Fuel Consumption with Manual versus Automatic Transmissions

T. Bennett

August 23, 2015

Executive Summary

We estimate that a car with a manual transmission gets 2.7(95% CI x to y) more miles per gallon than a car with an automatic transmission after adjustment for vehicle weight, number of carburetors, and quarter mile time.

Data Source

This analysis is based on the `mtcars` dataset included in the `datasets` package in base R. It contains data on the 1973-74 models of 32 different automobiles including fuel consumption and 10 aspects of automobile design. This dataset (not shown) contains 32 rows and 11 fields without missing data.

Exploratory Data Analyses

Miles per gallon (`mpg`) is the primary outcome. Transmission type (`am`) is the primary predictor of interest. 13 of 32 have a manual transmission and 19 of 32 have a manual transmission.

A bivariate plot does suggest that manual transmissions are associated with better gas mileage. See Appendix [Figure 1][].

A t-test of `mpg` using unequal variances bears that out: cars with automatic transmissions get 17.1 miles per gallon and cars with manual transmissions get 24.4 miles per gallon, t-test $p = 0.001$.

Potential Confounders

Several other variables might confound the Transmission Type - Mileage relationship. Based on my working knowledge of automobile engines, any of the other 9 variables in the `mtcars` dataset might be hypothesized to be important. The bivariate plots of the other 8 variables against `mpg` shown in Appendix [Figure 2][fig2] suggest that any of them could reasonably be a predictor.

To avoid including two covariates that are either highly correlated themselves or highly correlated with the primary predictor of interest, I created a `pairs` plot (not shown, code in Appendix [Figure 3][fig3]).

On the `pairs` plot, number of cylinders, weight, and displacement are highly correlated. I decided to include displacement and weight in the model selection process because they are continuous variables that reflects physical properties of the car. Number of cylinders is also reasonably highly correlated with transmission type, our primary predictor of interest. Horsepower is also highly correlated with number of cylinders, but not quite as highly correlated with the other variables, so I will consider it during model selection.

Linear Regression including all potential predictors (Model 1)

In this full model, we estimate that cars with manual transmissions get 2.6(95% CI x to y) more miles per gallon than cars with automatic transmissions after adjustment for 8 other variables. However, this comparison is not statistically significant. Weight is the variable that appears to potentially be an important predictor of fuel consumption, as each

1000 pounds of additional weight is associated with 3.7(95% CI x to y) fewer miles per gallon, $p =$ []. This model fits the data reasonably well (adjusted R^2 = [0.82]).

Assessing variance inflation

Our concern about the correlation between weight, horsepower, and displacement appears well-founded, as those variables have the highest variance inflation factors (VIFs). Horsepower and displacement were poor predictors of fuel consumption in the full model, therefore we will remove those variables and re-fit the model.

Model without horsepower or displacement (Model 2)

Model 2 still fits the data reasonably well (adjusted R^2 is the same, [0.82]), but the VIFs for the variables remaining in the model are now similar. The estimated effect of a manual transmission has changed only marginally (2.5 mpg instead of 2.6) and remains not statistically significant. The weight effect is slightly smaller (3.0 instead of 3.7), but is now statistically significant.

Highly parsimonious model (Model 3)

Because vehicle weight was the only significant predictor in either previous model, I considered a model containing only transmission type and vehicle weight.

This model is likely too parsimonious, as the adjusted R^2 decreased to [0.74]. As you might expect, a nested likelihood ratio (LR) test comparing the highly parsimonious model to the model without horsepower or displacement was significant, $p = 0.002$. Interestingly, transmission type has no association with fuel consumption in this model.

Adding variables back to the parsimonious model (Model 4)

Of the remaining variables, quarter mile time (perhaps a proxy for several vehicle and tuning properties) and number of carburetors were the closest to significance in the previous models. Therefore, we fit a model that added them back to the parsimious model.

Model 4 has the best adjusted R^2 of any considered thus far ([0.84)]). It also has no difference by LR test from Model 2, which contained several more variables. The variance inflation factors of Model 4 are the lowest of any model evaluated thus far. Model 4 will the basis for our final model.

Evaluation of Residuals

The residual plot from Model 4 (Appendix [Figure 4][fig4]) suggests that it fits the data reasonably well. The Residuals vs Fitted panel shows that, as might be expected in such a small data set, the model does not fit as well where the data are sparse, i.e. for very low or very high fuel consumption cars. Points 3, 17, and 18 have residuals that are 2 standard deviations from the regression line and have at least medium leverage. Point 17, in particular (see the Scale-Location plot), might be quite influential to the fit.

Diagnostics

Point 17 does appear to be quite influential, as it has by far the largest Cook's Distance and the largest `dfbeta` and `dffit` value of any point. We will evaluate the model with that point removed. After re-fitting the model with point 17 removed, the updated residual plots (Appendix [Figure 5][fig 5] suggest a better fit. This will be our final model.

Estimates from Final Model

We estimate that a car with a manual transmission gets 2.7(95% CI x to y) more miles per gallon than a car with an automatic transmission after adjustment for vehicle weight, number of carburetors, and quarter mile time.

Appendix

Figure 1 - MPG, by transmission type

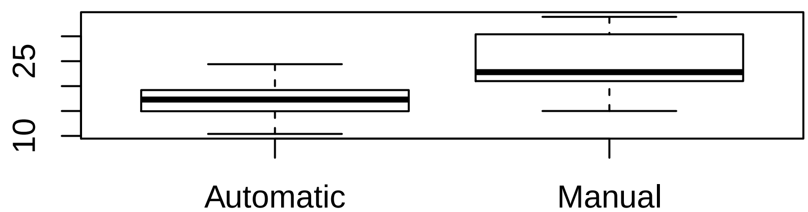


Figure 2 - Bivariate Plots

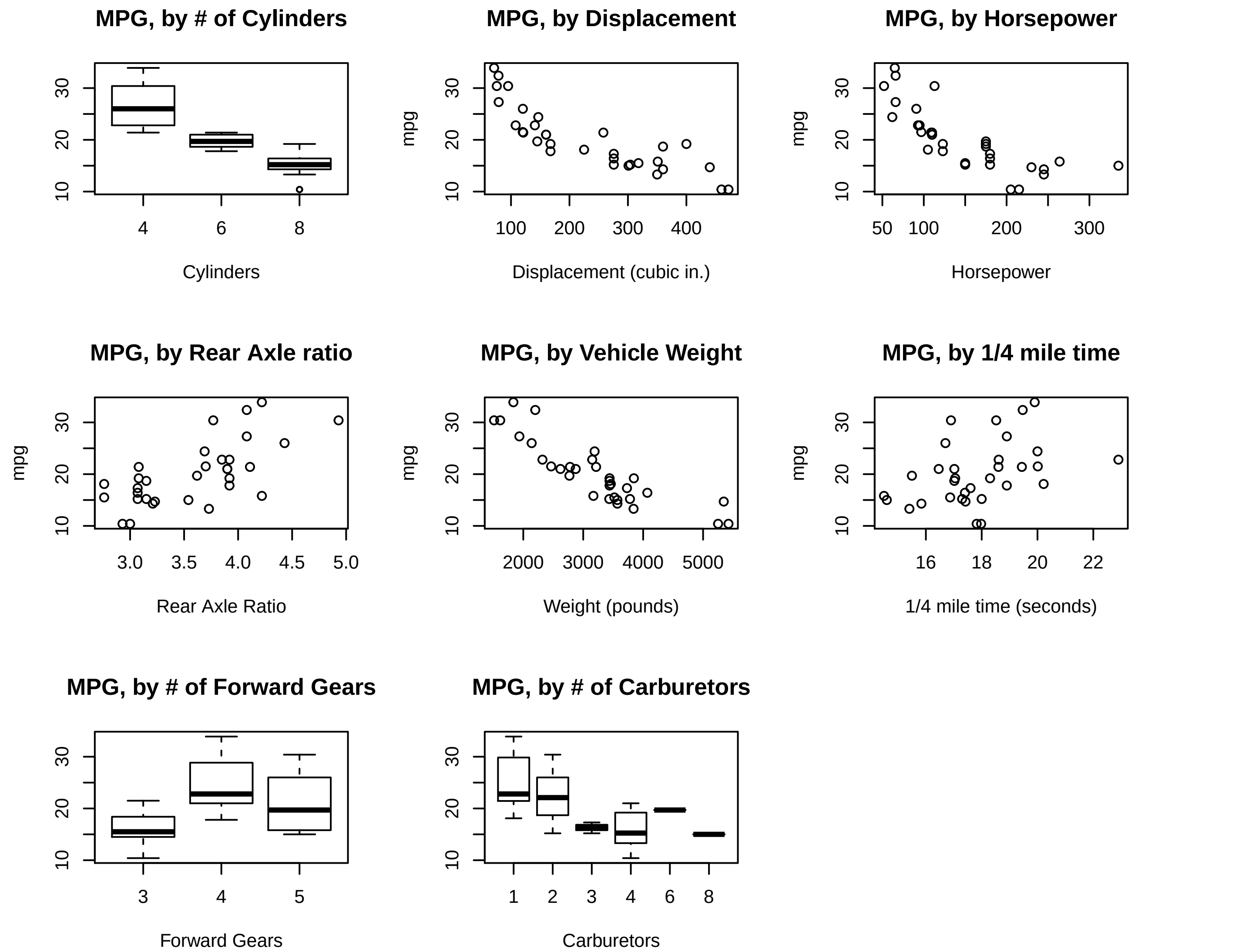


Figure 3 code

```
data(mtcars)
p <- ggpairs(mtcars, columns = 2:11, lower = list(continuous = "smooth"), params = c(method = "loess"))
```

Figure 4 - Residual plots from Model 4

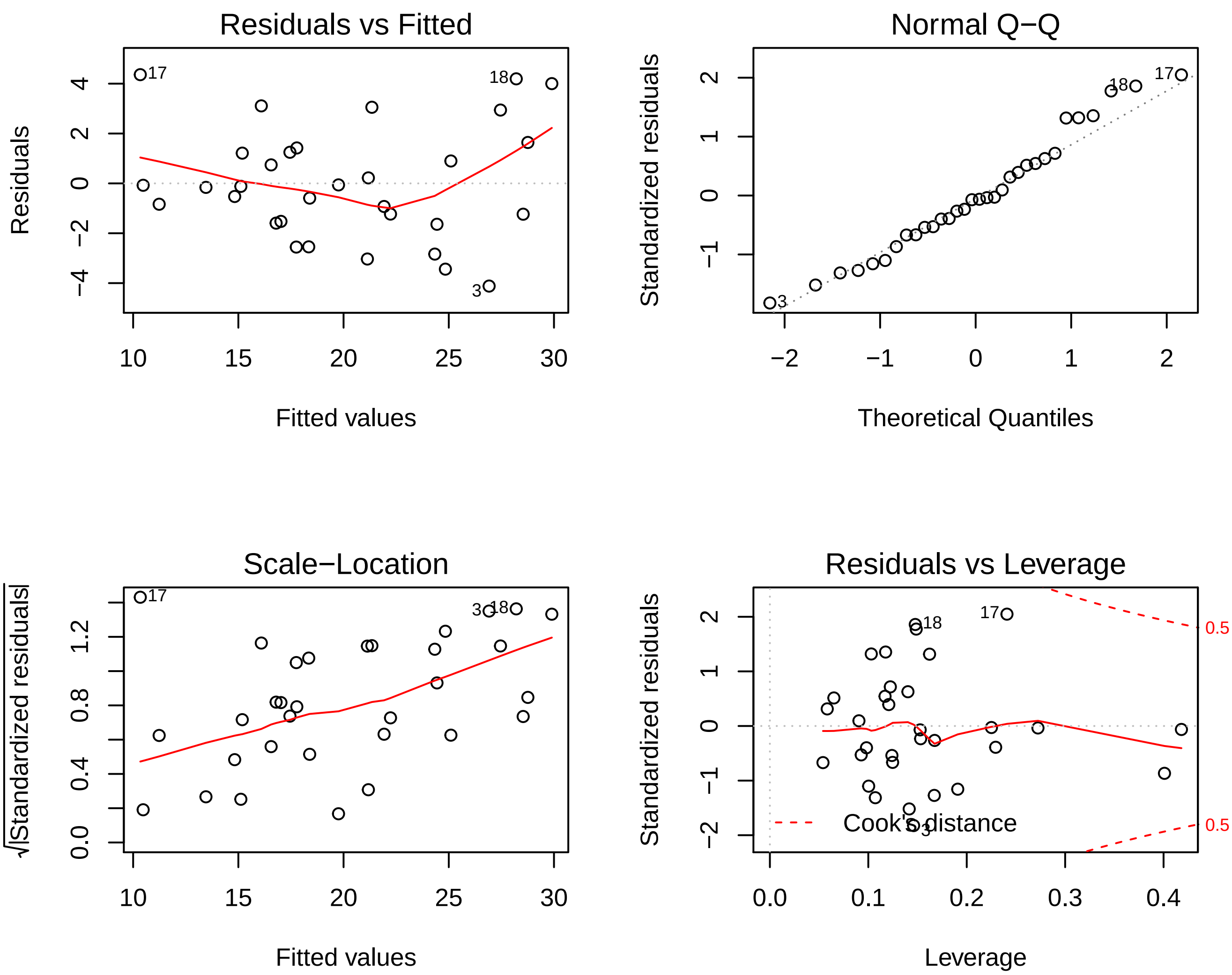
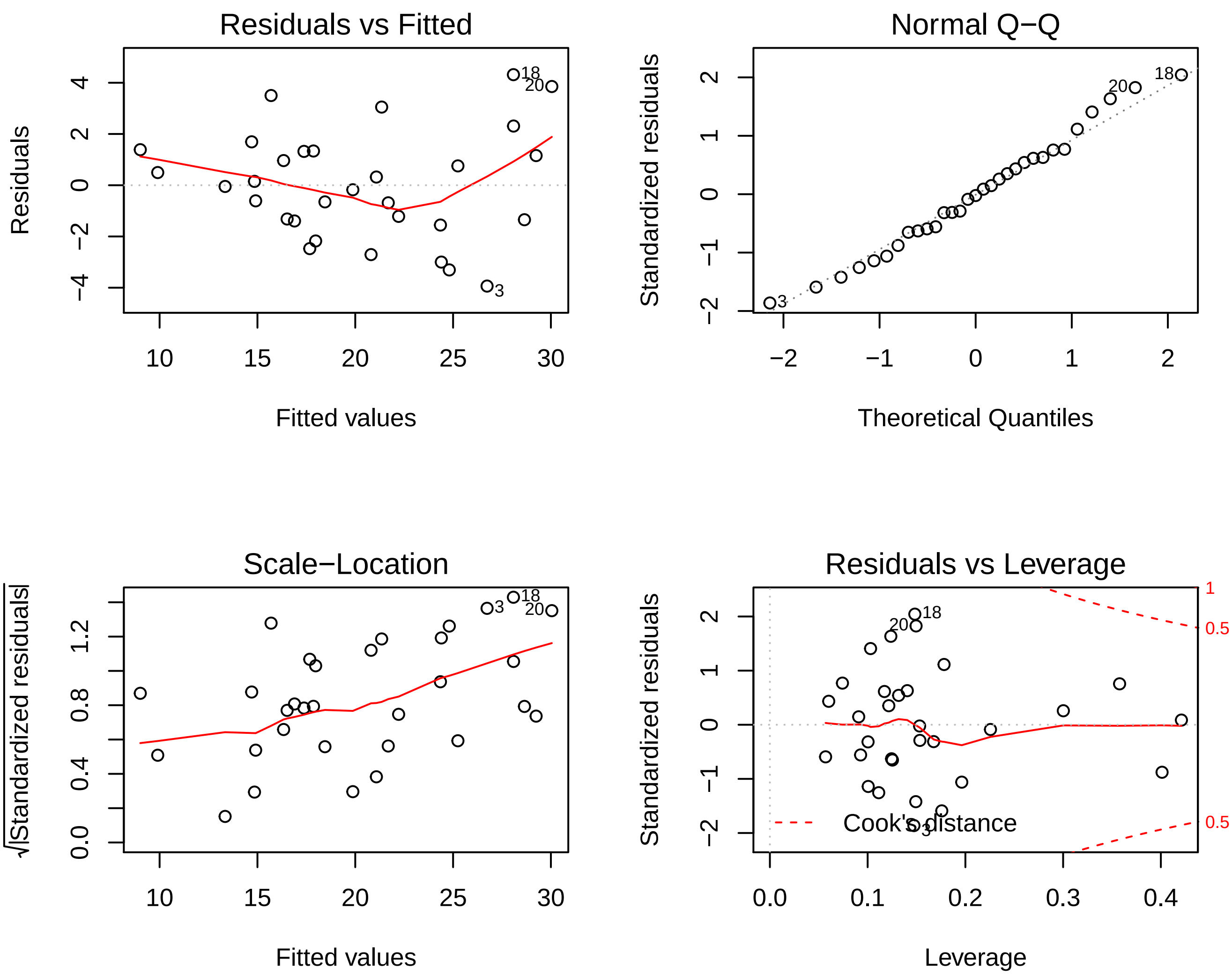


Figure 5 - Residual plots from Final Model



Reference

The code (Project.Rmd) for this analysis can be found here.