Practical No. 07

Aim: Build a linear regression model to predict Miles per Gallon (mpg) of cars using various independent variables such as Cylinder, Weight, Acceleration, Model Year, Origin etc. Use the dataset **auto_mpg.csv.**

S/W Required: Python 3.9, Jupyter Notebook

Theory:

Regression models describe the relationship between variables by fitting a line to the observed data. Linear regression models use a straight line, while logistic and nonlinear regression models use a curved line. Regression allows you to estimate how a dependent variable changes as the independent variable(s) change.

Simple linear regression is used to estimate the relationship between two quantitative variables. You can use simple linear regression when you want to know:

- How strong the relationship is between two variables (e.g. the relationship between rainfall and soil erosion).
- The value of the dependent variable at a certain value of the independent variable (e.g. the amount of soil erosion at a certain level of rainfall).

Assumptions of simple linear regression:

Simple linear regression is a parametric test, meaning that it makes certain assumptions about the data. These assumptions are:

1. Homogeneity of variance (homoscedasticity): the size of the error in our prediction doesn't change significantly across the values of the independent variable.
2. Independence of observations: the observations in the dataset were collected using statistically valid sampling methods, and there are no hidden relationships among observations.
3. Normality: The data follows a normal distribution.

Linear regression makes one additional assumption:

1. The relationship between the independent and dependent variable is linear: the line of best fit through the data points is a straight line (rather than a curve or some sort of grouping factor).

How to perform a simple linear regression:

The formula for a simple linear regression is:

$$y = \beta_0 + \beta_1 X + \epsilon$$

- **y** is the predicted value of the dependent variable (**y**) for any given value of the independent variable (**x**).
- $\boldsymbol{\beta_0}$ is the **intercept**, the predicted value of **y** when the **x** is 0.
- $\boldsymbol{\beta_1}$ is the regression coefficient – how much we expect **y** to change as **x** increases.
- **x** is the independent variable (the variable we expect is influencing **y**).
- **e** is the **error** of the estimate, or how much variation there is in our estimate of the regression coefficient.

Linear regression finds the line of best fit line through your data by searching for the regression coefficient (B1) that minimizes the total error (e) of the model.

While you can perform a linear regression by hand, this is a tedious process, so most people use statistical programs to help them quickly analyze the data.

Code/Program:

Conclusion:

Thus, we have successfully studied how to build a linear regression model for a given dataset.