The Impact of Transmission Type on Fuel Economy

Jonathan Labin

October 21, 2015

Executive Summary

This report attempts to determine which transmission type is better for fuel economy: Manual or Automatic. This analysis is performed using the mtcars dataset and develops a regression model to determine the fuel economy difference between the two types of transmission. While the overall trends produced by the model indicate a difference, the variability and small sample size results in a confidence intervals such that the effect of transmission cannot be clearly distinguished.

Load the data

This dataset is provided with R in the datasets library. See ?mtcars for more information.

```r
library(datasets)
data(mtcars)
```

Exploratory Analysis

We are interested determining a model which predicts fuel economy (mpg) based on a subset of the remaining variables. As a first cut, let's explore the correlation between the remaining variables and mpg. The flowing lists the absolute values from the mpg row of the correlation matrix.

```r
sort(round(abs(cor(mtcars)["mpg",]),3), decreasing=T)
```

```
##   mpg    wt   cyl  disp    hp  drat    vs    am  carb  gear  qsec
## 1.000 0.868 0.852 0.848 0.776 0.681 0.664 0.600 0.551 0.480 0.419
```

Clearly mpg itself is exactly correlated. Beyond that, it appears that weight (wt), the numer of cylinders (cyl) and the engine displacement (disp) are the highest correlated to economy. It is expected that in general, the heavier the vehicle, the lower the fuel economy might be. Displacement and number of cylinders are both a type of measures of engine size and so only one should be included in our model. You might expect that a vehicle tuned to provide more horsepower (hp) may be less fuel efficient but this variable would also be very much dependent on engine size.

We also see that our target variable, transmission type (am) has a reletively low correlation to economy. Further exploration of the effect of this variable in the context of some of the others will be required.

The vs and am variables in this dataset are really factor variables where the flag values are not nessesarily numerically related so we convert these variables into factors.

```r
mtcars$vs <- factor(mtcars$vs, labels = c("V-engine", "Straight"))
mtcars$am <- factor(mtcars$am, labels = c("Automatic", "Manual"))
```

Figure 1 and Figure 2 shown in the appendix show the dominating relationships between fuel economy and the highest two correlated variables (wt, cly) and how those relationships differ for each transmission type (am). Clearly, our models must take these highly influential variables into account in order to accurately model the effect of transmission type.

Fit Models

Let's explore building some models of this data. The following models each begin with our target variable (am) annd add additional variables to the model. We will then decide which variables are likely to be appropriate for our model with a alpha = 0.05 cutoff for the resulting p values.

```
fit1 <- lm(mpg ~ am, mtcars)
fit2 <- update(fit1, mpg ~ am + wt)
fit3 <- update(fit2, mpg ~ am * wt)
fit4 <- update(fit3, mpg ~ am * wt + factor(cyl))
fit5 <- update(fit4, mpg ~ am * wt + factor(cyl) + drat)
round(anova(fit1, fit2, fit3, fit4, fit5)$`Pr(>F)`,4)
```

```
## [1]     NA 0.0000 0.0004 0.0209 0.9429
```

The above p-values indicate that the addition of weight (wt) and the combined wt:am factor both have a reasonably significant effect in our model. Including the number of cylinders (cyl) is approaching our cuttoff but including drat is well outside usefulness here. Figure 3 in the appendix shows the parallel prediction lines for model fit2. Figure 4 in the appendix shows the prediction lines with different slopes for generated by model fit3. The remaining models would require more advanced visualizations in order to view the additional dimensions. We select fit4 which includes the interaction term wt and the a term for the cyl variable.

To confirm that our model is appropriately representing the variation in the data, Figure 5 provides the residual plots for the model. These plots do not appear to reveal any systematic patterns in the residual data. The error tearms appear to be normal in the Normal Q-Q plot.

Above, we selected fit4 = mpg ~ am + wt + factor(cyl) + am:wt as an appropriate model for our data. This section inspects this model to make inferences about the relationship between transmission type and fuel efficiency. First, let's look at the summary of the model coefficients:

```
summary(fit4)$coef
```

```
##                 Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)    29.774836  2.8403415 10.482836 7.870715e-11
## amManual       11.568790  4.0877912  2.830083 8.853842e-03
## wt             -2.398713  0.8439884 -2.842116 8.603904e-03
## factor(cyl)6 -2.709777  1.3573517 -1.996370 5.646509e-02
## factor(cyl)8 -4.776110  1.5558306 -3.069814 4.964603e-03
## amManual:wt    -4.067981  1.3974151 -2.911075 7.295503e-03
```

According to the exact values from this model, while holding all other variables constant, we expect a 2.4 drop in mpg for every 1,000 lbs increase in vehicle weight for automatic transmission vehicles and a 6.47 drop in mpg for every 1,000 lbs increase in vehicle weight for manual transmission vehicles. However, since the lines cross, neigher transmission dominates across all vehicle weights. Instead, in weights below around 2844 lbs, manual transmission vehicles result in higher fuel economy and automatic transmission vehicles produce better economy for vechicles above that weight.

However, all of the previous reasoning considers only the exact values. When considering confidence intervals (at the 95% level), we expect between 0.67 and 4.13 drop in mpg for automatic transmission vehicles and between 1.88 and 11.06 drop in mpg manual transmission vehicles.

These intervals contain a signfficant overlap even at the 95% confidence level. From this it can not be concluded that transmission type has a signfficant effect on the fuel economy of the vehicles. Perhaps with a larger dataset, the confidence intervals could be tightened to distinguish a difference.

Appendix: Figures

Figure 1.

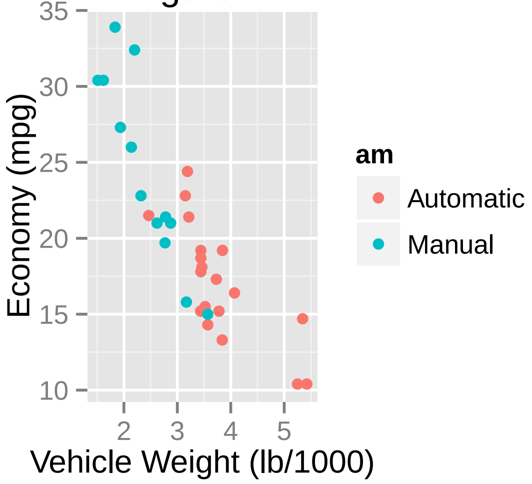


Figure 2.

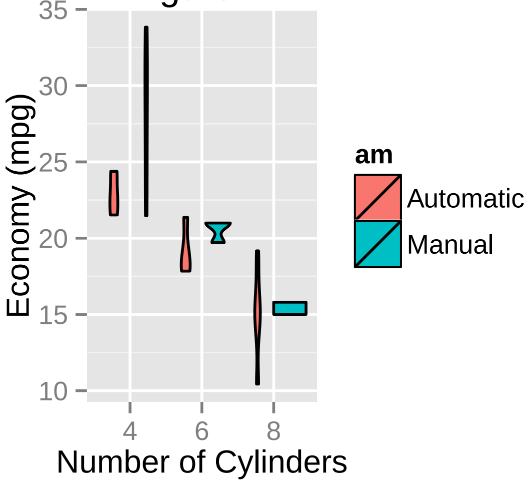


Figure 3.

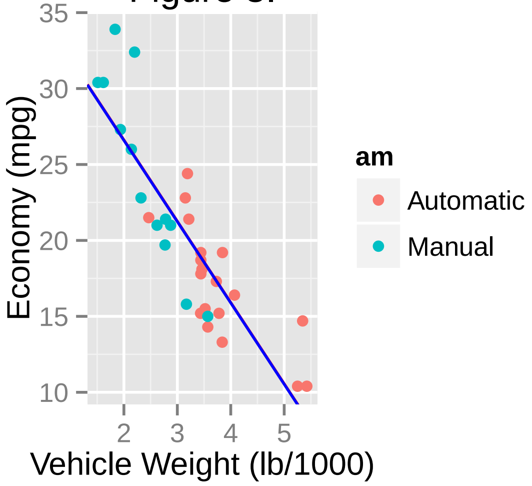


Figure 4.

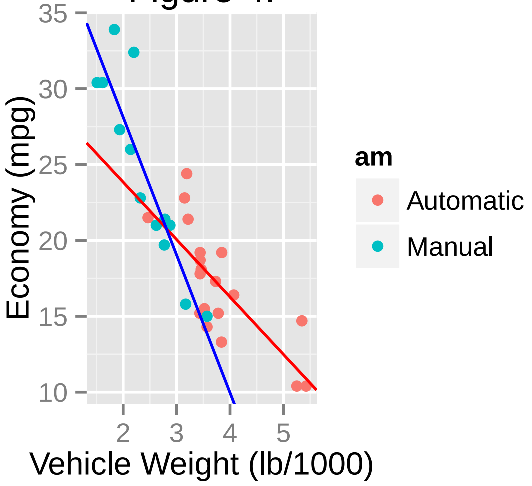
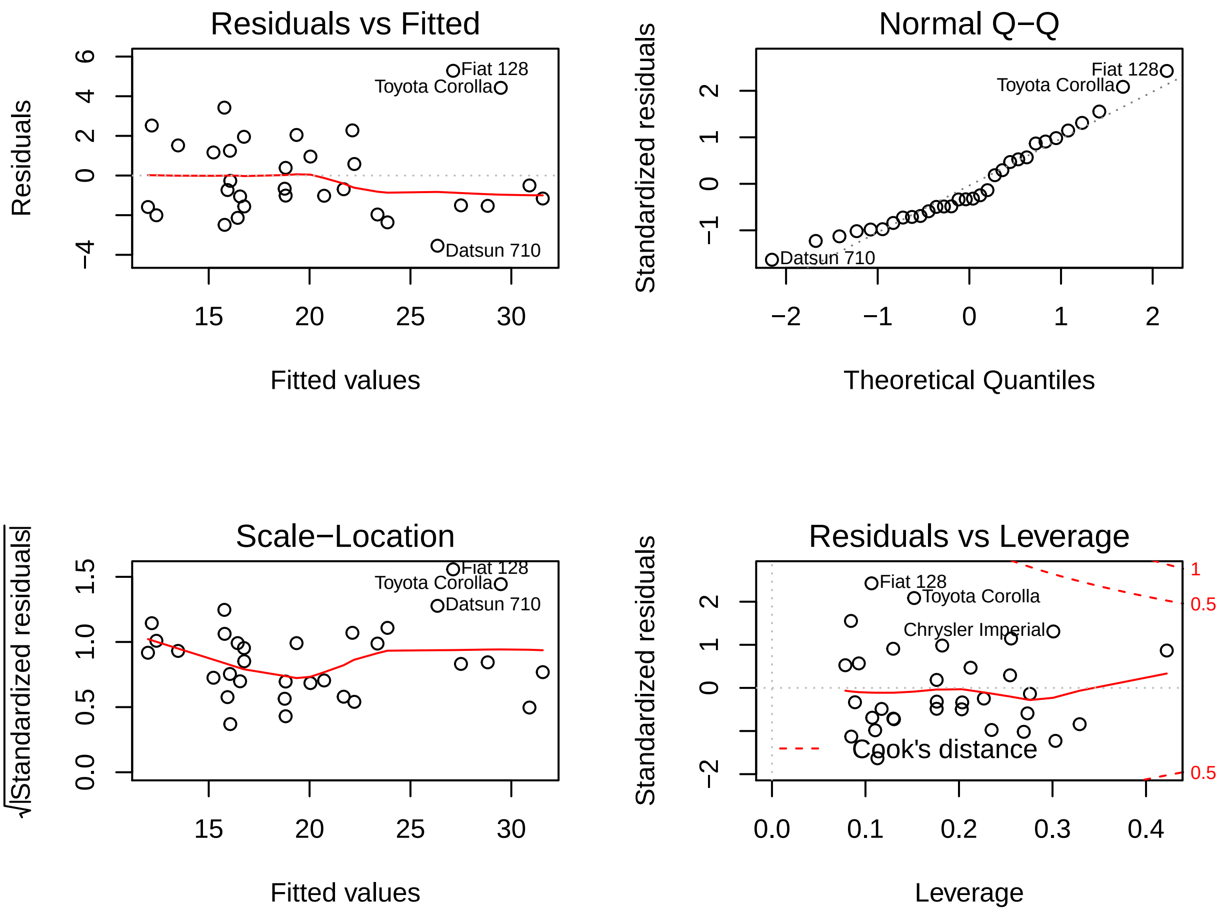


Figure 5. Residual Analysis (model fit4)

lm(mpg ~ am + wt + factor(cyl) + am:wt)



Residuals vs Fitted
Residuals
Fitted values
−4
0
2
4
6
15
20
25
30
Fiat 128
Toyota Corolla
Datsun 710
Normal Q−Q
Standardized residuals
Theoretical Quantiles
−1
0
1
2
−2
−1
0
1
2
Fiat 128
Toyota Corolla
Datsun 710
Scale−Location
√|Standardized residuals|
Fitted values
0.0
0.5
1.0
1.5
15
20
25
30
Fiat 128
Toyota Corolla
Datsun 710
Residuals vs Leverage
Standardized residuals
Leverage
−2
0
1
2
0.0
0.1
0.2
0.3
0.4
Fiat 128
Toyota Corolla
Chrysler Imperial
Cook's distance
1
0.5
0.5