Module - 5

STATISTICS

Mean → $\text{mean} = \frac{\Sigma x_i}{n} = \bar{x}$
$(\bar{x})$

Median → median is central value when data
(M) is arranged (sorted) in an order.

n is odd → $\left(\frac{n+1}{2}\right)^{th}$ observation

n is even → $\frac{\left(\frac{n}{2}\right)^{th} + \left(\frac{n+1}{2}\right)^{th}}{2}$

mean of $\left(\frac{n}{2}\right)^{th}$ and $\left(\frac{n+1}{2}\right)^{th}$

Mode → size of variables which occurs most frequently

Dispersion →

The extent / scatterness of values in the data around mean →

Measure of dispersion →

Deviation →

- Mean deviation → The mean of absolute value from deviation.

mean deviation (from central a) $= \frac{\Sigma |x_i - a|}{n}$

(about mean) $= \frac{\Sigma |x_i - \bar{x}|}{n}$

(median) $= \frac{\Sigma |x_i - M|}{n}$

$x \quad x_1 \quad x_2 \cdots x_n$
$f \quad f_1 \quad f_2 \cdots f_n$

continueous frequency

- $\bar{x}$ = mean
$$= \frac{\Sigma x_i f_i}{\Sigma f_i}$$

- Mean deviation = $\frac{\Sigma f_i |x_i - \bar{x}|}{\Sigma f_i}$

- Mean deviation about median = $\frac{\Sigma |x_i - M| \times f_i}{\Sigma f_i}$

variance →

$$\sigma^2 = \frac{1}{n} \Sigma (x_i - \bar{x})^2$$

$$\sigma^2 = \frac{1}{n} (x_i^2 - (\bar{x})^2)$$

• standard deviation →

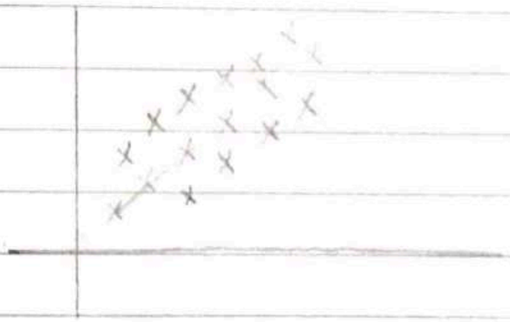
positive root of variance.
(square root)

Types of corelations.

• +ve corelation -

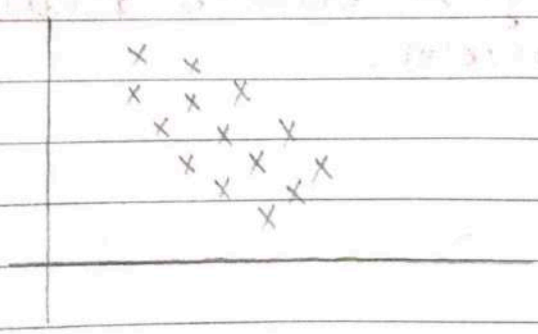
change in one variable changes value of another in same direction

eg.
no. of units and electricity bill.



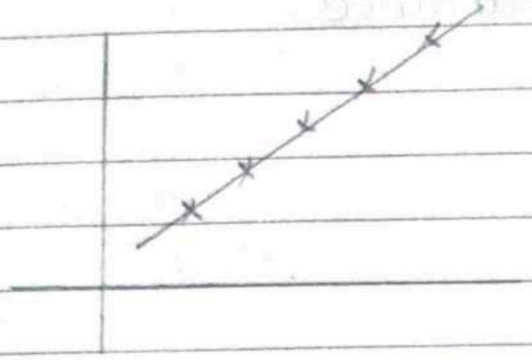
• -ve corelation -

change in one variable changes value of another in opposite direction.

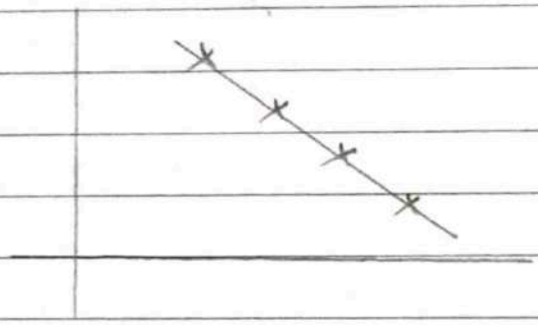


eg.
pressure - volume

• Perfectly positive corelation-



• perfectly negative corelation-



methods to find corelation-

i) Graphical method-

By plotting graph between relative values of variables.

Page No.:
Date:

ii) Algebric method -

r (coeff. of corelation)

$-1 < r < 1$ → if +ve → +ve corelation
→ if -ve → -ve corelation

$$r = \frac{\frac{\sum x_i y_i}{n} - (\bar{x})(\bar{y})}{\sigma_x \sigma_y}$$

iii) Karl pearson method

$$r = \frac{\frac{\sum x_i y_i}{n} - \bar{x}\,\bar{y}}{\sqrt{\frac{1}{n}\sum x_i^2 - (\bar{x})^2}\ \sqrt{\frac{1}{n}\sum y_i^2 - (\bar{y})^2}}$$

corelation coeff. is independent of change of origin and scale.

Que. calculate karl person's coeff. of follo. data.

x	28	45	40	38	35	33	40	32	36	33
y	23	34	33	34	30	26	28	31	36	35

→

$A = x - 33$	$B = y - 34$	AB	A^2	B^2
-5	-11	55	25	121
12	0	0	144	0
7	-1	-7	49	1
5	0	0	25	0
2	-4	-8	4	16
0	-8	0	0	64
7	-6	-42	49	36
-1	-3	3	1	9
3	2	6	9	4
0	1	0	0	1
30	-30	7	306	252

$$r = \frac{\frac{\sum AB}{n} - (\bar{A})(\bar{B})}{\sqrt{\frac{1}{n}\sum A^2 - (\bar{A})^2}\sqrt{\frac{1}{n}\sum B^2 - (\bar{B})^2}}$$ 0.5185

Page No.: Date: YOUVA

Que. A psycological test of inteligence and engg. ability of 10 students is taken. Here is the record of ungrouped data showing intelegence ratio and engg. ability ratio calculate r.

→

	IR (x)	ER	$A = x - 100$	$B = y - 100$	AB	A^2	B^2
A	105	101	5	1	5	25	1
B	104	103	4	3	12	16	9
C	102	100	2	0	0	4	0
D	101	98	1	-2	-2	1	4
E	100	95	0	-5	0	0	25
F	99	96	-1	-4	4	1	16
G	98	104	-2	4	-8	4	16
H	96	92	-4	-8	32	16	64
I	93	97	-7	-3	21	49	9
J	92	94	-8	-6	48	64	36
			-10	-20	112	180	180

corelation coeff. is independent of change of origin

$$r = \frac{\frac{\sum AB}{n} - (\bar{A})(\bar{B})}{\sqrt{\frac{1}{n}\sum A^2 - (\bar{A})^2}\sqrt{\frac{1}{n}\sum B^2 - (\bar{B})^2}} = 0.5963$$

Page No.:
Date:
YOUVA

Que 3 To find r of bivariate data following results where obtained

$n = 25$

$\Sigma x = 125$

$\Sigma y = 100$

$\Sigma x^2 = 650$

$\Sigma y^2 = 460$

$\Sigma xy = 508$

At the time of checking it was discovered that 2 pairs of (x,y), $(8,12)$, $(6,8)$ were wrongly recorded as $(6,14)$, $(8,6)$. Find correct corelation

→

correct Σx = incorrect Σx - (sum of incorrect values of x) + (sum correct values of x)

$= 125 - (8+6) + (8+6)$

correct $\Sigma x = 125$

correct Σy = incorrect Σy - (sum of incorrect values of y) + (sum of correct values of y)

$= 100 - (20) + (20)$

correct $\Sigma y = 100$

correct Σx^2 = incorrect $\Sigma x^2 - (6^2+8^2) + (8^2+6^2)$
$= 650$

correct Σy^2 = incorrect $\Sigma y^2 - (12^2+6^2) + (12^2+8^2)$
$= 436$

correct Σxy = incorrect $\Sigma xy - ((6\times14)+(8\times6))$
$+ ((6\times14)+(6\times8))$

$= 520$

$$\gamma = \frac{\frac{\Sigma xy}{n} - \bar{x}\bar{y}}{\frac{1}{n}\Sigma x^2 - (\bar{x})^2} = 0.6667$$

* Regression -

If variables are correlated then regression is method of estimating the value of one variable when the value of other variable is known.

* Linear regression -

Obtaining the best fitted line (Deviation should be minimum)

1) Regression line of y on x.
(Deviation measured along y-axis)
2) Regression line of x on y.
(Deviation measured along x-axis)

• linear Regression / Regression line-

1) Regression line y on x :

$$y = a + bx$$

$$b = \text{slope} = b_{yx} = r\frac{\sigma y}{\sigma x} = \text{regression coeff.}$$

$$a = \bar{Y} - b\bar{X}$$

$$\therefore y = (\bar{Y} - b\bar{X}) + bx$$

$$\therefore y - \bar{Y} = b(x - \bar{X}) \Rightarrow y - \bar{y} = b_{yx}(x - \bar{x})$$

2) Regression line x on y :

$$x = a + by$$

$$b = \text{slope} = b_{xy} = r\frac{\sigma x}{\sigma y} = \text{regression coeff.}$$

$$a = \bar{x} - b\bar{y}$$

$$x = (\bar{x} - b\bar{y}) + by$$

$$x - \bar{x} = b(y - \bar{y})$$

$$x - \bar{x} = b_{xy}(y - \bar{y})$$

- $b_{yx} = r \dfrac{\sigma_y}{\sigma_x}$ $\qquad$ • $b_{xy} = r \dfrac{\sigma_x}{\sigma_y}$

$$= \frac{\dfrac{\Sigma xy}{n} - \bar{x}\bar{y}}{\sigma_x \sigma_y} \cdot \frac{\sigma_y}{\sigma_x} \qquad = \frac{\dfrac{\Sigma xy}{n} - \bar{x}\bar{y}}{\sigma_x \sigma_y} \cdot \frac{\sigma_x}{\sigma_y}$$

$$= \frac{\dfrac{\Sigma xy}{n} - \bar{x}\bar{y}}{\dfrac{\Sigma x^2}{n} - (\bar{x})^2} \qquad = \frac{\dfrac{\Sigma xy}{n} - \bar{x}\bar{y}}{\dfrac{\Sigma y^2}{n} - (\bar{y})^2}$$

- $\tan\theta = \left| \dfrac{m_1 - m_2}{1 + m_1 m_2} \right|$

- slope point form $\rightarrow$ $y - y_1 = m(x - x_1)$

- $y = a + bx$
  $x = a + by$

- $b_{yx} = r \dfrac{\sigma_y}{\sigma_x}$ $\qquad$ $b_{xy} = r \dfrac{\sigma_x}{\sigma_y}$

- $b_{xy} \cdot b_{yx} = r^2$

  $r = \sqrt{b_{xy} \cdot b_{yx}}$

$$r = \frac{\frac{\Sigma AB}{n} - \bar{A}\bar{B}}{\sqrt{\frac{\Sigma A^2}{n} - (\bar{A})^2}\ \sqrt{\frac{\Sigma B^2}{n} - (\bar{B})^2}}$$

byx or bxy → change of origin

$$byx = \frac{\frac{\Sigma AB}{n} - \bar{A}\bar{B}}{\frac{\Sigma A^2}{n} - (\bar{A})^2}$$

Regression → y on x

$$y - \bar{y} = b(x - \bar{x})$$

$$B - \bar{B} = byx(A - \bar{A})$$

- Interpretation of Regression coeff.

y on x

$$y - \bar{y} = byx(x - \bar{x}) \quad \text{or} \quad y = a + byx\, x$$

M T W T F S S	
Page No.:	YOUVA
Date:	

Que. Given

$x - 4y = 5$

$x - 16y = -64$ are regression lines.

1) Find regression coeff. for regration y on x and x on y
2) r
3) $\bar{X}, \bar{Y}$
4) $6y$, if $6x = 8$

→ by looking at the eqn we cant decide which of the eqn is regression line of y on x or x on y [consider line of y on x and 2nd is x on y]

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

i.e. $b_{yx} = \frac{1}{4}$

$$x - 16y = -64$$
$$x = 16y - 64$$

Here, $b_{xy} = 16$

our assumption is wrong

eqn ① is regression line of x on y

eqn ② is regression line of y on x

$x - 4y = 5$

$x = 4y + 5$

i.e. $b_{xy} = 4$

$x - 16y = -64$

$16y = x + 64$

$y = \frac{x}{16} + \frac{64}{16}$

i.e $b_{yx} = 1/16$

$r = \sqrt{byx \times bxy} = \sqrt{4 \times \frac{1}{16}} = 0.5$ $\left[\because byx \text{ and } bxy \text{ both +ve}\right]$

Two regression line intersect in $\bar{x}$ and $\bar{y}$

$$\begin{aligned} x - 4y &= 5 \\ \ominus\, x \,\bar{\oplus}\, 16y &= \bar{\oplus}\, 64 \\ \hline 12y &= 69 \\ y &= \frac{69}{12} = \frac{23}{4} = \bar{y} \end{aligned}$$

$$\begin{aligned} x &= 4y + 5 \\ &= 4 \times \frac{23}{4} + 5 \\ x &= 28 \\ \bar{x} &= 28 \end{aligned}$$

$byx = r \frac{6y}{6x}$

$\Rightarrow \frac{1}{16} = \frac{1}{2} \times \frac{6y}{8} \Rightarrow 6y = 1$

Que. Height of fater and son are given below

Ht of father	150	152	155	157	160	161	164	166
Ht of son	154	156	158	159	160	162	161	164

1) find the equation of 2 lines of regression
2) calculate height of son when height of father is 154 cm.
3) 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
		⇓	⇓	⇓	⇓	⇓
		-15	-6	135	251	78

$\bar{A} = \frac{\Sigma A}{n} = \frac{-15}{8}$

$\bar{A} = -1.8750$

$\bar{B} = \frac{\Sigma B}{n} = \frac{-6}{8}$

$\bar{B} = -0.75$

Regression is independent of change of origin

$$byx = \frac{\frac{\Sigma AB}{n} - \bar{A}\bar{B}}{\frac{\Sigma A^2}{n} - (\bar{A})^2} = 0.5552$$

$$bxy = \frac{\frac{\Sigma AB}{n} - \bar{A}\bar{B}}{\frac{\Sigma B^2}{n} - (\bar{B})^2} = 1.6837$$

Regression line of y on x is

$$B - \bar{B} = byx(A - \bar{A})$$
$$B - (-0.75) = 0.5552(A - (-1.8750))$$
$$B + 0.75 = 0.5552(A + 1.8750)$$
$$y - 160 + 0.75 = 0.5552(x - 160 + 1.8750)$$

$$y = 0.5552x + 71.459 \quad \text{— (I)}$$

Regression line of x on y is

$$A - \bar{A} = bxy(B - \bar{B})$$

$$(x - 160) + 1.8750 = 1.6837[(y - 160) + 0.75]$$

$$x = 1.6837y - 110.0042 \quad \text{— (II)}$$

from equation (I) (To find value of y always use equatio no. (I))

$$y = 0.5552x + 71.459$$
$$y = 0.5552(154) + 71.459$$
$$y = 156.9598 \approx 157\,cm$$

$\therefore$ Height of son is 157 cm

Page No.: Date: YOUVA

Now, $r = \sqrt{byx \; bxy} = \sqrt{(0.5552)(1.6837)}$
$= 0.9668$

Que. The following results were obtained from marks in APM and Maths.

	Marks in APM	Marks in Math's
Mean	47.5	39.5
S D	16.8	10.8

Here,
$r = 0.95$
Find both regression equation and estimate y when $x = 30$

→

Given : $\bar{x} = 47.5$ $\quad \bar{Y} = 39.5$ $\quad r = 0.95$
$\sigma x = 16.8$ $\quad \sigma y = 10.8$

$$byx = r \frac{\sigma y}{\sigma x} = 0.6102$$

$$bxy = r \frac{\sigma x}{\sigma y} = 1.4778$$

Now Regression line of y on x

$$y - \bar{y} = byx(x - \bar{x})$$

$$y - 39.5 = 0.6102(x - 47.5)$$
$$y = 0.6107x + 10.4918 \quad \text{— ①}$$

Regression lin of x on

$$x - \bar{x} = bxy(y - \bar{y})$$
$$x = 1.4778y - 10.8731 \quad \text{— ②}$$

Page No.:
Date:

from ①

$y = 0.6102(30) + 10.4918$

$y = 28.8128$

Que. If θ is angle between two regression lines show that

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

explain significance when $r=0$, $r=\pm 1$

→ The regression line of y on x is

$$y - \bar{y} = b_{yx}(x - \bar{x})$$

Here, $m_1 = b_{yx}$

Now regression line of x on y is

$$x - \bar{x} = b_{xy}(y - \bar{y})$$

$$y - \bar{y} = \frac{1}{b_{xy}}(x - \bar{x})$$

Here, $m_2 = \frac{1}{b_{xy}}$

we know that $\tan\theta = \left|\frac{m_1 - m_2}{1 + m_1 m_2}\right|$

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

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

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

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

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

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

$$\theta = \frac{\pi}{2}$$

Two lines are perpendicular to each other

If $r = \pm 1$

$$\theta = 0^\circ$$

Two lines are parallel to each other

Page No.:
Date:

- Fitting of curve / non-linear regression model:
  value of $r = 0$

- fitting of second degree curve -

  suppose $\{(x, y), \ldots\}$
  are observations

  $y =$ dependent, $x =$ independend

  general eqn of 2nd degree curve

  $y = a + bx + cx^2$

  error/ deviatio
  $y_i - \hat{y}_i$

  $y_i \rightarrow$ observed value

  $\hat{y}_i \rightarrow$ estimated value

  $\hat{y}_i = a + bx_i + cx_i^2$

- sum of square of deviation $\rightarrow$

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

  $S = \Sigma (y_i - a - bx_i - cx_i^2)$ —— ⊛

  $\frac{\partial S}{\partial a} = 0$, $\frac{\partial S}{\partial b} = 0$, $\frac{\partial S}{\partial c} = 0$

$$\frac{\partial s}{\partial a} = \frac{\partial}{\partial a} \Sigma (y_i - a - bx_i - cx_i^2)^2$$

$$\frac{\partial s}{\partial a} = \Sigma 2 (y_i - a - bx_i - cx_i^2)(-1) = 0$$

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

$$\Sigma y_i - \Sigma a - \Sigma bx_i - \Sigma cx_i^2 = 0$$

$$\Sigma y_i = na + b\Sigma x_i + c\Sigma x_i^2 \quad - \text{①}$$

$$\frac{\partial s}{\partial b} = \frac{\partial}{\partial b} \Sigma (y_i - a - bx_i - cx_i^2)^2$$

$$= \Sigma 2 (y_i - a - bx_i - (x_i^2)(-x_i) = 0$$

$$(\Sigma y_i - \Sigma a - \Sigma bx_i - \Sigma cx_i^2)(x_i) = 0$$

$$\Sigma x_i y_i = a\Sigma x_i + b\Sigma x_i^2 + c\Sigma x_i^3 \quad - \text{②}$$

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

equation becomes

$$\Sigma x_i^2 y_i = a\Sigma x_i^2 + b\Sigma x_i^3 + c\Sigma x_i^4 \quad - \text{③}$$

- if question is in format

$$y = a + bx \rightarrow \text{linear}$$

$$S = \Sigma (y_i - a - bx_i)^2$$

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

$$\Sigma y = na + b\Sigma x$$

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

- if curve is of exponential type →

$$y = ab^x$$

$$\log y = \log a + \log b^x$$

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

$$V = A + Bx$$

$$\Sigma V = nA + B\Sigma x$$

$$\Sigma xV = A\Sigma x + B\Sigma x^2$$

$(A = \log a$

$a = e^A$

$B = \log b$

$b = e^b)$

Que The profit (in lakhs) earned by company in 5 years is

year (x)	1	2	3	4	5
profit (y)	24	27	32	38	45

Fit 2^{nd} degree curve $y = a + bx + cx^2$
Also estimate profit in 7th year.

→

Take $u = x - 3$

so equation of curve becomes

$$y = a + bu + cu^2$$

normal equations →

$$\Sigma y = na + b\Sigma u + c\Sigma u^2 \quad -①$$
$$\Sigma uy = a\Sigma u + b\Sigma u^2 + c\Sigma u^3 \quad -②$$
$$\Sigma u^2y = a\Sigma u^2 + b\Sigma u^3 + c\Sigma u^4 \quad -③$$

x	y	$u = x-3$	u^2	u^3	u^4	uy	u^2y
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
	166	0	10	0	34	53	341

From ①, ② and ③

$166 = 5a + 10c$

$53 = 10b \Rightarrow b \Rightarrow 5.3$

$341 = 10a + 34c$

$$\begin{array}{r} 10a + 20c = 332 \\ - \; 10a + 34c = 341 \\ \hline 0 - 14c = -9 \end{array}$$

$c = 0.6429$

$$a = \frac{166 - 10c}{5}$$

$a = 31.9143$

$y = a + bu + cu^2$

$= 31.9143 + 5.3u + 0.6429\,u^2$

$= 31.9143 + 5.3(x-3) + 0.6429(x-3)^2$

$= 31.9143 + 5.3x - 15.9 + 0.6429x^2 + 5.7861 - 3.8574x$

$y = 21.8003 + (1.4426)x + (0.6429)x^2$

put $x = 7$

$y = 63.4007$ lakhs

Que. The population of the state is given below →

year (x)	1951	1961	1971	1981	1991
(y) (population in million)	140	170	200	250	300

→ 1) fit the curve $y = ab^x$

2) fit $y = a + bx$

3) obtain best fit

4) estimate population in year 2020

→

Take $u = \dfrac{x - 1971}{10}$

$$y = ab^x$$

$$\log_e y = \log_e a + x \log_e b$$

$$V = A + xB \quad \Rightarrow \text{where } A = \log a$$

$$B = \log b$$

$$V = A + uB \qquad V = \log y$$

x	y	u	u^2	$V = \log_e y$	uV
1951	140	-2	4	4.9416	
1961	170	-1	1	5.1358	
1971	200	0	0	5.2983	
1981	250	+1	1	5.5215	
1991	300	+2	4	5.7038	
		0	10	26.6010	1.9101

eqns → $\Sigma V = nA + B\Sigma u$ — (1)

$\Sigma uV = A\Sigma u + B\Sigma u^2$ — (2)

From ① and ②

$26.6010 = 5A$

$A = 5.3202$

and $B = 0.19101$

$a = e^{A} = 204.4248$

$b = e^{B} = 1.2105$

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

$= (204.4248)(1.2105)^{\left[\frac{x-1971}{10}\right]}$ — ③

$\hat{y}_i$	$y_i - \hat{y}_i$	$(y_i - \hat{y}_i)^2$
139.5096	140 − 139.5096 = 0.4904	0.2405
168.8763	1.1237	1.2627
204.4248	−4.4248	+19.5789
247.4562	2.5438	6.4702
299.5458	0.4542	0.2063
		$\Sigma = 27.7593$

we have to fit

$y = a + bx$

$y = a + bu$ where $u = \frac{x-1971}{10}$

eqn are

$\Sigma y = na + b\Sigma u$

$\Sigma uy = a\Sigma u + b\Sigma u^2$

x	y	u	u^2	uy
1951	140	-2	4	-280
1961	170	-1	1	-170
1971	200	0	0	0
1981	250	1	1	250
1991	300	2	4	600
	1060	0	10	400

$1060 = 5a$

$a = 212$

$b = 40$

$y = 212 + 40 \times u$

$y = 212 + 40 \times \dfrac{x - 1971}{10}$

$y = 4x - 7672$

$\hat{y}_i$	$y_i - \hat{y}_i$	$(y_i - \hat{y}_i)^2$
132	$140 - 132 = 8$	64
172	-2	4
212	-12	144
252	-2	4
292	8	64
	0	Σ 280

Here $(y_i - \hat{y}_i)^2_{\text{exponential}} < (y_i - \hat{y}_i)^2_{\text{linear}}$

exponential is better than linear

put $x = 2000$ in eqn. 3

$\therefore y = 355.7390$ million