**Introduction**

The goal of this project is to determine which of the independent variables or attributes is contributing to higher gas mileage so the company can design a more fuel-efficient car. We use a regression model since the value we are the predicting/ output is a continuous value. Initially we clean our dataset to get optimum results and then we divide them into test data set and train data set and then we use a regression model to predict the miles per gallon and check the accuracy of the model using R squared value and AIC values. We then determine which of the attributes are contributing the most to miles per gallon and what changes we can make to increase this value.

**Data Cleaning**

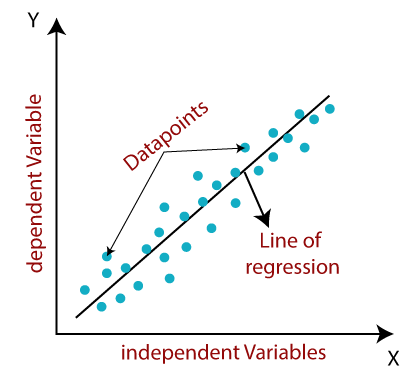
Data cleaning is one of the most important steps in the data analysis process and to get accurate results. For the purposes of this project, we have been given a car dataset which contains all the features considered to be important for the determination of miles per gallon. Our dataset has no missing values, so we now check for values which are inconsistent. In the horsepower column we found values that have a question mark. These values have no meaning and is not contributing to the dataset or the predictions in any way. So, we remove these rows from the dataset. We now plot graphs for all the remaining attributes i.e., Cylinders, displacement, horsepower, weight, acceleration, model year and US made. These graphs seem to normal distribution and no outliers. While we do see some outliers and huge standard deviations, these are relevant and right data, so these rows are not removed from the dataset. Our regression model only accepts numerical values. So, now we check if any of the columns are not in numerical format. It can be observed that Horsepower column is in object format, so we convert this column into an integer column. We can now input this data into the model and predict our dependent variable.

**Data Loading**

Initially we load the car data using scikit learn which is a machine learning library.

We then manually divide the data into training and testing sets. We use the training dataset to train the model and then we use the same model on our testing dataset so we can compare the results from training data and testing data to check the accuracy of the model.

**Linear Regression model**

It is a statistical method that is used for predictive analysis. Linear regression makes predictions for continuous/real or numeric variables. The linear regression model provides a sloped straight line representing the relationship between the variables [1]

There are 2 types of linear regression, simple and multiple linear regression. Simple for when we have a single independent variable and multiple for when we have multiple independent variables. These are the variables we use to predict our dependent variable which is prices in this case. The formula would be y=mx+c. We can use sklearn or statsmodels library in python for our model.

For our model, we input all the 7 attributes to determine the most and least important attributes that are contributing to predicting the miles per gallon value. Please refer Appendix A result 1 for the results. The P value is least for Weight, Model year and US made which means these have the highest significance in determining the miles per gallon value. We now look at the coefficient values. Negative value indicates negative impact and positive value indicates positive value. For weight and US made we see a negative value which means US made cars tend to have less Miles per gallon which can be investigated. The company can also look at ways to reduce the weight which is also impacting the MPG in a negative way. The model year however has a positive impact on miles per gallon which is mostly because the CAFE standards are given and improved every year by congress. We can see that we have a very high R square value which doesn’t always mean the model is accurate, so we also consider the AIC value which is 1650. This model looks accurate, but we can try to improve the accuracy by doing forward or backward selection.

**Backward Selection**

In the backward selection method, all the predictor variables we chose are added into the model. Then, the variables that do not (significantly) predict anything on the dependent measure are removed from the model one by one until removing a variable does not add anything to the model anymore. The ranking in our model is shown below. We determined this result from the p value when we input all 7 variables into the model.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 1 | 2 | 3 | 4 | 5 |
| Weight  US made  Model year | Displacement | Horsepower | Cylinders | Acceleration |

We now remove the variable which is least significant i.e., Acceleration. The results are in appendix A result 2. We can observe that the R squared value is 0.826 and AIC 1659. We now remove the next least significant variable, Cylinders. Appendix A result 3 shows the results of this model. We can observe that the AIC value has improved and is now 1658. So, we continue to remove the least significant variables one after the other. The AIC value reduces from 1658 to 1644 and then when another variable is removed to 1640. The results are shown in Appendix A result 4 and 5. When another variable is removed we can see that the AIC value increases. So we stop the backward selection here and choose **US made, Model Year and Weight** as the variables with most significance and as the model with most accurate prediction

**Forward Selection**

In the forward selection method, we use a model/method that looks at all the predictor variables you selected and picks the one that predicts the most on the dependent measure. That variable is added to the model. This is repeated with the variable that then predicts the most on the dependent measure. This little procedure continues until adding predictors does not add anything to the prediction model anymore.

We use SelectKBest for the feature selection and chi2 for the score we use to select these features.

1. SelectKBest: The SelectKBest class just scores the features using a function (in this case chi2 but could be others) and then "removes all but the k highest scoring features"
2. Chi2: The chi2 function from the sklearn feature selection package returns the chi-square statistic and the p-value.

We can give the k value as 1 to determine the feature with highest score. We get weight and when we apply it to our model, we get results as shown in Appendix A result 6. We now input k=2 and find the next significant feature. According to the model, we get the next significant feature as displacement and then we add Horsepower and Cylinder as indicated by the model. The results are in Appendix A results 7,8 and 9. Then we add Acceleration and see that the model is not improving as shown in Appendix A result 10. Hence, we stop here and consider 4 features: **Weight, Displacement, Horsepower, and Cylinder**. This model has a R square value of 0.717 and AIC value is 1774.

**Conclusion**

From the 2 selections methods above, we see in backward selection we have 3 attributes,

**US made, Model Year and Weight** and in the forward selection we have 4 attributes, **Weight, Displacement, Horsepower, and Cylinder.** When we have to provide a report to the car makers we considers attributes which they have control over and those they can change to increase the MPGs. US made cars tend to have less efficiency but logically it is impossible that a place of manufacturing has that much impact over the MPGs. Every year, EPA sets a fuel economy requirement for car markers which is also not under their control. With every year, the fuel economy increases (MPG increases) because of the requirement given by the government. Weight is the common feature in both the forward and backward selection. We can see that as the weight of the vehicle increases the MPGs tend to reduce from the negative coefficient values. Detroit's Big Three automakers, General Motors, Ford Motor Co and Chrysler parent Stellantis lag other automakers in fuel economy performance as they sell a rising number of large trucks and SUVs, a government report released on Friday shows [2]. The next significant feature we see is Horsepower. Higher the horsepower, more fuel gets consumed. So, we see a negative value for the feature. Higher number of cylinders also decreases the MPG. Hence, we see a negative coefficient value. The cylinders, displacement and horsepower are all interrelated and effect the MPG. The displacement has least significance compared to all the other variables. Most of the highest-mpg vehicles in the industry right now are hybrids—meaning they combine internal combustion engines and electric motors to save fuel. Most also have electronic continuously variable transmissions (eCVTs), which can save energy by regulating engine speed [3]. There are a lot of ways manufacturers can create smaller engines that provide an enjoyable acceleration without chugging down gasoline as larger engines do [4]. Sleeker body designs that reduce drag can improve fuel economy. Sedans are usually more aerodynamic than large, boxy SUVs and trucks. Improvements in this area can result in less wind noise and sleeker, sportier designs [3].

**References**

1. [Linear Regression in Machine learning - Javatpoint](https://www.javatpoint.com/linear-regression-in-machine-learning)
2. [Detroit's Big Three automakers lag industry on fuel economy (msn.com)](https://www.msn.com/en-us/money/companies/u-s-vehicle-fuel-economy-hits-new-high-in-2020-epa/ar-AAQUhtn)
3. New Ways Carmakers Are Getting You More MPG [New Ways Carmakers Are Getting You More MPG - Consumer Reports](https://www.consumerreports.org/fuel-economy-efficiency/new-ways-carmakers-are-getting-you-more-mpg/)
4. [2022 Cadillac CT4-V Blackwing vs. 2022 BMW M3 Competition: Which Sports Sedan Is Sharper? (motorbiscuit.com)](https://www.motorbiscuit.com/2022-cadillac-ct4-v-blackwing-vs-2022-bmw-m3-competition-sports-sedan-sharper/)

**Appendix A: Results**

*Result 1*

*Table

Description automatically generated*

*Result 2*

*Table

Description automatically generated*

*Result 3*

*Table

Description automatically generated*

*Result 4*

*Table

Description automatically generated*

*Result 5*

*Table

Description automatically generated*

*Result 6*

*Table

Description automatically generated*

*Result 7*

*A picture containing table

Description automatically generated*

*Result 8*

*Table

Description automatically generated*

*Result 9*

*Table

Description automatically generated*

*Result 10*

*Table

Description automatically generated*