# MAT 303 Module Two Problem Set Report

Interaction Terms and Qualitative Predictors

Kayla Sacks

kayla.sacks@snhu.edu

Southern New Hampshire University

## 1. Introduction

For this statistical analysis I will be using multiple variables from the *mtcars* data set to create two separate models to evaluate how different variables affect a car’s fuel economy. The first model will focus on the affect horsepower (hp), quarter mile time (qsec), rear axle ratio (drat), interaction between horsepower and quarter mile time (hp \* qsec), and horsepower and rear axle ratio (hp \* drat) have on fuel economy. The second model will also include the number of cylinders (cyl) and only the interaction term for horsepower and quarter mile time.

There are multiple analyses that I will be performing using these variables. These include performing multiple regressions using interaction terms, find the fitted values and residuals as well plot the residuals against the fitted values, create a q-q plot to evaluate the normality of the residuals, then find the 95% prediction and confidence intervals. The car maker can use the results of the analyses when designing cars to increase fuel economy.

## 2. Data Preparation

The original data set that I will be using has 32 rows and 12 columns, or variables. I will only be using 5 of the variables for my analysis. The important variables that I will be using are:

* mpg - Fuel economy in miles per gallon, This will be the dependent variable for the analysis.
* hp – Horsepower, this is the first predictor variable. This variable is used in both interaction terms.
* qsec – Quarter mile time, this is the second predictor variable.
* drat – Rear axle ratio, this is the third predictor variable.
* cyl – The number of cylinders in the car’s engine, this is the final predictor variable. This is a qualitative variable which can be either 4, 6, or 8. cyl6 and cyl8 will be used for this qualitative variable with 4 cylinders being the base. This will be used in the second model.

I will be using an interaction term between horsepower and quarter mile time as well as one between horsepower and rear axle ratio.

## 3. Model with Interaction Term

### Correlation Analysis

The correlation matrix for the Pearson Correlation Coefficients for miles per gallon, horsepower, quarter mile time, and rear axle ratio is below.

*A table with numbers and symbols

Description automatically generated*

From the correlation matrix we can see the correlation strength and direction between the dependent variable and predictor variables:

* The correlation coefficient between fuel economy and horsepower is -0.7762, showing a strong negative correlation.
* The correlation coefficient between fuel economy and quarter mile time is 0.4187, showing a moderate positive correlation.
* The correlation coefficient between fuel economy and rear axle ratio is 0.6812, showing a strong positive correlation.

### Reporting Results

The general form for this regression model which includes two interaction terms is:

The prediction equation for the regression model is:

Where is the predicted value of the fuel economy (mpg), is horsepower (hp), is quarter mile time (qsec), and is rear axle ratio (drat). The following regression model output was created with these variables:

A screenshot of a computer

Description automatically generated

Once the regression model was created for fuel economy by using horsepower, quarter mile time, and rear axle ratio as predictors as well as the interaction terms for horsepower and quarter mile time, and horsepower and rear axle ratio, I then updated the prediction model using the output values to be:

The bata estimates show how the response variable will increase, or decrease, each time a predictor variable increase by one. The (R-squared) is 0.8207 and the (Adjusted R-squared) is 0.7862 for this model. The R-squared value tells us that the regression model explains 82% of the variation in fuel economy using the predictor variables.

By using this model, we can estimate what the fuel economy will be for a car with 160 horsepower for each unit increase in quarter mile time. Since we have the interaction term and want to find the change in fuel economy for each (1) unit increase in quarter mile time, we will only use a part of the equation.

This indicates that for every 1.0 increase in quarter mile time for a car that has 160 horsepower, the fuel economy (mpg) will decrease by 1.4824.

We can use the same steps to now estimate the change in fuel economy of a car with 160 horsepower for each (1) unit increase in rear axle ratio:

This indicates that for every 1.0 increase in rear acle ratio for a car that has 160 horsepower, the fuel economy (mpg) will increase by 0.3516.

After calculating the fitted values and residuals, the scatterplot for Residuals against Fitted Values was created and can be viewed below.

A graph with red dots

Description automatically generated

The residuals against fitted values plot above does not show any patterns or clustering so we can that it meets the requirements of homoscedasticity. Next, I created the normal q-q plot from the residuals.

*A graph of a normal q-q plot

Description automatically generated*

Based on the Normal Q-Q plot above, we can assume a normal distribution as even though there is some deviation from the diagonal line, it is not significant enough to consider it non-normal without further analysis with a larger sample size.

### Evaluating Model Significance

To evaluate the model is at a 5% level of significance I first conducted the overall F-test starting with the null hypothesis which states that there is no relationship between the dependent variable and predicter variables. The mathematical equation for the null hypothesis is:

Next is the alternate hypothesis which states that there is at least one predictor variable has a relationship to the dependent variable with the following mathematical equation:

The p-value for the test is 6.098e-9 is lower than the significance level of 0.05, which means we will reject the null hypothesis in favor of the alternate hypothesis. This also means that there is a significant linear relationship between the dependent variable (fuel efficiency) and one or more of the predictor variables.

With the overall F-test competed, next is the individual beta test next using the null hypothesis and alternate hypothesis with :

During the individual test, was the only beta value over the .05 level of significance with a p-value of 0.084. This means that there is not sufficient evidence to reject the null hypothesis for the interaction term of horsepower and rear axle ratio which suggests that the term is not linearly related to fuel efficiency. The next highest p-value was 0.040 for which is lower than the .05 level of significance. The conclusion of this test shows that all predictor variables other than the interaction term of horsepower and rear axle ratio have a statistically significant relationship to fuel efficiency.

### Making Predictions Using the Model

I will use the prediction model equation to find the predicted fuel efficiency for a car that has 175 horsepower( ), 14.2 quarter mile time (), and 3.91 rear-axel ratio ():

The value above calculated from the prediction model equation is close to the value from the model which is 21.5285.

The 95% prediction interval for an individual response tells us that we can be 95% certain that a car’s fuel efficiency will fall within the interval of (15.0897, 27.9674) if it uses the same measurements used in the prediction model equation above.

The 95% confidence interval for the mean for fuel economy tells us that we can be 95% certain that the average fuel efficiency for multiple cars that meet the same measurements used in the prediction model equation above will be in the interval of (18.5881, 24.469).

## 4. Model with Interaction Term and Qualitative Predictor

### Reporting Results

The general form of the regression model including one interaction and one qualitative predictor term:

The prediction equation of this regression model:

Where is the predicted value of the fuel economy (mpg), is horsepower (hp), is quarter mile time (qsec), and are both dummy indicator variables for number of cylinders(cyl). The reference level for number of cylinders is 4 so will correspond to a car with 4 cylinders. indicates 6 cylinders (cyl6) while indicates 8 cylinders (cyl8). The output for the regression of the second model can be seen below:

A screenshot of a computer

Description automatically generated

After creating the second regression model for fuel economy by using horsepower, quarter mile time, interaction term for horsepower and quarter mile time, and number of cylinders we can insert the outputs into the prediction model equation.

The (R-squared) is 0.8327 and the (Adjusted R-squared) is 0.8005 for this model. The R-squared value tells us that the regression model explains 83% of the variation in fuel economy using the predictor variables.

After obtaining the fitted values and residuals from the model 2 data set, the Residuals against Fitted Values and Normal Q-Q plots were created and can be seen below.

A diagram with red dots

Description automatically generated

A graph of a normal q-q plot

Description automatically generated

As the Residuals against Fitted Values plot does not show any noticeable pattern, we can say that there is an assumption of homoscedasticity. From the normal q-q we can see that there are no significant deviations so we can say that the residuals are normally distributed.

### Evaluating Model Significance

I conducted the F-test for model2 regression model to find if the model is significant at a 5% level of significance by using the null hypothesis and alternate hypothesis which are the same as for the first model that was created.

Although the hypotheses are the same between the models, the p-value for model2 is 2.529e-9. This p-value is lower than the 0.05 level of significance so the null hypothesis is rejected. We can conclude from this test there is a significant linear relationship between the dependent variable (fuel efficiency) and one or more of the predictor variables.

Now we will move to the individual beta test next using the null hypothesis and alternate hypothesis with :

For model2 there are only two of the five variables that meet the 5% level of significance which is the first dummy variable for number of cylinders (cyl6) with a p-value of 0.0118 and the interaction term for horsepower and quarter mile time with a p-value of 0.0246. The variables for horsepower, quarter mile time, and the second dummy variable for number of cylinders(cly8) do not meet the 5% level of significance and therefore fail to reject the null hypothesis. We can conclude with this test that only the first dummy variable and the interaction term have a significant linear relationship with fuel efficiency.

### Making Predictions Using the Model

With the completion of model2, we can now predict the fuel economy for a car that has 175 horsepower, 14.2 quarter mile time, and 6 cylinders using the prediction model equation. Note that since the number of cylinders is a qualitative variable, the first dummy value will be 1 as it indicates a car with 6 cylinders the second dummy value, will be 0.

The value above calculated from the prediction model equation is close to the value from the model prediction which is 21.3424.

The 95% prediction interval for an individual response tells us that we can be 95% certain that a car’s fuel efficiency will fall within the interval of (14.8764, 27.8085) if it uses the same measurements used in the prediction model equation above.

The 95% confidence interval for the mean for fuel economy tells us that we can be 95% certain that the average fuel efficiency for multiple cars that meet the same measurements used in the prediction model equation above will be in the interval of (17.9965, 24.6884).

The prediction intervals consider uncertainty in the estimation which is why it is wider than the confidence intervals.

## 5. Conclusion

Based on the analysis and with the assumption that the sample size is of sufficient size, I would recommend the first model that was created. Although the R-squared value of model 1 was .82 whereas the model 2 value was .83,I don’t believe the .01 difference is significant enough to choose model two over model one. The main reason that I chose model one was due to the outcomes of the individual t-tests. Model one used five values, two of which are interaction terms with four of the five rejecting the null hypothesis in favor of the alternate hypothesis which means that those four variables have a statistically significant relationship with fuel economy. On the other hand, of the five values from model 2, one being an interaction term and two being dummy variables for number of cylinders, only two of the variables are shown to have a statistically significant relationship. The model 2 variables were also a bit tricky since the first of the dummy variables (cyl6) rejected the null hypothesis while the second (cyl8) did not. With this we can assume that in the model 2 that cars with 4 or 6 cylinders are more statistically significant when analyzing fuel economy than cars with 8 cylinders.

The result of this analysis is important as it can help car manufacturers find and understand what individual predictor variables have the most affect on fuel economy as well as how the predictor variables interactions with each other affects fuel economy. Using this information and testing, the manufacturers can design and build new cars that get better fuel economy.