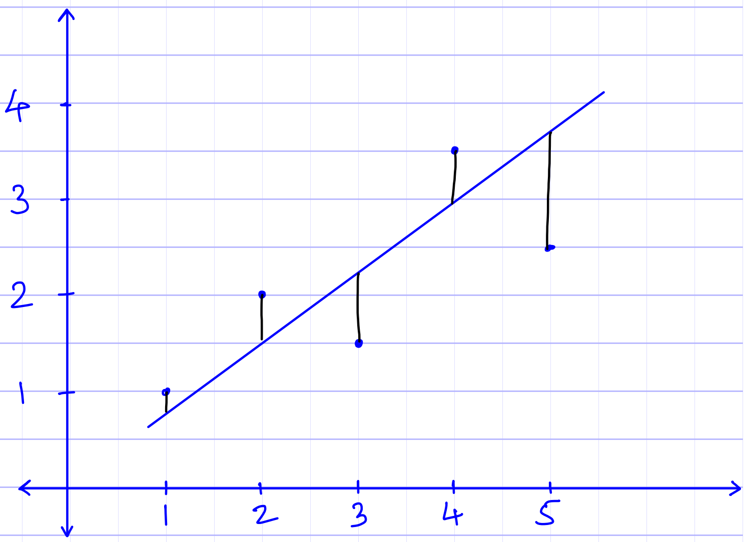
STATISTICS

STATISTICS

• Linear Regression : (Regression Line)

→ If variables are correlated, then Reg^n is the method of estimating value of one variable when value of other variable is known.

x	y
1	1
2	2
3	1.5
4	3.5
5	2.5



1) Regression line of Y on X
2) Regression line of X on Y

① Regression line of Y on X.

→ minimize dev^n of point measured along Y axis.

→ $Y = a + bX$ or $Y - \overline{Y} = b_{yx}(X - \overline{X})$

here, $b = \text{slope} = b_{yx} = r \cdot \dfrac{\sigma_y}{\sigma_x}$ = regression coefficient

& $a = y \text{ intercept} = \overline{Y} - b.\overline{X}$

→ Used for estimating value of Y, when X is given.

② Regression line of X on Y.

→ minimize dev^n of point measured along X axis.

→ $X = a + by$ or $X - \overline{X} = b_{xy}(Y - \overline{Y})$

here, $b = \text{slope} = b_{xy} = r \cdot \dfrac{\sigma_x}{\sigma_y}$ = regression coefficient

& $a = \overline{X} - b \cdot \overline{Y}$

→ Used for estimating value of X, when Y is given.

$$\boxed{b_{yx} = \dfrac{\dfrac{\Sigma xy}{n} - \overline{X} \cdot \overline{Y}}{\dfrac{\Sigma(x^2)}{n} - (\overline{X})^2}}$$

$$\boxed{b_{xy} = \dfrac{\dfrac{\Sigma xy}{n} - \overline{X} \cdot \overline{Y}}{\dfrac{\Sigma(y^2)}{n} - (\overline{Y})^2}}$$

- $\tan\theta = \left| \dfrac{m_1 - m_2}{1 + m_1 m_2} \right|$, $y - y_1 = m(x - x_1)$
- Two reg^n line intersect at $(\overline{X}, \overline{Y})$
- $r = \sqrt{b_{xy} \cdot b_{yx}}$ if b_{xy} & b_{yx} are −ve, r is also −ve.

  * • • • *

- reg^n coefficient is independent of change of origin but not the change of scale.
- reg^n line depends on change of origin & scale.

Interpretation of regression coefficient :

- Unit change in X will make changes by b_{yx} units in Y variable.

  if $b_{yx} > 0$, then increase in X (by 1 unit) will be associated by b_{yx} units (Increase) in Y.

  if $b_{yx} < 0$, then increase in X (by 1 unit) will be associated by b_{yx} units (Decrease) in Y.

Q1: Given that, $x - 4y = 5$ & $x - 16y = -64$ are the regression lines.

① Regression Coefficient for regⁿ line y on x & x on y.

② obtain r

③ obtain $\bar{X}$, $\bar{Y}$

④ find σ_y if $\sigma_x = 8$

By looking at equⁿ, we can not decide which of the equⁿ is regⁿ line of Y on X or X on Y.

let's consider, equⁿ ① is regⁿ line Y on X & ② is X on Y.

$$x - 4y = 5 \quad \ldots ①$$

$$y = \frac{x}{4} - \frac{5}{4}$$

$$\boxed{\therefore b_{yx} = 1/4}$$

$$x - 16y = -64 \quad \ldots ②$$

$$x = 16y - 64$$

$$\boxed{\therefore b_{xy} = 16}$$

$$\therefore r = \sqrt{b_{yx} \cdot b_{xy}} = +2$$

but, $-1 < r < 1$

$\therefore$ Assumption is Wrong.

∴ 1 is X on Y & 2 is Y on X.

$x - 4y = 5$	$x - 16y = -64$
$\therefore x = 4y + 5$	$\therefore y = \frac{x}{16} - 4$
$\boxed{b_{xy} = 4}$	$\boxed{\therefore b_{yx} = \frac{1}{16}}$

$$\therefore r = \sqrt{4 \times \frac{1}{16}} = \frac{1}{2}$$

$$\boxed{\therefore r = +0.5}$$ both +, ∴ +

On Solving 1 & 2,

$$\boxed{\therefore \overline{X} = 28 \quad \& \quad \overline{Y} = \frac{23}{4}}$$

$\sigma_x = 8$, $b_{yx} = r \cdot \frac{\sigma_y}{\sigma_x}$

$$\therefore \frac{1}{16} = \frac{1}{2} \times \frac{\sigma_y}{8}$$

$$\boxed{\therefore \sigma_y = 1}$$

Q2: Height of a father & son are given below

(X) Father	150	152	155	157	160	161	164	166
(Y) Son	154	156	158	159	160	162	161	164

Find equ^n of 2 lines of reg^n. Calculate height of son when height of father is 154 cm. Also, find r.

X	Y	A = X - 160	B = Y - 160	AB	A^2	B^2
150	154	-10	-6	60	100	36
152	156	-8	-4	32	64	16
155	158	-5	-2	10	25	4
157	159	-3	-1	3	9	1
160	160	0	0	0	0	0
161	162	1	2	2	1	4
164	161	4	1	4	16	1
166	164	6	4	24	36	16
		$\Sigma = -15$	$\Sigma = -6$	$\Sigma = 135$	$\Sigma = 251$	$\Sigma = 78$

$$\bar{A} = \frac{-15}{8} \quad , \quad \bar{B} = \frac{-6}{8}$$

$$\bar{A} = -1.8750 \quad , \quad \bar{B} = -0.7500$$

Reg^n Coeff. is indepd. of change of origin.

$$\therefore \quad b_{yx} = \frac{\frac{\Sigma AB}{n} - \bar{A}\cdot\bar{B}}{\frac{\Sigma A^2}{n} - (\bar{A})^2} = \frac{\frac{135}{8} - (-1.8750)(0.7500)}{\frac{251}{8} - (-1.8750)^2}$$

$$= \frac{16.8750 - 1.4063}{31.3750 - 3.5156}$$

$$= \frac{15.4687}{27.8594}$$

$$\boxed{\therefore \quad b_{yx} = 0.5552}$$

Now,

$$b_{xy} = \frac{\frac{\Sigma AB}{n} - \bar{A}\cdot\bar{B}}{\frac{\Sigma B^2}{n} - (\bar{B})^2} = \frac{16.8750 - 1.4063}{9.7500 - 0.5625}$$

$$\boxed{b_{xy} = 1.6837}$$

$\therefore$ Reg^n line of Y on X is

$$B - \bar{B} = b_{yx} \cdot (A - \bar{A})$$

$$B - (-0.7500) = 0.5552 \times [A - (-1.8750)]$$

$$\therefore y - 160 + 0.7500 = 0.5552 [x - 160 + 1.8750]$$

$$\boxed{y = 0.552x + 71.4590} \quad \ldots \text{(I)}$$

Now, Reg^n line of X on Y is

$$\therefore A - \bar{A} = b_{xy} (B - \bar{B})$$

$$\therefore x - 160 - (-1.8750) = 1.6837 [y - 160 - (-0.7500)]$$

$$\boxed{\therefore x = 1.6837y - 110.0042} \quad \ldots \text{(II)}$$

Given, $x = 154$, $Y = ?$

from (I), $y = 0.5552 \times x + 71.4590$

putting values,

$$\boxed{y = 156.9598}$$

$$y \approx 157 \text{ cm}$$

$$\boxed{\therefore \text{height of son is } 157 \text{ cm}}$$

$$\therefore \quad r = \sqrt{b_{yx} \times b_{xy}}$$

$$= \sqrt{0.5552 \times 1.6837}$$

$$\therefore \quad \boxed{r = 0.9668}$$

Q3 : Following results were obtained from marks in APM & maths

	APM	Maths
mean	47.5	39.5
S.D.	16.8	10.8

here, $r = 0.95$, Find both reg^n equations & estimate Y when $x = 30$

$$\bar{X} = 47.5 \quad , \quad \bar{Y} = 39.5$$

$$\sigma_x = 16.8 \quad , \quad \sigma_y = 10.8$$

$$\therefore \quad b_{yx} = r \cdot \frac{\sigma_y}{\sigma_x} \quad , \quad b_{xy} = r \cdot \frac{\sigma_x}{\sigma_y}$$

$$\boxed{b_{yx} = 0.6107} \quad , \quad \boxed{b_{xy} = 1.4778}$$

$\therefore$ Reg^n line of Y on X is

$$\therefore \quad Y - \bar{Y} = b_{yx} (X - \bar{X})$$

$$\therefore \quad Y - 39.5 = 0.6107 \, (x - 47.5)$$

$$\boxed{\therefore \quad y = 0.6107 x + 10.4918} \quad \ldots\ldots \text{(I)}$$

$\therefore$ Reg^n line of X on Y is

$$\therefore \quad X - \bar{X} = 1.4778(Y - \bar{Y})$$

$$\therefore \quad x - 47.5 = 1.4778(y - 39.5)$$

$$\boxed{\therefore x = 1.4778 y - 10.8731} \quad \ldots \text{(II)}$$

from (I), $\quad y = 0.6107 \times 30 + 10.4918$

$$\boxed{y = 28.8128}$$

Q4: If θ is angle bet^n two reg^n lines, Show that

$$\tan\theta = \left(\frac{1-r^2}{r}\right)\frac{\sigma_x \cdot \sigma_y}{\sigma_x^2 + \sigma_y^2}$$; Explain significance when $r = 0$ & $r = \pm 1$

$\rightarrow$ Reg^n line of Y on X is $\quad Y - \bar{Y} = b_{yx}(X - \bar{X})$

$$\therefore \quad m_1 = b_{yx} \quad \ldots \{\text{slope point form}\}$$

Reg^n line of X on Y is $\quad X - \bar{X} = b_{xy}(Y - \bar{Y})$

$$\therefore \quad m_2 = \frac{1}{b_{xy}} \quad \ldots [\text{slope point form}]$$

We know, $\quad \tan\theta = \left|\frac{m_1 - m_2}{1 + m_1 m_2}\right|$

$$\therefore \quad \tan\theta = \left|\frac{b_{yx} - 1/b_{xy}}{1 + \frac{b_{yx}}{b_{xy}}}\right|$$

$$\therefore \tan\theta = \left| \frac{r^2 - 1}{b_{xy} + b_{yx}} \right|$$

$$\therefore \tan\theta = \left| \frac{r^2 - 1}{r \cdot \frac{\sigma_x}{\sigma_y} + r \cdot \frac{\sigma_y}{\sigma_x}} \right|$$

$$\therefore \tan\theta = \left| \frac{r^2 - 1}{r\left(\frac{\sigma_x^2 + \sigma_y^2}{\sigma_x \cdot \sigma_y}\right)} \right|$$

$$\boxed{\tan\theta = \frac{1 - r^2}{r} \cdot \frac{\sigma_x \cdot \sigma_y}{\sigma_x^2 + \sigma_y^2}}$$

when, $r = 0$, $\tan\theta \to \infty$

$\therefore \theta = \pi/2$

$\boxed{\therefore \text{Two lines are perpendicular to each other.}}$

when, $r = \pm 1$, $\tan\theta \to 0$

$\therefore \theta = 0$

$\boxed{\therefore \text{Two lines are parallel to each other.}}$

Fitting of Curves : (Non linear Regression Model)

- Fitting of second degree curve :

Suppose $\{(x_i, y_i), i = 1, 2, 3, \dots n\}$ are observations.

y = dependent & x = independent

General equation of 2^{nd} degree curve is ,

$$y = a + bx + c.x^2$$

Let, y_i be observed value & $\hat{y}_i$ be estimated value

$$\therefore \hat{y}_i = a + bx_i + c.x_i^2$$

also, error or deviation $= y_i - \hat{y}_i$

$$\therefore S = \sum (y_i - \hat{y}_i)^2$$

$$\therefore S = \sum (y_i - a - bx_i - c.x_i^2)^2 \quad \dots\dots \quad (*)$$

for S to be minimum, $\frac{\partial s}{\partial a} = 0$, $\frac{\partial s}{\partial b} = 0$, $\frac{\partial s}{\partial c} = 0$

$$\therefore \frac{\partial s}{\partial a} = \frac{\partial}{\partial a} \sum y_i - a - b.x_i - cx_i^2)^2$$

$$\therefore \frac{\partial s}{\partial a} = \sum 2.(y_i - a - bx_i - cx_i^2)(-1)$$

$$\therefore 0 = \sum (y_i - a - bx_i - cx_i^2)$$

$$\therefore 0 = \sum y_i - \sum a - b \sum x_i - c \sum x_i^2$$

$$\therefore 0 = \sum y_i - an - b \sum x_i - c \sum x_i^2$$

$$\boxed{\therefore \sum y_i = na + b\sum x_i + c\sum x_i^2} \quad \ldots\ldots\ldots \text{I}$$

$\frac{\partial s}{\partial b}$ Now, $0 = \sum x_i \cdot y_i - a\sum x_i - b\sum x_i^2 - c\sum x_i^3$

$$\boxed{\therefore \sum x_i \cdot y_i = a\sum x_i + b\cdot\sum x_i^2 + c\cdot\sum x_i^3} \quad \ldots\ldots \text{II}$$

Now, $0 = \sum x_i^2 \cdot y_i - a\sum x_i^2 - b\sum x_i^3 - c\cdot\sum x_i^4$

$\frac{\partial s}{\partial c}$

$$\boxed{\therefore \sum x_i^2 \cdot y_i = a\sum x_i^2 + b\cdot\sum x_i^3 + c\cdot\sum x_i^4} \quad \ldots\ldots\ldots \text{III}$$

---

for $y = a\cdot b^x$,

$\therefore \log y = \log a + x\cdot\log b$

$\therefore V = A + B\cdot x$

$$\boxed{\sum V = n\cdot A + B\sum x} \quad \ldots\ldots \text{I}$$

$$\boxed{\sum x V = A\sum x + B\sum x^2} \quad \ldots\ldots \text{II}$$

$A = \log a \quad \therefore a = e^A \quad \& \quad B = \log b \quad \therefore b = e^B$

---

for linear, put $c = 0$ & take only first 2 equations from second degree.

Q 1: Profit (in lakhs) earned by company in 10 years is tabulated.

year	1	2	3	4	5
profit	24	27	32	38	45

Fit 2^{nd} degree curve $y = a + bx + cx^2$.

Also, estimate profit in 7^{th} year

→ take, $u = x - 3$

$\therefore$ equ^n becomes, $y = a + bu + cu^2$

Normal equ^n are $\Sigma y = na + b\Sigma u + c\Sigma u^2$ —— (I)

$\Sigma y.u = a.\Sigma u + b\Sigma u^2 + c\Sigma u^3$ —— (II)

$\Sigma y.u^2 = a\Sigma u^2 + b\Sigma u^3 + c\Sigma u^4$ —— (III)

x	y	$u=x-3$	u^2	u^3	u^4	$u.y$	$u^2.y$
1	24	−2	4	−8	16	−48	96
2	27	−1	1	−1	1	−27	27
3	32	0	0	0	0	0	0
4	38	1	1	1	1	38	38
5	45	2	4	8	16	90	180
	$\Sigma=166$	$\Sigma=0$	$\Sigma=10$	$\Sigma=0$	$\Sigma=34$	$\Sigma=53$	$\Sigma=341$

$$166 = 5a + 10c$$
$$53 = 10b$$
$$341 = 10a + 34c$$

Solving

$$\boxed{\therefore b = 5.3}$$

$$\boxed{\therefore a = 31.9142}$$

$$\boxed{\therefore c = 0.6429}$$

$$\therefore y = a + bu + cu^2$$
$$= 31.9142 + 5.3(x-3) + 0.6429(x-3)^2$$
$$= 31.9142 + 5.3x - 15.9 + 0.6429(x^2 - 6x + 9)$$
$$\therefore y = 21.8003 + 1.4426 \cdot x + 0.6429 \cdot x^2$$

$$\boxed{\therefore y = 21.8003 + 1.4426 \cdot x + 0.6429 \cdot x^2}$$

& for $x = 7$, $y = 63.4007$ lakh

Q2: Population of the state is given below.

year	1951	1961	1971	1981	1991
Population in Million	140	170	200	250	300

1) Fit $y = ab^x$ 2) Fit $y = a + bx$ 3) Obtain best fit

4) Estimate Population in year 2000

1) $y = a b^{x}$

$\ln(y) = \ln(a) + x \cdot \ln(b)$

$A = \ln(a)$, $B = \ln(b)$

$\therefore a = e^{A}$, $b = e^{B}$

$\therefore V = A + Bx$

take, $u = \dfrac{x - 1971}{10}$

equ^n becomes

$\therefore V = A + B \cdot u$ ——— (*)

Normal equ^{n}s are, $\Sigma V = n \cdot A + B \Sigma u$ ——— (I)

$\Sigma u \cdot V = A \cdot \Sigma u + B \Sigma u^2$ ——— (II)

x	y	$u = \frac{x-1971}{10}$	u^2	$V = \ln y$	$u \cdot V$
1951	140	-2	4	4.9416	-9.8832
1961	170	-1	1	5.1358	-5.1358
1971	200	0	0	5.2983	0
1981	250	1	1	5.5215	5.5215
1991	300	2	4	5.7038	11.4076
	$\Sigma = 1060$	$\Sigma = 0$	$\Sigma = 10$	$\Sigma = 26.6010$	$\Sigma = 1.9101$

$$26.6010 = 5 \cdot A$$

$$\boxed{\therefore A = 5.3202}$$

$$1.9101 = A \times 0 + B \cdot 10$$

$$\boxed{\therefore B = 0.19101}$$

$$\boxed{\therefore a = 204.4248 \quad , \quad b = 1.2105}$$

$$\therefore y = (204.4248)(1.2105)^{u}$$

$$\therefore y = (204.4248)(1.2165)^{\left[\frac{x-1971}{10}\right]} \quad \ldots\ldots \text{(III)}$$

	$\hat{y}_i$	$y_i - \hat{y}_i$	$(y_i - \hat{y}_i)^2$
Put, $x=1951$	139.5096	0.4904	0.2405
Put, $x=1961$	168.8763	1.1237	1.2627
Put, $x=1971$	204.4248	-4.4248	19.5789
Put $x=1981$	247.4562	2.5438	6.4709
Put $x=1991$	299.5458	0.4542	0.2063
		$\Sigma = 0.1873$	$\Sigma = 27.7593$

Now, We have to fit $y = a + b.x$

$$y = a + b.u \quad \text{when} \quad u = \frac{x - 1971}{10}$$

Normal equ^n are

$$\Sigma y = n.a + b.\Sigma u \quad \text{———} \quad (IV)$$

$$\Sigma u.y = a\Sigma u + b.\Sigma u^2 \quad \text{———} \quad (V)$$

x	y	$u = \frac{x-1971}{10}$	u^2	$u.y$
1951	140	-2	4	-280
1961	170	-1	1	-170
1971	200	0	0	0
1981	250	1	1	250
1991	300	2	4	1200
	$\Sigma = 1060$	$\Sigma = 0$	$\Sigma = 10$	$\Sigma = 400$

from (IV) & (V),

$$\boxed{\therefore a = 212 \quad , \quad b = 40}$$

$$\therefore y = 212 + 40.u$$

$$\therefore y = 212 + 40 \times \left(\frac{x - 1971}{10}\right)$$

$$\boxed{\therefore y = 4x - 7672}$$

$\hat{y}_i$	$y_i - \hat{y}_i$	$(y_i - \hat{y}_i)^2$
132	8	64
172	−2	4
212	−12	144
252	−2	4
292	8	64
		$\sum (y_i - \hat{y}_i)^2 = 280$

here, $(y_i - \hat{y}_i)^2$ of exponential $<$ $(y_i - \hat{y}_i)^2$ of linear

$\therefore$ Exponential fit is better than linear.

Now, for $x = 2000$, $y = 204.4248 \times (1.1205)^{\frac{2000-1971}{10}}$

$\therefore y = 355.7390$ million