RegressionModels Assignment

Executive Summary

In this analysis, we look at the mtcars data set and want to answer the questions: “Is an automatic or manual transmission better for MPG?”, “Quantify the MPG difference between automatic and manual transmissions?” The result of the “best” interaction model of this analysis is to be interpreteted as follows, and answers both questions. For a constant weight and horse power, a car with manual transmission have the difference of ‘11.555 - 3.578 * wt’ in miles per gallon more than automatic transmitted cars. That effectively means, that lighter cars have more mpg when being a manual transmission, and heavier cars will have a higher MPG with automatic transmission. The break-even value is at around around 3229 lbs.

Loading and Processing the data

```
#load data mtcars data set
library(datasets)
data(mtcars)
#str(mtcars) # not compiled for space reasons
```

All variables are numeric variable, although specific variable should represent a specific type, like cyl, V/S, am, gear, or carb. These variable are changed to factor variables. Before, a correlation matix is saved for later.

```
# save correlation matrix
correlation <- cor(mtcars)
#load data mtcars data set
mtcars$cyl <- factor(mtcars$cyl); mtcars$vs <- factor(mtcars$vs);mtcars$gear <- factor(mtcars$gear);
mtcars$carb <-factor(mtcars$carb);mtcars$am <- factor(mtcars$am)
#str(mtcars) # not compiled for space reasons
```

Exploratory Analysis

Let’s show the first three rows of the data set plus its dimensions.

```
# explore data set
head(mtcars,3)
dim(mtcars)
```

With a boxplot in the appendix (plot1), we visualize the distribution of MPG for both automatic (am = 0) and manual transmission (am = 1). One can see that for all examples, the MPG values for automatic transmission are less spread with an average value below 20, while the manual transmission is wider spread with an average value above 20. The exact mean values are

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

The mean MPG of automatic transmission is -7.245 less than manual transmission.

Statistical Inference

Next we check whether the difference between both transmissions is significant with a null hypothesis.

```
t.test(mtcars$mpg ~ mtcars$am)$p.value
```

```
## [1] 0.001373638
```

The p-value is 0.00137, so that the null hypothesis can be rejected. Hence the difference between the MPG of both transmissions is significant.

Regression Analysis

The simple linear model is fitted for MPG of only the variable am.

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

Let's include all variales.

```
full_model <- lm(mpg ~ ., data = mtcars)
#summary(full_model)
```

The full model has an adjusted R-squared value of 0.779, which means it can cover roughly 78 % of the variance of the MPG variable. That is a nicely high value, yet none of the coefficients are significant at to a 0.05 level. Looking at the correlation of MPG with all variables (see plot2 in Appendix)

```
#correlation[1,]
```

We see that MPG is less correlated with qsec, vs, gear and carb (if taking am for granted.). Let us run a model with all higher correlation values.

```
highCorModel <- lm(mpg ~ wt + cyl + disp + hp + am, data = mtcars)
#summary(highCorModel)
```

The full model has an adjusted R-squared value of 0.8344, and some coefficients are significant to a 0.05 level. When removing the non significant variables being cyl and disp, one ends up with an intuitive problem, where weight and the horse power having a big influence on MPG.

```
IntuitionModel <- lm(mpg ~ wt + hp + am, data = mtcars)
#summary(IntuitionModel)
```

The adjusted R-squared value is 0.8227, and only am1 is not significant to a 0.05 level of significance.

When plotting the mpg as a function of weight or horse power depending on the transmission type (see plot 3 and 4 in appendix), it becomes clear that there is both a linear trend and a cluster type dependence for manual and automatic transmission. In order to investigate this cluster behavior we include variables as an interaction term with am.

```
InteractionModel <- lm(mpg ~ wt + hp + am + wt:am + hp:am, data = mtcars)
#summary(InteractionModel)
```

The adjusted R-squared value is 0.8561, and only the interaction between hp and am is not significant to a 0.05 level. Leaving this term out leads to

```
Interaction2Model <- lm(mpg ~ wt + hp + am + wt:am, data = mtcars)
#summary(Interaction2Model)
```

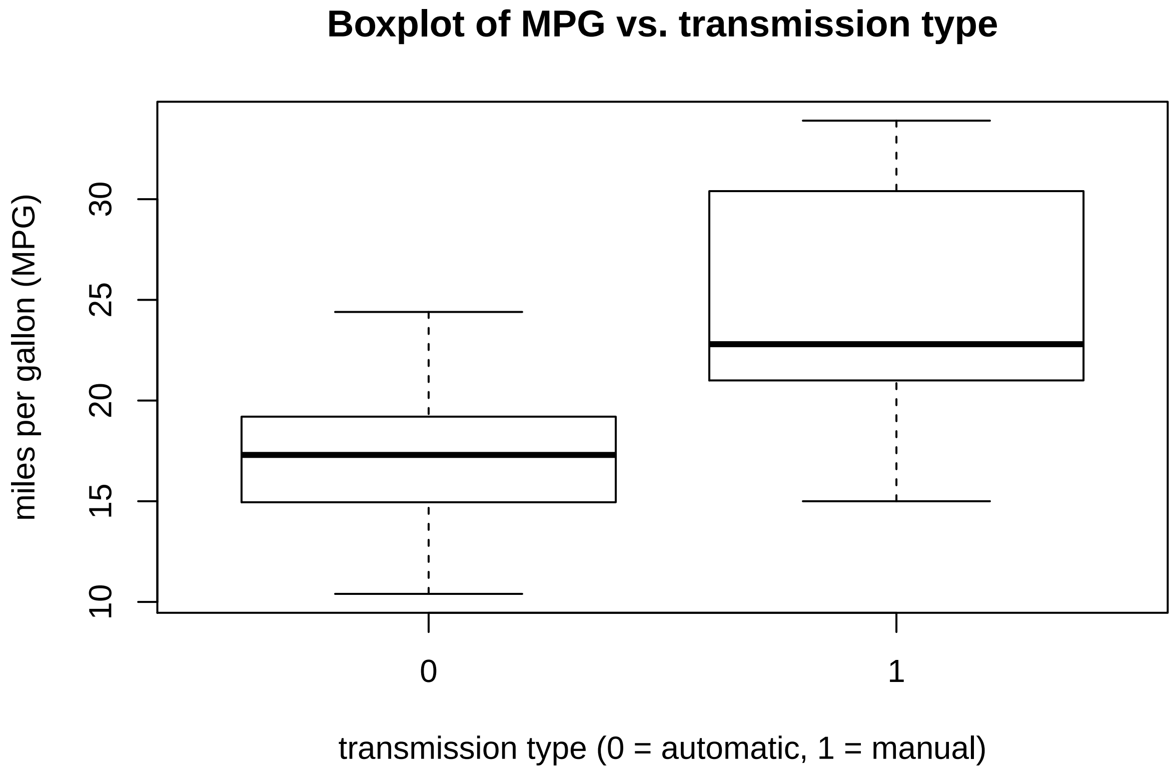
The adjusted R-squared value is 0.8503, and all fitting parameters to a 0.05 level of significance.

When comparing the residuals for both models (see plot 5 and 6 in appendix), the Interaction1model with higher adjusted R-squared value also has the residuals normally distributed and shows homoskedasticity in a cleaner manner than Interaction2Model. Overall the InteractionModel seems the best model.

Appendix

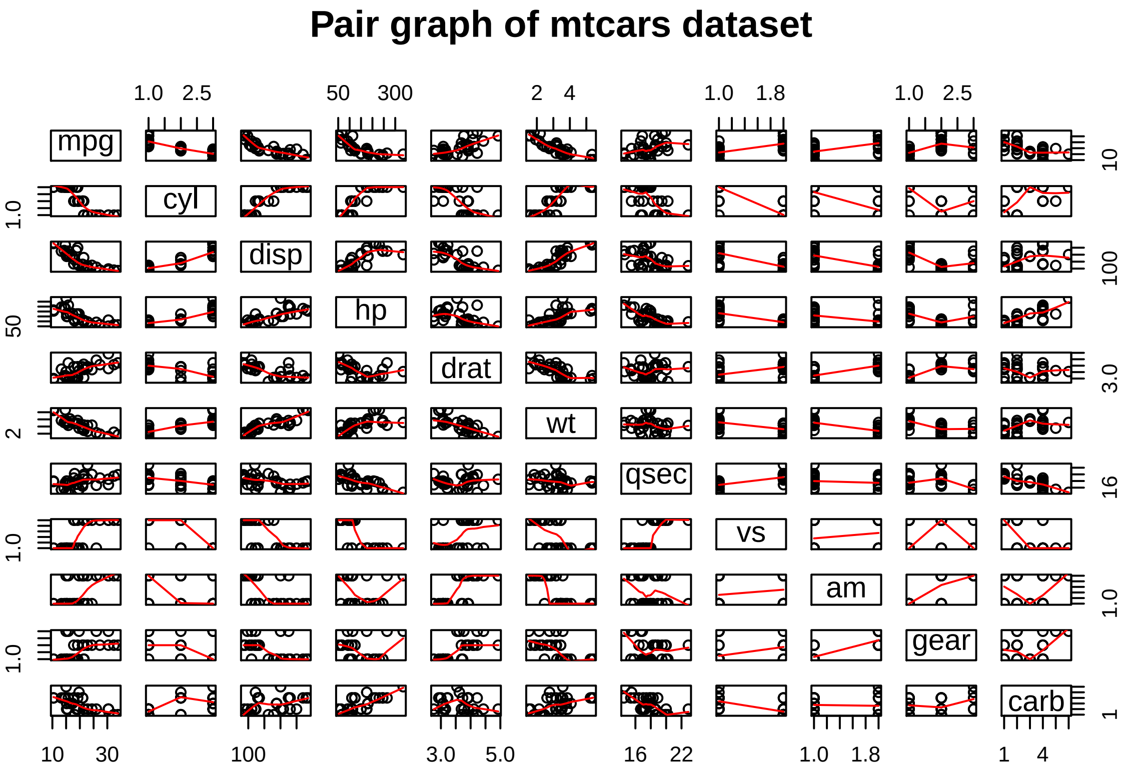
Plot 1: Boxplot of MPG for both transmission types

```
boxplot(mtcars$mpg ~ mtcars$am, xlab="transmission type (0 = automatic, 1 = manual)",
ylab="miles per gallon (MPG)",
main="Boxplot of MPG vs. transmission type")
```



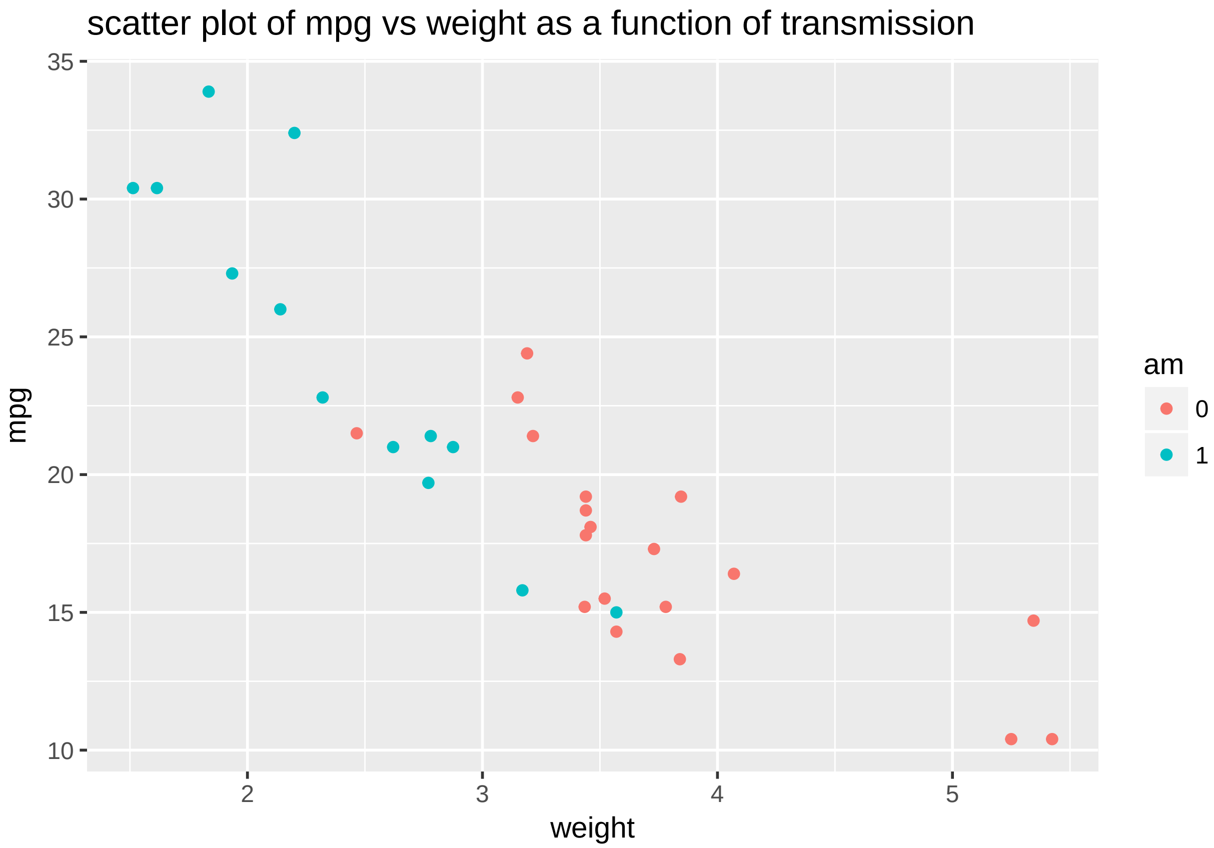
Plot 2: Pair graph of all variables

```
pairs(mtcars, panel = panel.smooth,main="Pair graph of mtcars dataset")
```



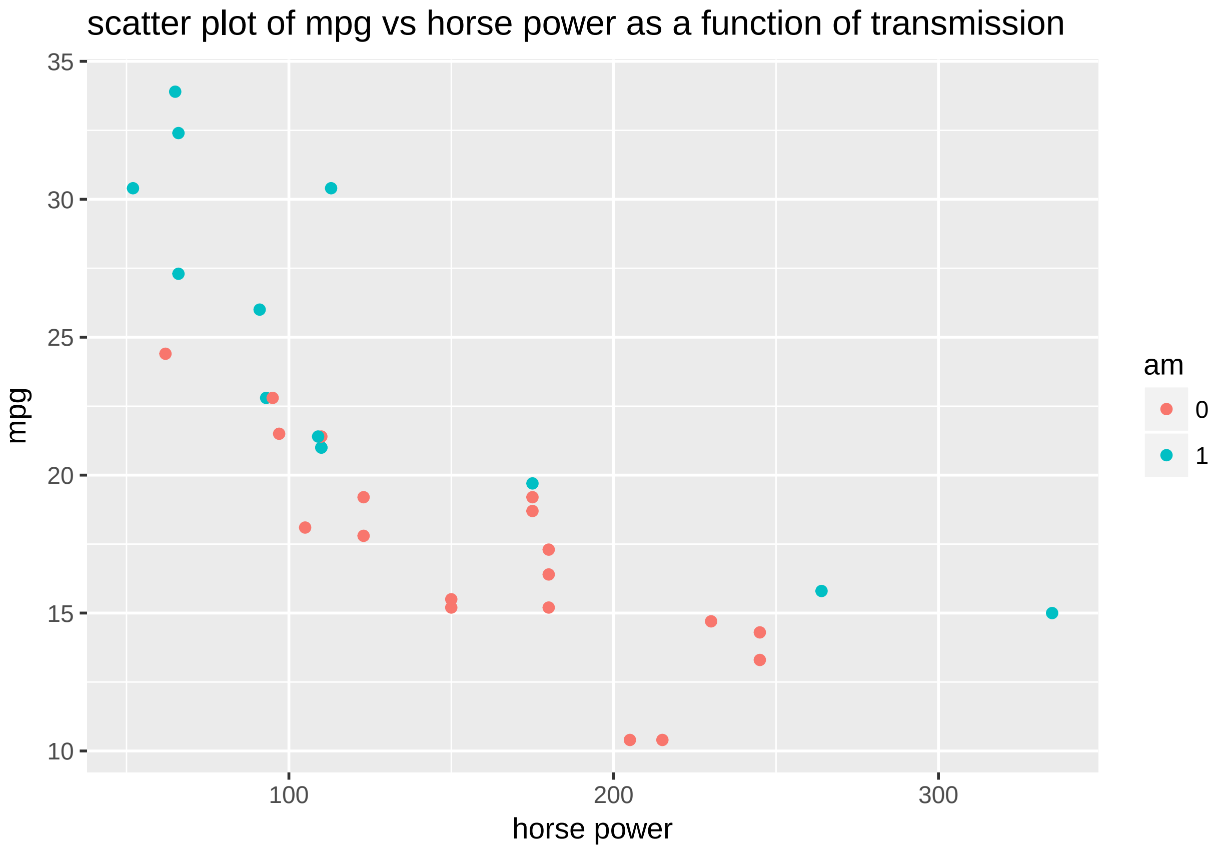
Plot 3: scatter plot of mpg vs weight as a function of transmission

```r
library(ggplot2)
ggplot(mtcars, aes(x=wt, y=mpg, group=am, color=am)) + geom_point() +
xlab("weight") + ggtitle("scatter plot of mpg vs weight as a function of transmission")
```



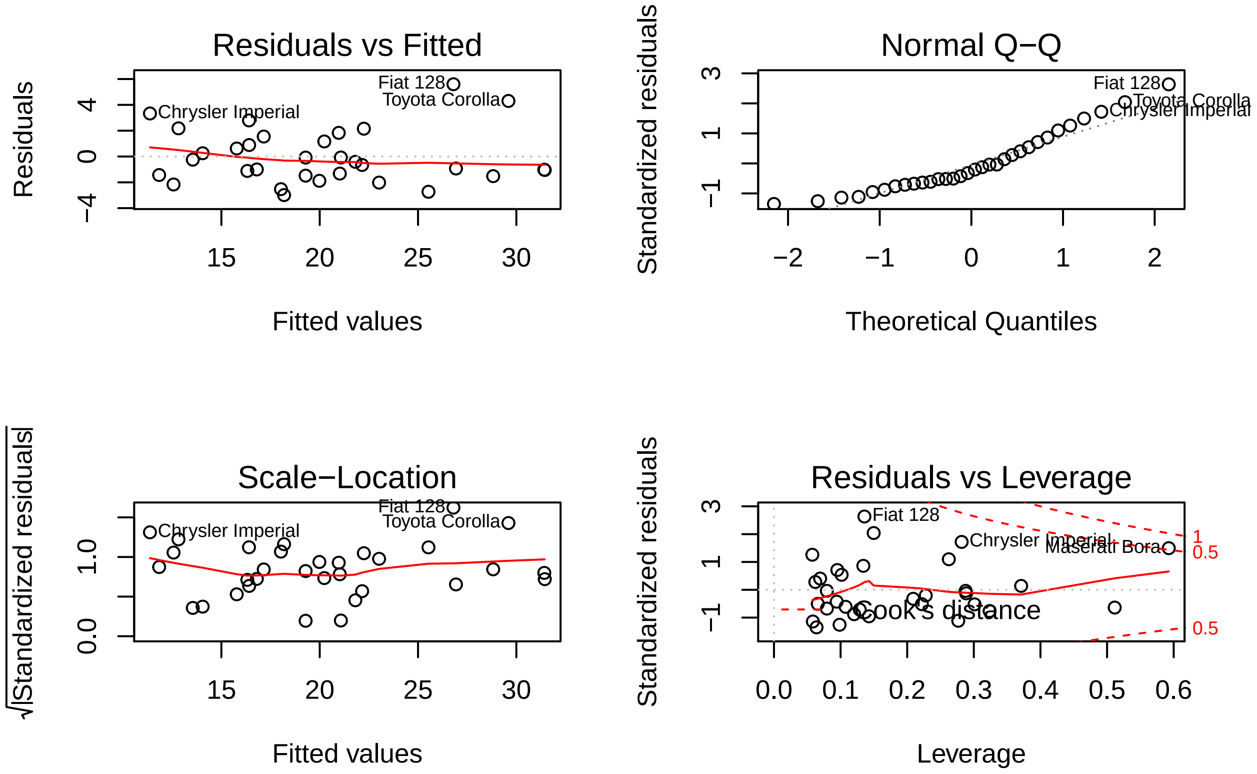
Plot 4: scatter plot of mpg vs horse power as a function of transmission

```
ggplot(mtcars, aes(x=hp, y=mpg, group=am, color=am)) + geom_point() +
xlab("horse power") + ggtitle("scatter plot of mpg vs horse power as a function of transmission")
```

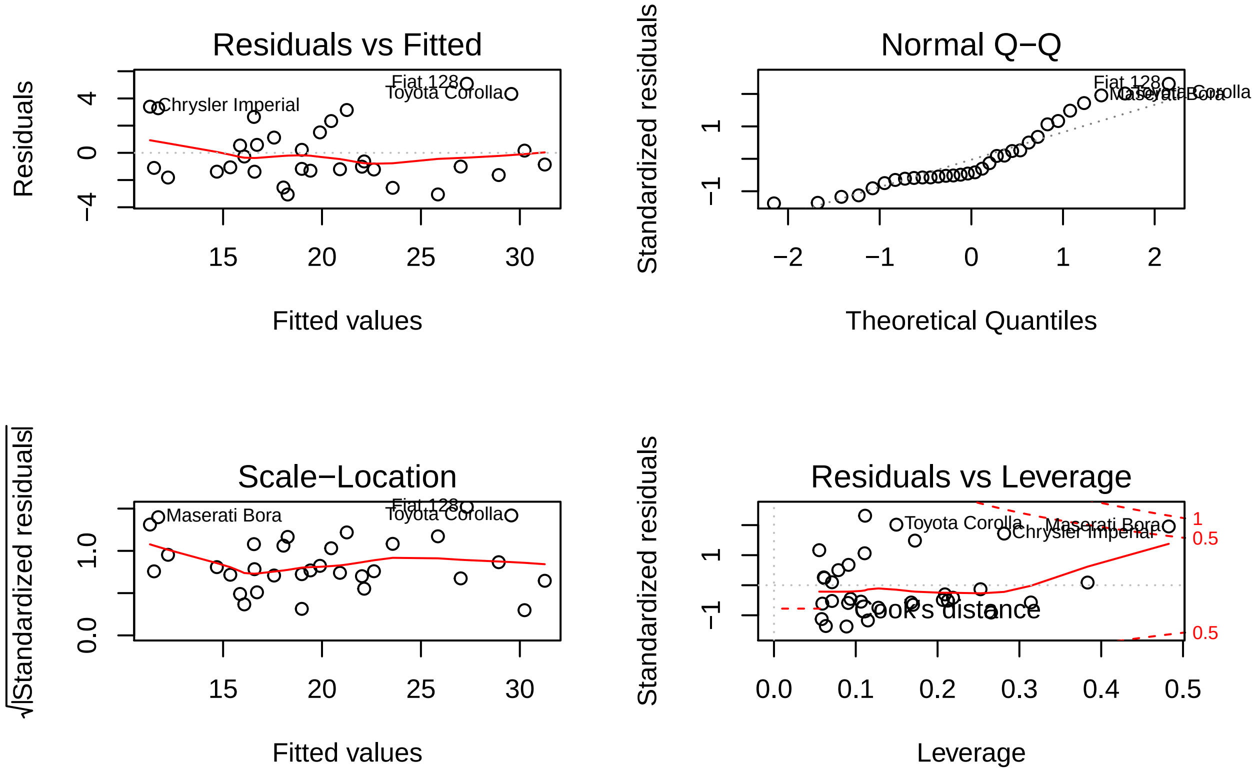


plot 5 und 6: residual plots of interaction1 and interaction2

```
par(mfrow = c(2,2))
plot(InteractionModel)
```



```r
par(mfrow = c(2,2))
plot(Interaction2Model)
```


Residuals vs Fitted
Residuals
Fitted values
Chrysler Imperial
Fiat 128
Toyota Corolla
−4
0
4
15
20
25
30
Normal Q−Q
Standardized residuals
Theoretical Quantiles
Fiat 128
Toyota Corolla
Maserati Bora
−1
1
−2
−1
0
1
2
Scale−Location
√|Standardized residuals|
Fitted values
Maserati Bora
Fiat 128
Toyota Corolla
0.0
1.0
15
20
25
30
Residuals vs Leverage
Standardized residuals
Leverage
Toyota Corolla
Maserati Bora
Chrysler Imperial
Cook's distance
1
0.5
0.5
−1
1
0.0
0.1
0.2
0.3
0.4
0.5