# MAT 303 Module Two Problem Set Report

Interaction Terms and Qualitative Predictors

[Luna Saccoia]

[Luna.Saccoia@snhu.edu]

Southern New Hampshire University

## 1. Introduction

*The statistical analyses being performed here have to do with calculating the relationship between a vehicle’s fuel efficiency and multiple variables. We are calculating the relationship between fuel efficiency against weight, horsepower, and gear ratio for our first model. For our second model we are calculating the relationship between fuel efficiency against weight, horsepower, and number of cylinders.*

*These results might be used by car manufacturers to determine how fuel efficient a vehicle will be when it is being designed. They can predict how each of these variables individually and together can affect fuel efficiency.*

*The analyses I will be running in this problem set are two multiple regression models, each with different variables being compared against fuel efficiency in ‘miles per gallon’. I will also be finding the fitted values, finding the residuals, plotting the fitted values against the residuals, creating a q-q plot to test assumptions of normality of the residuals, and finding the confidence intervals for each model. We will take all of this information and use it to draw conclusions about whether or not the variables used in each model have a statistically significant effect on fuel efficiency in miles per gallon.*

## 2. Data Preparation

*The variables that we will be working with for this problem set are mpg (miles per gallon), wt (weight), hp (horsepower), and drat (rear axle ratio). We are looking to see, based on our calculations, if weight, horsepower and rear axle ratio have a statistically significant affect on fuel efficiency.*

*There are 12 columns in the full data set, for different variables relating to the vehicles. There are 32 rows, one for each vehicle whose data is included in the data set.*

## 3. Model with Interaction Term

### Correlation Analysis

*We calculated the Pearson Correlation Coefficients between economy and weight; fuel economy and horsepower; and fuel economy and rear axle ratio.*

* *For fuel economy and weight we have a strong negative correlation based on an value of*

*.*

* *For fuel economy and horsepower we have a moderate negative correlation based on an value of .*
* *For fuel economy and rear axle ratio we have a moderate positive correlation based on an value of .*

### Reporting Results

*The equation for the multiple regression model with fuel efficiency as the response variable (denoted as ) and weight, horsepower and rear axle ratio as the predictor variables (denoted as ) is:*

*Our popular regression parameters are our values. Our is our intercept.*

*The model equation for our regression model with weight, horsepower and rear axle ratio as our predictors (denoted as respectively); including the interactive terms for weight and rear axle ration and for weight and horsepower is:*

*Our is our intercept. Our values are our popular regression parameters from our multiple regression model.*

*The value is . and our adjusted () value is . The value is the coefficient of multiple determination. This value tells us that approximately 89% of the variation in fuel economy can be explained using a model that uses predictors for weight, horsepower, rear axle ratio, weight against rear axle ratio and weight against horsepower.*

*For this model, we can estimate the change in fuel economy of a car with a weight of 3.50 for each 1.0 unit increase in horsepower by using the equation*

*Because of the interaction term, and the fact that we want to predict how much fuel economy will change based on this interaction we only use the following portion of the equation:*

*This means that for a car with a weight of 3.50, fuel economy will decrease by 0.022385 for every 1 unit increase in horsepower. This is done by plugging in our popular regression parameters for horsepower and weight against horsepower into the equation for respectively, 3.50 as our value since were using a car with a weight of 3.50 for this exercise, and 1 as our value since we want to know the change per 1 unit increase of this variable.*

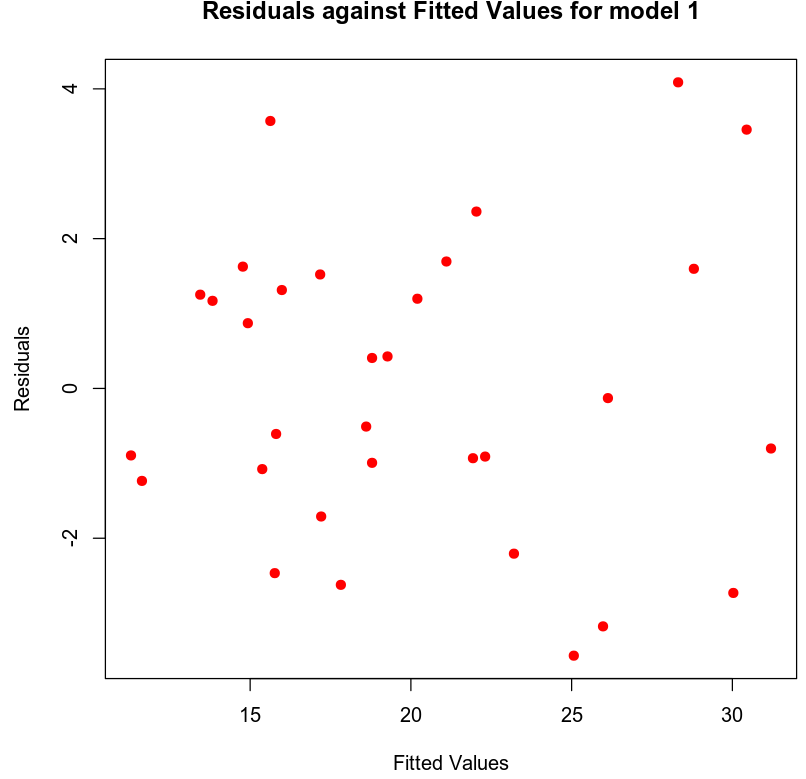
*We can use the same equation setup to estimate the change in fuel economy of a car with a weight of 3.50 for each 1 unit increase in real axle ratio, by using the following equation:*

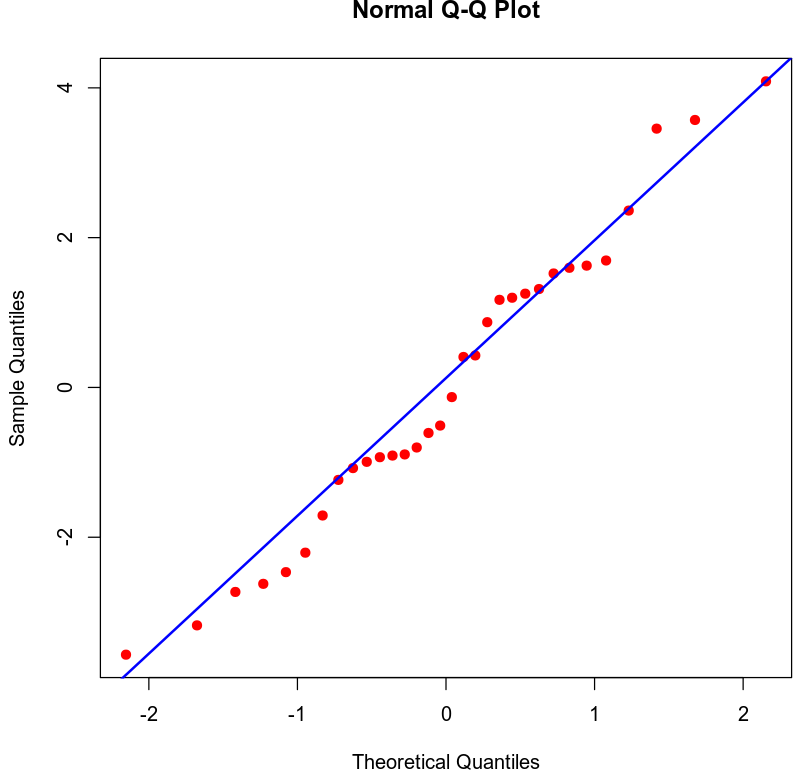
*Where our terms are now our rear axle ratio and our interactive term for weight against rear axle ratio. Our terms are now weight and the unit of increase in rear axle ratio.*

*Our equation is now:*

*This means that for a car with a weight of 3.50, fuel economy will increase by 0.52288 for every 1 unit increase in rear axle ratio.*

*Next we obtain the fitted values and residuals using the model for the data set and use them to create a plot for residuals against fitted values and a normal Q-Q plot, as seen below:*

**

**

*Based on the plot for residuals against fitted values, there is no discernible pattern and therefore has an assumption of homoscedasticity. I would conclude that the Q-Q plot has normally distributed residuals, as there is no significant deviation of the plots from the line.*

### Evaluating Model Significance

*In evaluating model significance for the regression model, where we want to see if the model is significant at a 5% level of significance, we perform an overall F-test by identifying our null hypothesis and alternative hypothesis in the following format:*

*Our P-value is 1.092e-11, which is far below our level of significance of 0.05. This means we should reject our null hypothesis in favor of our alternative hypothesis. The conclusion of this test tells us that a statistically significant relationship exists between fuel economy and at least one of our predictor variables.*

*In evaluating which terms in the model are significant at a 5% level of significance, we perform an individual beta test for each variable. The following are our null hypothesis and our alternative hypothesis:*

*Our P-value for weight is 0.02624, our P-value for rear axle ratio is 0.25886, our P-value for horsepower is 0.00146, our P-value for weight against rear axle ratio is 0.24447, and our P-value for weight against horsepower is 0.00595.*

*Using these P-values we see that P-value for weight, horsepower and weight against horsepower are all below our 0.05 level of significance. So we can conclude that these variables each have a statistically significant relationship with our fuel economy. Because our P-values for rear axle ratio and weight against rear axle ratio are not below our level of significance of 0.05, we can conclude that these variables do not have a statistically significant relationship with fuel economy.*

### Making Predictions Using the Model

*Next we will make some predictions using the regression model. For our scenario, we will use the hypothetical values of weight = 2.965, horsepower = 210, and rear axle ratio = 2.91. Our equation for this is:*

*(This value is from the model. When I did the math by hand, I got a vastly different number (~33.3114) and I am unsure why)*

*Our is our intercept, our values are our popular regression parameters for weight, horsepower, rear axle ratio, weight against rear axle ratio, and weight against horsepower respectively. Our are our hypothetical values for weight, horsepower and rear axle ratio respectively.*

*The 95% prediction interval for an individual response for fuel economy of this car is.*

*The prediction interval for an individual response tells us we can be 95% certain that a car’s fuel economy (in miles per gallon) will fall within these lower and upper bounds, if the car’s weight is 2.965, its horsepower is 210 and it has a rear axle ratio of 2.91. This is taking into account uncertainty related to  varying according to the regression error, , as well as sampling uncertainty related to estimating the regression parameters.*

*Taking the regression error into account as well as sampling uncertainty relating to estimating the regression parameters is the reason the prediction interval for an individual response is wider than the confidence interval for the mean. (Zybooks, 2016)*

*The 95% confidence interval for the mean for fuel economy of this car is.*

*This confidence interval for the mean tells us we can be 95% certain that the average fuel economy (in miles per gallon) of a group of cars will be within these lower and upper bounds; if they have a weight of 2.965, a horsepower of 210 and have a rear axle ratio of 2.91.*

*This is taking into account sampling uncertainty related to estimating the regression parameters. Because it does not also take the regression error into account, the confidence interval is not as wide as the prediction interval.(Zybooks, 2016)*

## 4. Model with Interaction Term and Qualitative Predictor

### Reporting Results

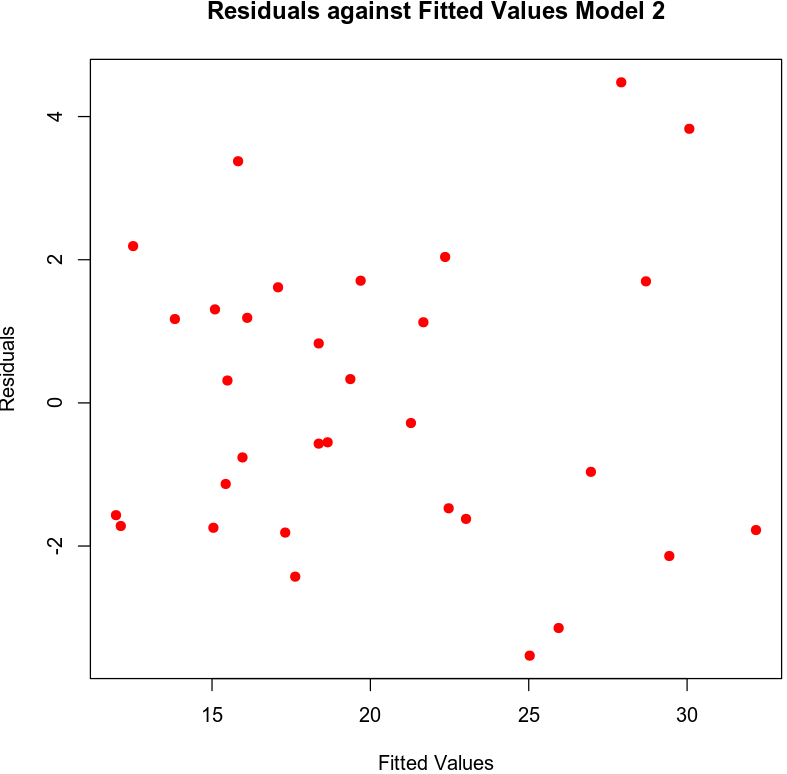
*The equation for the multiple regression model for fuel economy using weight, horsepower, the interaction term for weight against horsepower, and the number of cylinders is:*

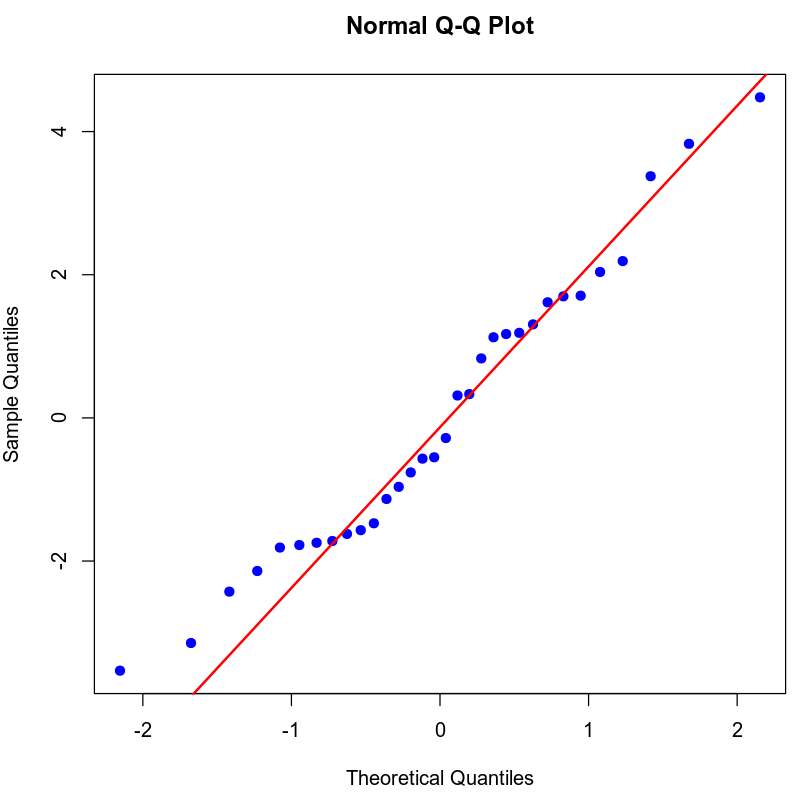
*Our is our intercept. Our values are our popular regression parameters from our model for weight, horsepower, the dummy variable for 6 cylinders, the dummy variable for 8 cylinders, and the interaction term for weight against horsepower respectively. Our values are weight, horsepower, and our dummy value denoting the qualitative value for number of cylinders ( for 6 cylinders and for 8 cylinders.*

*The equation for our regression model is:*

*Our value of . Our Adjusted R-Squared value is . Since our of is our coefficient of determination, we can conclude that approximately 89% of the variation in fuel economy can be explained by the model using predictors for weight, horsepower, number of cylinders and the interaction term of weight against horsepower.*

*We obtained our fitted values and residuals using model 2 and created a plot for Residuals against fitted values and a Normal Q-Q plot. I am including screenshots of them below:*

**

**

*Based on these plots I would say there is an assumption of homoscedasticity based on the ‘Residuals against fitted values’ plot showing no discernible pattern. I would conclude that the Q-Q plot has normally distributed residuals, as there is not significant deviation from the line with the exception of one outlier.*

### Evaluating Model Significance

*In evaluating model significance for the regression model, where we want to see if the model is significant at a 5% level of significance, we perform an overall F-test by identifying our null hypothesis and alternative hypothesis in the following format:*

*Our P-value is 1.503e-11, which is far below our level of significance of 0.05. This means we should reject our null hypothesis in favor of our alternative hypothesis. The conclusion of this test tells us that a statistically significant relationship exists between fuel economy and at least one of our predictor variables.*

*In looking to see which terms in the model are significant at a 5% level of significance, we will perform individual beta tests on each variable. For this our null and alternative hypotheses are as follows:*

*Our P-value for weight is , our P-value for horsepower is , our P-value for 6 cylinders is , our P-value for 8 cylinders is , and our P-value for the interaction term of weight against horsepower.*

*Our P-values for weight, horsepower and for our interaction term of weight against horsepower are all below our level of significance of 0.05. As such, we should reject our null hypothesis in favor of our alternative hypothesis and conclude that each of these variables has a statistically significant relationship with fuel economy.*

*Our P-values for 6 cylinders and 8 cylinders are not below our level of significance of 0.05 and therefore, we should not reject our null hypothesis. We can conclude that no statistically significant relationship exists between either of these variables and fuel economy.*

### Making Predictions Using the Model

*Using the second model (model 2), we can predict the fuel economy for a car that has a weight of 2.965, a horsepower of 210, and 6 cylinders. The model tells us the predicted fuel economy of a vehicle with these hypothetical variables is mpg.*

*The 95% prediction interval for an individual for fuel economy of this car is .*

*The prediction interval for an individual response tells us we can be 95% certain that a car’s fuel economy (in miles per gallon) will fall within these lower and upper bounds, if the car’s weight is 2.965, its horsepower is 210 and it has 6 cylinders. This is taking into account uncertainty related to  varying according to the regression error, , as well as sampling uncertainty related to estimating the regression parameters.*

*Taking the regression error into account as well as sampling uncertainty relating to estimating the regression parameters is the reason the prediction interval for an individual response is wider than the confidence interval for the mean. (Zybooks, 2016)*

*The 95% confidence interval for the mean for fuel economy of this car is .*

*This confidence interval for the mean tells us we can be 95% certain that the average fuel economy (in miles per gallon) of a group of cars will be within these lower and upper bounds; if they have a weight of 2.965, a horsepower of 210 and have 6 cylinders.*

*This is taking into account sampling uncertainty related to estimating the regression parameters. Because it does not also take the regression error into account, the confidence interval is not as wide as the prediction interval.(Zybooks, 2016)*

## 5. Conclusion

*Based on the analysis I have performed, as long as the sample size was sufficiently large, I am honestly unsure which model I would prefer to use. They both include variables that seem to have no statistically significant relationship with fuel economy. The first model’s variables for rear axle ratio and the interaction term of weight against rear axle ration both do not have a statistically significant relationship with fuel economy. The second model’s variables for 6 cylinders and 8 cylinders both do not have a statistically significant relationship with fuel economy.*

*As such I personally would use a model that measures just weight, horsepower, and an interaction term of weight against horsepower. This is looking at the conclusion that these are the only variables that have a significant relationship with fuel economy and these variables are shared in both models. I would love to try a model with these variables and add other variables from the data set, with some more interaction variables, to see if any variables can be added that would also have a statistically significant relationship with fuel economy.*

*The practical importance of the analyses that were performed is that they can help car manufacturers test which attributes of a vehicle have a significant effect on fuel economy and whether or not two attributes interacting with each other have a significant effect. This can be useful if a car manufacturer is, for example, looking to see how lowering weight will affect fuel economy. They can look at how weight and horsepower interact with each other to ensure that the resulting improvement in fuel economy doesn’t negatively affect horsepower when lowering the weight of the vehicle.*

## 6. Citations

Zybooks MAT 303: Applied Statistics II for Science, (2016, August).

Retrieved March 16, 2020, from https://learn.zybooks.com/zybook/SNHUMAT303v1