TWO DIMENSIONAL RANDOM VARIABLES

Suppose that we are performing an experiment *E* that results into a sample space *S*. Let us associate two functions, $X = X(s)$ and $Y = Y(s)$ to this sample space *S* so that these assign a real number to each element $s \in S$. Then (*X*, *Y*) is called a two dimensional random variable.

(*X*, *Y*) is called a two dimensional discrete random variable if the possible values of (*X*, *Y*) are finite or countably infinite.

(*X*, *Y*) is a two dimensional continuous random variable if (*X*, *Y*) can assume all values in a subset of *X*-*Y* plane

Probability Mass Function of a two Dimensional Discrete Random Variable

Let (X, Y) be a two dimensional discrete random variable taking values $\{(x_i, y_j), i, j = 1, 2, 3, \dots\}$. Let us associate a number $p_{ij} = p(x_i, y_j)$ to each of the values (x_i, y_j) representing the probability $P(X = x_i, Y = y_j)$. We say that (x_i, y_j, p_{ij}) is a joint probability distribution for (X, Y) and p as the joint *pmf* for (X, Y) if the following conditions are satisfied.

(i) $p(x_i, y_j) \geq 0 \ \ for\ all\ \ (x, y)$

(ii) $\sum_{j=1}^{\infty} \sum_{i=1}^{\infty} p(x_i, y_j) = 1$

Example: Probability mass function for a two dimensional random variable is represented in a tabular form. Following table gives the *pmf* of a two dimensional random variable (X, Y).

$X\downarrow\ \ Y\rightarrow$	0.5	1	1.5	3
1	1/12	1/24	1/6	1/24
2	1/24	1/24	1/12	1/6
3	1/6	1/12	1/24	1/24

One can infer from this table that, $P(X = 1, Y = 0.5) = 1/12$, $P(X = 2, Y = 3) = 1/6$ etc.

Probability Density Function of a two Dimensional Continuous Random Variable

Let (*X*, *Y*) be a two dimensional random variable taking all the values in some given region *R* of *X*-*Y* plane. The joint *pdf* of (*X*, *Y*) is defined as a function $f(x,y)$ satisfying the following properties.

(i) $f(x,y) \geq 0 \; for \; all \; (x,y) \in R$

(ii) $\iint_R f(x,y)dx\,dy = 1$

Please note that second property here implies that total volume under the surface $z = f(x,y)$ is 1?

Example: Find the value of *c* such that the function $f(x, y) = c, 5 < x < 10, 4 < y < 9; 0, elsewhere$, represents a legitimate *pdf*. Also, find $P(X \geq Y)$.

Let us first find the value of *c* using above mentioned second condition. We have to calculate *c* such that,

$$\int_4^9 \int_5^{10} c \, dx \, dy = 1$$

This gives 25*c* = 1 and in turn we get *c* = 1/25.

Thus the value of *c* is 1/25. This value is also such that it satisfies first condition. Thus, the legitimate *pdf* is given by,

$$f(x,y) = \frac{1}{25}, 5 < x < 10, 4 < y < 9;$$

$$= 0, elsewhere$$

Let us now calculate $P(X \geq Y)$.

$P(X \geq Y) = 1 - P(X < Y)$ and $P(X < Y)$ can be calculated as,

$$P(X < Y)$$

$$= \int_5^9 \int_x^9 \frac{1}{25}\ dy\ dx$$

$$= \frac{1}{25} \int_5^9 (9 - x) dx$$

$$= \frac{1}{25} |9x - \frac{x^2}{2}|_5^9$$

$$= \frac{1}{25} |9x - \frac{x^2}{2}|_5^9$$

$$= \frac{8}{25}$$

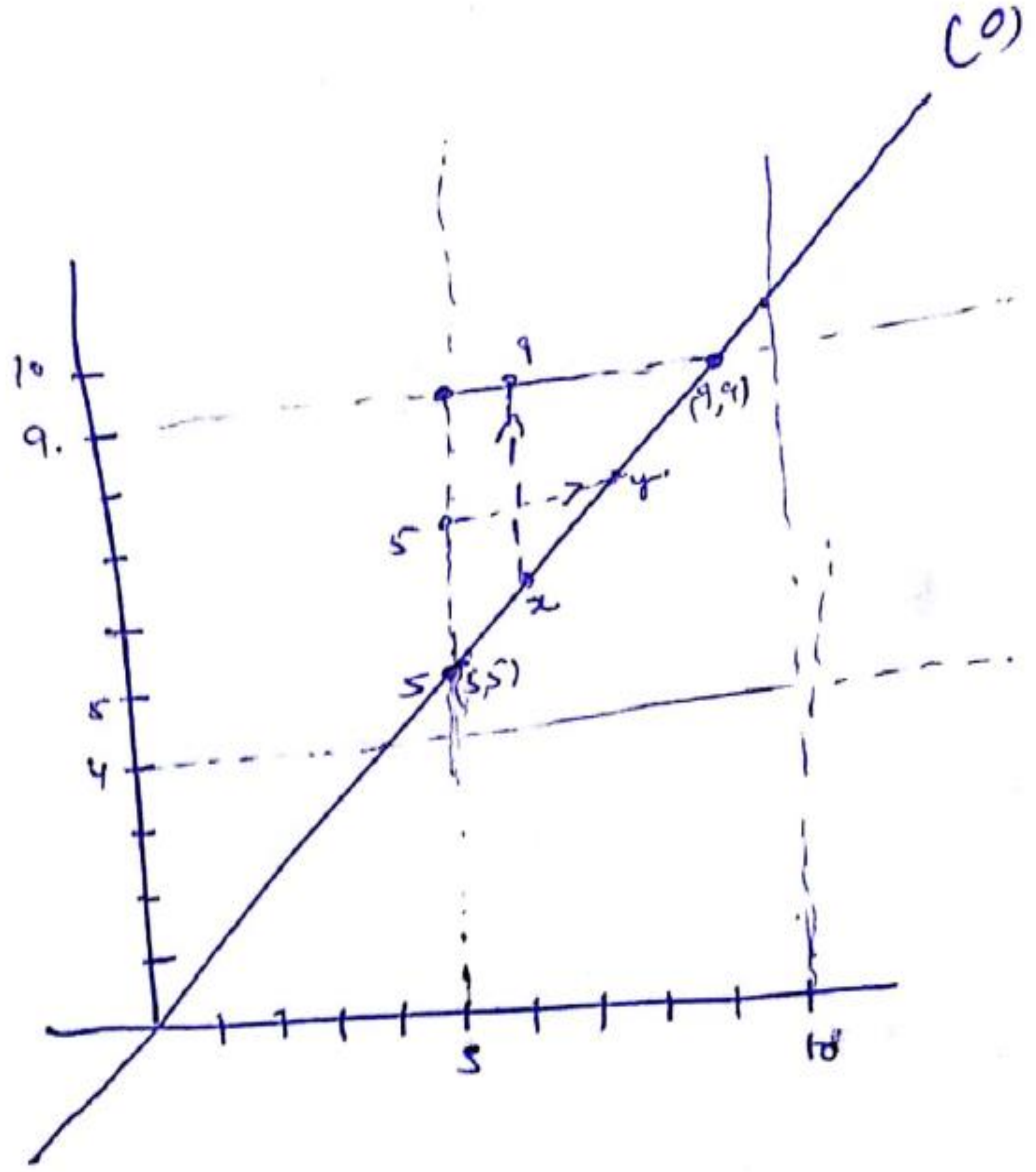
As such $P(X \geq Y)$ = $\frac{17}{25}$.

Here

$$P(X<Y) = \int_5^9 \int_x^9 \frac{1}{25} \, dy \, dx$$

$$= \frac{1}{25} \int_5^9 (9-x) \, dx$$

$$= \frac{1}{25} \left[9 \cdot x - \frac{x^2}{2} \right]_5^9$$



$$= \frac{8}{25}$$

So $P(X \geq Y) = 17/25$;

Similarly

we can also write -

$$P(X < Y) = \int_5^9 \int_5^y \frac{1}{25} \, dx \, dy$$

$$= \frac{1}{25}\int_5^9 (y-5)\, dy$$

$$= \frac{1}{25}\left[\frac{y^2}{2} - 5y\right]_5^9 \quad = \frac{8}{25}$$

$$\therefore \; P(X > Y) = 1 - 8/25 = 17/25$$

Note: The kind of distribution in which we take $f(x, y) = c$ over a given region R, is called as two dimensional uniform distribution. Here, the value of constant c shall be given by $\frac{1}{area\ of\ Region\ R}$.

MARGINAL PROBABILITY DISTRIBUTIONS

Suppose that we are given a two dimensional random variable (X, Y) and its probability distribution. With this distribution, we can associate two one dimensional distributions. These distributions are defined individually for the variable X and Y. These distributions are called marginal distributions. We define these distributions for discrete and continuous cases separately.

(i) Discrete Case

For the discrete case, we define marginal probability distribution of X as,

$$P(X = x_i) = p(x_i)$$

$$= P(X = x_i, Y = y_1 \; or \; X = x_i, Y = y_2 \; or \; \ldots)$$

$$= \sum_{j=1}^{\infty} p(\, x_i, y_j)$$

Similarly, marginal probability distribution of Y is defined as,

$$P\left(Y = y_j\right) = q\left(y_j\right)$$

$$= \sum_{i=1}^{\infty} p(\, x_i, y_j)$$

Let us take an example to illustrate the concept. Let us take the two dimensional random variable (X, Y) following the *pmf* as given below.

$X\downarrow \quad Y\rightarrow$	0.5	1	1.5	3
1	1/12	1/24	1/6	1/24
2	1/24	1/24	1/12	1/6
3	1/6	1/12	1/24	1/24

Using the above mentioned definitions, we can find the marginal distribution function of *X* as,

$X = x_i$	1	2	3
p_i	1/3	1/3	1/3

Thus X follows a uniform distribution. *A discrete distribution is called a uniform distribution if $P(X = x_i)$ is constant for all x_i.*

Marginal distribution function of *Y* shall be:

$Y = y_j$	0.5	1	1.5	3
q_j	7/24	1/6	7/24	¼

Please note that this is not a uniform distribution?

(i) **Continuous Case**

Let us be given the joint *pdf* of (*X*, *Y*) as $f(x, y)$, then marginal *pdf* of *X*, $g(x)$ is defined as,

$$g(x) = \int_{-\infty}^{\infty} f(x, y)dy$$

and marginal *pdf* of *Y*, $h(y)$ is defined as,

$$h(y) = \int_{-\infty}^{\infty} f(x, y)dx$$

Example: consider that the *pdf* of a two dimensional random variable (X, Y) is given by,

$$f(x, y) = \frac{1}{25}, 5 < x < 10, 4 < y < 9;$$

$$= 0, elsewhere.$$

We have shown that this represents a legitimate probability density function. Marginal *pdf* of X is given by,

$$g(x) = \int_{-\infty}^{\infty} f(x, y) dy$$

$$= \int_{4}^{9} \frac{1}{25} dy$$

$$= \frac{1}{5}$$

As such marginal *pdf* of X is given by

$$g(x) = \frac{1}{5}\ , 5 < x < 10;$$

$$= 0, elsewhere.$$

Using a similar integration will can get the marginal *pdf* of Y as,

$$h(y) = \frac{1}{5}\ , 4 < y < 9;$$

$$= 0, elsewhere.$$

CONDITIONAL PROBABILITY DISTRIBUTIONS

We can also define conditional distribution of X given Y and that of Y given X for the two situations when we deal with discrete variables and when we deal with continuous variables *(i) Discrete Case*

We define, for the case when we are interested in finding the probability of $x_i|y_j$,

$$p(x_i|y_j)$$

$$= P(X = x_i|Y = y_j)$$

$$= \frac{p(x_i, y_j)}{q(y_j)}, \quad q(y_j) > 0$$

Similarly, we define for the case when we are interested in finding the probability of $y_j|x_i$,

$$q(y_j|x_i)$$

$$= P(Y = y_j|X = x_i)$$

$$= \frac{p(x_i, y_j)}{p(x_i)}, p(x_i) > 0$$

(i) Continuous Case

Let us consider a two dimensional random variable (X, Y) with its joint *pdf* as $f(x, y)$. Let $g(x)$ and $h(y)$ be the marginal probability density functions of *X* and *Y*, respectively. Then conditional distribution of *X* given *Y* = *y* is defined as,

$$g(x|y) = \frac{f(x,y)}{h(y)}, h(y) > 0.$$

And the conditional distribution of *Y* given $X = x$ is defined as,

$$h(y|x) = \frac{f(x,y)}{g(x)}, g(x) > 0.$$

1

<u>Example</u>

$$f(x,y) = \begin{cases} 10\,xy^2 & 0 < x < y < 1 \\ 0 & \text{elsewhere} \end{cases}$$

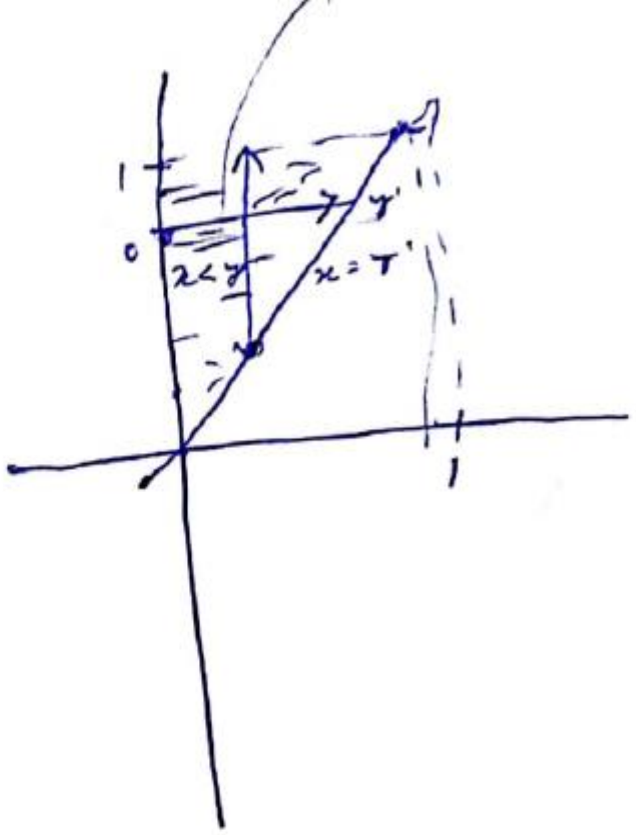
Check this is a valid p.d.f w.r.t. $(x<y)$

Solution :

$$\int_0^1 \int_0^y 10xy^2 \, dy \cdot dx$$

$$= \int_0^1 5y^4 \, dy = 1$$

Hence valid p.d.f.



$$g(x) = \int f(x,y) \, dy$$

Now
Marginal prob. distribution $g(x) = \int_x^1 10xy^2 \, dy$

$$= \frac{10}{3} x (1 - x^3) \quad ; \quad 0 < x < 1$$

$$0 \quad ; \quad \text{else where}$$

similarly

$$h(y) = \int_0^y 10 x y^2 \, dx.$$

$$= 5y^4 \quad ; \quad 0 < y < 1$$

$$0 \quad ; \quad \text{elsewhere.}$$

so the conditional distribution is

$$g(x|y) = \frac{f(x,y)}{h(y)}, \quad g(y) > 0$$

$$= \frac{10xy^2}{5y^4} \qquad \begin{matrix} 0 < x < y \\ 0 < y < 1 \end{matrix}$$

$$0 \qquad \text{elsewhere.}$$

Similarly

$$h(y|x) = \frac{f(x,y)}{g(x)}$$

$$= \begin{cases} \dfrac{10xy^2}{\frac{10}{3}x(1-x^3)} & x < y < 1 \\ & 0 < x < 1 \\ 0 & \text{elsewhere.} \end{cases}$$

$$= \begin{cases} \dfrac{3y^2}{(1-x^3)} & x < y < 1 \\ & 0 < x < 1 \\ 0 & \text{elsewhere.} \end{cases}$$

Now find the following probabilities

(i) $P\left(X < \frac{1}{4}\right)$ (ii) $P\left(Y > \frac{3}{4}\right)$ (iii) $P\left(0 < X+Y < \frac{1}{2}\right)$

(iv) $P\left(X < \frac{1}{2} \,\middle|\, Y = \frac{3}{4}\right)$ (v) $P\left(Y < \frac{1}{2} \,\middle|\, X = \frac{1}{4}\right)$

(vi) $P\left(0 < X < \frac{1}{2},\ \frac{1}{4} < y < \frac{3}{4}\right)$

Solution

(i) $P\left(X < \frac{1}{4}\right) = \int_0^{1/4} \frac{10}{3} x(1-x^3) \cdot dx$

Use marginal prob. distribution

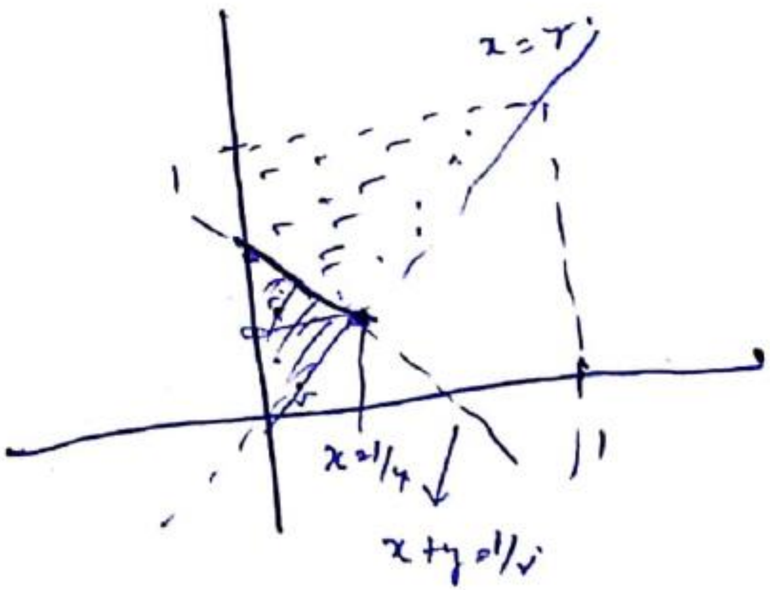
$$= \frac{10}{3}\left(\frac{1}{32} - \frac{1}{5.4^5}\right)$$

can be simplified

(ii) $P\left(Y > \frac{3}{4}\right) = \int_{3/4}^{1} 5y^4 \, dy$

$= 1 - \left(\frac{3}{4}\right)^5$ Ans.

(iii) $\displaystyle\iint_{0<x+y<1/2} 10\,xy^2\,dx\,dy$



$$= \int_0^{1/4}\int_x^{\frac{1}{2}-x} 10\,xy^2\,dy\,dx$$

$$= \frac{10}{3}\int_0^{1/4} x\left\{\left(\tfrac{1}{2}-x\right)^3 - x^3\right\}dx$$

Can be calculated.

4

$P\left(X < \frac{1}{2} \middle| y = 3/4\right) = ?$ (conditional dist.)

$$g(x \mid y = 3/4) = \frac{2x}{\frac{9}{16}} = \frac{32}{9} x \qquad 0 < x < 3/4$$

$$0 \qquad \text{elsewhere}$$

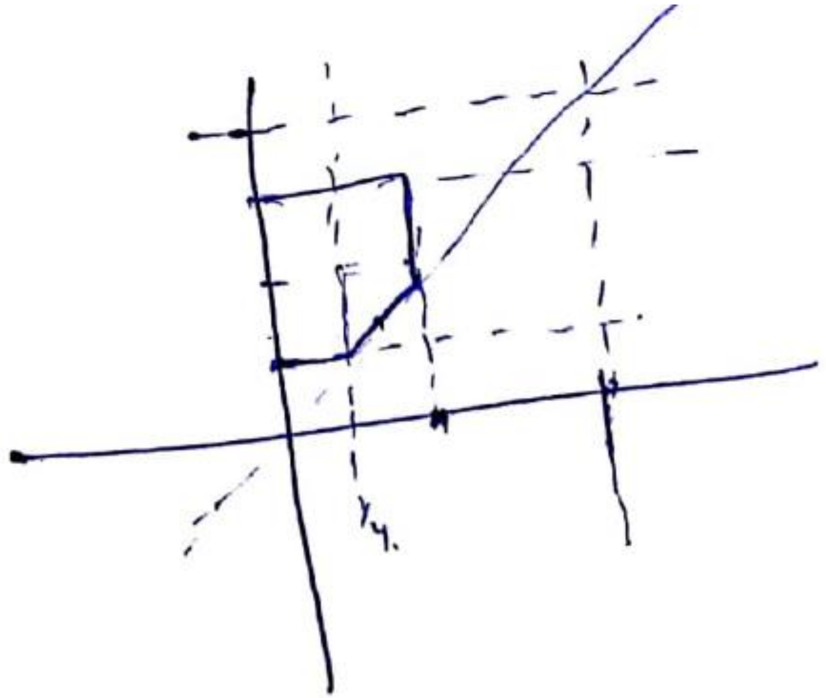
$$P\left(X < \frac{1}{2} \middle| y = 3/4\right) = \int_0^{1/2} \frac{32}{9} x \qquad = \frac{4}{9}$$

(V) $P(Y < 1/2 \mid X = 1/4) = ?$ (conditional Again)

$$h(y \mid x = 1/4) = \frac{3y^2}{1 - \left(\frac{1}{4}\right)^3} = \begin{cases} \frac{64}{21} y^2 ; & \frac{1}{4} < y < 1 \\ 0 ; & \text{elsewhere} \end{cases}$$

$$P(Y = \tfrac{1}{2} \mid x = 1/4) = \int_{1/4}^{1/2} \frac{64}{21} \cdot y^2 \, dy = \frac{1}{9}$$ Ans

(vi) $P\left(0 < X < \frac{1}{2}, \frac{1}{4} < Y < \frac{3}{4}\right)$



$$= \int_{1/4}^{3/4}\int_{0}^{1/4} 10xy^2\,dx\,dy + \int_{1/4}^{1/2}\int_{x}^{3/4} 10xy^2\,dy\,dx$$

= can be solved.

Independent Random Variables

Let (X, Y) be a two dimensional random variable. We say that *X* and *Y* are two independent random variables if and only if,

$p(x_i, y_j) = p(x_i).q(y_j)$ for all i and j, when (X, Y) is a discrete random variable

and

$f(x, y) = g(x).h(y)$ for all (*x*, *y*), when (X, Y) is a continuous random variable.

Independence of Random variables

Example :-

let $f(x,y) = 1 \quad 0 < x < 1,\ 0 < y < 1$

$\quad\quad 0 \quad$ elsewhere

here $g(x) = \begin{cases} 1 & 0 < x < 1 \\ 0 & \text{e.w} \end{cases}$ $\quad \because \int_0^1 f(x,y)\cdot dy = 1$

$h(y) = \begin{cases} 1 & 0 < y < 1 \\ 0 & \text{elsewhere} \end{cases}$ same here

Here $f(x,y) = g(x)\cdot h(y)$

so X & Y are independent Random variable

<u>Example 2</u>

$P(X=1, Y=1) = \frac{1}{4}$; $P(X=1, Y=0) = 1/4$

$P(X=0, Y=1) = \frac{1}{4}$, $P(X=0; Y=0) = 1/4$

Here

$P(X=0) = 1/2$ $\qquad$ $P(X=1) = 1/2$

$P(Y=0) = 1/2$ $\qquad$ $P(Y=1) = 1/2$.

$$P(x_i, y_j) = P(X_i) \cdot P(Y_j) \quad \forall \; x_i; y_j$$

$$\frac{1}{4} = \frac{1}{2} \cdot 1/2$$

So X & Y are ~~&~~ independent distribution

On the other hand, if we see the yesterday's example

$$f(x,y) \neq g(x)\cdot h(y)$$

$$10xy^2 \neq \frac{10}{3}x(1-x^3)\cdot 5y^4$$

So in yesterday's example, X, & Y are not independent.

Use of independent distribution is that sometime we know the individual distribution, but then in case of independent distribution, we can multiply and get the joint Distribution.

Expectation in case of Joint distribution

let $g(x, y)$ be a function of x & y

we define

$$E\, g(X, Y) = \sum_{(x_i, y_j) \in \mathcal{X} \times \mathcal{Y}} \sum g(x_i, y_j)\, p(x_i, y_j)$$

if X & Y are discrete with pmf $p(x_i, y_j)$

(provided given series is absolute convergent)

In case of (X, Y) continuous with joint pdf $f(x, y)$, we define

$$E\, g(X, Y) = \iint g(x, y)\, f(x, y)\, dx\, dy$$

provided integral is absolute convergent

In general, $g(X,Y)$ can be $X+Y$, XY, etc.

Product moments

$$\mu'_{r,s} = E(X^r Y^s) \rightarrow (r,s)^{th} \text{ non central product moment}$$

$$\mu'_{1,1} = E(XY)$$

$$\mu'_{1,0} = E(X) = \mu_X$$

$$\mu'_{0,1} = E(Y) = \mu_Y$$

$\mu_{r,s}$ is defined as

$$= E(X-\mu_X)^r (Y-\mu_Y)^s \rightarrow (r,s)^{th} \text{ central product moment}$$

$r=1, s=1$

$$\mu_{1,1} = E(X-\mu_X)(Y-\mu_Y) \longrightarrow$$ this is called covariance between X & Y

$$= E(XY - X\mu_Y - \mu_X Y + \mu_X\mu_Y)$$

$$= E(XY) - \mu_Y\mu_X - \cancel{\mu_X\mu_Y} + \cancel{\mu_X\mu_Y}$$

$$\mu_{1,1} = E(XY) - E(X)E(Y) \longrightarrow$$ covariance

(4)

If X & Y are independent, then

$$E(X^r Y^s) = E(X^r) \cdot E(Y^s)$$

Similarly $E(X-\mu_X)^r (Y-\mu_Y)^s = E(X-\mu_X)^r \cdot E(Y-\mu_Y)^s$

To see this, we will see following result

Theorem:- Let X & Y be independent r. v.

then $$E[f_1(X)\cdot f_2(Y)] = E[f_1(X)]\cdot E[f_2(Y)]$$

provided Expectation exists

Proof:-

Suppose X & Y are continuous with joint $f(x,y)$ (p.d.f.) and marginal pdfs $g(x)$ & $h(y)$ & $f(x,y) = g(x)\, h(y)\ \forall (x,y)$

Now $$E[f_1(X)\cdot f_2(Y)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_1(x)\cdot f_2(y)\cdot f(x,y)\, dx\, dy$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_1(x) \cdot f_2(y) \cdot g(x) \cdot h(y)\, dx\, dy$$

$$= \left(\int_{-\infty}^{\infty} f_1(x) \cdot g(x) \cdot dx \right) \left[\int_{-\infty}^{\infty} f_2(y) \cdot h(y) \cdot dy \right]$$

$$= E[f_1(x)] \cdot E[f_2(y)]$$

proved

Similarly explanation can be given for discrete random variables X & Y.

It means
If X & Y are independent then
Co-variance of X, Y i.e $Cov(X,Y) = 0$

using this, we define

The coefficient of correlation between X & Y

$$\rho_{X,Y} = \frac{Cov(X,Y)}{s.d.(X) \cdot sd(Y)} = \frac{\sigma_{XY}}{\sigma_X \cdot \sigma_Y}$$

$\sigma_X^2 = var(X)$; $\sigma_Y^2 = var(Y)$

coefficient of correlation gives the measure of linear relationship between X & Y.

Now consider r.v U & V with

$E(U)=0$; $E(U^2)=1$, $E(V)=0$, $E(V^2)=1$

Consider $E(U-V)^2 \geq 0$ (Expectation of non negative term)

$\Rightarrow E[U^2+V^2-2UV] \geq 0$

$\Rightarrow 1+1-2E[UV] \geq 0$

$\Rightarrow E[UV] \leq 1$

Similarly

$$E(U+V)^2 \geq 0$$

$$\Rightarrow E[U^2 + V^2 + 2UV] \geq 0$$

$$\Rightarrow 1 + 1 + 2E[UV] \geq 0$$

$$\Rightarrow E[UV] \geq -1 \quad \text{—— (11)}$$

$$\Rightarrow -1 \leq E[UV] \leq 1$$

Now to check, when the equality holds

$E[UV] = 1$ if $E[U-V]^2 = 0$

This will be possible iff $P[U = V] = 1$

Similarly $E[UV] = -1$ if $E[U+V]^2 = 0$

$\Rightarrow$ iff $P[U = -V] = 1$

Now for any random variables X & Y,

let $E(X) = \mu_X$, $E(Y) = \mu_Y$, $Var(X) = \sigma_X^2$,

$Var(Y) = \sigma_Y^2$

Define $U = \dfrac{X - \mu_X}{\sigma_X}$; $V = \dfrac{\cancel{Y} - \mu_Y}{\sigma_Y}$

$$E[U] = E\left(\frac{X - \mu_X}{\sigma_X}\right) = 0 \qquad E[U^2] = \frac{E(X - \mu_X)^2}{\sigma_X^2}$$

$$= \frac{\sigma_X^2}{\sigma_X^2} = 1$$

Similarly $E(V) = 0$, $E(V^2) = 1$

So

$$-1 \le E(UV) \le 1 \qquad \text{—— } ②$$

so $E(UV) = E\left(\frac{X-\mu_X}{\sigma_X}\right)\left(\frac{Y-\mu_Y}{\sigma_Y}\right)$

Numerator is $\operatorname{cov}(X,Y)$ so

$$E(UV) = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} = \rho_{X,Y}$$

so for any random variable X, Y

$$-1 \le \rho_{X,Y} \le 1$$

$$\rho_{X,Y} = 1 \iff P\left(\frac{X-\mu_X}{\sigma_X} = \frac{Y-\mu_Y}{\sigma_Y}\right) = 1$$

or $P(X = aY + b) = 1$ when $a > 0$

$$\rho_{X,Y} = -1 \iff P\left(\frac{X-\mu_X}{\sigma_X} = -\frac{Y-\mu_Y}{\sigma_Y}\right) = 1$$

or $P(X = aY + b) = 1$ if $a < 0$

we can write now

X & Y are perfectly linearly related in +ve direction

$\rho(X = aY + b) = 1$; $a > 0$

$\rho(X = aY + b) = 1$, $a < 0$

then we say X & Y are perfectly linearly related in −ve direction

In general any value between −1 to 1 gives us degree of linear relationship.

are

us .

If $\rho_{X,Y} = 0$, we say that X, & Y are uncorelated.

Uncorelated doesn't means ~~or~~ independent.

But if X & Y are independent, then they are uncorelated.

Theorem : If X & Y are independent, then $\rho_{X,Y} = 0$, but ~~the~~ the converse of this is not true.

Proof: $\Rightarrow$ if X, & Y are independent then

$$Cov(X,Y) = 0$$

$$\Rightarrow$$

$$\rho_{XY} = 0$$

$\Leftarrow$ let us see through Example.

$X \backslash Y$	-1	0	1	$g(x)$
0	0	$1/3$	0	$1/3$
1	$1/3$	0	$1/3$	$2/3$
$h(y)$	$1/3$	$1/3$	$1/3$	

$$E(X) = 0 \cdot 1/3 + 1 \cdot 2/3 = 2/3$$

$$E(Y) = -1 \cdot 1/3 + 0 \cdot 1/3 + 1 \cdot 1/3 = 0$$

$$E(XY) = 0(-1) \cdot 0 + (0)(0)(1/3) + 0(1) \cdot 0$$
$$+ 1 \cdot (-1) \cdot 1/3 + 1(0) \cdot 0 + 1 \cdot 1 \cdot 1/3$$

$$= 0$$

$$Cov(X,Y) = E(XY) - E(X)E(Y) = 0$$

$\rho_{X,Y} = 0 \rightarrow$ uncorrelated

but ~~$p_X(0) = 1/3$ $p_Y(0)$~~ $g(0) = 1/3$; $h(0) = 1/3$

$$f(0,0) = 1/3$$

$$f(0,0) \neq g(0) \cdot h(0)$$

So not independent but uncorrelated.

Example! - Let $f(x,y) = x + y \quad 0 < x < 1,\ 0 < y < 1$ (11)

$\qquad\qquad\qquad 0 \quad$ elsewhere

$$E(XY) = \int_0^1\int_0^1 x.y\,(x+y)\,dx\,dy$$

$$= \int_0^1 \left[y.\frac{x^3}{3} + y^2.\frac{x^2}{2}\right]_0^1 dy$$

$$= \int_0^1 \left(\frac{y}{3} + \frac{y^2}{2}\right) dy = \frac{1}{6} + \frac{1}{6} = 1/3$$

$$g(x) = \int_0^1 (x+y)\cdot dy = \begin{cases} x + 1/2 & 0 < x < 1 \\ 0 & \text{elsewhere} \end{cases}$$

similarly $$h(y) = \begin{cases} y + 1/2 & 0 < y < 1 \\ 0 & \text{elsewhere} \end{cases}$$

$$E(x^2) = \int_0^1 x^2 \cdot (x + 1/2)\, dx \; ;$$

$$= \frac{5}{12}$$

$$= E(Y^2)$$

$$E(X) = \int_0^1 x\,(x + 1/2)\, dx$$

$$= 7/12$$

$$= E(Y)$$

$$V(X) = E(X^2) - [E(X)]^2$$

$$= \frac{5}{12} - \frac{49}{144} = \frac{11}{144}$$

$$\therefore \quad \rho_{X,Y} = \frac{\frac{1}{3} - \left(\frac{7}{12}\right)^2}{\frac{11}{144}} = \boxed{-\frac{1}{11}}$$

It means there is negative low degree of correlation between variables.

$$f(x,y) = \begin{cases} 2 & 0 < y < x < 1 \\ 0 & ew \end{cases}$$

$$g(x) = \int_0^x 2 \cdot dy = \begin{cases} 2x & 0 < x < 1 \\ 0 & ew \end{cases}$$

$$h(y) = \int_y^1 2\, dx = \begin{cases} 2(1-y) & 0 < y < 1 \\ 0 & ew \end{cases}$$

$$E(X) = \int_0^1 2x^2 \cdot dx = 2/3 \quad ; \quad E(X^2) = \int_0^1 2 \cdot x^3 \cdot dx = 1/2$$

$$VAR(X) = 1/2 - \frac{4}{9} = \frac{1}{18}$$

$E(Y) =$

$$\hookrightarrow \int_0^1 2y(1-y)\,dy = 1 - 2/3 = 1/3$$

$$E(y^2) = \int_0^1 2 \cdot y^2 (1-y)\,dy = \frac{2}{3} - 1/2 = 1/6$$

$$VAR(y) = \frac{1}{6} - \frac{1}{9} = \frac{1}{18}$$

$$E(XY) = \int_0^1 \int_0^x 2xy\,dy\,dx.$$

$$= \int_0^1 x^3 \cdot dx = 1/4$$

$$COV(X,Y) = E(XY) - E(X) \cdot E(Y) = \frac{1}{4} - \frac{2}{3} \cdot \frac{1}{3}$$

$$= \frac{1}{36}$$

$$\rho_{X,Y} = \frac{1/36}{\frac{1}{18}} = \boxed{1/2} \rightarrow$$

Moderate degree of +ve linear relationship between X & Y.

Example:

Suppose that a two dimensional random variable (X, Y) is uniformly distributed over the region $\{(x, y) | -2 < x < 2, -2 < y < 4\}$. Find the correlation coefficient between X and Y.

Let us first find the *pdf* of this two dimensional random variable (X, Y). We know that for a uniformly distributed random variable (X, Y), the *pdf* will be of the form of $f(x, y) = c$, c being a constant. As such, we have to find c such that, $\int_{-2}^{2}\int_{-2}^{4} c\, dy\, dx = 1$. This will give us the value of c as $c = \frac{1}{24}$.

As such, *pdf* of (X, Y) is:

$$f(x, y) = 1/24, \quad -2 < x < 2, -2 < y < 4;$$

$$= 0, \quad \text{elsewhere.}$$

Now, we need to find marginal distributions of *X* and *Y* in order to find the correlation coefficient between *X* and *Y.*

Here, marginal distribution of *X* is obtained as:

$$g(x) = \int_{-2}^{4} \frac{1}{24} dy = \frac{1}{4}, -2 < x < 2$$

Similarly marginal distribution of *Y* is obtained as:

$$h(y) = \int_{-2}^{2} \frac{1}{24} dx = \frac{1}{6}, -2 < y < 4$$

One can note that, for this problem, $g(x)$ and $h(y)$ are again two uniform distributions defined for two one-dimensional variables. Using the theory of uniform distribution of one dimensional variables, we can obtain,

$$E(X) = \frac{-2+2}{2} = 0,\ E(Y) = \frac{-2+4}{2} = 1,$$

$$V(X) = \frac{(2-(-2))^2}{12} = \frac{4}{3} \text{ and } V(Y) = \frac{(4-(-2))^2}{12} = 3$$

Also,

$$E(XY) = \int_{-2}^{2}\int_{-2}^{4} xy\,dy\,dx$$

$$= \int_{-2}^{2} x\left|\frac{y^2}{2}\right|_{-2}^{4} dx$$

$$= \int_{-2}^{2} x(8-2)dx$$

$$= 6\left|\frac{x^2}{2}\right|_{-2}^{2}$$

$$= 6(2-2)$$

$$= 0$$

As such, correlation coefficient is:

$$\rho = \frac{E(XY) - E(X)E(Y)}{\sqrt{V(X)V(Y)}}$$

$$= \frac{0 - 0 * 1}{\sqrt{\frac{4}{3} * 3}} = 0$$

It means X and Y are uncorrelated

Correlation Coefficient of a Random Sample

Let $(X_1, Y_1),\ (X_2, Y_2), \ldots, (X_n, Y_n)$ be a random sample of size *n* > 2 from a bivariate distribution. Then the statistic,

$$R = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

is called the sample correlation coefficient between the two variables *X* and *Y*. Here, $\bar{X}$ is the sample mean for variable *X* and $\bar{Y}$ is the sample mean corresponding to variable *Y*.

Example: Following sample of size 5 is given. Find the correlation coefficient between X and Y.

x_i	1	2	3	4	5
y_i	2	5	4	8	6

We can calculate that $\bar{X} = 3$ and $\bar{Y} = 5$. Also, $\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) = 11$, $\sum_{i=1}^{n}(X_i - \bar{X})^2 = 10$ and $\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = 20$.

This gives,

$$R = \frac{11}{\sqrt{200}} = 0.7778.$$

As such, there is a high positive correlation between the variables X and Y.

CURVE FITTING USING PRINCIPLE OF LEAST SQUARE

We usually study two or more variables in a hope that we will be able to extract some association between them and this association will, in turn, help us in estimating the value of a variable that depends on one or more variables that are being studied. The methods that help us in such a prediction are called regression methods

Linear Regression using Principle of Least Squares

Let us be given some data in the form of $(x_i, y_i), i = 1, 2, \dots, n$. Here, variable *Y* is depending upon variable *X*. We can observe that this kind of data may be available to us in a variety of situations. Following pairs of (*X*, *Y*) are some of Following situations.

X	Y
The flight time of a space craft	Distance from earth
Amount of irrigation water	The yield of crop
Height of a student	Weight of a student
Percentage of marks in 10^{th} standard	Percentage of marks in 12^{th} standard
Percentage of marks in entrance examination	Percentage of marks in final examination
CGPA of a student after 2^{nd} semester	CGPA of a student after 8^{th} semester
…	…

The regression problem is to find a relationship between X and Y based on the given values $(x_i, y_i), i = 1, 2, \dots, n$ so that we can estimate the value of Y for those values of X that are not there in the given data.

Let us understand the linear regression first and we will then take this further to non-linear regression.

Let us assume that we have the data in the form of $(x_i, y_i), i = 1, 2, \dots, n$ and we wish to fit a linear curve to this data.

This curve will give us a relation of the form,

$$Y = a + bX.$$

This relationship involves two variables, namely, *a* and *b*. If we somehow know the values of these variables, this linear relationship between *X* and *Y* shall be completely defined. Let us comprehend the principle of least squares that is used to find the values of *a* and *b*.

For known values of *a* and *b*, we can find value of dependent variable *Y* for a given value of independent variable *X*. We can carry out this process even for those values that are there in the given data. Let us denote these by $\hat{Y}$, these are nothing but the estimated values of *Y* obtained from the assumed linear relationship between *X* and *Y*.

As such, we are given the data,

$$(x_i, y_i), i = 1, 2, \dots, n$$

and assuming the linear relationship,

$$Y = a + bX$$

we have the estimated data,

$$(x_i, \hat{y}_i) = (x_i, a + bx_i), i = 1, 2, \dots, n.$$

for some values of *a* and *b*.

Let us consider *E* as,

$$E = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

As such *E* is sum of squares of errors in estimated values. *Principle of least squares states that values of a and b are determined in such a way that this squared sum of errors is least*.

Here,

$$E = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - (a + bx_i))^2.$$

We use the theory of optimization to find the values of *a* and *b*. This theory states that *E* will be minimum for such values of *a* and *b* that are obtained by,

$$\frac{\partial E}{\partial a} = 0 \text{ and } \frac{\partial E}{\partial b} = 0$$

$\frac{\partial E}{\partial a} = 0$ gives,

$$\sum_{i=1}^{n} y_i = an + \ b \sum_{i=1}^{n} x_i \text{ and}$$

$\frac{\partial E}{\partial b} = 0$ gives,

$$\sum_{i=1}^{n} x_i y_i = a \sum_{i=1}^{n} x_i + \ b \sum_{i=1}^{n} x_i^2$$

These equations, called *normal equations* are used to calculate the values of *a* and *b*.

Example: Let us find the line of regression for the following data.

i	1	2	3	4	5
x_i	1	2	3	4	5
y_i	2	5	4	8	6

We have to calculate $\sum_{i=1}^{n} x_i,\ \sum_{i=1}^{n} y_i, \sum_{i=1}^{n} x_i y_i\ and\ \sum_{i=1}^{n} x_i^2$ for obtaining normal equations. Let us again consider the above data. We can obtain

i	1	2	3	4	5	$\sum$
x_i	1	2	3	4	5	15
y_i	2	5	4	8	6	25
$x_i y_i$	2	10	12	32	30	86
x_i^2	1	4	9	16	25	55

As such, the normal equations are,

$$5a + 15b = 25$$

and $$15a + 55b = 86$$

Solving these equations we get, $a = 1.7\ and\ b = 1.1$. Thus the line of regression for the above data is,

$$y = 1.7 + 1.1x$$

This equation is also called the line of regression of *Y* on *X*. This line can be used to predict the value of *Y* for given value of *X*. For example, when *x* = 1.5, we can predict the

value of *y* as *y* = 1.7 + 1.1*1.5 = 3.35. Also, when *x* is 3.5, we can predict that *y* will be 5.55.

We can also obtain the line of regression of *X* on *Y* following the very similar steps. The normal equations for such a line will be (by exchanging the roles of *X* and *Y* in normal equations),

$$5a + 25b = 15$$

and $$25a + 145b = 86$$

Solving these equations, we will get,

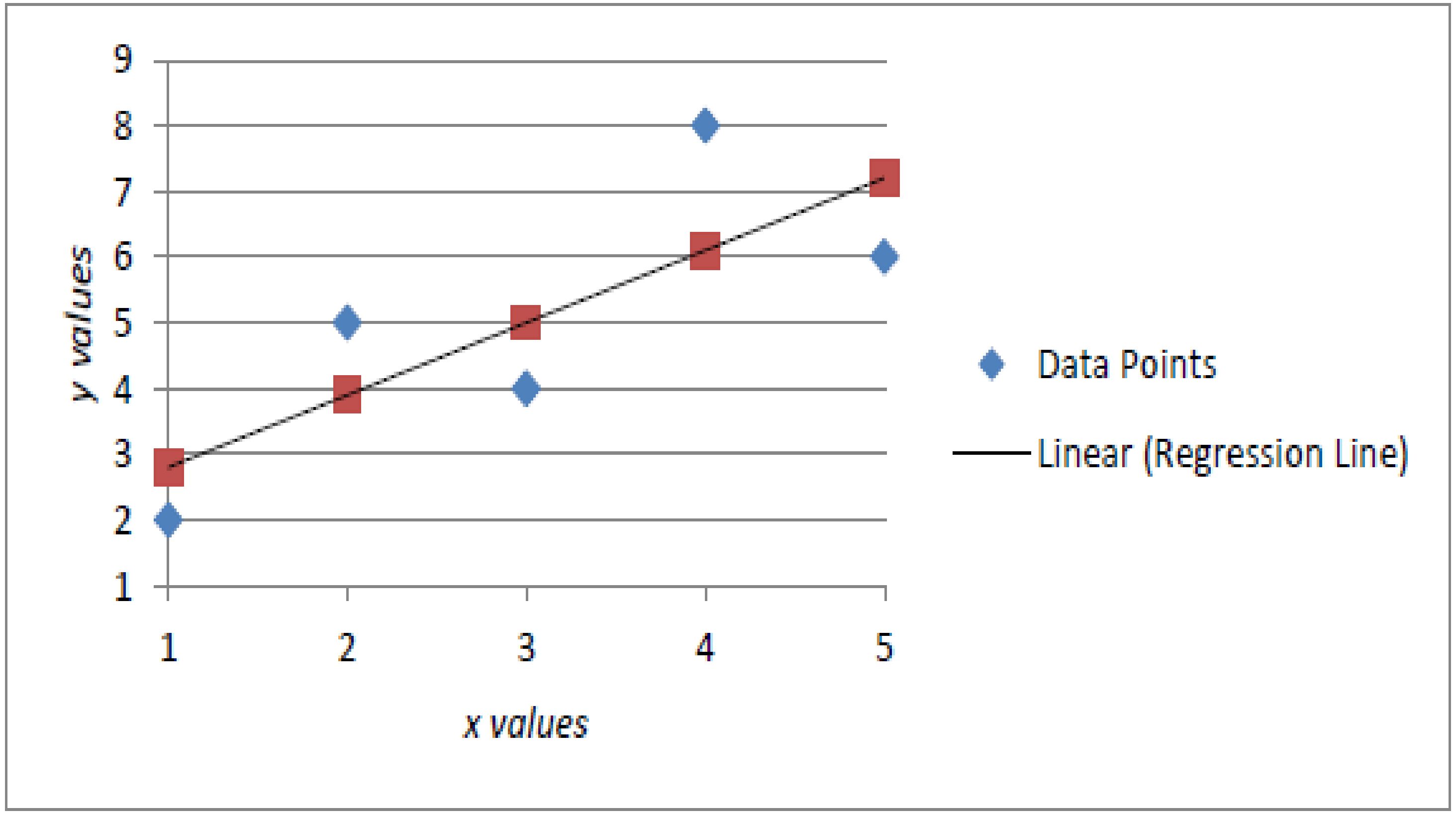
$$a = 0.25 \ and \ b = 0.55.$$

This gives the line of regression of *X* on *Y* as,

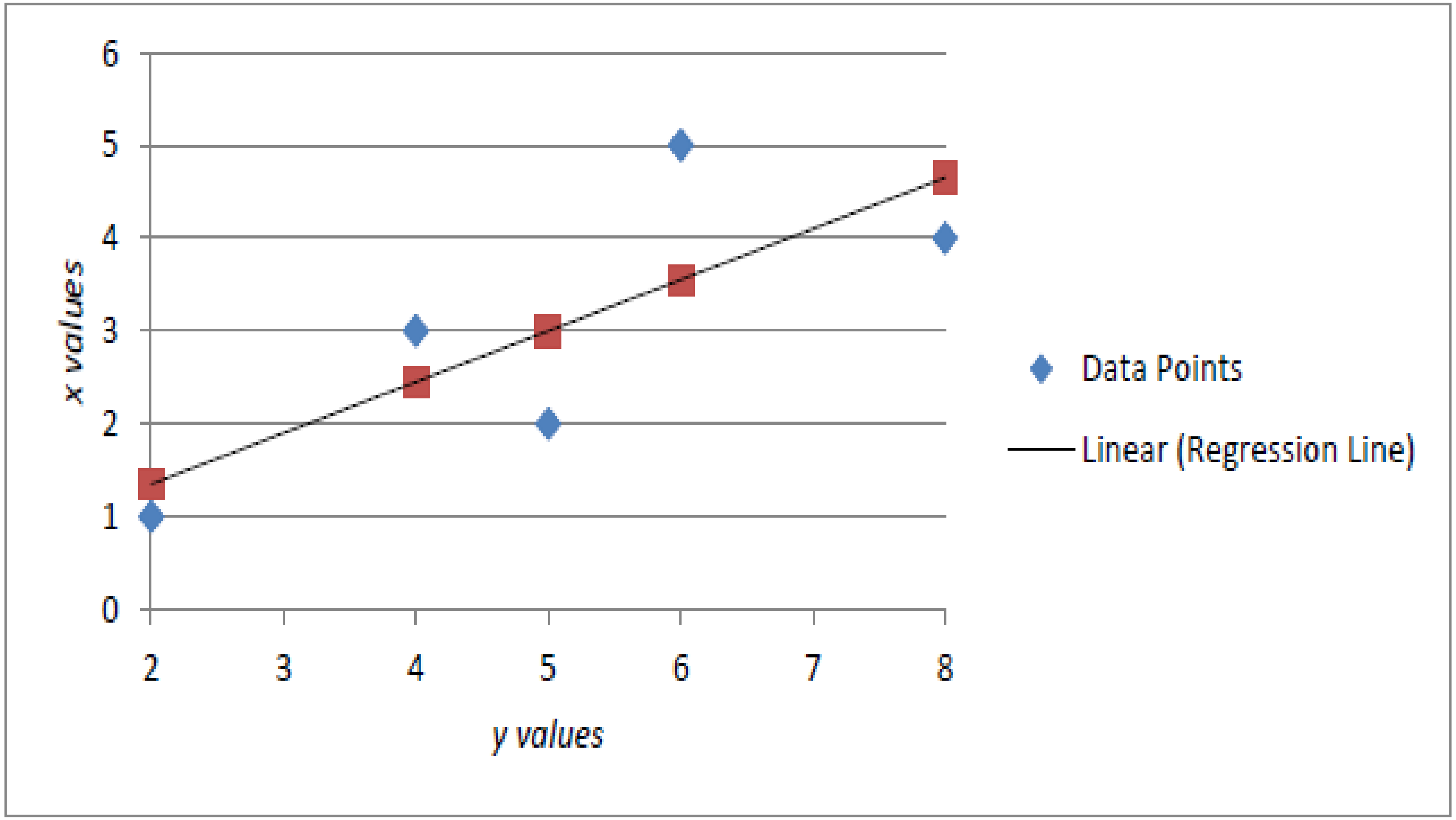
$$x = 0.25 + 0.55y.$$

This equation of line should be used to predict the values of *X* for given values of *Y*.

Let us plot these lines and also the given data.



Line of Regression of Y on X



Line of Regression of *X* on *Y*

Let us understand a few basic concepts about these lines of regression. If we consider the line of regression $Y = a + bX$, then we can obtain,

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

and,

$$a = \bar{y} - b\bar{x} \quad \text{(Dividing first normal equation by n)}$$

where,

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \text{ and } \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

As per equation

$$a = \bar{y} - b\bar{x},$$

we can say that the regression line pass through the $(\bar{x}, \bar{y})$. Similarly line on x on y also pass through $(\bar{x}, \bar{y})$. it means $(\bar{x}, \bar{y})$ is an intersection point of both the line.

As such, the line of regression of *Y* on *X* is,

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

We can similarly obtain the line of regression of *X* on *Y* as,

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

The slopes of two regression equations, namely, $\frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n(\bar{x})^2}$ and $\frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} y_i^2 - n(\bar{y})^2}$ are called coefficient of regression of *Y* on *X* and of *X* on *Y*, respectively.

Also

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

Dividing by n

$$b = \frac{\frac{1}{n}\sum x_i y_i - \bar{x}\bar{y}}{\frac{1}{n}\sum x_i^2 - (\bar{x})^2}$$

$$= \frac{E(XY) - E(X)\cdot E(Y)}{\sigma_X^2} = \frac{\mu_{11}}{\sigma_X^2}$$

Therefore regression line Y on X is

$$(y - \bar{y}) = \frac{\mu_{11}}{\sigma_X^2}(x - \bar{x})$$

Now $\rho = \frac{\mu_{11}}{\sigma_X \sigma_Y}$ $\mu_{11} \to \mathrm{cov}(X,Y)$ (as per earlier result)

$$\Rightarrow \quad (y-\bar{y}) = \rho \cdot \frac{\sigma_y}{\sigma_x}(x-\bar{x}) \rightarrow \text{r. line of Y on X}$$

Similarly

$$(x-\bar{x}) = \rho \frac{\sigma_x}{\sigma_y}(y-\bar{y}) - \text{r. line for X on Y}$$

we will write $b_{yx} \rightarrow$ regression coefficient for line Y on X

$b_{xy} \rightarrow$ regression coefficient for line X on Y

$$\therefore \quad b_{yx} \cdot b_{xy} = \rho^2$$

$$\Rightarrow \rho = \pm\sqrt{b_{yx} \cdot b_{xy}}$$

Ex:. obtain the equations of two lines of regression for the following data

X:	65	66	67	67	68	69	70	72
Y:	67	68	65	68	72	72	69	71

$U = X - 68$; $V = Y - 69$

Then by preparing the Table.

X	Y	$U = X - 68$	$V = Y - 69$	U^2	V^2	UV
65	67	-3	-2	9	4	6
66	68	-2	-1	4	1	2
67	65	-1	-4	1	16	4
67	68	-1	-1	1	1	1
68	72	0	3	0	9	0
69	72	1	3	1	9	3
70	69	2	0	4	0	0
72	71	4	2	16	4	8
	Σ	0	0	36	44	24

Now $\bar{U} = 0$; $\bar{V} = 0$; $\sigma_U^2 = 4.5$; $\sigma_V^2 = 5.5$, $\mathrm{Cov}(U,V) = \mu_{11} = 3$

$$\rho(U,V) = 0.6$$

Since correlation coefficient is independent of change of origin, we get $\rho = \rho(X,Y) = \rho(U,V) = 0.6$

$$U = \frac{X - 68}{\underset{h}{1}} \Rightarrow \bar{U} = \bar{X} - 68$$

$$\Rightarrow \bar{X} = 68,$$

$$\bar{V} = \frac{\bar{Y} - 69}{\underset{k}{1}} \Rightarrow \bar{Y} = 69$$

$$\sigma_X = h\sigma_U \Rightarrow \sigma_X = \sigma_U \Rightarrow \sqrt{4.5}$$

$$\sigma_Y = k\sigma_V \Rightarrow \sigma_Y = \sigma_V \Rightarrow \sqrt{5.5}$$

Hence line of regression Y on X is

$$Y - \bar{Y} = r \frac{\sigma_Y}{\sigma_X} (X - \bar{X})$$

$$\Rightarrow \quad Y - 69 = 0.6 \times \frac{2.35}{2.12} (X - 68)$$

$$\Rightarrow Y = 0.665 X + 23.78$$

Similarly line of regression X on Y is

$$X - \bar{X} = r \frac{\sigma_X}{\sigma_Y} (Y - \bar{Y})$$

$$\Rightarrow \quad X = 0.54 Y + 30.74$$

Regression Curves

This is worth noting here that principle of least squares can also be used to fit a curve of degree two (or more) to the given data. Let us again consider the data given in the form, $(x_i, y_i), i = 1, 2, \dots, n$ and let us fit a quadratic curve to this data.

Let the relationship between dependent and independent variables be described by,

$$Y = a + bX + cX^2$$

We thus consider a quadratic relationship between these two variables. Following the similar procedure as we did for linear regression, we can here obtain the normal equations as,

$$\sum_{i=1}^{n} y_i = an + b\sum_{i=1}^{n} x_i + c\sum_{i=1}^{n} x_i^2$$

$$\sum_{i=1}^{n} x_i y_i = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 + c\sum_{i=1}^{n} x_i^3$$

and

$$\sum_{i=1}^{n} x_i^2 y_i = a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i^3 + c\sum_{i=1}^{n} x_i^4$$

These are three linear equations in three unknowns $a, b \, and \, c$ that can be solved to get the quadratic relationship.

Example: Let us consider the example discussed earlier. The given data is,

i	1	2	3	4	5
x_i	1	2	3	4	5
y_i	2	5	4	8	6

For fitting a quadratic regression curve to this data, we obtain,

I	1	2	3	4	5	$\sum$
x_i	1	2	3	4	5	15
y_i	2	5	4	8	6	25
x_i^2	1	4	9	16	25	55
x_i^3	1	8	27	64	125	225
x_i^4	1	16	81	256	625	979
$x_i y_i$	2	10	12	32	30	86
$x_i^2 y_i$	2	20	36	128	150	336

Using this table, we obtain the normal equations as,

$$5a + 15b + 55c = 25,$$

$$15a + 55b + 225c = 86$$

and

$$55a + 225b + 979c = 336.$$

Solving these equations for a, b and c, we obtain the quadratic regression curve as,

$$y = -0.80 + 3.24x - 0.36x^2.$$

Arguing in the same manner, we can also obtain the regression curves of higher degree.

Fitting an Exponential Curve

Let us now comprehend a method that can be used to fit a curve of the form $y = a\,x^b$ to the given data $(x_i, y_i), i = 1, 2, \dots, n$. Here, $y = a\,x^b$ gives $\log(y) = \log(a) + b\log(x)$. This now becomes a linear regression problem. As such, we transform the given data $(x_i, y_i), i = 1, 2, \dots, n$ to $(\log\ x_i, \log\ y_i), i = 1, 2, \dots, n$ and then fit a line of regression to the transformed data. This line will be of the form $\log y = \log a + b \log x$. We can use this relationship to find the exponential curve $y = a\,x^b$.

COEFFICIENT OF DETERMINATION

Once we have obtained a least square regression line $y = a + bx$, we can consider to find how good does this line fit to the given data. For a given point x_i, we will get the estimated value, using linear fit, as,

$$\hat{y}_i = a + b\, x_i$$

We can note that the difference $|y_i - \hat{y}_i|$ between the observed values and predicted values should be small for a good fit.

Further,

$$|y_i - \bar{y}| = |(y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})|$$

Let us square both the sides and then summing over i, we get,

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + 2\sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \bar{y})$$

Here, the third term of right hand size can be proved to be zero using the following arguments,

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \bar{y})$$

$$= \sum_{i=1}^{n}(y_i - a - bx_i)(a + bx_i - \bar{y})$$

$$= a\sum_{i=1}^{n}(y_i - a - bx_i) + b\sum_{i=1}^{n}x_i(y_i - a - bx_i) - \bar{y}\sum_{i=1}^{n}(y_i - a - bx_i)$$

$$= 0$$

(Since we define a and b in such a way that the summations in above expressions are zero. These in fact form the normal equations.)

As such, we have,

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$

We can note that $(y_i - \bar{y})$ is the deviation of i^{th} observation from sample mean. As such, left hand side is the sum of squares of such deviations from mean. This sum is called total variation. Also, $(\hat{y}_i - \bar{y})$ is the difference between the predicted value and the sample mean. This is the quantity that is explained by the regression line and as such, $\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$ is called the explained variance. The quantity $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ is the sum of squares of residuals and this is called unexplained variance.

We thus have,

$$Total\,Variation = Unexplained\,Variation + Explained\,Variation$$

The coefficient of determination is defined as,

$$Coefficient\ of\ Determination = \frac{Explained\,Variation}{Total\,Variation}$$

$$= \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Thus, coefficient of variation should lie between 0 and 1. If the value of coefficient of variation is near to 1, it implies that the line of regression explains better the variation in data and thus is a good fit to the data.

Example: Let us consider the example of fitting regression line to the data,

I	1	2	3	4	5
x_i	1	2	3	4	5
y_i	2	5	4	8	6

We have obtained the line of regression of Y on X *as*,

$$y = 1.7 + 1.1x.$$

Using this line, we can obtain,

$$Coefficient\ of\ Determination = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

$$= \frac{12.1}{20}$$

$$= 60.5\%.$$

As such, the regression line $y = 1.7 + 1.1x$ explains only 60.5% of the variation in the give data.