Assignment: Supervised Learning - Regression Models and Performance Metrics

Assignment Code: D-AG-008
Total Marks: 200
Instructions: Each question carries 20 marks

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

Answer:

Simple Linear Regression (SLR) is a fundamental statistical and machine learning technique used to model the linear relationship between two continuous variables: one independent variable (predictor) and one dependent variable (response). The method fits a straight line through a set of data points to represent this relationship.

Purpose of Simple Linear Regression:

- To understand the relationship between two variables
- To predict the value of the dependent variable based on the independent variable
- To quantify the strength and direction of the relationship
- To establish a baseline model for more complex regression techniques
- To identify trends in data and make informed predictions

Simple Linear Regression is widely used in economics, finance, healthcare, and engineering for forecasting and trend analysis.

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

Answer:

Simple Linear Regression relies on several critical assumptions for validity and accuracy:

1. **Linearity:** The relationship between independent and dependent variables must be linear. The data should follow a straight-line pattern.
2. **Independence:** Observations must be independent of each other. There should be no autocorrelation between residuals, particularly important in time-series data.
3. **Homoscedasticity:** The variance of the residuals (errors) should be constant across all levels of the independent variable. This means the spread of data points around the regression line should be uniform.

4. **Normality:** The residuals should be approximately normally distributed. This ensures that confidence intervals and hypothesis tests are valid.
5. **No Multicollinearity:** In cases with multiple predictors, there should be no high correlation between independent variables (though this technically applies to multiple regression).
6. **No Measurement Error:** It is assumed that the independent variable is measured without error.

Violations of these assumptions can lead to biased predictions and inaccurate statistical inferences.

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

Answer:

The mathematical equation for Simple Linear Regression is:

$$y = \beta_0 + \beta_1 x + \epsilon$$

Explanation of each term:

- y (Dependent Variable): The response variable or target variable that we want to predict.
- β_0 (Intercept): The constant term, representing the value of y when $x = 0$. It is the y-intercept where the regression line crosses the y-axis.
- β_1 (Slope or Regression Coefficient): The coefficient that indicates the change in y for each unit change in x. A positive slope indicates a positive relationship, while a negative slope indicates a negative relationship.
- x (Independent Variable): The predictor variable used to explain or predict changes in the dependent variable.
- ϵ (Error Term or Residual): The random error component representing the difference between the observed value and the predicted value. It captures the variability in y not explained by x.

Example: If $y = 50 + 3x + \epsilon$, then $\beta_0 = 50$ and $\beta_1 = 3$, meaning for every one-unit increase in x, y increases by 3 units on average.

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

Answer:

Real-World Example: Predicting House Prices Based on Square Footage

One of the most practical applications of Simple Linear Regression is in real estate pricing:

- **Independent Variable (x):** Square footage of the house (in square feet)
- **Dependent Variable (y):** Price of the house (in dollars)

Application:

A real estate agent wants to predict house prices based on their size. By collecting historical data on house prices and their corresponding square footage, Simple Linear Regression can establish the relationship between these variables.

Example Data:

- A 1,000 sq ft house sells for $200,000
- A 1,500 sq ft house sells for $250,000
- A 2,000 sq ft house sells for $300,000

Using SLR, the model might produce:

$$\text{Price} = 100,000 + 100 \times \text{Square Footage}$$

This means each additional square foot adds approximately $100 to the house price.

Benefits:

- Quickly estimate prices for new listings
- Identify underpriced or overpriced properties
- Support investment decisions
- Provide clients with valuation insights

Other Real-World Examples:

- Predicting salary based on years of experience
- Estimating fuel consumption based on car weight
- Forecasting sales revenue based on marketing spend
- Predicting student test scores based on study hours

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

Answer:

The method of least squares is the fundamental technique used to estimate the coefficients (β_0 and β_1) in Simple Linear Regression. It is an optimization method that minimizes the sum of squared errors (residuals) between the observed and predicted values.

Core Principle:

The least squares method finds the best-fitting line by minimizing the sum of squared vertical distances between the actual data points and the regression line. Mathematically, it minimizes:

$$\text{SSE (Sum of Squared Errors)} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - (\beta_0 + \beta_1 x_i))^2$$

Where:

- y_i = observed value
- $\hat{y}_i$ = predicted value
- n = number of observations

Estimation Formulas:

The least squares estimates for the coefficients are:

$$\beta_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

Where:

- $\bar{x}$ = mean of independent variable
- $\bar{y}$ = mean of dependent variable

Why Least Squares?

1. **Mathematical elegance:** Produces closed-form solutions
2. **Statistical optimality:** Provides unbiased estimates under certain conditions
3. **Computational efficiency:** Relatively simple to calculate
4. **Interpretability:** Results are easy to understand and explain

Advantages:

- Ensures unique solution (if assumptions are met)
- Minimizes prediction error
- Theoretical foundation for statistical inference

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

Answer:

Logistic Regression:

Logistic Regression is a supervised learning algorithm used for binary classification problems. Despite its name, it is fundamentally different from Linear Regression. It predicts the probability that an instance belongs to a particular class (typically coded as 0 or 1), rather than predicting a continuous value.

Mathematical Form:

$$P(y = 1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}} = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$$

This is known as the logistic function or sigmoid function, which outputs values between 0 and 1.

Key Differences Between Linear and Logistic Regression:

Feature	Linear Regression	Logistic Regression
Purpose	Predicts continuous values	Predicts class probabilities (0 or 1)
Problem Type	Regression problem	Classification problem
Output Range	Any real number (-∞ to +∞)	Probability (0 to 1)
Method	Least squares method	Maximum likelihood estimation
Function	Linear equation: $y = \beta_0 + \beta_1 x$	Sigmoid/logistic function
Assumptions	Linearity, homoscedasticity, normality	Independence of observations, linearity in log-odds
Use Cases	House price prediction, temperature forecasting	Disease diagnosis, email spam detection, loan approval
Error Distribution	Normally distributed residuals	Binomial distribution
Decision Boundary	Continuous line	Probabilistic threshold (usually 0.5)

When to Use Each:

- **Linear Regression:** When the target variable is continuous (e.g., predicting weight, age, temperature)
- **Logistic Regression:** When the target variable is categorical with two classes (e.g., yes/no, spam/not spam, fraud/legitimate)

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

Answer:

1. Mean Absolute Error (MAE)

MAE measures the average absolute difference between actual and predicted values:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

- **Range:** 0 to ∞ (lower is better)
- **Interpretation:** Average prediction error in the same units as the dependent variable
- **Advantage:** Easy to interpret and not overly sensitive to outliers
- **Disadvantage:** Treats all errors equally regardless of magnitude

2. Mean Squared Error (MSE)

MSE measures the average of the squared differences between actual and predicted values:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

- **Range:** 0 to ∞ (lower is better)
- **Interpretation:** Average squared prediction error
- **Advantage:** Penalizes larger errors more heavily; differentiable for optimization
- **Disadvantage:** Sensitive to outliers; units are squared, making interpretation harder

3. R-squared (Coefficient of Determination)

R-squared indicates the proportion of variance in the dependent variable explained by the independent variable:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{\text{SSE}}{\text{SST}}$$

- **Range:** 0 to 1 (higher is better, though can be negative for poor models)
- **Interpretation:** Percentage of variance explained by the model
- **Advantage:** Unitless and easy to interpret; represents model explanatory power
- **Disadvantage:** Doesn't indicate if the model is good; always increases with more variables

Comparison Table:

Metric	MAE	MSE	R-squared
Units	Same as y	Squared units of y	Unitless
Sensitivity to Outliers	Low	High	Medium
Interpretability	Very intuitive	Less intuitive	Very intuitive
Best Used When	Outliers present	Outliers should penalize heavily	Comparing model fit

---

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

Answer:

Purpose of R-squared:

R-squared (also called the coefficient of determination) serves multiple important purposes in regression analysis:

1. Measuring Model Fit Quality

R-squared quantifies how well the regression model explains the variation in the dependent variable. It represents the proportion of variance in y that is predictable from the independent variable(s).

- **R^2 = 0.95:** The model explains 95% of the variance in the data
- **R^2 = 0.50:** The model explains 50% of the variance
- **R^2 = 0.20:** The model explains only 20% of the variance

2. Comparison and Model Selection

R-squared allows researchers to compare different regression models:

- Higher R^2 generally indicates a better fit
- Helps identify which independent variable(s) are more predictive
- Supports decisions about including or excluding variables

3. Assessing Predictive Power

R-squared helps evaluate the predictive power of the model:

- A high R^2 suggests the model can make reasonably accurate predictions
- A low R^2 indicates high prediction uncertainty and model unreliability

4. Communication of Results

It provides a simple, intuitive metric for communicating model performance to non-technical stakeholders:

- Easy to explain: "The model explains 87% of the variation in the target variable"
- Provides quick assessment of model adequacy

Interpretation Guidelines:

- **R^2 > 0.70:** Strong explanatory power (good fit)
- **R^2 = 0.50-0.70:** Moderate explanatory power (fair fit)
- **R^2 < 0.50:** Weak explanatory power (poor fit)

Important Caveats:

- **R^2 always increases or stays the same** when adding more variables, regardless of their relevance
- **High R^2 does not guarantee** causation or that the model is appropriate
- **Context matters:** In some fields (e.g., social sciences), even R^2 = 0.30 may be acceptable
- **R^2 can be negative** for models that perform worse than a horizontal line through the mean

Adjusted R-squared:

To address the issue of R^2 always increasing with more variables, adjusted R-squared penalizes the addition of unnecessary variables:

$$\text{Adjusted } R^2 = 1 - \frac{(1 - R^2)(n - 1)}{n - p - 1}$$

Where n is the number of observations and p is the number of predictors.

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

Answer:

Import necessary libraries

```
import numpy as np
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split
import pandas as pd
```

Sample Data: Square footage and house prices

X = np.array([[1000], [1500], [2000], [2500], [3000], [3500], [4000], [4500]])
y = np.array([200000, 250000, 300000, 350000, 400000, 450000, 500000, 550000])

Create and fit the Linear Regression model

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

Extract slope and intercept

slope = model.coef_[0]
intercept = model.intercept_

Print results

print("Simple Linear Regression Results")
print("="*40)
print(f"Slope (β_1): {slope:.2f}")
print(f"Intercept (β_0): {intercept:.2f}")
print("="*40)

Display the regression equation

print(f"\nRegression Equation:")
print(f"y = {intercept:.2f} + {slope:.2f} × x")

Make predictions

print(f"\nPredictions:")
X_test = np.array([[2200]])
y_pred = model.predict(X_test)
print(f"Predicted price for 2200 sq ft: ${y_pred[0]:,.2f}")

Calculate R-squared

r_squared = model.score(X, y)
print(f"\nR-squared (R^2): {r_squared:.4f}")

Additional predictions

print(f"\nAdditional Predictions:")
test_values = np.array([[1200], [2800], [3500]])
predictions = model.predict(test_values)
for square_feet, price in zip(test_values, predictions):
print(f"{square_feet[0]} sq ft → ${price:,.2f}")

Expected Output:

Simple Linear Regression Results

Slope (β_1): 100.00
Intercept (β_0): 100000.00

Regression Equation:
y = 100000.00 + 100.00 × x

Predictions:
Predicted price for 2200 sq ft: $320,000.00

R-squared (R^2): 1.0000

Additional Predictions:
1200 sq ft → $220,000.00
2800 sq ft → $380,000.00
3500 sq ft → $450,000.00

Alternative Code with Train-Test Split:

More robust example with train-test split

from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_squared_error, r2_score
import numpy as np

Create sample dataset

np.random.seed(42)
X = np.random.randn(100, 1) * 1000 + 2500
y = 100000 + 100 * X.flatten() + np.random.randn(100) * 10000

Split data into training and testing sets

X_train, X_test, y_train, y_test = train_test_split(
X, y, test_size=0.2, random_state=42
)

Create and fit the model

model = LinearRegression()
model.fit(X_train, y_train)

Print slope and intercept

print(f"Slope (β_1): {model.coef_[0]:.2f}")
print(f"Intercept (β_0): {model.intercept_:.2f}")

Evaluate the model

y_pred = model.predict(X_test)
mse = mean_squared_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)

print(f"Mean Squared Error: {mse:.2f}")
print(f"R-squared: {r2:.4f}")

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

Answer:

Understanding Regression Coefficients:

In Simple Linear Regression, the two main coefficients are the **intercept (β_0)** and the **slope (β_1)**. Interpreting these correctly is essential for understanding the model's meaning and making informed decisions.

1. The Slope (β_1) Coefficient

Definition: The slope represents the change in the dependent variable y for a one-unit increase in the independent variable x.

Interpretation Guidelines:

- **Positive slope ($\beta_1 > 0$):** There is a positive relationship between x and y. As x increases, y tends to increase proportionally.
  Example: If $\beta_1 = 50$, then for each additional year of work experience, salary increases by $50,000 on average.

- **Negative slope ($\beta_1 < 0$):** There is a negative relationship between x and y. As x increases, y tends to decrease.
  Example: If $\beta_1 = -0.5$, then for each additional advertising unit, customer complaints decrease by 0.5 on average.
- **Zero slope ($\beta_1 = 0$):** There is no linear relationship between x and y.
- **Large magnitude:** A steeper slope means stronger relationship and greater impact of x on y.

2. The Intercept (β_0) Coefficient

Definition: The intercept is the predicted value of y when $x = 0$. It is the y-intercept where the regression line crosses the y-axis.

Interpretation Guidelines:

- **Practical meaning:** The intercept often represents a baseline or starting value.
  Example: In the equation Price = 100,000 + 100 × Square Footage, the intercept of $100,000 represents the base price (possibly reflecting land value) independent of house size.
- **May not be meaningful:** In many real-world cases, $x = 0$ is not realistic or observable in the data range.
  Example: For predicting salary based on years of experience, an intercept represents salary at 0 years of experience, which may not be meaningful in the actual job context.
- **Centering solution:** When the intercept is not interpretable, researchers sometimes center the data around the mean to make the intercept more meaningful.

3. Combined Interpretation

The complete regression equation provides the full relationship:

$$\hat{y} = \beta_0 + \beta_1 x$$

Example 1: House Price Prediction

Regression equation: $\text{Price} = 50,000 + 150 \times \text{Square Footage}$

- **β_0 = 50,000:** The baseline house price is $50,000 (possibly representing land value and base construction costs)
- **β_1 = 150:** Each additional square foot adds $150 to the house price
- **Interpretation:** A 2,000 sq ft house is predicted to cost:
  $50,000 + 150(2,000) = \$350,000$

Example 2: Student Performance

Regression equation: $\text{Test Score} = 40 + 5 \times \text{Study Hours}$

- **β_0 = 40:** A student with 0 study hours would be expected to score 40 points (baseline score)
- **β_1 = 5:** Each additional hour of study increases the test score by 5 points on average
- **Interpretation:** A student studying 10 hours is predicted to score: $40 + 5(10) = 90$ points

4. Statistical Significance

Beyond interpretation, consider:

- **Statistical Significance:** Is the coefficient significantly different from zero? (Determined using p-values and confidence intervals)
- **Practical Significance:** Even if statistically significant, is the effect size practically meaningful?
- **Confidence Intervals:** The range of plausible values for the coefficient (e.g., β_1 could be between 145 and 155)

5. Effect Size and Standardized Coefficients

- **Unstandardized coefficients:** Expressed in original units (e.g., dollars, years)
- **Standardized coefficients (beta):** Allow comparison of relative importance when variables have different scales
  Formula: $\beta_{standardized} = \beta_{unstandardized} \times \frac{SD_x}{SD_y}$

Common Interpretation Mistakes to Avoid:

- ✕ **Confusing correlation with causation:** A significant β_1 shows association, not necessarily causation
- ✕ **Extrapolation beyond data range:** Predictions outside the observed x range are unreliable
- ✕ **Ignoring context:** Always consider domain knowledge and reasonableness of interpretations
- ✕ **Overinterpreting intercepts:** When $x = 0$ is outside the data range, the intercept may not be interpretable

Summary Table:

Aspect	Meaning	Example
Positive β_1	Direct relationship	Height increases with age
Negative β_1	Inverse relationship	Price decreases with age
Large $\lvert\beta_1\rvert$	Strong effect	Small change in x causes large change in y
β_0	Baseline/intercept	Predicted y when $x = 0$
$\beta_0 = 0$	Line passes through origin	Possible in physically meaningful models

---

End of Assignment Solutions

All questions answered comprehensively. Total marks possible: 200 (20 marks per question)