Regression Analysis

What are Regression Lines?

If we take the case of two variables X and Y, we shall have two regression lines as regression line of X on Y and the regression line of Y on X.

The regression line of Y on X gives most probable values of Y for given values of X and the regression line of X on Y gives most probable values of X for given values of Y. Thus we have two regression lines.

Under what conditions can there be one regression line?

When there is either perfect positive or perfect negative correlation between the two variables, the two correlation lines will coincide i.e. we will have one line.

The further the two regression lines are from each other, the lesser is the degree of correlation and the nearer the two regression lines to each other, the higher is the degree of correlation. If variables are independent, r is zero and the lines of regression are at right angles i.e. parallel to X – axis and Y – axis.

What are regression equations?

Regression Equations are algebraic expressions of regression lines. Since there are two regression lines there are two regression equations–the regression equation of X on Y is used to describe the variations in the values of X for given change in Y and the regression equation of Y on X is used to describe the variation in the values of Y for given changes in X.

Why is the line of 'best fit'?

The line of regression is the line which gives the best estimate to the value of one variable for any specific value of other variable. Thus the line of regression is the line of "best fit" and is obtained by the principles of least squares.

What is the general form of the regression equation of Y on X?

The general form of the linear regression equation of Y on X is expressed as follows:

$Y_e = a+bX$, where

Y_e = dependent variable to be estimated

X = independent variable.

In this equation a and b are two unknown constants (fixed numerical values) which determine the position of the line completely. The constants are called the parameters of the line. If the value of either one or both of them are changed, another line is determined.

What is the general form of the regression equation of Y on X?

The parameter ‘a’ determines the level of the fitted line (i.e. the distance of the line directly, above or below the origin). The parameter ‘b’ determines the slope of the line i.e. the change in Y for unit change in X.

To determine the values of ‘a’ and ‘b’, the following two normal equations are to be solved simultaneously.

$\sum Y = Na + b\sum X$, $\sum XY = a\sum X + b\sum X^2$.

What is the general form of the regression equation of X on Y?

The general form of the regression equation of X on Y is expressed as follows:

$$X = a+bY$$

To determine the values of ‘a’ and ‘b’, the following two normal equations are to be solved simultaneously.

$\sum X = Na+b\sum Y$, $\sum XY = a\sum Y+b\sum Y^2$.

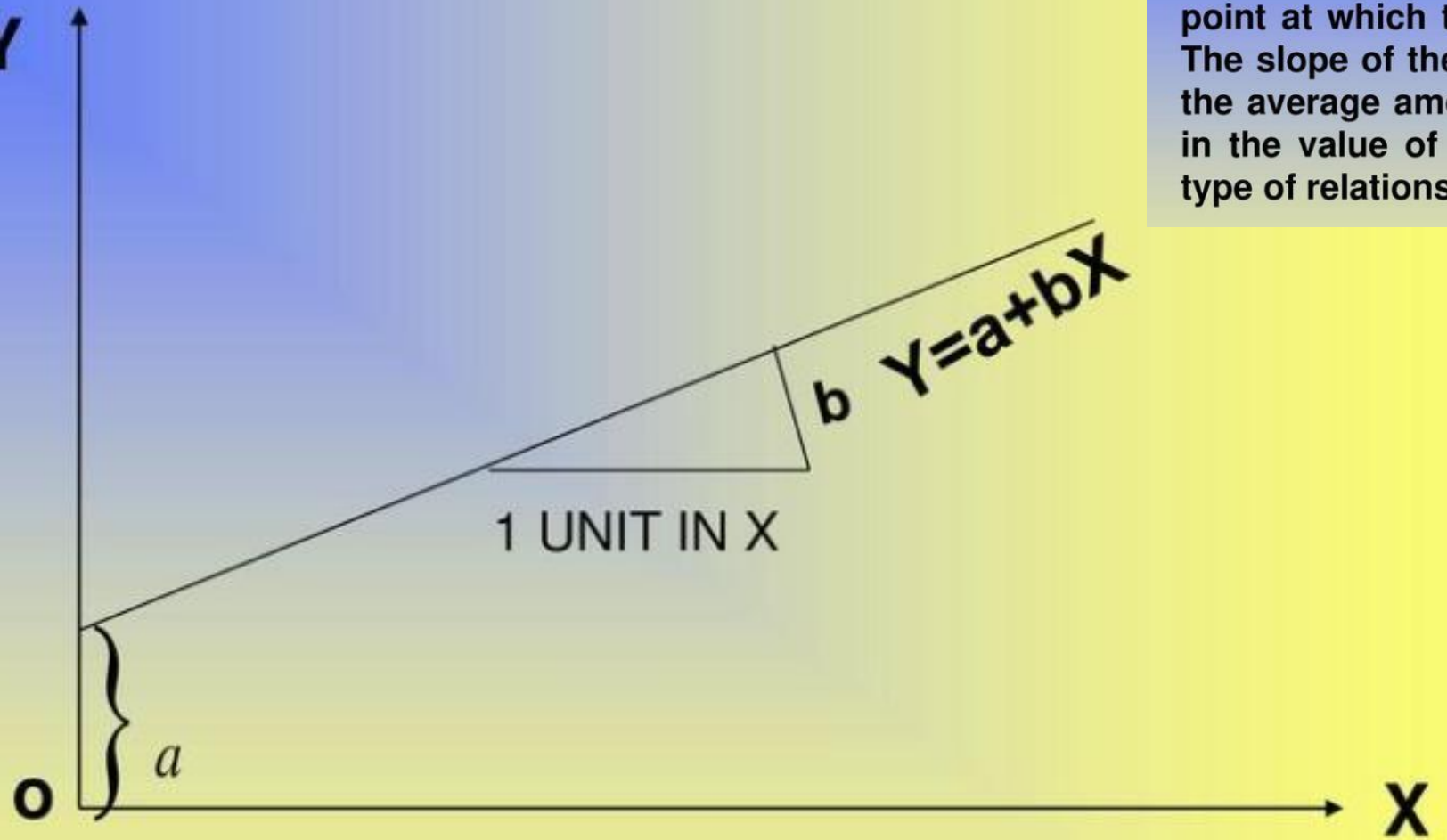
Continued……

What is the general form of the regression equation of X on Y?

These equations are usually called the normal equations. In the equations $\sum X$, $\sum Y$, $\sum XY$, $\sum X^2$, indicate totals which are computed from the observed pairs of values of the two variables X and Y to which the least squares estimating line is to be fitted and N is the total number of observed pairs of values.

The geometrical presentation of the linear equation , $Y = a+bX$ is shown in the diagram below:

It is clear from this diagram, the height of the line tells the average value of Y at a fixed value of X. When X=0, the average value of Y is equal to a . The value of a is called the Y- intercept since it is the point at which the straight line crosses the Y- axis. The slope of the line is measured by b, which gives the average amount of change of Y per unit change in the value of X. The sign of b also indicates the type of relationship between Y and X.



How are the values of ‘a’ and ‘b’ obtained to determine a regression completely?

The values of ‘a’ and ‘b’ are obtained by “the method of least squares” which states that the line should be drawn through the plotted points in such a manner that the sum of the squares of the vertical deviations of the actual Y values from the estimated Y values is the least, or in other words, in order to obtain a line which fits the points best, $(Y–Y_e)^2$ should be minimum. Such a line is known as line of best fit.

The number of pairs is n

Find the sums : $\sum X, \sum Y, \sum XY, \sum X^2, \sum Y^2$

i) **Regression Lines Y on X:**

$$byx = \frac{n\sum(XY) - (\sum X \ \sum Y)}{[n\sum(X^2) - (\sum X)^2]}, \qquad ayx = \frac{\sum Y}{n} - byx\,\frac{\sum X}{n}$$

Equation of Regression Line Y on X: $Y = byx\,X + ayx$

We can estimate Y for any *xvalue* using this equation

$$Estimated\ Y = byx * \ xvalue \ + ayx$$

i) Regression Lines X on Y:

$$bxy = \frac{n\sum(XY) - (\sum X \ \sum Y)}{[n\sum(Y^2) - (\sum Y)^2]}, \qquad axy = \frac{\sum X}{n} - \ bxy \ \frac{\sum Y}{n}$$

Equation of Regression Line X on Y is **X = bxy Y + axy**

Determining correlation coefficient r

$$if\ byx\ and\ bxy\ are\ positive\ then \quad r = \sqrt{byx * bxy}$$

$$if\ byx\ and\ bxy\ are\ negative\ then \quad r = -\sqrt{byx * bxy}$$

What are the properties of regression co–efficients?

- If one of the regression co–efficients is greater than unity, the other must be less than unity, as the value of the co–efficient correlation cannot exceed unity.

Example: if b_{xy} =1.2 and b_{yx} =1.4,

$$r \text{ would be } \sqrt{1 \cdot 2 \times 1 \cdot 4}$$

$$= 1 \cdot 29 \text{ which is not possible.}$$

What are the properties of regression co–efficient?

- Both the regression co–efficients will have the same sign i.e. they will be either positive or negative.
- The co–efficient of correlation will have the same sign as that of regression co–efficient.

Example: $If\ b_{xy} = -0 \cdot 2\ and\ b_{yx} = -0 \cdot 8,$

$$r = -\sqrt{0 \cdot 2 \times 0 \cdot 8} = -0 \cdot 4$$

What are the properties of regression co-efficient?

- **The average value of the two regression co–efficients would be greater than the value of co–efficient of correlation. Symbolically,** $\frac{b_{xy} + b_{yx}}{2} > r$

Example:

$$If\ b_{xy} = 0 \cdot 8\ and\ byx = 0 \cdot 4,\ the\ average\ of$$

$$the\ two\ values\ would\ be\ \frac{0 \cdot 8 + 0 \cdot 4}{2} = 0 \cdot 6.$$

$$The\ value\ of\ r\ would\ be\ \sqrt{0 \cdot 8 \times 0 \cdot 4}$$

$$= 0 \cdot 566\ which\ is\ less\ than\ 0 \cdot 6.$$

Example:

The following data give the hardness (X) and tensile strength (Y) of 7 samples of metal in certain units. Find the linear regression equation of Y on X.

X:	146	152	158	164	170	176	182
Y:	65	78	77	89	82	85	86

Calculate the regression equations of X on Y and Y on X from the following data

X:	1	2	3	4	5
Y:	2	5	3	8	7

From the following data, obtain the two regression equations and estimate the value of Sales when the purchases were 75, Also estimate the value of Purchases when the Sales were 100

Sales : 91 97 108 121 67 124 51 73 111 57

Purchases: 71 75 69 97 70 91 39 61 80 47

Also determine the pearson's correlation coefficien

Solution:

	Sales (x)	Purchases (y)	xy	x^2	y^2
	91	71	6461	8281	5041
	97	75	7275	9409	5625
	108	69	7452	11664	4761
	121	97	11737	14641	9409
	67	70	4690	4489	4900
	124	91	11284	15376	8281
	51	39	1989	2601	1521
	73	61	4453	5329	3721
	111	80	8880	12321	6400
	57	47	2679	3249	2209
∑	900	700	66900	87360	51868

1) Calculations for the equation Y on X

n= 10

$$byx = \frac{10*66900 \quad -900*700}{[10*87360 \; - \; (900*900)]}$$

$$byx = \frac{39000}{63600}$$

byx= 0.6132

$$ayx = \frac{700}{10} \; - \; 0.613 \; \frac{900}{10}$$

ayx= 14.812

EQUATION OF THE LINE Y ON X

Y= 0.6132 X + 14.812

$$byx = \frac{n\sum(XY) - (\sum X \; \sum Y)}{[n\sum(X^2) - (\sum X)^2]}$$

$$ayx = \frac{\sum Y}{n} - byx \; \frac{\sum X}{n}$$

$$Y = byx \; X + ayx$$

2) Calculations for the equation X on Y

n= 10

	Sales (x)	Purchases (y)	xy	x^2	y^2
Σ	900	700	66900	87360	51868

$$bxy = \frac{10*66900 - 900*700}{[10*51868 - (700*700)]}$$

$$bxy = \frac{39000}{28680}$$

bxy= 1.3598

$$axy = \frac{900}{10} - 1.36\ \frac{700}{10}$$

axy= -5.186

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

Note : Numerator is same

$$axy = \frac{\sum X}{n} - bxy\ \frac{\sum Y}{n}$$

EQUATION OF THE LINE X ON Y

X= 1.3598 Y-5.186

$$\mathbf{X = bxy\ Y + axy}$$

Estimation

EQUATION OF THE LINE **Y** ON X

Y= 0.6132 X + 14.812

EQUATION OF THE LINE **X** ON Y

X= 1.3598 Y-5.186

To Estimate the value of Sales when the purchases were 75, We have to estimate **Sales i.e X** , When purchases Y= 75 so we use the regression equation X **on Y**

EQUATION OF THE LINE **X** ON Y

X= 1.3598 Y-5.186

$$\text{Estimated X} = \text{bxy} * \dot{y}value + \text{axy}$$

X= 1.3598 * 75 + -5.186 X=? Y= 75

X= 96.799 = 97

To Estimate the value of **Purchases** when the Sales were 100, We have to estimate **Purchases i.e Y** , When sales X= 100 so we use the regression equation **Y on X**

EQUATION OF THE LINE **Y** ON X

Y= 0.6132 X + 14.812

Y= 0.6132 *100 + 14.812 Y=? X= 100

X= 76.132 = 76

Determining correlation coefficient r

$$if\ byx\ and\ bxy\ are\ positive\ then \quad r = \sqrt{byx * bxy}$$
$$if\ byx\ and\ bxy\ are\ negative\ then \quad r = -\sqrt{byx * bxy}$$

Here byx and bxy are positive so we take the positive square root

r= sqrt(byx*bxy)

r= sqrt(0.6132*1.3598)

r = 0.913143

byx= 0.613

bxy= 1.36

byx= -3

bxy= -4

prod 12

3.464

Exercise

(a) *The two regression coefficients.* (b) *The two regression equations.*

(c) *The coefficient of correlation between the marks in Economics and Statistics.*

(d) *The most likely marks in Statistics when marks in Economics are 30.*

Marks in Economics	:	25	28	35	32	31	36	29	38	34	32
Marks in Statistics	:	43	46	49	41	36	32	31	30	33	39

Example:

The following figures relate to advertisement expenditure and corresponding sales

Advertisement (in lakhs of Taka)	60	62	65	70	73	75	71
Sales (in crores of Taka)	10	11	13	15	16	19	14

Estimate

i) The sales for advertisement expenditure of Tk. 80 lakhs and

ii) The advertisement expenditure for a sales target of Tk. 25 crores

Example:

The following data relate to advertising expenditure (in lakhs of Taka) and their corresponding sales (in crores of Taka):

Advertising Expenditure	10	12	15	23	20
Sales	14	17	23	25	21

Estimate i) the sales corresponding to advertising expenditure of Tk 30 lakhs and

ii) the advertising expenditure for a sales target of Tka. 35 crores.

Numeric al

Obtain the two lines of regression from the following data and estimate the blood pressure when age is 50 years. Can we also estimate the blood pressure of a person aged 20 years on the basis of this regression equation? Discuss.

Age (in years)	56	42	72	39	63	47	52	49	40	42	68	60
Blood Pressure	127	112	140	118	129	116	130	125	115	120	135	133

$y = \beta_0 + \beta_1 x$

$\beta_0 = 87.07$

$\beta_1 = 0.7223$

$r = 0.9230$

$y = 123.185$

Numeric al

The following data relate to the scores obtained by 9 salesmen of a company in an intelligence test and their weekly sales in thousand rupees:

Salesmen	:	A	B	C	D	E	F	G	H	I
Test Scores	:	50	60	50	60	80	50	80	40	70
Weekly Sales (₹ 000)	:	30	60	40	50	60	30	70	50	60

(*a*) Obtain the regression equation of sales on intelligence test scores of the salesman.

(*b*) If the intelligence test score of a salesman is ₹65,000 what would be his expected weekly sales?

$y = \beta_0 + \beta_1 x$

$\beta_0 = 5$

$\beta_1 = 0.75$

$y = 53.75$

Numeric al

In a correlation study, the following values are obtained:

	X	Y
Mean	65	67
Standard Deviation	2·5	3·5
Coefficient of Correlation	0·8	

Find the two regression equations that are associated with the above values.

$$x - \bar{x} = r\frac{\sigma x}{\sigma y}(y - \bar{y})$$

$$x = 26.72 + 0.5714\,y$$

$$y - \bar{y} = r\frac{\sigma y}{\sigma x}(x - \bar{x})$$

$$y = 1.12x - 5.8$$

Numeric
al

Two random variables have the regression equations:

$$3X + 2Y - 26 = 0$$
$$6X + Y - 31 = 0$$

Find the mean value and the coefficient of correlation between X and Y. If the variable of $X = 25$, find the standard deviation of Y from the data given above.

xп$_0$ =
4
y0=
7
r = -0.5

$\sigma y = 15$

Numeric al

Given the following data for the sales of car of an automobile company for six consecutive years. Predict the sales for next two consecutive years.

Years	2013	2014	2015	2016	2017	2018
Sales	110	100	250	275	230	300