Regression Models Course Project

Daniel Alaiev

October 19, 2015

Executive Summary

This analysis aims to answer two questions for *Motor Trend* magazine. The data used for this analysis are from the *mtcars* data set. The data was extracted from the *1974 Motor Trend* US magazine. It contains fuel consumption and 10 automobile design and performance aspects for 32 automobiles (1973/74 models). The questions and their results are briefly described below.

Is an automatic or manual transmission better for MPG?

It **depends** on the weight of the car. Lighter cars benefit from having manual transmissions while heavier cars don't, holding *1/4 Mile Time* constant.

Quantify the MPG difference between automatic and manual transmissions.

When the transmission is **Automatic** the *Miles/Gallon* intercept is 9.72 and a one unit increase in *Weight (lb/1000)* results in a -2.94 decrease in *Miles/Gallon*, holding *1/4 Mile Time* constant.

When the transmission is **Manual** the *Miles/Gallon* intercept gets a bump to 23.8 but the rate of change in *Miles/Gallon* based on a unit increase in *Weight (lb/1000)* gets slashed to a total of -7.08, holding *1/4 Mile Time* constant.

Main Analysis

A pairs plot[1] that contains densities and correlations reminds us of a few important points. The highly variable distributions and small sample sizes may distort the accuracy of any model. Also, since the sample was not selected using a randomized process this means that the sample is biased. This also affects the accuracy of any model.

The **first** and simplest regression is an OLS model with *Miles/Gallon* as the dependent and *Trans Type* as the independent variable. This is done after renaming the columns and binary variables for better readability. It appears that on average, a *manual transmission* will yield 24.39 *Miles/Gallon* and an *automatic transmission* will yield 17.15 *Miles/Gallon*. Even though the errors appear to be normally distributed[2], this is a classic case of Simpson's Paradox. There are many other linear variables that affect *Miles/Gallon* that are offsetting each other in the residuals. With the correlations from the pairs plot as well as some common sense, it is easy to see that there are more covariates than just *Trans Type*.

The **second model** is the most computatationally intensive model. It uses all of the variables to explain *Miles/Gallon*. This results in *none* of the coefficients being significant at the 0.05 level. The average probability of the hypothesis that the different coefficients have a population mean of zero, or no effect on *Miles/Gallon* under this model, is 0.53. The regression coefficient for *Weight lb/1000* has an estimated probability of 0.06 that it comes from a distribution with a true mean of zero. Although this looks better than the average, this is still quite high. The highest probability from the regression model belongs to *Number of Cylinders*. The population coefficient has an estimated 0.92 probability of being from a distribution that is centered around zero. These results *do not* indicate

that all of the covariates don't matter. They are the result of variance inflation that is due to the high collinearity of the predictors. This underlines the fact that not all of the data are useful in predicting *Miles/Gallon* and that a logically sound model should be built by considering these findings.

There are many approaches to model selection and for the sake of simplicity the **final model** will be constructed using some simple judgment. A correlation table[3] will be used to avoid variance inflation due to collinearity. Some basic knowledge about cars will also be used to think about possible interactions amongst variables. Below is a list of the variables that have been selected for the final model and the rationale behind the choices.

- *Trans Type* must be included in the regression since it is part of the question.
- The *Gross Horsepower*, *Number of Carburetors*, and *Engine Type* are highly correlated to the *1/4 Mile Time*. Time to cover a distance in a car is a function of the three and therefore it will be used as a more complete predictor. From a mechanical standpoint, *Displacement* and *Rear Axel Ratio* also contribute to *1/4 Mile Time*. Including it in the regression lets us exclude the others along with their high sampling variability and potential biases.
- *Weight* has the strongest linear relationship with *Miles/Gallon*, the correlation coefficient is -0.87. From a logical standpoint, there is good reason to believe that the effect of *Weight* on *Miles/Gallon* changes with the *Trans Type*. Crude 1970s manual transmissions perform poorly in stop-and-go traffic for heavy cars. The clutches tend to be heavy which tends to disincentivize the driver from shifting on time. This happens at the expense of mileage. Lighter cars tend to have lighter clutches and easier manual gearboxes, which leads to more timely shifting and better mileage.

The **final model**[5] output is reproduced and summarized below.

```
##                                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)                           9.7231     5.8990  1.6482   0.1109
## `1/4 Mile Time`                       1.0170     0.2520  4.0354   0.0004
## `Weight (lb/1000)`                   -2.9365     0.6660 -4.4090   0.0001
## `Trans Type`Manual                   14.0794     3.4353  4.0985   0.0003
## `Weight (lb/1000)`:`Trans Type`Manual  -4.1414     1.1968 -3.4603   0.0018
```

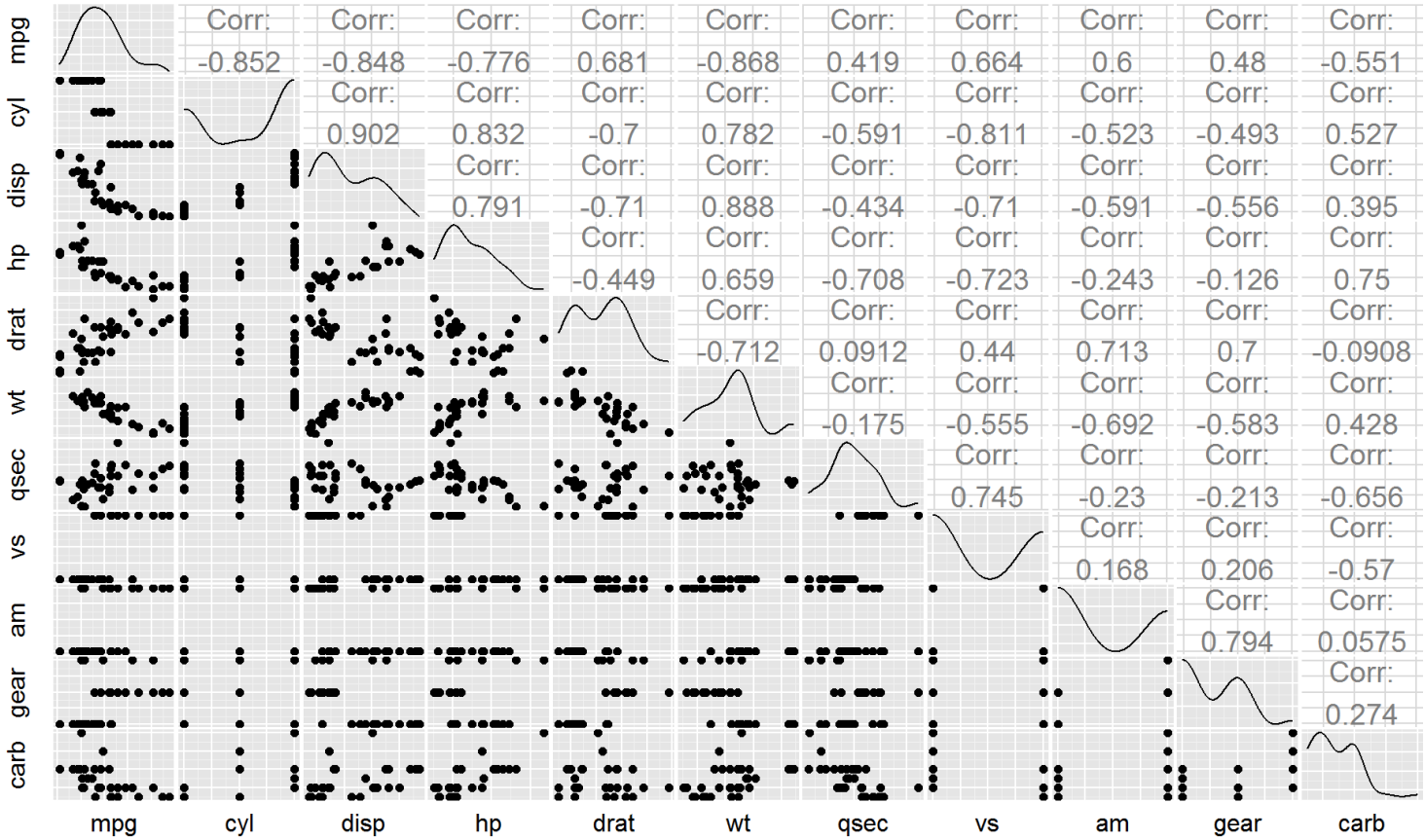
Both the slope and intercept change in this model depending on the *Transmission Type*. When the transmission is **Automatic** the intecept is 9.72, a one unit increase in the *1/4 Mile Time* results in a 1.02 increase in *Miles/Gallon*, and a one unit increase in *Weight (lb/1000)* results in a -2.94 decrease in *Miles/Gallon*. When the transmission is **Manual** the intercept is 23.8, a one unit increase in the *1/4 Mile Time* results in a 1.02 increase in *Miles/Gallon*, and a one unit increase in *Weight (lb/1000)* results in a -7.08 decrease in *Miles/Gallon*. All of the estimates are statistically significantly different than zero, except for the intercept.

It should be mentioned that the residual distribution doesn't appear to be normal[4]. The model's intercept is also not very significant with a probability of 0.11 that it comes from a distribution with a population mean of zero. The intercept's meaning doesn't have much weight in this model because none of the predictors can be zero by construction. Again, the small and highly variable sample can also distort the significance and meaning of the coefficients. The bias in the sample can also artificially inflate the adjusted coefficient of determination, 0.88. Even after being adjusted for the number of covariate parameters, this seems a bit too high. There is likely a mechanistic and nonlinear relationship between some of the variables but this is beyond the scope of this analysis. Future analyses could also adjust for the high variability of the samples and model the overall sampling bias.

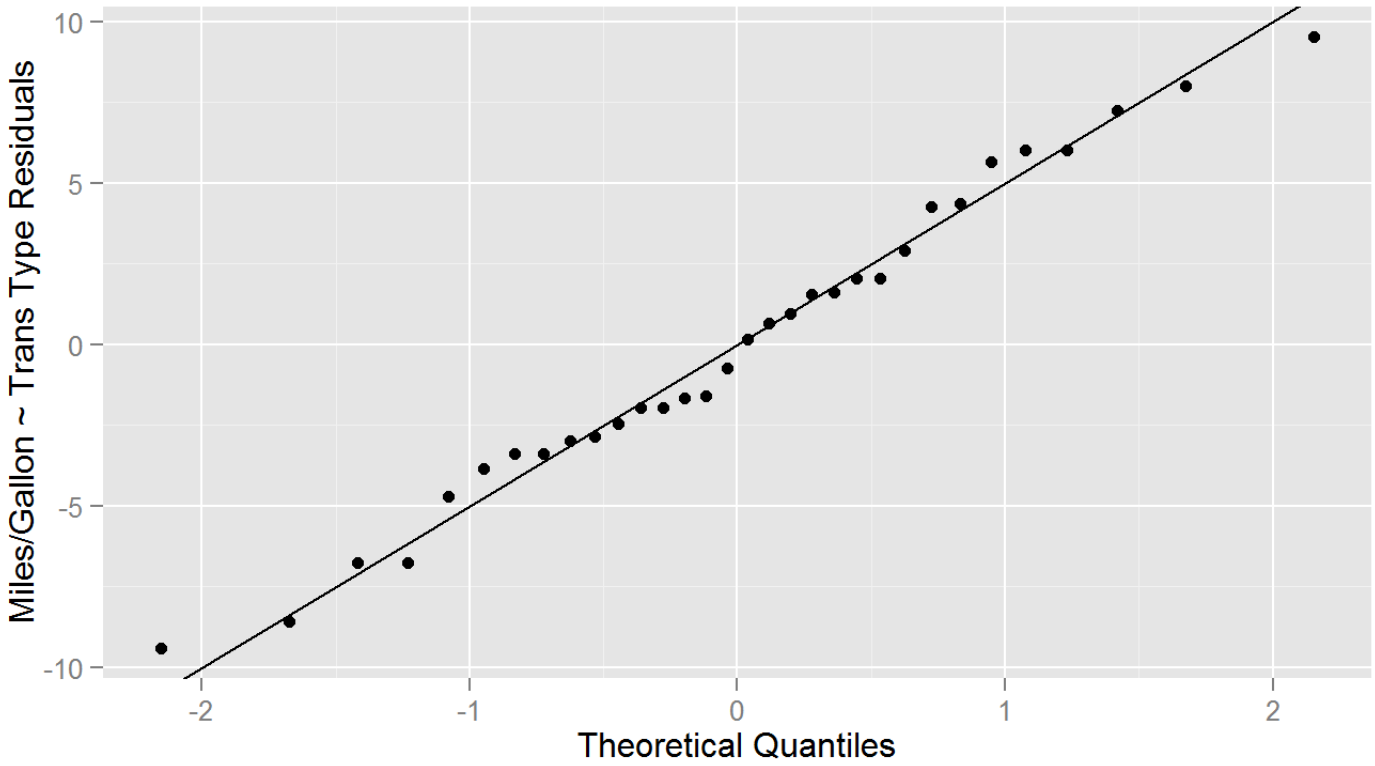
The code for this analysis is in the .RMD file.

Appendix

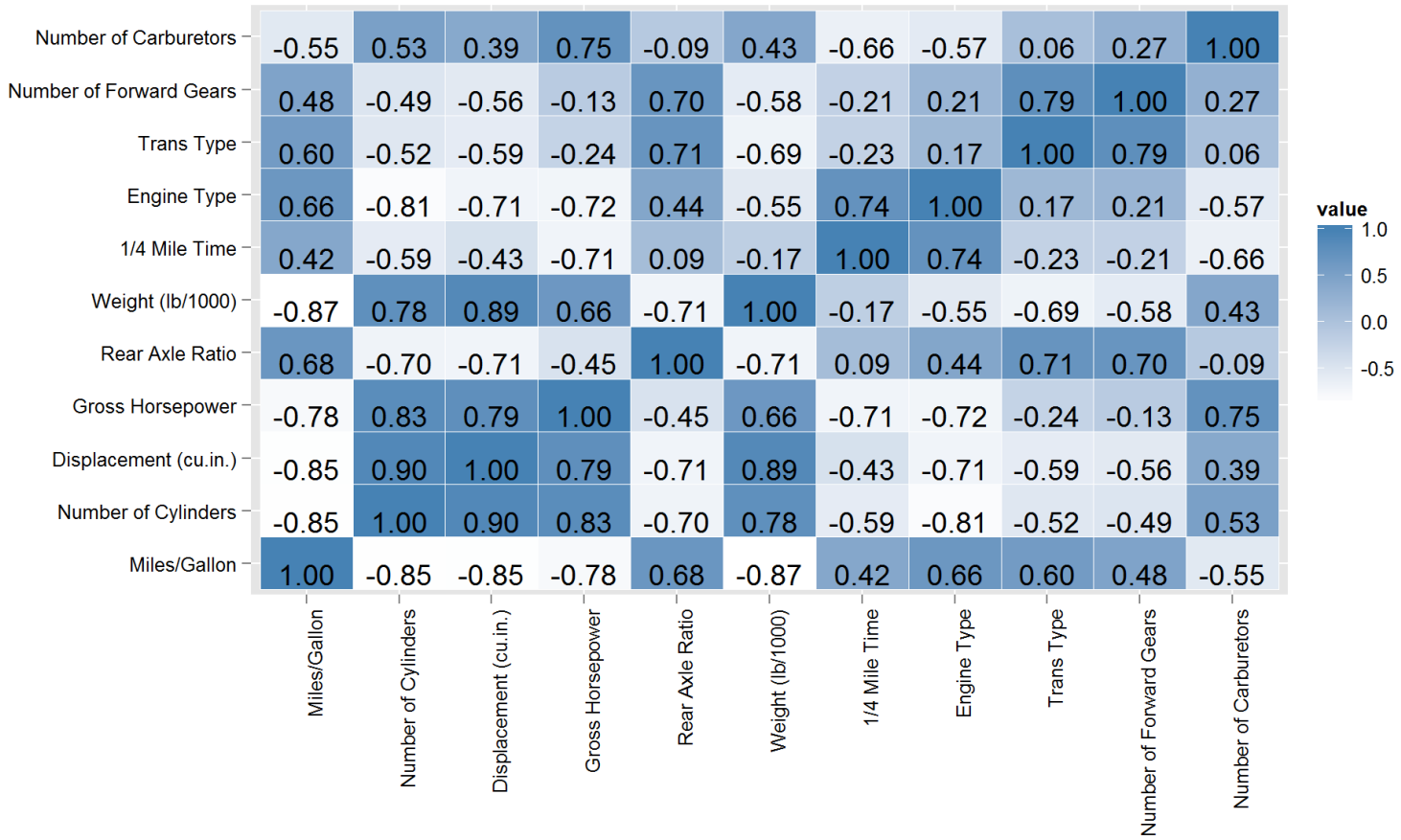
Pairs Plot[1]



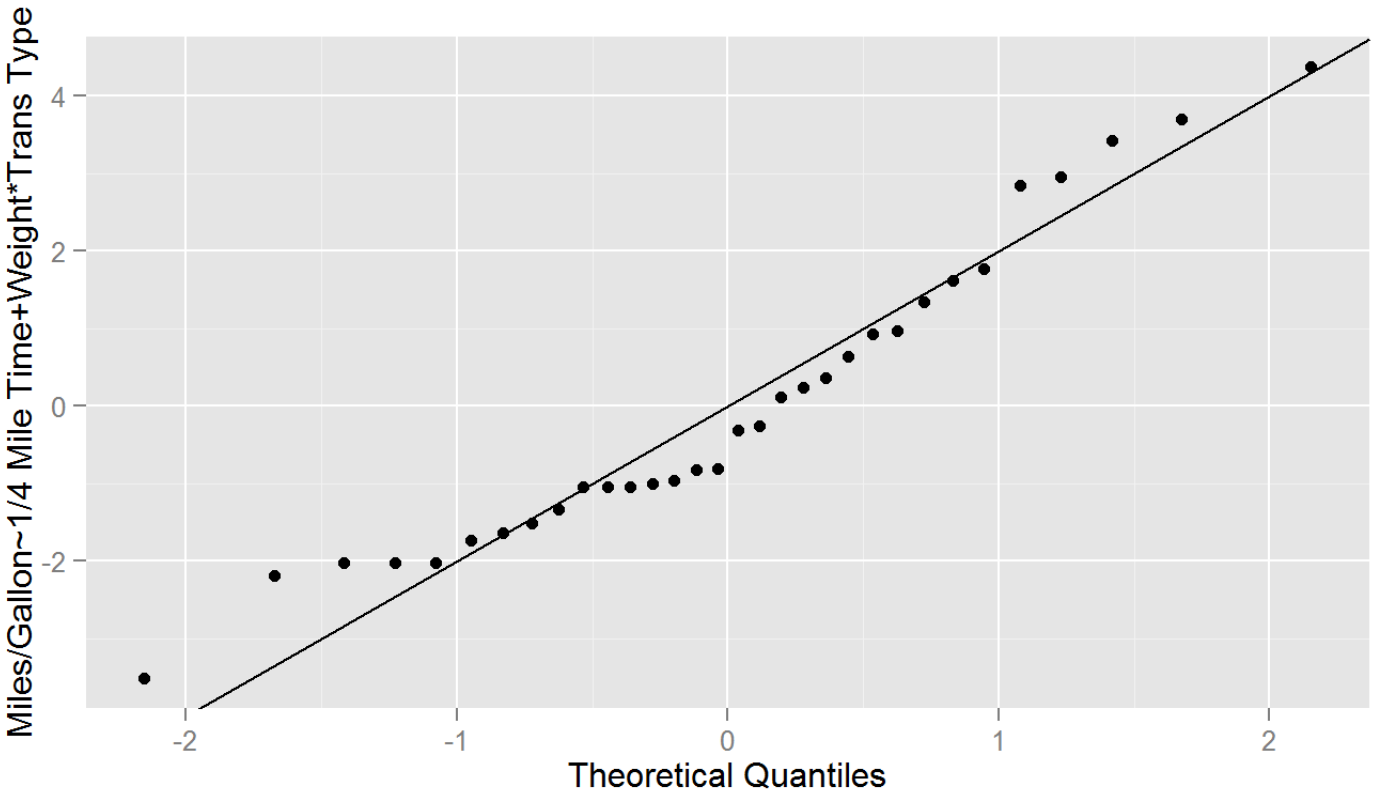
Simple OLS Residuals[2]



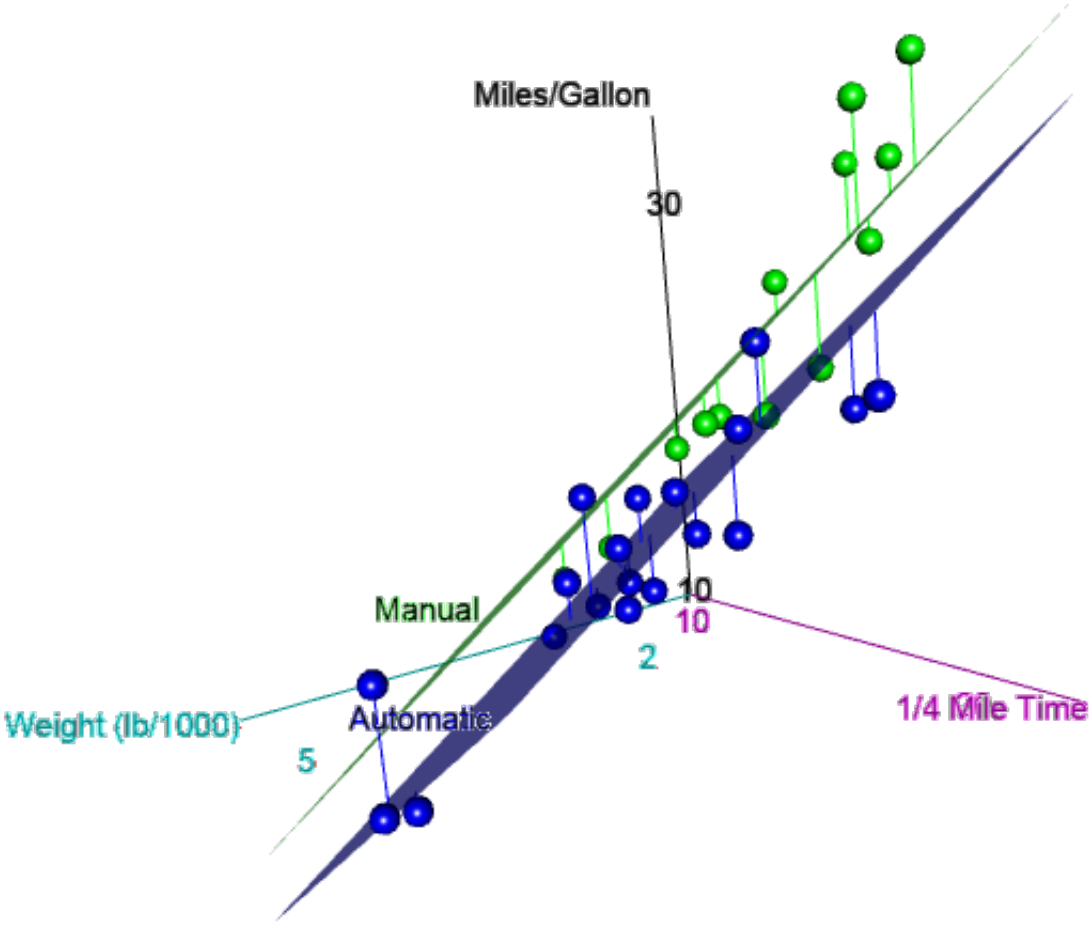
Full Correlation Matrix[3]



Final Model Residuals[4]



Final Model Plot[5]


Miles/Gallon
30
10
Manual
Automatic
10
2
1/4 Mile Time
Weight (lb/1000)
5