Is Automatic or Manual Transmission Better for MPG?

Steve Senior

21 January 2015

Executive Summary

This project tries to answer the question: "is an automatic or manual transmission better for mpg?" and to quantify the mpg difference between automatic and manual transmissions. It uses the mtcars dataset. Fitting a simple linear model with only the type of transmission (manual versus automatic) as predictor and mpg as the response suggests that there is a statistically significant relationship between type of transmission and mpg (with manual cars getting more miles per gallon). This relationship remains adjusting for weight and quarter mile time. The estimated difference is about 3 mpg (95% confidence interval 0.05 to 5.8 mpg)

Exploratory Data Analysis

To begin with it is worth looking at the distribution of mpg whether mpg is different for cars with manual and automatic transmissions. The mean mpg for manual cars is 24.3923 and the mean mpg for automatic cars is 17.1474. Plot 1 in the annex shows a boxplot of miles per gallon for manual and automatic transmission cars. Both the plot and means suggest that manual cars get more miles per gallon than automatics.

However, the mtcars dataset contains 10 other variables, many of which could be related to mpg and which could vary systematically with the type of transmission. A quick way to look at this is to do a pairs plot of the data like plot 2 in the annex. Looking at the top row only, you can see that lots of the other variables appear to correlate with mpg, and vary with the type of transmission. Fortunately, that's why we have regression models!

Model Selection

To answer the questions above, the response variable is miles per gallon (mpg). MPG is a continuous variable which is bounded from below (you can't have negative mpg, even in a humvee). However the lowest mpg in the data set is `10.4`, so the lower bound doesn't affect the data too badly. As a result, I'm only going to use simple linear regression, rather than any generalised linear model that could better model the constraints on the data. I have also not considered interaction variables: the high number of possible interactions means I'd never fit it all on to two pages.

The simplest model is to use am as a single predictor variable (plus an intercept term):

```
fit1 <- lm(mpg ~ am, data = mtcars)
summary(fit1)$coeff
```

```
##             Estimate Std. Error t value  Pr(>|t|)
## (Intercept)   17.147      1.125  15.247 1.134e-15
## ammanual       7.245      1.764   4.106 2.850e-04
```

This model gives a statistically significant relationship between am and mpg. The intercept term gives the expected mpg for an automatic transmission car (i.e. am = 'automatic' is the reference level). The estimated increase in mpg for manual cars is 7.2449. A 95% confidence interval for this increase is 3.6415, 10.8484. Note that this confidence interval does not include the null value of 0. The p-value associated with the am variable is 2.8502 × 10-4 So *if* we were only considering the am variable, we would reject the null hypothesis that there is no relationship between transmission type and conclude that manual transmissions get higher mpg than automatics. *However*, as noted above, it looks like we need to control for a number of confounding variables like weight, number of cylinders, engine size and so on.

There are a lot of different ways to approach this problem. Models can be built step-wise, either by adding variables one at a time or by starting with all the variables and eliminating them one at a time. Or they can be built based on some theory about how the varibles relate to oneanother. Since we're interested to see if the relationship between am and mpg holds up, another approach would be to create ten further models adding each other variable to see if it knocks out the effect of am. Because I don't have a lot of space, I'm going to start by throwing all the variables in, then using step-wise elimination to whittle this down (I know there are lots of criticisms of this approach). I'll then look at if the relationship between am and mpg is still significant. Note that just using all the variables results in none of them being statistically significant, because each variable increases the variance of the estimates and there are only 32 data points in this data set.

```
fit_all <- lm(mpg ~ ., data = mtcars)
fit2 <- step(fit_all)
```

```
summary(fit2)$coeff
```

```
##             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
## ammanual       2.936     1.4109   2.081 4.672e-02
```

Helpfully, am is one of the variables that survives the stepwise model fitting, so we don't need to add it back in to see if it's still significant. In this model, the expected difference between automatic and manual cars *holding the other covariates constant* is 2.9358 (the expected mpg is still higher for manual transmission cars). A 95% confidence interval is 0.0457, 5.8259. Note that the confidence interval does not contain the null value of zero and the associated p-value is 0.0467, which is less than the benchmark cut-off of 5%. Consequently, we can still reject the null hypothesis and conclude that manual transmission is better for mpg.

Residuals and Diagnostics

Plot 3 in the annex gives a series of diagnostic plots for this model. Of note the first and third panels suggest that the residuals don't vary in any obvious systematic way with the fitted values (though there may be something odd going on in the middle of the fitted values range). The second panel shows that the residuals are approximately normally distributed. Further diagnostics could be investigated by looking at measures of leverage (hat values) and measures of how much each data point is driving the coefficients (dfbetas, dffits and cooks.distance). Just looking at plot 3, the Merc 230 and Chrysler Imperial look like they could be exerting excess influence. Re-fitting the model excluding these (and possibly other data points) would be one way to check that they don't change the conclusions that I've reached here.

Conclusion

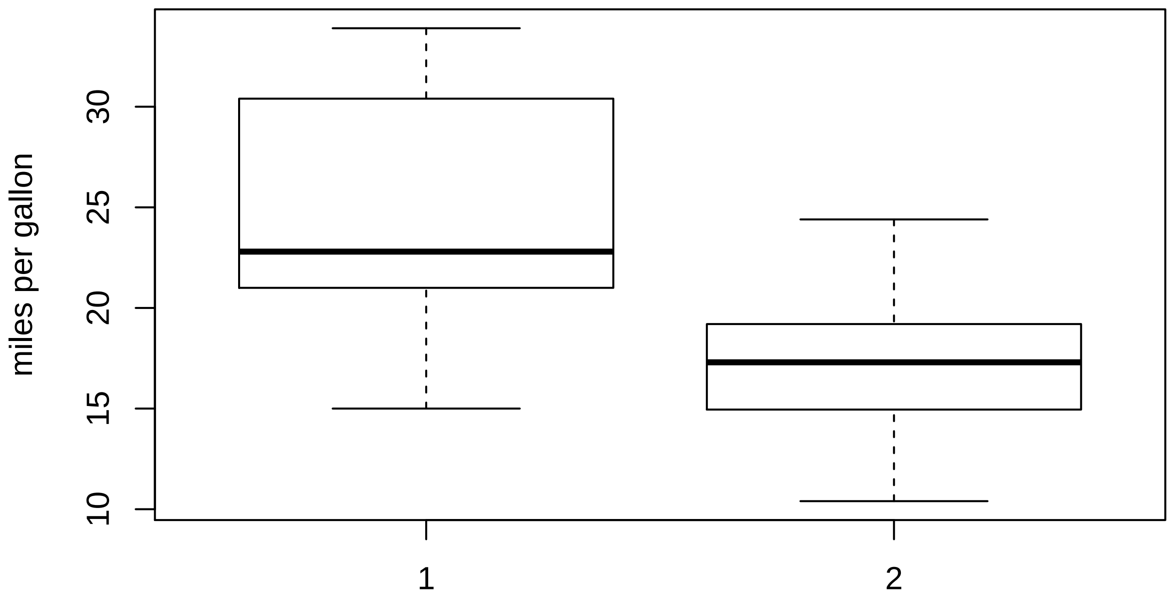
Based on the analyses here, it appears that:

- manual transmissions are better than automatics for mpg; and
- adjusting for weight and quarter mile time, the estimated difference is 2.9358 (95% CI) 0.0457, 5.8259.

However, further analysis would be necessary to see whether this effect is removed by a combination of other variables. It is also not possible to say that transmission type *causes* differences in mpg, since this is only an observational study.

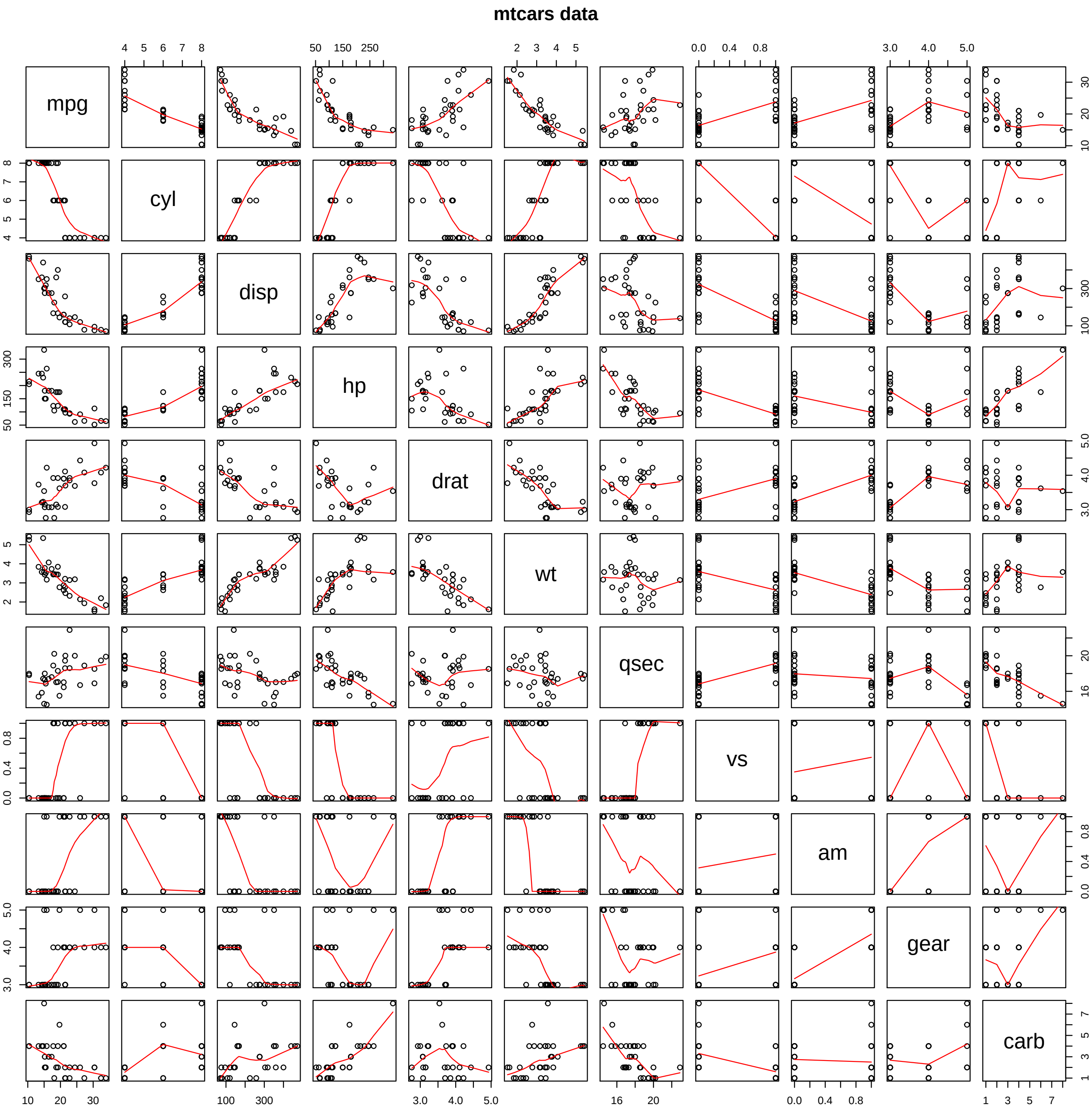
Annex: Figures

Plot 1: MPG in automatic versus manual transmission cars



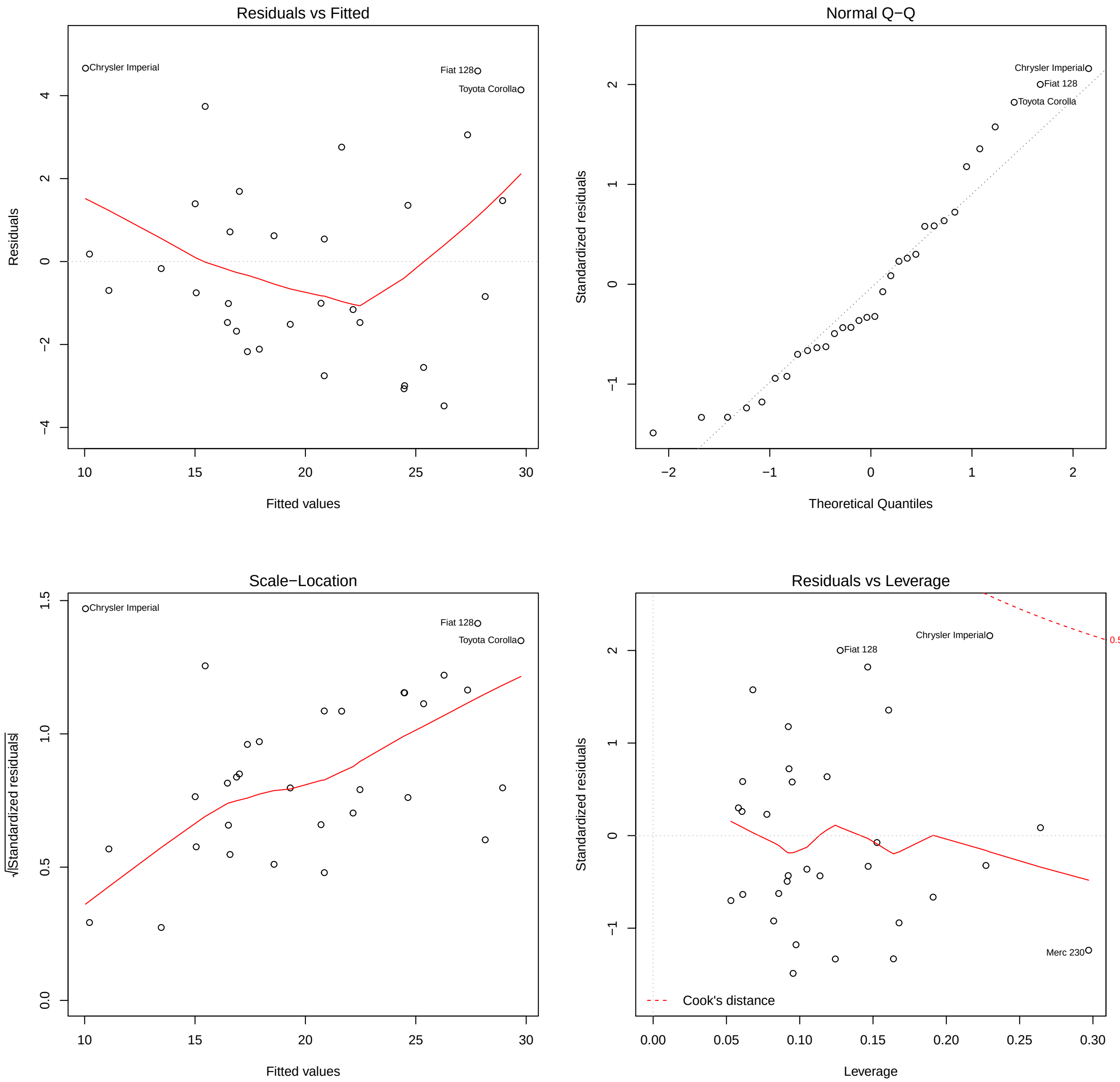
1.pdf

Plot 2: Pairs plot of variables in the mtcars dataset



2.pdf

Plot 3: Diagnostic plots for fit2



3.pdf