Module 2: Regression

Dr. A. N. Basugade
M.Sc. Ph.D
Head, Department of Statistics,
GopalKrishnaGokhaleCollege,Kolhapur
Email – arunb1961@gmail.com

Module 2 Regression

Overview

- Introduction
- Lines of Regression
- Equations of lines of Regression
- Derivation of Equation of line of regression of y on x using least square method.
- Properties of Regression coefficients.

Introduction:

When two variables are highly correlated then we can’t estimate value of one variable knowing the value of other variable. This can be done by analysis of regression.

A method of estimating the value of one variable knowing the value of other variable is known as **Regression.**

The term regression was first used by Galton. He has studied the relation between heights of fathers and heights of sons and found that sons of tall fathers have little less height than their fathers & sons of short fathers have little more height than their fathers. This shows that the average heights of sons of tall & short fathers will tend to the general average height. This process is termed as **regression by Galton**.

Lines of Regression:

As we know, if two variables are highly correlated then the plotted points in the scatter diagram lies on a narrow strip. We can draw a line passing through these points such that

i) The line will pass through maximum number of points

ii) Remaining points are very close to the line from both the sides.

iii) The sum of distances of the points from the line will be minimum.

Such a line is called as line of best fit or line of Regression.

- **Types of lines of Regression:**

Since we can minimize the distances of the points from the line in two ways; one along with the x-axis & other along with the y-axis hence there are two regression lines i) regression line of y on x & ii) regression line of x on y.

i) Regression line of y on x: If we minimize the distances of the points from the line along with the y-axis then we get a line of

Regression of y on x & its equation is y = a + bx.

ii) Regression line of x on y: If we minimize the distances of the points from the line along with the x-axis then we get a line of

Regression of x on y & its equation is x = a + by.

The values of a and b are obtained by using least square method.

After estimating the values of a 7 b we get the regression equations

as

i) regression equation of y on x

$$(y-\bar{y})=b_{yx}(x-\bar{x})$$

ii) regression equation of y on x

$$(x-\bar{x})=b_{xy}(y-\bar{y})$$

Where, b_{yx} & b_{xy} are regression coefficients of y on x & x on y respectively and $b_{yx}=r\frac{\sigma_y}{\sigma_x}$ & $b_{xy}=r\frac{\sigma_x}{\sigma_y}$

Note:

1) To estimate value of y for given value of x, we use the regression equation of y on x.

2) To estimate value of x for given value of y, we use the regression equation of x on y.

3) Correlation coefficient is geometric mean of the regression coefficients.

i. e. $\boldsymbol{r} = \sqrt{\boldsymbol{b}_{yx}\ \boldsymbol{b}_{xy}}$

The Method of Least Square

Let y = $a + bx$ be the equation of the line required. To find the line of regression of y on x we minimize the sum of the absolute distances of the points like $P\ (x_i,\ y_i)$ from the line measured along the y-axis. If Q is the point on the line corresponding to $P\ (X_i,\ y_i)$ we have to minimize the absolute distance PQ, Since Q lies on y = $a + bx$, its y-coordinate is $a + bx_i$

$$|PQ| = |y_i - a - bx_i|$$

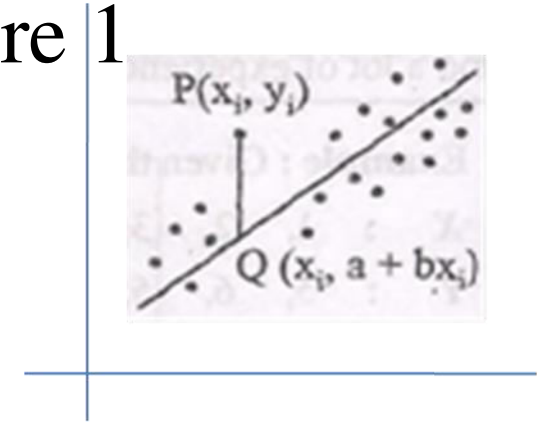
For minimizing $|PQ|$ we minimize its squares. Hence, if S denotes the sum of the squares of these distances then

$$S = \sum f_i\ (y_i - a - bx_i)^2$$

We have to find a and b such that S is minimum, the conditions for which are $\frac{ds}{da} = 0 \quad and \quad \frac{ds}{db} = 0$

i. e. $\sum f_i\ (y_i - a - bx_i) = 0$---(1) and $\sum f_i\ x_i\ (y_i - a - bx_i) = 0$--- (2)

Figure 1



From (1)we get $\sum f_i y_i - a \sum f_i - b \sum f_i x_i = 0$

Thus $N\bar{y} - aN - bN\bar{x} = 0$ i.e. $\bar{y} = a + b\bar{x}$ ----(3)

Which shows that the reg. line will passes through $(\bar{x}, \bar{y})$

From (2) we get $\sum f_i x_i y_i - a \sum f_i x_i - b \sum f_i x_i^{2}) = 0$ --- (4)

As we know that $\sum f_i x_i y_i = Nr\sigma_x\sigma_y + N\bar{x}\bar{y}$ and $\sum f_i x_i^2 = N\sigma_x^2 + N\bar{x}^2$

From (4) we have $Nr\sigma_x \sigma_y + N\bar{x}\bar{y} = aN\bar{x} + bN\sigma_x^2 + bN\bar{x}^2$

Thus $r\sigma_x \sigma_y + \bar{x}\bar{y} = a\bar{x} + b\sigma_x^2 + b\bar{x}^2$ -----(5)

Multiply equation (3) $\bar{x}$ by and subtracting it from (5) we get

$$b = r\frac{\sigma_y}{\sigma_x}$$

Since the line will pass through $(\bar{x}, \bar{y})$ and its slope is $b = r\frac{\sigma_y}{\sigma_x}$

Its equation is $$(y-\bar{y}) = r\frac{\sigma_y}{\sigma_x}(x-\bar{x})$$

Similarly the equation of line of regression of x on y is

$$(x-\bar{x}) = r\frac{\sigma_x}{\sigma_y}(y-\bar{y})$$

➢ Properties of Regression Coefficients:

1) Correlation coefficient is geometric mean of the regression coefficients.

Proof: We know that $b_{yx} = r\frac{\sigma_y}{\sigma_x}$ and $b_{xy} = r\frac{\sigma_x}{\sigma_y}$

Thus $b_{yx}\, b_{xy} = r\frac{\sigma_y}{\varpi_x}\; r\frac{\sigma_x}{\sigma_y}$ $= r^2$ Hence the result

2) A.M. of regression coefficients is greater than or equal to correlation coefficient. i.e. $1/2(\; b_{yx} + b_{xy} \;) \geq \; r$

Substituting the values of coefficients in above equation we get

$$1/2(r\frac{\sigma_y}{\varpi_x} + r\frac{\sigma_x}{\sigma_y}) \geq \; r$$

$$\frac{\sigma_y}{\varpi_x} + \frac{\sigma_x}{\sigma_y}) \geq \; 2$$

Thus

$$\sigma_x^2 - 2\sigma_x\sigma_y + \sigma_y^2 \geq 0$$

$$(\sigma_x - \sigma_y)^2 \geq 0$$

Which is trivially true hence the property .

3) One coefficient is greater than one then other must be less than one.

Proof: Since r lies between -1 to 1 hence $r^2 \leq 1$

b_{xy} $b_{yx} \leq 1$ i.e. $b_{yx} \leq 1/ b_{xy}$ hence the property.

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

Proof: Let change the origin of x & y by an amount a & b, and the scales by an amount h & k respectively therefore

u =(x-a)/h & v = (y-b)/k We know $r_{uv} = r_{xy}$

and $\sigma_u = \sigma_x /h$ and $\sigma_v = \sigma_y /k$

$b_{uv} = r_{uv}\, \sigma_u / \sigma_v = r_{xy}\, (\sigma_x /h)/ \sigma_y /k = (k/h)\, b_{xy}$

Similarly $b_{uv} = (h/k)\, b_{yx}$

Summary:

At the end of this module student must be able to

- **Define regression**
- **Explain Lines of Regression**
- **Derive Equation of line of regression of y on x using least square method.**
- **Prove some Properties of Regression coefficients.**