An Analysis of the Effect of Transmission Type on the Fuel Consumption of Vehicles

Abstract

An analysis of the mtcars dataset is presented, in order to determine whether a manual or an automatic transmission would be better for a vehicle's miles per gallon (mpg) reading. While an initial analysis suggests that a manual transmission would be better, further linear models disprove this hypothesis. It will be shown that the transmission actually has a very minimal effect on the mpg reading, and that the weight of the vehicle, engine displacement and number of cylinders have a far greater effect. A new model is contructed that better explains the mpg variations, and this is confirmed by examining the residuals of the model.

Introduction

With fuel consumption of vehicles being a concern for car owners, it is of interest to know what factors influence this. It is proposed that the type of tranmission - automatic or manual - that a vehicle has may affect the miles per gallon (mpg) reading that a vehicle has. This theory is put to the test and the results are discussed in this report. The chief questions to be answered are which transmission is better for mpg readings, and what is the quantifiable difference in mpg between the two transmission types.

Initial Exploratory Analysis

The mtcars dataset is loaded into R and a boxplot is generated. This will clearly show the mpg ranges separated by their transmission types. It will also give a good first indication of the relationship between the two, if any exists.

We can also generate a correlation table which will indicate how closely correlated all the variables in the dataset are with each other. In this case, the only row of interest is the first one, as this shows how each variable correlates individually with mpg.

It is worth noting that the transmission type value has only two values - 0 and 1. The 0 identifies an automatic transmission, while the 1 identifies a manual transmission.

The inital boxplot shows that the cars with a manual transmission have a higher range of mpg values, while those with an automatic transmission have lower mpg values. The mean value of the manual tranmssion fuel consumption is also higher than that of the automatic transmission fuel consumption. This would seem to indicate that a manual transmission would give a better mpg value, which would in turn result in better long-term fuel consumption. We will now develop a model to confirm or reject this hypothesis.

For the relevant R code and plot, see Appendix section A.1.

Linear Modelling

A simple linear model is fitted to the data, relating the mpg to the transmission type. The R squared value, which indicates how well the model fits the given data, is here calculated to be 0.3598. This can be interpreted as approximately 36 %. The implication of this is that only 36 % of the variation in the mpg value can be accounted for by the transmission type. This is important because there are other variables given in the mtcars dataset, such as weight and engine displacement, and these could also have an effect on the fuel consumption of the vehicle.

Despite the low R squared value, the model does show that there is a quatifiable difference between the mpg values recorded for manual and automatic transmissions. In practical statistics terms, the difference is the difference in the mean values, which here is 7.124. However, this is only applicable for this specific model and does not provide any information as to which might be preferable.

It is important to note that, given the number of variables in the available dataset, a large number of models can be constructed that could be used to show different effects on mpg. To identify which would be the best choices, we can refer back to the correlation table that was generated during the inital analysis of the data. This shows that the variables apart from transmission type that have the best correlation with mpg, indicated by the highest correlation co-efficients, are weight, displacement and number of cylinders. These three variable will therefore be used to build additional models for comparison purposes.

When the R squared values of the above models are compared, it can be seen that the best fit is obtained by including all three other variables. This model produces an R squared value of 83 %. This implies that there is a compounding effect of all these variables on the fuel consumption of the vehicles under study.It also shows a greatly decreased quantifiable difference in the means for the manual and automatic transmission values - it is now 0.12.

As an aside, it is interesting to note that, when looking at the individual effect of the new variables on mpg, the weight appears to have the biggest impact with an R squared value of 75 %.

From this, we can conclude that, while there is an effect on the mpg reading of a vehicle based on whether it uses a manual or automatic transmission, that effect is not significant enough to identify which one is better. There are other factors that must be considered and therefore the relationship is not as simple as originally suggested by the exploratory analysis. We must therefore reject our initial hypothesis that a manual transmission is better than an automatic transmission.

For the relevant R code and plot, see Appendix section A.2.

Residuals

An investigation of the residuals of the best fit model will check whether or not it really is a good fit. The plots show that the residuals appear to be normally distributed, although there are a few outliers that might merit further investigation. Overall, this model can be concluded to be a better fit for the data.

For the relevant R code and plot, see Appendix section A.3.

Conclusion

The data from the mtcars dataset was analysed to identify which transmission type, manual or automatic, would result in a better miles per gallon reading. An initial exploratory analysis suggested that a manual transmission would be better, and that there was a quantifiable difference between the mpg readings of the different classes of vehicles. However when the data were modelled using a linear model, it was found the the transmission actually has a minimal effect on the mpg reading. Addional variables were included in the model in and effort to better model the effects. It was noted that the vehicle weight has the single biggest effect on mpg reading, while the best fit model was one that included weight, engine displacement and number of cyclinders. A further investigation of the residuals for this model confirmed this.

Appendix

This section contains the R code, outputs and plots used for this analysis of the mtcars dataset.

A.1 - Inital Exploratory Analysis

```
# load the necessary libraries
library(datasets)
library(ggplot2)
```

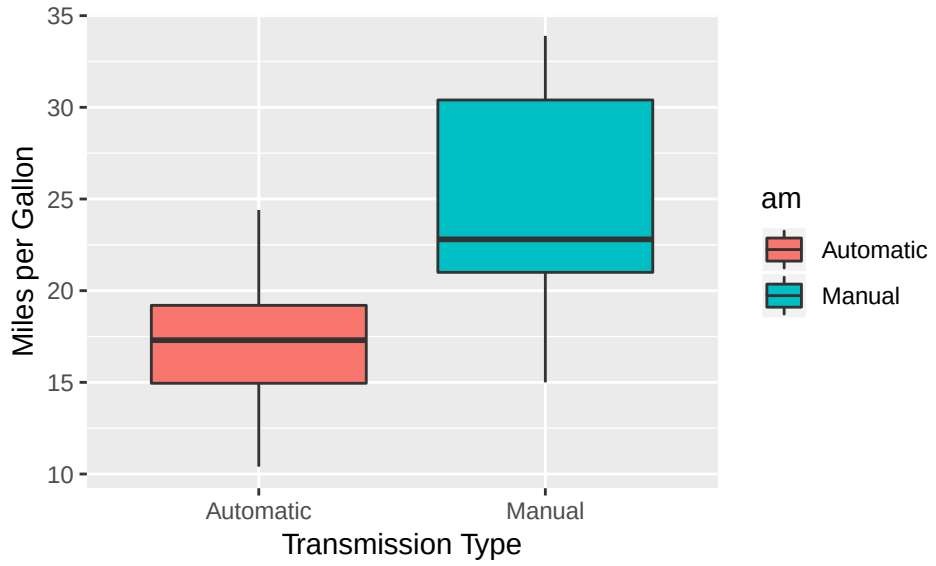
```
## Warning: package 'ggplot2' was built under R version 3.5.1
```

```
# load the mtcars dataset and convert the numeric transmssion type variable to a factor
data(mtcars)
mtcars1 <- mtcars
mtcars1$am <- factor(mtcars1$am, labels = c("Automatic", "Manual"))

# calculate the mean of the mpg for each transmission type
aggregate(mpg ~ am, data = mtcars1, mean)
```

```
##          am      mpg
## 1 Automatic 17.14737
## 2    Manual 24.39231
```

```
# plot a boxplot to get an initial idea of the ranges of the mpg values for
# each transmission type
car_plot <- ggplot(data = mtcars1, aes(x=am, y=mpg)) +
            geom_boxplot(aes(fill = am)) +
            labs(x="Transmission Type", y="Miles per Gallon")
car_plot
```



A.2 - Linear Modelling

```
# fit a simple linear model to the data
fit01 <- lm(mpg ~ am, data = mtcars1)
summary(fit01)
```

```
##
## Call:
## lm(formula = mpg ~ am, data = mtcars1)
##
## 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
# generate the correlation co-efficients from the original data
carsCorr <- cor(mtcars)
round(carsCorr[1,],2)
```

```
##   mpg   cyl  disp    hp  drat    wt  qsec    vs    am  gear  carb
##  1.00 -0.85 -0.85 -0.78  0.68 -0.87  0.42  0.66  0.60  0.48 -0.55
```

```r
# model including transmission and weight
fit02 <- lm(mpg ~ am + wt, data = mtcars1)
summary(fit02)
```

```
##
## Call:
## lm(formula = mpg ~ am + wt, data = mtcars1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5295 -2.3619 -0.1317  1.4025  6.8782
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 37.32155    3.05464  12.218 5.84e-13 ***
## amManual    -0.02362    1.54565  -0.015    0.988
## wt          -5.35281    0.78824  -6.791 1.87e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.098 on 29 degrees of freedom
## Multiple R-squared:  0.7528, Adjusted R-squared:  0.7358
## F-statistic: 44.17 on 2 and 29 DF,  p-value: 1.579e-09
```

```r
# model including transmission, weight and displacement
fit03 <- lm(mpg ~ am + wt + disp, data = mtcars1)
summary(fit03)
```

```
##
## Call:
```

```
## lm(formula = mpg ~ am + wt + disp, data = mtcars1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4890 -2.4106 -0.7232  1.7503  6.3293
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 34.675911   3.240609  10.700 2.12e-11 ***
## amManual     0.177724   1.484316   0.120   0.9055
## wt          -3.279044   1.327509  -2.470   0.0199 *
## disp        -0.017805   0.009375  -1.899   0.0679 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.967 on 28 degrees of freedom
## Multiple R-squared:  0.781,  Adjusted R-squared:  0.7576
## F-statistic: 33.29 on 3 and 28 DF,  p-value: 2.25e-09
```

```
# model including transmission, weight, displacement and cylinders
fit04 <- lm(mpg ~ am + wt + disp + cyl, data = mtcars1)
summary(fit04)
```

```
##
## Call:
## lm(formula = mpg ~ am + wt + disp + cyl, data = mtcars1)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.318 -1.362 -0.479  1.354  6.059
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 40.898313   3.601540  11.356 8.68e-12 ***
## amManual     0.129066   1.321512   0.098  0.92292
## wt          -3.583425   1.186504  -3.020  0.00547 **
## disp         0.007404   0.012081   0.613  0.54509
## cyl         -1.784173   0.618192  -2.886  0.00758 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.642 on 27 degrees of freedom
## Multiple R-squared:  0.8327, Adjusted R-squared:  0.8079
## F-statistic: 33.59 on 4 and 27 DF,  p-value: 4.038e-10
```

```
# model including transmission and displacement
fit05 <- lm(mpg ~ am + disp, data = mtcars1)
summary(fit05)
```

```
##
## Call:
## lm(formula = mpg ~ am + disp, data = mtcars1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.6382 -2.4751 -0.5631  2.2333  6.8386
```

```
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 27.848081   1.834071  15.184 2.45e-15 ***
## amManual     1.833458   1.436100   1.277    0.212
## disp        -0.036851   0.005782  -6.373 5.75e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.218 on 29 degrees of freedom
## Multiple R-squared:  0.7333, Adjusted R-squared:  0.7149
## F-statistic: 39.87 on 2 and 29 DF,  p-value: 4.749e-09
```

```r
# model including transmission, displacement and cylinders
fit06 <- lm(mpg ~ am + disp + cyl, data = mtcars1)
summary(fit06)
```

```
##
## Call:
## lm(formula = mpg ~ am + disp + cyl, data = mtcars1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.0863 -1.7831 -0.4842  1.5987  6.6358
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 32.91686    2.77914  11.844 2.03e-12 ***
## amManual     1.92873    1.33973   1.440   0.1611
## disp        -0.01559    0.01065  -1.463   0.1545
## cyl         -1.61822    0.69937  -2.314   0.0282 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3 on 28 degrees of freedom
## Multiple R-squared:  0.7761, Adjusted R-squared:  0.7522
## F-statistic: 32.36 on 3 and 28 DF,  p-value: 3.06e-09
```

```r
# model including transmisison and cylinders
fit07 <- lm(mpg ~ am + cyl, data = mtcars1)
summary(fit07)
```

```
##
## Call:
## lm(formula = mpg ~ am + cyl, data = mtcars1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.6856 -1.7172 -0.2657  1.8838  6.8144
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  34.5224     2.6032  13.262 7.69e-14 ***
## amManual      2.5670     1.2914   1.988   0.0564 .
## cyl          -2.5010     0.3608  -6.931 1.28e-07 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.059 on 29 degrees of freedom
## Multiple R-squared:  0.759,  Adjusted R-squared:  0.7424
## F-statistic: 45.67 on 2 and 29 DF,  p-value: 1.094e-09
```

```
# model including weight only
fit08 <- lm(mpg ~ wt, data = mtcars1)
summary(fit08)
```

```
##
## Call:
## lm(formula = mpg ~ wt, data = mtcars1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5432 -2.3647 -0.1252  1.4096  6.8727
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  37.2851     1.8776  19.858  < 2e-16 ***
## wt           -5.3445     0.5591  -9.559 1.29e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.046 on 30 degrees of freedom
## Multiple R-squared:  0.7528, Adjusted R-squared:  0.7446
## F-statistic: 91.38 on 1 and 30 DF,  p-value: 1.294e-10
```

```
# model including displacement only
fit09 <- lm(mpg ~ disp, data = mtcars1)
summary(fit09)
```

```
##
## Call:
## lm(formula = mpg ~ disp, data = mtcars1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.8922 -2.2022 -0.9631  1.6272  7.2305
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 29.599855   1.229720  24.070  < 2e-16 ***
## disp        -0.041215   0.004712  -8.747 9.38e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.251 on 30 degrees of freedom
## Multiple R-squared:  0.7183, Adjusted R-squared:  0.709
## F-statistic: 76.51 on 1 and 30 DF,  p-value: 9.38e-10
```

```
# model including cylinders only
fit10 <- lm(mpg ~ cyl, data = mtcars1)
summary(fit10)
```

```
##
## Call:
## lm(formula = mpg ~ cyl, data = mtcars1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.9814 -2.1185  0.2217  1.0717  7.5186
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  37.8846     2.0738   18.27  < 2e-16 ***
## cyl          -2.8758     0.3224   -8.92 6.11e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.206 on 30 degrees of freedom
## Multiple R-squared:  0.7262, Adjusted R-squared:  0.7171
## F-statistic: 79.56 on 1 and 30 DF,  p-value: 6.113e-10
```

A.3 - Residuals

```
## plot the residuals of the best model
par(mfrow = c(2,2))
plot(fit04)
```