Experiment 4 – Linear Regression

Aim: Implementation of Linear Regression for Single Variate and Multi-variate

Theory:

Univariate Linear Regression:

Univariate linear regression is a statistical method used to model the relationship between a single independent variable and a dependent variable. The fundamental assumption is that the relationship between the variables can be approximated by a linear equation. In other words, it assumes a straight-line relationship between the independent and dependent variables. The general form of a univariate linear regression equation is:

$y=mx+b$

Here,
y is the dependent variable,
x is the independent variable,
m is the slope of the line (representing the strength and direction of the relationship),
b is the y-intercept (the value of y when x is zero).
The goal of univariate linear regression is to estimate the values of m and b that minimize the sum of squared differences between the observed and predicted values of the dependent variable.

Multivariate Linear Regression:

Multivariate linear regression extends the concept of univariate linear regression to multiple independent variables. It models the relationship between a dependent variable and two or more independent variables. The general form of the multivariate linear regression equation is:

$y = b0 + b1x1 + b2x2 + ... + bnxn$

Here,
y is the dependent variable,
b0 is the y-intercept,
b1, b2, ..., bn are the coefficients associated with the independent variables x1, x2, ..., xn.

The goal is to estimate the values of the coefficients that minimize the difference between the observed and predicted values of the dependent variable. Multivariate linear regression is particularly useful when analyzing complex relationships involving multiple factors. It provides a way to quantify the impact of each independent variable on the dependent variable while accounting for the influence of other variables.

Implementation: For this experiment we were required to perform the following:

Part A:
Program Single variate using inbuilt functions.
Predict for unseen samples
Plot the regression
Part B:
Program Multi variate using inbuilt functions.
Predict for unseen samples

We have used a popular dataset for this experiment called “student_marks.csv”.

Husain Exp4.ipynb

```
[9] import pandas as pd
    import numpy as np
    from sklearn.model_selection import train_test_split
    from sklearn.metrics import mean_absolute_error
    from sklearn.linear_model import LinearRegression
    import matplotlib.pyplot as plt
```

```
[4] df = pd.read_csv("student_marks.csv")
```

```
df
```

	number_courses	time_study	Marks
0	3	4.508	19.202
1	4	0.096	7.734
2	4	3.133	13.811
3	6	7.909	53.018
4	8	7.811	55.299
...	...	...	...
95	6	3.561	19.128
96	3	0.301	5.609
97	4	7.163	41.444
98	7	0.309	12.027
99	3	6.335	32.357

100 rows × 3 columns

```
[6] df.corr()
```

	number_courses	time_study	Marks
number_courses	1.000000	0.204844	0.417335
time_study	0.204844	1.000000	0.942254
Marks	0.417335	0.942254	1.000000

```
[7] X = df.iloc[:,1].values
    y = df.iloc[:,-1].values
```

```
[10] L = LinearRegression()
     X_train, X_test, y_train, y_test=train_test_split(X, y, test_size=0.2, random_state=13)
```

```
X_train = X_train.reshape(-1,1)
X_train.shape
```

```
(80, 1)
```

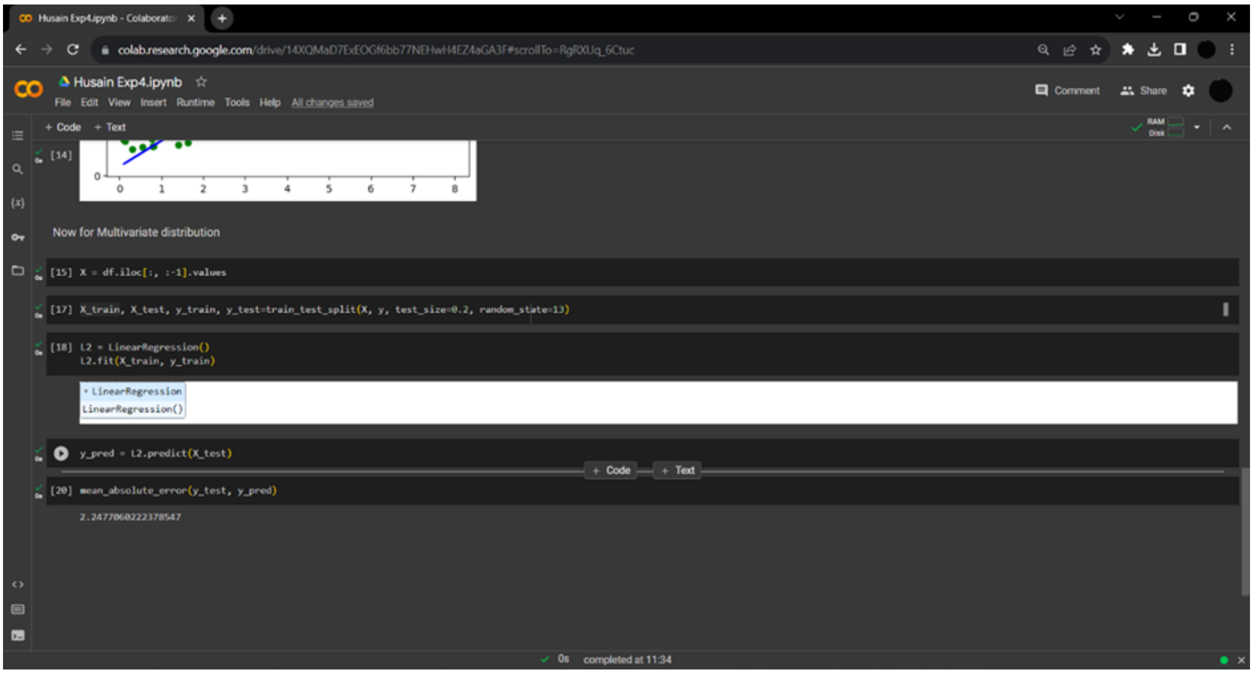
```
[12] L.fit(X_train, y_train)
```

```
▸ LinearRegression
LinearRegression()
```

```
[13] y_preds = L.predict(X_test.reshape(-1,1))
```

```
[14] plt.scatter(X_train, y_train, color = "green")
     plt.plot(X_train, L.predict(X_train), color="blue")
     plt.title("Regressor analysis")
     plt.show()
```

Now for Multivariate distribution



Conclusion: In both univariate and multivariate linear regression, the assumptions of linearity, independence, homoscedasticity (constant variance of errors), and normality of errors are crucial for the validity of the model and the interpretation of its results. These methods are widely used in various fields such as economics, finance, biology, and social sciences for predictive modelling and understanding relationships between variables.