Decoding Car Performance and Efficiency: A Data-Driven Approach

---

1 The Issues

Numerous characteristics of several cars, including their displacement, horsepower, weight, acceleration, and miles per gallon, are represented in this dataset. It consists of numerical values, where each row corresponds to a particular car. The features give details about the car's weight, performance, and engine.

The miles per gallon (mpg) metric of a car's fuel efficiency can be used to assess the relationship between its many components. It can also be used to create a predicted model for fuel efficiency depending on the car's weight or horsepower, among other characteristics. The dataset can also be used to spot any anomalies or outliers in the data that might point to measurement errors or other problems.

1. Is at least one of the predictors useful in predicting the response?
2. Do all the predictors help to explain the response, or is only a subset of the predictors useful?
3. How well does the model fit the data?
4. Given a set of predictor values, what response value should we predict, and how accurate is our prediction?

2 Findings

The regression analysis shows that "displacement" and "weight" have negative correlations with "mpg", while "horsepower" has a significant negative correlation with "mpg". "Acceleration" has a positive correlation with "mpg", but it is not statistically significant. The overall model explains a significant amount of the variance in "mpg", but the residuals are not normally distributed.

3 Discussions

The model only includes four independent variables, which may not capture all the factors that influence the dependent variable "mpg". There may be other variables, such as engine size, aerodynamics, or driving conditions, that could have a significant impact on "mpg" but are not included in the model and the independent variables in the model may be highly correlated with each other, which can lead to instability in the coefficients and unreliable results. It is important to check for multicollinearity among the independent variables before interpreting the results.

The model may be sensitive to outliers, which are observations that are significantly different from the rest of the data. Outliers can have a large influence on the regression coefficients and may distort the results. The relationship between the independent variables and the dependent variable may not be linear, and a linear regression model may not capture this relationship. It may be necessary to use a more complex model, such as a polynomial regression or a non-parametric regression, to capture the true relationship.

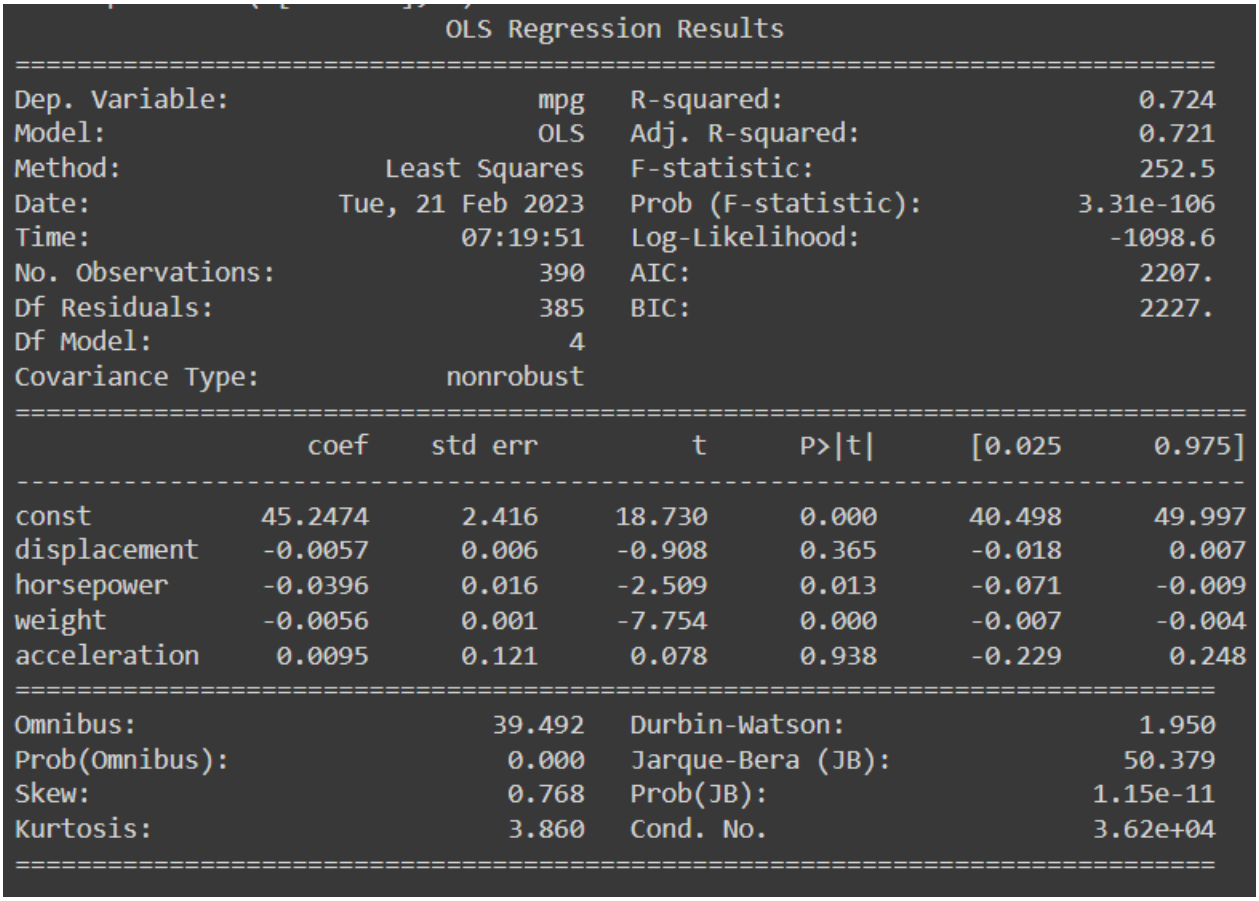
The model is based on a sample of only 390 observations, which may not be representative of the entire population. It is important to consider the sample size and the sampling method when interpreting the results. There may be feedback loops or reverse causation between the independent and dependent variables, which can violate the assumptions of the regression model and lead to biased results. For example, it is possible that "mpg" affects "weight" or "horsepower" rather than the other way around.

In summary, it is important to consider these potential issues when interpreting the results of the regression analysis and to exercise caution when making any conclusions or decisions based on the results.

4 Appendix A: Method

Multi-linear regression is a statistical method used to analyze the relationship between a dependent variable and two or more independent variables. In the case of an auto dataset with variables such as displacement, horsepower, weight, acceleration, and mpg, multi-linear regression can be used to determine the relationship between these variables and fuel efficiency.

Initially, collect the numeric data given in the website and clean it by removing any missing values. Then do scatter plots to visualize the data between different variables. Use the multiple linear regression model to analyze the data and make predictions out of it. Using Python, I calculated the coefficients of the model and the goodness of fit like the R-squared value. After interpreting the data, we can try to understand the relationship between the



OLS Regression Results

Dep. Variable:	mpg	R-squared:	0.724
Model:	OLS	Adj. R-squared:	0.721
Method:	Least Squares	F-statistic:	252.5
Date:	Tue, 21 Feb 2023	Prob (F-statistic):	3.31e-106
Time:	07:19:51	Log-Likelihood:	-1098.6
No. Observations:	390	AIC:	2207.
Df Residuals:	385	BIC:	2227.
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	45.2474	2.416	18.730	0.000	40.498	49.997
displacement	-0.0057	0.006	-0.908	0.365	-0.018	0.007
horsepower	-0.0396	0.016	-2.509	0.013	-0.071	-0.009
weight	-0.0056	0.001	-7.754	0.000	-0.007	-0.004
acceleration	0.0095	0.121	0.078	0.938	-0.229	0.248

Omnibus:	39.492	Durbin-Watson:	1.950
Prob(Omnibus):	0.000	Jarque-Bera (JB):	50.379
Skew:	0.768	Prob(JB):	1.15e-11
Kurtosis:	3.860	Cond. No.	3.62e+04

Figure 1: Ordinary Least Squares (OLS) regression analysis

variables and make predictions for a new set of data.

5 Appendix B: Results

Figure 1 is a summary of an Ordinary Least Squares (OLS) regression analysis, which is a statistical method for modeling the relationship between a dependent variable (in this case, mpg, or miles per gallon) and more independent variables (displacement, horsepower, weight, and acceleration).

The summary provides information about the overall fit of the model (R-squared and adjusted R-squared), as well as the statistical significance of each independent variable (represented by the "coef" column, which shows the estimated regression coefficients, and the "P¿—t—" column, which shows the p-values for testing the null hypothesis that the corresponding coefficient is zero).

The regression equation is:

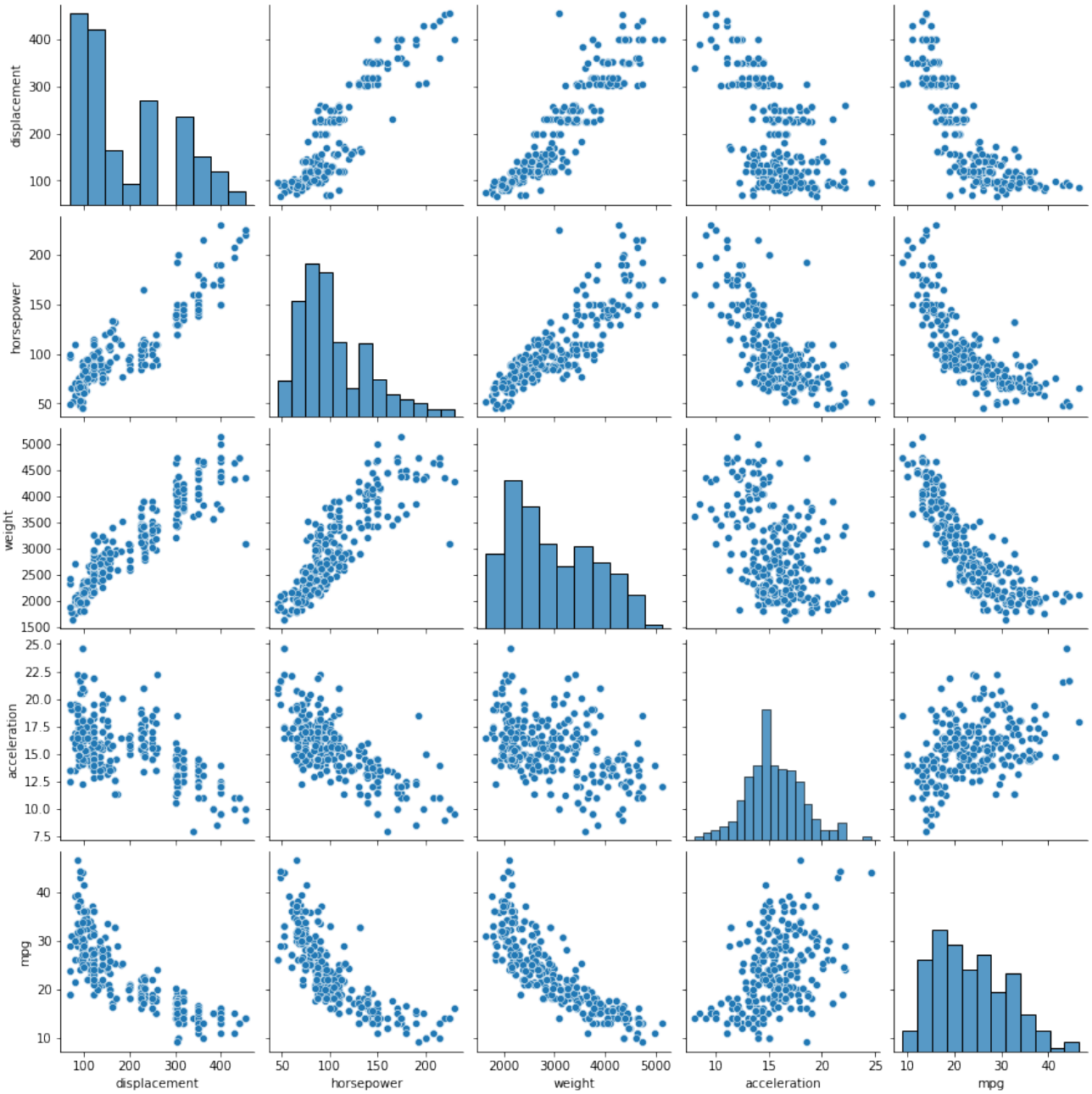


Figure 2: Correlation plots

mpg = 45.2474 - 0.0057(displacement) - 0.0396(horsepower) - 0.0056(weight) + 0.0095(acceleration)

According to the summary, the model has an R-squared value of 0.724, which means that about 72.4 percent of the variation in mpg can be explained by the independent variables in the model. The adjusted R-squared value is slightly lower at 0.721, which is a more conservative estimate of the model's fit that takes into account the number of independent variables and the sample size.

The F-statistic of 252.5 has a very small p-value (3.31e-106), which indicates that at least one of the independent variables in the model is statistically significant in predicting mpg. The individual p-values for each coefficient suggest that weight and horsepower are statistically significant at the 5 percent level, but displacement and acceleration are not.

The coefficients for weight and horsepower are negative, which means that an increase in weight or horsepower is associated with a decrease in mpg. The coefficient for acceleration is positive, but it is not statistically significant, so we cannot conclude that there is a meaningful relationship between acceleration and mpg.

The summary also provides information about the residuals of the model (Omnibus, Durbin-Watson, Jarque-Bera, Skew, and Kurtosis), which can be used to assess the normality and homoscedasticity of the errors. Finally, the summary includes information about the sample size (390) and the degrees of freedom for the residuals (385).

Is at least one of the predictors useful in predicting the response?

the p-value is less than a chosen significance level (usually 0.05), we reject the null hypothesis and conclude that at least one of the predictors is useful in predicting the response variable. In other words, there is evidence of a linear relationship between the independent variables and the dependent variable.

Therefore, if the p-value associated with the F-statistic in the model summary output is less than 0.05, we can conclude that at least one of the predictors is useful in predicting the response. If the p-value is greater than 0.05, we cannot reject the null hypothesis and conclude that none of the predictors are useful in predicting the response.

Do all the predictors help to explain the response, or is only a subset of the predictors useful?

To determine if all the predictors help to explain the response or if only a subset of the predictors are useful, we can look at the coefficients and their associated p-values in the multivariate linear regression model.

The coefficient estimates measure the strength and direction of the linear relationship between each predictor and the response variable. A positive coefficient indicates a posi-

tive relationship (as the predictor variable increases, the response variable increases) and a negative coefficient indicates a negative relationship (as the predictor variable increases, the response variable decreases).

The p-value associated with each coefficient measures the evidence against the null hypothesis that the coefficient is equal to zero, i.e., that the predictor is not useful in explaining the response variable. A p-value less than the chosen significance level (usually 0.05) indicates strong evidence against the null hypothesis and suggests that the predictor is useful in explaining the response variable.

Therefore, we can look at the coefficient estimates and their associated p-values to determine which predictors are useful in explaining the response variable. If a predictor has a coefficient estimate that is statistically significant (i.e., has a p-value less than 0.05), we can conclude that it is useful in explaining the response variable. If a predictor has a coefficient estimate that is not statistically significant (i.e., has a p-value greater than 0.05), we cannot conclude that it is useful in explaining the response variable.

It is also possible for multiple predictors to be useful in explaining the response variable, but with different strengths of association. In this case, we can compare the magnitude of the coefficient estimates and their associated standard errors to assess the relative strength of each predictor's association with the response variable.

How well does the model fit the data?

Since r value is 0.724 it is a good fit.

Given a set of predictor values, what response value should we predict, and how accurate is our prediction?

Given a set of predictor values, what response value should we predict, and how accurate is our prediction?

Given a set of predictor values, we can use the multivariate linear regression model to predict the corresponding response value. The predicted response value can be obtained by plugging in the predictor values into the regression equation.

To assess the accuracy of our prediction, we can use the predicted response value and compare it to the actual response value. One common metric for assessing prediction accuracy is the root mean squared error (RMSE), which measures the average difference between the predicted and actual response values. Lower values of RMSE indicate better prediction accuracy.

Another common metric is the coefficient of determination (R2), which measures the proportion of variation in the response variable that is explained by the predictor variables. Higher values of R2 indicate that the predictor variables are better able to explain the

variation in the response variable, and therefore better predict the response variable.

It is important to note that prediction accuracy is affected by the quality and quantity of the data used to train the model. Therefore, it is recommended to evaluate the model using a separate validation dataset or cross-validation technique to obtain an unbiased estimate of the prediction accuracy.

6 Appendix C: Code

Python code is used to plot and find out all the descriptive statistical values that are needed for necessary predictions and analysis. Google colab research cloud is used to execute and get the necessary results.

```
import pandas as pd
import seaborn as sns
import statsmodels.api as sm

# Read the Excel file into a DataFrame
df = pd.read_excel('https://mth522.files.wordpress.com/2023/01/auto_data_bijja

# Define the independent and dependent variables
X = df[['displacement', 'horsepower', 'weight', 'acceleration']]
y = df['mpg']

# Add a constant term to the independent variables
X = sm.add_constant(X)

# Fit the linear regression model
model = sm.OLS(y, X).fit()

# Print the model summary
print(model.summary())
sns.pairplot(df);
```