An analysis of factors which affect automobile MPG.

Executive Summary

The Motor Trend mtcars dataset was analysed to explore the relationship between automobile MPG and a set of variables with specific reference to transmission type. After some exploratory analysis and the selection of an appropriate linear regression model it was found that manual transmissions exhibit a better (higher) MPG compared to automatics. We found that, while holding other variables constant, manual transmissions were more fuel efficient by 2.936 MPG. According to the model, the MPG of manual transmissions would fall in the interval of a 0.04573 to 5.826 MPG improvement with 95% confidence.

Introduction

Motor Trend has compiled data on 32 automobiles in the USA from 1973/74. These data are comprised of fuel consumption along with 10 aspects of automobile design and performance. A sample of the data follows:

```
##                 mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4      21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Hornet 4 Drive 21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
```

With the aid of this data, Motor Trend set out to investigate two questions:
Is an automatic or manual transmission better for MPG?
Quantify the MPG difference between automatic and manual transmissions.

Exploratory Analysis

Transmission type was plotted versus MPG to determine if there was any basic relationship (Appendix A). It shows manual transmissions appear to be more fuel efficient. Next a sequence of pair plots we generated (Appendix B). Green points for manual and red for automatic transmissions. These give indication that a few features may relate to MPG and exhibit some sort of linear relationship. In order to probe further we fitted a linear model with all available features to MPG to get an understanding of the coefficients as follows:

```
## $coefficients
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.30337   18.71788  0.6573  0.51812
## cyl         -0.11144    1.04502 -0.1066  0.91609
## disp         0.01334    0.01786  0.7468  0.46349
## hp          -0.02148    0.02177 -0.9868  0.33496
## drat         0.78711    1.63537  0.4813  0.63528
## wt          -3.71530    1.89441 -1.9612  0.06325
## qsec         0.82104    0.73084  1.1234  0.27394
## vs           0.31776    2.10451  0.1510  0.88142
## am1          2.52023    2.05665  1.2254  0.23399
## gear         0.65541    1.49326  0.4389  0.66521
## carb        -0.19942    0.82875 -0.2406  0.81218
```

From the coefficients it appears that weight(wt) is most significant, with a p-value of 0.06325, followed by transmission type(am) and 1/4 mile time(qsec). In order to validate this observation we performed a step-wise model selection via the stepAIC function in the MASS package to determine the best model as shown below:

```r
library(MASS); step <- stepAIC(fitAll, direction="both"); selectedFit <- step$call; selectedFit;
```

```
## lm(formula = mpg ~ wt + qsec + am, data = mtcars)
```

Step-wise model selection selects wt, qsec and am as the most significant variables confirming our initial thoughts. Lastly we used anova to determine if a better fitting model could be obtained using interactions between features with the following formula: *lm(formula = mpg ~ wt * qsec * am, data = mtcars)*

Results
The coefficients of the step-wise selected model are as follows:

```
## $coefficients
##             Estimate Std. Error t value  Pr(>|t|)
## (Intercept)    9.618     6.9596   1.382 1.779e-01
## wt            -3.917     0.7112  -5.507 6.953e-06
## qsec           1.226     0.2887   4.247 2.162e-04
## am1            2.936     1.4109   2.081 4.672e-02
```

All 3 features have p-values < 0.05 and thus have significant influence on the model. Now let's look at the formula identified with anova to see if the model could be improved using interactions:

```
## $coefficients
##             Estimate Std. Error  t value Pr(>|t|)
## (Intercept) -11.8073    57.7960 -0.20429   0.8398
## wt            3.8670    17.0915  0.22625   0.8229
## qsec          2.2148     3.1734  0.69793   0.4919
## am1          -5.5842    65.4915 -0.08527   0.9328
## wt:qsec      -0.3805     0.9467 -0.40189   0.6913
## wt:am1        4.0906    20.5487  0.19907   0.8439
## qsec:am1      1.1768     3.6310  0.32409   0.7487
## wt:qsec:am1  -0.5009     1.1648 -0.43006   0.6710
```

None of the p-values are near 0.05. We thus reverted back to the step-wise model and generated its residual diagnostic plots (Appendix C) and its confidence intervals below:

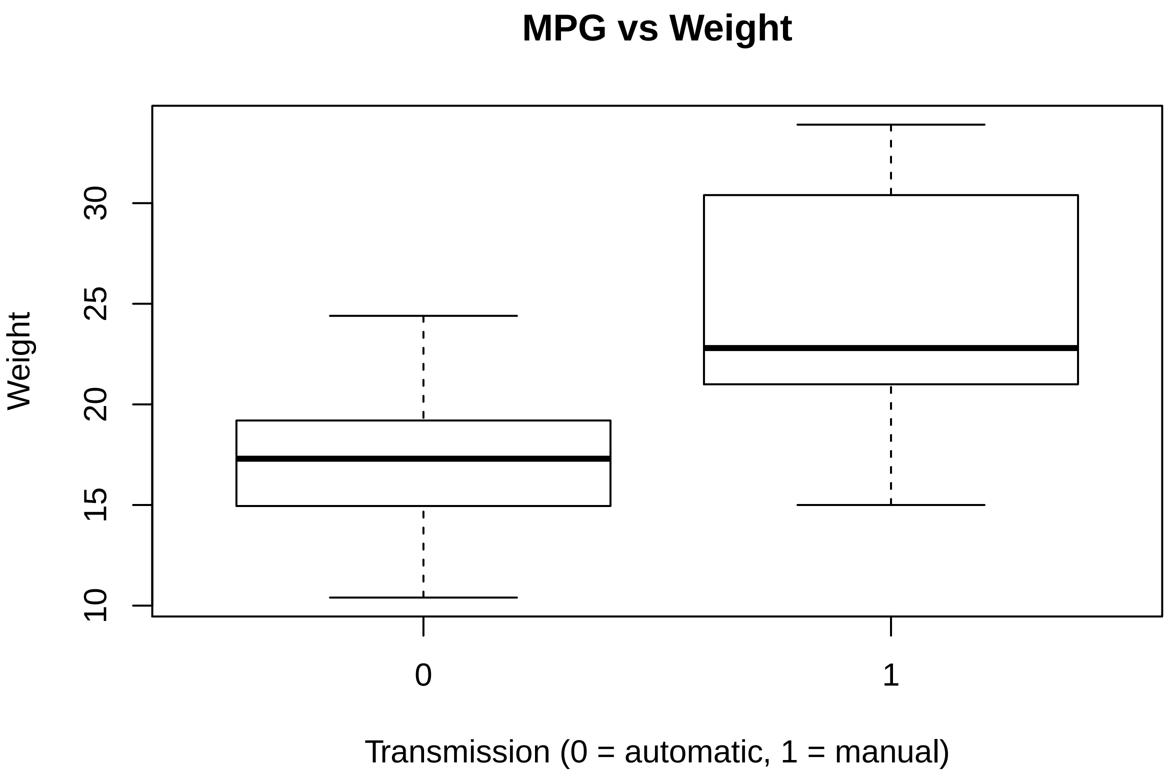
```
confint(lm(formula = mpg ~ wt + qsec + am, data = mtcars))
```

```
##                2.5 % 97.5 %
## (Intercept) -4.63830 23.874
## wt          -5.37333 -2.460
## qsec         0.63457  1.817
## am1          0.04573  5.826
```

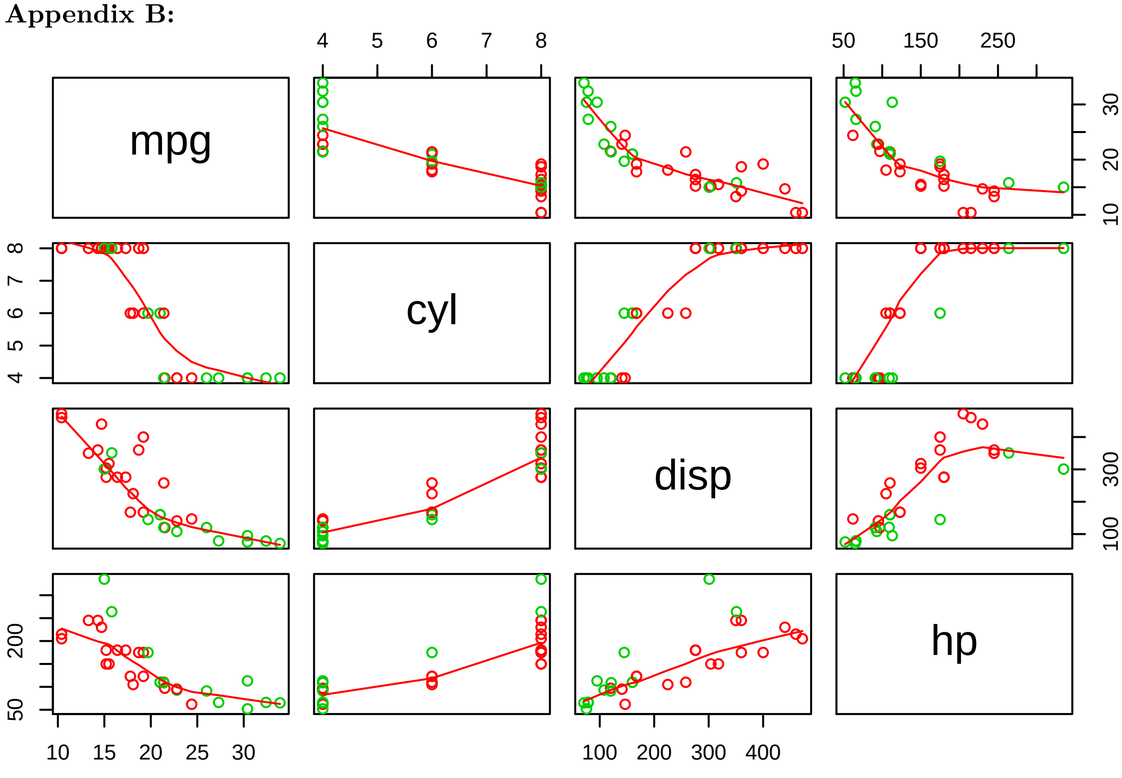
Conclusion
The model obtained via step-wise selection was chosen as the most appropriate. Anova gave indications that a model with feature interaction produces a better fit but looking at the p-values of that model it was felt this was due to over-fitting the model to the data. Looking at the residual diagnostic plots the model appears to be of average fit. The residuals have a fairly random distribution which is good and the Q-Q plot shows a distribution tending toward normal, although it does drift a little on the upper end which one would need to keep in mind.

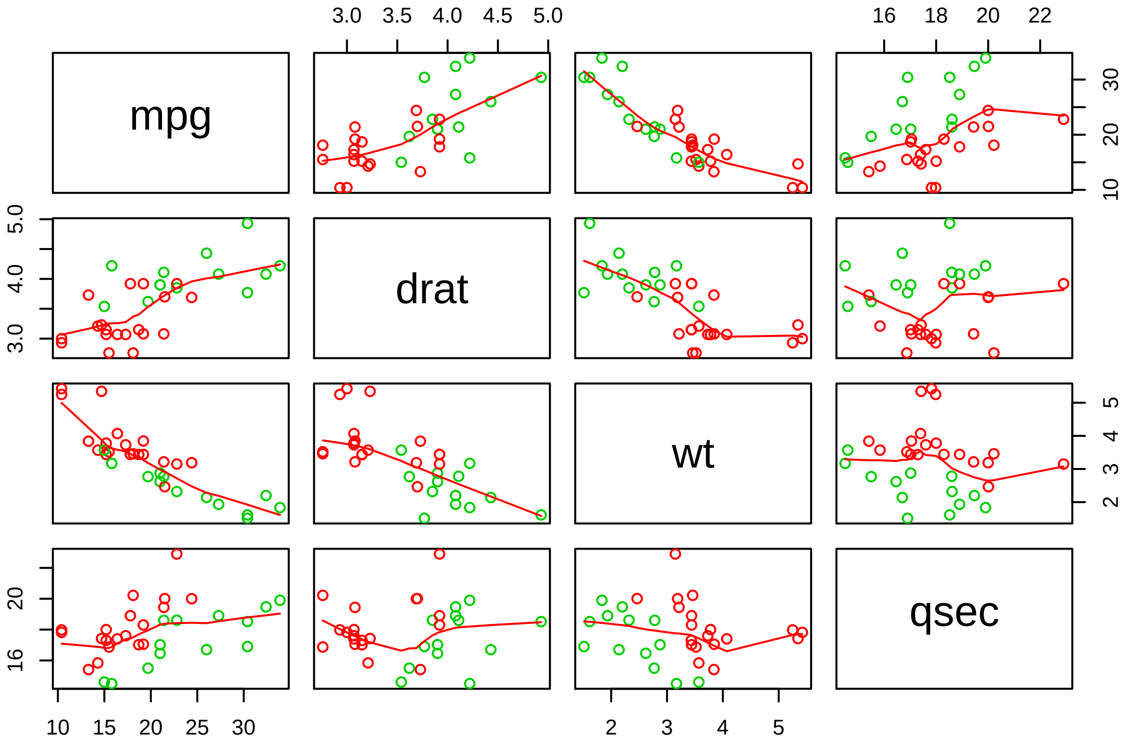
From the coefficients we can see that if we hold wt and qsec constant, the model indicates manual transmissions will have a improvement of 2.936 MPG over automatics. Looking at the confidence intervals we can also see that we have 95% confidence that manual transmissions will have a higher MPG in the interval of 0.04573 to 5.826 MPG more than automatic transmissions. For the given model we can therefore be fairly certain that manual transmissions will be to some degree more fuel efficient.

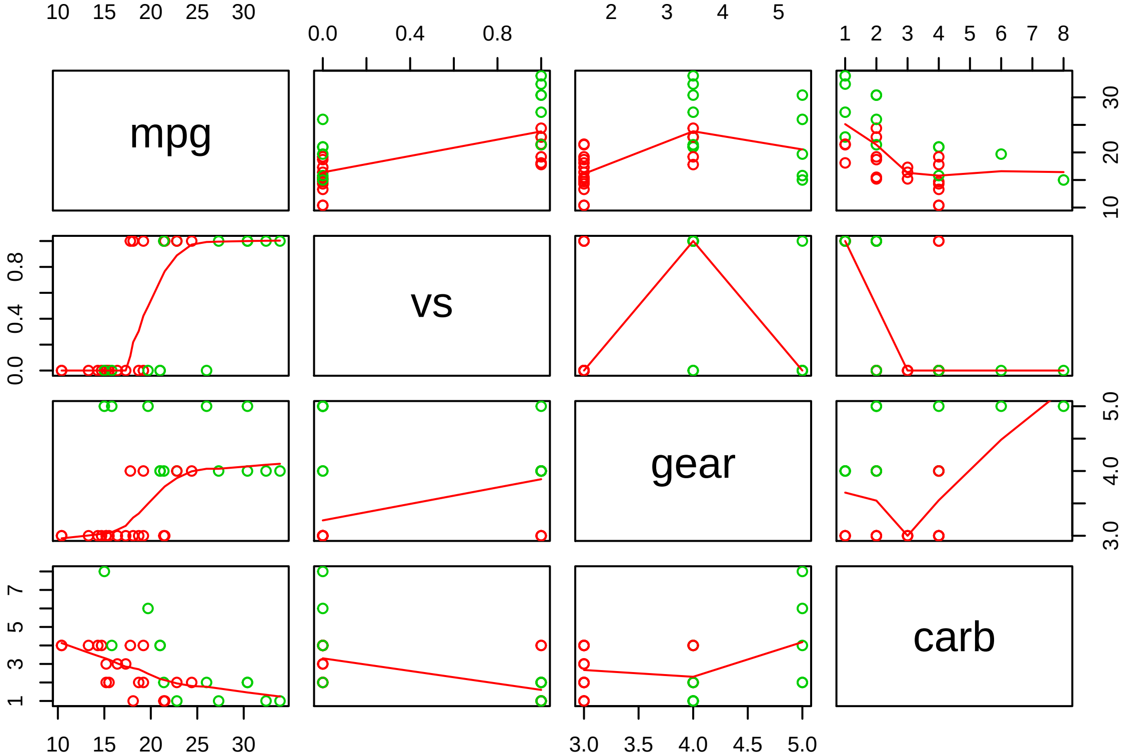
Appendix A:



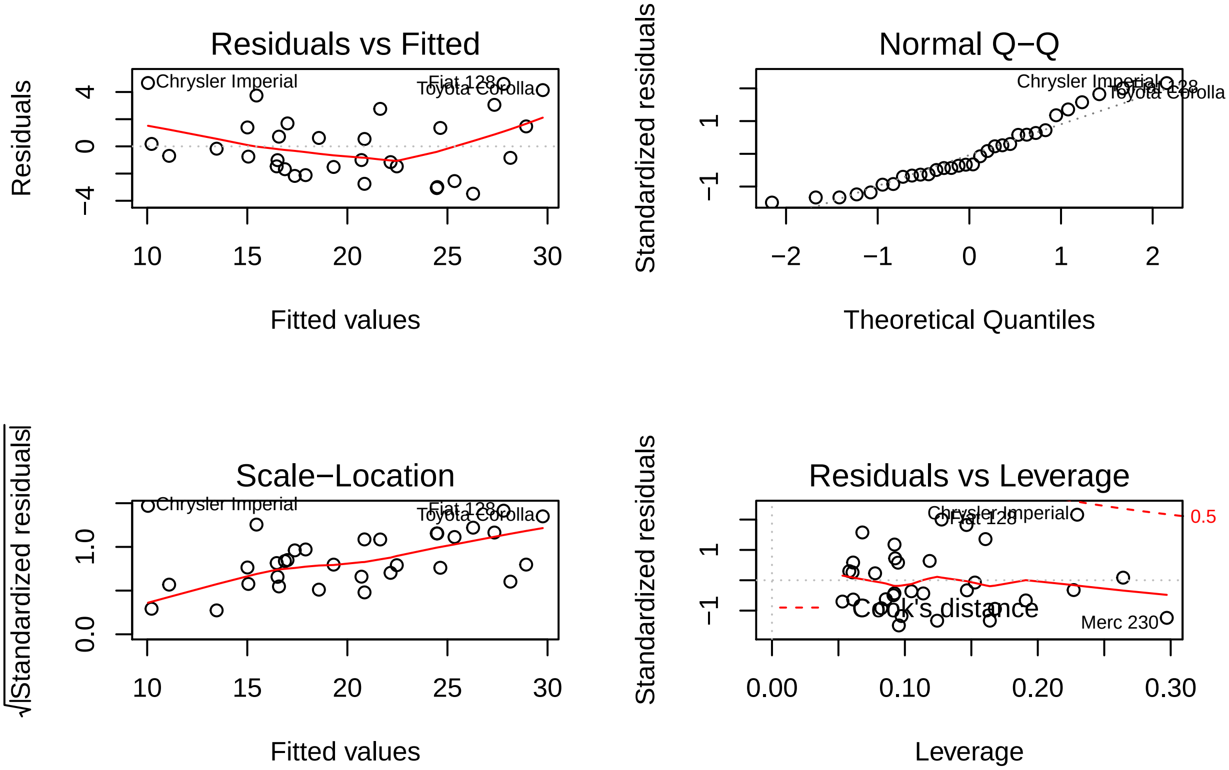
Appendix B:




mpg
drat
wt
qsec


mpg
vs
gear
carb

Appendix C:



Residuals vs Fitted
Residuals
Fitted values
Chrysler Imperial
Fiat 128
Toyota Corolla
Normal Q−Q
Standardized residuals
Theoretical Quantiles
Scale−Location
√|Standardized residuals|
Residuals vs Leverage
Leverage
Cook's distance
Merc 230
0.5