Linear Regression

Question 1: Explain Linear Regression

Interviewer's Expectation: The candidate should provide a clear and concise explanation of linear regression, including its main objective.

Answer:
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 from the independent variables. The model is represented as:

$Y=\beta 0+\beta 1X1+\beta 2X2+...+\beta nXn+\epsilon Y=\beta 0+\beta 1X1+\beta 2X2+...+\beta nXn+\epsilon$

where YY is the dependent variable, $XiXi$ are the independent variables, $\beta i\beta i$ are the coefficients to be estimated, and $\epsilon\epsilon$ is the error term, assumed to be normally distributed.

Question 2: What is the significance of the R-squared statistic in linear regression?

Interviewer's Expectation: The interviewer looks for an understanding of how R-squared measures the goodness of fit of a linear regression model.

Answer:
R-squared is a statistic that measures the proportion of variance in the dependent variable that can be explained by the independent variables in the model. It is a value between 0 and 1, where 0 means no explained variability and 1 means perfect explanation. High R-squared values indicate that the model explains a significant portion of the variance, whereas low values suggest poor explanatory power. However, it's important to note that a high R-squared does not imply causality, nor does it indicate whether the regressions estimates and predictions are unbiased.

Question 3: Describe how you would check for multicollinearity in your predictors and the steps you would take if you find it.

Interviewer's Expectation: The interviewer wants to know if you understand multicollinearity and its effects, and if you can handle it practically.

Answer:
To check for multicollinearity, I would start by looking at the correlation matrix of the predictors to spot any high correlations between variables. Following that, I would calculate the Variance Inflation Factor (VIF) for each predictor; a VIF greater than 10 indicates significant multicollinearity. If multicollinearity is found, I would consider several approaches:

1. **Removing variables**: Eliminate one or more of the highly correlated predictors.
2. **Principal Component Analysis (PCA)**: Transform the predictors into a smaller number of uncorrelated components.
3. **Regularization methods**: Use techniques like Ridge or Lasso regression that can handle multicollinearity by adding a penalty to the loss function.

Question 4: How do you interpret the coefficients of a linear regression model?

Interviewer's Expectation: The candidate should clearly explain what the coefficients represent in the context of model prediction.

Answer:
In a linear regression model, each coefficient represents the change in the dependent variable for a one-unit change in the corresponding independent variable, holding all other variables constant. The coefficient $\beta 1$ $\beta 1$ associated with variable $X 1 X 1$ suggests that for each unit increase in $X 1 X 1$, the dependent variable $Y Y$ increases by $\beta 1 \beta 1$ units, assuming other variables remain constant. A positive coefficient implies a direct relationship, while a negative coefficient implies an inverse relationship.

Question 5: What are some assumptions underlying linear regression and how do you test them?

Interviewer's Expectation: Demonstrating knowledge of the assumptions and practical testing methods.

Answer:
The main assumptions of linear regression include linearity, independence, homoscedasticity, normal distribution of errors, and no multicollinearity. Here's how I test them:

1. **Linearity**: Check through scatter plots of observed vs. predicted values, or residual vs. predictor plots.

2. **Independence**: For data like time series, check for autocorrelation using the Durbin-Watson statistic.
3. **Homoscedasticity**: Look at a plot of residuals versus predicted values; patterns or funnels suggest heteroscedasticity.
4. **Normal Distribution of Errors**: Use a Q-Q plot to compare the distribution of residuals to a normal distribution.
5. **No Multicollinearity**: Calculate the Variance Inflation Factor (VIF) for each predictor.