EXP3

Aim: Multiple Linear Regression in Python
Theory:

Multiple Linear Regression (MLR) is a statistical technique that models the relationship between a dependent (response) variable and two or more independent (predictor) variables. The objective is to predict the value of the dependent variable using the values of the independent variables.

Formula for Multiple Linear Regression:

The general form of a multiple linear regression equation is:

$Y=\beta 0+\beta 1X1+\beta 2X2+\ldots+\beta nXn+\epsilon$

Where:

- Y is the dependent (response) variable.
- β0 is the intercept (constant) term.
- β1,β2,…,βn are the coefficients (weights) of the independent variables.
- X1,X2,…,Xn are the independent (predictor) variables.
- ϵ is the error term (residuals), accounting for variability in Y not explained by the independent variables.

Key Concepts:

1. **Intercept (β0)**:
   - Represents the predicted value of the dependent variable when all independent variables are zero.
2. **Coefficients (β1,β2,…,βn)**:
   - Represent the change in the dependent variable for a one-unit change in the corresponding independent variable, assuming all other variables are held constant.
3. **Error Term (ϵ)**:
   - Accounts for the difference between the observed and predicted values, representing the factors not captured by the model.

Steps in Multiple Linear Regression:

1. **Data Preparation**:
   - Collect data for the dependent and independent variables.
   - Ensure that the data is clean (no missing values, outliers, or incorrect data).
2. **Model Fitting**:
   - The regression model is fitted to the data using a method called **Ordinary Least Squares (OLS)**, which minimizes the sum of squared differences between the observed and predicted values.
3. **Model Evaluation**:
   - After fitting the model, we evaluate its performance using metrics such as **R² (R-squared)**, **Mean Squared Error (MSE)**, and **coefficients**.

- **R² (R-squared)**: Measures how well the model fits the data. An R² value closer to 1 indicates a good fit.
- **Mean Squared Error (MSE)**: Measures the average squared difference between the actual and predicted values.

4. **Prediction**:

- Once the model is trained, it can be used to make predictions on new data by plugging the values of the independent variables into the regression equation.

Numerical:

X1	1	2	3	4	5
X2	2	3	4	5	6
Y	4	5	6	7	8

S1: Model Equation

Y=β0+β1X1+β2X2

S2: Calculate the Mean of X1, X2, and Y

X1 = (1+2+3+4+5)/5= 3
X2= (2+3+4+5+6)/5 = 4
X3= (4+5+6+7+8)/5 =6

S3 : Calculate the Covariance and Variance

$$\text{Cov}(X_1, Y) = \frac{1}{n}\sum_{i=1}^{n}(X_{1i} - \overline{X_1})(Y_i - \overline{Y})$$

$$\text{Cov}(X_2, Y) = \frac{1}{n}\sum_{i=1}^{n}(X_{2i} - \overline{X_2})(Y_i - \overline{Y})$$

$$\text{Cov}(X_1, X_2) = \frac{1}{n}\sum_{i=1}^{n}(X_{1i} - \overline{X_1})(X_{2i} - \overline{X_2})$$

$$\text{Var}(X_1) = \frac{1}{n}\sum_{i=1}^{n}(X_{1i} - \overline{X_1})^2$$

$$\text{Var}(X_2) = \frac{1}{n}\sum_{i=1}^{n}(X_{2i} - \overline{X_2})^2$$

Covariance Calculations:

1. $\text{Cov}(X_1, Y)$

$$\text{Cov}(X_1, Y) = \frac{1}{5}[(1-3)(4-6) + (2-3)(5-6) + (3-3)(6-6) + (4-3)(7-6) + (5-3)(8-6)]$$

$$= \frac{1}{5}[(2)(2) + (-1)(-1) + (0)(0) + (1)(1) + (2)(2)] = \frac{1}{5}[4+1+0+1+4] = \frac{10}{5} = 2$$

2. $\text{Cov}(X_2, Y)$

$$\text{Cov}(X_2, Y) = \frac{1}{5}[(2-4)(4-6) + (3-4)(5-6) + (4-4)(6-6) + (5-4)(7-6) + (6-4)(8-6)]$$

$$= \frac{1}{5}[(-2)(-2) + (-1)(-1) + (0)(0) + (1)(1) + (2)(2)] = \frac{1}{5}[4+1+0+1+4] = \frac{10}{5} = 2$$

3. $\mathrm{Cov}(X_1, X_2)$

$$\mathrm{Cov}(X_1, X_2) = \frac{1}{5}[(1-3)(2-4) + (2-3)(3-4) + (3-3)(4-4) + (4-3)(5-4) + (5-3)(6-4)]$$

$$= \frac{1}{5}[(2)(2) + (-1)(-1) + (0)(0) + (1)(1) + (2)(2)] = \frac{1}{5}[4+1+0+1+4] = \frac{10}{5} = 2$$

4. $\mathrm{Var}(X_1)$

$$\mathrm{Var}(X_1) = \frac{1}{5}[(1-3)^2 + (2-3)^2 + (3-3)^2 + (4-3)^2 + (5-3)^2]$$

$$= \frac{1}{5}[4+1+0+1+4] = \frac{10}{5} = 2$$

5. $\mathrm{Var}(X_2)$

$$\mathrm{Var}(X_2) = \frac{1}{5}[(2-4)^2 + (3-4)^2 + (4-4)^2 + (5-4)^2 + (6-4)^2]$$

$$= \frac{1}{5}[4+1+0+1+4] = \frac{10}{5} = 2$$

S4: Calculate the coefficients

$$\beta_1 = \frac{\mathrm{Cov}(X_1, Y)\mathrm{Var}(X_2) - \mathrm{Cov}(X_2, Y)\mathrm{Cov}(X_1, X_2)}{\mathrm{Var}(X_1)\mathrm{Var}(X_2) - \mathrm{Cov}(X_1, X_2)^2}$$

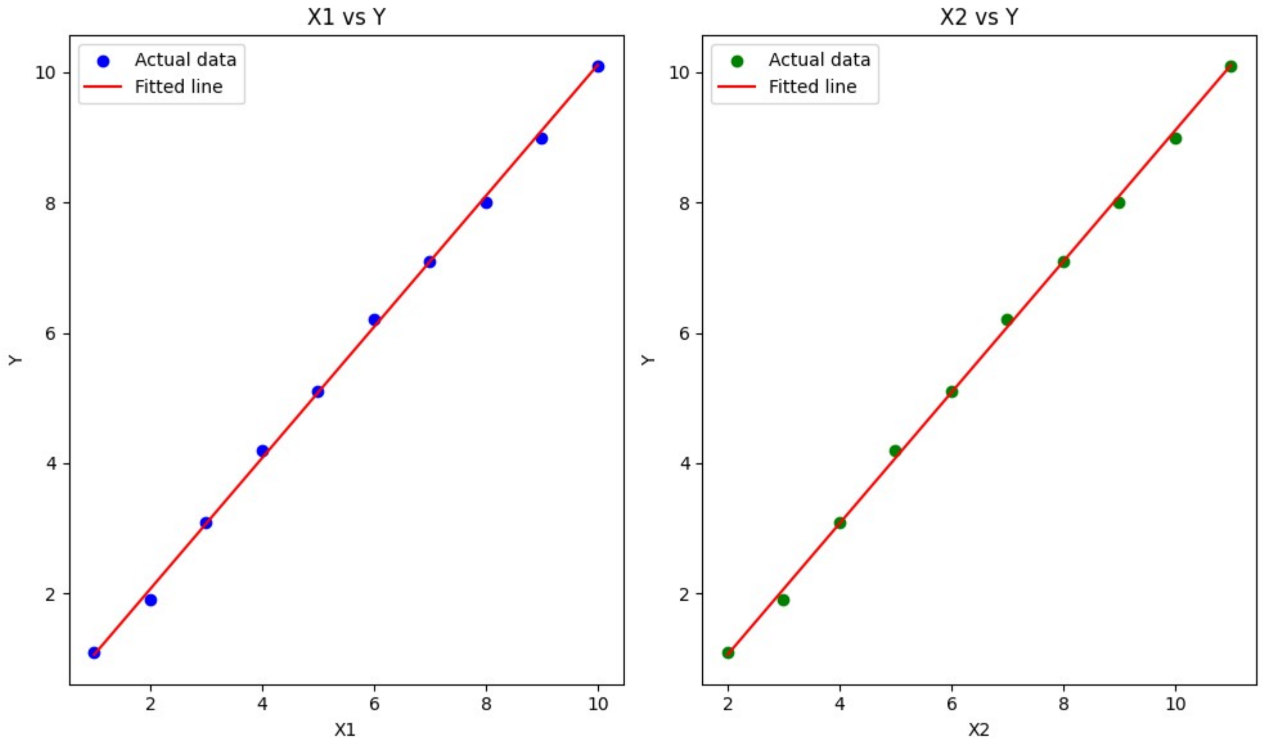
$$\beta_1 = \frac{(2)(2) - (2)(2)}{(2)(2) - (2)(2)} = \frac{4-4}{4-4} = \frac{0}{0}$$

Code:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error, r2_score
data = {
'X1': [1, 2, 3, 4, 5, 6, 7, 8, 9, 10], # Feature 1
'X2': [2, 3, 4, 5, 6, 7, 8, 9, 10, 11], # Feature 2
'Y': [1.1, 1.9, 3.1, 4.2, 5.1, 6.2, 7.1, 8.0, 9.0, 10.1] # Target variable
}
df = pd.DataFrame(data)

X = df[['X1', 'X2']] # Independent variables (Multiple features)
Y = df['Y'] # Dependent variable

X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.3, random_state=0)

model = LinearRegression()

model.fit(X_train, Y_train)
```

```
Y_pred = model.predict(X_test)

print("Coefficients:", model.coef_)
print("Intercept:", model.intercept_)

mse = mean_squared_error(Y_test, Y_pred)
r2 = r2_score(Y_test, Y_pred)

print("Mean Squared Error:", mse)
print("R^2 Score:", r2)

plt.figure(figsize=(10, 6))
plt.subplot(1, 2, 1)
plt.scatter(df['X1'], df['Y'], color='blue', label='Actual data')
plt.plot(df['X1'], model.predict(df[['X1', 'X2']]), color='red', label='Fitted line')
plt.xlabel('X1')
plt.ylabel('Y')
plt.title('X1 vs Y')
plt.legend()

plt.subplot(1, 2, 2)
plt.scatter(df['X2'], df['Y'], color='green', label='Actual data')
plt.plot(df['X2'], model.predict(df[['X1', 'X2']]), color='red', label='Fitted line')
plt.xlabel('X2')
plt.ylabel('Y')
plt.title('X2 vs Y')
plt.legend()

# Show both plots
plt.tight_layout()
plt.show()
```

Output:



Conclusion:

In the practical implementation of multiple linear regression, we learned that the relationship between multiple independent variables and a dependent variable can be modeled using regression coefficients. It is crucial to ensure that the independent variables are not highly correlated (to avoid multicollinearity) as this can affect the accuracy of the model. By interpreting the regression coefficients, we gain insights into how each independent variable influences the dependent variable. However, careful data preprocessing and validation are necessary for the model to provide reliable predictions.