Regression Analysis of mtcars data

Julian Hatwell

Introduction

This is a report based on the mtcars data in R. In particular it seeks to answer the questions:

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

The data was extracted from the 1974 Motor Trend US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973-74 models).

Note

For brevity, not all the code is rendered in this document. It is available in a separate R Markdown Document.

Contents

Executive Summary

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

A correlogram of all the available variables revealed correlations and relationships that gave cause to suspect the validity of these initial findings.

Applying the principles linear modeling, it was discovered that the vehicles' overall weight (wt) was the defining variable with a strong negative trend (supporting all reasonable expectations).

This negative trend was more pronounced for Manual trans than Automatic. In other words, the mpg performance of heavier manual cars would be much worse than heavier automatic cars.

The reason this was not apparent in the first pass 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.

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, implying a less fuel hungry engine. However any further speculation on the mechanism is outside of the scope of this research.

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

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

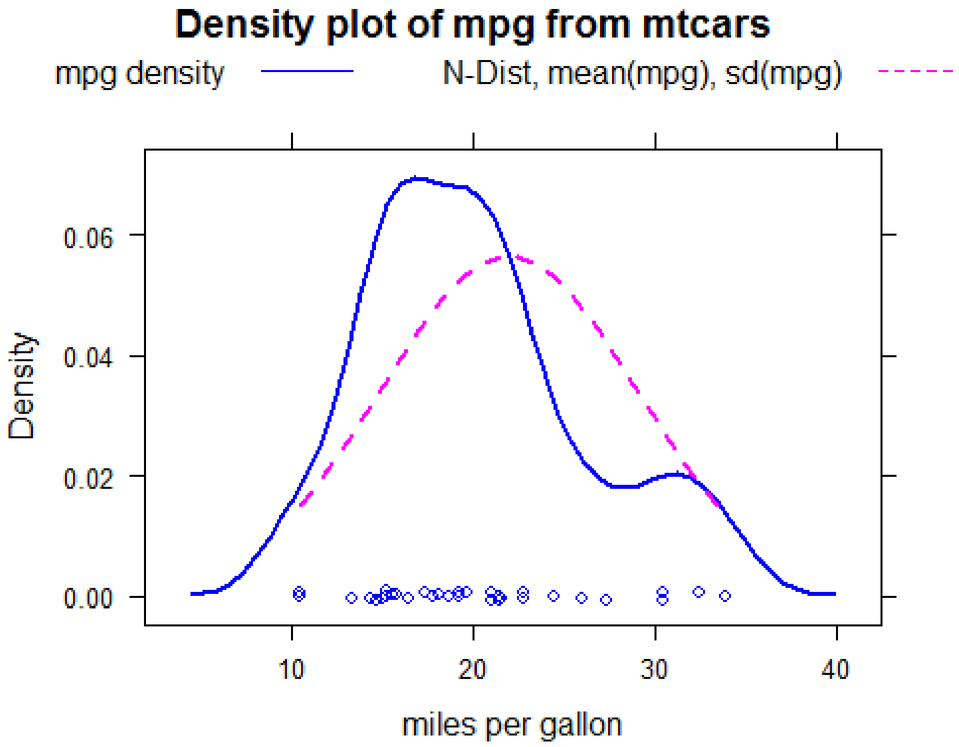
In closing, any advice for the car buyer who is conscious of fuel economy would be to prefer a lighter vehicle above all and then select a manual car which scored at the low end of the range on acceleration times. Of course this becomes a trade off between fuel economy and performance for the driver/buyer.

Needless to say that any inferences drawn from this data relate to cars manufactured in the 1970's. A great deal has changed since then and it would be wrong to apply these conclusions to cars in 2015.

Appendices

Exploratory Data Analysis

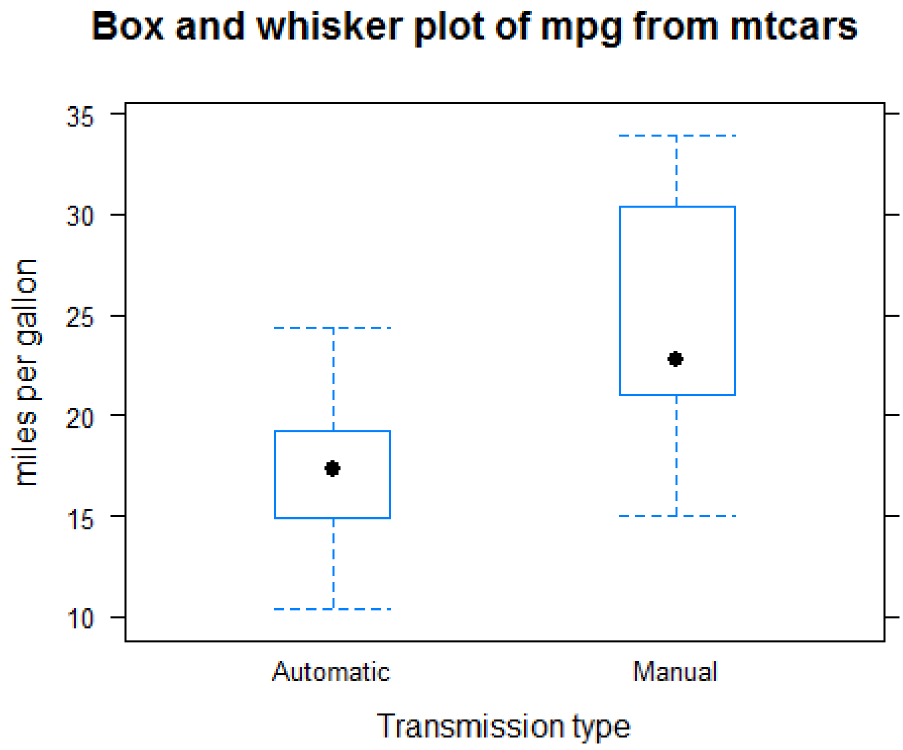
A density plot was created to examine the distribution of mpg data. A reference curve of the normal distribution with identical mean and sd is superimposed.



The mpg data is approximately normally distributed. Some deviation may be a result of the small data set, or there may be some systemic reason common to the car manufacturing process.

This investigation is primarily interested in which transmission type gives the best mpg value. The first thing to do is see what a simple comparison reveals.

A boxplot was generated to look at the two vehicle types side by side:



There does appear to be a difference between the two transmission types, with manual covering somewhat higher range of values than transmission.

A t-test was run of the mpg of automatic vs manual transmission to confirm and quantify the difference:

```
## 
##  Welch Two Sample t-test
## 
## data:  mpg by trans
## t = -3.7671, df = 18.332, p-value = 0.001374
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.280194  -3.209684
## sample estimates:
## mean in group Automatic    mean in group Manual 
##                17.14737                24.39231
```

Zero is not contained in the confidence interval and the null hypothesis is rejected. Cars with different transmission types have significantly different mpg ratings.

It can be inferred with >95% confidence that the difference in mean mpg of automatic vehicles in the sample is between the confidence intervals of the t-test.

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.

The next section details the process of finding a well fitting model.

Regression Model Selection step by step

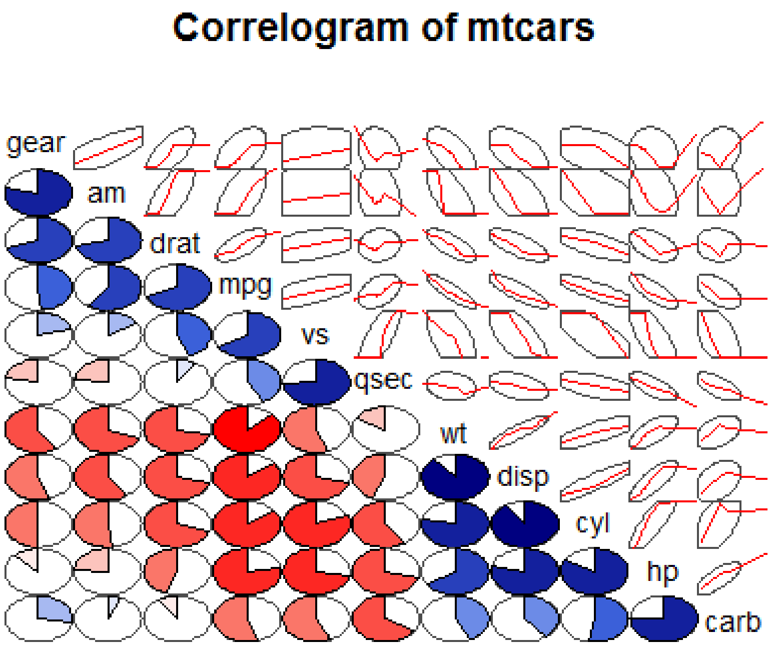
The first model to be explored is the known relationship between mpg and transmission. This will be the benchmark when looking for the best fitting model.

```
lm1 <- lm(mpg~trans, data = mtcars)
lm1$coefficients

## (Intercept) transManual
##   17.147368    7.244939

summary(lm1)$r.squared

## [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 which is a good deal little tighter than the t-test confidence interval above.

However, it can be seen from the R squared value of 0.3598 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:



This is useful as it helps to visualise the relationship of variables to mpg and also to each other.

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 and weight and 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:

```
lm2 <- lm(mpg~wt + disp + hp, data = mtcars)
lm2$coefficients

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

summary(lm2)$r.squared

## [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.

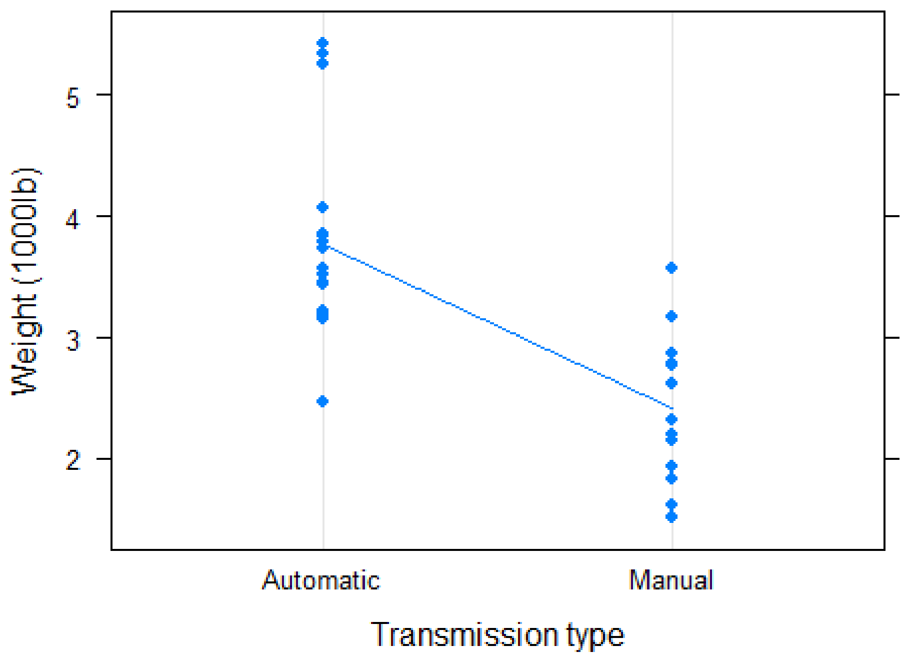
Furthermore, this R squared statistic of 0.81 explains far more of the variance than the transmission model.

This warrants a closer look at the relationship between weight and transmission type.

A t.test is run to quantifies the correlation between wt and trans. Also a dot plot is generated to show how the weights are distributed. A reference line is added to show the marginal difference (difference in averages between the two groups).

```
##
##  Welch Two Sample t-test
##
## data:  wt by trans
## t = 5.4939, df = 29.234, p-value = 6.272e-06
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.8525632 1.8632262
## sample estimates:
## mean in group Automatic    mean in group Manual
##                3.768895                2.411000
```

Distribution of vehicle weights by transmission



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:

```
lm3 <- lm(mpg~wt * trans, data = mtcars)
lm3$coefficients
```

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

```
summary(lm3)$r.squared
```

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

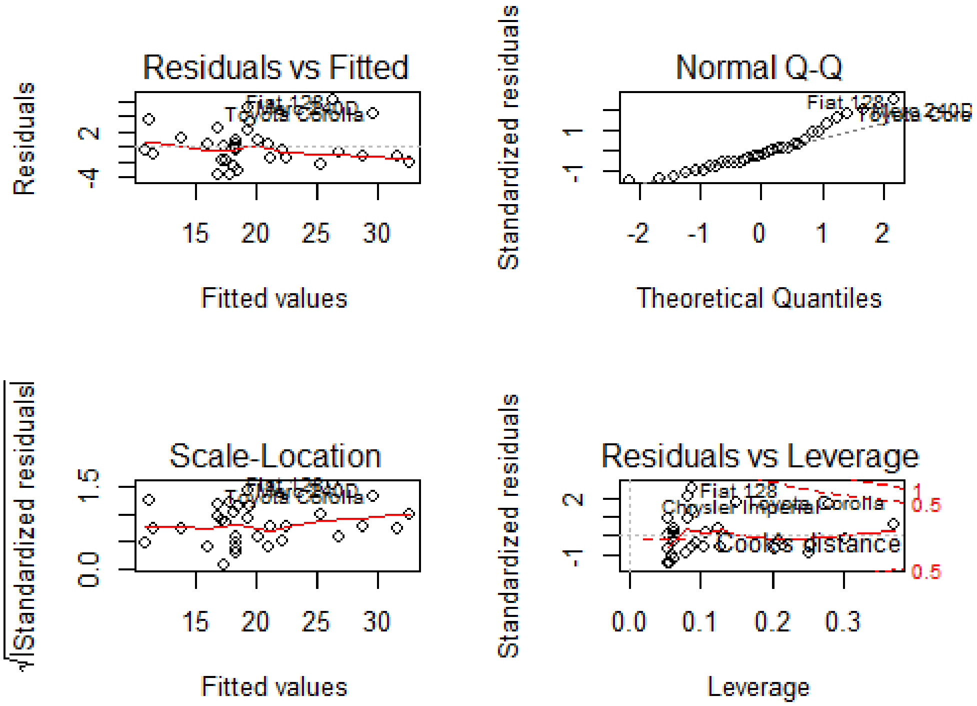
This model now explains 83% of the variation and is a small improvement on the wt only model.

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:

Diagnostic plots have been rendered 1/4 size for brevity. The reader is encouraged to run them full size from the R Mark Down document.



There is immediate concern from the residuals versus fitted plot that the majority of the residuals fall under the fitted line, implying that the model is systematically predicting values of mpg that are too high. This is almost certainly explained by the highlighted outliers "Merc 240D", "Fiat 128" and "Toyota Corolla."

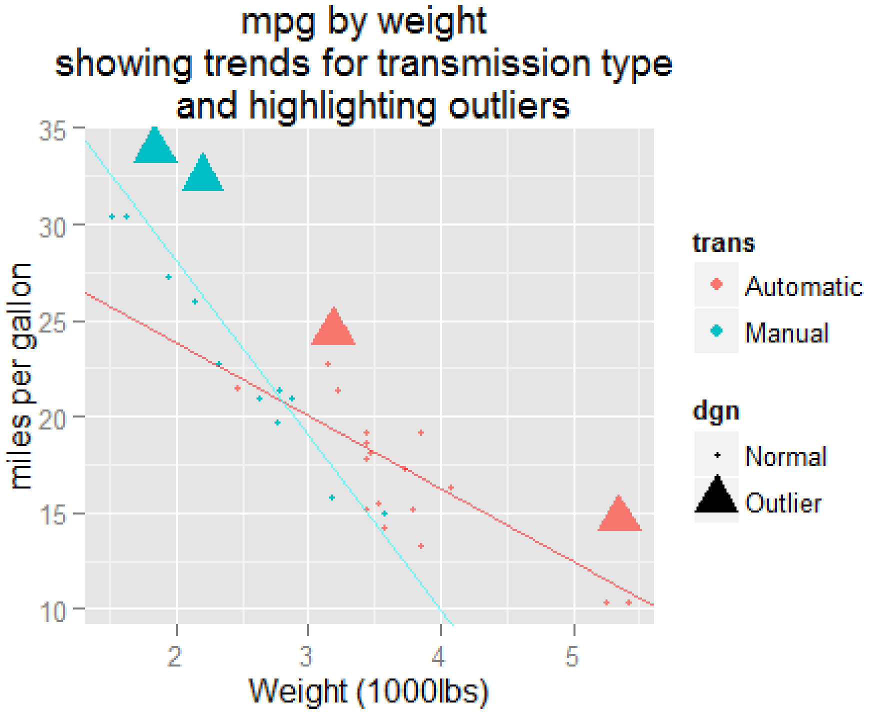
These same points appear again showing skew in the upper quantiles of the Normal Q-Q plot.

To determine what effect these are having on the model, it's worth revisiting the exploratory data analysis and find out more about these specific points.

```
mtcars$dgn <- factor(1 * rownames(mtcars) %in% c("Fiat 128", "Toyota Corolla", "Merc 240D",
"Chrysler Imperial"),labels = c("Normal", "Outlier"))

mtcars[mtcars$dgn == "Outlier", "trans"]
```

```
## [1] Automatic Automatic Manual    Manual
## Levels: Automatic Manual
```

There are two from each transmission type. The following plot reveals where they sit in relation to the other data points. Using the coefficients from the current linear model, two reference lines are added to show how mpg is dropping off with wt much more rapidly for manual vehicles.



Removing them is potentially going to make prediction more accurate, but is unlikely to affect the overall trends, which still hold. It is difficult to justify any particular threshold for selectively removing the points as they're not disproportionately far out of the range.

Two get any improvement on the model, it's worth revisiting the two remaining variables. Perhaps there is an obvious pattern that can be explored.

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
## Fiat X1-9         18.90  Normal
## Datsun 710        18.61  Normal
```

There is definitely something going on with the qsec variable as the main outliers are grouped together near the top.

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:

```
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
```

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

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.