Session - 53

→ multicallinearity:- (MC)

↳ it occurs when two or more independent variable(i/p column) in a multiple reg column are highly correlated (0.8, 0.9).

· Problem with MC -

$$lpa = \beta_0 + \beta_1 cgpa + \beta_2 iq$$

$\beta_1 \Rightarrow$ If we keep iq constant and increase CGPA by 1 then how much lpa will get ↑

But if cgpa & iq are correleted than if ~~iq~~ CGPA ↑ then iq ↑ so $\beta_0, \beta_1, \beta_2$ become unreliable

· If we are doing inferences than ~~no~~ mc is bad ~~&~~ as $\beta_1, \beta_2, \beta_0$ are not reliable

· If we are doing predictⁿ than no problem

$x_1 \quad x_2 \mid y$ x_1 & x_2 are releted

$$x_1 = a_0 + a_1 x_2 + \eta$$

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

$$\beta_0 + \beta_1 a_0 + \beta_1 a_1 x_2 + \eta \beta_1 + \beta_2 x_2 + \varepsilon$$

$$y = (\beta_0 + \beta_1 a_0) + (\beta_1 a_1 + \beta_2) x_2 + (\eta \beta_1 + \varepsilon)$$

Since y is only fⁿ of x_2 so no problem

- Generally in correl^n -

$$X_1 = a_0 + a_1 X_2 + error$$

due to this error we can't find exact value of X_1.

but if $X_1 = a_0 + a_1 x_2$
means no error, than this situatn is called perfect multicollinearity

- Lets assume Perfect MC -

CGPA	Percent	lpa
8	80	3
6	60	4

$$lpa = \beta_0 + \beta_1 cgpa + \beta_2 percent + error$$

we want to find β -

$$\beta = (X^T X)^{-1} X^T Y$$

↳ while solving it we have to find $\det(X^T X)$ which is zero. means its a singular metrix means we can't calculate inverse
So we can't find β.
Hence in perfect MC we can't find β. and most of data is not perfect MC

- what exactly happens in MC?
  - → Difficult to tell which indep. variable has most significant affect on dep. column
  - → infleted standard error (Big SE)
  - → unstable & unreliable estimates (If we change small values in i/p then lots of change)

Page No.
Date

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

$var(\beta) = SE(\beta)\ \sigma^2 (x^T x)^{-1}$

$= \sigma^2 \begin{bmatrix} \overset{a}{-} & & \\ & \overset{b}{-} & \\ & & -c \end{bmatrix}$

↳ var. covar. matrix

$a \Rightarrow var(\beta_0)$ $\quad SE(\beta_0) = \sqrt{Var(\beta_0)}$

$b = var(\beta_1)$

$c = var(\beta_2)$ $\quad SE(\beta) = \sqrt{dig(\sigma^2 (x^T x)^{-1})}$

=> if Strong MC then $\det(x^T x)$ will be very small due to which inverse will be very high hence Standard Error will also be high

→ Types of multicollinearity -

i) Structural MC -

↳ it arises due to the way in which variables are defined

eg $x \mid y$

↳ $x^0 \ x^1 \ x^2 \mid y$

due to polynomial reg. Structural MC arises as $x^0 \ x^1 \ x^2$ are dep correlated

2) Data driven MC -

↳ Indep. variables are highly correlated due to specific data being analysed

more area of flat => more washroom.

· How to detect MC –

1) Correlation

$$x_1 = a_1 x_2 + a_0 + \text{Error}$$

Correlatn b/w x_1 & x_2

check values > 0.8 or 0.9

2) Variance Inflation factor (VIF)

A B C | D

1) we make A, B as input & C as output and apply L.R & find R^2 score
2) A, C as input & B as o/p
3) B, C " " A "

$$\text{VIF} = \frac{1}{1-R^2}$$ } if VIF > 5 or 10 then M.C exist else not

3) Condition Number –

matrix $X^T X$

$$\text{Con. No} = \frac{\text{Largest Eigen value}}{\text{Smallest Eigen value}}$$

Cond^n no. tells about ill conditioning of matrix. means if small change in error cause large change in Sq^n

Cond^n no > 30 means MC

· How to remove MC –

- → Callect more date
- → Remove one of the highly correlated variable
- → combine correlated variable
- → Use partial least squars (PLS) regression