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:

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

- **One independent variable (X) – the predictor or input**
- **One dependent variable (Y) – the outcome or response**

It assumes that the relationship between these two variables can be represented by a straight line.

Purpose of Simple Linear Regression

The main purposes of SLR are:

1. **To understand the relationship**
   It helps determine whether a linear relationship exists between X and Y.

2. **To measure the strength and direction**
   - **If $\beta_1 > 0 \rightarrow$ positive relationship**
   - **If $\beta_1 < 0 \rightarrow$ negative relationship**
3. **To make predictions**
   Once the line is estimated, we can predict Y for any given value of X.
4. **To quantify impact**

   It tells us how much the dependent variable changes when the independent variable changes by one unit.

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

Answer:

Simple Linear Regression (SLR) relies on several important assumptions. If these assumptions are violated, the model's estimates or predictions may become unreliable.

Here are the key assumptions:

1 Linearity

There must be a linear relationship between the independent variable (X) and the dependent variable (Y).

- **The relationship should form a straight-line pattern.**
- **If the relationship is curved, SLR is not appropriate.**

---

2 Independence of Errors

The residuals (errors) must be independent of each other.

3 Homoscedasticity (Constant Variance)

The variance of the errors should be constant across all values of X.

4 Normality of Errors

The residuals should be normally distributed.

- **This assumption is mainly important for hypothesis testing and confidence intervals**

5 No Perfect Multicollinearity (Automatically Satisfied in SLR)

Since Simple Linear Regression has only one independent variable, multicollinearity is not an issue (it becomes relevant in multiple regression).

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

Answer:

The mathematical equation for a Simple Linear Regression (SLR) model is:

$Y=\beta 0+\beta 1X+\varepsilon=\beta 0+\beta 1X+\varepsilon$

Explanation of Each Term

1️⃣ Y — Dependent Variable

- **Also called the response variable.**
- **It is the outcome we want to predict or explain.**

2️⃣ X— Independent Variable

- **Also called the predictor or explanatory variable.**
- **It is the variable used to explain changes in Y.**

3️⃣ β0 — Intercept

- **The value of Y when X=0**
- **It represents the point where the regression line crosses the Y-axis.**
- **Sometimes it may not have a practical meaning (if X=0 is unrealistic).**

4️⃣ β1 — Slope (Regression Coefficient)

- **Measures the change in Y for a one-unit increase in X.**
- **If:**
    - **β1>0: Positive relationship**
    - **β1<0: Negative relationship**
- **It indicates the strength and direction of the relationship.**

5️⃣ ε— Error Term

- **Represents the difference between the observed value and the predicted value.**

- **Captures other factors affecting YYY that are not included in the model.**
- **Assumed to have:**
    - **Mean = 0**
    - **Constant variance**
    - **Normal distribution (for inference)**

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

Answer:

A classic real-world example of simple linear regression is predicting house prices based on house size.

Example: House Price vs. Square Footage

A real estate analyst wants to understand how the size of a house (in square feet) affects its selling price.

- **Independent variable (X): Square footage**
- **Dependent variable (Y): House price**

By collecting data from recently sold homes, the analyst might find that as square footage increases, the price tends to increase in a roughly linear way.

The simple linear regression model would look like:

Price=β0+β1×(Square Footage)

- **β_0 (intercept): Estimated price of a house with 0 square feet (theoretical starting** point)

- β_1 (slope): Average increase in price per additional square foot

Practical Use

If the model estimates:

Price=50,000+150×(Square Footage)

Then:

- **A 2,000 sq ft house would be predicted to cost 50,000+150×2000**
- **=350,000**

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

Answer:

The method of least squares is the technique used in linear regression to find the "best-fitting" straight line through a set of data points.

🔹 The Main Idea

In simple linear regression, we model the relationship between:

y=β0+β1x

But real data points don't fall perfectly on a straight line. So we choose the line that minimizes the total error between:

- **The actual values y**
- **The predicted values y^**

“Least Squares” Mean?

For each data point, we calculate the residual (error):

Residual=yi−y^

Instead of minimizing the raw errors (which could cancel out), we:

1. **Square each residual**
2. **Add them up**
3. **Choose the line that makes this total as small as possible****

This total is called the Sum of Squared Errors (SSE):

SSE=∑(yi−y^i)2sq

The regression line is chosen so that this value is as small as possible.

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

Answer:

Logistic Regression is a statistical method used for classification problems, where the outcome variable is categorical, usually binary (e.g., Yes/No, 0/1, Pass/Fail).

Instead of predicting a continuous value, logistic regression predicts the probability that an observation belongs to a particular class.

- ### **Model Form**

It uses the logistic (sigmoid) function to transform a linear combination of inputs into a probability:

p=1/1+e−(β0+β1x)p

Where:

- **p = probability of the event occurring (between 0 and 1)**
- **β0+β1x = linear combination of predictors**

The output is always between 0 and 1, making it suitable for probability estimation.

Feature	Linear Regression	Logistic Regression
Type of problem	Regression	Classification
Output variable	Continuous (e.g., price, height)	Categorical (e.g., 0/1)
Output range	Any real number (-∞ to +∞)	Between 0 and 1
Model equation	y=β0+β1x	p=11+e−zp
Error minimization	Minimizes Sum of Squared Errors (Least Squares)	Maximizes Likelihood (Maximum Likelihood Estimation)
Graph shape	Straight line	S-shaped (sigmoid curve)

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

Answer:

The three common evaluation metrics used for regression models:

1 Mean Absolute Error (MAE)

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

MAE=1/n∑∣yi−y^i∣
It tells :

- **On average, how far predictions are from actual values.**
- **Errors are measured in the same units as the target variable.**

2 Mean Squared Error (MSE)

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

MSE=1/n∑(yi−y^i)2

It tells :

- **Penalizes larger errors more heavily because of squaring.**

3 R-squared (Coefficient of Determination)

Definition:
Measures the proportion of variance in the dependent variable that is explained by the model.

R2=1−Residual Sum of Squares/Total Sum of Squares

It tells :

- **Value ranges from 0 to 1.**
- **Higher values indicate better model fit.**

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

Answer:

Purpose of the R-squared Metric in Regression Analysis

R-squared (R^2), also called the coefficient of determination, measures how well a regression model explains the variability of the dependent variable.

- Main Purpose is:

R^2 tells us:

> **What proportion of the total variation in the dependent variable is explained by the independent variable(s).**

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:

A simple example of Python code to fit a Simple Linear Regression model using scikit-learn and print the slope and intercept.

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

# Sample dataset (X must be 2D for scikit-learn)
# Example: Hours studied vs Exam score
X = np.array([1, 2, 3, 4, 5]).reshape(-1, 1)  # Independent variable
y = np.array([50, 55, 65, 70, 75])            # Dependent variable

# Create the model
model = LinearRegression()

# Fit the model
model.fit(X, y)

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

# Print results
print("Slope (Coefficient):", slope)
print("Intercept:", intercept)
```

Output:

```
Slope (Coefficient): 6.500000000000001
Intercept: 43.5
```

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

Answer:

In a simple linear regression model, the equation is:

y=β0+β1xy

There are two coefficients to interpret:

1️⃣ Intercept (β_0)

🔹 **Meaning:**

The intercept represents the expected value of y when x = 0.

🔹 **Interpretation:**

It is the point where the regression line crosses the y-axis.

🔹 **Example:**

If the model is:

Score=40+5(Hours Studied)
The intercept (40) means:

> A student who studies 0 hours is predicted to score 40.

2️⃣ Slope (β_1)

🔹 **Meaning:**

The slope represents the **average change in y for a one-unit increase in x**.

🔹 **Interpretation:**

- If $\beta_1 > 0$ → Positive relationship (y increases as x increases)
- If $\beta_1 < 0$ → Negative relationship (y decreases as x increases)
- If $\beta_1 = 0$ → No linear relationship

🔹 **Example:**

If the slope is 5:

> For each additional hour studied, the score increases by 5 points on average

- **Intercept (β_0)** → Starting value
- **Slope (β_1)** → Rate of change