Regression Models: Peer Assessment

Executive Summary

This sheet explores the relationship between miles-per-gallon (MPG) and other variables in the mtcars data set. First, the analysis attempts to determine whether an automatic or manual transmission is better for MPG. At second part of document we find the best fit model for mpg outcome.

Answer for main question - Is an automatic or manual transmission better for MPG? The cars with manual transmission have a better fuel consumption and bigger mileage per gallon. The mean MPG difference between transmissions is equal 7.245 MPG.

Quick view on data

Load the data and quick view on it.

```
data (mtcars)
dim (mtcars)
```

```
## [1] 32 11
```

```
head (mtcars, 3)
```

```
##                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
```

The dataset has 32 objects of 11 variables.

Boxplot show us the strong difference in miles per gallon between automatic and manual transmission (see Appendix 1)

Linear model based on transmisison

```
fit <- lm (mpg ~ factor (am), mtcars)
summary (fit)
```

```
##
## Call:
## lm(formula = mpg ~ factor(am), data = mtcars)
##
## 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 ***
## factor(am)1    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
```

The beta0 coefficient (Intercept) is equal 17.147 and beta1 coefficient is equal 7.245. It means that the average MPG for cars with automatic transmission is 17.147 MPG, and the average MPG for cars with manual transmissions is beta0 plus beta1 = 24.392 MPG. The mean MPG difference between transmissions equal beta1, 7.245 MPG - the cars with manual transmission have a better fuel consumption and bigger mileage per gallon.

For residual plot for this model, please, see Appendix (2).

The 95% confidence intervals for beta0 and beta1:

```
sumCoef <- summary (fit)$coefficients
sumCoef [1,1] + c(-1, 1) * qt(.975, df = fit$df) * sumCoef [1, 2] # beta0
```

```
## [1] 14.85062 19.44411
```

```
sumCoef [2,1] + c(-1, 1) * qt(.975, df = fit$df) * sumCoef [2, 2] # beta1
```

```
## [1]  3.64151 10.84837
```

Multivariable linear model

Ok, let's try to find the best fit linear model using one of my favorite packages - caret (http://topepo.github.io/caret/index.html). Caret packages suggests a different feature selection methods, such as recursive feature selection, Genetic Algorithms, Simulated Annealing. Apply the simplest, RFE method.

First, create an rfeControl object, that specifies the details of the feature selection. Argument 'function' is set to 'lmFuncs', method to repeated cross validation and repeats is equal five. Don't forget to set seed for reproducible research.

```
library ('caret')
set.seed (1)
ctrl <- rfeControl (functions = lmFuncs, method = "repeatedcv",
                    repeats = 5, verbose = FALSE)
```

Second, call the rfe function, that fits the model. X is predictors (first column of data frame is MPG) and Y is an outcome. The 'sizes' arguments corresponds to the number of features that should be retained.

```
lmProfile <- rfe (x = mtcars [-1], mtcars [ , 1], sizes = 1:10, rfeControl = ctrl)

predictors (lmProfile)
```

```
## [1] "wt"   "am"   "qsec" "hp"
```

The last command tell us that the best model is with predictors weigth, A/M tansmission, 1/4 mile time and horsepower. Look at summary model.

```
lmfit <- lmProfile$fit
lmfit
```

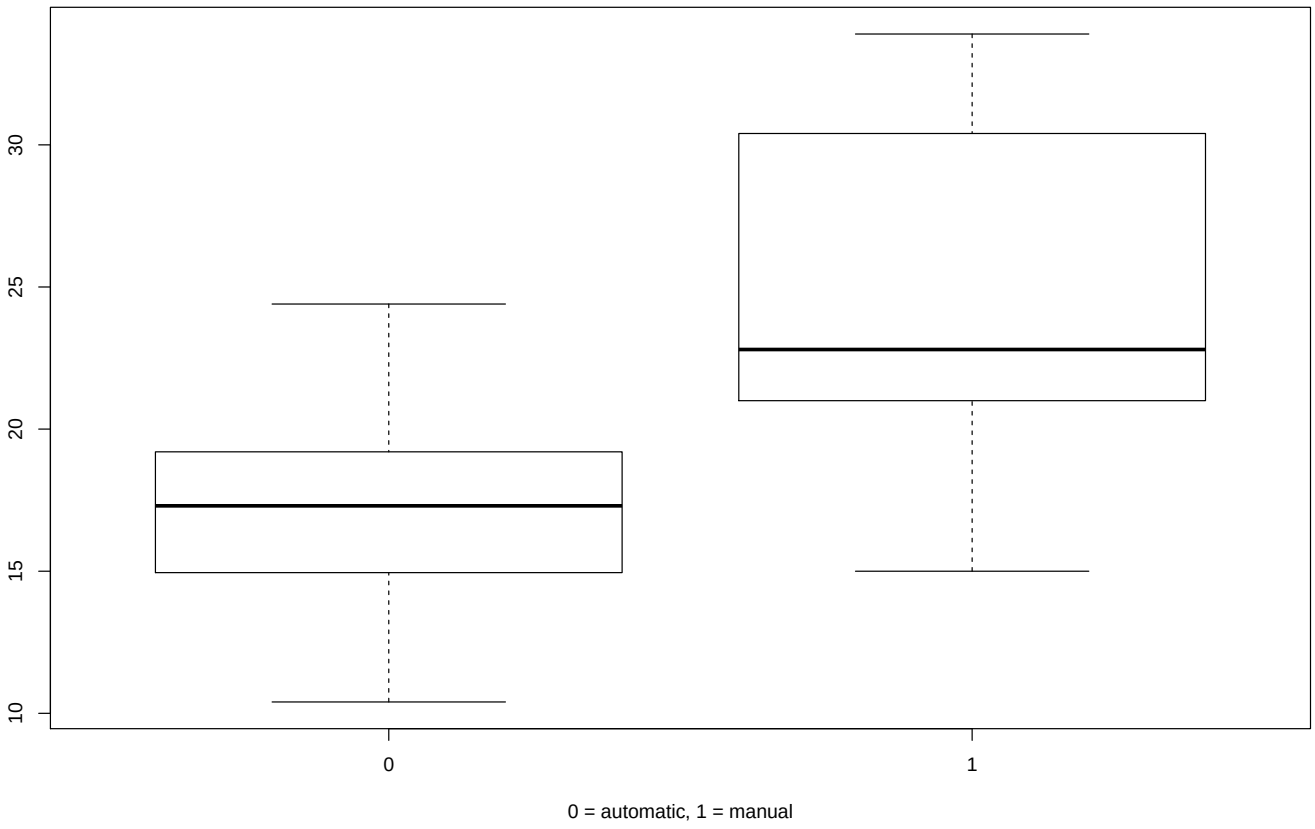
```
##
## Call:
## lm(formula = y ~ ., data = tmp)
##
## Coefficients:
## (Intercept)           wt           am         qsec           hp
##    17.44019     -3.23810      2.92550      0.81060     -0.01765
```

Diagnostic plots for fitting model. (see Appendix). Residuals vs Fitted looks good without any pattern. Normal QQ plot is on the line.

Appendix

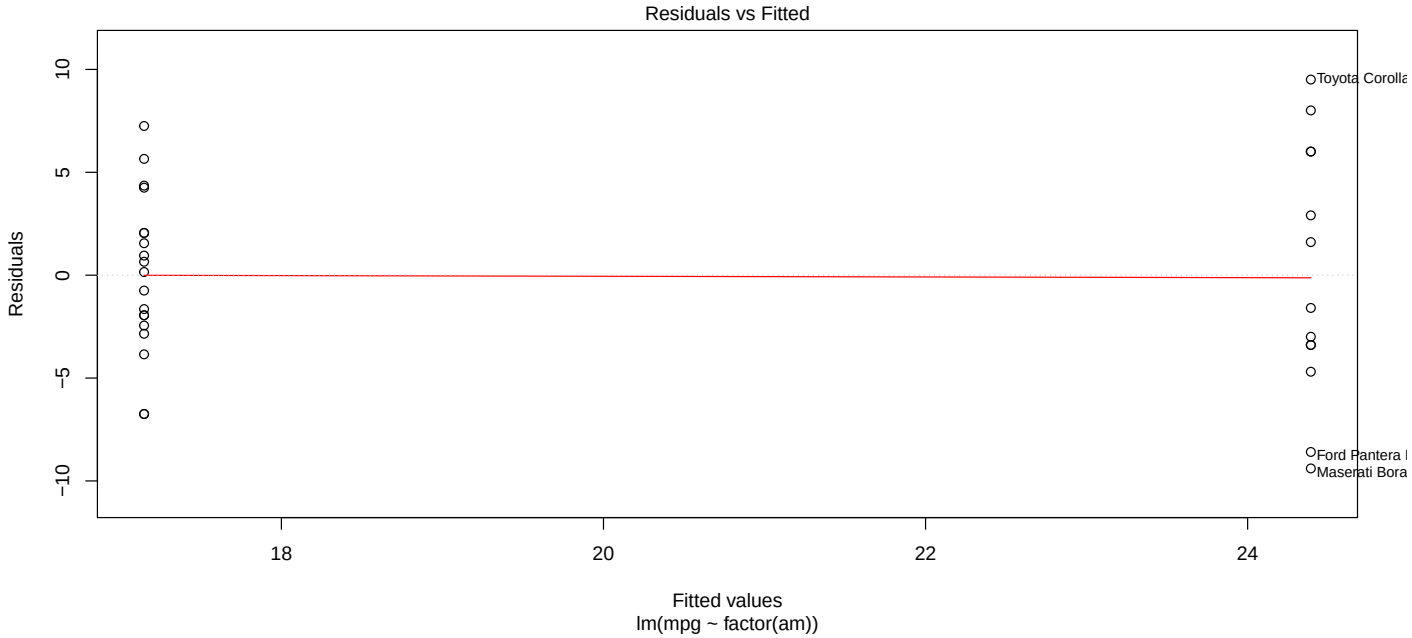
1.Boxplot MPG vs transmission.

```
with (mtcars, boxplot (mpg ~ am, xlab = "0 = automatic, 1 = manual"))
```



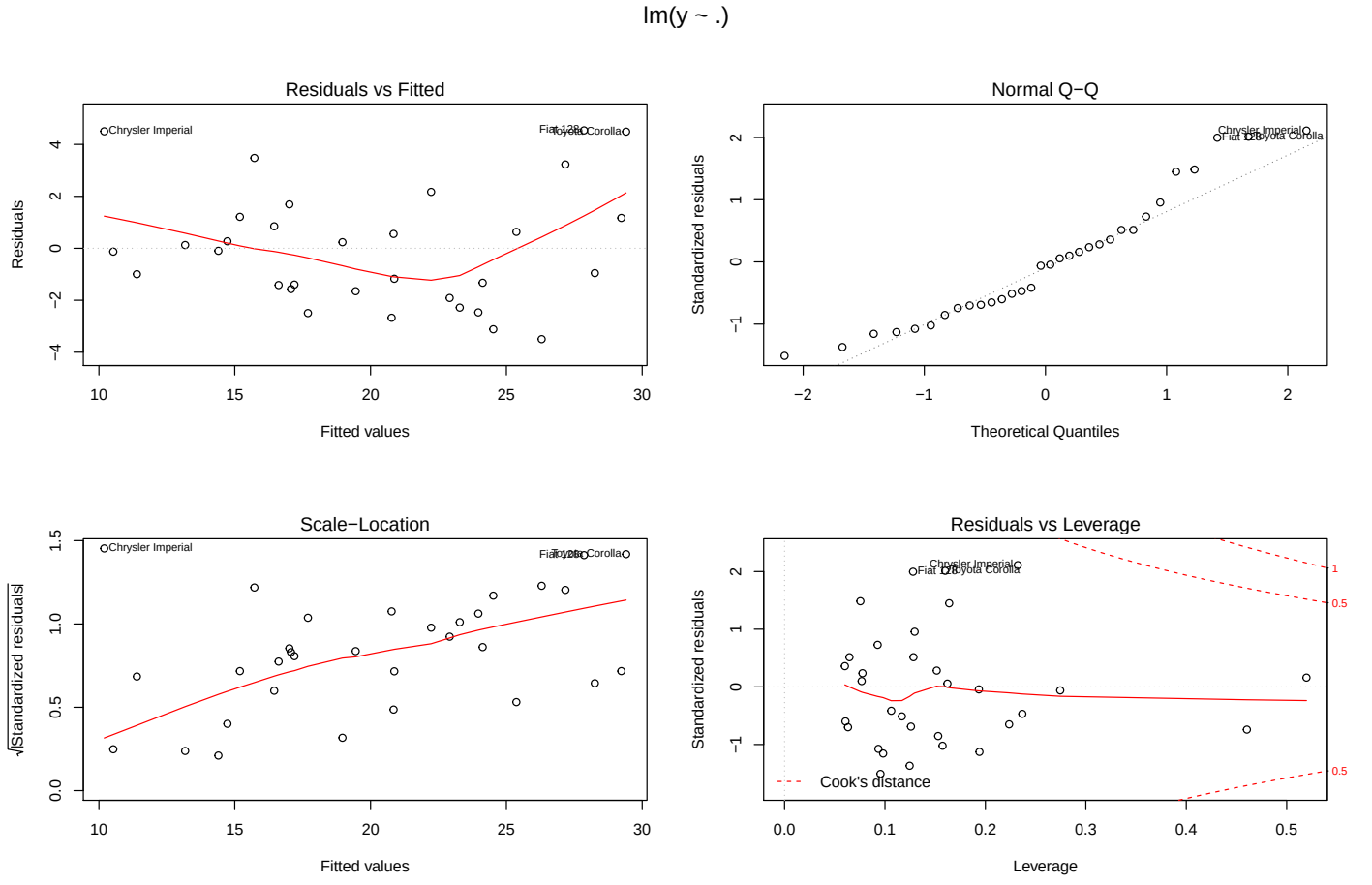
2. Residual plot for single linear model.

```
plot (fit, which = 1)
```



3. Diagnostic plots for fitting model. (see Appendix).

```
par (mfrow = c (2, 2), oma = c (0, 0, 2, 0))
plot (lmfit)
```



4. Best fit model feauture selection. Root Mean Squared Error (RMSE).

```
trellis.par.set(caretTheme())
plot(lmProfile, type = c("g", "o"))
```