[Academic Script]

Regression

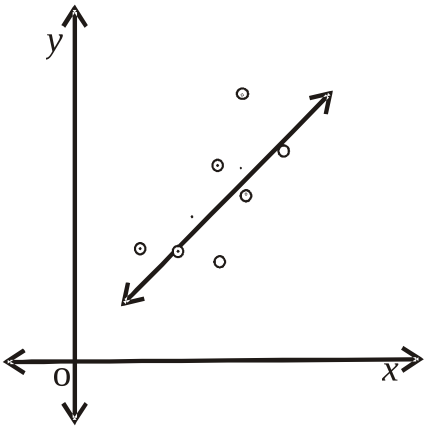
Subject:	Business Economics
Course:	B. A. (Hons.), 1st Semester, Undergraduate
Paper No. & Title:	Paper – 102 Statistics for Business Economics
Unit No. & Title:	Unit – 3 Multivariate Analysis
Lecture No. & Title:	Lecture – 2 Regression

1. Introduction

When there are simultaneous changes in the values of two variables and change in the value of one variable is due to change in the value of other variable then it can be said that there exists correlation between two variables and the value of correlation coefficient expresses the degree and direction of the relationship between variables. But it cannot give the estimate value one variable for the given value of other variable. For example if the data regarding the population and per capita income of certain region is known then by using correlation the nature and the strength of the relationship can be studied but if one wants to predict the value of per capita income for the given population then some mathematical relation has to be obtained between the variables, such relation is called regression. In order to obtain relationship between two or more correlated variables regression analysis is to be studied. ***A mathematical or functional relationship between two correlated variables is called regression.*** This relation is used to estimate the value of one variable for the given value of other variable. The variable for which the value is given is called independent variable and the variable for which the value is obtained by using the relation is called dependent variable.

2. Regression line

A simple method of studying relationship between two variables is the scatter diagram method.



In the above figure a scatter diagram is drawn for the data. Now in order to fixed up the relationship one has to draw a line, for that there are two concept (i) draw a line which covers maximum number of points, which is not reliable because different person draw a line in different manner (ii) draw a line which passes from nearer to all points in a scatter diagram, i.e. a line which minimize the distance between the points (actual values) and the line (estimated values). The distance or difference between the points and line is called error in estimation based on this, a principle is obtained to fixed up a line is that "draw a line which minimize the distance between the points and a line" or equivalently " draw a line which minimize the error sum of squares" this principle is called least square principle. ***A line obtained by applying least square principle is called is called regression line or a line of average relationship or best fitted line.*** When least square principle is applied by taking deviations parallel to y axis then a regression line of y on x is obtained which is given as $y = a + b_{yx}x$ where a and b are constants to be determined by using least square method and x is independent (cause) variable and y is dependent (effect) variable. If the least square principle is applied by taking deviations parallel to x axis then a regression line of x on y is obtained which is given as $x = a + b_{xy} y$ where a and b are constants and y is independent (cause) variable and x is dependent (effect) variable. Here it should be noted that in regression a functional relation is established in between the correlated variables which is nothing but cause and effect relationship. In practice it may happen that some time a variable is treated as cause (independent variable) and some time it may be treated as effect (dependent variable) so it is usual practice to obtain two regression lines for the given data.

3. Derivation of Regression line of y on x

Let us consider the equation of regression line of y on x as $y = a + bx + e$ where e is the error term and for the given value of $x$the estimated value of y is$\hat{y} = a + bx$. Here the constants a and b is determined by using least square principle, i.e., in such a way the error sum of squares is minimum

i.e. $e = \sum(y - \hat{y})^2$ is minimum.

$e = \sum(y - a - bx)^2$ is minimum.

For the minimization $\frac{\partial e}{\partial a} = 0$ and $\frac{\partial e}{\partial b} = 0$gives following normal equations

$$\sum y = na + b\sum x \quad ......(1)$$
$$\sum xy = a\sum x + b\sum x^2 .....(2)$$

Solving these equations for a and b following formulas can be obtained.

$$b = b_{yx} = \frac{Cov(x, y)}{S_x^2} \qquad and \qquad a = \overline{y} - b_{yx}\overline{x}$$

By substituting these values in the equation $y = a + bx$ the fitted or regression equation of y on x is given as $y - \overline{y} = b_{yx}(x - \overline{x})$

Similarly the fitted equation or regression equation of x on y is given as $x-\bar{x}=b_{xy}(y-\bar{y})$

4. Regression coefficients

In the regression equation of$y$ on x, $y=a+b_{yx}x$ the coefficient of x give the value of the regression coefficient of y on x and is denoted by b_{yx} it also indicate the rate of change or slope of the equation and the constant a indicate y - intercept. The regression equation of$x$ on y, $x=a+b_{xy}y$ the coefficient of y gives the value of the regression coefficient of x on y and is denoted by. ***A numerical measure which shows probable change in the values of dependent variable for a unit change in the value of independent variable is called regression coefficient.*** It shows the marginal value of the equation.

The regression coefficient y on x is denoted by b_{yx} and it indicate the probable change in the value of y for a unit change in the value of$x$. Some important formulae are as under:

1. $b_{yx}=\dfrac{Cov(x,y)}{S_x^2}$

2. $b_{yx}=r\dfrac{S_y}{S_x}$

3. $b_{yx}=\dfrac{\sum(x-\bar{x})(y-\bar{y})}{nS_x^2}$

4. $b_{yx}=\dfrac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2}$

5. $b_{yx}=\dfrac{\sum xy-n\bar{x}\bar{y}}{\sum x^2-n\bar{x}^2}$

6. $b_{yx}=\dfrac{n\sum xy-(\sum x)(\sum x)}{n\sum x^2-(\sum x)^2}$;

7. $b_{yx}=\dfrac{n\sum uv-(\sum u)(\sum v)}{n\sum x^2-(\sum x)^2}\times\dfrac{C_y}{C_x}$;$where\ \ u=\dfrac{x-A}{C_x},\ \ u=\dfrac{y-B}{C_y}$

Also the regression coefficient of x on y is denoted by b_{xy} and it indicate the probable change in the value of x for a unit change in the value of y. Some important formulae are as under:

1. $b_{xy} = \dfrac{Cov(x,y)}{S_y^2}$

2. $b_{xy} = r\dfrac{S_y}{S_x}$

3. $b_{xy} = \dfrac{\sum (x-\overline{x})(y-\overline{y})}{nS_y^2}$

4. $b_{xy} = \dfrac{\sum (x-\overline{x})(y-\overline{y})}{\sum (y-\overline{y})^2}$

5. $b_{xy} = \dfrac{\sum xy - n\overline{x}\,\overline{y}}{\sum y^2 - n\overline{y}^2}$

6. $b_{xy} = \dfrac{n\sum xy - (\sum x)(\sum x)}{n\sum x^2 - (\sum x)^2}$;

7. $b_{xy} = \dfrac{n\sum uv - (\sum u)(\sum v)}{n\sum y^2 - (\sum y)^2} \times \dfrac{C_x}{C_y}$ $;where\ u = \dfrac{x-A}{C_x},\ u = \dfrac{y-B}{C_y}$

(a) Important properties of Regression coefficients:

1. Correlation coefficient is a geometric mean of regression coefficients. i.e. $r = \pm\sqrt{b_{xy} \times b_{yx}}$
2. Sign of correlation coefficient and the regression coefficients are always same and depend on the sign of covariance.
3. The value of one regression coefficient may be more than one but at the same time the value of other regression coefficient must be less than one so that teir product must be less than one.
4. In the case of perfect correlation both the regression coefficients are reciprocal to each other.

5. Regression coefficients are independent of change of origin but not of scale.

(b) Important properties of Regression lines:

1. In the case of perfect correlation two regression lines are co-inside and vice-versa. i.e. in this case there is only one regression line.
2. In the case of no linear correlation two regression lines are perpendicular to each other and vice-versa.
3. As the angle between two regression lines increases, correlation between the variable decreases and vice-versa.
4. Two regression lines intersect each other at the point $(\bar{x}, \bar{y})$ i.e. the value of $\bar{x}$ and the value of $\bar{y}$ satisfies both the equations of regression.
5. Slope of the regression line y on x is b_{yx} and that of x on y is $\frac{1}{b_{xy}}$

Notes:

1. Two regression lines never be parallel to each other, either they co-inside or intersect each other at the point $(\bar{x}, \bar{y})$.
2. In practice, a regression line of dependent variable on independent variable is obtained. For example there exists positive correlation between blood pressure and age of a person and practically blood pressure of a person depend upon the age then one has to obtain a regression line of blood pressure on age of person.

5. Standard error of estimates

By using the fitted equation of regression line, one can obtain the estimated value of dependent variable for the given value of independent variable. In order to measure the reliability of the fitted equation, standard error of estimate is used. The standard error of the estimates measures the variability or scatter of the observed values around the regression line. It can be calculated by using the following formula:

$$SE = \sqrt{\frac{\sum(y-\hat{y})^2}{n-2}}$$

Where,

n is number of pairs of observations.

y is the actual or given value of dependent variable,

$\hat{y}$ is the estimated values obtained from the fitted equation that corresponds to each value of independent variable.

OR Short cut formula for calculating standard error of estimates

$$SE = \sqrt{\frac{\sum y^2 - a\sum y - b\sum xy}{n-2}}$$

Where,

n is number of pairs of observations.

x is the values of independent variable.

y is the actual or given value of dependent variable.

a is the y- intercept of the regression equation of y on x.

b is the slope of the regression equation of y on x.

As the value of standard error is large it indicates more scatterness around the regression line, inversely as the value of standard error is zero then it indicate no scatterness or dispersion. Also this standard error is used to obtain confidence limits for the estimates, which is given as under;

$\hat{y} \pm Z_\alpha\, SE$ in the case of sample when $n \geq 30$

$\hat{y} \pm t_\alpha\, SE$ in the case of sample when $n < 30$

6. Examples with Solution

Illustration: Information regarding the amount spends by a multinational company in advertisement and the profit for the company is given below:

Amount spent on advertisement (in crores) (x)	5	11	4	5	3	2
Profit (In crores) (y)	31	40	30	34	25	20

The management of company believes that the profit depends up on the amount spend on advertisement.

1. The company plan to spend rupees 9 crores on advertisement then estimate the profit for the next year.
2. Also interpret the values of $y-$ intercept and slope of the regression line.
3. Determine the sum of error in estimation and error sum of squares.
4. Estimates the standard error of estimates.
5. Also determine the confidence limits for the estimates. $t_\alpha = 2.776$

Solution:

Since the managements believe that the profit (y) depends on the amount spend on advertisement (x) hence the regression equation of y on x is to be obtained.

x	y	x^2	xy
5	31	25	155
11	40	121	440
4	30	16	120

5	34	25	170
3	25	9	75
2	20	4	40
$\sum x = 30$	$\sum y = 180$	$\sum x^2 = 1000$	$\sum xy = 200$

Here

$$\bar{x} = \frac{\sum x}{n} = \frac{30}{6} = 5 \qquad and \qquad \bar{y} = \frac{\sum y}{n} = \frac{180}{6} = 30$$

Also, the regression coefficient of y on $x$is

$$\begin{aligned} b_{yx} &= \frac{n\sum xy - (\sum x)(\sum x)}{n\sum x^2 - (\sum x)^2} \\ &= \frac{6(1000) - (30)(180)}{6(200) - (30)^2} \\ &= \frac{600}{300} \\ &= 2 \end{aligned}$$

Hence the regression equation of profit (y) on amount spend $(x)$is

$$\begin{aligned} & y - 30 = 2(x - 5) \\ & \therefore y = 2x + 20 \end{aligned}$$

1. Next year when rupees 9 crores is spend on advertisements, the estimated value of profit is $\hat{y} = 2(9) + 20 = 38$ corers.

2. The value of $y-$ intercept is 20 which indicate if the company will not spend any amount on advertisement then also there is a profit of rupees 20 crores. Also the slope is 2 which indicates the rate of change or probable change in the value of profit y for a unit change in the amount spend on advertisement x.

3. Now the estimated values of dependent variable y can be obtained as under:

x	y	$\hat{y} = 2x + 20$	$(y - \hat{y})$	$(y - \hat{y})^2$
5	31	30	1	1
11	40	42	- 2	4
4	30	28	2	4
5	34	30	4	16
3	25	26	- 1	1
2	20	24	- 4	16
$\sum x = 30$	$\sum y = 180$	$\sum \hat{y} = 180$	$\sum(y - \hat{y}) = 0$	$\sum(y - \hat{y})^2 = 42$

From the table, the error sum is zero. $i.e. \sum(y - \hat{y}) = 0$

Sum of squares of error is 42 $\sum(y - \hat{y})^2 = 42$

The estimated value of standard error of estimates is as under

$$SE = \sqrt{\frac{\sum(y - \hat{y})^2}{n - 2}}$$

$$SE = \sqrt{\frac{42}{4}}$$

$$SE = 3.24$$

The confidence limits for the estimates is

$$\hat{y} \pm t_{\alpha} SE = 38 \pm 2.776(3.24)$$

$$= 38 \pm 8.99$$

$$= [29.01,\ 46.99]$$

7. Summary

- A mathematical or functional relationship between two variables is called regression.
- The least square principle is “draw a line which minimize the distance between the points and a line”
- A line obtained by applying least square principle is called regression line or a line of average relationship or best fitted line.
- A numerical measure which shows probable change in the values of dependent variable for a unit change in the value of independent variable is called regression coefficient.
- In practice a regression line of dependent variable on independent variable is obtained.
- In the case of perfect correlation two regression lines are co-inside and vice-versa.
- In the case of no linear correlation two regression lines are perpendicular to each other and vice-versa.
- Two regression lines intersect each other at the point$(\bar{x}, \bar{y})$
- The standard error of the estimates measures the variability or scatter of the observed values around the regression line.