Interview Questions:

Q1.What assumptions does linear regression make?
Ans.Linear regression assumes a linear relationship between independent and dependent variables, meaning changes in predictors lead to proportional changes in the outcome. It assumes independence of observations, homoscedasticity (constant variance of errors), and that residuals are normally distributed. It also expects no multicollinearity among predictors. Violating these assumptions can lead to biased or inefficient estimates and unreliable predictions.

Q 2.How do you interpret the coefficients?
Ans. In linear regression, each coefficient represents the expected change in the target variable for a one-unit increase in the corresponding predictor, assuming all other variables are held constant. A positive coefficient indicates a direct relationship, while a negative coefficient indicates an inverse relationship. The intercept represents the predicted value of the target when all predictors are zero.

Q 3.What is R^2 score and its significance?
Ans. The R^2 score, or coefficient of determination, indicates how well a linear regression model explains the variability of the target variable. It ranges from 0 to 1, where 0 means the model explains none of the variation and 1 means it explains all of it. For example, an R^2 of 0.85 means that 85% of the variation in the dependent variable is explained by the model. A higher R^2 generally suggests a better fit, but it doesn't guarantee accuracy or imply that the model is appropriate—especially if assumptions are violated or if overfitting occurs.

Q4.When would you prefer MSE over MAE?
Ans. You would prefer MSE (Mean Squared Error) over MAE (Mean Absolute Error) when you want to penalize larger errors more heavily. MSE squares the errors, so it gives more weight to large deviations, making it useful when outliers are especially important or costly. It’s also mathematically convenient for optimization in many machine learning algorithms. However, it’s more sensitive to outliers than MAE, which treats all errors equally.

Q 5.How do you detect multicolinearity?
Ans. Multicollinearity is detected by checking how strongly independent variables are correlated with each other. Common methods include examining the correlation matrix for high pairwise correlations and calculating the Variance Inflation Factor (VIF) for each predictor. A VIF value above 5 (or 10, depending on context) typically indicates multicollinearity. High multicollinearity can make coefficient estimates unstable and reduce the interpretability of the model.

Q 6.What is the difference between simple and multiple regression?
Ans.Simple regression involves one independent variable to predict a dependent variable, showing the direct linear relationship between them. In contrast, multiple regression uses two or more independent variables to predict the dependent variable, allowing for the analysis of more complex relationships. Multiple regression provides a more comprehensive view but also requires checking for issues like multicollinearity and overfitting.

Q7.Can linear regression be used for classification?
Ans. Linear regression is not suitable for classification because it predicts continuous values, not class labels. While it might be used to separate classes in a binary setting by applying a threshold (e.g., above 0.5 = class 1), this approach is unreliable as it doesn't constrain outputs between 0 and 1 or handle probabilities properly. Logistic regression is the preferred method for classification tasks, as it models the probability of class membership using a sigmoid function.

Q8.What happens if you violate regression assumptions?

Ans. Violating regression assumptions can lead to biased, inefficient, or misleading results. For example, if the relationship between variables isn't linear, the model may underfit. Non-constant error variance (heteroscedasticity) can distort confidence intervals and significance tests. Non-normal residuals affect the validity of hypothesis testing, while multicollinearity inflates standard errors, making it hard to assess the effect of individual predictors. Violating independence can result in underestimated standard errors and overconfident predictions. Overall, such violations weaken the model's reliability and interpretability.