Week 4 Assignment

Executive Summary

In this project we study the mtcars dataset and invetsigate the following questions:

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

We dtermined the factors that contribute most to MPG are Transmission, Weight and Quarter Mile Time (qsec). Our model explains 85% of the total variation in MPG. Evidence suggests that in general Manual cars get more miles per gallon.

Data Processing

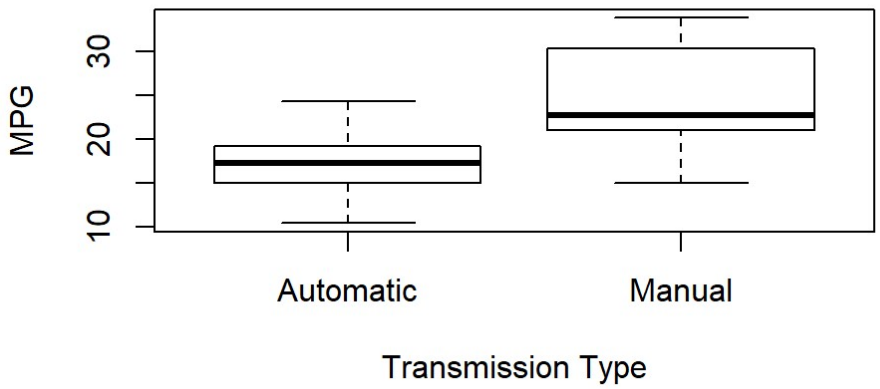
We first load the data mtcars as well as any additional libraries we will need. We also set the discrete 'am' variable to a factor type since this tells us if the cars is automatic (1) or manual (0).

```
library(dplyr)
library(car)
data(mtcars)
mtcars$am<-as.factor(mtcars$am)
```

Exploratory Data Analysis

A boxplot suggests that manual cars get more miles per gallon than automatic.

```
mpgdata<-select(mtcars, c("mpg","am"))
levels(mpgdata$am) <- c("Automatic", "Manual")
plot(x=mpgdata$am, y=mpgdata$mpg, ylab = "MPG", xlab = "Transmission Type")
```



The p-value for the coefficent for am in the following simple linear model is significant at the 5% level (against the null hypothesis it is zero) and therefore there is evidence to include the transmission predictor in the model. However, we haven't taken account of other variables, and therefore could be susceptible to Simpson's Paradox, whilst also at risk of excluding other predictors that could refine the model.

```
summary(lm(data=mpgdata, mpg~.))$coeff
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 17.147368   1.124603 15.247492 1.133983e-15
## amManual     7.244939   1.764422  4.106127 2.850207e-04
```

```
summary(lm(data=mpgdata, mpg~.))$r.squared
```

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

Modelling

Let's run a Variance Inflation Factor diagnostic.

```
fit<-lm(data=mtcars, mpg~.)
vif(fit)
```

```
##       cyl      disp        hp      drat        wt      qsec        vs
## 15.373833 21.620241  9.832037  3.374620 15.164887  7.527958  4.965873
##        am      gear      carb
##  4.648487  5.357452  7.908747
```

Given that a lot of the variables are over 10, we see that there is clearly some collinearity going on. This suggests that we cannot simply use all the predictors in the model (like in fit). We instead use R's function step() to build a model which will discern which predictors to use.

```
betterFit<-step(lm(data = mtcars, mpg ~ .), trace=0)
summary(betterFit)
```

```
##
## 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 ***
## am1           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
```

We should include the predictors Weight(wt) and Quarter Mile Time (qsec) as well as transmission according to the step() method. Note that this new model explains 85% of the variation while the model with just Transmission explains 36%.

Appendix

We include the following residual plots:

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