# MAT 303 Module Two Problem Set Report

Interaction Terms and Qualitative Predictors

Andrea Plunkett Jackson

andrea.jackson11@snhu.edu

Southern New Hampshire University

## 1. Introduction

The statistical analyses being performed in this problem set are focused on exploring the relationship between a vehicle’s fuel economy (measured in miles per gallon, mpg) and several variables. Specifically, we are examining the relationship between fuel economy and horsepower (hp), fuel economy and quarter mile time (qsec), and fuel economy and rear axle ratio (drat).

These results could be of significant value to automobile manufacturers. By understanding how these variables individually and collectively impact fuel economy, manufacturers can make informed decisions during the design process to optimize fuel efficiency.

Regarding the specific analyses I’m conducting, I will be running multiple regression models. The first model compares fuel efficiency (in miles per gallon) against weight, horsepower, and gear ratio. The second model also uses multiple regression but compares fuel efficiency against weight, horsepower, and the number of cylinders. In addition to these, I will be calculating fitted values, finding residuals, plotting the fitted values against the residuals, creating a q-q plot to test the assumptions of normality of the residuals, and determining the confidence intervals for each model. All of this information will be used to draw conclusions about whether or not the variables used in each model have a statistically significant effect on fuel economy. This comprehensive analysis will provide a robust understanding of the factors influencing fuel economy in vehicles.

## 2. Data Preparation

In this problem set, I’m focusing on several key variables that are believed to influence a vehicle’s fuel economy. These variables include miles per gallon (mpg), horsepower (hp), quarter mile time (qsec), and rear axle ratio (drat). My goal is to determine whether these variables - horsepower, quarter mile time, and rear axle ratio - have a significant impact on fuel economy.

Regarding the structure of the data set, it consists of 12 columns, each representing a different variable related to the vehicles. The data set includes 32 rows, with each row corresponding to a specific vehicle. This comprehensive data set will allow me to conduct a thorough analysis of the factors influencing fuel economy.

## 3. Model with Interaction Term

### Correlation Analysis

In the course of this analysis, I computed the Pearson Correlation Coefficients, a measure of the linear correlation between two variables. The results of these calculations are as follows:

* For fuel economy and horsepower, we have a strong negative correlation with a Pearson Correlation Coefficient (R) of -0.7762. This suggests that as horsepower increases, fuel economy tends to decrease.
* For fuel economy and quarter mile time, we have a moderate positive correlation with an R value of 0.4187. This indicates that vehicles with faster quarter mile times tend to have better fuel economy.
* For fuel economy and rear axle ratio, we have a moderate positive correlation with an R value of 0.6812. This suggests that vehicles with higher rear axle ratios tend to have better fuel economy.

These correlations provide valuable insights into how these variables might influence a vehicle’s fuel economy.

### Reporting Results

The general form of the regression model is:

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

mpg=​+ + ​ \* qsec + ​ \* drat + \* hp \* qsec + ​ \* hp \* drat

Where:

mpg is the miles per gallon (fuel efficiency)

hp is the horsepower

qsec is the quarter mile time

drat is the rear axle ratio

The prediction equation, based on the outputs obtained from the R script, is:

mpg=−14.529137 + 0.352800 \* hp + 1.509555 \* qsec + 5.666624 \* drat −0.018723 \* hp \* qsec −0.033246 \* hp \* drat

The interaction terms represent the combined effect of horsepower with quarter mile time and rear axle ratio on fuel efficiency. This allows me to capture not just the individual effects of these variables, but also how they work together to influence fuel efficiency.

The R2 value for the model is 0.8207, and the adjusted R2 value is 0.7862. These values suggest that the model explains a substantial portion of the variability in mpg.

For a car with 160 horsepower, each unit increase in quarter mile time would result in a change in fuel economy of 0.352800 − 0.018723 \* 160 hp which equals approximately -2.65 mpg, and each unit increase in rear axle ratio would result in a change in fuel economy of 0.352800 −0.033246 \* 160 hp which equals approximately -5.0674 mpg. This means that for a car with 160 horsepower, each unit increase in quarter mile time would decrease the fuel economy by about 2.65 miles per gallon, and each unit increase in rear axle ratio would decrease the fuel economy by about 5.07 miles per gallon. These results provide valuable insights into how changes in quarter mile time and rear axle ratio might affect the fuel economy of a car with a given horsepower.

The scatterplot of residuals against fitted values provides a visual representation of the residuals and their relationship with the predicted fuel economy. This offers valuable insights into the accuracy of the model’s predictions and guides further refinements to improve its predictive accuracy. It also underscores the potential complexities in predicting fuel economy, which may be influenced by a range of factors including horsepower, quarter mile time, and rear axle ratio. These factors may not follow a simple linear relationship with fuel economy, as indicated by our model that includes interaction terms. This complexity is a reminder of the multifaceted nature of vehicle fuel efficiency and the challenges in predicting it accurately.

A graph with red dots

Description automatically generated

A graph with red dots

Description automatically generated

Based on the scatter plot of residuals against fitted values, the residuals appear to be somewhat evenly distributed across different fitted values, with no clear pattern or funnel shape. This would indicate that the variance of the residuals is constant, satisfying the assumption of homoscedasticity without violation.

Based on the Normal Q-Q Plot the residuals seem to deviate from the reference line at both ends of the plot, suggesting potential deviations from normality, especially in the tails. This could indicate potential outliers or heavy tailenders in the distribution of residuals. However, the majority of points do follow the trend of the reference line, indicating some adherence to normality.

### Evaluating Model Significance

In evaluating the significance of a multiple regression model and its predictors, specifically horsepower (hp), quarter mile time (qsec), and rear axle ratio (drat), in predicting fuel efficiency (mpg), I am using a 5% significance level for these tests.

For the overall model, the equation for our overall F-test is:

: ​= ​ = ​ ​= ​= = 0

​: at least one = 0 for i=1,2,3,4,5

I am checking if at least one predictor variable significantly affects fuel efficiency. The P-value obtained for the F-statistic is extremely small (6.098e-09), which is much less than the 5% significance level. This leads me to reject the null hypothesis (that all predictor variables have no effect) and accept the alternative hypothesis (that at least one predictor variable does have an effect). In simpler terms, this means that either horsepower, quarter mile time, rear axle ratio, or their interactions have a significant impact on fuel efficiency.

Next, I individually test each predictor variable for significance. The equation for these tests is:

: ​= 0 for some i=1,2,3,4,5

​: = 0

From the model summary, the p-value for hp is 0.01175, for qsec is 0.04043, for drat is 0.03262, and for hp:qsec is 0.00307. All these p-values are less than 0.05, so we reject the null hypotheses for these terms and conclude that hp, qsec, drat, and the interaction term hp:qsec are significant at a 5% level of significance.

The p-value for hp:drat is 0.08405, which is greater than 0.05. Therefore, we fail to reject the null hypothesis for this term, suggesting that the interaction term hp:drat is not significant at a 5% level of significance.

This suggests that horsepower, quarter mile time, rear axle ratio, and the interaction between horsepower and quarter mile time have a significant relationship with fuel efficiency.

The 95% confidence intervals for the parameter estimates of these variables are as follows:

* For horsepower (hp), the coefficient estimate is 0.352800, suggesting that as horsepower increases, fuel efficiency increases, assuming other variables are held constant.
* For quarter mile time (qsec), the coefficient estimate is 1.509555, suggesting that as quarter mile time increases, fuel efficiency also increases, assuming other variables are held constant.
* For rear axle ratio (drat), the coefficient estimate is 5.666624, suggesting that as rear axle ratio increases, fuel efficiency also increases, assuming other variables are held constant.
* For the interaction term between horsepower and quarter mile time (hp:qsec), the coefficient estimate is -0.018723, suggesting that the effect of horsepower on fuel efficiency changes when considering quarter mile time.

These results provide valuable insights into the relationships between these predictors and fuel efficiency and can guide further refinements of the model.

### Making Predictions Using the Model

Based on the regression model, I can now make some predictions. For our scenario, I will use the hypothetical values of horsepower = 175, quarter mile time = 14.2, and rear axle ratio = 3.91. The equation for this is:

mpg=​+ + ​ \* qsec + ​ \* drat + \* hp \* qsec + ​ \* hp \* drat

Where:

mpg is the miles per gallon (fuel efficiency)

hp is the horsepower

qsec is the quarter mile time

drat is the rear axle ratio

The prediction equation, based on the outputs obtained from the R script, is:

mpg=−14.529137 + 0.352800 \* hp + 1.509555 \* qsec + 5.666624 \* drat −0.018723 \* hp \* qsec −0.033246 \* hp \* drat

The predicted fuel economy (mpg) for a car that has 175 horsepower, 14.2 quarter mile time, and 3.91 rear-axle ratio is 21.5285 mpg.

The 95% prediction interval for the fuel economy of this car is from 15.0897 mpg to 27.9674 mpg. The 95% prediction interval for an individual response for fuel economy of this car is (15.0897, 27.9674).

The prediction interval for an individual response tells me I 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 horsepower is 175, its quarter mile time is 14.2 and it has a rear axle ratio of 3.91.

The 95% confidence interval for the average fuel economy of cars with these specifications is from 18.5881 mpg to 24.469 mpg. The 95% prediction interval for an individual response for fuel economy of this car is (15.0897, 27.9674).

This means I can be 95% confident that the true average fuel economy for cars with 175 horsepower, 14.2 quarter mile time, and 3.91 rear-axle ratio lies within this range. This interval is narrower than the prediction interval because it estimates where the mean of the population likely falls, rather than predicting a new observation.

## 4. Model with Interaction Term and Qualitative Predictor

### Reporting Results

The general form of the regression model is:

The regression model for fuel economy, denoted as mpg (miles per gallon), is constructed using several predictors. These predictors include horsepower (hp), quarter mile time (qsec), an interaction term between horsepower and quarter mile time (hp\*qsec), and the number of cylinders (cyl). The number of cylinders is a qualitative predictor and is represented by two dummy variables, cyl6 and cyl8.

The general form of this regression model can be expressed as follows:

E(mpg)=

In this equation:

E(mpg) denotes the expected value of fuel economy.

hp represents horsepower.

qsec stands for quarter mile time.

cyl6 and cyl8 are dummy variables for the number of cylinders, representing vehicles with 6 and 8 cylinders respectively.

This model allows me to predict the fuel economy based on the given predictors from model 2.

mpg=24.505565+0.141850(hp)+0.531630(qsec)−0.012526(hp \* qsec) −4.408372 (cyl6) −4.580823(cyl8)

The R2 value for our model is 0.8327, and the adjusted R2 value is 0.8005. These values indicate that the model accounts for a significant portion of the variability in fuel economy (mpg). Specifically, the model explains approximately 83.27% of the variability in fuel economy. These statistics suggest that this regression model fits well to explain variations in fuel economy based on given predictors including an interaction term between horsepower and quarter mile time while considering different numbers of cylinders as a qualitative predictor.

The scatterplot of residuals against fitted values offers a visual exploration of the residuals and their correlation with the predicted fuel economy. This visualization provides crucial insights into the precision of the model’s predictions and can guide further enhancements to improve its predictive accuracy. It also highlights the potential intricacies in predicting fuel economy, which may be influenced by a variety of factors including horsepower, quarter mile time, and the interaction term between horsepower and quarter mile time. These factors may not exhibit a straightforward linear relationship with fuel economy, as suggested by the model that incorporates interaction terms. This complexity serves as a reminder of the multifaceted nature of vehicle fuel efficiency and the challenges inherent in accurately predicting it. The number of cylinders, a qualitative predictor in the model, further adds to this complexity. With an R2 value of 0.8327 and an adjusted R2 value of 0.8005, the model accounts for a significant portion of the variability in fuel economy, indicating its effectiveness despite these complexities.

A diagram with red dots

Description automatically generated

A graph with blue dots

Description automatically generated

In the scatterplot, the residuals are plotted against the fitted values. If the assumption of homoscedasticity holds true, I would expect to see a random scatter of points, with no discernible pattern or trend, and with a roughly constant spread across the range of fitted values.

In the above plot, the points appear to be randomly scattered around the horizontal axis and there doesn’t seem to be any clear pattern or trend in the residuals as I move along the fitted values. The spread of residuals also appears to be fairly constant across different fitted values. There’s no clear evidence of a funnel shape (indicating increasing or decreasing variance with fitted values). Therefore, based on this visual inspection, it seems reasonable to conclude that the assumption of homoscedasticity is not violated for this model.

Observing the Normal Q-Q Plot, it’s noticeable that the residuals tend to stray from the reference line at both ends of the plot. This suggests possible deviations from normality, particularly in the tails. These deviations could be indicative of potential outliers or heavy-tailed distributions in the residuals. Despite these deviations, a significant number of points adhere to the trend of the reference line, demonstrating some level of normality. Therefore, while there may be slight deviations, the assumption of normality for the residuals in the model is largely upheld.

### Evaluating Model Significance

In assessing the significance of the multiple regression model and its predictors, specifically horsepower (hp), quarter mile time (qsec), the interaction term for horsepower and quarter mile time (hp:qsec), and number of cylinders (cyl), in predicting fuel economy (mpg), I am using a 5% significance level for these tests.

Overall F-test:

* Null Hypothesis (): All regression coefficients are equal to zero indicating none of the predictors contribute to the model.
* Alternative Hypothesis (​): At least one regression coefficient is not equal to zero indicating at least one predictor contributes to the model.
* : ​= ​ = ​= = 0

​: at least one ≠ 0 for i=1,2,…..,n

* The p-value for the overall F-statistic is given as 2.526e-09, which is much smaller than the 5% level of significance.
* Conclusion: We reject the null hypothesis at the 5% level of significance. This suggests that the model is significant, meaning that at least one of the predictors (horsepower, quarter mile time, interaction term for horsepower and quarter mile time, and number of cylinders) significantly contributes to predicting the fuel economy.

Individual Beta Tests:

* Null Hypothesis (): The coefficient of the predictor is equal to zero indicating the predictor does not contribute to the model.
* Alternative Hypothesis (​): The coefficient of the predictor is not equal to zero indicating the predictor contributes to the model.
* : ​= 0 for some i=1,2,…..,n

​: ≠0

* Looking at the p-values for the individual predictors:

hp: p-value = 0.0848

qsec: p-value = 0.4828

cyl6: p-value = 0.0118

cyl8: p-value = 0.0847

hp:qsec: p-value = 0.0246

* Conclusion: At the 5% level of significance, the predictors cyl6 and hp:qsec are significant as their p-values are less than 0.05. The predictors hp, qsec, and cyl8 are not significant at the 5% level as their p-values are greater than 0.05.

This analysis suggests that the number of cylinders (specifically 6 cylinders) and the interaction between horsepower and quarter mile time have a significant relationship with fuel economy. However, horsepower, quarter mile time, and the number of cylinders (specifically 8 cylinders) do not significantly affect fuel economy at a 5% level of significance.

### Making Predictions Using the Model

Based on the regression model, I can now make some predictions. For the scenario, I will use the hypothetical values of horsepower = 175, quarter mile time = 14.2, and number of cylinders = 6. The equation for this is:

E(mpg)=

Where:

mpg is the miles per gallon (fuel efficiency)

hp is the horsepower

qsec is the quarter mile time

cyl6 and cyl8 are dummy variables for the number of cylinders, representing vehicles with 6 and 8 cylinders respectively.

The prediction equation, based on the outputs obtained from the R script, is:

Mpg = 24.505565 + 0.141850 ∗ hp + 0.531630 ∗ qsec −0.012526 ∗ (hp ∗ qsec) − 4.408372 ∗ cyl6 − 4.580823 ∗ cyl8

The predicted fuel economy (mpg) for a car that has 175 horsepower, 14.2 quarter mile time, and 6 cylinders is 21.3424 mpg.

The 95% prediction interval for the fuel economy of this car is from 14.8764 mpg to 27.8085 mpg. . The 95% prediction interval for an individual response for fuel economy of this car is (14.8764, 27.8085).

The prediction interval for an individual response tells me I 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 horsepower is 175, its quarter mile time is 14.2 and it has 6 cylinders.

The 95% confidence interval for the average fuel economy of cars with these specifications is from 17.9965 mpg to 24.6884 mpg. The 95% prediction interval for an individual response for fuel economy of this car is (17.9965, 24.6884). This means I can be 95% confident that the true average fuel economy for cars with 175 horsepower, 14.2 quarter mile time, and 6 cylinders lies within this range. This interval is narrower than the prediction interval because it estimates where the mean of the population likely falls, rather than predicting a new observation.

## 5. Conclusion

Based on the statistical analyses conducted, and assuming a sufficiently large sample size, I can confidently recommend a model. Both models analyzed include variables that do not seem to have a statistically significant relationship with fuel economy. Specifically, in the second model, the variables for horsepower, quarter mile time, and the number of cylinders (specifically 8 cylinders) do not have a statistically significant relationship with fuel economy.

Given these findings, I would recommend a model that includes just horsepower, quarter mile time, and an interaction term of horsepower against quarter mile time. This recommendation is based on the conclusion that these are the only variables that have a significant relationship with fuel economy, and these variables are shared in both models. It would be interesting to further refine this model by adding other variables from the data set, along with additional interaction variables, to see if any other variables could also have a statistically significant relationship with fuel economy.

The practical importance of these analyses is that they can assist car manufacturers in determining which attributes of a vehicle significantly affect fuel economy and whether or not the interaction of two attributes has a significant effect. This can be particularly useful if a car manufacturer is interested in understanding how increasing horsepower will affect fuel economy. They can examine how horsepower and quarter mile time interact with each other to ensure that any improvement in fuel economy resulting from an increase in horsepower does not negatively impact the quarter mile time of the vehicle. This analysis provides valuable insights that can guide decision-making in car manufacturing and design.

## 6. Citations

zyBooks. (n.d.). Learn.zybooks.com. Retrieved July 14, 2024, from <https://learn.zybooks.com/zybook/MAT-303-16699.202456-1/chapter/2/section/1>

‌