MPG Analysis

Roberto Garrote Bernal

September 2015

1 Executive summary

The objective of this report is to analyze, for a collection of cars, the relationship between a set of variables (in particular the kind of transmission, am) and miles per gallon (mpg). The main interest is to know if an automatic or manual transmission is better for mpg, and quantify the difference in terms of mpg between automatic and manual transmissions.

The conclusion is that there is a relationship between the kind of transmission and mpg. However, a linear relationship of mpg with respect to am alone cannot be established, and additional variables like weight and 1/4 mile time are needed to better model mpg.

2 Exploratory data analysis

We can observe that the average MPG for cars with manual transmission (24.39) is much higher than the average MPG for cars with automatic transmission (17.15). We must test if the difference of the means is statistically significant ($H_0 : \mu_0 - \mu_1 = 0$).

Using the method Welch Two Sample t-test we obtain a p-value 0.0013736, so the alternative hypothesis is true (means are different with a confidence over 95%) and we can expect a relationship and search for it.

Let's center the continuous variables, including mpg, to obtain more meaningful value for the intercept in the regression models.

A linear regression model between mpg and am is apparently strong as the estimated parameters for intercept and slope are very significant (p-values are 0.01376 and 2.9×10^{-4}, respectively, much lower than 0.05). However, this will be true for any factor with just two levels.

The percentage of total variability of MPG explained by the linear relationship with the predictor $R^2 = 0.36$ is low and the residuals plot (see figure 1) confirms that the model fit is not adequate.

So we need to look for other predictors of mpg. To figure out which ones, let's start by looking at the first row (mpg) in figure 4. We can observe that associated with increasing mpg are:

- Less cyclinders (cyl)
- Less displacement (disp)
- Less horsepower (hp)
- Less weight (wt)
- Less carburetors (carb)
- More rear axle ratio (drat)
- More 1/4 mile time (qsec)
- S (vs)

Number of forward gears (gear) may influence mpg in opposite directions depending if transmission (am) is automatic or manual.

There may be relationships between am and other variables, as you can observe in the column am in figure 4, and between other pairs of variables.

3 Model selection

In order to define the linear model that better explain the mpg values of the population, I have try two possibilities:

1. To start with just one predictor (for example, those numeric variables with higher correlation with mpg), and then introduce additional predictors one by one. If the coefficients are statistically significant, the standard error reduces and the adjusted R^2 increases, then keep that variable in the model.
2. To start with all the variables as predictors (excluding mpg, of course), and then eliminate one variable in each step, that one whose coefficient is the lowest statistically significant.

With the first one, the search space is higher, while with the second one, the process is deterministic.

Using the second approach, I have try the following models:

```
##  Start with all -> eliminate not fitted, one by one (higher p)
summary(lm(mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear+ carb, mtcars))
summary(lm(mpg ~       disp + hp + drat + wt + qsec + vs + am + gear+ carb, mtcars))
summary(lm(mpg ~       disp + hp + drat + wt + qsec +      am + gear+ carb, mtcars))
summary(lm(mpg ~       disp + hp + drat + wt + qsec +      am + gear      , mtcars))
summary(lm(mpg ~       disp + hp + drat + wt + qsec +      am             , mtcars))
summary(lm(mpg ~       disp + hp        + wt + qsec +      am             , mtcars))
summary(lm(mpg ~              hp        + wt + qsec +      am             , mtcars))
summary(lm(mpg ~                          wt + qsec +      am             , mtcars))
```

In the final model, am is a predictor. The intercept coefficient (-1.193) is the difference with the mpg mean (20.091), for automatic transmission when the values for the rest of the predictors are its respective means. The am1 coefficient (2.936) is the increase of mpg for an hypothetical car with manual transmission when the rest of the predictors are the same.

The residuals for this model are shown in figure 2, which are much lower than with the model $mpg = \beta_0 + \beta_1 am$ (figure 1).

The residuals have mean 0 and are aproximately normally distributed according to the QQ plot in figure 3, so we can calculate confidence intervals for intercept and am1 coefficients.

For intercept, the confidence interval is [-2.67, 0.28], which contains 0. For am1, the confidence interval is [0.05, 5.83], which does not contain 0, so we can conclude that, when the other predictors do not change, the manual transmission increases the miles per gallon around 2.936 (95% confidence).

4 Appendix (supporting figures)

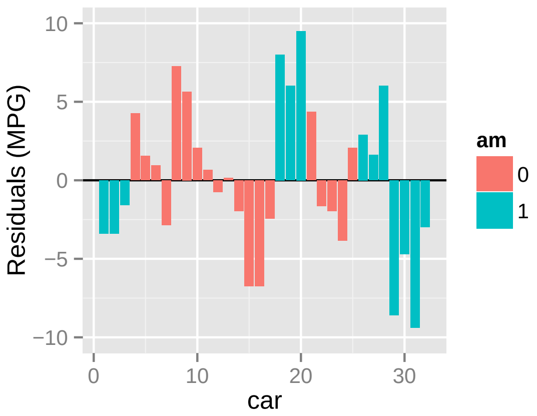

Figure 1: Residuals of linear model $mpg = \beta_0 + \beta_1 am$

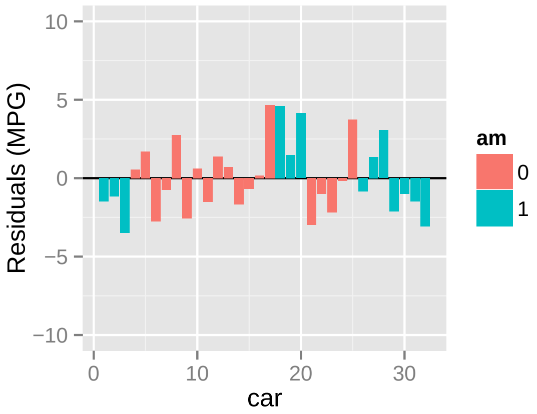

Figure 2: Residuals of linear model $mpg = \beta_0 + \beta_1 wt + \beta_2 qsec + \beta_3 am1$

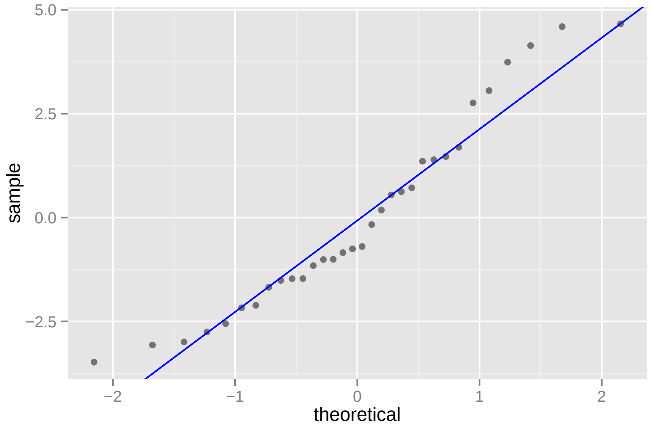

Figure 3: QQ plot of residuals of linear model $mpg = \beta_0 + \beta_1 wt + \beta_2 qsec + \beta_3 am1$

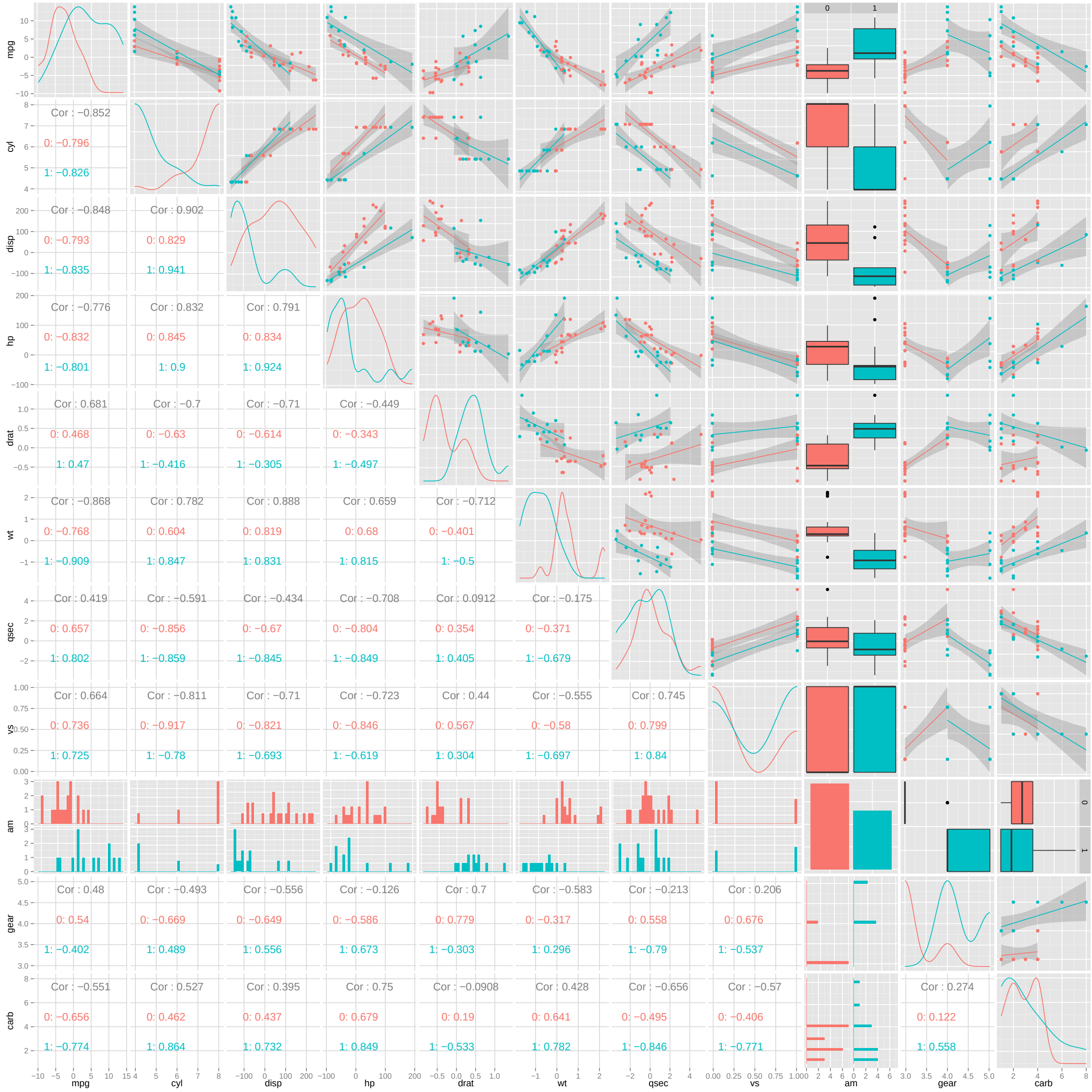

Figure 4: Exploratory data analysis tool