1. Explain the linear regression algorithm in detail.

The Linear algorithm in detail

Regression: - is the technique to find the dependent variable by calculating the independent variable, regression is a basic and commonly used type of predictive analysis.

Simple linear regression:

When the dependent variable(x) is dependent on only one independent variable(y) , It is assumed that the two variables are linearly related. Hence, we try to find a linear function that predicts the response value(y) as accurately as possible as a function of the feature or independent variable(x).

Liner Equation: y = a_0 + a_1 * X

y_n=Dependent variable

X=Independent Variable

a-0=Y_intercept

a_1=Slope Coefficient for independent variable

The motive of the linear regression algorithm is to find the best values for a_0 and a_1

Cost Function

The cost function helps us to figure out the best possible values for a_0 and a_1 which would provide the best fit line for the data points

$$minimize\frac{1}{n}\sum_{i=1}^{n}(pred_i - y_i)^2$$

$$J = \frac{1}{n}\sum_{i=1}^{n}(pred_i - y_i)^2$$

Here yi is the actual value and predi is the predicted value of y, by summation of squaring their difference we get the value of Residual Square sum which is the error, by minimizing this function we can get the slope coefficient for x.

Multiple liner regression:

When the dependent variable(x) is dependent on more than one independent variable(y), The goal of multiple linear regression (MLR) is to model the linear relationship between the independent variables and dependent variable.

Liner Equation: y_n= a_0+a_1*X1+a_2*X2+ --------a_n*Xn y_n=Dependent variable

X1,X2,--Xn= Independent Variable

a-0=Y_intercept

a_1,a_2------a_n=Slope Coefficient for each independent variable

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

Assumptions of Linear Regression Regarding Residual

a) Normality assumption: It is assumed that the error terms are normally distributed. If the residuals are not normally distributed, their randomness is lost, which implies that the model is not able to explain the relation in the data. And the Model is not considered as good model.

b) Zero mean assumption: It is assumed that the residuals have a mean value of zero, i.e., the error terms are normally distributed around zero.

c) Constant variance assumption: It is assumed that the residual terms have the same variance. This assumption is also known as the assumption of homogeneity or homoscedasticity.

d) Independent error assumption: It is assumed that the residual terms are independent of each other, i.e. their pair-wise co-variance is zero. This means that there is no correlation between the residuals and the predicted values, or among the residuals themselves.

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

Coefficient of Correlation

It is the measure of strength of relationship between the movement of two different variables. Its values are observed between -1 and 1 where -1 indicates that the two variable suppose x1 and x2 are not hooding any sort of relationship, if x1 is increasing then x2 must be decreasing, i.e. x1 and x2 are moving in opposite direction, while if the correlation coefficient is 1 of x1 and x2 then it indicated that both the variable are moving in the same direction. If the correlation coefficient is 0 then there is no relationship between x1 and x2.

Coefficient of Correlation

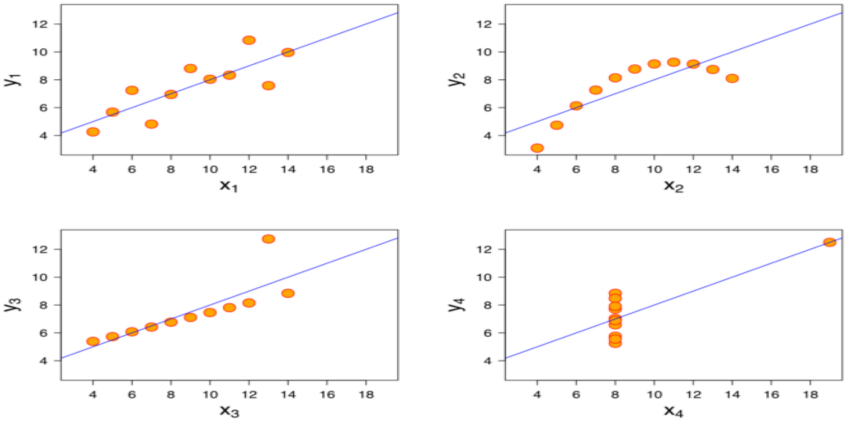
It is a statistical measurement that examines how differences in one variable can be explained by the difference in a second variable, when predicting the outcome of a given event. It can also be explaining as the value which explains the variance on the predicted linear model

This coefficient is commonly known as R-squared (or R2), and is sometimes referred to as the "goodness of fit."

This measure is represented as a value between 0.0 and 1.0, where a value of 1.0 indicates a perfect fit

4. Explain the Anscombe's quartet in detail.

Ascombe's quartet



This is the plot Ascombe focused aiming to explain the importence of visiualizing the data then just trusting the statisticks result.

Here the four graph is of four different data sets whoes Standar deviation, Mean and Sum were same. But when we look the graph it is telling the different story.

Dataset I appears to have clean and well-fitting linear models.

Dataset II is not distributed normally.

In 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.

This quartet emphasizes the importance of visualization in Data Analysis. Looking at the data reveals a lot of the structure and a clear picture of the dataset.

5. What is Pearson's R?

Pearson's Correlation Coefficient(r)

Correlation is a technique for investigating the relationship between two quantitative, continuous variables, for example, speed and distance covered . Pearson's correlation coefficient (r) is a measure of the strength of the association between the two variables. The value lies between -1 and 1.

Positive Correlation: both variables change in the same direction.

Neutral Correlation: No relationship in the change of the variables.

Negative Correlation: variables change in opposite directions.

The performance of some algorithms can deteriorate if two or more variables are tightly related, called multicollinearity.

r = -1	data lie on a perfect straight line with a negative slope
r = 0	no linear relationship between the variables
r = +1	data lie on a perfect straight line with a positive slope

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

Scaling

Scaling is to bring all the dependent and independent variable on the same scale.

Normalization

Normalization is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1. It is also known as Min-Max scaling.

Here's the formula for normalization:

$$x = \frac{x - mean(x)}{sd(x)}$$

Normalization equation

Here, Xmax and Xmin are the maximum and the minimum values of the feature respectively.

When the value of X is the minimum value in the column, the numerator will be 0, and hence X' is 0

On the other hand, when the value of X is the maximum value in the column, the numerator is equal to the denominator and thus the value of X’ is 1

If the value of X is between the minimum and the maximum value, then the value of X’ is between 0 and 1

Standardization

Standardization is another scaling technique where the values are centered around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero and the resultant distribution has a unit standard deviation.

Here’s the formula for standardization:

$$x = \frac{x - min(x)}{max(x) - min(x)}$$

Standardization equation

Feature scaling: Mu is the mean of the feature values and Feature scaling: Sigma is the standard deviation of the feature values. Note that in this case, the values are not restricted to a particular range.

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

Variance Inflation Factors

Variance inflation factors show the degree to which a regression coefficient will be affected because of the variable's redundancy with other independent variables. As the squared multiple correlation of any predictor variable with the other predictors approaches unity, the corresponding VIF becomes infinite.

$$\mathrm{VIF_i} = \frac{1}{1 - R_i^2}$$

If R_squared value becomes 1, i.e. the predicted model perfectly fits the actual values and in such suituation the VIF will be infine.

8. What is the Gauss-Markov theorem?

Gauss-Markov theorem

The Gauss-Markov theorem states that if all the assumptions of Linear Regession are satisfied by the linear regression model then the variance will be as smaller as possible.

The Gauss-Markov theorem famously states that OLS is BLUE. BLUE is an acronym for the following: Best Linear Unbiased Estimator

Best refers to the minimum variance or the narrowest sampling distribution.

9. Explain the gradient descent algorithm in detail.

Gradient Descent in Linear Regression

At theoretical level, It is an algorithim of minimising the function, where the function is the cost function. The cost function is the summation of the square of the error in the model.

In the gradient descent algorithm we minimize the cost function using calculus to get the best possible slope-coefficient and y-intercept which give the most accurate model.

linear regression model is defined as follows:

y = b+ m * x

p(i)=y

error = p(i) – y(i)

Where p(i) is the prediction for the i'th instance in our dataset and y(i) is the i'th output variable for the instance in the dataset.

$$\text{Error}_{(\text{m,b})} = \frac{1}{N}\sum_{i=1}^{N}(y_i - (mx_i + b))^2$$

$$\frac{\partial}{\partial \text{m}} = \frac{2}{N}\sum_{i=1}^{N} - x_i(y_i - (mx_i + b))$$

$$\frac{\partial}{\partial \text{b}} = \frac{2}{N}\sum_{i=1}^{N} - (y_i - (mx_i + b))$$

By doing partial diffriantiation w.r.t b and m , and equating the equation to zero we can find the respective value of b and m and we will get the best predicted model.

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

Q-Q PLot

Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.

This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.

It is used to check if:

The two data sets- i. come from populations with a common distribution

ii. have common location and scale

iii. have similar distributional shapes

iv. have similar tail behavior