REGRESSION

Meaning:

A study of measuring the relationship between associated variables, wherein one variable is dependent on another independent variable, called as Regression. It is developed by Sir Francis Galton in 1877 to measure the relationship of height between parents and their children.

Regression analysis is a statistical tool to study the nature and extent of functional relationship between two or more variables and to estimate (or predict) the unknown values of dependent variable from the known values of independent variable.

The variable that forms the basis for predicting another variable is known as the Independent Variable and the variable that is predicted is known as dependent variable. For example, if we know that two variables price (X) and demand (Y) are closely related we can find out the most probable value of X for a given value of Y or the most probable value of Y for a given value of X. Similarly, if we know that the amount of tax and the rise in the price of a commodity are closely related, we can find out the expected price for a certain amount of tax levy.

Uses of Regression Analysis:

1. It provides estimates of values of the dependent variables from values of independent variables.
2. It is used to obtain a measure of the error involved in using the regression line as a basis for estimation.
3. With the help of regression analysis, we can obtain a measure of degree of association or correlation that exists between the two variables.
4. It is highly valuable tool in economies and business research, since most of the problems of the economic analysis are based on cause and effect relationship.

Distinction between Correlation and Regression

Sl No	Correlation	Regression
1	It measures the degree and direction of relationship between the variables.	It measures the nature and extent of average relationship between two or more variables in terms of the original units of the data
2	It is a relative measure showing association between the variables.	It is an absolute measure of relationship.
3	Correlation Coefficient is independent of change of both origin and scale.	Regression Coefficient is independent of change of origin but not scale.
4	Correlation Coefficient is independent of units of measurement.	Regression Coefficient is not independent of units of measurement.
5	Expression of the relationship between the variables ranges from −1	Expression of the relationship between the variables may be in any

	to +1.	of the forms like: $Y = a + bX$ $Y = a + bX + cX^2$
6	It is not a forecasting device.	It is a forecasting device which can be used to predict the value of dependent variable from the given value of independent variable.
7	There may be zero correlation such as weight of wife and income of husband.	There is nothing like zero regression.

Regression Lines and Regression Equation:

Regression lines and regression equations are used synonymously. Regression equations are algebraic expression of the regression lines. Let us consider two variables: X & Y. If y depends on x, then the result comes in the form of simple regression. If we take the case of two variable X and Y, we shall have two regression lines as the regression line of X on Y and regression line of Y on X. The regression line of Y on X gives the most probable value of Y for given value of X and the regression line of X on Y given the most probable value of X for given value of Y. Thus, we have two regression lines. However, when there is either perfect positive or perfect negative correlation between the two variables, the two regression line will coincide, i.e. we will have one line. If the variables are independent, r is zero and the lines of regression are at right angles i.e. parallel to X axis and Y axis.

Therefore, with the help of simple linear regression model we have the following two regression lines

1. Regression line of Y on X: This line gives the probable value of Y (Dependent variable) for any given value of X (Independent variable).
   Regression line of Y on X : $Y - \dot{Y} = b_{yx}(X - \dot{X})$
   OR : $Y = a + bX$

2. Regression line of X on Y: This line gives the probable value of X (Dependent variable) for any given value of Y (Independent variable).
   Regression line of X on Y : $X - \dot{X} = b_{xy}(Y - \dot{Y})$
   OR : $X = a + bY$

In the above two regression lines or regression equations, there are two regression parameters, which are "a" and "b". Here "a" is unknown constant and "b" which is also denoted as "b_{yx}" or "b_{xy}", is also another unknown constant popularly called as regression coefficient. Hence, these "a" and "b" are two unknown constants (fixed numerical values) which determine the position of the line completely. If the value of either or both of them is changed, another line is determined. 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).

If the values of constants "a" and "b" are obtained, the line is completely determined. But the question is how to obtain these values. The answer is provided by the method of least squares. With the little algebra and differential calculus, it can be shown that the following two ***normal equations,*** if solved simultaneously, will yield the values of the parameters "a" and "b".

Two normal equations:

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

This above method is popularly known as direct method, which becomes quite cumbersome when the values of X and Y are large. This work can be simplified if instead of dealing with actual values of X and Y, we take the deviations of X and Y series from their respective means. In that case:

Regression equation Y on X:

$Y = a + bX$ will change to $(Y - \dot{Y}) = b_{yx}(X - \dot{X})$

Regression equation X on Y:

$X = a + bY$ will change to $(X - \dot{X}) = b_{xy}(Y - \dot{Y})$

In this new form of regression equation, we need to compute only one parameter i.e. "b". This "b" which is also denoted either "b_{yx}" or "b_{xy}" which is called as regression coefficient.

Regression Coefficient:

The quantity "b" in the regression equation is called as the regression coefficient or slope coefficient. Since there are two regression equations, therefore, we have two regression coefficients.

1. Regression Coefficient X on Y, symbolically written as "b_{xy}"
2. Regression Coefficient Y on X, symbolically written as "b_{yx}"

Different formula's used to compute regression coefficients:

Method	Regression Coefficient X on Y	Regression Coefficient Y on X
Using the correlation coefficient (r) and standard deviation (σ)	$b_{xy} = r\frac{\sigma x}{\sigma y}$	$b_{yx} = r\frac{\sigma y}{\sigma x}$
Direct Method: Using sum of X and Y	$b_{xy} = \frac{N\sum XY - \sum X\sum Y}{N\sum Y^2 - (\sum Y)^2}$	$b_{yx} = \frac{N\sum XY - \sum X\sum Y}{N\sum X^2 - (\sum X)^2}$
When deviations are taken from arithmetic mean	$b_{xy} = \frac{\sum xy}{\sum y^2}$ where $x = X - \dot{X}$ and $y = Y - \dot{Y}$	$b_{yx} = \frac{\sum xy}{\sum x^2}$ where $x = X - \dot{X}$ and $y = Y - \dot{Y}$

Properties of Regression Coefficients:

1. The coefficient of correlation is the geometric mean of the two regression coefficients. Symbolically $r = \sqrt{bxy * byx}$

2. If one of the regression coefficients is greater than unity, the other must be less than unity, since the value of the coefficient of correlation cannot exceed unity. For example if bxy = 1.2 and byx = 1.4 "r" would be = $\sqrt{1.2 * 1.4} = 1.29$, which is not possible.
3. Both the regression coefficient will have the same sign. i.e. they will be either positive or negative. In other words, it is not possible that one of the regression coefficients is having minus sign and the other plus sign.
4. The coefficient of correlation will have the same sign as that of regression coefficient, i.e. if regression coefficient have a negative sign, "r" will also have negative sign and if the regression coefficient have a positive sign, "r" would also be positive. For example, if $b_{xy} = -0.2$ and $b_{yx} = -0.8$ then $r = -\sqrt{0.2 * 0.8} = \mathbf{-0.4}$
5. The average value of the two regression coefficient would be greater than the value of coefficient of correlation. In symbol (bxy + byx) / 2 > r. For example, if bxy = 0.8 and byx = 0.4 then average of the two values = (0.8 + 0.4) / 2 = 0.6 and the value of r = r = $\sqrt{0.8 * 0.4} = 0.566$ which less than 0.6
6. Regression coefficients are independent of change of origin but not scale.

Illustration 01:

Find the two regression equation of X on Y and Y on X from the following data:

X	:	10	12	16	11	15	14	20	22
Y	:	15	18	23	14	20	17	25	28

Solution:

Calculation of Regression Equation

X	Y	X^2	Y^2	XY
10	15	100	225	150
12	18	144	324	216
16	23	256	529	368
11	14	121	196	154
15	20	225	400	300
14	17	196	289	238
20	25	400	625	500
22	28	484	784	616
120	**160**	**1,926**	**3,372**	**2,542**
$\sum X$	**$\sum Y$**	**$\sum X^2$**	**$\sum Y^2$**	**$\sum XY$**

Here N = Number of elements in either series X or series Y = 8

Now we will proceed to compute regression equations using normal equations.

Regression equation of X on Y: **X = a + bY**

The two normal equations are:

$$\sum X = Na + b\sum Y$$

$$\sum XY = a\sum Y + b\sum Y^2$$

Substituting the values in above normal equations, we get

120 = 8a + 160b ..... (i)

2542 = 160a + 3372b ..... (ii)

Let us solve these equations (i) and (ii) by simultaneous equation method

Multiply equation (i) by 20 we get 2400 = 160a + 3200b

Now rewriting these equations:

2400 = 160a + 3200b

2542 = 160a + 3372b

(-) (-) (-)

-142 = -172b

Therefore now we have -142 = -172b, this can rewritten as 172b = 142

Now, $b = \frac{142}{172} = 0.8256$ (rounded off)

Substituting the value of b in equation (i), we get

120 = 8a + (160 * 0.8256)

120 = 8a + 132 (rounded off)

8a = 120 - 132

8a = -12

a = -12/8

a = -1.5

Thus we got the values of a = -1.5 and b = 0.8256

Hence the required regression equation of X on Y:

X = a + bY => **X = -1.5 + 0.8256Y**

Regression equation of Y on X: **Y = a + bX**

The two normal equations are:

$\sum Y = Na + b\sum X$

$\sum XY = a\sum X + b\sum X^2$

Substituting the values in above normal equations, we get

160 = 8a + 120b ..... (iii)

2542 = 120a + 1926b ..... (iv)

Let us solve these equations (iii) and (iv) by simultaneous equation method

Multiply equation (iii) by 15 we get 2400 = 120a + 1800b

Now rewriting these equations:

2400 = 120a + 1800b

2542 = 120a + 1926b

(-) (-) (-)

-142 = -126b

Therefore now we have -142 = -126b, this can rewritten as 126b = 142

Now, $b = \frac{142}{126} = 1.127$ (rounded off)

Substituting the value of b in equation (iii), we get

160 = 8a + (120 * 1.127)

160 = 8a + 135.24

8a = 160 - 135.24

8a = 24.76

a = 24.76/8

a = 3.095

Thus we got the values of a = 3.095 and b = 1.127

Hence the required regression equation of Y on X:

$Y = a + bX$ => **$Y = 3.095 + 1.127X$**

Illustration 02:

After investigation it has been found the demand for automobiles in a city depends mainly, if not entirely, upon the number of families residing in that city. Below are the given figures for the sales of automobiles in the five cities for the year 2019 and the number of families residing in those cities.

City	No. of Families (in lakhs): X	Sale of automobiles (in '000): Y
Belagavi	70	25.2
Bangalore	75	28.6
Hubli	80	30.2
Kalaburagi	60	22.3
Mangalore	90	35.4

Fit a linear regression equation of Y on X by the least square method and estimate the sales for the year 2020 for the city Belagavi which is estimated to have 100 lakh families assuming that the same relationship holds true.

Solution:

Calculation of Regression Equation

City	X	Y	X^2	XY
Belagavi	70	25.2	4900	1764
Bangalore	75	28.6	5625	2145
Hubli	80	30.2	6400	2416
Kalaburagi	60	22.3	3600	1338
Mangalore	90	35.4	8100	3186
	375 ΣX	**141.7** ΣY	**28,625** ΣX^2	**10,849** ΣXY

Regression equation of Y on X: **$Y = a + bX$**

The two normal equations are:

$\Sigma Y = Na + b\Sigma X$

$\Sigma XY = a\Sigma X + b\Sigma X^2$

Substituting the values in above normal equations, we get

141.7 = 5a + 375b ..... (i)

10849= 375a + 28625b ..... (ii)

Let us solve these equations (i) and (ii) by simultaneous equation method

Multiply equation (i) by 75 we get 10627.5 = 375a + 28125b

Now rewriting these equations:

10627.5	=	375a	+	28125b	
10849	=	375a	+	28625b	
(-)		(-)		(-)	.
-221.5	=			-500b	

Therefore now we have -221.5 = -500b, this can rewritten as 500b = 221.5

Now, $b = \frac{221.5}{500} = 0.443$

Substituting the value of b in equation (i), we get

141.7	=	5a	+	(375 * 0.443)
141.7	=	5a	+	166.125
5a	=	141.7	-	166.125
5a	=	-24.425		
a	=	-24.425/5		
a	=	-4.885		

Thus we got the values of a = -4.885 and b = 0.443

Hence, the required regression equation of Y on X:

Y = a + bX => **Y = -4.885 + 0.443X**

Estimated sales of automobiles (Y) in city Belagavi for the year 2020, where number of families (X) are 100(in lakhs):

Y = -4.885 + 0.443X

Y = -4.885 + (0.443 * 100)

Y = -4.885 + 44.3

Y = 39.415 ('000)

Means sales of automobiles would be 39,415 when number of families are 100,00,000

Illustration 03:

From the following data obtain the two regression lines:

Capital Employed (Rs. in lakh):	7	8	5	9	12	9	10	15
Sales Volume (Rs. in lakh):	4	5	2	6	9	5	7	12

Solution:

Calculation of Regression Equation

X	Y	X^2	Y^2	XY
7	4	49	16	28
8	5	64	25	40
5	2	25	4	10
9	6	81	36	54
12	9	144	81	108
9	5	81	25	45
10	7	100	49	70
15	12	225	144	180
75	**50**	**769**	**380**	**535**
ΣX	**ΣY**	**ΣX^2**	**ΣY^2**	**ΣXY**

Regression line/equation of X on Y:

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

$\dot{X} = \frac{\sum X}{n} = \frac{75}{8} = 9.375$

$\dot{Y} = \frac{\sum Y}{n} = \frac{50}{8} = 6.25$

Regression coefficient of X on Y:

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

$b_{xy} = \frac{(8 * 535) - (75 * 50)}{(8 * 380) - (50)^2}$

$= \frac{4280 - 3750}{3040 - 2500}$

$= \frac{530}{540} = \underline{0.9815}$

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

⇨ $X - 9.375 = 0.9815 (Y - 6.25)$

⇨ $X - 9.375 = 0.9815Y - 6.1344$

⇨ $X = 9.375 - 6.1344 + 0.9815Y$

⇨ **$X = 3.2406 + 0.9815Y$**

Regression line/equation of Y on X:

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

$\dot{X} = \frac{\sum X}{n} = \frac{75}{8} = 9.375$

$\dot{Y} = \frac{\sum Y}{n} = \frac{50}{8} = 6.25$

Regression coefficient of Y on X:

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

$b_{xy} = \frac{(8 * 535) - (75 * 50)}{(8 * 769) - (75)^2}$

$= \frac{4280 - 3750}{6152 - 5625}$

$= \frac{530}{527} = \underline{1.0057}$

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

⇨ $Y - 6.25 = 1.0057 (X - 9.375)$

⇨ $Y - 6.25 = 1.0057X - 9.4284)$

⇨ $Y = 6.25 - 9.4284 + 1.0057X$

⇨ **$Y = -3.1784 + 1.0057X$**

Illustration 04:

From the following information find regression equations and estimate the production when the capacity utilisation is 70%.

	Average (Mean)	Standard Deviation
Production (in lakh units)	42	12.5
Capacity Utilisation (%)	88	8.5
Correlation Coefficient (r)	0.72	

Solution:

Let production be variable X and capacity utilisation be variable Y. Regression equation of production based on based on capacity utilisation shall be given by X on Y and regression equation of capacity utilisation of production shall be given by Y on X, which can be computed as given below:

Given Information: $\dot{X} = 42$ $\dot{Y} = 88$ $\sigma_x = 12.5$ $\sigma_y = 8.5$ $r = 0.72$

Regression coefficient of X on Y:

$b_{xy} = r\frac{\sigma x}{\sigma y} = 0.72 * \frac{12.5}{8.5} = 1.0588$

Regression Equation of X on Y:

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

⇨ $X - 42 = 1.0588 (Y - 88)$

⇨ $X = 42 - 93.1744 + 1.0588Y$

⇨ **$X = -51.1744 + 1.0588Y$**

Regression coefficient of Y on X:

$b_{yx} = r\frac{\sigma y}{\sigma x} = 0.72 * \frac{8.5}{12.5} = 0.4896$

Regression Equation of Y on X:

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

⇨ $Y - 88 = 0.4896 (X - 42)$

⇨ $Y = 88 - 20.5632 + 0.4896X$

⇨ **$Y = 67.4368 + 0.4896X$**

Estimation of the production when the capacity utilisation is 70% is regression equation X on Y, where Y = 70

Regression Equation of X on Y:

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

$X = -51.1744 + 1.0588Y$

$= -51.1744 + (1.0588 * 70)$

$= -51.1744 + 74.116$

$= \mathbf{22.9416}$

Therefore, the estimated production would be **22,94,160** units when there is a capacity utilisation of 70%.

Illustration 05:

The following data gives the age and blood pressure (BP) of 10 sports persons.

Name	:	A	B	C	D	E	F	G	H	I	J
Age (X)	:	42	36	55	58	35	65	60	50	48	51
BP (Y)	:	98	93	110	85	105	108	82	102	118	99

i. Find regression equation of Y on X and X on Y (Use the method of deviation from arithmetic mean)

ii. Find the correlation coefficient (r) using the regression coefficients.

iii. Estimate the blood pressure of a sports person whose age is 45.

Solution:

Calculation of Regression Equation

Name	Age (X)	BP (Y)	x=X-$\dot{X}$ x=X-50	y=Y-$\dot{Y}$ y=Y-100	x^2	y^2	xy
A	42	98	-8	-2	64	4	16
B	36	93	-14	-7	196	49	98
C	55	110	5	10	25	100	50
D	58	85	8	-15	64	225	-120
E	35	105	-15	5	225	25	-75
F	65	108	15	8	225	64	120
G	60	82	10	-18	100	324	-180
H	50	102	0	2	0	4	0
I	48	118	-2	18	4	324	-36
J	51	99	1	-1	1	1	-1
	500 ΣX	**1,000** ΣY	**0** Σx	**0** Σy	**904** Σx^2	**1,120** Σy^2	**-128** Σxy

$$\dot{X} = \frac{\Sigma X}{n} = \frac{500}{10} = 50 \qquad \dot{Y} = \frac{\Sigma Y}{n} = \frac{1000}{10} = 100$$

Regression coefficients can be computed using the following formula:

$$b_{xy} = \frac{\Sigma xy}{\Sigma y^2} \qquad b_{yx} = \frac{\Sigma xy}{\Sigma x^2}$$

where $x = X - \dot{X}$ and $y = Y - \dot{Y}$

Regression coefficient of X on Y:

$$b_{xy} = \frac{\sum xy}{\sum y^2} = \frac{-128}{1120} = -0.1143$$

Regression coefficient of Y on X:

$$b_{yx} = \frac{\sum xy}{\sum x^2} = \frac{-128}{904} = -0.1416$$

Regression equation of X on Y:

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

$\Rightarrow X - 50 = -0.1143 (Y - 100)$

$\Rightarrow X - 50 = -0.1143Y + 11.43$

$\Rightarrow X = 50 + 11.43 - 0.1143Y$

$\Rightarrow$ **$X = 61.43 - 0.1143Y$**

Regression equation of Y on X:

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

$\Rightarrow Y - 100 = -0.1416 (X - 50)$

$\Rightarrow Y - 100 = -0.1416X + 7.08$

$\Rightarrow Y = 100 + 7.08 - 0.1416X$

$\Rightarrow$ **$Y = 107.08 - 0.1416X$**

Computation of coefficient of correlation using regression coefficient:

$$r = \sqrt{bxy * byx} = -\sqrt{0.1143 * 0.1416} = -\sqrt{0.01618488} = -0.1272$$

Therefore, we have low degree of negative correlation between age and blood pressure of sports person.

Estimation of the blood pressure (Y) of a sports person whose age is X=45 can be calculated using regression equation Y on X:

Regression equation of Y on X:

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

$\Rightarrow Y = 107.08 - 0.1416X = 107.08 - (0.1416 * 45) = 107.08 - 6.372 =$ **<u>100.708</u>**

It means estimated blood pressure of a sports person is 101 (rounded off) whose age is 45.

Illustration 06:

There are two series of index numbers, ***P*** for price index and ***S*** for stock of commodity. The mean and standard deviation of *P* are 100 and 8 and *S* are 103 and 4 respectively. The correlation coefficient between the two series is 0.4. With these data, work out a linear equation to read off values of *P* for various values of *S*. Can the same equation be used to read off values of *S* for various values of *P*?

Solution:

Let us assume that *P*=Price Index be variable X an *S*=Stock of Commodity be variable Y. Linear equation to read off values of *P* for various values of S would be regression equation of X on Y. Regression coefficient is to be computed using mean and standard deviation.

From the problem we can list out the given information:

$\dot{X} = 100$ $\quad \dot{Y} = 103$ $\quad \sigma_x = 8$ $\quad \sigma_y = 4$ $\quad r = 0.4$

Regression equation of X on Y:

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

$\Rightarrow (X - \dot{X}) = r \frac{\sigma x}{\sigma y} (Y - \dot{Y})$

⇨ $(X - 100) = (0.4 * \frac{8}{4})(Y - 103)$

⇨ $(X - 100) = 0.8\,(Y - 103)$

⇨ $(X - 100) = 0.8Y - 82.4$

⇨ $X = 100 - 82.4 + 0.8Y$

⇨ **$X = 17.6 + 0.8Y$**

Linear equation to read off values of *P* for various values of *S* is $X = 17.6 + 0.8Y$

To read off values of S for various values of P we need regression equation of Y on X and therefore above linear equation cannot be used. Hence, the following regression equation of Y on X be computed:

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

⇨ $(Y - \dot{Y}) = r\frac{\sigma y}{\sigma x}(X - \dot{X})$

⇨ $(Y - 103) = 0.4 * \frac{4}{8}(X - 100)$

⇨ $(Y - 103) = 0.2\,(X - 100)$

⇨ $Y - 103 = 0.2X - 20$

⇨ $Y = 103 - 20 + 0.2X$

⇨ **$Y = 83 + 0.2X$**

Hence, the linear equation to read off values of S for various values of *P* is $Y = 83 + 0.2X$

Review of Correlation and Regression Analysis:

In correlation analysis, when we are keen to know whether two variables under study are associated or correlated and if correlated what is the strength of correlation. The best measure of correlation is proved by Karl Pearson's Coefficient of Correlation. However, one severe limitation of this method is that it is applicable only in case of a linear relationship between two variables. If two variables say X and Y are independent or not correlated then the result of correlation coefficient is zero.

Correlation coefficient measuring a linear relationship between the two variables indicates the amount of variation one variable accounted for by the other variable. A better measure for this purpose is provided by the square of the correlation coefficient, known as "coefficient of determination". This can be interpreted as the ratio between the explained variance to total variance:

$r^2 = \frac{\text{Explained variance}}{\text{Total Variance}}$ Similarly, Coefficient of non-determination = $(1 - r^2)$.

Regression analysis is concerned with establishing a functional relationship between two variables and using this relationship for making future projection. This can be applied, unlike correlation for any type of relationship linear as well as curvilinear. The two lines of regression coincide i.e. become identical when r= -1 or +1 in other words, there is a perfect negative or positive correlation between the two variables under discussion if r = 0, then regression lines are perpendicular to each other.

* * * * *