1. What is Linear Regression?

Linear regression is a statistical method used for modeling the relationship between a dependent variable and one or more independent variables by fitting a linear equation to observed data. The simplest form is simple linear regression, where we model the relationship between two variables as: $Y=\beta 0+\beta 1X+\epsilon Y=\beta_0+\beta_1X+\epsilon$ Where:

- YY is the dependent variable.
- XX is the independent variable.
- $\beta 0\beta_0$ and $\beta 1\beta_1$ are coefficients.
- $\epsilon\epsilon$ represents the error term, accounting for the variability in YY not explained by XX.

A practical example is predicting house prices (YY) based on their sizes (XX). As house size increases, we expect the price to increase, provided other factors are constant.

2. What are the assumptions made in linear regression?

Linear regression assumes:

- **Linearity**: The relationship between the predictors and the target variable is linear.
- **Independence**: Observations are independent of each other.
- **Homoscedasticity**: Constant variance of error terms irrespective of the value of predictors.
- **Normality**: For any fixed value of XX, YY is normally distributed.
- **No multicollinearity**: If multiple predictors are used, they should not be too highly correlated.

Violating these assumptions can lead to inefficiency and bias in the model's estimates.

3. How do you interpret the coefficients of a linear regression model?

In the equation $Y=\beta 0+\beta 1X+\epsilon Y=\beta_0+\beta_1X+\epsilon$:

- $\beta 0\beta_0$ (intercept) is the expected value of YY when all XX are 0.
- $\beta 1\beta_1$ (slope) represents the change in YY associated with a one-unit change in XX. A positive $\beta 1\beta_1$ indicates that as XX increases, YY also increases, and

vice versa for a negative $\beta 1 \beta_1$. If $\beta 1 \beta_1$ is close to zero, XX has little impact on YY.

4. What is the difference between simple and multiple linear regression?

Simple linear regression involves two variables (one independent and one dependent variable). In contrast, multiple linear regression involves more than one independent variable influencing the dependent variable. Multiple regression is used to examine more complex relationships by controlling for various factors simultaneously.

5. How do you assess the performance of a linear regression model?

Performance is commonly assessed using:

- **R-squared**: The proportion of variance in the dependent variable that is predictable from the independent variables. High R-squared values indicate a model that closely fits the data.
- **Adjusted R-squared**: Adjusts the R-squared value for the number of predictors in the model. It is useful for comparing models with different numbers of predictors.

6. How would you handle categorical variables in a linear regression model?

Categorical variables must be converted into numerical values. This is commonly done through one-hot encoding, which creates a new binary column for each category or target encoding

7. Can you explain how you would diagnose and remedy multicollinearity in a regression model?

Multicollinearity can be diagnosed using Variance Inflation Factor (VIF) – values greater than 10 indicate high multicollinearity. Remedies include removing highly correlated predictors, combining them into a single predictor, or using regularization techniques.

8. How does regularization help in linear regression?

Regularization adds a penalty term to the loss function used to estimate the coefficients, which helps to prevent overfitting by keeping the coefficients small.

Common methods are Ridge (adds squared magnitude of coefficient as penalty term) and Lasso (adds absolute value of coefficients as penalty term).

9. Describe a time you used linear regression in a past project.

In a project aiming to predict retail sales, linear regression was used to model sales as a function of marketing spend and seasonality. Challenges included dealing with holiday effects, which were addressed by including dummy variables for holidays in the model.

10. How would you handle non-linear relationships using a linear model?

Non-linear relationships can be modeled using polynomial terms (e.g., $X2$,$X3X_2$,X_3) or transformations of the variables (like log, square root), which can be included as predictors in the linear model.

12. Given a dataset with missing values, how would you prepare it for a linear regression analysis?

Handle missing data by imputation (replacing missing values with the mean, median, or mode of the column), or by removing rows or columns with a high percentage of missing values.

13. What diagnostics would you perform on the residuals of a linear regression model?

Check residuals for:

- **Normality**: Using Q-Q plots or statistical tests (Read About them)
- **Constant variance**: Plotting residuals vs. fitted values.
- **Independence**: Checking for autocorrelation in the residuals.

14. How do you deal with outliers in your dataset when performing linear regression?

Identify outliers using visual methods like scatter plots or statistical methods like the Interquartile Range (IQR). Once identified, you can remove them or use robust regression methods.

15. Can you explain stepwise regression and its pros and cons?

Stepwise regression includes forward selection, backward elimination, and both. It automates the process of model selection by adding or removing predictors based on their statistical significance. However, it can lead to models that are overfitted to the sample data and may not generalize well.

These detailed answers should give a thorough understanding of various aspects of linear regression, suitable for preparing for interviews or deepening your understanding of the topic.

ONLY FOR CURIOUS : Mathematics

Derivation of the Ordinary Least Squares (OLS) Estimator

Objective:

The objective of OLS is to find the parameter estimates that minimize the sum of the squared residuals. For a simple linear regression model $Y = \beta_0 + \beta_1 X + \epsilon$, where ϵ are the error terms, the residual for each observation i is $e_i = y_i - (\beta_0 + \beta_1 x_i)$. The sum of squared residuals (SSR) is given by:

$SSR = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$

Minimization:

To find the values of β_0 and β_1 that minimize the SSR, we take the partial derivatives of SSR with respect to β_0 and β_1, and set them to zero.

1. **Partial derivative with respect to β_0:**

$\frac{\partial}{\partial \beta_0} SSR = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0$

$\sum_{i=1}^{n} y_i - n\beta_0 - \beta_1 \sum_{i=1}^{n} x_i = 0$

$n\beta_0 + \beta_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$

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

where $\bar{y}$ and $\bar{x}$ are the sample means of Y and X respectively.

2. **Partial derivative with respect to β_1:**

$\frac{\partial}{\partial \beta_1} SSR = -2 \sum_{i=1}^{n} x_i (y_i - \beta_0 - \beta_1 x_i) = 0$

$\sum_{i=1}^{n} x_i y_i - \beta_0 \sum_{i=1}^{n} x_i - \beta_1 \sum_{i=1}^{n} x_i^2 = 0$

$\beta_0 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i y_i - \beta_1 \sum_{i=1}^{n} x_i^2$

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