Motor Trend Analysis using Regression Models

Executive Summary

This report examines the **mtcars** data set of motor details, to explore the relationship between a set of variables and miles per gallon (MPG) (outcome). The report answers the following two questions:

1. “Is an automatic or manual transmission better for MPG”
2. “Quantify the MPG difference between automatic and manual transmissions”

Exploratory Data Analysis showed the average MPG of Manual Transmission cars is 7.245 MPG higher than for Automatic Transmission Cars; this was confirmed using Simple Linear Regression. Obviously, the simplicity of this model has very little value, as it does not consider the possible confounding effect of the other regressors. Multiple Regression Analysis was then used to adjust this figure by taking into account multiple predictor variables and to identify the *best model* which highlighted that alongside **manual transmission**, **weight** and **qsec ¼ mile time** were significant in my model.

The outcome indicated that the overall difference in MPG between automatic and manual transmission cars was actually only **2.9358**. Further detailed analysis would be required to understand whether some of the influential observations, such as the outlier **Chrysler Imperial** significantly skewed this overall analysis.

Data Processing

First we load the specified data set (**mtcars**) and convert the **am** variable into a categorical variable.

```
data(mtcars); mtcars$am <- factor(mtcars$am, labels=c("Automatic","Manual"))
```

Exploratory Data Analysis

The **mtcars** data frame has 32 observations on 11 variables; the data frame specification is detailed at https://stat.ethz.ch/R-manual/R-devel/library/datasets/html/mtcars.html and in **Appendix A (Section A1 - Exploratory Data Analysis)** below.

This section of the appendix also includes a boxplot detailing mpg for each of automatic and manual transmission.

The average (mean) MPG of Manual Transmission cars is 7.245 MPG higher than for Automatic Transmission Cars:

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

```
##          am      mpg
## 1 Automatic 17.14737
## 2    Manual 24.39231
```

This was further confirmed by running a t.test (**t.test(mpg~am, data=mtcars)**). With the p-value being 0.001374 we can reject the null hypothesis and state that there is significant difference in the transmission means.

Regression Analysis

Simple Linear Regression

We start by examining a single predictor variable **am** on the response (outcome) variable **mpg** using: **r.initial <- lm(mpg ~am, data=mtcars)**. The summary details of this regression are detailed in **Appendix A - Section A2 - Simple Regression Summary (initial model)**.

The value of the Adjusted R-squared is 0.3385 which tells us that 33.85% of the variation in MPG is down to Manual Transmission. This means that at least 64% of the variation is unexplained. The p-value of the *amManual* coefficient is 0.000285 which indicates a high likelihood that this is significant.

In an attempt to enhance the model we will use Multiple Linear Regression.

Multiple Linear Regression

Multiple Linear Regression examines the impact of having multiple predictor variables. We will start with all variables (full model): **r.full <- lm(mpg ~., data=mtcars)**. The summary details of this regression are detailed in Appendix A - Section A3.

The Adjusted R-squared (which shows the measure of the quality of the model) illustrates that this regression model explains 80.66% of the variance of y, with the remaining 19.44 unexplained. The F-statistic p-value *3.793e-07* is under the 0.05 which indicates a meaningful model. On examining the coefficients all are above 0.05 which indicates in this model they are insignificant; only **wt** is marginally significant at 0.06.

Appendix A (Section A3) illustrates diagnostic regression plots for the full model. The Residuals vs Fitted plot has a some-what parabolic shape indicating model is incomplete. Both the Scale-Location and Residuals vs Leverage show some points scattered away from centre, which suggests some points have excessive leverage.

In an attempt to further improve the full model above, we now apply stepwise regression method to select the best variables.

```
r.best <- step(r.full, direction="backward", trace=0); summary(r.best)
```

```
##
## Call:
## lm(formula = mpg ~ wt + qsec + am, 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
## 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
```

This has successfully resulted in increasing the value of the Adjusted R-squared to by 83.36% and identified that the key significant predictor variables are **wt, qsec and amManual**, and shows that Manual Transmission accounts for only 2.9358 difference on MPG over Automatic Transmission. Appendix A (Section A4) illustrates diagnostic regression plots for the best model. The points in the Normal Q-Q plot are more or less on the line.

We now run a test to identify the best model's most outlying observation, using **outlierTest(r.best)**, which identifies **Chrysler Imperial** as the outlier.

Observations that may have the most influence on the *best* regression model were highlighted using **influence.measures(r.best)**. The following observations were identified: Cadillac Fleetwood, Lincoln Continental and Chrysler Imperial.

On further investigation these ,may highlight potential problems with the data set as all three interestingly have the same characteristics of Automatic Transmission, high Weight and lowest MPG. If further time was available I would recommend further analysis to determine whether these three observations are useful to the model or damage the analysis.

We now compare the models to determine whether they are significantly different. I have chosen to compare the initial model using Anova (Analysis of Variance - **anova(r.initial, r.best)**) against the best model; the output is detailed within Appendix A - Section A5:

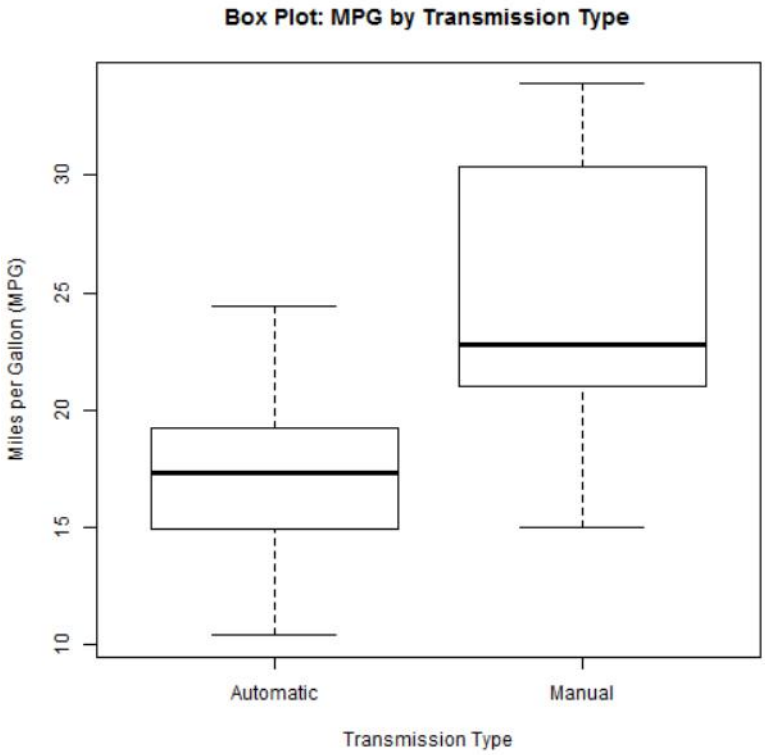
We can see that the p-value is **1.55e-09** which shows the two models are significantly different.

APPENDIX

A1 - Exploratory Data Analysis

Motor Trend Car Road Tests (**mtcars** data frame)

The data was extracted from the 1974 Motor Trend US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973-74 models). [, 1] mpg Miles/(US) gallon [, 2] cyl Number of cylinders [, 3] disp Displacement (cu.in.) [, 4] hp Gross horsepower [, 5] drat Rear axle ratio [, 6] wt Weight (lb/1000) [, 7] qsec ¼ mile time [, 8] vs V/S [, 9] am Transmission (0 = automatic, 1 = manual) [,10] gear Number of forward gears [,11] carb Number of carburetors



A2 - Simple Regression Summary (initial model)

```
r.initial <- lm(mpg ~am, data=mtcars); summary(r.initial)
```

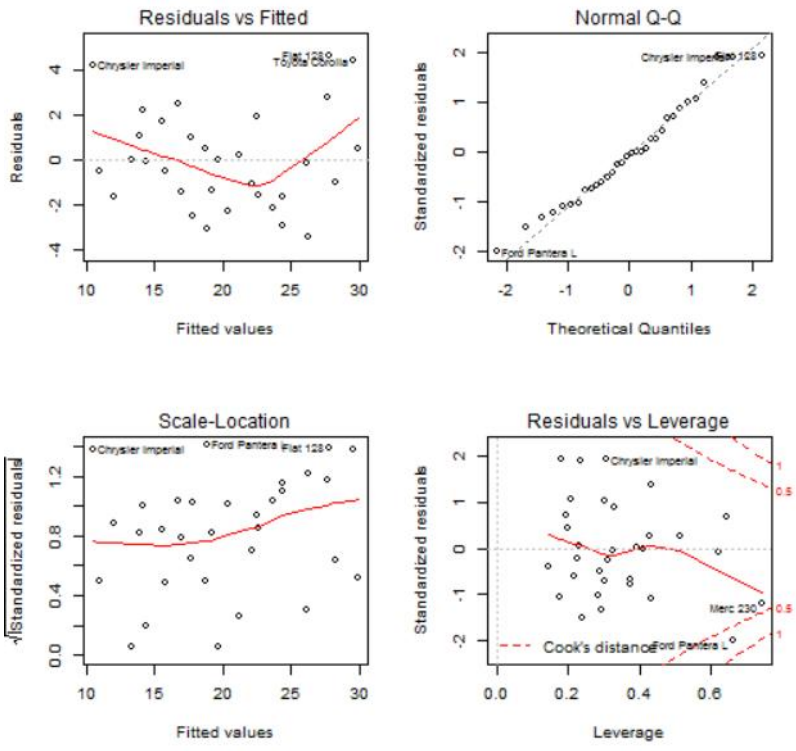
```
##
## Call:
## lm(formula = mpg ~ am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.3923 -3.0923 -0.2974  3.2439  9.5077
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   17.147      1.125  15.247 1.13e-15 ***
## amManual       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
```

A3 - Full Model Summary (all variables) and Diagnostic Reports

```
r.full <- lm(mpg ~., data=mtcars); summary(r.full)
```

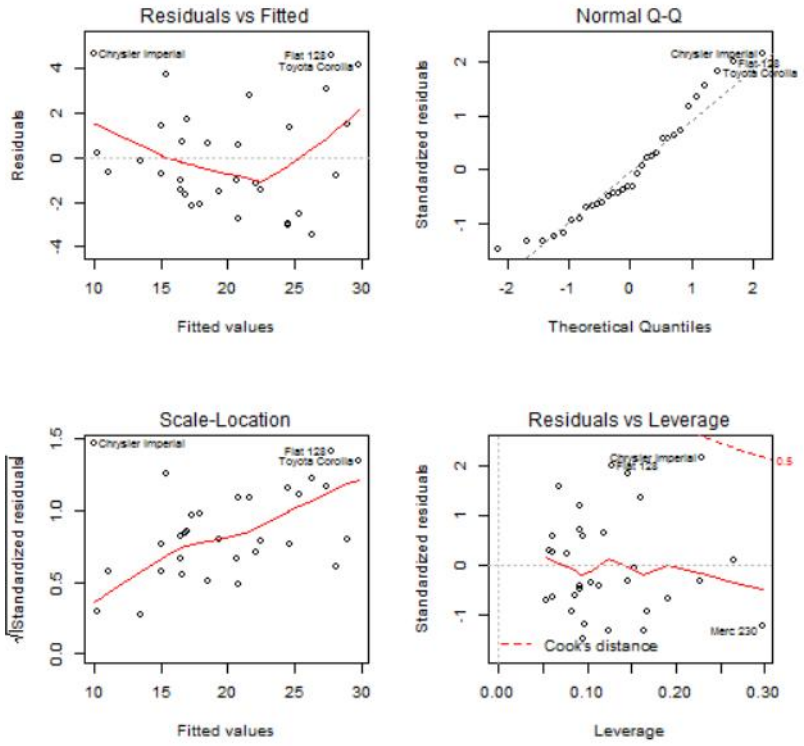
```
##
## Call:
## lm(formula = mpg ~ ., data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4506 -1.6044 -0.1196  1.2193  4.6271
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.30337   18.71788   0.657   0.5181
## cyl         -0.11144    1.04502  -0.107   0.9161
## disp         0.01334    0.01786   0.747   0.4635
## hp          -0.02148    0.02177  -0.987   0.3350
## drat         0.78711    1.63537   0.481   0.6353
## wt          -3.71530    1.89441  -1.961   0.0633 .
## qsec         0.82104    0.73084   1.123   0.2739
## vs           0.31776    2.10451   0.151   0.8814
## amManual     2.52023    2.05665   1.225   0.2340
## gear         0.65541    1.49326   0.439   0.6652
## carb        -0.19942    0.82875  -0.241   0.8122
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.65 on 21 degrees of freedom
## Multiple R-squared:  0.869,  Adjusted R-squared:  0.8066
## F-statistic: 13.93 on 10 and 21 DF,  p-value: 3.793e-07
```

```
par(mfrow=c(2, 2)); plot(r.full);
```



A4 - Diagnostic Reports on Best Model

```
par(mfrow=c(2, 2)); plot(r.best)
```



A5 - Analysis of Variance betwen Initial and Best Models

```
anova(r.initial, r.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     30 720.90
## 2     28 169.29  2    551.61 45.618 1.55e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```