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 correlation matrix was evaluated - the results are summarise in figure one. 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

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.

Model Selection

In order to understand how effective the propsoed 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

When we add the rest of the terms in, the im

Appendix - Plots

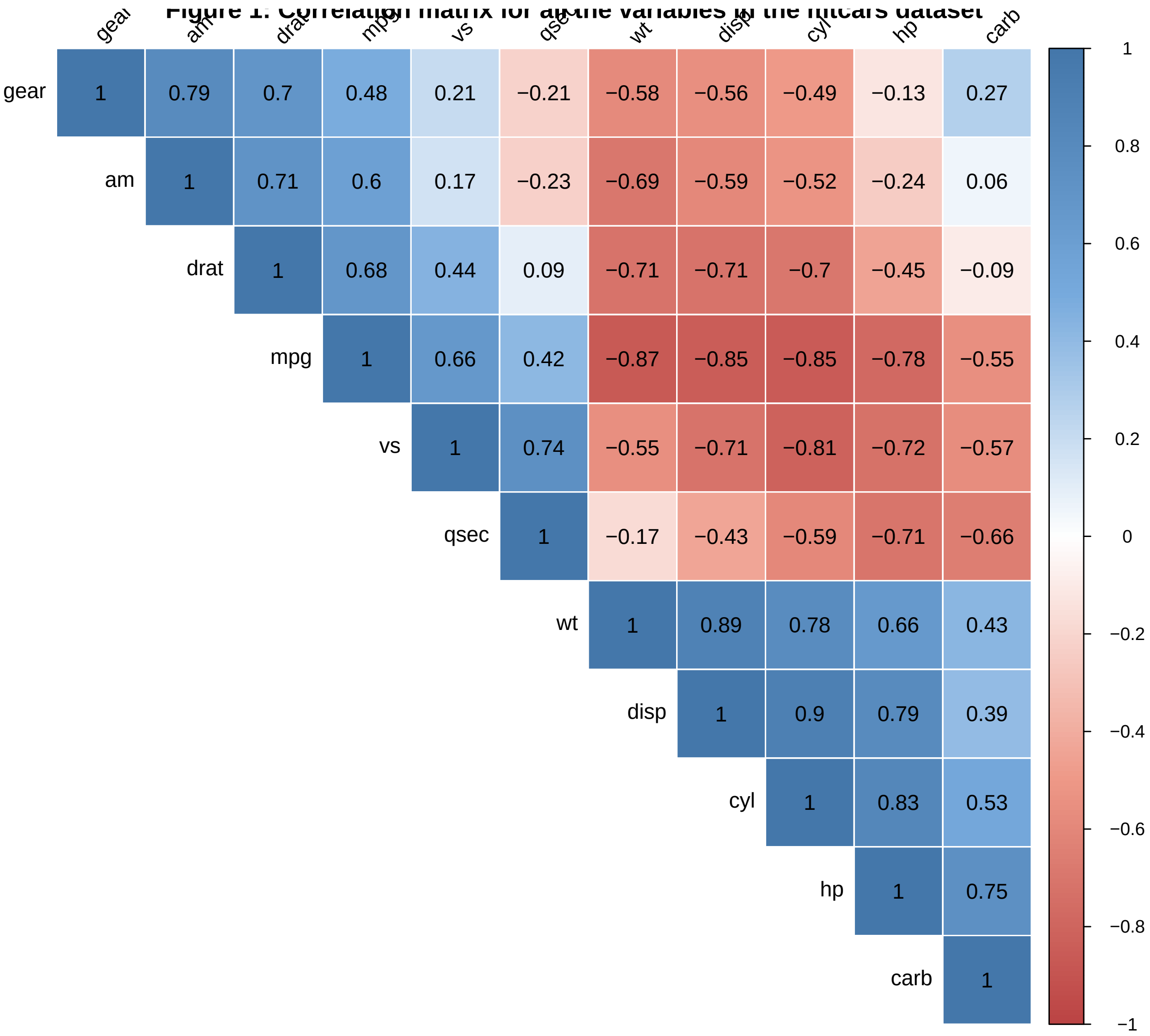
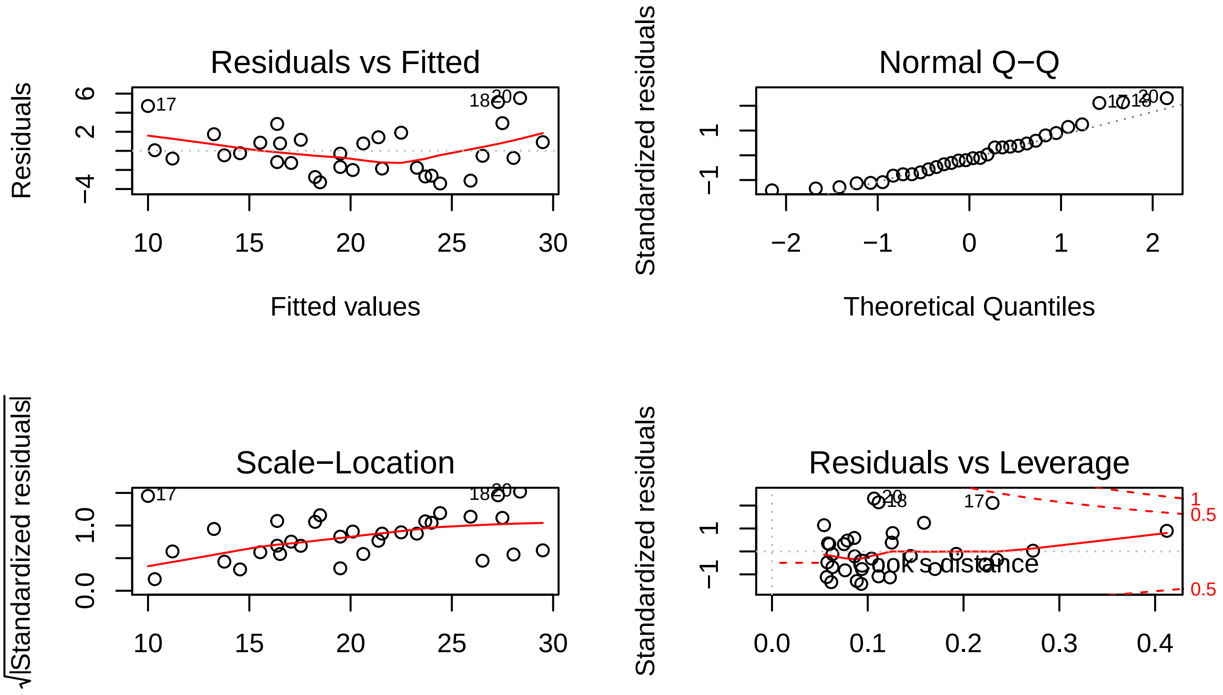


Figure 1: Correlation matrix for all the variables in the mtcars dataset

	gear	am	drat	mpg	vs	qsec	wt	disp	cyl	hp	carb
gear	1	0.79	0.7	0.48	0.21	−0.21	−0.58	−0.56	−0.49	−0.13	0.27
am		1	0.71	0.6	0.17	−0.23	−0.69	−0.59	−0.52	−0.24	0.06
drat			1	0.68	0.44	0.09	−0.71	−0.71	−0.7	−0.45	−0.09
mpg				1	0.66	0.42	−0.87	−0.85	−0.85	−0.78	−0.55
vs					1	0.74	−0.55	−0.71	−0.81	−0.72	−0.57
qsec						1	−0.17	−0.43	−0.59	−0.71	−0.66
wt							1	0.89	0.78	0.66	0.43
disp								1	0.9	0.79	0.39
cyl									1	0.83	0.53
hp										1	0.75
carb											1

lm(mpg ~ am + wt + hp)


Residuals vs Fitted
Residuals
6
2
−4
10
15
20
25
30
Fitted values
17
18
20
Normal Q−Q
Standardized residuals
1
−1
−2
−1
0
1
2
Theoretical Quantiles
17
18
20
Scale−Location
√|Standardized residuals|
1.0
0.0
10
15
20
25
30
Fitted values
17
18
20
Residuals vs Leverage
Standardized residuals
1
−1
0.0
0.1
0.2
0.3
0.4
Leverage
20
18
17
Cook's distance
1
0.5
0.5