Business Mathematics & Statistics

NCWEB Hansraj Centre

E-Class on 14/03/2020

Topics discussed:

Concept of Linear Regression

Method of Least Squares

Relationship between Correlation and Regression Coefficients

Difference Between Correlation And Regression

By Ms.Komal Chhikara
Bcom Sem II
Section A

THE CONCEPT OF REGRESSION

Correlation tells whether exists a relationship between two variable or not but it does not reflect cause and effect relationship between two variables. Therefore, we cannot predict the value of one variable for a given value for other variable. This limitation is removed by regression analysis.

In regression analysis, the relationship between variable are expressed in the form of a mathematical equation. It is assumed that one variable is cause and the other is the effect.

In regression analysis we have two types of variables: i) dependent (or explained) variable, and ii) independent (or explanatory) variable.

As the name (explained and explanatory) suggests the dependent variable is explained by the independent variable.

For example, it can be postulated that as household income increases, expenditure also increases. Here, consumption expenditure is the dependent variable and household income is the independent variable.

Usually we denote the dependent variable as Y and the independent variable as X.

The statistical method which is used for prediction is called regression analysis. And, when the relationship between the variables is linear, the technique is called simple linear regression . Hence, the technique of regression goes one step further from correlation and is about relationships that have been true in the past as a guide to what may happen in the future.

The objective of simple linear regression is to represent the relationship between two variables with a model of the form shown below:

$$Y = \beta_0 + \beta_1 X + e_i$$

wherein

Y	=	value of the dependent variable,
β_0	=	Y-intercept,
β_1	=	slope of the regression line,
X	=	value of the independent variable,
e_i	=	error term (i.e., the difference between the actual Y value and the value of Y predicted by the model.
i	=	represents the observation number, ranges from 1 to n. Thus Y_3 is the third observation of the dependent variable and X_6 is the sixth observation of the independent variable.

If we consider the two variables (X variable and Y variable), we shall have two regression lines. They are:

i) Regression of Y on X

ii) Regression of X on Y.

The first regression line (Y on X) estimates value of Y for given value of X. The second regression line (X on Y) estimates the value of X for given value of Y. These two regression lines will coincide, if correlation between the variable is either perfect positive or perfect negative.

The alternative simplified expression for the above is:

$$bxy = r\frac{\sigma_x}{\sigma_y} = \frac{\Sigma XY - \frac{(\Sigma X)(\Sigma Y)}{N}}{\Sigma Y^2 - \frac{(\Sigma Y)^2}{N}}$$

Question. From the following 12 months sample data of a company, estimate the regression lines.

(Rs. in lakh)

Advertisement Expenditure:	0.8	1.0	1.6	2.0	2.2	2.6	3.0	3.0	4.0	4.0	4.0	4.6
Sales:	22	28	22	26	34	18	30	38	30	40	50	46

Solution

Advertising (X)	Sales (Y)	X^2	Y^2	XY
0.8	22	0.64	484	17.6
1.0	28	1.00	784	28.0
1.6	22	2.56	484	35.2
2.0	26	4.00	676	52.0
2.2	34	4.84	1156	74.8
2.6	18	6.76	324	46.8
3.0	30	9.00	900	90.0
3.0	38	9.00	1,444	114.0
4.0	30	16.00	900	120.0
4.0	40	16.00	1600	160.0
4.0	50	16.00	2,500	200.0
4.6	46	21.16	2,116	211.6
ΣX=32.8	**ΣY=384**	$\Sigma X^2 = 106.96$	ΣY^2=13368	ΣXY-1150.0

i) We know the regression equation of Y on X is:

$$\hat{Y} - \bar{Y} = byx\,(X - \bar{X})$$

$$\bar{Y} = \frac{384}{12} = 32;\ \bar{X} = \frac{32.8}{12} = 2.733$$

$$byx = \frac{\Sigma XY - \frac{(\Sigma X)(\Sigma Y)}{N}}{\Sigma X^2 - \frac{(\Sigma X)^2}{N}}$$

$$= \frac{1{,}150 - \frac{(32.8)(384)}{12}}{106.96 - \frac{(32.8)^2}{12}} = 100.4/17.31 = 5.8$$

Now Y on X equation is $\hat{Y} - \bar{Y} = byx\,(\hat{X} - \bar{X})$

$\hat{Y} - 32 = 5.8\,(X - 2.733)$

$\hat{Y} = 5.8\,X - 15.85 + 32 = 5.8\,X + 16.15$

Or $\hat{Y} = 16.15 + 5.8X$

ii) We know the regression equation of X on Y is

$$\hat{X} - \bar{X} = bxy\,(Y - \bar{Y})$$

$$bxy = r\frac{\sigma_x}{\sigma_y} = \frac{\Sigma XY - \frac{(\Sigma X)(\Sigma Y)}{N}}{\Sigma Y^2 - \frac{(\Sigma Y)^2}{N}} = \frac{1{,}150 - \frac{(32.8)(384)}{12}}{13368 - \frac{(328)^2}{12}} = \frac{100.4}{1{,}080} = 0.093$$

Now X on Y equation is :

$$\hat{X} - 2.733 = 0.093\,(Y - 32)$$

$$\hat{X} - 2.733 = 0.093Y - 2.976$$

$$\hat{X} = 2.733 - 2.976 - 0.093Y$$

$$\hat{X} = -0.243 + 0.093Y$$

We have the values of $\bar{X} = 2.733$ and $\bar{Y} = 32$

Method of Least Squares

The basic equation of least square method that y on x equation is:

$$\hat{Y} = a + bx \text{ and x on y equation is } \hat{X} = a + by.$$

We can obtain the values of the coefficient a and b of the least square regression line through the following equations:

$$\sum Y = Na + b\sum x \ldots\ldots\ldots \text{(i)}$$

$$\sum XY = a\sum X + b\sum X^2 \ldots\ldots\ldots \text{(ii)}$$

Illustration. Assume that quantity of agricultural production depends on the amount of rainfall and fit a linear regression to the data given.

Rainfall (in mm)	:	60	62	65	71	73	75	81	85	88	90
Agricultural Production (in tones)	:	33	37	38	42	42	45	49	52	55	57

In this case dependent variable (Y) is quantity of agricultural production and independent variable (X) is amount of rainfall.
The regression equation to be fitted is

$$Y_i = a + bX_i$$

For the above equation we find out the normal equations by the method of least squares.

X	Y	X^2	XY	$\hat{Y}$	$Y - \hat{Y}$
60	33	3600	1980	33.85	-0.85
62	37	3844	2294	35.34	1.66
65	38	4225	2470	37.57	0.43
71	42	5041	2982	42.03	-0.03
73	42	5329	3066	43.51	-1.51
75	45	5625	3375	45.00	0.00
81	49	6561	3669	49.46	-0.46
85	52	7225	4420	52.43	-0.43
88	55	7744	4840	54.66	0.34
90	57	8100	5130	56.15	0.85
$\Sigma X = 750$	$\Sigma Y = 450$	$\Sigma X^2 = 57294$	$\Sigma XY = 34526$	$\Sigma \hat{Y} = 450$	$\Sigma e_i = 0$

Now we will solve the following equation

$$\sum Y = Na + b\sum x \ldots\ldots\ldots \text{(i)}$$

$$\sum XY = a\sum X + b\sum X^2 \ldots\ldots\ldots \text{(ii)}$$

By substituting values from the above table in the above normal equation (i) and (ii), we will get following

$$450 = 10a + 750b \ldots\ldots\ldots\ldots \text{(iii)}$$

$$34,526 = 750a + 57,294b \ldots\ldots\ldots \text{(iv)}$$

Before substituting the above values of the two equations (iii & iv) we have to adjust the value connected with either a or b coefficient as equal by the value of suitable multiplier.

Here, if we multiply the equation (iii) with the value 75 we may equalize the value connected with coefficient a, we will get:

$$450 = 10a + 750b \times 75$$

$$33,750 = 750a + 56,250b \text{ (adjusted of iii)}$$

$$(-)\ 34,526 = 750a + 57,294b \text{ (as (iv)}$$

$$-\ 776 = .\ -\ 1,044\ b$$

$$\text{Now, } b = \frac{-776}{-1.044} = 0.743$$

We will find the value of coefficient a by considering the equation (iii) above

i.e.

$450 = 10a + 750\ (0.743)$

$450 = 10a + 557.25$

$-10a = 557.25 - 450$

$a = 107.25/-10 = -10.73$

So the regression line is $\hat{Y} = -10.73 + 0.743\ X$.

Coefficient b is called the regression coefficient. This coefficient reflects the amount of increase in Y when there is a unit increase in X. In regression equation the coefficient b = 0.743 implies that if rainfall increase by 1 mm. agricultural production will increase 0.743 thousand tonne.

RELATIONSHIP BETWEEN CORRELATION AND REGRESSION COEFFICIENTS

The following points about the regression should be noted:

1) The geometric mean of the two regression coefficients (byx and bxy) gives coefficient of correlation. That is, $r = \pm\sqrt{(bxy)\ (byx)}$

2) Both the regression coefficients will always have the same sign (+ or –).

3) Coefficient of correlation will have the same sign as that of regression coefficients. If both are positive, then r is positive. In case both are negative, r is also negative. For example, bxy = –1.3 and byx = –0.65,

then r is:

$\pm\sqrt{-1.3x\ -\ 0.65}$ = – 0.919 but not + 0.919

4) Regression coefficients are independent of change of origin, but not of scale.

DIFFERENCE BETWEEN CORRELATION AND REGRESSION

1) Correlation coefficient 'r' between two variables (X and Y) is a measure of the direction and degree of the linear relationship between them, which is mutual. Whereas regression analysis aims at establishing the functional relationship between the two variables under study, and then using this relationship to predict the value of the dependent variable for any given value of the independent variable. It also reflects upon the nature of the variables (i.e., which is the dependent variable and which is independent variable).

2) Correlation need not imply cause and effect relationship between the variables under study. But regression analysis clearly indicates the cause and effect relationship between the variables. The variable corresponding to cause is taken as independent variable and the variable corresponding to effect is taken as dependent variable.

3) Correlation coefficient ‘r’ is a relative measure of the linear relationship between X and Y variables and is independent of the units of measurement. It is a number lying between ±1. Whereas the regression coefficient byx (or bxy) is an absolute measure representing the change in the value of the variable Y (or X) for a unit change in the value of the variable X (or Y).

4) There may be spurious (non-sense) correlation between two variables which is due to pure chance and has no practical relevance. For example, the correlation between the size of shoe and the income of a group of individuals. There is no such thing as spurious regression.

5) Correlation analysis is confined only to study of linear relationship between the variables and, therefore, has limited applications. Whereas regression analysis has much wider applications as it studies linear as well as non-linear relationships between the variables.

Thank You