1. Explain the linear regression algorithm in detail.
2. What are the assumptions of linear regression regarding residuals?
3. What is the coefficient of correlation and the coefficient of determination?
4. Explain the Anscombe's quartet in detail.
5. What is Pearson's R?
6. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
7. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
8. What is the Gauss-Markov theorem?
9. Explain the gradient descent algorithm in detail.
10. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Answers

1. Explain the linear regression algorithm in detail.

Ans: The linear regression is well known and best algorithm of statistics and machine learning

Where we can find the best fit line from the given data also derives the relationship between independent and dependent variable which is done by the sum of residual squares where we calculate the distance of each point from its best fit line and R2 is the sum of all these square, the regression means the output variable is predicted to be continuous variable

There are 2 type of linear regression

Simple linear regression

Multiple linear regression

Simple linear regression where the R2 for one independent variable and dependent variable

The multiple linear regression you the R2 with multiple independent variable and one dependent variable where if one independent variable coefficients changes the others are kept constant.

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

Ans: There are certain assumptions for residuals in linear regression the following are the assumptions:

a)Normality Assumptions :

The errors or the residuals follow a normal distribution, however is the sample size increases the Normality assumption is not required. When we take a repeated sampling from our population for large samples sizes, the distribution of the ordinary least squares estimates of regression coefficients follow a

normal distribution which is also considered the inferential procedures and result is consequence of central limit theorem.

b)Zero mean assumptions : In this the residuals have a mean value of zero that is the error terms are normally distributed around zero

c) Constant variance assumptions: It is assumed that the residual values have same variance this is also known as assumptions of homogeneity or homoscedasticity. Where the residuals are equally distributed around the regression line

d) No Multicollinearity: there is no collinearity with the independent variables with each other , so there is no or little multicollinearity the best way to check it is pair plot or the heap map.

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

Ans: The coefficient of correlation is the R value which you get from the summary table of the Regression output. The R square is called coefficient of determination which you get by multiplying R by R times. In other words coefficient of determination is square of coefficient of correlation. The R square gives the percentage of variation in Y which is explained by all the X variables and it lies in between 0 to 1 as it is a square value, higher the better. The R square is preferred for both simple and multilinear regression. R values lies in between -1, 1 and have let you know the variables are negatively or positively correlated or if they are not correlated at all.

4. Explain the Anscombe's quartet in detail.

Ans: Anscombe's Quartet comprises four datasets, each containing eleven (x,y) pairs. The essential thing to note about these datasets is that they share the same descriptive statistics. But things change completely, when they are graphed. Each graph tells a different story irrespective of their similar summary statistics.

When we plot these four datasets on an x/y coordinate plane, we can observe that they show the same regression lines as well but each dataset is telling a different story.

Dataset I have clean and well-fitting linear models.

Dataset II is not distributed normally.

Dataset III the distribution is linear, but the calculated regression is thrown off by an outlier.

Dataset IV shows that one outlier is enough to produce a high correlation coefficient.

Anscombe's Quartet reminds us that graphing data prior to analysis is good practice, outliers should be removed when analyzing data, and statistics about a data set do not fully depict the data set in its entirety.

5. What is Pearson’s R?

Ans: Pearson's r is a numerical summary of the strength of the linear association between the variables. The Pearson’s correlation coefficient lies between -1 to 1:

r = 1 means the data is perfectly linear with a positive slope

r = -1 means the data is perfectly linear with a negative slope

r = 0 means there is no linear association

r > 0 < 5 means there is a weak association

r > 5 < 8 means there is a moderate association

r > 8 means there is a strong association

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

Ans: Scaling means transforming the data into a specific range like 0 to 1, 0 to 100 depending on the method you use. You use scaling to compare different variables on equal footing that would let us know if there is any relation between the variables

Normalization typically means rescales the values into a range of [0,1]. Standardization typically means rescales data to have a mean of 0 and a standard deviation of 1 (unit variance).

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

Ans: Variance inflation factor (VIF) is used to check the presence of multicollinearity in a data set.

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well) and if there is perfect correlation with other variables so it is better to remove the variables which are perfectly correlated.

8. What is the Gauss-Markov theorem?

Ans: The Gauss Markov theorem tells us that if a certain set of assumptions are met, the ordinary least squares estimate for regression coefficients gives you the best linear unbiased estimate (BLUE) possible.

The Assumptions are as follows:

Linearity: the parameters we are estimating using the OLS method must be themselves linear.

Random: our data must have been randomly sampled from the population.

Non-Collinearity: the regressor’s being calculated aren’t perfectly correlated with each other.

Exogeneity: the regressor’s aren’t correlated with the error term.

Homoscedasticity: no matter what the values of our regressor’s might be, the error of the variance is constant.

9. Explain the gradient descent algorithm in detail.

Ans: Gradient descent is an optimization algorithm used to find the values of parameters of a function that minimizes cost and finds the values of the estimators corresponding to the optimized cost values

The graph of the gradient function is like a ball rolling a graph, the ball moves in the direction of greatest gradient which is maxima and comes rest to the flat surface which is minima

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

Quantile-Quantile plots are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile.
if we run a statistical analysis that assumes our dependent variable is Normally distributed, we can use a Normal Q-Q plot to check that assumption. It's just a visual check, not an air-tight proof, so it is somewhat subjective. But it allows us to see at-a-glance if our assumption is plausible, and if not, how the assumption is violated and what data points contribute to the violation.

A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight. It is use to fit a linear regression model, check if the points lie approximately on the line,