Analysis of MPG as a Function of Transmission Type

Executive Summary

This is the final project for course 7 in the Johns Hopkins Data Science Specialization on Coursera. I examined the mtcars data set in order to answer the following questions:

- Is an automatic or manual transmission better for MPG?
- Quantify the MPG difference between automatic and manual transmissions?

I will demonstrate in this paper that a manual transmission appears to be better for MPG.

Exploratory analysis

I began my examining a pairwise plot of mtcars (appendix, fig. 1). Looking at the mpg row of plots, it appears that there could be a linear relationship between mpg and weight. I also suspect cyl, however, it would be just as likely that cyl could be a confounder with weight, increasing in lockstep to compensate for a heavier car. It's difficult to tell how strong the relationship is between mpg and am, or transmission.

Examining this plot also led me to treat cyl, vs, and am as factor variables.

Additionally, a quick look at a linear model with mpg as the response and factor(am) as the predictor reveals the following.

```
##             Estimate Std. Error t value  Pr(>|t|)
## (Intercept)   17.147      1.125  15.247 1.134e-15
## factor(am)1    7.245      1.764   4.106 2.850e-04
```

```
## [1] 0.3385
```

Remembering that 0 indicates an automatic transmission and 1 a manual transmission, and that we have a positive slope, this preliminary model indicates that a manual transmission is, in fact, more fuel efficient than an automatic. This would seem to agree with figure 1. The p-value is very low, indicating that I can make an inference that there very little chance that the relationship is due to chance.

However, the adjusted R squared is only about .338, meaning this relationship may not be that strong in this model. There may also be confounders, which I would need to identify before coming to firm conclusions.

Creating the model

The method I used was to systematically add variables one at a time, checking to see which one had the most significant p-value, since its relationship with mpg was least likely to have occured by chance alone. So on the first iteration, I checked all 11 variables as predictors of mpg and took the most significant.

The most significant was wt, or weight. For the second iteration, I now checked lm(mpg ~ wt + xn), where n was each of the remaining 10 variables. Again, I checked each of the 10 models to see which had the most significant p-value. However, for this interation, and subsequent iterations, I not only continued to make sure my p-value was <.05, I also did an Anova test to check the same thing.

What I found was that after adding three variables, adding additional variables were no longer significant, either in terms of p-values or Anova tests. In order to keep to the two page limit, here is just the final model from each of the subsequent rounds of iterations.

```r
model1<-lm(mpg~wt,data=mtcars)
model2<-lm(mpg~wt+qsec,data=mtcars)
model3<-lm(mpg~wt+qsec+factor(am),data=mtcars)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ wt
## Model 2: mpg ~ wt + qsec
## Model 3: mpg ~ wt + qsec + factor(am)
##   Res.Df RSS Df Sum of Sq     F  Pr(>F)
## 1     30 278
## 2     29 196  1      82.9 13.70 0.00093 ***
## 3     28 169  1      26.2  4.33 0.04672 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Initially, I have reason to feel good about model3. It has passed significance tests along the way and has an adjusted R square of .834, but there are a couple of checks to make. First, the variance inflation factor. The reader will note below that the numbers for vif are all well below 5, leading me to believe that my covariates are not too closely related to each other.

```
##         wt       qsec factor(am)
##      2.483      1.364      2.541
```

Second, if one examines the residual plot (figure 2), there are no noteworthy patterns. For example, they will note that the residuals are homoscedastic, a sign of consistent variance in the residuals. Between the variance inflation factor and the residual plot, we have reason to believe this is a good model.

I did experiment with the use of an interaction term in model. However, it did not appear to improve fit dramatically. Perhaps just as importantly, as Dr. Caffo has pointed out, it is helpful to make sure a linear model is both parsimonious and interpretable, which we have with just three terms. Here is my final model.

```
##             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
## factor(am)1    2.936     1.4109   2.081 4.672e-02
```

Interpretation and Final Answers

As is show in the summary above, weight, acceleration, and, yes, transmission type, were all found to be statistically significant with alpha=.05. Therefore, we can proceed to intepret the coefficients with 95% confidence.

To answer the initial questions set out for this project:

- We still have a positive slope on the factor(am) term, making manual the better choice for mpg.
- The coefficient on factor(am) is 2.936, meaning that when holding weight and acceleration constant, on average, the MPG will improve 2.936 when going from automatic to manual transmission.

As for the other coefficients, the qsec has a coefficient of 1.226, so when holding weight constant and for a given transmission type, we can expect to see an average improvement of 1.226 MPG when the quarter mile time improves by 1 second. The weight coefficient is -3.917, so for a given acceleration and type of transmission, we can expect MPG to decrease an average of 3.917 for every 1000 lbs added to the car's weight.

Appendix

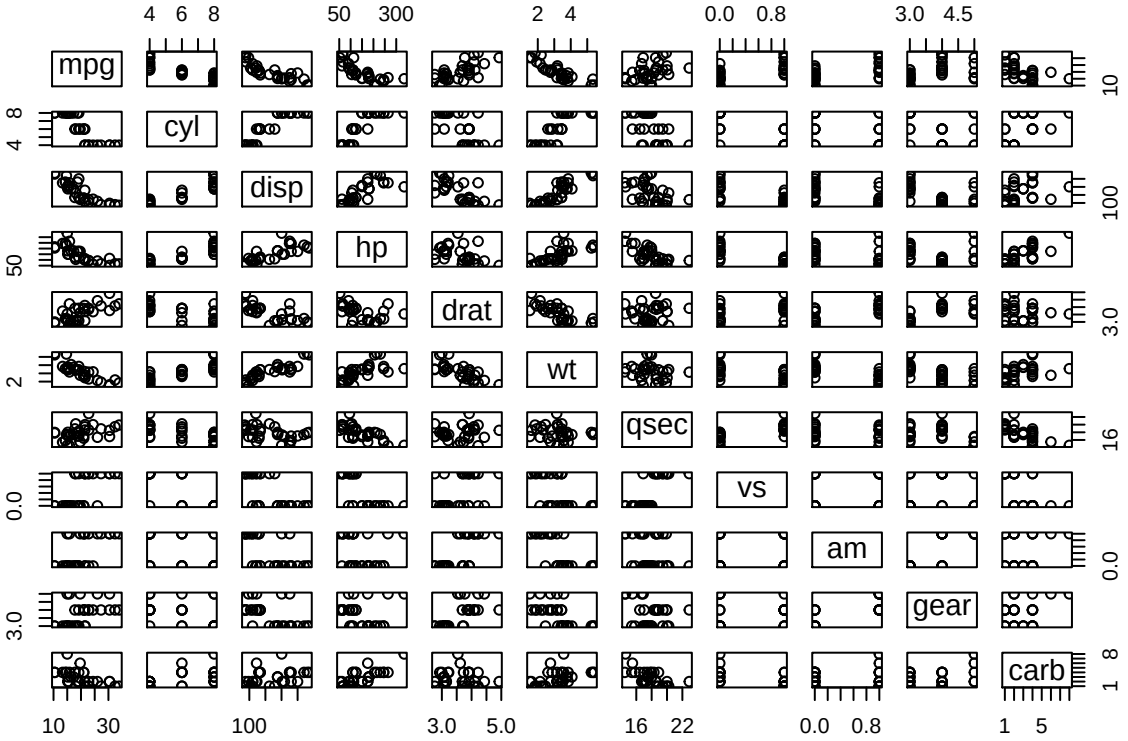

Figure 1

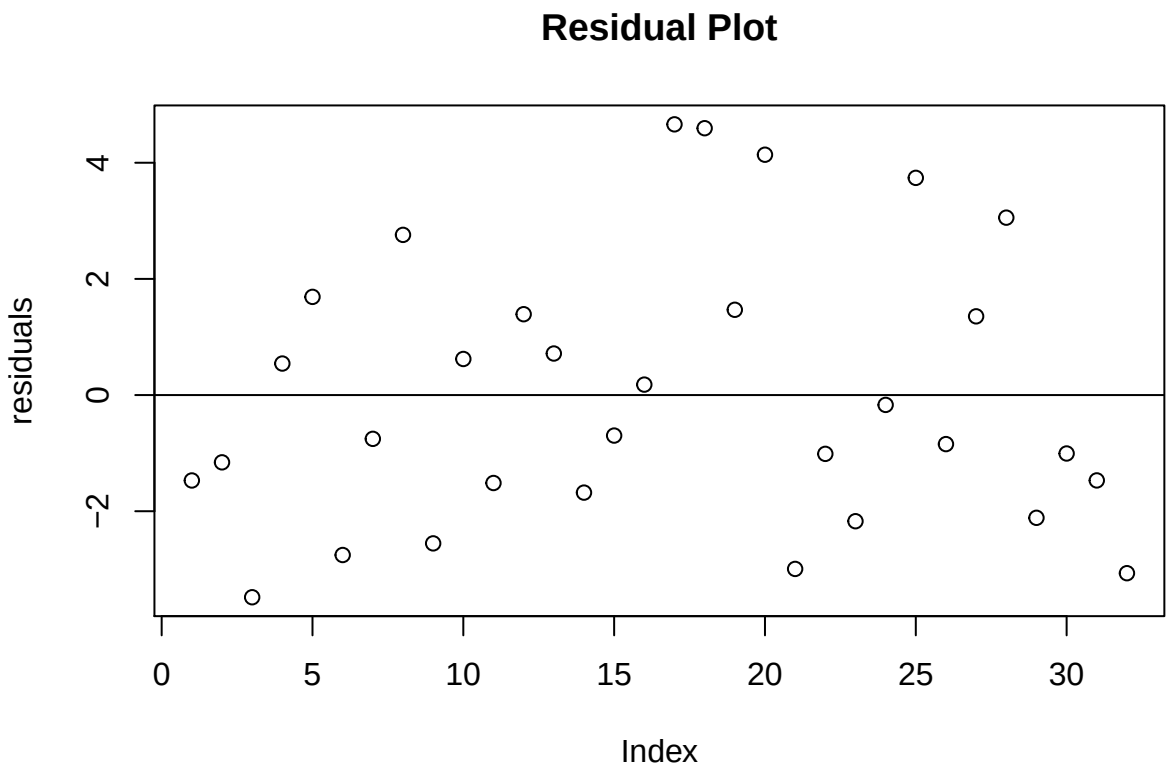

Figure 2