Transmission Types and MPG

BKC

May 11, 2015

### Executive Summary

### Analysis

Exploratory Analysis: Since we are specifically interested in how transmission type affects gas mileage, I used a boxplot (Figure 1) and a histogram (Figure 2) for exploratory analysis. Figure 1 shows us that there is a clear difference between the median and IQR for transmission type. Automatic transmissions appear to yield lower miles per gallon than manual transmissions. Figure 2 shows us that the distribution for automatic transmission and mpg is more clearly unimodal than the distribution for manual transmission.

Model Selection: Next, I determined which variables to include to get the strongest model. The data set available to answer this question has 11 variables, including miles per gallon (mpg). I performed stepwise model selection, specifically: backwards elimination using adjusted R2 and backwards elimination using p-value.

Backwards Elimination with Adjusted R2: During this process, I started with a model that included all variables -- mpg as the reponse and the other 10 as predictors. I recorded the R2 for this initial model (0.8066423) and then dropped one variable at a time and compared the adjusted R2 of each smaller model. Of these smaller models, I chose the model with the highest increase in adjusted R2 and then repeated the process, dropping one variable at a time, until none of the smaller models resulted in an increase.

Furthermore, I went through this process twice. Since transmission type (am) is of specific interest, I first performed this backwards elimination by keeping the "am" variable in the model. I then repeated the entire backwards elimination, allowing the "am" variable to be a candidate for removal. Both processes produced the same model with the strongest R2 (0.8375334), and the "am" variable was included in both models. This gave me some reassurance that transmission type is a significant varible. The resulting model: final <- lm(mpg~disp+hp+wt+qsec+am,mtcars)

Backwards Elimination with P-values: I was also interested to understand how using p-value might change the resulting model. Similar to the process above, I started with a model that included all variables, examined the p-values for each slope and dropped the variable with the highest p-value. I then repeated this process with the smaller model, until all variables reflected p-values of significance (less than .05). This process resulted in a model that, again, includes the "am" variable, but did not match the model resulting from R2 elimination. The resulting model: fit<-lm(mpg~am+wt+qsec,mtcars)

Model Selection Decision: When deciding between the model from R2 elimination versus p-value elimination, I chose to use R2 elimination model. The p-value approach is useful to determine which predictors are statistically significant, but adjusted R2 is used for more reliable predictions.

Diagnostics: Next, I performed some diagnostics against the model I had selected. First I looked for any linear relationships that might exist between my response variable (mpg) and the other numerical variables in the model. For the model to be valid, there should be a linear relationship. This is checked by plotting the residuals against each predictor variable and confirming there is random scatter around 0. Figure 3-5 shows the residuals plotted against each of the three numerical variables in the model (hp, wt, qsec) and each reflects the random scatter we would hope to see. Finally, the Normal Q-Q Plot confirms that the residuals are, for the most part, normally distributed with the largest deviations only at the tails.

Final Analysis: After determining a model and performing diagnostics to confirm its validity, I examined the model's coefficients to try and quantify the mpg difference between transmission types.

final <- lm(mpg~am+disp+hp+wt+qsec,mtcars)  
summary(final)

##   
## Call:  
## lm(formula = mpg ~ am + disp + hp + wt + qsec, data = mtcars)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.5399 -1.7398 -0.3196 1.1676 4.5534   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 14.36190 9.74079 1.474 0.15238   
## am 3.47045 1.48578 2.336 0.02749 \*   
## disp 0.01124 0.01060 1.060 0.29897   
## hp -0.02117 0.01450 -1.460 0.15639   
## wt -4.08433 1.19410 -3.420 0.00208 \*\*  
## qsec 1.00690 0.47543 2.118 0.04391 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.429 on 26 degrees of freedom  
## Multiple R-squared: 0.8637, Adjusted R-squared: 0.8375   
## F-statistic: 32.96 on 5 and 26 DF, p-value: 1.844e-10

The coefficient for am is interpreted as the increase or decrease in the mean comparing those in the group to those not.

“Is an automatic or manual transmission better for MPG” "Quantify the MPG difference between automatic and manual transmissions"

Did the student interpret the coefficients correctly? Did the student do some exploratory data analyses? Did the student fit multiple models and detail their strategy for model selection? Did the student answer the questions of interest or detail why the question(s) is (are) not answerable? Did the student do a residual plot and some diagnostics? Did the student quantify the uncertainty in their conclusions and/or perform an inference correctly? Was the report brief (about 2 pages long) for the main body of the report and no longer than 5 with supporting appendix of figures? Did the report include an executive summary? Was the report done in Rmd (knitr)?

### Appendix



