Experiment No. 1

Aim: To preprocess data and implement Linear Regression and Multiple Linear Regression using NumPy, Pandas, Matplotlib and Scikit-learn in Python.

Platform used: Google Colab

Theory:

Linear Regression

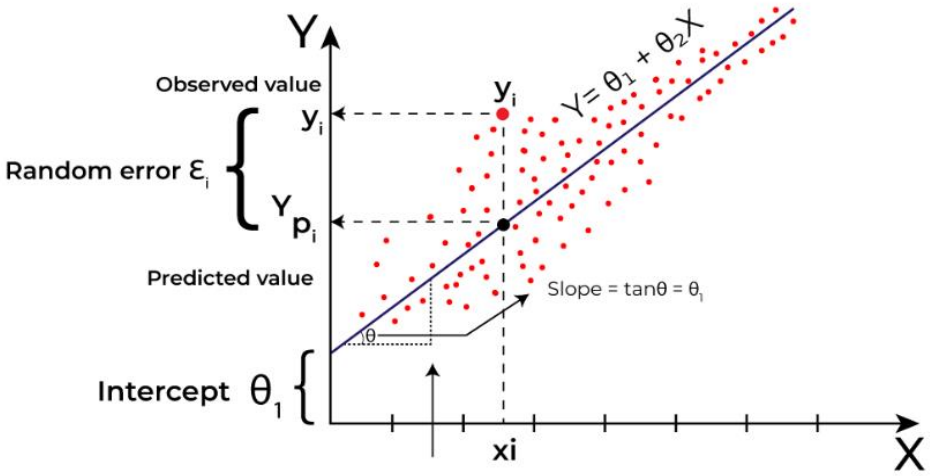
Linear regression is a type of supervised machine-learning algorithm that learns from the labelled datasets and maps the data points with most optimized linear functions which can be used for prediction on new datasets. It assumes that there is a linear relationship between the input and output, meaning the output changes at a constant rate as the input changes. This relationship is represented by a straight line.

For example, we want to predict a student's exam score based on how many hours they studied. We observe that as students study more hours, their scores go up. In the example of predicting exam scores based on hours studied. Here

Independent variable (input): Hours studied because it's the factor we control or observe.

Dependent variable (output): Exam score because it depends on how many hours were studied.

We use the independent variable to predict the dependent variable.



Linear Regression

Here Y is called a dependent or target variable and X is called an independent variable also known as the predictor of Y. There are many types of functions or modules that can be used for regression.

A linear function is the simplest type of function. Here, X may be a single feature or multiple features representing the problem.

Equation of the Best-Fit Line in Linear Regression

For simple linear regression (with one independent variable), the best-fit line is represented by the equation: **y = mx + b**

where:

y is the predicted value (dependent variable)

x is the input (independent variable)

m is the slope of the line (how much y changes when x changes)

b is the intercept (the value of y when x = 0)

The best-fit line will be the one that optimizes the values of m (slope) and b (intercept) so that the predicted y values are as close as possible to the actual data points.

Multiple regression

Multiple linear regression is a statistical method used to model the relationship between a dependent variable and two or more independent variables. It aims to find the best-fitting linear equation that predicts the dependent variable's value based on the independent variables' values. This technique is widely used in various fields for prediction, pattern detection, and understanding the impact of multiple factors on a single outcome.

The Equation:

A multiple linear regression model can be represented by the following equation:

$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \varepsilon$

Where:

Y is the dependent variable.

β_0 is the y-intercept (the value of Y when all independent variables are zero).

$\beta_1, \beta_2, ..., \beta_p$ are the coefficients for each independent variable ($X_1, X_2, ..., X_p$).

$X_1, X_2, ..., X_p$ are the independent variables.

ε is the error term (representing the unexplained variation in Y).

Conclusion: Thus, we successfully performed data preprocessing and implemented Linear Regression and Multiple Linear Regression using NumPy, Pandas, Matplotlib and Scikit-learn in Python.