March 23, 2020

Dr. SUSHIL KUMAR AZAD

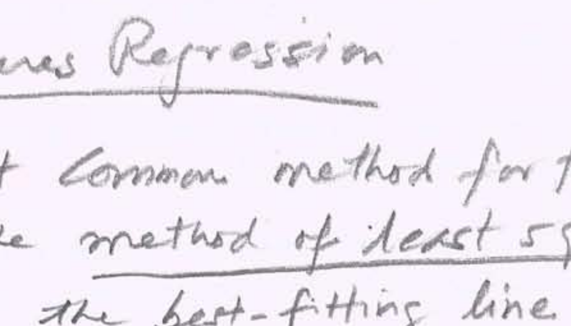
Linear regression – In statistics, linear regression is a linear approach to modeling the relationship between a scalar response (i.e. the output variables) (Y) and one or more explanatory variables (i.e. the input variables) (X_i, $i = 1, 2, \dots, n$). The case of one explanatory variable (X) is called simple linear regression (SLR).

A linear regression line has an equation of the form $Y = a + bX$, where X is the explanatory variable and Y is the dependent variable.

SUSHIL KUMAR AZAD

Least squares Regression

The most common method for fitting a regression line is the method of least squares. The method calculates the best-fitting line for the observed data by minimizing the sum of the squares of the vertical deviations from each data point to the line, S (if a point lies on the fitted line exactly, then its vertical deviation is 0).



$$S = \sum_i (y_i - y_{c_i})^2 = \sum (y_i - a - bx_i)^2,$$

for a given data set consisting of n points (x_i, y_i), $i = 1, \dots, n$.

The difference between the actual value of the dependent variable (y) and the value predicted by the

Continued - - -

model(y_c), $r_i = y_i - y_{ci}$,
$= y_i - a - bx_i$,

is called, residual. The least-squares method find the optimal parameter values (a & b) by minimizing the sum, S, of squared residuals:

$$S = \sum_{i=1}^{n} r_i^2 .$$

Calculating the parameters a & b:

For Min(S), we must have

$$\frac{\partial S}{\partial a} = 0 \ \& \ \frac{\partial S}{\partial b} = 0$$

$$\Rightarrow \quad \sum y = na + b\sum x, \ \& \quad \text{---(i)}$$

$$\sum xy = b\sum x + a\sum x^2 \quad \text{---(ii)}$$

The equations (i) & (ii) are called Normal equations.

Solving Equations (i) & (ii), we get

$$a = \frac{\sum y \sum x^2 - \sum x \sum xy}{n\sum x^2 - (\sum x)^2}$$

and

$$b = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2}$$

In the equation, $Y_c = a + bx$, if b is positive, the regression (or trend) line will be upward and if b is negative, the regression line will be downward.

Continued -

SUSHIL KUMAR AZAD

The Conditional Expectation function and regression

We know the conditional expectation function (CEF) is simply $E(Y|X)$, for a joint distribution (X, Y). Since X & Y are random, so is CEF.

We have the conditional mean of Y, given $X=x$

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

$$= \frac{\int_{-\infty}^{\infty} y \, f(x, y)\,dy}{f_1(x)}$$

assuming the random variables of the continuous type.

$\Rightarrow$ $E(Y|x)$ is a function of x, say $u(x)$.

In case $u(x)$ is a linear function of x, say $u(x) = a + bx$, we say that the conditional mean of Y is linear in x (or that Y has a linear conditional mean).

<u>Question 1.</u> What could we infer about the relationship between the dependent variable Y and the CEF, $E(Y|X)$.

Let $\quad Y = E(Y|X) + \varepsilon$

Then, using the properties of expectations, we can show that $E(\epsilon|X) = 0$, i.e., mean independent and ε is uncorrelated with any function of X.

SUSHIL KUMAR DZAD

Question 2. How would we find that the optimum choice of $u(X)$ is exactly the CEF?, that is,

As discussed above, we try to find a function, $u(X)$ that minimizes the squared mean error, i.e.,

$$\min E[(Y - u(X))^2] \quad \text{---} (iii)$$

Consider

$$(Y - u(X))^2 = \left((Y - E(Y|X)) + (E(Y|X) - u(X))\right)^2$$

$$= (Y - E(Y|X))^2 + 2(Y - E(Y|X))(E(Y|X) - u(X))$$
$$+ (E(Y|X) - u(X))^2$$

In the second term on the right side, $(Y - E(Y|X))$ is simply ε and a function of X multiplied with ε would still give an expectation of zero.

The first term on the right side does not factor in the min problem (iii).

Hence, the equation (iii) can be simplified to minimising last term on the right side, $[E(Y|X) - u(X)]^2$, which is only minimised when $u(X) = \text{CEF}$.

//

When $u(x) = a + bx$ the constants a and b have simple values which we show in the following theorem.

Theorem 1. Let (X, Y) have a joint distribution with means and variances of X and Y, μ_1, μ_2 and σ_1^2, σ_2^2, respectively, and let ρ be the correlation coefficient between X and Y. If $E(Y|X)$ is linear in X, then

Continued

SUSHIL KUMAR AZAD

$$E(Y|X) = \mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x - \mu_1) \quad \text{---(iv)}$$

and
$$E(\mathrm{var}(Y|X)) = \sigma_2^2(1-\rho^2) \quad \text{---(v)}$$

Proof. We assume that X, Y are of continuous type.

Let $E(Y|X=x) = a+bx$.

$$\Rightarrow \frac{\int_{-\infty}^{\infty} y f(x,y)\,dy}{f_1(x)} = a+bx$$

$$\Rightarrow \int_{-\infty}^{\infty} y f(x,y)\,dy = (a+bx) f_1(x) \quad \text{----(vi)}$$

Integrating on both sides w.r.t. 'x', we get

$$\int_{-\infty}^{\infty}\left(\int_{-\infty}^{\infty} y f(x,y)\,dy\right)dx = \int_{-\infty}^{\infty}(a+bx) f_1(x)\,dx$$

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

$$\Rightarrow E(Y) = a.1 + b.E(X).$$

or
$$\mu_2 = a + b\mu_1 \quad \text{---(vii)}$$

Now, ~~Con~~ Let (vi) $\times x$ & integrate w.r.t. 'x' on both sides, we get

$$\int_{-\infty}^{\infty} x \int_{-\infty}^{\infty} y f(x,y)\,dy\,dx = \int_{-\infty}^{\infty}(a+bx)\,x f_1(x)\,dx$$

$$\Rightarrow E(XY) = a\,E(X) + b\,E(X^2)$$

or
$$\rho\sigma_1\sigma_2 + \mu_1\mu_2 = a\mu_1 + b(\sigma_1^2 + \mu_1^2) \quad \text{---(viii)}$$

SUSHIL KUMAR AZAD

Solving equations (vii) & (viii), we get

$$a = \mu_2 - \rho\frac{\sigma_2}{\sigma_1}\mu_1, \text{ and}$$

$$b = \rho\frac{\sigma_2}{\sigma_1}$$

(Note that these values are same as obtained by LSM, on page (2))

SUSHIL KUMAR AZAD

$$\therefore \quad E(Y/x) = a + b\,x$$

$$= \left(\mu_2 - \rho\frac{\sigma_2}{\sigma_1}\mu_1\right) + \rho\frac{\sigma_2}{\sigma_1}x$$

$$\Rightarrow E(Y/x) = \mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x - \mu_1)$$

$$\therefore \quad E(Y/X) = \mu_2 + \rho\frac{\sigma_2}{\sigma_1}(X - \mu_1),$$ which

is the first result, (iv),

Now, consider the conditional variance of Y is given by

$$\text{var}(Y/x) = \int_{-\infty}^{\infty}\left[y - E(Y/x)\right]^2 f_{2|1}(y|x)\,dy$$

$$= \int_{-\infty}^{\infty}\left[y - \mu_2 - \rho\frac{\sigma_2}{\sigma_1}(x - \mu_1)\right]^2 \frac{f(x,y)}{f_1(x)}\,dy$$

$$= \frac{\int_{-\infty}^{\infty}\left[(y - \mu_2) - \frac{\rho\sigma_2}{\sigma_1}(x - \mu_1)\right]^2 f(x,y)\,dy}{f_1(x)}$$

continued.

SUSHIL KUMAR AZAD

$$\Rightarrow f_1(x)\,\mathrm{var}(Y|x) = \int_{-\infty}^{\infty} \left[(y-\mu_2) - \rho\frac{\sigma_2}{\sigma_1}(x-\mu_1)\right]^2 f(x,y)\,dy$$

$$\Rightarrow \int_{-\infty}^{\infty} \mathrm{var}(Y|x)\, f_1(x)\,dx$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \left[(y-\mu_2) - \frac{\rho\sigma_2}{\sigma_1}(x-\mu_1)\right]^2 f(x,y)\,dy\,dx$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \left[(y-\mu_2)^2 - 2\frac{\rho\sigma_2}{\sigma_1}(y-\mu_2)(x-\mu_1) + \frac{\rho^2\sigma_2^2}{\sigma_1^2}(x-\mu_1)^2\right] \times f(x,y)\,dy\,dx.$$

$$= E\left[(Y-\mu_2)^2\right] - 2\rho\frac{\sigma_2}{\sigma_1}E\left[(X-\mu_1)(Y-\mu_2)\right] + \frac{\rho^2\sigma_2^2}{\sigma_1^2}E\left[(X-\mu_1)^2\right]$$

$$= \sigma_2^2 - 2\rho\frac{\sigma_2}{\sigma_1}\times\rho\sigma_1\sigma_2 + \frac{\rho^2\sigma_2^2}{\sigma_1^2}\times\sigma_1^2$$

$$= \sigma_2^2 - 2\rho^2\sigma_2^2 + \rho^2\sigma_2^2$$

$$= \sigma_2^2(1-\rho^2)$$

$$\Rightarrow E(\mathrm{var}(Y/X)) = \sigma_2^2(1-\rho^2)$$

which is the desired result, (v). //

Example 21. Let $f(x,y) = \begin{cases} 2, & 0<x<y,\ 0<y<1 \\ 0, & \text{elsewhere} \end{cases}$

be the joint pdf of X and Y. Show that

(a) the conditional means are, respectively, $(1+x)/2,\ 0<x<1$ and $y/2,\ 0<y<1$,

(b) the correlation coefficient of X and Y is $\rho = \frac{1}{2}$

(c) the variance of the conditional distribution of Y, given $X=x$, is $(1-x)^2/12,\ 0<x<1$,

(d) the variance of the conditional distribution of X, given $Y=y$, is $y^2/12,\ 0<y<1$

Solution: The marginal pdfs of X and Y are

$$f_1(x) = \int_x^1 f(x,y)\,dy = \int_x^1 2\,dy = \begin{cases} 2(1-x), & 0<x<1 \\ 0 & \text{elsewhere} \end{cases}$$

& $$f_2(x) = \int_0^y f(x,y)\,dx = \begin{cases} 2y, & 0<y<1 \\ 0 & \text{elsewhere} \end{cases}$$

$\therefore$ The conditional pdfs of X, given $Y=y$ and of Y, given $X=x$, are

$$f_{1|2}(x/y) = \frac{f(x,y)}{f_2(y)} = \frac{1}{y}, \quad 0<y<1$$

and $$f_{2|1}(y/x) = \frac{f(x,y)}{f_1(x)} = \frac{1}{(1-x)}, \quad 0<x<1$$

Continued ...

(9)

(a) Now, the conditional mean of X, given $Y=y$ is

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

$$= \int_0^{y} \frac{x}{y}\, dx = \frac{y}{2}, \quad 0<y<1$$

& the conditional mean of Y, given $X=x$, is

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

$$= \int_x^1 y \cdot \frac{1}{1-x}\, dy$$

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

(b) we have

$$\mu_1 = E(X) = \int_0^1 x f_1(x)\, dx$$

$$= \int_0^1 2x(1-x)\, dx$$

$$= 2 \cdot \left| \frac{x^2}{2} - \frac{x^3}{3} \right|_0^1 = 2 \times \frac{1}{6} = \frac{1}{3}$$

and $\mu_2 = E(Y) = \int_0^1 y\, f_2(y)\, dy$

$$= 2 \left. \frac{y^3}{3} \right|_0^1 = \frac{2}{3},$$

Next, $E(X^2) = \int_0^1 x^2 f_1(x)\, dx = 2\left| \frac{x^3}{3} - \frac{x^4}{4} \right|_0^1 = \frac{1}{6}$

$$\Rightarrow \quad \sigma_1^2 = \frac{1}{6} - \left(\frac{1}{3}\right)^2 = \frac{1}{18},$$

and $E(Y^2) = \int_0^1 y^2 f_2(y)\, dy = \frac{2}{4} = \frac{1}{2}$

$$\Rightarrow \quad \sigma_2^2 = E(Y^2) - \mu_2^2 = \frac{1}{2} - \frac{4}{9} = \frac{1}{18}$$

SUSHIL KUMAR AZAD

Now, $E(XY) = \int_0^1 \int_0^y xy\, f(x,y)\,dx\,dy$

$$= \int_0^1 y \left(\int_0^y 2x\,dx \right) dy$$

$$= \int_0^1 y \cdot y^2\,dy = \frac{1}{4}$$

$$\Rightarrow Cov(X,Y) = E(XY) - \mu_1 \mu_2$$

$$= \frac{1}{4} - \frac{1}{3} \times \frac{2}{3}$$

$$= \frac{9-8}{36} = \frac{1}{36}$$

$$\Rightarrow \rho = \frac{Cov(X,Y)}{\sigma_1 \sigma_2} = \frac{(1/36)}{\sqrt{(1/18)}\sqrt{(1/18)}} = \frac{1}{2}$$

Hence the Correlation coefficient between X and Y, ρ is $\frac{1}{2}$. //

Alternatively, Since $E(Y/x) = \frac{1+x}{2}$, $0<x<1$,

i.e. $E(Y|x)$ is a linear function of x,

therefore by theorem

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

$$= \frac{2}{3} + \rho\left(x - \frac{1}{3}\right) \quad \text{--(ix)}$$

Now $\frac{1+x}{2} = \frac{1}{2} + \frac{1}{2}x = \frac{1}{2} + \frac{1}{2}\left(x - \frac{1}{3}\right) + \frac{1}{6}$

$$\Rightarrow E(Y/x) = \frac{1+x}{2} = \frac{2}{3} + \frac{1}{2}\left(x - \frac{1}{3}\right) \quad \text{--(x)}$$

Comparing (ix) & (x), we get, $\boxed{\rho = \frac{1}{2}}$ //

(c) The variance of the conditional distribution of Y, given $X=x$, is

$$var(Y/x) = \int_x^1 [y - E(Y/x)]^2 f_{2|1}(y|x)\,dy$$

$$= \int_x^1 \left[y - \frac{(1+x)}{2}\right]^2 \frac{1}{(1-x)}\,dy$$

$$= \frac{1}{(1-x)} \int_x^1 \left[y^2 - (1+x)y + \frac{(1+x)^2}{4}\right] dy$$

$$= \frac{1}{(1-x)} \left\{ \left[\frac{1}{3} - \frac{(1+x)}{2} + \frac{(1+x)^2}{4}\right] - \left[\frac{x^3}{3} - \frac{(1+x)x^2}{2} + \frac{(1+x)^2 x}{4}\right] \right\}$$

$$= \frac{1}{12(1-x)} \left\{ \left[4 - 6 - 6x + 3(1+2x+x^2)\right] - \left[4x^3 - (6x^2+6x^3) + 3x + 6x^2 + 3x^3\right] \right\}$$

$$= \frac{1}{12(1-x)} \cdot \left[1 - 3x^2 - 3x + x^3\right]$$

$$= \frac{1}{12(1-x)} \times (1-x)^3$$

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

$$\Rightarrow var(Y/x) = \frac{(1-x)^2}{12}, \quad 0 < x < 1,$$

which is the result

(d) The variance of the conditional distribution of X, given $Y=y$, is

$$var(X/y) = \int_0^y [x - E(X|y)]^2 f_{1|2}(x|y)\,dx$$

SUSHIL KUMAR AZAD

$$\Rightarrow \operatorname{var}(X/y) = \int_0^y \left[x - \frac{y}{2}\right]^2 \frac{1}{y}\,dx$$

$$= \frac{1}{y}\int_0^y \left(x^2 - xy + \frac{y^2}{4}\right) dx$$

$$= \frac{1}{y}\left|\frac{x^3}{3} - \frac{x^2 y}{2} + \frac{y^2 x}{4}\right|_0^y$$

$$= \frac{1}{y}\left\{\left[\frac{y^3}{3} - \frac{y^3}{2} + \frac{y^3}{4}\right] - [0]\right\}$$

$$= y^2\left(\frac{1}{3} - \frac{1}{2} + \frac{1}{4}\right)$$

$$= \frac{y^2}{12}, \quad 0 < y < 1$$

$\therefore$ $\operatorname{var}(X/y) = \dfrac{y^2}{12}$, $0 < y < 1$, which is the result.

SUNIL KUMAR ARYA