Is an Automatic or Manual Transmission Better for Vehicle MPG?

Jeff Spoelstra

2016-05-13

Executive Summary

This report describes an analysis of the mtcars data set in the R datasets package exploring the research question of whether an automatic or manual transmission is better for overall vehicle miles per gallon (mpg) fuel efficiency. The results showed that there is a correlation between mpg and transmission type, but other variables may be better predictors. The appendix at the end of this report contains supporting information.

Exploratory Data Analysis

The following description of the data set comes from the mtcars codebook.

The mtcars data set consists of eleven attributes (variables) of automobile design and performance for 32 different 1973-1974 model vehicles (observations) of various classes from different manufacturers (domestic U.S. and foreign). The data was compiled from 1974 *Motor Trend* magazine. The variables include miles per gallon, number of cylinders, engine displacement, transmission type, and several others.

Over all of the vehicles, mpg ranges from 10.4 to 33.9. Of the 32 vehicles, 13 (40.6%) have a manual rather than automatic transmission. There are no missing values in the data.

Hypothesis Testing

The target of the analysis was to find a possible correlation between vehicle tranmission type and mpg. The null hypothesis is that mpg is independent of transmission type. To reject this hypothesis would require evidence of a significant and quantifiable relationship between them.

Fitting a simple linear model with mpg as the outcome and transmission type as the only explanatory variable shows a statistically significant relationship with a beta1 coefficient of 7.24494 and $p<0.05$ (0.00029). However, the model isn't a particularly good fit with a r-squared value of 0.3598 and wide 95% confidence intervals (see below).

```
##               2.5 % 97.5 %
## (Intercept) 14.8506 19.444
## am           3.6415 10.848
```

The variances are high as well; 14.6993 for the automatic transmission residuals, 38.02577 for the manual transmission residuals, and 23.25473 overall.

See Figure 1 and Figure 2 in the Appendix for model fit and residuals charts. The sloped blue line on Figure 1 shows the regression line for the model. The horizontal green lines on Figure 2 indicate two standard deviations away from the mean of zero separately for automatic transmission residuals and manual transmission residuals.

Looking at the full set of variables in the data set, there are several that could confound the target relationship, in particular vehicle weight, number of cylinders, engine displacement, engine horsepower, engine design (V-style or straight-line), number of carburators, number of gears in the transmission, and rear differential

ratio. In a step-wise manner each variable was added to the model and a new linear model created until all variables were combined in one model; resulting in a total of nine models. An ANOVA analysis was then performed with all the models.

The ANOVA results showed that the best performing model appears to be one with cylinder count included with transmission type. The r-squared value is a more favorable 0.75901 as is the overall variance of 8.75362.

However, looking at the p values of the beta coefficients, the value for cylinder count is nearly zero (0) while the value for transmission type is just above 0.05 (0.05635) - thus adding cylinder count has effectively rendered transmission type irrelevant to the model.

Figure 3 in the Appendix shows a chart of the residuals. The horizontal green lines indicate two standard deviations away from the mean of zero for all residuals. Note that there appears to be some grouping affect on the residuals at the left and right ends of the X axis probably induced by the binary nature of the transmission type variable.

Further review of the ANOVA results suggests vehicle weight as an important factor, and, in fact, a model relating mpg to just cylinder count and vehicle weight appears to perform best with both beta coefficient p values far below 0.05. Additionally, the r-squared value is higher (0.83023), and the residuals variance is lower (8.75362), than any of the other models tested. The 95% confidence intervals are narrower than the original simple model, too (see below).

```
##                2.5 %   97.5 %
## (Intercept) 36.1787 43.19380
## cyl         -2.3559 -0.65966
## wt          -4.7390 -1.64292
```

The confidence interval for cylinder count comes dangerously close to including zero, but a model tested without it (relating mpg to just vehicle weight) performed less well than the model with both variables.

Figure 4 in the Appendix shows a chart of the residuals for this model. The horizontal green lines indicate two standard deviations away from the mean of zero for all residuals. With no binary explanatory variables, the residuals of the best fitting model show a more normal distribution (see Figure 5), albeit a litte skewed; perhaps because the sample of 32 vehicles was not equally distributed across the possible number of cylinders. For example, there are 14 eight-cylinder vehicles in the data, but only 7 six-cylinder vehicles. The vertical green lines on Figure 5 indicate one standard deviation from the mean of zero.

Conclusion

Based on the results obtained from the simple model of mpg as a function of just transmission type, it is possible to reject the null hypothesis and accept the alternative that mpg is related to transmission type. Even so, using transmission type alone isn't the best model to fit the data, and can, in fact, be eliminated altogether in favor of number of cylinders and vehicle weight instead.

Appendix

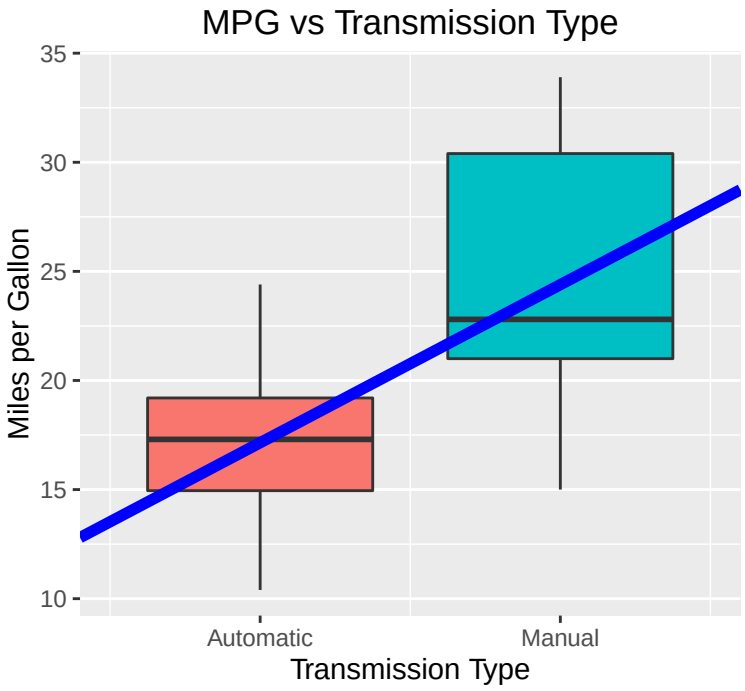

Figure 1

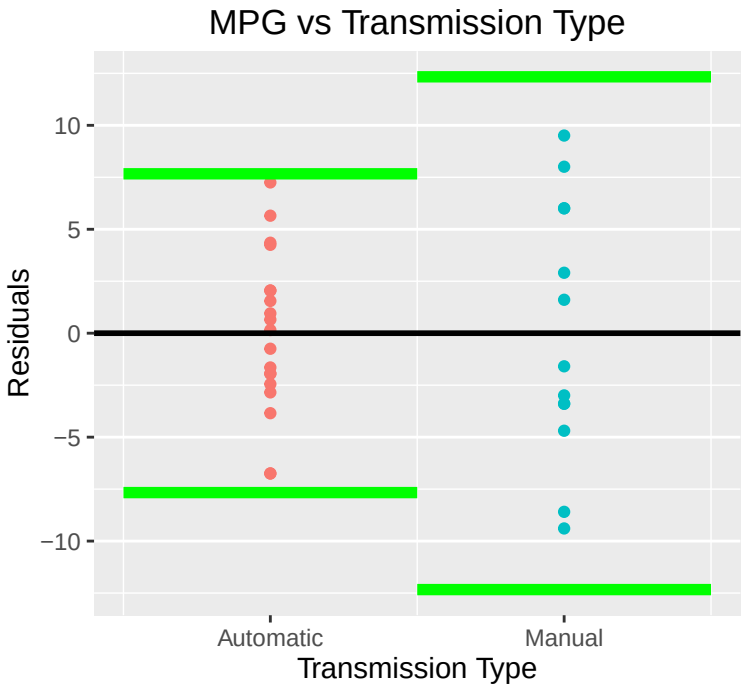

Figure 2

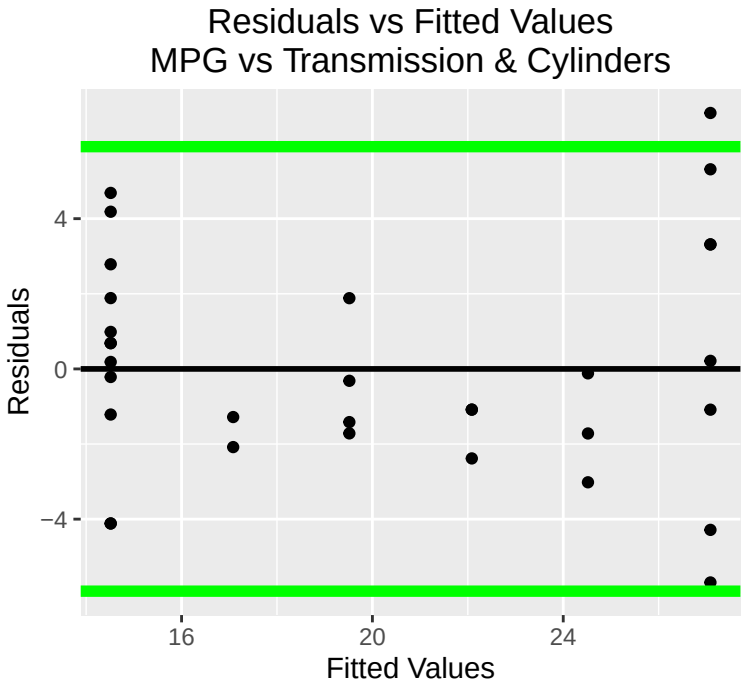

Figure 3

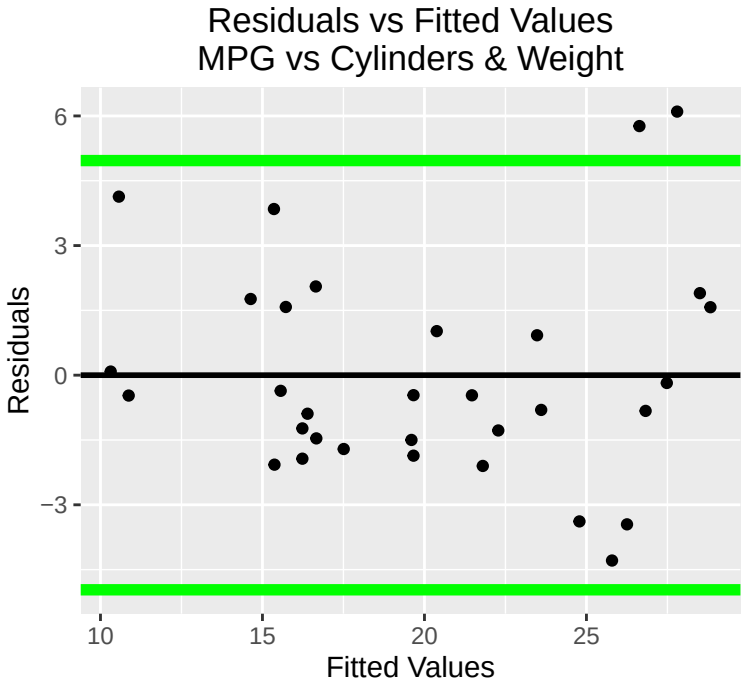

Figure 4

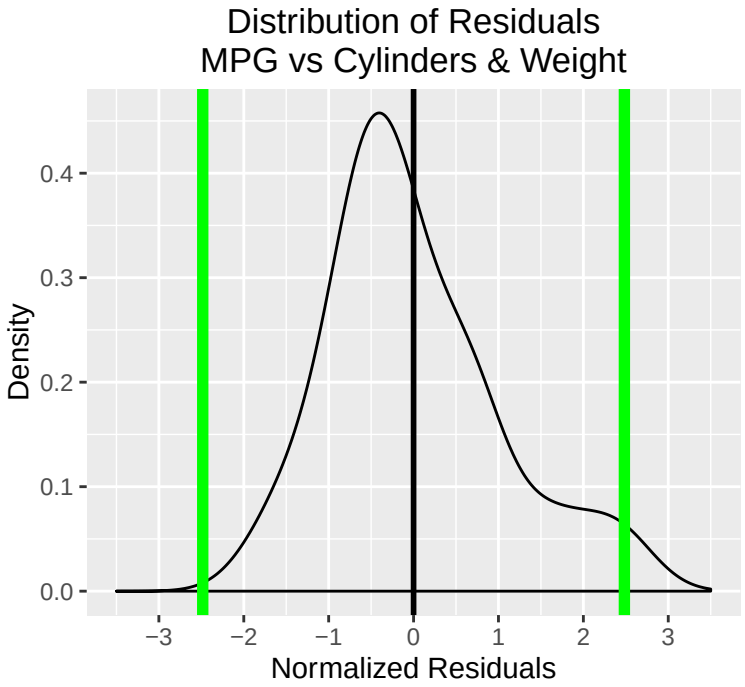
Distribution of Residuals
MPG vs Cylinders & Weight
Density
0.4
0.3
0.2
0.1
0.0
−3
−2
−1
0
1
2
3
Normalized Residuals

Figure 5