Chapter: Regression 7

Md. Abdul Alim
2-11-2010

Regression: Regression is a mathematical measure of expressing the average of relationship between two or more variables in terms of the original units of the data. In a regression analysis there are two types of variables. The variable whose is influenced or is to be predicted is called dependent variable, regressed predicted or explained variable and the variable which influences the values or is used for prediction is called independent variable or regressor or predictor or explanator. These relationships between two variables can be considered between say rainfall and agricultural production, price of an output and the overall cost of product, consumer expenditure and disposable income.

Regression equation: Regression equations are algebraic expression of the regression lines. Since there are two regression lines, the regression equation of X on Y is said to describe the variation in the values of X for given changes 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.

The regression equation of Y on X is expressed as follows

$$Y = a + bX$$

The regression equation of X on Y is expressed as follows

$$X = a + bY$$

Regression line: If the variables in a bivariate distribution are related we will find that points in the scatter diagram will cluster around some curve called the "Curve of regression". If the curve is straight line of, it is called the line of regression and there is said to be linear regression between the variables, otherwise regression is said to be curvilinear. The line of regression is the line which gives the best estimate to the value of one variable for any specific value of the other variable.

Q. What are the regression coefficients?

In the line of regression of Y on X

$$Y = a + bX$$

The coefficient 'b' which is the slope of the line of regression of Y on X is called the coefficient of regression of Y on X. It represents the increment in the value of the dependent variable Y for a unit change in the value of the independent variable X. For notational convenience, coefficient of regression of Y on X is denoted by b_{yx}.

Regression coefficient of Y on X is

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

and the intercept

$$a = \bar{Y} - b\bar{X} = \frac{\sum Y}{n} - b\frac{\sum X}{n}$$

Similarly in the regression equation of X on Y

$$X = a + bY$$

Regression coefficient of X on Y is

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

and the intercept

$$a = \bar{X} - b\bar{Y} = \frac{\sum X}{n} - b\frac{\sum Y}{n}$$

Note: Interpret a and b for the regression equation $Y = a + bX$

**** The slope b represents the estimated average change in Y when X increases by one unit.**

**** The intercept a represents the estimated average value of Y when X equals zero.**

Example # 1: From the following data obtain the regression equations of Y on X :

Sales (X)	91	97	108	121	67	124	51	73	111	57
Purchase (Y)	71	75	69	97	70	91	39	61	80	47

Solution: We know that, the regression equation of Y on X is expressed as follows

$Y = a + bX$

Again, Regression coefficient of Y on X is

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

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

*

$$b_{yx} = \frac{\sum XY - \frac{\sum X \sum Y}{n}}{\sum X^2 - \frac{(\sum X)^2}{n}} = \frac{66900 - \frac{900 \times 700}{10}}{87360 - \frac{(900)^2}{10}} = 0.613207547 = 0.613$$

$$a = \bar{Y} - b\bar{X} = \frac{\sum Y}{n} - b\frac{\sum X}{n} = \frac{700}{10} - 0.613 \times \frac{900}{10} = 14.81$$

Regression equation of Y on X is $Y = 14.81 + 0.613X$

Example # 2: The following data give the ages and blood pressure of 10 women

Age (X)	56	42	36	47	49	42	60	72	63	55
Blood pressure (Y)	147	125	118	128	145	140	155	160	149	150

Md. Moyazzem Hossain

a) Find the correlation coefficient between X and Y

b) Determine the least squares regression equation of Y on X.

c) Estimate the blood pressure of a women whose age is 45 years

Solution:

a) We know that, Correlation coefficient between X and Y is given by

$$r=\frac{\sum XY-\frac{\sum X\sum Y}{N}}{\sqrt{\left[\left\{\sum X^2-\frac{(\sum X)^2}{N}\right\}\left\{\sum Y^2-\frac{(\sum Y)^2}{N}\right\}\right]}}$$

Age (X)	Blood Pressure (Y)	X^2	Y^2	XY
56	147	3136	21609	8232
42	125	1764	15625	5250
36	118	1296	13924	4248
47	128	2209	16384	6016
49	145	2401	21025	7105
42	140	1764	19600	5880
60	155	3600	24025	9300
72	160	5184	25600	11520
63	149	3969	22201	9387
55	150	3025	22500	8250
$\sum X = 522$	$\sum Y = 1417$	$\sum X^2 = 28348$	$\sum Y^2 = 202493$	$\sum XY = 75188$

$$* \; r=\frac{75188-\frac{522\times1417}{10}}{\sqrt{\left[\left\{28348-\frac{(522)^2}{10}\right\}\left\{202493-\frac{(1417)^2}{10}\right\}\right]}}=0.891678842$$

b) We know that, the regression equation of Y on X is expressed as follows

$$Y=a+bX$$

Again, Regression coefficient of Y on X is

$$b_{yx}=\frac{\sum XY-\frac{\sum X\sum Y}{n}}{\sum X^2-\frac{(\sum X)^2}{n}}=\frac{75188-\frac{522\times1417}{10}}{28348-\frac{(522)^2}{10}}=1.110040015=1.11$$

$$a=\overline{Y}-b\overline{X}=\frac{\sum Y}{n}-b\frac{\sum X}{n}=\frac{1417}{10}-1.11\times\frac{522}{10}=83.75591124$$

Regression equation of Y on X is

$$Y=83.756+1.11X$$

Md. Moyazzem Hossain
Lecturer Dept. of Statistics, I

c) When $X = 45$ then

$Y = 83.756 + 1.11 \; 45 = 133.706$

Hence the most likely blood pressure of women of 45 years is 134.

Compare the correlation analysis with regression analysis.

Correlation	Regression
1) Correlation coefficient is symmetric i.e. $r_{xy} = r_{yx}$	1) Regression co-efficients are not symmetric in X and Y i.e. $b_{xy} \neq b_{yx}$
2) Correlation co-efficient r_{xy} is a relative measure of the linear relationship between X and Y and is independent of the units of the measurement. If is a pure number lying between ± 1.	2) The regression co-efficient $b_{yx}\left(b_{xy}\right)$ are absolute measures representing the change in the value of the variable Y (X) for a unit change in the variable X (Y).
3) Correlation analysis has limited applications as it is confined only to the study of linear relationship between the variables.	3) Regression analysis studies linear as well as non-linear relationship between the variables and therefore has much wider applications.

Uses:

- The relation can be used for predictive purpose.
- Regression analysis is widely used in statistical estimation of demand curves, supply curves, production functions; cost functions, consumption function etc.

Md. Moyazzem Hossain
Lecturer Dept. of Statistics, I