# MAT 303 Module Two Problem Set Report

Interaction Terms and Qualitative Predictors

Nicholas Kreuziger

nicholas.kreuziger@snhu.edu

Southern New Hampshire University

## 1. Introduction

The data set being explored is from the 1974 Motor Trend US Magazine. This data set touches upon fuel consumption and 10 aspects of automobile design and performance for 32 automobiles. The models of these cars are from the years 1973-1974. The results may be used to determine what variables are most important for a car’s fuel economy. Car makers are interested in the variables that contribute the most to Fuel Economy. The analyses being run will be a Correlation Analysis and a First Order Multiple Regression Analysis with Qualitative Variables and Interaction Terms.

## 2. Data Preparation

There are 11 variables in this data set in a 32 Row and 12 column table. The important variables identified as the predictor variables are weight (wt.), horsepower (hp), rear axle ratio (drat) and cylinders (cyl). Weight is measured in units of 1,000 lbs. If the Weight variable is 2.5 it represents 2,500 lbs. Horsepower, when ft\*lbf and rpm is available, is calculated by the equation . The unit for horsepower is ft∙lbf/min, 3 horsepower of force moves 3 lb. per foot of force per minute. Rear Axle Ratio is measured by the number of Driveshaft Rotations it takes to rotate the Rear Tire once. This is measured by Drive Shaft Rotations : Rear Tire Rotations; In example a Rear Axle Ratio of 3 equates to 3:1 and means it takes 3 drive shaft rotations to turn the Rear Tire once. Cylinders are a measure of how many cylinders are in the engine.

Weight, Horsepower, Rear Axle Ratio and Cylinders are measured differently and fall into two separate categories. Quantitative and Qualitative. Quantitative data is measured numerically. In example, I can measure my weight to be 158.3 pounds. Qualitative data cannot be measured in such a concrete fashion, it is the quality of something or a possessed trait. Using the grocery store example, if I buy 12 eggs, I have 1 dozen eggs. I cannot buy 1.2 or 1.5 dozen eggs. The measure of a dozen is Qualitative. Weight, Horsepower and Rear Axle Ratio are all Quantitative data points. Cylinders are a Qualitative Data point since you can only have 4, 6 or 8 Cylinders. A car cannot have 4.2 Cylinders.

## 3. Model with Interaction Term

### Correlation Analysis

Calculating a Pearson Correlation Coefficient Matrix of the variables’ mpg, weight, horsepower and rear axle ratio helps to define the linear relationship the scatterplots visualize. Zybooks MAT 303: Applied Statistics II for Science Section 3.3 Table 3.3.1 defines a Strong correlation as anything above 0.80, and a Moderate Correlation between 0.40 and 0.80.

Table 3.3.1: Strength of correlation.

|  |  |
| --- | --- |
| Value of |R| | Strength of correlation |
| 0<|R|≤0.40 | Weak |
| 0.40<|R|≤0.80 | Moderate |
| 0.80<|R|≤1.00 | Strong |

Taking these definitions of Correlation Strength into account, we can analyze fuel efficiency and its independent correlation to weight, horsepower, and rear axle ratio.

|  |  |  |
| --- | --- | --- |
| Variable | Pearson Correlation Coefficient | Strength of Correlation |
| Weight | -0.8677 | Strong, Negative |
| Horsepower | -0.7762 | Moderate, Negative |
| Rear Axle Ratio | 0.6812 | Moderate, Positive |

### Reporting Results

The general form of the regression model for fuel economy using weight, horsepower and rear axle ratio as predictor variables including interaction terms for weight and horsepower and weight and rear axle ratio is stated as follows. Included is a definition of the variables.

Y = Fuel Efficiency (mpg)

X1 = Weight (wt)

β1 = Weight Coefficient

X2 = Horsepower (hp)

β2 = Horsepower Coefficient

X3 = Rear Axle Ratio (drat)

β3 = Rear Axle Ratio Coefficient

β4 = Weight and Horsepower Interaction Coefficient

β5 = Weight and Rear Axle Ratio Interaction Coefficient

Y = β0 + β1X­1 + β2X2 + β3X3 + β4X1X2 + β5X1X­3

This multiple regression model has a summary as follows:

Table

Description automatically generated

This results in a multiple regression equation of:

Y = 75.68 – 16.13X1 – 0.16X2 – 5.45X3 + 0.04X1X2 + 1.71X1X­3

The *R-Squared* value for this model is 89.07% with an adjusted R-Squared *Adjusted R-Squared* value of 86.97%. R-Squared is the variation in the Dependent Variable (MPG) explained by the independent variables (Weight , Horsepower, Rear Axle Ratio, 89.07% of variation in MPG is explained by Weight and Horsepower. Adjusted R-Squared is the variation explained by only the independent variables that significantly help in explaining the dependent variable, Adjusted R-Squared penalizes adding independent variables that do not help in predicting the dependent variable. Adjusted R-Squared is saying that, when considering the correlation of your independent variables to the dependent variables, this model accounts for 86.97% of the variation in MPG.

* *For this model, estimate the change in fuel economy of a car with weight 3.50 for each unit increase in horsepower. Explain your answer.*
* *Now estimate the change in fuel economy of a car with weight 3.50 for each unit increase in rear axle ratio. Explain your answer.*

Using this model, we can predict the change in fuel economy for a car with a fixed weight of 3.50 when a unit of Horsepower is added or subtracted. To do so the term X1 (weight) must be factored in the multiple regression model. Thus,

*Model*

Y = 75.68 – 16.13X1 – 0.16X2 – 5.45X3 + 0.04X1X2 + 1.71X1X­3

*Terms with X2*

– 0.16X2 + 0.04X1X2

*Substitute X2 = 1 and X1 = 3.50*

-0.16(1) + 0.04(1)(3.50)

-0.16 + 0.14

-0.02

For every unit of Horsepower with a fixed weight of 3.50 the fuel efficiency will decrease by 0.02 miles per gallon.

This model can also be used to predict the change in fuel economy for a car with a fixed weight of 3.50 for a unit increase in rear axle ratio. As follows,

*Model*

Y = 75.68 – 16.13X1 – 0.16X2 – 5.45X3 + 0.04X1X2 + 1.71X1X­3

*Terms with X3*

– 5.45X3 + 1.71X1X­3

*Substitute X3 = 1 and X1 = 3.50*

-5.45(1) + 1.71(3.50)(1)

-5.45 + 5.985

0.535

For every unit of Rear Axle Ratio with a fixed weight of 3.50 the fuel efficiency will increase by 0.535 miles per gallon.

Evaluating this model, we will use a scatterplot of Residual against Fitted Values and a Q-Q Plot.

Chart, scatter chart

Description automatically generated

A scatterplot plotting Residual values against Fitted Values evaluates the Mean of Zero and Constance Variance assumptions of a multiple regression model. The Mean of Zero assumption assumes that the residual values in a model are zero (the response variable is linear in response to predictor variables), if they are not then the model may be nonlinear. Constant variance assumes that the residuals for each of the predictor variables should have equal or similar variance, called homoscedasticity. Heteroscedasticity is a condition where unequal and unrelated variance occurs. The scatterplot for our model does not appear to have any plot patterns obstructing linearity nor patterns in variance.

Chart, scatter chart

Description automatically generated

Above is a Q-Q plot, a Q-Q plot is a visualization of the distribution of errors. The Q stands for Quantiles, which are continuous intervals with equal probabilities across a probability distribution. A normalized set of errors will follow a linear pattern. The Q-Q plot for this model indicates normally distributed residuals. The 2nd standard deviation of distribution appears to be straying from the Q-Line, indicating the data is right-skewed (as demonstrated by the histogram below). The model tends to overestimate MPG by 1 to 4 units.

Chart, histogram

Description automatically generated

### Evaluating Model Significance

Evaluating this model’s significance at a 5% level, we will be using the ANOVA (Analysis of Variance) F-Test. An F-test applied to a multiple regression model is used to determine if the overall model and its variables are collectively influencing the response variable in a statistically significant manner. After determining the results of the F-Test, if a linear relationship exists you can evaluate individual relationships in the model with a multiple regression individual t-test. This is useful when you want to troubleshoot or assess the value of certain predictor variables in a model.

Table

Description automatically generated

The Overall F-Test is testing if the response variable has a linear relationship with at least one of the predictor variables. The Significance Level of our test is 5%, mathematically α = 0.05.

The Null hypothesis, H0 is that the predictor variables have no linear relationship with the response variable. Mathematically stated, the beta coefficient of predictor variables Weight and Horsepower are equal to 0.

H0 : β1 = β2 = β3 = β4 = β5 = 0

The Alternative hypothesis, Ha is that at least one predictor variable has a linear relationship with the response variable MPG. Mathematically stated, the beta coefficient of predictor variables Weight and Horsepower are not equal to 0.

Ha : βi ≠ 0 for *i* = 1, 2, 3, 4, 5

The P-Value of the Overall F-Test is 1.092 e-11. Since P-Value < Significance Level the null hypothesis is rejected in favor of the alternative hypothesis. At least one predictor variable has a linear relationship to the response variable MPG.

Individual t-tests can be used once the F-Test has determined the model has a predictor variable with a linear relationship to the response variable. Every individual t-test has a similar Null and Alternative Hypothesis. We will be testing with a significance level of 5%, mathematically α = 0.05.

The Null Hypothesis, H0, is that the predictor variable coefficient βi (i = 1, 2, 3, 4, 5 for each individual test) = 0. The Alternative Hypothesis, Ha , is that the predictor variable coefficient βi  (i = 1, 2, 3, 4, 5 for each individual test) ≠ 0. The t-test results indicate weight, horsepower, wt:hp are all statistically significant. The variables rear axle ratio (drat) and weight:rear axle ratio (wt:drat) are not statistically significant.

### Making Predictions Using the Model

A prediction interval is used to predict what range a future individual observation may fall in all outcomes. A confidence interval is used to predict the range the average of the response variable will fall. In short, a prediction interval tries to account for all possible values with the model for a single observation while a confidence interval tries to predict where the average of the probability distribution will fall given specific values for the predictor variables. Confidence intervals will always be narrower than Prediction intervals since a Confidence interval is the estimation of a range of an average while the predictor interval is the range of values for a single instance.

Using this model we can predict the fuel economy for a car with 2.965 weight, 210 horsepower and 2.91 rear-axle ratio will have 19.8443 mpg fuel efficiency.

The 95% prediction interval for the car in question is

A picture containing table

Description automatically generated

The prediction interval lower and upper ranges are 13.5295 mpg and 26.159 mpg. If we were to observe a car with a weight of 2.965, horsepower of 210 and a 2.91 rear axle ratio we would observe MPG values within this lower and upper range 95% of the time.

Table

Description automatically generated with low confidence

The confidence interval lower and upper ranges are 17.813 to 22.3072. If we were to observe a car with a weight of 2.965, horsepower of 210 and a 2.91 rear axle ratio repeatedly, we would arrive at an average MPG value between 17.813 to 22.3072. This range is predicted to be a part of the true population 95% of the time.

## 4. Model with Interaction Term and Qualitative Predictor

### Reporting Results

The model above could use some improvement based upon the results of the individual t-tests following the F-Test. The dataset has a qualitative value of cylinders that we will use while also removing the variables deemed to be statistically insignificant. A cylinder in a car engine is a chamber that is used to combust fuel into energy, the cylinder harnesses the power of that combustion. Hence the phrase “internal combustion engine”. We will also retain the variables that were statistically significant in the F-Test and Subsequent T-tests (weight, horsepower, weight:horsepower). We will call this model 2.

When using Qualitative variables dummy variables must be used. These are variables that indicate 1 or 0 for the presence of that quality. Which means X3 = Dummy Cylinders Variable 1 and X4 = Dummy Cylinders Variable 2, β4 = Dummy Cylinders Variable 1s coefficient and β5 = Dummy Cylinders Variable 2 coefficient. The chart below demonstrates the meaning of X3 and X4 values using 4 cylinders as the reference level. For ease of labelling, we will label the variables and their associated coefficients with the cylinder they pertain to when marked 1.

|  |  |  |
| --- | --- | --- |
| Cylinders | X3 | X4 |
| 4 | 0 | 1 |
| 6 | 1 | 0 |
| 8 | 0 | 0 |

Y = Fuel Efficiency (mpg)

X1 = Weight (wt)

β1 = Weight Coefficient

X2 = Horsepower (hp)

β2 = Horsepower Coefficient

β3 = Weight and Horsepower Interaction Coefficient

β4 = 6 Cylinders Coefficient (cyl)

X3 = 6 Cylinders Variable

β5 = 4 Cylinders Coefficient

X4 = 4 Cylinders Variable

Y = β0 + β1X­1 + β2X2 + β3X1X2 + β4X3 + β5X4

Model 2 Statistical Results are:

Table

Description automatically generated

This results in a model of:

Y = 47.33 - 7.31X­1 - 0.10X2 + 0.02X1X2 - 1.26X3 - 1.45X4

The *R-Squared* value for this model is 88.8% with an adjusted R-Squared *Adjusted R-Squared* value of 86.64%. R-Squared is the variation in the Dependent Variable (MPG) explained by the independent variables (Weight , Horsepower, cylinders), 88.8% of variation in MPG is explained by Weight , Horsepower, Weight:Horsepower and Cylinders data. Adjusted R-Squared is the variation explained by only the independent variables that significantly help in explaining the dependent variable, Adjusted R-Squared penalizes adding independent variables that do not help in predicting the dependent variable. Adjusted R-Squared is saying that, when considering the correlation of your independent variables to the dependent variables, this model accounts for 86.64% of the variation in MPG.

Chart, scatter chart

Description automatically generated

A scatterplot plotting Residual values against Fitted Values evaluates the Mean of Zero and Constance Variance assumptions of a multiple regression model. The Mean of Zero assumption assumes that the residual values in a model are zero (the response variable is linear in response to predictor variables), if they are not then the model may be nonlinear. Constant variance assumes that the residuals for each of the predictor variables should have equal or similar variance, called homoscedasticity. Heteroscedasticity is a condition where unequal and unrelated variance occurs. The scatterplot for our model does not appear to have any plot patterns obstructing linearity nor patterns in variance.

Chart, line chart

Description automatically generated

Above is a Q-Q plot, a Q-Q plot is a visualization of the distribution of errors. The Q stands for Quantiles, which are continuous intervals with equal probabilities across a probability distribution. A normalized set of errors will follow a linear pattern. The Q-Q plot for this model indicates normally distributed residuals. There does appear to be one outlier worth exploring. The histogram below demonstrates that the model is regularly predicting mpg one unit above or one unit below the observed value.

Chart, histogram

Description automatically generated

### Evaluating Model Significance

Table

Description automatically generated

The Overall F-Test is testing if the response variable has a linear relationship with at least one of the predictor variables. The Significance Level of our test is 5%, mathematically α = 0.05.

The Null hypothesis, H0 is that the predictor variables have no linear relationship with the response variable. Mathematically stated, the beta coefficient of predictor variables Weight and Horsepower are equal to 0.

H0 : β1 = β2 = β3 = β4 = 0

The Alternative hypothesis, Ha is that at least one predictor variable has a linear relationship with the response variable MPG. Mathematically stated, the beta coefficient of predictor variables Weight and Horsepower are not equal to 0.

Ha : βi ≠ 0 for *i* = 1, 2, 3, 4

The P-Value of the Overall F-Test is 1.503 e-11. Since P-Value < Significance Level the null hypothesis is rejected in favor of the alternative hypothesis. At least one predictor variable has a linear relationship to the response variable MPG.

Individual t-tests can be used once the F-Test has determined the model has a predictor variable with a linear relationship to the response variable. Every individual t-test has a similar Null and Alternative Hypothesis. We will be testing with a significance level of 5%, mathematically α = 0.05.

The Null Hypothesis, H0, is that the predictor variable coefficient βi (i = 1, 2, 3, 4 for each individual test) = 0. The Alternative Hypothesis, Ha , is that the predictor variable coefficient βi  (i = 1, 2, 3, 4 for each individual test) ≠ 0. The t-test results indicate weight, horsepower, wt:hp are all statistically significant. The qualitative variables for cylinders did not result in additional significance in the model.

### Making Predictions Using the Model

A prediction interval is used to predict what range a future individual observation may fall in all outcomes. A confidence interval is used to predict the range the average of the response variable will fall. In short, a prediction interval tries to account for all possible values with the model for a single observation while a confidence interval tries to predict where the average of the probability distribution will fall given specific values for the predictor variables. Confidence intervals will always be narrower than Prediction intervals since a Confidence interval is the estimation of a range of an average while the predictor interval is the range of values for a single instance.

Using this model we can predict the fuel economy for a car with 2.965 weight, 210 horsepower and 6 cylinders is 17.6286.

The 95% prediction interval for the car in question is

A picture containing text

Description automatically generated

The prediction interval lower and upper ranges are 12.3883 mpg and 22.869 mpg. If we were to observe a car with a weight of 2.965, horsepower of 210 and six cylinders we would observe MPG values within this lower and upper range 95% of the time.

Diagram

Description automatically generated with medium confidence

The confidence interval lower and upper ranges are 14.9904 to 20.2669. If we were to observe a car with a weight of 2.965, horsepower of 210 and 6 cylinders repeatedly, we would arrive at an average MPG value between 14.9904 to 20.2669. This range is predicted to be a part of the true population 95% of the time.

## 5. Conclusion

In conclusion, let us summarize in broad strokes the results of the two models using the table below.

|  |  |  |  |
| --- | --- | --- | --- |
| Model | R^2 | Statistically Significant Coefficients | Statistically Insignificant Coefficients |
| 1 | 89.07% | 3 | 2 |
| 2 | 88.80% | 3 | 2 |

Models 1 and 2 both accounted for just shy of 90% of the variance of their perspective variables. Each had the same amount of statistically significant and insignificant coefficients. It is difficult to conclude the better model from this data alone. However, if we peer at the histograms once more,

**Model 1**

Chart, histogram

Description automatically generated

**Model 2**

Chart, histogram

Description automatically generated

It boils down to a choice between over-estimating mpg by 1-4 consistently or over/under-estimating mpg by 1-2 consistently. In this regard my opinion errs towards utilizing Model 2, Model 2 will consistently be within a 2-mpg residual.

Practically speaking, the importance of these analyses is discovering factors that directly predict fuel efficiency. This topic relates to many different sects of industry and consumer preference. These Analyses also introduce the possibility of an anomaly in car manufacturing directly impacting fuel efficiency. The first model is unimodal and right skewed. The second model is bimodal. My biggest takeaway is that more models must be experimented with to arrive at a consistent linear relationship between all involved variables.