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

Ans. Linear Regression is a type of **supervised learning**. LR is done on continues numeric data. A simple LR explains the relationship between the dependent and the independent variable by fitting a **straight line equation**.
The equation for straight line is – **Y = mx + c**
where m is the **slope** which denotes the relationship between y and x, and c is the **intercept** which is the value of y when x =0.
The **equation** for **simple LR** is as following – $\mathbf{E[Y] = \beta_0 + \beta_1 X}$.
There can be multiple factors that effect a dependent variable. These are called Multiple Linear regression.
The **equation for multiple linear regression** is – **E[Y] = B0 + B1*X1 + B2*X2 …….. Bn*Xn**

Example of LR: WhatsApp revenue based on number of WhatsApp users, house price prediction etc.

From Machine Learning perspective, LR is the simplest and most used model. He steps involved in LR are – 1. We plot a **graph** between dependent variable and independent variable.

2. We try to plot straight line and then **check for correlation**.
3. We **split the data** into train and test data.
4. Create **dummy data.**
5. **Scale** all the continues variables.
6. Then we do a **predictive analysis** – that is build model using training data and predict using test data. We can choose either sklearn model or stats model.
7. Then we keep on removing variables whose **p-values and vif are high** on the train data in order to get the **best fit** line.
8. We also calculate the **residual** value which is yi – y predicted value. The for formula for which is - $\sum_i (y_i - (\beta_0 + \beta_1 X))^2$
9. **Gradient descent** choses Bo and B1 in such a manner that the **error is decreased**.
10. In the end we check the **r square and adjusted r square** value of both test and train data.
11. **80%** and above r square is considered good and there should not be more than **5%** difference in r square and adjusted r square.

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

Ans. The assumptions of Linear regression are as follows –

1. **Linearity –** The relationship between X and the mean Y is linear.
2. **Homoscedasticity –** The variance of residual is same for any value , that is the error terms have constant variance.
3. **Independence –** Error terms are independent of each other.
4. **Normality –** The error terms are normally distributed.

NOTE – No assumptions are made on the distribution of X or Y.

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

Ans. Coefficient of determination –

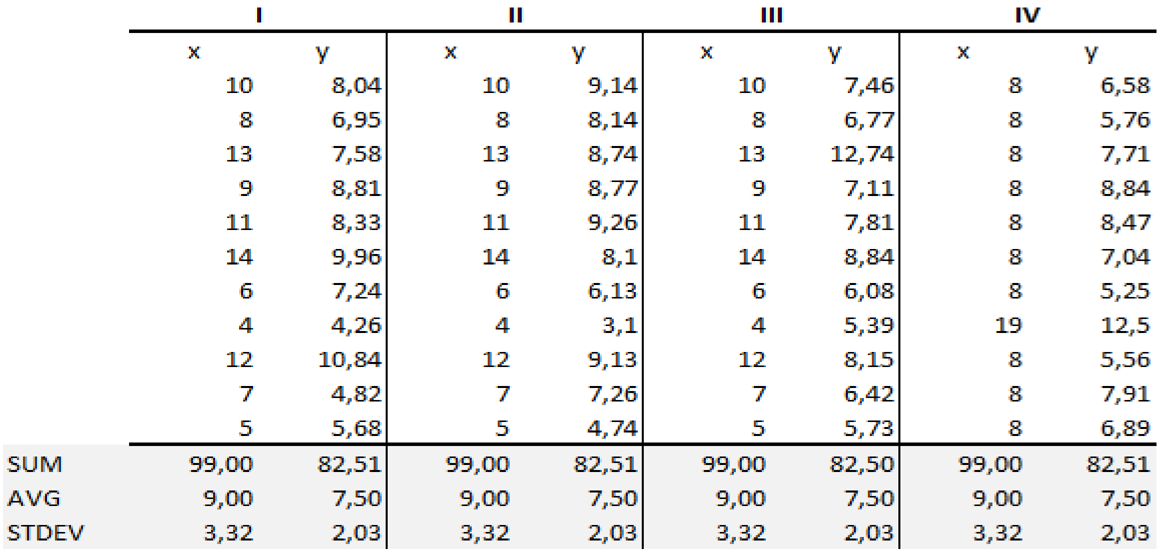
1. **R square** is called coefficient of determination.
2. Multiply R times R to get the R square value.
3. The Coefficient of Determination is the square of Coefficient of Correlation.
4. R square **lies between 0 and 1**. The percentage variation in y which is explained by all the x variables together is called r square.
5. R square **never negative** as it is a square of a number.
6. **High r square** value is considered **better**.
7. It is easy to explain the R square in terms of regression. It is not so easy to explain the R in terms of regression.

Coefficient of Correlation –

1. **R** is called coefficient of correlation.
2. It is relationship between two variables.
3. It can go between **-1 and 1**. +1 indicates perfect positive fit, -1 means that the two variables are opposite.
4. They increase and decrease together and have perfect correlation.
5. For multiple linear regression R is computed but is difficult to explain the value of r in multiple linear regression. R square is a better term.
6. We can explain R square for both simple linear regressions also for multiple linear regressions.

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

Ans. Anscombe's quartet has 4 data sets which has almost all **identical descriptive statics**, and still has different distributions and **is graphed differently**. Each dataset consists of eleven x,y points.



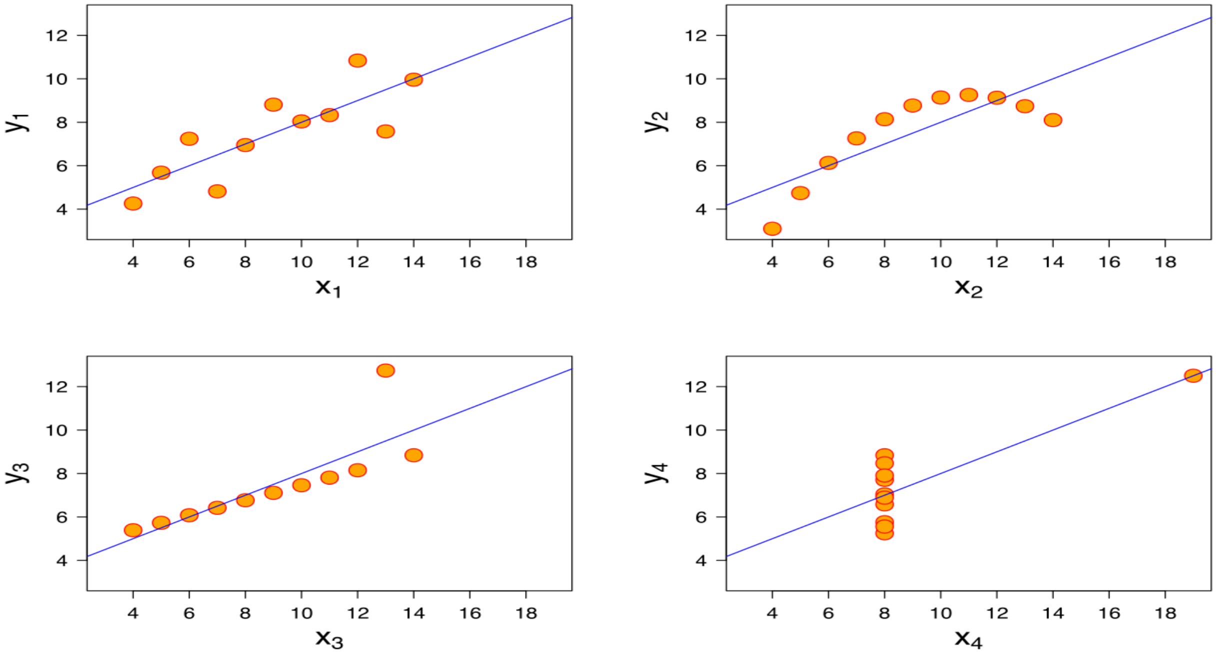
	I		II		III		IV	
	x	y	x	y	x	y	x	y
	10	8,04	10	9,14	10	7,46	8	6,58
	8	6,95	8	8,14	8	6,77	8	5,76
	13	7,58	13	8,74	13	12,74	8	7,71
	9	8,81	9	8,77	9	7,11	8	8,84
	11	8,33	11	9,26	11	7,81	8	8,47
	14	9,96	14	8,1	14	8,84	8	7,04
	6	7,24	6	6,13	6	6,08	8	5,25
	4	4,26	4	3,1	4	5,39	19	12,5
	12	10,84	12	9,13	12	8,15	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 statics shown above has **mean and variance** which is identical for x and y across the groups.

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

When we plot these, we see that they have same regression line even though the dataset is different. The following are the explanation for the graph shown below –

1. Dataset I appear to have clean and well-fitting linear models.
2. Dataset II is not distributed normally.
3. In Dataset III the distribution is linear, but the calculated regression is thrown off by an outlier.
4. Dataset IV shows that one outlier is enough to produce a high correlation coefficient.



The quarter is often used to show **the importance** of looking at a set of **data graphically** before analyzing a relationship and the basic static properties for describing the realistic datasets.

5. What is Pearson's R?

Ans. Measure of the strength of the association between two variables is called Pearson's R.
It is known as Pearson's correlation coefficient. Steps to find out Pearson's r are –

1. We need to plot the **scatter plot** to figure the linearity between the variables. The Pearson's R is can not be calculated if the relationship between variables is not linear.
2. If the scatter plots are closer **to straight line** it indicates a **higher strength**.
3. The interval range is from -**1 to 1**.
4. If
    a. R = 1 -> perfect straight line with negative slope.
    b. R = 0 -> no linear relationship between variables.

    c. R = -1 -> perfect straight line with negative slope.

5. **Positive** correlation indicates that both variables **increases or decreases at same time**, whereas **negative** correlation showcases that **if one variable increase then other decreases** and vice versa.
6. **T-test** is used to find out the whether there is a coefficient of correlation is different from zero in order to find out the association between the variables.
7. **Example:** To find out if there is any relation between is normal or hyperventilating breath given dataset of how long a student can hold the breath.

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

Ans. Scaling is a technique **to standardize the independent features** present in the data in a fixed range. It is a part of data pre-processing.

In a dataset different features have **different units,** in order to **standardize** all the data so it can be used for further calculations in the model, scaling is performed.

The difference between Standardize scaling and Normalize scaling is as following –

Standardize - Standardize replaces values by their Z score. Standardization transforms data to mean od zero and standard deviation of 1. The formula for which is

$$z_i = \frac{x_i - \bar{x}}{s}$$

This redistributes the features with their mean μ = 0 and standard deviation σ =1**.**

Normalizing – The distribution will have values between -1 and 1 with mean μ = 0. Normalization means scale a variable to have values between 0 and 1.

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

The terms normalization and standardization are sometimes used interchangeably but are different.

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

Ans. VIF is calculated to find out the collinearity between the variables he reason for infinite VIF value indicates that **the corresponding variable** may be expressed by a **linear combination** of other variables.

8. What is the Gauss-Markov theorem?

Ans. The GMT states that, if the error in the LR model are not correlated and have equal variance and expectation value zero, the OLS will have lowest sampling variance within the class of linear unbiased estimator. In other words GMT states that **if LR model** satisfied the **first six assumptions**, then **OLS regression** produces unbiased estimates that have the **smallest variance** of all possible linear estimator.

Advantages of GMT –
1.GMT helps us test linear dependency.
2.GMT helps solving matrix equation.
3. GMT helps us determine the reason for null values in matrix.
4.GMT also helps in generating a solution set of general linear system.

The errors need not be normal nor do they need to be independent and identical

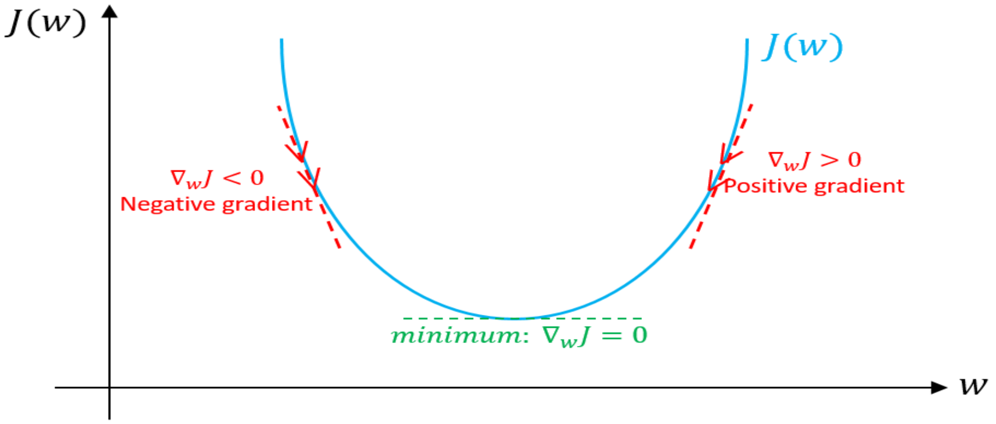
9. Explain the gradient descent algorithm in detail.

Ans. Gradient Descent is the most common **optimization algorithm** in *machine learning* and *deep learning*. It is a first-order optimization algorithm. This means it only takes into account the first derivative when performing the updates on the parameters. On each iteration, we update the parameters in the opposite direction of the gradient of the objective function *J(w)* w.r.t the parameters where the gradient gives the direction of the steepest ascent. The size of the step we take on each iteration to reach the local minimum is determined by the learning rate α. Therefore, we follow the direction of the slope downhill until we reach a local minimum.

Procedure –

1. Initialize weight *w* and bias *b* to any random numbers.
2. Pick a value for the learning rate α. The learning rate determines how big the step would be.
   If α is very small, it would take long time to converge and become computationally expensive.
   If α is large, it may fail to converge and overshoot the minimum.

Therefore, plot the cost function against different values of α and pick the value of α that is right before the first value that didn't converge so that we would have a very fast learning algorithm that converges. Below is the graph of Gradient descent and how algorithm uses the derivation of the loss to the follow downhill its minimum.



There are two main types of Gradient descent –
1.**Batch Gradient Descent** – Batch gradient descent refers to calculating the derivative from all the training data before calculating an update.

2.**Stochastic –** Stochastic gradient descent refers to calculating the derivative from each training data instance and calculating the update immediately.

In the end we can conclude that **optimization is essential** part of machine learning and Gradient descent is a simple optimization procedure that you can use with many machine learning algorithms.

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

Ans. In statistics, a Q–Q plot is a probability plot, which is a graphical method for comparing two probability distributions by plotting their quantiles against each other. If the data is **normally** distributed, the points in the **QQ**-**normal plot** lie on a straight diagonal line. We can add this line to our **QQ plot** with the command qq line(x) , where x is the vector of values. The deviations from the straight line are minimal. This indicates **normal** distribution.

Importance –
1. 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.
2. Also, it helps to determine if two data sets come from populations with a **common distribution**.
3. **Example:** A 45 degree angle is plotted on the Q Q plot, if the two data sets come from a common distribution, the points will fall on that reference line.

Other Advantages of QQ plot are as following –

1. It can be used with sample sizes also
2. 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.

QQ Plot in LR - 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.