Project 2: Model Stock Prices Using Linear Regression & test all the 5 Assumption of LR Model Using statistical Methods

Linear Regression → statistical model

- Linear Regression is a way to find the relationship between 2 or more things.

→ If one thing changes (year of working exp), how does the salary changes.

- working exp
- salary

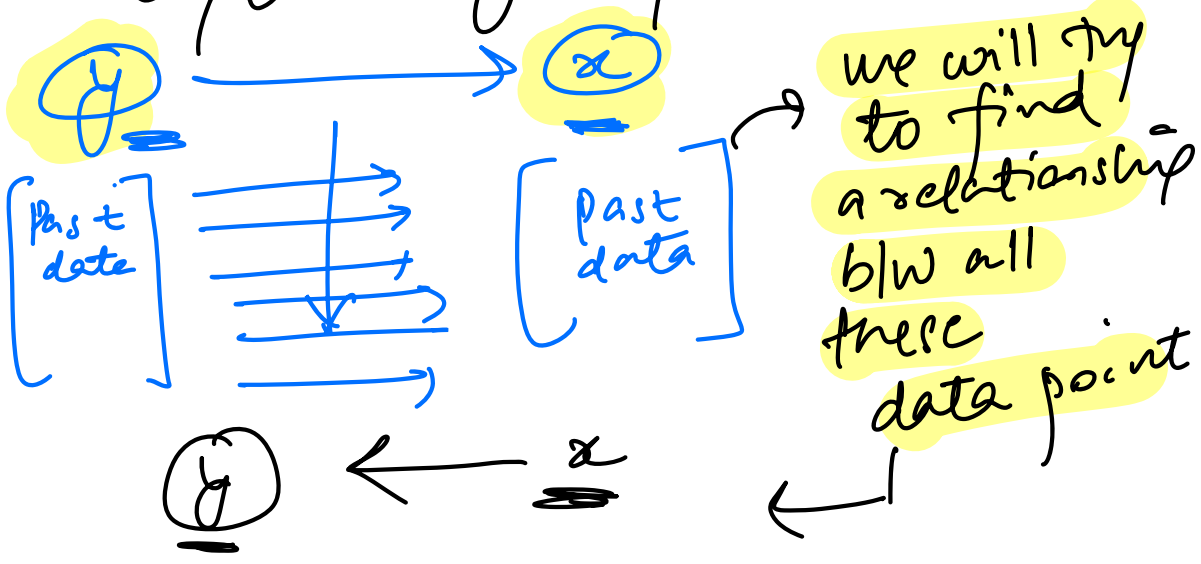
(x) ← relationship blw these 2 variable..
(y) ←

[If year of working exp is higher → salary high (True)]

→ pattern → past data that we observed

- LR helps you predict a value based on past data

→ predicting salary using year of working exp



Linear Regression Eqn ⇒

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

$x = 0$

- y = salary
- x = years of working exp
- β_0 = intercept → starting point of y when $x = 0$
- β_1 = coefficient of x

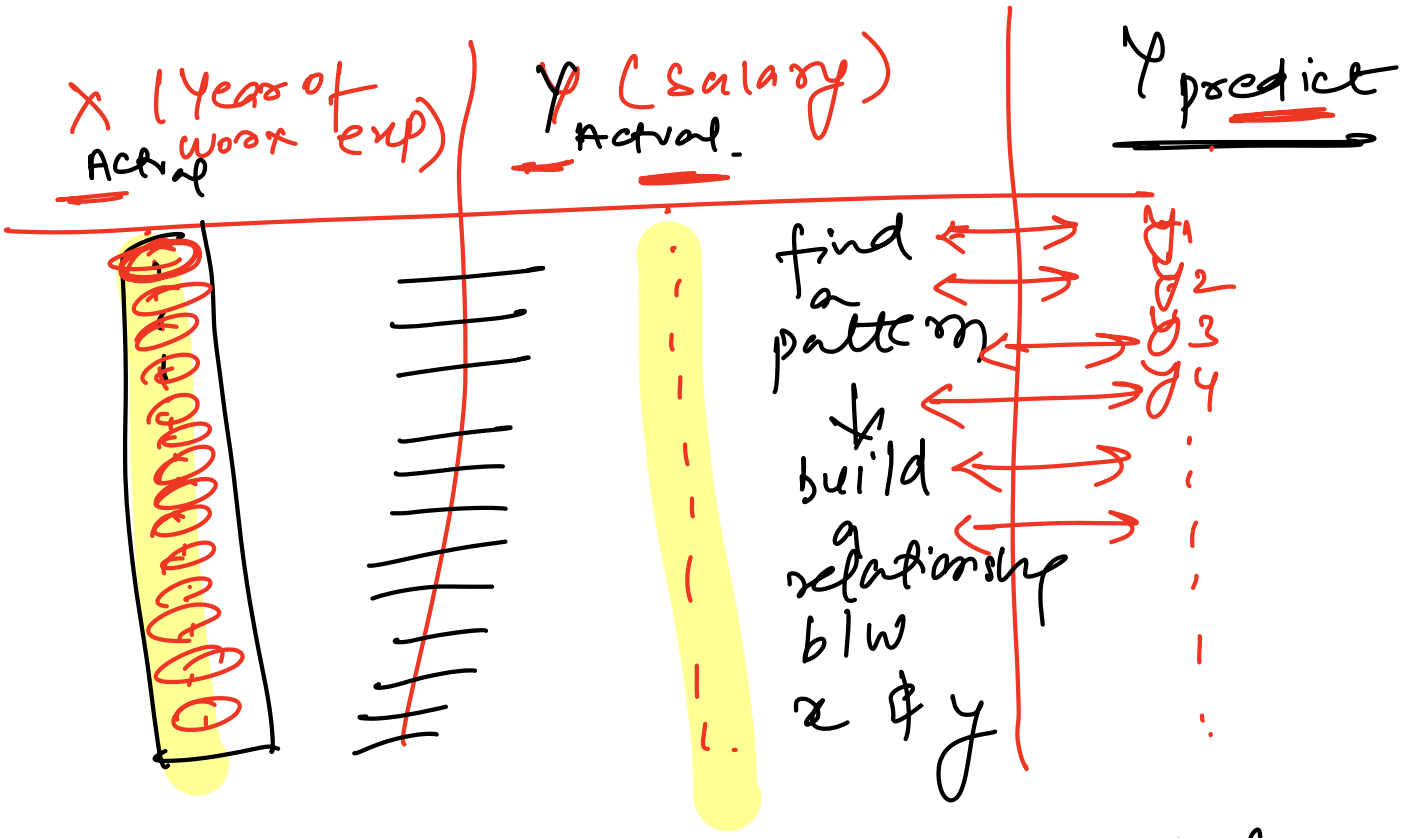
NOTE : Linear regression tries to draw a "straight line" through the

data point → Best fitted line

Goal → Predict Y

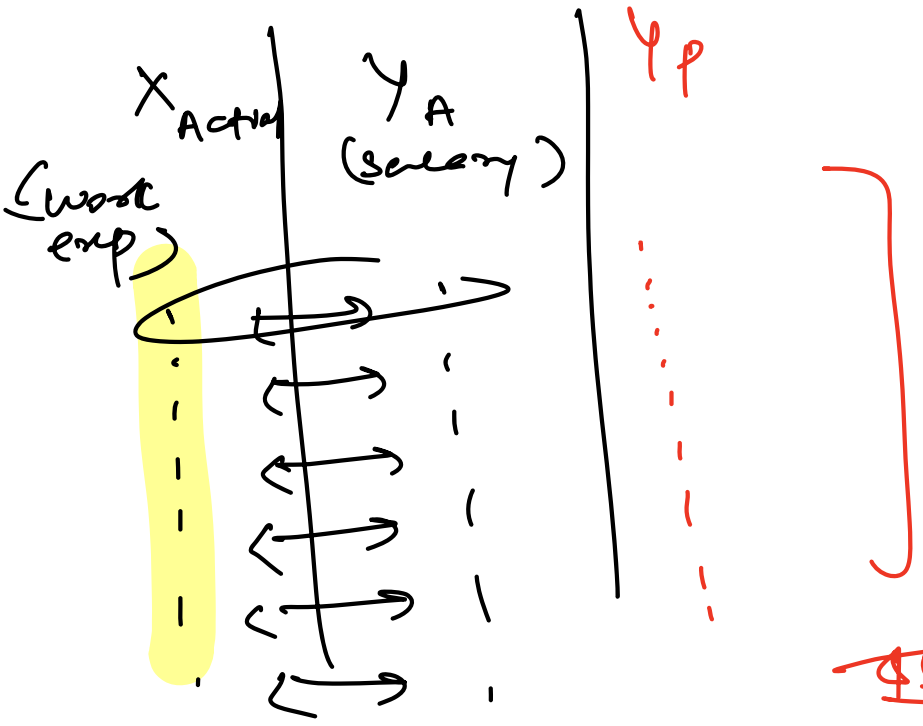
LR find the best fitted line that predict the output as close as possible to the real value.

"Best" Means → minimize the error b/w
→ what we predicted
→ what actually happen.



- When we run a LR model
  1st • Train the model

500 working professionals → Salary
→ working exp



How close
Y_A & Y_P are.

Y_A | Y_P → close
Best

Y_A | Y_P → Bad
far off
$100,000
$50,000

[We run a LR model to find a relationship b/w x & y]

$[y = \beta_0 + \beta_1 x]$ → LR Model

Using all these 500 data point

$\beta_0 = 0.5$
$\beta_1 = 2$

LR Model

$$y = 0.5 + 2x$$

$x = 2 \rightarrow y_1$
$x = 3 \rightarrow y_2$
$x = 3.5 \rightarrow y_3$

Linear Regression

→ LR is a statical method which is used to find a relationship b/w a dependent variable & one or more independent variable.

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

$y \rightarrow$ dependendent var
$x \rightarrow$ independent var.

y (salary) → x (year of work exp)

→ y is dependent on 'x'

→ x (working exp) → was not dependent on Anything (indep)

MBA

Salary → year of exp (1), technical skillset (2), which university you graduated (3)

depen

y = salary

1 → x_1, 2 → x_2, 3 → x_3

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

$\beta_0 \rightarrow$ intercept

$\beta_1, \beta_2, \beta_3$ → Coeff of x_1, x_2, x_3

ε → error term

↳ captures the variation that is not explained by the model.

Types

- One Independent Var (x) → $y = \beta_0 + \beta_1 x$
  - ↳ Simple Linear Regression
- More than one Indep. Var ($x_1, x_2 \ldots$) → $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$
  - ↳ Multiple Linear Regression.

$y = \beta_0 + \beta_1 x$ ← $x = x_1$

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$

How do you calculate coefficient

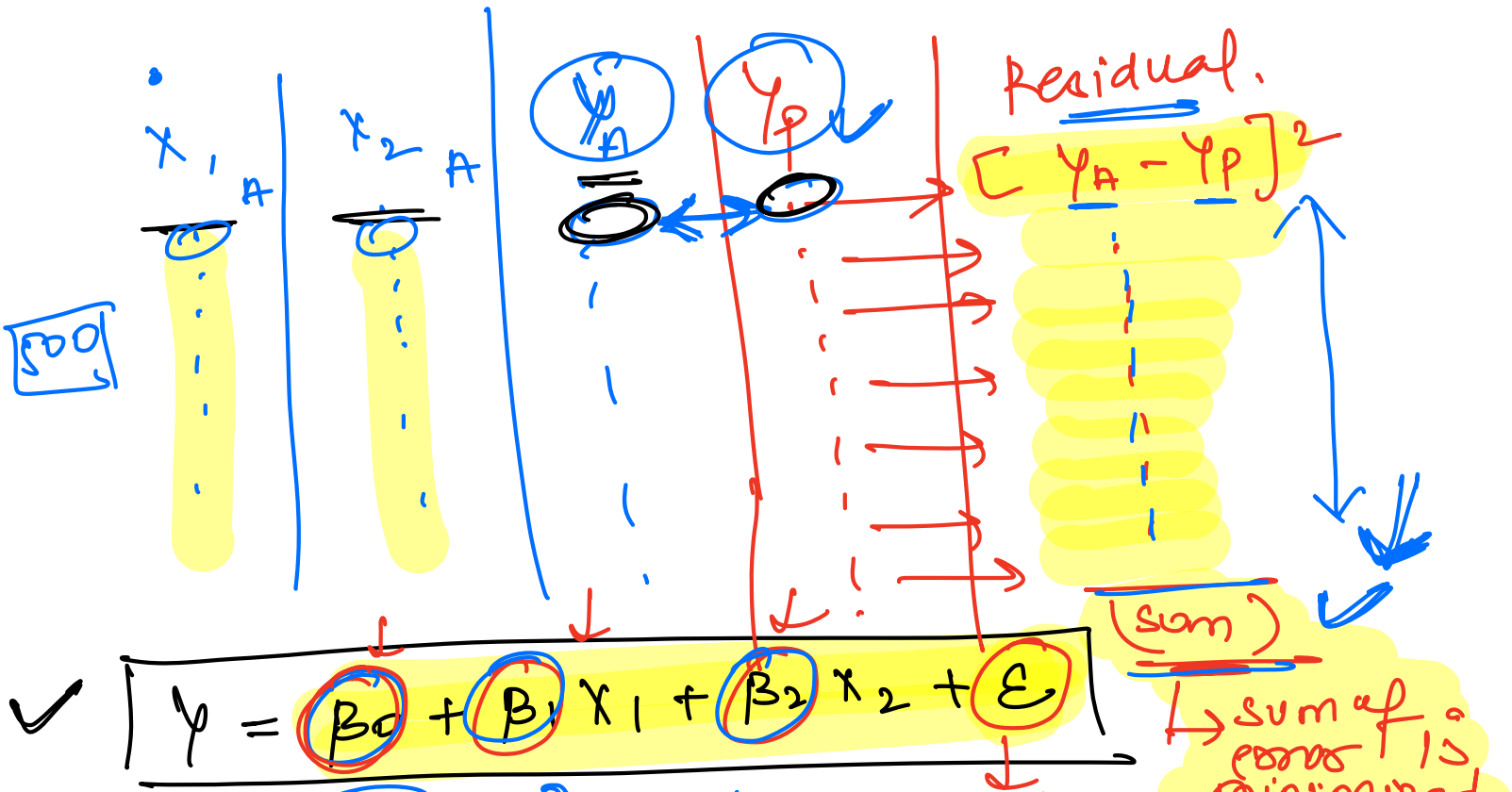
$\beta_0, \beta_1, \beta_2 \ldots \beta_n$

β_0 ↙ Intercept

$\beta_1, \beta_2 \ldots$ ↘ Coeff.

↳ The way we estimate the parameters is in such a way that the linear regression model minimizes the sum

of square errors [Residual]
b/w the Actual value & the predicted value.



$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

Error term

$\beta_0 = 0.7$
$\beta_1 = 0.5$
$\beta_2 = 1.5$
$\varepsilon = 0.05$

$$y = 0.7 + 0.5 x_1 + 1.5 x_2 + 0.05$$

final LR model

β_0
β_1
β_2
ε

least square error

minimize

$(y_A - y_P)^2$

→ The technique to find the parameter of LR model is called least square error method.

Summarize

- statistical model relationship b/w y & x ($x_1, x_2 \ldots$)
- LR tries to find best fitted line.
- Types → simple LR (x) [A vs P]
  → Multiple LR ($x_1, x_2 \ldots$)
- $B_0, B_1 \ldots$ → least square error Method

Assumption of Linear Regression Model

** • what are the Ass. of LR Model
** • How do you test all the Ass. of LR Model.

- LR has 5 assumption

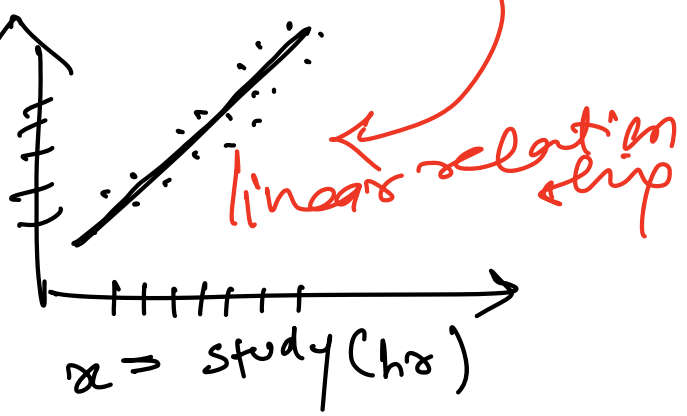
a) Linearity
b) Independence of Error

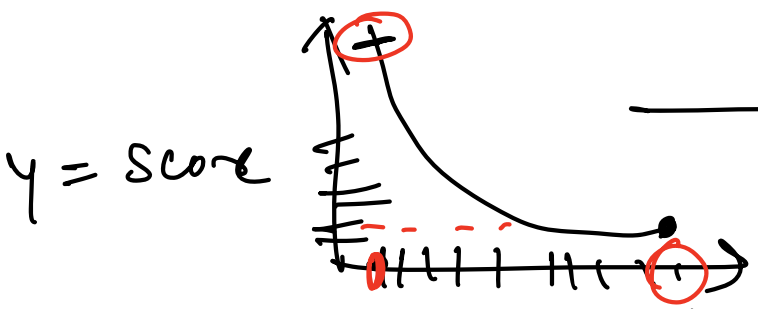
c) Homoscedasticity
d) Normal Distribution of Error
e) No multi collinearity

(a) Linearity: There should be a linear relationship b/w your dependent & independent variable

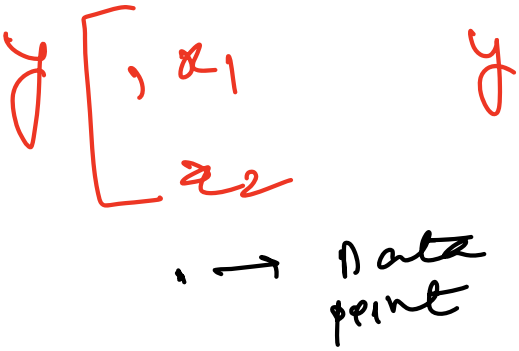
straight line

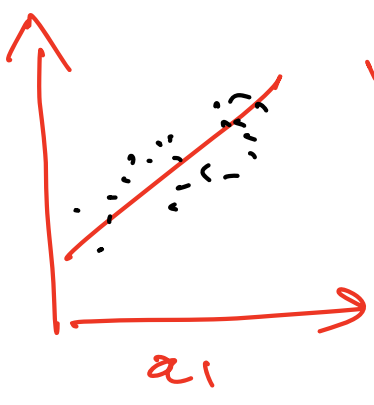
y & x
↳ linear relationship

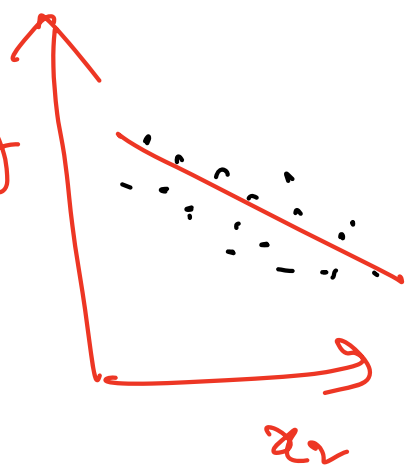




→ Does this make sense
↳ 22 No



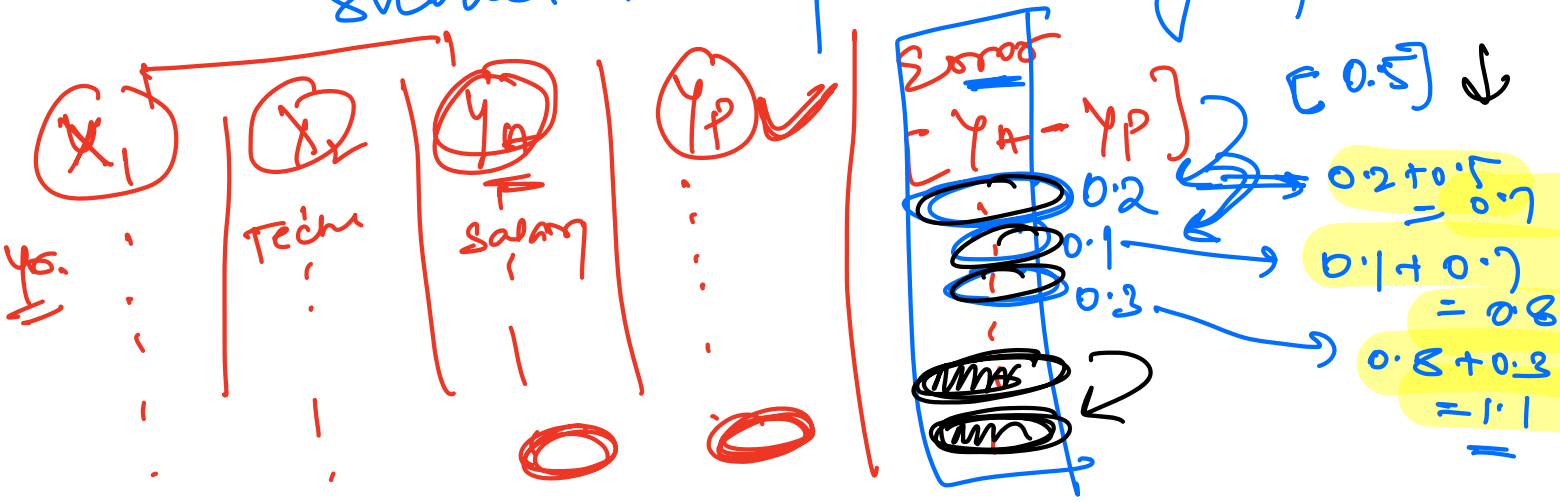




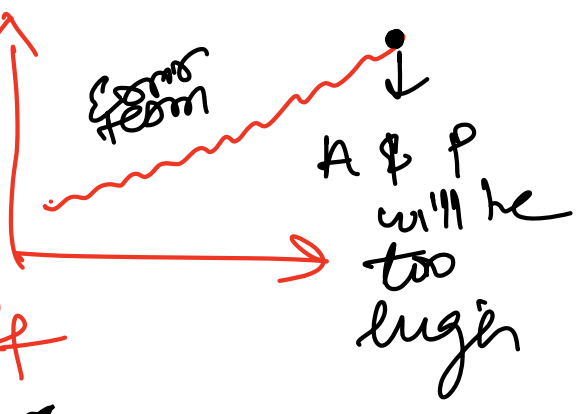
② Independence of Errors (Residual)

- The residual (errors) are independent of each other.

Mean: The error (difference b/w A & P) should not follow any pattern



- If its dependent



If error is high
↳ A & P are too off
↳ No LR

→ All the error (Diff A vs P) should be independent of each other

Eg: By mistake one person's salary prediction is too high ⇒ Diff A & P

This should not affect the prediction of next person's salary.

Error higher

③ Homoskedasticity (Constant variance)

The variance of residual is constant across all levels of independent variable.

Means: The spread of the error should be roughly the same across all values of i/p.

Eg: When you are predicting the salary for someone who has 1 yr of work exp & someone with 20 yr of work exp, the error in prediction should be similar

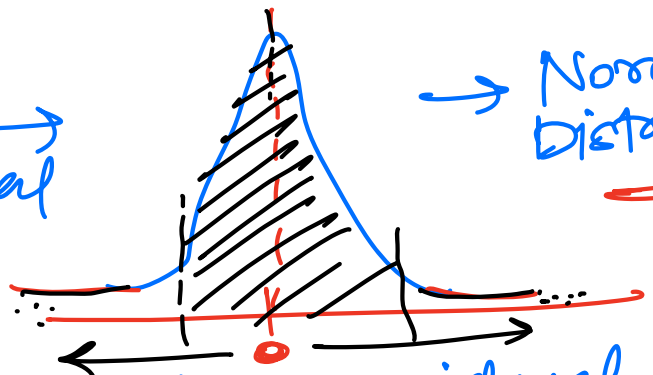
↳ small for one

↳ Big for the other

Residual

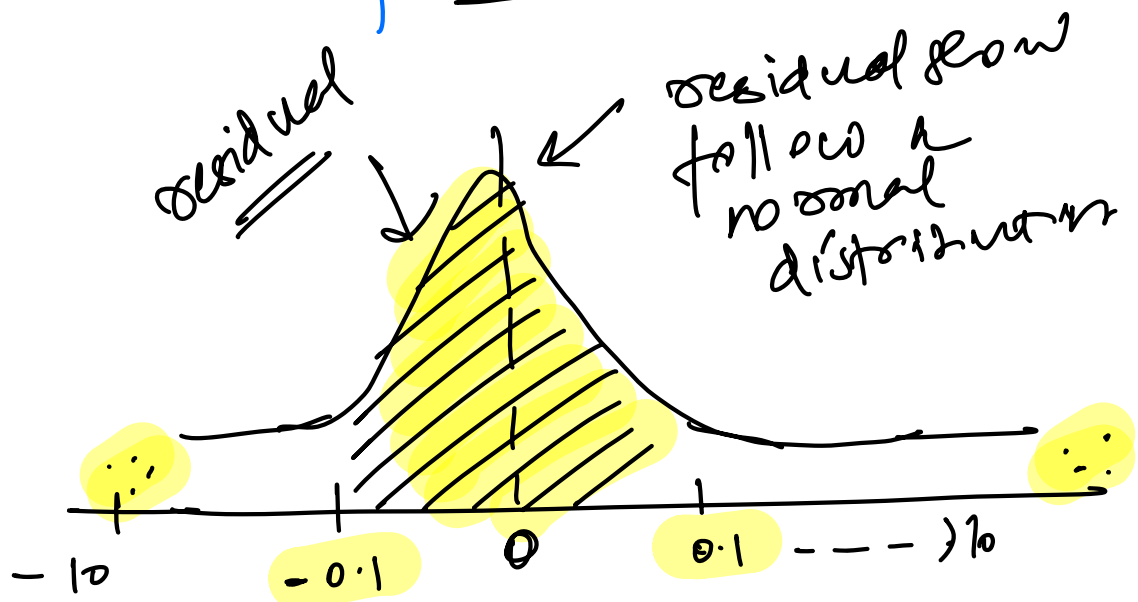
(4) Normal Distribution of Errors

✓ The residual should be normally distribution



Means : The residual (dif A & P) should follow a bell shape curve.

→ most errors should be small few errors can be large.



Error Term

Residual

$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$

error term

✓ error which the model is not able to explain

✓ Diff b/w A & P

✓ Residual

→ $(A-P)^2$ +ve → Log Normal Dist

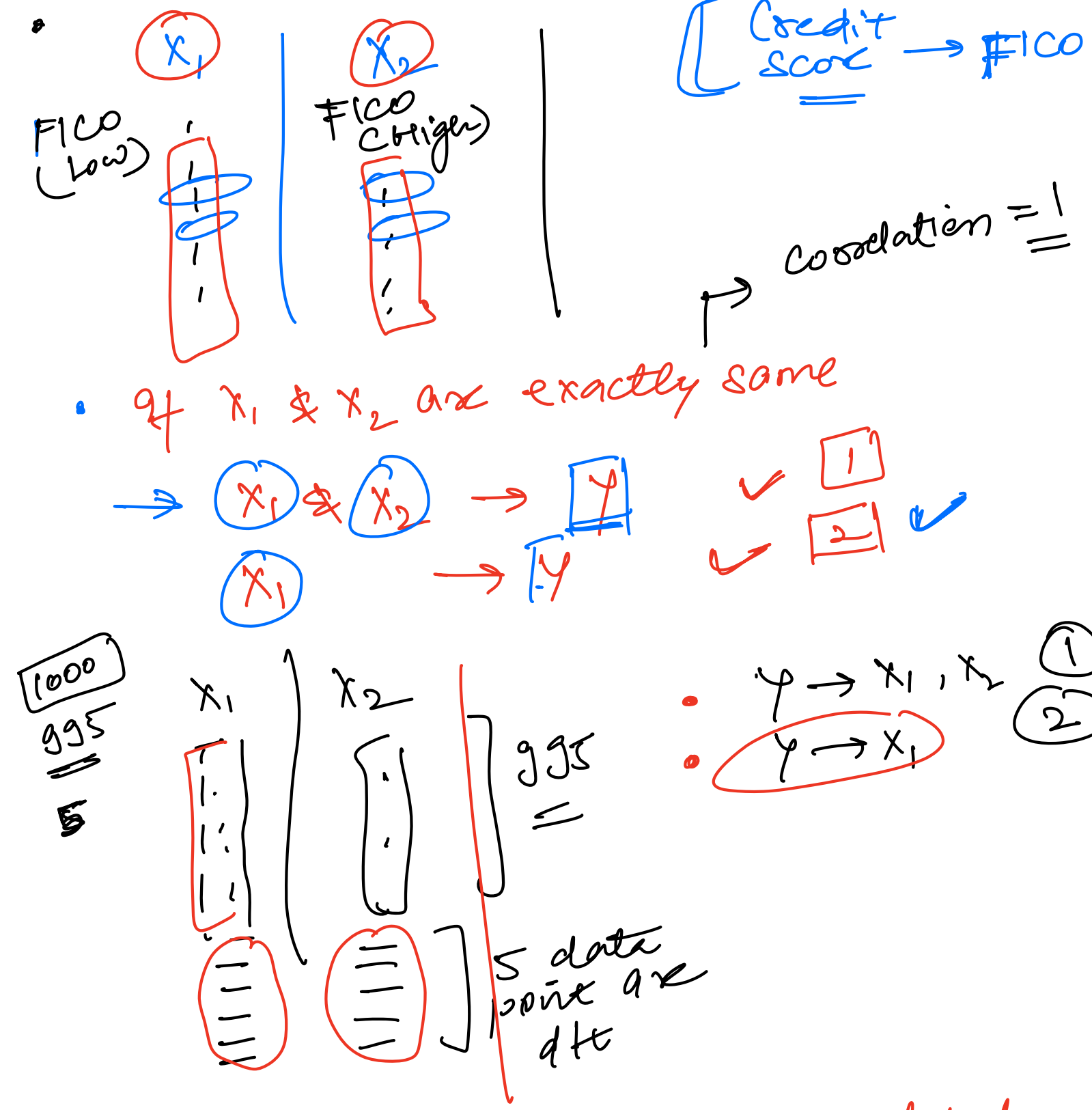
parameter

least square error Method

Plot → Residual ⇒ (A-P) → Normal distⁿ

5 No Multicollinearity

- The independent var ($x_1, x_2, x_3 \ldots$) should not be highly correlated with each other



- When i/p var are highly correlated
  - ↳ it makes no sense in having those var in our LR model
- We need to make sure i/p variable should not be closely related

to each other

x_1 x_2

same

House size

House price

Size of each room + size of all remaining space

If the i/p var are closely aligned with each other → it makes no sense of having both of them together

↳ Drop 1 var

SLR → x

MLR → x_1 x_2

Summarize

① Linearity : linear relationship b/w dep & indep var
(y) ($x_1, x_2 \ldots$).

② Indepence of Errors : The residual should be independent of each other

③ Homoscedasticity : The variance of residual should be const

④ Normal Dist^n of Errors → The residual should be normally distributed

⑤ No multicollinearity → The independent var ($x_1, x_2, \ldots$) should not be highly correlated with each other.