LINEAR REGRESSION SUBJECTIVE QUESTIONS

1. Explain the linear regression algorithm in detail.

Linear regression is a type of machine learning algorithm that finds the best straight line to represent the relationship between a dependent variable and one or more independent variables. It can be used to predict the value of the dependent variable based on the independent variables. Linear regression is used in many fields to understand and predict the behaviour of a particular variable.

When the number of independent features is one, it is known as Univariate Linear Regression, and in the case of more than one feature, it is known as Multivariate Linear Regression.

The goal of the algorithm is to find the best linear equation that can predict the value of the dependent variable based on the independent variables. The equation provides a straight line that represents the relationship between the dependent and independent variables. The slope of the line indicates how much the dependent variable changes for a unit change in the independent variable(s).

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

Linear regression makes several assumptions about the residuals. According to web search results, these assumptions include:

- *Independence:* Residuals should be independent of each other and of the predictor variables.
- *Constant variance (homoscedasticity):* Residuals should have constant variance at any value of the predictor variables.
- *Zero mean:* Residuals should have a mean of zero.
- *Normality:* Residuals should follow a normal distribution.

If one or more of these assumptions are violated, then the results of our linear regression may be unreliable or even misleading

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

The *coefficient of correlation*, also known as the R-value, measures the strength and direction of a linear relationship between two variables (x and y) with possible values between -1 and 1. A positive correlation indicates that two variables rise and fall together, while a negative correlation indicates that one variable goes up while the other goes down. If there is no linear correlation or a weak linear correlation, the R-value is close to 0.

The *coefficient of determination*, also known as R square or R^2, is the square of the coefficient of correlation. It measures the proportion of the variability in y that is accounted for by the linear relationship between x and y. Its value is usually between 0 and 1 and it indicates the strength of a Linear Regression model. The higher the R^2 value, the less scattered the data points are, so it is a good model. The lower the R^2 value, the more scattered the data points are.

4. Explain Anscombe's quartet in detail.

Anscombe's quartet comprises four datasets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed. Each dataset consists of eleven (x, y) points. They demonstrate both the importance of graphing data when analyzing it and the effect of outliers and other influential observations on statistical properties.

The first scatter plot (top left) appears to be a simple linear relationship, corresponding to two variables correlated where y could be modeled as Gaussian with mean linearly dependent on x. The second graph (top right); while a relationship between the two variables is obvious, it is not linear, and the Pearson correlation coefficient is not relevant. In the third graph (bottom left), the modeled relationship is linear, but should have a different regression line (a robust regression would have been called for). The calculated regression is offset by the one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.816. Finally, the fourth graph (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.

The quartet is still often used to illustrate the importance of looking at a set of data graphically before starting to analyze according to a particular type of relationship and the inadequacy of basic statistic properties for describing realistic datasets.

5. What is Pearson's R?

Pearson's R, also known as the Pearson correlation coefficient, is a measure of linear correlation between two sets of data. It is a number between -1 and 1 that measures the strength and direction of the relationship between two variables. A positive correlation indicates that two variables rise and fall together, while a negative correlation indicates that one variable goes up while the other goes down. If there is no linear correlation or a weak linear correlation, the R-value is close to 0.

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

Scaling is a data preprocessing technique that involves transforming the values of features or variables in a dataset to a similar scale. This is done to ensure that all features contribute equally to the model and to prevent features with larger values from dominating the model. Scaling is essential when working with datasets where the features have different ranges, units of measurement, or orders of magnitude.

There are several common feature scaling techniques, including normalization and standardization. Normalization, also known as Min-Max Scaling, is used to transform features to be on a similar scale by scaling the range to [0, 1] or sometimes [-1, 1]. Standardization, also known as Z-Score Normalization, is the transformation of features by subtracting from the mean and dividing by standard deviation. This transforms your data so the resulting distribution has a mean of 0 and a standard deviation of 1.

The main difference between normalization and standardization is that normalization changes the range of your data while standardization changes the shape of the distribution of your data.

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

The value of VIF (Variance Inflation Factor) can be infinite when there is perfect correlation between two or more independent variables. This happens when the R-squared value approaches 1, which leads to the denominator of the VIF formula becoming 0 and the overall value becoming infinite. It denotes perfect correlation in variables.

8. What is the Gauss-Markov theorem?

The Gauss-Markov theorem states that under certain conditions, the ordinary least squares (OLS) estimator of the coefficients of a linear regression model is the best linear unbiased estimator (BLUE). This means that the OLS estimator has the smallest variance among those that are unbiased and linear in the observed output variables.

The conditions for the Gauss-Markov theorem to hold are that the errors in the linear regression model are uncorrelated, have equal variances and expectation value of zero. If these conditions are met, then the OLS estimator is the most efficient unbiased linear estimator.

9. Explain the gradient descent algorithm in detail.

Gradient descent is an optimization algorithm used to find the values of parameters (coefficients) of a function that minimizes a cost function. It is commonly used to train machine learning models and neural networks.

The algorithm works by iteratively adjusting the values of the parameters in the direction of the negative gradient of the cost function with respect to the parameters. The size of the steps taken in the direction of the negative gradient is determined by the learning rate, a hyperparameter that controls how quickly the algorithm converges to a minimum.

The goal of gradient descent is to find the values of the parameters that minimize the cost function. This is done by iteratively updating the values of the parameters until the algorithm converges to a minimum, at which point the cost function is close to or equal to zero.

There are several variations of gradient descent, including batch gradient descent, stochastic gradient descent, and mini-batch gradient descent. These variations differ in how they calculate the gradient of the cost function and how they update the values of the parameters.

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

A Q-Q plot (quantile-quantile plot) is a graphical tool used to compare two probability distributions by plotting their quantiles against each other. In linear regression, Q-Q plots are often used to assess whether the residuals of a model are normally distributed.

The use and importance of a Q-Q plot in linear regression lie in its ability to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal distribution. This is important because many statistical tests and methods, including linear regression, assume that the residuals are normally distributed.

If the residuals are normally distributed, then the points on the Q-Q plot should approximately lie on a straight line. If the points deviate significantly from a straight line, then this may indicate that the residuals are not normally distributed and that the assumptions of the linear regression model may not be met.