1. Explain the linear regression algorithm in detail.

The purpose of the regression is to examine two things. Firstly, it determines whether a set of predictor variables perform well in predicting an outcome variable. Secondly, the linear regression algorithm determines the variables that are significant predictors of the outcome variable and in what way do they impact the outcome variable; this is indicated by the magnitude and sign of the beta estimations. These estimations explain the relationship between one dependent variable and one or more independent variables.

In simple terms , Linear regression is finding the best linear relationship between the independent and dependent variables. It is mostly done by the Sum of Squared Residuals Method. The objective is to obtain a line that best fits the data. The best fit line is the one for which total prediction error are as small as possible. Error is the distance between the point to the regression line.

Suppose Y is a dependent variable, and X are independent variables . The regression equation would be:

$$\hat{Y}_t = b_0 + b_1 X_{1t} + b_2 X_{2t} + \ldots + b_k X_{kt}$$

Where bo is the intercept and b1,b2 … bk are the coefficients of the features.

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

Linear regression consists of 5 key assumptions:

- linear relationships: We assume theres a linear relation between dependent and independent variables.
- Homoscedasticity: the data should be homoscedastic and heteroscedasticity is not present in the data. A random variable is said to be heteroscedastic when different subpopulations have different variabilities (standard deviation).
- No auto-correlation: Autocorrelation occurs when the residuals are not independent from each other. In other words when the value of y(x+1) is not independent from the value of y(x).
- No multicollinearity : The independent variables don't have linear relationships between them.
- Multivariate normality : All variables should be multivariate normal. A vector is said to be k-variate normally distributed if every linear combination of its k components has a univariate normal distribution. It can be checked through Q-Q plot.

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

The coefficient of determination, "r^2", is the ratio of the explained variation to the total variation. It represents the percent of the data that is the closest to the line of best fit; it explains all the variations. The coefficient of correlation, the "R" value, is a statistical measure of the degree that changes to the value of one variable predict change to the value of the other. In the positive

correlated variables, the value increases or decreases in together, whereas in the negative correlated variables, the value of one increases as the value of the other decreases.

4. Explain the Anscombe's quartet in detail.

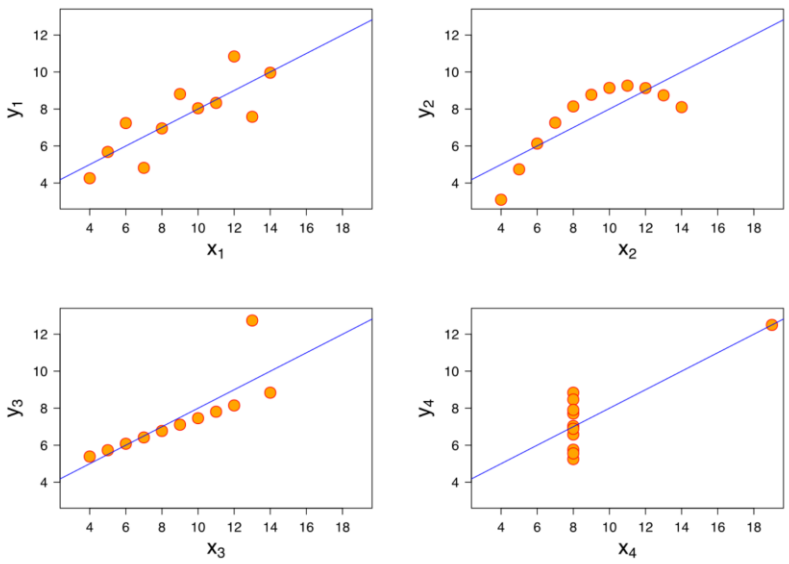
Anscombe's Quartet was developed by statistician Francis Anscombe. It comprises four datasets, each containing eleven (x,y) pairs. The essential thing to note about these datasets is that they share the same descriptive statistics. But things change completely, when they are graphed. Each graph tells 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

Quartet's Summary Stats

- Mean of x is 9 and mean of y is 7.50 for each dataset.
- 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.

Statistics wise these looks same , but when we plot it, it changes completely.



5. What is Pearson's R?

The Pearson's R is a measure of the strength of the linear relationship between two variables. It ranges from -1 to 1; the R of -1 designates a perfect negative linear relationship between variables, an R of 0 designates no linear relationship between variables, and a r of 1 designates a perfect positive linear relationship between the variables.

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

Scaling is a method used to formalize the range of independent variables or features of data. Although scaling is not mandatory, it helps handle disparities in units and helps reduce computational expenses during long processes. This method helps improve the performance of the model and reduces the values from varying widely.

- Normalized scaling rescales the value into a range of [0,1]. This is a good technique to use when the distribution of the data is unknown or when the distribution is not Gaussian.
- Standardized scaling rescales data to have a mean of 0 and a standard deviation of 1. This scaling assumes that the data has a Gaussian distribution, however, this does not have to be true but the technique is more effective if the attribute distribution is Gaussian.

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

If any feature have a perfect fit against other variables , ie R2 = 1, the VIF goes to infinite. Below is the formula for VIF.

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

If R2 = 0 , the eqn would be 1/0 and hence it will go to infinite.

8. What is the Gauss-Markov theorem?

The Gauss–Markov theorem states that in a linear regression model in which the errors are uncorrelated, have equal variances and expectation value of zero, the best linear unbiased estimator (BLUE) of the coefficients is given by the ordinary least squares (OLS) estimator, provided it exists.

The Gauss–Markov assumptions concern the set of error random variables.

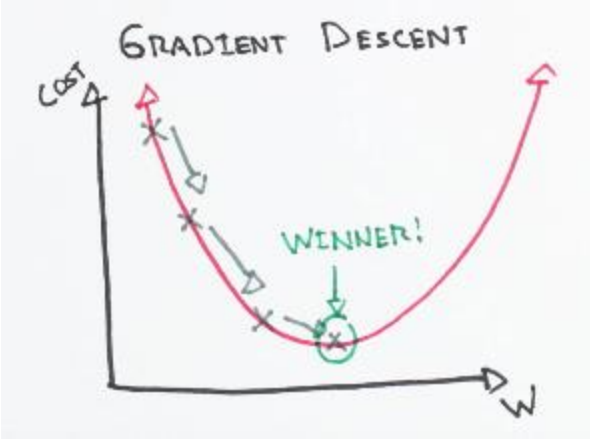
- They have mean zero
- They are homoscedastic, that is all have the same finite variance
- Distinct error terms are uncorrelated

9. Explain the gradient descent algorithm in detail.

Gradient descent is an optimization algorithm used to minimize a function by iteratively moving in the direction of steepest descent as defined by the negative of the gradient. In linear regression, We use gradient descent to update the coefficients of our model.

Below is an example of how GD works. It takes steps in downhill direction given by negative gradient. After each step we calculate the gradient again and move towards the negative side.We continue this process iteratively until we get to the bottom of our graph, or to a point where we can no longer move downhill–a local minimum.

The size of these steps are called learning rates and A Loss Functions tells us “how good” our model is at making predictions for a given set of parameters. The cost function has its own curve and its own gradients. The slope of this curve tells us how to update our parameters to make the model more accurate.



There are two parameters in our cost function that we control, m(weight) and b(bias). Since we need to consider the impact each one has on the final prediction, we need to use partial derivatives. We calculate the partial derivatives of the cost function with respect to each parameter and store the results in a gradient.

Given the cost function:

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

The gradient can be calculated as:

$$f'(m,b) = \begin{bmatrix} \frac{df}{dm} \\ \frac{df}{db} \end{bmatrix} = \begin{bmatrix} \frac{1}{N}\sum -2x_i(y_i - (mx_i + b)) \\ \frac{1}{N}\sum -2(y_i - (mx_i + b)) \end{bmatrix}$$

To solve for the gradient, we iterate through our data points using our new m and b values and compute the partial derivatives. This new gradient tells us the slope of our cost function at our current position (current parameter values) and the direction we should move to update our parameters. The size of our update is controlled by the learning rate.

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

Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions. It is used to compute the theoretically expected value for every data point based on the distribution in question.

If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x.

In Linear Regression, Q-Q plot is used to assess if your residuals are normally distributed. They compare the distribution of your data to a normal distribution by plotting the quartiles of your data against the quartiles of a normal distribution. If your data are normally distributed then they should form an approximately straight line