# MAT 303 Module Two Problem Set Report

Interaction Terms and Qualitative Predictors

Sarah Steinbaum

Sarah.steinbaum@snhu.edu

Southern New Hampshire University

## 1. Introduction

The following statistical analyses will be used to determine a vehicle’s fuel efficiency in miles per gallon based on variables such as horsepower, quarter mile time for acceleration, rear axle ratio for torque, and the number of cylinders of the engine. The data set consists of 32 vehicles, one for each row, and 12 variables, one for each column. Two multiple regression models will be used to allow us to determine if there is a linear relationship between fuel efficiency and these variables. If a relationship does exist, we will conduct individual beta tests to determine which variables or terms are statistically significant. The analyses will also entail finding the fitted values, residuals, plotting the fitted values against the residuals through a scatterplot, creating Q-Q plots, and finding confidence and prediction intervals. The results of these analyses will allow vehicle manufacturers to determine the balance between the vehicle’s features and fuel efficiency in order to produce a product that will increase vehicle sales.

## 2. Data Preparation

The important variables in this data set include fuel econmy (mpg) in US miles per gallon, horsepower (hp), quarter mile time (qsec) which measures the vehicle’s acceleration, and rear axle ratio (drat) which indicates the number of rotations the of drive shaft, the higher the rear axle ratio, the higher the torque. The data set also includes the number of cylinders (cyl) which is a qualitative predictor variable. The number of cylinders can be 4, 6, or 8. The data set contains 12 columns, one variable for each column and 32 rows each resprenting a vehicle.

## 3. Model with Interaction Term

### Correlation Analysis

The Pearson Correlation Coefficients illustrates the relationship between fuel economy, horsepower, quarter mile time, and rear axle ratio. The correlation coefficient between fuel economy and horsepower is -0.7762 which indicates a moderately negative relationship. The correlation coefficient between fuel economy and quarter mile time is 0.4187. The correlation coefficient between fuel economy and rear axle ratio is 0.6812 which indicates a moderate positive correlation.

### Reporting Results

The general form of this regression model:

The prediction equation of this regression model:

In this regression model, y represents fuel efficiency, represents horsepower, represents quarter mile time, and represents rear axle ratio. The interaction term for horsepower and quarter mile time is represented by and the interaction term for horsepower and rear axle ratio is represented by . Once the R script has ran, we can place the beta estimates into the equation:

The value of R-squared is 0.8207. R-squared is the coefficient of multiple determination which means that roughly 82% of the variation in the response variable, fuel economy, is explained by the predictors horsepower, quarter mile time, and rear axle ratio as well as the interaction terms in the regression model above. The Adjusted R-squared is the adjusted coefficient of multiple determination and its value is 0.7862.

In order to estimate the change in fuel economy of a car with 160 horsepower for each unit increase in quarter mile time, we need to use a portion of the model as the equation involves interaction terms. Since we are evaluating the variable quarter mile time (qsec) which is represented by , we need to use the beta estimates for and .

= (1.509555 – 0.018723(160))

= -1.486125 mpg

This means that the fuel economy of a car with 160 horsepower is estimated to decrease by 1.486125 for every unit increase in quarter mile time. In order to estimate the change in fuel economy for a car with 160 horsepower for each unit increase in rear axle ratio, we will use a portion of the model as this equation also involves interaction terms. Since we are evaluating the variable rear axle ratio (drat) which is represented by , we need to use the beta estimates for and .

= (160))

= 0.273024 mph

This means that the fuel economy of a car with a horsepower of 160 will increase by 0.273024 mpg for every unit increase in rear axle ratio.

Chart, scatter chart

Description automatically generated

Chart, line chart, scatter chart

Description automatically generated

The charts above can assist in making assumptions about the homoscedasticity and normality of the residuals. The first chart is Residuals against Fitted Values which uses the predictor variables and portrays constant variance with no discernable pattern, indicating homoscedasticity. The second chart is the Normal Q-Q Plot which portrays that the values are normally distributed as most are on or near the linear trendline, although there is a slight deviation as the values increase.

### Evaluating Model Significance

An overall F-test can be conducted in order to determine if the model is significant at a 5%, or 0.05, level of significance. The null hypothesis is that no linear relationship exists between the response variable, fuel economy, any of the predictor variables or interaction terms. The alternative hypothesis is that a relationship exists with at least one predictor variable or term.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable | Hypothesis | Test Statistic | P-value | Conclusion |
| Fuel economy  (mpg) | = ….. = = 0  for *i* = 1,2,…,*n* | *t* = - 0.900 | *p* = 6.098E-09 | Reject the null hypothesis. |

The P-value is 6.098E-09 which is less than the level of significance of 5% or 0.05. This indicates that the model is statistically significant and there is sufficient evidence to reject the null hypothesis. There is a linear relationship between the response variable, fuel economy, and at least one predictor variables and interaction terms.

In order to find what predictor variables and interaction terms in the model are significant at a level of significance of 5% or 0.05, we will conduct individual beta tests. The null hypothesis is that no relationship exists between the response variable, fuel economy, and the predictor variables or terms. The alternative hypothesis is that a linear relationship does exist.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable/  Term | Hypothesis | Test Statistic | P-value | Conclusion |
| Horsepower  (hp) | =0 for *i* =1,2,…*n*  : for *i* =1,2,…,*n* | *t* = 2.710 | *p* = 0.01175 | Reject the null hypothesis |
| Quarter mile time (qsec) | =0 for *i* =1,2,…*n*  : for *i* =1,2,…,*n* | *t* = 2.157 | *p* = 0.04043 | Reject the null hypothesis |
| Rear axle ratio  (drat) | =0 for *i* =1,2,…*n*  : for *i* =1,2,…,*n* | *t* = 2.257 | *p* = 0.03262 | Reject the null hypothesis |
| Horsepower: quarter mile time | =0 for *i* =1,2,…*n*  : for *i* =1,2,…,*n* | *t* = - 3.256 | *p* = 0.00307 | Reject the null hypothesis |
| Horsepower: rear axle ratio | =0 for *i* =1,2,…*n*  : for *i* =1,2,…,*n* | *t* = - 1.796 | *p* = 0.08405 | Fail to reject the null hypothesis |

The interaction term horsepower: rear axle ratio is greater than the level of significance of 5% or 0.05 which indicates that this term is not statistically significant. However, the predictor variables horsepower, quarter mile time, rear axle ratio, and the interaction term horsepower: quarter mile time are less the 5% or 0.05 level of significance which provides us with sufficient evidence to reject the null hypothesis. These terms are statistically significant which means that there is a relationship with the response variable, fuel economy.

### Making Predictions Using the Model

The predicted fuel economy for a car that has a 175 horsepower, 14.2 quarter mile time and 3.91 rear axle ratio is 21.5278133 mpg and is expressed in the equation below:

= 21.5278133 mpg

The 95% prediction interval for the fuel economy of this same car is (15.0897 mpg, 27.9674 mpg). This indicates that there is a 95% likelihood that a vehicle with these factors will have a fuel economy that falls within these bounds. The 95% confidence interval for the fuel economy of this car is (18.5881 mpg, 24.469 mpg). This means that we are 95% confident that the average fuel economy will fall between this range if the sample were conducted repeatedly using these particular features.

## 4. Model with Interaction Term and Qualitative Predictor

### Reporting Results

The general form of the regression model:

+

The prediction equation of the regression model:

+

In the equations above, y represents fuel economy (mph), represents horsepower (hp), represents quarter mile time (qsec), represents the interaction term for horsepower and quarter mile time, and and represent the number of cylinders (cyl) which is a qualitative predictor. The number of cylinders can be either 4, 6, or 8. In this equation, represents 6 cylinders and represents 8 cylinders. Once the R script has ran, we can place the beta estimates into the equation:

– 4.580823

R-squared is the coefficient of determination and its value for this mode is 0.8327. This means that roughly 83% of the variation for fuel economy can be explained by the predictor variables and the qualitative predictor variables. The value for the Adjusted R-squared 0.8005.

Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

Based on the charts above, assumptions can be made regarding homoscedasticity and normalcy of the residuals. The Residuals against Fitted Values chart displays values in no discernable pattern which is a great indicator for homoscedasticity. The Normal Q-Q Plot displays values that are normally distributed as most are on or near the linear trendline.

### Evaluating Model Significance

In order to determine if the model is statistically significant at a 5%, or 0.05, level of significance, we can conduct an overall F-test. The null hypothesis is that no relationship exists between the response variable, fuel economy, and the predictor variables, interaction terms, or qualitative variables. The alternative hypothesis is that a relationship does exist between the response variable and at least one of the variables or terms.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable | Hypothesis | Test Statistic | P-value | Conclusion |
| Fuel economy  (mpg) | = ….. = = 0  for *i* =1,2,…,*n* | *t* = 1.858 | *p* = 2.562E-09 | Reject the null hypothesis |

The p-value is 2.562E-09 which is less than the level of significance of 5%. This indicates that this model is statistically significant and that a relationship does exist between at least one of the variables or terms. There is enough evidence to reject the null hypothesis.

In order to find which variables and/or terms in the model are significant at a level of significance of 5%, we will conduct individual beta tests. The null hypothesis is that the null hypothesis is that no relationship exists between the response variable and any of the predictor variables, interaction terms, or qualitative variables. The alternative hypothesis is that a relationship does exist.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable/  Term | Hypothesis | Test Statistic | P-value | Conclusion |
| Horsepower  (hp) | =0 for *i* =1,2,…*n*  : for *i* =1,2,…,*n* | *t* = 1.792 | *p* = 0.0848 | Fail to reject the null hypothesis |
| Quarter mile time  (qsec) | =0 for *i* =1,2,…*n*  : for *i* =1,2,…,*n* | *t* = 0.712 | *p* = 0.4828 | Fail to reject the null hypothesis |
| Horsepower: Quarter mile time | =0 for *i* =1,2,…*n*  : for *i* =1,2,…,*n* | *t* = - 2.386 | *p* = 0.0246 | Reject the null hypothesis |
| Number of cylinders (6) | =0 for *i* =1,2,…*n*  : for *i* =1,2,…,*n* | *t* = -2.708 | *p* = 0.0118 | Reject the null hypothesis |
| Number of cylinders (8) | =0 for *i* =1,2,…*n*  : for *i* =1,2,…,*n* | *t* = - 1.792 | *p* = 0.0847 | Fail to reject the null hypothesis |

The interaction term horsepower: quarter mile time, and the qualitative predictor number of cylinders for 6 cylinders both have a p-value of less than the 5% or 0.05 level of significance. This indicates they are statistically significant and provide us with enough evidence to reject the null hypothesis. We can now determine that there is a statistical relationship between fuel economy and horsepower: quarter mile time as well as number of cylinders for 6 cylinders.

### Making Predictions Using the Model

The predicted fuel economy for a car that has 175 horsepower, 14.2 quarter mile time and 6 cylinders would be 21.3424 and is expressed in the equation below:

= 21.3424 mpg

The 95% prediction interval for the fuel economy of this same car is (14.8764 – 27.8085 mpg) which indicates that the likelihood that a vehicle with these characteristics will have a fuel economy that falls between these bounds. The 95% confidence interval for the fuel economy of this car is (17.9965 – 24.6884 mpg). This means that we are 95% confident that the average fuel economy will fall between this range if the sample were conducted repeatedly using these features. Prediction intervals generally have larger sample sizes because they include the random variation of values as well as uncertainty in the population mean, making them wider compared to confidence intervals.

## 5. Conclusion

In conclusion, we conducted an analysis using two different multiple regression models in order to determine if certain characteristics of a vehicle are correlated to fuel efficiency. In the first model, we conducted an overall F-test and found that there is a correlation between fuel efficiency and at least one of the predictor variables or interaction terms. Next, we conducted individual beta tests to find which variables or terms have an impact of fuel efficiency. We determined that the predictor variables horsepower, quarter mile time, rear axle ratio, and the interaction term horsepower: quarter mile time are statistically significant which means that there is a relationship with the response variable, fuel economy.

The second model added a qualitative predictor variable, number of cylinders, using 6 and 8 cylinders. We conducted an overall F-test to determine if the predictor variables horsepower and quarter mile time, the interaction term horsepower: quarter mile time, and the qualitative predictor number of cylinders for 6 and 8 cylinders have an impact on fuel efficiency. We determined that there is a relationship with at least one variable or term. Next, we conducted individual beta tests to determine which variable or term has an impact on fuel efficiency. The interaction term horsepower: quarter mile time, predictor variable quarter mile time, and qualitative predictor variable number of cylinders for 6 cylinders each have a p-value of less than the 5% level of significance which indicates they are statistically significant and have a correlation to fuel economy. The first model has a R-squared value of 0.8207 and the second model has a R-squared value of 0.8327. The first model has a residual standard error of 2.787 and the second model’s residual standard error was 2.692. Both models’ residual standard error was based on 26 degrees of freedom.

Based on the analysis that were performed and assuming that the sample size is sufficiently large, I would recommend the first model as it is slightly more accurate based on its residual standard error and R-squared values. Overall, these analyses can help automotive manufacturers determine what characteristics of a vehicle impact its fuel efficiency. This can allow the manufacturers to produce a car with the proper balance of horsepower, quarter mile time, and rear axle ratio needed to maintain or improve a vehicle’s fuel efficiency in order to create a vehicle based on customer needs. This can improve sales and create a larger return on investment for the manufacturers.

## 6. Citations

Berrier, J. (2016). MAT 303: Applied Statistics 2 for Science. Zyante Inc. (zyBooks.com)