1. **Explain the linear regression algorithm in detail.**

Linear regression is a modeling technique where we try to explain relationship between a dependent variable and one or more independent variable using linear equation. The underlying assumption is that the variables are linearly dependent. It is supervised learning technique. We make the model on train set and evaluate using test set.

Regression between one dependent variable and one independent variable is called simple linear regression and is represented by the line equation.

$$y = B_0 + B_1 x \qquad \text{--- 1}$$

where m and c are slope and intercept respectively. We calculate value of m and c for the line that represents the data called best-fit line. For this we use minimize the square error. Error is the square of difference between the actual y value and the predicted y value added together for every point.

$$e_i = y_i - y_{pred} \qquad \text{-- 2}$$

2nd equation can be written as.

$$e_i = y_i - B_0 - B_1 x_i \qquad \text{-- 3}$$

$$e = e_1 + e_2 + \ldots + e_n \qquad \text{-- 4}$$

e is called Residual Sum of Squares (RSS).RSS is not sufficient to evaluate the model as it is not a normalized quantity. R squared is used to evaluate the model.

R squared = RSS/TSS.

$$TSS = \sum_{i=0}^{n} (y_i - \bar{y}) \qquad \text{-- 5}$$

R squared is normalized i.e. between 0 and 1 and is used to evaluate the fitness of model.

After the model has been evaluated for fitness. We also have to check for residual error in the model. The two properties it should follow are
1. The errors should be randomly distributed.
2. There should be no pattern between the errors and the actual y values.

For multiple linear regression. the line gets converted to a hyperplane and the equation becomes :

$$y = B_0 + B_1x_1 + B_2x_2 + \cdots + B_nx_n \quad \text{-- 6}$$

Where x1,x2, …, xn are independent variables and B1,B2, …, Bn are the intercepts.

There are two more considerations.
1. Number of features (independent variables) to select. Adding more variables will increase the complexity of the model and may suffer from overfitting.
2. Multicollinearity - relations between different features.

Multicollinearity is represented by VIF (Variance Inflation Factor)

$$\text{VIF} = \frac{1}{1-R_i^2} \quad \text{-- 7}$$

Less VIF means less collinearity of the feature with other features.
To evaluate different multiple linear regression models, adjusted R square is calculated for evaluation of different models with different selected features and different sample sizes. The idea is to penalize the model with high number of variables.

$$\text{Adjusted } R^2 = \frac{(1-R^2)(N-1)}{N-p-1} \quad \text{-- 8}$$

p is number of features selected and N is sample size.

2. What are the assumptions of linear regression regarding residuals?

Three assumptions are made regarding residuals :
1. The residuals are normally distributed.
2. Sum of residuals are zero
2. There is no pattern between the residuals and the predictors.

3. **What is the coefficient of correlation and the coefficient of determination?**

   Coefficient of determination (R square) is the square of coefficient of correlation (R)

4. **Explain the Anscombe's quartet in detail.**

   Anscombe's quartet is group of four data sets such that the summary statistics (mean, variance, correlation, R square) for the those datasets are same but when plotted shows that the shape of graphs are completely different.

   It show the importance of EDA on the data- set and to not relying only on the summary statistics before making any machine learning models.

5. **What is Pearson's R?**

   Pearson's R is the measure of correlation between two variables X and Y. It ranges between +1 and -1. Positive value shows positive correlation i.e. if one increased, other increases. Negative value shows negative correlation i.e. if one increased , other decreases.

6. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?**

   scaling is a way of standardizing the values of dataset. It makes the calculation faster and makes it easy to interpret values of coefficients.
   In normalized scaling , the data point is scaled to the range [0,1] or [1,-1].
   In standardized scaling, the data point is scales such that the mean is 0 and standard deviation is 1.

7. **You might have observed that sometimes the value of VIF is infinite. Why does this happen?**

   infinite VIF means that the feature is completely described by another feature i.e. R square of the two features is 1.

8. **What is the Gauss-Markov theorem**

   Gauss-Markov theorem states that the ordinary least squares (OLS) estimation is the best linear unbiased estimate (BLUE) if it follow the following assumptions:

   1. Linearity : the parameters we are estimating are linear.

2. Random : Data is randomly sampled.
3. Non-Collinearity : Selected features are not correlated.
4. Exogeneity : there is no pattern between selected features and error terms.
5. Homoscedasticity: variance of error is constant.

9. Explain the gradient descent algorithm in detail.

Gradient descent is an optimization algorithm. It takes a cost function and maximize or minimize the value depending on the requirement. Linear regression model uses gradient descent to calculate the intercept and slope of the regression line.
Gradient descent is a first order iterative method. It first starts from a position and then moves towards the minima/maxima by iteratively adding values . First we calculate the gradient of the cost function :

$$\frac{d}{d\theta}J(\theta)$$

where $J(\theta)$ is the cost function. For one variable the equation is as follows.

$$\theta_1 = \theta_0 + \eta\frac{dJ}{d\theta} \text{ on } \theta = \theta_0$$

η is the learning factor. It determines how fast the cost function reaches the optimal solution. If η is small it will take many iterations to reach solution. If η is large, the iteration will skip pass the solution and will oscillate in the consecutive iterations till it reaches the solution.

10. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Q - Q (Quantile - Quantile) plot is a plot where quantiles of x is plotted against the quantiles of y. They are used to check whether data fits a normal distribution.