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

Answer:

Simple Linear Regression (SLR) is a statistical and machine learning technique used to understand and model the relationship between **one independent variable (input)** and **one dependent variable (output)**. It assumes that the relationship between these two variables can be represented using a **straight line**.

The main idea behind simple linear regression is to find a line that best fits the given data points in such a way that the difference between the predicted values and the actual values is minimized. This line helps us describe how changes in the independent variable affect the dependent variable.

The **purpose of Simple Linear Regression** includes:

- To **analyze the strength and direction** of the relationship between two variables.
- To **predict future values** of the dependent variable based on new values of the independent variable.
- To **quantify the effect** of the independent variable on the dependent variable.
- To **identify trends and patterns** in data.

Simple Linear Regression is widely used because it is easy to understand, computationally efficient, and forms the foundation for more advanced regression techniques. It is commonly applied in economics, business forecasting, social sciences, and basic data analysis tasks.

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

Answer:

Simple Linear Regression is based on several important assumptions that must be satisfied for the model to produce reliable and accurate results. These assumptions define the conditions under which the regression estimates are valid.

The key assumptions of Simple Linear Regression are:

1. **Linearity**
   There must be a linear relationship between the independent variable and the dependent variable. This means that changes in the independent variable should result in proportional changes in the dependent variable.

2. **Independence of Errors**
   The residuals (errors) should be independent of each other. This implies that the value of one error does not influence another error.
3. **Homoscedasticity**
   The variance of the errors should remain constant across all values of the independent variable. If the variance changes, the model may produce biased predictions.
4. **Normality of Errors**
   The residuals should be approximately normally distributed, especially important for hypothesis testing and confidence interval estimation.
5. **No Outliers with High Influence**
   Extreme values should not excessively influence the regression line, as they can distort the relationship between variables.

When these assumptions are satisfied, Simple Linear Regression provides unbiased, efficient, and interpretable results.

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

Answer:

The mathematical equation of a Simple Linear Regression model is:

$y=\beta0+\beta1x+\epsilon y = \beta_0 + \beta_1 x + \epsilon y=β0+β1x+ϵ$

Explanation of each term:

- **yyy**:
  The dependent variable (target variable) that we are trying to predict.
- **xxx**:
  The independent variable (input feature) used to make predictions.
- **β0\beta_0β0 (Intercept)**:
  The value of yyy when $x=0x = 0x=0$. It represents the starting point of the regression line on the y-axis.
- **β1\beta_1β1 (Slope)**:
  The rate of change of yyy with respect to xxx. It indicates how much yyy changes when xxx increases by one unit.
- **ϵ\epsilonϵ (Error term)**:
  Represents the difference between the actual value and the predicted value. It accounts for randomness and unobserved factors.

This equation forms the foundation of regression analysis and helps in understanding the relationship between variables in a mathematical form.

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

Answer:

A real-world example of Simple Linear Regression is **predicting house prices based on house size**.

- **Independent Variable (x):** House size in square feet
- **Dependent Variable (y):** House price

In this case, Simple Linear Regression can be used to analyze how the price of a house increases as the size of the house increases. By fitting a regression line to historical data, we can predict the price of a new house based on its size.

Other real-world applications include:

- Predicting salary based on years of experience
- Estimating sales based on advertising expenditure
- Forecasting electricity consumption based on temperature

These examples demonstrate how Simple Linear Regression helps in decision-making and future predictions using historical data.

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

Answer:

The **method of least squares** is a mathematical approach used to estimate the parameters (slope and intercept) of a linear regression model. Its main objective is to find the regression line that best fits the data points.

The method works by minimizing the **sum of the squared residuals**, where a residual is the difference between the actual value and the predicted value.

Mathematically, it minimizes:

∑(yi−y^i)2\sum (y_i - \hat{y}_i)^2∑(yi−y^i)2

Where:

- yiy_iyi is the actual value
- y^i\hat{y}_iy^i is the predicted value

Squaring the errors ensures that both positive and negative errors contribute equally and penalizes larger errors more strongly. The least squares method ensures that the fitted line provides the most accurate predictions possible under the linear model assumption.

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

Answer:

Logistic Regression is a supervised learning algorithm used for **classification problems**, especially when the dependent variable is binary (such as Yes/No, 0/1, True/False).

Unlike Linear Regression, which predicts continuous values, Logistic Regression predicts the **probability** that an observation belongs to a particular class using the **sigmoid function**.

Key differences between Logistic Regression and Linear Regression:

Aspect	Linear Regression	Logistic Regression
Output	Continuous values	Probabilities (0–1)
Problem Type	Regression	Classification
Function Used	Linear equation	Sigmoid function
Range of Output	(-∞, +∞)	(0, 1)

Logistic Regression is widely used in medical diagnosis, spam detection, and fraud detection.

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

Answer:

Three commonly used evaluation metrics for regression models are:

1. **Mean Squared Error (MSE)**
   It measures the average of the squared differences between actual and predicted values. Lower MSE indicates better model performance.
2. **Root Mean Squared Error (RMSE)**
   It is the square root of MSE and represents error in the same units as the dependent variable, making it easier to interpret.
3. **Mean Absolute Error (MAE)**
   It calculates the average of absolute differences between actual and predicted values and is less sensitive to outliers.

These metrics help in comparing regression models and selecting the best-performing one.

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

Answer:

R-squared (R^2) is a statistical metric that measures how well the independent variable explains the variation in the dependent variable.

Its value ranges from **0 to 1**:

- **0** means the model explains no variance
- **1** means the model explains all variance

R-squared helps in:

- Evaluating model goodness of fit
- Comparing multiple regression models
- Understanding the explanatory power of the model

A higher R-squared value generally indicates a better-fitting model.

Question 9: Write Python code to fit a simple linear regression model using scikit-learn and print the slope and intercept.

Answer:

```
from sklearn.linear_model import LinearRegression
```

```
import numpy as np

# Sample data
X = np.array([[1], [2], [3], [4], [5]])
y = np.array([2, 4, 6, 8, 10])

# Create model
model = LinearRegression()

# Fit model
model.fit(X, y)

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

Output:

```
Slope: 2.0
Intercept: 0.0
```

This output indicates a perfect linear relationship where y increases by 2 units for every 1 unit increase in x.

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

Answer:

In a simple linear regression model, the coefficients explain the relationship between the independent variable and the dependent variable. These coefficients help us understand how changes in the input variable affect the output variable.

The simple linear regression equation is:

y=β0+β1x+ϵy = \beta_0 + \beta_1 x + \epsilony=β0+β1x+ϵ

Where β0\beta_0β0 is the intercept and β1\beta_1β1 is the slope.

The **intercept (β0\beta_0β0)** represents the value of the dependent variable when the independent variable is zero. It indicates the starting point of the regression line on the y-axis. In some real-world cases, the intercept has a clear meaning, while in others it mainly helps in positioning the regression line correctly.

The **slope (β1\beta_1β1)** shows how much the dependent variable changes when the independent variable increases by one unit. If the slope is positive, it means that an increase in

the independent variable leads to an increase in the dependent variable. If the slope is negative, it indicates an inverse relationship between the two variables.

The magnitude of the slope indicates the strength of the relationship. A larger absolute value of the slope means a stronger effect of the independent variable on the dependent variable.

Overall, interpreting the coefficients helps in understanding the direction, strength, and practical impact of the relationship between variables in a simple linear regression model.