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

Ans. Linear regression is a type of supervised learning algorithm which is based on finding out the best relationship between some independent variables and a target variable. Linear regression is commonly used for predictive analysis and modeling. For example, it can be used to quantify the relative impacts of age, gender, and diet (the predictor variables) on height (the outcome variable).

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

Ans. The assumptions of linear regression regarding residuals are:

- The error terms are normally distributed around zero (with zero mean)
- There is a constant variance between the error terms (Assumption of homoscedasticity)
- The errors terms are independent of each other.

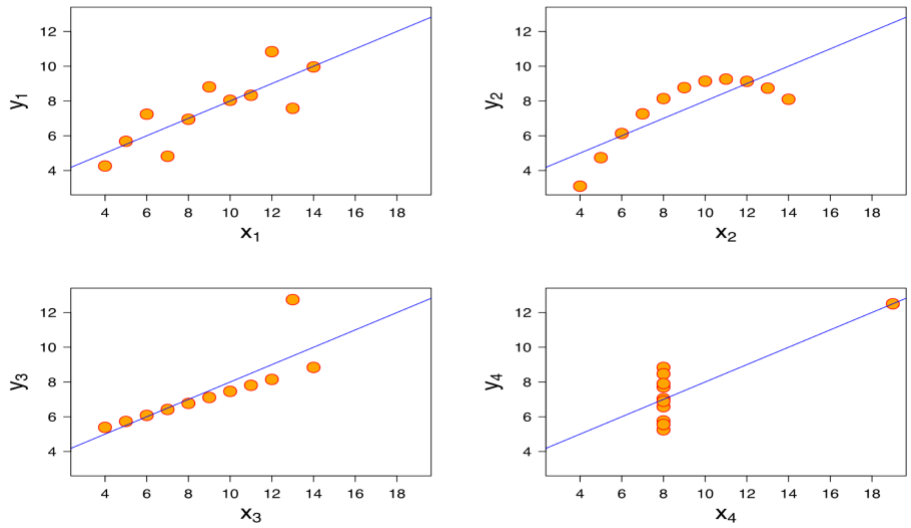
3. **What is the coefficient of correlation and the coefficient of determination?**

Ans. Coefficient of Correlation is a measure of the strength of the straight-line or linear relationship between two variables. A correlation coefficient of 1 means that for every positive increase in one variable, there is a positive increase of a fixed proportion in the other. For example, shoe sizes go up in (almost) perfect correlation with foot length. A correlation coefficient of -1 means that for every positive increase in one variable, there is a negative decrease of a fixed proportion in the other. For example, the amount of gas in a tank decreases in (almost) perfect correlation with speed. Zero means that for every increase, there isn't a positive or negative increase. The two just aren't related.

Coefficient of determination is the proportion of the variance in the dependent variable that is predictable from the independent variable(s).It is denoted by R2. It is a measurement used to explain how much variability of one factor can be caused by its relationship to another related factor. This correlation, known as the "goodness of fit," is represented as a value between 0.0 and 1.0.Higher the value, the better model explains our dependent variable.

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

Ans. 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. It is used to demonstrate both the importance of graphing data before analyzing it and the effect of outliers and other influential observations on statistical properties.



i) The first scatter plot (top left) appears to be a simple linear relationship, corresponding to two variables correlated where y could be modelled as gaussian with mean linearly dependent on x.

ii) The second graph (top right) is not distributed normally; while a relationship between the two variables is obvious, it is not linear, and the Pearson correlation coefficient is not relevant. A more general regression and the corresponding coefficient of determination would be more appropriate.

iii) In the third graph (bottom left), the distribution 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.

i**v)** 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.

5. What is Pearson's R?

Ans. It is a statistic that measures linear correlation between two variables X and Y. It has a value between +1 and −1, where 1 is total positive linear correlation, 0 is no linear correlation, and −1 is total negative linear correlation. It is the covariance of the two variables divided by the product of their standard deviations.

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} \quad \text{(Eq.1)}$$

where cov is the Covariance, X and Y are some random variables and σX and σY is the Std. deviation of X and Y respectively.

It is generally the first step in studying the relationship between two continuous variables. We can use scatterplot to see if there is any correlation.Also, we can draw heatmaps to see the correlation between multiple variables.

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

Ans. Scaling is a method used to normalize the range of independent variables or features of data. Since the range of values is different for different variables, objective functions like gradient descent will not work properly without scaling. Moreover, since each variable is on a different scale, it is very difficult to compare the coefficient of each independent variable.

Normalization scaling also known as the Min-Max scaling rescales the features in a range of [0-1]. The general formula is given by: x`=(x-min(x))/(max(x)-min(x)) where x is the original value and min(x) is the minimum value and max(x) is the maximum value. This is used when distribution is not Gaussian, responds well if standard deviation is small. This scaler is sensitive to outliers.

Standardized scaling also known as Z scaling makes the values of each feature in the data have zero-mean and one variance. Its formula is given by : x`=x-mean(x)/ σ
It assumes that data has normally distributed features and will scale them to zero mean and 1 standard deviation. After applying the scaler all features will be of same scale .

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

Ans. 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) i.e if there is a perfect correlation, then VIF is infinity. We should remove the variable if the VIF in infinity to avoid the problem of multicollinearity.

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.

There are five Gauss Markov assumptions (also called conditions):

i) Linearity: the parameters we are estimating using the OLS method must be themselves linear.
ii) Random: our data must have been randomly sampled from the population.
iii) Non-Collinearity: the regressors being calculated aren't perfectly correlated with each other.
iv) Exogeneity: the regressors aren't correlated with the error term.
v) 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.

Ans. 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 best used when the parameters cannot be calculated analytically (e.g. using linear algebra) and must be searched for by an optimization algorithm.

Gradient Descent Procedure

The procedure starts off with initial values for the coefficient or coefficients for the function. These could be 0.0 or a small random value.

coefficient = 0.0

The cost of the coefficients is evaluated by plugging them into the function and calculating the cost.

cost = f(coefficient) or cost = evaluate(f(coefficient))

The derivative of the cost is calculated. The derivative is a concept from calculus and refers to the slope of the function at a given point. We need to know the slope so that we know the direction (sign) to move the coefficient values in order to get a lower cost on the next iteration.

delta = derivative(cost)

Now that we know from the derivative which direction is downhill, we can now update the coefficient values. A learning rate parameter (alpha) must be specified that controls how much the coefficients can change on each update.

coefficient = coefficient – (alpha * delta)

This process is repeated until the cost of the coefficients (cost) is 0.0 or close enough to zero to be good enough.

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

Ans. The quantile-quantile (q-q) 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. By a quantile, we mean the fraction (or percent) of points below the given value. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value.

A 45-degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line. The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions.

Q-Q plot in Linear regression is used to assess if your residuals are normally distributed. Basically what you are looking for here is the data points closely following the straight line at a 45% angle upwards (left to right). 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.