1. Explain the linear regression algorithm in detail.

Ans: - Linear regression is a method of finding the best linear relationship between the independent and dependent variables.

In technical terms, linear regression is a machine learning algorithm that finds the best linear-fit relationship on any given data, between independent and dependent variables. It is mostly done by the Sum of Squared Residuals Method.

The key point in Linear Regression is that the dependent variable must be a continuous/real value. However, the independent variables can be measured on continuous or categorical values.

Linear regression algorithm has mainly two objectives:

Model the relationship between the input and output variables. Such as the relationship between Income and expenditure, experience and Salary, etc.

Forecasting new observations. Such as Weather forecasting according to temperature, Revenue of a company according to the investments in a year, etc.

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

Ans: -

- The mean of residuals is zero
    - It is assumed that the residuals have a mean value of zero, i.e., the error terms are normally distributed around zero.
- Homoscedasticity of residuals or equal variance
    - It is assumed that the residual terms have the same (but unknown) variance, σ2.
- No autocorrelation of residuals
    - It is assumed that the residual terms are independent of each other, i.e. their pair-wise covariance is zero.
- The X variables and residuals are uncorrelated
    - It is assumed that there is no correlation between the X variables and error terms.
- The residuals should be normally distributed.
    - It is assumed that the error terms, ε(i), are normally distributed.

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

Ans: - Correlation coefficient (R) measures linear relationship between two variables, while coefficient of determination (R-squared) measures explained variation.

For example; height and weight of individuals are correlated. If the correlation coefficient is R = 0.8 means there is high positive correlation. That means that both height and weight of individuals increase/decrease together (positive) and their relationship (linear) is strong.

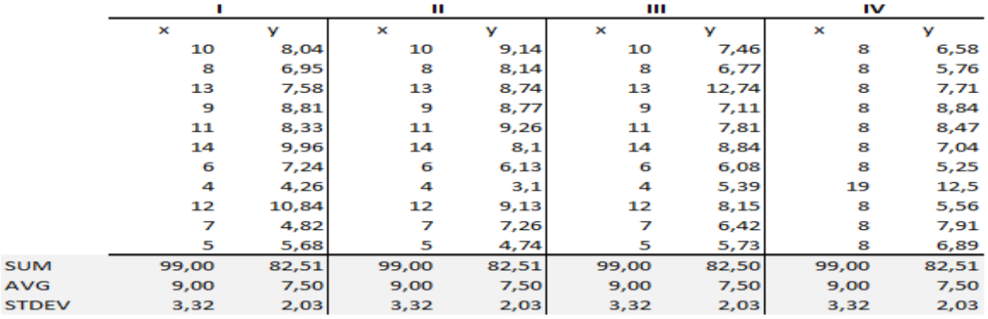
But height of individuals may also be affected by other factors like age, genetics, food intake, amount and type of exercise, location etc.

So, we if try to predict height by using weight as a single predictor, coefficient of determination is 0.64 (equals to square of correlation coefficient here).

It shows that 0.64 (or 64%) of variation in height can be explained by weight and remaining 36% of variation in height may be due to other factors which affect height of individuals like age, genetics, food intake, amount and type of exercise, location etc.

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. When we plot them, each graph tells a different story irrespective of their similar summary statistics.

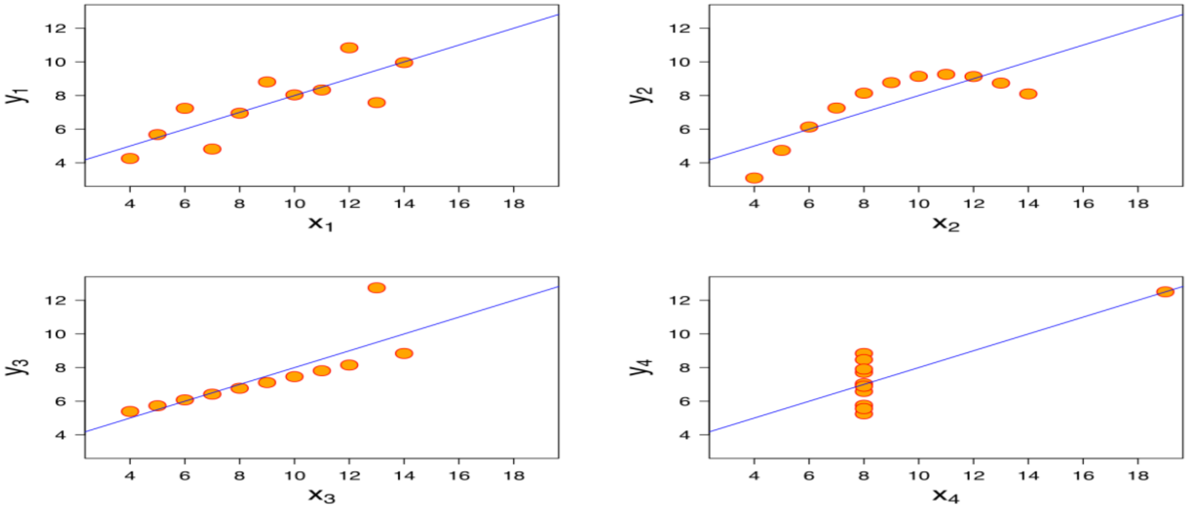


	I		II		III		IV	
	x	y	x	y	x	y	x	y
	10	8,04	10	9,14	10	7,46	8	6,58
	8	6,95	8	8,14	8	6,77	8	5,76
	13	7,58	13	8,74	13	12,74	8	7,71
	9	8,81	9	8,77	9	7,11	8	8,84
	11	8,33	11	9,26	11	7,81	8	8,47
	14	9,96	14	8,1	14	8,84	8	7,04
	6	7,24	6	6,13	6	6,08	8	5,25
	4	4,26	4	3,1	4	5,39	19	12,5
	12	10,84	12	9,13	12	8,15	8	5,56
	7	4,82	7	7,26	7	6,42	8	7,91
	5	5,68	5	4,74	5	5,73	8	6,89
SUM	99,00	82,51	99,00	82,51	99,00	82,50	99,00	82,51
AVG	9,00	7,50	9,00	7,50	9,00	7,50	9,00	7,50
STDEV	3,32	2,03	3,32	2,03	3,32	2,03	3,32	2,03

The summary statistics show that the means and the variances were identical for x and y across the groups.

- Mean of x is 9 and mean of y is 7.50 for each dataset.
- Similarly, the variance of x is 11 and variance of y is 4.13 for each dataset.
- The correlation coefficient between x and y is 0.816 for each dataset.

But, 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 appear 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?

Ans: - The Pearson product-moment correlation coefficient is a measure of the strength of a linear association between two variables. Basically, a Pearson product-moment correlation attempts to draw a line of best fit through the data of two variables, and the Pearson correlation coefficient, r, indicates how far away all these data points are to this line of best fit.

The Pearson's correlation coefficient varies between -1 and +1 where:

r = 1 means the data is perfectly linear with a positive slope (i.e. both variables tend to change in the same direction)

r = -1 means the data is perfectly linear with a negative slope (i.e. both variables tend to change in different directions)

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: - Feature Scaling is a technique to standardize or normalize 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 feature 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.

Normalization: This technique re-scales a feature or observation value with distribution value between 0 and 1.

$$X_{new} = \frac{X_i - \min(X)}{\max(x) - \min(X)}$$

Standardization: This technique re-scales a feature value so that it has distribution with 0 mean value and variance equals to 1.

$$X_{new} = \frac{X_i - X_{mean}}{\text{Standard Deviation}}$$

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

Ans: - 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.

VIF = 1/(1-R^2)

If all the independent variables are orthogonal to each other, then VIF = 1.0. If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables.

8. What is the Gauss-Markov theorem?

Ans: - The Gauss Markov theorem says that, if certain set of assumptions (mentioned below) are met in a linear model, then, the ordinary least squares (OLS) estimator of the coefficients of a linear regression model is the best linear unbiased estimator (BLUE),

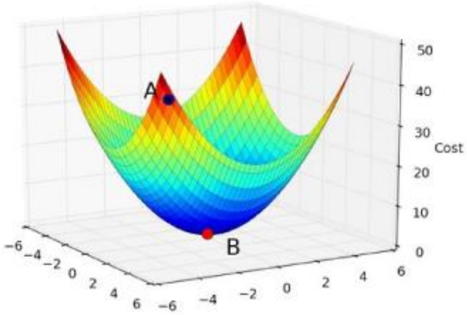
i.e. the estimator that has the smallest variance among those that are unbiased and linear in the observed output variables.

Assumptions: -

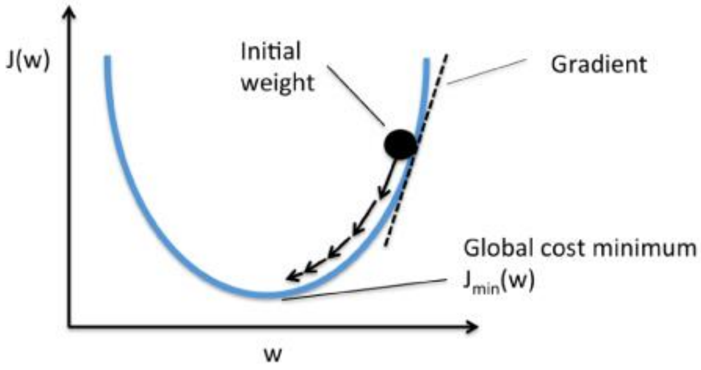
- *Linearity:* - The parameters we are estimating using the OLS method must be themselves linear.
- *Random:* - Our data must have been randomly sampled from the population.
- *Non-Collinearity:* - The regressors being calculated aren't perfectly correlated with each other.
- *Exogeneity:* - The regressors aren't correlated with the error term.
- *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 which is mainly used to find the minimum of a function. In machine learning, gradient descent is used to update parameters in a model.
For example, a ball is placed it on an inclined plane (at position A in below picture). As per laws, it will start rolling until it travels to a gentle plane where it will be stationary (at position B as shown in the figure below).



This is exactly what happens in gradient descent. The inclined is the cost function when it is plotted and the role of gradient descent is to provide direction and the velocity (learning rate) of the movement in order to attain the minima of the function i.e. where the cost is minimum.



The cost function can be calculated with below formula.

$$J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)}\right)^2$$

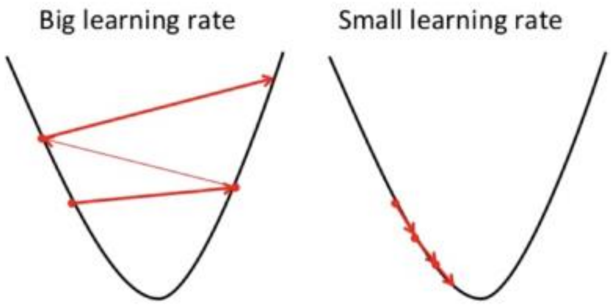
Here, h is the linear hypothesis model, h=Θ0 + Θ1x, y is the true output, and m is the number of data points in the training set.

Learning Rate: -

If we start by initializing Θ0 and Θ1 to any two values, say 0 for both, the algorithm is as follows:

$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta_0, \theta_1) \qquad (\text{for } j = 0 \text{ and } j = 1)$$

Where α, alpha, is the learning rate, or how rapidly do we want to move towards the minimum.



Here we calculate the partial derivative of the cost function. It helps us to know the direction (sign) in which the coefficient values should move so that they attain a lower cost.

Once we know the direction from the derivative, the solution is updated to the new value where the cost function has a lower value.

Repeat until convergence

$$\theta_j = \theta_j - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}).\ x_j^{(i)} \quad \text{for j} = 1,2,\ldots,n$$

Types of Gradient Descent Algorithms: -

Batch Gradient Descent – Where all the training examples are processed for each iteration of gradient descent. It gets computationally expensive if the number of training examples is large.

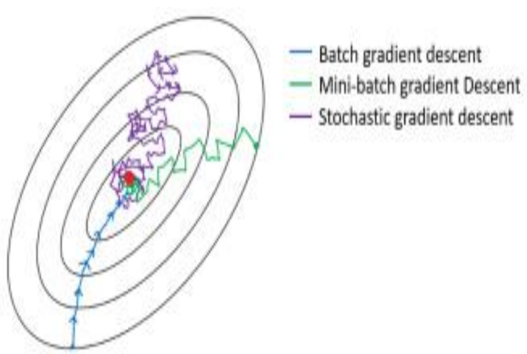
This is when batch gradient descent is not preferred, rather a stochastic gradient descent or mini-batch gradient descent is used.

Stochastic Gradient Descent – Where samples are selected at random for each iteration instead of selecting the entire data set.

When the number of training examples is too large, it becomes computationally expensive to use batch gradient descent, however, Stochastic Gradient Descent uses only a single sample, i.e., a batch size of one, to perform each iteration.

The sample is randomly shuffled and selected for performing the iteration. The parameters are updated even after one iteration where only one has been processed. Thus, it gets faster than batch gradient descent.

Mini Batch gradient descent - This type of gradient descent is considered to be faster than both batch gradient descent and stochastic gradient descent. Even if the number of training examples is large, it processes it in batches in one go. Also, the number of iterations is lesser in spite of working with larger training samples.



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

Ans: - Quantile-Quantile plot is a graphical tool to help us assess if a set of data possibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution.
It helps to determine if two data sets come from populations with a common distribution.
Also, this helps in case 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.

Advantages: -

- It can be used with sample sizes also
- Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.