Analysis of mtcars Dataset

Synopsis

Motor Trend, a magazine about the automobile industry, is interested in exploring the relationship between a set of variables and miles per gallon (MPG) (outcome) in a data set of a collection of cars (mtcars). 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).

In this project, we will answer the following questions:

- Is an automatic or manual transmission better for MPG?
- What is the MPG difference between automatic and manual transmissions?

Loading the Data

Load in required libraries.

```
library(ggplot2)
library(dplyr)
```

We first load in the dataset and observe its first few rows and structure.

```
data(mtcars)
head(mtcars)
```

```
##                    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
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

```
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 ...
```

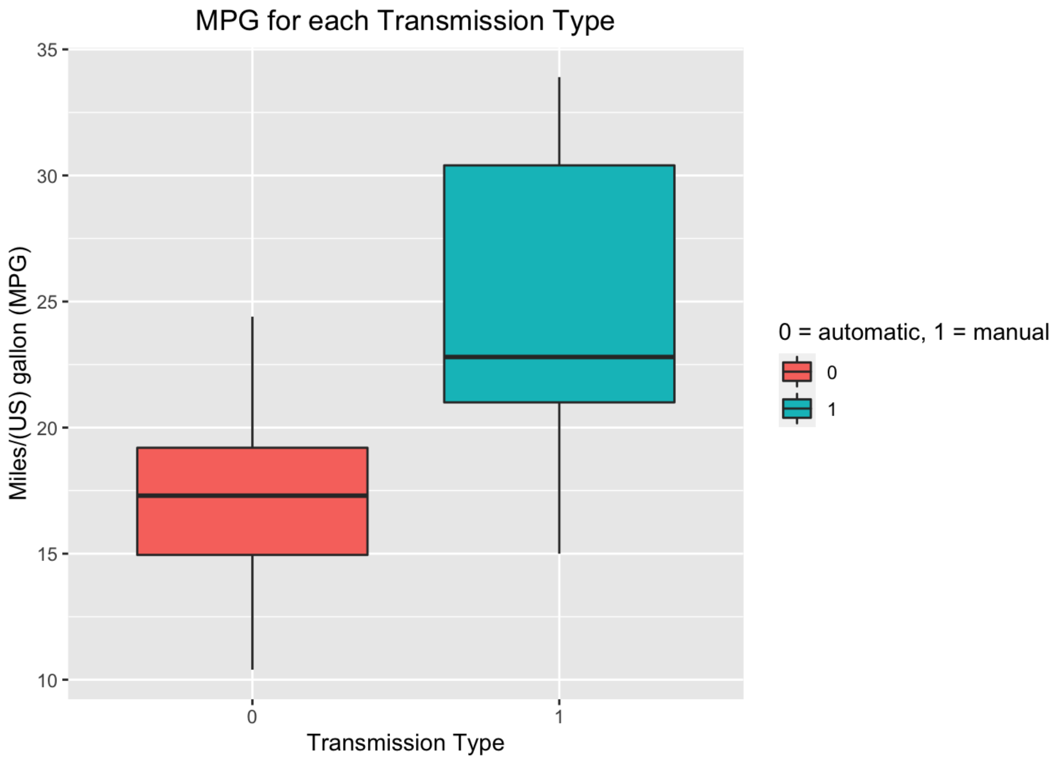
Let’s convert some of the variables to categorical.

```
mtcars$vs <- as.factor(mtcars$vs)
mtcars$am <- as.factor(mtcars$am)
```

Exploratory Data Analysis

Now, we create a boxplot to compare the MPG values for automatic and manual transmissions.

```
ggplot(mtcars, aes(x=am, y=mpg, fill=am)) + geom_boxplot() + xlab("Transmission Type") + ylab("Miles/(US) ga
llon (MPG)") + ggtitle("MPG for each Transmission Type") + guides(fill=guide_legend(title="0 = automatic, 1
= manual")) + theme(plot.title = element_text(hjust = 0.5))
```



The median MPG for manual transmissions is higher, indicating that cars with manual transmission generally have higher MPG values than cars with automatic transmission. To further support this observation, we can perform a t test as shown in the Appendix.

Model Fitting

Simple Linear Regresssion

Let's run a simple linear regression, using transmission type to predict MPG.

```
# Model 1
fit <- lm(mpg ~ am, data= mtcars)
summary(fit)
```

```
##
## 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 ***
## am1            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 t test has a p-value of 0.000285 < 0.05, so we reject the null hypothesis at 95% confidence level and conclude that the difference in mean MPG of 7.245 between automatic and manual transmission is statistically significant.

However, the R-squared value is only 0.3598, meaning that only about 36% of variation in the response variable (MPG) can be explained by the predictor variable (am). This indicates that there could be other predictor variables that affect MPG.

Multiple Linear Regression

Let's try some multiple linear regression models, with each model adding some new predictor variables. We then run ANOVA to perform nested model testing, to determine if the inclusion of extra predictor variables is necessary.

```
# Model 2
fit2 <- lm(mpg ~ am + cyl + hp, data= mtcars)
# Model 3
fit3 <- lm(mpg ~ am + cyl + hp + wt + disp, data= mtcars)
anova(fit, fit2, fit3)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + cyl + hp
## Model 3: mpg ~ am + cyl + hp + wt + disp
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 720.90
## 2     28 220.55  2    500.34 39.8754 1.199e-08 ***
## 3     26 163.12  2     57.43  4.5772   0.01981 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the ANOVA table, we observe that from Model 1 to Model 2, the p-value = 1.199e-08 < 0.05, indicating that the inclusion of `cyl` and `hp` significantly improved the model. Similarly, from Model 2 to Model 3, the p-value = 0.01981 < 0.05, indicating that the inclusion of `wt` and `disp` further improved Model 2. This suggests that we should use Model 3 to fit the data.

We check the accuracy of Model 3 as follows:

```
summary(fit3)
```

```
##
## Call:
## lm(formula = mpg ~ am + cyl + hp + wt + disp, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5952 -1.5864 -0.7157  1.2821  5.5725
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 38.20280    3.66910  10.412 9.08e-11 ***
## am1          1.55649    1.44054   1.080  0.28984
## cyl         -1.10638    0.67636  -1.636  0.11393
## hp          -0.02796    0.01392  -2.008  0.05510 .
## wt          -3.30262    1.13364  -2.913  0.00726 **
## disp         0.01226    0.01171   1.047  0.30472
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.505 on 26 degrees of freedom
## Multiple R-squared:  0.8551, Adjusted R-squared:  0.8273
## F-statistic:  30.7 on 5 and 26 DF,  p-value: 4.029e-10
```

The R-squared is 0.8551, indicating that about 85.5% of variation in the response variable can be explained by the predictor variables, which is a very good result.

We also check the validity of the model using residual analysis in the Appendix.

Results

From our analysis, we were able to determine that manual transmission is better for MPG than automatic transmission since the difference in mean MPG of 7.542 is statistically significant.

MPG is also not just determined by transmission type, but also determined by multiple variables. The model chosen offers a good fit to the data but it may not be the most accurate one. Other techniques could have been used to select the best predictors for the model.

Appendix

T test

We want to determine if cars with manual transmission have higher MPG values than cars with automatic transmission.

First, we find out the mean MPGs for automatic and manual transmission.

```
auto <- mtcars %>% filter(am == 0) %>% select(mpg)
manual <- mtcars %>% filter(am == 1) %>% select(mpg)
mean(auto$mpg)
```

```
## [1] 17.14737
```

```
mean(manual$mpg)
```

```
## [1] 24.39231
```

We observe that manual transmission has a higher mean MPG than automatic transmission. To determine if this difference is significant, we can perform a t test as shown below:

```
t.test(auto, manual, var.equal = FALSE)
```

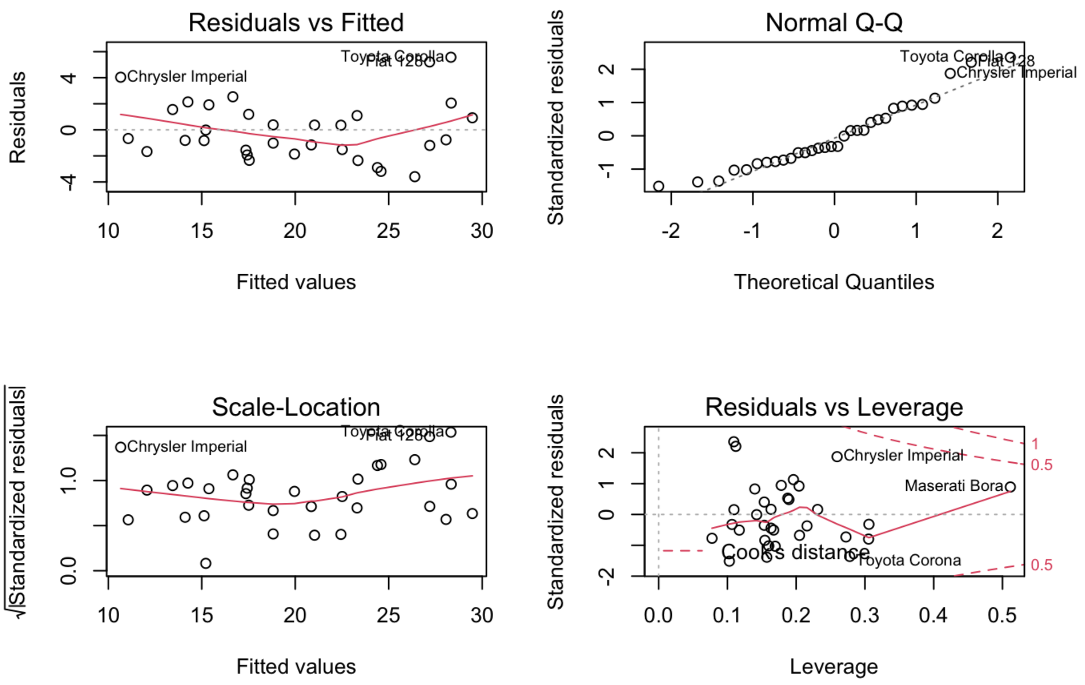
```
##
##  Welch Two Sample t-test
##
## data:  auto and manual
## t = -3.7671, df = 18.332, p-value = 0.001374
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.280194  -3.209684
## sample estimates:
## mean of x mean of y
##  17.14737  24.39231
```

The p-value = 0.001374 < 0.05, so we reject the null hypothesis that the means are equal at 95% level of confidence. This means that the difference in mean MPG for automatic and manual transmission is significant, indicating that cars with manual transmission have higher MPG values than cars with automatic transmission.

Residual Analysis

We plot the residuals as follows:

```
par(mfrow = c(2,2))
plot(fit3)
```



From the residuals vs fitted values plot, it has relatively equally spread residuals around the horizontal line without a distinct pattern. This indicates that a linear model is valid.

From the QQ plot, most of the points lie close to the line and there is no substantial departure from the line, indicating that the data is approximately normally distributed.