Assignment Code: D-AG-008

Supervised Learning: Regression Models and Performance Metrics

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 data.

The basic form of the SLR equation is:

$Y=a+bX$Y=a+bX

Where:

- **Y** = dependent variable (what you want to predict)
- **X** = independent variable (the predictor)
- **a** = intercept (value of Y when X = 0)
- **b** = slope (how much Y changes for a one-unit change in X)

Purpose of Simple Linear Regression

1. **To understand relationships**
   It helps determine whether and how strongly one variable affects another (for example, how study hours affect exam scores).
2. **To make predictions**
   Once the relationship is established, SLR can be used to predict future or unknown values of Y based on given values of X.
3. **To quantify impact**
   The slope tells us the direction and magnitude of change—whether Y increases or decreases as X changes, and by how much.
4. **To simplify data analysis**
   It provides a simple, easy-to-interpret model to summarize trends in data.

In short, **SLR helps explain, predict, and analyze the relationship between two variables using a straight line**—simple, effective, and very exam-friendly 😄📈

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

Simple Linear Regression (SLR) works properly only when certain **key assumptions** are satisfied. These assumptions ensure that the model is reliable and the results are valid.

Key Assumptions of Simple Linear Regression

1. **Linearity**
   The relationship between the independent variable (X) and the dependent variable (Y) is linear.
   👉 Changes in X cause proportional changes in Y.
2. **Independence of Errors**
   The residuals (errors) are independent of each other.
   👉 One observation's error should not influence another's.
3. **Homoscedasticity**
   The variance of the errors is constant across all values of X.
   👉 The spread of residuals should remain roughly the same (no funnel shapes!).
4. **Normality of Errors**
   The residuals are normally distributed.
   👉 Especially important for hypothesis testing and confidence intervals.
5. **No Perfect Multicollinearity**
   Since SLR has only **one independent variable**, this assumption is automatically satisfied.
   👉 (More relevant in multiple regression, but often mentioned for completeness.)

In simple words 🧠

- The relationship should be straight-line
- Errors should be random and independent
- Error spread should be constant
- Errors should be normally distributed

If these assumptions hold, **SLR becomes a powerful and trustworthy tool**—if not, the model starts lying (politely, but still lying 😉).

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

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

$Y=\beta 0+\beta 1X+\varepsilon$Y=β0+β1X+ε

Explanation of Each Term

- Y**Y** — *Dependent variable*
  This is the outcome or response variable we want to explain or predict.
- X**X** — *Independent variable*
  This is the predictor or explanatory variable used to explain changes in YY.
- $\beta 0$**β0** — *Intercept*
  It represents the expected value of YY when $X=0$X=0.
  In practical terms, it is where the regression line cuts the Y-axis.
- $\beta 1$**β1** — *Slope (Regression coefficient)*
  It measures the change in YY for a one-unit increase in XX.
  - If $\beta 1>0$β1>0: positive relationship
  - If $\beta 1<0$β1<0: negative relationship
- ε**ε** — *Error term (Random disturbance)*
  This captures all other factors affecting YY that are not included in the model, such as randomness, measurement errors, or omitted variables.

In short 📌

The equation says:

> **Y is explained by a constant part ($\beta 0$β0), a linear effect of X ($\beta 1X$β1X), and some unavoidable noise (εε).**

Clean, simple, and exam-ready ✔️

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

A clear **real-world example** of applying **Simple Linear Regression (SLR)** is:

📚 Example: Study Hours vs Exam Scores

- **Independent variable (X):** Number of hours a student studies
- **Dependent variable (Y):** Exam score obtained

Using simple linear regression, we can model the relationship as:

$\text{Exam Score}=\beta 0+\beta 1(\text{Study Hours})+\varepsilon$Exam Score=β0+β1(Study Hours)+ε

How SLR Helps in This Case

1. **Understand the relationship**
   It shows whether studying more hours leads to higher exam scores.
2. **Quantify the effect**
   The slope ($\beta 1$β1) tells how much the exam score increases for each additional hour of study.
3. **Make predictions**
   Teachers or students can estimate expected exam scores based on study time.

Other real-world examples (just in case 😄)

- Advertising cost → Sales revenue
- Temperature → Electricity consumption
- Years of experience → Salary

So basically, **whenever one variable influences another in a roughly straight-line way**, simple linear regression is your go-to tool 📈✔️

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

The **method of least squares** is a standard technique used in **linear regression** to estimate the best-fitting regression line for a given set of data.

What It Does

The method chooses the regression line such that the **sum of the squared differences** between the **actual values** and the **predicted values** is **as small as possible**.

These differences are called **residuals**.

Mathematical Idea

For a simple linear regression model:

$Y=\beta0+\beta1X$Y=β0+β1X

The least squares method minimizes the objective function:

$\sum(Yi-Y\hat{}i)2$∑(Yi−Y^i)2

Where:

- YiYi = observed (actual) values
- $Y\hat{}i$Y^i = predicted values from the regression line
- $Yi-Y\hat{}i$Yi−Y^i = residual (error)

Why "Squared" Errors?

- Squaring avoids positive and negative errors canceling out
- It penalizes larger errors more heavily
- It leads to a unique, mathematically convenient solution

Purpose in Simple Words 🧠

The method of least squares **finds the line that comes closest to all data points overall**, making prediction errors as small as possible.

Think of it as telling the regression line:

> *"Miss the points if you must, but miss them as little as possible!"* 😄📉

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

What is Logistic Regression?

Logistic Regression is a statistical and machine-learning method used when the **dependent variable is categorical**, most commonly **binary** (e.g., Yes/No, Pass/Fail, 0/1).

Instead of predicting a raw numeric value, logistic regression **predicts the probability** that an outcome belongs to a particular class.

The model looks like this:

$P(Y=1)=11+e-(\beta0+\beta1X)$P(Y=1)=1+e−(β0+β1X)1

This equation uses a **logistic (sigmoid) function** to squeeze predictions between **0 and 1**, which makes sense for probabilities.

How Logistic Regression Differs from Linear Regression

Aspect	Linear Regression	Logistic Regression
Type of output	Continuous values	Probabilities (0 to 1)
Dependent variable	Numeric	Categorical (usually binary)
Model form	Straight line	S-shaped (sigmoid curve)
Prediction range	−∞−∞ to +∞+∞	Between 0 and 1
Use case	Predict quantities	Predict classes
Error method	Least Squares	Maximum Likelihood

Simple Example

- **Linear Regression:**
  Predicting **house price** based on size
- **Logistic Regression:**
  Predicting whether an email is **spam or not spam**

In plain words 🧠

- **Linear Regression answers:** *“How much?”*
- **Logistic Regression answers:** *“Which one?” or “What’s the chance?”*

Different goals, different tools—using linear regression for classification is like using a ruler to measure emotions… ambitious, but not ideal 😄📏✔️

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

Three **common evaluation metrics for regression models** are used to measure how well the predicted values match the actual values:

---

1. Mean Absolute Error (MAE)

MAE measures the **average absolute difference** between actual and predicted values.

MAE=1𝑛∑∣𝑌𝑖−𝑌^𝑖∣MAE=n1∑∣Yi−Y^i∣

Why it's useful:

- Easy to understand
- Treats all errors equally
- Lower MAE = better model

---

2. Mean Squared Error (MSE)

MSE calculates the **average of squared errors** between actual and predicted values.

MSE=1𝑛∑(𝑌𝑖−𝑌^𝑖)2MSE=n1∑(Yi−Y^i)2

Why it's useful:

- Penalizes large errors more heavily
- Commonly used in optimization
- Lower MSE = better model

---

3. Root Mean Squared Error (RMSE)

RMSE is the **square root of MSE**, bringing the error back to the original unit of Y.

RMSE=MSERMSE=MSE

Why it's useful:

- Easy to interpret
- Sensitive to large errors
- Widely reported in practice

In short 📊

- **MAE:** average size of errors
- **MSE:** average squared errors (harsh on big mistakes 😬)
- **RMSE:** standard deviation of prediction errors

Together, these metrics give a clear picture of a regression model's performance ✔️

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

The **purpose of the R-squared (R2R2) metric** in regression analysis is to measure **how well the regression model explains the variability in the dependent variable**.

What R-squared Tells Us

R2R2 represents the **proportion of variance in the dependent variable (Y)** that is explained by the independent variable(s) in the model.

R2=Explained VariationTotal VariationR2=Total VariationExplained Variation

Its value ranges from **0 to 1** (or **0% to 100%**).

How to Interpret R-squared

- **R2=0R2=0** → The model explains none of the variability
- **R2=1R2=1** → The model explains all the variability
- **Higher R2R2** → Better fit to the data

Example:
If R2=0.80R2=0.80, then **80% of the variation in Y is explained by the model**, while the remaining 20% is due to other factors or randomness.

Why R-squared Is Useful

- Evaluates **goodness of fit**
- Helps compare different regression models
- Indicates how meaningful the predictors are

Important Note ⚠️

A high R2R2 **does not guarantee** a good or correct model—it doesn't imply causation and can be misleading if assumptions are violated or irrelevant variables are included.

In simple words 🧠

R-squared answers:

> *“How much of what’s happening can my model explain?”*

Helpful, but don’t worship it 😄📈

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.)

Here’s a **simple, clean Python example** using **scikit-learn** to fit a **Simple Linear Regression** model and print the **slope** and **intercept**.

I’m using a small sample dataset so the output is clear and reproducible.

```
# 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)
y = np.array([2, 4, 5, 4, 5])

# 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
```

Explanation (quick one 😄)

- **Slope (0.6):** For every 1-unit increase in X, Y increases by 0.6
- **Intercept (2.2):** Predicted value of Y when X = 0

That’s a full, exam-ready example—code + output + meaning ✔️📈

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

In a **Simple Linear Regression (SLR)** model, the coefficients explain **how the dependent variable changes in response to the independent variable**.

The model is:

$Y=\beta 0+\beta 1X$Y=β0+β1X

Interpretation of the Coefficients

1. Intercept ($\beta 0$β0)

- It is the **expected value of Y when X = 0**.
- Represents the **baseline level** of the dependent variable.
- Sometimes it has no practical meaning (if X = 0 is outside the data range), but it is still mathematically important.

Example:
If $\beta 0=5$β0=5, then when $X=0$X=0, the predicted value of YY is 5.

---

2. Slope ($\beta 1$β1)

- It represents the **average change in Y for a one-unit increase in X**.
- Shows the **direction** and **strength** of the relationship:
  - $\beta 1>0$β1>0: Positive relationship
  - $\beta 1<0$β1<0: Negative relationship

Example:
If $\beta 1=2$β1=2, then for every 1-unit increase in X, Y increases by 2 units.

In simple words 🧠

- **Intercept:** where the line starts
- **Slope:** how steep the line is and which direction it goes

Together, they tell the story of how X influences Y—no drama, just math 😄📈✔️