Report: Motor Trend Analysis of MPG Covariates by Transmission Type
Source code used for analysis available at: https://github.com/cs79/MotorTrendAnalyses (https://github.com/cs79/MotorTrendAnalyses)

Executive Summary

In this report we explore data on a variety of car models with associated parameters, and construct linear models of fuel efficiency (MPG) as it relates in various ways to transmission type (automatic or manual). Broadly, we find that manual transmission cars are more fuel-efficient than automatics by expectation of about 7 MPG on average, and through guided model construction we find that the primary feature through which expected MPG is significantly influenced by transmission type is engine displacement. Models and summary outputs are all available in the source code on GitHub (link above).

Exploration

We start with exploration to gain some basic understanding of the MPG covariates and the relationships that may exist in the data. As our questions of interest relate to potential differences in cars' MPG by car transmission type (automatic or manual), we start with these two variables. Plotting MPG against transmission type and fitting a linear model shows a significant difference between the slope and zero, which is clear in **Figure 1** in the Appendix. The slope here is interpreted as the change in expected MPG, going from an automatic transmission to a manual transmission. We find based on the model that we would expect to see about a 7 MPG fuel efficiency increase, on average, moving from an automatic to manual transmission. Knowing this, we can dig deeper into the data to look more specifically at what is driving this effect.

To help us try to find what covariates might have the most significant impact on the relationship between transmission type and MPG, we can look at a pairs plot of all of the data we have available, keeping an eye out for two things: what variables have a noticeable relationship to MPG, and in which cases is that relationship informed by transmission type? The pairs plot in **Figure 2** of the Appendix, while harder to discern in the smaller report output, indicates that displacement, rear axle ratio, quarter mile time, number of forward gears, number of carburetors, and possibly number of cylinders appear to show the type of relationship we are interested in: related to MPG, but with a discernable visual separation in the data as they relate to MPG by transmission type.

Prior to our model building, we may already be able to tell that some of these features are less interesting than others. For example, a car's quarter mile time would seem to be logically correlated with features like engine size and number of cylinders, and possibly others as well – it is more of an effect in the data than a cause. There may be other variable combinations with significant information overlaps.

Model Selection

There are two broad approaches we could take to building models, based on our exploration so far. The first approach involves focusing on the variables which most significantly affect MPG in a multivariate analysis and then modeling the interaction of transmission type with those variables to see its effect on MPG through them. The second approach would be to focus on the non-MPG variables that have the most significant interaction with the transmission variable, and then model based on those. We will use the first approach to keep our focus on the outcome of interest (MPG), with the understanding that the second may provide additional insights if performed as well.

We use a winnowing approach to determine a model with "most significant" features, based on our exploration. Our base model includes the features we identified during exploration (cyl, disp, drat, qsec, gear, and carb in the dataset). We then winnow the base model variables model down, discarding the least significant coefficients, number of cylinders and quarter mile time – unsurprisingly two that could be ruled out

either in exploration of the pairs plot or based on logical reasoning. Winnowing again through the same process, our next model (which now has two significant coefficients, number of carburetors and displacement) causes us to discard rear axle ratio, leaving us with a model that has ONLY significant coefficients (carb, disp, and gear in the dataset). We can confirm the appropriateness of our winnowing based on the subset of variables identified during exploration using ANOVA; P-values for adding back terms that got winnowed out through our process are not significant. Using the model suggested by our winnowing process with the "most significant" coefficients impacting MPG, we can include interaction terms for each covariate with transmission type. This final winnowed model for our analysis describes the top features in our dataset informing MPG, with their effects on MPG separated by transmission type. The interpretation of the model is below.

Analysis: Interpreting our Model

The coefficients for our final winnowed model describe six slopes: three estimated slopes for the effect of carb, disp, and gear for automatic transmission cars on MPG, and three additional slope change estimates, which when added to the first three yield the estimated slopes for the same effects, but for manual transmission cars. Another way to interpret this second set of slopes is as the respective expected change in slope for the three estimator variables (carb, disp, gear) per unit change in the "am"" variable (which, as "am" is a dummy variable indicating transmission type, means the difference in estimated slopes between the two transmission types). The interpretation of an individual slope coefficient for a given transmission type is the expected change in MPG per unit change in that estimator, holding the other estimators equal.

In the final winnowed model, for example, the interpretation for the carb variable estimates would be that for an automatic transmission car, holding other variables in our model constant, we would expect to see a -2.1 unit decrease in MPG per unit increase in the number of carburetors. In a manual transmission car, we add the am:disp slope estimate of 0.9 to our automatic transmission slope, yielding an expected -1.2 unit decrease in MPG per unit increase in the number of carburetors. The other slope coefficients are interpreted in the same manner.

Interestingly, we note that in our winnowed model with interactions, the only significant difference slope coefficients when considering a split on transmission type is displacement (which, on its own, is a hugely significant predictor of MPG). Fitting a simple model with just one interaction term, disp*am, results in a model with all highly significant coefficients – it may be enough to model MPG in this manner for our purposes. **Figure 3** in the Appendix highlights the relationship between MPG, displacement, and transmission type succinctly.

Conclusions and Ex-post Diagnostics

To answer our questions, our analysis shows that a manual transmission is, on average, better for MPG efficiency; we quantified this earlier using the simple linear model relating just these two variables and saw that the MPG increase switching from automatic to manual transmission is about 7.2. We can also quantify the certainty of our estimate using the Standard Error from our model; we are 95% certain that the true value of the difference is between 4.3 and 10.2 for these data. We can calculate quantities for our final winnowed model and the simplified interaction model similarly, using the interpretation of coefficients from the section above.

Finally, we can run a quick diagnostic on our models using residual plots. These can be seen in **Figure 4**. All of the plots are appropriately random, indicating that the model variables are not being confounded.

Appendix

Figure 1: Basic Model Plot: MPG by Transmission Type

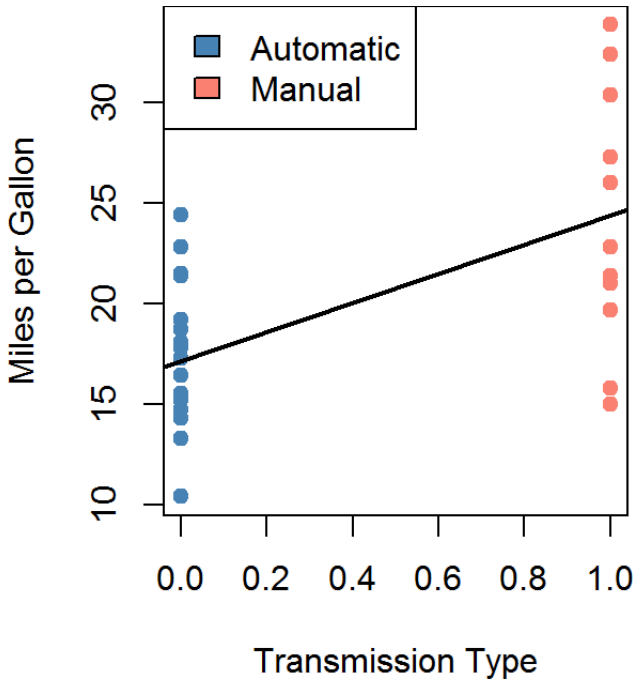

Figure 2: Exploratory Pairs Plot

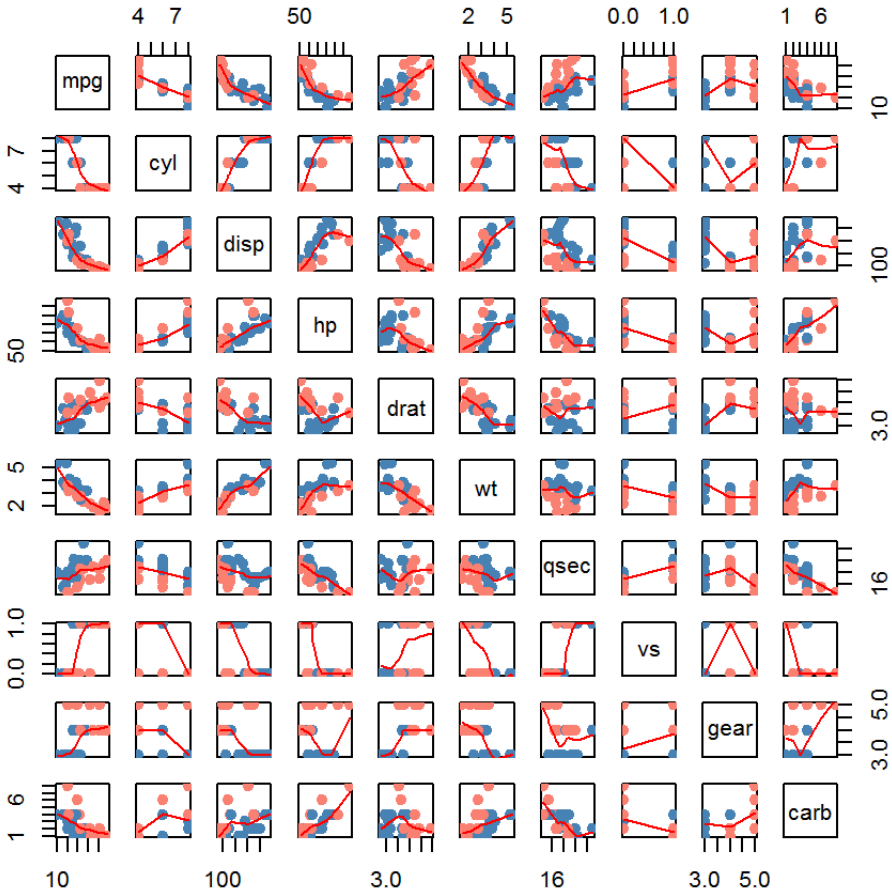

Figure 3: MPG by Engine Displacement and Transmission Type

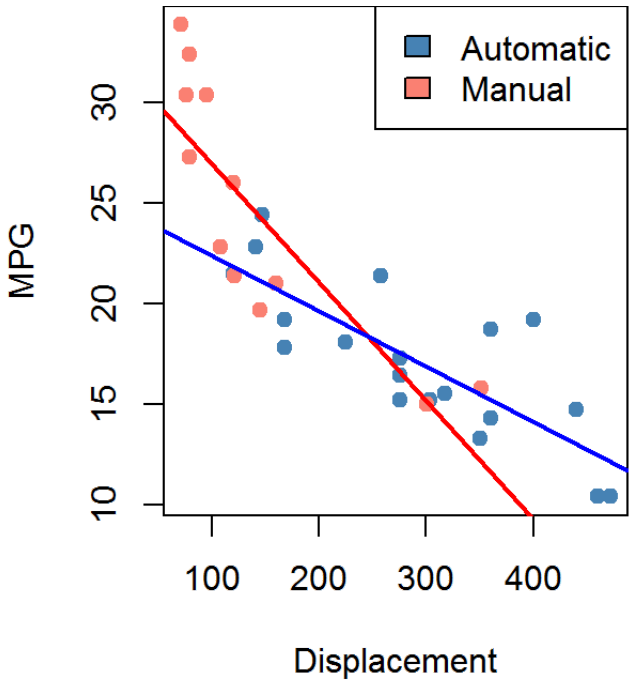

Figure 4: Residual Plots