# MAT 303 Module Two Problem Set Report

Multiple Regression with Interaction and Qualitative Terms

Brandon Hobbs

Brandon.Hobbs@snhu.edu

Southern New Hampshire University

## Introduction

The data set being analyzed consists of 32 rows and 12 columns. Each row contains data about a particular car, e.g., *mpg*, *cyl*, *disp*, etc. See Figure 1 for the first few rows for data.

The data will be used to build a multiple-regression model with the purpose of predicting miles per gallon (mpg) from the other available data.  
  
First, the data in the csv-file will be ingested into a dataframe so the R-language may be used for the stated purpose. Next, it will be plotted to provide a sense of the data and then the regression models, and their appropriateness, will be calculated. Finally, the model will be used to make predictions.

Table

Description automatically generated

**Figure 1: First 5 Rows of Data used for Analysis**

## Data Preparation

To begin the analysis the data, all 32 rows and 12 columns, were imported into a data frame for consumption in the R-language. Of particular interest are the *mpg*, *drat, qsec,* *cyl*, and *hp* columns. Two regression models will be created.

One model will try and predict the miles-per-gallon (mpg) as a function of rear-axel ratio (drat), quarter-mile times (qsec), horsepower (hp). Two interaction terms are included: one between *hp* and *drat*, and one between *hp* and *qsec*.

The second model will try and predict the miles-per-gallon (mpg) as a function of quarter-mile times (qsec), number of engine cylinders (cyl), horsepower (hp), and an interaction term between *hp* and *qsec.*

## Multiple Regression Model with Interaction Terms

This model will try and predict the miles-per-gallon (mpg) as a function of rear-axel ratio (drat), quarter-mile times (qsec), horsepower (hp). Two interaction terms are included: one between *hp* and *drat*, and one between *hp* and *qsec*.

### Correlation Analysis To determine how strongly *mpg* is dependent on *drat, qsec,* or *hp* the Pearson correlation coefficient (*R*) was computed. A positive correlation between two variables means that as one variable increases, the other variable increases as well. A negative correlation between two variables means that as one variable increases, the other variable decreases. Figure 2 provides guidance on how to interpret the magnitude of this value.

*Table

Description automatically generated*

**Figure 2: Pearson Correlation Coefficient Magnitude and its Interpretation**

Table 1 contains the Pearson correlation matrix which may be interpreted using Figure 2.

**Table 1: Pearson Correlation Matrix**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | ***mpg*** | ***hp*** | ***qsec*** | ***drat*** |
| ***mpg*** | 1 | -0.7762 | 0.4187 | 0.6812 |
| ***hp*** | -0.7762 | 1 | -0.7082 | -0.4488 |
| ***qsec*** | 0.4187 | -0.7082 | 1 | 0.0912 |
| ***drat*** | 0.6812 | -0.4488 | 0.0912 | 1 |

*hp* is negatively correlated, i.e., as *hp* increases mpg decreases, *drat and qsec* are positively correlated, that is as the axel ratio or quarter-mile times increase so does *mpg*. Moreover, *hp* is more strongly correlated than *drat and qsec* but all could be, according to Figure 2, classified as moderate.

### Reporting Results

The previous section showed that *drat, qsec,* and *hp* are moderately correlated to fuel efficiency. Due to this fact a linear regression model was created with *drat, qsec,* and *hp* as the predictor variables and *mpg* as the response variable. This model will be of the form:

With the final model being:

With *hp* as X1, *qsec* as X2, *drat* as X3. This model as mentioned earlier also includes 2 interaction terms, i.e., *hp/qsec* and *hp/drat*. This model says, for example, that *mpg* will increase 0.353 units if the horsepower value increases one unit and all other terms are held constant.

This model has a coefficient of determination (*R2*) of 0.821 – meaning that 82.1% of the variability in mpg is explained by the predictor variables. The model also has an adjusted *R2* of 0.786. The adjusted *R2* tends to only increase when a worthwhile predictor variable is added. This value should not be used in isolation but could be used if a new predictor variable, e.g., weight (*wt*), was added to the model. The adjusted *R2* could evaluate it was a valuable addition.

To further determine if the model was relevant an F-test is used. An F-test is run to determine if there is indeed an association between the predictor variables and the response variable. First, the null hypothesis (*H0*) and alternative hypothesis (*Ha*) are created:

*H0: β1 = β2 =…= βn = 0*

*Ha: At least one βi ≠ 0 for i = 1 to n*

The null hypothesis states that *β1* through *β5* are zero; meaning there is no correlation between *drat*, *qsec*, *hp,* the interaction terms, and *mpg*. The alternative states at least one beta term, *β1* through *β5,*are not zero; meaning there is a correlation between at least one of *drat*, *hp, qsec,* the interaction terms*,* and *mpg*. This will be evaluated against an α of 5% or a 95% confidence interval. Table 2 shows the F-Test statistic and its associated P-value:

**Table 2: Hypothesis Test for the Overall F-Test**

| **Statistic** | **Value** |
| --- | --- |
| Test Statistic | 41.52 |
| P-value | 3.081E-9 |

The P-value confirms that the null value may be rejected, 3.081E-9 << 0.05; thus, at least one variable is linearly correlated to *mpg*. Moreover, this further confirms that the model shown above is valid at the 95% confidence level.

What the F-test does not reveal is how many of the predictor variables are relevant or which ones. To determine which variables are relevant an individual t-test is conducted on each variable. Each t-test will have a similar null hypothesis and alternative hypothesis. The null hypothesis and alternative hypothesis will be of this form:

*H0: βi =0*

*Ha: βi ≠ 0 for i = 1…n*

As before, the null hypothesis states that *βi* is zero; meaning there is no correlation between its predictor variable and *mpg*. The alternative states that *βi* is not zero; meaning there is a correlation between its predictor variable and *mpg*. Based on these hypotheses the P-values can be used to determine statistical relevance, see Table 3.

**Table 3: T-test for Individual Predictor Variables**

| **Variable** | **P-Value** |
| --- | --- |
| *hp* | 0.01175 |
| *qsec* | 0.04043 |
| *drat* | 0.03262 |
| *hp:qsec* | 0.00307 |
| *hp:drat* | 0.08405 |

All P-values are less than the 5% significance level, i.e., P-value << 0.05, except the *hp:drat* interaction term. Therefore, all variables, except the single interaction term, are shown to have a statistically relevant. Removing the *hp:drat* interaction term and recomputing the model should be evaluated but will be left for another day.

The beta values used in the model represent the statistical mean of the possible range of beta values. Using a 95% confidence range (α = 0.05), as before, means someone can be 95% confident that the mean value is between the lower and upper limits. Table 4 gives the lower and upper limits of these values.

**Table 4: 95% Confidence Interval for the Model Parameters**

|  |  |  |
| --- | --- | --- |
|  | **Lower Limit** | **Upper Limit** |
| ***(Intercept)*** | -47.6982 | 18.6399 |
| ***hp*** | 0.0852 | 0.6204 |
| ***qsec*** | 0.0710 | 2.9481 |
| ***drat*** | 0.5063 | 10.8270 |
| ***hp:qsec*** | -0.0305 | -0.0069 |
| ***hp:drat*** | -0.0713 | 0.0048 |

Table 4 further confirms the relevance of each term. Any range that contains 0 should be considered as not statistically significant; e.g., *hp:drat*, and any not containing 0 should be considered statistically significant.

The final test to determine whether linear regression is appropriate is to plot the residuals versus the fitted values, Figure 3. This plot allows someone to diagnostically examine if the necessary linear regression assumptions are valid for the sample data. These assumptions are:

* Mean of zero
* Independence
* Normality
* Constant variance

Chart, scatter chart

Description automatically generated

**Figure 3: Fitted Value versus Residuals**

The fitted values are the model’s predictions, and the residuals are the difference between the actual data and the model’s prediction. Figure 3 confirms the assumptions of mean of zero, data is centered around a y-value of zero, and constant variance, there are no “fan-shaped” patterns. Constant variance is also known as homoscedasticity – Figure 3 confirms the data is homoscedastic.

An independence test is not needed as there are no time varying variables in the data set.

Normality is confirmed with a QQ plot, Figure 4.

Chart, line chart, scatter chart

Description automatically generated

**Figure 4: QQ Plot of Residuals**

A QQ plot confirms normality if the residuals lie on the blue line. If the points deviate “significantly” from the diagonal line, then the assumption of normality is violated. The residuals do not appear to deviate significantly and thus normality is confirmed.

### Making Predictions Using the Model

With the new model created and confirmed as relevant it is useable for predictions. As an example, a car with 175 hp, a rear-axel ratio of 3.91, and a 14.2 second quarter-mile time is predicted to have a 21.53 mpg.

However, as before, the value of 21.53 is the mean and someone can be 95% confidant that the actual average mpg exists in the range of 18.59 to 24.47 mpg for all cars with those *hp, qsec,* and *drat* figures.

Due to the uncertainty in estimating the mean value and the random variation in what was already observed someone could be certain, to 95%, that a single vehicle with the same specs will have a mpg between 15.09 and 27.97 mpg. This wider range is known as the prediction interval. The prediction interval is wider because it considers the variability of the individual points around the predicted mean in addition to the uncertainty in sampling.

## Multiple Regression Model with Interaction Terms and Qualitative Predictors

## This second model will try and predict the miles-per-gallon (mpg) as a function of quarter-mile times (qsec), number of engine cylinders (cyl)(a qualitative predictor), horsepower (hp), and an interaction term between *hp* and *qsec*.

### Reporting Results

The first step in building the model is to relevel the qualitative predictor, *cyl*. In the data there are three possible values of *cyl* so only two variables are needed to represent the data – a 1 indicates that a quality exists. Table 5 shows these values:

**Table 5: Qualitative Predictors with 4-Cylinder as the Reference**

|  |  |  |
| --- | --- | --- |
|  | ***cyl6*** | ***cyl8*** |
| **4-cylinder** | 0 | 0 |
| **6-cylinder** | 1 | 0 |
| **8-cylinder** | 0 | 1 |

A linear regression model was created with *qsec,* *cyl*, and *hp* as the predictor variables and *mpg* as the response variable. This model will be of the form:

With the final model being:

With *hp* as X1, *qsec* as X2, qualitative variable [0,1] *cyl6* as X3, qualitative variable [0,1] *cyl8* as X4. This model also includes the interaction term *hp/qsec*. This model says, for example, that mpg will drop 4.41 units if the engine changes from a 4- to a 6-cylinder and all other terms are held constant.

This model has a coefficient of determination (*R2*) of 0.833 – meaning that 83.3% of the variability in mpg is explained by the predictor variables. The model also has an adjusted *R2* of 0.801. The adjusted *R2* tends to only increase when a worthwhile predictor variable is added. Since the adjusted *R2* of this model is greater than the previous model it suggests the new variables were valuable additions.

To further determine if the model was relevant an F-test is used. An F-test is run to determine if there is indeed an association between the predictor variables and the response variable. First, the null hypothesis (*H0*) and alternative hypothesis (*Ha*) are created:

*H0: β1 = β2 =…= βn = 0*

*Ha: At least one βi ≠ 0 for i = 1 to n*

From before, the null hypothesis states that *β1* through *β5* are zero; meaning there is no correlation between the predictor variables and *mpg*. The alternative states at least one beta term, *β1* through *β5,*are not zero; meaning there is a correlation between at least one predictor variable and *mpg*. This will be evaluated against an α of 5% or a 95% confidence interval. Table 6 shows the F-Test statistic and its associated P-value:

**Table 6: Hypothesis Test for the Overall F-Test**

| **Statistic** | **Value** |
| --- | --- |
| Test Statistic | 25.88 |
| P-value | 2.526E-9 |

The P-value confirms that the null value may be rejected, 2.526E-9 << 0.05; thus, at least one variable is linearly correlated to *mpg*. Moreover, this further confirms that the model shown above is valid at the 95% confidence level.

What the F-test does not reveal is how many of the predictor variables are relevant or which ones. To determine which variables are relevant an individual t-test is conducted on each variable. Each t-test will have a similar null hypothesis and alternative hypothesis. The null hypothesis and alternative hypothesis will be of this form:

*H0: βi =0*

*Ha: βi ≠ 0 for i = 1…n*

As before, the null hypothesis states that *βi* is zero; meaning there is no correlation between its predictor variable and mpg. The alternative states that *βi* is not zero; meaning there is a correlation between its predictor variable and mpg. Based on these hypotheses the P-values can be used to determine statistical relevance, see Table 7.

**Table 7: T-test for Individual Predictor Variables**

| **Variable** | **P-Value** |
| --- | --- |
| *hp* | 0.0848 |
| *qsec* | 0.4828 |
| *cyl6* | 0.0118 |
| *cyl8* | 0.0847 |
| *hp:qsec* | 0.0246 |

Many of the P-values are greater than the 5% significance level, i.e., P-value >> 0.05. Therefore, even though the adjusted *R2* value is closer to 1.0 than the last model, this model should be evaluated carefully. A third model, similar to this one, but without the *hp*, *qsec*, *cyl8* (all P-values > 0.05) should be created and tested.

Moreover, this model may be suffering from overfitting – not evaluated at this time. That is to say, if a vehicle has a lot of horsepower (large *hp* value) and an 8-cylinder engine it probably has a low qsec time (all linearly related). Thus, the interaction term and a few of the other predictor variables are already describing the same behavior.

The beta values used in the model represent the statistical mean of the possible range of beta values. Using a 95% confidence range (α = 0.05), as before, means someone can be 95% confident that the mean value is between the lower and upper limits. Table 8 gives the lower and upper limits of these values.

**Table 8: 95% Confidence Interval for the Model Parameters**

|  |  |  |
| --- | --- | --- |
|  | **Lower Limit** | **Upper Limit** |
| ***(Intercept)*** | -2.5988 | 51.6099 |
| *hp* | -0.0209 | 0.3046 |
| *qsec* | -1.0033 | 2.0665 |
| *cyl6* | -7.7541 | -1.0626 |
| *cyl8* | -9.8342 | 0.6726 |
| *hp:qsec* | -0.0233 | -0.0017 |

Table 8 further confirms the relevance of each term. Any range that contains 0 should be considered as not statistically significant, e.g., *cyl8*, and any not containing 0 should be considered statistically significant.

The final test to determine whether linear regression is appropriate is to plot the residuals versus the fitted values, Figure 5. This plot allows someone to diagnostically examine if the necessary linear regression assumptions are valid for the sample data. These assumptions are:

* Mean of zero
* Independence
* Normality
* Constant variance

Chart, scatter chart

Description automatically generated

**Figure 5: Fitted Value versus Residuals**

The fitted values are the model’s predictions, and the residuals are the difference between the actual data and the model’s prediction. Figure 5 confirms the assumptions of mean of zero, data is centered around a y-value of zero, and constant variance, there are no “fan-shaped” patterns. Constant variance is also known as homoscedasticity – Figure 5 confirms the data is homoscedastic.

An independence test is not needed as there are no time varying variables in the data set.

Normality is confirmed with a QQ plot, Figure 6.

Chart, scatter chart

Description automatically generated

**Figure 6: QQ Plot of Residuals**

A QQ plot confirms normality if the residuals lie on the blue line. If the points deviate “significantly” from the diagonal line, then the assumption of normality is violated. The residuals do not appear to deviate and thus normality is confirmed.

### Making Predictions Using the Model

With the new model created and confirmed as relevant it is useable for predictions. As an example, a car with 175 hp, a 6-cylinder engine, and a 14.2 second quarter-mile time is predicted to have 21.34 mpg.

However, as before, the value of 21.34 is the mean and someone can be 95% confidant that the actual average mpg exists in the range of 17.99 to 24.69 mpg for all cars with those *hp, qsec,* and engine size figures.

Due to the uncertainty in estimating the mean value and the random variation in what was already observed someone could be certain, to 95%, that a single vehicle with the same specs will have a mpg between 14.88 and 27.81 mpg. Again, this wider range is known as the prediction interval, see above for explanation.

## Conclusion

This report details the creation of two multi-parameter linear regressions. Both models attempt to predict fuel efficiency from different variables. To determine if the model was statistically relevant an F-test and individual t-tests were conducted on both models.

Both the F- and t-tests showed that the beta values predicted were relevant within a 95% confidence range. Because both tests confirmed the statistical relevance, the model is recommended for usage.

Moreover, the data was examined for its appropriateness in this type of linear regression. The residuals versus fitted values plot and the QQ-plot confirmed that the underlying assumptions were valid and thus the data was appropriate for this type of analysis.

Since the model and the data’s quality were confirmed a model of this type could be valuable to a rental company so they might pre-compute the expected MPG for any new asset to see how it impacts fleet costs. The rental company could also compute the MPG expected for existing assets and compare to this to the actual MPG as an indicator of "health".

## Citations

Hobbs, B. (2022). *MAT 303 module one summary report*. [Unpublished report]. SNHU.