Motor Trend - analysis of MPG data for automatic and manual transmissions

Regression Models Assignment (Part of Coursera - Johns Hopkins University - Data Science Specialization)

Author: Khadar vali

Executive Summary

The assignment examines a data set of a collection of cars and explores the relationship between a set of variables and miles per gallon (MPG).

The analysis comprises of these stages:

- Data Processing for the analysis.

- Exploratory Data Analysis of the processed data. This initially examines MPG and am (transmission type) and then looks for other variables that might influence any potential the MPG/am relationship. “qsec” (acceleration time) and “wt” (weight) were also found to have an influence.

- Model selection and examination. Look at simple linear regression model and a multi-variable regression model. Whilst the linear model provided a fit, a multi-variate model was found to be stronger option.

The analysis is focused on answering two questions:

1 - Is an automatic or manual transmission better for MPG?

2 - Quantify the MPG difference between automatic and manual transmissions.

- Conclusion. Both models support the conclusion that the cars in the data set with manual transmissions have higher miles per gallon (MPG) than those cars in the data set with automatic transmission systems. Manual transmission delivers 2.94 more mpg than automatic transmission (using the multi-variate model). However, other variables (weight and acceleration time) do have significant influence on this correlation and further investigation and multi-variate modelling is recommended.

Data Processing

```r
library(dplyr)

# load data

data("mtcars")

# prepare variable "am" (automatic and manual) - change to factor

mtcarsFact <- mtcars

mtcarsFact$am <- as.factor(mtcarsFact$am)

levels(mtcarsFact$am) <- c("Automatic", "Manual")
```

Exploratory Data Analysis

(Code and print-out of Exploratory Data Analysis results - see Appendix A)

There are 32 observations (rows) and each row is for a car type. There are 11 variables (columns).

The two variables of interest “am” (transmission type) and “MPG” (miles per gallon). Plot these two variables against each other to see if a visual analysis indicates a possible relationship. (Code and violin plot of “mpg” and “am” - see Appendix B)

The violin plot indicates that manual transmission has a higher mpg than automatic transmission. However, this is based on 32 observations and so is a relatively small sample size. A t-test is also carried out to test the significance of the relationship. The t-test is for the null hypothesis “there is no correlation between transmission type and mpg”.

```
# t-test for null hypothesis

tTest1 <- t.test(mpg~am,data=mtcarsFact); tTest1$p.value

## [1] 0.001373638
```

The p-value is 0.001373638. This is less than 0.05 and therefore the null hypothesis is rejected. The alternative hypothesis - a significant difference (correlation) of mpg between automatic and manual transmissions is now examined.

Model Selection

Linear regression model.

The t-test indicates there could be a significant difference between mpg for the two transmission types. The first model to be applied is a linear regression model - this will test the significance found in the above t-test and the associated ‘adjusted r-squared’ value will indicate if the linear model is optimal.

```
# Linear model fit and call adjusted R-squared

initialLinear <-lm(mpg~am, data=mtcarsFact)

summary(initialLinear)$adj.r.squared

## [1] 0.3384589
```

The adjusted r squared value of the linear model is 0.3384589 (it explains 33.8% of the variation) - this is rather low and so other models should be examined.

Model selection - Multi-variate

As a linear model is possibly not optimal, a multi-variate model is applied to see if it provides a better fit.

First, use the step function to find number of variables that are optimal. (Code and print-out of step function - see Appendix C)

The variables that provide an optimal fit are: “am”, “qsec”(acceleration time) and “wt” (weight). We generate a pairs plot of these optimal variables. (Code and pairs plot of “mpg” and optimal variables “am”, “qsec” and “wt”. - see Appendix D) The pairs plot shows that there a number of other correlations in addition to the “am” (transmission type) and “mpg” variables that has been established. These variables and correlations should be explored in a multi-variate model.

Multi-variate Model fit

(Code and print-out of multivariate model fit - see Appendix E)

The adjusted R squared value of 0.8336 is a significant improvement. Manual transmission delivers 2.9358 (round to 2.94) more mpg than automatic transmission The p-value is also below 0.05.

Diagnostics

A plot is generated to examine the model and to check the diagnostics. (Code and plot - see Appendix F)

Residual vs fitted plot gives is broadly horizontal line and no distinct pattern. This indicates a lack of any non-linear relationships (i.e. a broadly linear relationship) but there may be other variables that provide an influence.

Normal Q-Q plot shows no outliers and the results broadly following the dashed line. A normal distribution is the conclusion from this plot.

Scale-Location plot shows a random distribution but the line is not horizontal. Conclusion - variance would appear to be equal but it is very possible there are other influences (variables).

Residuals vs Leverage. This plot indicates there are no significant outliers (nothing top and bottom right outside the Cook’s distance)

Conclusion

To answer the two questions outlined in the executive summary:

Q1 - Manual transmission delivers a significantly higher mpg than automatic. Manual transmission is therefore ‘better’ in terms of mpg.

Q2 - Manual transmission delivers 2.94 more mpg than automatic transmission (using the multi-variate model). This has an adjusted r squared of 0.83 and a p-value below 0.05.

It may be possible to improve on this with a further multi-variate model that also includes interactions between the variables using qsec (acceleration) and weight (wt). (However, the limits of report length (around 2 pages) and appendices (around 3 pages) for this assignment mean that this additional investigation is not included.)

Appendix

Appendix A (Exploratory Data Analysis)

```
# Use "head" and "dim" to explore basic structure of data-set.

head(mtcarsFact)

##                    mpg cyl disp  hp drat    wt  qsec vs        am gear c
arb

## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0    Manual    4
4

## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0    Manual    4
4

## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1    Manual    4
1

## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1 Automatic    3
1

## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0 Automatic    3
2

## Valiant           18.1   6  225 105 2.76 3.460 20.22  1 Automatic    3
1

dim(mtcarsFact)

## [1] 32 11
```

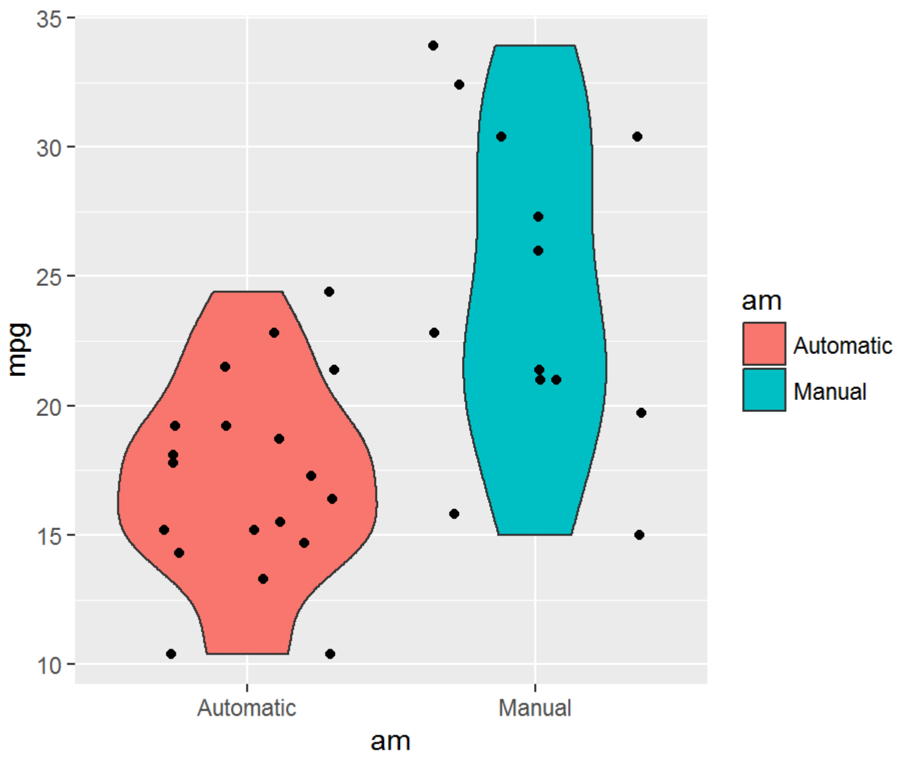
Appendix B (Variables Violin Plot - Exploratory Data Analysis)

```
library(ggplot2)

# Plot two variables "am” and “MPG”. Use violin and jitter plots to give ad
ditional information on the distribution of the variables.

plot1 <- ggplot(mtcarsFact, aes(am, mpg))

plot1 + geom_violin(aes(fill = am)) + geom_jitter(height = 0)
```



Appendix C (Step Function)

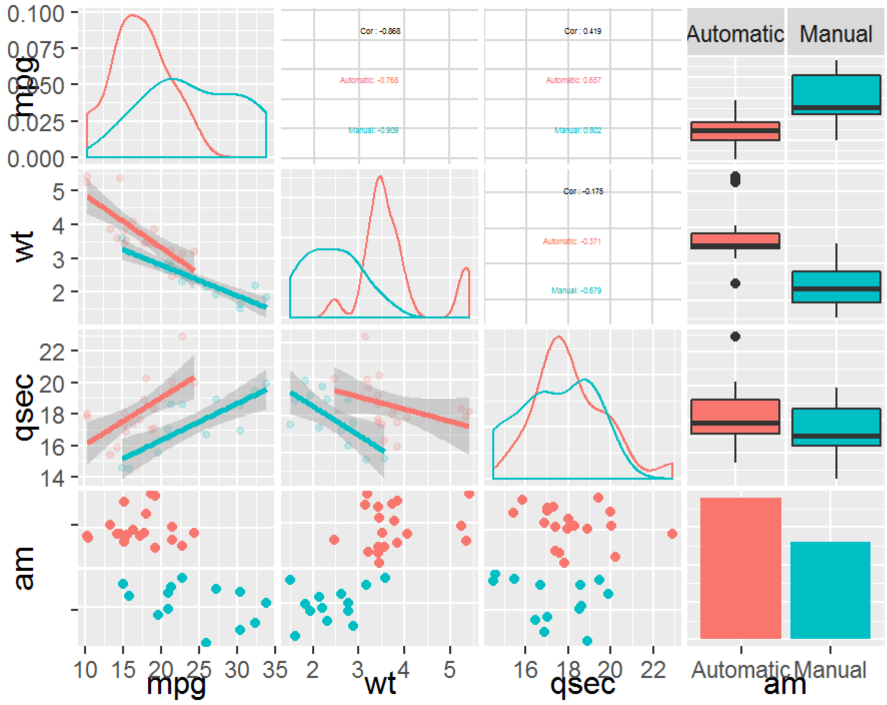
```
# step function to find variables that are optimal.
mtOptimalVar <- step(lm(data = mtcarsFact, mpg ~.), direction = "both", tra
ce=0)
summary(mtOptimalVar)
```

```
##
## Call:
## lm(formula = mpg ~ wt + qsec + am, data = mtcarsFact)
##
## 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
## wt           -3.9165     0.7112  -5.507 6.95e-06 ***
## qsec          1.2259     0.2887   4.247 0.000216 ***
## amManual      2.9358     1.4109   2.081 0.046716 *
## ---
## 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
```

Appendix D (Optimal Variables - Pairs Plot)

```
# pairs plot of "mpg" and optimal variables "am", "qsec" and "wt"
library(ggplot2)
library(GGally)
mtcarsFact[, c(1,6,7, 9)]  %>%
    ggpairs (
        mapping = ggplot2::aes(color = am), upper =list(continuous = wrap("cor", size = 1)), lower = list(continuous = wrap("smooth", alpha=0.2, size=1), combo = wrap("dot"))
    )
```



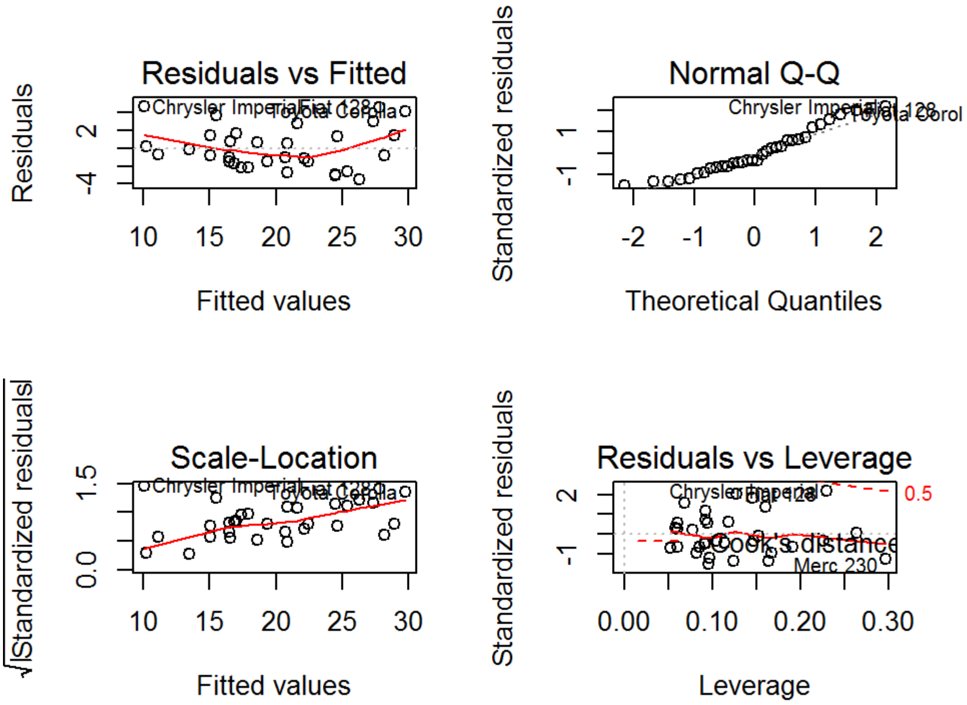
Appendix E (multi-variate Model)

```
# multi-variate model fit
mvModel <- lm(formula = mpg ~ am + wt + qsec, data=mtcarsFact)
summary(mvModel)
##
## Call:
```

```
## lm(formula = mpg ~ am + wt + qsec, data = mtcarsFact)
##
## 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
```

Appendix F (Diagnostics Plot)

```
# Diagnostics Plot
par(mfrow = c(2,2))
plot(mvModel)
```


Residuals vs Fitted
Residuals
2
-4
10
15
20
25
30
Fitted values
Chrysler Imperial
Fiat 128
Toyota Corolla
Normal Q-Q
Standardized residuals
1
-1
-2
-1
0
1
2
Theoretical Quantiles
Chrysler Imperial
Fiat 128
Toyota Corol
Scale-Location
√|Standardized residuals|
1.5
0.0
10
15
20
25
30
Fitted values
Chrysler Imperial
Fiat 128
Toyota Corolla
Residuals vs Leverage
Standardized residuals
2
-1
0.00
0.10
0.20
0.30
Leverage
Chrysler Imperial
Fiat 128
Cook's distance
Merc 230
0.5