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.*

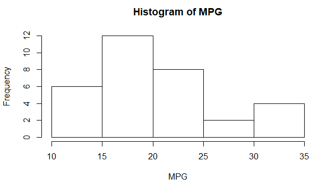
**Introduction:**

*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.

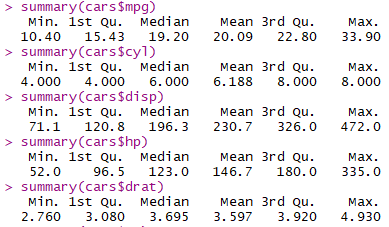
**Data:**

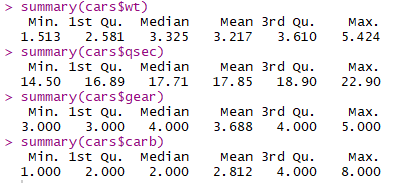
This dataset was collected from ETH Zurich’s website at the <https://stat.ethz.ch/R-manual/R-devel/library/datasets/html/mtcars.html> and contains a total of 32 observations. There are 11 Variables from the data set: miles per gallon, number of engine cylinders, displacement, horsepower, rear axle ratio, 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:**

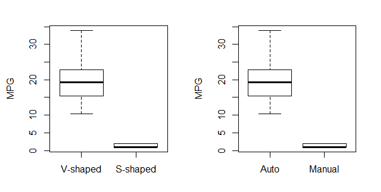


The above histogram of miles per gallon (MPG) values shows a pretty decent normal distribution with a slight skew to the right. The median of the variable is 19.20 and the range of MPG is 10.40 to 33.90.

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.

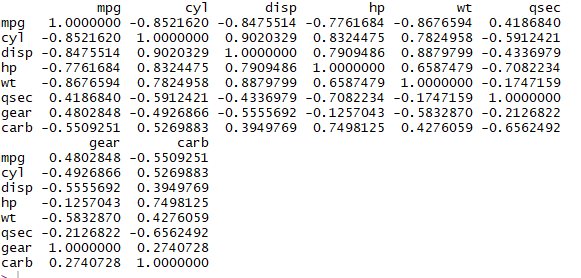
**Model Selection and Diagnostics:**

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*

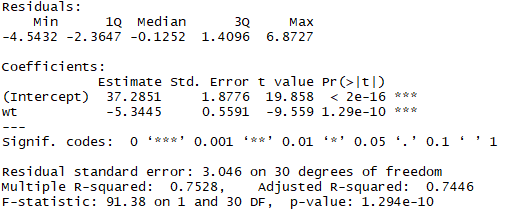
For my first model, I used a simple linear regression the variable most correlated to our response variable (mpg), weight(wt). Since it has the strongest correlation, it inherently will produce the best model for predicting mpg with a simple linear regression.

Below is the AIC output of the model:



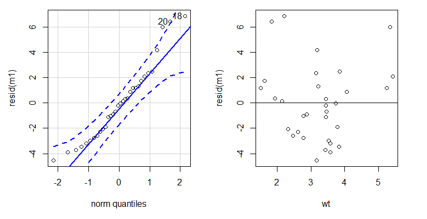
Since we have no other AIC’s to compare to, this is the best model.

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. There seems to be a nonlinear pattern forming for the residuals though, which violates SVA 3.

*Model 2: mpg by weight and horsepower*

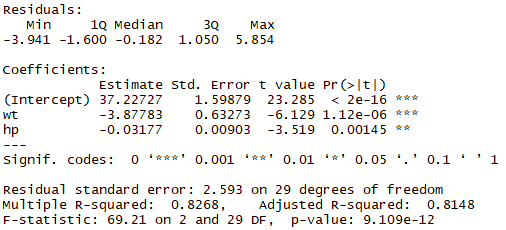
For my second model, I used a multiple linear regression with both weight and horsepower.

Below is the AIC output of the model:

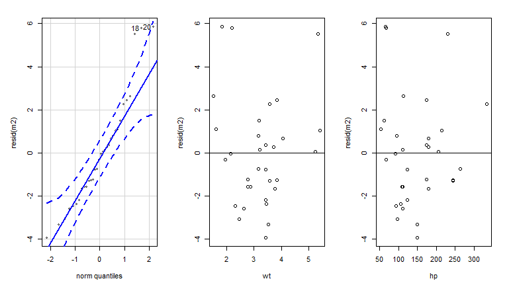


Our AIC for model 2 is an improvement from our first model significantly, providing evidence that this is a better model.

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. There seems to be a slight linear positive linear correlation with both residuals.

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

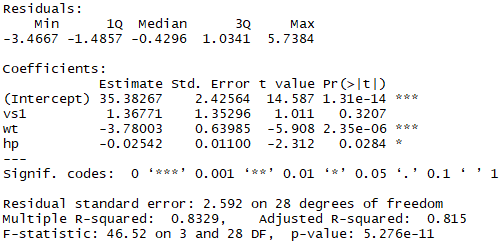
For my third model, I used a multiple linear regression with both weight horsepower by engine type.

Below is the AIC output of the model:



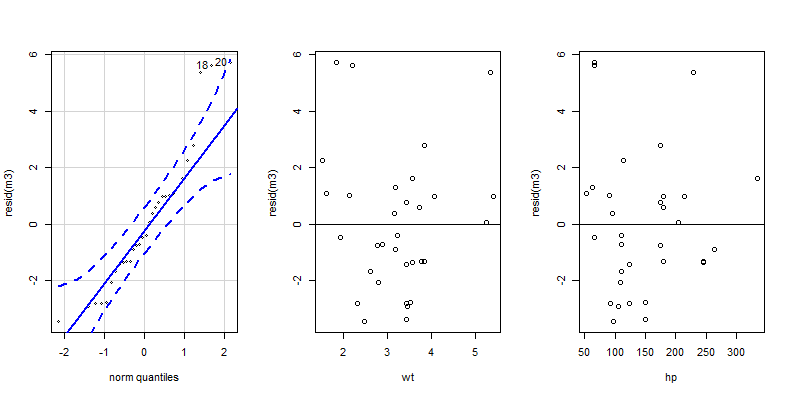
The AIC for model 3 indicates that this model is better than our model 1, but very slightly worse than our model 2

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. There again seems to be a slight linear positive linear correlation with both residuals.

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

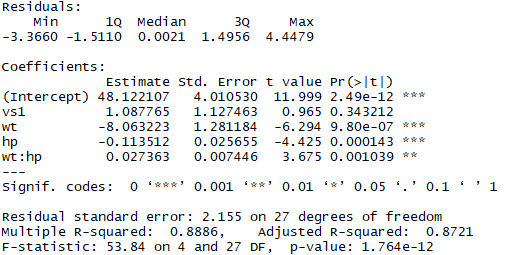
For my fourth model, I used a multiple linear regression with both weight and horsepower as interaction terms by engine type.

Below is the AIC output of the model:



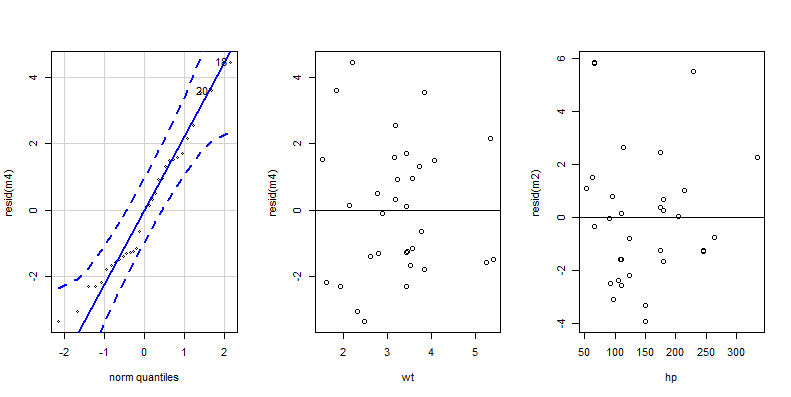
The AIC for model 4 indicates that this model is the best compared to our other models, being 10 less than model 2’s AIC value.

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. There is also still a slight positive linear correlation for both residuals.

Model Selection:

The first model (the simple linear regression of mpg on weight) did have a statistically significant f-statistic p-value, indicating that the model was useful. However, it had the highest AIC value, indicating that it was the weakest of our models. It slightly violates normality in the upper quartile but fails to meet the second and third SVA as there seems to be no linear relationship between the residuals and no equal variance.

The second model (multiple linear regression of mpg against weight and horsepower) also had a statistically significant f-statistic p-value. All slope coefficient p-values were statistically significant. The AIC value showed a significant improvement form our first model. It also slightly violates normality in the upper quartile but violates SVA 2 with the residuals in the upper and lower ranges of both numeric variables. There seems to be a weak positive linear relationship between the residuals.

The third model (multiple linear regression of mpg against weight and horsepower by engine type) again had a statistically significant f-statistic p-value. All slope coefficient p-values were statistically significant. The AIC value increased slightly from model 2, indicating that it is slightly weaker of a model. It again also slightly violates normality in the upper quartile but violates SVA 2 with the residuals in the upper and lower ranges of weight. The variance of the residuals improve in the lower range for this model. There seems to be a weak positive linear relationship between the residuals.

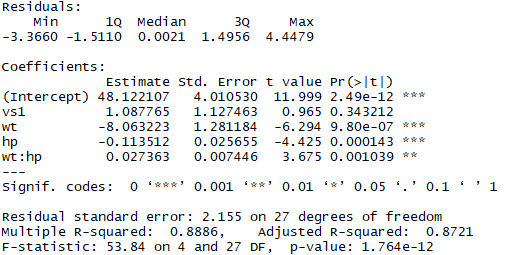
The fourth model (multiple linear regression of mpg by engine type with weight and horsepower and interaction terms) had the best f-statistic p-value. All coefficient p-values from the model summary were statistically significant except for the p-value for S-type engines, however the p-value is still small. The AIC value for this model is the smallest, indicating this may be the best model. This model completely satisfies normality given the quantile plot, unlike any of the other models. The issue with SVA 2 and 3 remain the same; variance seems to skew in the upper and lower ranges for both variables and there seems to be a weak positive correlation.

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 model in r is defined as:

Again the summary statistics for our model are:

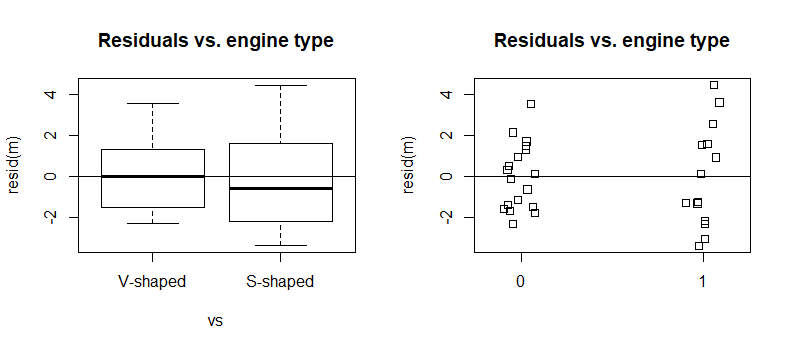


The fitted model equation for our chosen model is:

Summary Interpretation:

Based on the model output the predicted mean mpg is 48.12 when weight and horsepower equal 0 and the engine type is a V-shape. A p-value of 2.49e-12 is small enough to reject the null hypothesis that the slope is zero. If the engine is an S-type, the mpg increases by 1.08 however, the p-value is 0.34 which is >0.05, meaning it may be difficult to reject the null-hypothesis that the slope coefficient is equal to 0, but the value is still relatively small. When weight increases by 1 unit, mpg is decreased by -8.06. The p-value for weight is 9.80e-7 which is significant enough to reject the null hypothesis that the coefficient for weight is 0. This makes sense since added weight requires more energy to move a given car, thus this would make a car less fuel efficient. When horsepower increases by 1 unit, mpg is decreased by 0.11 and this is significant sine the p-value is 0.000143 which is small enough to reject the null hypothesis that the coefficient for horsepower is 0. Generally, the more horsepower an engine has, its is more geared towards trying to maximize performance over fuel efficiency. An interaction term if 0.03 is multiplied to both weight and horsepower; this is significant since the p-value is 0.001039 which is small enough to reject the null hypothesis that the coefficient for this term is 0.

Below is a more detailed analysis of this models SVAs by engine type:



From our model testing, we already know that this model does not violate normality. From the quantile plot, we saw that all of the residuals fit into the normal line. Also from the model testing we know that our models residuals for the numeric variables do not have equal variance in the upper and lower ranges, but have some slight positive linear correlation. From looking at the residuals by engine type, We can see that variance for the residuals of V-shaped engines are for the most part evenly distributed except. However, S-shaped residuals are skewed to the right, indicating that our model may have more trouble predicting mpg values for S-shaped engines.

Predictions:

I used the model to predict mpg for 4 separate cars. The results of this test are below compared to the actual values from the dataset:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Test** | **hp** | **wt** | **vs** | **Actual mpg** | **Confidence interval** | **Prediction interval** |
| 1 | 110 | 2.62 | 0 | 21 | fit lwr upr 22.39613 20.62237 24.16989 | fit lwr upr  22.39613 17.63114 27.16112 |
| 2 | 245 | 3.57 | 0 | 14.3 | fit lwr upr  15.45889 13.93847 16.97931 | fit lwr upr  15.45889 10.78229 20.13549 |
| 3 | 93 | 2.32 | 1 | 22.8 | fit lwr upr  25.8504 24.61551 27.08529 | fit lwr upr  25.8504 21.25868 30.44212 |
| 4 | 66 | 2.2 | 1 | 32.4 | fit lwr upr  27.95208 26.51015 29.39401 | fit lwr upr  27.95208 23.30041 32.60376 |

Based on this test. We can see that predictions for V-shaped engines are fairly accurate. Both mean predictions for V-shaped engines are within 1.5 units away from the actual value. Also, the actual mpg value for the given test is within the 95% confidence interval. However, our predictions for S-shaped engines are significantly worse. Test 3’s mean prediction value was off by 3.04 units and test 4’s was off by 4.45 units. Also the 95% confidence interval does not contain the actual mpg value within its range. This is probably explained by the skew of residuals of S-type engines we observed in our SVA analysis.

Conclusion:

We know that the selected model we used a good indicator for predicting mpg given engine type, horsepower, and weight. We know from the model summary output we have a good f-statistic p-value at 1.764e-12 which is significant enough to reject null hypothesis that the model is not useful. Also This is also the smallest p-value out of all of our models. The r-squared value is 0.8886 which shows that the data is closely fitted to the regression line. The model does not violate normality, however it demonstrates some issues with variance of the residuals in the upper and lower ranges of the numeric variables. It also only have a faint linear association with the residuals. Based on the residuals by engine type, it is clear that there is a skew to residuals for S-shaped engines, indicating that predictions for S-shaped engines may be problematic. During predictions, we saw that the model was great at predicting mpg for V-shaped engines, but predictions for S-shaped engines were far off from the actual mpg values. Our S-shaped confidence interval prediction did not even include the actual mpg within the 95% confidence interval. This means our model is really only useful for accurately predicting mpg for V-shaped engines. To generalize would mean we would expect to see the same association in the larger population that this sample represents. Generalization is not justified here. This dataset was taken in 1974; Technology has changed drastically and some of these factors may influence MPG differently than before. One notable example is transmission type, which was left out. Modern day automatic transmissions tend to get way better MPG than manual counterparts even though our testing said otherwise. So in this sample of 33 cars, there is a statistically significant association between MPG (response) and horsepower, weight, and engine type (explanatory variables). We cannot infer cause and effect since this data is using older vehicles and technology has changed so much since the test date.

**Bibliography:**

https://stat.ethz.ch/R-manual/R-devel/library/datasets/html/mtcars.html