CHAPTER

REGRESSION

Regression is used to denote estimation or prediction of the average value of one variable for a specified value of the other variable. One of the variables is called **independent** or the **explained variable** and the other is called **dependent** or the **explaining variable**.

"Regression is the measure of the average relationship between two or more variables in terms of the original units of the data." *M.M. Blair*

The estimation or prediction is done by means of suitable equation derived on the basis of available bivariate data. Such an equation is known as *Regression equation* and its geometrical representation is called *Regression curve*.

I. Regression Equation of X on Y is

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

$$= r\frac{\sigma x}{\sigma y}(Y - \bar{Y})$$ [It estimates X for given Value of Y]

II. Regression Equation Y on X is

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

$$= r\frac{\sigma y}{\sigma x}(X - \bar{X})$$ [It estimates Y for a given value of X]

X = Value of x

$\bar{X}$ = Mean of x

σx = Standard deviation of x series

r = Correlation coefficient

Y = Value of Y

$\bar{Y}$ = Mean of Y

σy = Standard deviation of y series

b = Slope or coefficient of regression

• **Regression Lines:**

If a bivariate data are plotted as points on graph paper, it will be found that the concentration point follows a certain pattern showing the relationship between the variables. When the trend points are found to be linear, we determine the best fitting straight line by *Least Square Method. Such straight lines which are used to obtain best estimates of one variable for given values of the other, are called* ***regression lines.***

If two variables are linearly related, then that relation can be expressed as $Y = bx + a$.

Where 'b' is the slope of the line relating Y to X and 'a' is the 'Y' intercept of that line.

A *line of regression* is the straight line which gives the best fit in the least square sense to given sets of data.

Regression Coefficient:

I. The regression coefficient (b) is an expression of how much (*on the average*) one dependent variable (Y) may be expected to change per unit change in some other independent variable (X).

II. It is denoted by letter 'b'.

III. The regression coefficient of Y on X is

$$= byx = r\frac{\sigma y}{\sigma x}\frac{(S.D.of\ Y\ series)}{(S.D.of\ X\ series)}.$$

IV. The regression coefficient of X on Y is

$$= bxy = r\frac{\sigma x}{\sigma y}\frac{(S.D.of\ X\ series)}{(S.D.of\ Y\ series)}.$$

Types of Regression:

(a) Simple regression:

I. Here the dependent variable (criterion) is a function of a single independent variable (predictor).

II. The score of the dependent variable is predicted from the given scores of the single predictor.

Example: Height of person on his weight.

(b) Multiple regression:

I. Here the dependent variable (criterion) is a function of two or more predictors.

II. The scores are predicted from the scores of more than one predictor.

III. It may be linear or nonlinear.

Example: Thyroid calcitonin on combination of thyroxine secretion & serum calcium.

(c) Linear regression:

I. Here the dependent variable (criterion) is linearly correlated with the predictor (independent variable).

II. The scores of the dependent variables are predicted by working out an equation for a straight line, depending on the linear association between the two.

The statistical analysis employed to find out the exact position of the straight line is known as linear regression analysis.

(d) Nonlinear regressions:

If the criterion (dependent variable) has a nonlinear correlation with the predictor (independent variable), the scores of the criterion have to be predicted in terms of a curved line like a sigmoid or hyperbolic or exponential curve, according to their form of association.

Properties of Regression:

1. It is an expression of the dependent variable (criterion) as a function of the independent variable (predictor).
2. A regression can be worked out only when there is a significant correlation between the dependent (criterion) and the independent (predictor) variable.
3. Regression predicts only a probable score of the criterion on a given score of the predictor.
4. When a pair of variables correlated with one another, regression can be worked in two ways. viz. (*i*) a regression variable X as criterion on variable Y as predictor and (*ii*) another regression of variable Y as criterion on variable X as predictor.
5. A regression equation is worked out using a statistic called the regression coefficient.

• **Method of Studying Regression:**

There are two methods: (*a*) Graphic method and (*b*) Algebraic method.

(*a*) Graphic method:

I. The points are plotted on a graph paper representing pairs of values of concerned variables.

II. In this diagram independent variable is taken on the horizontal axis and dependent variable on the vertical axis.

III. These points give a picture of a scatter diagram. A regression line may be drawn in between these points by free hand or by a scale rule.

(*b*) Algebraic method:

I. A regression line is a straight line fitted to the data by the method of least squares.

II. It indicates the best probable mean value of one variable corresponding to the mean value of the other.

III. There are always two regression lines constructed for the relationship between the two variables. *viz.*, X and Y.

Thus one regression line shows regression of X upon Y and the other shows regression Y upon X.

Linear regression:

Let the equation in general terms:

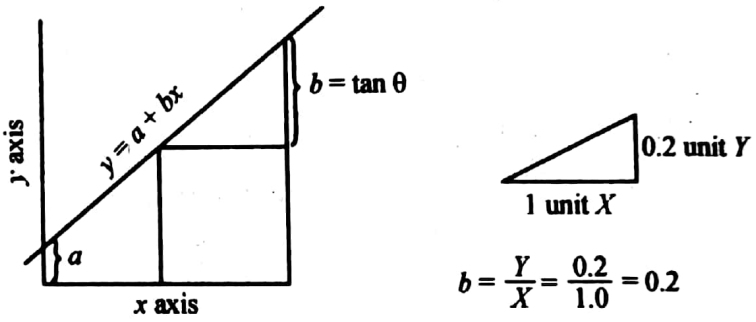

Fig. 14.1. Intercept & Slope for regression line.

I. $Y = a + bx$ where y and x represent two variables, 'a' is the y intercept or distance between the x axis and the point where the line crosses the y axis.

II. 'b' is the slope or increase in the y value per unit change in x value.

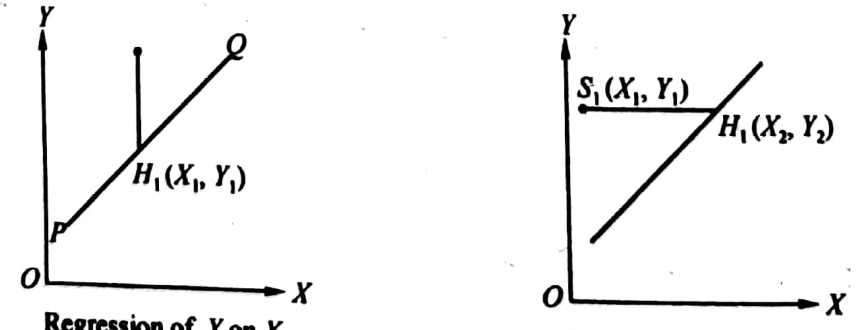

Fig. 14.2. (*a*) & (*b*) show regression Y on X and X on Y respectively.

(*A*) I. If the line of regression is so chosen that the sum of squares of deviation parallel to the axis of y is minimized [Fig. 14.2 (*a*)], it is called the line of regression of Y on X and it gives the best estimate of Y for any given value of X.

II. Its equation is $y = a + bx$.

III. The slope of the line b in the equation is known as the regression coefficient. It shows that y changes b times as fast as x.

IV. Symbolically the regression coefficient of y on x is b_{yx}.

(*b*) I. If the line of regression is so chosen that the sum of squares of deviations parallel to the axis of x is minimized [Fig. 14.2 (*b*)], it is called the line of regression of X on Y and it gives the best estimate of x for any value of y.

II. The regression equation in this case is $x = a + by$.

III. The regression coefficient of x on y is b_{xy}.

Regression Y on X	Regression X on Y
Line of regression $Y = mx + c$ The coefficient of x *i.e.*, m represent the regression coefficient of Y on X *i.e.*, myx.	Line of regression $X = my + c$ The coefficient of y *i.e.*, m represent the regression coeffcient of X on Y *i.e.*, mxy.

• Computation of Linear Regression:

(List of formulae)

1. Regression Y on X:

(*a*) Regression equation:

$$y = mx + c$$

$$y - \bar{y} = byx(x - \bar{x})$$

(*b*) Regression coefficient: $bxy = \dfrac{\sigma y}{\sigma x}$

(*i*) When the deviations are taken from the mean:

$$dx = x - \bar{x}$$

$$dy = y - \bar{y}$$

$$byx = \frac{\sum dxdy}{\sum d_x^2} \text{ or } byx = \frac{\sum xy - n(\bar{x}\bar{y})}{\sum x^2 - n(\bar{x})^2}$$

(*ii*) When the deviations are taken from the assumed mean:

$u = x - a \qquad v = y - b$

(a and b assumed mean of X and Y series respectively)

$$byx = \frac{\sum uv - \dfrac{\sum u \sum v}{n}}{\sum u^2 - \dfrac{(\sum u)^2}{n}}.$$

(*iii*) When the original values (raw scores) are used:

$$byx = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2} = \frac{\sum xy - \dfrac{(\sum x)(\sum y)}{n}}{\sum x^2 - \dfrac{(\sum x)^2}{n}}.$$

2. Regression X on Y:

(*a*) Regression equation: $X = my + c$

$$X - \bar{X} = bxy\,(Y - \bar{Y}).$$

(*b*) Regression coefficient: $byx = \frac{\sigma x}{\sigma y}$

I. When the deviations are taken from the mean:

$$dx = X - \bar{X} \quad dy = Y - \bar{Y}$$

$$bxy = \frac{\sum dx \sum dy}{\sum d^2 y} \text{ or } bxy = \frac{\sum xy - x(\bar{x}.\bar{y})}{\sum y^2 - n(\bar{Y})^2}$$

II. When the deviations are taken from assumed mean:

$$u = x - a \quad v = y - b$$

$$bxy = \frac{\sum uv - \frac{\sum u \sum v}{n}}{\sum v^2 - \frac{(\sum v)^2}{n}}$$

III. When the original values are used (i.e., raw score):

$$bxy = \frac{n\,xy - \sum x \sum y}{n\sum y^2 - \sum y^2} \text{ or } bxy = \frac{\sum xy - \left(\sum x \sum y\right)n}{\sum y^2 - \left(\sum y\right)^2 n}$$

• **Properties of Regressi... Co...ff...**

Correlation	Regression
1. Correlation is the relationship between two or more variables which vary in sympathy with the other in the same or the opposite direction.	1. Regression is a mathematical measure showing the average relationship between two variables.
2. Here both the variables *i.e.*, x and y are random variables.	2. Here x is a random variable and y is fixed. Sometimes both the variables may be random variables.
3. It finds out the degree of relationship between two variables [not cause and effect of the variable].	3. It indicates the cause and effect relationship between the variables.
4. It is used for testing and verifying the relationship between two variables.	4. It is used for prediction of one value in respect to the other given value.
5. The coefficient correlation is a relative measure. The range of relationship lies between ± 1.	5. Regression coefficient is an absolute figure. If we know the value of independent variable, we can find the value of dependent variable.
6. It has limited application because it is confined only to linear relationship between the variables.	6. It has wide application as it studies linear and non-linear relationship between the variables.
7. If the coefficient correlation is positive, then the two variables are positively correlated and vice-versa.	7. The regression coefficient explains that the decrease in one variable is associated with the increase in the other variable.

Example 1: Compute *byx* for the following data:

x, y : (5, 2), (7, 4), (8, 3), (4, 2), (6, 4).

Solution:

x	y	xy	x^2
5	2	10	25
7	4	28	49
8	3	24	64
4	2	08	16
6	4	24	36
30	15	94	190

$$\sum x = 30 \quad \sum y = 15 \quad \sum xy = 94 \quad \sum x^2 = 190 \quad n = 5$$

$$byx = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{(\Sigma x)^2}{n}} = \frac{94 - \frac{30 \times 15}{5}}{190 - \frac{(30)^2}{5}} = \frac{94 - 90}{190 - 180} = \frac{4}{10} = \frac{2}{5} = 0.4$$

$\therefore$ $bxy = 0.4$.

Example 2: Find the two regression equations from the following pairs of observations on X and Y: (1, 2), (2, 3), (3, 5), (4, 6), (5, 4). Hence find the predicted value of Y when $X = 2.5$ and the predicted value of X when $Y = 4.5$.

Solution:

X	Y	x^2	y^2	XY
1	2	1	4	2
2	3	4	9	6
3	5	9	25	15
4	6	16	36	24
5	4	25	16	20
15	20	55	90	67

$\sum X = 15 \quad \sum Y = 20 \quad n = 5 \quad \sum XY = 67 \quad \bar{X} = \frac{15}{5} = 3 \quad \bar{Y} = \frac{20}{5} = 4$

$\sum X^2 = 55 \quad \sum Y^2 = 90$

$$bxy = \frac{\sum XY - \frac{\sum X \sum Y}{n}}{\sum Y^2 - \frac{(\sum Y)^2}{n}} = \frac{67 - \frac{15 \times 20}{5}}{90 - \frac{(20)^2}{5}} = \frac{67 - 60}{90 - 80} = \frac{7}{10} = 0.7$$

$$byx = \frac{\sum XY - \frac{\sum X \sum Y}{n}}{\sum X^2 - \frac{(\sum X)^2}{n}} = \frac{67 - \frac{15 \times 20}{5}}{55 - \frac{(15)^2}{5}} = \frac{67 - 60}{55 - 45} = \frac{7}{10} = 0.7$$

The regression equation Y on X is

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

or $\quad Y - 4 = 0.7\,(X - 3)$

or $\quad Y = 0.7X\ 2.1 + 4$

$$= 0.7X + 1.9 \qquad \ldots(1)$$

The regression equation X on Y is

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

$$X - 3 = 0.7\,(Y - 4)$$

or $\quad X = 0.7Y - 2.8 + 3$

$$X = 0.7Y + 0.2 \qquad \ldots(2)$$

If $x = 2.5$, then from Equation (1), we get

$$Y = 0.7 \times 2.5 + 1.9$$

$$Y = 1.75 + 1.9 = 3.65$$

$\therefore \quad Y = 3.65$

If $y = 4.5$, then from Equation (2), we get

$$X = 0.7 \times 4.5 + 0.2$$

$$X = 3.15 + 0.2 = 3.35$$

$\therefore \quad X = 3.35.$

Example 3: Calculate regression coefficients *byx* and *bxy* for the following data:

$$\sum X = 55 \quad \sum Y = 88 \quad \sum XY = 586 \quad \sum X^2 = 385 \quad \sum Y^2 = 1114 \quad n = 10$$

Solution:

$$bxy = \frac{\sum XY - \frac{\sum X \sum Y}{n}}{\sum Y^2 - \frac{\sum Y^2}{n}} = \frac{586 - \frac{55 \times 88}{10}}{1114 - \frac{(88)^2}{10}}$$

$$= \frac{586 - 484}{1114 - 774.4} = \frac{102}{339.6} = \frac{1020}{3396} = 0.3$$

$$byx = \frac{\sum XY - \frac{\sum X \sum Y}{n}}{\sum X^2 - \frac{(\sum X)^2}{n}} = \frac{586 - \frac{55 \times 88}{10}}{385 - \frac{(55)^2}{10}} = \frac{586 - 484}{385 - 302.5}$$

$$= \frac{102}{82.5} = \frac{1020}{825} = 1.236 = 1.2.$$

Example 4: Find the correlation coefficient in each of the following cases: (*i*) $byx = 0.4$ and $bxy = 0.9$, (*ii*) $byx = 1.6$ and $bxy = 0.4$, (*iii*) $byx = -0.3$ and $bxy = -1.2$.

Solution: (*i*) $r = \sqrt{byx \times bxy} = \sqrt{0.4 \times 0.9} = \sqrt{0.36} = \pm 0.6$

(*ii*) $r = \sqrt{1.6 \times 0.4} = \sqrt{0.64} = \pm 0.8$

(*iii*) $r = \sqrt{-0.3 \times -1.2} = \sqrt{0.36} = \pm 0.6$

Here *byx* & *bxy* are both negative. Hence *r* must be negative *i.e.*, –0.6.

Example 5: From the following data, calculate (*a*) correlation coefficient and (*b*) standard deviation of *Y* (σ_y). $X = 0.85Y$ $Y = 0.89X$, $\sigma_x = 3$

Solution:

(*a*) $r = \sqrt{bxy \times byx}$ $r = \sqrt{0.85 \times 0.89} = \sqrt{0.7565} = 0.869$

(*b*) $bxy = r \times \frac{\sigma_x}{\sigma_y}$

or $0.85 = 0.869 \times \frac{3}{\sigma_y}$

or $\sigma_y = \frac{0.869 \times 3}{0.85} = \frac{2.607}{0.85} = 3.067.$

Example 6: Find the regression coefficients *byx* and *bxy* of *Y* on *X* on *Y* respectively, if standard deviations of *X* and *Y* are 4 and 3 respectively and correlation coefficient between *X* and *Y* is 0.8.

Solution:

$\sigma_x = 4$ $\sigma_y = 3$ $r = 0.8$

$bxy = r \frac{\sigma_x}{\sigma_y}$ $byx = r \frac{\sigma_Y}{\sigma_X}$

$bxy = 0.8 \times \frac{4}{3}$ $byx = 0.8 \times \frac{3}{4}$

$bxy = 1.066$ $byx = 0.6$

$bxy = 1.067.$

Example 7: The correlation coefficient between *X* and *Y* is 0.60. If the variance of *x* = 225, the variance of *Y* = 400, mean of *X* = 10 and mean of *Y* = 20, find the equation of the regression lines of (*i*) *Y* on *X* and (*ii*) *X* on *Y*.

Solution:

Variance of X i.e., $\sigma_x^2 = 225$ $\therefore$ $\sigma_x = \sqrt{225} = 15$

Variance of Y i.e., $\sigma_y^2 = 400$ $\therefore$ $\sigma_y = \sqrt{400} = 20$

$r = 0.60$ $\bar{X} = 10$ $\bar{Y} = 20$

$$byx = r\frac{\sigma_y}{\sigma_x} = 0.6 \times \frac{20}{15} = 0.8 \quad bxy = 0.6 \times \frac{15}{20} = 0.45$$

(*i*) Regression equation Y on X i.e.,

$$Y - \bar{Y} = byx\,(X - \bar{X})$$
$$Y - 20 = 0.8\,(X - 10)$$
$$Y = 0.8X - 8 + 20$$
$$Y = 0.8X + 12$$

(*ii*) Regression equation of X on Y i.e.,

$$X - \bar{X} = bxy\,(Y - \bar{Y})$$
$$X - 10 = 0.45\,(Y - 20)$$
$$= 0.45Y - 9$$
$$X = 0.454 - 9 + 10 = 0.45Y + 1.$$

Example 8: You are given the following results of two variables X and Y:

$\bar{X} = 36$ $\bar{Y} = 85$ $\sigma_x = 11$ $\sigma_y = 8$ $r(X, Y) = 0.66$.

Find the two regression equations and estimate the value of X when $Y = 75$.

Solution:

$$bxy = r\frac{\sigma_x}{\sigma_y} = 0.66 \times \frac{11}{8} = \frac{7.26}{8} = 0.9075$$

$$byx = r\frac{\sigma_y}{\sigma_x} = 0.66 \times \frac{8}{11} = \frac{5.28}{11} = 0.48$$

The regression equation X on Y is

$$X - \bar{X} = bxy(Y - \bar{Y})$$
$$X - 36 = 0.9075\,(Y - 85)$$
$$= 0.9075Y - 77.1375$$
$$X = 0.9075Y + 36 - 77.1375$$
$$= 0.9075Y - 41.1375$$

The regression equation Y on X is

$$Y - \bar{Y} = byx\,(X - \bar{X})$$
$$Y - 85 = 0.48\,(X - 36)$$
$$Y = 0.48X - 17.28 + 85$$
$$Y = 0.48X + 67.72$$

When $Y = 75$ the value of X will be

$$X = 0.9075\,Y - 41.1375$$
$$= 0.9075 \times 75 - 41.1375$$
$$= 68.0625 - 41.1375$$
$$= 26.925$$