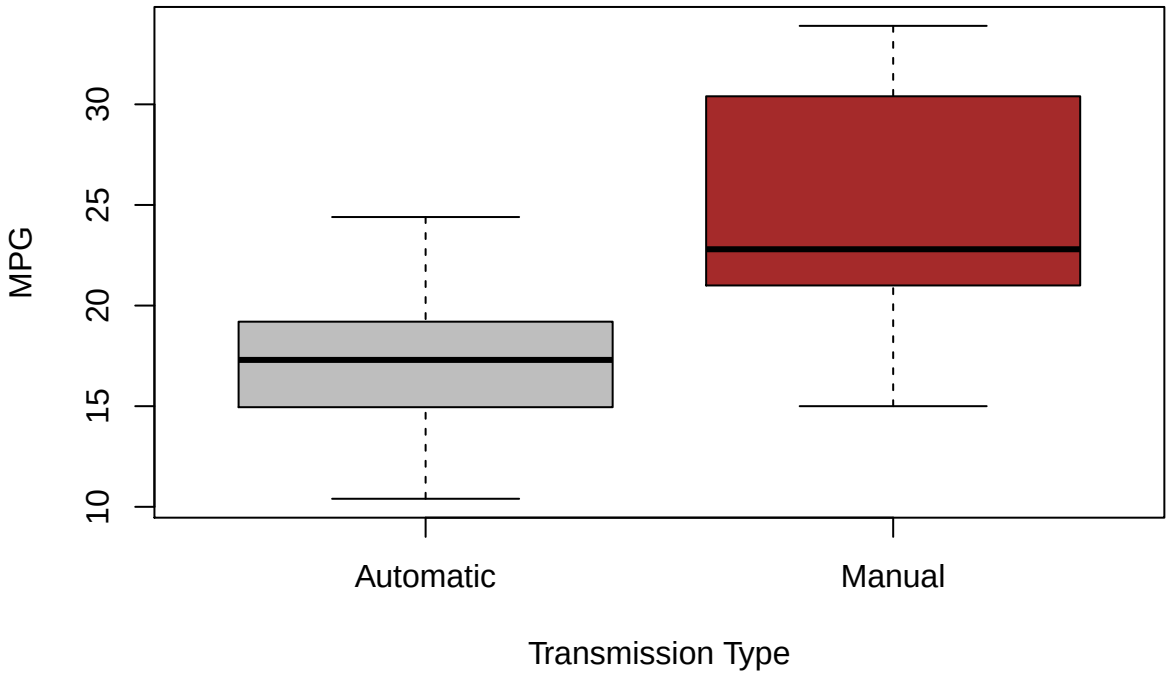
1. It is not generally true to say that, for mpg, manual transmission is better than automatic transmission as there are other factors, e.g. weight, which needs to be taken into account.
2. For the second question, we use the model to calculate the average factor by which automatic transmission mpg is lower than manual transmission one:

```
automatic <- row.names(mtcars[mtcars$am == "Automatic", ])
manual <- row.names(mtcars[mtcars$am == "Manual", ])
p <- predict(best)
mean(p[automatic]) / mean(p[manual])
```

```
## [1] 0.703
```

Appendix

Fig. 1 − MPG vs. Transmission Type



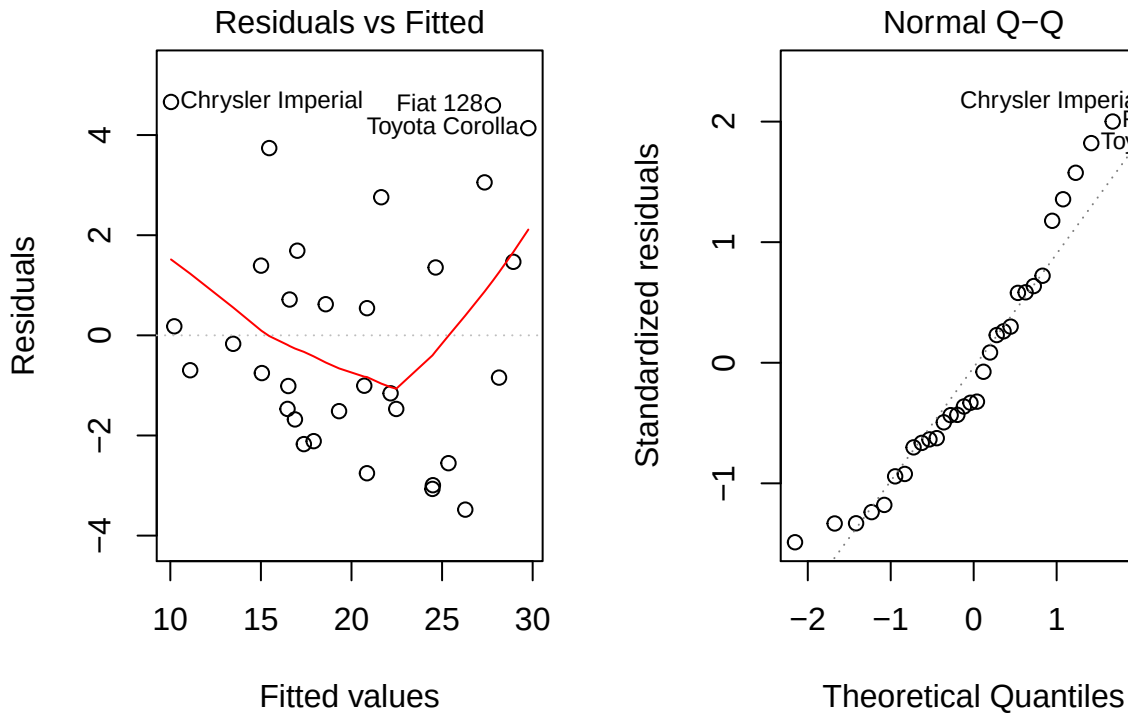


Figure 2 - Diagnostics