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

Answer:
Linear regression is one of the easiest statistical models in ML which is used to predict real values based upon continuous variables. We establish a relationship between independent variables and dependent variables by fitting a best line.

Types of Linear Regression model

1. Linear regression
The Simple Linear Regression model can be represented using the below equation:

$y=\beta0+\beta1x + \varepsilon$

Where β0 is the intercept of the fitted line and β1 is the coefficient for the independent variable x and ε is the error term.

2. Multiple Linear regression
It represents the relationship between two or more independent input variables and a response variable. Multiple linear regression is needed when one variable might not be sufficient to create a good model and make accurate predictions.

$Y=\beta0+\beta1X1+\beta2X2+...+\beta pXp+\epsilon$

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

Answer:
Assumptions of linear regression

- There is a linear relationship between X and Y
- Error terms are normally distributed with mean zero(not X, Y)
- Error terms are independent of each other:
- Error terms have constant variance (homoscedasticity)
- The error terms must be normally distributed.
- The independent variables should not be correlated.
- There should be no correlation between the residual (error) terms.

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

Answer:

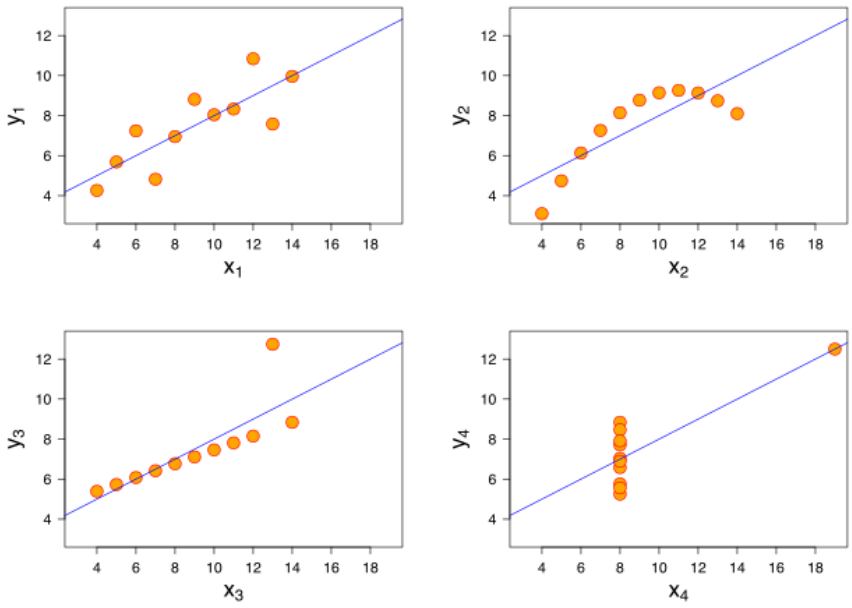
The correlation coefficient R of a model (say with variables x and y) takes values between −1 and 1. It describes how x and y are correlated.

- If x and y are in perfect unison, then this value will be positive 1
- If x increases while y decreases in exactly the opposite manner, then this value will be −1
- 0 would be a situation where there is no correlation between x and y

However, this R value is only useful for a simple linear model (just an x and y). Once we consider more than one independent variable (now we have x1, x2, ...), it is very hard to understand what the correlation coefficient means.
Tracking which variable contributes what to the correlation is not so clear.
This is where the **coefficient of determination** denoted by R2 value comes into play. It is simply the square of the correlation coefficient. It takes values between 0 and 1, where values close to 1 imply more correlation (whether positively or negatively correlated) and 0 implies no correlation.

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

Answer:
Anscombe's quartet comprises four data sets 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.



- Dataset X1 appears to have clean and well-fitting linear models.
- Dataset X2 is not distributed normally.
- Dataset X3 the distribution is linear, but the calculated regression is thrown off by an outlier.
- Dataset X4 shows that one outlier is enough to produce a high correlation coefficient.

5. What is Pearson's R?

Answer:
A correlation or simple linear regression analysis can determine if two numeric variables are significantly linearly related. A correlation analysis provides information on the strength and direction of the linear relationship between two variables, while a simple linear regression analysis estimates parameters in a linear equation that can be used to predict values of one variable based on the other.

The **Pearson correlation coefficient**, r, can take on values between -1 and 1. The further away r is from zero, the stronger the linear relationship between the two variables. The sign of r corresponds to the direction of the relationship. If r is positive, then as one variable increases, the other tends to increase. If r is negative, then as one variable increases, the other tends to decrease. A perfect linear relationship (r=-1 or r=1) means that one of the variables can be perfectly explained by a linear function of the other.

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

Answer:
Machine learning algorithms like linear regression, logistic regression, neural network, etc. that use gradient descent as an optimization technique require data to be scaled.

Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units. If scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values. It basically helps to normalise the data within a particular range. Sometimes, it also helps in speeding up the calculations in an algorithm.

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

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.

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

Answer:
VIF is an index that provides a measure of how much the variance of an estimated regression coefficient increases due to collinearity. In order to determine VIF, we fit a regression model between the independent variables.

An **infinite** VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables.

$VIF_i = 1/1-R^2_i$ where R^2_i is the coefficient of determination of a regression model where the *i*th factor is treated as a response variable in the model with all of the other factors. VIF_i can range from one to infinity. Values equal to one imply orthogonality, while values greater than one indicate a degree of collinearity between the *i*th factor and one or more other factors. The square root of the VIF indicates how much larger the standard error is (and therefore, how much larger the confidence intervals will be), compared to a factor that is uncorrelated with the other factors.

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

Answer:
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.

A theorem that proves that if the error terms in a *multiple regression* have the same variance and are uncorrelated, then the estimators of the parameters in the model produced by *least squares estimation* are better (in the sense of having lower dispersion about the mean) than any other unbiased linear estimator.

Gauss Markov Assumptions

1. Linearity: the parameters we are estimating using the OLS method must be themselves linear.
2. Random: our data must have been randomly sampled from the population.
3. Non-Collinearity: the regressors being calculated aren't perfectly correlated with each other.
4. Exogeneity: the regressors aren't correlated with the error term.
5. Homoscedasticity: no matter what the values of our regressors might be, the error of the variance is constant.

9. Explain the gradient descent algorithm in detail.

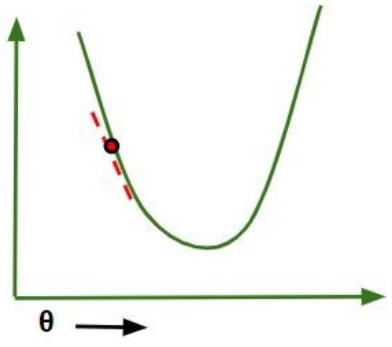
Answer:
Gradient descent is an optimization algorithm used to find the values of parameters (coefficients) of a function (f) that minimizes a cost function (cost).

Gradient descent is a first-order iterative optimization algorithm for finding a local minimum of a differentiable function. To find a local minimum of a function using gradient descent, we take steps proportional to the negative of the gradient (or approximate gradient) of the function at the current point. But if we instead take steps proportional to the positive of the gradient, we approach a local maximum of that function; the procedure is then known as gradient ascent.
To find the optimal betas of the line to fit best the data and one way of finding it is to follow optimisation methods such as Gradient Descent.

Gradient Descent Algorithm For Linear Regression

$$J(\theta 0,\theta 1)=\sum Ni=1(yi-yi(p))2$$ – cost function

$$\theta 1=\theta 0-\eta\partial\partial\theta J(\theta)$$ - Gradient Descent



Where η is known as the learning rate, which defines the speed at which we want to move towards negative of the gradient.

The way to find the optimal betas or thetas is know as Gradient Descent.

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

Answer:
The Q-Q (quantile-quantile) plot is a graphical technique for determining if two data sets come from populations with a common distribution.

A Q-Q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.

The advantages of the Q-Q plot are:

- The sample sizes do not need to be equal.
- Many distributional aspects can be simultaneously tested.

To formed Q-Q plot:

- Vertical axis: Estimated quantiles from data set 1
- Horizontal axis: Estimated quantiles from data set 2

Importance of Q-Q plot in 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.