MotorTrend MPG Regression Analysis

Wally Thornton

September 27, 2015

Executive Summary: In this project, we are asked by the editor of MotorTrend magazine to answer two questions: 1. Is an automatic or manual transmission better for mpg, and 2. Quantify the mpg difference between automatic and manual transmissions. As will be shown in this knitr document, the simple answers are manual and 7.24 mpg. But we'll see that the predictive value of transmission type is fairly low and therefore run three additional models: 1. the highest-correlating single predictor, 2. beginning with all predictors and iteratively reducing non-significant factors, and 3. our theory that horsepower and weight are the key predictors. We will show that the more complex (but more accurate) answer to the questions is that transmission type is a poor predictor of mpg. The magazine would better serve its readers with an article about more weight and horsepower leading to lower mpg values.

EDA: The `mtcars` dataset has 32 cars of myriad makes and models, with a mean mpg of 20.09 and standard deviation of mpg as 6.03, so there's quite a bit of variance in the values (a histogram of the values for mpg is in the Appendix, Figure 1).

The `mtcars` description file defines `mpg` as miles/US gallon and `am` as the transmission type, with 0 signifying automatic and 1 manual. Since we are examining the influence (or lack thereof) of the transmission type on miles per gallon, let's look at the breakdown of the two types in the data set:

```
## 0=Auto, 1=Manual
##  0  1
## 19 13
```

So 59.4% of the cars have automatic transmissions. A density plot in the Appendix (Figure 2) shows that there is a noticeable difference in mpg by transmission type, which the correlation between the two variables confirms: 0.6. While this correlation is not strong enough to convince us, the magazine editor believes transmission type is causal, so we run our first model with a linear regression, using `am` as the predictor and `mpg` as the outcome, resulting in:

```
##              Estimate Std. Error   t value                 Pr(>|t|)
## (Intercept) 17.147368   1.124603 15.247492 0.00000000000000001133983
## am           7.244939   1.764422  4.106127 0.00028502074393506769
```

Interpretation of Coefficients: β_0 (or the y intercept) is 17.15 while β_1 (or the slope of x) is 7.24. Since automatic transmissions in this dataset are set to the value 0 in `am`, β_0 is the mean mpg for an automatic transmission and β_1 is the predicted gain in MPG for the manual transmission cars.

With the p-value of the slope well below our pre-selected α of 0.05, we could conclude that the transmission type does affect mpg, with the move from automatic to manual adding 7.24 miles per gallon, all else held constant. (A scatterplot with fitted regression line is shown in the Appendix, Figure 3.) However, R^2 is 0.34, which means that this model only explains 34% of the variance of the data, with the remainder due to other variables. Relatedly, the standard error of this model is 4.9, which is quite high for this dataset. We have confounders to find.

With multiple potentially important variables in the dataset (and with many perhaps derivatives of others, such as `qsec`), we run a correlation matrix first to narrow down our choices (in Appendix, Figure 4). The results show that while `am` has a very loose correlation with `mpg` (0.6), other variables are much more correlated (e.g., `wt`: -0.87). We also see that some predictors exhibit high colinearity (and are logically connected, like displacement and number of cylinders).

Model Strategy: We want to be parsimonious with our model while returning the best-performing model, so we'll pursue three paths. First, a couple single-predictor models using variables with the highest correlation to `mpg`. Second, running a model with all predictors and iteratively removing the variables with highest p-values. And third, our hypothesis that `wt` and `hp` are the best predictors (explained below). The winning model will be that with the highest adjusted R^2, balanced with a low standard error.

The higheset correlation to `mpg` is `wt`, followed by `cyl`.

```
##              Estimate Std. Error   t value                     Pr(>|t|)
## (Intercept) 37.285126   1.877627 19.857575 0.00000000000000000008241799
## wt          -5.344472   0.559101 -9.559044 0.00000000012939587013505 30
```

```
##             Estimate Std. Error   t value                    Pr(>|t|)
## (Intercept) 37.88458  2.0738436 18.267808 0.00000000000000000008369155
## cyl         -2.87579  0.3224089 -8.919699 0.00000000061126871425809 8
```

Weight, by itself, is a much better predictor than `am`, with `mpg` declining 5.3 for every 1,000 pounds increase in weight. With an adjusted R^2 of 0.745, this leaves 25.5% of the variance unexplained, which is much better than using `am` alone, and the RSE has dropped dramatically, to 3.05. (Not surprisingly, results for `cyl` are slightly worse than for `wt`).

When we **plot the residuals for `wt` as the predictor (shown in Appendix, Figure 5)**, we see exactly what we want to see: broad dispersion, balanced distributed on both sides of the regression line and no discernible patterns. Although we'd like to see the residuals closer to the regression line (meaning a better fit), this plot shows our best fit line for `wt` is about in the middle of the data points, there is little to no bias in our data and there is no sign of heteroscedasticity. We conclude that a linear model is appropriate for this data.

Our second approach is to run a linear regression with all the variables and then stepwise remove the least significant until we're left with only p-values < 0.05:

```
##                Estimate Std. Error   t value                      Pr(>|t|)
## (Intercept) 37.22727012 1.59878754 23.284689 0.00000000000000000002565459
## wt          -3.87783074 0.63273349 -6.128695 0.00000111964713620004645522
## hp          -0.03177295 0.00902971 -3.518712 0.00145122853156942643350347
```

This strategy leaves us with `wt` and `hp`, with much improved values of 0.81 for R^2 and 2.59 for standard error.

And this leads to our third and final model. At its most fundamental, driving a car is the acceleration and deceleration of a mass using an engine of some sort as the primary means of propulsion. Miles per gallon is one measure of the efficiency of these actions, and the mass and propulsive aspects of a vehicle are best represented in this dataset by weight and horsepower, respectively. (Torque is an even more important factor in acceleration, but the data do not include these values.) Horsepower (`hp`) and weight (`wt`) were also the survivors of our last model's defactorization, and our theory is that the interaction between these two will do an even better job of predicting mpg:

```
##                Estimate Std. Error   t value                Pr(>|t|)
## (Intercept) 49.80842343 3.60515580 13.815887 0.00000000000005005761
## wt          -8.21662430 1.26970814 -6.471270 0.00000051992872795832
## hp          -0.12010209 0.02469835 -4.862758 0.00004036243020675190
## wt:hp        0.02784815 0.00741958  3.753332 0.00081083073737062529
```

Conclusion: The adjusted R^2 of this last model is 0.87, which is pretty good for this dataset and has a residual standard error of only 2.15. These values are far superior to the other models above in predicting mpg, with only 13% of the variance left unexplained (Figure 6 in the Appendix shows that the linear model is appropriate, the data is fairly normal and there is no sign of heteroscedasticity). Therefore, we can tell our editor that weight, horsepower and the interaction between the two best predict miles per gallon. But we will caution him that there are still fairly significant factors absent from our model, **amounting to 13% of the variance of the residuals, which is the uncertainty in our model**, and that further work is necessary to find these confounders.

Appendix

Fig 1: Count of MPG Values of Cars

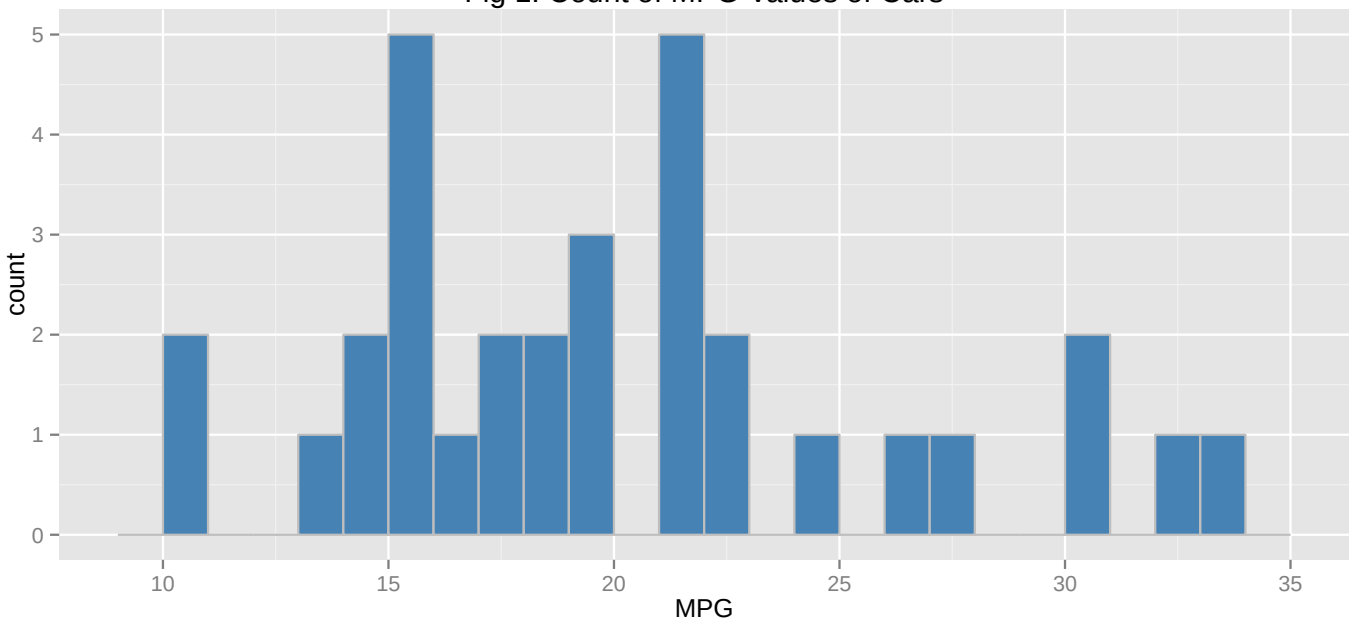

Fig 2: Densities of MPG Values by Transmission Type

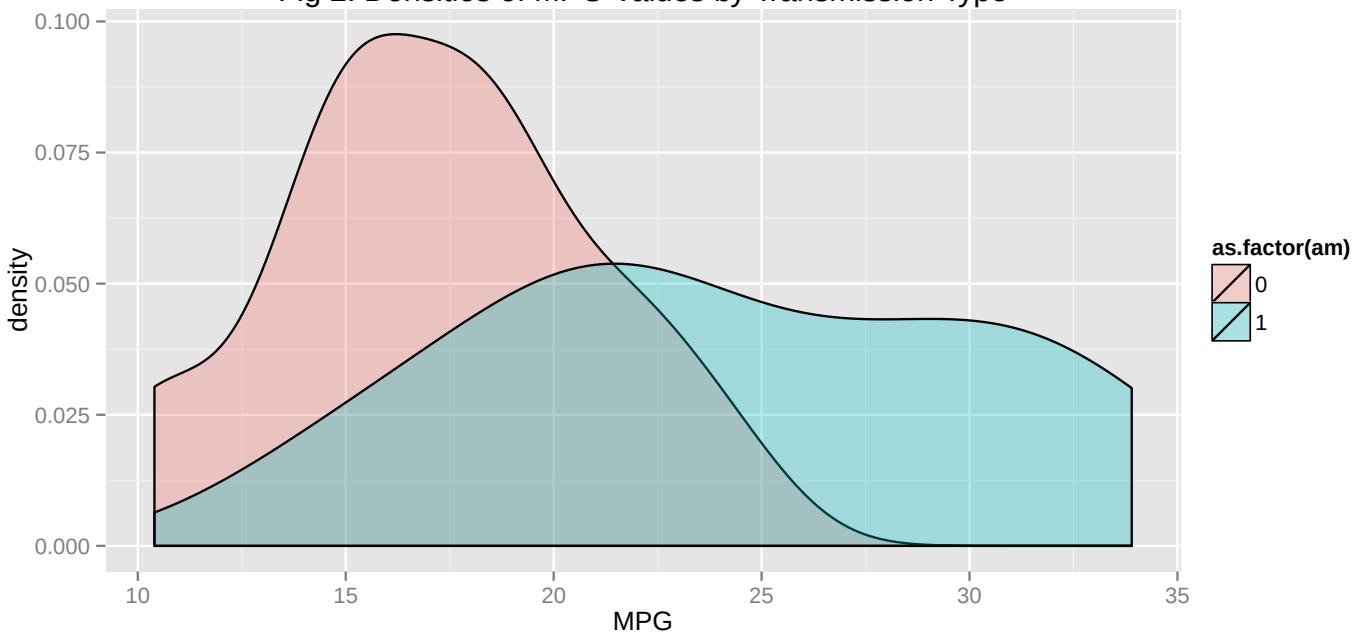

Fig 3: Linear Regression Line for MPG Predicted by Transmission Type

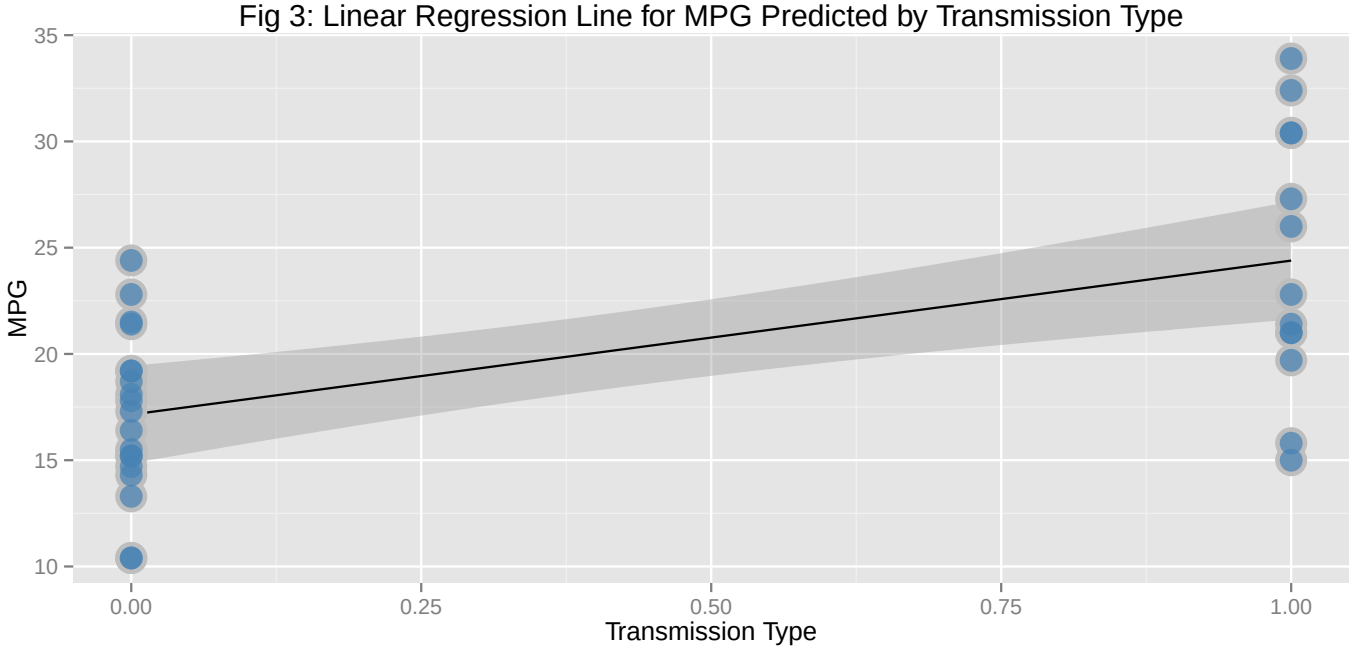

Fig 4: Correlation Matrix of All Factors

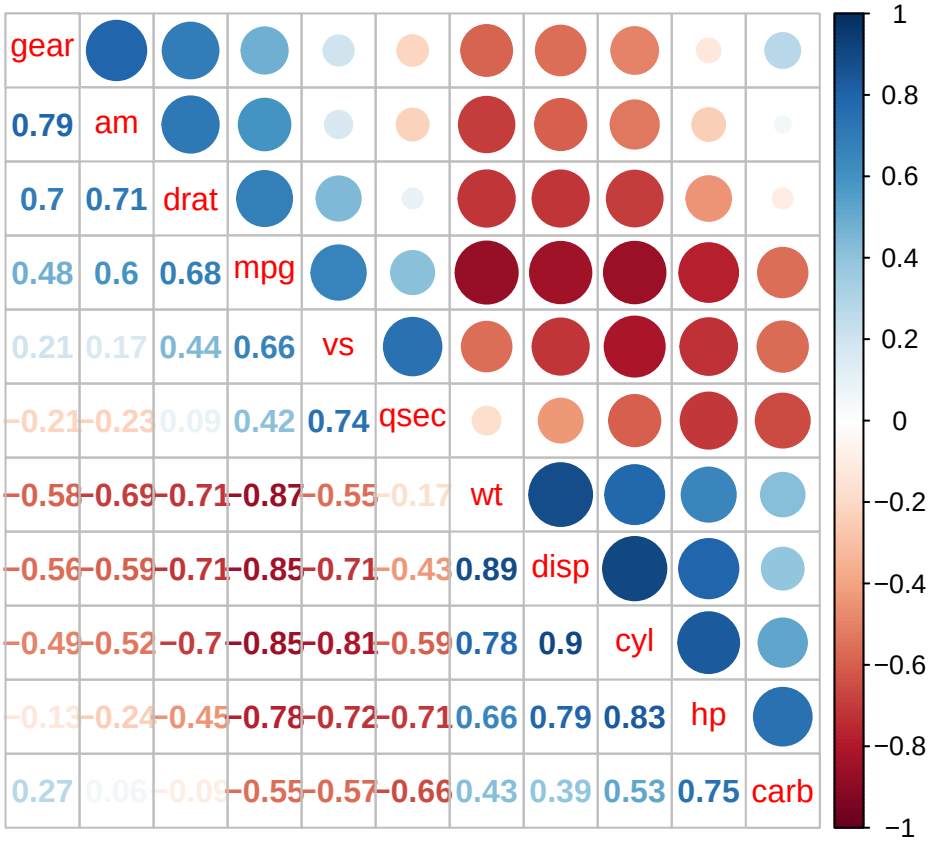

Fig 5: Plot of mpg~wt Residuals

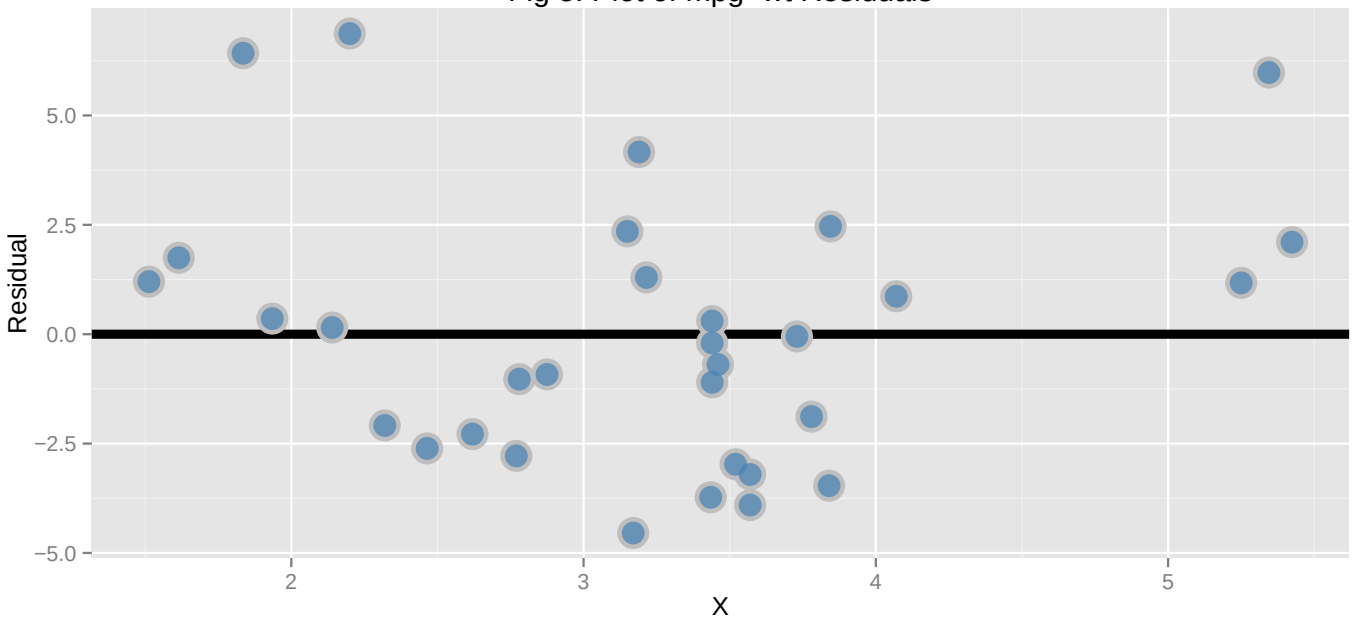

Fig 6: Plots Diagnosing Goodness of Final Model