Motor Car Trend Analysis

Executive Summary

The data extracted from the 1974 ‘Motor Trend’ magazine is analysed to explore the relationship between the variables in the data set and the cars’ efficiency, measured in miles per gallon (MPG). In particular the the analysis looks at:

1. Whether an automatic or manual transmission better for MPG

2. Quantifying the MPG difference between automatic and manual transmissions

Exploratory data analysis was conducted, followed by hypothesis testing, which confirmed that manual transimission is better than automatic for MPG. Both single and multivariate linear regression models were then created and the one with the highest adjusted R-squared value used to infer that MPG will increase by 1.81 mpg in cars with manual transmission compared to automatic transmission (adjusted by horse power, number of cylinders and weight).

Exploratory Data Analysis

The mtcars data set is loaded and 5 of the variables transformed into factor variables to facilitate further analysis (see appendix).

A violin plot of MPG for automatic and manual transmissions is then constructed (see appendix). It appears to show that automatic cars have a lower MPG than manual cars.

Hypothesis Testing

The violin plot appears to show that automatic cars have lower MPG but it is possible that this is due to random chance - the sample could have happened to include a group of automatic cars with low MPG and a group of manual cars with high MPG. To test this a t-test was conducted.

```r
t.test(mpg ~ am, data= mtcars, var.equal = FALSE)$p.value
```

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

The p-value from this t-test is very low so we reject the null hypothesis - manual cars have higher MPG than automatic, the confidence interval shows a 95% chance the difference is between 3.2 and 11.3 mpg.

Linear Regression

A single variable regression of the mpg variable against the am variable has an R^2 value of 0.36 and hence only accounts for 36% of the variance of the MPG variable.

```r
summary(lm(mpg~am, mtcars))$r.squared
```

```
## [1] 0.3597989
```

Using a multivariable regression may yield a higher Rˆ2 value but I don't want to overfit by adding unnecessary variables. A stepwise selection method shows the best variables that best explain MPG.

```
summary(step(lm(mpg~., mtcars),trace = 0))$r.squared
```

```
## [1] 0.8658799
```

The stepwise selection shows that the best model for mpg includes the cyl, hp, wt and am variables (see appendix). The Rˆ2 for this model is 0.87, so 87% of the variance in mpg is now explained by the model.

To confirm that the cyl, hp, wt variables are significant for predicting mpg we can compare the last two models using anova.

```
anova(lm(mpg~am,mtcars),lm(mpg~am+cyl+hp+wt,data=mtcars))$"Pr(>F)"[2]
```

```
## [1] 1.688435e-08
```

The p-value from this anova test is very low so we reject the null hypothesis - the cyl hp and wt variables are significant for the predicting mpg.

We can draw the following conclusions:

1. MPG will decrease by 3.03 and 2.16 mpg if the number of cylinders increases from 4 to 6 and 8 respectively (adjusted by hp, wt and am).

2. MPG will decrease by 0.03 mpg for every hp increase (adjusted by cyl, wt and am).

3. MPG wil decrease by 2.50 mpg for every 1000 lb of increase in weight (adjusted by cyl, wt and am).

4. MPG will increase by 1.81 mpg in cars with manual transmission compared to automatic transmission (adjusted by hp, cyl and wt).

Residual and Diagnostics

The regression diagnostic plots show that the residuals are normally distributed (see Residual vs Fitted and Normal Q-Q plots), have constant variance (see Scale-Location plot) and that no outliers are present as all values fall well within the 0.5 bands (see Residuals vs Leverage plot).

Appendix

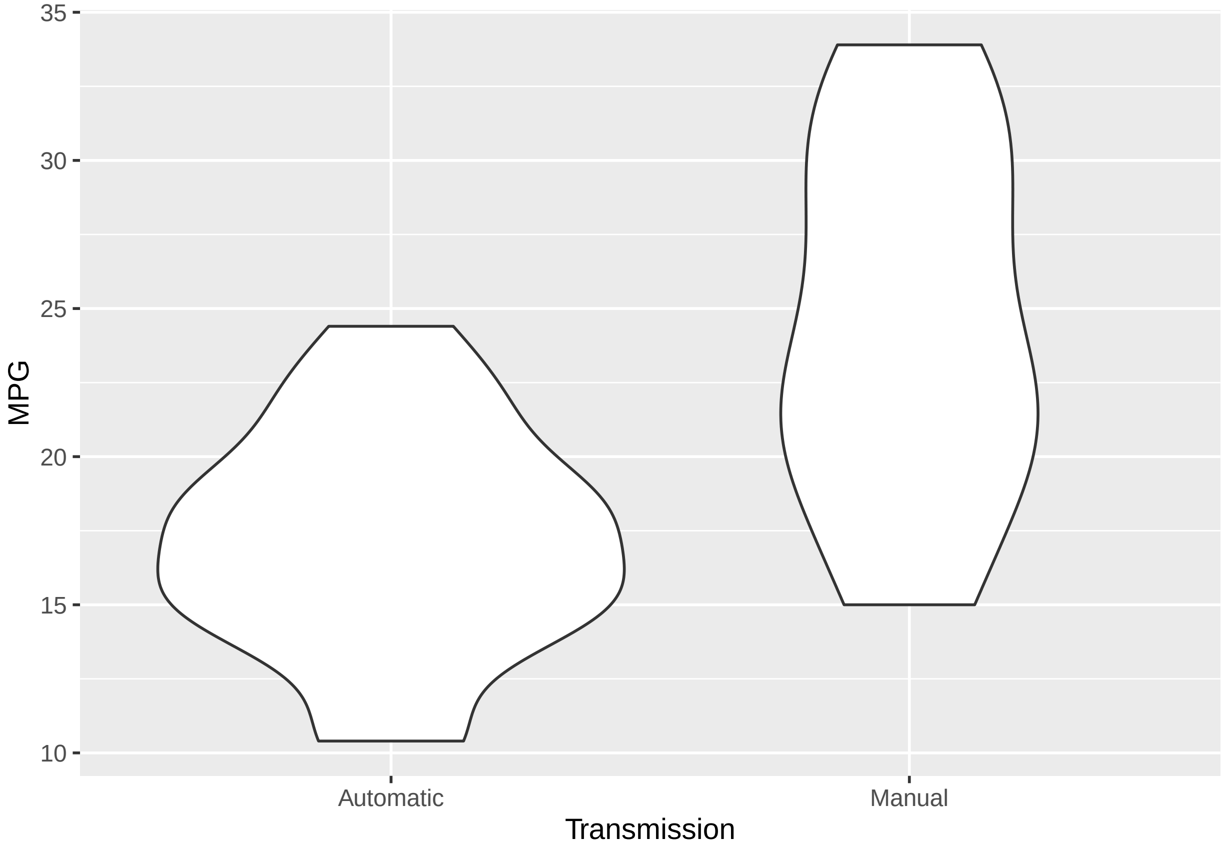
Exploratory Data Analysis

```
data(mtcars)
str(mtcars)
```

```
## 'data.frame':    32 obs. of  11 variables:
##  $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
##  $ disp: num  160 160 108 258 360 ...
##  $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
##  $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
##  $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
##  $ qsec: num  16.5 17 18.6 19.4 17 ...
##  $ vs  : num  0 0 1 1 0 1 0 1 1 1 ...
##  $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
##  $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
##  $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

```
mtcars$cyl <- factor(mtcars$cyl)
mtcars$vs <- factor(mtcars$vs)
mtcars$am <- factor(mtcars$am, labels = c("Automatic", "Manual"))
mtcars$gear <- factor(mtcars$gear)
mtcars$carb <- factor(mtcars$carb)
```

```
library(ggplot2)
ggplot(mtcars, aes(y=mpg, x=am)) + geom_violin() + xlab("Transmission") + ylab("MPG")
```



Hypothesis Testing

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

```
##
##  Welch Two Sample t-test
```

```
## 
## data:  mpg by am
## t = -3.7671, df = 18.332, p-value = 0.001374
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.280194  -3.209684
## sample estimates:
## mean in group Automatic    mean in group Manual 
##                17.14737                24.39231
```

Linear Regression

```r
summary(lm(mpg~am, mtcars))
```

```
## 
## 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
```

```r
summary(step(lm(mpg~., mtcars),trace = 0))
```

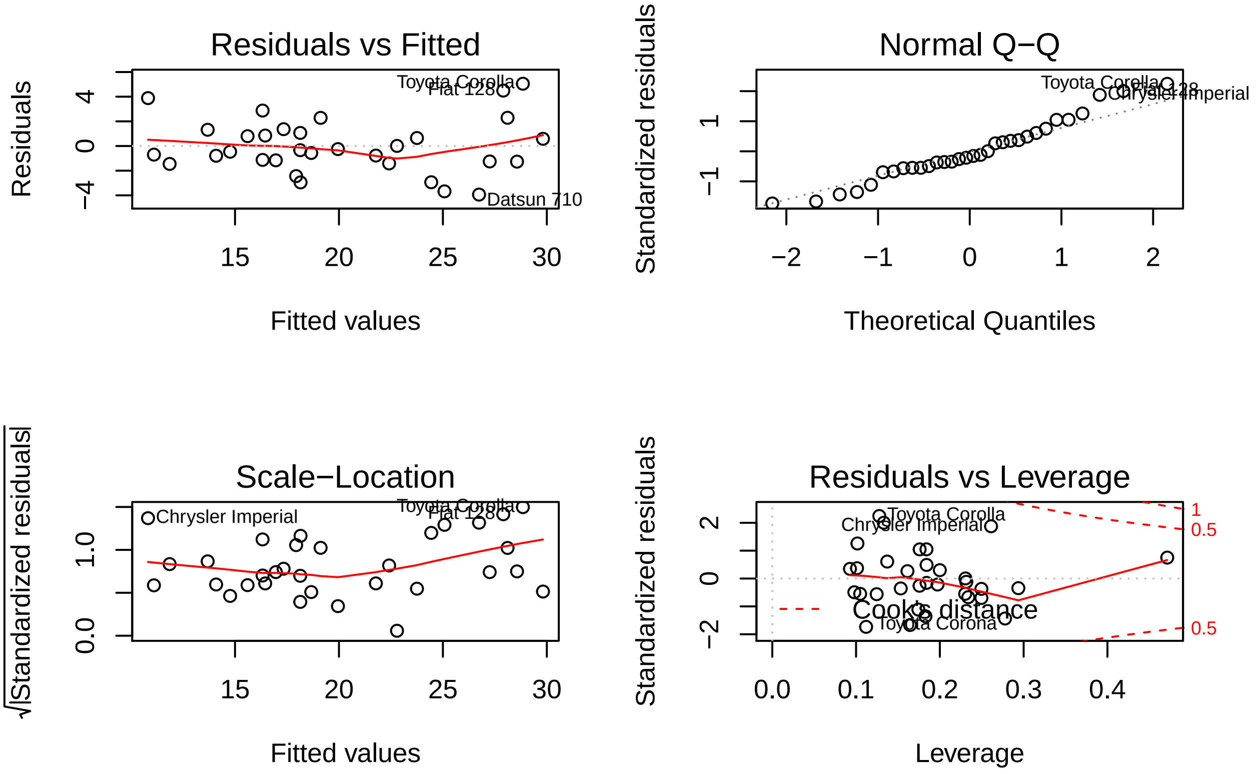
```
## 
## Call:
## lm(formula = mpg ~ cyl + hp + wt + am, data = mtcars)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -3.9387 -1.2560 -0.4013  1.1253  5.0513 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 33.70832    2.60489  12.940 7.73e-13 ***
## cyl6        -3.03134    1.40728  -2.154  0.04068 *  
## cyl8        -2.16368    2.28425  -0.947  0.35225    
## hp          -0.03211    0.01369  -2.345  0.02693 *  
## wt          -2.49683    0.88559  -2.819  0.00908 ** 
## amManual     1.80921    1.39630   1.296  0.20646    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.41 on 26 degrees of freedom
## Multiple R-squared:  0.8659, Adjusted R-squared:  0.8401
## F-statistic: 33.57 on 5 and 26 DF,  p-value: 1.506e-10
```

```
anova(lm(mpg~am,mtcars),lm(mpg~am+cyl+hp+wt,data=mtcars))
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + cyl + hp + wt
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     30 720.90
## 2     26 151.03  4    569.87 24.527 1.688e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Residual and Diagnostics

```
par(mfrow=c(2,2))
plot(lm(mpg~am+cyl+hp+wt,data=mtcars))
```


Residuals vs Fitted
Residuals
-4
0
4
Toyota Corolla
Fiat 128
Datsun 710
15
20
25
30
Fitted values
Normal Q−Q
Standardized residuals
-1
1
Toyota Corolla
Fiat 128
Chrysler Imperial
−2
−1
0
1
2
Theoretical Quantiles
Scale−Location
√|Standardized residuals|
0.0
1.0
Chrysler Imperial
Toyota Corolla
Fiat 128
15
20
25
30
Fitted values
Residuals vs Leverage
Standardized residuals
-2
0
2
Toyota Corolla
Chrysler Imperial
Toyota Corona
Cook's distance
1
0.5
0.5
0.0
0.1
0.2
0.3
0.4
Leverage