Question 1 : What is Simple Linear Regression (SLR)? Explain its purpose.

Simple Linear Regression (SLR) is a statistical method used to model the relationship between one independent variable (X) and one dependent variable (Y) by fitting a straight line to the observed data.

Mathematical form

Y=β0+β1X+εY = \beta_0 + \beta_1 X + \varepsilonY=β0+β1X+ε

Purpose of Simple Linear Regression

1. Understand the relationship
   It helps determine how changes in the independent variable affect the dependent variable.

2. Prediction
   SLR is used to predict the value of the dependent variable for a given value of the independent variable.

3. Quantify the effect
   The slope (β1\beta_1β1) measures how much the dependent variable changes for a one-unit change in the independent variable.

4. Trend analysis
   It identifies and explains linear trends in data.

Question 2: What are the key assumptions of Simple Linear Regression?

Key Assumptions of Simple Linear Regression (SLR)

For Simple Linear Regression to give reliable and valid results, the following assumptions must be satisfied:

1. **Linearity**
   The relationship between the independent variable (X) and the dependent variable (Y) is linear.
   👉 This means the change in Y is proportional to the change in X.

2. **Independence of Errors**
   The residuals (errors) are independent of each other.
   👉 One observation's error should not influence another's (important in time-series data).

3. **Homoscedasticity**
   The variance of residuals is constant across all values of X.
   👉 The spread of errors should be roughly the same at low and high values of X.

4. **Normality of Errors**
   The residuals are normally distributed with a mean of zero.
   👉 This is mainly required for valid hypothesis testing and confidence intervals.

5. **No Perfect Multicollinearity**
   In SLR, this means the independent variable has variation and is not constant.
   👉 (Multicollinearity is more relevant in multiple regression.)

6. **Zero Mean of Errors**
   The expected value of the error term is zero:
   E(ε)=0E(\varepsilon) = 0E(ε)=0
   👉 Ensures the model is unbiased.

Question 3: Write the mathematical equation for a simple linear regression model and explain each term.

Mathematical Equation of Simple Linear Regression (SLR)

Y=β0+β1X+εY = \beta_0 + \beta_1 X + \varepsilonY=β0+β1X+ε

Explanation of Each Term

- **YYY** – *Dependent (response) variable*
  The variable we want to predict or explain (e.g., house price, exam score).

- **XXX** – *Independent (predictor) variable*
  The variable used to explain or predict YYY (e.g., area of a house, hours studied).

- **β0\beta_0β0 (Intercept)**
  The expected value of YYY when X=0X = 0X=0.
  It shows where the regression line cuts the Y-axis.

- **β1\beta_1β1 (Slope coefficient)**
  The change in YYY for a one-unit increase in XXX.
  - If β1>0\beta_1 > 0β1>0: positive relationship
  - If β1<0\beta_1 < 0β1<0: negative relationship

- **ε\varepsilonε (Error term)**
  Represents random variation or unexplained factors affecting YYY that are not captured by XXX.

Question 4: Provide a real-world example where simple linear regression can be applied.

Real-World Example of Simple Linear Regression

Example: Predicting Exam Score Based on Hours Studied

A school wants to understand how **study time affects students' exam performance**.

- **Independent variable (X):** Number of hours studied
- **Dependent variable (Y):** Exam score

The Simple Linear Regression model is:

Exam Score=β0+β1(Hours Studied)+ε\text{Exam Score} = \beta_0 + \beta_1 (\text{Hours Studied}) + \varepsilonExam Score=β0+β1(Hours Studied)+ε

Application

- Data is collected on students' study hours and their exam scores.
- A straight line is fitted to show the relationship between study time and score.
- The model is used to **predict a student's score** based on how many hours they study.

Question 5: What is the method of least squares in linear regression?

Method of Least Squares in Linear Regression

The **method of least squares** is a mathematical technique used to **estimate the best-fitting regression line** in linear regression.

Basic Idea

It chooses the values of the intercept (β0\beta_0β0) and slope (β1\beta_1β1) such that the **sum of the squared differences between the observed values and the predicted values is minimized**.

Mathematical Explanation

For each data point:

- **Observed value:** yiy_iyi
- **Predicted value:** y^i=β0+β1xi\hat{y}_i = \beta_0 + \beta_1 x_iy^i=β0+β1xi
- **Residual (error):** ei=yi−y^ie_i = y_i - \hat{y}_iei=yi−y^i

The least squares method minimizes:

∑i=1n(yi−y^i)2\sum_{i=1}^{n} (y_i - \hat{y}_i)^2i=1∑n(yi−y^i)2

Question 6: What is Logistic Regression? How does it differ from Linear Regression?

Logistic Regression is a statistical and machine learning technique used for **classification problems**, especially when the **dependent variable is binary** (0/1, Yes/No, True/False).

Instead of predicting a continuous value, it predicts the **probability** that an observation belongs to a particular class using the **logistic (sigmoid) function**.

Logistic Regression Model

$P(Y=1 \mid X)=\frac{1}{1+e^{-(\beta_0+\beta_1 X)}}$

Where the output lies between **0 and 1**.

Difference Between Logistic Regression and Linear Regression

Aspect	Linear Regression	Logistic Regression
Purpose	Predict continuous values	Predict categorical outcomes
Dependent Variable	Continuous (e.g., price, score)	Binary (e.g., pass/fail)
Output	Any real number	Probability between 0 and 1
Model Equation	$Y=\beta_0+\beta_1X+\varepsilon$	$P(Y=1)=\frac{1}{1+e^{-z}}$
Function Used	Straight line	Sigmoid (logistic) function
Error Minimization	Least Squares	Maximum Likelihood Estimation
Interpretation	Change in Y per unit X	Change in log-odds per unit X

Question 7: Name and briefly describe three common evaluation metrics for regression models.

1. Mean Absolute Error (MAE)

- **Definition:** The average of the absolute differences between actual and predicted values.

MAE=1n∑i=1n∣yi−y^i∣\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i|MAE=n1i=1∑n∣yi−y^i∣

- **Interpretation:**
  Shows the average error in the same units as the target variable.

- **Advantage:**
  Easy to understand and less sensitive to outliers.

---

2. Mean Squared Error (MSE)

- **Definition:** The average of the squared differences between actual and predicted values.

MSE=1n∑i=1n(yi−y^i)2\text{MSE} = \frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2MSE=n1i=1∑n(yi−y^i)2

- **Interpretation:**
  Penalizes larger errors more heavily.

- **Disadvantage:**
  Units are squared, making interpretation less intuitive.

---

3. R-squared (R2R^2R2) – Coefficient of Determination

- **Definition:** Measures the proportion of variance in the dependent variable explained by the model.

R2=1−SSresSStotR^2 = 1 - \frac{\text{SS}_{res}}{\text{SS}_{tot}}R2=1−SStotSSres

- **Interpretation:**
  Values range from 0 to 1 (closer to 1 means better fit).

- **Advantage:**
  Indicates how well the model explains the data.

Question 8: What is the purpose of the R-squared metric in regression analysis?

Purpose of the R-squared (R2R^2R2) Metric in Regression Analysis

The **R-squared (Coefficient of Determination)** metric is used to **measure how well a regression model explains the variability of the dependent variable**.

Key Purposes of R-squared

1. **Explains Variance**
   R2R^2R2 shows the **proportion of total variation in the dependent variable (Y)** that is explained by the independent variable(s) (X).
   - Example: R2=0.80R^2 = 0.80R2=0.80 means **80% of the variation in Y** is explained by the model.
2. **Measures Goodness of Fit**
   It indicates how well the regression line fits the observed data.
   - Higher R2R^2R2 → better model fit
   - Lower R2R^2R2 → weaker explanatory power
3. **Model Comparison**
   Helps compare regression models predicting the same dependent variable.
   - The model with a higher R2R^2R2 generally explains the data better.

Mathematical Formula

R2=1−SSresSStotR^2 = 1 - \frac{\text{SS}_{res}}{\text{SS}_{tot}}R2=1−SStotSSres

Where:

- SSres\text{SS}_{res}SSres = sum of squared residuals
- SStot\text{SS}_{tot}SStot = total sum of squares

Question 9: Write Python code to fit a simple linear regression model using scikit-learn and print the slope and intercept. (Include your Python code and output in the code box below.)

```
# Import required libraries
import numpy as np
from sklearn.linear_model import LinearRegression

# Sample data (independent variable X and dependent variable y)
X = np.array([1, 2, 3, 4, 5]).reshape(-1, 1)  # Feature
y = np.array([2, 4, 5, 4, 5])                 # Target

# Create and fit the model
model = LinearRegression()
model.fit(X, y)

# Print slope and intercept
print("Slope (Coefficient):", model.coef_[0])
print("Intercept:", model.intercept_)
```

Output

```
Slope (Coefficient): 0.6
Intercept: 2.2
```

Question 10: How do you interpret the coefficients in a simple linear regression model?

Interpretation of Coefficients in Simple Linear Regression

A simple linear regression model is written as:

Y=β0+β1X+εY = \beta_0 + \beta_1 X + \varepsilonY=β0+β1X+ε

Each coefficient has a clear and practical interpretation:

---

1. Intercept (β0\beta_0β0)

- **Meaning:** The expected value of the dependent variable YYY when the independent variable X=0X = 0X=0.
- **Interpretation:** It represents the **baseline level** of YYY.
- **Example:**
  If β0=30\beta_0 = 30β0=30, then when hours studied = 0, the predicted exam score is 30.

⚠️ Note: The intercept is meaningful only if X=0X = 0X=0 is within a realistic range.

---

2. Slope (β1\beta_1β1)

- **Meaning:** The average change in YYY for a **one-unit increase in XXX**.
- **Interpretation:**
  - β1>0\beta_1 > 0β1>0: Positive relationship (Y increases as X increases)
  - β1<0\beta_1 < 0β1<0: Negative relationship (Y decreases as X increases)
- **Example:**
  If β1=5\beta_1 = 5β1=5, then each additional hour of study increases the exam score by **5 marks on average**.