March 24, 2020

Regression and Correlation (Continued) Dr. SUSHIL KUMAR AZAD

Let a bivariate (joint) distribution of continuous random variables, given by the joint pdf $f(x, y)$.

In the last lecture, we learned that $E(Y/x)$ as a function of x is defined as <u>the Regression of Y on X</u> and $E(X/y)$ as a function of y is defined as <u>the Regression of X on Y</u>.

Let us assume that both $E(Y/x)$ and $E(X/y)$ are linear, i.e.,

$$E(Y/x) = \int_{-\infty}^{\infty} y f_{2/1}(y/x)\, dy = \alpha + \beta x, \quad \text{---}(i)$$

and

$$E(X/y) = \int_{-\infty}^{\infty} x f_{1/2}(x, y)\, dx = \gamma + \delta y. \quad \text{---}(ii)$$

Then, $\int_{-\infty}^{\infty} f_1(x) E(Y/x)\, dx$, ($f_1(x)$ is the marginal pdf of X)

$$= \int_{-\infty}^{\infty} f_1(x)(\alpha + \beta x)\, dx$$

$$= \alpha + \beta E(X), \quad \left(\text{as } \int_{-\infty}^{\infty} f_1(x)\, dx = 1\right)$$

$$\Rightarrow \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y f_1(x) f_{2/1}(y/x)\, dy\, dx$$

$$= \alpha + \beta E(X) \quad (\text{from } (i))$$

$$\Rightarrow \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y f(x, y)\, dy\, dx = \alpha + \beta E(X)$$

$$\Rightarrow \quad E(Y) = \alpha + \beta E(X) \quad \text{---}(iii)$$

Which shows that the regression line of Y on X passes through the mean of X and Y, i.e., $E(X)$ and $E(Y)$.

In a similar manner, using (ii), we get

$$E(X) = \gamma + \delta\, E(Y) \quad \text{---} (iv)$$

that is, the regression line of X on Y also passes through $(E(X), E(Y))$.

Hence, <u>the regression lines intersect at a point $(E(X), E(Y))$.</u>

SUSHIL KUMAR AZAD

Again, multiplying (i) by $x f_1(x)$ and integrating with respect to x, we have

$$\int_{-\infty}^{\infty} x f_1(x)\, E(Y/x)\, dx = \int_{-\infty}^{\infty} x f_1(x)(\alpha + \beta x)\, dx$$

$$\Rightarrow \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x y f_1(x) f_{2/1}(y/x)\, dy\, dx = \alpha \int_{-\infty}^{\infty} x f_1(x)\, dx + \beta \int_{-\infty}^{\infty} x^2 f_1(x)\, dx$$

$$\Rightarrow \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x y f(x, y)\, dy\, dx = \alpha\, E(X) + \beta\, E(X^2)$$

$$\Rightarrow \quad E(XY) = \alpha\, E(X) + \beta\, E(X^2). \quad \text{---} (v)$$

Similarly, multiplying (ii) by $y f_2(y)$ and integrating with respect to y, we have

$$E(XY) = \gamma\, E(Y) + \delta\, E(Y^2) \quad \text{--} (vi)$$

Equations (iii) and (v) are known as <u>normal equations</u> for the regression of Y on X.

Solving for α and β, we get

$$\beta = \frac{E(XY) - E(X)E(Y)}{E(X^2) - [E(X)]^2} = \frac{\text{cov}(X, Y)}{\sigma_X^2}$$

$$= \rho \frac{\sigma_Y}{\sigma_X}, \quad \rho = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} \quad \text{----- (vii)}$$

and $\alpha = E(Y) - \beta E(X)$

$$= \mu_2 - \frac{\rho \sigma_Y}{\sigma_X} \mu_1 \quad \text{--- (viii)}$$

where β is called <u>the regression coefficient of Y on X</u>.

Similarly, equations (iv) and (vi) the normal equations for the regression of X on Y. Solving for γ and δ, we get

$$\delta = \frac{E(XY) - E(X)E(Y)}{E(Y^2) - [E(Y)]^2}$$

$$= \frac{\text{Cov}(X, Y)}{\sigma_Y^2} = \rho \frac{\sigma_X}{\sigma_Y} \quad \text{--- (ix)}$$

and $\gamma = E(X) - \delta E(Y)$

$$= \mu_1 - \rho \frac{\sigma_X}{\sigma_Y} \mu_2 \quad \text{--- (x)}$$

where δ is called <u>the regression coefficient of X and Y</u>.

SUSHIL KUMAR AZAD

From (i), (vii) & (viii), we get

$$E(Y/x) = \left(\mu_2 - \frac{\rho\sigma_Y}{\sigma_X}\mu_1\right) + \frac{\rho\sigma_Y}{\sigma_X}x$$

$$\Rightarrow \boxed{E(Y/x) = \mu_2 + \frac{\rho\sigma_Y}{\sigma_X}(x-\mu_1)},$$ which is the regression line of Y on X

which is the first part of Theorem 1. of last lecture.

Similarly, from (ii), (ix) & (X), we get

$$\boxed{E(X/y) = \mu_1 + \frac{\rho\sigma_X}{\sigma_X}(y-\mu_2)},$$

which is the regression line of X on Y.

From (vii) & (ix), we get

$$\beta = \frac{Cov(X,Y)}{Var(X)} \quad \&$$

$$\delta = \frac{Cov(X,Y)}{Var(Y)}$$

$$\Rightarrow \beta\delta = \frac{[Cov(X,Y)]^2}{Var(X)\,Var(Y)}$$

$$\Rightarrow \sqrt{\beta\delta} = \frac{Cov(X,Y)}{\sqrt{Var(X)\,Var(Y)}} = \frac{Cov(X,Y)}{\sigma_X\sigma_Y} = \rho,$$ the correlation coefficient between X and Y,

$$\Rightarrow \rho = \sqrt{\beta\delta}$$

that is, ρ is the geometric mean of the regression coefficients when both the regression are linear and is called the product moment correlation coefficient between X and Y.

SUSHIL KUMAR AZAD

March 24, 2020
DR. SUSHIL KUMAR AZAD

Example 22. Let the random variables X and Y have the joint density function:

$$f(x, y) = \begin{cases} 24xy, & x>0,\ y>0,\ x+y<1 \\ 0, & \text{elsewhere} \end{cases}$$

Find the regression equation of Y on X.

Solution. We know that the conditional mean of Y, given $X=x$, denoted by $E[Y|x]$ or $\mu_{Y/x}$ is given by

$$\mu_{Y/x} = \int_{-\infty}^{\infty} y\, f_{2|1}(y|x)\, dy$$

We have, $f_1(x)$ (marginal density of X)

$$= \int_0^{1-x} f(x,y)\, dx = \begin{cases} 12x(1-x)^2, & 0<x<1 \\ 0 & \text{elsewhere} \end{cases}$$

and given $X=x$, $f_{2|1}(y|x) = \dfrac{f(x,y)}{f_1(x)} = \dfrac{2y}{(1-x)^2}$, $0<y<1-x$.

$$\therefore \mu_{Y/x} = \int_0^{1-x} y \cdot \frac{2y}{(1-x)^2}\, dy$$

$$= \frac{2}{(1-x)^2}\left[\frac{y^3}{3}\right]_0^{1-x}$$

$$= \frac{2}{3}(1-x), \quad 0<x<1$$

$\Rightarrow$ the regression equation of Y on X, is

$$\mu_{Y/x}\left(E(Y/x)\right) = \frac{2}{3}(1-x), \quad 0<x<1. \quad //$$

Example 23. Find the regression equation of X on Y of joint distribution given by Example 22. Also find $P\left(0<X<\frac{1}{2} \,/\, Y=\frac{5}{8}\right)$.

Solution. We have, the marginal density of Y

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

$$= \int_0^{1-y} 24xy\, dx$$

$$\Rightarrow \quad f_2(y) = \begin{cases} 12y(1-y)^2, & 0<y<1 \\ 0 & \text{elsewhere.} \end{cases}$$

Sushil Kumar Azad

Given $Y = y \in (0, 1)$, we have

$$f_{1|2}(x|y) = \frac{f(x, y)}{f_2(y)}$$

$$= \frac{24xy}{12y(1-y)^2} = \frac{2x}{(1-y)^2}, \quad 0<x<1-y.$$

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

$$= \int_{-\infty}^{\infty} x . f_{1|2}(x|y)\, dx$$

$$= \int_0^{1-y} \frac{2x^2}{(1-y)^2}\, dx$$

$$\Rightarrow \mu_{X|y} = \frac{2}{3}(1-y), \quad 0<y<1.$$

which is the required result,

that is, given $Y = y \in (0, 1)$, the regression equation of X on Y is

$$\mu_{X|y} = \frac{2}{3}(1-y), \quad 0<y<1$$

Continued

Next, $P(0 < X < \frac{1}{2} \mid Y = \frac{5}{8})$

$$= \int_0^{1-5/8} f_{1/2}(x \mid y = \tfrac{5}{8})\, dx$$

$$= \int_0^{3/8} \frac{2x}{(1-5/8)^2}\, dx$$

$$= \int_0^{3/8} \frac{2x}{9/64}\, dx$$

$$= \frac{64}{9} \cdot [x^2]_0^{3/8}$$

$$= \frac{64}{9} \times \frac{9}{64} = 1,$$ which is the result. //

Sushil Kumar Azad

Example 24. Given the two random variables X and Y that have the joint density:

$$f(x, y) = \begin{cases} \dfrac{e^{-x/y}\, e^{-y}}{y} & 0 < x < \infty,\ 0 < y < \infty \\ 0 & \text{otherwise} \end{cases}$$

Find the regression equation of X on Y.

<u>Solution.</u> Consider the conditional density

$$f_{X/Y}(x \mid y) = \frac{f(x, y)}{f_Y(y)}, \text{ where}$$

$$f_Y(y) = \int_{-\infty}^{\infty} f(x, y)\, dx = \int_0^{\infty} \frac{1}{y} e^{-x/y} e^{-y}\, dx$$

$$= e^{-y} \int_0^{\infty} \left(\frac{1}{y}\right) e^{-x/y}\, dx = e^{-y} \left| e^{-x/y} \right|_0^{\infty}$$

$$\Rightarrow f_Y(y) = \begin{cases} e^{-y} & \text{if } 0 < y < \infty \\ 0 & \text{elsewhere} \end{cases}$$

$$\therefore f_{X/Y}(x/y) = \frac{1}{y} e^{-x/y}, \text{ and}$$

$$E(X \mid Y = y) = \int_0^{\infty} \frac{x}{y} e^{-x/y}\, dx = y,$$ which is the required regression equation of X on Y.

Example 25. From a partially destroyed, the two regression lines available are $X+2Y=4$ and $X+3Y=5$. Find the means μ_1, μ_2, ρ_{XY} and σ_X, ~~given that $\sigma_Y = 2$.~~

Solution: Let us assume that $X+2Y=4$ be the regression of Y on X, then $Y = -\frac{1}{2}X + 2$, $\Rightarrow \beta = -\frac{1}{2}$

and $X+3Y=5$ be the regression of X and Y, then we have $X = -3Y+5 \Rightarrow \delta = -3$

$$\Rightarrow \rho_{XY} = \pm\sqrt{\beta\delta} = \pm\sqrt{\left(-\tfrac{1}{2}\right)(-3)} = \pm\sqrt{\tfrac{3}{2}} > 1 \left(\text{or} < -1\right)$$

a contradiction, as $-1 \le \rho_{XY} \le 1$.

Therefore, the above assumption is wrong.

So, let $X+2Y=4$ be the regression line of X on Y, and $X+3Y=5$ be the regression line of Y on X.

Then, we have $X = -2Y+4$

and $Y = -\frac{1}{3}X + \frac{5}{3}$

$$\Rightarrow \beta = -\tfrac{1}{3} \ \& \ \delta = -2$$

$$\therefore \rho_{XY} = \pm\sqrt{\beta\delta} = \pm\sqrt{\left(-\tfrac{1}{3}\right)(-2)} = \pm\sqrt{\tfrac{2}{3}} < 1 \ (\text{or} > -1)$$

Since both β and δ are negative, we have

$$\boxed{\rho_{XY} = -\sqrt{\tfrac{2}{3}}}$$

SUSHIL KUMAR AZAD

Further, we know that two regression lines intersect at their means (μ_1, μ_2).

By solving the equations

$$Y = -\frac{1}{3}X + \frac{5}{3}$$

and $X = -2Y + 4$

we get $\boxed{\begin{array}{l}\mu_1 (= E(X)) = 2 \text{ and} \\ \mu_2 (= E(Y)) = 1.\end{array}}$

Now, $\beta = \rho \dfrac{\sigma_Y}{\sigma_X} = -\dfrac{1}{3}$

$$\Rightarrow -\sqrt{\frac{2}{3}}\,\frac{\sigma_Y}{\sigma_X} = -\frac{1}{3}$$

$$\Rightarrow -\sqrt{\frac{2}{3}}\cdot\frac{2}{\sigma_X} = -\frac{1}{3}$$

$$\Rightarrow \sigma_X = 6\sqrt{\frac{2}{3}} = 4.9$$

//

SUSHIL KUMAR AZAD