**Homework 1:**

Names :

* Mohamed Amine Belyamani
* Bhavin Vinod Jain

# **Section 2.4 Question 1:**

1. **Flexible Learning Method** **is Better** since there is many observations in our dataset, so having an overfitted model is unlikely, this will also lower the bias since the model will be more flexible,
2. **Flexible Method is Worse** due to the small number of observations, which poses the risk of overfitting. Also, any small changes in the data can cause the model to change due to high variance with the large number of predictors
3. **Flexible Learning Method** **is Better,** because having a more flexible model will pick more features of the non-linearity than a non-flexible model,
4. **Flexible Method is Worse,** due to the high irreducible error , which could lead to overfitting, although having a high variance doesn’t give us much information, if the data is large enough then a flexible method would be a better method with the risk of overfitting,

# **Section 2.4 Question 3:**

1. \

Diagram

Description automatically generated

* **Var():** Irreducible error for approximating a real-world problem
* **Variance**: since it is defined as the extent in which our model would change if we would use a different dataset, as the model flexibility increases, any small changes to the dataset would large changes in
* **Bias:** defined as the error introduced to the model by approximating to a real-world problem, if our model is more flexible this would lower the bias
* **Training MSE:** As flexibility increases, the model is more fitted to the real world problem, which in term lower the Training MSE,
* **Test MSE:** As Flexibility increases, the test MSE decreases to a minimum point, after that as the relevant variance start increasing we see an increase in the Test MSE, that increase comes from overfitting the data,

# **Section 2.4 Question 6:**

A **Parametric Approach** involve a 2-step approach by first assuming the model’s form , this model represents a relationship between the prediction and the predictors using parameters (A linear assumption for example change into looking for . This assumption reduces the problem of estimating entirely into trying to estimate the parameters.

On the other hand, a **Non-Parametric Approach** do not make any assumptions on the form of but seeking an estimate of f that can fit the data as large as possible. This means that it doesn’t rely on any distribution within the data,

**Advantages:**

* **Parametric Approach:**
* Simplify the complexity of estimating the model into try to estimate its parameters
* Have more statistical power because of the distribution in the data
* **Non-Parametric Approach:**
* More Robust, since it is a distribution free method, the model cannot be affected by outliers

**Disadvantages:**

* **Parametric Approach:**
* the model will usually not match the true unknown form of the model
* **Non-Parametric Approach:**
* the dataset needs to be large to increase of an accurate model estimate

# **Section 2.4 Question 8:**



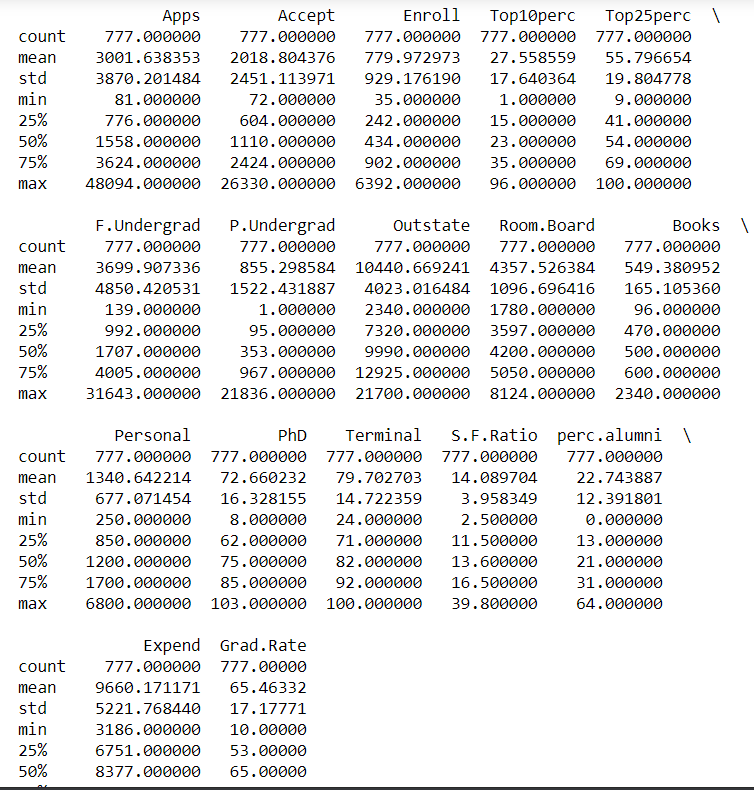
A screenshot of a computer

Description automatically generated with medium confidence

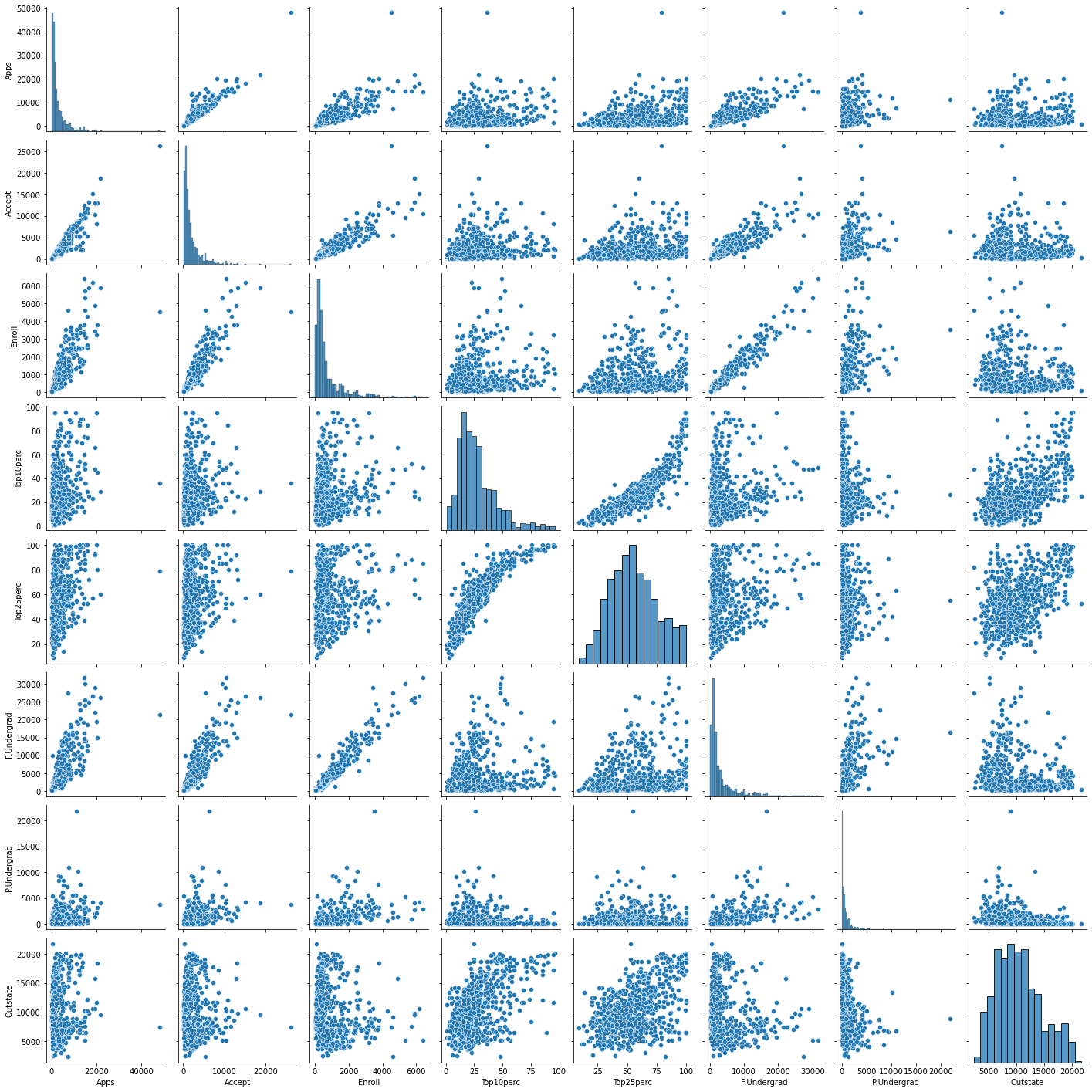
1. f

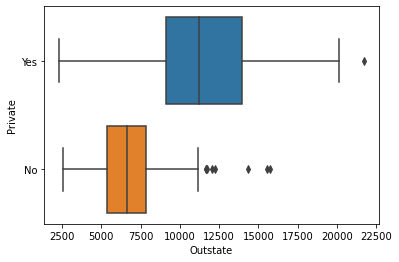
Graphical user interface

Description automatically generated

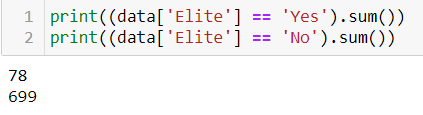


sns.pairplot(data.iloc[: , :10])

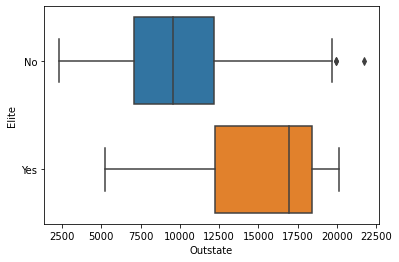


1. 





sns.boxplot(data = college, x= 'Outstate', y= 'Elite')





Text, application

Description automatically generated

A picture containing window, building

Description automatically generated

**Bins defined = 10**

**Default Histogram : No bins were defined**

1. We observe that the tuition for the Elite category schools is higher than the non-elite ones. Given the quality of Elite category schools it might be justifiable. Also, the number of Elite category institutions are 78 which is far less than the non-elite ones with the count of 699 thus the acceptance rates of non-elite schools are very low. The charge of the Private schools is higher than the public schools. There is difference in the quartile values of the Elite vs Outstate and Private vs Outstate boxplots.

# **Section 2.4 Question 9:**

1. **Quantitative and Qualitative Data:** Looking at the first rows the data set after importing it we can see which data is quantitative and which is qualitative,

Graphical user interface, text, application

Description automatically generated

* ***Qualitative Data:*** origin, name (dummy number refer to a country)
* ***Quantitative Data:*** mpg, cylinders, horsepower, weight, acceleration, year

1. **Range of Quantitative Variables :**

Range is given by the difference between the maximum values and the minimum values, creating a new data frame for quantitative predictors only, we can calculate the max and min of every variable and conclude the range of each variable,

Figure below show the range of the quantitative variables for the Auto dataset:



Table

Description automatically generated

1. **Mean and Standard Deviation:**



Table

Description automatically generated

1. **Mean and Standard Deviation (Dataset Edit):**



Text

Description automatically generatedTable

Description automatically generated

A picture containing text, crossword puzzle

Description automatically generated

* We notice some strong positive and negative between data
* Weight vs Acceleration for example shows no correlation relationship between the two data
* Some predictors show linear correlation while other show quadratic correlation (i.e. mpg vs horsepower)

Square

Description automatically generated

* Since the origin data is a categorical data, we can visualize the data depending on the origin of the car , 1 seems to be the USA, 2 is Europe (EU) and 3 is Japan, having this categorization, we can compare the different observations for each origin for more analysis,

A collage of plants

Description automatically generated with low confidence

* The Boxplot of displacement vs origin shown below shows that the USA median at a high level of 250 displacement and well separated than the EU or Japan,

Chart, box and whisker chart

Description automatically generated

* Showing data as a scatter plot below with the origin as our hue, we can clearly see that the US have more displacements

Chart, scatter chart

Description automatically generated

1. Predicting mpg:

* From the scatter plot, we notice that there are several predictors for mpg
* Origin also seems like a good predictor
* Year and acceleration also seem to be good predictor for mpg

Chart, box and whisker chart

Description automatically generated

Chart, scatter chart

Description automatically generated

Looking at the scatter plots above of displacement, horsepower, and weight, we can clearly see a quadratic relationship with the dependent variable (mpg), this shows us that we have a number of independent variables that can help predict mpg

# **3. Section 3.7 Question 1:**

* Newspaper p-value is large and positive at 0.8599, meaning that there is not really any relationship between newspaper advertising and sales,
* Since the p-value for TV is zero, which suggests that there is strong evidence that TV advertising is related to sales. i.e., by fixing the amount of radio and newspaper advertising, a rise in TV advertising can lead to a higher number of sales,
* Since the p-value for radio is very small, which suggests that there is strong evidence that radio advertising is related to sales. i.e., by fixing the amount of TV and newspaper advertising, a rise in radio advertising can lead to a higher number of sales,

# **Section 3.7 Question 5:**

Let’s consider the form,

Where,

We can rewrite the form above,

We can write the form as,

Where,

# **Section 3.7 Question 5:**

From Table (3.4) we have,

A picture containing text, clock

Description automatically generated

For a simple linear regression model, we can write it as the following,

for , and replacing with its expression, we get,

this means that the least square lines always pass through the point ().

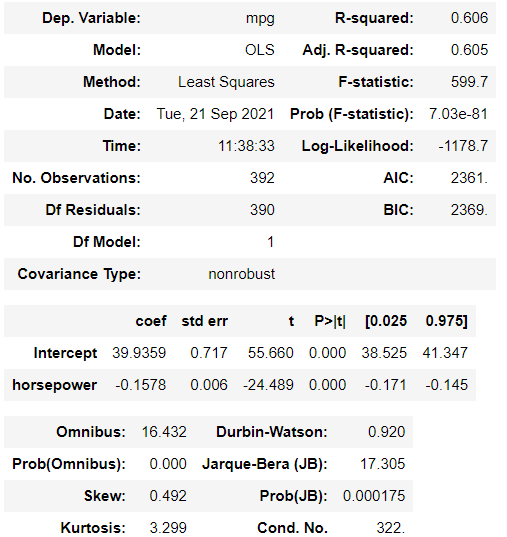
1. **Section 3.7, page 121-122, question 8**

**a.** Fitting linear regression model.

model = smf.ols('mpg ~ horsepower', data= data)

model = model.fit()

Summary:

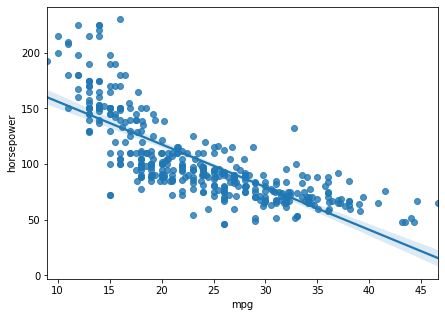


1. Yes, there is a relationship between predictor and response.
2. The R square value is 0.606, the closer it is to 1 the stronger the relationship. Also, the p- value of f-statistic is very small.
3. Since the coefficient of horsepower is –0.1578 which is negative the relationship is negative. The high the value of horsepower the less will be the mpg efficiency.
4. Mpg = (-0.1578)(horsepower) + 39.9359

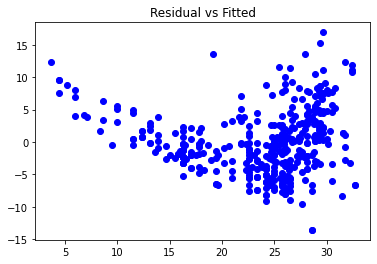
Mpg = (-0.1578)(98) + 39.9359

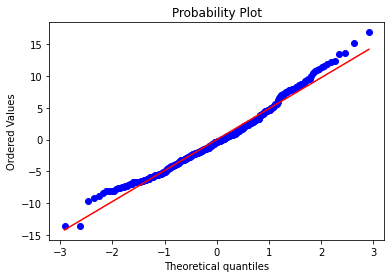
Mpg = 24.4715

b.



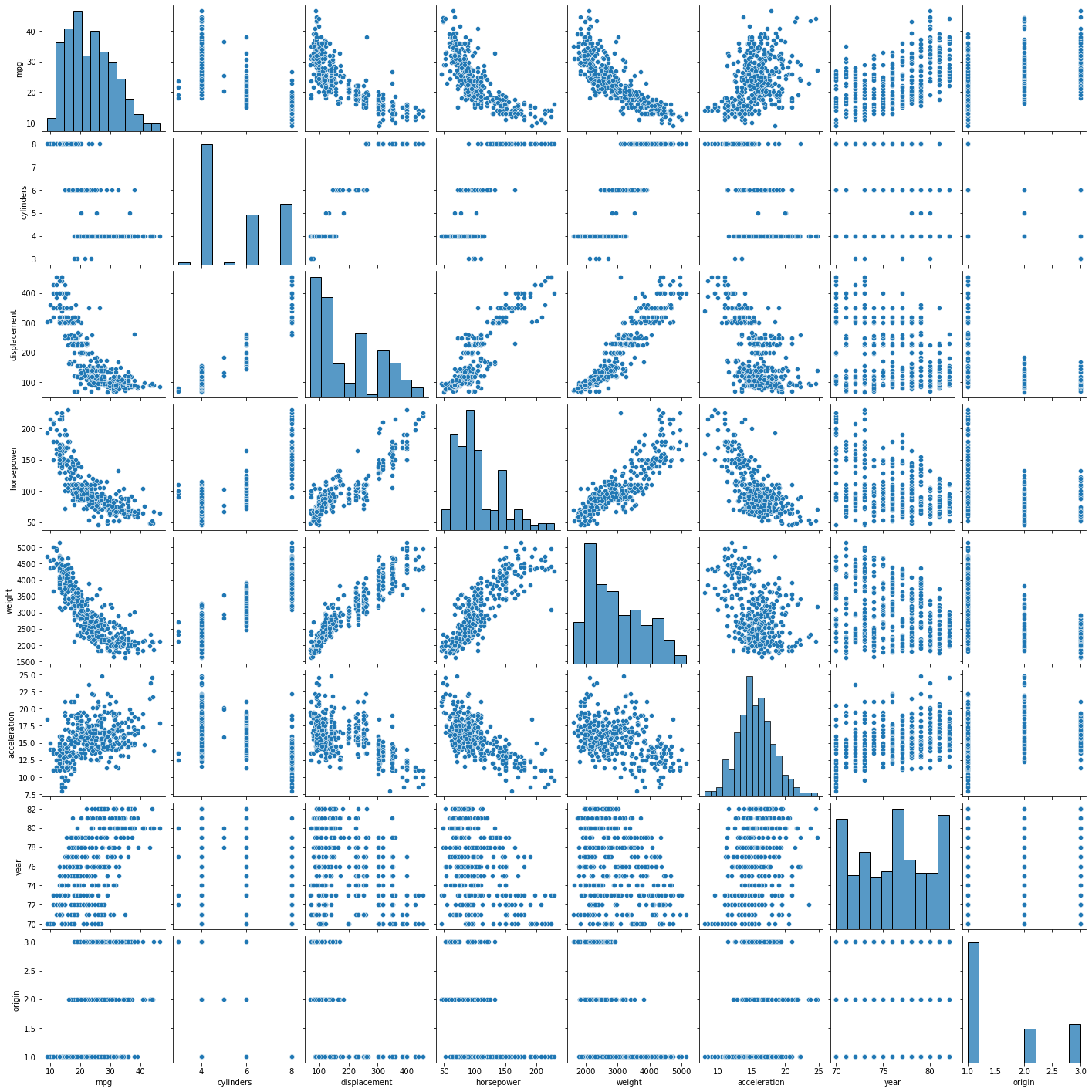
1. Diagnostic plots

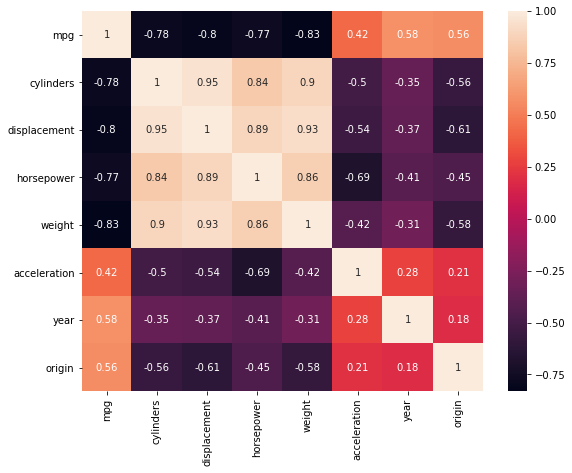




The U shape of the Residual vs Fitted plot indicate that there is non-linearity in the data. The probability plot confirms that fits look normal.

1. **Sec 3.7 Q9**

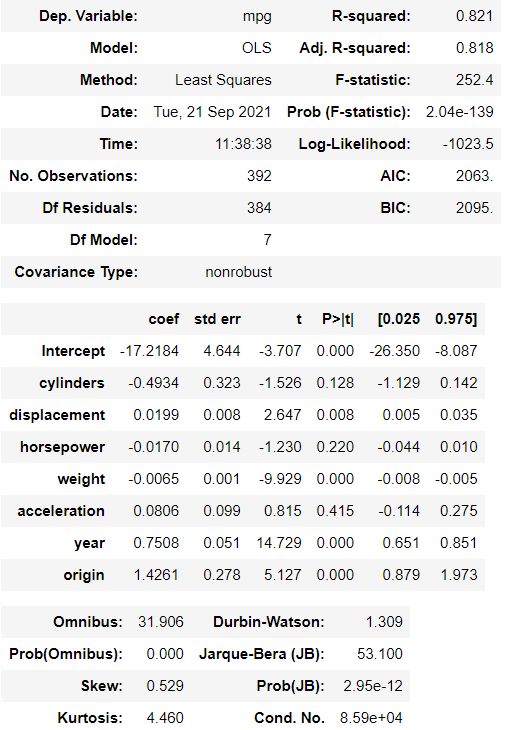




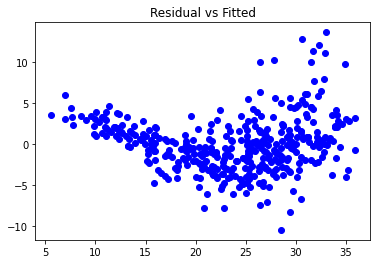
1. model2 = smf.ols('mpg ~ cylinders + displacement + horsepower + weight + acceleration + year + origin', data= data)

model2 = model2.fit()

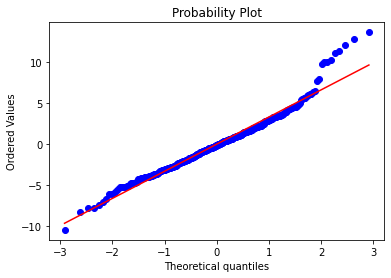
model2.summary()



* 1. Yes, there is relationship between predictors and response(mpg).
  2. Based on the p-value, all the predictors are statistically significant except for cylinders, horsepower and acceleration.
  3. The coefficient of year variable shows that an increase in fuel efficiency of 1 year is equivalent to an increase in average annual mpg of 0.7507727 by keeping all other predictors constant.

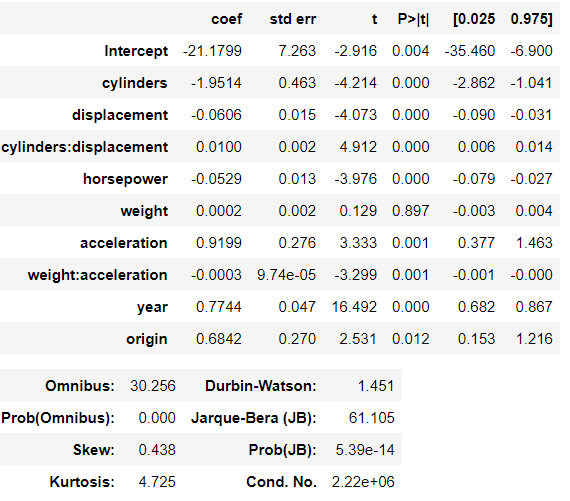


The U shape of the Residual vs Fitted plot indicate that there is non-linearity in the data.



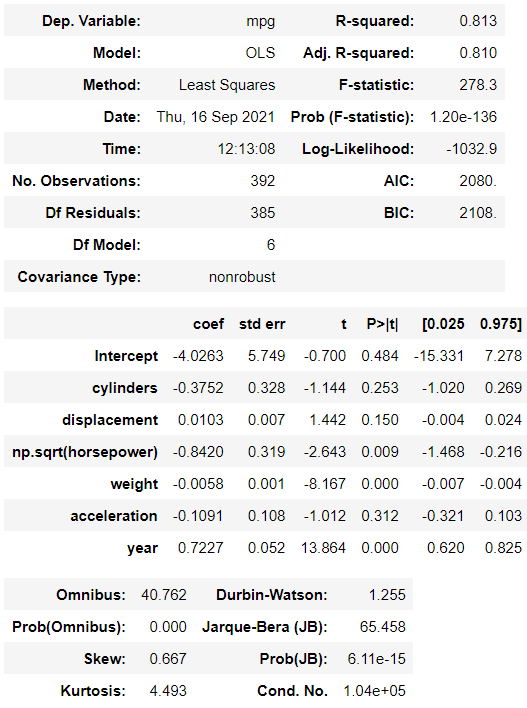
The probability plot confirms that fits looks normal. There are some outlier points in the data.

1. model4 = smf.ols('mpg ~ cylinders \* displacement + horsepower + weight \* acceleration + year + origin', data= data)

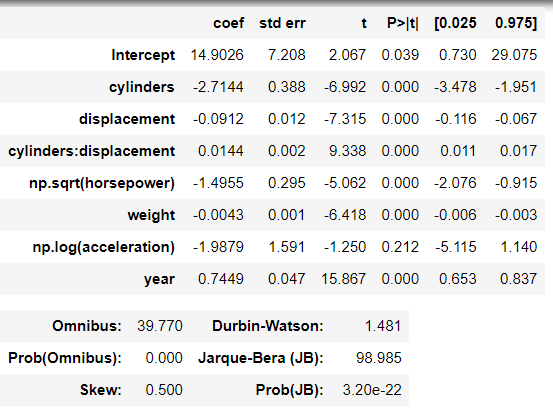


Looking at the p-values, we see that the interaction between cylinders and displacement, weight and acceleration is statistically significant.

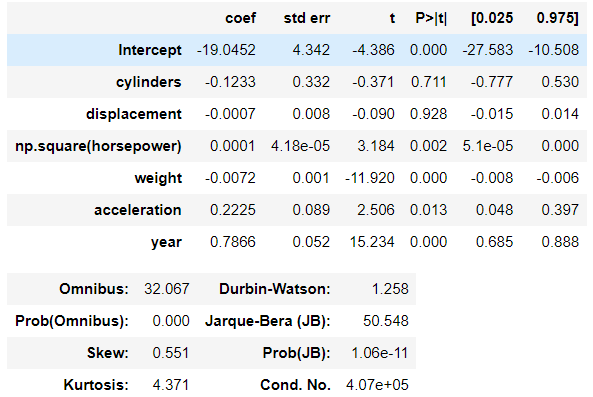
1. Experimenting with different transformations of the variables



Square root of horsepower is statistically significant as it has a low p-value of 0.009 but it is not statistically significant in the original LR model.



The log transformation of acceleration variable showed some improvement as the p-value got low but still the value is higher than 0.05 which shows it isnt statistically significant.



Taking the square of variable horsepower made the variable more statistically significant with the p-vale 0.002, compared to the original model and square root of the variable.