Regression equations

introduction

In the last topic, we learned how measure the association or relationship between two variables using correlation coefficient. After establishing the relationship between two variables, the next logical question would be by knowing one variable (X or Y), can I estimate the other variable. This process of finding one variable by knowing the other variable is called estimation/prediction/forecasting. Regression is one such statistical technique used for estimation/prediction/forecasting.

Prediction or estimation is one of the major problems in almost all spheres of human activity. The estimation of future production, consumption, prices, sales, profits, investment etc are paramount importance to a business organization. Population estimates and population projections are indispensible for efficient planning of an economy. Regression analysis is one of the scientific techniques for making such predictions. ‘regression analysis is a mathematical measure of the average relationship between two or more variables in terms of the original units of the data’.

In regression analysis there are two types of variables. The variable whose value is influenced or is to be estimated is called **dependent variable** and the other variable which influences the values or is used for estimation is called **independent variable**. In regression analysis independent variable is also known as regressor or predictor or explanator while the dependent variable is also known as regressed or explained variable.

Regression lines

Line of regression is the line which gives the best estimate of one variable for any given value of the other variable. In case of two variables x and y, we will have two lines of regression: one of y on x and the other of x on y.

Line of regression of **y on x** is the line which gives the best estimate for the value of y for any specified value of x.

Line of regression of **x on y** is the line which gives the best estimate for the value of x for any specified value of y.

The term best fit in interpreted in accordance with the principle of least squares which consists in minimizing the sum of the squares of the errors of estimates i.e. the deviations between the given observed values of the variable and their corresponding estimated values as given by the line of best fit.

The lines of regression are

Y on x represented as y = a + b x

a is called intercept and b is called slope of the line

X on y represented as x = c + d y

c is called intercept and d is called slope of the line

Normal equations

Regression line of y on x is $\mathbf{y = a + b x}$

We need two equations to solve two unknowns **a** and **b**. the two equations are called normal equations. They are

$$\Sigma y = n\,\mathbf{a} + \mathbf{b}\,\Sigma x$$

$$\Sigma xy = \mathbf{a}\,\Sigma x + \mathbf{b}\,\Sigma x^2$$

from the given data we can find Σy, Σx, Σxy and Σx^2 and substitute in the equations to solve for unknowns **a** and **b**.

Regression line of x on y is $\mathbf{x = c + d\ y}$

We need two equations to solve two unknowns **c** and **d**. the two equations are called normal equations. They are

$$\Sigma x = n\,\mathbf{c} + \mathbf{d}\,\Sigma y$$

$$\Sigma xy = \mathbf{c}\,\Sigma y + \mathbf{d}\,\Sigma y^2$$

from the given data we can find Σy, Σx, Σxy and Σy^2 and substitute in the equations to solve for unknowns **c** and **d**.

example

find two regression lines for the following data

x	y	x^2	y^2	xy
2	7	4	49	14
3	9	9	81	27
4	10	16	100	40
5	14	25	196	70
6	15	36	225	90
$\Sigma x = 20$	$\Sigma y = 55$	$\Sigma x^2 = 90$	$\Sigma y^2 = 651$	$\Sigma xy = 241$

$\Sigma y = n\,\mathbf{a} + \mathbf{b}\,\Sigma x$

$\Sigma xy = \mathbf{a}\,\Sigma x + \mathbf{b}\,\Sigma x^2$

$55 = 5a + 20b$

$241 = 20a + 90b$

Multiplying equation (1) by 4 and solving we get

$220 = 20a + 80b$

$241 = 20a + 90b$

$-21 = -10b \Rightarrow \mathbf{b = 2.1}$

substituting b in equation (1) , we get

55 = 5a + (20 x 2.1)

55 = 5a + 42 $\Rightarrow$ **5a = 13 $\Rightarrow$ a = 13/5 = 2.6**

therefore the regression line of y on x is

y = 2.6 + 2.1 x

Alternate approach

The formula for calculation slope 'b' is

$$b = \frac{\Sigma xy - n\bar{x}\,\bar{y}}{\Sigma x^2 - n\,\bar{x}^2} = \frac{241 - 5 \times 4 \times 11}{90 - 5\,(4)^2} = 2.1$$

Formula for calculation intercept 'a' is **$a = \bar{y} - b\,\bar{x} = 11 - (2.1 \times 4) = 2.6$**

Regression line of x on y

$\Sigma x = n\,\mathbf{c} + \mathbf{d}\,\Sigma y$

$\Sigma xy = \mathbf{c}\,\Sigma y + \mathbf{d}\,\Sigma y^2$

$20 = 5c + 55\,d$

$241 = 55\,c + 651\,d$

Multiplying equation (1) by 11 and solving we get

$220 = 55\,c + 605\,d$

$241 = 55\,c + 651\,d$

$-21 = -46\,d \Rightarrow \mathbf{d = 0.457}$

substituting ‘d’ in equation (1) , we get

20 = 5c + (55 x 0.457)

20 = 5c + 25.135 ⇒ **5c = -5.135 ⇒ c = -5.135/5 = -1.027**

therefore the regression line of x on y is

x = -1.027 + 0.457 y

Alternate approach

The formula for calculation slope ‘d’ is

$$d = \frac{\Sigma xy - n\bar{x}\bar{y}}{\Sigma y^2 - n\bar{y}^2} = \frac{241 - 5 \times 4 \times 11}{651 - 5(11)^2} = 0.457$$

Formula for calculation intercept ‘c’ is **$c = \bar{x} - d\bar{y} = 4 - (0.457 \times 11) = -1.027$**

Properties of regression coefficients

- The coefficient of correlation is the geometric mean of two regression coefficients. Symbolically, **r =√ (b x d)**
- As the coefficient of correlation cannot exceed 1, in case one of the regression coefficient is **greater than 1**, then the other must be **less than 1.**
- Both the regression coefficients will have the same sign, either positive (+) or negative (-). If one regression coefficient is +, then the other will also be +
- The correlation coefficient and the regression coefficients will have the **same sign**. If the regression coefficients are +, the correlation coefficient will also be + and vice versa.