Transmission Effect on MPG

Lucas Qualmann

9/29/2019

Executive Summary

This report shows that a manual transmission gets better gas mileage than an automatic transmission after using horsepower as a confounding variable. Even with an outlier data point working against this hypothesis, the model confirmed this conclusion. This report highlights the steps taken and analysis done to confirm the model is a good fit. All of the R code used in this analysis isn't included in this document as it was not wanted under the criteria for this assignment.

Exploratory Data Analysis

First, we want to make sure there is some perceived relationship between mpg and transmission type, so we plot a boxplot showing mpg based on transmission type (see appendix). This plot shows that manual transmissions have a higher mpg rate than automatic transmissions, so that is the hypothesis our model will seek to prove.

Next we want to get an idea of what variables were most correlated to each other and mpg to know what variables would be most important to use for our model since the correlation between transmission type and mpg might be explained more by other variables than just transmission (see appendix for table). As a result of this table, we found disp, hp, wt, and carb to be highly correlated to each other. Drat and qsec were another pair of correlations followed by vs and am.

Model Selection

Based on our exploratory data analysis and knowledge of what car items should have a strong effect on mpg, we will narrow down our potential list of variables in addition to transmission type to cyl (# of cylinders), disp (displacement), hp (horsepower), wt(weight in 1,000 lbs), carb (# of carburetors), and vs (engine shape). After making a model with all of these variables, disp was eliminated as a possibility because it had the highest standard deviation of variables (highest correlation to the other variables in the model). In the next model run, vs and cyl were eliminated as they had the highest p-values. This left us with 3 variables to use for our model: hp, wt, and carb.

We'll build 8 linear models containing all of the combinations of the 3 remaing variables: 1 model of just transmission type, 3 models of transmission plus 1 variable, 3 models of transmission plus 2 variables, and 1 model of transmission plus 3 variables. We'll then use the anova function to compare the models. Based on the anova function, the model with the best F and P-values is mpg ~ factor(am) + hp - 1. So our model will look at transmission type and horsepower.

```
mdl2 <- lm(mpg ~ factor(am) + hp - 1, data = mtcars)
model <- mdl2
summary(model)
```

```
##
## Call:
## lm(formula = mpg ~ factor(am) + hp - 1, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.3843 -2.2642  0.1366  1.6968  5.8657
##
```

```
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## factor(am)0 26.584914   1.425094  18.655  < 2e-16 ***
## factor(am)1 31.861999   1.282279  24.848  < 2e-16 ***
## hp          -0.058888   0.007857  -7.495 2.92e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.909 on 29 degrees of freedom
## Multiple R-squared:  0.9825, Adjusted R-squared:  0.9807
## F-statistic: 543.4 on 3 and 29 DF,  p-value: < 2.2e-16
```

Diagnostics, Residuals, and Confidence Intervals

Next we'll do some model analysis. First we'll look at a plot of the model (see appendix). The first chart in this plot of 4 charts is residuals vs fitted plot. If there are any patterns in this chart, that's a sign there is a pattern that our model has missed. In this case, no strong pattern is developing, so there isn't anything to worry about here. The other three charts show us if there are any data points which look like outliers which could affect our analysis. There does appear to be one data point, the Maserati Bora, which could be affecting our analysis. We'll do a couple more pieces of analysis to confirm.

Looking at both a hatvalues plot (leverage of points) and dfbetas (slope change without a point) plot show the Maserati Bora as a highly leveraged and influential point for our model compared to the other points (see appendix for plots). While this could be messing up our model and have us want to remove the data point, when we look at the values for the Maserati Bora, we see it has a mpg of 15.0 and a manual transmission. This means that the data point is actually working against our hypothesis that manual transmissions have better gas mileage than automatic transmissions. Because of this, we will keep it to avoid biasing the model in a way which would help our hypothesis.

Finally, we want to look at confidence intervals. Since we've broken out both transmission types in our model, the coefficient of each variable is the mean of that transmission type over all horsepower. The mean for an automatic transmission is 26.6 mpg, and the mean for a manual transmission is 31.9 mpg. Let's look at the 95% confidence intervals for both of those values to make sure the type of transmission is significantly affecting the mpg.

```
##    2.5 %   97.5 %
## 23.67027 29.49956
```

Above: confidence interval of automatic transmission mpg.

```
##    2.5 %   97.5 %
## 29.23944 34.48455
```

Above: confidence interval of manual transmission mpg.

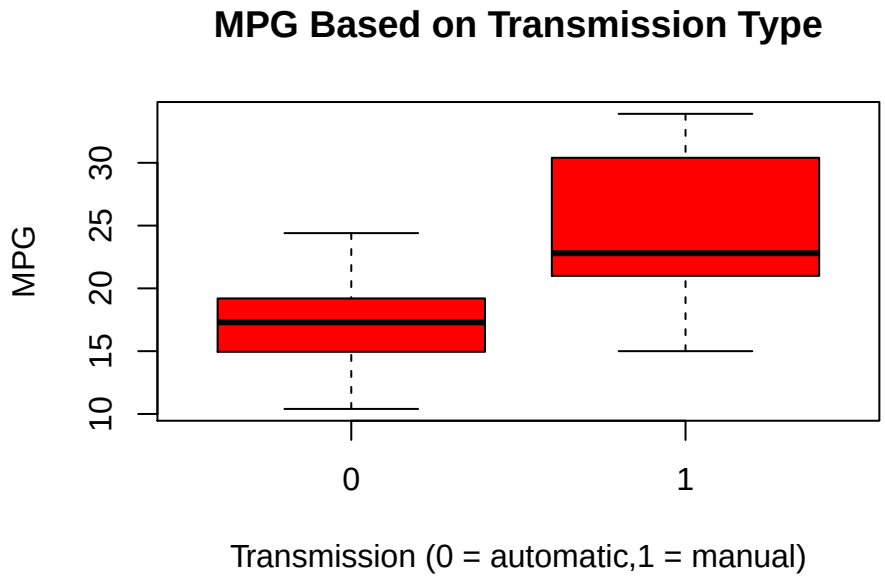
As you can see, there is only a slight overlap in the confidence interval values of the means between manual and automatic transmission (the overlap is small enough we don't need to explore it any further to make sure it is insignificant as it obviously is). Therefore, we can view the model as statistically significant.

Conclusion

We are able to conclude that the type of transmission does have an effect on mpg when including horsepower as a confounding variable. MPG is on average 5.3 mpg higher with a manual transmission than an automatic transmission.

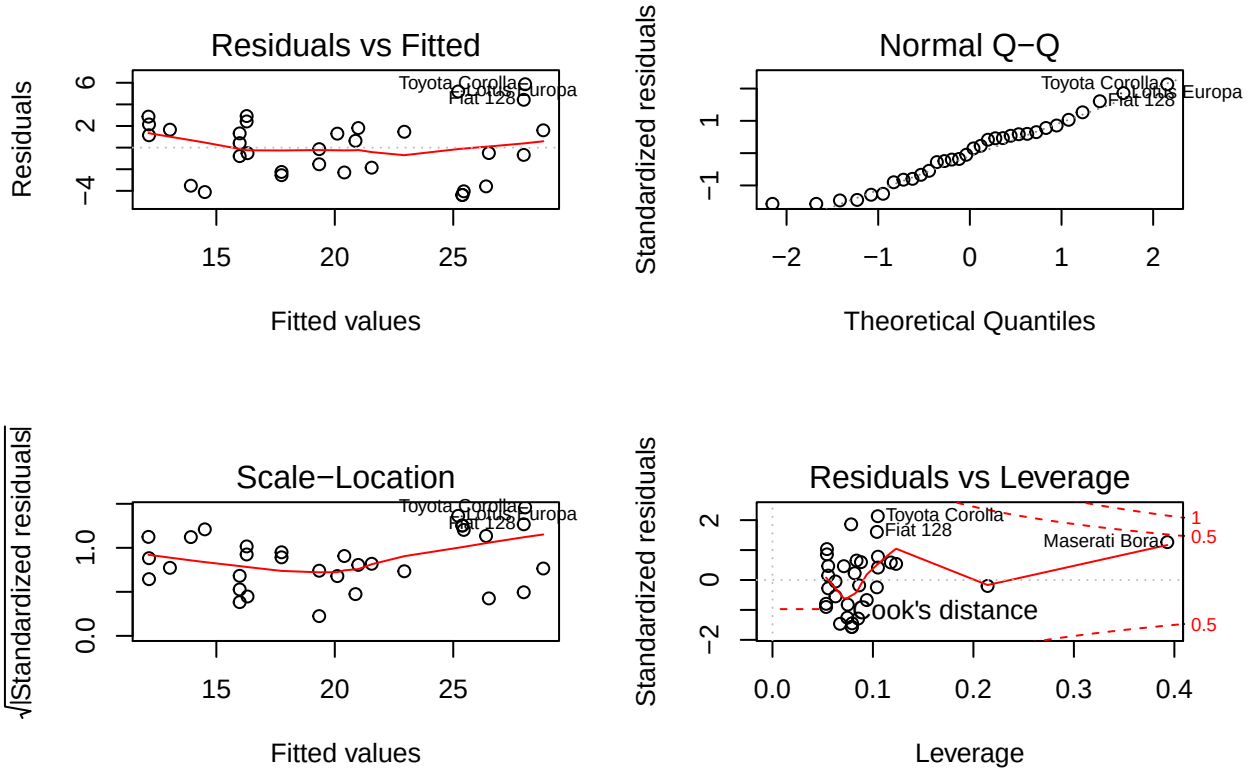
Appendix

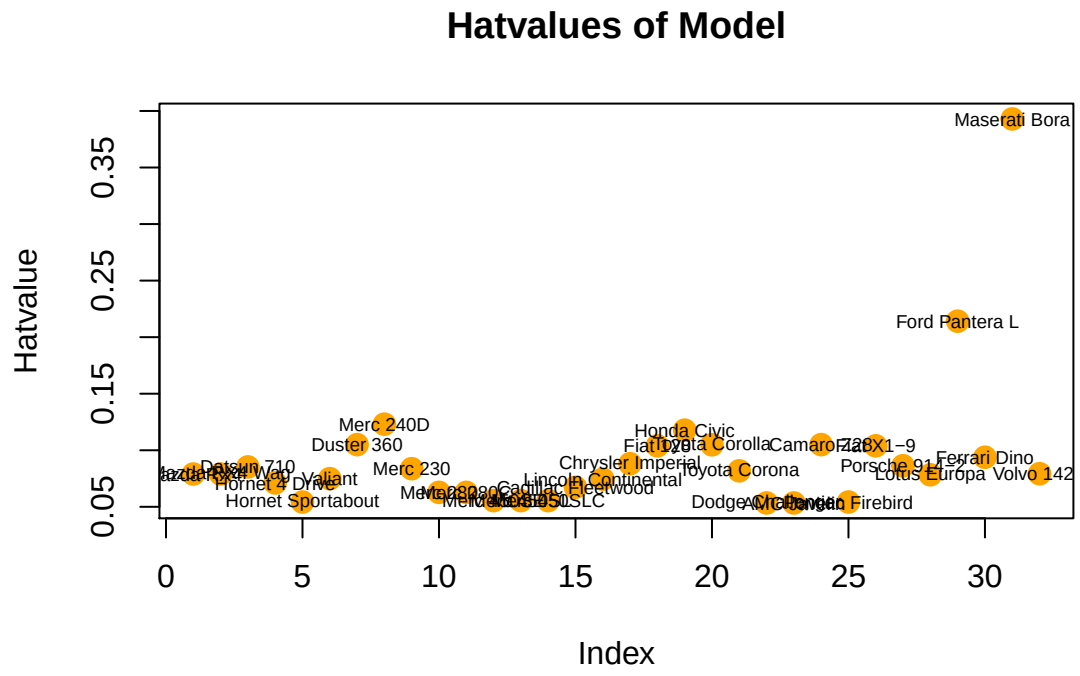
Exploratory Data Analysis

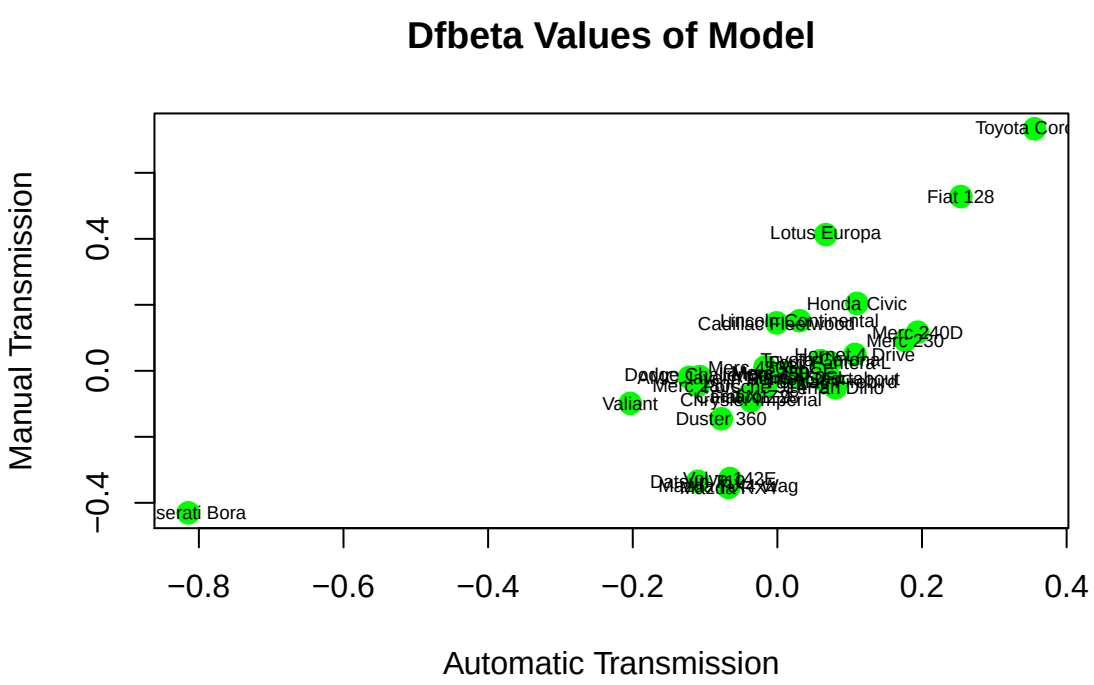



```
##             mpg        cyl       disp         hp        drat         wt
## mpg   1.0000000 -0.8521620 -0.8475514 -0.7761684  0.68117191 -0.8676594
## cyl  -0.8521620  1.0000000  0.9020329  0.8324475 -0.69993811  0.7824958
## disp -0.8475514  0.9020329  1.0000000  0.7909486 -0.71021393  0.8879799
## hp   -0.7761684  0.8324475  0.7909486  1.0000000 -0.44875912  0.6587479
## drat  0.6811719 -0.6999381 -0.7102139 -0.4487591  1.00000000 -0.7124406
## wt   -0.8676594  0.7824958  0.8879799  0.6587479 -0.71244065  1.0000000
## qsec  0.4186840 -0.5912421 -0.4336979 -0.7082234  0.09120476 -0.1747159
## vs    0.6640389 -0.8108118 -0.7104159 -0.7230967  0.44027846 -0.5549157
## am    0.5998324 -0.5226070 -0.5912270 -0.2432043  0.71271113 -0.6924953
## gear  0.4802848 -0.4926866 -0.5555692 -0.1257043  0.69961013 -0.5832870
## carb -0.5509251  0.5269883  0.3949769  0.7498125 -0.09078980  0.4276059
##             qsec         vs          am       gear        carb
## mpg   0.41868403  0.6640389  0.59983243  0.4802848 -0.55092507
## cyl  -0.59124207 -0.8108118 -0.52260705 -0.4926866  0.52698829
## disp -0.43369788 -0.7104159 -0.59122704 -0.5555692  0.39497686
## hp   -0.70822339 -0.7230967 -0.24320426 -0.1257043  0.74981247
## drat  0.09120476  0.4402785  0.71271113  0.6996101 -0.09078980
## wt   -0.17471588 -0.5549157 -0.69249526 -0.5832870  0.42760594
## qsec  1.00000000  0.7445354 -0.22986086 -0.2126822 -0.65624923
## vs    0.74453544  1.0000000  0.16834512  0.2060233 -0.56960714
## am   -0.22986086  0.1683451  1.00000000  0.7940588  0.05753435
## gear -0.21268223  0.2060233  0.79405876  1.0000000  0.27407284
## carb -0.65624923 -0.5696071  0.05753435  0.2740728  1.00000000
```

Diagnostics, Residuals, and Confidence Intervals






Dfbeta Values of Model
Manual Transmission
Automatic Transmission
0.4
0.0
−0.4
−0.8
−0.6
−0.4
−0.2
0.0
0.2
0.4
Toyota Corc
Fiat 128
Lotus Europa
Honda Civic
Merc 240D
Merc 230
Drive
Valiant
Duster 360
serati Bora