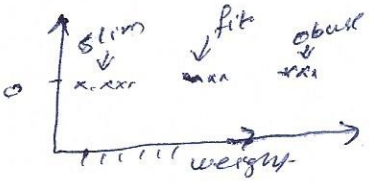
Module-5 -

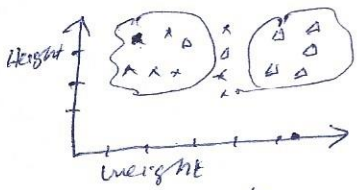
Univariate data - Data consisting of only one variable. ie only one variable is measured for each observation. ie only age or only weight.

Univariate data analysis - By drawing histograms.
- by drawing Pie charts.

When plotting with the single feature we can classify the data into categories. ie Just with the help of weight, it is possible to classify. But using only one feature, there may be overlapping in classification. So we are using bivariate or multivariate data analysis.



Univariate analysis



Bivariate analysis.

Based on the obtained points, it is possible to classify but the points may overlap.

Suppose there are 3 features, age and dob, then we have to use 3D or 4D diagrams. But it is not practically possible. So this multivariate data analysis is performed.

Bivariate data

- It is the data consisting of only 2 variables (quantitative and/or qualitative)
- 3 cases,
  - Both can be quantitative.
    eg. age & heights of persons.
  - One qualitative & one quantitate.
    eg: Type of tyre thread compared to stopping distance.

- Both qualitative.

eg: Gender of students compared to types of degrees.

Used to - determine the relationship b/w 2 variables

- find the dependency of one on the other.

- Determine the correlation b/w the data.

Bivariate data can be considered as two measurements on each observation.

ie There are 2 variables, x & y.

for eg.

	age (x)	height (y)
1	27	5'6"
2	⋮	⋮
⋮	⋮	⋮
n	21	6'0"

(age & height of n persons)

In a bivariate data, - both variables can be quantitative. (as in previous example)

- one qualitative & one quantitative

eg:- Type of tyre thread compared to stopping distance.

- Both qualitative.

eg:- Gender of students compared to types of degrees.

Bivariate data are - used to determine the relationship b/w 2 variables.

- to find the dependency of one with the other.

- to determine the correlation b/w the data.

more examples of bivariate data,

House price in $1000s.	Square Feet
245	1400
312	1600
279	1700
308	1850
199	1100
219	1550
405	2350
324	2450
319	1425
255	1700.

n = 10.

Shelf Space	Sales
5	140
10	260
15	280
20	310.

n = 4

ie when 2 variables are measured on a single experimental unit, the resulting data are called bivariate data.

x is called independent variable or Predictor or covariant

y is called dependent variable or outcome or response.

- You can describe each variable individually and also can explore the relationship b/w 2 variables.
- Bivariate data can be described with graphs or numerical measures.
- A single measurement is a pair of numbers (x, y) that can be plotted using a 2D graph called a scatter plot.

Scatter diagram

Scatter diagram method is a simple representation which is used in commerce & statistics to find correlation b/w 2 variables.

- These 2 variables are plotted along the X and Y axis on a 2D graph and the pattern represents the association b/w

these given variables.

- Study of such a graphical representation involving 2 variables and using such a diagram is known as scatter diagram analysis.

Using scatter diagram we can determine,

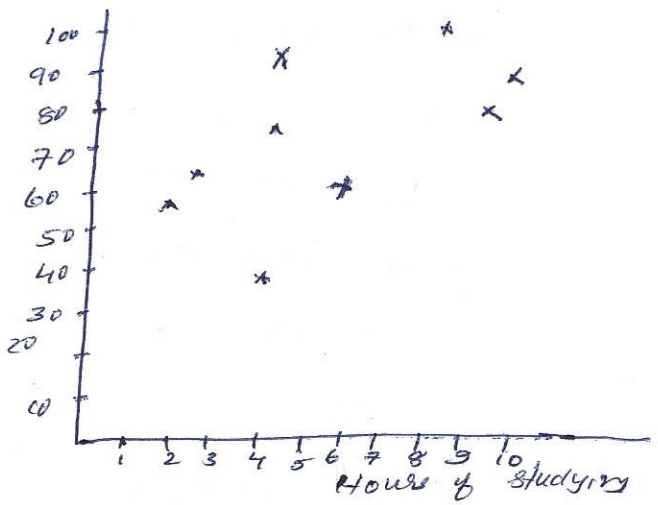
- the pattern is linear or non-linear,
- strength of the relationship b/w variables
- the presence of unusual observations, clusters and outliers.

Let $(x_1, y_1)(x_2, y_2) \ldots (x_n, y_n)$ be the set of observations obtained in a study of a population or sample, in which 2 characteristics are considered.

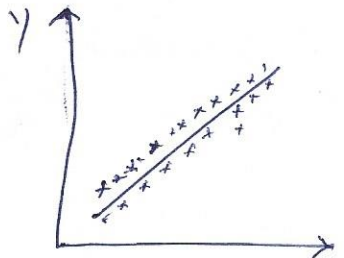
Scatter diagram is a diagram obtained by plotting points with co-ordinates $(x_1, y_1)(x_2, y_2) \ldots (x_n, y_n)$. ~~is called a scatter diagram~~ It consists of 'n' points scattered over the (x, y) plane.

fig eg:

Hours of studying	exam score
9	53
4.5	35
5	91
5	72
6	60
3	62
10	85
9.5	78
8	99



From the scatter diagram, it is often possible to visualize a smooth curve approximating the data. Such a curve is called an approximating curve.



fig(1)

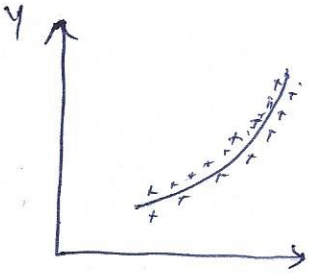


fig (2)

fig (3)

In fig (1), the data appear to be approximated well by a straight line, ie there exists a linear relationship between the variables.

In fig(2), a relationship exists b/w the variables but it is not a linear relationship. It is called a nonlinear relationship.

In fig (3), no relationship exists b/w the variables.

Curve fitting.

The general problem of finding equations of approximating curves that fit given sets of data is called curve fitting. The type of equation is often suggested from the scatter diagram.

In fig(1) we could use a straight line,

$$y = a + bx$$

In fig(2) we could use a parabola or quadratic curve,

$$y = a + bx + cx^2.$$

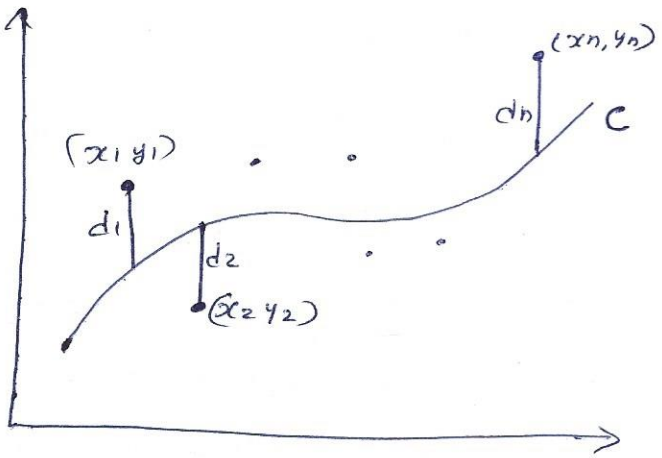
Regression

One of the main purposes of curve fitting is to estimate one of the variables (dependent variable) from the other (independent variable). The process of estimation is often referred to as regression. If y is to be estimated from x by means of some equation

we call the equation a regression equation of y on x and the corresponding curve a regression curve of y on x.

Method of least squares

- Generally more than one curve of a given type will appear to fit a set of data.
- To avoid individual judgments in constructing lines, Parabolas, or other approximating curves, it is necessary to agree on a definition of a 'best-fitting line', 'best-fitting Parabola' etc.



Consider the above figure ~~with~~ in which the data points are (x_1, y_1) (x_2, y_2) ... (x_n, y_n).

- for a given value of x, say x_1, there will be a difference b/w the value y_1 and the corresponding value as determined from the curve C. Let this difference be d_1, which is sometimes referred to as a deviation, error or residual and may be +ve, -ve or zero.

Similarly, corresponding to the values $x_2, \ldots x_n$, we obtain the deviations

$d_2, d_3 \ldots d_n$.

- A measure of the goodness of fit of the curve C to the set of data is provided by the quantity $d_1^2 + d_2^2 + \ldots + d_n^2$. If this is small the fit is good, if it is large, the fit is bad.

So the definition

- Of all curves in a given family of curves approximating a set of n data points, a curve having the property that $d_1^2 + d_2^2 + \ldots + d_n^2$ is a minimum is called a best-fitting curve in the family.

- A curve having this property is said to fit the data in the least squares sense and is called a least squares regression curve or simply least square curve.

- A line having this property is called a least squares line.

- A Parabola with this property is called a least squares Parabola etc.

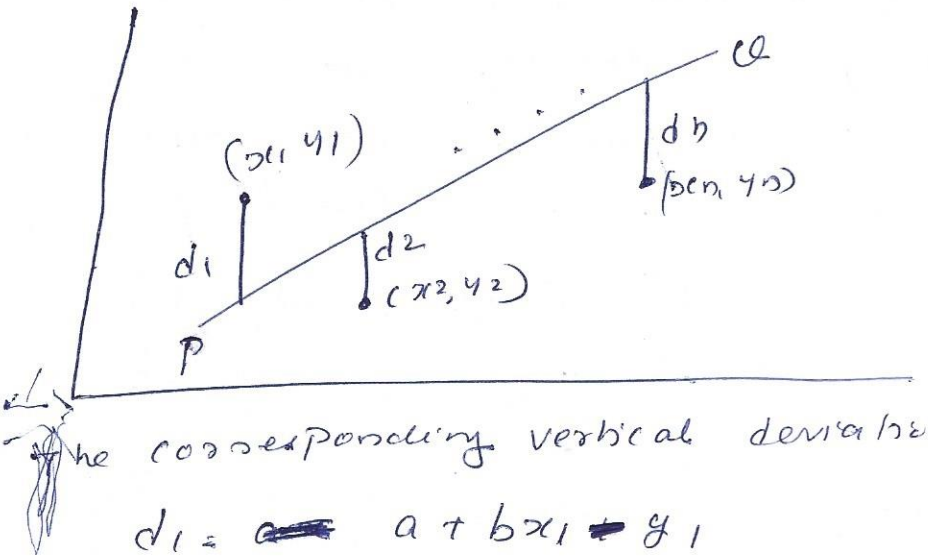
Here x is the independent variable, and y is the dependent variable.

If x is the dependent variable & y is the independent variable, then we have to interchange the x & y axes.

fitting of a straight line - by method of least squares

Let $(x_1, y_1)\ (x_2, y_2) \ldots (x_n, y_n)$ be the observations.

The values of y on the least squares line corresponding to $x_1, x_2 \ldots x_n$ are (referring to the following figure) $a+bx_1,\ a+bx_2 \ldots a+bx_n$.



The corresponding vertical deviations are,

$$d_1 = a + bx_1 - y_1$$

$$d_2 = a + bx_2 - y_2$$

$$d_n = a + bx_n - y_n.$$

Then the sum of the squares of the deviations is

$$d_1^2 + d_2^2 + \ldots d_n^2 = (a+bx_1-y_1)^2 + (a+bx_2-y_2)^2 + \ldots + (a+bx_n-y_n)^2.$$

ie $\Sigma d^2 = \Sigma (a+bx-y)^2$

This is a function of a and b.

ie $E(a,b)$.

These $d_1, d_2, \ldots$ are called the residual errors

$E(a,b) = \Sigma(a+bx-y)^2$.

A necessary condition for this to be a minimum is that $\frac{\partial E}{\partial a} = 0$ & $\frac{\partial E}{\partial b} = 0$ since

$$\frac{\partial \epsilon}{\partial a} = \Sigma \frac{\partial}{\partial a}(a+bx-y)^2 = \Sigma\, 2\,(a+bx-y)$$

$$\frac{\partial \epsilon}{\partial a} = \Sigma \frac{\partial}{\partial b}(a+bx-y)^2 = \Sigma\, 2x\,(a+bx-y)$$

ii $\Sigma\, 2(a+bx-y) = 0$ & $\Sigma\, 2x(a+bx-y) = 0$

ii $\Sigma(a+bx-y) = 0$ & $\Sigma\, x(a+bx-y) = 0$.

$\Sigma y = an + b\Sigma x$ & $\Sigma xy = a\Sigma x + b\Sigma x^2$.

iii The least squares line approximating the set of points $(x_1, y_1)(x_2 y_2) \ldots (x_n y_n)$ has the equation,

$$y = a + bx \quad \text{---} \quad (1)$$

where the constants a & b are determined by solving simultaneously the equations,

$$\left.\begin{aligned} \Sigma y &= an + b\Sigma x \\ \Sigma xy &= a\Sigma x + b\Sigma x^2 \end{aligned}\right\} (2)$$

which are called the normal equations for the least-squares line.

The values of a & b obtained from (2) are given by,

$$\left.\begin{aligned} a &= \frac{(\Sigma y)(\Sigma x^2) - (\Sigma x)(\Sigma xy)}{n\Sigma x^2 - (\Sigma x)^2} \\ b &= \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{n\Sigma x^2 - (\Sigma x)^2} \end{aligned}\right\} (3)$$

b can also be written as,

$$b = \frac{\Sigma (x-\bar{x})(y-\bar{y})}{\Sigma (x-\bar{x})^2}$$

where $\bar{x}$ & $\bar{y}$ are the mean ie $\bar{x} = \frac{\Sigma x}{n}, \bar{y} = \frac{\Sigma y}{n}$

Problem ① Fit a straight line of the form $y = a + bx$ for the following data.

x	1	3	4	6	8	9	11	14
y	1	2	4	4	5	7	8	9

The equation of the line is $y = a + bx$.

The normal equations are,

$$\Sigma y = an + b\Sigma x$$
$$\Sigma xy = a\Sigma x + b\Sigma x^2.$$

x	y	x^2	xy
1	1	1	1
3	2	9	6
4	4	16	16
6	4	36	24
8	5	64	40
9	7	81	63
11	8	121	88
14	9	196	126
$\Sigma x = 56$	$\Sigma y = 40$	$\Sigma x^2 = 524$	$\Sigma xy = 364$

here $n = 8$

The normal equations are,

$$40 = a \times 8 + b \times 56 \; ; \quad 364 = a \times 56 + b \times 524$$

i.e. $$8a + 56b = 40 \quad \text{—} ①$$
$$56a + 524b = 364 \quad \text{—} ②$$

① $\times 7 \rightarrow$ $$56a + 392b = 280 \quad \text{—} ③$$

② $-$ ③ $\rightarrow$ $$132b = 84$$

(10) $$b = 84/132 = 7/11$$

Substituting the value of b in (1)

$$8a + 56 \times \frac{7}{11} = 40$$

$$8a = 40 - \frac{392}{11} = \frac{440 - 392}{11} = \frac{48}{11}$$

$$\text{or } a = \frac{48}{11 \times 8} = \frac{6}{11}$$

So the required least square line is

$$y = a + bx$$

$$\text{i.e. } y = \frac{6}{11} + \frac{7}{11}x$$

Prob 2 Fit a straight of the form $y = a + bx$ for the following data,

x	1	3	5	7	8	10
y	8	12	15	17	18	20

x	y	x^2	xy
1	8	1	8
3	12	9	36
5	15	25	75
7	17	49	119
8	18	64	144
10	20	100	200
$\Sigma x = 34$	$\Sigma y = 90$	$\Sigma x^2 = 248$	$\Sigma xy = 582$.

$$n = 6$$

The normal equations are,

$$\Sigma y = an + b\Sigma x$$

$$\Sigma xy = a\Sigma x + b\Sigma x^2$$

i.e. $90 = 6a + 34b$; $582 = 34a + 248b$.

i.e. $6a + 34b = 90$ — (1)

$34a + 248b = 582$ — (2)

(1) × 17 → $102a + 578b = 1530$. — (3)

(2) × 3 → $102a + 744b = 1746$ — (4)

(4) − (3) →

$$0 + 166b = 216$$

$$b = \frac{216}{166} = \frac{108}{83} = \underline{1.3012}$$

Substituting (b) in (1)

$$6a + 34 \times \frac{108}{83} = 90.$$

$$6a + 44.24 = 90$$

$$6a = 90 - 44.24 = 45.76$$

$$a = \frac{45.76}{6} = \underline{7.63}$$

~~Substi~~

So the required least square line is

$$y = a + bx$$

i.e. $y = 7.63 + 1.3012 \times x$

i.e. $y = 7.63 + 1.3x$

Prob(3) Fit a straight line of the form, $y = a + bx$ to the following data.

x :	0	1	2	3	4
y :	0	1.8	3.3	4.5	6.3

Prob(4) In the following data S denotes Son's height in cms & F denotes fathers height in cms. Fit a straight line of the form $S = a + bF$

S :	142	168	156	173	175	176	177
F :	155	160	163	175	178	179	180