Regression Analysis of MPG by Transmission Type

Justin Papreck

May 18, 2018

Executive Summary

It is advantageous to consider regression analysis to answer such questions as to which the type of transmission, automatic or manual, has better milage per gallon, and whether there are other factors must be considered when answering this question. In this analysis, it is initially observed that the manual transmission vehicles have higher milage per gallon than do the automatic transmission vehicles, however, when considering the influence of the other data collected, this may not be a meaningful conclusion. Following exploratory analysis it was determined that comparing the transmission alone as a predictor of milage per gallon was insufficient, and a subsequent model was created. This model included several other variables selected using an AIC stepwise regression algorithm. This model demonstrated a much better predictor for milage per gallon and was significanly better than the model just examining transmission. A subsequent residual analysis did not indicate any aberrant patterns, so this model seems to explain 84% of the variance from the data.

Data

The data analyzed were obtained from the 1974 Motor Trend magazine containing 11 aspects of automobile performance and design for 32 automobiles, acquired from the dataframe *mtcars*. The data included fuel efficiency (MPG), number of cylinders in the engine, displacement (cubic inches), gross horsepower, rear axle ratio, vehicle weight (lb/1000), quarter mile time (seconds), engine type (V or straight), transmission type (automatic or manual), number of forward gears, and number or carburetors. Since we are interested in the influence of the transmission type, the binary data for transmission (mtcars$am) was changed to a factor variable.

Exploratory Analysis

Figures shown in Appendix.

```
summary(fit)$r.squared  ## R-squared value for model
```

```
## [1] 0.3597989
```

```
summary(fit)$coef  ## Linear model coefficients for model
```

```
##                     Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)        17.147368   1.124603 15.247492 1.133983e-15
## TransmissionManual  7.244939   1.764422  4.106127 2.850207e-04
```

As can be seen in the first figure and from the data in the regression table, there seems to be a significantly higher fuel efficiencty, by 7.24 MPG in the Manual transmission than in the vehicles with the automatic transmissions. While the p-value falls below 0.05, suggesting that this finding is significant, the R-Squared value is 0.360, suggesting that only 36% of the variance is accounted for with this correlation. The second figure shows the a correlation matrix, highlighting the potential influence factors within other aspects of the data collected. The data here suggest that the cylinders, displacement, horsepower, rear-axle ratio, weight, engine type, and transmission type may all be influencial to the milage per gallon.

Statistical Inference

To analyze the nature of the relationship between transmission type and milage per gallon, we must consider the other variables that may be contributors to the milage per gallon.

1a. H_0: There is no relationship between the miles per gallon and the transmission.

2a. H_α: There is some relationship between the miles per gallon and the transmission.

The above data would suggest that we can reject the null hypothesis, hoewever, as mentioned, only 36% of the variance can be explained by this model. A second set of hypotheses must be established:

1b. H_0: There is no difference between the linear model with only tranmission type than the model containing other variables.

2b. H_α: There is a difference between the different models suggesting that the other variables influence the milage per gallon.

Linear Regression

Three models were tested: the original model comparing milage per gallon and transmission type, a model looking at all of the aforementioned influencial data, and a model that utilizes a step function to test each of the combinations of each of those variables. Ultimately, the step function output was tested against the original function for analysis of our second hypothesis.

The third model returned by the AIC Stepwise Algorithm was as follows:

```
## lm(formula = mpg ~ cyl + hp + wt + am, data = cars)
```

```
## (Intercept)        cyl6        cyl8          hp          wt         am1
## 33.70832390 -3.03134449 -2.16367532 -0.03210943 -2.49682942  1.80921138
```

In comparing the three models, there are notable differences in the R-squared values:

```
## [1] 0.3384589
```

```
## [1] 0.829248
```

```
## [1] 0.8400875
```

The adjusted R-squared for the original model can explain 33.8% of the variance in the data, however in the models including all variables and a selection of those variables both yielded R-squared values much closer to 1, explaining 82.9% and 84.0% of the variance respectively. To finally analyze the second hypothesis, an ANOVA test was used (See Appendix).

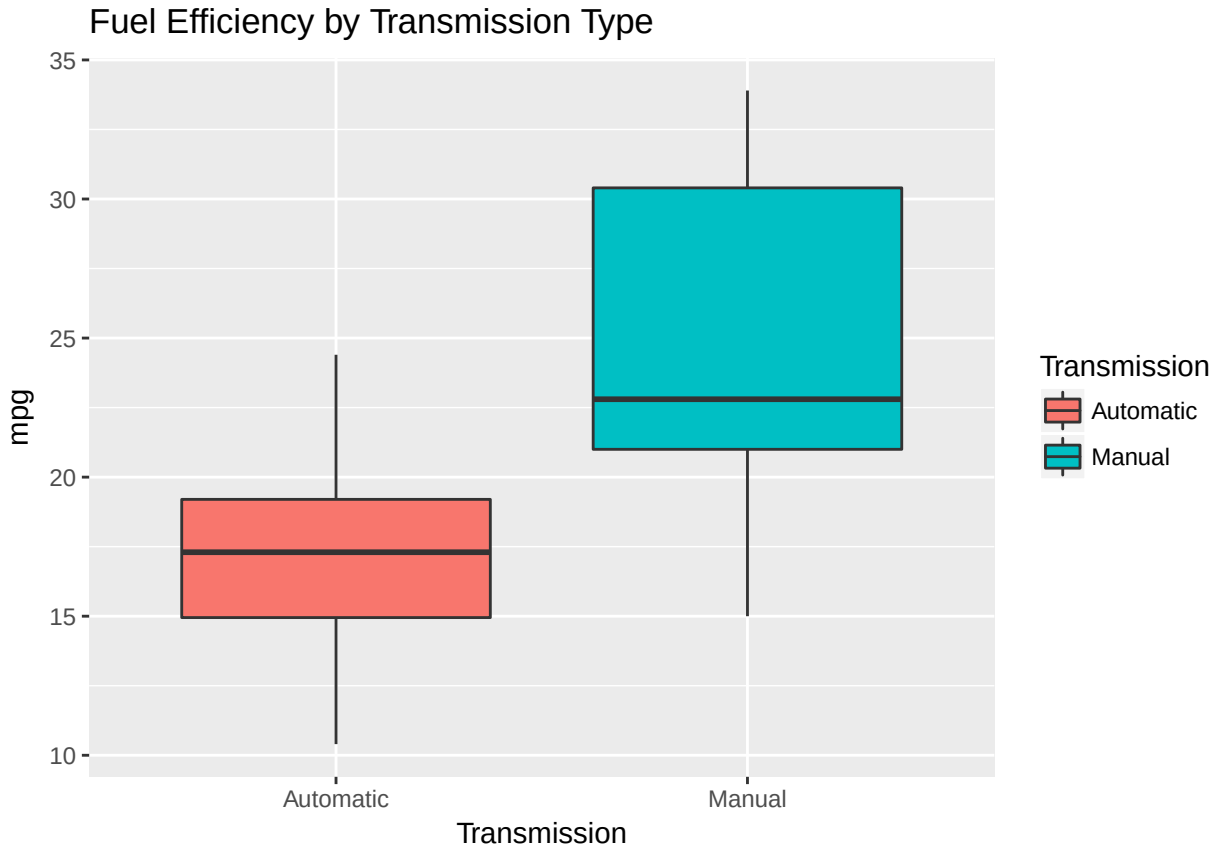
The ANOVA analsysis yielded an F-statistic of 24.5 and p-value on the order of 10^{-8}. The high F-statistic suggests that there is a difference between the two models, and the low p-value suggests that there is little uncertainty in this claim, thus we reject the second hypothesis and accept the alternative hypothesis, that there is a significant difference between the model only transmission and that testing the other variables with respect to milage per gallon.

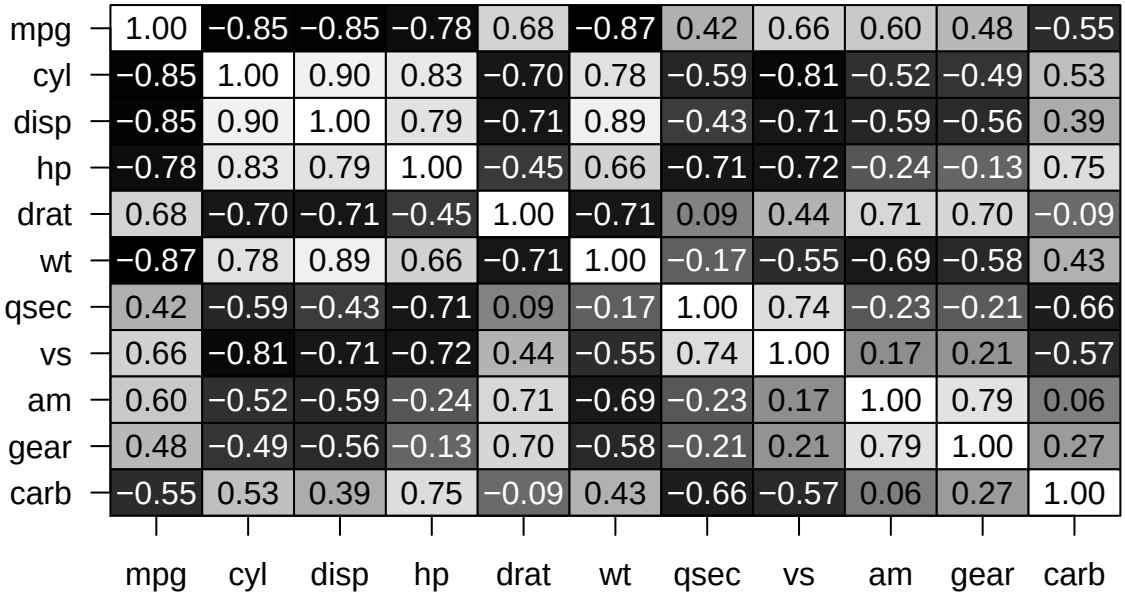
Residual Analysis

Finally, to check that there are no aberrant patterns emerging from the residuals, a plot of the final stepwise model was analyzed. The residuals do not have any visual discrepancies or emerging patterns. The residuals versus leverage all fall within Cook's distance suggesting that none of the values influence the regression fit.

Appendix

Exploratory Analysis Figures





ANOVA Analysis of Original Model and Stepwise Regression Model

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ cyl + hp + wt + am
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     30 720.90
## 2     26 151.03  4    569.87 24.527 1.688e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Residual Analysis