Regression

It is the statistical tool to measure the average relationship between two or more variables in terms of others.

In regression analysis one of the two variables, say x is regarded as an independent variable, the other variable y as the dependent variable or vice versa. We are interested in the dependence of y on x or dependence of x on y.

For Example: (i) dependence of blood pressure (y) on the age (x)

(ii) dependence of weight of a person (y) on the age (x).

In general, we specify $x_1, x_2, \cdots, x_n$ and then observe the corresponding values $y_1, y_2, \cdots y_n$, so that we get a bivariate sample (x_i, y_i); $i = 1, 2, 3, \cdots n$. We are interested in finding a mathematical measure of the average relationship between the two (or more in case of multivariate data) variables in terms of the original units of the data.

Bivariate Regression

Let X and Y be two random variables, which are jointly distributed. Then the regression equation of Y on X is defined as

$$\mu_{Y|X} = E(Y|X=x)$$

$$\Rightarrow \mu_{Y|X} = \begin{cases} \sum_y \frac{y\,P(x,y)}{P_1(x)} & \text{for discrete case} \\ \int_{-\infty}^{\infty} \frac{y\,f(x,y)}{f_1(x)}\,dy, & \text{for cont'} \text{ case.} \end{cases}$$

where $P_1(x)$ & $f_1(x)$ are marginal ~~P~~pmf and pdf of random variable X respectively.

Similarly, the regression equation of X on Y is defined as

$$\mu_{X|Y} = E(X|Y=y)$$

$$\Rightarrow \mu_{X|Y} = \begin{cases} \sum_x \frac{x\,P(x,y)}{P_2(y)} & \text{for discrete case} \\ \int_{-\infty}^{\infty} \frac{x\,f(x,y)}{f_2(y)}\,dx, & \text{for cont'} \text{ case.} \end{cases}$$

where $P_2(y)$ is pmf of discrete r.v. X and $f_2(y)$ is pdf of continuous r.v. X.

Thm Let μ_1, μ_2 are the mean and σ_1^2, σ_2^2 are the variances of X and Y respectively and ρ is the correlation coefficient between X and Y. If the regression of X or Y is linear, then Prove that

$$\mu_{X|Y} = \mu_1 + \rho \frac{\sigma_1}{\sigma_2}(y - \mu_2)$$

Proof: We shall Prove it for Contⁿ case. Similarly we can Prove it for discrete case.

let regression eqⁿ of X on Y is linear, then

$$\mu_{X|Y} = a + by$$

$$\Rightarrow \int_{-\infty}^{\infty} x f(x|y)\, dx = a + by$$

$$\Rightarrow \int_{-\infty}^{\infty} x \frac{f(x,y)}{f_2(y)}\, dx = a + by$$

$$\Rightarrow \int_{-\infty}^{\infty} x f(x,y)\, dx = (a+by) f_2(y) \quad \text{---------- (1)}$$

Integrate eqⁿ (1) w.r.t. to 'y' we have.

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x f(x,y)\, dx\, dy = \int_{-\infty}^{\infty} (a+by) f_2(y)\, dy$$

$$\Rightarrow \int_{-\infty}^{\infty} x \left\{ \int_{-\infty}^{\infty} f(x,y)\, dy \right\} dx = a\int_{-\infty}^{\infty} f_2(y)\, dy + b\int_{-\infty}^{\infty} y f_2(y)\, dy$$

$$\Rightarrow \int_{-\infty}^{\infty} x f_1(x)\,dx = a\cdot 1 + b E(Y) \qquad (4)$$

$$\Rightarrow E(X) = a + b E(Y) \qquad \text{---------- (2)}$$

Now multiply eqn (1) by y, we have.

$$y \int_{-\infty}^{\infty} x f(x,y)\,dx = y(a+by) f_2(y)$$

$$\Rightarrow \int_{-\infty}^{\infty} xy f(x,y)\,dx = (ay + by^2) f_2(y)$$

Now integrate it with respect to y, we have.

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy f(x,y)\,dx\,dy = a\int_{-\infty}^{\infty} y f_2(y)\,dy + b\int_{-\infty}^{\infty} y^2 f_2(y)\,dy$$

$$\Rightarrow E(XY) = aE(Y) + bE(Y^2) \qquad \text{---------- (3)}$$

$\because E(X) = \mu_1$, $E(Y) = \mu_2$.

$V(X) = \sigma_1^2 \Rightarrow E(x^2) - (E(x))^2 = \sigma_1^2$

$\Rightarrow E(x^2) - \mu_1^2 = \sigma_1^2$.

$\Rightarrow E(x^2) = \mu_1^2 + \sigma_1^2$.

Similarly, $E(Y^2) = \mu_2^2 + \sigma_2^2$.

using these eqns in eqn (2) & (3), we have

$$a + b\mu_2 = \mu_1 \qquad \text{---------- (4)}$$

$$a\mu_2 + b(\mu_2^2 + \sigma_2^2) = E(XY) \qquad \text{---------- (5)}$$

Now. $\rho = \frac{Cov(X,Y)}{\sigma_1 \sigma_2} = \frac{E(XY) - E(X)E(Y)}{\sigma_1 \sigma_2}$

$\Rightarrow \rho = \frac{E(XY) - \mu_1 \mu_2}{\sigma_1 \sigma_2}$

$\Rightarrow E(XY) = \mu_1 \mu_2 + \rho \sigma_1 \sigma_2$ ——— (6)

using it in eqn (5), we have.

$a\mu_2 + b(\mu_2^2 + \sigma_2^2) = \mu_1 \mu_2 + \rho \sigma_1 \sigma_2$.

——— (7)

Multiplying eqn (4) by μ_2, we have.

$a\mu_2 + b\mu_2^2 = \mu_1 \mu_2$. ——— (8)

Now eqn (7) − (8) $\Rightarrow$

$b\sigma_2^2 = \rho \sigma_1 \sigma_2 \Rightarrow b = \rho \frac{\sigma_1}{\sigma_2}$ ——— (9)

using it in eqn (4), we have.

$a + \rho \frac{\sigma_1}{\sigma_2} \mu_2 = \mu_1 \Rightarrow a = \mu_1 - \rho \frac{\sigma_1}{\sigma_2} \mu_2$.

——— (10)

using eqn (9) and (10) in eqn (1), we have.

$\mu_{X|Y} = \left(\mu_1 - \rho \frac{\sigma_1}{\sigma_2} \mu_2\right) + \rho \frac{\sigma_1}{\sigma_2} y$

$\Rightarrow \mu_{X|Y} = \mu_1 + \rho \frac{\sigma_1}{\sigma_2} (y - \mu_2)$

Note: If the regression of Y on X is linear, then

$$\mu_{Y|X} = \mu_2 + \rho \frac{\sigma_2}{\sigma_1}(x - \mu_1)$$

Regression lines (Alternative definition)

If we take, $E(X) = \mu_1 = \bar{x}$

& $E(Y) = \mu_2 = \bar{y}$

$V(X) = \sigma_x^2$, $V(Y) = \sigma_y^2$.

then the eqn of the line of regression of y on x is given by

$$y - \bar{y} = \rho \frac{\sigma_y}{\sigma_x}(x - \bar{x}) \quad \left[\text{Just replace } \mu_{Y|X} \text{ by } y\right].$$

Similarly, eqn of line of regression of x on y is given by

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

Q. Let the joint density of (X, Y) is

$$f(x,y) = \begin{cases} 8xy, & 0 < x < y < 1 \\ 0, & \text{otherwise.} \end{cases}$$

(a) find correlation coefficient of X and Y

(b) find both the regression lines.

(7)

Sol^n (a) Marginal pdf's

$$f_1(x) = \int_x^1 8xy\,dy = (4xy^2)_x^* = 4x - 4x^3 \quad \text{for } 0 \le x \le 1$$

$$\text{and } f_2(y) = \int_0^y 8xy\,dx = [4x^2y]_0^y = 4y^3 \quad , 0 \le y \le 1$$

Now, $E(X) = \int_0^1 x(4x-4x^3)\,dx = 8/15.$

$$E(Y) = \int_0^1 y(4y^3)\,dy = 4/5$$

$$E(X^2) = \int_0^1 x^2(4x-4x^3)\,dx = 1/3$$

$$E(Y^2) = \int_0^1 y^2(4y^3)\,dy = 2/3$$

$$\therefore V(X) = E(X^2) - (E(X))^2 = 1/3 - \frac{64}{225} = \frac{11}{225}$$

$$\therefore \sigma_X = \sqrt{V(X)} = \frac{1}{15}\sqrt{11}$$

$$V(Y) = E(Y^2) - (E(Y))^2 = 2/3 - \frac{16}{25} = \frac{2}{75}$$

$$\therefore \sigma_Y = \sqrt{V(Y)} = \sqrt{2/75} = \sqrt{\frac{6}{225}} = \frac{1}{15}\sqrt{6}$$

$$E(XY) = \int_0^1 \int_{xy}^{y} (8xy)\,dx\,dy. = \int_0^1 \int_0^y 8x^2y^2\,dx\,dy.$$

$$= \int_0^1 8y^5/3\,dy = \left(\frac{8y^6}{(3)(6)}\right)_0^1 = \frac{8}{18} = 4/9$$

Now, $\rho = \dfrac{E(XY) - E(X)E(Y)}{\sigma_X \sigma_Y} = \dfrac{4/9 - (8/15)\cdot(4/5)}{\frac{1}{15}\sqrt{11} \cdot \sqrt{\frac{2}{75}}}$

$$= \frac{225\left(\frac{4}{9} - \frac{32}{75}\right)}{\sqrt{66}} = \frac{4}{\sqrt{66}} = 0.49$$

Now Regression line of Y on X is

$$\mu_{Y|X} = \mu_2 + \rho \frac{\sigma_2}{\sigma_1}(x - \mu_1)$$

$$\Rightarrow \mu_{Y|X} = 0.8 + (0.49) \cdot \frac{\frac{1}{15}\sqrt{6}}{\frac{1}{15}\sqrt{11}}(x - 0.53)$$

$$\Rightarrow \mu_{Y|X} = 0.8 + 0.267(x - 0.53)$$

Similarly, Regression line of X on Y is

$$\mu_{X|Y} = \mu_1 + \rho \frac{\sigma_1}{\sigma_2}(Y - \mu_2)$$

$$\Rightarrow \mu_{X|Y} = 0.53 + 0.898(y - 0.8)$$