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.

From the calculated correlation values for the continuous variables, we see some positive correlation between fuel consumption and the rear axel ratio and 1/4 mile time. We see strong negative correlation between fuel consumption and Displacement, Horsepower, and weight. There is a weaker negative correlation between fuel consumption and the number of carburetors.

Interestingly (and maybe not surprisingly) fuel consumtions appears to depend in almost all the other factors!

Model Selection

In order to find the optimal fit, we proceed to fit nested models, we will then proceed to evaluate an ANOVA for these fits and use the output of the F-Test to identify the necessary terms.

```
fit1 <- lm(mpg ~  am , mtcars)
fit2 <- lm(mpg ~ am + wt, mtcars)
fit3 <- lm(mpg ~ am + wt + disp, mtcars)
fit4 <- lm(mpg ~ am + wt + disp + cyl, mtcars)
fit5 <- lm(mpg ~ am + wt + disp + cyl + drat, mtcars)
fit6 <- lm(mpg ~ am + wt + disp + cyl + drat + qsec, mtcars)
fit7 <- lm(mpg ~ am + wt + disp + cyl + drat + qsec + vs, mtcars)
fit8 <- lm(mpg ~ am + wt + disp + cyl + drat + qsec + vs + gear, mtcars)
fit9 <- lm(mpg ~ am + wt + disp + cyl + drat + qsec + vs + gear + carb, mtcars)
fit10 <- lm(mpg ~ am + wt + disp + cyl + drat + qsec + vs + gear + carb + hp, mtcars)
anova(fit1, fit2, fit3, fit4, fit5, fit6, fit7, fit8, fit9, fit10)
```

```
## Analysis of Variance Table
##
```

```
## Model  1: mpg ~ am
## Model  2: mpg ~ am + wt
## Model  3: mpg ~ am + wt + disp
## Model  4: mpg ~ am + wt + disp + cyl
## Model  5: mpg ~ am + wt + disp + cyl + drat
## Model  6: mpg ~ am + wt + disp + cyl + drat + qsec
## Model  7: mpg ~ am + wt + disp + cyl + drat + qsec + vs
## Model  8: mpg ~ am + wt + disp + cyl + drat + qsec + vs + gear
## Model  9: mpg ~ am + wt + disp + cyl + drat + qsec + vs + gear + carb
## Model 10: mpg ~ am + wt + disp + cyl + drat + qsec + vs + gear + carb +
##     hp
##    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      28 246.56  1     31.76  4.6405   0.04427 *
## 4      26 182.87  2     63.69  4.6522   0.02267 *
## 5      25 182.67  1      0.19  0.0284   0.86786
## 6      24 157.22  1     25.46  3.7193   0.06887 .
## 7      23 157.20  1      0.02  0.0025   0.96055
## 8      21 156.79  2      0.40  0.0296   0.97090
## 9      20 152.21  1      4.59  0.6701   0.42318
## 10     19 130.05  1     22.16  3.2371   0.08789 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
fit1 <- lm(mpg ~  am , mtcars)
fit2 <- lm(mpg ~  am + wt + hp , mtcars)
fit3 <- lm(mpg ~  am + wt + hp + cyl + disp , mtcars)
fit4  <- lm(mpg ~  am + wt + hp + cyl + disp + gear + carb , mtcars)
fit5  <- lm(mpg ~  am + wt + hp + cyl + disp + gear + carb + vs + qsec , mtcars)
anova(fit1, fit2, fit3, fit4, fit5 )
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + wt + hp
## Model 3: mpg ~ am + wt + hp + cyl + disp
## Model 4: mpg ~ am + wt + hp + cyl + disp + gear + carb
## Model 5: mpg ~ am + wt + hp + cyl + disp + gear + carb + vs + qsec
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 720.90
## 2     28 180.29  2    540.61 41.2701 7.968e-08 ***
## 3     25 150.41  3     29.88  1.5208    0.2398
## 4     22 148.35  3      2.06  0.1047    0.9563
## 5     20 130.99  2     17.36  1.3252    0.2881
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
fit1 <- lm(mpg ~  am , mtcars)
fit2 <- lm(mpg ~  am + wt + hp , mtcars)
fit3 <- lm(mpg ~  am + wt + hp + cyl + disp , mtcars)
fit4  <- lm(mpg ~  am + wt + hp + cyl + disp + gear + carb , mtcars)
fit5  <- lm(mpg ~  am + wt + hp + cyl + disp + gear + carb + vs + qsec , mtcars)
anova(fit1, fit2, fit3, fit4, fit5 )
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + wt + hp
## Model 3: mpg ~ am + wt + hp + cyl + disp
## Model 4: mpg ~ am + wt + hp + cyl + disp + gear + carb
## Model 5: mpg ~ am + wt + hp + cyl + disp + gear + carb + vs + qsec
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 720.90
## 2     28 180.29  2    540.61 41.2701 7.968e-08 ***
## 3     25 150.41  3     29.88  1.5208    0.2398
## 4     22 148.35  3      2.06  0.1047    0.9563
## 5     20 130.99  2     17.36  1.3252    0.2881
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Appendix - Plots

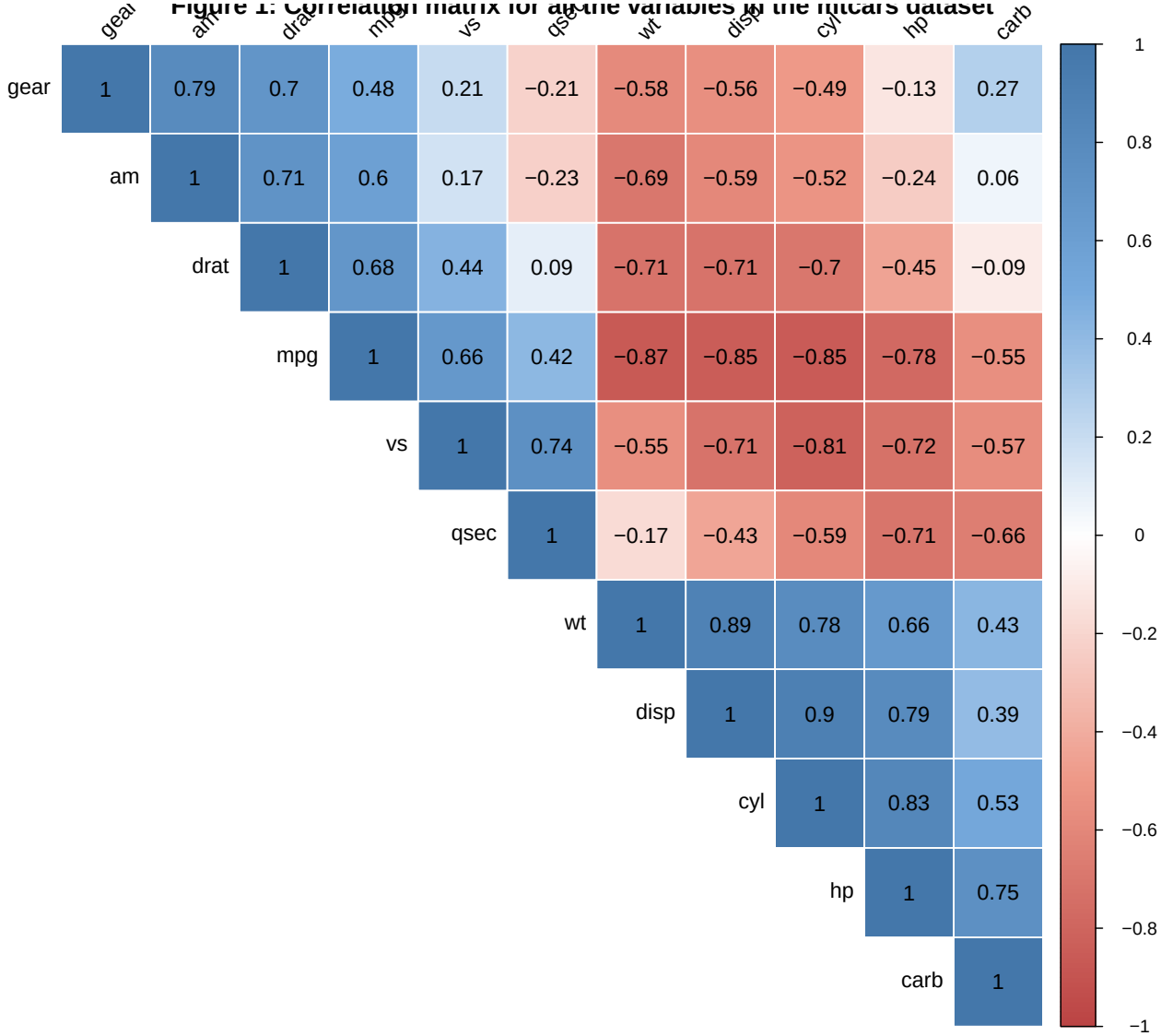

Figure 1: Correlation matrix for all the variables in the mtcars dataset