# Title: Regression Models: Peer Assessment One

*Dr Paul Fergus*

*25/01/2015*

## Executive Summary

In this weeks edition of Motor Trend we present a study that looks at the mtcar dataset and the difference between transmission type and miles per gallon (MPG), if any.

## Synopsis

The results are interesting. They show that manual transmission cars get more miles per gallon (on average 2.9 MPG) when compared with automatic cars.

## Exploratory Data Analysis

The mtcars dataset is introduced with some exploratory data analysis. The dataset contains 32 observations and 11 variables. The variables are miles per gallon, number of cylinders, displacement, gross horsepower, rear axle ratio, weight, quarter mile time, vs, transmission, number of forward gears, and number of caruretors. A more detailed summary of the statics for the mtcars dataset can be found in the appendix.

The two main variables of interest in this article are mpg and am; Figure one (found in the appendix) shows a box plot and the relationship between mpg and am. From Figure One and the statistics below we can see that automatic transmission has a mean of 17.15 and a standard deviation of 3.833. While manual transmission has a mean of 24.39 and a standard deviation of 6.16.

```
##   am      mpg
## 1  0 17.14737
## 2  1 24.39231
```

```
##   am      mpg
## 1  0 3.833966
## 2  1 6.166504
```

## Regression Analysis

Regression analysis is applied at this point to determine the best model fit using all of the variables in the dataset.

### Model Selection

#### Linear Regression

The analysis starts with a simple linear model using mpg and am as these variables are of particular interest to us. Looking at the coefficients and intercpets we find that on average, automatic cars are capable of 17.15

miles to the gallon, while for manual this is 7.24 more or 24.39 miles to the gallon. These set of results do not provide us with any additional information beyond what we obtained from the exploratory data analysis. What we can take from this analysis is that the R-squared value is relatively low (0.36), which means that our model only accounts for 36% of the variance. We can conclude that additional variables are required to produce a better model (i.e. to describe more of the variance within the dataset).

#### Multivariate Linear Regression

The previous set of results make a strong case for the use of additional variables in conjunction with our initial set (mpg and am). Determining the best variable set is found using the original model and the step function. A summary of the findings can be found in the appendix. Running the analyis we find that the variable combination that accounts for the most variance are mpg, wt, qsec and am. This is shown in the new model where the R-squared value is 0.86 (0.84 when adjusted), i.e. the variable combination accounts for 86% of the variance. Using anova we can not compare the two models and use the p-value to determine whether the wt, qsec variables contribute to the overall accuracy of the model.

```
anova(cars.lm.model, cars.lm.model.best.mv)
```

```
## 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
```

It is clear from the results that the variables wt and qsec, in conjunction with mpg and am, contribute to the accuracy of the model.

## Residual Plot and Diagnostics

Figure Two in the Appendix shows the residuals for the model using the wt and qsec variables in conjunction with the mpg and am varaibles. These help to understand non-normality and determine if there are any signs of heteroskedasticity.

T conclude Figure Two suggests that our model fits particully well with our data and exhibits resonable normality (this is not a perfect match however) as evident in the qqplot. This said there appear to be outliers in the data. Given that the dataset is small, we would be at a disadvantage if we drop any of the observations. If we were to do this we could use SMOTE to perhaps oversample the dataset. The residual verses fitted seems to support the independece condition and the scale-location plot shows that there is constant variance.

More importnatly, we find that manual transmission cars get more miles per gallon when compared with automatic cars. The difference is 2.9 MPG.

### Appendix

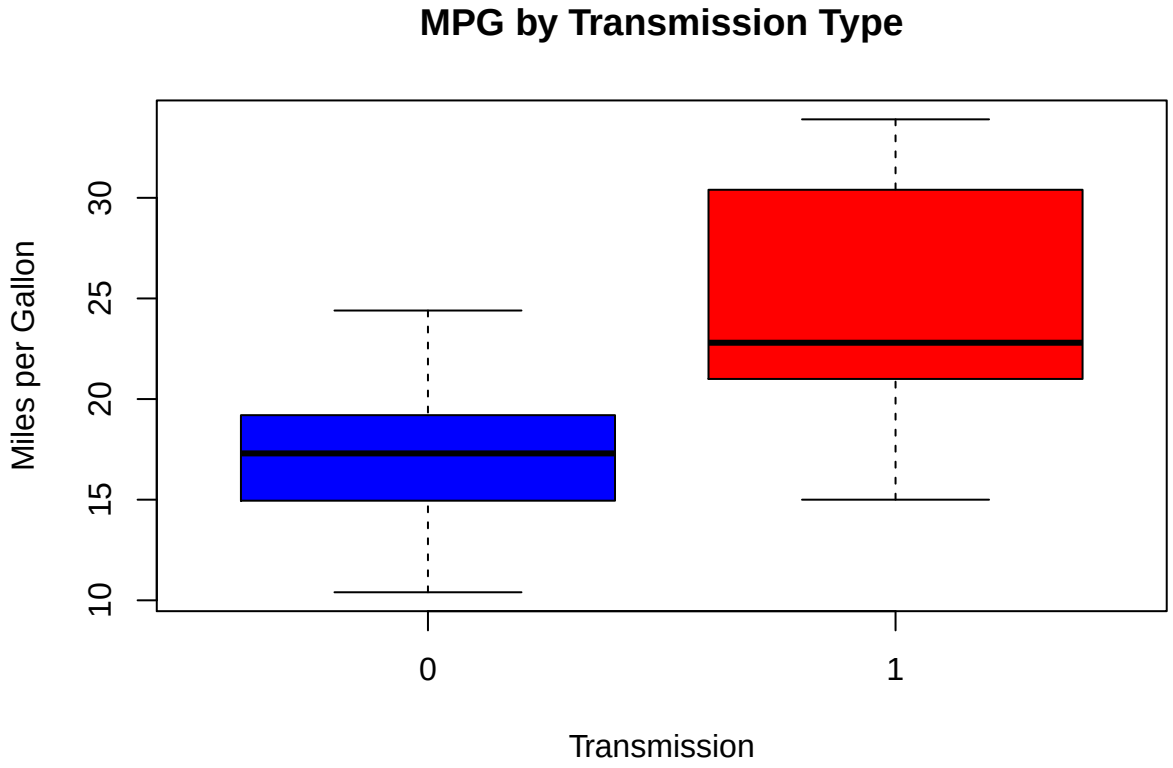
Summary statistics for mtcars.

```r
summary(mtcars)
```

```
##       mpg             cyl             disp             hp
##  Min.   :10.40   Min.   :4.000   Min.   : 71.1   Min.   : 52.0
##  1st Qu.:15.43   1st Qu.:4.000   1st Qu.:120.8   1st Qu.: 96.5
##  Median :19.20   Median :6.000   Median :196.3   Median :123.0
##  Mean   :20.09   Mean   :6.188   Mean   :230.7   Mean   :146.7
##  3rd Qu.:22.80   3rd Qu.:8.000   3rd Qu.:326.0   3rd Qu.:180.0
##  Max.   :33.90   Max.   :8.000   Max.   :472.0   Max.   :335.0
##       drat             wt             qsec             vs
##  Min.   :2.760   Min.   :1.513   Min.   :14.50   Min.   :0.0000
##  1st Qu.:3.080   1st Qu.:2.581   1st Qu.:16.89   1st Qu.:0.0000
##  Median :3.695   Median :3.325   Median :17.71   Median :0.0000
##  Mean   :3.597   Mean   :3.217   Mean   :17.85   Mean   :0.4375
##  3rd Qu.:3.920   3rd Qu.:3.610   3rd Qu.:18.90   3rd Qu.:1.0000
##  Max.   :4.930   Max.   :5.424   Max.   :22.90   Max.   :1.0000
##        am              gear            carb
##  Min.   :0.0000   Min.   :3.000   Min.   :1.000
##  1st Qu.:0.0000   1st Qu.:3.000   1st Qu.:2.000
##  Median :0.0000   Median :4.000   Median :2.000
##  Mean   :0.4062   Mean   :3.688   Mean   :2.812
##  3rd Qu.:1.0000   3rd Qu.:4.000   3rd Qu.:4.000
##  Max.   :1.0000   Max.   :5.000   Max.   :8.000
```

Figure One - Box Plots.

```r
boxplot(mpg~am, data=mtcars,
        col = c("blue", "red"),
        xlab = "Transmission",
        ylab = "Miles per Gallon",
        main = "MPG by Transmission Type")
```




Summary of simple linear model

```
summary(cars.lm.model)
```

```
##
## 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 ***
## am             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
```

Summary of the multivariant linear model

```
summary(cars.lm.model.best.mv)
```

```
##
```

```
## 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 ***
## am            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
```

Figure Two - Residual Plots.

```
par(mfrow = c(2,2))
plot(cars.lm.model.best.mv)
```