Assignment Code: D-AG-008

Supervised Learning: Regression Models and Performance Metrics | **Solution**

Instructions: Carefully read each question. Use Google Docs, Microsoft Word, or a similar tool to create a document where you type out each question along with its answer. Save the document as a PDF, and then upload it to the LMS. Please do not zip or archive the files before uploading them. Each question carries 20 marks.

Total Marks: 200

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

Answer:

Definition:
Simple Linear Regression (SLR) is a statistical method used to model the relationship between two variables —

- **one independent variable (X) and**
- **one dependent variable (Y).**

It tries to find a straight line (best fit line) that best describes how the dependent variable changes as the independent variable changes.

Purpose of Simple Linear Regression:

1. **Prediction:**
   To predict the value of one variable (Y) based on the value of another variable (X).
   Example: **Predicting house price (Y) based on area in square feet (X).**

2. **Relationship Understanding:**
   To understand the strength and direction of the relationship between X and Y.
   ***If b > 0*, then as X increases, Y also increases.**
   ***If b < 0*, then as X increases, Y decreases.**

3. **Trend Analysis:**
   To identify trends or patterns between two variables over time or in data samples.

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

Answer:

1. Linearity

- **The relationship between the independent variable (X) and the dependent variable (Y) is linear.**
- **Mathematically:**
  Y=a+bX+e
- 🔹 ***Meaning:* As X increases or decreases, Y changes proportionally along a straight line.**

---

2. Independence of Errors

- **The residuals (errors) should be independent of each other.**
- 🔹 ***Meaning:* The error for one observation should not influence the error for another.**
- **Commonly checked using the Durbin–Watson test (especially in time-series data).**

---

3. Homoscedasticity

- **The variance of the errors should be constant across all values of X.**
- **🔹 *Meaning:* The spread of residuals should be roughly the same for all predicted values.**
- **If not constant, it's called heteroscedasticity, which can distort results.**

---

4. Normality of Errors

- **The residuals (errors) should be normally distributed.**
- **🔹 *Meaning:* When plotted on a histogram, residuals should form a bell-shaped curve.**
- **Important for making valid confidence intervals and hypothesis tests.**

---

5. No Multicollinearity

- **In simple linear regression, there's only one independent variable, so this assumption is automatically satisfied.**
- **(It becomes important in multiple regression.)**

---

6. No Significant Outliers

- **There should be no extreme outliers that can heavily influence the regression**

line.

- **Outliers can distort the slope and intercept.**

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

Answer:

Mathematical Equation of a Simple Linear Regression Model

Y=a+bX+eY = a + bX + eY=a+bX+e

Explanation of Each Term:

Term	Meaning	Description
Y	**Dependent Variable**	**The variable we want to predict or explain (also called the response variable).**
X	**Independent Variable**	**The variable used to predict Y (also called the predictor or explanatory variable).**
a	**Intercept (Constant Term)**	**The value of Y when X = 0. It shows where the regression line crosses the Y-axis.**
b	**Slope (Regression Coefficient)**	**Shows the change in Y for a one-unit change in X. If b > 0, Y increases as X increases; if b < 0, Y decreases as X increases.**
e	**Error Term (Residual)**	**Represents the difference between actual and predicted values of Y. It accounts for all factors that affect Y but are not included in X.**

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

Answer:

Real-World Example of Simple Linear Regression

Example:

Predicting a Student's Exam Score Based on Hours Studied

Scenario:

A teacher wants to know how the number of hours studied (X) affects the exam score (Y) of students.
After collecting data from several students, the teacher applies Simple Linear Regression.

Regression Equation:

Y=a+bX
Suppose the model comes out as:
Exam Score=35+6X

Interpretation:

- **Intercept (a = 35):**
  If a student studies 0 hours, the expected exam score is 35 marks.

- **Slope (b = 6):**
  For every 1 extra hour of study, the exam score increases by 6 marks.

Usefulness:

- **The teacher can predict a student's exam score based on how many hours they study.**
  For example, if a student studies 5 hours:
  $Y=35+6(5)=65$
  The predicted score is 65 marks.

- **It helps understand the relationship between effort (hours studied) and performance (exam score).**

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

Answer:

Method of Least Squares in Linear Regression

The Method of Least Squares is a mathematical technique used to find the best-fitting line in a simple linear regression model.
It works by minimizing the sum of the squares of the errors (residuals) between the actual values and the predicted values.

Mathematical Idea:

In the regression equation:

$Y=a+bX+e$

we want to find the values of a (intercept) and b (slope) such that the sum of squared errors is minimum.

Formula:

We minimize:

$$SSE = \sum (Y_i - \hat{Y_i})^2$$

Where:

- Y_i = actual value
- $\hat{Y_i} = a + bX_i$ = predicted value
- $(Y_i - \hat{Y_i})$ = error (residual)

📊 The Least Squares Estimates:

$$b = \frac{n(\sum XY) - (\sum X)(\sum Y)}{n(\sum X^2) - (\sum X)^2}$$

$$a = \bar{Y} - b\bar{X}$$

Where:

- n = number of observations
- $\bar{X}$ = mean of X values
- $\bar{Y}$ = mean of Y values
-

🎯 Purpose:

- **To find the line of best fit that makes the predicted values as close as possible to the actual values.**
- **Ensures that positive and negative errors do not cancel out (because they're**

squared).

- **Provides the most accurate and unbiased estimates of the slope and intercept.**

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

Answer:

Definition:
👉 **Logistic Regression is a statistical method used to predict a categorical (usually binary) outcome from one or more independent variables.**

It is mainly used when the dependent variable (Y) has two possible outcomes, such as:

- **Yes / No**
- **Pass / Fail**
- **0 / 1**
- **Spam / Not Spam**

🔄 Difference Between Linear and Logistic Regression

Feature	Linear Regression	Logistic Regression
Dependent Variable (Y)	**Continuous (e.g., marks, salary)**	**Categorical (usually binary: 0 or 1)**
Output	**Direct numeric value**	**Probability between 0 and 1**
Equation	**Y=a+bX**	**p=1/(1+e−^(a+bX)**

Curve Type	**Straight line**	**S-shaped (Sigmoid curve)**
Error Measurement	**Mean Squared Error (MSE)**	**Log-Loss (Cross-Entropy)**
Use Case Example	**Predicting house prices**	**Predicting if a student passes or fails**

🎯 Example:

- **Linear Regression: Predicting a student's *score* based on study hours.**
  → Output: 67.5 marks

- **Logistic Regression: Predicting whether a student *passes (1)* or *fails (0)* based on study hours.**
  → Output: Probability = 0.85 → Pass

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

Answer:

Three Common Evaluation Metrics for Regression Models

When we build a regression model (like linear regression), we need to check how well it predicts the target values.
Here are three commonly used metrics to evaluate its performance

---

1 Mean Absolute Error (MAE)

MAE=1/n∑i=1n∣Yi−Yi^∣
Meaning:
It measures the average of the absolute differences between the actual values (YiY_iYi) and predicted values (Yi^\hat{Y_i}Yi^).

Interpretation:

- **Lower MAE → better model accuracy.**
- **It gives an idea of average prediction error in actual units (e.g., marks, prices).**

---

2 Mean Squared Error (MSE)

MSE=1/n∑i=1-> n(Yi−Yi^)
Meaning:
It measures the average of the squared differences between actual and predicted values.

Interpretation:

- **Penalizes larger errors more than smaller ones because errors are squared.**
- **Lower MSE → better performance.**

---

3 R-squared (Coefficient of Determination)

R2=1−SSres/SStot

Where:

- **SSres=∑(Yi−Yi^)2 (residual sum of squares)**
- **SStot=∑(Yi−Y¯)2 (total sum of squares)**

Meaning:
It tells how well the regression line fits the data — i.e., the proportion of variance in Y explained by X.

Interpretation:

- **$R^2=1$: Perfect fit**
- **$R^2=0$: No relationship between X and Y**

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

Answer:

Purpose of the R-squared Metric in Regression Analysis

🔹 Definition:

R-squared (R^2) — also called the Coefficient of Determination — measures how well the independent variable(s) explain the variation in the dependent variable.

It tells us how good the regression model is at fitting the data.

🧮 Formula:

$R^2=1-SS_{res}/SS_{tot}$

Where:

- **$SS_{res}=\sum(Y_i-\hat{Y}_i)^2$ → Residual sum of squares**
- **$SS_{tot}=\sum(Y_i-\bar{Y})^2$ → Total sum of squares**

🎯 Purpose and Interpretation:

- **R^2 shows the proportion of the variance in Y that can be explained by X using the regression model.**
- **It ranges from 0 to 1:**
  - **$R^2 = 1$: Perfect fit — all data points lie exactly on the regression line.**
  - **$R^2 = 0$: The model doesn't explain any variation in Y (completely useless).**
  - **$R^2 = 0.8$: 80% of the variation in Y is explained by X; only 20% is unexplained.**

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.)

Answer:

```python
# Answer:

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

# Example data
# X = independent variable (hours studied)
# Y = dependent variable (exam scores)
X = np.array([1, 2, 3, 4, 5]).reshape(-1, 1)    # X should be 2D for sklearn
Y = np.array([40, 50, 60, 65, 80])
```

```
# Create a Linear Regression model
model = LinearRegression()

# Fit the model to the data
model.fit(X, Y)

# Get the slope (coefficient) and intercept
slope = model.coef_[0]
intercept = model.intercept_

# Print results
print("Slope (b):", slope)
print("Intercept (a):", intercept)

#output:-
Slope (b): 9.500000000000002
Intercept (a): 30.499999999999993
```

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

Answer:

Interpreting the Coefficients in a Simple Linear Regression Model

The simple linear regression equation is:

Y=a+bX

Where:

- **Y → Dependent variable (the one you want to predict)**

- X → Independent variable (the predictor)
- a → Intercept (constant term)
- b → Slope (regression coefficient)

🔹 1️⃣ Intercept (a):

- It represents the predicted value of Y when X = 0.
- In other words, it's where the regression line crosses the Y-axis.

Example:
If the equation is

Exam Score=35+6X

→ When a student studies 0 hours (X = 0), the predicted score is 35 marks.

🔹 2️⃣ Slope (b):

- It represents the change in Y for a one-unit increase in X.
- Shows the strength and direction of the relationship between X and Y:
  - b > 0 → Positive relationship (as X increases, Y increases).
  - b < 0 → Negative relationship (as X increases, Y decreases).

Example:
In the same equation

Exam Score=35+6X

→ For every 1 extra hour studied, the exam score increases by 6 marks.