□ What assumptions does linear regression make?

Linear regression assumes:

- Linearity: The relationship between the independent variables and the dependent variable is linear.
- Independence: Observations are independent of each other.
- Homoscedasticity: The variance of residuals (errors) is constant across all levels of the independent variables.
- Normality: The residuals are normally distributed (primarily for inference, like hypothesis testing).
- No multicollinearity: Independent variables are not highly correlated with each other (in multiple regression).
- No autocorrelation: Residuals are not correlated with each other (important in time-series data).

□ How do you interpret the coefficients?

In linear regression, the coefficients represent the change in the dependent variable for a one-unit change in the independent variable, holding all other variables constant.

- The intercept (β_0) is the predicted value of the dependent variable when all independent variables are zero.
- Slope coefficients (β_1, β_2, etc.) indicate the direction and magnitude of the relationship: positive for direct relationships, negative for inverse. For example, if $\beta_1 = 2.5$ for a variable like "years of experience," it means the dependent variable (e.g., salary) increases by 2.5 units for each additional year, assuming other factors are fixed.

□ What is R^2 score and its significance?

The R^2 (R-squared) score, also known as the coefficient of determination, measures the proportion of the variance in the dependent variable that is predictable from the independent variables. It ranges from 0 to 1 (or 0% to 100%), where:

- 0 indicates the model explains none of the variability (as good as predicting the mean).
- 1 indicates the model explains all the variability perfectly. Its significance lies in evaluating model fit: higher R^2 suggests better explanatory power, but it doesn't imply causation, and it can be misleading in overfitted models or when comparing models with different numbers of predictors (use adjusted R^2 for that).

□ When would you prefer MSE over MAE?

Mean Squared Error (MSE) penalizes larger errors more heavily than Mean Absolute Error (MAE) because it squares the residuals. Prefer MSE when:

- You want to emphasize and reduce large outliers or extreme errors in predictions.
- The problem involves optimization where differentiability is key (e.g., in gradient descent algorithms).

- The data has a normal distribution of errors, as MSE aligns with maximum likelihood estimation under normality. MAE is better for robustness to outliers or when all errors should be treated equally.

⍰ How do you detect multicollinearity?

Multicollinearity occurs when independent variables are highly correlated, leading to unstable coefficient estimates. Detection methods include:

- Correlation matrix: Check pairwise correlations; values above 0.8-0.9 signal issues.
- Variance Inflation Factor (VIF): Calculate VIF for each variable; VIF > 5-10 indicates high multicollinearity.
- Condition number: From the eigenvalue decomposition of the feature matrix; a high condition number (>30) suggests problems.
- Tolerance: Inverse of VIF; low tolerance (<0.1-0.2) flags multicollinearity. If detected, remedies include removing variables, combining them (e.g., PCA), or using regularization (e.g., Ridge regression).

⍰ What is the difference between simple and multiple regression?

Aspect	**Simple Linear Regression**	**Multiple Linear Regression**
Number of Predictors	One independent variable (e.g., $Y = \beta_0 + \beta_1 X + \varepsilon$)	Two or more independent variables (e.g., $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \varepsilon$)
Purpose	Models the relationship between two variables	Models the relationship while controlling for multiple factors
Complexity	Easier to interpret and visualize (e.g., scatter plot with line)	More complex; requires checking for multicollinearity and interactions
Use Case	Basic predictions, like height predicting weight	Real-world scenarios, like house price based on size, location, and age

⍰ Can linear regression be used for classification?

Technically yes, but it's not ideal. Linear regression outputs continuous values, so for binary classification, you could threshold predictions (e.g., >0.5 as class 1). However, it violates assumptions like bounded outputs (predictions can be outside [0,1]) and homoscedasticity. Better alternatives include logistic regression (for binary) or softmax regression (for multi-class), which model probabilities properly and handle non-linear decision boundaries via transformations.

⍰ What happens if you violate regression assumptions?

Violating assumptions can lead to:

- Biased or inefficient coefficient estimates (e.g., non-linearity causes systematic errors).
- Invalid inference: Unreliable p-values, confidence intervals, or hypothesis tests (e.g., non-normality affects t-tests).

- Poor predictions: Heteroscedasticity or autocorrelation inflates error variance, reducing model accuracy.
- Unstable models: Multicollinearity causes high variance in coefficients, making them sensitive to data changes. Remedies include transformations (e.g., log for non-linearity), robust methods (e.g., weighted least squares), or switching models (e.g., GLM for non-normality). Always check residuals with plots like Q-Q or scatter for diagnostics.