# MAT 303 Module One Problem Set Report

Multiple Regression

Zachary Tankersley

Zachary.tankersley@snhu.edu

Southern New Hampshire University

## Introduction

The data set being explored is comprised of data referencing the fuel efficiency and other performance metrics of several models of vehicles. It is stored in a .csv formatted file. The results will be analyzed to study fuel economy and predict fuel economy. The results could be used to help make better business and design/engineering decisions that may hinge on the fuel efficiency of a car manufacturer’s products, or the products produced by competitors. The results could also be used to steer marketing campaigns and drive sales. A first-order regression model will be built to study the relationships between weight, horsepower, and fuel efficiency.

## Data Preparation

There are three variables in this data set that will be of concern: the MPG, weight, and horsepower of the vehicles. All other variables can be considered extraneous. Miles per Gallon will be the dependent variable, with weight and horsepower being the independent variables. In the source data set, there are 33 rows and 12 columns, however, the first row contains the row header titles so in practice there are only 32 rows and 12 columns.

## Multiple Regression Model

### Correlation Analysis

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

Side by side here you can see the scatterplot of fuel economy compared against weight, and the scatterplot of fuel economy against horsepower respectively. In both cases, and as is expected, the correlation is very much negative in both cases. As one who is familiar with the laws of conservation of energy may assume, the more weight or horsepower one adds to a car, the more inefficient the vehicle becomes. If you look at the lightest cars with the lowest horsepower, they tend to be more highly efficient. The Pearson Correlation Coefficient tests run on these two relationships indicates a -0.87 and -0.78 correlation respectively and would confirm a strong negative correlation.

### Reporting Results

The general form of the multiple regression model using MPG as the dependent variable and weight and horsepower as dependent variables is as follows:

**MPG = β0 + (β1\*Weight) + (β2\*Horsepower)**

The actual model equation after running a summary analysis on the formula above with proper intercepts and coefficients (rounded to 4 decimal places) is:

**MPG = 37.2273 – (3.8778\*Weight) – (0.0318\*Horsepower)**

The R-Squaredvalue is 0.8268 and the *Adjusted R-Squared* value is 0.8148 which hints at a fairly decent fit for this model when applied to this data set. All things taken into consideration, and only using two independent variables, a fit of 82/82% seems fairly high. The beta coefficients in this case being roughly -3.8778, and -0.0317 respectively, both hint at the negative influences that weight and horsepower have on a vehicle, with weight being a much more negative influence than horsepower. This would certainly fit with real-world examples; a dump truck gets horrible gas mileage, while a very light, high-horsepower car can still get fairly decent mileage. In regression modeling, fitted values are the *predicted* values given by the regression model when you input the data points in the given data set. Residual values are the differences between the actual data points and the fitted values. When comparing residuals and fitted values, you can see that conditions for homoscedasticity are met, with very little clustering. When analyzing a quantile-quantile plot for normalcy you can see a very tight grouping along the distribution line up until about the 1.5 mark or 95th percentile where you can start seeing potential outliers. Taking all of these factors into consideration, these results seem to be “normal” and expected, especially when compared to real-world conditions.

### Evaluating Model Significance

When looking at the summary analysis of the model itself, the P-value of 9.109e-12 is very much lower than the 5%level of significance which would cause the rejection of the null hypothesis in favor of the alternative hypothesis stating that a *significant* relationship *does* exist between fuel efficiency and one of the independent variables (weight/horsepower). When examining the weight and horsepower variables using individual T-tests, we come up with 1.12e-06 and 0.00145 respectively, which are both less than our 0.05 level of significance, again causing rejection of the null hypotheses in favor of the alternative hypotheses. This suggests that both weight *and* horsepower have a statistically significant relationship with fuel efficiency. When creating a 95% confidence interval for weight and horsepower we are shown values of -2.8027 and -0.0164 which again are both lower than the 5% significance level, causing rejection of the null hypotheses, suggesting that both horsepower and weight are statistically significantly related to fuel efficiency.

### Making Predictions Using the Model

The predicted fuel efficiency for a car that has a weight of 2.95 and a horsepower of 179 according to our model would be 20.0955mpg, giving a residual of +/-2.60441 when compared to the given actual value of 22.7mpg. The 95% prediction interval for this car would be 20.1003 and the 95% confidence interval would be the same. These two values are very close to the value of 20.0955 given by the model. With that said, the upper and lower bounds of both predictions vary from 14.645 and 25.5556 for prediction intervals, and 18.8249 and 21.3758 for confidence intervals. The prediction interval range seems to be wider than the confidence interval range possibly because it is more uncertain. You are *predicting* the range, whereas the confidence interval is based on the data itself and less susceptible to predictive variances.

## Conclusion

Assuming the sample size of 32 cars *is* indeed sufficiently large, I would recommend using this model for the majority of cases. If a vehicle manufacturer crams a high-horsepower/high-efficiency motor in a very light car you would get strange results that may not be accurate, but for the majority of vehicles made, this model would probably work fairly well. The predictions increase/and decrease based on weight and horsepower as one would expect in the real world. Once you get above the 95th percentile or *perhaps* below the 5th percentile, you start to see outliers in your results as should probably be expected in many models. The significance levels are much lower than the industry “norm” of 5% indicating that the null hypotheses in each case can be rejected. Concurrently, this indicates that the weight *and* horsepower independent variables do have a statistically significant influence and relationship with fuel efficiency (MPG). Furthermore, independent testing on both independent variables in comparison with fuel efficiency shows that both do have significant relationships on their own. Both of these attributes likely have some combined impact on fuel efficiency. These results can be used by a manufacturer to predict how fuel efficient a car might be in the design process if weight and horsepower are known. Or it could be used to establish a baseline fuel efficiency that a vehicle should have, and then used as an anomaly detection method. Practically, or maybe pragmatically speaking, the analysis says three things to me:

1. The more a vehicle weighs, the more fuel it will use
2. The more horsepower a vehicle has, the more fuel it will use.
3. We should probably all drive a Toyota Corolla.