Regression

Regression In analysing data, we find that it is frequency desirable to learn something about the relationship between two variables. For example, we may be interested in studying the relationship between blood pressure and age, height and weight. The nature of relationship between variables such as these may be examined by **Regression analysis**. Regression analysis is helpful in ascertaining the probable form of relationship between variables and to predict or estimate the value of one variable corresponding to a given value of other variable.

In regression analysis the two variables are related as independent and dependent.

Dependent Variable: The variable to be estimated is called dependent variable or we can say the variable whose value is influenced or is to be predicted, is called a dependent variable.

Independent Variable: The variable which is known, is called independent variable or another way the variable which influences the value is called an independent variable.

Types of regression analysis The regression analysis can be two types: simple and multiple.

Simple Regression: The regression analysis confined to the study of only two variables at a time is termed as simple regression.

Multiple Regression: The regression analysis confined to the studying more than two variables at a time is know as multiple regression.

Linear Regression: When observations from two variables are plotted as a graph, and if the points so obtained fall in a straight line, then relationship is linear and it is said that there is linear regression between variables. however, if the line is not a straight line, the regression is termed as non-liner regression.

Regression Equation: For a liner regression, the equation for a dependent variable Y against independent variable X can be given as follow: $\mathbf{Y = a + b\,X}$.

Here, value of a and b are constant and are fixed for a particular line. The constant a is known as intercept and denotes the value of Y when the value of X is zero. The constant b measures the slope of the line and a is called regression coefficient. If the value of a and b are known, Y can be obtained for any corresponding value of X. The values of a and b are calculated by the following equation:

Regression Equation: The linear regression model $Y = a + b\,X$. The normal equations are

$$\sum y_i = n\,a + b\sum x_i \text{ - - - -(1)} \qquad \sum x_i\,y_i = a\sum x_i + b\sum x_i^2 \text{ - - - -(2)}$$

Solve above two equation and find the value of a and b for given values $\sum y_i, \sum x_i, \sum x_i\,y_i, \sum x_i^2$. These values of a and b put in $Y = a + b\,X$ we can fit a linear regression between X and Y.

Ex: Find out regression equation from the following data.

X	13.4	15.1	15.3	16.8	17.5	19.2	21.2
Y	2.1	2.3	2.3	2.6	2.7	3	3.3

Sol: A linear regression equation $Y = a + b\,X$. So, first find the values of a and b.

X	13.4	15.1	15.3	16.8	17.5	19.2	21.2	$\sum X = 118.5$
Y	2.1	2.3	2.3	2.6	2.7	3	3.3	$\sum Y = 18.3$
$X\,Y$	28.14	34.73	35.19	43.68	47.25	57.6	69.96	$\sum X\,Y = 316.55$
X^2	179.56	228.01	234.09	282.24	306.25	368.64	449.44	$\sum X^2 = 2048.23$

Put all these values in normal equations, we get

$18.3 = 7\,a + 118.5\,b$ - - -(1) $316.55 = 118.5\,a + 2048.23\,b$ - - -(2)

Multiply in equation (1) by 118.5 and in equation (2) by 7, then subtract we get

$829.5\,a + 14042.25\,b = 2168.55$
$829.5\,a + 14337.61\,b = 2215.85$

$-\,295.36\,b = -47.3$ so, $b = 0.16$

Put this value of b in equation (1) to find value of a so, $7\,a + 118.5 * 0.16 = 18.3$ solve we get $a = -0.094$

A linear regression equation for this given data is **Y = -0.094 + 0.16 X**

The line of regression of Y on X is given by

$$Y - \bar{y} = r\,\frac{\sigma_y}{\sigma_x}\,(X - \bar{x})$$

or

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

where b_{YX} is Regression coefficient of Y on X. It represents the increment in the value of dependent variable Y corresponding to a unit change in the value of independent variable X.

The line of regression of X on Y is given by

$$X - \bar{x} = r\,\frac{\sigma_x}{\sigma_y}\,(Y - \bar{y})$$

or

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

where b_{XY} is Regression coefficient of X on Y indicates the change in the value of variable X corresponding to a unit change in the value of variable Y.

Properties of Regression coefficients

(i) Correlation Coefficient is the geometric mean between the regression coefficients.

$b_{XY}\, b_{YX} = r\,\frac{\sigma_x}{\sigma_y} \times r\,\frac{\sigma_y}{\sigma_x} = r^2$ so

$r = \pm\sqrt{b_{XY}\, b_{YX}}$

(ii) If one of the regression coefficient is greater than unity, the other must be less than unity.

Let $b_{YX} > 1$ then $\frac{1}{b_{YX}} < 1$ and we know that $r^2 \le 1$ so

$b_{XY}\, b_{YX} \le 1$ hence $b_{XY} \le \frac{1}{b_{YX}} < 1$.

Ex: Obtain the equations of two lines of regression for the following data. Also Obtain the estimate of X for $Y = 70$.

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

Sol:

X	Y	X^2	Y^2	XY
65	67	4225	4489	4355
66	68	4356	4624	4488
67	65	4489	4225	4355
67	68	4489	4624	4556
68	72	4624	5184	4896
69	72	4761	5184	4968
70	69	4900	4761	4830
72	71	5184	5041	5112
544	552	37028	38132	37560

$$\bar{X} = \frac{\sum X}{n} = \frac{544}{8} = 68, \quad \bar{Y} = \frac{\sum Y}{n} = \frac{552}{8} = 69$$

$$\sigma_X = \sqrt{\tfrac{37028}{8} - 68^2} = \sqrt{4.5} = 2.12, \qquad \sigma_Y = \sqrt{\tfrac{38132}{8} - 69^2} = \sqrt{5.5} = 2.35.$$

$Cov(X,Y) = \frac{37560}{8} - 68 \times 69 = 3,$

$$r = \frac{cov(x,y)}{\sigma_x\, \sigma_y} = \frac{3}{2.12 \times 2.35} = 0.6$$

The line of regression of Y on X is given by

$Y - \bar{y} = r\,\frac{\sigma_y}{\sigma_x}\,(X - \bar{x})$

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

$Y = 0.665\,X + 23.78$

The line of regression of X on Y is given by

$X - \bar{x} = r\,\frac{\sigma_x}{\sigma_y}\,(Y - \bar{y})$

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

$X = 0.54\,Y + 30.74$

To estimate X for given Y, we use the line of regression of X on Y. If $Y = 70$, estimated value of X is given by $\hat{X} = 0.54 \times 70 + 30.74 = 68.54$.

Short-Cut Method

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

X	9	8	20	2	7	1	9	7	9	8
Y	6	3	7	8	8	9	10	1	3	5

Sol:

X	Y	U	V	U^2	V^2	$U\,V$
9	6	0	-2	0	4	0
8	3	-1	-5	1	25	5
20	7	11	-1	121	1	-11
2	8	-7	0	49	0	0
7	8	-2	0	4	0	0
1	9	-8	1	64	1	-8
9	10	0	2	0	4	0
7	1	-2	-7	4	49	14
9	3	0	-5	0	25	0
8	5	1	-3	1	9	3
TOTAL		-10	-20	244	118	3

$U = x - 9, V = y - 8, \quad \bar{U} = \dfrac{\sum U}{n} = \dfrac{-10}{10} = -1, \quad \bar{V} = \dfrac{\sum V}{n} + \dfrac{-20}{10} = -2,$

$\bar{x} = a + \bar{U} = 9 - 1 = 8, \quad \bar{y} = b + \bar{V} = 8 - 2 = 6,$

$Cov(x,\,y) = Cov(U,\,V) = \dfrac{1}{n}\sum U\,V - \bar{U}\,\bar{V} = \frac{3}{10} - (-1)(-2) = -1.7$

$$\sigma_x^2 = \sigma_u^2 = \sqrt{\frac{1}{n}\sum U^2 - \bar{U}^2} = \sqrt{\frac{244}{10} - (-1)^2} = 23.4,$$

$$\sigma_y^2 = \sigma_v^2 = \sqrt{\frac{1}{n}\sum V^2 - \bar{V}^2} = \sqrt{\frac{118}{10} - (-2)^2} = 7.8$$

The line of regression of Y on X is given by

$Y - \bar{y} = \frac{Cov(x,y)}{\sigma_x^2}\,(X - \bar{x})$

$Y - 6 = \frac{-1.7}{23.4}(X - 8)$

$Y = 6.58 - 0.07\,X$

The line of regression of X on Y is given by

$X - \bar{x} = \frac{Cov(x,y)}{\sigma_y^2}\,(Y - \bar{y})$

$X - 8 = \frac{-1.7}{7.8}(Y - 6)$

$X = 9.31 - 0.22\,Y$

Ex: In a partially destroyed laboratory, record of an analysis of correlation data, the following results only are legible: Variance of $X = 9$.

Regression Equations: $8\,X - 10\,Y + 66 = 0,\ 40\,X - 18\,Y = 214.$

What are

(i) The mean values X and Y.

(ii) The correlation coefficient between X and Y.

(ii) The Standard deviation of Y.

Sol: Since both lines of regression pass through the point $(\bar{X}, \bar{Y})$, we have

$8\,\bar{X} - 10\,\bar{Y} + 66 = 0$ and $40\,\bar{X} - 18\,\bar{Y} = 214$. Solving, we get $\bar{X} = 13$ and $\bar{Y} = 17$.

Let $8\,X - 10\,Y + 66 = 0$ and $40\,X - 18\,Y = 214$ be the lines of regression of Y on X and X on Y respectively.

These equations can be put in the form:

$$Y = \frac{8}{10}X + \frac{66}{10} \text{ and } X = \frac{18}{40}Y + \frac{214}{40}$$

b_{YX} = Regression coefficient of Y on $X = \frac{8}{10}$

b_{XY} = Regression coefficient of X on $Y = \frac{18}{40}$.

Hence $r^2 = b_{YX}\, b_{XY} = \frac{8}{10} \times \frac{18}{40} = \pm 0.6$

But Since both the regression coefficient are positive, we take $r = 0.6$.

$b_{YX} = r\,\frac{\sigma_y}{\sigma_x} \Rightarrow \frac{8}{10} = 0.6\,\frac{\sigma_y}{3} \Rightarrow \sigma_y = 4$.

Ex: For the regression lines $4\,X - 5\,Y + 33 = 0,\ 20\,X - 9\,Y = 107$.

(i) The mean values X and Y.

(ii) The correlation coefficient between X and Y.

(ii) The Standard deviation of Y given that the variance of X is 9.

Sol: Since both lines of regression pass through the point $(\bar{X}, \bar{Y})$, we have $4\,\bar{X} - 5\,\bar{Y} + 33 = 0$ and $20\,\bar{X} - 9\,\bar{Y} = 107$. Solving, we get $\bar{X} = 13$ and $\bar{Y} = 17$.

Let $8\,X - 10\,Y + 66 = 0$ and $40\,X - 18\,Y = 214$ be the lines of regression of Y on X and X on Y respectively.

These equations can be put in the form:

$$Y = \frac{4}{5}X + \frac{33}{5} \text{ and } X = \frac{9}{20}Y + \frac{107}{20}$$

b_{YX} = Regression coefficient of Y on $X = \frac{4}{5}$

b_{XY} = Regression coefficient of X on $Y = \frac{9}{20}$.

Therefore, the coefficient of correlation between X and Y is $r = \sqrt{b_{YX}\, b_{XY}} = \sqrt{\frac{4}{5} \times \frac{9}{20}} = \pm 0.6$

But Since both the regression coefficient are positive, we take $r = 0.6$.

$b_{YX} = r\,\frac{\sigma_y}{\sigma_x}$

$\frac{4}{5} = 0.6\,\frac{\sigma_y}{3} \Rightarrow \sigma_y = 4.$

References

- S.C. Gupta, V.K. Kapoor, Fundamentals of Mathematical Statistics, Sultan Chand & Sons.
- A. Kumar, A. Chaudhary, Text Book Statistical Methods, Krishna's Educational Publishers.
- Syed Qaim Akbar Rizvi, Text Book Non Parametric Methods & Regression Analysis, Krishna's Educational Publishers.
- K.S. Negi, Biostatistics, Aitbs Publishers.
- V.B. Rastogi, Fundamentals of Biostatistics, ANE Books.