Linear Regression Interview Question

What is linear regression?

- Linear regression is a statistical method for modeling the relationship between a dependent variable and one or more independent variables by fitting a linear equation to observed data.

Explain the difference between simple linear regression and multiple linear regression.

- Simple linear regression involves one independent variable, while multiple linear regression involves two or more independent variables.

What are the assumptions of linear regression?

- Linearity: The relationship between the independent and dependent variables is linear.
- Independence: Observations are independent of each other.
- Homoscedasticity: Constant variance of errors.
- Normality: Errors are normally distributed.
- No multicollinearity: Independent variables are not highly correlated.

How do you interpret the coefficients in a linear regression model?

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

What is the purpose of the intercept term in linear regression?

- The intercept term represents the value of the dependent variable when all independent variables are zero.

How do you evaluate the performance of a linear regression model?

- Common metrics include R-squared, adjusted R-squared, Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE).

What is R-squared, and how is it interpreted?

- R-squared is the proportion of the variance in the dependent variable that is predictable from the independent variables. It ranges from 0 to 1, with higher values indicating a better fit.

What is multicollinearity, and why is it a problem in linear regression?

- Multicollinearity occurs when independent variables are highly correlated, which can make it difficult to determine the individual effect of each variable on the dependent variable and can lead to unreliable coefficient estimates.

How can you detect multicollinearity?

- Common methods include checking the correlation matrix for high correlations, calculating Variance Inflation Factor (VIF), and examining eigenvalues of the feature matrix.

What is the difference between L1 and L2 regularization in linear regression?

- L1 regularization (Lasso) adds the absolute value of the magnitude of coefficients as a penalty term to the loss function, promoting sparsity. L2 regularization (Ridge) adds the squared magnitude of coefficients as a penalty term, promoting small coefficients.

Explain the concept of gradient descent in the context of linear regression.

- Gradient descent is an optimization algorithm used to minimize the cost function by iteratively adjusting the model parameters in the direction of the steepest descent of the cost function.

What is the normal equation, and how is it used in linear regression?

- The normal equation is a closed-form solution to linear regression, derived by setting the derivative of the cost function with respect to the model parameters to zero and solving for the parameters.

How do you handle outliers in linear regression?

- Outliers can be handled by removing them, transforming variables, using robust regression techniques, or employing regularization methods.

What is heteroscedasticity, and how can it be addressed?

- Heteroscedasticity refers to the non-constant variance of errors. It can be addressed by transforming variables, using weighted least squares, or applying robust standard errors.

Explain the difference between linear regression and logistic regression.

- Linear regression is used for predicting continuous outcomes, while logistic regression is used for predicting binary or categorical outcomes by modeling the probability of a class membership.