**1. What assumptions does linear regression make?**

Linear regression relies on several key assumptions to produce valid and unbiased estimates:

* **Linearity**: The relationship between the independent variables and the dependent variable is linear.
* **Independence**: Observations are independent of each other.
* **Homoscedasticity**: Constant variance of residuals (errors) across all levels of the independent variables.
* **Normality of errors**: The residuals (errors) are normally distributed.
* **No multicollinearity**: Independent variables should not be too highly correlated with each other.

**2. How do you interpret the coefficients?**

* Each **coefficient** represents the **change in the dependent variable (Y)** for a **one-unit change in the independent variable (X)**, **holding all other variables constant**.
* The **sign** of the coefficient (+/-) shows the direction of the relationship.
* The **magnitude** shows how much impact that variable has on the target.

Example:  
If Y = 2 + 3X1 - 4X2, then:

* For each 1-unit increase in X1, Y increases by 3 (if X2 stays constant).
* For each 1-unit increase in X2, Y decreases by 4.

**3. What is R² score and its significance?**

* **R² (R-squared)** measures the proportion of variance in the dependent variable that is explained by the independent variables.
* **Range**: 0 to 1 (or negative if the model performs worse than a horizontal line).

**Interpretation:**

* R² = 0.75 → 75% of the variance in Y is explained by the model.
* A higher R² generally means a better fit, but it does **not guarantee** a good model (watch out for overfitting).

**4. When would you prefer MSE over MAE?**

* **MSE (Mean Squared Error)** penalizes **larger errors more** due to squaring. Use it when:
  + Large errors are especially undesirable.
  + You want to heavily penalize outliers.
  + The problem benefits from a differentiable loss function (e.g., in gradient-based optimization).
* **MAE (Mean Absolute Error)** treats all errors equally. Use it when:
  + You want a more **robust metric** (less sensitive to outliers).
  + You're concerned with **typical errors**, not squared ones.

**5. How do you detect multicollinearity?**

* **Correlation matrix**: High correlation between independent variables is a sign.
* **Variance Inflation Factor (VIF)**:
  + VIF > 5 (or >10 in some cases) suggests multicollinearity.
* **High standard errors** of coefficients or unstable coefficient estimates when adding/removing features.

**6. What is the difference between simple and multiple regression?**

* **Simple Linear Regression**:
  + One independent variable.
  + Formula: Y = β₀ + β₁X + ε
* **Multiple Linear Regression**:
  + Two or more independent variables.
  + Formula: Y = β₀ + β₁X₁ + β₂X₂ + ... + βₙXₙ + ε

**7. Can linear regression be used for classification?**

* **Technically no**, but:
  + **Linear regression is not ideal for classification** tasks because it predicts continuous values.
  + If forced, you could map predictions to classes (e.g., if output > 0.5, predict class 1), but this is **not recommended**.
  + For binary classification, use **logistic regression** instead—it models probabilities and is better suited.

**8. What happens if you violate regression assumptions?**

* **Linearity violation** → Model is biased, predictions are unreliable.
* **Independence violation** (e.g., in time series) → Underestimated standard errors, false significance.
* **Homoscedasticity violation** → Inefficient estimates, misleading confidence intervals.
* **Normality violation** → Affects inference (confidence intervals, p-values), but less of a concern for large samples.
* **Multicollinearity** → Inflated standard errors, hard to interpret coefficients, unstable predictions.