# MAT 303 Module Two Problem Set Report

Interaction Terms and Qualitative Predictors

Zachary Tankersley

Zachary.tankersley@snhu.edu

Southern New Hampshire University

## 1. Introduction

For this analysis, we will be using the mtcars dataset containing different data points concerning certain qualities of different vehicles from different manufacturers. The results of this analysis can be used to gauge the impact that certain modifications to the different aspects of these cars may have on the fuel efficiency of the vehicles. Example data points would be final drive gear ratios, number of cylinders, weight, or horsepower. For this analysis, a regression model will be built using interaction terms and qualitative variables that will attempt to explain the impact that relationships between variables may have on the response variable (MPG).

## 2. Data Preparation

For this data set, we will be taking a closer look at the weight (wt), horsepower (hp), the rear axle ratio (drat), and the number of cylinders (cyl). In combination with those data points, we will use the interaction terms weight to horsepower (wt:hp), weight to rear axle ratio (wt:drat), and weight to horsepower (wt:hp). There are 33 rows and 12 columns in the source data, although the first row contains header values that only name the columns, making the row count 32 in practice.

## 3. Model with Interaction Term

### Correlation Analysis

Using analysis techniques, we can calculate the Pearson correlation coefficients between the variables as follows:

1. The correlation coefficient between MPG and weight is -0.867659376517228 which indicates that MPG and weight have a very strong negative correlation.
2. The correlation coefficient between MPG and horsepower is -0.776168371826586 which indicates that MPG and horsepower have a strong negative correlation.
3. The correlation coefficient between MPG and rear axle ratio is -0.681171907806749 which indicates that MPG and rear axle ratio have a strong negative correlation.

As can be seen in these results, each of the variables has at *least* a strong negative correlation with fuel efficiency.

### Reporting Results

Beginning with a basic analysis of the data and variables in question, we can represent the regression model and include the variables in their general form, in the following manner:

Once we run the summary analysis function using the R interpreter, we can build the real model which appears as:

With the summary analysis function, we also receive an R-Squared value of 0.8907 and an Adjusted R-Squared value of 0.8697 which both indicated a good “fit” for the chosen variables in the model, and the theory explains 89% of the variance in the dependent variable with this combination of independent variables. The adjusted value has a small difference but wouldn’t necessarily be enough to indicate that one of our variables or interaction terms is not a good fit. The coefficient for horsepower comes out to be (-0.16480 + 0.04069x1) and when we substitute a car with a weight of 3.50, we get -0.022385, meaning that for every unit increase in horsepower, fuel efficiency decreases by about 0.02 miles per gallon. The coefficient for real axle ratio (drat) comes out to be (-5.44987 + 1.70650x1) and when substituting in a car of weight 3.50 we get 0.52288, which means that for every unit increase in real axle ratio, we see an MPG increase of about 0.52. When using a general plot function and comparing a scatter plot of residual values vs fitted values and a normal quantile-quantile plot we can see a mostly very randomized scatter plot and a QQ plot where values follow the linear model fairly tightly all the way through. This would hint at strong normality in the residual values, and also confirms homoscedasticity.

### Evaluating Model Significance

When critiquing the model and trying to determine if the model has significance at a 5% level, we carry out an overall F-test and identify the null and alternative hypotheses and state them in the following manner:

With the P-value returned of 1.092e-11, which is significantly lower than the significance of 0.05, we can confidently reject the null hypotheses in favor of the alternative hypotheses; a statistically significant relationship exists between one of the predictor variables and MPG. To determine which individual variables are significant in this model at a 5% level we perform an individual beta test on each variable and compare them to the null and alternative hypotheses as follows:

The P-values for the variables are shown to be 0.02624 for weight, 0.00146 for horsepower, 0.25886 for axle ratio, 0.24447 for weight to axle ratio, and 0.00595 for weight to horsepower. The P-values for axle ratio and weight to axle ratio are above the significance level of 0.05 and so cause us to reject the favor of the null hypothesis and reject those variables as *not* statistically significant relationships to fuel efficiency. All other variables pass the test and fit our model well.

### Making Predictions Using the Model

When using the model to make predictions, we can test and substitute values for a car that has 2.965 weight, 210 horsepower, and 2.91 rear axle ratio. When we do this test, we are returned an estimated MPG of ~17.45198. The 95% prediction interval for the MPG for this car is [13.2984,21.6056], meaning 95% of predicted values should fall within that range, and the 95% confidence interval is [15.5853, 19.3186], meaning that 95% of our current population should fall within that range.

## 4. Model with Interaction Term and Qualitative Predictor

### Reporting Results

To analyze the data set using an interaction term between weight and horsepower still, but now including the number of cylinders as a qualitative predictor, we can represent the general form of the regression model as follows:

Once we run the summary analysis function using the R interpreter, we can build the real model as follows:

Also using the summary analysis function, we receive an R-Squared value of 0.8869 and an Adjusted R-Squared of 0.8702 which both indicate a good “fit” for the chosen variables in the model and the theory explains about 88-89% of the variance in the dependent variable with this combination of independent variables. The adjusted value is slightly smaller but not alarmingly so. Examining the residual vs fitted value scatterplot and QQ plot created below, it would appear that the data *is* homoscedastic, *and* its normality holds. The residual vs fitted values plot has no clear groupings (although a slight right skew exists) and the QQ plot values do not deviate from the regression line a great deal.

### Evaluating Model Significance

When evaluating the significance model and looking at the results from the previous summary analysis function we can see that the model has a P-value of 2.156e-12 which is much lower than the 5% significance threshold. When comparing the null and alternative hypotheses (which should look the same as model 1), we can reject the null hypothesis in favor of the alternative and say that a significant relationship exists between at least one predictor variable and fuel efficiency. To determine which variables are significant we will perform individual beta tests and receive the following results: the P value for weight is 2.93e-05, for horsepower, it is 0.00114, for cylinders it is 0.47887, and for weight to horsepower it is 0.00322. When we compare these values to our null and alternative hypotheses, we can safely include all values *except* for cylinders, which fails the beta test causing us to reject cylinders as a statistically significant variable in this model.

### Making Predictions Using the Model

When using the second model to make predictions, we can test and substitute values just as before, for a car that has the same weight and horsepower, but with 6 cylinders instead. When we perform this substitution into our model, we can predict with 95% accuracy that the MPG will be ~18.0084. The 95% prediction interval for the MPG for this car is [13.9283, 22.0885], meaning 95% of predicted values will fall within that range. The 95% confidence interval is [16.2864, 19.7304], meaning that 95% of our current population should fall within that range. Prediction intervals are usually wider due to the uncertainty of the predictive nature of a regression model. The confidence interval can be somewhat considered to be the measurement wherein 95% of sample means fall within, whereas the prediction interval is the range that the model “predicts” 95% of new observations will fall between.

## 5. Conclusion

It is very difficult for me to say at this stage which model I would recommend over the other. On the one hand, the R-Squared is slightly better on model 1, however, the Adjusted R-Squared is slightly better on model 2. The P-value on the second model is lower than that of the first model, meaning more significance in theory. The beta test on the cylinder variable in the second model summary is outside of our desired range by quite a bit more than either axle ratio or axle ratio to weight in the first model. With all of that said, there seems to be more of a skew on the residual vs fitted scatter plot for the second model. I am unsure which of these factors is more critical than the others if at all, but my instincts are telling me that the model using cylinders is likely to be *more* accurate in predicting MPG than the model using gear ratios. Both models have variables that do not seem to fit the significance thresholds. With real-world experience in high horsepower cars, some very high horsepower 6-cylinder cars get great gas mileage for what they are, while their high horsepower v8/v10/v12 counterparts get dismal mileage, at best (see average MPG of any dodge 8-cylinder car). With all of these things said, I would like to eventually test and see whether or not the inclusion of *both* cylinders and axle ratios wouldn’t build a better model. Hopefully, one of these models or a distilled version could be used for reduction in emissions or better fuel efficiency for manufacturers, or profits. Marketing campaigns for automakers like to focus on miles per gallon so having a predictive model that can help vehicle designs to be efficient *before* they come off the finish line may help to improve profit ratios.