**Module 2 Project**

**Generalized Linear Regression**

**ALY 6020 Predictive Analytics**

**Instructor:** [**Justin**](https://northeastern.instructure.com/courses/33166/users/56608) **Grosz**

**Northeastern University**



[**Dhanashree Rajendra Chavan**](https://northeastern.instructure.com/courses/33166/users/82134)

**College of Professional Studies, Northeastern University, Boston, MA 02215**

Table of Contents

[Introduction 3](#_Toc78132514)

[Analysis 3](#_Toc78132515)

[Conclusion 7](#_Toc78132516)

[References 7](#_Toc78132517)

[Appendix 8](#_Toc78132518)

# **Introduction**

The dataset we are working on for this assignment is related to car attributes and it contains information about different used cars. Our business goal is to help a well-known car manufacturer to figure out which attributes of a car may lead to higher gas mileage so that the manufacturer can design a more fuel-efficient automobile and improve his car sales. For this purpose, we will be building machine learning model using different techniques, analyzing different car attributes and how they are related to the miles per gallon attribute, compare the models, and decide which model gives the most accurate result.

# **Analysis**

Before building the model, we need to understand and analyse the data thoroughly and clean the data if there are any errors as our model will work efficiently when we are having clean data.

So as a part of data cleaning, I first imported the dataset using appropriate libraries in python and checked how many records we have in the dataset. Our dataset consists of 398 records and 8 columns, we will be exploring each column deeply in this analysis. Further I checked if there are any duplicate records in the dataset, as duplicate records may affect our model results, but I found that there are no duplicate records in the dataset.

In real world scenarios, data may be missing due to errors or corrupt data and if the missing data is not handled correctly the results of our analysis may not be correct. Hence I further checked if our dataset has any missing values. I found that there are no missing values in any column in the dataset, however the ‘Horse Power’ attribute had erroneous values such as ‘?’ and there were 6 records with such data entry issues. So there are two issues with this attribute, one is erroneous value and another on is this attribute should be a numeric field, however in the dataset this is currently a non-numeric field. So we need to resolve this data issue. For this purpose I replaced ‘?’ with null values, converted the attribute to a numeric field and later to handle this missing data I replaced the null value with mean of HorsePower attribute. Besides HorsePower all other attributes were having proper datatype.

Further working with the numerical data there were no missing values but we need to check if there are any outliers in the data as they would impact our model. To detect the outliers, I used the boxplot chart as it would clearly help us to find out the outliers. I found HorsePower attribute had a very few outliers, which were not extreme from the upper boundary of outlier, hence I decided to keep them in the dataset.

Further deeply analysing the data I found that:

1. Our target variable ‘Miles Per Gallon’ is approximately normally distributed.
2. Most of the automobile car contains 4 cylinders.
3. Many cars were manufactured in 1973 than other years.
4. Most of the cars are manufactured in United States.

Before building the model, I also checked how other attributes of car are co-related with Miles Per Gallon attribute and here is what I found (as shown in Figure 1 in Appendix):

1. If the number of cylinders in the car are less, the car will have higher miles per gallon capacity.
2. If the displacement of the car is less, the car will have higher miles per gallon capacity.
3. If the HorsePower required for the car is less, the car will have higher miles per gallon capacity.
4. If the car is lighter in weight, the car will have higher miles per gallon capacity.

However, just looking at the correlations doesn't work as our model might include several variables which might be statistically significant. Hence we will now build a multiple linear regression model model to know which attributes helps the Manufacturer to increase their sales of cars.

After implementing linear regression on the dataset, we were able to get a bunch of attributes who contribute the most to higher miles per gallon. I would like to mention those attributes which had the most impact:

1. Displacement
2. Weight
3. Model Year
4. US Made

I came to this conclusion after analyzing the p-value from the summary of the regression model I ran as shown in Figure 1. in Appendix. I figured out the predictors which had the p-value less than the level of significance which is 0.05, which means the above attributes are statistically significant. From the model results in the Figure 1. in the Appendix we can see that we got the accuracy of 82.3%. I calculated this accuracy based on the summary statistics generated from the model, and used the adjusted R-squared metric. Adjusted R-squared is used to determine how reliable the correlation is and how much it is determined by the addition of independent variables.

So, from the model results, we can observe that the manufacturer needs to focus on the above significant car attributes and their coefficients values to know the dependency (in the Figure 1. in the Appendix) to build an efficient car which would have higher miles per gallon capacity. As higher the coefficient value indicates the attribute is more positively or negatively impacting the target attribute.

Using the p-value to eliminate insignificant attributes is not always advisable because the results may not be always accurate. As a result, I used Feature Selection techniques, which is a method of identifying the key attributes that are most relevant in predicting the outcome variable. I have used the Backward Elimination and Forward Selection techniques separately to give those key attributes.

First, I implemented the Forward Selection technique in which we start with having no feature in the model. In each iteration, we keep adding the feature which best improves our model till an addition of a new variable does not improve the performance of the model. After implementing this technique, I found that Weight of car, Year in which the car was designed, Displacement, Acceleration and where the car was designed in US or outside US plays a major role in designing a car with high MPG. I built a new regression model with these features and got the accuracy from the Adjusted R square value as 82.1% (Figure 3 in Appendix).

Later I implemented a Backward Elimination technique, in which we first run model iteratively using all the predictors and start eliminating predictors in each iteration who has p-value less than significant level i.e., 0.05 and here are the final attributes we got and building final regression model:

1. Displacement
2. Weight
3. Model Year
4. If the car is made is car

According to this technique above attributes play significant role in higher MPG, and we got an accuracy of 82.0% for this model.

Hence comparing all three models as shown below, we can observe that our generalized linear regression model implemented without any feature selection technique gave the highest accuracy of 82.3% as compared to other model so we will ask the manufacturer to use the GLM model and focus on the attributes Displacement, Weight, Model Year, US Made to design an efficient car with higher MPG. We will also ask them to focus on coefficient values of those attributes, as higher the coefficient value indicates the attribute is more positively or negatively impacting the target attribute.

|  |  |
| --- | --- |
| Model | Accuracy |
| Generalized Linear Regression Model (GLM) | 82.3% |
| Regression with Forward Selection | 82.1% |
| Regression with Backward Elimination | 82.0% |

# **Conclusion**

So, after analyzing all the attributes of car and building a generalized linear regression model which gave highest accuracy amongst all models which is 82.3%, we came to a conclusion that Displacement, Weight of the car, Year in which car was designed, if the car was designed in US or not are significant factors which would help the manufacturer to design a fuel-efficient car with higher MPG. So based on these factors now management can accordingly change the car's design, business strategy, and other factors to increase the car sales.

# **References**

Himanshi. *Forward Feature Selection: Implementation of Forward Feature Selection*. Retrieved from. <https://www.analyticsvidhya.com/blog/2021/04/forward-feature-selection-and-its-implementation/>

Srinidhi. *Backward Elimination for Feature Selection in Machine Learning*. Retrieved from. <https://towardsdatascience.com/backward-elimination-for-feature-selection-in-machine-learning-c6a3a8f8cef4>

# **Appendix**

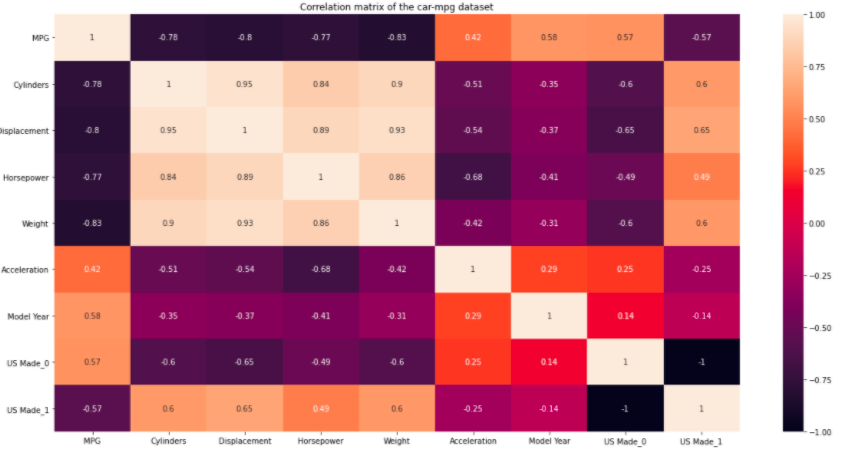


Figure 1

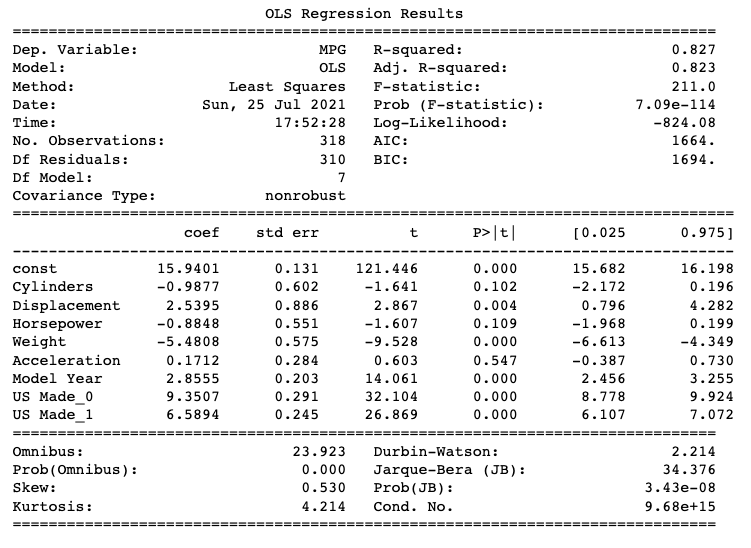


Figure 2

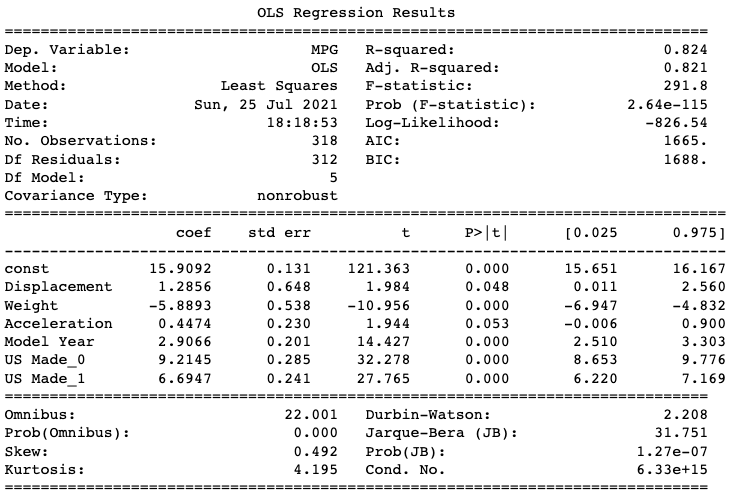


Figure 3

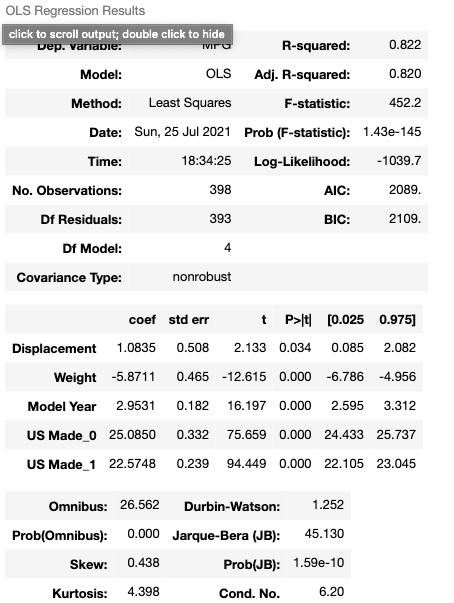


Figure 4