Regression Analysis - I

↳ It is statistical method used to examine the relationship b/w one dependent var. and one or more ind. variable.

↳ Its goal is to understand how dep. variable changes when one or more dep. variable altered

Flow of :-

- Define the research quest.
- Collect & prepare data
- Visualize the data
- Check assumptions
- Fit linear regression model
- Interpret the model (Regression Analysis)
- Validate the model
- Report results

- Statistical inference
  - ↳ by using Sample we try to conclude about population

CGPA	IQ	Package LPA
-	-	-
-	-	-
-	-	-

$f(\text{Lpa}) \rightarrow \text{Lpa} = f(\text{CGPA}, \text{IQ}) + \varepsilon$ (irreducible error)

Calculate it

parametric
↳ we assume it is following linear or other regression
↳ $\text{Lpa} = \beta_0 + \beta_1 \text{CGPA} + \beta_2 \text{IQ}$
↳ while calculating we get values $b_0, b_1, b_2 \neq \beta_0, \beta_1, \beta_2$

non-parametric
↳ we don't assume anything rather start from scratch

$\therefore f'(\text{CGPA}, \text{IQ})$

$\therefore$ error $\Rightarrow f() - f'()$
↳ reducible error

$f()$ → True relation of x & y for population
$f'()$ → estimated " " " " given sample

: Prediction :- Suppose we have CGPA & IQ & LPA. Now we want to predict other people's LPA

lpa = $\beta_0 + \beta_1$ (cgpa) + β_2 (iq)

→ any person come we put its cgpa & iq and get lpa

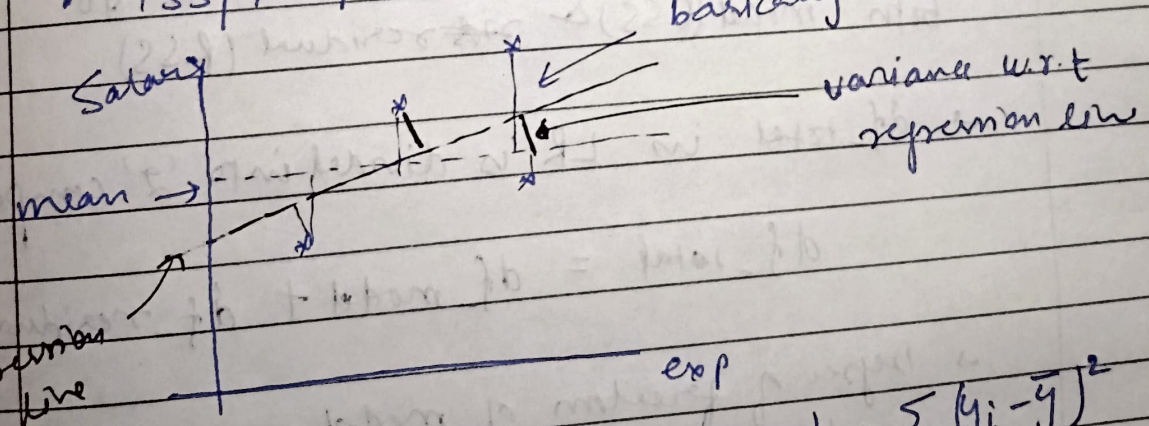
- Inference → In this we do relationship study. like relat^n b/w lpa & cgpa : now it changes etc

→ Linear regression is good in inference if inference good then prediction week & vice versa

- F-statistics & Prob(F-statistics) :) They tell us there is any relat^n b/w X(input columns) & Y (o/p col^n)

for this we have to do hypothesis test called F-test for overall significance

- TSS/RSS/ESS



TSS (Total Sum of squares) = $\sum (y_i - \bar{y})^2$

↳ It tells overall variance in the data

RSS (Residual Sum of Squares)
$= \sum (y_i - \hat{y}_i)^2$

↳ It tells variance present in data w.r.t regression line (means even after passing the reg. we can't tell any variance occur)

- ESS (Explained Sum of Squares)
$= TSS - RSS$

$TSS = ESS + RSS$

- Degree of freedom -
↳ #rows - 1 $= n-1$
df_total $= n-1$

↳ It represent overall variability (TSS) in dataset that can be attributed to both model (ESS) & ~~risk~~ residual (RSS)

→ df_total in LR is divided into '2' components -

df_total = df_model + df_residual

→ Degree of freedom of model -
↳ Equal to no. of independent variables in

x_1 x_2 | y → $k=2$

→ Degree of freedom of residual -
↳ n - no. of estimated parameters including intercept

Page No.
Date

$= n-(k+1)$

⇒ F-test for overall significance is a statistical test used to find weather there is relation by columns.

exp | Salary ⇒ weather salary is affected by experience or not