Simple Linear Regression: Step-by-Step Implementation

Step 1: Define the Dataset

The dataset consists of an independent variable (X) and a dependent variable (Y):

$$X = [1, 2, 3, 4, 5, 6, 7, 8, 9, 10]$$

$$Y = [2.3, 4.5, 6.7, 8.1, 10.4, 12.3, 14.5, 16.2, 18.8, 20.1]$$

This data is organized into a Pandas DataFrame for processing.

Step 2: Separate Independent and Dependent Variables

The independent variable (X) is extracted as a 2D array:

$$X = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \\ 8 \\ 9 \\ 10 \end{bmatrix}$$

The dependent variable (Y) is:

$$Y = [2.3, 4.5, 6.7, 8.1, 10.4, 12.3, 14.5, 16.2, 18.8, 20.1]$$

Step 3: Visualize the Data

A scatter plot of X vs Y is created to observe the linear relationship. The Python code:

```
plt.scatter(X, Y, color='blue', label='Data points')
plt.title('Scatter Plot of X vs Y')
plt.xlabel('X (Independent Variable)')
plt.ylabel('Y (Dependent Variable)')
```

The plot shows a clear positive linear relationship.

Step 4: Fit the Linear Regression Model

The Simple Linear Regression model is defined as:

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

Using Python's `sklearn.linear_model.LinearRegression`, the model is fitted to the data.

Step 5: Get Model Parameters

The regression parameters are:

- **Intercept (β_0)**: The predicted value of Y when $X = 0$.
- **Slope (β_1)**: The change in Y for a one-unit increase in X.

The calculated values are:

$$\beta_0 = 0.38, \quad \beta_1 = 2.00$$

The equation of the fitted line is:

$$Y = 0.38 + 2.00X$$

Step 6: Make Predictions

The predicted Y values (Y_{pred}) are computed using the fitted model:

$$Y_{\text{pred}} = \beta_0 + \beta_1 X$$

For each X, substitute the values:

$$Y_{\text{pred}} = [2.38, 4.38, 6.38, 8.38, 10.38, 12.38, 14.38, 16.38, 18.38, 20.38]$$

Step 7: Evaluate the Model

The model is evaluated using:

- **R-squared (R^2)**: Proportion of variance in Y explained by X.

$$R^2 = 0.99$$

- **Mean Squared Error (MSE)**: Average squared difference between actual and predicted values.

$$\text{MSE} = 0.14$$

The high R^2 and low MSE indicate an excellent fit.

Step 8: Visualize the Regression Line

The regression line is plotted along with the data points. The Python code:

```
plt.scatter(X, Y, color='blue', label='Data points')
plt.plot(X, Y_pred, color='red', label='Regression line')
plt.title('Simple Linear Regression')
plt.xlabel('X (Independent Variable)')
plt.ylabel('Y (Dependent Variable)')
```

The plot shows the fitted line passing through the data points, confirming the model's accuracy.

Results

- The linear regression equation is:

$$Y = 0.38 + 2.00X$$

- The model explains 99% of the variance in Y ($R^2 = 0.99$).
- The mean squared error is minimal ($MSE = 0.14$).

Compile Solution

```
Intercept (B0): 0.43
Slope (B1): 1.99
R² (Coefficient of Determination): 1.00
Mean Squared Error (MSE): 0.05
```