Analysis of Impact of Transmission Type on Vehicle Mileage

presented by **Brad Mager**

Summary

We wish to determine if an automatic or manual transmission is better for miles per gallon (MPG), and to quantify the MPG difference between the two transmission types. A preliminary analysis seems to indicate that manual transmissions perform much better than automatic. However, when we look at the other factors involved, it becomes clear that the weight of the car is the most important factor. In fact, we can predict the MPG of a car through a linear regression model using just a combination of second-degree polynomials of the weight and horsepower.

Exploratory Analysis

We can start by looking at a simple box plot that compares transmission type to MPG, as shown on the left side of Figure 1 in the Appendix. It initially appears that transmission type plays a major role in mileage. However, as the right side of Figure 1 shows, transmission type correlates strongly with weight. This makes sense, as one can hypothesize that people who like to drive large cars want automatic transmissions, or that smaller, sporty cars typically have manual transmissions.

In fact, simply using weight as a predictor with a cut-off of 3,000 lbs, we can achieve 91% accuracy in predicting what kind of transmission a car has. For cars above that weight, we predict automatic transmission, and below that weight we predict manual.

We should therefore re-phrase the questions about transmission types and MPG, and instead ask *what* are the main factors that affect MPG, and can we quantify those? The correlation matrix of all the factors included with the data set suggests we should consider number of cylinders, displacement, horsepower, and weight in our model, as their correlation coefficients with MPG are greater than 0.75. However, cylinders and displacement are highly correlated with each other *and* with weight, so they may not add much to the model.

Fitting Models

The simplest model involves regressing weight (wt) or horsepower (hp) onto MPG. However, the graph of weight measured against MPG suggests a polynomial model, while log of horsepower may work better for that variable. In addition, we can try adding transmission type (trans) as a predictor to see if that improves on the other models. Thus, we can try four models:

1. $MPG = \beta_0 + \beta_1 wt$
2. $MPG = \beta_0 + \beta_1 wt + \beta_2 wt^2$
3. $MPG = \beta_0 + \beta_1 wt + \beta_2 wt^2 + \beta_3 log(hp)$
4. $MPG = \beta_0 + \beta_1 wt + \beta_2 wt^2 + \beta_3 log(hp) + \beta_4 trans$

The line fits of the first two models are shown in Figure 2, clearly showing that the quadratic model fits the data points better than the straight line. As a side note, we can also see how well the log model for horsepower fits the data, as shown in Figure 3.

Residual Analysis

Now let's quantify how well these models do, so we can choose from among them. The summary() function in R provides the residual standard error, which measures how far off the model is, and the R-squared value, which tells us the percentage of variance explained by the model. Results for the four models are:

Model	Residual S.E.	R-squared
1	3.046	0.7446
2	2.651	0.8066
3	2.106	0.8779
4	2.142	0.8737

Adding the transmission type (Model 4) doesn't do as well as Model 3, with the polynomial of weight and log of horsepower, though it doesn't hurt the model much. Since the models are nested, i.e., each builds upon the previous one, we can perform an analysis of variance (ANOVA) on them, with the following results:

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	30 278.32				
2	29 203.75	1	74.576	16.2583	0.0004065
3	27 124.16	1	79.583	17.3497	0.0002854
4	26 123.85	1	0.314	0.0684	0.7956399

Again, it's clear that adding in transmission type does nothing in building an accurate model to predict mileage, since the P-value for the transmission coefficient is 0.796, indicating it is not significant. Based on the above, Model 3 is the best fit for this data — it has the lowest residual sum of squares and highest R-squared values. We can check it with a residual plot, as shown in Figure 4, which has no discernible pattern. Summing up the residuals gives a value of $1.554312e-15$, a tiny number that shows the model is a good fit.

The final model is: $MPG = 44.60 - 19.61wt + 6.10wt^2 - 5.02 \log hp$

Conclusion

Although the original concern was about how transmission type affects mileage, this may not be the right question to ask about this data set. We can clearly see that weight and horsepower play a much greater role, and can build a regression model based on those predictors. The best model includes a second-degree polynomial of weight combined with the log of horsepower, and including transmission type does not improve the results.

Appendix

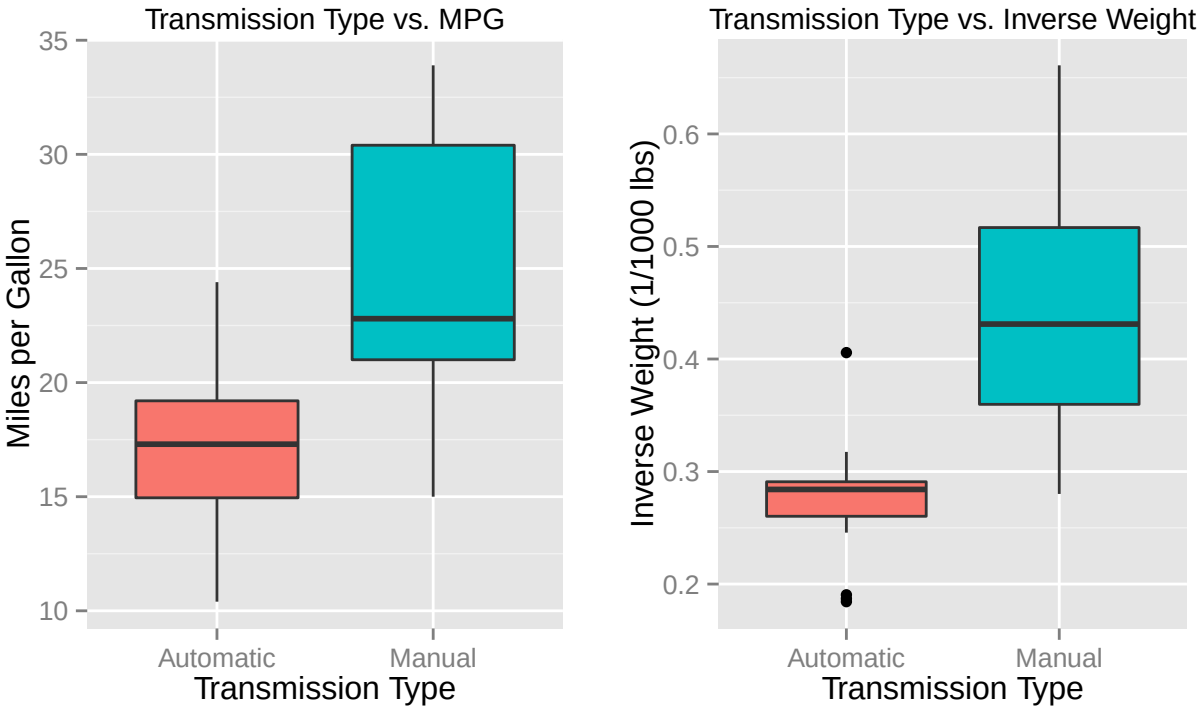

Fig. 1—*Comparison of transmission types. Left: Manual transmissions appear to be associated with higher MPG. Right: Manual transmissions are just as likely to be associated with lower weight.*

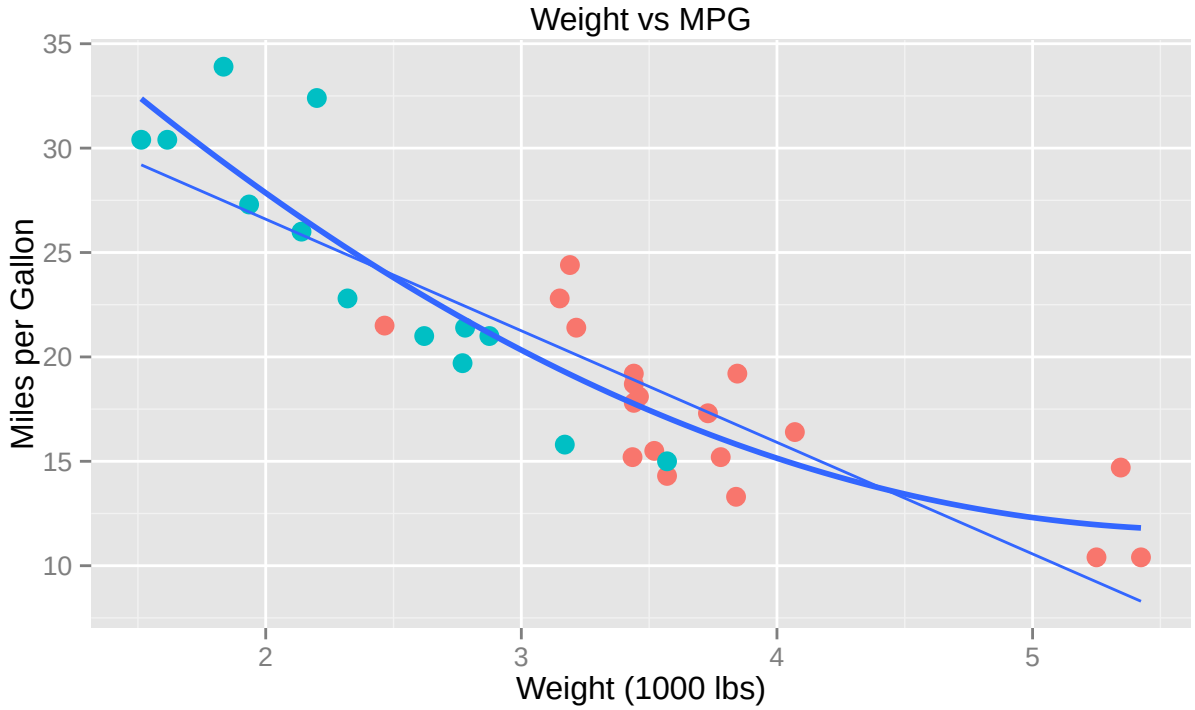

Fig. 2—*Weight vs. MPG. Automatic transmission points shown in red, manual in teal.*

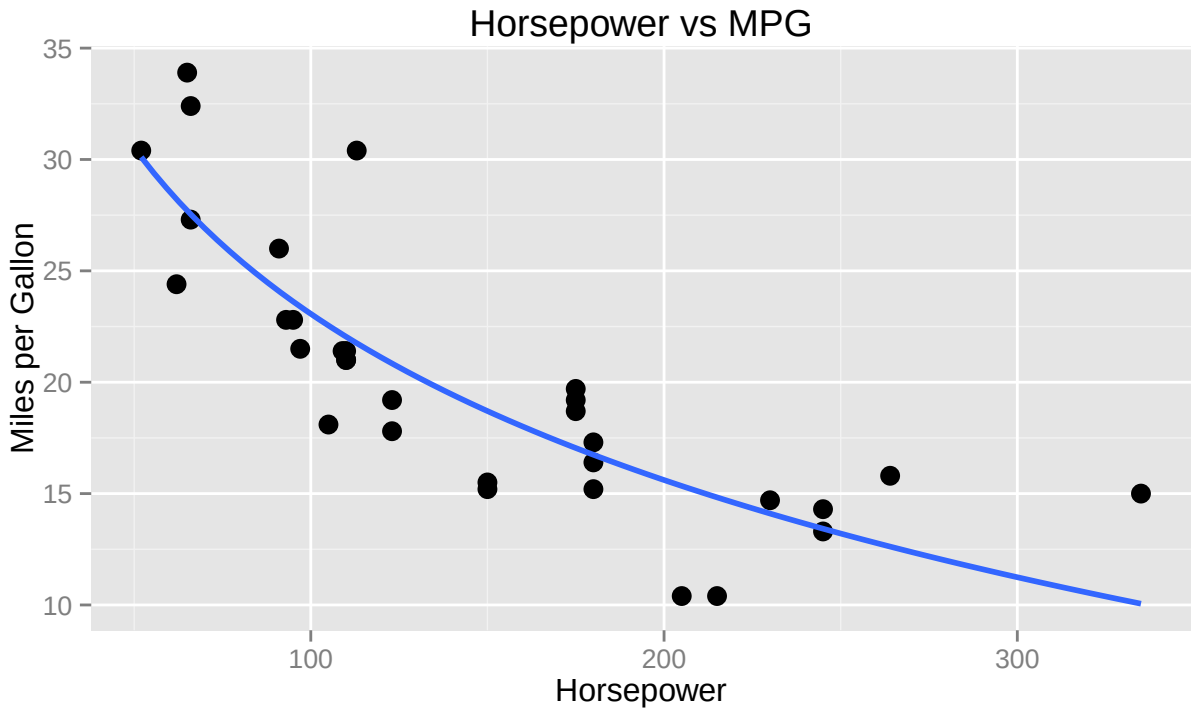

Fig. 3—*Horsepower vs. MPG. A log fit seems to work well with this data.*

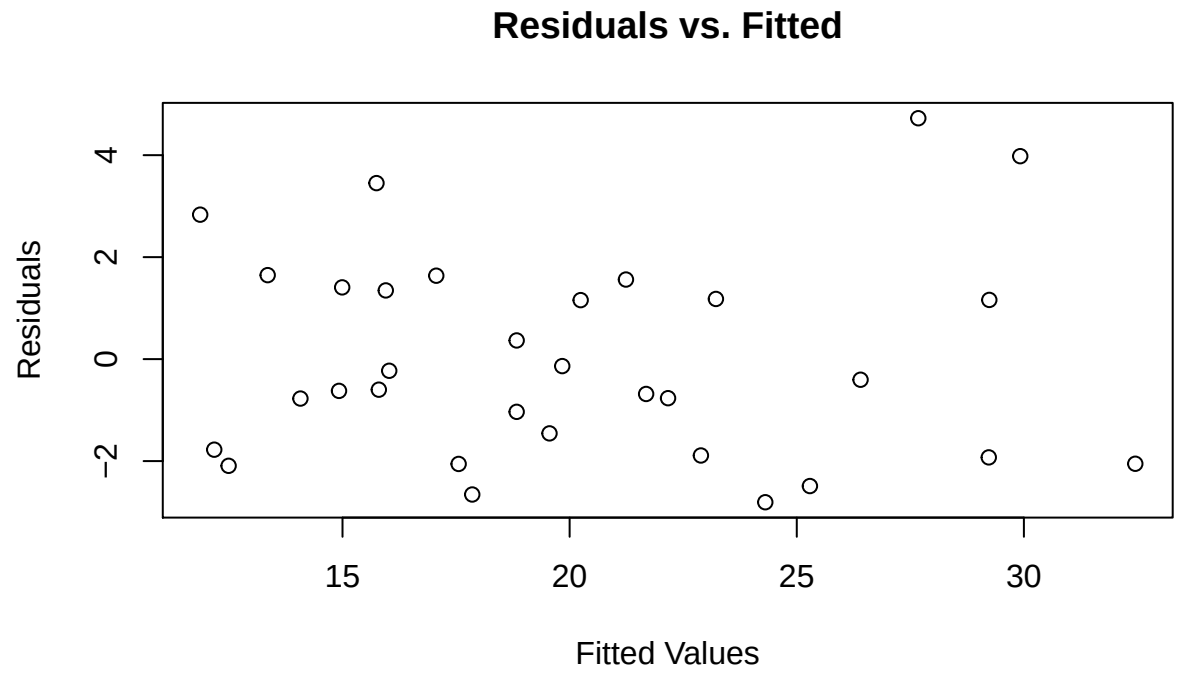

Fig. 4—*Residuals vs. Fitted Values for Model 4. There is no discernable pattern here.*