L- 9: Predictive Analytics & Revision

innovate achieve lead

Agenda

- Review of last session
- Introduction to regression
- Method of least squares
- Simple linear regression

Covariance of X and Y

$$E(x) = \sum x\, p(x)$$

$$= \int x f(x)\, dx$$

$$Cov(X,Y) =$$

$$= \left[E\left(x-\mu_x\right)\left(\gamma-\mu_y\right) \right]$$

joint p.d.f

$$= \sum_x \sum_y (x-\mu_x)(y-\mu_y)\, P(x,y)$$

if discrete

joint prob. density fn

$$= \iint (x-\mu_x)(y-\mu_y)\, f(x,y)\, dx\, dy$$

if continuous

$$Cov(x, y) = \frac{\sum_{i=1}^{n} (x - \mu_x)(y - \mu_y)}{n - 1}$$

Correlation

X → income

Y → expenditure

$n - 1$

n

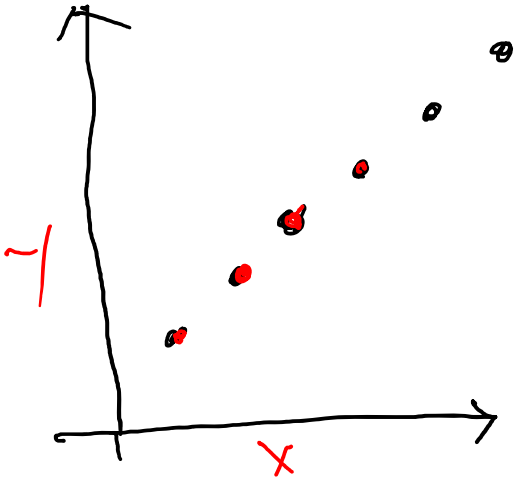
$(n-1)$

And also

$\Rightarrow$ Farmer has an impression that if he uses more fertilizers, then the crop yield increases.

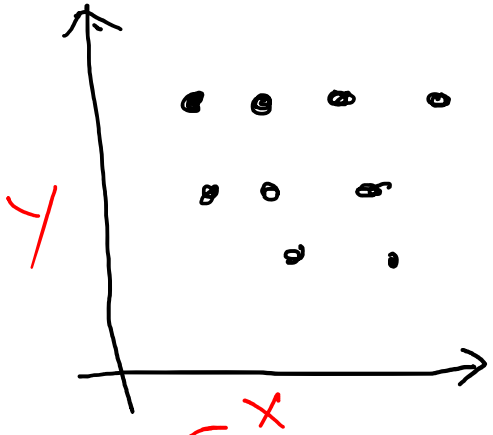
We need to validate this?

How $\longrightarrow$?



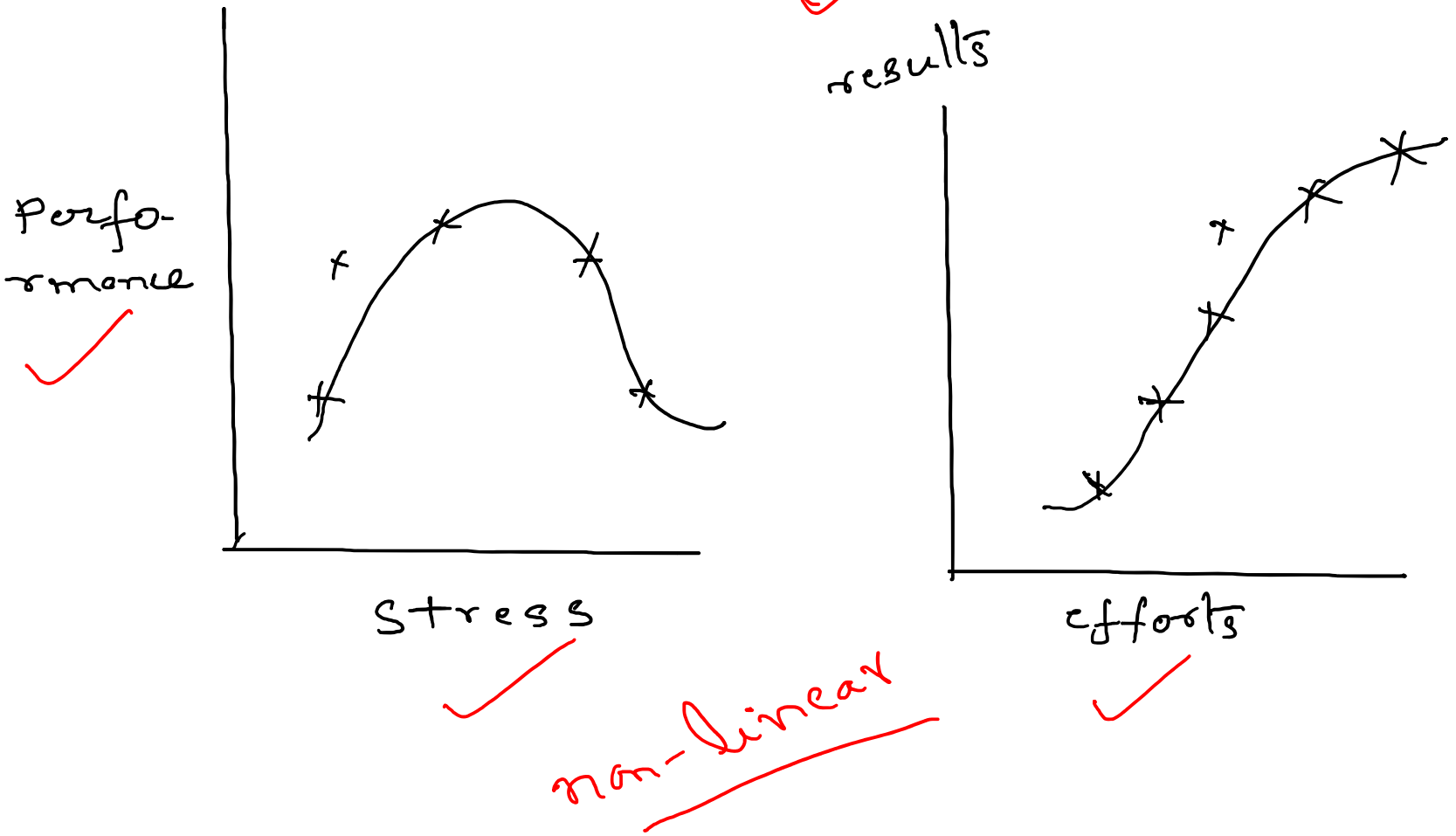
positive correlation

Negative correlation



No correlation

Linear


Perfo-
rmance
Stress
results
efforts
non-linear

Coefficient of correlation:

$$r = \frac{cov(x,y)}{\sigma_x \sigma_y} = \frac{\Sigma XY}{\sqrt{\Sigma x^2 \cdot \Sigma y^2}}$$

where $X = x - \bar{x}$

$Y = y - \bar{y}$

$x^2 = (x - \bar{x})^2$

$y^2 = (y - \bar{y})^2$

Coefficient of Correlation

$r = 1 \Rightarrow$ Perfect and positive relation

$r = -1 \Rightarrow$ " " negative relation

$r = 0 \Rightarrow$ No relation

$0 < r < 1 \Rightarrow$ Partial positive relation

$-1 < r < 0 \Rightarrow$ " negative "

$-1 \leq r \leq 1$

Example - 1

x	1	2	3	4	5	6	7	8	9
y	10	11	12	14	13	15	16	17	18

$$\bar{x} = \frac{\Sigma x}{n} = \frac{45}{9} = 5$$

$$\bar{y} = \frac{\Sigma y}{n} = \frac{126}{9} = 14$$

x	$X = x-5$	X^2	y	$Y = y-14$	Y^2	XY
1	−4	16	10	−4	16	16
2	−3	9	11	−3	9	9
3	−2	4	12	−2	4	4
4	−1	1	14	0	0	0
5	0	0	13	−1	1	0
6	1	1	15	1	1	1
7	2	4	16	2	4	4
8	3	9	17	3	9	9
9	4	16	18	4	16	16
		60			60	59

$$r = \frac{\Sigma XY}{\sqrt{\Sigma X^2 \Sigma Y^2}}$$

$$= \frac{59}{\sqrt{60 \times 60}}$$

$$= 0.9833$$

$r = 0.9833$

→ Corrected

→ +ve correlated

x	$X = x-5$	X^2	y	$Y = y-14$	Y^2	XY
1	−4	16	10	−4	16	16
2	−3	9	11	−3	9	9
3	−2	4	12	−2	4	4
4	−1	1	14	0	0	0
5	0	0	13	−1	1	0
6	1	1	15	1	1	1
7	2	4	16	2	4	4
8	3	9	17	3	9	9
9	4	16	18	4	16	16
		60				59

$$\therefore \quad \text{cov}(x,y) = \frac{\Sigma XY}{n-1} = \frac{59}{8} = 7.375$$

Coefficient of Determination

innovate achieve lead

r is coeff. of Correlation

r^2 is coeff of determination

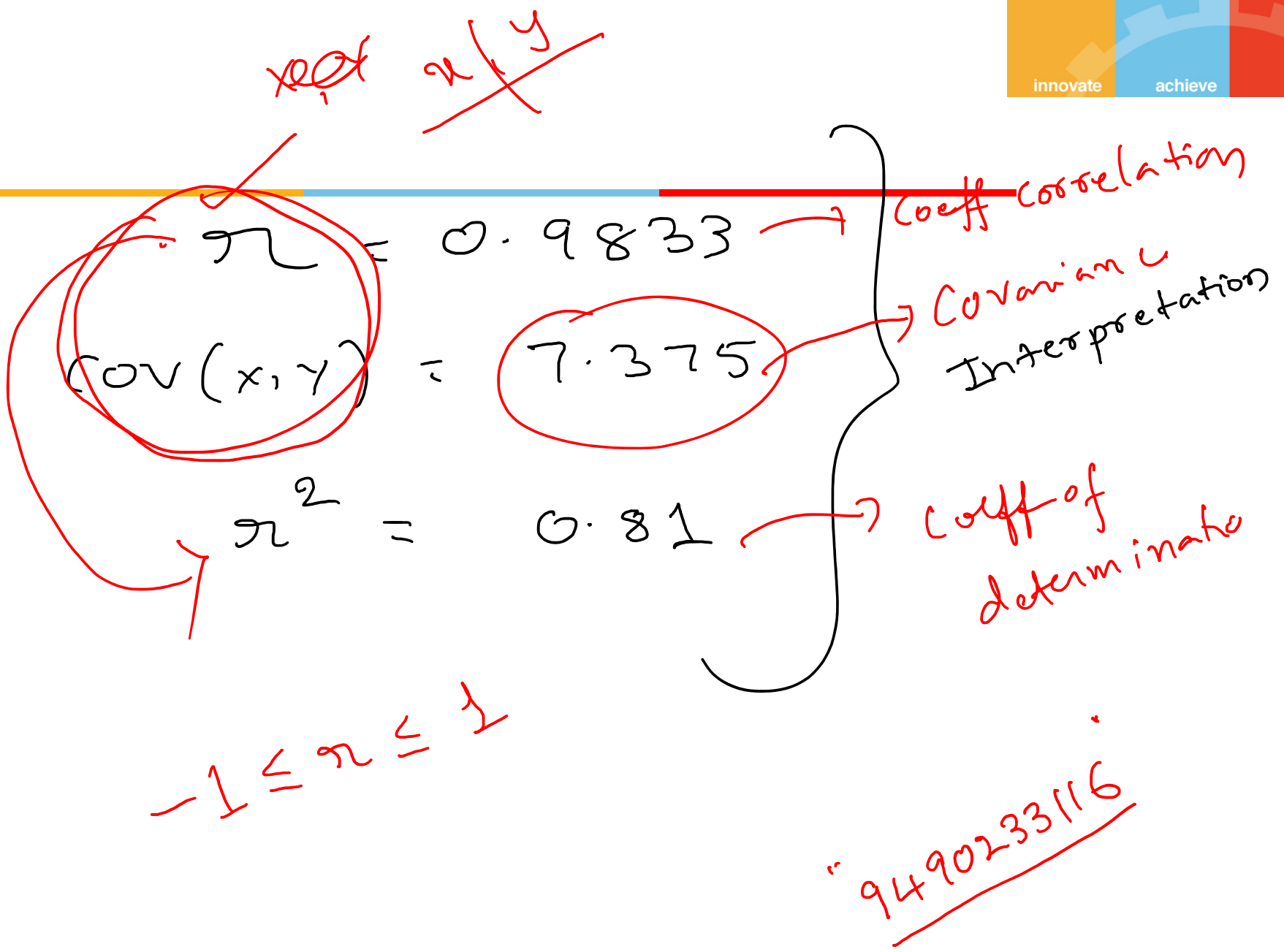
↓

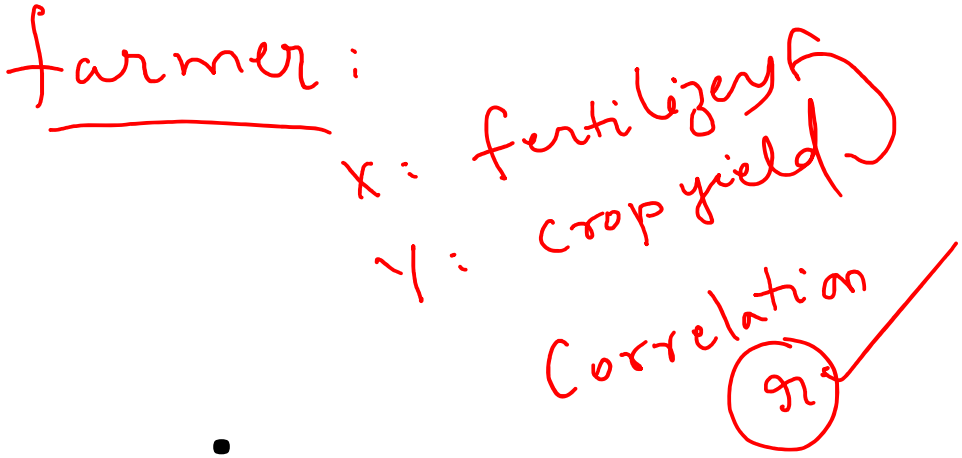
Indicates the extent to which variation in one variable is explained by the variation in the other.

$r = 0.9 \Rightarrow r^2 = 0.81$

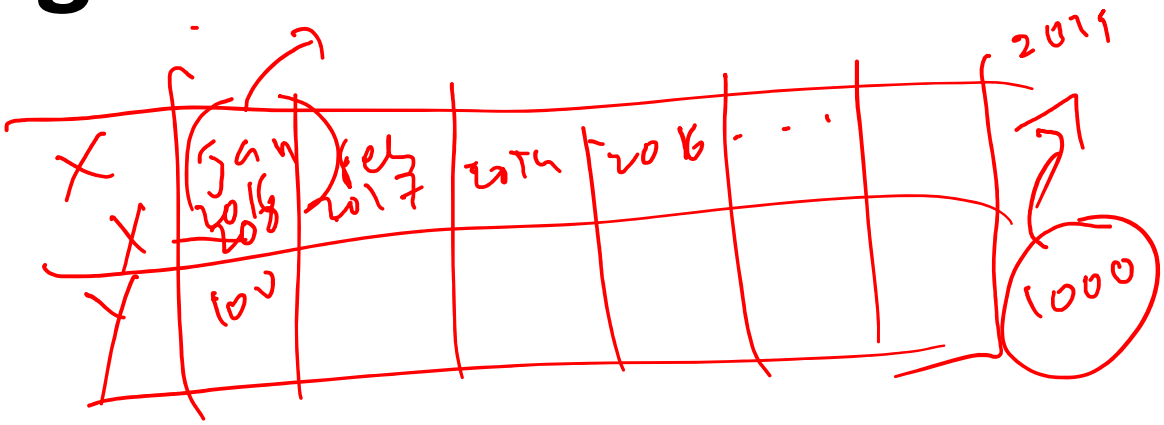
i.e 81% of the variation in y due to variation in 'x'

remaining 19% is due to some other factors.


xeq x/y
r = 0.9833
Coeff correlation
Cov(x,y) = 7.375
Covariance
Interpretation
r² = 0.81
coeff of determinate
-1 ≤ r ≤ 1
" 9490233116


farmer :
x : fertilizer
y : crop yield
Correlation
92

Regression


X
Y
Jan 2016
Feb 2017
2018
100
1000

Regression :-

Correlation (r)

x	1	2	3	4	5
y	1	4	9	16	25

$y = x^2$ when $x = 7$; $y = 49$

when $x = 7$: $y = ?$

5.4 5.6 5.8

x	1	2	3	4	5
y	1	6	2	5	4

when $x = 7$, $y = ?$

$y = 2 \ldots$

Correlation

→ Measuring strength or degree of the relationship between two variables

→ no estimation

→ both variables are independent

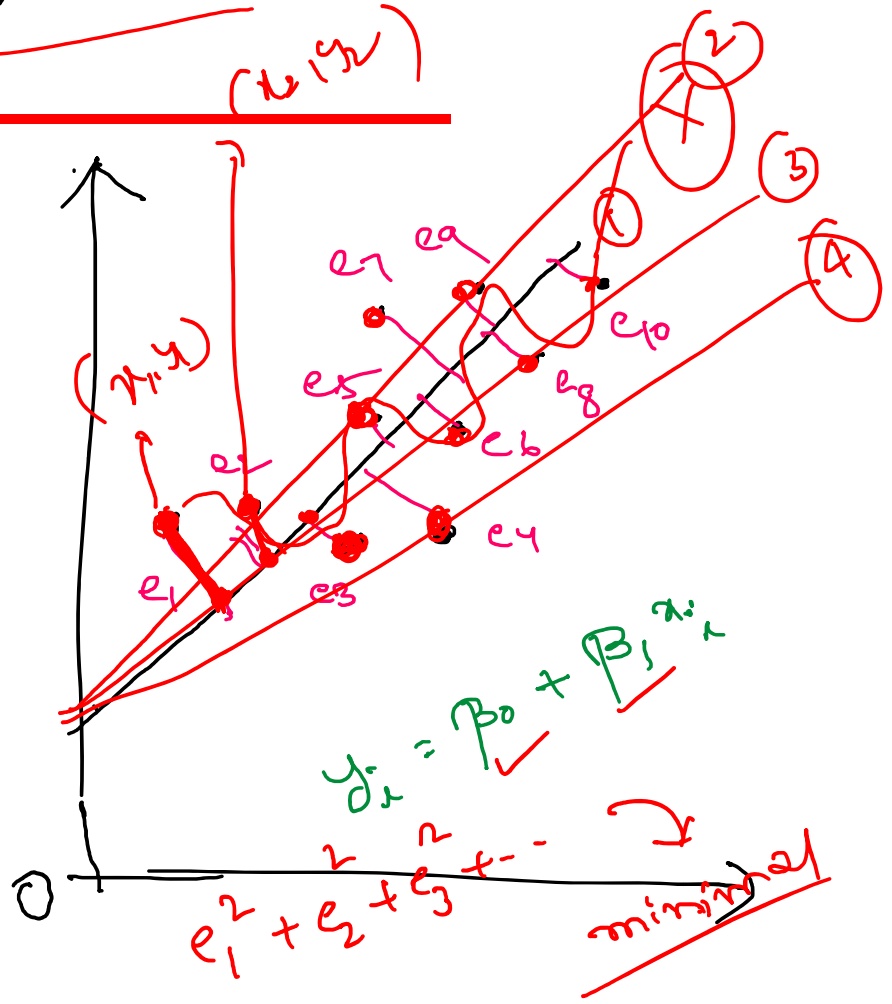
Regression

→ Having an algebraic equation between two variables

→ estimation

→ one is dept variable and other indepn Variables

Method of Least squares



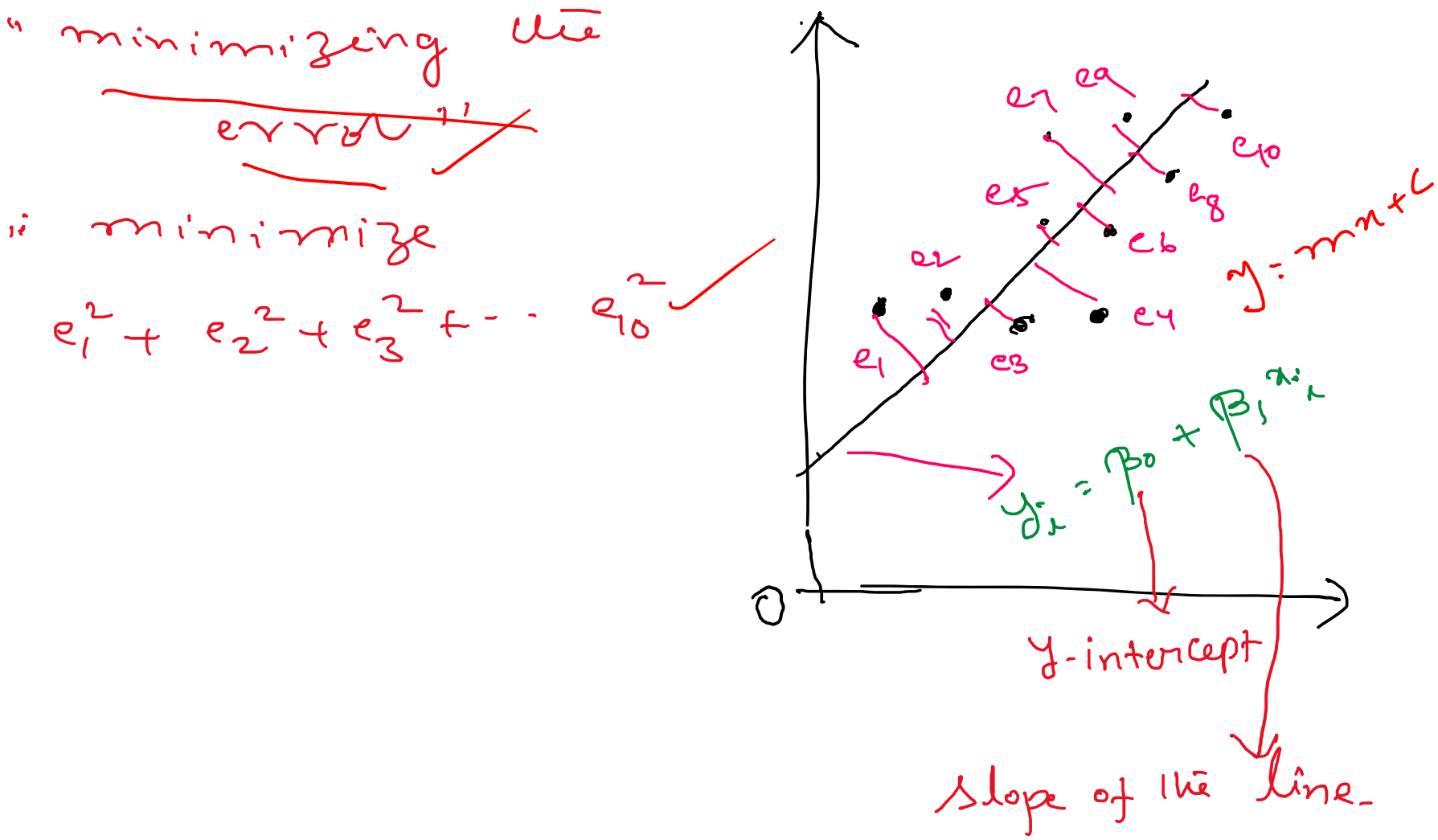
y : Dependent Variable

x : Independent Variable

predictor variable

response Variable

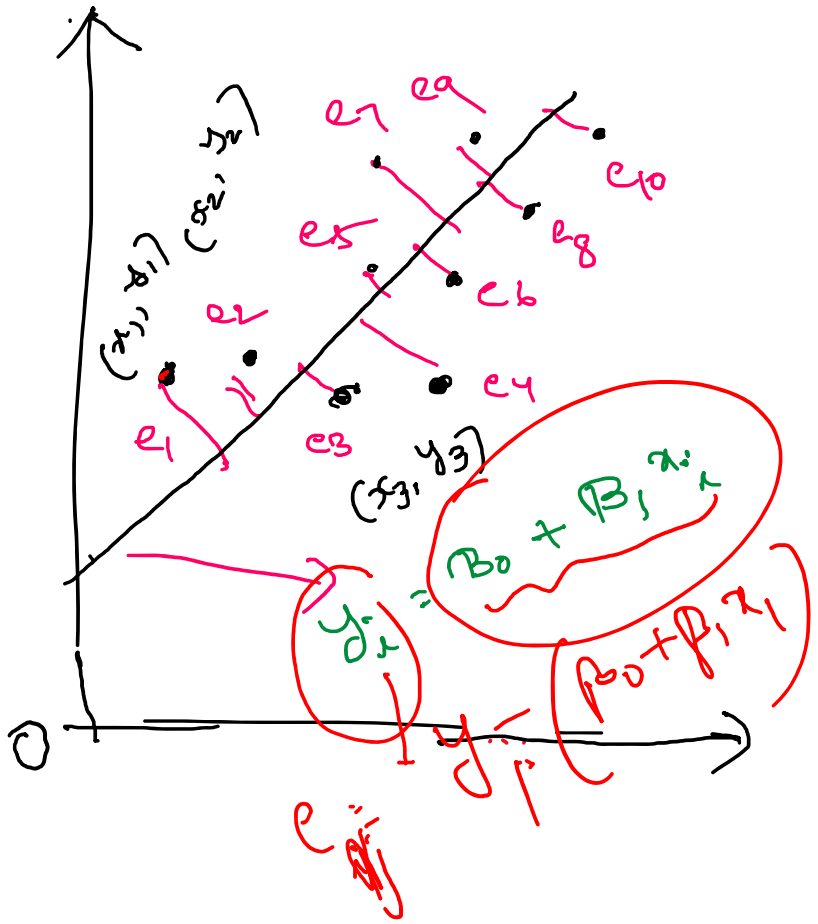
Method of Least squares



Method of Least squares

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

we need to choose β_0 and β_1 which minimizes the error.



Method of Least squares

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

$$\frac{\partial S}{\partial \beta_0} = 0 \Rightarrow 2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)(-1)$$

$$\Rightarrow \sum_{i=1}^{n} y_i = n\beta_0 + \beta_1 \sum_{i=1}^{n} x_i$$

$$\frac{\partial S}{\partial \beta_1} = 0 \Rightarrow 2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)(2)(-x_i)$$

$$\Rightarrow \sum_{i=1}^{n} x_i y_i = \beta_0 \sum x_i + \beta_1 \sum x_i^2$$

on solving these, we get β_0 & β_1 which minimizes error.

Linear regression

$$y = \beta_0 + \beta_1 x$$

$$\Sigma y = \beta_0 n + \beta_1 \Sigma x$$

$$\Sigma xy = \beta_0 \Sigma x + \beta_1 \Sigma x^2$$

Normal equations. β_0, β_1

$$y = \beta_0 + \beta_1 x$$

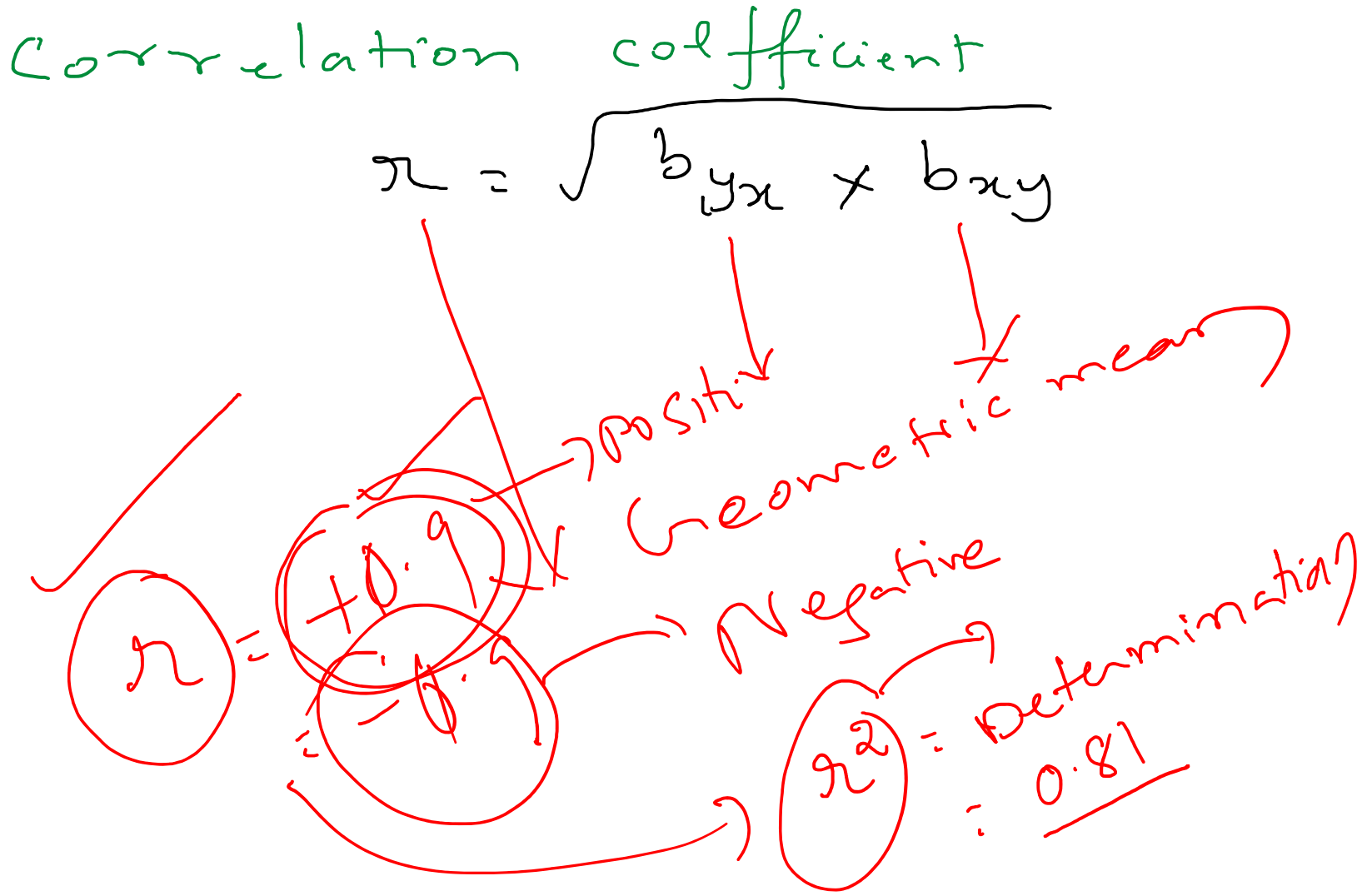
Regression Coefficients

$y = a + bx$ — regression line of y on x

b_{yx} : Regression coeff of y on x

$x = c + dy$ — regression line of x on y

b_{xy} : regression coeff of ~~y on x~~ x on y


Correlation coefficient
$r = \sqrt{b_{yx} \times b_{xy}}$
Positive
Geometric mean
Negative
r = +0.9 (or) -0.9
r^2 = Determination
= 0.81

Example :-

company	Advt Expt	Sales Revenue
A	1	1
B	3	2
C	4	2
D	6	4
E	8	6
F	9	8
G	11	8
H	14	9

$$y = a + bx$$

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

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

Example :-

Sales Revenue y	Advt. expt. x	x^2	xy
1	1	1	1
2	3	9	6
2	4	16	8
4	6	36	24
6	8	64	48
8	9	81	72
8	11	121	88
9	14	196	126
$\Sigma 40$	$\Sigma 56$	$\Sigma 524$	$\Sigma 373$

$y = \beta_0 + \beta_1 x$

$\Sigma y = n\beta_0 + \beta_1 \Sigma x$

$\Sigma xy = \beta_0 \Sigma x + \beta_1 \Sigma x^2$

$\Rightarrow 40 = 8\beta_0 + 56\beta_1$

$373 = 56\beta_0 + 524\beta_1$

on solving

$\beta_0 = 0.072$

$\beta_1 = 0.704$

i.e. $y = (0.072) + (0.704)x$

i.e. $y = (0.072) + (0.704)x$

when $x = 0.075$, then

$y = (0.072) + (0.704)(0.075)$

$= 0.1248 \simeq 12.48\%$

innovate achieve lead

Example:

Consider the following data

x	1	2	4	0
y	0.5	1	2	0

Fit a linear regression line

Estimate y when $x = 5$.

x	y	xy	x^2
1	0.5	0.5	1
2	1	2	4
4	2	8	16
0	0	0	0
$\Sigma = 7$	$\Sigma 3.5$	$\Sigma 10.5$	$\Sigma 21$

$$y = \beta_0 + \beta_1 x$$

$$\Sigma y = n\beta_0 + \beta_1 \Sigma x$$

$$\Sigma xy = \beta_0 \Sigma x_1 + \beta_1 \Sigma x^2$$

$$3.5 = 4\beta_0 + \beta_1 (7)$$

$$10.5 = 7\beta_0 + \beta_1 (21)$$

on solving these

$$\beta_0 = 0$$

$$\beta_1 = 0.5$$

i.e $y = 0 + (0.5)x$

When $x = 5$, $y = (0.5)5$

$= 0.25$

Linear regression

(Multiple regression)

Example:-

No	Size	No of rooms	No of floors	Age of home	Price Lakhs
1	2000	5	2	45	4000
1	1400	3	1	40	2000
1	1600	3	2	30	3000
1	800	2	1	35	2000

x_1 x_2 x_3 x_4 y

House → Price

y → n

increment my part

20 years
1 floor
2 rooms
1200 sqft

$y =$

Multiple Linear Regression

$y = \beta_0 + \beta_1 x$ → $\Sigma y = \beta_0 n + \beta_1 \Sigma x$, $\Sigma xy = \beta_0 \Sigma x + \beta_1 \Sigma x^2$

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$$

$$\Sigma y = \beta_0 n + \beta_1 \Sigma x_1 + \beta_2 \Sigma x_2 + \beta_3 \Sigma x_3 + \beta_4 \Sigma x_4$$

$$\Sigma x_1 y = \beta_0 \Sigma x_1 + \beta_1 \Sigma x_1^2 + \beta_2 \Sigma x_1 x_2 + \beta_3 \Sigma x_1 x_3 + \beta_4 \Sigma x_1 x_4$$

$$\Sigma x_2 y = \beta_0 \Sigma x_2 + \beta_1 \Sigma x_1 x_2 + \beta_2 \Sigma x_2^2 + \beta_3 \Sigma x_2 x_3 + \beta_4 \Sigma x_2 x_4$$

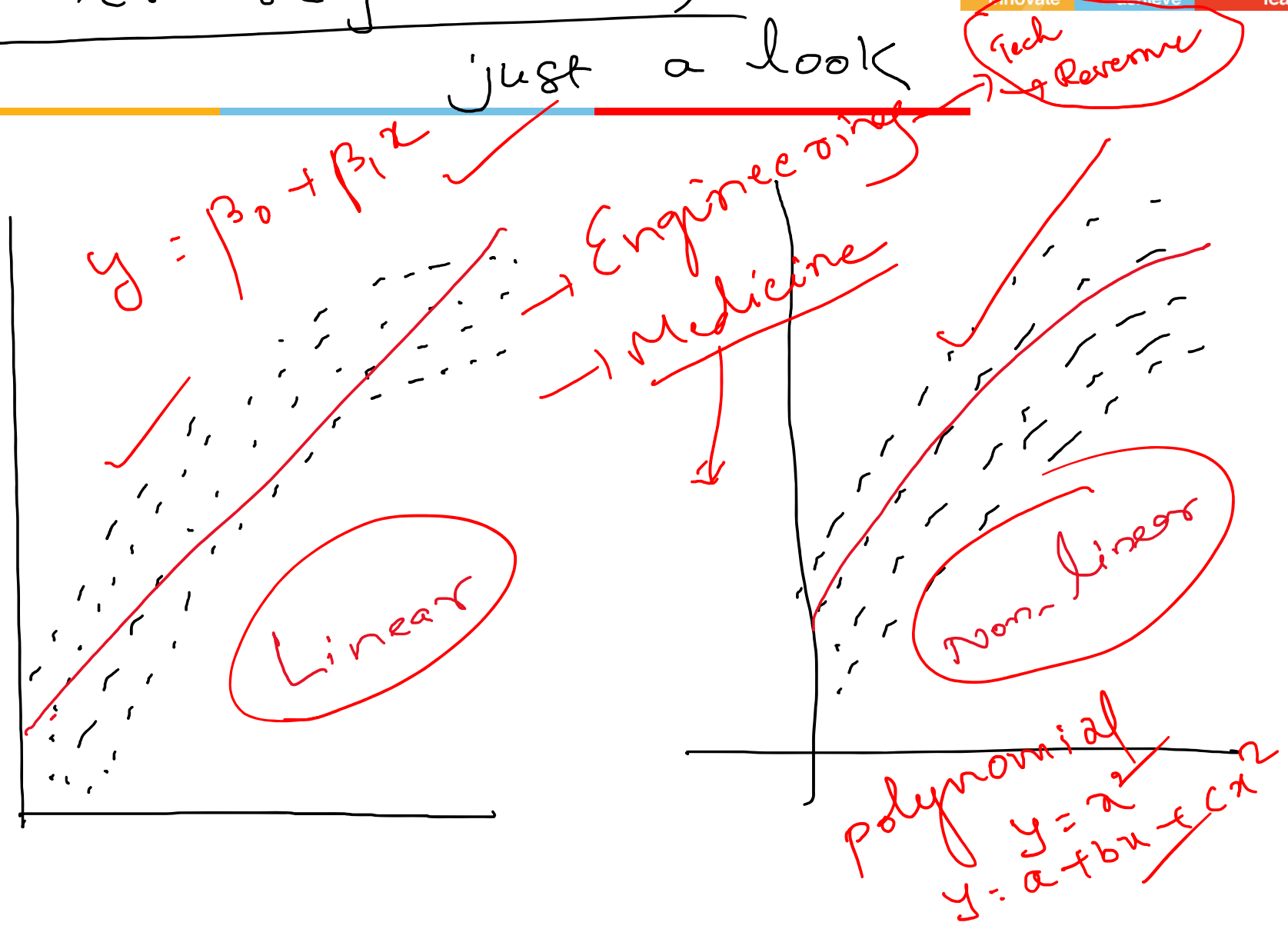
$$\Sigma x_3 y = \beta_0 \Sigma x_3 + \beta_1 \Sigma x_1 x_3 + \beta_2 \Sigma x_2 x_3 + \beta_3 \Sigma x_3^2 + \beta_4 \Sigma x_3 x_4$$

$$\Sigma x_4 y = \beta_0 \Sigma x_4 + \beta_1 \Sigma x_1 x_4 + \beta_2 \Sigma x_2 x_4 + \beta_3 \Sigma x_3 x_4 + \beta_4 \Sigma x_4^2$$

Solve $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4$

Other regressions

just a look



Suppose $y = ae^{bx}$ → exponential curve

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

$Y \quad A \quad X$

i.e. $Y = A + bX$ → non-linear eqn → linear eqn reg

$\sum Y = An + b \sum X \rightarrow (1)$

$\sum XY = A \sum X + b \sum X^2 \rightarrow (2)$

$A = ?$ → a

b

Hence, we get $y = ae^{bx}$

Suppose $y = ax^b$ → Non Linear → Power Curve

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

$Y = A + bX$

$\sum Y = An + b\sum X$

$\sum XY = A\sum X + b\sum X^2$

$A \rightarrow a$

$b \rightarrow b$

$y = ax^b$

Matrix Approach:

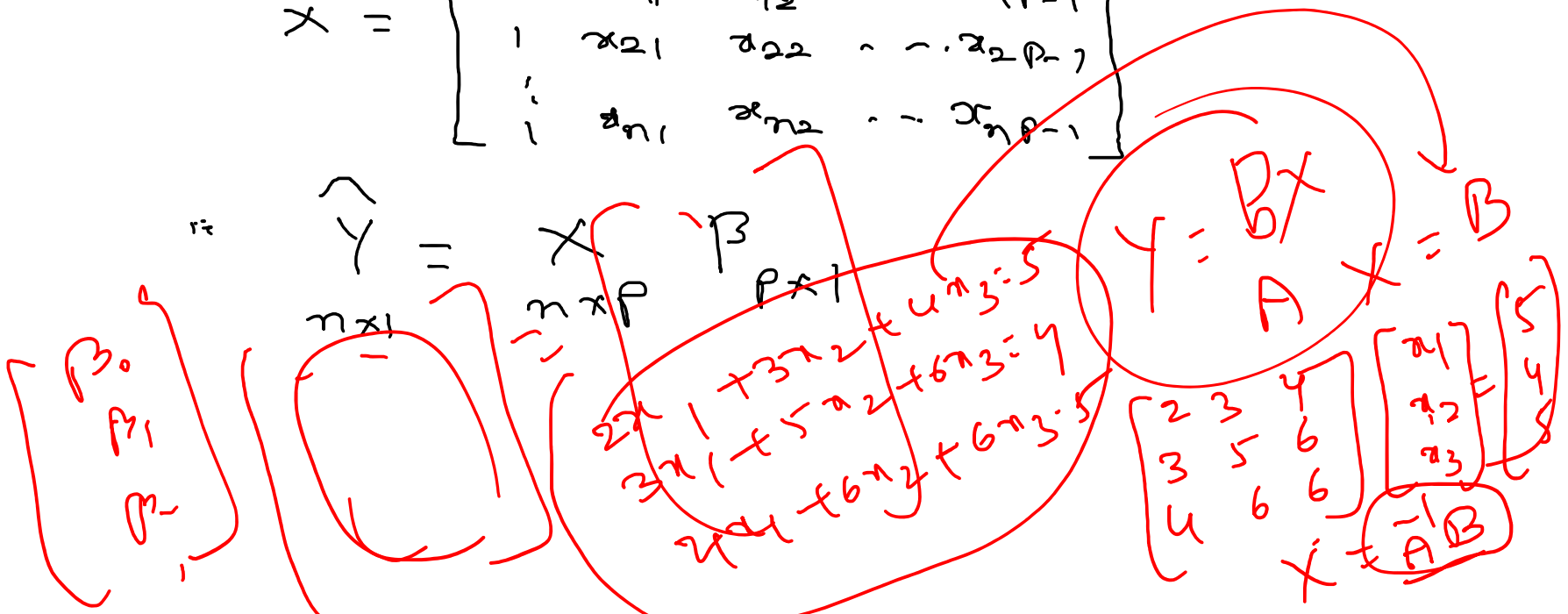
Let $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_{p-1} x_{p-1}$

observations $y_i = 1, 2, \ldots n$ → by a vector Y

unknowns $\beta_0, \beta_1 \cdots \beta_{p-1}$ → " " β

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1,p-1} \\ 1 & x_{21} & x_{22} & \cdots & x_{2,p-1} \\ \vdots & & & & \\ 1 & x_{n1} & x_{n2} & \cdots & x_{n,p-1} \end{bmatrix}$$

i.e. $\underset{n\times 1}{\hat{Y}} = \underset{n\times p}{X} \; \underset{p\times 1}{\beta}$



Find β to minimize

$$S(\beta) = \sum_{i=1}^{m} \left(y_i - \beta_0 - \beta_1 x_1 - \beta_2 x_2 \cdots \right)^2$$

$$= \| y - x\beta \|^2 = \| y - \hat{y} \|^2$$

Diff S wrt to each β we get linear eqns

$$x^T x \hat{\beta} = x^T y \longrightarrow \text{normal eqns}$$

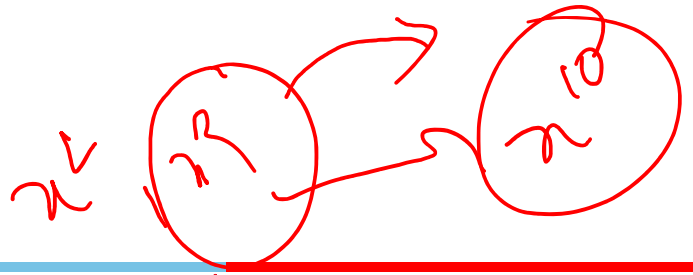
If $x^T x$ is non-singular, the soln is

$$\hat{\beta} = (x^T x)^{-1} x^T y$$

β_0 β_1

$AX = B$

$X = A^{-1}B$



Computationally, it is sometimes unwise even to form the normal equations because the multiplications involved in forming $X^T X$ can introduce undesirable round-off error.

→ If $X^T X$ is non-invertible ... ?

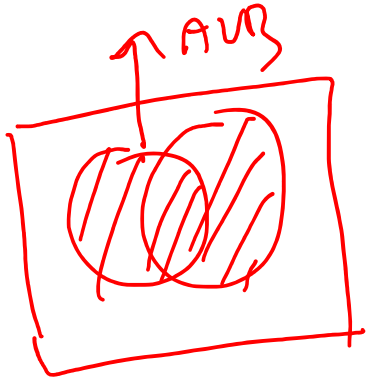
✓ Redundant features

✓ too many features → scaling

Revision

→ probability → $P(A \cup B)$

$P(A \cap B)$



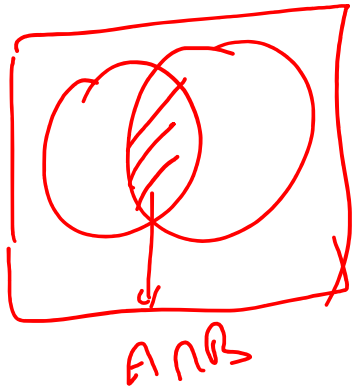
$P(A \cup B) = P(A) + P(B) - P(A \cap B)$

→ Conditional probability:

$P(A/B) = \frac{P(A \cap B)}{P(B)}$

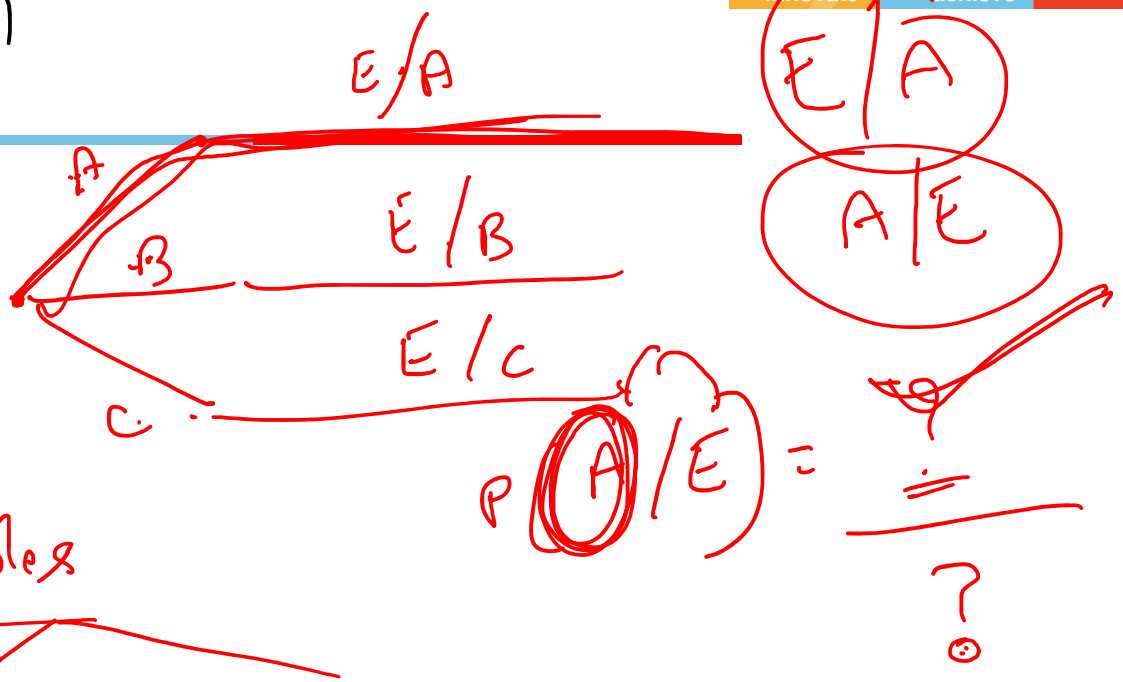
$P(B/A) = \frac{P(A \cap B)}{P(A)}$ i.e. $P(A \cap B) = P(A/B)P(B)$

$= P(B/A)P(A)$



Revision

→ Bayes theorem :-



→ Random variables

Discrete

$P(x)$

i) $0 \le P(x) \le 1$

ii) $\sum P(x) = 1$

Mean $= E(X) = \sum x P(x)$

Continuous

$f(x)$

i) $0 \le f(x) \le 1$

ii) $\int f(x)\,dx = 1$

$= \int x f(x)\,dx$

Variance : $E(x-\mu)^2$

$= E(X^2) - \mu^2$

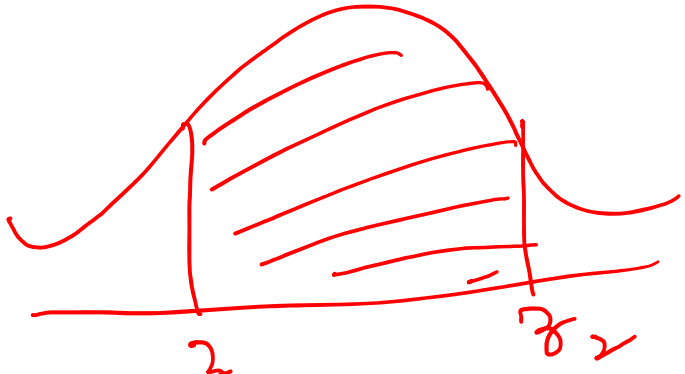
$= E(X^2) - [E(X)]^2$

Revision

→ Binomial dist $P(x) = {}^{n}C_{x}\, p^{x} q^{n-x}$, $x = 0, 1, 2 \ldots n$

poisson dist $P(x) = \dfrac{e^{-\lambda} \lambda^{x}}{x!}$, $x = 0, 1, 2 \ldots \infty$

→ Normal distribution :-



$P(30 \le x \le 50)$

$z = \dfrac{x - \mu}{\sigma}$

$P(z_1 \le z \le z_2)$

$= F(z_2) - F(z_1)$

→ Testing of Hypothesis → Mean → one → Z, t
→ two → Z, t
→ Proportion → χ^2 - distribution

Thanks