Derek Albosta

Professor Dove

MATH 260

11/29/18

**Modeling MPG Based on Various Vehicle Specifications**

**Abstract:**

*This study examines a data set from the 1974 American Car Magazine Motor Trend on various car data they collected during trial tests. I was interested in examining how car features can impact the efficiency of gasoline usage. I Looked at 4 different models: a simple linear regression with weight, a multiple linear regression with horsepower and weight, a multiple linear regression with weight horsepower and engine type as a categorical variable, and a multiple linear regression with engine type and weight with horsepower as an interaction term.*

*Based on model testing, the best model was the multiple linear regression with engine type as a categorical variable and horsepower with weight as an interaction term. The model had the lowest AIC value of 146.5263 and lowest f-statistic p-value of 1.764e-12.*

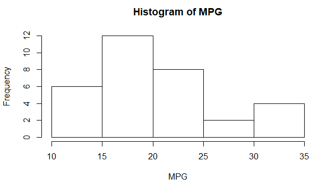
**Introduction:**

*Motor Trend*, a popular American car enthusiast magazine, tested various cars and collected data on each car’s individual performance and specifications. Variables for the dataset were obtained by testing cars on a professional race track with a professional driver, where and representative from *Motor Trend* would record their observations. This dataset was generated in 1974.

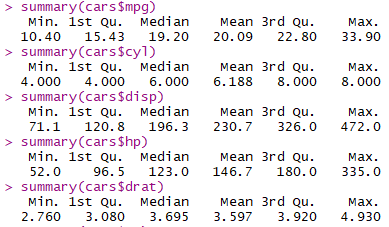
**Data:**

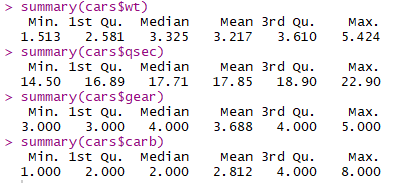
This dataset was collected from ETH Zurich’s website at the url <https://stat.ethz.ch/R-manual/R-devel/library/datasets/html/mtcars.html> and contains a total of 32 observations. There are 11 Variables from the data set: miles per gallon, number of engine cylinders, displacement, horsepower, rear axle ratio, weight, ¼ mile time, engine type, transmission, number of forward gears, and number of carburetors. I looked at 3 variables for testing : engine type, horsepower, and weight. The engine type is the only categorical variable where the variable type is indicated by a 0 or 1. For engine type (vs) 0 is a v-shaped engine while 1 is a straight engine. Horsepower is the gross horsepower of a given vehicle and weight is measured in a scale of 1000lb.

**Analysis of response variable:**

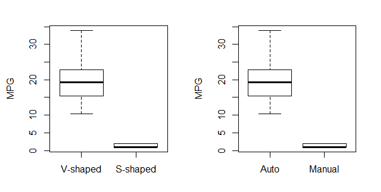


The above histogram of miles per gallon (MPG) values shows a pretty decent normal distribution with a slight skew to the right. The median of the variable is 19.20 and the range of MPG is 10.40 to 33.90.

Below is the statistical summaries for all numeric variables:



Below is a boxplot summary of our categorical variables:



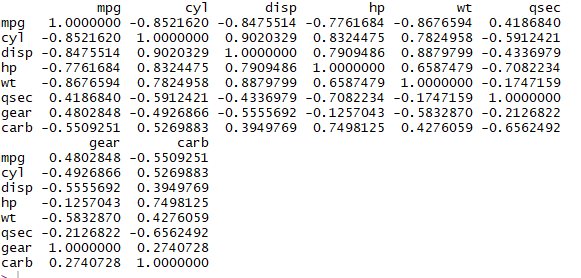
We can see they plots are exactly correlated, so we will throw one of these categorical variables out later.

**Model Selection and Diagnostics:**

Below is a scatterplot matrix of all our variables:



In the above scatterplot matrix, we can see several strong linear relationships between some of our variables. Displacement and horsepower seem to have a strong correlation. Weight and displacement have a strong relationship as well. The correlation matrix is as follows:



In the above correlation matrix, we can see a better picture of the linear relationships between our variables. Displacement has a very strong linear relationship with number of cylinders. Also horsepower has a pretty strong correlation with number of cylinders. This makes sense since more pistons generally mean an engine is more powerful. So out of these 3 variables we will keep horsepower for our model. Although weight has a decent correlation with horsepower, it is strongly correlated to our response variable, so this will also be kept. Quarter mile time, gear, and number of carburetors all have a decent correlation with our response variable but they are significantly weaker than the correlation of our other variables, so we will leave them out of our model.

In total, we are looking at two numeric variables: horsepower and weight; Considering the similarities between our boxplot of categorical variables, we are only keeping one of them for our model. I have selected engine type to be the categorical variable.

**Models:**

*Model 1: mpg by weight*

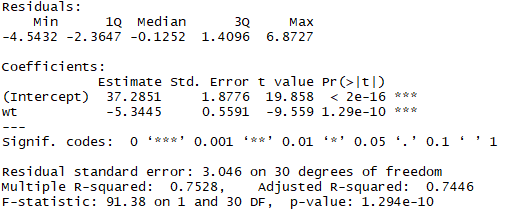
For my first model, I used a simple linear regression the variable most correlated to our response variable (mpg), weight(wt). Since it has the strongest correlation, it inherently will produce the best model for predicting mpg with a simple linear regression.

Below is the AIC output of the model:



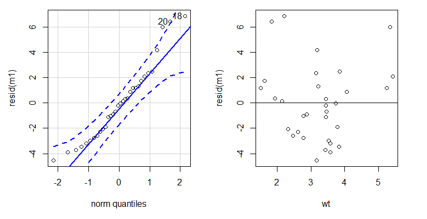
Since we have no other AIC’s to compare to, this is the best model.

Below is the summary output the our model:



Based on the summary, we have a good f-statistic p-value at 1.294e-10 which is significantly smaller than 0.05 indicating that we can reject the null hypothesis and this model is statistically significant. The r-squared value is 0.7528 which is fairly high indicating a good relationship between these variables.

SVAs:



Our normal quantile plot is good besides a slight departure in the upper quartile, therefore we are not violating normality. Our residuals show equal variance in the center, but seems to fall off in the lower and upper ranges of weight.

*Model 2: mpg by weight and horsepower*

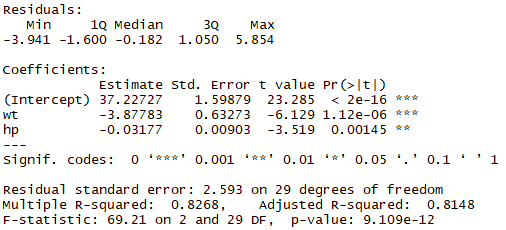
For my second model, I used a multiple linear regression with both weight and horsepower.

Below is the AIC output of the model:

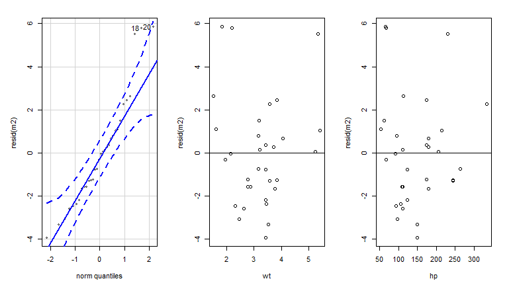


Our AIC for model 2 is an improvement from our first model significantly, providing evidence that this is a better model.

Below is the summary output for the model:



Based on the summary, we have a good f-statistic p-value at 1.294e-12 which is significantly smaller than 0.05 indicating that we can reject the null hypothesis and this model is statistically significant. This p-value is also smaller than our first model, indicating an improvement. The r-squared value is 0.8268 which is also an improvement from our first model.

SVAs:

Our normal quantile plot is good besides a slight departure in the upper quartile, still showing a similar result from model 1. The variance of our residuals seem the same for model one for weight, but the variance for horsepower shows improvement in the lower ranges. The variance for horsepower get skewed in the upper range.

*Model 3: mpg by weight and horsepower on engine type*

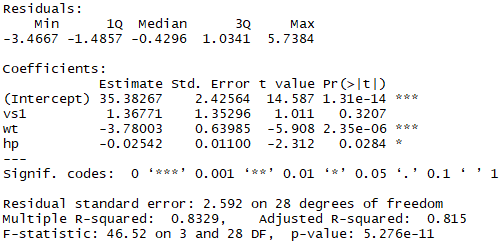
For my third model, I used a multiple linear regression with both weight horsepower by engine type.

Below is the AIC output of the model:



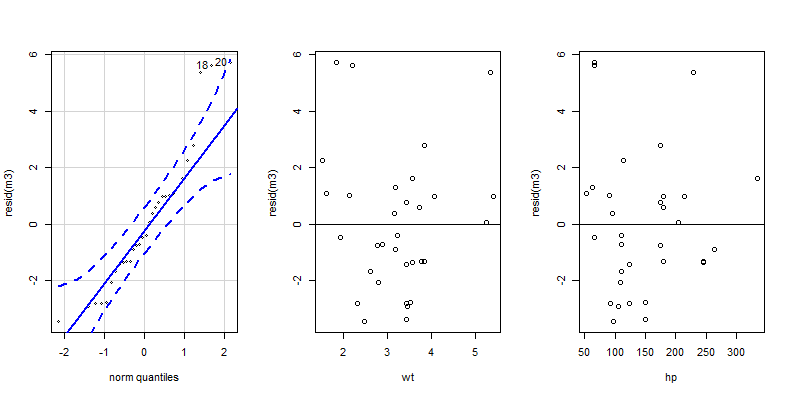
The AIC for model 3 indicates that this model is better than our model 1, but very slightly worse than our model 2

Below is the summary output of the model:



Based on the summary, we have a good f-statistic p-value at 5.265e-12 which is significantly smaller than 0.05 indicating that we can reject the null hypothesis, however this p-value is still larger than model 2’s. The r-squared value is 0.8329 which is a slight improvement from model 2’s r-squared value.

SVA’s:



Our normal quantile plot is good besides a slight departure in the upper quartile, still showing a similar result from our previous models. However the variance for our numeric variables look better. There is a slight improvement for the residuals in the lower range of weight and the variance for horsepower looks overall tighter than model 2’s

*Model 4: mpg by horsepower and weight as interaction terms on engine type*

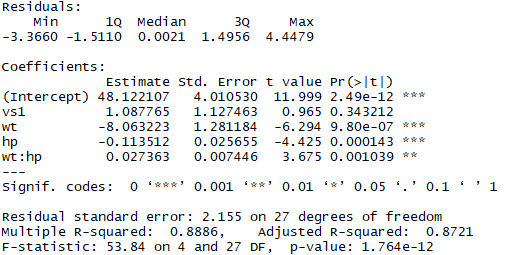
For my fourth model, I used a multiple linear regression with both weight and horsepower as interaction terms by engine type.

Below is the AIC output of the model:



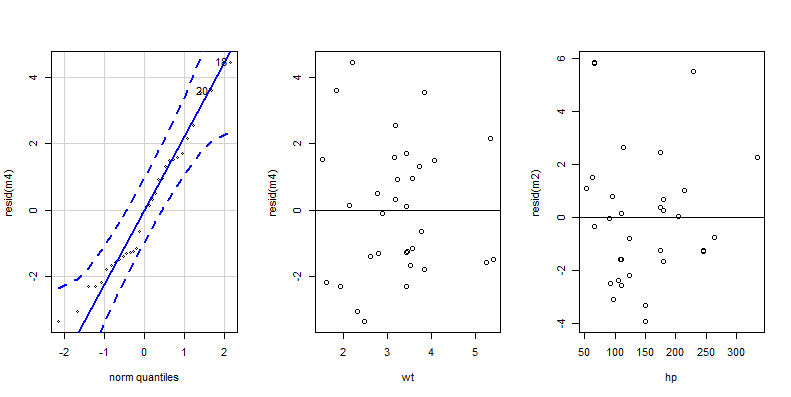
The AIC for model 4 indicates that this model is the best compared to our other models, being 10 less than model 2’s AIC value.

Below is the summary output of the model:



Based on the summary, we have a good f-statistic p-value at 1.764e-12 which is significantly smaller than 0.05 indicating that we can reject the null hypothesis. This is also the smallest p-value out of all of our models. The r-squared value is 0.8886 which is a slight improvement from model 3’s r-squared value and is the highest r-squared compared to all models.

SVA’s:



Our normal quantile plot perfect as all values are inside the normal lines. This is the only model with no questions of violating normality. Variance for weight against the residuals are better than compared to our previous models in the upper and lower ranges. Variance for the horsepower residuals still seem to diverge in the upper ranges.

Based on the fact that model 4 has a perfect quantile plot, the lowest AIC value, the lowest f-statistic p-value and best variance’s for the residuals. This is the model that will be used for predictions.

**Summary of Statistical Findings**

The fitted model equation for our chosen model is:

\