Analysis on Auto data using Multiple Linear Regression

1 The Issues

The data set is obtained as a subset of the Auto data mentioned in "An Introduction to Statistical Learning with Applications in R"

We address the following questions:
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

In this analysis, we were trying to figure out what factors impact how many miles per gallon (mpg) a car can get.

We looked at four different factors that could potentially influence mpg: displacement, horsepower, weight, and acceleration. We used a type of statistical analysis called "OLS Regression" to see which factors were most strongly related to mpg.

What we found was that two factors, weight and horsepower, had the biggest impact on a car's fuel efficiency. Specifically, we found that as a car gets heavier or has more horsepower, its mpg tends to go down. This makes sense, since heavier cars and more powerful engines require more fuel to move around.

On the other hand, we didn't find a strong relationship between mpg and the other two factors we looked at, displacement and acceleration. This means that having a larger engine or being able to accelerate quickly doesn't necessarily mean that a car will be less fuel-efficient.

However, there are likely other factors that also play a role in determining a car's fuel efficiency.

3 Discussions

While this analysis found that weight and horsepower were strongly correlated with fuel efficiency, it's important to remember that correlation doesn't necessarily imply causation. There may be other factors that are driving the relationship between weight/horsepower and mpg, and it's possible that changing weight or horsepower wouldn't necessarily result in a significant change in fuel efficiency.

As I mentioned earlier, this analysis only looked at four factors that might affect fuel efficiency. There are likely other variables that could play a role in determining a car's mpg, such as the type of fuel it uses, its aerodynamics, or how well it's maintained.

OLS Regression relies on a number of assumptions about the data, such as that the relationship between the dependent and independent variables is linear, that the errors are normally distributed, and that there is no multicollinearity (i.e., that the independent variables are not highly correlated with each other). It's possible that some of these assumptions may not hold for this dataset, which could impact the accuracy of the results.

4 Appendix A: Method

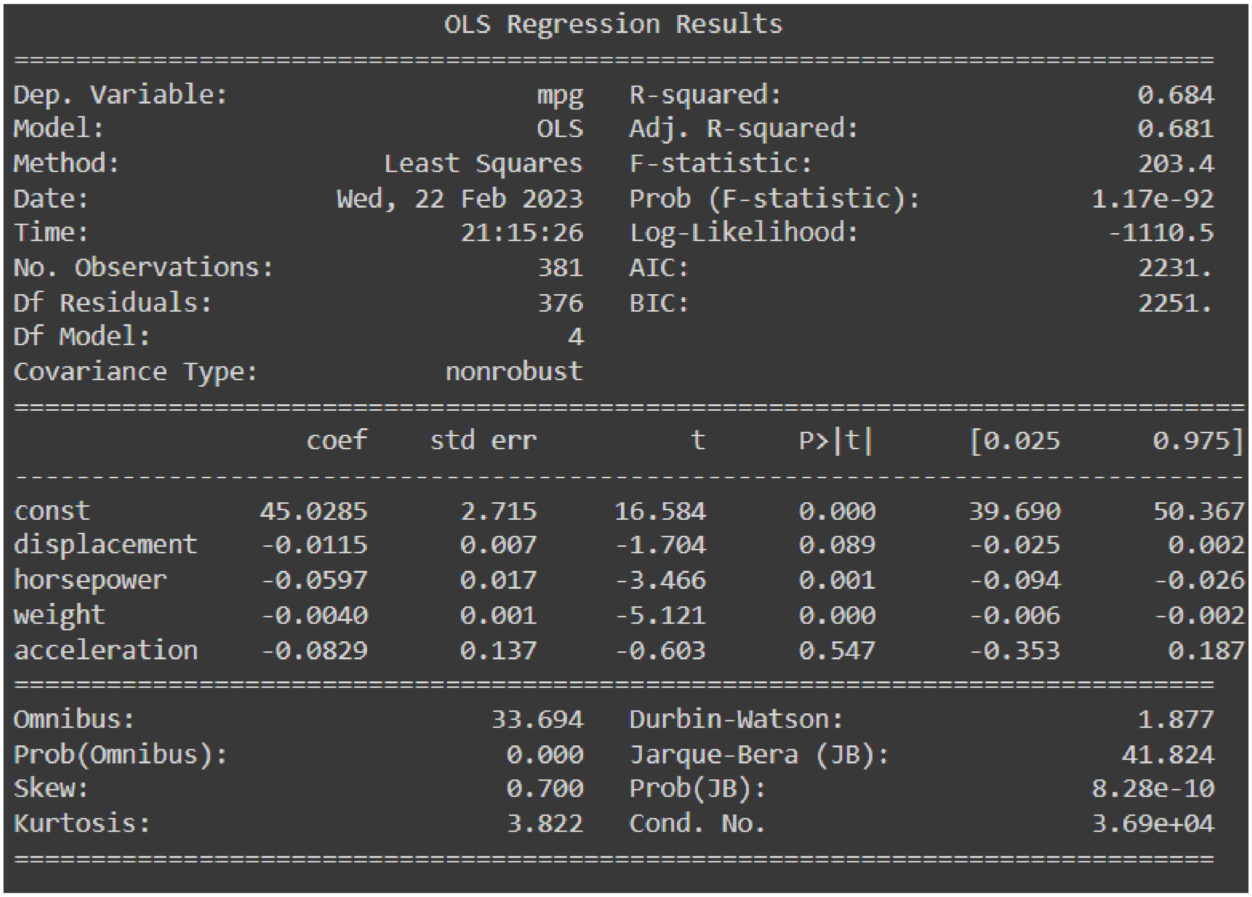
The data set is obtained as a subset of the Auto data mentioned in "An Introduction to Statistical Learning with Applications in R", Chapter 3, page 123. For this data set, there are 4 predictor variables: displacement, horsepower, weight, and acceleration, and one predicted (= response) variable: mpg.

Multi linear regression is a statistical method used to analyze the relationship between a dependent variable and two or more independent variables. I performed a linear regression analysis on a dataset of automobile performance and fuel efficiency data using the OLS method.

Data was downloaded as a comma separated (.csv) file and imported into colab. The independent variables are defined as displacement, horsepower, weight, and acceleration, and the dependent variable is defined as mpg. A constant term is added to the independent variables, and the linear regression model is fit to the data using the OLS() function from the statsmodels library. The model's coefficients are estimated using the fit() method, and a summary of the regression results is printed using the summary() method. Additionally, generated a scatterplot matrix of the dataset using the pairplot() function from the seaborn library, which shows the relationship between each pair of variables in the dataset.

5 Appendix B: Results

The results in Figure 1 were obtained from a linear regression model using the OLS (ordinary least squares) method. The R-squared value of 0.684 indicates that the model explains 68.4 percentage of the variance in the data. This means that the model fits the data relatively well. The adjusted R-squared value of 0.681 is similar and indicates that the model is not overfitting the data.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    mpg   R-squared:                       0.684
Model:                            OLS   Adj. R-squared:                  0.681
Method:                 Least Squares   F-statistic:                     203.4
Date:                Wed, 22 Feb 2023   Prob (F-statistic):           1.17e-92
Time:                        21:15:26   Log-Likelihood:                -1110.5
No. Observations:                 381   AIC:                             2231.
Df Residuals:                     376   BIC:                             2251.
Df Model:                           4
Covariance Type:            nonrobust
================================================================================
                   coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
const           45.0285      2.715     16.584      0.000      39.690      50.367
displacement    -0.0115      0.007     -1.704      0.089      -0.025       0.002
horsepower      -0.0597      0.017     -3.466      0.001      -0.094      -0.026
weight          -0.0040      0.001     -5.121      0.000      -0.006      -0.002
acceleration    -0.0829      0.137     -0.603      0.547      -0.353       0.187
==============================================================================
Omnibus:                       33.694   Durbin-Watson:                   1.877
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               41.824
Skew:                           0.700   Prob(JB):                     8.28e-10
Kurtosis:                       3.822   Cond. No.                     3.69e+04
==============================================================================
```

Figure 1: Results of Ordinary Least Squares (OLS)

The F-statistic of 203.4 and the associated p-value of 1.17e-92 indicate that the overall model is statistically significant.

The coefficients for each independent variable show the effect that each variable has on the dependent variable, holding all other variables constant. The constant term of 45.0285 indicates the predicted value of mpg when all independent variables are equal to zero.

The coefficient for displacement is negative, but the p-value of 0.089 is not significant, meaning that it is not clear if displacement has a significant effect on mpg.

The coefficient for horsepower is also negative, and the associated p-value of 0.001 is significant, indicating that horsepower has a significant negative effect on mpg.

The coefficient for weight is negative, and the associated p-value of 0.000 is significant, indicating that weight has a significant negative effect on mpg.

The coefficient for acceleration is negative, but the p-value of 0.547 is not significant, indicating that acceleration is not a significant predictor of mpg in this model.

The other statistics in the results provide information on the goodness of fit of the model, including the AIC and BIC values, the omnibus test for normality of residuals, the Durbin-Watson test for autocorrelation of residuals, and the Jarque-Bera test for overall model fit.

Also plotted scatterplot matrix of the dataset using the pairplot() function from the seaborn library, which shows the relationship between each pair of variables in the dataset.

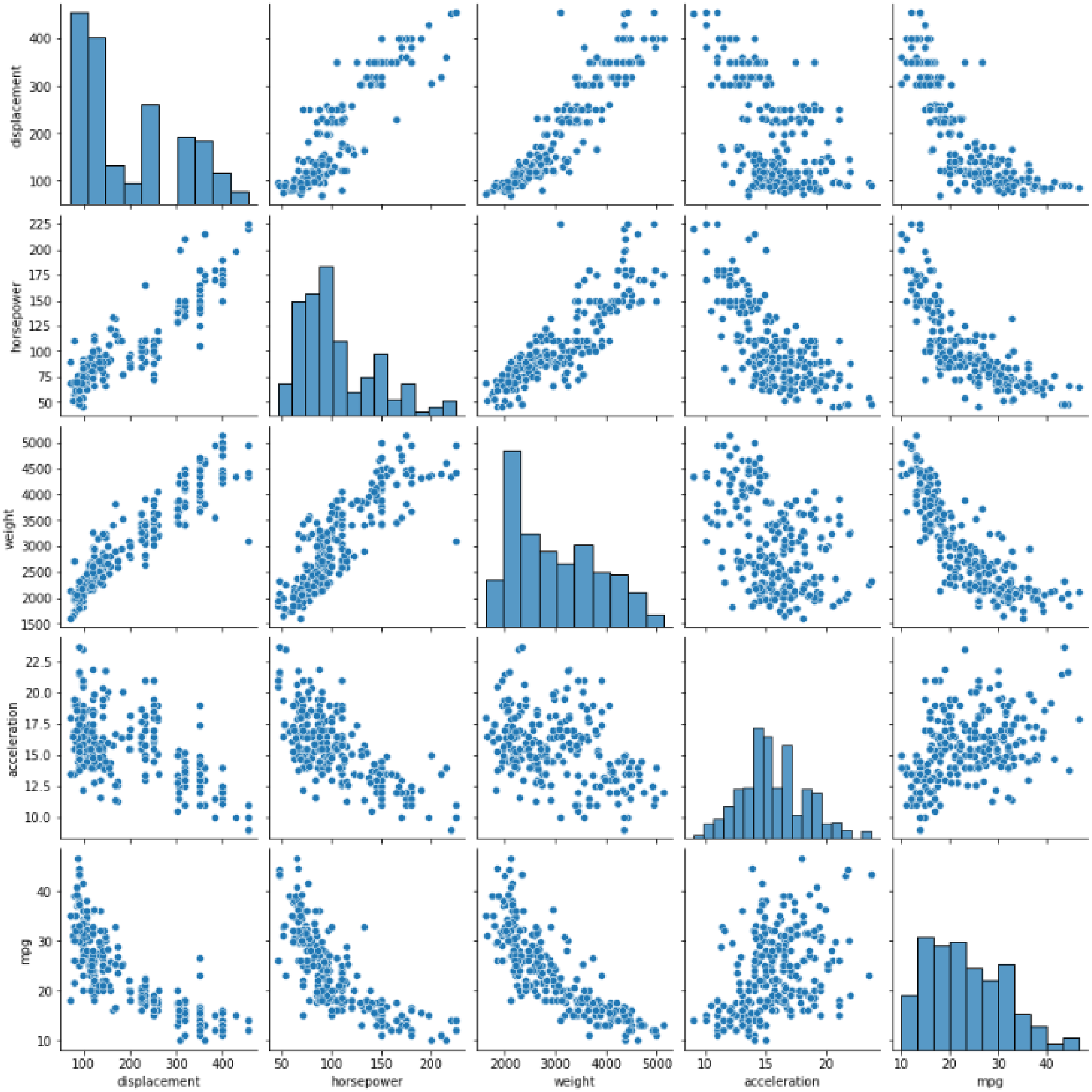


Figure 2: Scatterplot matrix which shows the relationship between each pair of variables in the dataset

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

Yes, at least one of the predictors is useful in predicting the response as the F-statistic has a very low p-value (1.17e-92) which suggests that the model as a whole is significant and at least one predictor is useful.

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

Not all the predictors help to explain the response, as the p-value of the acceleration predictor is high (0.547) which suggests that it is not significant and does not help in explaining the response as compared to the other predictors.

3. How well does the model fit the data?

The R-squared value of 0.684 suggests that the model explains 68.4 percentage of the variability in the response variable (mpg). This indicates that the model has moderate predictive power and is a relatively good fit for the data.

4. 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 predict the response value (mpg) using the coefficients of the predictors. The accuracy of our prediction can be assessed using the standard error of the regression or the confidence interval for the predicted value.

The regression equation for the above analysis can be written as:
mpg = 45.0285 - 0.0115(displacement) - 0.0597(horsepower) - 0.0040(weight) - 0.0829(acceleration)

6 Appendix C: Code

```
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_banala_shashank.xls')

# 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);
```