# MAT 303 Module One Problem Set Report

Multiple Regression

Catherine Taft

catherine.taft@snhu.edu

Southern New Hampshire University

## Introduction

I’m a data analyst for a car company studying factors that may improve fuel economy for our vehicles, which is remarkable because I’ve never had a driver’s license and know very little about cars. The data set we have includes variables for MPG, number of cylinders, horsepower, axel ratio, weight, number of gears, and a few other things.

This is a fairly obvious thing to do – to consider what factors might influence something like fuel economy. That’s an important consideration when purchasing a car, given that gas is a huge expense, which is one of the reasons I opted to never own a car. So the value of this would be to examine the relationship between these predictor variables and the response variable of MPG, to see what adjustments could be made in order to make the car as fuel-efficient as possible.

We’ll be doing some visualizations at first and then running a multiple regression model to examine the relationships of these variables.

## Data Preparation

Being as I mentioned a person who doesn’t do cars, the variables that stood out to me are the number of cylinders, horsepower, and weight. These are things that even to a bike person such as myself stand out as being obviously influential on fuel efficiency. Cylinders are where the fuel combustion occurs that gives the car its power, so obviously more of those means more fuel consumption. Horsepower is the unit used to measure the power of the car, so high horsepower would indicate more fuel consumption. Weight will obviously affect fuel efficiency, as it simply takes more power to drag around a heavier object.

The output from the code where it prints the dataframe shows us that there’s 32 rows and 12 columns. The rows are organized by individual vehicles.

## Multiple Regression Model

### Correlation Analysis

First we created some scatterplots looking at the relationship between MPG and weight, and then MPG and horsepower. Common sense says we can expect a negative correlation between these variables but let’s look and make sure.

First the plot of MPG against weight:

Chart, scatter chart

Description automatically generated

Indeed here we can see a negative correlation (meaning as one variable increases, the other decreases) between these two variables. That makes sense because as I mentioned before, the heavier a thing gets, the more force is required to drag it around, and force with cars is generated through fuel consumption.

Now let’s look at the plot for MPG against horsepower:

Chart, scatter chart

Description automatically generated

Again we can see a negative correlation although this has quite a bit more variance. But in general that makes sense – horsepower measures power and power is derived from fuel consumption. You could imagine engineering decisions that could influence the extent to which a car gets X amount of horsepower at Y MPG that would lead to some variance here, whereas the weight thing is just basic physics.

The Pearson’s Correlation Coefficient is a number between -1 and 1 that indicates the strength and direction of a linear relationship between two variables. If a Pearson R of 0 indicates no relationship at all, -1 indicates a perfect negative correlation and 1 indicates a perfect positive correlation.

So we’ll run get the Pearson matrix for MPG, weight and horsepower to look at the strength of the correlation between these variables.

Table

Description automatically generated

We can see that the coefficient for MPG to weight is -0.8677 and the one for MPG to horsepower is -0.7762. That confirms what we saw in the scatterplots, which is that both have a negative correlation, and that weight is more tightly correlated to MPG than horsepower is.

### Reporting Results

The general equation for multiple regression is Y = β0 + β1X1 + β2X­­2­ +…+ βnX­­n which is basically an expanded form of the version of the slope-intercept form that’s used to calculate simple linear regression, which only features one predictor variable. β0 is the Y-intercept of the regression line basically, and β1 etc is the slope, and each X is a predictor variable. So for this scenario we’d use the following information:

Text, table

Description automatically generated

And our equation would end up being.

I’ll be honest, I’m not exactly sure what it means where it says “create a multiple regression model for MPG using weight and horsepower as independent variables” because I thought that’s what I just did, but maybe it means use specific numbers. So let’s pretend our car has a weight of 3500 pounds and 20 horsepower. I don’t know from good horsepower but let’s go with it. Our equation would then be: MPG = 37.22727 – 3.87783(3.50) -0.03177(20). This would come out to roughly 23MPG with those values. I assume that’s bad. However, we could use any other values we wanted to see how we could tweak this to wind up with a more satisfactory result.

The ­R-Squared value or the correlation coefficient represents the proportion of the variance that’s predictable with the given values. In other words, what percentage of the results fall along the regression line. Anything outside of that would be considered outliers. The adjusted R-Squared value is basically a modified version of that that has been adjusted for the number of predictor variables in the model. The higher the R-Squared value, the stronger the correlation, although in some cases a very high R-Squared can indicate overfitting. Here for our model the R-Squared value is 0.8268 and the Adjusted R-Squared is 0.8148 which indicates a strong correlation.

The beta values are the coefficients for the predictor variables, indicating that for every unit increase in those variables, the result would increase or decrease by that amount. So for weight, the coefficient is -3.878, so for every increase of 1,000lbs in the car, MPG would decrease (because it’s a negative value) by 3.878. The same logic follows for the horsepower coefficient, or any other.

A fitted value is the model’s prediction of the mean of the response variable. This is found by substituting the relevant number in as the X value in the regression equation. A residual is the difference between the actual value and the estimated one.

In the code we get the fitted values and residuals for the dataset which I won’t put here, but then we can make a plot of the residuals against the fitted values, as well as a QQ plot.Chart, scatter chart

Description automatically generatedChart, line chart, scatter chart

Description automatically generated

A model is homoscedactic if the residuals for all the variables have constant variance. The residuals plot meets the qualifications of homoscedacity because it is more or less randomly spread throughout the plot without any patterns or clustering. The QQ plot demonstrates a normal distribution because nearly all of the data, with the exception of a few significant outliers towards the top, lie across the regression line.

### Evaluating Model Significance

The overall F-test is used to determine if any of the variables in a multiple regression model are redundant or have a slope of 0, therefore not contributing to the model. The null hypothesis is that all of the slopes are 0 for every predictor variable. The alternative hypothesis is that at least one of the slopes does not equal 0 and therefore contributes meaning to the model. In math terms:

The p-value we got earlier (see chart above) which is 9.109^-12 which is basically 0. So at a 5% significance level, the p-value is obviously less than 0.05 so we can reject the null hypothesis that there is no linear relationship between any of the variables.

Now to determine which variables are significant to the model we can carry out individual F-tests. The null hypothesis is that the slope of the predictor variable is 0 and therefore not statistically significant, and the alternative hypothesis is that it is not 0 and therefore is statistically significant. Or:

In the same chart above we can see the p-values for each variable. For weight it’s 1.12^-6 and for horsepower it’s 0.00145. In both cases they are less than the 0.05 significance level, so we can reject the null hypothesis and say that these are both statistically significant variables.

Table

Description automatically generated

Here we have the 95% confidence intervals for those parameters: for weight [-5.1719, -2.5837] and for horsepower [-0.0502, -0.0133]. That means we can say that 95% of the time the mean value of the parameters will fall within those intervals.

### Making Predictions Using the Model

Using the same regression equation we used above to find MPG of that car with abysmal horsepower, we can now estimate the MPG for a car with a weight of 2,950 lbs (2.95) and a horsepower of 179. That would be MPG = 37.22727 – 3.87783(2.95) -0.03177(179) which evaluates to about 20.1MPG. It’s interesting that that gives worse fuel efficiency than the ridiculous model I gave above. I suppose very low horsepower just means much lower fuel consumption.

If a car actually gets an average of 22.7 MPG, the residual would be the difference between the actual value and the estimated value, so 22.7-20.1 = 2.6.

So now we can calculate the 95% prediction interval for the above value:

Table

Description automatically generated

That comes to [14.645, 25.5556] which means we can expect that 95% of the time, a car’s MPG will fall within those bounds if the weight and horsepower values are as specified in the equation.

The confidence interval at 95% is as follows:

Table

Description automatically generated with medium confidence

This tells us that 95% of the time the mean MPG of a group of cars will fall between [18.8249, 21.3758] if all the parameters are as specified in the equation.

The prediction interval is wider than the confidence interval because it has to inherently include more uncertainty.

## Conclusion

If the sample size were sufficiently large I would recommend this model (for lack of anything more detailed and complex, I suppose) because it allows for the adjustment of various variables to influence the target variable in a way that is meaningful and actionable.

The results of this particular model demonstrate that there is a distinct linear relationship between MPG and both horsepower and weight. If you increase weight or increase horsepower, the fuel efficiency of the car will go down. If you decrease them, it’ll go up. By adjusting the coefficients of the predictor variables you can see the influence on the response variable.

Doing this in a development environment would allow you to adjust various factors within your equation to try to approach a target value for your response variable. In this case since fuel efficiency is such an important factor when purchasing a car, this would be really useful. Obviously a real car company would use more than two variables, but the principle is the same.