**Building the Car of the Future**

**Saad Khan**

**Introduction**

The challenge at hand is to determine the characteristics that lead to increased gas mileage in order to assist a struggling automaker in designing an energy-efficient vehicle. Developing a linear regression model that precisely forecast a car's miles per gallon (MPG) based on its characteristics such as the number of cylinders, displacement, horsepower, weight, acceleration, model year, and place of origin is the goal. The first section covers appropriate methods for data purification, including how to handle missing numbers, deal with outliers, verify the data's integrity and consistency, and transform or normalize the data.

Part 2 requires developing a linear regression model to forecast MPG based on vehicle characteristics and identifying the key variables influencing fuel efficiency through statistical methods. To create a model that can accurately predict MPG and identify the most significant predictors of fuel efficiency, Part 3 entails optimizing the linear regression model using any of the selection techniques to improve its performance and determine which attributes contribute to higher MPG over others.

The provided dataset, which includes details on fuel consumption and other characteristics for a variety of car models from the late 1970s and early 1980s, is a well-known and often used dataset in machine learning and data analysis. The MPG rating, weight, displacement, number of cylinders, horsepower, model year, acceleration, and origin are among the eight variables that comprise the 398 observations. Because of its practical importance and modeling issues, the dataset has also sparked several academic publications and discussions in the field of statistics and data analysis.

**Data Cleaning**

**Part 1**

Several processes are involved in the data cleaning process to guarantee high-quality data before the dataset is ready for analysis. The dataset's missing values are first found and appropriately handled. In order to allow Python's data analysis packages to manage missing values, the dataset expresses missing values using "?" characters that are substituted with NaN values. It's also important to take into account the unknown "?" observations in the "Horsepower" variable, which had an Object data type.

The pd.to\_numeric() function from the Pandas library was utilized to solve this problem. It transforms a DataFrame column from a non-numeric datatype, such as texts or objects, to a numeric datatype, such as integers or floats. Values in strings that should be numbers can be converted into actual numbers with this function. The median value of the column was used in place of the "?" observations. The distribution was skewed, therefore even though the missing values were low, we decided to use the median rather than the mean to maintain data integrity and make sure the analysis was supported by as much data as feasible.

After addressing any outliers that can significantly affect the outcome, the data is ready for analysis. Three variables were found to have outliers: MPG, Horsepower, and Acceleration. A box plot was made to analyze all the numerical variables. With the help of the interquartile range (IQR) approach, a total of eighteen outliers were found. After more examination, it was found that the outliers appeared to be within the range of common automobile scenarios and were not considerably skewed. For example, automobiles on the market typically have between 180 and 200 horsepower on average (Copilot, 2022), however sports cars can have more horsepower. Due to the short size of the dataset, all outliers were kept since there was insufficient data to support their removal. The box plot utilized to find the outliers is displayed in [[FIG 1]](#fig1).

I constructed histograms [[FIG 2]](#fig2) and performed data analysis to make sure the data is normally distributed. Each variable had its own histogram, and when the MPG histogram [[FIG 3]](#fig3) was overlaid with a normalization curve, it was clear that the data was in the expected format. Thus, nothing needed to be changed in this area. For the cylinder variable, I again made a frequency plot [[FIG 4]](#fig4) and saw that 51.25% of the autos had four cylinders. It was surprising to see that there were more cars with eight cylinders than six.

**Linear Regression Model**

**Part 2**

Prior to building the model, I wanted to inspect if there does exist any multicollinearity. For that purpose, I created the correlation plot [[FIG 5]](#fig5) so as to study the relationships between the variables.

According to the linear regression assumption, the output displays the findings of the analysis of the relationships between the various variables in the dataset. The dependent variable of mpg was shown to have a negative correlation with the variables that includes displacement, horsepower, weight, and cylinders. This suggests that an increase in any of these variables will result in a decrease in the mpg number. However, the non-multicollinearity assumption of linear regression is broken by the significant positive correlation that exists between these independent variables. This indicates that there is multicollinearity between these variables, which may have an impact on our regression model's performance and accuracy. In order to eliminate some of these factors, feature selection must be done.

I used VIF to find multicollinearity in order to address the multicollinearity issue. Cylinders, displacement, and weight all have VIF values more than 10. That amounts to 10.67, 22.96, and 10.52 in that order. Since there is multicollinearity between them according to the VIF criterion, I eliminated those variables by leaving them out of the linear regression model I was creating and created a new dataset called "new\_sdc."

Then I split this dataset (new\_sdc) in 80:20, to trains (using 80% dataset) and to test (using 20% dataset) and developed the linear regression model. After which I fitted the model, and below are some insights about the result [[FIG 6]](#fig6).

Horsepower, acceleration, model year, and origin are the four independent factors, while MPG is the dependent variable. The model's strong R-squared score of 0.767 means that the independent variables account for 76.7% of the variability in MPG. When other independent variables are held constant, the influence of the independent variables on the dependent variable is displayed by their coefficients. With other independent variables held equal, the coefficient for Horsepower is -4.5094, meaning that, on average, MPG falls by 4. 5094 units for every unit increase in Horsepower. Likewise, holding other independent factors equal, MPG falls by 0.9862, rises by 2.5539, and falls by 2.3521 units for every unit increase in Acceleration, Model Year, and Origin.

Since the P-values for each independent variable are less than 0.05, they are all significant predictors of the dependent variable. Additionally, the model's low Skew and Omnibus values show that the residuals are regularly distributed. Nonetheless, the residuals may exhibit autocorrelation because the Durbin-Watson value is less than 2. There does not appear to be a significant multicollinearity issue, based on the Cond. No. value of 3.12.

The findings show a valid relationship between the coefficients, indicating that a rise in horsepower will lead to a corresponding rise in fuel consumption and a subsequent drop in the vehicle's MPG. In the same way, increased acceleration causes the transmission to use more gas, which lowers MPG. The correlation between Model Year and fuel efficiency is positive, suggesting that newer cars are more fuel-efficient. However, the dataset does not include other variables that support this assumption.

**Model Optimization using Selection Technique**

**Part 3**

In order to avoid multicollinearity in part 2 of the linear regression modeling process, I eliminated a few variables; nevertheless, in order to make the model better, I started over with all the variables using a method known as forward addition. To obtain the results, I utilized the 'Sequential Feature Selector' module from the 'mlxtend' package. With an AIC value of 1703 [[FIG 7]](#fig7), the first five variables I utilized were 'Displacement', 'Horsepower', 'Weight', 'Acceleration' and 'Model Year'. Afterwards, I narrowed the selection down to four variables: 'Horsepower', 'Weight', 'Acceleration' and 'Model Year'. These variables had a lower AIC value of 1702 [[FIG 8]](#fig8).

After this I tried the three variables, namely 'Horsepower', 'Weight' and 'Model Year' and it had an AIC value as 1701 [[FIG 9]](#fig9) and the r-squared value as 0.799. Later I also tried only two valriables (namely 'Horsepower' and 'Model Year') which had the AIC values as 1879 [[FIG 10]](#fig10), so I stopped there since a high AIC does not indicate a good model.

with an AIC value of 1701, the model containing the four variables 'Horsepower', 'Weight' and 'Model Year' was therefore the most useful and effective for me. Since the coefficients for 'Horsepower' and 'Weight' are all negative, it is likely that increasing any of these factors will lower MPG. Conversely, a positive coefficient of Model Year suggests that greater MPG values are linked to newer model years. The F-statistic of 415.9 with a relatively low p-value suggests that the model is statistically significant overall, and the R-squared value of 0.799 demonstrates that the model explains a significant percentage of the variance in MPG.

The Jarque-Bera test shows that the residuals are normally distributed, and the Durbin-Watson value of 2.088 implies that there is no significant autocorrelation in the residuals.

**Explanation about selected Model**

**Part 4**

Based solely on Horsepower, Weight, and Model Year, the model with the lowest AIC score (1701) and the highest R-squared (0.799) achieves the best fit and forecast accuracy. Compared to the previous models, this three-feature model more accurately predicts MPG. Out of these three characteristics, weight and horsepower have negative coefficients, which means that larger values of these characteristics correlate with lower MPG, as would be expected. The positive coefficient for model year indicates that newer models typically have higher MPG, most likely as a result of advancing engine technologies over time.

In conclusion, the project goal is met and predicted accuracy is optimized in the optimized model that includes features for horsepower, weight, and model year.

**Conclusion**

In conclusion, this study successfully developed a linear regression model to forecast vehicle MPG depending on important parameters like weight, horsepower, and model year. With an R-squared of 0.799, the final optimized model that included only these three features was able to explain a significant portion of the variance in MPG.

The way we improved our model was important. We used a method called sequential feature selection to figure out which factors were most useful and to deal with issues where some factors were too connected. We discovered that more powerful and heavier cars tended to have lower MPG, which makes sense because they need more fuel. On the other hand, newer cars were linked to higher MPG, probably because of better fuel efficiency in modern engines.

In the end, we successfully created a model to predict MPG based on car features. The model, which uses three main factors, gives us useful information about what affects fuel efficiency. It's reliable and could help the carmaker design more fuel-efficient vehicles by concentrating on reducing weight and horsepower without sacrificing performance. To make the model even better, we could consider adding more details like engine size, drivetrain setup, aerodynamics, and other factors.

In summary, this analysis and modeling work identified the optimal predictive model for vehicle MPG and revealed horsepower, weight, and model year as the most influential attributes. These findings can guide engineering and design decisions to improve fuel efficiency. Additional real-world testing and validation would be needed before implementing any changes based on the model.

In summery, we found the best model to predict how many miles a car can go on a gallon of gas. The most important things that affect this are horsepower, weight, and how new the car is. These findings can help make decisions about how to build cars that use fuel more efficiently. Before making any actual changes based on the model, we'd need to test and make sure it works in real-world situations.