Data Science and Artificial Intelligence

Machine Learning

Regression

Lecture No. 07

By- SIDDHARTH SABHARWAL SIR

Recap of Previous Lecture

- **Topic** — Hat matrix
- **Topic** — Multicollinearity
- **Topic** — Assumption in LR
- **Topic** — Outlier
- **Topic**

Topics to be Covered

- Topic: Assumption in LR
- Topic: LR advantage disadvantage
- Topic: LR Space-Time Complexity
- Topic
- Topic

About the Faculty

- AIR 1 GATE 2021, 2023 (ECE).
- AIR 3 ESE 2015 ECE.
- M.Tech from IIT Delhi in VLSI.
- Published 2 papers in field of AI-ML.
- Paper 1 : Feature Selection through Minimization of the VC dimension.
- Paper 2 : Learning a hyperplane regressor through a tight bound on the VC dimension.

By- SIDDHARTH SABHARWAL SIR

Be the change that you wish to see in the world.

— MAHATMA GHANDI

BRIAN TRACY
INTERNATIONAL

REVISION

Hat Matrix

Model

$$\left(y = \underbrace{\beta_0 + \beta_1 x^1 + \beta_2 x^2 \,----}\right)$$

when we put values of x we get predicted value

$$\underbrace{X\beta} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_N \end{bmatrix}$$

$$\left\{ X\beta = \hat{Y} \right\}$$

- $\hat{Y} = X\beta$

$$= X\left[(X^T X)^{-1} X^T Y\right]$$

Pred. Value

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

Actual Value

Hat matrix

$$\left[X(X^T X)^{-1} X^T\right]$$

REVISION

Outlier and its effect

* data Point with huge noise

• LR is affected by outlier

Linear Regression

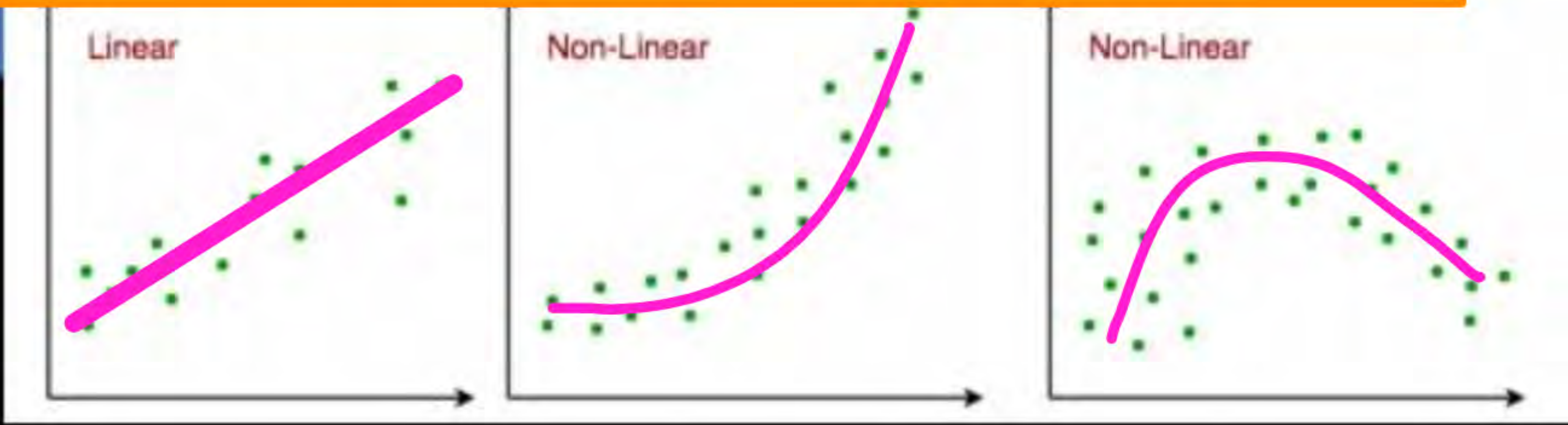
Assumptions in Linear Regression

Linear regression needs to meet a few conditions in order to be accurate and dependable solutions.

1. **Linearity:** The independent and dependent variables have a linear relationship with one another. This implies that changes in the dependent

If LR is applied on data with NL Relation b/w y and x

then model will be Underfit/Poor model.



2. No multicollinearity

- In LR we need that dimensions shd be independent of each other
- In LR the model

$$y = \beta_0 + \beta_1 x^1 + \beta_2 x^2 + - - - -$$

β_1 : Coeff of 1st Dim - -

β_2 : " " 2nd " - - -

What is multicollinearity

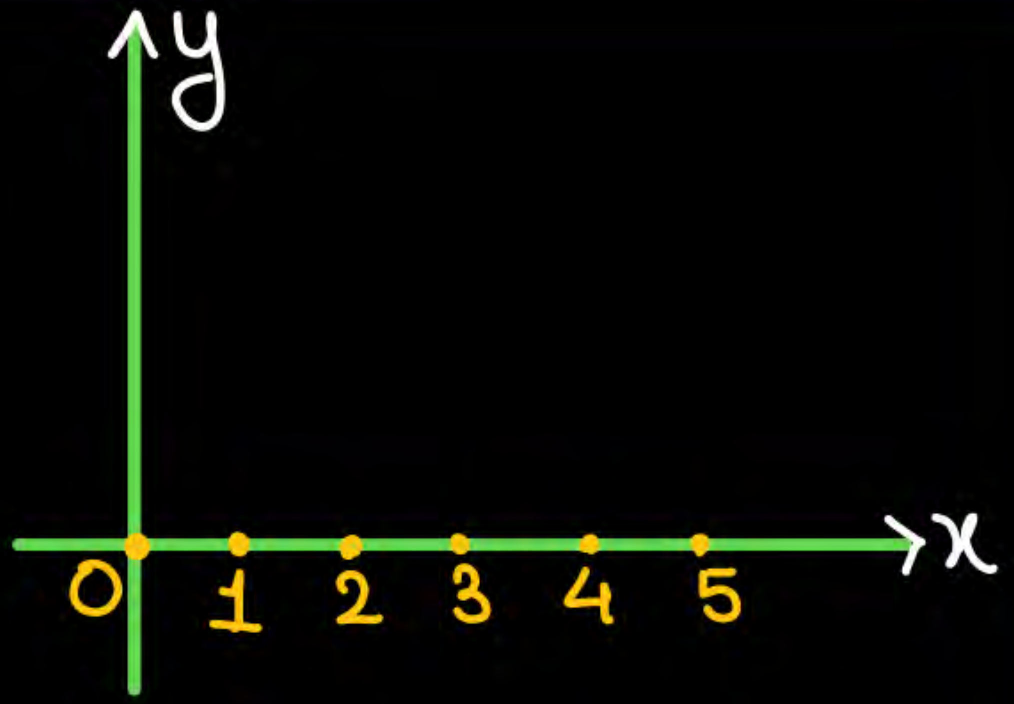
it means that dimensions are dependent on each other

- So in LR we get seperate coeff. for each dimension.
- If dimensions are dependent we get poor model.

Assumptions in Linear Regression

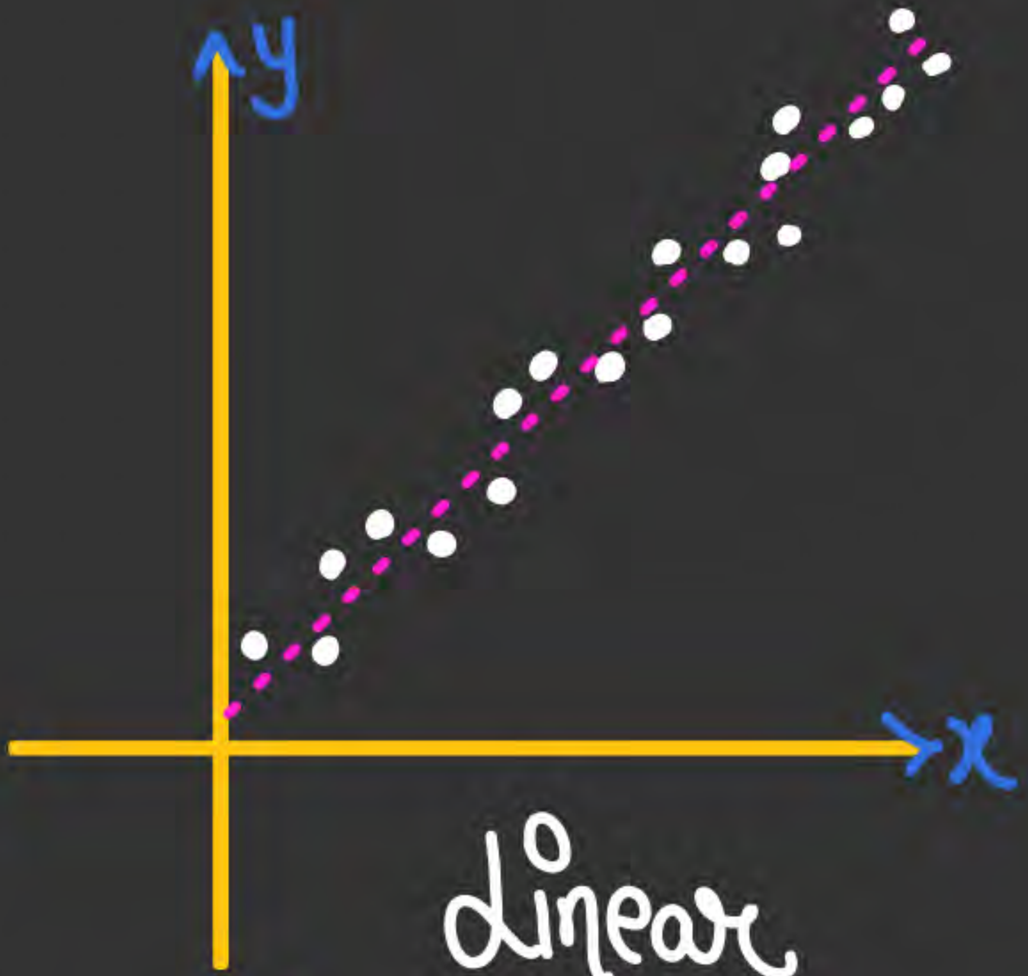
3. The datapoints should be independent of each other ⇒

To understand exact pattern of data, the data shd be collected Randomly and hence data points shd be independent of each other



4.

data Pattern

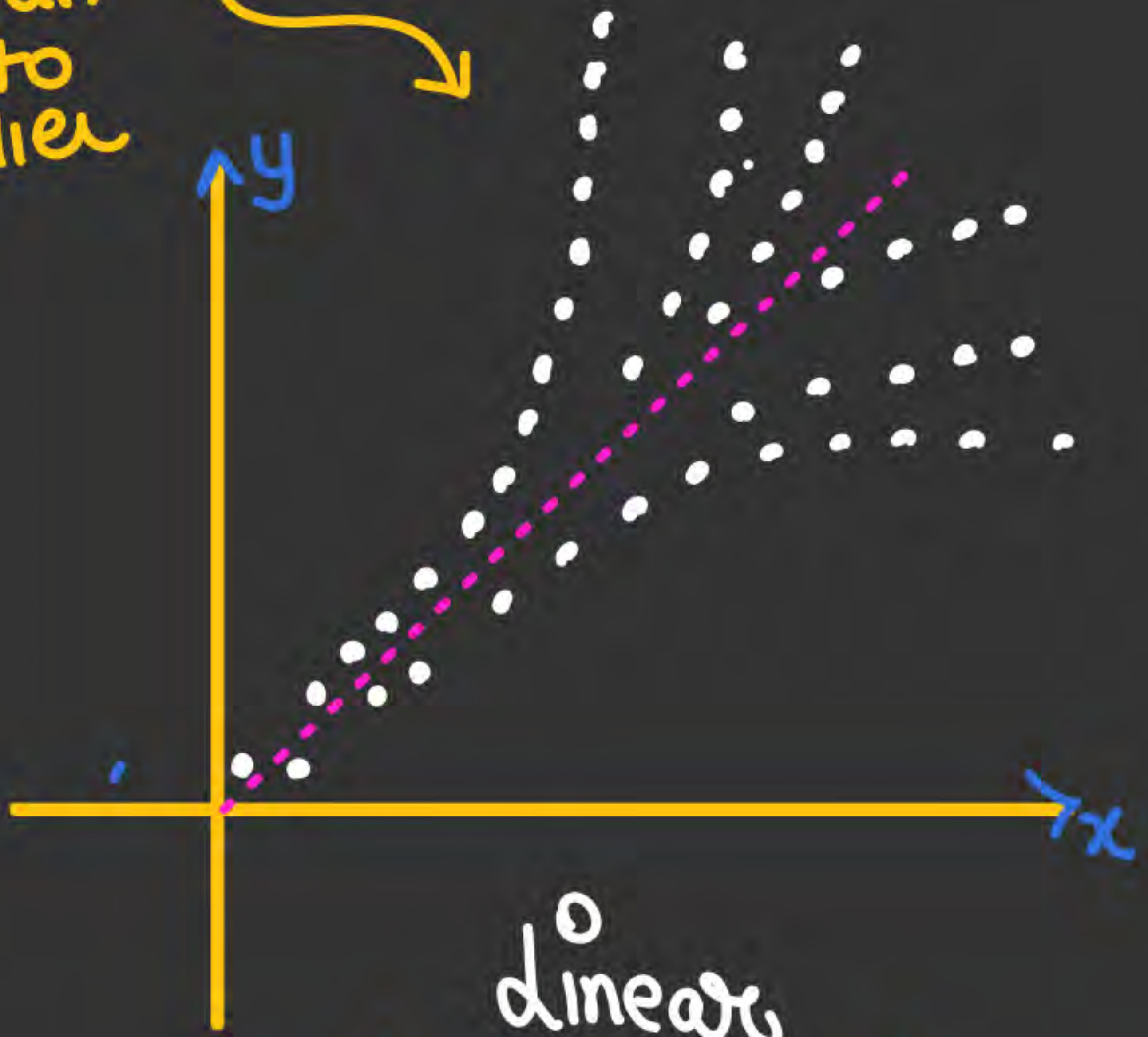


linear

→ Noise in data is independent of x

→ noise is homogeneous

• LR fail due to outlier



linear

→ Noise inc as x inc

→ Noise is Heterogeneous.

- Temp 50°

Rainfall

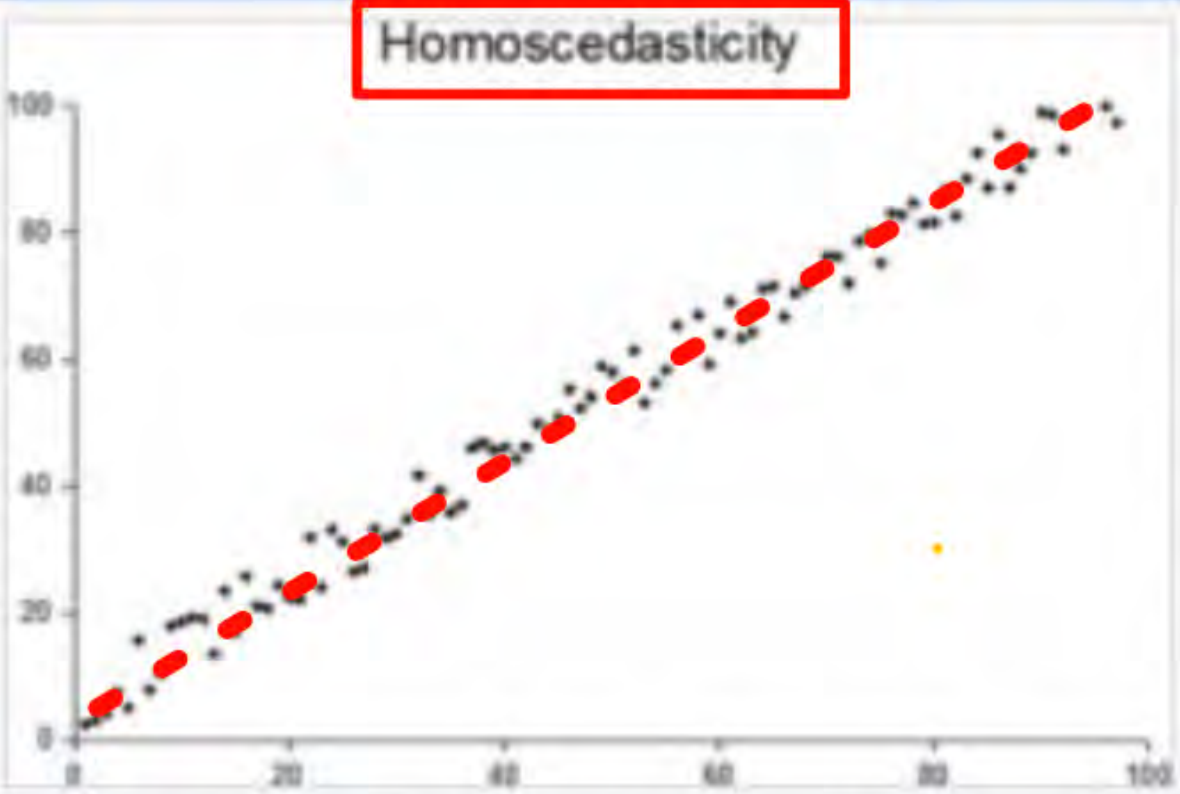
Pattern

- Y and x Ka relation is Pattern

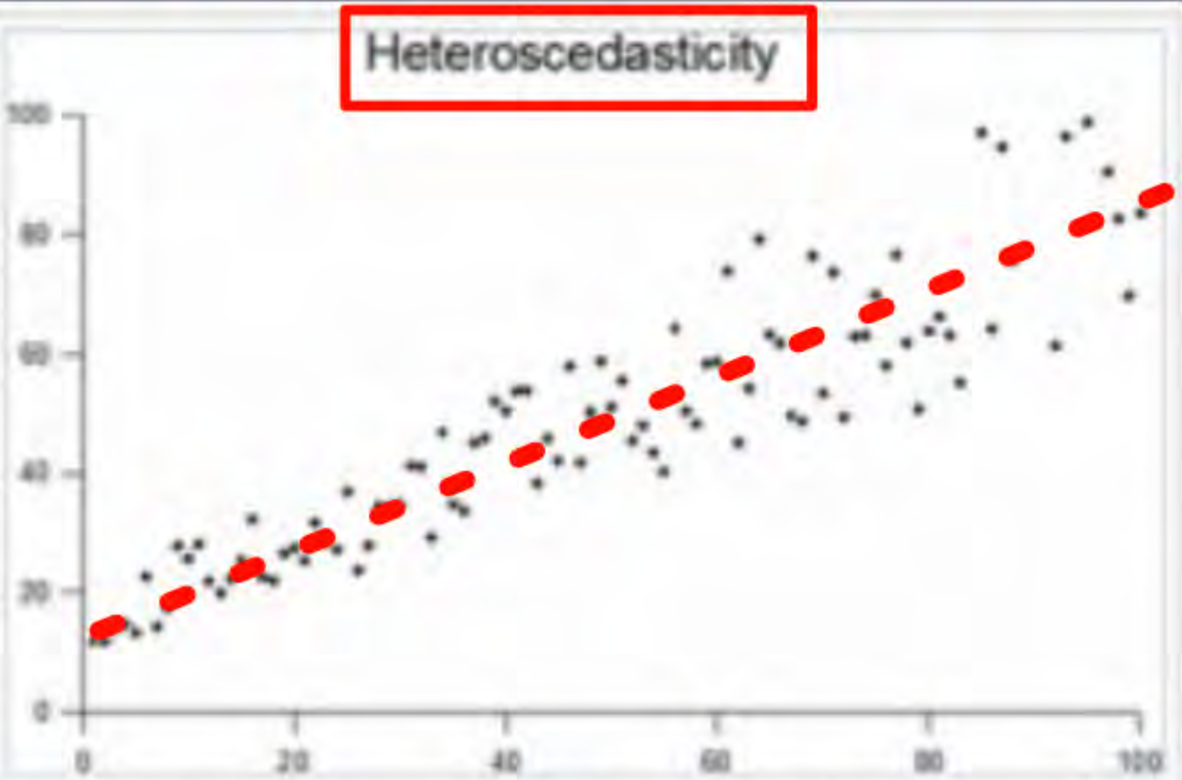
x	y
Temp	Rainfall
50°	
60°	
70°	
30°	

Assumptions in Linear Regression

4. Homoscedasticity: In LR we need Homoscedasticity.



Plot with random data showing homoscedasticity: at each value of *x*, the *y*-value of the dots has about the same variance.



Plot with random data showing heteroscedasticity: The variance of the *y*-values of the dots increase with increasing values of *x*.

Assumptions in Linear Regression

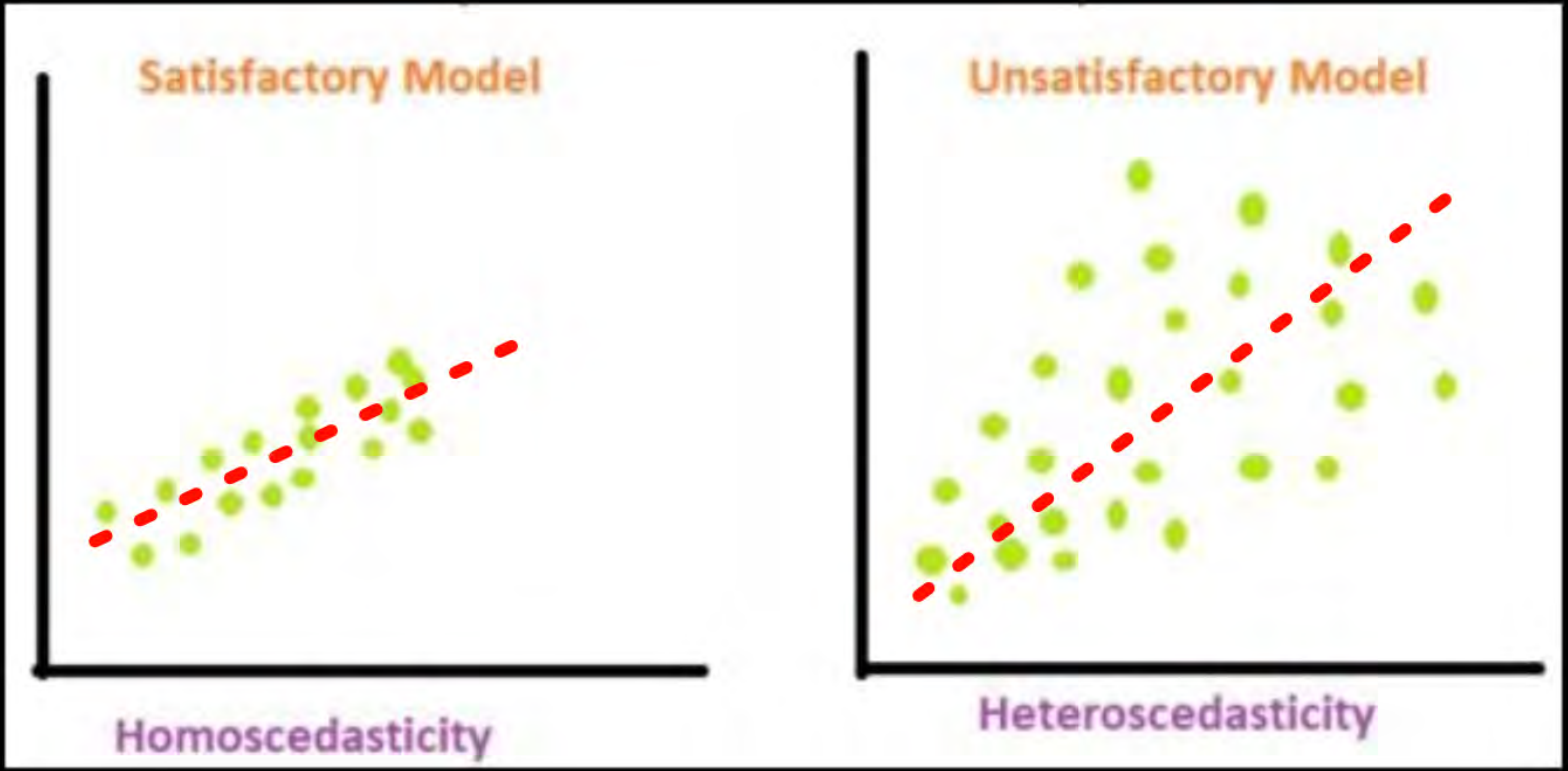
~~3~~ 4. **Homoscedasticity:** → • The noise in data shd be independent of x values.

(Name + meaning)

• Heteroscedasticity → noise dependent on x

→ creating too many outlier

Assumptions in Linear Regression



GAATE 2026 FODENGE

Assumptions in Linear Regression

Independent Variable → dimensions

$\boxed{\textbf{4. No multicollinearity:}}$ There is no high correlation between the independent variables. This indicates that there is little or no correlation between the independent variables. Multicollinearity occurs when two or more independent variables are highly correlated with each other, which can make it difficult to determine the individual effect of each variable on the dependent variable. If there is multicollinearity, then multiple linear regression will not be an accurate model.

How to find Correlation b/w 2 variable

3D data

x^1	x^2	x^3	y
a	b	c	y_1
d	e	f	y_2
g	h	i	y_3
J	K	L	y_4

To check multicollinearity we have to check Correlation b/w all dimension

$x^1 x^2$

$x^1 x^3$

$x^2 x^3$

Correlation Table

	x^1	x^2	x^3
x^1	$\rho_{x^1x^1}$	$\rho_{x^1x^2}$	$\rho_{x^1x^3}$
x^2	$\rho_{x^2x^1}$	$\rho_{x^2x^2}$	$\rho_{x^2x^3}$
x^3	$\rho_{x^3x^1}$	$\rho_{x^3x^2}$	$\rho_{x^3x^3}$

How to find Correlation b/w 2 variable

3D data

x^1	x^2	x^3	y
a	b	c	y_1
d	e	f	y_2
g	h	i	y_3
J	K	L	y_4

Correlation Matrix

$$\begin{array}{c|ccc} & x^1 & x^2 & x^3 \\ \hline x^1 & \rho_{x^1x^1} & \rho_{x^1x^2} & \rho_{x^1x^3} \\ x^2 & \rho_{x^2x^1} & \rho_{x^2x^2} & \rho_{x^2x^3} \\ x^3 & \rho_{x^3x^1} & \rho_{x^3x^2} & \rho_{x^3x^3} \end{array}$$

$$\rho_{x^1x^1} = \frac{Cov(x^1x^1)}{\sigma_{x^1}\sigma_{x^1}} = \frac{Var(x^1)}{\sigma_{x^1}^2} = 1$$

$$\rho_{x^1x^2} = \frac{Cov(x^1x^2)}{\sigma_{x^1}\sigma_{x^2}} = \rho_{x^2x^1}$$

How to find Corelation b/w 2 variable

3D data

x^1	x^2	x^3	y
a	b	c	y_1
d	e	f	y_2
g	h	i	y_3
J	K	L	y_4

Correlation Matrix

$$
\begin{array}{c|ccc}
 & x^1 & x^2 & x^3 \\
\hline
x^1 & 1 & \rho_{x^1x^2} & \rho_{x^1x^3} \\
x^2 & \rho_{x^2x^1} & 1 & \rho_{x^2x^3} \\
x^3 & \rho_{x^3x^1} & \rho_{x^3x^2} & 1
\end{array}
$$

- If this matrix is given, how to check Multicollinearity.

→ If except diagonal other values are close to zero →
- dimensions not Correlated
- No multi-collinearity

Correlation Matrix

How to find Correlation b/w 2 variable

3D data

x^1	x^2	x^3	y
a	b	c	y_1
d	e	f	y_2
g	h	i	y_3
J	K	L	y_4

	x^1	x^2	x^3
x^1	1	$\rho_{x^1x^2}$	$\rho_{x^1x^3}$
x^2	$\rho_{x^2x^1}$	1	$\rho_{x^2x^3}$
x^3	$\rho_{x^3x^1}$	$\rho_{x^3x^2}$	1

- If this matrix is given, how to check Multicollinearity.

→ If except diagonal other values are close to ± 1 $\Big\}$ → Multi-Collinearity

- dimensions are Correlated

find x

$\min f(x)$

find x

$\min \frac{1}{2} f(x)$

Variance Inflation Factor
VIF

Take x^1 as label and apply LR

x^1 Ko x^2 x^3 Ke term mein dikhunga

$$x^1 = \alpha_0 + \alpha_1 x^2 + \alpha_2 x^3$$

Label

data

Remove

x^1	x^2	x^3	y
a	b	c	y_1
d	e	f	y_2
g	h	i	y_3
j	k	l	y_4

Now

- If the model is predicting value of x^1 correctly $\rightarrow$ model is good $\rightarrow$ dimensions are correlated

Variance Inflation Factor
VIF

Label   Remove

data

x^1	x^2	x^3	y
a	b	c	y_1
d	e	f	y_2
g	h	i	y_3
j	k	l	y_4

Take x' as Label and apply LR

x^1 Ko x^2 x^3 Ke term mein dikhunga

$$\left(x^1 = \alpha_0 + \alpha_1 x^2 + \alpha_2 x^3\right)$$

Now

- If the model is predicting Value of x' Poorly $\longrightarrow$ model is Poor underfit $\longrightarrow$ dimensions are not Correlated.

Variance Inflation Factor

- Remove y from data
- Now take any one dimension as label (x^1)
- Now apply LR :- x^1 as label and other $x^2, x^3, x^4 --- x^D$ as dimensions
- Now we get model
  $$x^1 = (\alpha_0 + \alpha_1 x^2 + \alpha_2 x^3 ---)$$

Multicollinearity #1

- Case 1 : model is good, predicted value of x_1 = actual value
  $R^2 \Rightarrow$ Close to 1

No multicollinearity.

- Case 2 : model is poor.
  $R^2 \approx 0$

Variance Inflation Factor

- $VIF = \frac{1}{1-R^2}$

$R^2 \approx 1$, VIF = V. large

multicollinearity

$R^2 \approx 0$, $VIF = 1$

No multicollinearity

Multicollinearity $\neq 1$

- Case 1: model is good, predicted value of x_1 = actual value

$R^2 \Rightarrow$ Close to 1

- Case 2: model is poor.

$R^2 \approx 0$

No multicollinearity.

Assumptions in Linear Regression

Detecting Multicollinearity includes two techniques:

1. Correlation matrix
   if except diagonal elements
   → other terms are close to 0 → No M
   → " " " " " ±1 → M

2. $VIF = \frac{1}{1-R^2}$
   (i) VIF → large → M
   (ii) VIF → small → No M

Assumptions in Linear Regression

Correlation between two variables : $$\text{Correlation} = \rho = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y}$$

$\rho = 0 \longrightarrow$ Not Correlated

$\rho = \pm 1 \longrightarrow$ Highly Correlated

4 Assumption in LR

1. Linear Relation b/w y and x

2. No Multicollineaity

3. data points Shd be independent

4. Homo Scedasticity.

Considering data of P Dimensions

Lets Practice

Based on the data provided below, answer questions from (7-10). We consider a function we wish to minimize. $J(w) = \frac{1}{10}\Sigma_i^5(y^{(i)} - w_1x^{(i)} - w_0)^2$ where the constants $x^{(i)}$, $y^{(i)}$ are provided in the table below

$$J(\omega) = \frac{1}{10}\sum_{i=1}^{5}(y_i - \omega_1 x^1 - \omega_0)^2$$

i	$x^{(i)}$	$y^{(i)}$
1	0	1
2	1	3
3	2	5
4	3	8
5	4	9

Dataset

data point
↳ y and x both.

7) The dimension of w is ________.

Considering data of P Dimensions

Lets Practice

8) Start with the initial guess of $[w_0, w_1] = [0, 0]$. Take the value of learning rate = 1. The value of w_0 after 1 iterations of gradient descent will be ________.

2

$(\omega_0 \; \omega_1) = (0, 0)$

$\eta = 1$

ω_0 after 2 iteration

$\theta \cdot \omega$

What is Multicollinearity

done

- One crucial assumption in regression models is that independent variables should not correlate among themselves. This is essential for isolating the individual impact of each variable on the target variable, as indicated by regression coefficients.
- Multicollinearity arises when variables are correlated, making it challenging to discern their separate effects on the target variable.

Linear Regression

What is Multicollinearity

done

- **Example of multicollinearity :**

done

What is Multicollinearity

done

- Why this is a problem ?
- Because in regression we are looking at how the independent variables are individually effecting the output label.

What is Multicollinearity

- How to solve the problem of multicollinearity ?

- later skip

Linear Regression

What is Multicollinearity

done

- Multicollinearity creates a problem in the multiple regression model because the inputs are all influencing each other. Therefore, they are not actually independent, and it is difficult to test how much the combination of the independent variables affects the dependent variable, or outcome, within the regression model.

How do we measure Multicollinearity?

done

- A very simple test known as the VIF test is used to assess multicollinearity in our regression model. The variance inflation factor (VIF) identifies the strength of correlation among the predictors.
- VIF help in predicting that which variable in the data is more correlated with other variables

Linear Regression

How do we measure Multicollinearity?

done

Formula and Calculation of VIF

The formula for VIF is:

$$\text{VIF}_i = \frac{1}{1 - R_i^2}$$

where:

$R_i^2 =$ Unadjusted coefficient of determination for regressing the ith independent variable on the remaining ones

Advantage

Simplest ML algo.

LR has high interpretability

- Using LR we can determine the importance of each dimension in data

$y = 3x + 10$

$y = 3x^2 + 10x + 15$

$y = 4x^3 + 3x^2 + 10x + 15$

- $$\rho_{XY} = \frac{\mathrm{Cov}(x,y)}{\sigma_x \sigma_y} \Rightarrow \frac{\sum_{i=1}^{N}(x_i-\bar{x})(y_i-\bar{y})}{\sigma_x \sigma_Y} \longrightarrow \frac{\sum_{i=1}^{N} x_i y_i}{\sigma_x \sigma_Y}$$

x	y
-2	-4
-1	-2
0	0
1	2
2	4

$\bar{x}=0$

$\bar{y}=0$

$(\rho = +ve \approx 1)$

- $\rho_{XY} = \dfrac{Cov(x,y)}{\sigma_x \sigma_y}$ $\Rightarrow$ $\dfrac{\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})}{\sigma_x \sigma_Y}$ $\longrightarrow$ $\dfrac{\sum_{i=1}^{N} x_i y_i}{\sigma_x \sigma_Y} = \dfrac{-ve}{+ve+ve}$

x	y
-2	+4
-1	+2
0	0
1	-2
2	-4

$\bar{x} = 0$

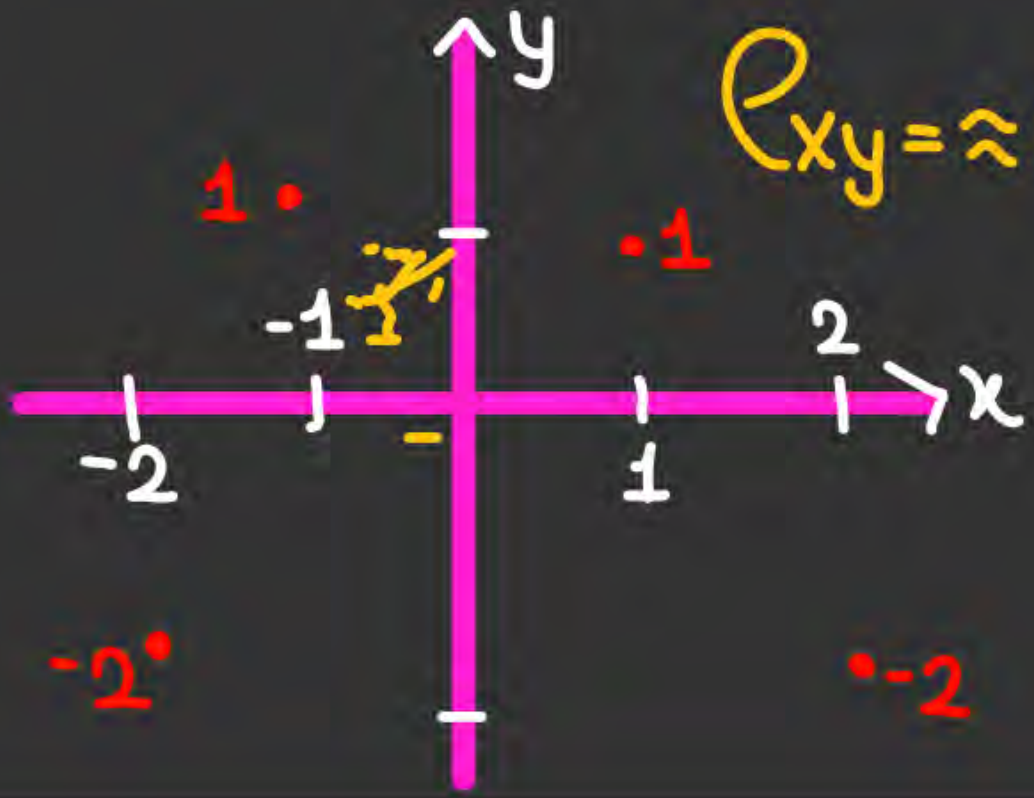
$\bar{y} = 0$

$\rho_{xy} = \approx -1$

- $\rho_{XY} = \dfrac{Cov(x,y)}{\sigma_x \sigma_y} \Rightarrow \dfrac{\sum_{i=1}^{N}(x_i-\bar{x})(y_i-\bar{y})}{\sigma_x \sigma_Y} \rightarrow \dfrac{\sum_{i=1}^{N} x_i y_i}{\sigma_x \sigma_Y} = \dfrac{-ve}{+ve+ve}$

$\bar{x}=0$
$\bar{y}=0$

x	y
+1	1
2	-2
-2	-2
-1	1



$\rho_{xy} = \approx 0.$

THANK - YOU