# MAT 303 Module Two Problem Set Report

Interaction Terms and Qualitative Predictors

Catherine Taft

cat.taft@snhu.edu

Southern New Hampshire University

## 1. Introduction

Again we’re doing multiple regression models to study the fuel efficiency of cars. We’re examining the relationship between weight, horsepower and the axle ratio. This time however, because axle ratio can negatively impact the fuel economy even as the fuel economy is being positively impacted by decreases in weight, we have to evaluate the equation with the interaction term describing this relationship.

These results could be used by the car manufacturer to create a more realistic and nuanced equation, because in reality when you have a machine as complicated as a car with so many components, there are going to be scenarios like this that are less straightforward. So being able to evaluate for these nuanced situations is important, and a lot more realistic than what we did previously.

In this problem set we’re going to do the same basic types of multiple regression models, look at correlation and the R-squared value, run an F-test then start working with the interaction term and make some predictions based off of that form of the model.

## 2. Data Preparation

For this problem set we’re analyzing fuel efficiency (mpg), weight (wt) and horsepower (hp) again but this time adding in axle ratio (drat) to look at the relationships between these variables. The other variables won’t be used in this analysis. This data set has 6 rows and 12 columns. As before, it’s arranged by vehicle.

## 3. Model with Interaction Term

### Correlation Analysis

Here we’ll calculate the Pearson Coefficient, which describes the strength and direction of the correlation between variables. The closer to -1, the closer to a perfect negative correlation. The closer to 1, the closer to a perfect positive correlation. A Pearson Coefficient close to 0 would indicate little or no correlation.

Table

Description automatically generated

The Pearson Coefficient between fuel economy and weight is -0.8677 which indicates a strong negative correlation, and if we recall last week, we actually plotted that, and indeed fuel efficiency decreases as weight increases, which is a negative correlation.

The coefficient between fuel economy and horsepower is -0.7762, which is a slightly weaker negative correlation, and if we recall the scatterplot from last week, it did have quite a bit more variance in it.

The coefficient between fuel economy and rear axle ratio is 0.6812, which is a positive correlation but not particularly strong.

### Reporting Results

First, we run the model to get the coefficients and other information:

Table

Description automatically generated

The general form of the regression model with three predictor variables (weight, horsepower and rear axle ratio) would be . But since we are using the interaction terms for weight and horsepower and weight and rear axle ration this would be, assuming the predictor variables are, in order: weight (, hp( ) and axle ratio (): *.*

So with all of above information for our specific variables filled in, the model equation for this scenario would be:

Y = 75.68431 – 16.12967 – 0.1648 – 5.44987 + 1.70650 + 0.04069

The R-squared value here is 0.8907, and the adjust R-squared is 0.8697. The R-squared value tells us how much of the variation in the model is explainable by these variables, so here it tells us that about 89% of the variation in fuel efficiency can be explained by a model with these variables. The adjusted R-squared serves the same purpose as the regular R-squared, except that it’s altered to reflect the presence of multiple predictor variables. So still the value of 0.8697 demonstrates a strong model.

The calculate the change in fuel economy of a car with a weight of 3500lbs (3.50) per single unit increase in horsepower we can slim down the full regression equation above to only work with the relevant variables and interaction term (recalling the variable order I established above) and then substitute in the coefficients against an HP value of 1:

Y = - 0.1648(1) + 0.04069(3.5)(1)

Y = -0.022385

That means that for every single unit increase in horsepower, for a car with a weight of 3500lbs, fuel efficiency will decrease by 0.022385 MPG.

To calculate the change in fuel economy of a 3500lb car for every unit increase in rear axle ratio we can use the same principle. We’ll isolate the portion of the equation that deals with these variables, substitute in the appropriate parameters and evaluate:

Y = -5.44987(1) + 1.7065(3.5)(1)

Y = 0.52288

So this means that for every unit increase in rear axle ratio, whatever that would be, for a car of 3500lbs will have its fuel efficiency increased by 0.52288 MPG.

Next we’ll find the fitted and residual values for the model and plot them:

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

Based on the plot of fitted against residual values, we don’t see any pattern here so that suggests homoscedacity. With the QQ plot, since the plotted data points all fall more or less along the regression line, we can assume the residuals have a normal distribution.

### Evaluating Model Significance

Now we’re going to evaluate the model significance by doing an overall F-test. The null hypothesis here will be that there is no linear relationship between any of the variables (i.e. the coefficients are all 0), and the alternative hypothesis is that there is a linear relationship with at least one of the variables (at least one coefficient is not 0). In math terms:

We’re evaluating this at a significance level of 0.05. Looking at the above output from the code, we can see the p-value is 1.092^-11 which is really close to 0 and certainly much less than 0.05, so we will reject the null hypothesis and accept the null hypothesis that at least one of these predictor variables has a linear relationship with the response variable of fuel efficiency.

Because the overall F-test just tells us if one or more predictor variables has a linear relationship to the response variable, you’d then perform individual F-tests to see which ones do or do not. The null hypothesis here is that the variable in question does not have a linear relationship to MPG (that its slope is 0) and the alternative hypothesis is that it does (that its slope is not 0).

Again looking at the chart of output from the code, the p-values of weight, horsepower and axle ratio are 0.02624, 0.00146, and 0.25886 respectively. The p-value of the interaction term for weight:horsepower is 0.00595, and the one for the interaction term for weight:axle ratio is 0.24447. So evaluating this at 0.05 again, we can see that the weight, horsepower, and weight:hp all have p-values below the significance level, so these are statistically relevant in the model. However, the values for axle ratio and the interaction term for weight:axle ratio are both larger than 0.05, so these do not have a linear relationship with fuel efficiency.

### Making Predictions Using the Model

Now we’re going to run a version of this model that has a car with a weight of 2,965lbs, 210 horsepower and a rear axle ratio of 2.91.

We can use the form of the equation we wrote before and substitute in our specific values:

where weight is , hp is and axle ratio is .

Y= 75.6841 – 16.12967(2.965) – 0.1648(210) – 5.44987(2.91) + 1.7065(2.965)(2.91) + 0.04069 (2.965)(210)

That evaluates to ~ 17.452.

Table

Description automatically generated

Above we have the prediction and confidence intervals at 95% for the specific car we were just evaluating, with those parameters. The prediction interval is [12.4462, 22.4577] which means that we can say with 95% certainty that this car’s MPG will fall within those values. The confidence interval is [15.2024, 19.7016], which means that 95% of the time, the mean fuel efficiency for a group of cars with the parameters that we specified above would fall within that interval.

## 4. Model with Interaction Term and Qualitative Predictor

### Reporting Results

The general form of the regression model for fuel economy using (in order) weight, hp, the qualitative predictors of number of cylinders (6 and 8), and the interaction term for weight:hp is:

We’ve pulled the information for the new model:

Table

Description automatically generated

So with this we can make the actual model equation for this scenario:

The R-squared value here is 0.888, and the Adjusted R-squared value is 0.8664. This means that about 87% of variation in MPG can be explained by the model using these specific predictors.

After this we get the fitted and residual values, and created a plot of those as well as a QQ plot:

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

There is not a pattern to the scatterplot of fitted values against residuals, so we can assume homoscedacity here. Also the residuals all fall more or less along the regression line in the QQ plot so we can assume they have a normal distribution.

### Evaluating Model Significance

Now we are going to evaluate the model at a 5% significance level. The null hypothesis for this hypothesis test is that there is no linear relationship between any of the variables in the model. The alternative hypothesis is that there is a linear relationship between at least one of the variables, and MPG. In other words at least one variable will have a coefficient (slope) that is not 0.

In the R output above we can find the P-value, which is 1.503^-11 which is much lower than 0.05, so we will reject the null hypothesis and accept the alternative hypothesis that there is a linear relationship between at least one of these variables, and MPG.

To determine which variables are statistically significant in the model we’ll carry out individual beta tests at a 5% significance level. The null hypothesis is that the variable has a coefficient of 0 and therefore does not have a linear relationship to MPG. The alternative hypothesis is that it has a coefficient other than 0 and does have a linear relationship.

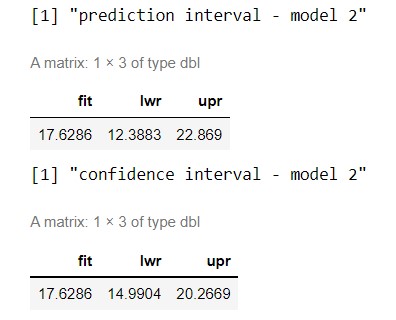
Again looking at the R output, we can find the individual p-values for each variable. I’ll just list them here. Weight: 0.000181, HP: 0.003274, 6 cylinders: 0.405685, 8 cylinders: 0.487246, and the one for the interaction term of weight:hp is 0.012865. Weight, HP and the interaction term are all below the 0.05 significance level, so we’ll reject the null hypothesis and accept the alternative hypothesis, meaning each of them is statistically relevant in this model. However, the p-values for 6 and 8 cylinders are both above the significance level so here we’ll accept the null hypothesis, meaning they are not statistically relevant in the model.

### Making Predictions Using the Model

To find the predicted fuel economy for a car with weight of 2.965k lbs, 210 horsepower and 6 cylinders we can substitute those values into the equation for this model:

This evaluates to ~11.98 MPG.

We ran the prediction and confidence intervals for this model:



The 95% prediction interval for this model is [12.3883, 22.869] which means that 95% of the time the MPG for a car with these specific parameters will fall within that value range. The 95% confidence interval is [14.9904, 20.2669] and this tells us that 95% of the time the averaage MPG of a group of cars with these specific parameters will fall within that value range.

Prediction intervals are wider than confidence intervals because the prediction interval has to account for sampling uncertainty as well as any potential variation in the individual parameters.

## 5. Conclusion

I’m not really sure which model I’d recommend, because they both included variables that were not statistically relevant, and if we removed those variables it would basically be the same model. I guess the first model had slightly higher R-squared values which would indicate that it’s a bit more reliable with more of the variation explained by the model, but even that difference is pretty negligible.

Based on these results it seems like weight and horsepower are the only two obviously influential variables here. The use of interaction terms here is a great idea and I think since the additions we tried were not useful to the model, I’d suggest looking for some additional variables to test to see if they contribute. However even just the use of the interaction terms provides are more nuanced and appropriate model.

The practical importance of the analyses we did is that since we’ve basically isolated variables that were cluttering the equation that contributed no meaning to the model. If a car company is trying to figure out how to influence the fuel economy of one of its cars, it would be really important to make sure that you’re aware of exactly what impacts that. If you focus your energy on engineering decisions that end up not having an impact on fuel economy at all when that’s your goal, that’s a waste of time and energy. Additionally, the inclusion of the interaction terms would make the model more accurate and nuanced, reflecting how the car’s components influence one another.