Analysis on the relationship between Transmission Type and MPG

Introduction

In this analysis I will try to find if there is a relationship between a set of variables and the miles per galons of a set of cars. To do this, we will use the *mtcars* dataset that is included in R.

The data set includes the following:

A data frame with 32 observations on 11 (numeric) variables.

- [, 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 (1000 lbs)
- [, 7] qsec 1/4 mile time
- [, 8] vs Engine (0 = V-shaped, 1 = straight)
- [, 9] am Transmission (0 = automatic, 1 = manual)
- [,10] gear Number of forward gears
- [,11] carb Number of carburetors

The course project asks two questions:

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

Exploratory Data Analysis

We begin by loading the necessary libraries and the data set:

```
library(ggplot2)
data(mtcars)
mtcars$cyl = as.factor(mtcars$cyl)
mtcars$vs = as.factor(mtcars$vs)
mtcars$am = as.factor(mtcars$am)
```

```
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
```

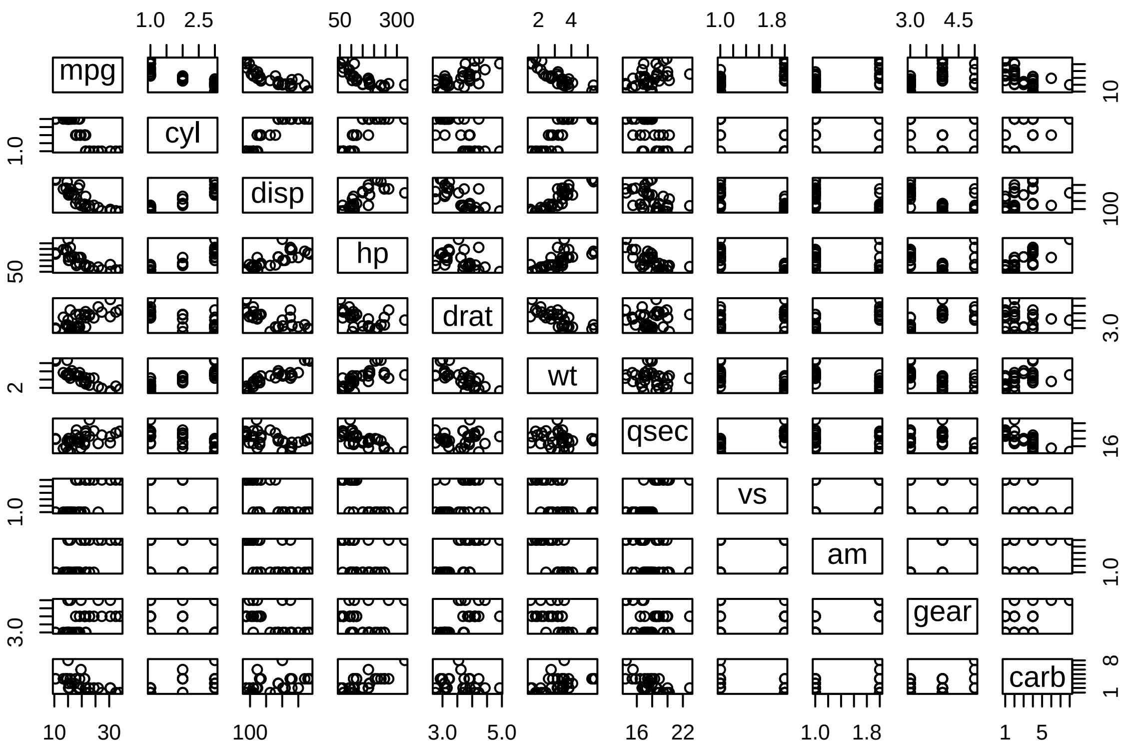
```
str(mtcars)
```

```
## '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 : Factor w/ 3 levels "4","6","8": 2 2 1 2 3 2 3 1 1 2 ...
## $ 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  : Factor w/ 2 levels "0","1": 1 1 2 2 1 2 1 2 2 2 ...
## $ am  : Factor w/ 2 levels "0","1": 2 2 2 1 1 1 1 1 1 1 ...
## $ gear: num 4 4 4 3 3 3 3 4 4 4 ...
## $ carb: num 4 4 1 1 2 1 4 2 2 4 ...
```

We can then use some graphics to find a visible relationship:

```
pairs(mtcars)
```

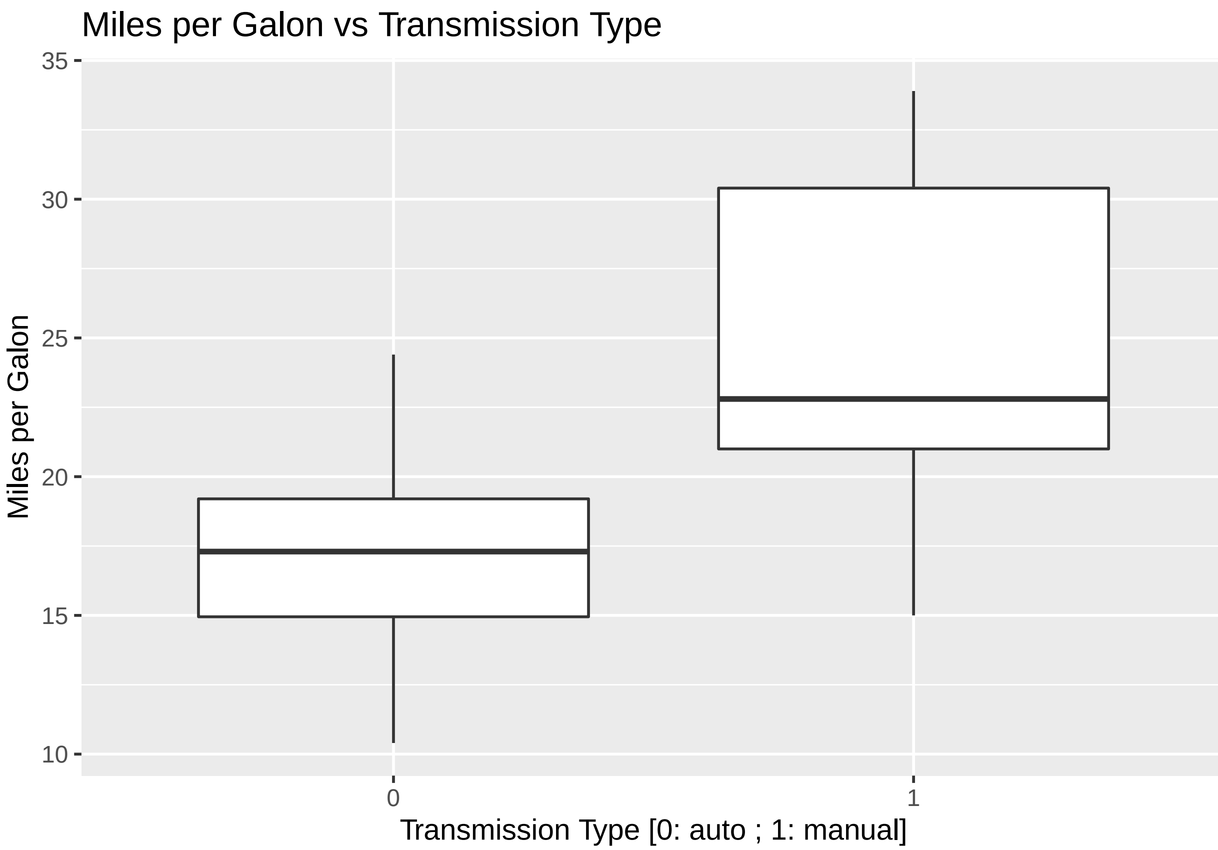


Seeing the pairs plot, we can identify the following for mpg:

- A negative relationship with disp, hp, cyl and wt.
- A positive relationship with drat and qsec.
- A clear difference between vs and am.

Since the scope of this study only includes the transmission type, we will only analyse the am variable

```
g <- ggplot(data = mtcars, aes(x = factor(am), y = mpg))
g + geom_boxplot() + ggtitle("Miles per Galon vs Transmission Type") + xlab("Transmission Type [0: auto
```



Using this boxplot, we could assume that there is a very clear difference between the fuel efficiency of automatic and manual cars, showing that manual transmissions lead to a better mileage. We wan't to test if there is a statistically significant difference in the gas mileage of both types of transmissions.

Statistical Analysis

To do this, we perform a T test to compare the medians

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

```
##
##  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 between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -11.280194  -3.209684
## sample estimates:
## mean in group 0 mean in group 1
##        17.14737        24.39231
```

Using a p-value of 0.025 (two-sided test) shows that there is a significant difference in the medians of both groups, meaning that manual cars do have a better gas mileage than those with automatic transmissions.

Regression Model

To find a proper regression model, we start with a basic linear model using mpg as the outcome and am as the regression:

```
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
```

We haven an adjusted R-squared of 0.3385, meaning that our model is not very accurate and that am only accounts for around 33% of the increase in mpg. We can test and overfit this model using all variables:

```
fit_all <- lm(mpg ~ ., data = mtcars)
summary(fit_all)
```

```
##
## Call:
## lm(formula = mpg ~ ., data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4734 -1.3794 -0.0655  1.0510  4.3906
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 17.81984   16.30602   1.093   0.2875
## cyl6        -1.66031    2.26230  -0.734   0.4715
## cyl8         1.63744    4.31573   0.379   0.7084
## disp         0.01391    0.01740   0.799   0.4334
## hp          -0.04613    0.02712  -1.701   0.1045
## drat         0.02635    1.67649   0.016   0.9876
## wt          -3.80625    1.84664  -2.061   0.0525 .
## qsec         0.64696    0.72195   0.896   0.3808
## vs1          1.74739    2.27267   0.769   0.4510
## am1          2.61727    2.00475   1.306   0.2065
## gear         0.76403    1.45668   0.525   0.6057
## carb         0.50935    0.94244   0.540   0.5948
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.582 on 20 degrees of freedom
## Multiple R-squared:  0.8816, Adjusted R-squared:  0.8165
## F-statistic: 13.54 on 11 and 20 DF,  p-value: 5.722e-07
```

Here we have a better R-squared, but using a p-value of 0.05 shows that none of the variables are significant, which is a sign of overfitting. At this point we use the *step* function to iteratively find a model that better fits our data:

```
fit_step <- step(fit_all, trace = F)
summary(fit_step)
```

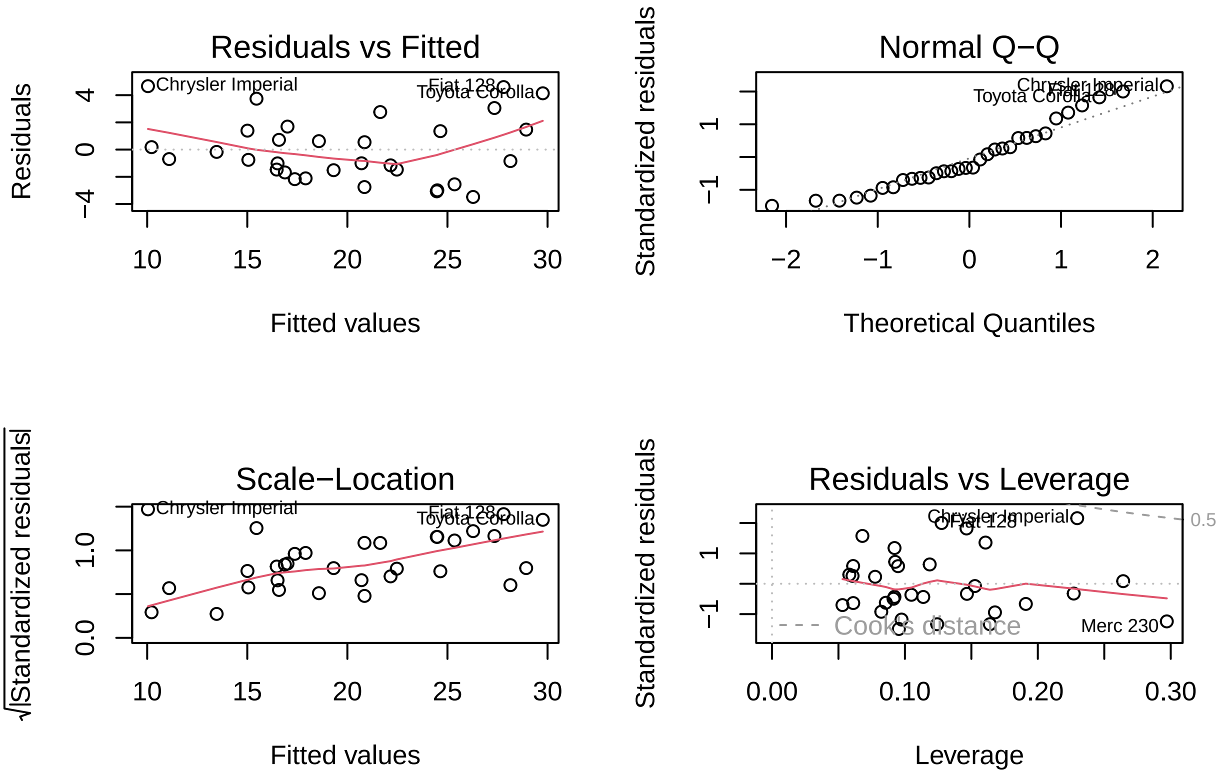
```
##
## Call:
## lm(formula = mpg ~ wt + qsec + am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4811 -1.5555 -0.7257  1.4110  4.6610
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.6178     6.9596   1.382 0.177915
## wt           -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
```

All variables in this model have a p-value bigger than 0.05, meaning they are statistically significant. The step function chose the variables *wt*,*qsec* and *am* for our new model. The confidence interval for this assumption is:

```
confint(fit_step)["am1",]
```

```
##      2.5 %     97.5 %
## 0.04573031 5.82594408
```

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



The Residuals vs. Fitted plot shows that the residuals are uncorrelated with the fitted values and the Normal Q-Q plot shows that the distribution is roughly normal

Conclusions

After analyzing the data we can conclude that there is a strong relationship between transmission type and gas mileage. Other significant variables are the Weight and the 1/4 mile time of the vehicle.