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

   In simple terms, linear regression is a method of finding the best straight line fitting to the given data, i.e., 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 Residual Sum of Squares Method.

   Linear regression models can be classified into two types depending upon the number of independent variables:
   - Simple linear regression: This is used when the number of independent variables is 1.
   - Multiple linear regression: This is used when the number of independent variables is more than 1.

   The equation of the best fit regression line $\mathbf{Y = \beta_0 + \beta_1 X}$ can be found by minimizing the cost function (RSS in this case, using the ordinary least squares method), which is done using the following two methods:
   - Differentiation
   - Gradient descent

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

   Below are the Assumptions about the residuals:
   - Normality assumption: It is assumed that the error terms, ε(i), are normally distributed.
   - Zero mean assumption: It is assumed that the residuals have a mean value of zero, i.e., the error terms are normally distributed around zero.
   - Constant variance assumption: It is assumed that the residual terms have the same (but unknown) variance, σ2 . This assumption is also known as the assumption of homogeneity or homoscedasticity.
   - Independent error assumption: It is assumed that the residual terms are independent of each other, i.e., their pair-wise covariance is zero

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

   Coefficient of correlation is “R” value which is given in the summary table in the Regression output. R square is also called coefficient of determination. Multiply R times R to get the R square value. In other words Coefficient of Determination is the square of Coefficient of Correlation.

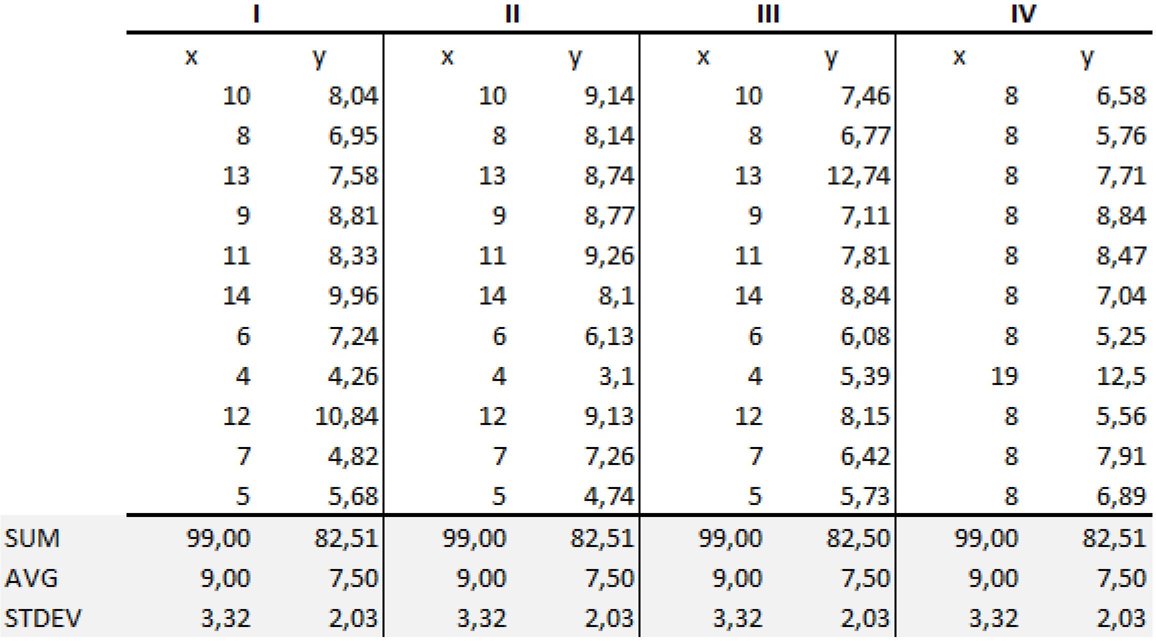
   R square or coeff. of determination shows percentage variation in y which is explained by all the x variables together. Higher the better. It is always between 0 and 1. It can never be negative – since it is a squared value.

   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: is the degree of relationship between two variables say x and y. It can go between -1 and 1. 1 indicates that the two variables are moving in unison. They rise and fall together and have perfect correlation. -1 means that the two variables are in perfect opposites. One goes up and other goes down, in perfect negative way. Any two variables in this universe can be argued to have a correlation value. If they are not correlated then the correlation value can still be computed which would be 0. The correlation value always lies between -1 and 1 (going thru 0 – which means no correlation at all – perfectly not related). Correlation can be rightfully explained for simple linear regression – because you only have one x and one y variable. For multiple linear regression R is computed, but then it is difficult to explain because we have multiple variables involved here. That's why R square is a better term. You can explain R square for both simple linear regressions and also for multiple linear regressions.

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, and I must emphasize COMPLETELY, when they are graphed. Each graph tells a different story irrespective of their similar summary statistics.



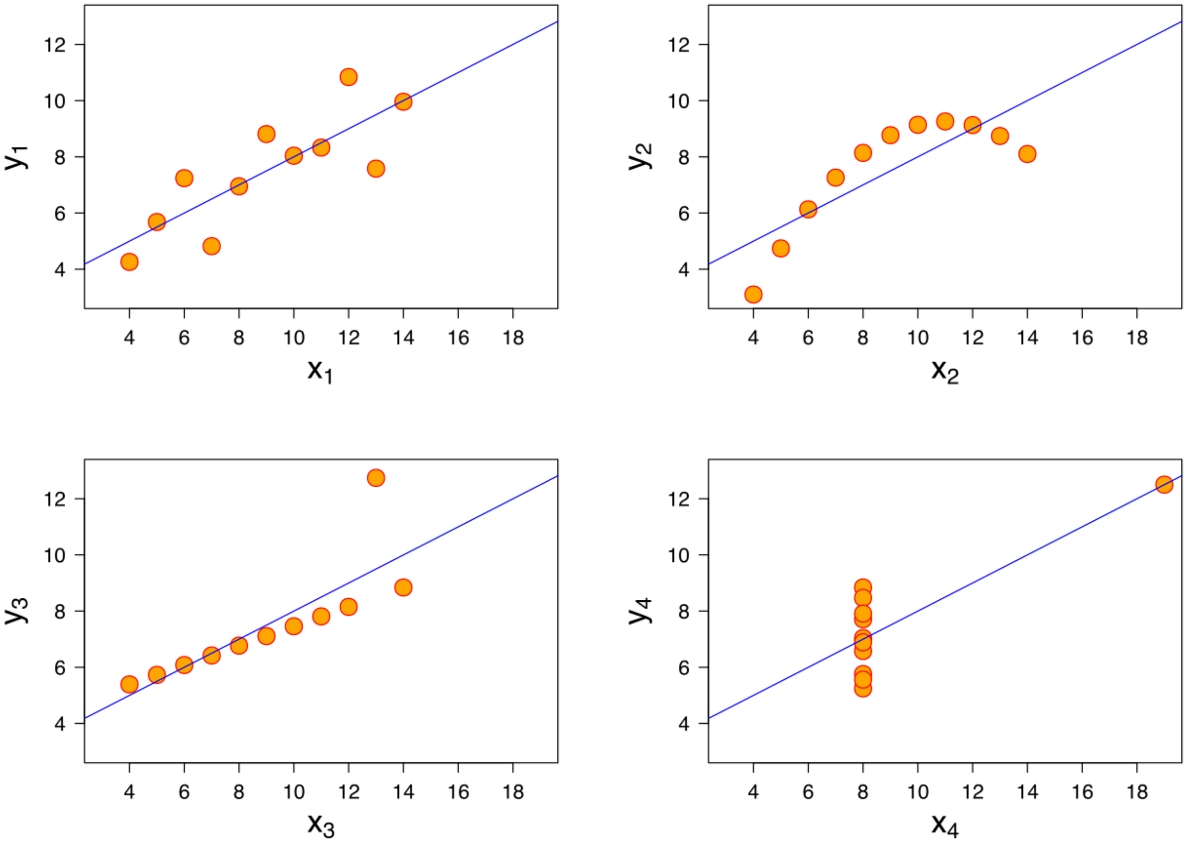
	I		II		III		IV	
	x	y	x	y	x	y	x	y
	10	8,04	10	9,14	10	7,46	8	6,58
	8	6,95	8	8,14	8	6,77	8	5,76
	13	7,58	13	8,74	13	12,74	8	7,71
	9	8,81	9	8,77	9	7,11	8	8,84
	11	8,33	11	9,26	11	7,81	8	8,47
	14	9,96	14	8,1	14	8,84	8	7,04
	6	7,24	6	6,13	6	6,08	8	5,25
	4	4,26	4	3,1	4	5,39	19	12,5
	12	10,84	12	9,13	12	8,15	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

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 (how strong a relationship is between two variables) between x and y is 0.816 for each dataset

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 appears 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?

In statistics, the Pearson correlation coefficient, also referred to as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC) or the bivariate correlation,is a measure of the linear correlation between two variables X and Y. According to the Cauchy–Schwarz inequality 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.

For a population

Pearson's correlation coefficient when applied to a population is commonly represented by the Greek letter ρ (rho) and may be referred to as the population correlation coefficient or the population Pearson correlation coefficient. Given a pair of random variables {X,Y}, the formula for ρ is:

$\rho\ X,Y = cov(X,Y)/\ \sigma X\ \sigma Y$

where:

- cov is the covariance
- σX is the standard deviation of X
- σY is the standard deviation of Y

The formula for ρ can be expressed in terms of mean and expectation. Since

$cov(X,Y) = E[(X - \mu x)(Y - \mu y)]$

$\rho\ X,Y = E[(X - \mu x)(Y - \mu y)]/\sigma X\ \sigma Y$

where:

- σY and σX are defined as above
- μx is the mean of X
- μy is the mean of Y
- E is the expectation

For a sample

Pearson's correlation coefficient when applied to a sample is commonly represented by r(xy) and may be referred to as the sample correlation coefficient or the sample Pearson correlation coefficient

$$r_{xy} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\ \sqrt{n\sum y_i^2 - (\sum y_i)^2}}.$$

where:

- N – sample size ,Xi,Yi – sample points indexed with i
- This formula suggests a convenient single-pass algorithm for calculating sample correlations, but, depending on the numbers involved, it can sometimes be numerically unstable.

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

Feature scaling is a method used to normalize the range of independent variables or features of data. In data processing, it is also known as data normalization and is generally performed during the data preprocessing step.

It is a step of Data Pre Processing which is applied to independent variables or features of data. It basically helps to normalise the data within a particular range. Sometimes, it also helps in speeding up the calculations in an algorithm.

Normalization typically means rescales the values into a range of [0,1]. Standardization typically means rescales data to have a mean of 0 and a standard deviation of 1 (unit variance).

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

The VIF gives how much the variance of the coefficient estimate is being inflated by collinearity. If the VIF for a variable is 16 then associated standard error is four times as large as it would be if its VIF was 1. In such a case, the coefficient would have to be 4 times as large to be statistically significant at a given significance the level.
The VIF can be conceived as related to the R-squared of a particular predictor variable regressed on all other includes predictor variables:
VIF of X1 = 1/(1 - R-squared of X1 on all other Xs).
If you only have 1 X or that X is orthogonal with all the other Xs; then
VIF = 1/(1-0) = 1 - so no variance inflation
If two Xs are perfectly correlated
VIF = 1/(1-1)= 1/0 = infinity that is the estimate is as imprecise as it can be.

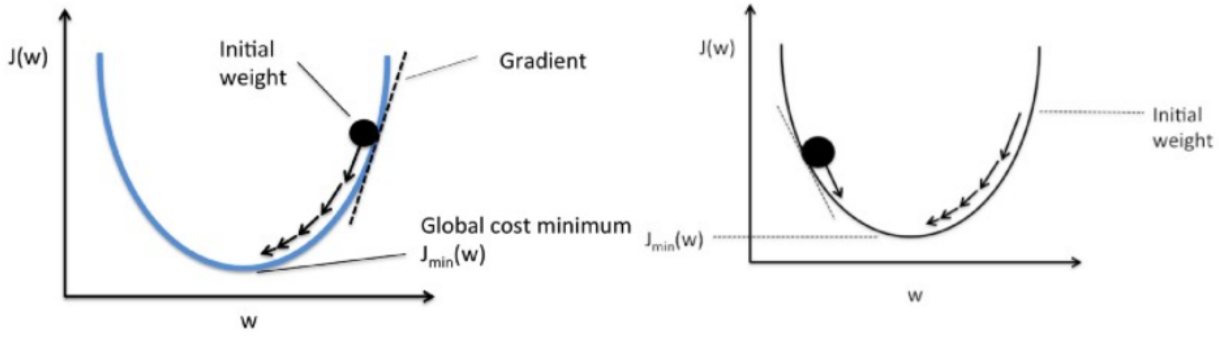
8. What is the Gauss-Markov theorem?

In statistics, 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.

9. Explain the gradient descent algorithm in detail.

Gradient descent is an optimisation algorithm. In linear regression, it is used to optimise the cost function and find the values of the βs (estimators) corresponding to the optimised value of the cost function.

Gradient descent works like a ball rolling down a graph (ignoring the inertia). The ball moves along the direction of the greatest gradient and comes to rest at the flat surface (minima).



Gradient Descent

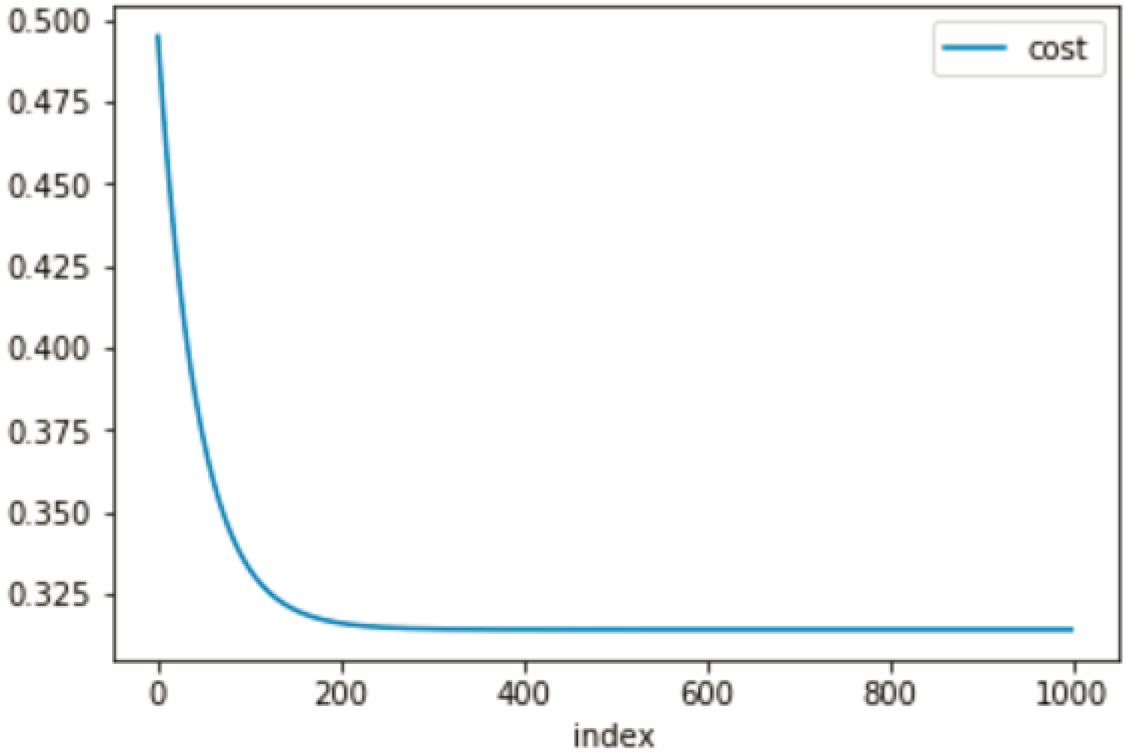
Mathematically, the aim of gradient descent for linear regression is to find the solution of

ArgMin J(Θ0,Θ1), where J(Θ0,Θ1) is the cost function of the linear regression. It is given by:

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

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

Gradient descent starts with a random solution, and then, based on the direction of the gradient, the solution is updated to the new value, where the cost function has a lower value.



Gradient Descent in Python

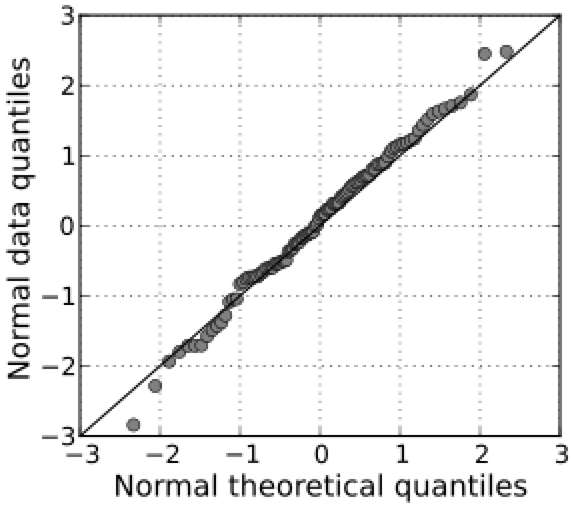
In the graph above, we can observe that after 200 iterations the cost function is getting flattened. Thus, we can get the global minimum before it completes 200 iterations.

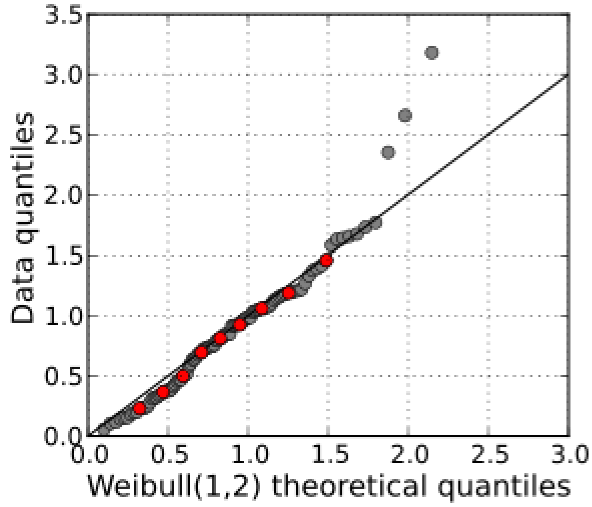
The learning rate and the number of iteration have a direct effect on the shape of the graph you get. Note that the graph may vary from the above shown as we change the learning rate and the number of iterations.

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

A Q–Q plot is a plot of the quantiles of two distributions against each other, or a plot based on estimates of the quantiles. The pattern of points in the plot is used to compare the two distributions.

The main step in constructing a Q–Q plot is calculating or estimating the quantiles to be plotted. If one or both of the axes in a Q–Q plot is based on a theoretical distribution with a continuous cumulative distribution function (CDF), all quantiles are uniquely defined and can be obtained by inverting the CDF. If a theoretical probability distribution with a discontinuous CDF is one of the two distributions being compared, some of the quantiles may not be defined, so an interpolated quantile may be plotted. If the Q–Q plot is based on data, there are multiple quantile estimators in use. Rules for forming Q–Q plots when quantiles must be estimated or interpolated are called plotting positions.





A simple case is where one has two data sets of the same size. In that case, to make the Q–Q plot, one orders each set in increasing order, then pairs off and plots the corresponding values. A more complicated construction is the case where two data sets of different sizes are being compared. To construct the Q–Q plot in this case, it is necessary to use an interpolated quantile estimate so that quantiles corresponding to the same underlying probability can be constructed.

More abstractly, given two cumulative probability distribution functions F and G, with associated quantile functions F' and G' (the inverse function of the CDF is the quantile function), the Q–Q plot draws the q-th quantile of F against the q-th quantile of G for a range of values of q. Thus, the Q–Q plot is a parametric curve indexed over [0,1] with values in the real plane R2.