# MAT 303 Module One Problem Set Report

Multiple Regression

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 dataframe for consumption in the R-language. Of particular interest are the *mpg*, *drat,* and *hp* columns. A regression model will be created with *drat* and *hp* as the predictor variables and *mpg* as the response variable.

This model will be used to try and predict the miles-per-gallon (mpg) as a function of rear-axel ratio (drat) and horsepower (hp).

## Multiple Regression Model

### Correlation Analysis

To begin, the mpg was plotted versus rear-axel ratio to see if any trends could be spotted visually, Figure 2, left. This was also carried out for horsepower, Figure 2, right.  
  
Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

**Figure 2: Left, Fuel Economy versus Rear-Axel Ratio; Right, Fuel Economy versus Horsepower**

Both predictor variables appear, visually, to be linearly correlated with mpg; however, *drat* appear to be positively correlated while *hp* negatively correlated.To determine how strongly mpg is dependent on *drat* 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 3 provides guidance on how to interpret the magnitude of this value.

*Table

Description automatically generated*

**Figure 3: Pearson Correlation Coefficient Magnitude and its Interpretation**

Table 1 contains the Pearson correlation matrix which may be interpreted using Figure 3.

**Table 1: Pearson Correlation Matrix**

|  |  |  |  |
| --- | --- | --- | --- |
|  | ***mpg*** | ***hp*** | ***drat*** |
| ***mpg*** | 1.0000 | -0.7762 | 0.6812 |
| ***hp*** | -0.7762 | 1.0000 | -0.4488 |
| ***drat*** | 0.6812 | -0.4488 | 1.0000 |

Table 4 confirms the suspicions from visual inspection. That is, *hp* is negatively correlated, i.e., as *hp* increases mpg decreases, and *drat* is positively correlated, that is as the axel ratio increases so does *mpg*. Moreover, *hp* is more strongly correlated than *drat* but both could be, according to Figure 3, classified as moderate.

### Reporting Results

The previous section showed that *drat* and *hp* are moderately correlated to fuel efficiency. Due to this fact a linear regression model was created with *drat* 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 and *drat* as X2.

This model has a coefficient of determination (*R2*) of 0.741 – meaning that 74.1% of the variability in mpg is explained by *hp* and *drat*. The model also has an adjusted *R2* of 0.723. 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 3rd predictor variable, e.g., weight (*wt*), was added to the model to evaluate if the new variables 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 = 0*

*Ha: At least one βi ≠ 0 for i = 1 to 2*

The null hypothesis states that *β1* and *β2* are zero; meaning there is no correlation between *drat*, *hp* and *mpg*. The alternative states that either, or both, *β1* or *β2* are not zero; meaning there is a correlation between *drat*, *hp* 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.

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

| **Variable** | **P-Value** |
| --- | --- |
| *hp* | 0.000 |
| *drat* | 0.000 |

Both P-values are less than the 5% significance level, i.e., 0.000 << 0.05, therefore, both variables are shown to have a statistically relevant.

Each beta variable is shown to be statistically relevant via the t-tests, but the -0.052 and 4.698 beta values in the model are not precisely known. The values instead represent the statistical mean of the possible range of 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)** | 0.4047 | 21.1750 |
| **hp** | -0.0708 | -0.0328 |
| **drat** | 2.2610 | 7.1353 |

Table 4 further confirms that each predictor variable is relevant as zero does not appear within any of the ranges.

The final test to determine whether linear regression is appropriate is to plot the residuals versus the fitted values, Figure 4. 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 4: 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 4 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 4 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 5.

Chart, scatter chart

Description automatically generated

**Figure 5: 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 120 hp and a rear-axel ratio of 3.15 is predicted to have a 19.37 mpg.

The residual for this fitted value is 1.13 (20.5 – 19.37) if the actual mpg is 20.5.

However, as before, the value of 19.37 is the mean and someone can be 95% confidant that the actual average mpg exists in the range of 17.57 to 21.18 mpg for all cars with those *hp* 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 12.64 and 26.10 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.

## Conclusion

This report details the creation of a multi-parameter linear regression which attempted to predict fuel efficiency from rear-axel ratio and horsepower. To determine if the model was statistically relevant an F-test and individual t-tests were conducted.

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 243 project one summary report*. [Unpublished report]. SNHU.

Hobbs, B. (2022). *MAT 243 project two summary report*. [Unpublished report]. SNHU.

Hobbs, B. (2022). *MAT 243 project three summary report*. [Unpublished report]. SNHU.