Regression Analysis of mtcars data

Julian Hatwell

Introduction

Using data was extracted from the 1974 Motor Trend US magazine this report will try to explain whether an automatic or manual transmission better for MPG and quantify the difference.

Executive Summary

Initially comparing miles per gallon (mpg) to transmission type (trans, either "Automatic" or "Manual") shows an apparently significant difference between the two classes of trans of between 4.7 and 9.8 mpg better for Manual trans compared to Automatic.

However, deeper investigation applying linear modeling revealed that the vehicles' overall weight (wt) was the defining variable with a strong negative trend which was more pronounced for Manual trans. In other words, the mpg performance of heavier manual cars would be much worse than heavier automatic cars.

The reason this was immediately obvious was because the dataset is not evenly distributed by weight. The manual cars included in the dataset are on average 1400lbs lighter than the Automatic cars.

From the linear modelling, it can be inferred that:

Automatic cars lose between 3.0 and 4.6 mpg per 1000lbs additional weight
Manual cars lose between 6.9 and 11.3 mpg per 1000lbs additional weight

Further refinement of the model was achieved by including the Quarter Mile Time (qsec) measure of vehicle acceleration. The measure is positively correlated with mpg. A higher qsec would indicate a slower acceleration.

Both automatic and manual cars have improved mpg ratings by a factor of 1.02 for every additional second of qsec.

Any advice for the car buyer who is conscious of fuel economy would be to prefer a lighter vehicle above all and then select a manual car which scored at the low end of the range on acceleration times.

Needless to say that any inferences drawn from this data relate to cars manufactured in the 1970's. It would be wrong to apply these conclusions to cars in 2015.

Exploratory Data Analysis

A boxplot was generated to look at the two vehicle types side by side. See Appendix A.

From the plot, it appears that manual trans has better mpg. However, there are a number of other variables in car design which may have an effect on mpg. It is necessary to determine which of these may be contributing or confounding the results. This can be done through comparison of linear models.

Regression Model Selection step by step

The first model to be explored is the known relationship between mpg and transmission. This will be the benchmark when looking for the best fitting model.

```
## (Intercept) transManual
##   17.147368    7.244939
```

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

The lm summary gives the trans coefficient as 7.245 with a standard error of 1.764 (not shown). This can be interpreted as a gain of between 5.5 to 9.0 mpg. However, it can be seen from the R squared value that this model only explains 36% of the variation and there is still a lot of residual variation not explained by the model.

In order to select the variables of most interest for modelling, a correlogram of the whole data set is produced. See Appendix B.

From the corrgram, mpg appears to be positively correlated with vs, drat, gear and qsec in addition to am.

Of these, gear and drat are also strongly correlated with am, and so it is not useful to include them in any further modelling.

vs is weakly correlated with am and qsec is weakly negatively correlated, so these may be of interest. They are initially put aside to see if a good enough fit can be developed without them.

Furthermore wt, disp, cyl, hp and carb all appear to have strong negative correlations with mpg and strong positive correlations with each other. This is definitely of interest but it is important to select only one of them. They all relate to car or engine size/weight.

cyl and carb are both factor (count) variables of specific engine components, which would naturally be expected to increase with overall engine size. This is reflected in the corrgram. These will be excluded from further investigation.

Regression modeling is used to determine which of the continuous variables of size/weight is the most significant for this analysis:

```
##   (Intercept)            wt          disp            hp
## 37.1055052690 -3.8008905826 -0.0009370091 -0.0311565508
```

```
## [1] 0.8268361
```

From the coefficient table it can be seen that wt has the strongest negative correlation. The others are not on the same scale and so will be excluded from further investigation.

The coefficient of -3.80 with a standard error of plus/minus 1.066191 (not shown) implies that for every 1000lb additional weight, a vehicle may lose up to 4.8mpg.

Furthermore, this R squared statistic of 82.7% explains far more of the variance than the transmission model.

This warrants a closer look at the relationship between weight and transmission type. A dot plot is generated to show how the weights are distributed. See Appendix C

This is very problematic for the simplistic model because, on average, the automatic cars in the dataset are over 1400lbs heavier than the manual cars. The manual cars in this sample will inevitably report a large improvement on mpg if weight is not factored into the model.

A model using wt interacting with trans is produced:

```
##    (Intercept)             wt    transManual wt:transManual
##      31.416055      -3.785908      14.878423      -5.298360
```

```
## [1] 0.8330375
```

This model now explains just over 83% of the variation.

To seek any further improvements to the model there are two options.

1. Add more variables to the model.
2. Perform some diagnostics on the existing model to determine if any systematic effects or influential outliers.

Option 1 is not very attractive at this point because only vs and qsec have not been included yet but are known to be weakly correlated. Option 2 is followed. See Appendix D for Residual Plots.

There are a small number (2-4) of vehicles that appear to have unexpectedly high mpg given their weight and so a plot is made to reveal where they sit in relation to the other data points. Also, using the coefficients from the current linear model, two lines are added to show how mpg is dropping off with wt much more rapidly for manual vehicles. See Appendix E.

It is decided that these points need to remain in the model. Removing them is potentially going to make prediction more accurate but is unlikely to affect the overall trends, which still hold.

Two improve on the model, the two remaining variables (vs and qsec) were investigated. It was found that 3 of the vehicles with unexpectedly high mpg were near the top for qsec. See Appendix F.

Two further models are created using qsec. The first is independent of trans and the second is interacting with trans. The results are tested with ANOVA. See Appendix G for the full results.

The R squared value has jumped up 6 percentage points to 89%-90%.

This is a significant increase with model lm4 improvement given the F statistic of 16.28 and very small P of 0.0005.

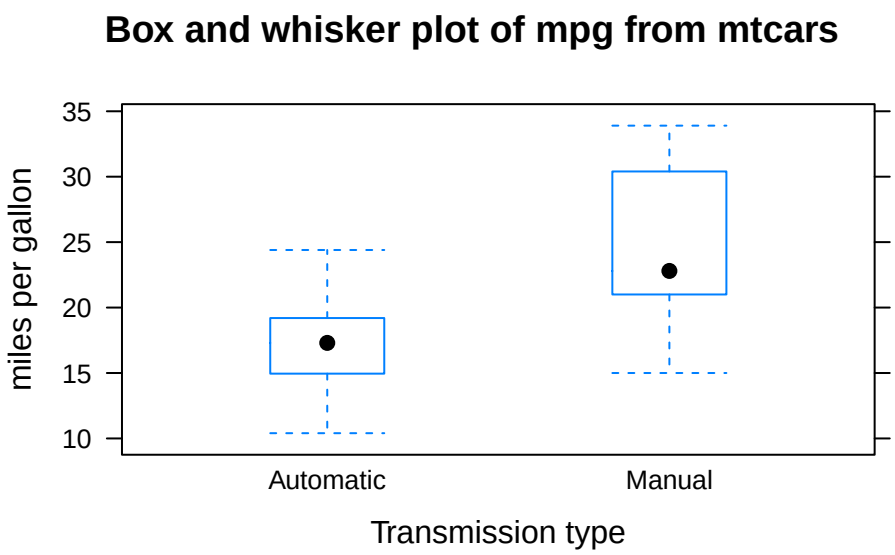
Any further improvement from lm5 is too small and scores very poorly ($p = 0.676$) and so lm5 is rejected in favour of the more parsimonious lm4.

Diagnostic plots show that the problems with the lm3 model have been reduced. These are not included here for brevity.

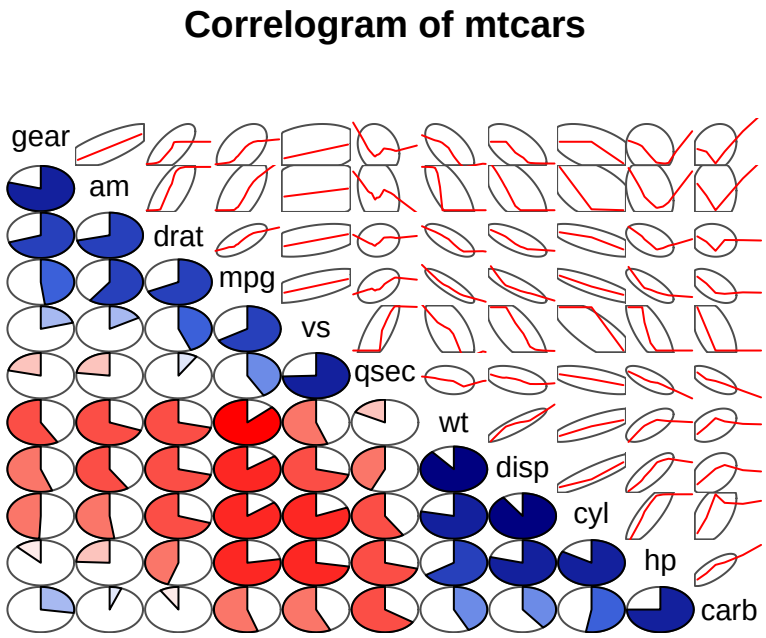
It is decided from this information that the preferred regression model is that of mpg depending on weight interacting with transmission type and favouring vehicles with a lesser acceleration.

Appendices

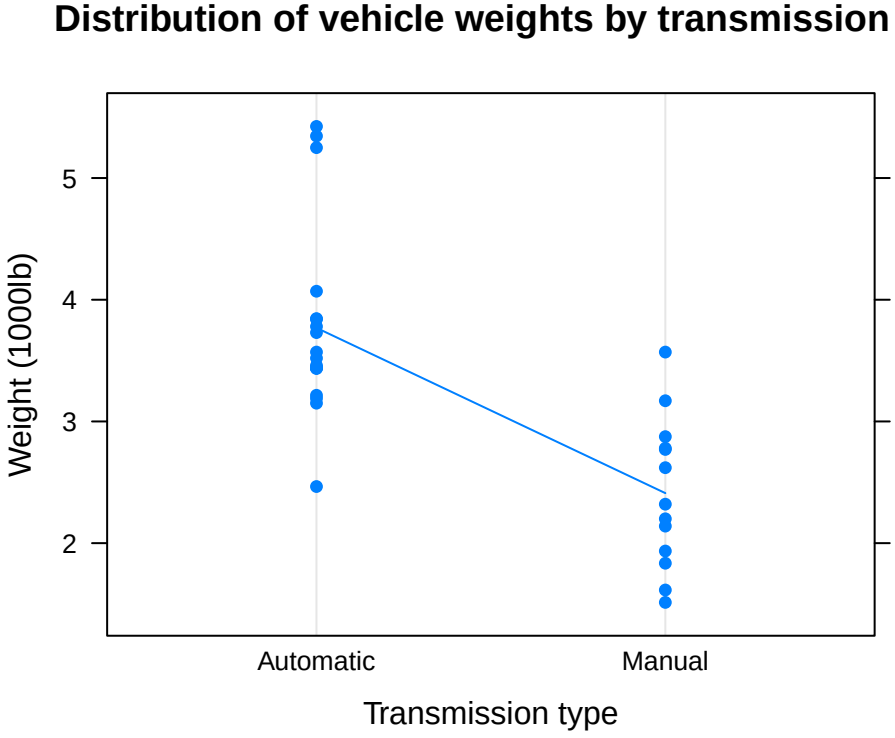
Appendix A.



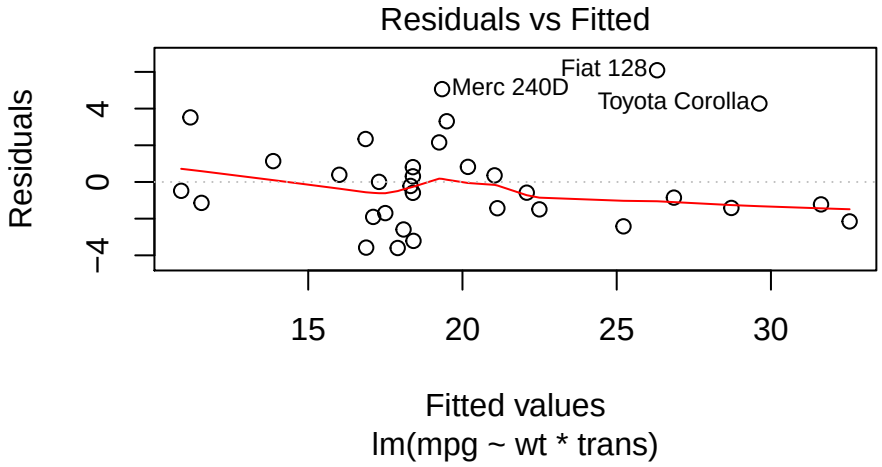
Appendix B.

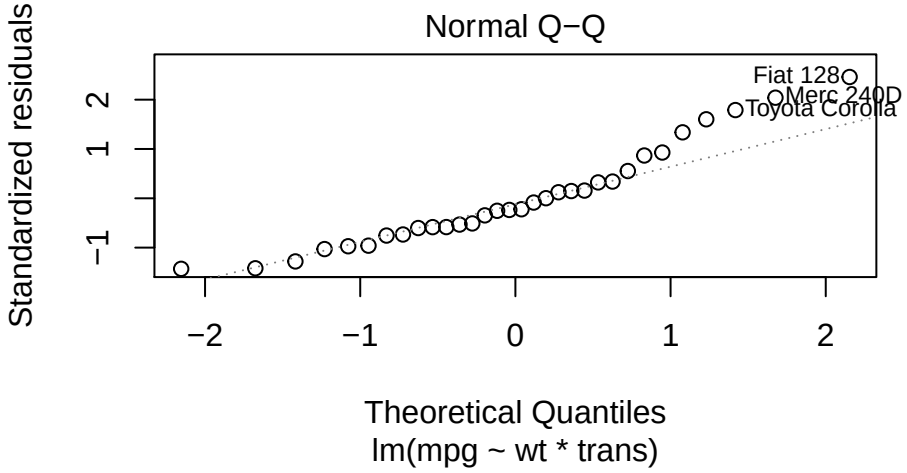


Appendix C.

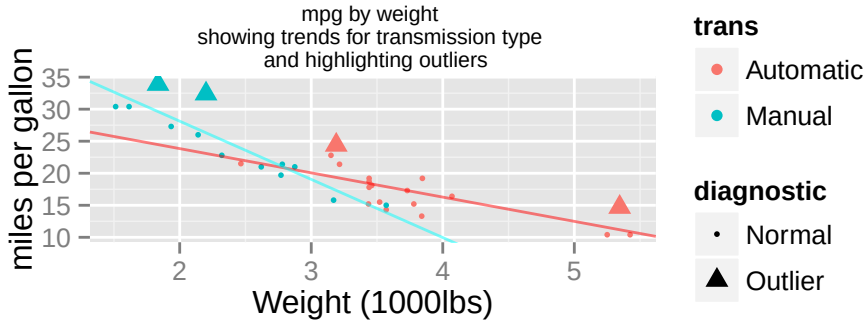


Appendix D.





Appendix E.



Appendix F.

For brevity, only the top results are shown for qsec as this turned out to be of interest:

```
##                  qsec     dgn
## Merc 230        22.90  Normal
## Valiant         20.22  Normal
## Toyota Corona   20.01  Normal
## Merc 240D       20.00 Outlier
## Toyota Corolla 19.90 Outlier
## Fiat 128        19.47 Outlier
## Hornet 4 Drive 19.44  Normal
## Merc 280C       18.90  Normal
```

Appendix G.

ANOVA of candidate linear models:

```
lm4 <- lm(mpg~wt * trans + qsec, data = mtcars)
lm4$coefficients
```

```
##    (Intercept)             wt    transManual           qsec wt:transManual
##       9.723053      -2.936531      14.079428       1.016974      -4.141376
```

```
summary(lm4)$r.squared
```

```
## [1] 0.8958514
```

```
lm5 <- lm(mpg~wt * trans + qsec * trans, data = mtcars)
lm5$coefficients
```

```
##      (Intercept)               wt      transManual             qsec
##       11.2489412       -2.9962762        8.9264577        0.9454396
##   wt:transManual transManual:qsec
##       -3.7580835        0.2355322
```

```
summary(lm5)$r.squared
```

```
## [1] 0.8965639
```

```
anova(lm3, lm4, lm5)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ wt * trans
## Model 2: mpg ~ wt * trans + qsec
## Model 3: mpg ~ wt * trans + qsec * trans
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     28 188.01
## 2     27 117.28  1    70.731 15.7891 0.0005009 ***
## 3     26 116.47  1     0.802  0.1791 0.6756302
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```