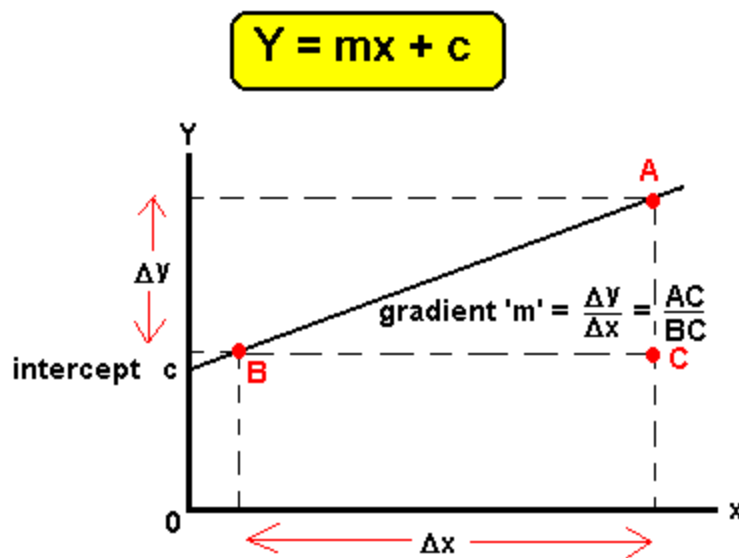


1. Explain the linear regression algorithm in detail.

Linear Regression is a machine learning algorithm based on **supervised learning**. It performs a **regression task**. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between dependent and independent variables, they are considering and the number of independent variables being used.

Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y (output). Hence, the name is Linear Regression.



This above equation is the basis of Linear Regression. The core idea is to obtain a line that best fits the data. The best fit line is the one for which total prediction error (all data points) are as small as possible. The error is the distance between the points to the regression line.

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

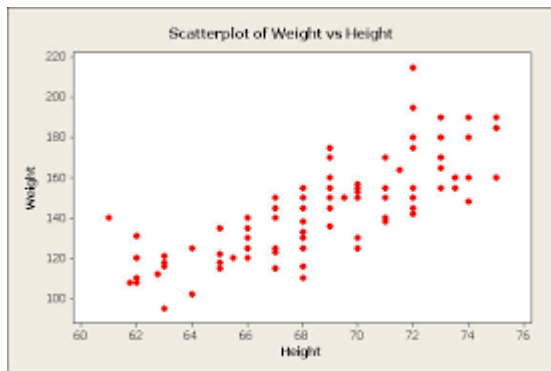
This assumption states that the variance of error terms are similar across the values of the independent variables. A plot of standardized residuals versus predicted values can show whether points are equally distributed across all values of the independent variables.

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

Correlation 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. What does that mean?

It means that both height and weight of individuals increase/decrease together (positive) and their relationship (linear) is strong. The scatter plot is something like:



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). How to interpret it?

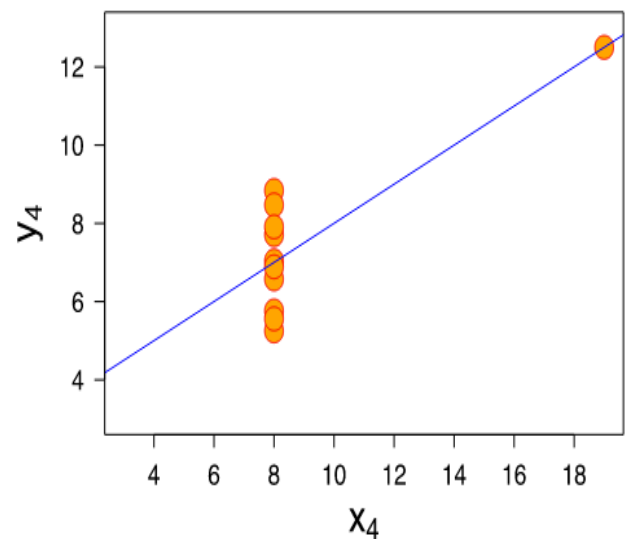
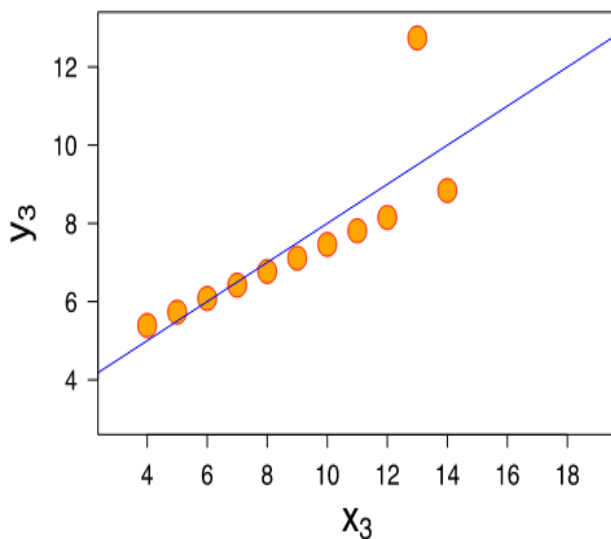
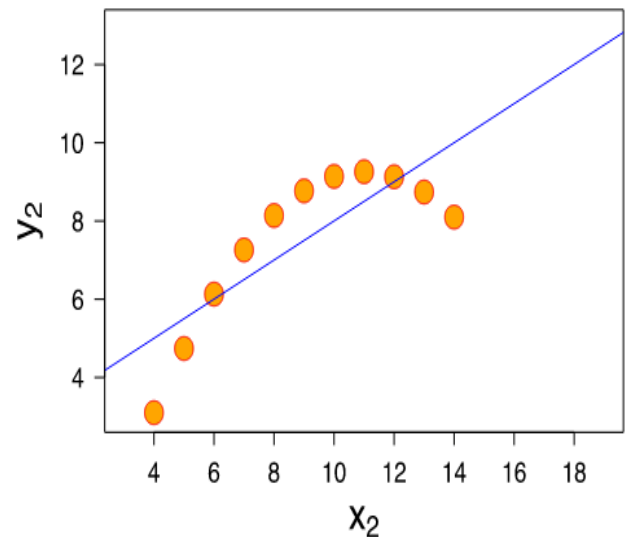
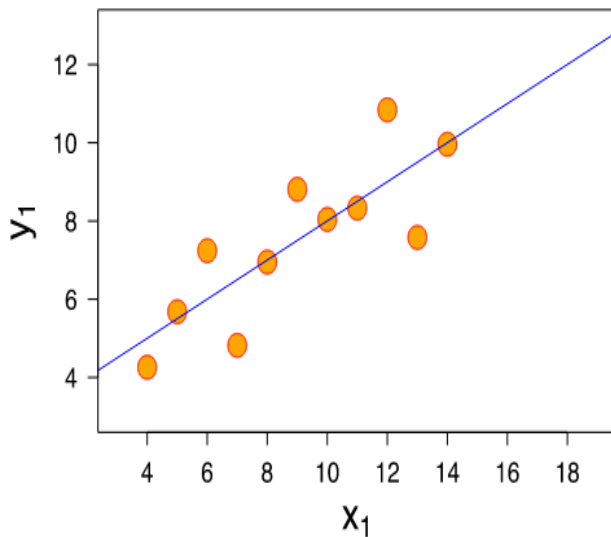
It shows that 0.64 (or 64%) of variation in height can be explained by weight. and remaining 36% of variation in height may be due to other factors which affect height of individuals like age, genetics, food intake, amount and type of exercise, location etc.

4. Explain the Anscombe's quartet in detail.

It's a group of four datasets that appear to be similar when using typical summary statistics, yet tell four different stories when graphed. Each dataset consists of eleven (x, y) pairs.

All the summary statistics you'd think to compute are close to identical:

- The average x value is 9 for each dataset
- The average y value is 7.50 for each dataset
- The variance for x is 11 and the variance for y is 4.12
- The correlation between x and y is 0.816 for each dataset
- A linear regression (line of best fit) for each dataset follows the equation $y = 0.5x + 3$.
- So far these four datasets appear to be pretty similar. But when we plot these four data sets on an x/y coordinate plane, we get the following results



Now we see the real relationships in the datasets start to emerge. Dataset I consists of a set of points that appear to follow a rough linear relationship with some variance. Dataset II fits a neat curve but doesn't follow a linear relationship (maybe it's quadratic?). Dataset III looks like a tight linear relationship between x and y , except for one large outlier. Dataset IV looks like x remains constant, except for one outlier as well.

Computing summary statistics or staring at the data wouldn't have told us any of these stories. Instead, it's important to visualize the data to get a clear picture of what's going on.

5. What is Pearson's R ?

Pearson's correlation coefficient (r) is a measure of the strength of the association between the two variables.

The first step in studying the relationship between two continuous variables is to draw a scatter plot of the variables to check for linearity. The correlation coefficient should not be calculated if the relationship is not linear. For correlation only purposes, it does not really matter on which axis the variables are plotted. However, conventionally, the independent (or explanatory) variable is plotted on the x -axis (horizontally) and the dependent (or response) variable is plotted on the y -axis (vertically).

The nearer the scatter of points is to a straight line, the higher the strength of association between the variables. Also, it does not matter what measurement units are used.

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

Feature scaling is an important technique in Machine Learning and it is one of the most important steps during the preprocessing of data before creating a machine learning model. This can make a difference between a weak machine learning model and a strong one. 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.

Why feature scaling is applied is that gradient descent converges much faster with feature scaling than without it.

Normalization Scaling: Normalization is the process of rescaling one or more attributes to the range of 0 to 1. This means that the largest value for each attribute is 1 and the smallest value is 0.

Standardized Scaling: It 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?

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. If there is perfect correlation, then $VIF = \infty$.

8. What is the Gauss-Markov theorem?

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

9. Explain the gradient descent algorithm in detail.

Gradient Descent is an optimization algorithm used for minimizing the cost function in various machine learning algorithms. It is basically used for updating the parameters of the learning model.

Types of gradient Descent:

1. **Batch Gradient Descent:** This is a type of gradient descent which processes all the training examples for each iteration of gradient descent. But if the number of training examples is large, then batch gradient descent is computationally very expensive. Hence if the number of training examples is large, then batch gradient descent is not preferred. Instead, we prefer to use stochastic gradient descent or mini-batch gradient descent.
2. **Stochastic Gradient Descent:** This is a type of gradient descent which processes 1 training example per iteration. Hence, the parameters are being updated even after one iteration in which only a single example has been processed. Hence this is quite faster than batch gradient descent. But again, when the number of training examples is large, even then it processes only one example which can be additional overhead for the system as the number of iterations will be quite large.
3. **Mini Batch gradient descent:** This is a type of gradient descent which works faster than both batch gradient descent and stochastic gradient descent. Here b examples where $b < m$ are processed per iteration. So even if the number of training examples is large, it is processed in batches of b training examples in one go. Thus, it works for larger training examples and that too with lesser number of iterations.

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

The **quantile-quantile (q-q) plot** is a graphical technique for determining if two data sets come from populations with a common distribution.

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. By a quantile, we mean the fraction (or percent) of points below the given value. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value.

A 45-degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line. The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions.

The advantages of the q-q plot are:

1. The sample sizes do not need to be equal.
2. Many distributional aspects can be simultaneously tested. For example, shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot. For example, if the two data sets come from populations whose distributions differ only by a shift in location, the points should lie along a straight line that is displaced either up or down from the 45-degree reference line.

The q-q plot is similar to a probability plot. For a probability plot, the quantiles for one of the data samples are replaced with the quantiles of a theoretical distribution.

The q-q plot is used to answer the following questions:

- Do two data sets come from populations with a common distribution?
- Do two data sets have common location and scale?
- Do two data sets have similar distributional shapes?
- Do two data sets have similar tail behavior?