**1)Explain the concept of regression analysis**

Regression analysis is a set of statistical methods used for the estimation of relationships between a dependent variable and one or more independent variables. It can be utilized to assess the strength of the relationship between variables and for modelling the future relationship between them.

**2)What is the difference between simple linear regression and multiple regression?**

* Linear regression is one of the most common techniques of regression analysis when there are only two variables.  linear regress only has one independent variable.
* Multiple regression is a broader class of regressions that encompasses linear and nonlinear regressions with multiple explanatory variables. multiple regression incorporates multiple independent variables.

**3) How can interpret the slope and intercept in a simple linear regression model?**

In a simple linear regression model represented as Y=MX+C. The slope (m) and intercept (c) in linear regression represent the parameters that determine the characteristics of the best-fit line.

M-The slope determines the rate of change of the dependent variable with respect to the independent variable. This signifies the change in the dependent variable (Y) for a one-unit change in the independent variable (X).

C- intercept represents the value of the dependent variable when the independent variable is zero. This is the value of the dependent variable (Y) when the independent variable (X) is zero. It represents the starting point of the regression line on the y-axis.

**4) What is the purpose of using multiple regression instead of simple linear regression?**

Multiple regression is used over simple linear regression when there are multiple independent variables that potentially influence the dependent variable. Purposes are,

* Modelling complex relationship
* Improves predictive accuracy.
* Handling multi variables
* Enhancing interpretability

**5) what is multicollinearity and why it's a concern in multiple regression?**

what is multicollinearity and why it's a concern in multiple regression.it can cause,

* Decrease predictive accuracy.
* Instability of coefficients
* Inflate standard error.
* Masking variable importance

**6) how to assess the goodness of fit in a regression model?**

some common methods used to evaluate the goodness of fit in regression are,

* R-squared
* Adjusted R-squared.
* MSE, MAE, RMSE.
* Residual analysis

**7) Explain assumptions of linear regression. How to check if these assumptions are met?**

Main assumptions are,

* Linearity
* Independence of errors
* Homoscedasticity
* Normality of residuals
* No perfect multicollinearity
* No autocorrelation of residuals
* Mean of residuals is zero.

Methods to check these assumptions are,

* Residual analysis
* Multicollinearity
* Homoscedasticity Test
* Linearity checking

**8) what is the role of regularization in regression models in regression models, and why might it be necessary?**

Regularization methods provide a means to control our regression coefficients, which can reduce the variance and decrease sample error. Regularization plays role in regression models by addressing overfitting and improving the model's predictive performance. It involves adding a penalty term to the regression model's objective function to prevent extreme parameter values. Two methods are,

* Lasso regularization
* Ridge regularization

It is used because,

* Prevent Overfitting
* Dealing with multicollinearity
* Balancing bias & variance
* Handling noisy data
* Improving model stability

**9) Describe the process of feature selection in the context of multiple regression**

Feature selection is the process of identifying and selecting a subset of input variables that are most relevant to the target variable. Perhaps the simplest case of feature selection is the case where there are numerical input variables and a numerical target for regression predictive modelling. Steps involved are,

1. Understand the data and problem.
2. Exploratory data analysis
3. Apply feature selection methods.
4. Evaluate Selected features.
5. Iterate and Refine
6. Final model Building

**10) How to handle outliers in regression analysis?**

1.Identifying outliers using visualization and z-score

2.Outlier handling

* Trimming/Remove the outlier.
* Quantile based flooring and capping.
* Mean/median imputation.

**11)** What is the difference between r-squared and adjusted r-squared? When to prefer one over the other?

1)R-Squared

R-squared measures the proportion of variance in the dependent variable that is explained by the independent variables in the model. It ranges between 0 and 1, where 1 indicates that all variability in the dependent variable is explained by the independent variables, while 0 suggests that none of the variability is explained.

2)Adjusted R-squared

Adjusted R-squared is a modified version of R-squared that adjusts for the number of predictors in the model. It penalizes the addition of unnecessary predictors that do not significantly improve the model's fit, addressing the issue of overfitting that R-squared doesn't account for.

* If simplicity and interpretability are important and there's no concern about overfitting due to the number of predictors, R-squared may suffice.
* If comparing models with different numbers of predictors or when addressing potential overfitting caused by including too many predictors, adjusted R-squared is more appropriate.