correlation (r) :- correlation describe the mutual relation ship or connection b/w actual and predicted value. It is defined as follows

$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2}} \Rightarrow r(x,y) \text{ Direct method}$$

shortcut method

$$r(x,y) = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sqrt{(\sum x_i^2 - n\bar{x}^2)(\sum y_i^2 - n\bar{y}^2)}}$$

for Ex :- if $n = 5$

from the given data set, we have values of x and y

x	0	1	2	3	4
y	1	5	10	22	38

$\bar{x} = 2 ; \bar{y} = 15.2$

x	y	xy	x^2	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$	y^2
0	1	0	0	-2	-14.2	28.4	0
1	5	5	1	-1	-10.2	10.2	25
2	10	20	4	0	-5.2	0	100
3	22	66	9	1	6.8	6.8	484
4	38	152	16	2	22.8	45.6	1444

$\sum x_i = 10$, $\sum y_i = 76$, $\sum x_i y_i = 243$, $\sum x_i^2 = 30$; $\sum y_i^2 = 2053$

$$r = \frac{28.4 + 10.2 + 0 + 6.8 + 45.60}{\{4+1+0+1+4\}\{(-14.2)^2 + (-10.2)^2 + (-5.2)^2 + (6.8)^2 + (22.8)^2\}}$$

$$r = \frac{91}{\sqrt{10 * 898.6}} \Rightarrow \frac{91}{94.8} \Rightarrow 0.9599$$

using shortcut method

$$r(x,y) = \frac{243 - 5 \times 2 \times 15.2}{\sqrt{(30 - 5 \times 4)(2053 - 5 \times 231.04)}} \Rightarrow \frac{91}{\sqrt{10 \times 897.8}} \Rightarrow \frac{91}{94.75}$$

$\Rightarrow 0.9604$

Least Square Method :→

In least square method our aim to calculate the value m (slope) and b (y-intercept) in the equation of line $y = mx + b$

where y = how far up

x = How far along

m = slope or gradient (How steep the line is)

b = the y intercept (where the line crosses the y-axis)

steps :- To find the line of best fit for N Points

step I :- find m (slop)

step II :- calculate intercept b;

step III :- Assemble the eqn of a line $y = mx + b$

Ex :→ Using the Method of least square, find an equation of the form $y = ax + b$ that fit the any data (given at Previous Page)

Sol^n :- Consider the normal eqn of least square fit of a straight line i.e.

$$\sum_{i=1}^{n} y_i = a \sum_{i=1}^{n} x_i + nb \quad -(1)$$

$$\Sigma x_i y_i = a \Sigma x_i^2 + b \Sigma x_i \quad -(2)$$

So after Putting values

$$10a + 5b = 76 \quad -(3)$$

$$30a + 10b = 243 \quad -(4)$$

$$20a + 10b = 152$$
$$30a + 10b = 243$$
$$\overline{-10a = -91}$$
$$a = 9.1$$
$$b = -3$$

Putting in equation (3)

$10 \times 9.1 + 5b = 76$

$5b = 76 - 91$

$5b = -15$

Prediction :- Numeric Prediction is the task of Predicting continuous value for a given input.

For Example :- we may wish to Predict the salary of college graduates with 10 years of work experience or the potential sales of a new product given its Price.

By far, the most widely used approach for numeric Prediction is regression. It is a statistical methodology that was developed by Sir Francis Galton (1822-1911) In facts many texts use the term "regression" and "Numeric Prediction" synonymously. However, as we have seen some classification technique (such as backpropogation, SVM, and KNN classifiers) can be adapted for Prediction.

Regression :- The term regression means "stepping back toward the Average."

Regression analysis :- Regression analysis can be used to model the relationship between one or more independent or predictor variables and a dependent response variable (which is continuous valued).

In the context of data Mining, the Predictor variables are the attributes of interest describing the tuple. In general variable of Predictor variable are known.

OR

Regression analysis is a mathematical Measurement of the Average relationship between Two or more Variable in terms of the original units of the data.

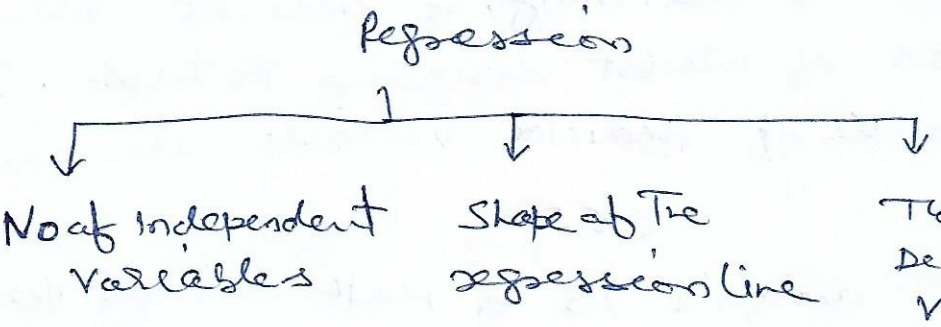
Regression Analysis is a form of Predictive modeling technique, which investigates the relationship b/w a dependent (target) and independent variable (Predictor).

This technique is used for forecasting, Time series modeling and finding the casual effect relationship between the variables.

for Ex:- Relationship between rash driving and number of road accidents by a driver is best studies through regression.

Regression analysis is an important tool for modelling and analysis data. Here, we fit a curve / line to the data points in such a manner that the difference b/w the distance of data points from the curve or line is minimised.

Types of Regression technique :-



Most commonly used regressions are:-

* Linear Regression
* logistic regression
* Polynomial Regression
* lasso regression
* Ridge Regression
* elastic regression
* step wise Regression

Linear Regression :→

If the variables in a bivariate distribution are related. we will find the point.

In the statistics, linear regression is a linear approach for modeling the relationship between a scalar dependent variable y and one or more explanatory variables (or independent variable) denoted x.

The case of one explanatory variable is called simple linear regression. for more than one explanatory variable, the process is called multiple linear regression (The term is distinct from multivariate linear regression, where multiple correlated dependent variables are predicted, rather than a single scalar variable.

Linear and logistic regressions are usually the first algorithms that people learn in predictive modeling.

In this technique, the dependent variable is continuous, independent variable can be continuous or discrete, and nature of regression line is linear.

Linear regression establish a relationship b/w dependent variable (y) and one or more independent variable (x) using a best fit straight line (also known as regression line)

It is represented by an eqn $y = ax + b + e$

where a is intercept, b is slop of the line and e is error term. This eqn can be used to Predict the value of target variable based on given Predictor variable (s)

* The difference between simple linear regression and multiple linear regression is that multiple linear regression has (>1) independent variable.

How to obtain best fit line (value of a and b):-

This task can be easily accomplished by least square method. It is the most common method used for fitting a regression line. It calculates the best fit line for the observed data by minimizing the sum of squres of the vertical deviations from each data point to the line, because the deviation are first squred, when added, There are no cancelling out between +ve and -ve values.

$\min_w \|x_w - \bar{X}\|_2^2$: Sum of squre $= \sum_{i=0}^{n} (x_i - \bar{x})^2$

where x_i = The i^{th} item in the set

$\bar{x}$ = The Mean of all items in the set

$(x_i - \bar{x})$ = The deviation of each item from the mean.

* We can evaluate the Model Performance using the metric R^2 also.

Line of best fit (least square Method) →

A line of best fit is a straight line that is the best approximation of the given set of data. It is used to study the nature of the relation between two variables (we are considering the two Dimensional case, here)

A line of best fit can ~~also~~ be roughly determined using an eyeball method by drawing a straight line on a scatter plot so that the number of points above the line and below the line is about equal (and the line passes through as many point as possible)

A More accurate way of finding the line of best fit is the least square method.

Use the following steps to find the eqn of Line of best fit for a set of ordered pairs $(x_1, y_1), (x_2, y_2), \ldots (x_n, y_n)$.

Step 1: Calculate the Mean of the x-value and y-values

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

Step 2:- the following formula gives the slope of line of best fit

$$m = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \quad \text{--- (1)}$$

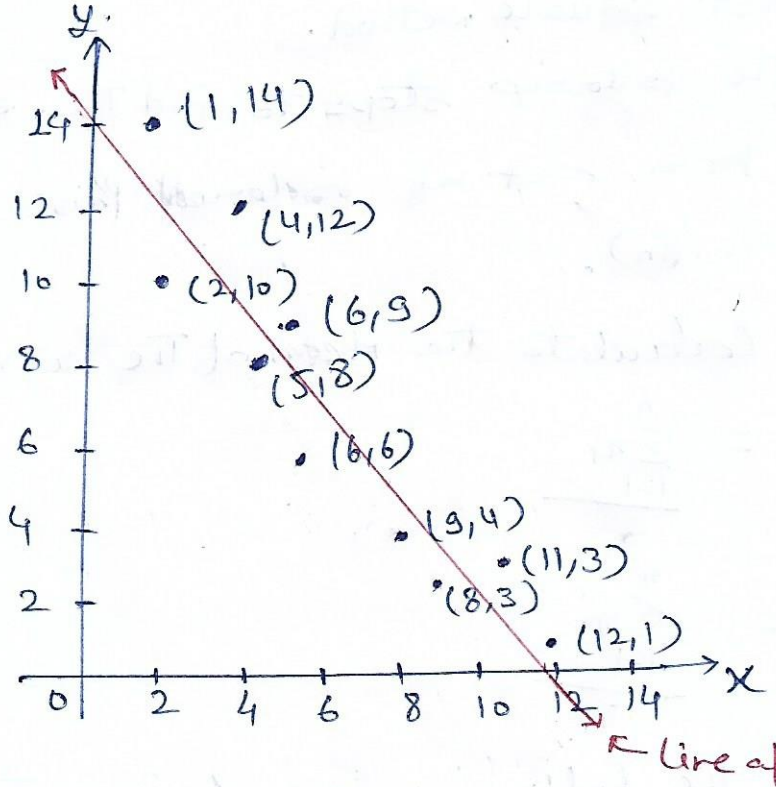
Step 3:- Compute the y-intercept of the line by using the formula

$$b = \bar{y} - m\bar{x} \quad \text{--- (2)}$$

Step 4:- use the slope m and y-intercept b to form the eqn of line.

Example:- use the least square method to determine the eqn of line of best fit for the data. Then plot the line

x	8	2	11	6	5	4	12	9	6	1
y	3	10	3	6	8	12	1	4	9	14



Calculate the means of the x-values and y-values

$$\bar{x} = \frac{8+2+11+6+5+4+12+9+6+1}{10} = 6.4$$

$$\bar{y} = \frac{3+10+3+6+8+12+1+4+9+14}{10} = 7$$

Now calculate $x_i - \bar{x}$, $y_i - \bar{y}$, $(x_i - \bar{x})(y_i - \bar{y})$, and $(x_i - \bar{x})^2$ for each i.

i	x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
1	8	3	1.6	-4	-6.4	2.56
2	2	10	-4.4	3	-13.2	19.36
3	11	3	4.6	-4	-18.2	21.16
4	6	6	-0.4	-1	0.4	0.16
5	5	8	-1.4	1	-1.4	1.96
6	4	12	-2.4	5	-12	5.76
7	12	1	5.6	-6	-33.6	31.36
8	9	4	2.6	-3	-7.8	6.76
9	6	9	-0.4	2	-0.8	0.16
10	1	14	-5.4	7	-37.8	29.16
					-131	118.4

Calculate the Slope

$$m = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \Rightarrow \frac{-131}{118.4} \Rightarrow -1.1$$

Calculate the y-intercept

use the formula to compute the y-intercept

$$b = \bar{y} - m\bar{x}$$

$$= 7 - (-1.1 \times 6.4)$$

$$= 7 + 7.04$$

$$\approx 14.0$$

use the slope and y-intercept to form the eqn of the line of best fit.

Therefore, the eqn is $y = -1.1x + 14.0$

Draw the line on the Scatterplot:-

Linear Regression :→ Straight line regression analysis involves a response variable, 'y' and a single predictor variable x. It is the simplest form of regression, and models y as a linear function of x. That is

$$y = b + wx$$

where the value of y is assumed to be constant and b and w are regression coefficients specifying the y intercept and slop of the line, respectively. The regression coefficients w and b, can also be thought of as weights, so that we can equivalently write

$$\boxed{y = w_0 + w_1 x}$$

These coefficient can be ~~solved for~~ by the method of least squares, which estimates the best fitting straight line as the one that minimize the error between the actual data and the estimate of the line.

* Model The relationship that salary may be related to The number of years of work experience with The eqn

$$y = w_0 + w_1 x$$

table salary data

x-years experience	y-salary (in $1000)
3	30
8	57
9	64
13	72
3	36
6	43
11	59
21	90
1	20
16	83

table (a)

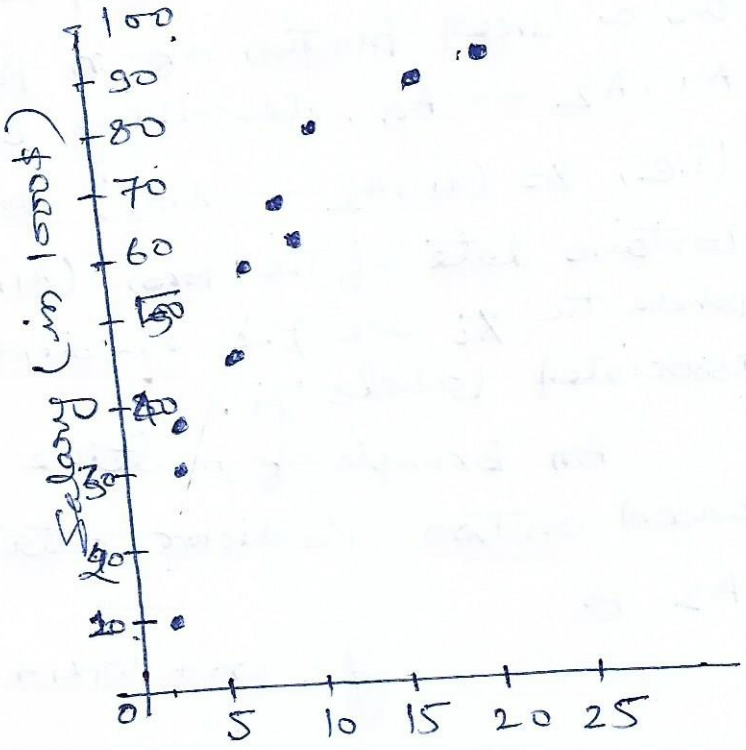


figure (b)

fig (b): shows The overall pattern suggest a linear relationship b/w x (year experience) and y (salary). Given The above data, we compute $\bar{x} = 9.1$ and $\bar{y} = 55.4$. Substitute these values into eqn (1) and (2), we get

$$w_1 = (3.5)$$
$$w_0 = (23.6)$$

Thus The eqn of least square line is estimated by $y = 23.6 + 3.5x$ .. Using This eqn, we can predict that salary of a college graduate with 10 years of experience is 58000$

Multiple Linear Regression :→

It is an extension of Straight line regression, so as to involve more than one Predictor Variable. It allows response variable y to be modeled as a linear function of n Predictor variable or attributes $A_1, A_2 \dots A_n$, describing a tuple, X.
(i.e, $X = (x_1, x_2 \dots x_n)$) our Train data set, D, contains data of the form $(x_1, y_1), (x_2, y_2) \dots (x_{|D|}, y_{|D|})$, where the x_i are the n-dimensional Train tuples with associated labels y_i.

An Example of multiple linear regression model based on Two Predictor attributes or variable A_1 and A_2 is

$$y = w_0 + w_1 x_1 + w_2 x_2 \quad -(3)$$

or $$y = w_0 + w_1 x_1 + w_2 x_2 \dots + w_3 x_3 \quad -(4)$$

where $x_1, x_2 \dots x_n$ are the value of attribute A_1, A_2 and $\dots A_n$, respectively.

The Method of least square can be ~~find~~ extended to solve for $w_0, w_1, w_2 \dots w_n$.

Non-linear Regression :→

How can we Model data that does not show a linear dependency ?

for example :- what if a given response variable and Predictor Variable have a relationship that may be modeled by a Polynomial function ? Instead of linear model, we can get more accurate model using a non-linear model such as Parabola or some other higher-order Polynomial. Polynomial regression is often of interest when There is just one Predictor variable. It can be Modeled by adding Polynomial terms to the basic linear model.

for EX :- $y = w_0 + w_1 x + w_2 x^2 + w_3 x^3$