Investigation into Fuel Efficiency in Cars

Executive summary

A statistical analysis was carried out to investigate how fuel consumption depends on a number of variables - the key question was whether the data showed that cars with manual transmissions are more efficient that those with automatic ones. The analysis showed that a number of other variables were also important, namely the weight of the vehicle and the power of the engine. It was suspected that several of the other variables did not need to be included as they implicitly depended on these variables, an assumption which the analysis bore out. The conclusion of the analysis is that in general manual vehicles tend to be more efficient than automatic ones, but the other factors mentioned are also important. For example a very light automatic vehicle with a less powerful engine is likely to be more efficient than a very heavy one with a very powerful engine.

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 that mpg is strongly correlated with wt, disp, cyl and hp. In turn, wt is strongly correlated with disp and cyl and 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 versus Manual gearbox makes, so we favour a parsimonious model over a more opaque one even if that is at the cost of some 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 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. These models are then compared using an Anova analysis.

```
## 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 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 are much less supportive of including these terms.

The one possible exception to the above statement 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 in the interests of parsimony.

Figure 3 shows the standard diagnostic plots for the lm function. The normal Q-Q plot shows that the assumption that the errors in the data are normal is a pretty good one. The residuals versus fitted plot shows that the residuals are pretty well distributed either side of 0 and there is no obvious pattern in them which is a good sign. The scale-location plot shows a potential problem in that it is tending to rise with increasing X which is a sign of heteroscedasticity. However there are some points highlighted that have a large Cook's distance that might be causing this - these points would clearly bear some further investigation.

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 (as am is categorical) 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. 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 within 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 this element of the discussion is slightly contrived, as 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!

Conclusions

- Manual vehicles tend to be more efficient that automatic ones. However other factors also govern efficiency and these also need to be taken into account.
- The fuel efficiency is inversely proportial to the weight of the vehicle.
- The fuel efficiency is inversely proportional to the power of the vehicles engine.

Appendix - Plots

Figure1: Scatterplot matrix for all the non-categorical variables in the mtcars dataset

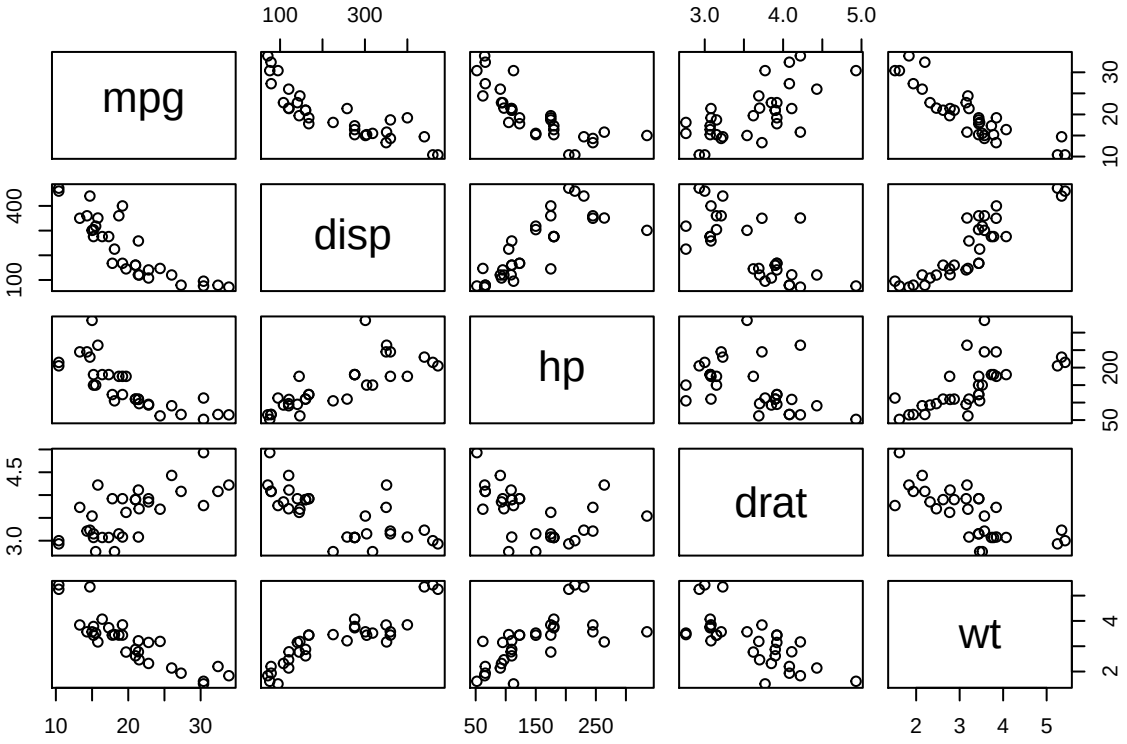

Table A: Fit parameters and Confidence Intervals

```
summary(fit3)$coefficients
```

```
##                Estimate  Std. Error   t value     Pr(>|t|)
## (Intercept) 34.00287512 2.642659337 12.866916 2.824030e-13
## am1          2.08371013 1.376420152  1.513862 1.412682e-01
## wt          -2.87857541 0.904970538 -3.180850 3.574031e-03
## hp          -0.03747873 0.009605422 -3.901830 5.464023e-04
```

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

```
## [1] 28.58963 39.41612
```

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

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

```
sumCoef[3,1]+c(-1,1)*qt(.975,df=fit3$df)*sumCoef[3,2]
```

```
## [1] -4.732324 -1.024827
```

```
sumCoef[4,1]+c(-1,1)*qt(.975,df=fit3$df)*sumCoef[4,2]
```

```
## [1] -0.05715454 -0.01780291
```

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

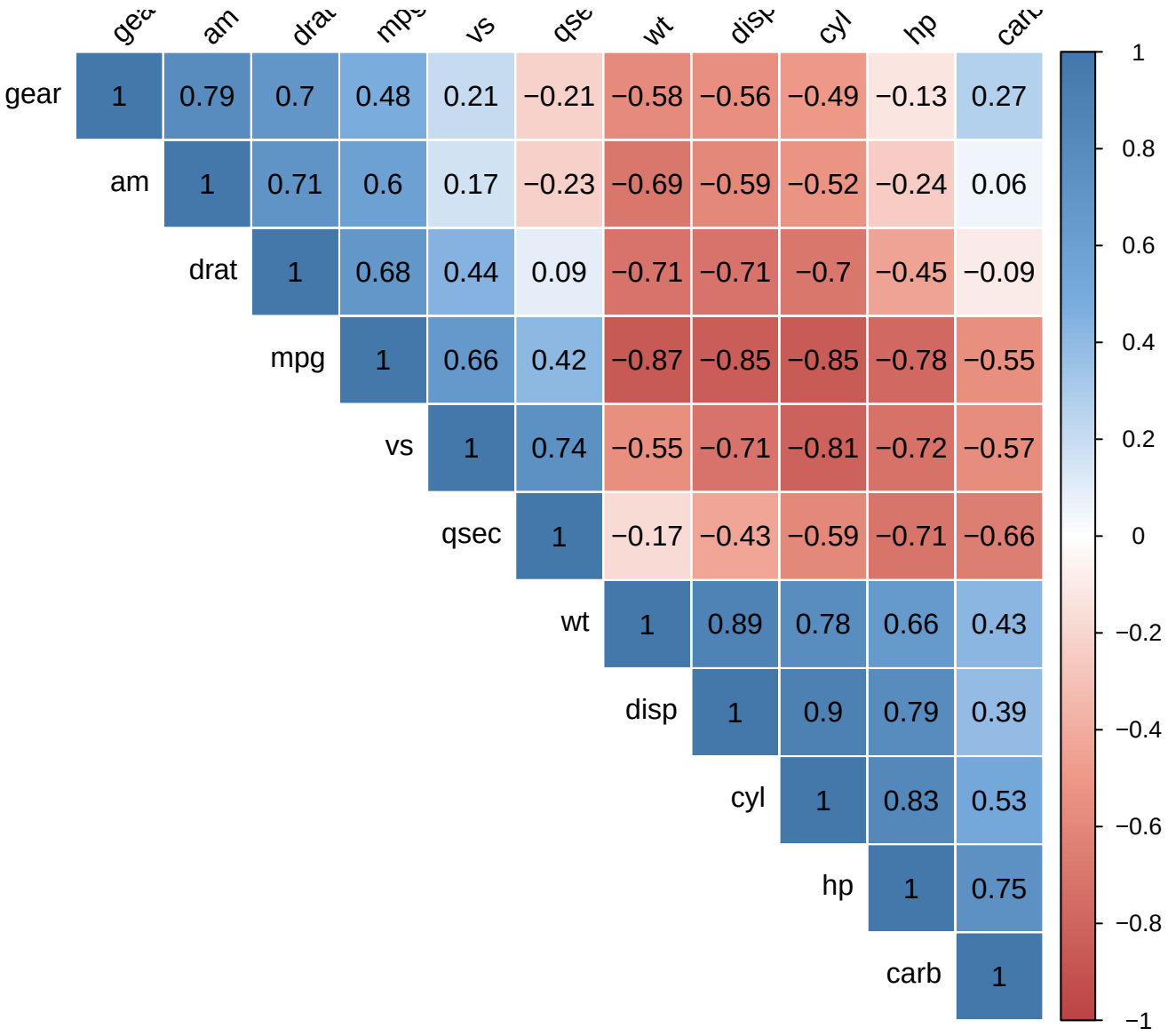

Figure 3: Standard R lm fit diagnostics for the fit that we chose to model mpg

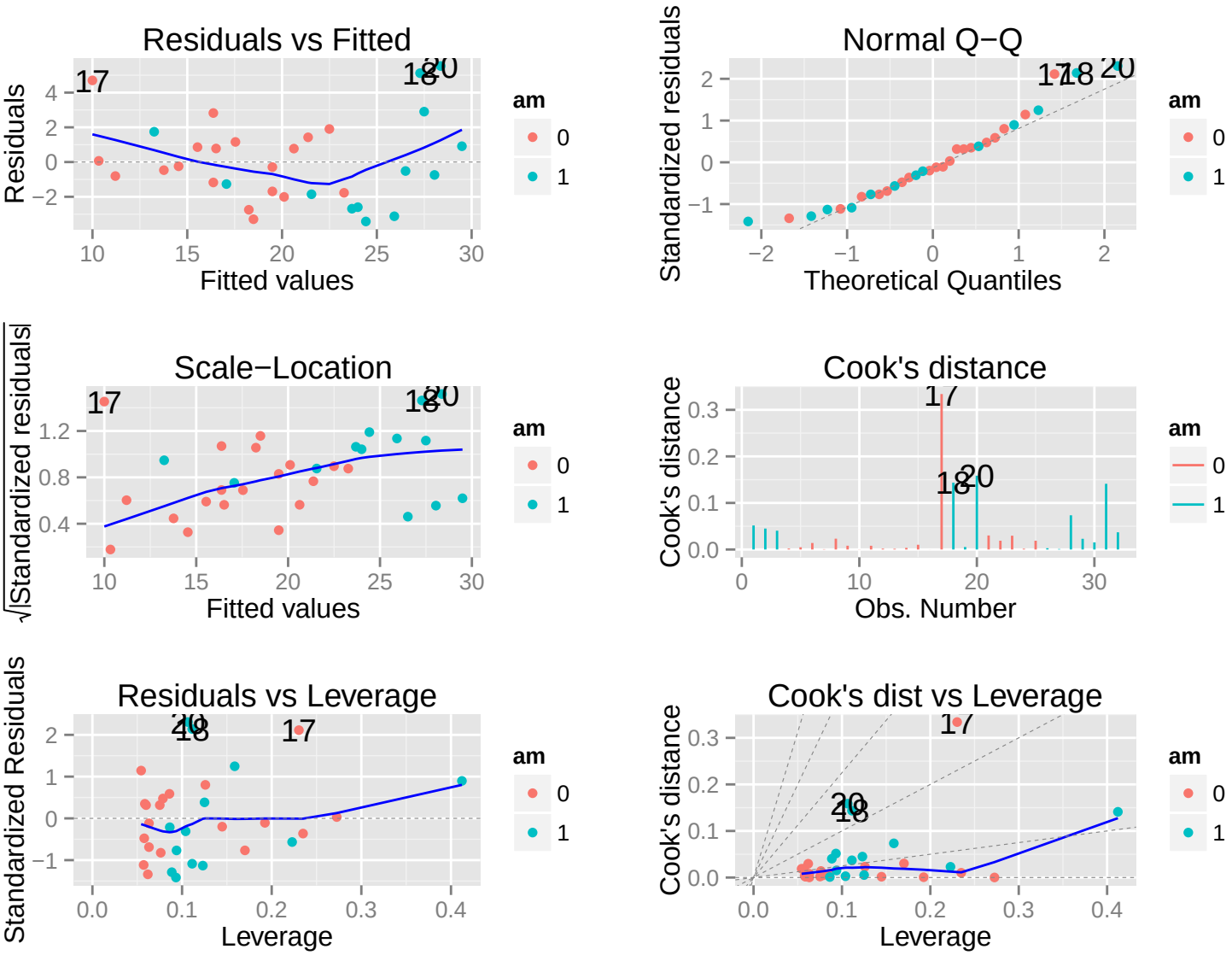

Figure 4: Boxplot showing how mpg depends on am in the mtcars dataset