Unit 3: Regression analysis

Regression

The literal meaning of regression is 'step back towards the average'.

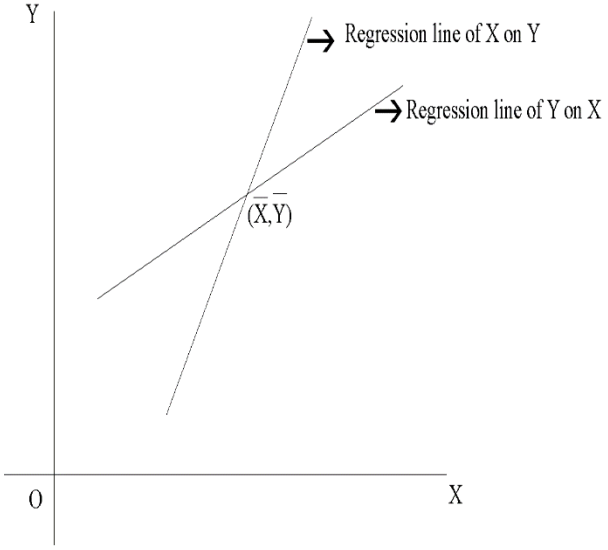
In regression analysis, there are two variables – dependent and independent. The value of the variable which is to be estimated or predicted is called dependent variable. The variable which is used for prediction is called independent variable. It is also called regressor or predictor or explanatory variables. The main objective of regression analysis is to establish a functional relationship between the dependent and independent variable and is used to predict the values of dependent variable.

Lines of regression

From the bivariate data we can plot scatter diagram as before and there we will find some points that will cluster round some curve and is known as curve of regression. If the curve is a straight line, the regression is said to be linear otherwise non linear or curvilinear. A linear relationship between two variables is described by a straight line through the points and is known as line of regression. A line of regression gives the best estimate of one unknown variable for any given value of a known variable. A line fitted by the method of least square is the best fit. In other words, the least squares method calculates the line that comes the closest to running through all of the data points i.e. this line is the one that passes through the center of the data points. Therefore, the least square method derives the equation of the line that relates the relationship between all of the data points with a minimum of error.

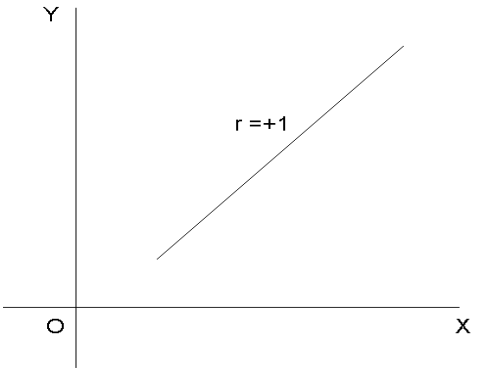
There are two lines of regression one is Y on X and other is X on Y. The line of regression Y on X is used to estimate the value of dependent variable Y for any given value of the independent variable X. Similarly, the line of regression of X on Y is used to estimate the value of dependent variable X for any given value of the independent variable Y.

The regression equation of Y on X is obtained by minimizing the sum of squares of error parallel to Y axis while in obtaining the regression equation of X on Y the sum of squares of errors are minimizing parallel to X axis.

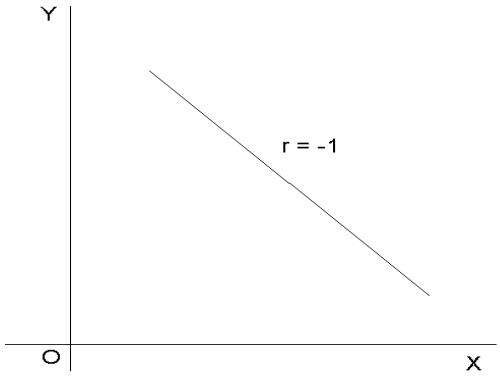


On plotting two lines of regression of the same graph, they intersect each other at the mean value of the variables i.e. on $(\bar{x}, \bar{y})$. From the angle enclosed by two lines of regression, we can get the idea about the correlation between variables. The lesser is the angle between two regression lines,

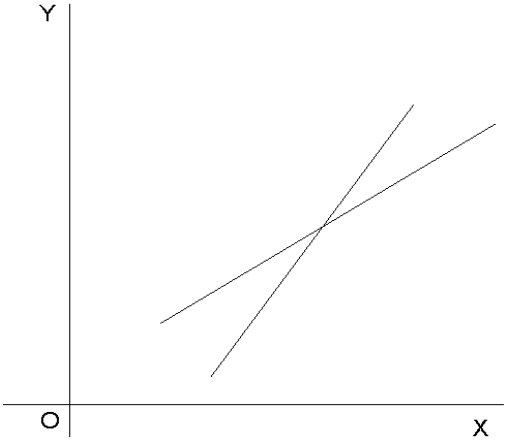
higher will be the correlation between variables and greater the angle between lines, lower will be the correlation. As the angle between two lines of regression is 90^o i.e. two lines are perpendicular to each other, then there will be no correlation which shows the variable are independent to each other. When there is a perfect correlation, then two lines of regression coincides each other and we will see only one line. This will be clearer after studying angle between two lines of regression.



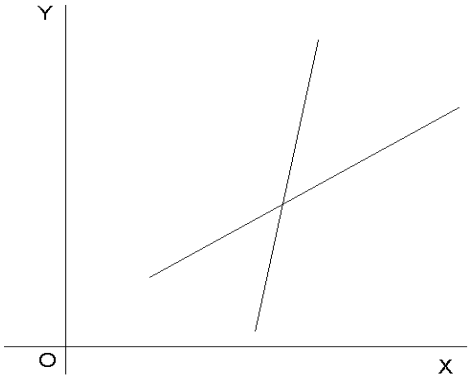
Perfect positive correlation



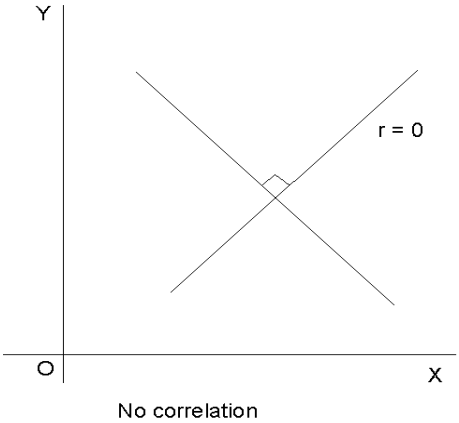
Perfectly negatively correlation



High degree of correlation



Low degree of correlation



No correlation

The linear regression model

Regression equation of Y on X

The line of regression of Y on X is,

$Y = a + bX$ ……..(i)

Where Y is dependent variable, X is independent variable, a is constant or y intercept, b is the regression coefficient of Y on X(slope of regression line on Y on X) and is also denoted by b_{yx}. The regression coefficient (b) is the average change in the dependent variable (*Y*) for a 1 unit change in the independent variable (*X*). It is the slope of the regression line.
The values of a and b are estimated by the method of least square (i.e. minimizing the sum of squares of error) which gives the normal equations as:
$\sum Y = na + b\sum X$ …….(ii)
or $\sum XY = a\sum X + b\sum X^2$ …….(iii)
Equation (ii) and (iii) are known as normal equations and solving these normal equations we will get the values of a and b.
However these values can be obtained by the following formula,

$$b_{YX} = \frac{\sum XY - n\overline{X}\overline{Y}}{\sum X^2 - n\overline{X}^2}$$ on simplification we will also get,

$$b_{YX} = \frac{n\sum XY - \sum X\sum Y}{n\sum X^2 - \left(\sum X\right)^2} \text{ and } a = \overline{Y} - b\overline{X}$$

Substituting the values of a and b in equation (i) we get the best fit of regression line of Y on X.

Regression equation of X on Y
The line of regression of X on Y is
$X = a' + b'Y$ ……..(1)
Where X is dependent variable, Y is independent variable, a' is constant or x intercept, b' is the regression coefficient of X on Y(slope of regression line on X on Y) and is also denoted by b_{xy}
The values of a' and b' are estimated by the method of least square (i.e. minimizing the sum of squares of error).
$\sum X = n\,a' + b'\sum Y$ …….(2)
or $\sum XY = a'\sum Y + b'\sum Y^2$ …….(3)
Equation (2) and (3) are known as normal equations and solving these normal equations we will get the values of a' and b'.
However these values can be obtained by the following formula,

$$b_{XY} = \frac{\sum XY - n\overline{X}\overline{Y}}{\sum Y^2 - n\overline{Y}^2}$$ on simplification we will also get,

$$b_{XY} = \frac{n\sum XY - \sum X\sum Y}{n\sum Y^2 - \left(\sum Y\right)^2} \text{ and } a' = \overline{X} - b'\overline{Y}$$

Substituting the values of a' and b' in equation (1) we get the best fit of regression line of X on Y.

Properties of regression coefficient

1. The range of regression coefficient is - ∞ to ∞.
2. Correlation coefficient is the geometric mean of regression coefficients.

   i.e. $r = \sqrt{b_{yx} \times b_{xy}}$

3. Both the regression coefficient has the same sign. If both regression coefficients are positive, then r will be positive and if both the regression coefficients are negative, then r will be negative.
4. If one regression coefficient is greater than one, other must be less than one i.e. product of two regression coefficients can not exceed one.

   i.e. $b_{yx} \times b_{xy} \leq 1$
5. Regression coefficient is independent of change in origin but not of scale.
6. If the variables X and Y are independent, the regression coefficients are zero. i.e if r = 0, then $b_{yx} = b_{xy} = 0$.
7. The arithmetic mean of regression coefficient is greater than the correlation coefficient.

   i.e. $\frac{1}{2}\left(b_{yx} + b_{xy}\right) \geq r$

Exercise

1. The following table gives the normal weight of a baby during the first six months of life.

Age in months	0	2	3	5	6
Weight in lbs	5	7	8	10	12

   Estimate the weight of a baby at the age of 4 months. **(8.877 lbs)**
2. During a laboratory experiment muscular contractions of a frog muscle were measured against different doses of a given drug. The height of the curves was considered as the response to the drug, the observations were as below:

	Serial number of the experiment				
	1	2	3	4	5
Dose of drug	0.3	0.4	0.6	0.8	0.9
Response to drug	54.0	59.0	60.0	65.0	70.0

   Calculate the response of the drug for a dose of 0.5. **(59.29)**
3. The following table gives the ages and blood pressure of 6 women.

Age in years	56	42	72	36	63	47
Blood pressure	147	125	160	118	149	128

   i. Compute the line of regression for estimating blood pressure.
   ii. Estimate the blood pressure of a woman whose age is 45 years.

   (Y = 1.2 X + 74.626; 128.626)
4. A study was reported in a medical journal suggesting that the peak heart rate of an individual can reach during intensive exercise decreases with age. A cardiologist wanted to do his own study. The next 9 patients were given a stress test on the tread mill at 6 miles per hour and their age (X) and their heart rates (Y) were recorded as follows:

X	30	30	40	20	20	45	30	45	50
Y	190	180	180	200	195	170	185	175	165

   Can we predict the peak heart rate of an over 80 year old man who is given a similar stress test? If so, what peak heart rate do you predict? **(144.51)**
5. Table below shows the systolic blood pressure (Y) recorded at various time in minutes (X) on an individual.

Time periods(X)	0	5	10	15	20
Diastolic blood pressure (Y)	72	66	70	64	66

   Calculate the two regression coefficients.
6. Calculate the two regression coefficient from the following data.

Age	30	33	38	42	50	57	62	69	71

BMI 28.3 25.3 22.7 39.6 28.7 27.1 26.6 28.9 26.7

7. The height of fathers and sons (in inches) is given in the following table. Find the two lines of regression and estimate the expected average height of the son when the height of the father is 67.5 inches.

Height of father	65	66	67	67	68	69	71	73
Height of son	67	68	64	68	72	70	69	70

(Y = 39.5484 + 0.4242X; X = 32.2875 + 0.525Y; 68.18 inches)

Multiple regression analysis

In simple regression analysis, we studied the linear relationship between only two variables, one independent and other dependent. Based on the relationship, we could predict the value of dependent variable for a given value of independent variable. Multiple regression analysis consists of the measurement of the relationship between the dependent variable and two or more independent variables. The procedure is similar to that of simple regression, with a difference that other independent variables are added to the regression equations.

The multiple regression equation of dependent variable X_1 on independent variables X_2 and X_3 is given by,

$X_1 = a + b_1X_2 + b_2 X_3$ …… (1)

a = y intercept

b_1 = the partial regression coefficeint of X_1 on X_2 keeping X_3 constant (also written as $b_{12.3}$)

b_2 = the partial regression coefficeint of X_1 on X_3 keeping X_2 constant (also written as $b_{13.2}$)

The values of a, b_1 and b_2 are estimated by the method of least square which gives the normal equations as,

$$\sum X_1 = na + b_1 \sum X_2 + b_2 \sum X_3 \text{ …… (2)}$$

$$\sum X_1X_2 = a\sum X_2 + b_1 \sum X_2^2 + b_2 \sum X_2X_3 \text{ ……. (3)}$$

$$\sum X_1X_3 = a\sum X_3 + b_1 \sum X_2X_3 + b_2 \sum X_3^2 \text{ …….. (4)}$$

Solving equations (2), (3) and (4) we get the values of a, b_1 and b_2. Substituting these values in equation (1), we get the fitted regression equation of X_1 on X_2 and X_3.

The multiple regression equation of dependent variable X_2 on independent variables X_1 and X_3 is given by,

$X_2 = a + b_1X_1 + b_2 X_3$ …… (i)

a = y intercept

b_1 = the partial regression coefficeint of X_2 on X_1 keeping X_3 constant (also written as $b_{21.3}$)

b_2 = the partial regression coefficeint of X_2 on X_3 keeping X_1 constant (also written as $b_{23.1}$)

The values of a, b_1 and b_2 are estimated by the method of least square which gives the normal equations as,

$$\sum X_2 = na + b_1 \sum X_1 + b_2 \sum X_3 \text{ …… (ii)}$$

$$\sum X_1X_2 = a\sum X_1 + b_1 \sum X_1^2 + b_2 \sum X_1X_3 \text{ ……. (iii)}$$

$$\sum X_2X_3 = a\sum X_3 + b_1 \sum X_1X_3 + b_2 \sum X_3^2 \text{ …….. (iv)}$$

Solving equations (ii), (iii) and (iv) we get the values of a, b_1 and b_2. Substituting these values in equation (i), we get the fitted regression equation of X_2 on X_1 and X_3.

Similarly we can obtain multiple regression equation of dependent variable X_3 on independent variables X_1 and X_2.

Exercise

1. Numerical data had been collected for 5 years, according to the values of the dependent variable X_1 = number of doctors per 10000 population ; independent variables X_2 = average annual income (in Lakh) and X_3 = density of health care clinics.

Obsevation No.	1	2	3	4	5
X_1	23	18	8	20	16
X_2	12	8	5	9	6
X_3	7	5	12	6	10

   Find the prediction equation which best fits and estimate the number of doctors whose average annual income is 10 lakh and desity of health care clinics is 8.

2. A developer of food for pigs would like to determine what relationship existis among the age of a pig when it starts receiving a newly developed food supplement, the initial weight of the pig, and the amount of weight it gains in 1 week period with the food supplement. The following information is the result of a study of eight piglets :

Piglet no.	Initial weight (pounds)	Initial age (weeks)	Weight gain
1	39	8	7
2	52	6	6
3	49	7	8
4	46	12	10
5	61	9	9
6	35	6	5
7	25	7	3
8	55	4	4

   How much might we expect a pig to gain in a week with the food supplement if it were 9 weeks old and weighed 48 pounds ?