MotorTrend MPG Regression Analysis

Wally Thornton

September 27, 2015

MotorTrend MPG Regression Analysis

Using regression models, this analysis seeks 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, an automatic is better for MPG, with the difference being: .

The `mtcars` dataset has 32 cars of myriad makes and models, with a mean mpg of 20.090625 and standard deviation of mpg as 6.0269481, 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  1
## 19 13
```

So 59.4% of the cars have automatic transmissions. A density plot (Figure 2) in the Appendix shows that there is a marked difference in mpg by transmission type, which the correlation between the two variables confirms: 0.5998324. Nevertheless, the magazine editor believes transmission type affects mpg 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 0.000000000000001133983
## am           7.244939   1.764422  4.106127 0.000285020743935067769
```

β_0 (or the y intercept) is 17.15 while β_1 (or the slope of x) is 7.24. Since automatic transmission in this dataset is 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 increasing mpg by 7.24. (A scatterplot with fitted regression line is shown in Appendix, Figure 3.) However, R^2 is 0.34, which means that this model only explains 34% of the variance of the data. This is evident when we plot the residuals (shown in Appendix, Figure 4), and see the broad dispersion. Relatedly, RSE 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 5). The results show that while `am` has a very loose correlation with `mpg` (0.5998324, other variables are much more correlated (e.g., `wt`: -0.8676594) so we will focus on those. We also see that some variables are highly correlated to each other (and logically connected, like displacement and number of cylinders) so we eliminate those and run a multiple regression model with those variables that are potentially predictive. We also want to be parsimonious with our model so we'll first look at just `wt` as a predictor and then run the model with all predictors, followed by systematically removing those with high p-values, with an eye toward increasing R^2 while keeping RSE as low as possible.

TODO: ADD SOME LOGIC AS TO WHY I'M CHOOSING CERTAIN VARIABLES

```
##              Estimate Std. Error   t value                  Pr(>|t|)
## (Intercept) 37.285126   1.877627 19.857575 0.00000000000000000008241799
## wt          -5.344472   0.559101 -9.559044 0.00000000001293958701350530
```

```
## NULL
```

```
##                 Estimate Std. Error    t value            Pr(>|t|)
## (Intercept) 34.209443370 2.82282610 12.1188632 0.000000000001979953
## wt          -3.046747000 1.15711931 -2.6330448 0.013829362001523989
## disp         0.002489354 0.01037681  0.2398959 0.812222918884354717
## hp          -0.039323213 0.01243358 -3.1626624 0.003842032133278817
## am           2.159270737 1.43517565  1.5045341 0.144053078426479519
```

Weight, by itself, is a much better predictor than `am` alone, with `mpg` declining 5.3 for every With an adjusted R^2 of 0.74, this leaves 26% of the variance unexplained, which is much better than using `am` alone, but not as good as the model with additional predictors with an R^2 of 0.71 and RSE of only 3.05. But even this can be improved, since hp is a function of displacement, cylinders, gearing and other confounding factors that aren't in the dataset. Horsepower and weight appear to have the greatest influence on mileage, so our final model focuses on the interaction between these two variables:

```
##                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
```

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.

TODO: come to conclusions, ensure initial questions are explicitly answered

---

Appendix

Fig 1: Count of MPG Values of Cars

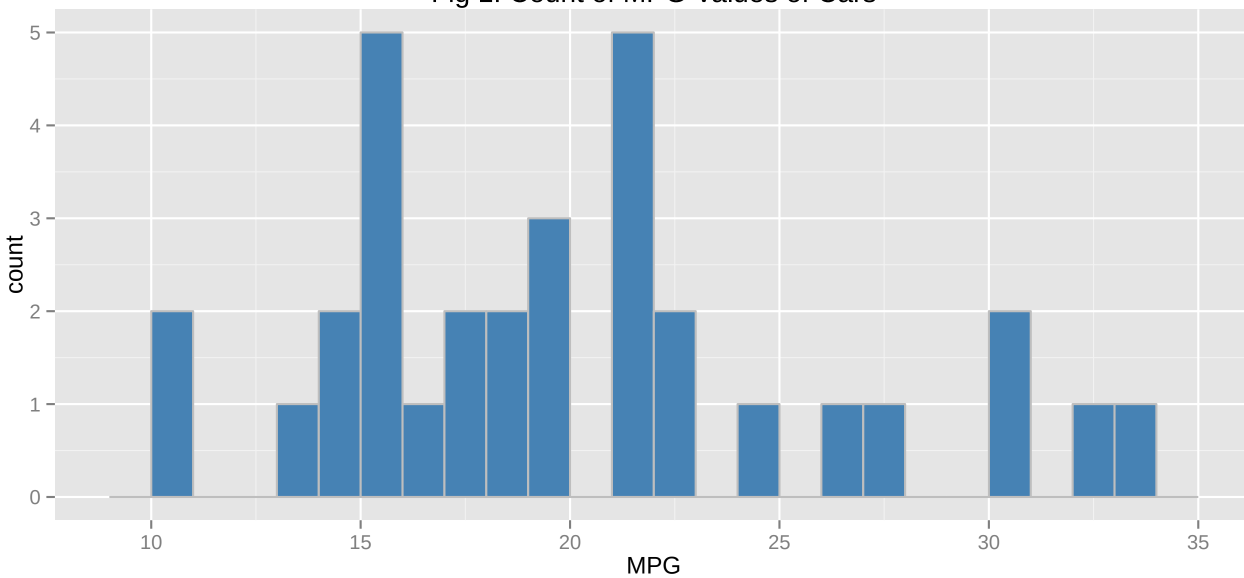

Fig 2: Densities of MPG Values by Transmission Type

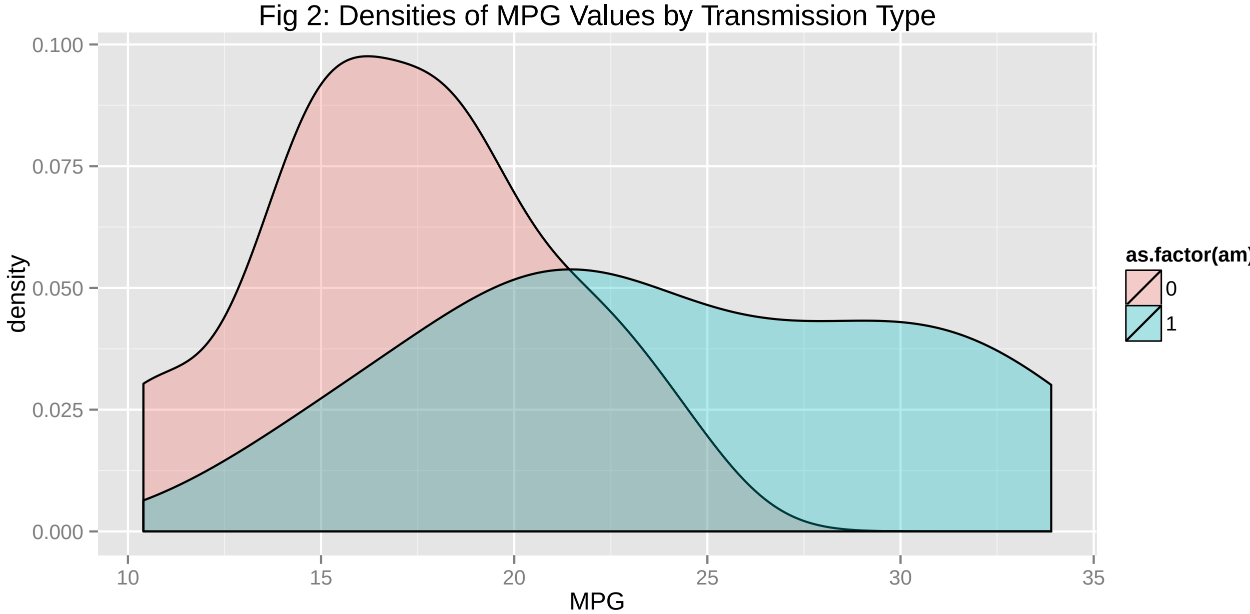

Fig 3: Linear Regression Line for MPG Predicted by Transmission Type

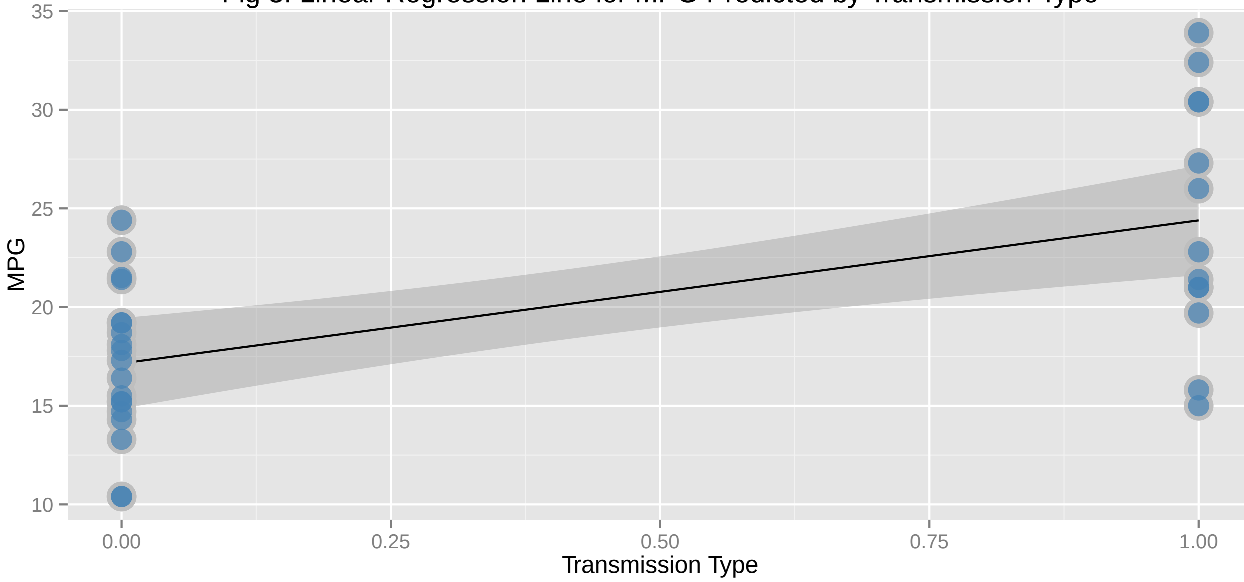

Fig 4: Plot of mpg~am Residuals

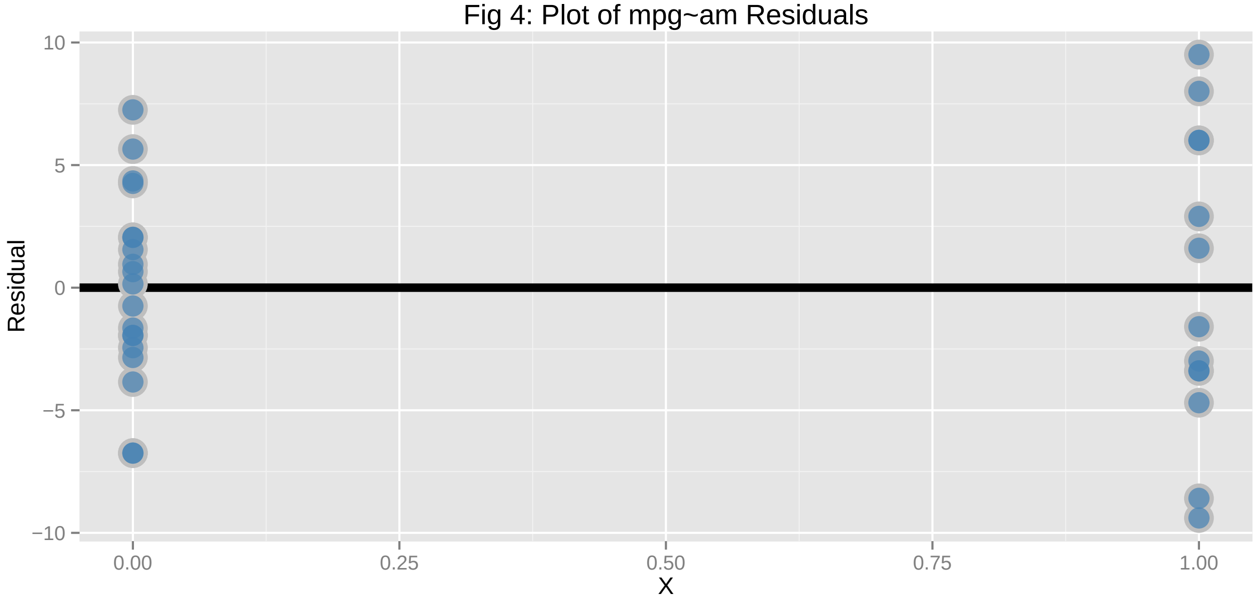

Fig 5: Correlation Matrix of All Factors