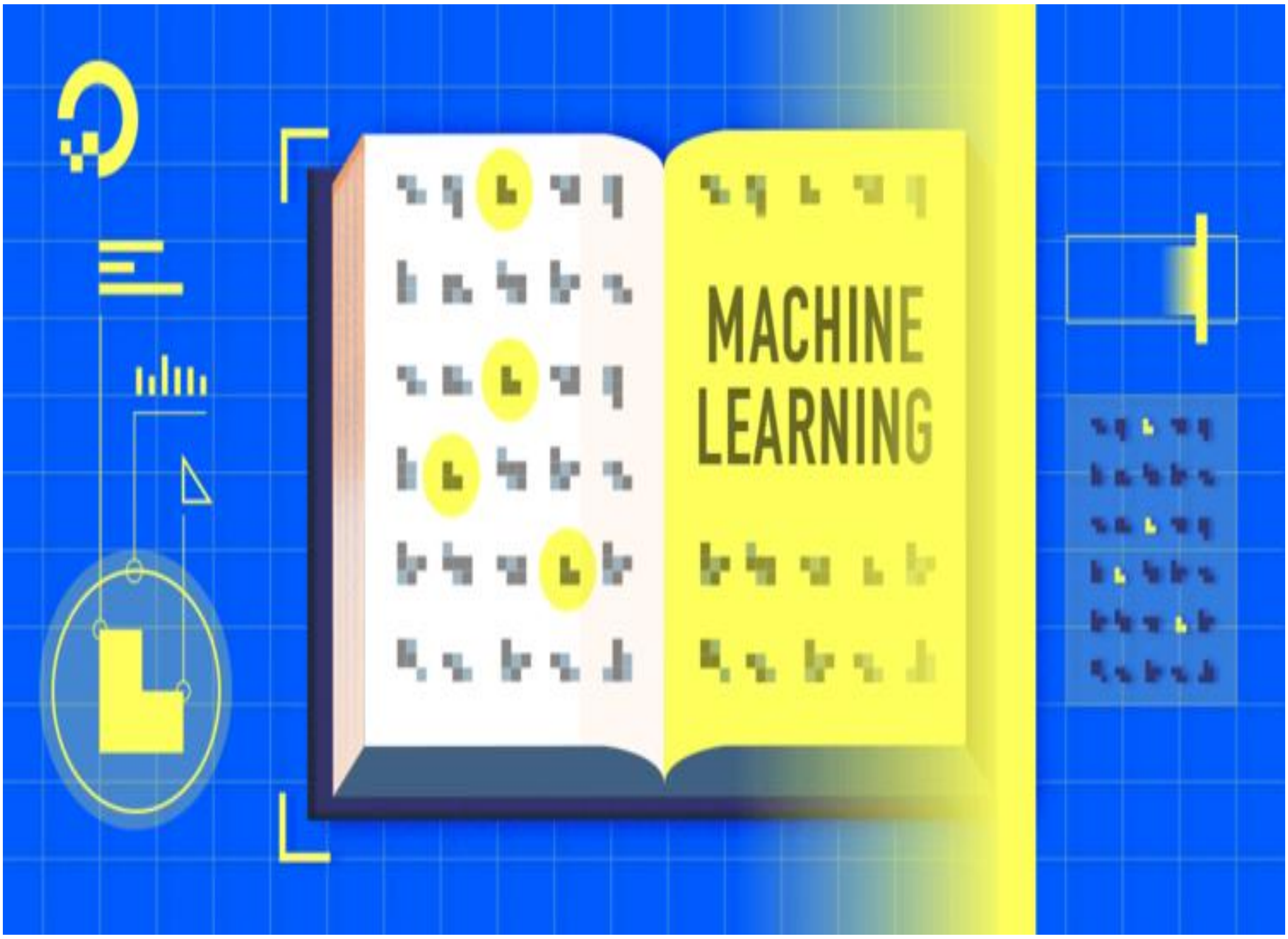
Linear Regression Assignment



Question#1 Explain the linear regression algorithm in detail.

Answer #1

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. It is a part of supervised machine learning.

Technically, linear regression is a machine learning algorithm that finds the best linear-fit relationship on any given data, between independent and dependent variables.

Linear Regression is used to predict the future results based on the existing/previous data.

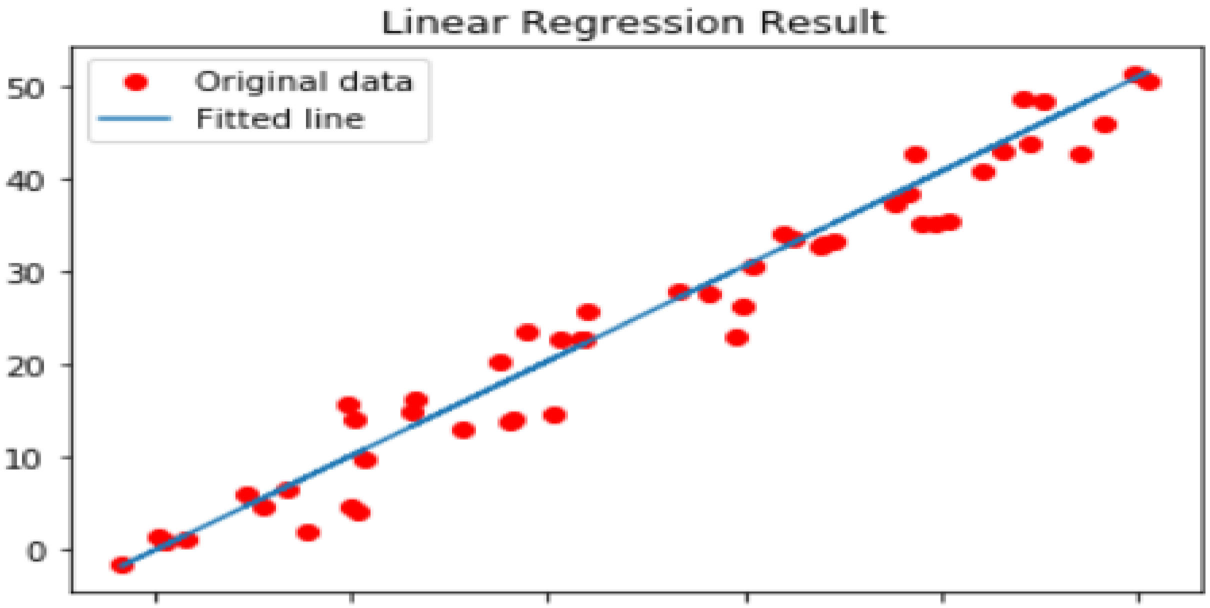
Broadly, there are two types or Linear Regressions. They are:

i. *Simple Linear Regression, also known as* ***SLR***

ii. *Multiple Linear Regression, also known as* ***MLR***

Simple Linear Regression

The most elementary type of regression model is the simple linear regression which explains the relationship between a dependent variable and one independent variable using a straight line. The straight line is plotted on the scatter plot of these two points.



Equation of Linear Regression => $Y = \beta_0 + \beta_1 X$,

Where, β_0 *denotes intercept and* β_1 *denotes slope*

Multiple Linear Regression:

Multiple linear regression is a statistical technique to understand the relationship between one dependent variable and several independent variables. The objective of multiple regression is to find a linear equation that can best determine the value of dependent variable Y for different values independent variables in X.

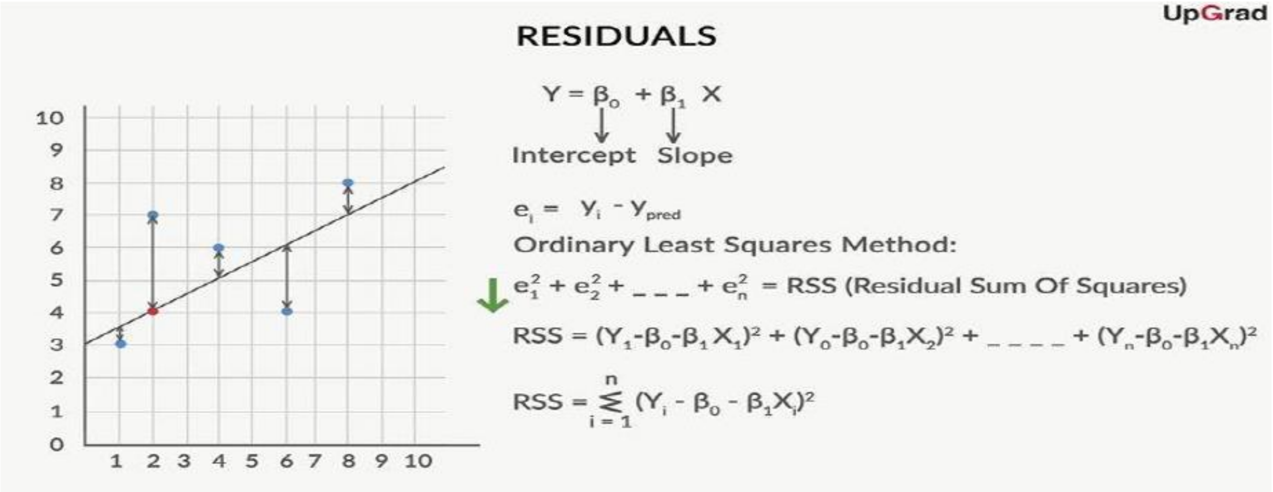
Equation of Multiple Linear Regression =>

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots\ldots\ldots \beta_p X_p + \text{€}$$

Few Important Concepts of Linear Regression:

1. **Best Fit Line**

The best fit line is a line that is obtained by plotting Scatter Plot between the independent and dependent variables, that expresses the relationship between the data points. Typically, best fit line is found by minimising the RSS(Residual Sum of Squares). Residual for any data points is calculated by subtracting the predicted value of dependent variable from actual value of dependent variable.



RESIDUALS

$$Y = \underset{\text{Intercept}}{\underset{\downarrow}{\beta_0}} + \underset{\text{Slope}}{\underset{\downarrow}{\beta_1}} X$$

$e_i = Y_i - Y_{pred}$

Ordinary Least Squares Method:

$$e_1^2 + e_2^2 + \_\,\_\,\_ + e_n^2 = \text{RSS (Residual Sum Of Squares)}$$

$$\text{RSS} = (Y_1 - \beta_0 - \beta_1 X_1)^2 + (Y_0 - \beta_0 - \beta_1 X_2)^2 + \_\,\_\,\_\,\_ + (Y_n - \beta_0 - \beta_1 X_n)^2$$

$$\text{RSS} = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

2. **R2 or Coefficient of Determination**

   R2 is used to find out what portion of the given data variance is explained by the developed model. Its value always varies from 0 to 1. Higher the values of R2, better the model fits the data.

   Mathematically, it is represented as: $R^2 = 1 - (RSS / TSS)$

   Where, RSS = Residual Sum of Square
   TSS = Total Sum of Square

3. **RSS(Residual Sum of Square)**

RSS is defined as total sum of errors, across the whole sample. A small RSS indicates a tight fit of the model.

$$RSS=\sum_{i=1}^{n}(y_i-(\alpha+\beta x_i))^2$$

4. **TSS(Total Sum of Square)**

TSS is the sum of errors from the mean of the predicted variable.

$$\mathbf{TSS}=\sum_{i=1}^{n}(y_i-\bar{y})^2$$

5. **Cost Function**

The cost function is used to figure out the best possible values for θ0 and θ1, which would provide the best fit line for the data points. As best possible values are needed for θ0 and θ1, we convert the search problem into the optimization problem, where we would like to minimize the errors between the predicted values and the actual values.

Formula to calculated cost function is given as :

$$J(\theta_0,\theta_1)=\sum_{i=1}^{N}(y_i-y_i(p))^2$$

6. **Gradient Descent**

Gradient Descent is a method of updating *θ0 and θ1 to reduce the cost function(aka, MSE). The idea is that we start with some values for θ0 and θ1, then we change these values iteratively.*

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

Question#2 What are the assumptions of Linear Regression regarding Residuals?

Answer #2

There are following four assumptions about the Residual in Linear Regression:

1. ***Normality assumption***

   it is assumed that the error terms $\varepsilon^{(i)}$*, are always normally distributed.*

2. ***Zero mean assumption***

   It is assumed that the residuals have a mean value of zero(0.0), that is the error terms are normally distributed around zero.

3. ***Constant variance assumption***

   The residual terms have the same variance, $\sigma2$ *. This assumption is also known as assumption of homogeneity or homoscedasticity.*

4. ***Independent error assumption***

   It is assumed that the residuals are independent of each other. That is the pair-wise co-variance is zero.

Question#3 What is the coefficient of correlation and the coefficient of determination?

Answer #3

Coefficient of Correlation

Coefficient of correlation in statistics, is a term, which indicates the relationship between dependent and independent variables. It is represented by "r" and it has a value ranging from -1 to 1.

When the coefficient of correlation is positive, it means, there is a positive relationship between dependent and independent variables and dependent variable increases, when there is an increase in the independent variable.

When the coefficient of correlation is negative, it means, there is a negative relationship between dependent and independent variables and dependent variable also decreases, when independent variable decreases.

If the coefficient of correlation is +0.80 or -0.80, it indicates that there is a strong relationship between dependent and independent variables. If it is +0.20 or -0.20, then there is a weak relationship between the variables. While coefficient value of 0.0 indicates, there is no relation.

Coefficient of Determination

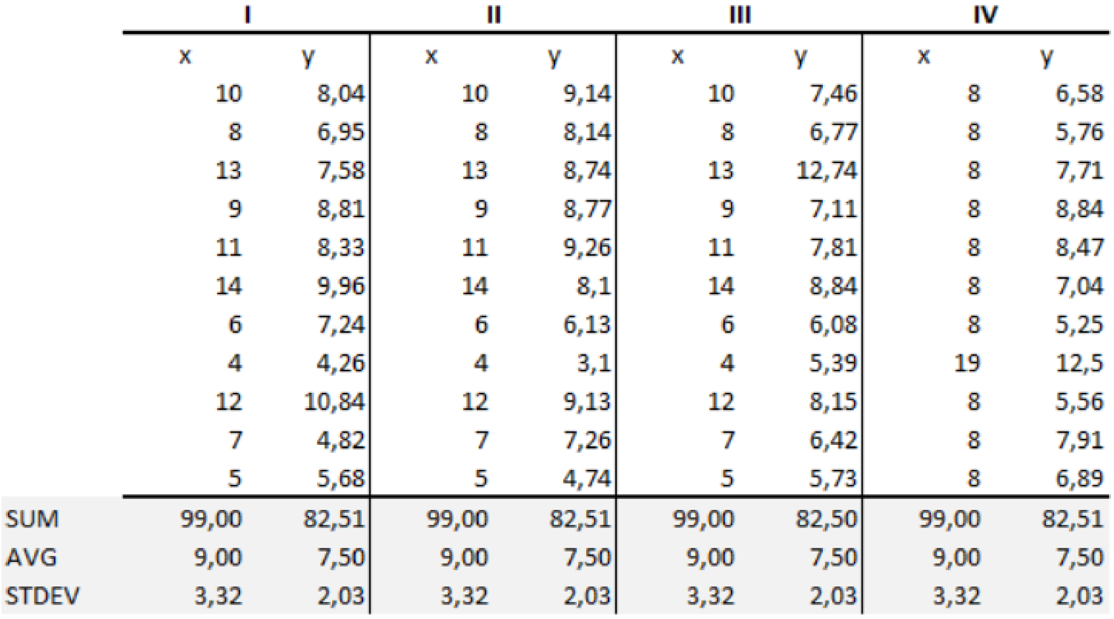
Coefficient of determination is a square of coefficient of correlation. This means, if the coefficient of correlation is 0.80, then the coefficient of determination will be 0.64 or 64%.

This coefficient of determination of 0.64, indicates that 64% of the change in the total of the dependent variable is associated with the change in the independent variable. The coefficient of correlation of 0.20, will result into r-squared of 4%, which indicates that only 4% change in the dependent variable is explained by the independent variable.

Question#4 Explain Anscombe's quartet in detail?

Answer #4

We know that statistics are great for describing general trends and aspects of data, but statistics alone can't fully depict any data set. Francis Anscombe realized this in 1973 and created several data sets, all with several identical statistical properties, to illustrate it. These data sets, collectively known as "Anscombe's Quartet". Each data set contains eleven(x,y) pairs. The important thing to note about these data sets is that they share the same descriptive statistics. But, things change completely, whey they are graphed. Each graph tells a different story, irrespective of the similar statistics description.



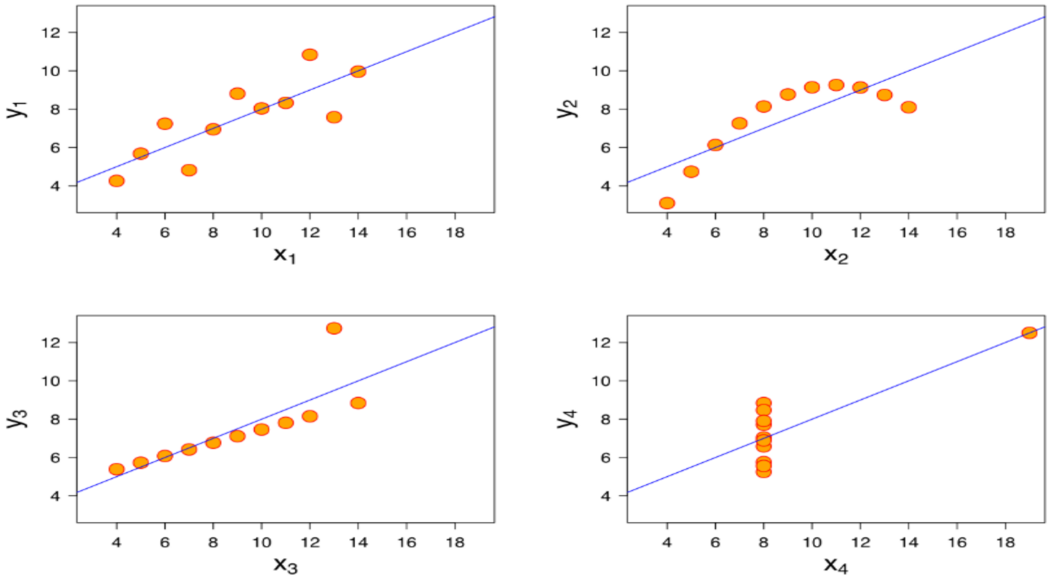
	I		II		III		IV	
	x	y	x	y	x	y	x	y
	10	8,04	10	9,14	10	7,46	8	6,58
	8	6,95	8	8,14	8	6,77	8	5,76
	13	7,58	13	8,74	13	12,74	8	7,71
	9	8,81	9	8,77	9	7,11	8	8,84
	11	8,33	11	9,26	11	7,81	8	8,47
	14	9,96	14	8,1	14	8,84	8	7,04
	6	7,24	6	6,13	6	6,08	8	5,25
	4	4,26	4	3,1	4	5,39	19	12,5
	12	10,84	12	9,13	12	8,15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

- *Mean of x is 9 and mean of y is 7.5 in each data set.*
- *Variance of x is 11 and variance of y is 4.13 for each data set.*
- *The coefficient of correlation between x and y is 0.816 for each data set*

When these four data sets are plotted on x/y coordinate plane, it can be observed that all of them show the same regression line, but each one is telling a different story.



- ✓ *Data set #1 appears to have clean and well fitted linear model*
- ✓ *Data set #2 is not distributed normally*
- ✓ *Data set #3 seems to be linear, but the calculated regression is thrown off by an outlier.*
- ✓ *Data set #4 shows that one outlier is enough to produce high correlation coefficient.*

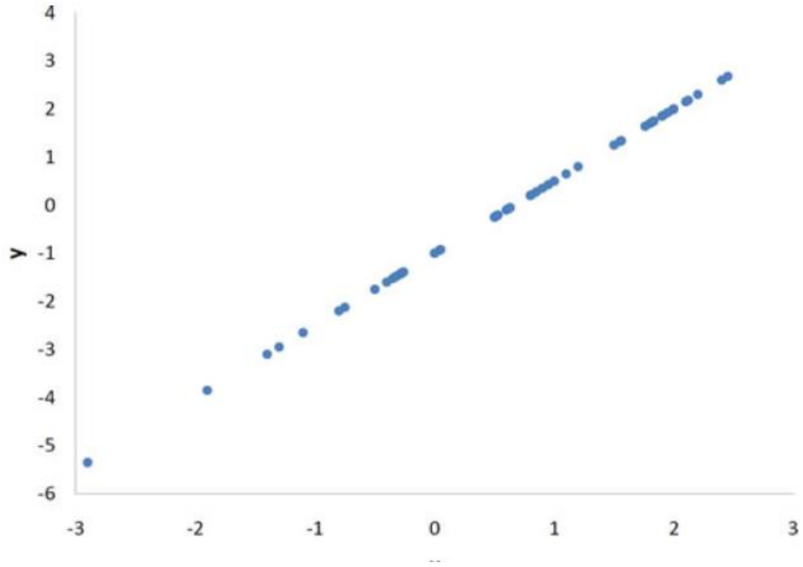
Anscombe's Quartet reminds us that graphing data prior to analysis is good practice, outliers should be removed when analyzing data, and statistics about a data set do not fully depict the data set in its entirety.

Question#5 What is Pearson's R?

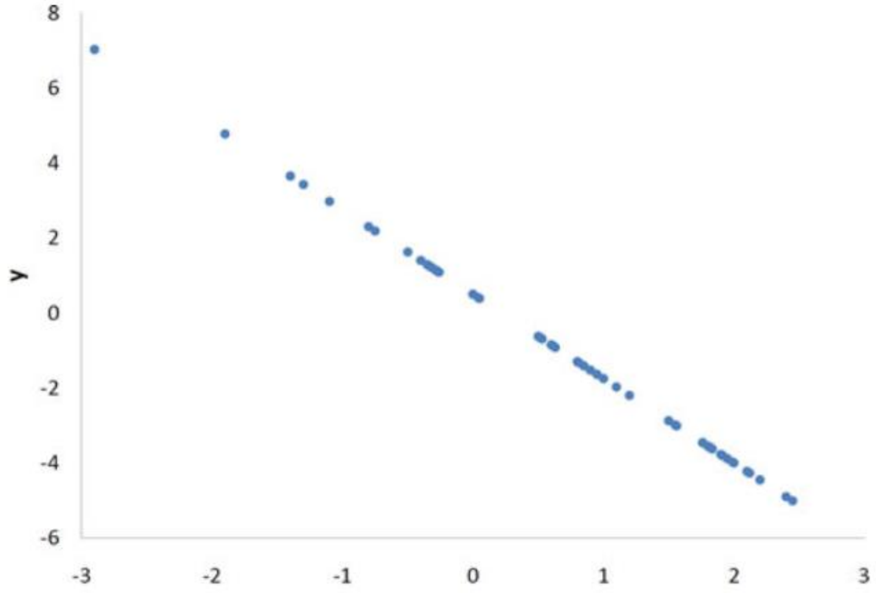
Answer #5

Pearson's R, also known as coefficient of correlation is a measure of the strength of the linear relationship between two variables. If the relationship between the two variables is not linear then the coefficient of correlation does not adequately represent the linear relationship of the two variables.

Pearson's R can range from -1 to 1. An r of -1 indicates, a perfect negative relationship between two variables, a value 0f +1, means a perfect positive relationship between two variables, while 0 indicates, there is no relationship between variables.



Perfect positive correlation



Perfect negative correlation

Question#6 what is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Answer #6

Scaling also known as data normalization is the method that is used to standardize the range of feature of data. As we know, the range of values of data may vary widely, it has become one of the mandatory steps of data pre-processing to scale the features, while building machine learning models or algorithms.

There are two method of scaling:

1. ***MinMax Scaling**:*

   The features are scaled in such a way that all the values lie between 0 and 1 using the maximum and minimum values in the data.

   Formula :

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

2. ***Standardized Scaling**:*

   In this method, the features are scaled such that their mean is 0 and standard deviation is 1

   Formula:

$$x = \frac{x - mean(x)}{sd(x)}$$

Why Scaling is performed?

When we have a lot of independent variables, a lot of them might be on a very different scales, which will lead a model to a very odd coefficients, that might be difficult to interpret. Therefore, we basically scale features because of the below reasons:

I. *Ease of interpretation*

II. *Faster convergence of gradient descent method.*

Difference between normalized and standardized scaling?

- *Normalization usually means to scale a variable to have a values between 0 and 1, while standardization transforms data to have a mean of zero and a standard deviation of 1.*
- *Normalized scaling does not change the values of Dummy variables, while standardize scaling does.*

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

Answer #7

VIF is a measure that shows the degree to which a variable will be affected because of the variable redundancy with other independent variables. As the squared multiple correlation of any independent variable with the other independent variables approaches unity, the correspondence VIF becomes infinite.

In other words, An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables.

Question#8 What is Gauss-Markov theorem?

Answer #8

According to Gauss-Markov theorem, if the certain set of assumptions are met, the OLS(Ordinary Least Square) estimate for regression coefficients gives the “Best Linear Unbiased Estimates”, also known as BLUE.

A theorem that proves that if the error terms in a multiple regression have the same variance and are uncorrelated, then the estimators of the parameters in the model produced by least squares estimation are better (in the sense of having lower dispersion about the mean) than any other unbiased linear estimator.

The Assumptions are:

1. ***Linearity***
   The parameters we are estimating using the OLS method must be themselves linear.
2. ***Random***
   Data must have been randomly sampled from the population.
3. ***Non-Collinearity:***
   The regressors being calculated aren’t perfectly correlated with each other.
4. ***Exogeneity***
   The regressors aren’t correlated with the error term.
5. ***Homoscedasticity***
   No matter what the values of the regressors might be, the error of the variance is constant.

Mathematically,

If *a linear regression model represented by*

$$yi = xi' \beta + \varepsilon i$$

and generated by the ordinary least squares estimate is the best linear unbiased estimate (BLUE) possible if

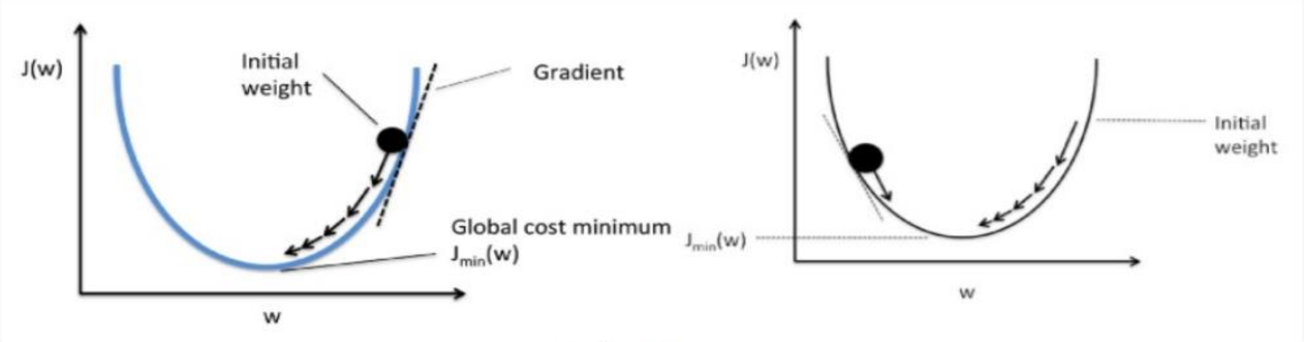
- *E{εi} = 0, i = 1, . . . , N*
- *{ε1……εn} and {x1…..,xN} are independent*
- *cov{εi, εj} = 0, i, j = 1,…., N I ≠ j.*
- *V{ε1 = σ2, i= 1, ….N*

Question#9 Explain the gradient -descent algorithm in detail?

Answer #9

Gradient descent is an optimization algorithm, which is used to optimize the cost function and find the value of cost function βs(estimators) corresponding to the optimized value of the cost function.

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



Gradient Descent

Mathematically, the aim of the gradient descent is to find the solution of the ArgMin J(Θ0,Θ1), where J(Θ0,Θ1) is the cost function of the linear regression.

Equation is :

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

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.

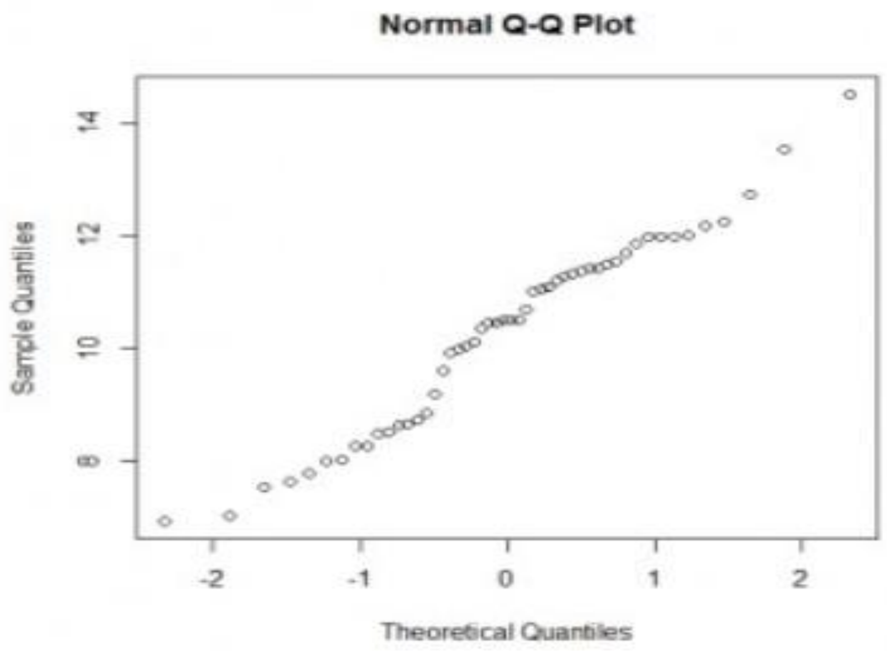
The update is:

$$\theta_j = \theta_j \; - \alpha \, \frac{1}{m} \; \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}).\, x_j^{(i)} \;\; \text{for j = 1,2,...,n}$$

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

Answer #10

The Q-Q plot, or quantile-quantile plot, is a graphical tool, used to assess if a set of data plausibly came from some theoretical distribution such as a Normal or exponential. *It is a scatterplot created by plotting two sets of quantiles(or percentiles) against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight. Here's an example of a Normal Q-Q plot when both sets of quantiles truly come from Normal distributions.*



Usage:

Below are some of the usages of Q-Q plot:

- *To verify if the two data sets come from population with a common distribution.*
- *Whether two data sets have common location and scale*
- *To verify, if the two data sets have the similar distributional shapes.*
- *To verify, if the two data sets have the similar tail behavior.*

Importance:

Check for common distribution:

When there are two data samples, it is often desirable to know if the assumption of a common distribution is justified. If so, then location and scale estimators can pool both data sets to obtain estimates of the common location and scale. If two samples do differ, it is also useful to gain some understanding of the differences. The q-q plot can provide more insight into the nature of the difference than analytical methods such as the chi-square, 2-sample tests etc.