ST. LAWRENCE HIGH SCHOOL

A JESUIT CHRISTIAN MINORITY INSTITUTION

FOR GOD AND COUNTRY

STUDY MATERIAL-3
SUBJECT – STATISTICS

Pre-test

Chapter: BIVARIATE ANALYSIS **Class: XII**

Topic:Regression **Date: 12.05.2020**

REGRESSION

PART 1

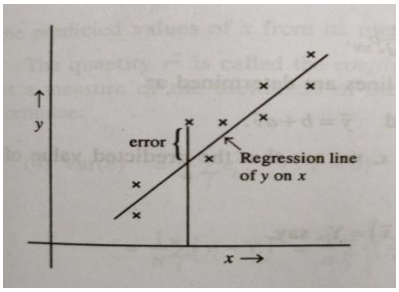
REGRESSION

Regression is the quantitative or exact mathematical relation between two variables.

In a scatter diagram let the plotted points be $(x_i , y_i) \; for \; i = 1(1)n.$ and the assumed best fitted line is Y = a + b X ……..(***)

From every point we draw lines parallel to Y-axis which intersect the assumed best fitted line at (x_i , Y_i).

Where $Y_i = a + b \; x_i$ (since the line is parallel to X-axis the abscissa remains same.)



$e_i = y_i - \; Y_i$

Total sum of squares E= $\sum e_i^2 = \; (y_i - \; Y_i \;)^2 = (y_i - \; a - bx_i)^2$

To minimize partially w.r.t. a and b

Differetiating partially w.r.t a and equating to 0,

$$\frac{\delta}{\delta a} \; ((y_i - \; a - bx_i)^2 = 0 \; \Longrightarrow 2 \sum (y_i - \; a - bx_i) \; (0 - 1 - 0) = 0$$

$$\Longrightarrow \; \sum y_i = na + b \; \sum x_i \; \ldots \ldots \ldots \ldots (*)$$

$\frac{\delta^2 E}{\delta a^2} = 2 \sum (\; 0 \; - \; 1 \; - 0)$ = 2n > 0

Differetiating partially w.r.t b and equating to 0,

$$\frac{\delta}{\delta b}\left((y_i - a - bx_i)^2 = 0 \implies 2\sum(y_i - a - bx_i)(0 - 0 - x_i) = 0\right.$$

$$\implies \sum x_i y_i = a \sum x_i + b \sum x_i^2 \text{ .....(**)}$$

$$\frac{\delta^2 E}{\delta a^2} = 2\sum(0 - 0 - x_i) = 2\sum x_i^2 > 0$$

(*)and (**) are known as normal equations.

Since the second order derivative of the normal equations are greater than zero, for common values of a and b which satisfies both the normal equations, the sum of squares of error becomes minimum.

To find the value of b from the nomal equation nX(**) - $(\sum x_i)$ (*)

$$\implies n\sum x_i y_i - n(\sum x_i)(\sum y_i) = nb \sum x_i^2 - b(\sum x_i)^2$$

By dividing both sides by n^2, we get, $\frac{1}{n}\sum x_i y_i - \bar{x}\,\bar{y} = b\left(\frac{1}{n}\sum x_i^2 - \bar{x}^2\right)$

$$\implies \text{cov}(x, y) = b\ s_x^2$$

$$\implies b = \frac{cov(x, y)}{s_x^2} = b_{yx}\ (say)$$

Now substituting the value of b in (***) we get,

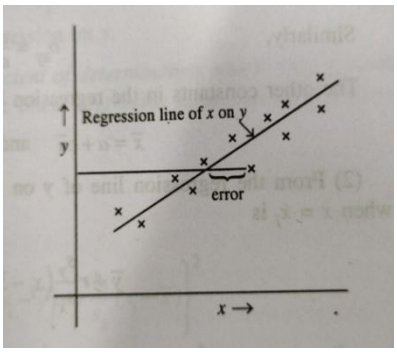
$$Y = a + b_{yx}x \ldots\ldots(1)$$

and from the property of arithmetic mean $\bar{y} = a + b_{yx}\bar{x} \ldots\ldots(2)$

So $a = \bar{y} - b_{yx}\bar{x}$

Either by substituting a and b or by subtracting (1) from (2),

$Y - \bar{Y} = b_{yx}(x - \bar{x})$ which is known as regression equation y on x. and b_{yx} is known as regression coefficient of y on x. . It gives the increment of y for unit increase of x.



Now instead of measuring errors parallel to Y-axis if we do the same parallel to X-axis and assume the best fitted line as X= a + bY and follow the same procedure then we get regression equation x on y which is $X - \bar{x} = b_{xy}(y - \bar{y})$.

Where $b_{xy} = \frac{cov\ (x\ ,y\ )}{s_y^2}$ is known as regression coefficient x on y. It gives the increment of x for unit increase of y.

The literal meaning of regress is deviation .

So the regression line y on x can be defined as :

The linear equation which gives the regressed or predicted value of the variable y for the same value value of x_i.

Similarly we can also define regression line x on y.

The linear equation which gives the regressed or predicted value of the variable x for the same value value of y_i.

Properties of regression lines and regression coefficients:

- Change of base or origin and scale
  If $u_i = a + bx_i$ and $v_i = c + dx_i$ for all i=1(1)n
  Then $r_{uv} = \frac{bd}{|b|\,|d|}\, r_{xy}$.

  Now $b_{vu} = r_{uv}.\frac{s_v}{s_u} = \frac{bd}{|b|\,|d|}\, r_{xy}.\frac{|d|s_y}{|b|s_x} = \frac{bd}{b^2}\, b_{yx}$
  The regression line v on u,
  $v - \bar{v} = b_{vu}(u - \bar{u}) \Longrightarrow d(Y - \bar{y}) = \frac{bd}{b^2}\, b_{yx}.b.(x - \bar{x})$

$\Longrightarrow Y - \bar{y} = b_{yx}(x - \bar{x})$ which is same as regression line y on x.

And regression line u on v becomes same as regression line x on y.

Hence the regression equation is independent of change of origin and scale.

- The regression lines intersect at $(\bar{x}, \bar{y})$. Since $x = \bar{x}$ and $y = \bar{y}$ satisfies euations of both the regression lines.

  $b_{yx} = \frac{cov(x,y)}{s_x^2} = \frac{cov(x,y)}{s_x s_y}\frac{s_y}{s_y}$ (by multiplying the numerator and the denominator by s_y)

  $\Longrightarrow b_{yx} = r_{xy}\frac{s_y}{s_x}$

  And similarly $b_{xy} = r_{xy}\frac{s_x}{s_y}$

- b_{yx}, b_{xy} and r_{xy} should be of same sign. (Since the sign of both the regression coefficients is same that of the correlation coefficient as standard deviation is always positive.)

Prepared by

Sanjay Bhattacharya