Automatic or Manual Transmissions - Which is best?

Sean Clarke

17 August 2015

Exploratory Data Analysis

In order to understand the relationships between the different variables in the dataset a matrix of scatterplots was created as can be seen in Figure 1. This provided us with a means of getting a feel for how the variables depend on one another. As a result of this analysis and looking at the help page for the mtcars dataset it is clear that cyl (number of cylinders), vs (engine setup), am (automatic or manual gearbox) and gear (number of forward gears) are categorical variables and so they were converted to factors. **Note: For the variable am, 0 indicates automatic transmission, 1 indicates manual transmission.**

Additionlly, a correlation matrix was evaluated - the results are summarised in figure 2. If we define the threshold for significance to be ±0.8 after rounding the figures to one decimal place, we see the following:

- mpg is strongly correlated with wt, disp, cyl and hp
- wt is strongly correlated with disp and cyl
- hp is strongly correlated with disp, cyl and carb

Model Selection

It is well known that introducing multiple terms that have strong dependencies between them does not improve the quality of the fit. Additionally in this piece of work we are not looking at making predictions but rather in understanding the effect that an automated Vs Manual gearbox makes, so we favour a parsimonious model over an opaque one even if that is at the cost of this accuracy.

To this end, we propose to choose the variables that we include in our model based on how strongly they are correlated with mpg and discard variables from the resulting set based on strong mutual correlation. The exception to this is am which we include as it is the main motivation for this study! This means that we initially propose to include am, wt and hp in the fit.

In order to understand how effective the proposed choice of variables to include in the fit is, we initially fit a model mpg ~ am and then build a series of nested models, initially including the variables that we intend to include and then those that we propose to leave out based on the correlation analysys that we did in the previous section. These models when then be fed into and Anova analysis in order to enable us to understand the impact of including each team.

```
## Analysis of Variance Table
##
## Model  1: mpg ~ am
## Model  2: mpg ~ am + wt
## Model  3: mpg ~ am + wt + hp
## Model  4: mpg ~ am + wt + hp + disp
## Model  5: mpg ~ am + wt + hp + disp + cyl
## Model  6: mpg ~ am + wt + hp + disp + cyl + drat
## Model  7: mpg ~ am + wt + hp + disp + cyl + drat + qsec
## Model  8: mpg ~ am + wt + hp + disp + cyl + drat + qsec + vs
## Model  9: mpg ~ am + wt + hp + disp + cyl + drat + qsec + vs + gear
## Model 10: mpg ~ am + wt + hp + disp + cyl + drat + qsec + vs + gear + carb
##    Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1      30 720.90
## 2      29 278.32  1    442.58 64.6588 1.553e-07 ***
## 3      28 180.29  1     98.03 14.3216  0.001254 **
## 4      27 179.91  1      0.38  0.0560  0.815428
## 5      25 150.41  2     29.50  2.1548  0.143406
## 6      24 150.10  1      0.31  0.0450  0.834189
## 7      23 141.21  1      8.89  1.2995  0.268478
## 8      22 139.02  1      2.18  0.3189  0.578872
## 9      20 134.00  2      5.02  0.3668  0.697741
## 10     19 130.05  1      3.95  0.5771  0.456767
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We clearly see that adding wt and hp into the model results in a large change in the Residual Sum of Squares(RSS), large F values and p values that are clearly insigficant which indicates that these terms definitely improve the fit. When we add the other terms into the fit, the RSS changes are much smaller and the F-values and p values dont suggest that we should include these terms.

The one possible exception to this is when we add in cyl, however even the effect of including cyl is four times as small as adding the proposed terms so we do not include it. We will further analyse the quality of our fit to validate this choice

Results Analysis

Qualitative analysys

The values of the mean of mpg for automatic (17.1) and manual (24.4) cars suggests that cars with a manual transmission have a 42.3% higher mpg value than automatic vehicles. The Boxplot shown in figure 4 also backs up this assertion - the mean value for mpg for manual vehicles is higher than the top of the third quartile for automatic vehicles.

Quantitative Analysis

The curve that we have fitted is as follows

$$mpg_i = \beta_0 + X_{i,am}\beta_{am} + X_{i,wt}\beta_{wt} + X_{i,hp}\beta_{hp}$$

where each $X_{i,am}$ is binary and equal to 0 if the relevant vehicle is automatic and 1 if it is manual. It can be shown that if we control for the other varaibles and consider $E[Y_i]$ for the cases where $X_{i,am}$ is equal to 1 and 0, then β_{am} can be thought of as the increase in the mean comparing those in the group (vehicles with Manual transmission) to those outside of it (vehicles with automatic transmission). We can evaluate a confidence interval to attempt to apply some statistical inference to this idea

```
sumCoef<-summary(fit3)$coefficients
sumCoef[2,1]+c(-1,1)*qt(.975,df=fit3$df)*sumCoef[2,2]
```

```
## [1] -0.7357587  4.9031790
```

Therefore, the fit shows us that the mean mpg for vehicles with manual transmission is wihtin 95% certainty, -0.74 to 4.9 miles/gallon higher than for this with automatic transmission. Interestingly the 95% Confidence Interval includes the possibility of the vehicle being automatic making it more efficient - this isnt as counter-intuitive as it might seem as there is considerable overlap in the box plots in figure 4.

Its worth also noting that in this case, you cant just “go from automatic to manual” or vice-versa, this would mean changing physical components which would undoubtedly affect the validity of the model!

Appendix - Plots

Figure1: Scatterplot matrix for all the variables in the mtcars dataset"

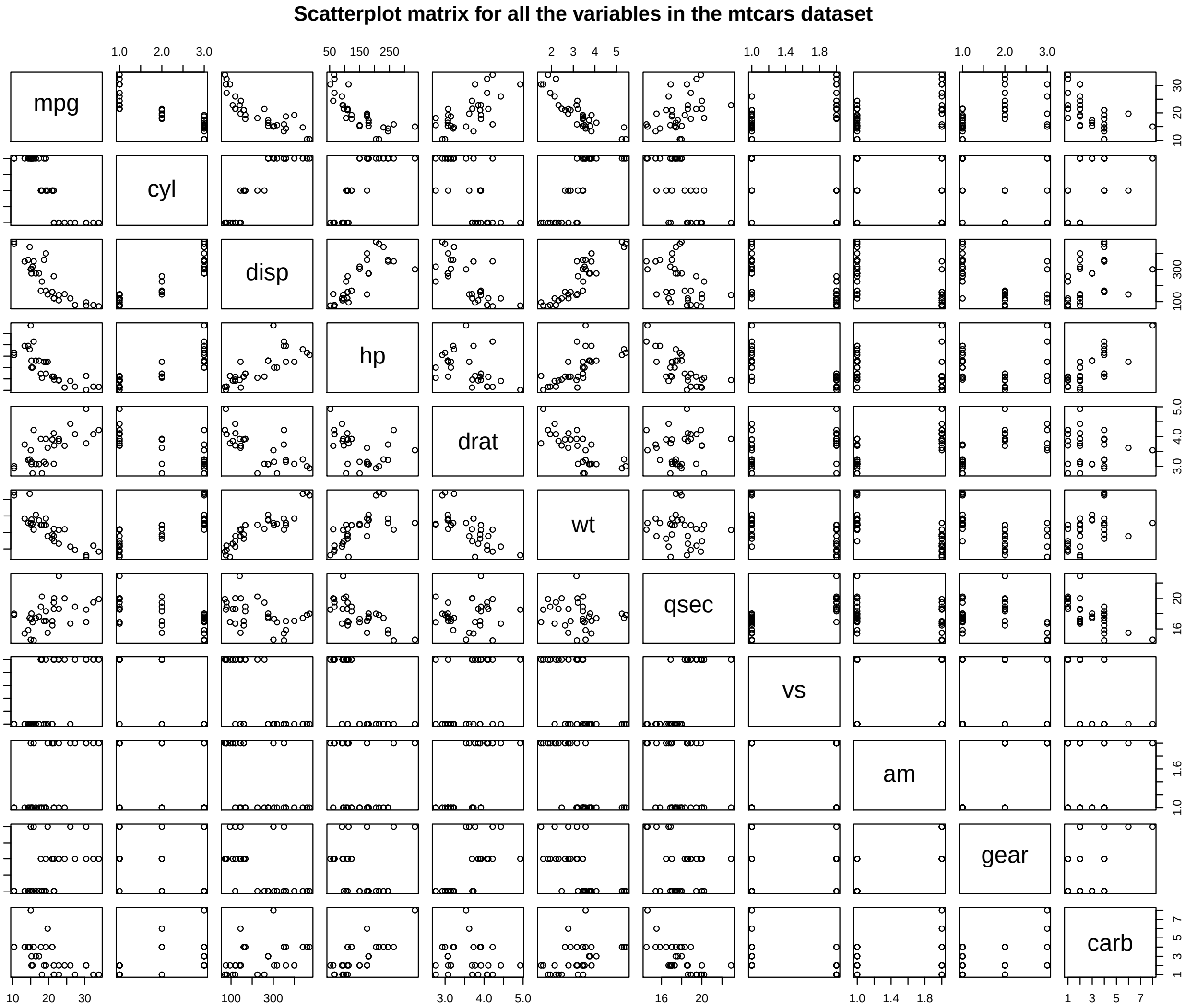


Figure 2: Correlation matrix for all the variables in the mtcars dataset

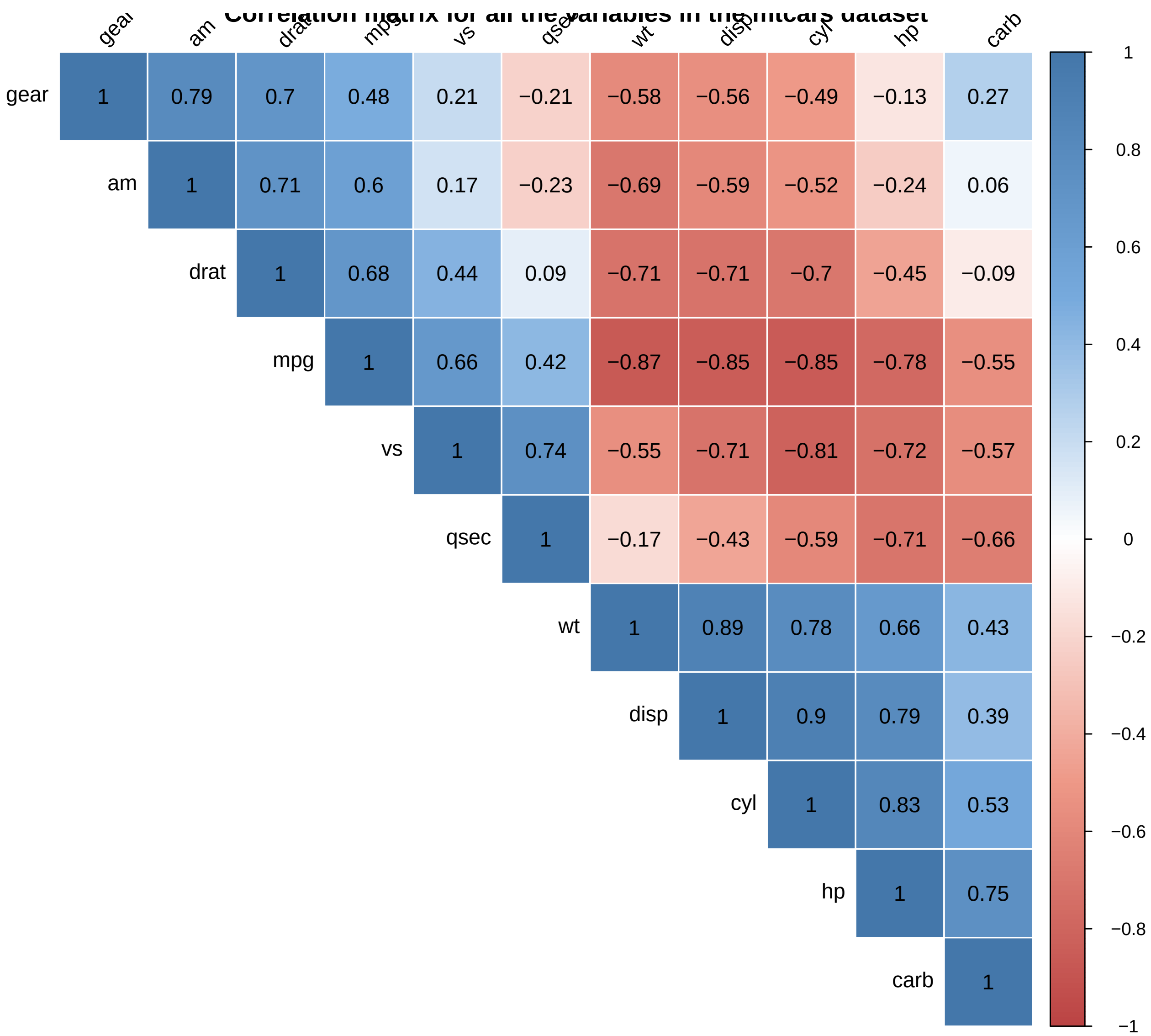


Figure 3: Standard R lm fit diagnostics for the fit that we chose to model mpg

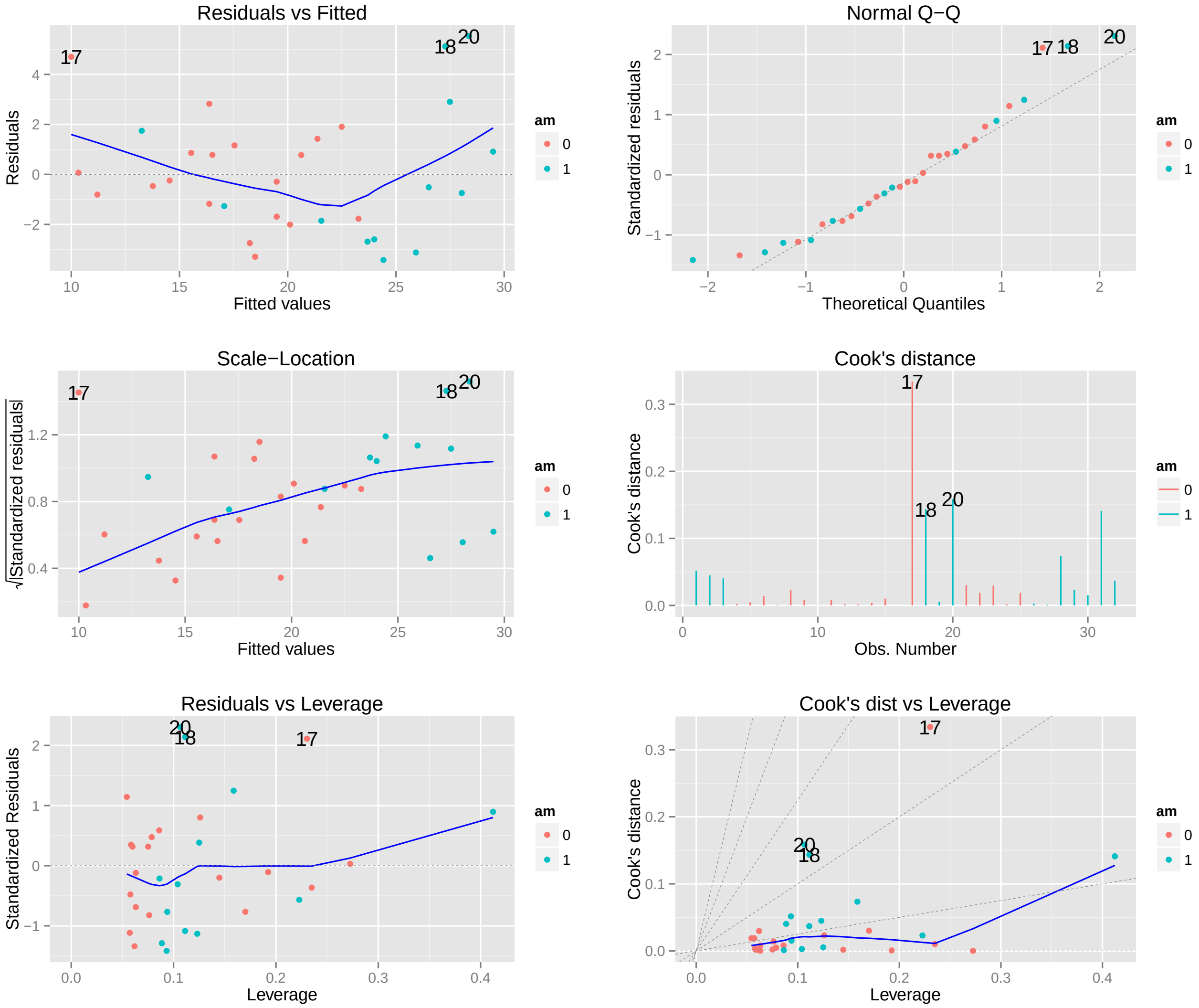
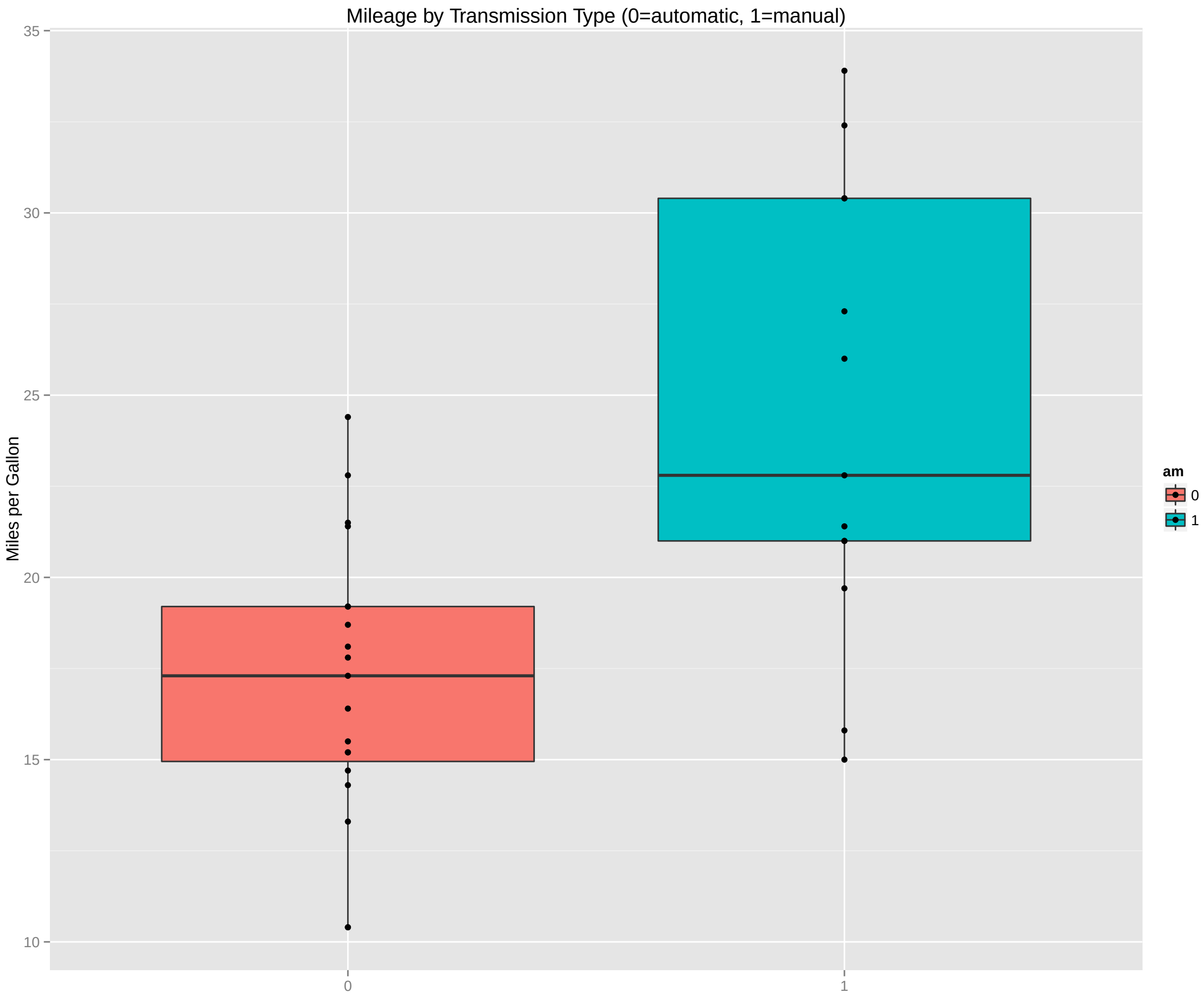


Figure 4: Boxplot showing how mpg depends on am in the mtcars dataset


Mileage by Transmission Type (0=automatic, 1=manual)
Miles per Gallon
am