Descriptive statistics

UNIT V
REGRESSION

The relationship between the two variables is called Regression, one variable is independent variable and the other variable is dependent variable.
The meaning of the word regression is returning or going back. The line which gives the average relationship between the variables is known as the regression line. The corresponding equation is the regression equation. The value of the dependent variable is estimated corresponding to any value of the independent variable by using the regression equation.

Methods of forming the Regression Equations

1. Regression Equations on the basis of Normal Equations
2. Regression Equations on the basis of $\overline{X}$, $\overline{Y}$, b_{XY} , b_{YX}

Regression Equations on the basis of $\overline{X}$, $\overline{Y}$, b_{XY} , b_{YX}

Regression equation of Y on X	Regression equation of X on Y
$Y-\overline{Y}=b_{YX}(X-\overline{X})$	$X-\overline{X}=b_{XY}(Y-\overline{Y})$
b_{YX} is called the regression coefficient of Y on X	b_{XY} is called the regression coefficient of X on Y
$b_{YX}=\frac{r\sigma_Y}{\sigma_X}$	$b_{XY}=\frac{r\sigma_X}{\sigma_Y}$
$b_{YX}=\frac{\sum xy}{\sum x^2}$	$b_{XY}=\frac{\sum xy}{\sum y^2}$
$b_{YX}=\frac{N\sum XY-\left(\sum X\right)\left(\sum Y\right)}{N\sum X^2-\left(\sum X\right)^2}$	$b_{XY}=\frac{N\sum XY-\left(\sum X\right)\left(\sum Y\right)}{N\sum Y^2-\left(\sum Y\right)^2}$

1. You are given the following data:

	X	Y
Arithmetic mean	36	85
Standard deviation	11	8
Correlation coefficient between X and Y	0.66	

(a) Find the two Regression equations
(b) Estimate the value of X when Y = 75

Solution:

Given; $\overline{X}$ = 36, $\overline{Y}$ = 85, σ_X =11, σ_Y = 8, r = 0.66

$$b_{YX}=\frac{r\sigma_Y}{\sigma_X}=\frac{0.66X8}{11}=\frac{5.28}{11}=0.48$$

$$b_{XY}=\frac{r\sigma_X}{\sigma_Y}=\frac{0.66X11}{8}=\frac{7.26}{8}=0.9075$$

(a) Regression equation of Y on X

$Y-\overline{Y}=b_{YX}(X-\overline{X})$

Y – 85 = 0.48(X – 36)

Y – 85 = 0.48X – 17.28

Y = 0.48X – 17.28 + 85

Y = 0.48X + 67.72

Regression equation of X on Y

$X-\overline{X}=b_{XY}(Y-\overline{Y})$

X – 36 = 0.9075(Y- 85)

X – 36 = 0.9075Y – 77.1375

X = 0.9075Y – 77.1375 +36

X = 0.9075Y – 41.1375

X = 0.9075Y – 41.14

(b) When Y = 75, X = ?

Sub Y = 75 in the Regression equation of X on Y

X = 0.9075Y – 41.14

X = 0.9075(75) – 41.14

X = 68.0625 – 41.14

X = 26.9225

X = 26.92

2. From the following information on values of two variables X and Y find the two regression lines and the correlation coefficient:

$N = 10,\ \sum X = 20,\ \sum Y = 40,\ \sum X^2 = 240,\ \sum Y^2 = 410,\ \sum XY = 200$

Solution:

$$\overline{X}=\frac{\sum X}{N}=\frac{20}{10}=2$$

$$\overline{Y}=\frac{\sum Y}{N}=\frac{40}{10}=4$$

$$b_{YX}=\frac{N\sum XY-\left(\sum X\right)\left(\sum Y\right)}{N\sum X^2-\left(\sum X\right)^2}=\frac{10(200)-(20)(40)}{10(240)-(20)^2}=\frac{2000-800}{2400-400}=\frac{1200}{2000}=0.6$$

$$b_{XY}=\frac{N\sum XY-\left(\sum X\right)\left(\sum Y\right)}{N\sum Y^2-\left(\sum Y\right)^2}=\frac{10(200)-(20)(40)}{10(410)-(40)^2}=\frac{2000-800}{4100-1600}=\frac{1200}{2500}=0.48$$

(a) Regression equation of Y on X

$Y-\overline{Y}=b_{YX}(X-\overline{X})$

Y – 4 = 0.6(X-2)

Y -4 = 0.6X – 0.6X2

Y -4 = 0.6X – 1.2

Y = 0.6X – 1.2 + 4

Y = 0.6X + 2.8

Regression equation of X on Y

$$X - \overline{X} = b_{XY}(Y - \overline{Y})$$

X – 2 = 0.48(Y- 4)

X -2 = 0.48Y – 0.48X4

X – 2 = 0.48Y – 1.92

X = 0.48Y – 1.92 +2

X = 0.48Y + 0.08

The correlation coefficient

$$r = \frac{N\sum XY - \left(\sum X\right)\left(\sum Y\right)}{\sqrt{N\sum X^2 - \left(\sum X\right)^2}\sqrt{N\sum Y^2 - \left(\sum Y\right)^2}}$$

(or)

$$r = \pm\sqrt{b_{XY} X b_{YX}}$$

$$r = +\sqrt{0.48 X\, 0.6}$$

$$r = +\sqrt{0.288}$$

r = 0.5367

3. Calculate the two regression equations from the following data:
Also estimate Y when X = 20.

X	Y	X^2	Y^2	XY
10	40	100	1600	400
12	38	144	1444	456
13	43	169	1849	559
12	45	144	2025	540
16	37	256	1369	592
15	43	225	1849	645
$\sum X = 78$	$\sum Y = 246$	$\sum X^2 = 1038$	$\sum Y^2 = 10136$	$\sum XY = 3192$

Solution:

$$\overline{X} = \frac{\sum X}{N} = \frac{78}{6} = 13,\quad \overline{Y} = \frac{\sum Y}{N} = \frac{246}{6} = 41$$

$$b_{YX} = \frac{N\sum XY - \left(\sum X\right)\left(\sum Y\right)}{N\sum X^2 - \left(\sum X\right)^2} = \frac{6(3192) - (78)(246)}{6(1038) - (78)^2} = \frac{19152 - 19188}{6228 - 6084} = \frac{-36}{144} = -0.25$$

$$b_{XY} = \frac{N\sum XY - \left(\sum X\right)\left(\sum Y\right)}{N\sum Y^2 - \left(\sum Y\right)^2} = \frac{6(3192) - (78)(246)}{6(10136) - (246)^2} = \frac{19152 - 19188}{60816 - 60516} = \frac{-36}{300} = -0.12$$

Regression equation of Y on X

$$Y - \overline{Y} = b_{YX}(X - \overline{X})$$

Y – 41 = -0.25(X – 13)

Y – 41 = -0.25X + 3.25

Y = -0.25X +3.25 + 41

Y = -0.25X + 44.25

Regression equation of X on Y

$$X - \overline{X} = b_{XY}(Y - \overline{Y})$$

X – 13 = -0.12 (Y – 41)

X – 13 = -0.12Y + 4.92
X = -0.12Y + 4.92 +13
X = -0.12Y + 17.92

When X = 20, Y = ?

Sub X = 20 in the Regression equation of Y on X
Y = -0.25X + 44.25
Y = -0.25(20) + 44.25
Y = -5 + 44.25
Y = 39.25

The correlation coefficient

$r = \pm\sqrt{b_{XY} X b_{YX}}$

$r = -\sqrt{-0.12X - 0.25}$

$r = -\sqrt{-0.03}$

r = -0.1732

4. From the data given below, Find
(a) The two regression equations
(b)The coefficient of correlation between the marks in Mathematics and Statistics
(c)The most likely marks in Statistics when the marks in Mathematics is 30

Solution:
Let the marks in Mathematics be X
Let the marks in Statistics be Y

X	Y	X^2	Y^2	XY
25	43	625	1849	1075
28	46	784	2116	1288
35	49	1225	2401	1715
32	41	1024	1681	1312
31	36	961	1296	1116
36	32	1296	1024	1152
29	31	841	961	899
38	30	1444	900	1140
34	33	1156	1089	1122
32	39	1024	1521	1248
$\sum X = 320$	$\sum Y = 380$	$\sum X^2 = 10380$	$\sum Y^2 = 14838$	$\sum XY = 12067$

$$\overline{X} = \frac{\sum X}{N} = \frac{320}{10} = 32, \quad \overline{Y} = \frac{\sum Y}{N} = \frac{380}{10} = 38$$

$$b_{YX} = \frac{N\sum XY - \left(\sum X\right)\left(\sum Y\right)}{N\sum X^2 - \left(\sum X\right)^2} = \frac{10(12067) - (320)(380)}{10(10380) - (320)^2} = \frac{120670 - 121600}{103800 - 102400} = \frac{-930}{1400} = -0.6643$$

$$b_{XY} = \frac{N\sum XY - (\sum X)(\sum Y)}{N\sum Y^2 - (\sum Y)^2} = \frac{10(12067) - (320)(380)}{10(14838) - (380)^2} = \frac{120670 - 121600}{148380 - 144400} = \frac{-930}{3980} = -0.2337$$

(a) Regression equation of Y on X

$$Y - \overline{Y} = b_{YX}(X - \overline{X})$$

Y – 38 = -0.6643(X – 32)

Y – 38 = -0.6643X + 21.2576

Y = -0.6643X +21.2576 +38

Y = -0.6643X + 59.26

(b) Regression equation of X on Y

$$X - \overline{X} = b_{XY}(Y - \overline{Y})$$

X – 32 = -0.2337(Y – 38)

X – 32 = -0.2337Y + 8.8806

X = -0.2337Y + 8.8806 + 32

X = -0.2337Y + 40.88

For finding the value of marks in Statistics (Y), when the marks in Mathematics is (X) = 30,

Substitute X = 30 in the Regression equation of Y on X

Y = -0.6643X + 59.26

Y = -0.6643(30) + 59.26

Y = -19.929 +59.26

Y = 39.33

The correlation coefficient

$$r = \pm\sqrt{b_{XY} X b_{YX}}$$

$$r = -\sqrt{-0.2337X - 0.6643}$$

$$r = -\sqrt{0.15524}$$

r = -0.3940

5. Height of the father and son are given below. Find the height of the son when the height of the father is 70 inches.

Father(inches) X	Son(inches) Y	X^2	Y^2	XY
71	69	5041	4761	4899
68	64	4624	4096	4352
66	65	4356	4225	4290
67	63	4489	3969	4221
70	65	4900	4225	4550
71	62	5041	3844	4402
70	65	4900	4225	4550
73	64	5329	4096	4672
72	66	5184	4356	4752
65	59	4225	3481	3835
66	62	4356	3844	4092
$\sum X = 759$	$\sum Y = 704$	$\sum X^2 = 52445$	$\sum Y^2 = 45122$	$\sum XY = 48615$

$$\overline{X} = \frac{\sum X}{N} = \frac{759}{11} = 69$$

$$\overline{Y} = \frac{\sum Y}{N} = \frac{704}{11} = 64$$

$$b_{YX} = \frac{N\sum XY - \left(\sum X\right)\left(\sum Y\right)}{N\sum X^2 - \left(\sum X\right)^2} = \frac{11(48615) - (759)(704)}{11(52445) - (759)^2} = \frac{534765 - 534336}{576895 - 576081} = \frac{429}{814} = 0.5270$$

$$b_{XY} = \frac{N\sum XY - \left(\sum X\right)\left(\sum Y\right)}{N\sum Y^2 - \left(\sum Y\right)^2} = \frac{11(48615) - (759)(704)}{11(45122) - (704)^2} = \frac{534765 - 534336}{496342 - 495616} = \frac{429}{726} = 0.5909$$

(a) Regression equation of Y on X

$$Y - \overline{Y} = b_{YX}(X - \overline{X})$$

Y - 64 = 0.5270(X – 69)

Y – 64 = 0.5270X – 69X0.5270

Y – 64 = 0.5270X – 36.363

Y = 0.5270X – 36.363 + 64

Y = 0.5270X + 27.637

Regression equation of X on Y

$$X - \overline{X} = b_{XY}(Y - \overline{Y})$$

X – 69 = 0.5909(Y-64)

X – 69 = 0.5909Y – 37.8176

X = 0.5909Y – 37.8176 + 69

X = 0.5909Y + 31.1824

For finding the value of Y when X = 70(father's height),
Substitute X = 70 in the Regression equation of Y on X

Y = 0.5270X + 27.637

Y = 0.5270(70) + 27.637

Y = 36.89 + 27.637

Y = 64.527

Y = 65

The correlation coefficient

$$r = \pm\sqrt{b_{XY} X b_{YX}}$$

$$r = +\sqrt{0.5270 X 0.5909}$$

$$r = +\sqrt{0.3114043}$$

r = 0.5580

Properties of Regression Lines and Coefficients

1. The two regression equations are generally different and are not to be interchanged in their usage.
2. The two regression lines intersect at $(\overline{X}, \overline{Y})$.
3. Correlation coefficient is the geometric mean of the two regression coefficients.
   $r = \pm\sqrt{b_{XY} X b_{YX}}$
4. The two regression coefficients and the correlation coefficient have the same sign.

5. Both the regression coefficients cannot be greater than 1 numerically simultaneously.
6. Regression coefficients are independent of change of origin but are affected by change of scale
7. Each regression coefficient is in the unit of the measurement of the dependent variable
8. Each regression coefficient indicates the quantum of change in the dependent variable corresponding to unit increase in the independent variable.

Difference between Correlation and Regression:

	Correlation	Regression
1	Correlation is the relationship between variables. It is expressed numerically.	Regression means going back. The average relation between the variables is given as an equation.
2	Between two variables, non is identified as independent or dependent.	One of the variables is independent variable and the other is dependent variable in any particular context.
3	Correlation does not mean causation. One variable need not be the cause and the other effect.	Independent variable may be the 'cause' and dependent variable be the'effect'.

6. Given the regression lines as 2x - y + 1= 0 and 3x -2y+7 =0

Find their point of intersection and interpret it. Also find the correlation Co-efficient between x and y.

Solution:

$3x + 2y = 26$ ---------- I

$6x + y = 31$ ------------ II

$3x + 2y = 26$

II X 2 $\quad 12x + 2y = 62$ --------III

III – I $\quad 9x = 36 \rightarrow x = 36/9 = 4$

Substitute x=4 in I, $3x4 + 2y =26$

$12 +2y = 26$

$2y = 26 – 12 = 14$

$Y = 14/2 = 7$

Point of Intersection is (4, 7)

$\overline{X} = 4, \quad \overline{Y} = 7$

Let $3x + 2y = 26$ be the regression equation of y on x

$\rightarrow$ $2y = 26-3y \Rightarrow y = 13- 3/2\, y \Rightarrow b_{yx} = -3/2$

Let $6x + y = 31$ be the regression equation of x on y

⇨ $6x = 31 – y \Rightarrow x = 31/6 – y/6$ =. $b_{yx} = - 1/6$

$r = \pm\sqrt{(bxy * \ byx)} = - \sqrt{(-1/6) \times (-3/2)} = - \sqrt{3/12} = - \sqrt{0.25}$

Curve Fitting:

There are a number of problems in Various fields for estimation of one variable with the help of one or more variables. For eg.,

1) The weight of a baby could be estimated from its age.
2) A students success in an University examination Could be predicted from his performance in a College examination.
3) The biologist relates the Size of a growing population to time
4) The meteorologist forecasts the path of a Storm.
5) The maintenance cost of an automobile is related to its age
6) The astronomer predicts the path of a satellite.

For all Such problems the curve fitting forms the basis. For fitting the curve we use the principle of least squares.

The form of the curve to fit a Statistical data Should be known to apply the principle of least Squares.

The principle of least squares will enable us to determine the parameters involved in the relationship Connecting the variables.

Principle of Least Squares.

Suppose the variables x and y are functionally related, i.e., the graph of the relation is a Specified curve. Let us further suppose that the relation is Suggested for predicting the Values of y for known values of x. Then y is a random variable and x is a mathematical ~~model~~.

variable. The variable x is then assumed to have an error associated with it. For any arbitrarily chosen value of x, the value of the random variable is determined from the suggested relation. This estimated value of y will have an error associated with it which is of random nature. The curve of best fit is that which makes the errors of estimation, known as residuals, as small as possible.

Let $y = f(x)$ be the relation suggested between the variables x and y.

Let (x_i, y_i), $i = 1, 2 \ldots n$ be n sample values.

Then $y - y_i$ is a residual.

The principle of Least Squares provides the technique for minimising the error.

The error of the estimate $y - y_i$ may be positive or negative. Therefore, for the best filting curve the sum of the absolute value of errors should be as small as possible

Since the absolute values is not convinent we ~~use~~ make the sum of squares of the errors as small as possible.

The parameters involved in $f(x)$ can be determined by the minimisation principle.

Filting a Straight line

Suppose $(x_1, y_1), (x_2, y_2), \ldots (x_n y_n)$ be n pairs of values and we have to ~~fit~~ determine the line of best fit for this data.

Let us assume $y = a + bx$ as a line of best fit.

Using the principle of least squares we can determine the parameters a and b.

The parameters a and b are determined by the equations

$$na + b\Sigma x = \Sigma y \quad \text{———} \quad (1)$$
$$a\Sigma x + b\Sigma x^2 = \Sigma xy \quad \text{———} \quad (2)$$

These equations are called normal equations.
Therefore, the line of best fit is $y = a + bx$,
Where a and b are given by solving the equations (1) & (2).

Fitting a Second degree Polynomial

Let $(x_1, y_1), (x_2, y_2) \dots (x_n, y_n)$ be n pairs of values of (x, y)

Suppose $y = a + bx + cx^2$ is suggested relationship between the variables x and y. The parameters a, b, and c are determined using the principle of least squares.
They are found by Solving the normal equations

$$na + b\Sigma x + c\Sigma x^2 = \Sigma x$$
$$a\Sigma x + b\Sigma x^2 + c\Sigma x^3 = \Sigma xy$$
$$a\Sigma x^2 + b\Sigma x^3 + c\Sigma x^4 = \Sigma x^2 y.$$

Fitting a curve of the form $y = ae^{bx}$ (exponential curve)

Let $y = ae^{bx}$, where a & b are constants be the suggested relationship between x and y.

Taking log on both sides we have,

$$\log y = \log a + bx$$
$$\text{ie} \quad Y = A + bx$$

After getting the values of A & b we get $a = A.L.(\log A)$ and substitute in $y = ae^{bx}$.

By the method of Least Squares find the best fitting Straight line to the data given below

x :	5	10	15	20	25
y :	16	19	23	26	30

Let the straight line be $y = ax + b$

The normal equations are

$$a\Sigma x + nb \cancel{\Sigma y} = \Sigma y$$

$$a\Sigma x^2 + b\Sigma x = \Sigma xy$$

x	y	x^2	xy
5	16	25	80
10	19	100	190
15	23	225	345
20	26	400	520
25	30	625	750
$\Sigma x = 75$	$\Sigma y = 114$	$\Sigma x^2 = 1375$	$\Sigma xy = 1885$

The normal equations are

$$75a + 5b = 114$$

$$1375a + 75b = 1885$$

Eliminate b,

multiply (1) by 15

$$1125a + 75b = 1710$$

(2) - (3) gives,

$$250a = 175 \text{ or } a = 0.7$$

Hence, $b = 12.3$

Hence, the best fitting line is

$$y = 0.7x + 12.3$$

Fit a Straight line to the data given below. Also estimate the value of y at $x = 2.5$

x :	0	1	2	3	4
y :	1	1.8	3.3	4.5	6.3

Let the best fit be $y = ax + b$

The normal equations are

$$\Sigma y = a\Sigma x + nb \quad \text{— I}$$

$$\Sigma xy = a\Sigma x^2 + b\Sigma x \quad \text{— II}$$

I ⇒

x	y	x^2	xy
0	1.0	0	0
1	1.8	1	1.8
2	3.3	4	6.6
3	4.5	9	13.5
4	6.3	16	25.2
10	16.9	30	47.1

$I \rightarrow 10a + 5b = 16.9$ — ①

$II \rightarrow 30a + 10b = 47.1$ — ②

① × 2; $20a + 10b = 33.8$ — ③

② − ③; $10a = 13.3$

$\Rightarrow a = \frac{13.3}{10} = 1.33$

Sub $a = 1.33$ in ①

$10 \times 1.33 + 5b = 16.9$

$5b = 16.9 - 13.3$

$b = \frac{3.6}{5} = 0.72$

Hence, the equation is

$$y = 1.33x + 0.72$$

y at $x = 2.5$; $y = 1.33(2.5) + 0.72 = 4.045$

3. Fit a parabola by the method of least squares, to the following data, also estimate y at $x = 6$.

x :	1	2	3	4	5
y :	5	12	26	60	97

Sol. Let $y = ax^2 + bx + c$ be the equation of best fit

Then the normal equations are

$$\Sigma y = a\Sigma x^2 + b\Sigma x + nc \quad \text{— ①}$$

$$\Sigma xy = a\Sigma x^3 + b\Sigma x^2 + c\Sigma x \quad \text{— ②}$$

$$\Sigma x^2 y = a\Sigma x^4 + b\Sigma x^3 + c\Sigma x^2 \quad \text{— ③.}$$

x	y	x^2	x^3	x^4	xy	x^2y
1	5	1	1	1	5	5
2	12	4	8	16	24	48
3	26	9	27	81	78	234
4	60	16	64	256	240	960
5	97	25	125	625	485	2425
15	200	55	225	979	832	3672

The equations ①, ② & ③ becomes

$$55a + 15b + 5c = 200 \quad \text{--- ④}$$

$$225a + 55b + 15c = 832 \quad \text{--- ⑤}$$

$$979a + 225b + 55c = 3672 \quad \text{--- ⑥}$$

Solving we get $a = 5.7143$, $b = -11.0858$ & $c = 10.4001$

Hence, the parabola is

$$y = 5.7143x^2 - 11.0858x + 10.4001.$$

When $y(x=6) = 149.6001$.

Fitting an exponential curve

Let $(x_i, y_i), i = 1, 2 \ldots n$ be the n sets of observations of related data and let $y = ab^x$ be the best fit for the data.

Taking log on both sides

$$\log_{10} y = \log_{10} a + x\log_{10} b$$

ie $Y = A + Bx$, where $Y = \log_{10} y$, $B = \log_{10} b$ & $A = \log_{10} a$

Using this linear fit we find A, b.

Hence a & b are known. Thus $y = ab^x$ is found.

Fit a curve of the form $y = ab^x$ to the given data

x :	1	2	3	4	5	6
y :	151	100	61	50	20	8

Sol. $y = ab^x$

$$\log_{10} y = \log_{10} a + x \log_{10} b$$

$$Y = A + Bx \quad \text{--- ①}$$

The normal equations are

$$\Sigma Y = nA + B\Sigma x \qquad \text{—— (2)}$$

$$\Sigma xY = A\Sigma x + B\Sigma x^2 \qquad \text{—— (3)}$$

x	y	$Y = \log y$	x^2	xY
1	151	2.1790	1	2.1790
2	100	2.000	4	4.0000
3	61	1.7853	9	5.3559
4	50	1.6990	16	6.7960
5	20	1.3010	25	6.5050
6	8	0.9031	36	5.4186
21		9.8674	91	30.2545

Using (2) and (3) we get

$$6A + 21B = 9.8674 \qquad \text{—— (4)}$$

$$21A + 91B = 30.2545 \qquad \text{—— (5)}$$

Solving, we get

$$A = 2.5010, \quad B = -0.2447$$

Since $\log_{10} a = A$, $a = A.L(A) = 316.9568$

$$b = A.L(B) = 0.5692$$

$\therefore$ The equation is

$$y = 316.9568(0.5692)^x.$$