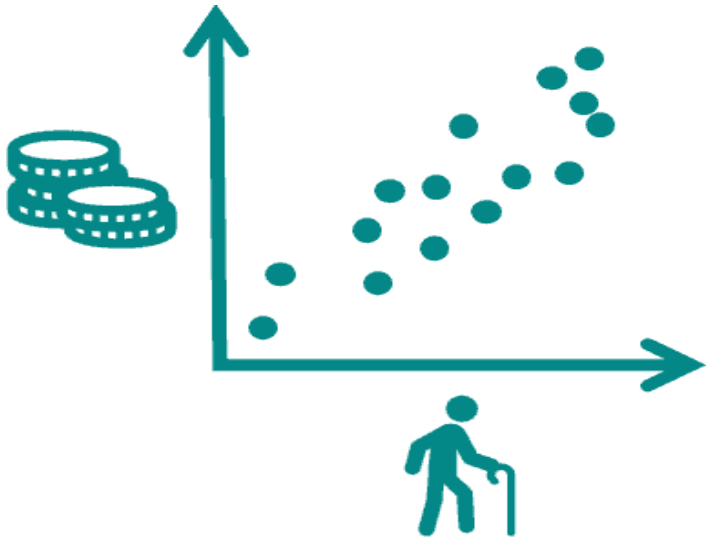
GITAM
DEEMED TO BE UNIVERSITY



B. Tech Computer Science & Engineering
(School of Technology)

SEMESTER –V
MATH2361: PROBABITY AND STATISTICS

UNIT-III:
Correlation, Regression and Estimation

Dr. Mallikarjuna Reddy Doodipala
Associate Professor
Department of Mathematics
GITAM University
Hyderabad Campus

UNIT-III: Correlation, Regression and Estimation

Correlation, correlation coefficient, rank correlation

Regression, lines of regression, regression coefficients

Curve fitting :principle of least squares and curve fitting (straight line, parabola and exponential curves)

Estimation: Parameter, statistic, sampling distribution, point estimation

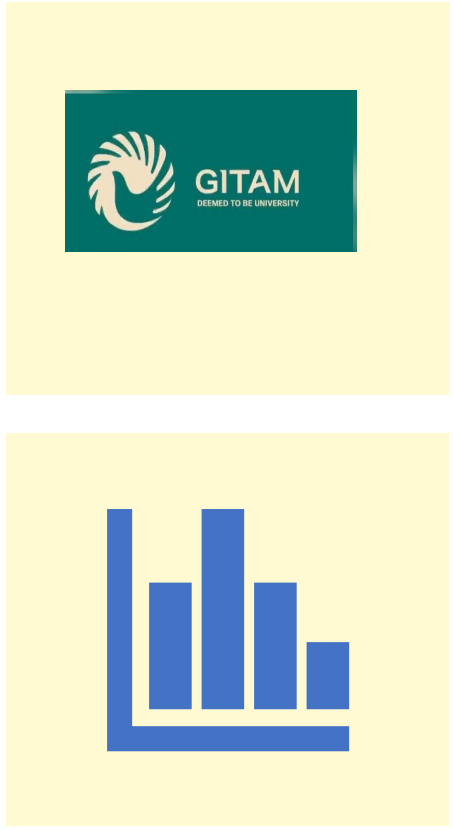
Properties of estimators, interval estimation

UNIT-III: Learning outcomes

After completion of this unit, the student will be able to

- Identify different trends in scatter plots, strengths of association between two numerical variables (L3).
- make use of the line of best fit as a tool for summarizing a linear relationship and predicting future observed values (L3).
- Estimate the value of a population parameter, computation of point and interval estimations (L3).

What is Correlation and Regression Analysis?



- Statistical tools for quantifying the strength of a *mathematical* relationship between two variables.
- Statistical tools for estimating values for the constants in a model equation.
- Statistical tools for evaluating how well a specific model equation fits your data.

Why Correlation and Regression?

Correlation

- The most used word in statistics to know the connection or an apparent relationship between two measurable characteristics is **correlation**.
- The relationship or the association between any two measured variables(or characteristics) is known as **correlation**.
- Two measured variables are said to be **correlated** if the change in one variable affects on the change in other variable.
- This effect or change may in same direction or in opposite direction.

Correlation-Examples

- The following are real life examples for some amount of correlation.

1. Height and Weight of individuals
2. Weight and Cholesterol level
3. Pressure and Volume of ideal gas
4. Temperature and Pulse rate
5. Amount of rain fall and yield

Types of correlation

- **Correlation** determines the relationship between two variables . But does not prove that variable causes the change in the other. The cause of change in the same or opposite direction due to some other reacting factor(s).
- **Types of correlation**: Simple, Partial and multiple
- Simple correlation: 1.Positive correlation
  2. Negative correlation
  3. Un correlation

Positive correlation:

- Two measured variables are said to be positively correlated if the change in one variable affects on the change in other variable (increase or decrease) in same direction

ex: 1. Height and weight of individuals.

2. Temperature and pulse rate

3. Amount of rain fall and yield.

Negative correlation

- Two measured variables are said to be negatively correlated if the change in one variable affects on the change in other variable (increase or decrease) in opposite direction

EX: 1.pressure and volume of ideal gas

2. Supply and demand of items

Un correlation :

- Two measured variables are said to be uncorrelated if the change in one variable does not affect the change in another variable in the same (or in opposite) direction.

Ex: heights and incomes of employees

Measures of correlation:
1.Pearson correlation coefficient:

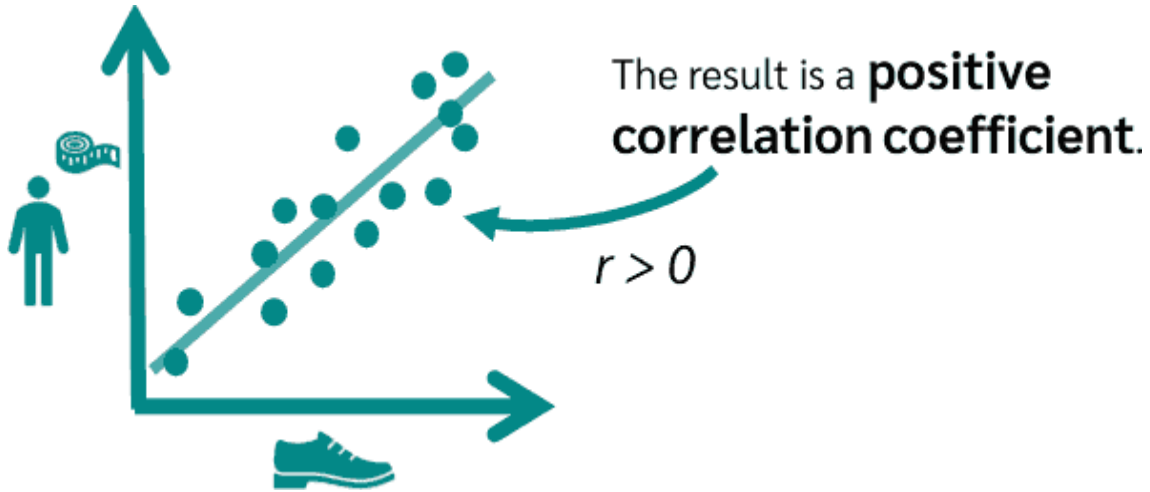
- The degree of relation ship between two sets of figures is measured Pearson formula called correlation coefficient and it is denoted by **r**

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

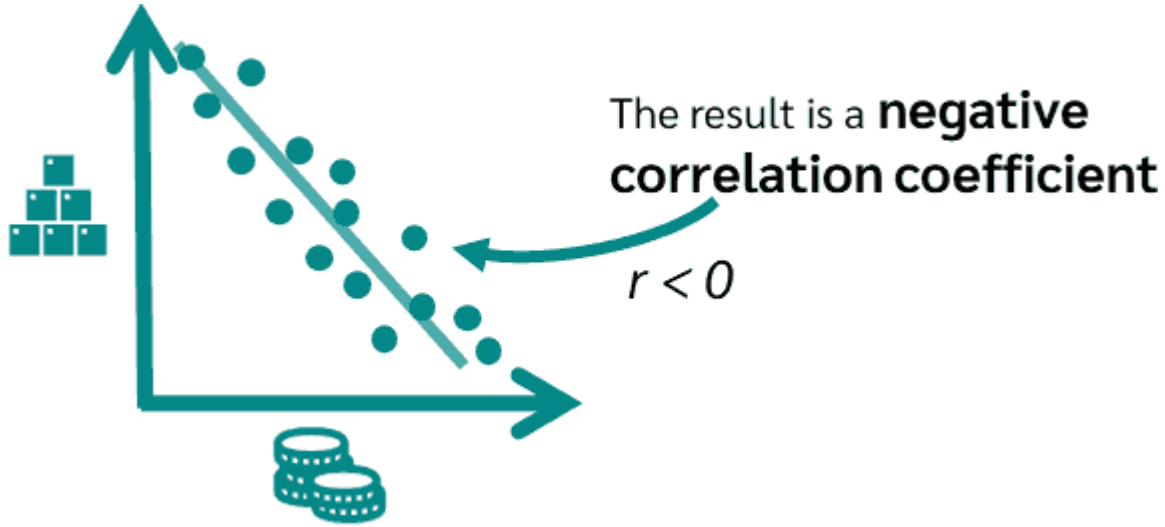
$$Correlation = \frac{Cov\,(x,y)}{\sigma x * \sigma y}$$ **or**

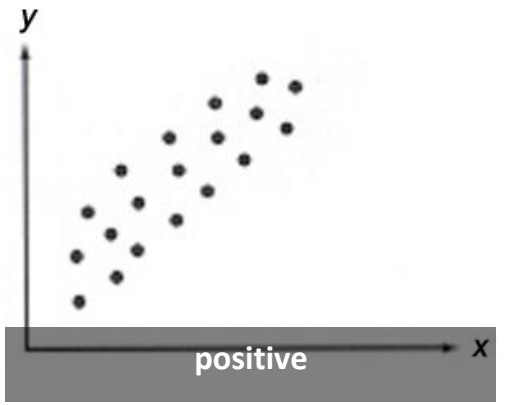
$$r = \frac{n\sum XY - \sum X.\sum Y}{\sqrt{n\sum X^2 - (\sum X)^2}\sqrt{n\sum Y^2 - (\sum Y)^2}}$$

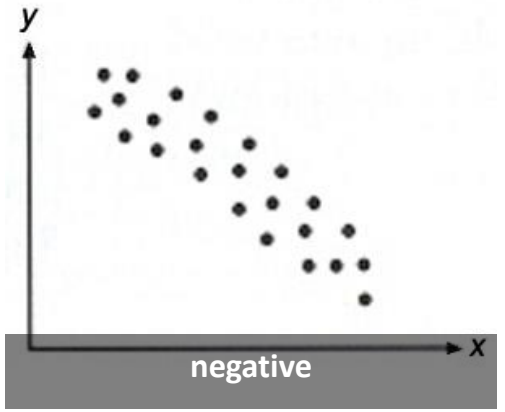
- **Limits: -1 ≤ r ≤ 1**
- If r>0 two variables are positively correlated
- If r<0 two variables are negatively correlated
- If r=0 two variables are uncorrelated

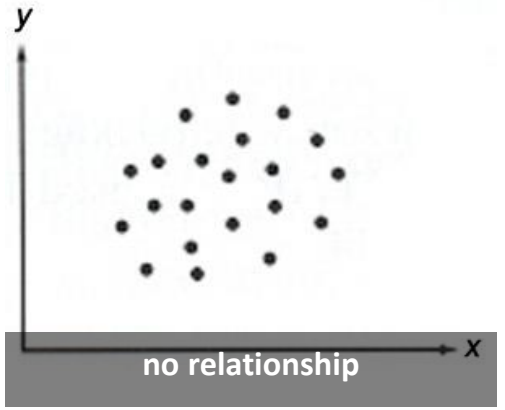


Measures of correlation 2.scattered diagram



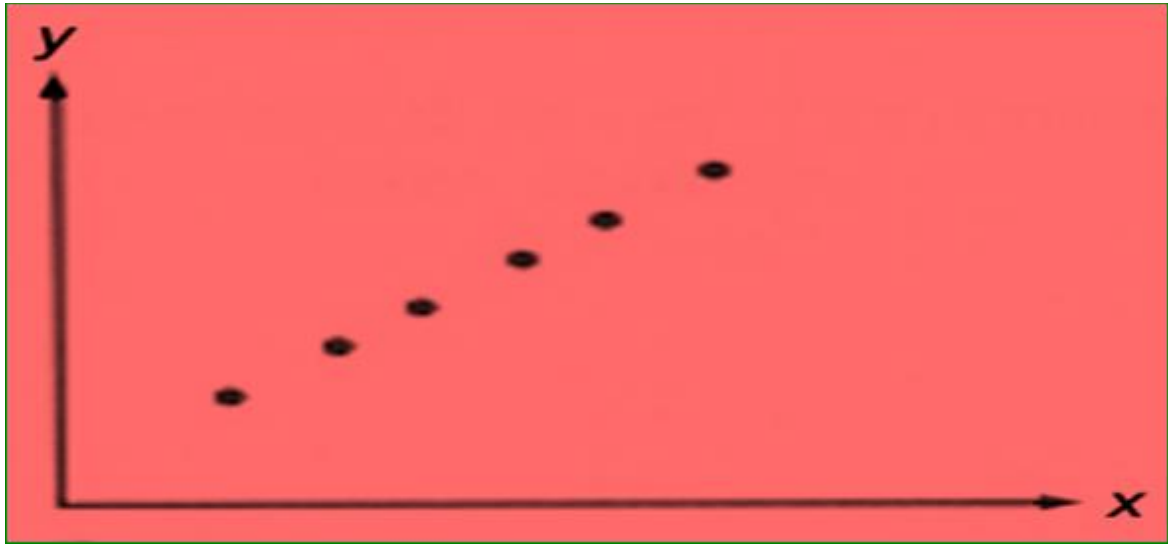






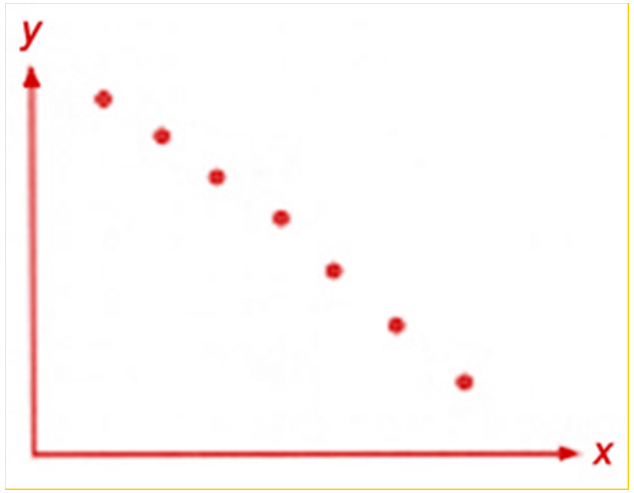
scattered plot (perfect relation)

***if* $r = 1$: a perfect positive relationship between y and x**

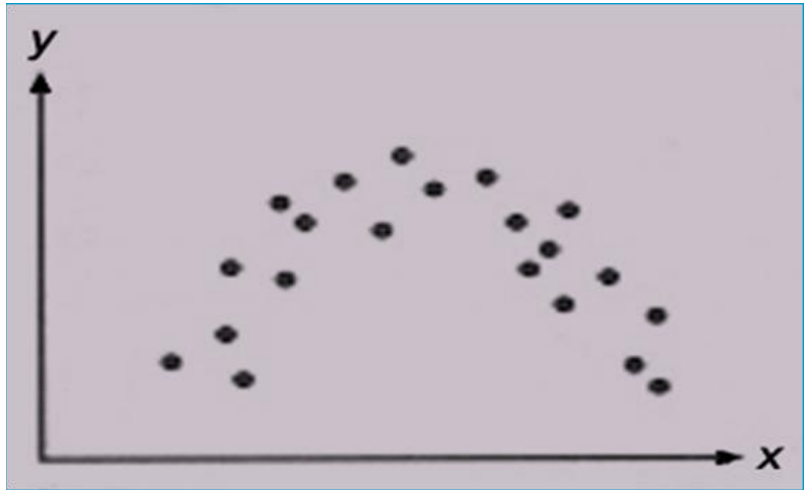


Measures of correlation
scattered plot (perfect relation)

- *if* $r = -1$: a perfect negative relationship between y and x



- *if* r near 0: little or no relationship between y and x



Sample problem:

- Find the correlation coefficient between height of fathers and their son with Following data.
- Hence give your conclusion.

Height of father in inch (X)	**65**	**66**	**67**	**67**	**68**	**69**	**70**	**72**
Height of son in inch (Y)	**67**	**68**	**65**	**68**	**72**	**72**	**69**	**71**

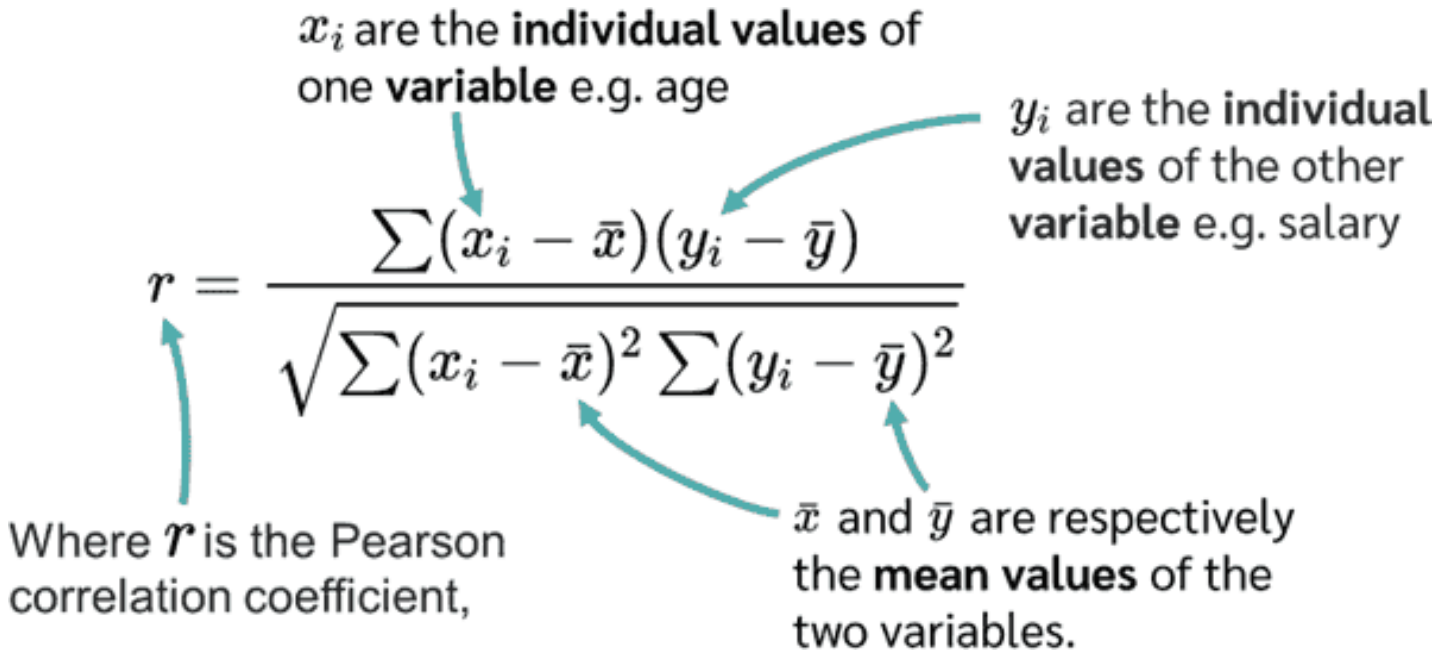
Calculations : correlation coefficient r

- From the table
- n= no of observations (subjects)=8
- Σx=544, Σy=552
- Σxy=37560
- ΣX2=37028
- Σ Y2=38132
- $r=\frac{n\sum XY-\sum X.\sum Y}{\sqrt{n\sum X^2-(\sum X)^2}\sqrt{n\sum Y^2-(\sum Y)^2}}$=0.603

S.NO	X	Y	X^2	Y^2	XY
1	65	67	4225	4489	4355
2	66	68	4356	4624	4488
3	67	65	4489	4225	4355
4	67	68	4489	4624	4556
5	68	72	4624	5184	4896
6	69	72	4761	5184	4968
7	70	69	4900	4761	4830
8	72	71	5184	5041	5112
TOTAL	**544**	**552**	**37028**	**38132**	**37560**

- **Since r > o we conclude that the father and their sons heights are positively correlated**

Calculations : correlation coefficient r



2.Calculations : correlation coefficient r

Maths (X)	Science (Y)	Xmean	Ymean	Xi - Xmean	Yi - Ymean	(Xi-Xm)(Yi-Ym)	(Xi-Xm)^2	(Yi-Ym)^2
98	88	84.71	86.43	13.29	1.57	20.87	176.62	2.46
87	92			2.29	5.57	12.76	5.24	31.02
91	95			6.29	8.57	53.91	39.56	73.44
75	82			-9.71	-4.43	43.02	94.28	19.62
81	74			-3.71	-12.43	46.12	13.76	154.5
68	78			-16.71	-8.43	140.87	279.22	71.06
93	96			8.29	9.57	79.34	68.72	91.58
					Total	**396.89**	**677.4**	**443.68**

Pearson Correlation Coefficient Formula

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

$$r = \frac{396.88}{SQRT(677.4*443.68)}$$

$$= 0.724$$

The performance in mathematics and science are positively correlated

Significance Study of Correlation : Probable error

The probable error is given by

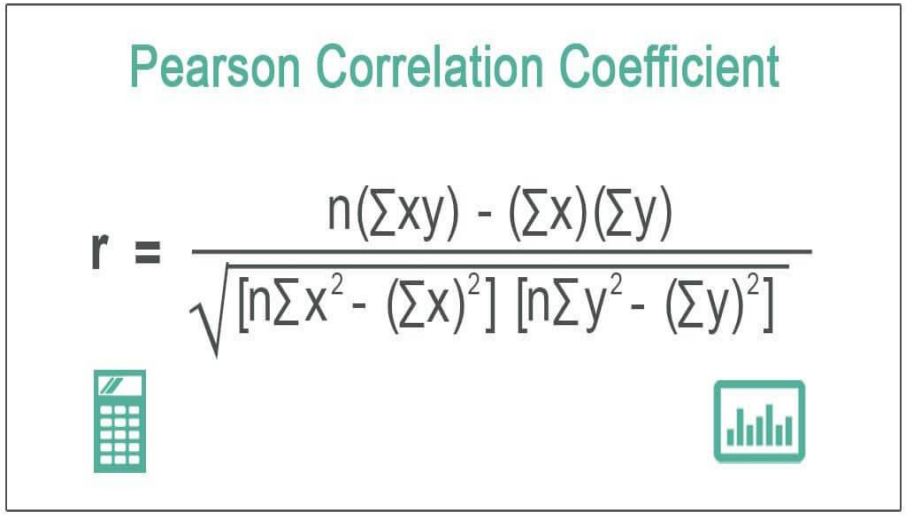
P.E(r)=0.6745xS.E(r)

if r<P.E(r) not significant

if r>=P.E(r) significant

Where

$SE(r)=\frac{1-r^2}{\sqrt{n}}$



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

Alternative Formula for Correlation Coefficient

- This formula used if the data sets are small otherwise the following formula is consistent

"by change of origin and scale“

("transformation to U and V)

.

$$r=\frac{n\sum XY-\sum X.\sum Y}{\sqrt{n\sum X^2-(\sum X)^2}\sqrt{n\sum Y^2-(\sum Y)^2}}$$

$$r=\frac{n\sum UV-\sum U.\sum V}{\sqrt{n\sum U^2-(\sum U)^2}\sqrt{n\sum V^2-(\sum V)^2}}$$

$$U=\frac{X-A}{h},\quad V=\frac{Y-B}{k}$$

where A, B are constants taken from X and Y
h,k are scales

Exercise to reader

x	8	6	12	14	16	10
y	15	10	20	25	30	20

Protein Content(gm)	2.0	3.0	4.0	5.0	5.5	6.0	7.0	7.5
weight(gm)	8	10	14	18	22	25	28	35

Age of husband(yrs)	23	27	28	29	30	31	33	36	38	40
age of wife(yrs)	18	23	23	24	25	26	30	31	33	35

Fertili zer (tons)	15	18	20	24	30	35	40	50
yield (tons)	85	93	95	105	120	130	150	160

3.Spearman's rank correlation coefficient

Let a group of n individuals be arranged in a order of merit in getting two characteristics of A and B.

Let (Xi, Yj) be the ranks of n individuals two characteristics of A and B

Using ranks between two characteristics of A and B spearman discovered a formula called Spearman correlation coefficient. It is defined as

$$\rho = 1 - \frac{6\sum d^2}{n(n^2-1)} \quad n \neq 1$$

$d = x - y, n =$ no.of individuals

also $-1 \leq \rho \leq 1$

Spearman's rank correlation coefficient

Assess significance: Same as Pearson.

- Randomize if possible. Less than 20
- Use t-Student distribution with n -2 degrees of freedom
- ρ: correlation coefficient, n: number of subjects

$$t = \frac{\rho}{\sqrt{(1-\rho^2)/(n-2)}}$$

- For large n it approaches normal Variate Z

Sample Problem

- $\rho = 1 - \frac{6\sum d^2}{n(n^2-1)}$
- from the table $\sum d^2 = 40$,
- $n = 8$

then

- $\rho = 1 - \frac{6\times 40}{8(64-1)} = 0.52$

Height of father in inch (X)	Height of son in inch (Y)	X	Y	X-Y=d	d^2
65	67	8	6	2	4
66	68	7	5	2	4
67	65	6	8	-2	4
71	70	2	3	-1	1
68	72	5	1	4	16
69	66	4	7	-3	9
70	69	3	4	-1	1
72	71	1	2	-1	1
				total	40

Repeated Ranks Concept

- **If the ranks are repeated in two individuals A and B then the spearman's formula discussed above is modified by using the correction factor(C.F). It is given by**

$$C.F = \frac{m(m^2 - 1)}{12}$$

here $m =$ no.of times observstion repeats

$$\rho = 1 - \frac{6\sum d^2 + C.F}{n(n^2 - 1)}$$ from the table

we compute $\sum d^2$, C.F

for X and Y if the ranks are repeated substitute all in the above formula.

Sample Problem on Repeated Ranks

Height of father in inch (X)	Height of son in inch (Y)	X	Y	X-Y=d	d^2
68	62	4	5	-1	1
64	58	6	7	-1	1
75	68	2.5	3.5	-1	1
50	45	9	10	-1	1
64	81	6	1	5	25
80	60	1	6	-5	25
75	68	2.5	3.5	-1	1
40	48	10	9	-1	1
55	50	8	8	0	0
64	70	6	2	4	16
	Total				72

"we conclude that father and son's heights are positively correlated"

$$C.F = \frac{m(m^2-1)}{12} = \frac{2(2^2-1)}{12} = 1/2 \text{ for 75 in X and 68 inY}$$

$$C.F = \frac{m(m^2-1)}{12} = \frac{3(3^2-1)}{12} = 2 \text{ for 68 in X then total C.F}$$

$$= 1/2 + 1/2 + 2 = 3$$

here m = no.of times observstion repeats

$$\rho = 1 - \frac{6\sum d^2 + C.F}{n(n^2-1)}$$ from the table

we compute $\sum d^2 = 72$

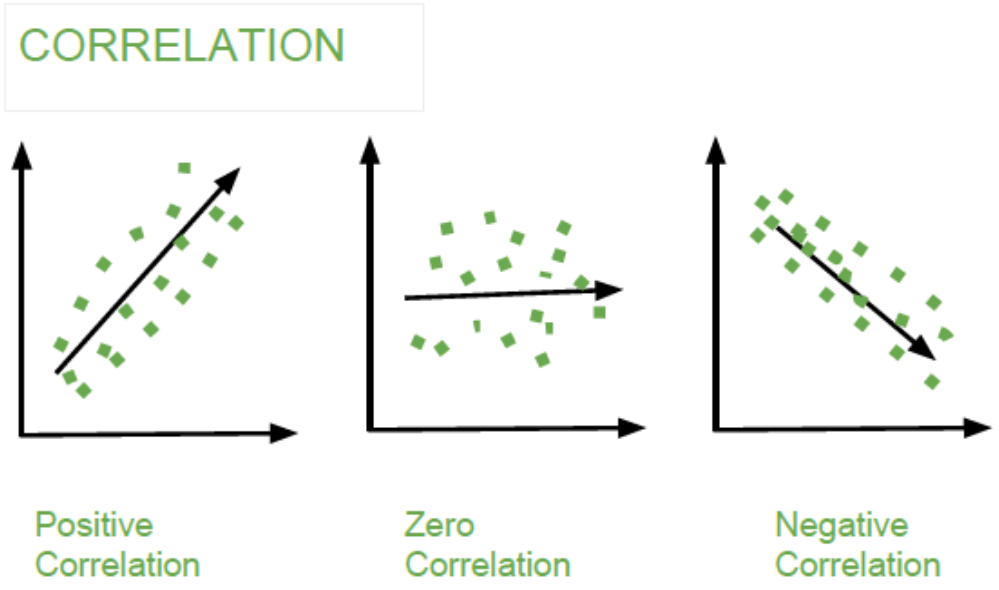
substitute all in the above formula.

Then we get

$$\rho = 1 - \frac{6(72+3)}{10(10^2-1)} = 0.542$$

Correlation Interpretation

- An interpretation of the size of the coefficient has been described by Cohen (1992) as:



Correlation coefficient value			Relationship
-0.3 to +0.3			Weak
-0.5 to -0.3	or	0.3 to 0.5	Moderate
-0.9 to -0.5	or	0.5 to 0.9	Strong
-1.0 to -0.9	or	0.9 to 1.0	Very strong

Regression

- Regression analysis is a statistical technique that attempts to explore and model the relationship between two or more variables.
- For example, an analyst may want to know if there is a relationship between **road accidents** and **the age of the driver**.
- Thus, Helpful in ascertaining the probable form of the relationship between variables and predict or estimate the value corresponding to a given value of another variable.

Regression

- **Regression:**

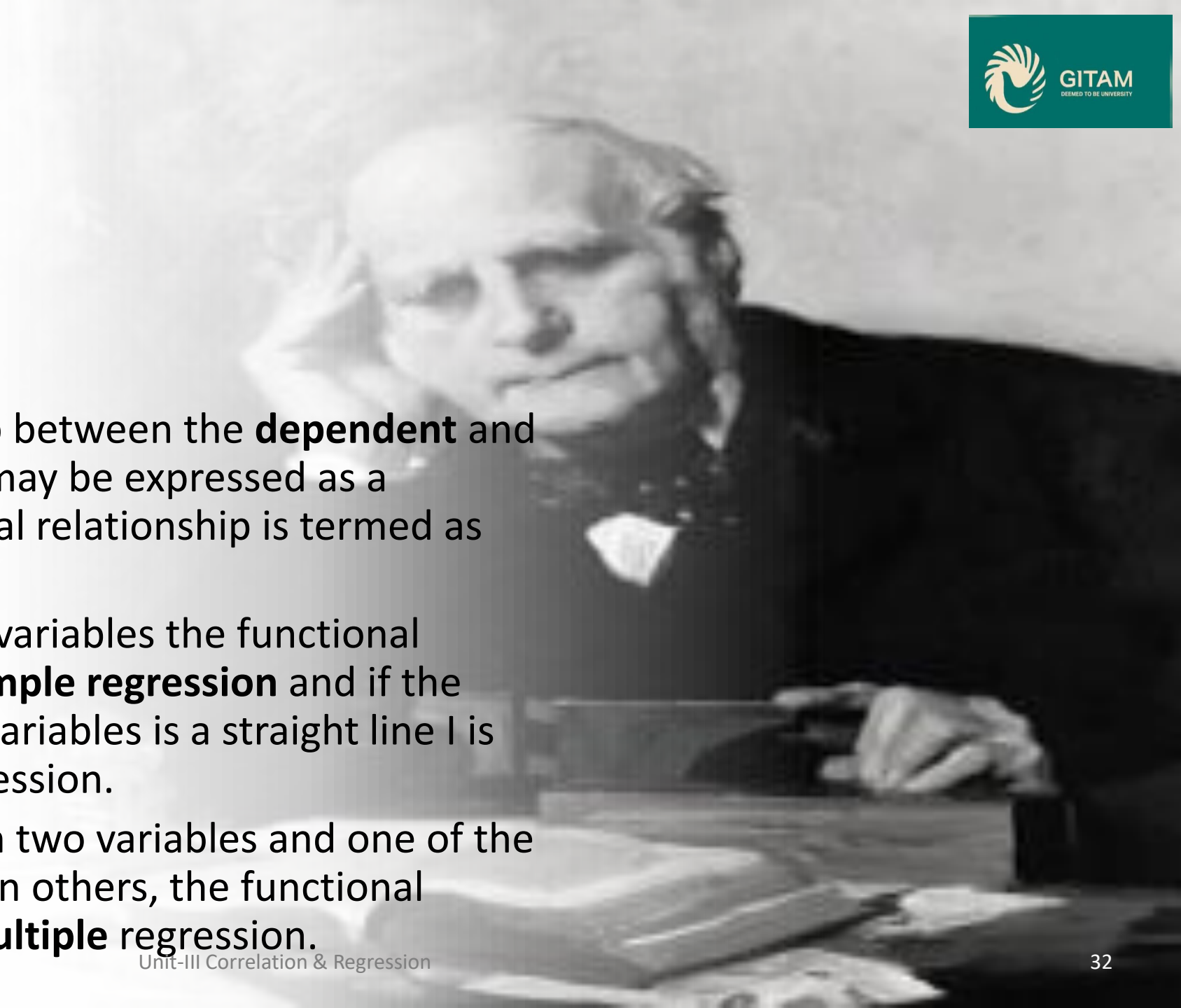
The literal meaning of regression is stepping back towards averages.

- At first the word regression was introduced by a famous statistician **Sir Francis Galton** in an experiment of inheritance.

Regression Analysis

- Regression is the functional relationship between two variables and of the two variables one may represent cause and the other may represent effect.
- The variable representing cause is known as independent variable and is denoted by X. The variable X is also known as predictor variable or regressor.
- The variable representing effect is known as dependent variable and is denoted by Y. Y is also known as predicted variable.

Types of regression



- In general, The relationship between the **dependent** and the **independent** variable may be expressed as a function and such functional relationship is termed as regression.
- When, there are only two variables the functional relationship is known as **simple regression** and if the relation between the two variables is a straight line I is known a simple linear regression.
- When, there are more than two variables and one of the variables is dependent upon others, the functional relationship is known as **multiple** regression.

Types of regression

- Two variables, *X* and *Y*, are int

where, *X* = Independent variable

Y = Dependent variable and vice versa

In simple linear regression, we consist of two regression equations or lines.

- -Regression equation of Y on X
- -Regression equation of X on Y

Regression line of Y on X

- Estimate the variable y we have regression line called regression model equation of y on x and it is given by

  $Y = a + bX$ (1)

- where a, b are constants. Equation (1) is general form of R.L of Y on X.

- The best trend or predicted value of Y obtained minimizing the total residual error by using Principle of least squares.

Regression line of X on Y

- To estimate the variable X we have a regression line called regression model equation of X on Y and it is given by

  $X = a+bY$ (2)

- where a, b are constants.
- Equation (2) is general form of R.L of X on Y

Fitting of Regression Line of Y on X

- To fit regression line of the form . Let an Analytic expression of the form
- Y=f(x) =a+bX, for all (x, y) to a set of ‘n’ points, i=1,2,3,…n. -----(*)

- The best trend values of Y obtained by minimizing the total residual error.
- i.e. The ‘best’ line has minimum error between line and data points.
- This error is minimized by **principle of least square**.
- This is also called the least squares approach, since square of the error is minimized.

$i.e\ Minimize\{Error = \sum_{i=1}^{n}[y_i - (Y')]^2\}$ is zero, say

Fitting of Regression Line of Y on X (proof not required)

$$Minimize\left\{Error = \sum_{i=1}^{n}[y_i - (a + bx_i)]^2\right\}$$

Take the derivative of the error with respect to a and b, set each to zero

$$\frac{\partial(Error)}{\partial a} = \frac{\partial\left(\sum_{i=1}^{n}\left[y_i - \left((a + bx_i)\right)\right]^2\right)}{\partial a} = 0$$

$$\frac{\partial(Error)}{\partial b} = \frac{\partial\left(\sum_{i=1}^{n}\left[y_i - \left((a + bx_i)\right)\right]^2\right)}{\partial b} = 0$$

$$\frac{\partial(Error)}{\partial a} = -2\sum_{i=1}^{n}\left[y_i - \left((a + bx_i)\right)\right] = 0$$

$$na + b\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \qquad \ldots.(1)$$

$$\frac{\partial(Error)}{\partial b} = -2\sum_{i=1}^{N} x_i\left[y_i - \left((a + bx_i)\right)\right] = 0$$

$$a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{N} x_i y_i \qquad \ldots.(2)$$

Eqn. (1), (2) are called normal (Legendre) eqns.
On solving two equations we get values of a, b.

$$b = \frac{\frac{1}{n}\sum xy - n\bar{x}\,\bar{y}}{\frac{1}{n}\sum x^2 - n\bar{x}^2} = \frac{\mathrm{cov}(x,y)}{\mathrm{var}(x)} = b_{yx}$$

$$a = \bar{y} - \frac{\mathrm{cov}(x,y)}{\mathrm{var}(x)}\bar{x}$$

Now using (*) the regression line of Yon X is given by

$$(Y - \bar{Y}) = b_{YX}\,(X - \bar{X})$$

$$b_{YX} = r\frac{\sigma_Y}{\sigma_X}$$

$$b_{yx} = \frac{n\sum XY - \sum X.\sum Y}{n\sum X^2 - (\sum X)^2}$$

$$\bar{X} = \frac{\sum X}{n}$$

$$\bar{Y} = \frac{\sum Y}{n}$$

The regression line X on Y

- The regression line X on Y is

$$(X - \bar{X}) = b_{XY}\,(Y - \bar{Y})$$

$$b_{XY} = r\frac{\sigma_X}{\sigma_y} = \frac{n\sum XY - \sum X.\sum Y}{n\sum Y^2 - (\sum Y)^2}$$

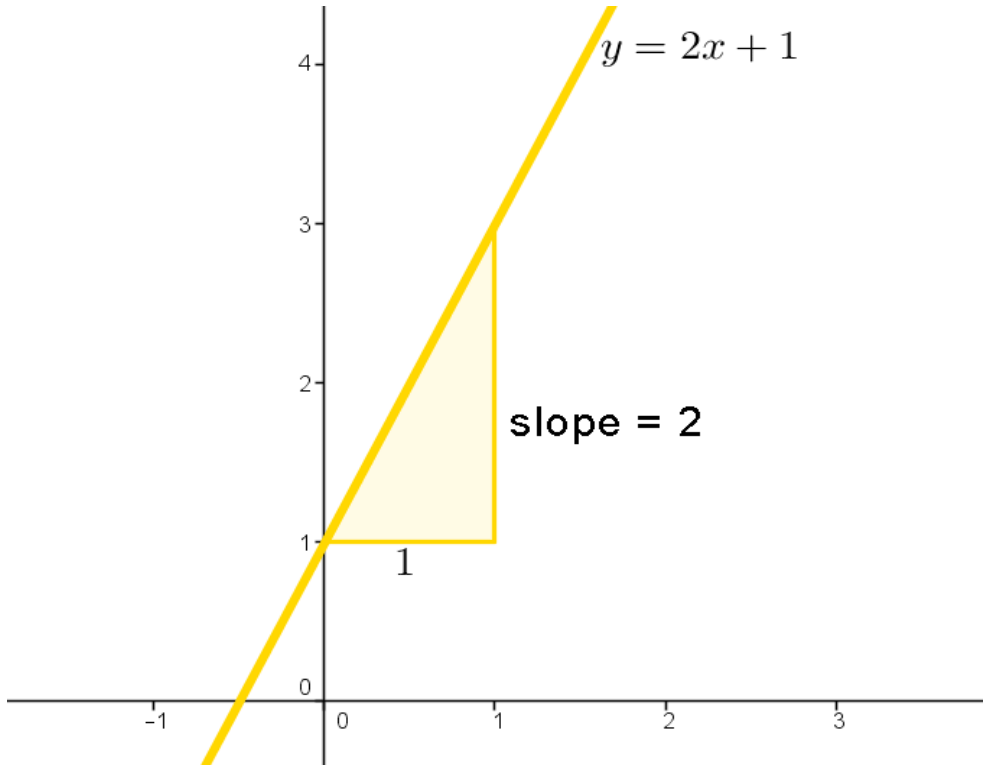
Called regression coefficient of Yon X. Also

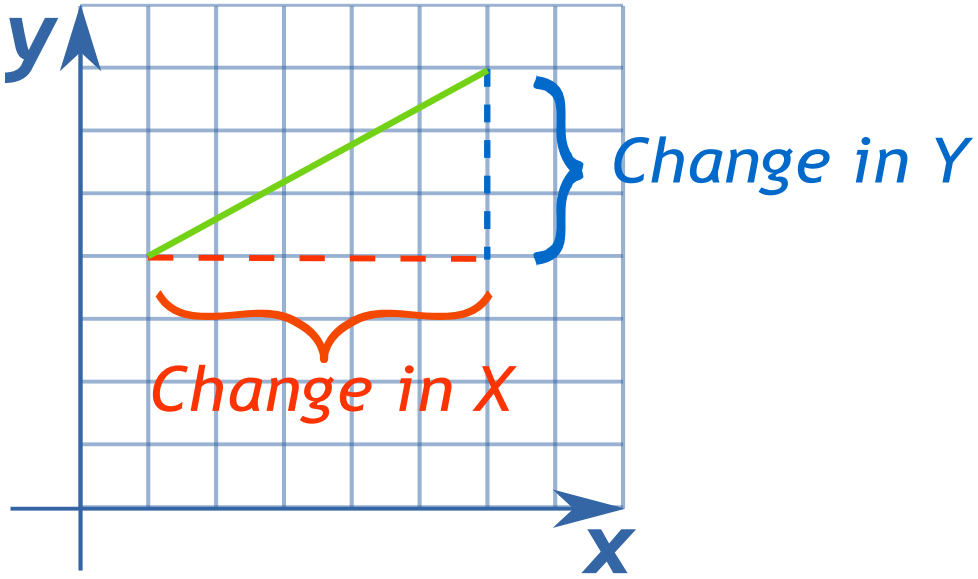
Where $\sqrt{b_{YX}\,b_{XY}} = r$ is the correlation coefficient.

σ_X , σ_Y are s.d for X and Y respectively.

Note: fitting procedure of RL of X on Y is left to Reader

Example Y=bx+a : Regression Line of Y on X





The ratio of the "vertical change" to the "horizontal change" between (any) two distinct points on a line.

Properties of Regression Coefficients

Two regression Coefficients b_{yx} and b_{xy} has the following five properties

1. If one regression coefficient is greater than unity other must be less than or equal to unity.
2. The GM of two regression coefficients is correlation coefficient 'r'

$$GM = \sqrt{b_{xy} X b_{yx}} = \pm r$$

3. The AM of two regression coefficients is greater than or equal to correlation coefficient 'r'

$$AM = \frac{b_{xy} + b_{yx}}{2} \geq r$$

4. Two regression coefficients are independent of change of origin but not scale.

$$b_{xy} = \frac{h}{k} b_{uv}, b_{yx} = \frac{k}{h} \quad \text{where } u = \frac{x-a}{h}, v = \frac{y-b}{h}$$

5. The angle between two regression lines is given by

$$\tan\theta = \frac{1-r^2}{r}\left\{\frac{\sigma_x \sigma_y}{{\sigma_x}^2 + {\sigma_y}^2}\right\}$$

Sample problem:

- **Find two regression lines**
  1. R.L of X on Y 2. R.L of Y on X between height of fathers and their son with Following data.

Height of father in inch (X)	65	66	67	67	68	69	70	72
Height of son in inch (Y)	67	68	65	68	72	72	69	71

Calculations for Regression Coefficients

S.NO	X	Y	X^2	Y^2	XY
1	65	67	4225	4489	4355
2	66	68	4356	4624	4488
3	67	65	4489	4225	4355
4	67	68	4489	4624	4556
5	68	72	4624	5184	4896
6	69	72	4761	5184	4968
7	70	69	4900	4761	4830
8	72	71	5184	5041	5112
TOTAL	**544**	**552**	**37028**	**38132**	**37560**

n = no of observations
From the table (subjects) = 8
$\Sigma Y2 = 38132$
$\Sigma X2 = 37028$
$\Sigma xy = 37560$
$\Sigma x = 544, \Sigma = 552$

$$byx = \frac{n\sum XY - \sum X.\sum Y}{n\sum X^2 - (\sum X)^2} = 0.67$$

$$bxy = \frac{n\sum XY - \sum X.\sum Y}{n\sum Y^2 - (\sum Y)^2} = 0.56$$

Continued…

- The two regression equations can also be written as follows.
- The regression line of Yon X is

$$(Y-\bar{Y}) = b_{YX}(X-\bar{X})$$

$$b_{YX} = r\frac{\sigma_Y}{\sigma_X} = \frac{n\sum XY - \sum X.\sum Y}{n\sum X^2 - \left(\sum X\right)^2}$$

Called regression coefficient of Yon X

- Where $\sqrt{b_{YX}\,b_{XY}} = r$ is the correlation coefficient.

σ_X , σ_Y are s.d for X and Y respectively

Continued…

- The regression line X on Y is

$(X-\bar{X}) = b_{XY}(Y-\bar{Y})$ $\bar{X}$=68,$\bar{Y}$ =69,bxy=0.55,byx=0.67

Now R.L of X on Y is x-68=0.55(y-69)… (1)

- The regression line of Yon X is

$(Y-\bar{Y}) = b_{YX}(X-\bar{X})$

Now R.L of Y on X is Y-69=0.67(x-68)…(2)

Problem-2

Two regression lines are given by 8x-10y+66=0, 40x-18y=214.

(a) Find the mean values of X and Y.

(b) Correlation coefficient 'r' (c) SD of Y if SD of X is 3

Sol: Given that 8x-10y+66=0----(1)

40x-18y-214=0---(2)

(a) On solving (1)&(2) we get X=13 and Y=17 are known as mean values of X and Y respectively.

i.e. $\bar{X} = X = 13, \bar{Y} = Y = 17$

(b) Eqn (1)&(2) are called RL of Yon X & X on Y respectively, then

Problem-2

(b) Two regression lines are given by
8x-10y+66=0,
40x-18y=214.
Eqn (1)&(2) are called RL of Yon X
& X on Y respectively, then

$$Y = \frac{8}{10}X + \frac{66}{10} \quad (3) \quad \text{and } X = \frac{18}{40}Y + \frac{214}{40} \quad (4)$$

$\therefore$ from (3) & (4) $b_{yx} = 8/10 = 4/5$

$b_{xy} = 10/40 = 9/20$

now the correlation coefficient r is

$$r = \sqrt{b_{xy}.b_{yx}} = \sqrt{9/20.}\,4/5 = 3/5 = 0.6$$

(c) To find SD of Y we know that SD of X is 3

$$b_{yx} = \frac{r\sigma_y}{\sigma_x}$$

$$\frac{4}{5} = \frac{3/5.\sigma_y}{3} = 4$$

$$\sigma_y = 4$$

Exercise to Reader

1. X =2.5+5.0Y estimate x at y=7

   Y=5.5+4.6X estimate Y at X=4.5 . Also find mean values

2. X:6 2 10 4 8 Find regression lines Y on X and X on Y

   Y:9 11 5 8 7 Estimate X at Y=6. Estimate Y at X =9

3. X:1 2 3 4 5 6 7

   Y: 2 4 7 6 5 6 3

4. Rain fall(inches) : 1 2 3 4 5 5 6 7 8 9

   Wheat yield(qui) : 1 3 2 5 5 4 7 6 9 8

5. Child age(yrs) : 1 2 3 4 5

   Weight(kg) : 4 6 9 11 12

last but One Slide

Any Questions? Suggestions?

Thank you

Feedback to mdoodipa@gitam.edu