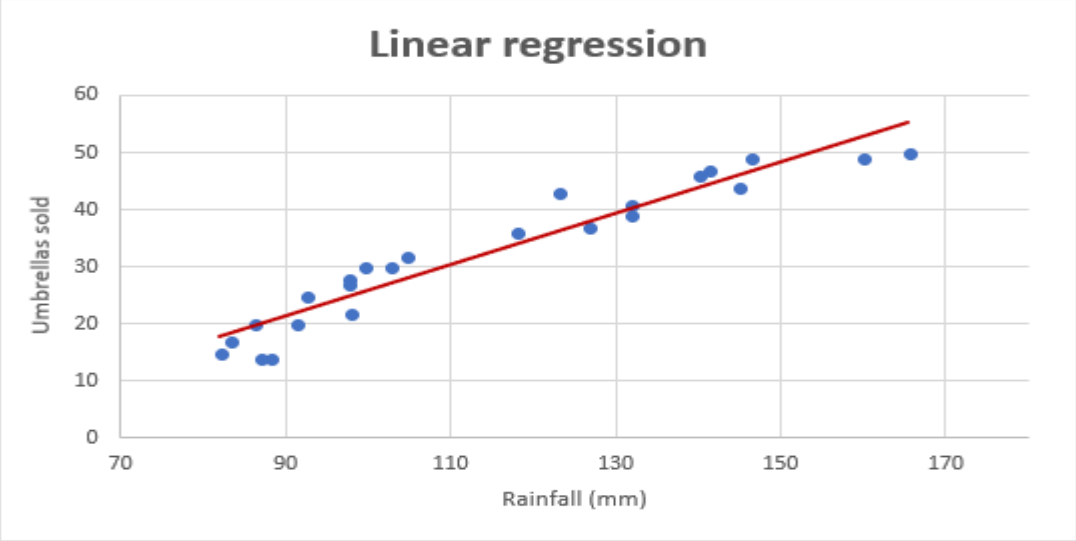
1. Explain the linear regression algorithm in detail.

Linear Regression is a machine learning algorithm and comes under the category of **supervised learning**. It performs a **regression task**. Regression is a method of modelling a target value based on independent predictors. Linear regression is a type of regression analysis where there is a linear relationship between the independent(x) and one or more dependent(y) variables.

For example: there is a linear relationship between rainfall (independent variable) and the number of umbrellas sold (dependent variable) as shown below:



In the figure above, x (input) is the rainfall and y (output) is the number of umbrellas sold. The regression line is the best fit line for our model, which is given by:

$$y = a+bx$$

Here, y is the dependent variable, x is the independent or predictor variable, a is the intercept and b is the slope.

Linear Regression is of two types:

Simple Linear Regression: When we have a single input variable (x) in linear regression, this is called simple linear regression.

For example: prediction of number of umbrellas sold using rainfall as the input variable, as explained above, is a simple linear regression.

Multiple Linear Regression: It is the extension of simple linear regression that predicts a response using two or more independent variables. **For example**, multiple linear regression can be used to understand whether exam score can be predicted based on revision time, lecture attendance and gender.

It is given by:

$$y = a+b_1x_1+b_2x_2+\ldots+b_px_p+e$$

where $x_1, x_2, \ldots, x_p$ are the predictor variables and e is the error term.

2. What are the assumptions of linear regression regarding residuals?
    1) **Linearity**: There is a **linear relationship** between parameters. A linear relationship suggests that a change in response Y due to one unit change in X^1 is constant, regardless of the value of X^1.
    Example ***Y = α + (β1*X1) + (β2*X2^2).***
    Though, the X2 is raised to power 2, the equation is still linear in beta parameters. So the assumption is satisfied in this case.
    2) **No Autocorrelation**: Error terms are independent of each other. In other words, there should be no correlation between the residual (error) terms. Absence of this phenomenon is known as Autocorrelation.
    3) The mean of residuals (error) terms is zero.
    4) The error terms must have constant variance. This phenomenon is known as **homoscedasticity**. The presence of non-constant variance is referred to heteroscedasticity.
    5) The error terms must be normally distributed.

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

Coefficient of correlation is a statistical measure that calculates the strength and direction of relationship between two variables. The value ranges between -1 to +1 and it is interpreted as follows:

- 1 indicates a strong positive relationship.
- -1 indicates a strong negative relationship.
- A result of zero indicates no relationship at all.

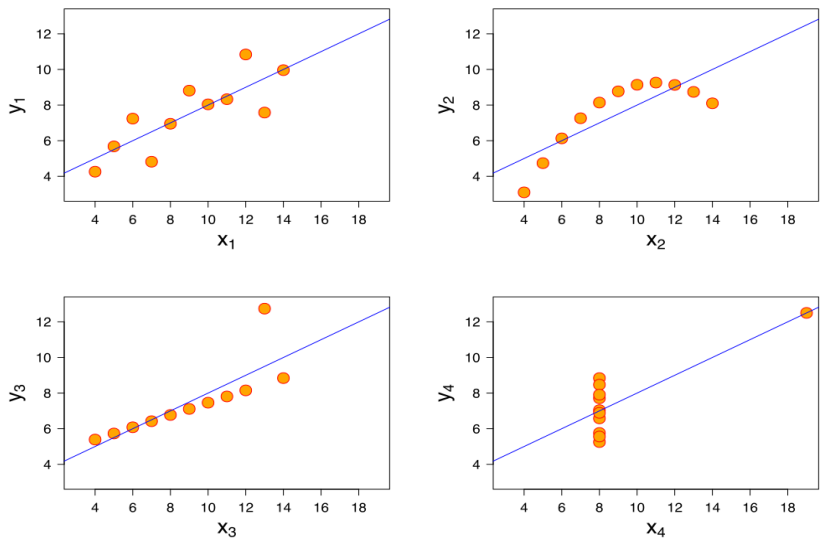
For example: height and body weight have a positive correlation.

Coefficient of determination, R^2 is a statistical measure that assesses the ability of a model to predict or explain an outcome in the linear regression. R^2 indicates the proportion of the variance in the dependent variable (Y) that is explained by linear regression and the independent variable. It ranges between value of 0 to 1, and a high R^2 value indicates that the model is a good fit for the data. It is the square of the coefficient of correlation.

For example there is a strong linear relationship between the number of stories a building has and its height. R^2 will be high in such case but not 100%.

4. Explain the Anscombe's quartet in detail.

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. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it and the effect of outliers and other influential observations on statistical properties.



All four of these data sets have the same variance in x, variance in y, mean of x, mean of y, and linear regression. But as it can be seen from the graphs, they are quite different from one another.

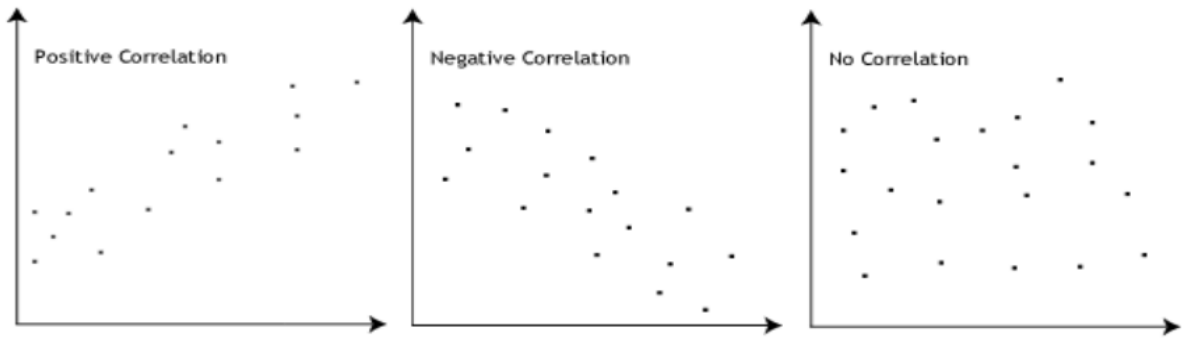
- Dataset I appears to have clean and well-fitting linear models.
- Dataset II is not distributed normally.
- In Dataset III the distribution is linear, but the calculated regression is thrown off by an outlier.
- Dataset IV shows that one outlier is enough to produce a high correlation coefficient.

This shows the **importance of visualization of data**. Statistics are great for describing general trends and aspects of data, but statistics alone can't fully depict any data set.

5. What is Pearson's R?

Pearson's R is the **Pearson product-moment correlation coefficient**, which is a measure of the strength of a linear association between two continuous variables. It gives information about the magnitude, or correlation, as well as the direction of the relationship. Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations.

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.



Examples:

- Height and weight of a person are positively correlated.
- Temperature and the cost of air conditioning are negatively correlated, as the temperature increases the cost of air conditioning decreases.
- Height and eye color of a person have no correlation.

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

   Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing.

   Why feature scaling?
   Many times the data is collected on different scales.
   For example, the age of employees in a company may be between 21-70 years, the size of the house they live is 500-5000 Sq feet and their salaries may range from 30000-80000. In this situation if you use a simple Multi linear regression model, the age feature will not play any role because it is several order smaller than other features. However, it may contain some important information that may be useful for the task. Here, we may **want to normalize the features** independently to the same scale, say [0,1], so they contribute equally while computing the distance.
   There are two most important scaling techniques - **Standardization and Normalization**.

   Normalization: ***Xchanged = X – Xmin / Xmax – Xmin***
   Normalization is actually used where we want to **scale down the population to [0, 1**].

   Standardization: ***Xnew = X - mean / sigma***

   With standardization we **can transform the data into the range** such that the new population has mean (average) = 0 and standard deviation = 1.

   Another difference between these is that, in normalized scaling, we change the range of the data while in standardized scaling we change the shape of the distribution of our data.

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

   If there is perfect correlation between variables, i.e. every change in the x variable is accompanied by the corresponding change in the y variable, then **VIF = infinity**.

   A perfect positive correlation is represented by the correlation coefficient value +1.0 and -1.0 indicates a perfect negative correlation, which would lead to infinite VIF value. It is uncommon to observe this type of relationship in actual data.

For example there is perfect linear relationship between temperatures in Celsius and temperatures in Fahrenheit. Correlation coefficient = 1 in this case and vif = Infinity.

8. What is the Gauss-Markov theorem?
   The Gauss-Markov theorem states that if your linear regression model satisfies the set of classical assumptions, then ordinary least squares (OLS) regression produces unbiased estimates that have the smallest variance of all possible linear estimators.

   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.

   The **Gauss Markov assumptions** guarantee the validity of ordinary least squares for estimating regression coefficients.

   Checking how well our data matches these assumptions is an important part of estimating regression coefficients. When we know where these conditions are violated, you may be able to plan ways to change your experiment setup to help your situation fit the ideal Gauss Markov situation more closely.

9. Explain the gradient descent algorithm in detail.

   Gradient descent is an optimization algorithm used to find the values of parameters (coefficients) of a function (f) that minimizes a cost function (cost). 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.

   Now there are many types of gradient descent algorithms. They can be classified by two methods mainly:

   - On the basis of data ingestion
     1. Full Batch Gradient Descent Algorithm

2. Stochastic Gradient Descent Algorithm

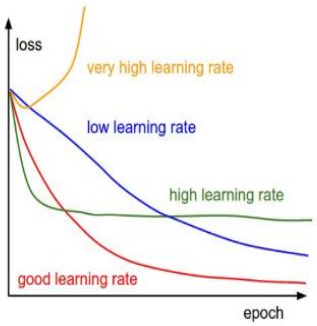
In full batch gradient descent algorithms, you use whole data at once to compute the gradient, whereas in stochastic you take a sample while computing the gradient.

Step by step to understand the Gradient Descent algorithm:

Step 1: Initialize the weights (a & b) with random values and calculate Error (SSE)

Step 2. Pick a value for the learning rate α. The learning rate determines how big the step would be on each iteration.

- 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.

- The most commonly used rates are: 0.001, 0.003, 0.01, 0.03, 0.1, 0.3.

Step 3: Calculate the gradient i.e. change in SSE when the weights (a & b) are changed by a very small value from their original randomly initialized value. This helps us move the values of a & b in the direction in which SSE is minimized.

Step 4: Adjust the weights with the gradients to reach the optimal values where SSE is minimized

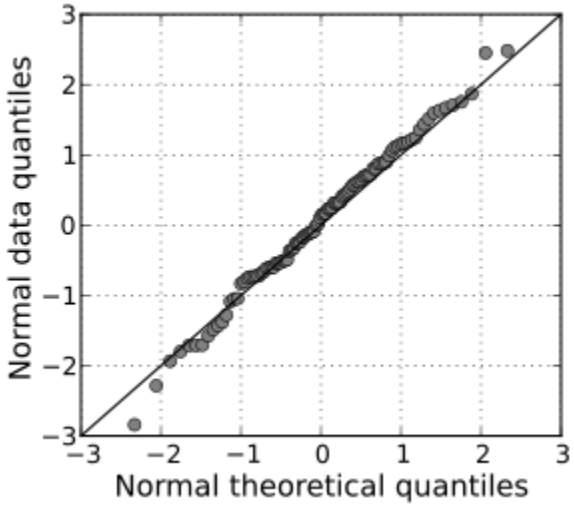
Step 5: Use the new weights for prediction and to calculate the new SSE

Step 6: Repeat steps 3 and 4 till further adjustments to weights doesn't significantly reduce the Error

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

In statistics, a **Q–Q (quantile-quantile) plot** are plots of two quantiles against each other. Q-Q plot will plot on the x-axis the quantiles of one variable and on the y-axis the quantiles of the other variable. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it.

A Q-Q plot is a way to determine whether a dataset matches a specified probability distribution. QQ-plots are often used to determine whether a dataset is normally distributed.



If the elements of a dataset perfectly match the specified probability distribution, the points on the graph will form a 45 degree line.

Use of QQ plot in linear regression:

This plot shows if residuals are normally distributed. If residuals are lined well on the straight line, then they are normally distributed.