

**Department of Electrical and Electronics Engineering**

**GE 461: Introduction to Data Science**

*Class Project I Report*

**Name and Surname:** Ali Aral Takak

**Student ID:** 22001758

**Department:** EEE

**Introduction**

The first project of the course focuses on linear regression, a simple yet effective method of supervised learning. Students are required to solve the problems **3.7.8** and **3.7.9** from the book *“An Introduction to Statistical Learning with Applications in Python and R”*, written by Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani and Jonathan Taylor.

Before moving further in the implementations and results of the problems, the report will briefly explain what linear regression is, with provided formulations and visualizations. Then, we will report our findings on the specified questions, analyze our results, and conclude the results.

**Overview of Linear Regression**

Linear regression is a statistical method that aims to model the relationship between two variables by fitting a linear equation to a set of observed data. In simple linear regression model, we have a quantitative response, denoted as , on the basis of a single predictor variable, [1]. Mathematically, one can define this relationship as:

Where is named as the intercept term, which eliminates origin bias from the linear model, and is named as the slope. Together, these terms are labeled as the **model coefficients** or **parameters** [1]. One can generate an approximation after using the training data in order to produce estimates, and define a new equation as:

In the equation provided above, represents our prediction of on the basis of , with the estimate coefficients and . Now, a new question arises from these definitions: We know that in practice, and are unknown. Hence, how can we learn these parameters in order to make reliable estimations?

Let us define some data. Let:

Represent observation pairs, in which each consists of a measurement of and a measurement of . We desire to obtain coefficient estimates and , so that the linear model that we will define will fit the data well. The most common method that is used is minimizing the least squares criterion [1]. Let

Denote the prediction for based on the value of . In this case, one can calculate the difference between the observed value and the true value as:

In this case, the total error for all observations, or the Residual Sum of Squares (RSS) can be calculated as:

One can rewrite the Residual Sum of Squares as:

Finally, one can determine the model coefficients as:

Where and are sample means, defined as follows:

Although simple linear regression is an effective method, in realistic models, there exists more than one predictor variable. In this scenario, we need to extend our model to be constructed from multiple predictors. Hence, we take a closer look to multiple linear regression model. Suppose that we have distinct predictor variables. Then, we can write the multiple linear regression model as:

In a similar manner, the regression coefficients are unknown and must be estimated. Given estimates , we can make predictions by:

The Residual Sum of Squares can be written as:

**Problem I**

The first problem involves the use of simple linear regression on the Auto data set, and expects students to answer the questions given below:

1. **Use the sm.OLS() function to perform a simple linear regression with mpg as the response and horsepower as the predictor. Use the summarize() function to print the results. Comment on the output. For example:**
2. **Is there a relationship between the predictor and the response?**

After fitting our, model, we have obtained the regression line which can be observed in the figure provided below:

A graph with a red line

AI-generated content may be incorrect.

Figure 1: Generated simple regression line.

Then, we have determined our model parameters and various statistical measurements as provided in the figure below:

A screenshot of a computer

AI-generated content may be incorrect.

Figure 2: Model parameters along with statistical measurements.

Hence, we can determine our simple linear regression model as:

1. **How strong is the relationship between the predictor and the response?**

In order to determine the strength of the relationship between the predictor and the response, we can control two variables, namely the R-squared value and the p-value. The p-value can be observed in *Figure 2,* and it is determined as for *horsepower.* This indicates that statistically, there is a strong relationship between mpg and horsepower parameters.

Additionally, we can check the R-squared value too. The figure given below displays the R-squared value:

A black background with white text

AI-generated content may be incorrect.

Figure 3: R-squared value between the parameters.

This also indicates that there is a strong relationship between our parameters, mpg and horsepower.

1. **Is the relationship between the predictor and the response positive or negative?**

In order to answer this question, we have two different observations to make. First of all, the horsepower coefficient, in other words, the slope of the model, is given as a negative coefficient. Hence, an increase in horsepower would gradually decrease the value of the miles per gallon. Second of all, investigating the generated plot confirms our intuition of negative impact on miles per gallon. Hence, the relationship between the predictor and the response is negative.

1. **What is the predicted mpg associated with a horsepower of 98? What are the associated 95% confidence and prediction intervals?**

We can determine the predicted miles per gallon value for a horsepower of 98 manually. Given our model:

When we plug in the value , we can obtain that . We can also inspect our plot to confirm our results. The figure given below displays the values at :

A screen shot of a graph

AI-generated content may be incorrect.

Figure 4: Regression line results at (x,y)=(98.01,24.49) [check right below corner].

The figure given below displays the calculations made via Python in order to determine the prediction at , and the corresponding 95% correspondence and prediction intervals:

A black screen with white numbers

AI-generated content may be incorrect.

Figure 5: Predictions and statistics for 98 horsepower.

**Problem II**

The second question involves the usage of multiple linear regression on the Auto data set, and expects students to answer the questions provided below:

1. **Produce a scatterplot matrix which includes all of the variables in the data set.**

The figures given below display the scatterplots pairwise for all variables in the data set:  
A diagram of a graph

AI-generated content may be incorrect.

Figure 6: Acceleration versus year scatterplot.

A graph of blue dots

AI-generated content may be incorrect.

Figure 7: Cylinders versus acceleration scatterplot.

A graph of a graph

AI-generated content may be incorrect.

Figure 8: Cylinders versus displacement scatterplot.

A graph of blue dots

AI-generated content may be incorrect.

Figure 9: Cylinders versus horsepower scatterplot.

A graph of a graph

AI-generated content may be incorrect.

Figure 10: Cylinders versus weight scatterplot.

A graph of blue dots

AI-generated content may be incorrect.

Figure 11: Cylinders versus year scatterplot.

A diagram of a scatter plot

AI-generated content may be incorrect.

Figure 12: Displacement versus acceleration scatterplot.

A diagram of a scatter plot

AI-generated content may be incorrect.

Figure 13: Displacement versus horsepower scatterplot.

A graph of a scatter plot

AI-generated content may be incorrect.

Figure 14: Displacement versus weight scatterplot.

A diagram of a scatter plot

AI-generated content may be incorrect.

Figure 15: Displacement versus year scatterplot.

A diagram of a graph

AI-generated content may be incorrect.

Figure 16: Horsepower versus acceleration scatterplot.

A diagram of a graph

AI-generated content may be incorrect.

Figure 17: Horsepower versus weight scatterplot.

A diagram of a scatter plot

AI-generated content may be incorrect.

Figure 18: Horsepower versus year scatterplot.

A diagram of a scatter plot

AI-generated content may be incorrect.

Figure 19: MPG versus acceleration scatterplot.

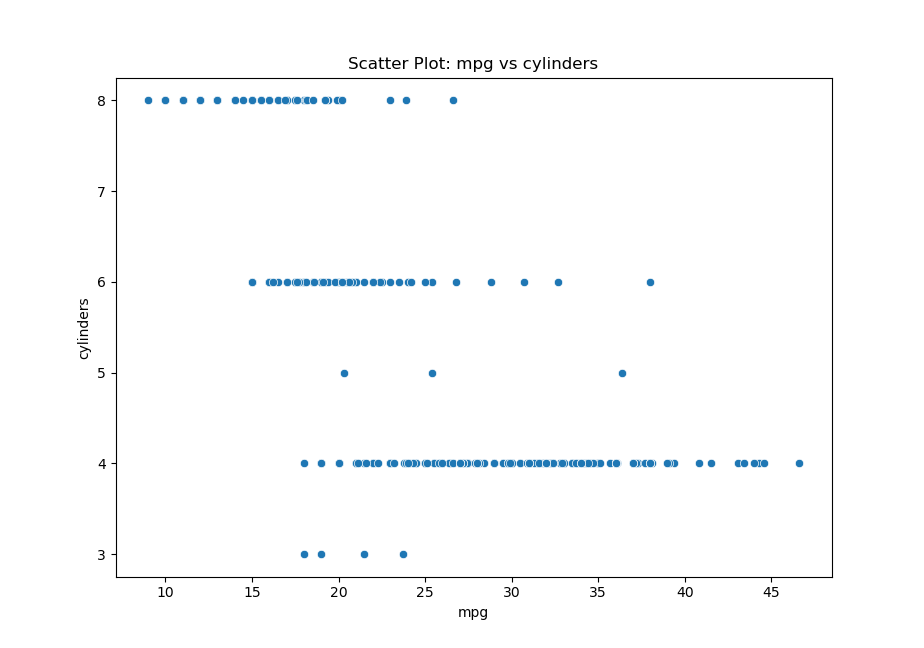


Figure 20: MPG versus cylinders scatterplot.

A diagram of a scatter plot

AI-generated content may be incorrect.

Figure 21: MPG versus displacement scatterplot.

A diagram of a graph

AI-generated content may be incorrect.

Figure 22: MPG versus horsepower scatterplot.

A diagram of a scatter plot

AI-generated content may be incorrect.

Figure 23: MPG versus weight scatterplot.

A diagram of a graph

AI-generated content may be incorrect.

Figure 24: MPG versus year scatterplot.

A diagram of a graph

AI-generated content may be incorrect.

Figure 25: Weight versus acceleration scatterplot.

A diagram of a graph

AI-generated content may be incorrect.

Figure 26: Weight versus year scatterplot.

1. **Compute the matrix of correlations between the variables using the DataFrame.corr() method.**

The figure provided below displays the heatmap of correlations between the variables:

A screenshot of a data set

AI-generated content may be incorrect.

Figure 27: Correlation heatmap of Auto dataset parameters.

1. **Use the sm.OLS() function to perform a multiple linear regression with mpg as the response and all other variables except name as the predictors. Use the summarize function to print the results. Comment on the output. For example:**
   1. **Is there a relationship between the predictors and the response? Use the anova\_lm() function from statsmodels to answer this question.**

The figure given below displays the results of the OLS on the Auto dataset with multiple variables:

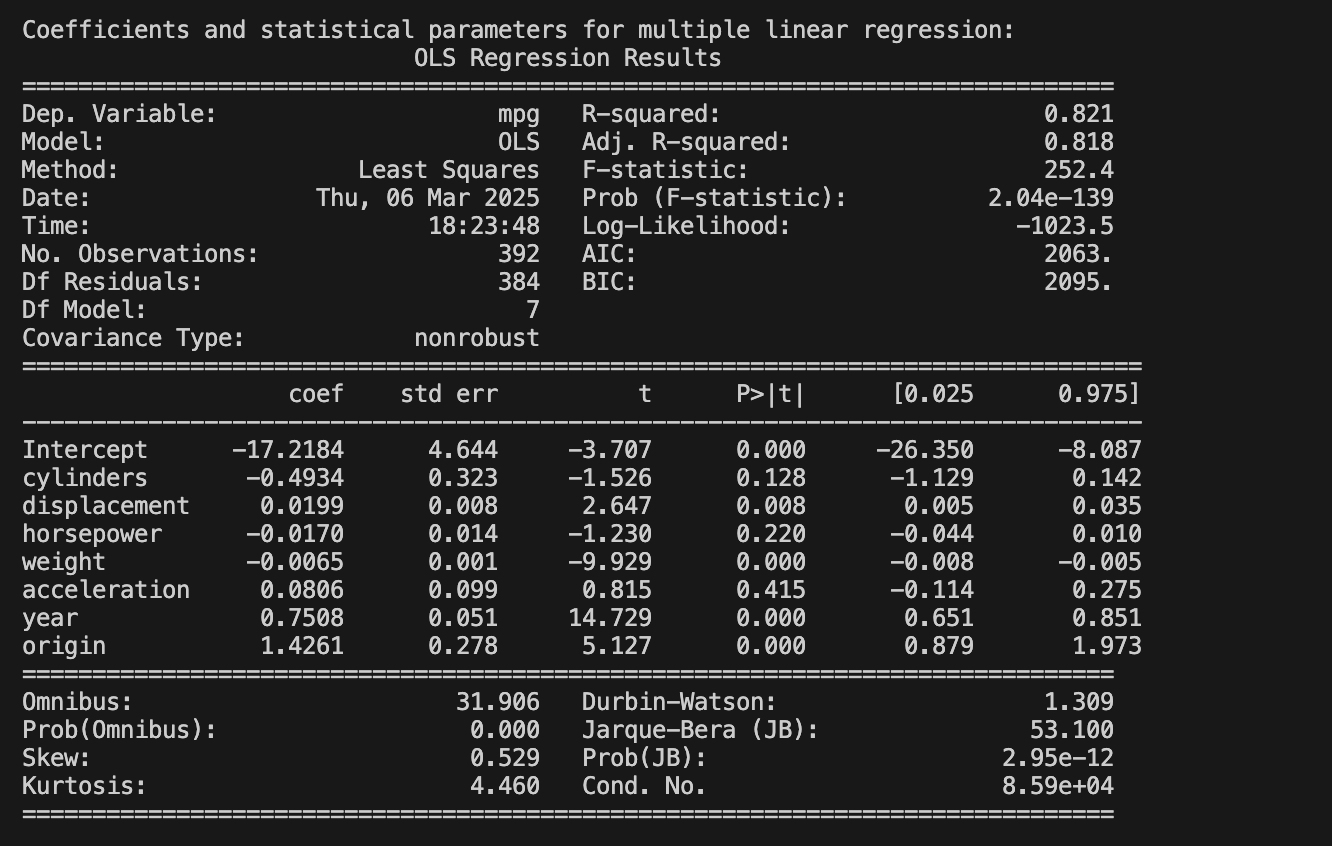


Figure 28: Results for the multiple linear regression.

The figure given below displays the results for Analysis of Variance (ANOVA):

A black screen with numbers and letters

AI-generated content may be incorrect.

Figure 29: ANOVA results.

* 1. **Which predictors appear to have a statistically significant relationship to the response?**

There are a few takeaways from the results provided above, which can be listed as:

* An R-squared result of shows that the variability in mpg is explained by the predictors.
* There exist few significant predictors with , which are:
  + Displacement: Positive coefficient of displacement suggests that higher displacement leads to higher mpg.
  + Weight: Negative coefficient suggests that heavier cars have lower mpg.
  + Year: Positive coefficient suggests that newer cars tend to have higher mpg.
  1. **What does the coefficient for the year variable suggest?**

The coefficient for the year variable suggests that newer cars tend to have higher mpg, whereas older cars tend to have lower mpg.

1. **Produce some of diagnostic plots of the linear regression fit as described in the lab. Comment on any problems you see with the fit. Do the residual plots suggest any unusually large outliers? Does the leverage plot identify any observations with unusually high leverage?**

The figure given below displays the residual plot for the multiple linear regression model:

A diagram of a plot

AI-generated content may be incorrect.

Figure 30: Residual plot for the multiple regression model.

Inspecting the plot provided above yields some points deviate significantly from the horizontal zero line, particularly at higher fitted values.

1. **Fit some models with interactions as described in the lab. Do any interactions appear to be statistically significant?**

The figure given below displays the results of the multiple linear regression with the interaction term, which is defined as :

A screenshot of a computer

AI-generated content may be incorrect.

Figure 31: Regression results with interaction term defined as horsepower\*acceleration.

It is evident that our interaction term is statistically significant, which can be proved by the statistical measurements given in the figure.

1. **Try a few different transformations of the variables, such as .**

The figure given below displays the results of the multiple linear regression with the provided transformations applied:

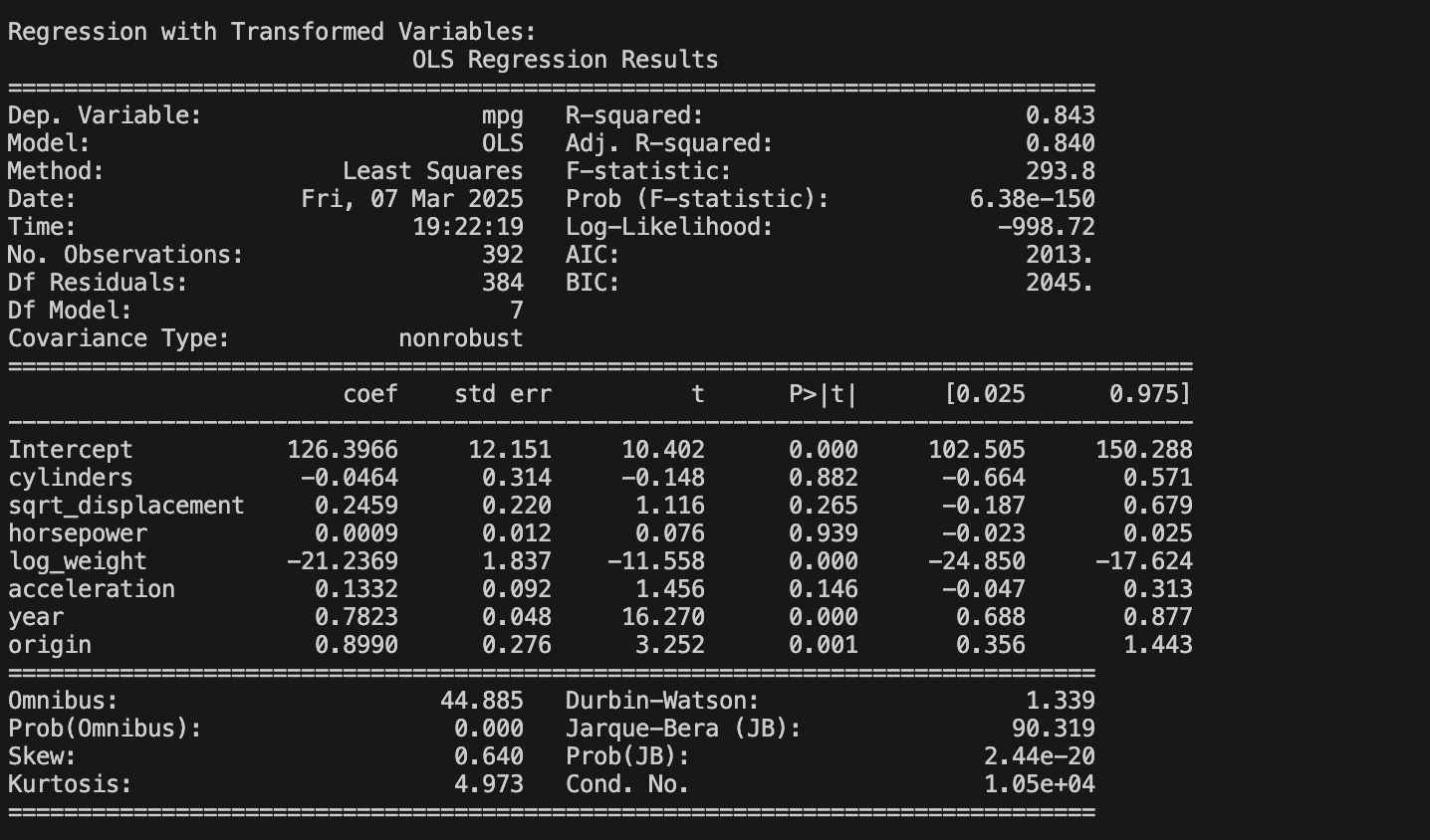


Figure 32: Regression results with transformed variables.

**Conclusion**

In this project, we explored linear regression using the Auto dataset and analyzed its results. We first applied simple linear regression to examine the relationship between miles per gallon (mpg) and horsepower, showing a strong negative correlation. Then, we extended our analysis to multiple linear regression, adding additional predictors to improve our model. The results showed that weight, displacement, and year had significant effects on mpg. We also investigated diagnostic plots, interactions, and transformations to refine our understanding of the model's performance.

**References**

**[1]** G. James, D. Witten, T. Hastie, R. Tibshirani, and J. Taylor, *An Introduction to Statistical Learning: With Applications in Python*. Cham, Cham: Springer International Publishing Springer, 2023.