Supervised Learning: Regression Models and Performance Metrics

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

Ans: Simple Linear Regression (SLR) is the simplest form of regression analysis that models the linear relationship between one independent (predictor) variable and one dependent (outcome) variable using a straight line. It estimates how changes in the input (X) relate to changes in the output (Y).

- It's used to **understand the relationship** between two variables.
- It helps **predict the value of the dependent variable** for a given value of the independent variable

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

Ans:

1. **Linearity** – The relationship between the independent and dependent variable should be straight-line (linear).
2. **Independence** – Observations (and errors) must be independent of one another.
3. **Homoscedasticity** – The variance of errors should be constant across all values of the predictor (no "fanning").
4. **Normality of Errors** – The residuals (differences between actual and predicted) should be roughly normally distributed.

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

Ans: **Simple Linear Regression Equation**

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

Term-by-Term Explanation

- Y: Dependent variable (target/outcome you're predicting)
- X: Independent variable (predictor/input).
- β_0: Intercept — expected value of Y when $X = 0$.
- β_1: Slope coefficient — change in Y for a one-unit increase in X
- ε: Error term — captures variation in Y not explained by X.

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

Ans:

Predicting test scores based on hours studied. For instance, you can model the relationship between **hours a student studies (independent variable)** and their **exam score (dependent variable)** to forecast expected scores for different study durations.

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

Ans: The **method of least squares** in linear regression is a mathematical optimization technique used to determine the **best-fit line** by **minimizing the sum of the squared differences** (errors) between the observed data points and the values predicted by the line. In other words, it finds the line where the total of all squared vertical distances between actual and predicted values is as small as possible, ensuring the most accurate linear approximation of the data.

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

Ans: **Logistic Regression** is a supervised machine learning classification algorithm that models the **probability** of a categorical outcome (often binary, like Yes/No) based on input features — it uses a **sigmoid function** to ensure predicted values fall between 0 and 1, which can then be thresholded into classes.

How it differs from Linear Regression:

- **Output Type:** Logistic regression predicts **probabilities** for categories; linear regression predicts a **continuous numeric value.**
- **Use Case:** Logistic is used for **classification** tasks (e.g., spam vs. not spam); linear is used for **regression** tasks (e.g., predicting price).
- **Function Shape:** Logistic uses an **S-shaped curve (sigmoid)** to map outputs to [0,1]; linear uses a **straight line**.

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

Ans:

- **Mean Absolute Error (MAE)** – Measures the average absolute difference between the predicted values and the actual values. Lower MAE means predictions are closer to real outcomes.
- **Mean Squared Error (MSE)** – Calculates the average of the squared differences between predicted and actual values, penalizing larger errors more. Lower MSE indicates better fit.

- **R-squared (R^2)** – Shows the proportion of variance in the target variable explained by the model; values closer to 1 mean a better fit.

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

Ans: In regression analysis, **R-squared** (also called the *coefficient of determination*) is a metric that shows **how much of the variation in the dependent variable is explained by the independent variable(s)** in your model. It ranges from **0 to 1**, where a higher value means a better fit of the model to the data.

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.)

Ans:

CODE:

```
import numpy as np
from sklearn.linear_model import LinearRegression

X = np.array([1, 2, 3, 4, 5]).reshape(-1, 1)
y = np.array([2, 4, 5, 4, 5])

model = LinearRegression()
model.fit(X, y)

print("Slope (Coefficient):", model.coef_[0])
print("Intercept:", model.intercept_)
```

OUTPUT:

```
Slope (Coefficient): 0.7
Intercept: 1.5
```

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

Ans:

- **Intercept (β_0):** This is the *predicted value of the dependent variable (Y)* when the independent variable (X) is **0**. It's the baseline level of Y at X = 0.
- **Slope (β_1):** This tells you how much **Y is expected to change** for *every one-unit increase in X*. If β_1 is positive, Y increases as X increases; if negative, Y decreases as X increases