# MAT 303 Module Two Problem Set Report

Interaction Terms and Qualitative Predictors

Justin Farquhar

Justin.farquhar@snhu.edu

Southern New Hampshire University

## 1. Introduction

For this analysis we are utilizing a data set containing the fuel economy of several cars containing important variables to see their effects on the vehicles fuel economy. These results could be used to help find correlations between various variables to see how they affect fuel economy, either on their own or in conjunction with one another. We will be performing analyses on some of the various variables in correlation with mpg, seeing how the variables correlate in different graph types allowing us to better visualize the data.

## 2. Data Preparation

Most of the variables could be considered important, but for these analyses specifically the most important are going to be miles per gallon, horsepower, rear axle ratio, quarter mile time, and number of cylinders. For miles per gallon, we are going to be comparing every other variable to see how this changes and affects the overall fuel economy of the vehicle. For horsepower, generally as a vehicle can accelerate quicker, the fuel economy is going to fall, but this may not always be the case. For rear axle ratio, this can affect how many times the wheel is able to rotate and how quickly it is able to. For quarter mile time, we would generally be thinking similarly to horsepower, as the quicker a vehicle is able to accelerate, the quicker their time should be, which would likely have a similar effect on fuel efficiency. Finally for number of cylinders, generally the more cylinders a vehicle has the more power it can output, likely resulting in similar results as horsepower and quarter mile time. There are a total of 32 rows and 12 columns throughout this data set.

## 3. Model with Interaction Term

### Correlation Analysis

*Describe the relationships between variables in the data set. Address the following questions in your analysis:*

*A table with numbers and letters

Description automatically generated*

The image provided above shows the Pearson Correlation matrix between the variables and miles per gallon. This shows that there is a negative linear relationship between horsepower and miles per gallon since the value is -0.7762. The closer to -1 the number is, the stronger the negative relationship is, eventually leading to a perfect linear relationship, so a relationship of -0.7762 is fairly strong. To then move onto quarter mile time and miles per gallon we have a value of 0.4187, which is a positive linear relationship, but is pretty close to a medium strength. Then to more onto rear axle ratio, we have a value of 0.6812, which is again a positive linear relationship, and a stronger relationship but only just past the middle.

### Reporting Results

General Form:

+

Prediction Model:

+

Prediction model equation using outputs from multiple regression model:

For this data, the value of R-squared is 0.8207 and the Adjusted R-squared is 0.7862. The R-squared value indicates that roughly 82.07% of the variability in miles per gallons is explained by these variables. The remaining percentage would be account for by the other variables that we did not account for in this analysis. The adjusted R-squared value being relatively close to R-squared indicates that the predictors that we added to the model are relevant and provide meaningful information.

To estimate the change in fuel economy of a car with 160 horsepower for each unit increase in quarter mile time we will have to adjust the formula slightly.

Looking at this, for a vehicle with 160 horsepower, each additional unit increase in quarter mile time is associated with an estimated decrease of approximately 1.4824 mpg.

To then estimate the change in fuel economy of a car with 160 horsepower for each unit increase in rear axle ratio we will again have the adjust the formula slightly.

Looking at this, for a vehicle with 160 horsepower, each additional unit increase in rear axle ratio is associated with an estimated increase of approximately 0.3546 mpg.

A graph with orange dots

Description automatically generated

A graph of a normal q-q plot

Description automatically generated

We can assess homoscedasticity based on the image of the scatter plot. Homoscedasticity means that the variance of the residuals should be constant across all levels of the fitted values, ideally being randomly scattered around zero without any clear pattern. In this case, we see some spread with some variation in residual size. We can visualize no clear patterns, confirming homoscedasticity. We then assess the normality of residuals based on the Q-Q plot above. For perfect normality of residuals, every plot point would fall along the line. Looking at the image, the points are relatively along the line, but deviate on the tails and a little along the middle. Mild deviations are common and may not impact model validity, but pronounced deviations could suggest that the residuals are not perfectly normally distributed.

### Evaluating Model Significance

Looking at the F-statistic p-value for this model we can see that our model is significant at a 5% level of significance since the value is 6.098e-09, which is much lower that 5%. We can also look at the individual p values for each variable and see that all would be considered significant at the 5% level except for horsepower:rear axle ratio which came out to 8.4%.

### Making Predictions Using the Model

A screenshot of a computer

Description automatically generated

To determine the predicted fuel economy for a car that has 175 horsepower, 14.2 quarter mile time and 3.91 rear-axle ratio we can plug these values into our model which we can see in the image above gets a value of 21.5285. This image also shows the 95% prediction interval for the fuel economy which gives us a lower and upper value that the fuel economy should fall between after accounting for other potential variables. We can then see the confidence interval below that, also showing a lower and upper value that the value should fall between.

## 4. Model with Interaction Term and Qualitative Predictor

### Reporting Results

General Form:

+

Prediction Model:

+

Prediction model equation using outputs from multiple regression model:

For this model, the value of R-squared is 0.8327 and the adjusted R-squared is 0.8005. The R-squared value indicates that roughly 83.27% of the variability in miles per gallons is explained by these variables. The remaining percentage would be account for by the other variables that we did not account for in this analysis. The adjusted R-squared value being relatively close to R-squared indicates that the predictors that we added to the model are relevant and provide meaningful information.

A graph with orange dots

Description automatically generated  
A graph with green dots

Description automatically generated

Looking at each of these graphs again, we can see that the model again adheres to the assumption of homoscedasticity as the points are varied around zero with no clear pattern. Looking at the Q-Q graph we can see fairly similar results to our previous model, where the tails are starting to trail off slightly, but for the most part the values in the middle are adhering to the line and following the normality of the residuals. Mild deviations are common and may not impact model validity, but pronounced deviations could suggest that the residuals are not perfectly normally distributed.

### Evaluating Model Significance

Looking at the F-statistic p-value for this model we can see that our model is significant at a 5% level of significance since the value is 2.536e-09, which is much lower that 5%. We can also look at the individual p values for each variable and see that all would be considered significant at the 5% level except for horsepower which came out to 8.48% and 8 cylinders which came out to 8.47%.

### Making Predictions Using the Model

A screenshot of a computer

Description automatically generated

To determine the predicted fuel economy for a car that has 175 horsepower, 14.2 quarter mile time and 6 cylinders, we can plug these values into our new model, which looking at the image above gives us a miles per gallon of 21.3424. This image also shows the prediction interval for this vehicle, giving a lower and upper limit for the value to be. There is also the confidence interval present, also including lower and upper limits for the value to be. The prediction interval is wider because it accounts for two sources of variability, the variability in estimating the mean and the natural variability of individual observations around the mean.

## 5. Conclusion

Assuming the sample size is sufficiently large, I believe the second model better accounts for homoscedasticity and sticking to the normalcy of the residuals should generally indicate a better fit. If residuals have constant variance across fitted values, this suggests that the model’s predictions are equally reliable across the range of data, rather than becoming more or less accurate at certain values. When residuals are normally distributed, it suggests that the model errors are symmetrically distributed around zero and follow a consistent pattern.

These results can further improve the car manufacturing pipeline to where companies can begin with creating the concepts that the car is going to include and see a prediction of the fuel efficiency before sufficiently testing it. This can allow them to adjust some of the features that may have more of a negative relationship with fuel efficiency to better adhere to the customer population they are trying to create the vehicle for, as some customers will value fuel efficiency much higher than others, while some may value the aspects of the actual vehicle and not care as much as to what the fuel efficiency is.