1. What is linear regression?

Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. The goal is to find a linear equation that best predicts the dependent variable based on the values of the independent variables.

---

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

In simple linear regression we have only one independent feature and in multiple linear regression we have more than 1 independent variable

---

3. What is the linear regression equation?

Y=β0+β1X1+β2X2+…+βnXn+ϵ where Y is the dependent variable, X1,X2,…,Xn are the independent variables, β0 is the intercept, β1,β2,…,βn are the coefficients, and ϵ\epsilonϵ is the error term.

---

4. What do the coefficients in a linear regression model represent?

The coefficients in a linear regression model represent the estimated effect of each independent variable on the dependent variable, holding all other variables constant. Specifically:

1. Intercept: This is the expected value of the dependent variable when all independent variables are zero.
2. Slope coefficients: Each of these represents the change in the dependent variable for a one-unit increase in the corresponding independent variable, assuming all other variables remain constant

---

5. How do you interpret the intercept in a linear regression model?

The intercept in a linear regression model represents the expected value of the dependent variable when all independent variables are set to zero.

---

5. What are the assumptions of linear regression?

Linearity: The relationship between the independent variables and the dependent variable is linear.

Independence: The observations are independent of each other.

Homoscedasticity: The variance of the residuals is constant across all levels of the independent variables.

Normality: The residuals are normally distributed.

No multicollinearity: The independent variables are not highly correlated with each other.
No autocorrelation: The residuals are not correlated with each other.

---

6. What is multicollinearity, and how can it affect a linear regression model?

Multicollinearity refers to a situation in linear regression where two or more independent variables are highly correlated with each other.it becomes challenging to determine the effect of single variable on dependent variable since increase in one independent variable will effect the another correlated independent variable. so will not get the exact relation between one variable with the dependent variable. also it will increase the risk of overfitting.

---

7. How can you detect and handle multicollinearity?

Methods to detect multicollinearity include:

Variance Inflation Factor (VIF), Correlation matrix

To address multicollinearity, you might:

Remove one of the correlated variables, Combine correlated variables, Use regularization techniques (like ridge regression)

---

8. What is heteroscedasticity,effect of heteroscedasticity and how can it be detected?, how to fix Heteroscedasticity.

Heteroscedasticity refers to the situation in a regression model where the variability of the residuals (the differences between observed and predicted values) is not constant across all levels of the independent variables.

Effects of heteroscedasticity:

Inefficient estimates: While coefficient estimates remain unbiased, they are no longer the most efficient (lowest variance) estimates.

Biased standard errors: This leads to incorrect confidence intervals and hypothesis tests.

Unreliable F-tests and t-tests: These tests become less reliable, potentially leading to incorrect conclusions about the significance of variables.

Detection methods:

Residual plots: Plot residuals against fitted values or independent variables.

Breusch-Pagan test: Tests whether the variance of the errors depends on the values of the independent variables.

The null hypothesis (H0): Signifies that Homoscedasticity is present.

The alternative hypothesis: (Ha): Signifies that the Homoscedasticity is not present.

Transform the dependent variable: We can alter the dependent variable using some technique. For example, we can take the log of the dependent variable.

---

9. What is autocorrelation, and how does it affect linear regression models?

Autocorrelation refers to the correlation between the error terms of a regression model at different time points or observations. It commonly occurs in time series data but can also appear in other types of data.

Effects of autocorrelation on linear regression models:

Biased standard errors:

Usually, autocorrelation leads to underestimated standard errors. Due to biased standard errors, confidence intervals and hypothesis tests become unreliable.

Detection methods:

Durbin-Watson test:

Tests for first-order autocorrelation. Values range from 0 to 4, with 2 indicating no autocorrelation

---

10. Explain R-squared and adjusted R-squared.

R-squared: An R-squared of 0.7 means that 70% of the variability in the dependent variable is explained by the model. R-squared always increases (or stays the same) when you add more variables to the model, even if these variables are not meaningful.

Adjusted R-squared It increases only if the new term improves the model more than would be expected by chance. It can decrease if a predictor improves the model less than expected by chance.

---

11. What is stepwise regression?

It's used for variable selection, particularly when dealing with a large number of potential independent variables. There are three main types of stepwise regression:

Forward Selection:

Starts with no variables in the model. Adds the most significant variable (usually based on F-statistics or p-values) at each step. Continues adding variables until no remaining variable meets the specified criteria for entry.

Backward Elimination:

Starts with all candidate variables in the model. Removes the least significant variable at each step. Continues removing variables until all remaining variables meet the specified criteria to stay in the model.

Bidirectional Elimination (Stepwise Regression):

Combines forward selection and backward elimination. Variables can be added or removed at each step. Reassesses variables added in earlier steps, potentially removing them if they become non-significant.

---

12. What is regularization in the context of linear regression?

Regularization in the context of linear regression is a technique used to prevent overfitting and improve the generalization of the model. It does this by adding a penalty term to the loss function, which discourages the model from relying too heavily on any individual feature. The two main types of regularization for linear regression are:

L1 Regularization (Lasso Regression):

Adds the absolute value of the magnitude of coefficients as a penalty term. Can lead to sparse models by shrinking some coefficients to exactly zero. Loss function: $L(\beta) = RSS + \lambda \Sigma|\beta j|$

L2 Regularization (Ridge Regression):

Adds the squared magnitude of coefficients as a penalty term. Shrinks all coefficients towards zero, but typically doesn't make them exactly zero. Loss function: $L(\beta) = RSS + \lambda \Sigma\beta j^2$

---

13. What is polynomial regression, and when would you use it?

Polynomial regression is an extension of linear regression that allows for modeling nonlinear relationships between the independent and dependent variables. It does this by including polynomial terms (squared, cubed, etc.) of the predictor variables in the regression equation.

---

14. How can you detect and mitigate overfitting in linear regression?

Train-Test Split Performance:

Significant difference between training and test set performance indicates overfitting. The model performs much better on the training data than on the test data.

R-squared vs. Adjusted R-squared:

A large difference between R-squared and adjusted R-squared suggests overfitting.

Mitigation Strategies: Regularization, Feature Selection, Increase Training Data