Data Science and Artificial Intelligence

Machine Learning

Regression

Lecture No. 06

By- SIDDHARTH SABHARWAL SIR

Recap of Previous Lecture

- Topic — Questions
- Topic — R^2 ⇒ Coeff of determination → Goodness of fit; → RSS; → TSS
- Topic — MSE
- Topic — RMSE
- Topic — *Residual plot
- Topic — H.W

Topics to be Covered

- Topic
- Topic
- Topic
- Topic
- Topic

effect of outlier

Assumption in L.R → Multicollinearity
→ VIF
→ Homo/Heteroscedasticity


Positivity
Optimism is the one quality more associated with success and happiness than any other.
BRIAN TRACY
BRIAN TRACY
INTERNATIONAL

REVISION

What is Coeff of Determination

$$R^2 = 1 - \frac{RSS}{TSS} \Rightarrow (0 \text{ to } 1)$$

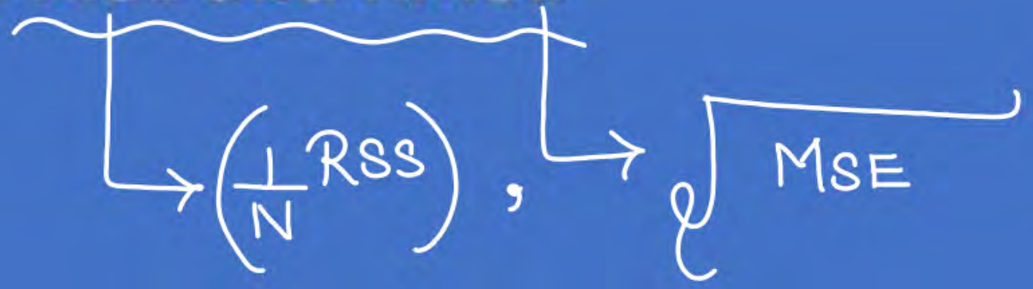
1 ⇝ overfitting

0 ⇝ the model = bekar

- goodness of fit
- It determine how much our model detect/learn the variation/pattern of y.

REVISION

What is MSE and RMSE



Basics of Machine Learning

What is Hat Matrix

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

$$\hat{Y} = X(X^TX)^{-1}X^TY$$

Hat matrix

If one dimensional data

- $y = wx \Rightarrow w = \dfrac{\sum_{i=1}^{N} x_i y_i}{\sum_{i=1}^{N} x_i^2}$

- $y = w_0 \Rightarrow w_0 = \dfrac{\sum_{i=1}^{N} y_i}{N}$

Linear Regression

Question 15: What is the purpose of the coefficient of determination (R-squared) in simple linear regression?

A. To determine the slope of the regression line
B. To measure the strength of the linear relationship
C. To calculate the p-value of the regression
D. To identify outliers in the dataset

How much the
The linear model
understand pattern of data

Linear Regression

Question 19: If the R-squared value in simple linear regression is 0.75, what does it indicate?

A. A strong linear relationship between the variables
B. A weak linear relationship between the variables
C. No linear relationship between the variables
D. The model is overfitting $\leftarrow R^2 = 1$

Simple LR

↳ Single dimension data

$y = \beta_0 + \beta_1 x$

Multiple L-R

↳ data is of more than one dimension

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

Linear Regression

Question 2: What does the coefficient of determination (R-squared) measure in multiple linear regression?

A. The correlation between predictor variables
B. The percentage of variance in the dependent variable explained by the model
C. The significance of the intercept term
D. The number of predictor variables in the model

Linear Regression

In multiple linear regression, what is the key difference between simple linear regression and multiple linear regression?

A) Simple linear regression has one independent variable, while multiple linear regression has two or more.

B) Simple linear regression uses categorical variables, while multiple linear regression uses continuous variables.

C) Simple linear regression is used for classification, while multiple linear regression is used for prediction.

D) There is no difference between simple and multiple linear regression.

Linear Regression

Which statistic is used to assess the strength and direction of the relationship between the dependent variable and each independent variable in multiple linear regression?

y $x^1, x^2, x^3 \ldots$

A) Mean absolute error (MAE)

B) R-squared (R^2) ✓✓

C) Standard error

D) Confidence interval

Linear Regression

What is the purpose of the residual plot in multiple linear regression analysis?

A) To visualize the relationship between independent variables.
B) To check for homoscedasticity and the presence of outliers.
C) To calculate the correlation coefficient (r).
D) To assess multicollinearity.

Linear Regression

What is the main purpose of the intercept term in a multiple linear regression model?

(C)

→ β_0 ⇒ Bias term.

A) It represents the slope of the regression line.

B) It is used to control for multicollinearity.

C) It represents the expected value of the dependent variable when all independent variables are zero.

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

D) It is not used in multiple linear regression.

What is the Centred data

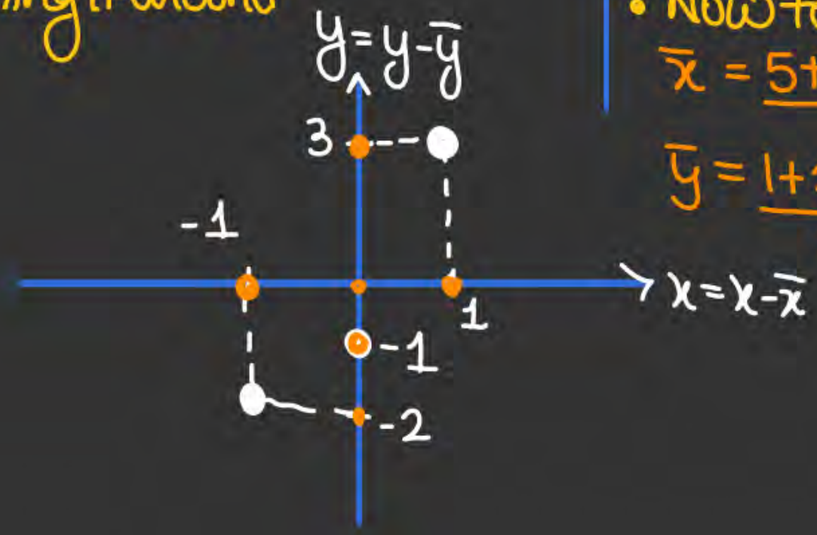
- So one Dimension data ⇒
- The data is plotted ⇒
- The Centre of the 2D Plane ⇒ origin
- Now Centering of data means, we pull the data and bring it around origin

data
(5,1) , (6,2) , (7,6)

- Now to Centre the data

$\bar{x} = \frac{5+6+7}{3} = 6$

$\bar{y} = \frac{1+2+6}{3} \Rightarrow 3$



So Centering of any data ⇒

(x_i, y_i) are datapoints

Step 1 ⇒ $\bar{x} \Rightarrow \frac{\sum x_i}{N}$

$\bar{y} = \frac{\sum y_i}{N}$

Step 2 ⇒ Now data

$(x_i - \bar{x}, y_i - \bar{y})$

This new Centred data ⇒ has mean of x and $y = 0$

- If we have a centred data ⇒ then in L.R the Bias or intercept term β_0 ⇒ Shd be ~~~ .

$$\beta_0 = \bar{y} - m\bar{x}$$

$$\beta_0 = 0 - 0$$

already done in Maths

$\Rightarrow Y, X$ are RV

$Y = aX + b$

$\Rightarrow \overline{Y} = a\overline{X} + b$

If $X_{new} = X_{old} - \overline{X_{old}}$

$\overline{X_{new}} = \overline{X_{old}} - \overline{X_{old}}$

$= 0$

2D data

(x_i^1, x_i^2, y_i)

→ Centering →

Step 1 . $\bar{y} = \frac{\sum y_i}{N}$, $\overline{x^2} = \frac{\sum x_i^2}{N}$, $\overline{x^1} = \frac{\sum x_i^1}{N}$

Step 2 ⇒ $y_{inew} = y_i - \bar{y}$

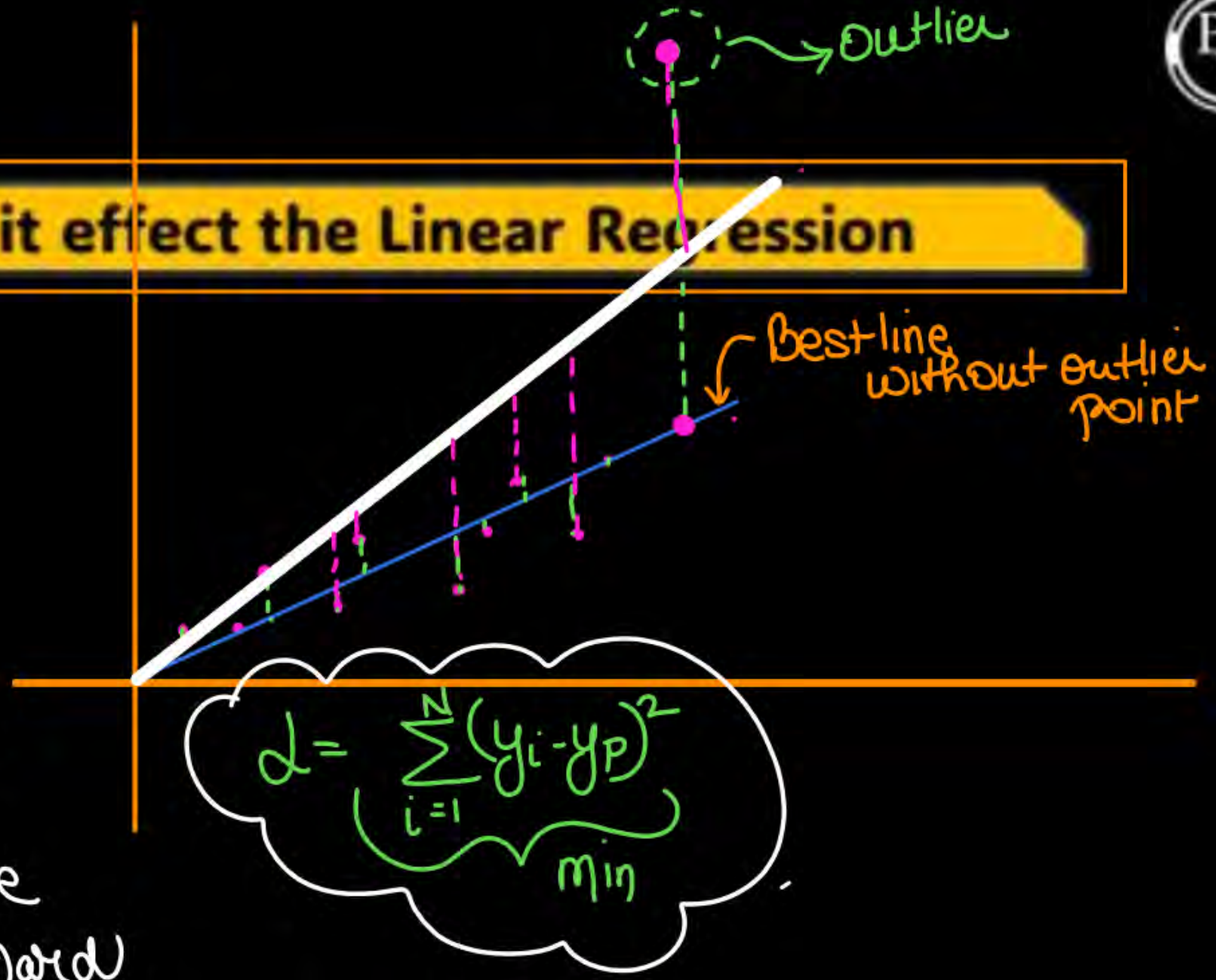
$x^2_{inew} = x_i^2 - \overline{x^2}$

$x^1_{inew} = x_i^1 - \overline{x^1}$

Linear Regression

What is an outlier and how it effect the Linear Regression

- datapoint which has huge noise and unfollow the datapattern



* So the L.R try to minimise the RSS, if the data has outlier then the L.R will be effected by this and the Resultant line will be shifted toward the outlier

So L.R algorithm donot neglect the outlier.

Assumptions in LR =>

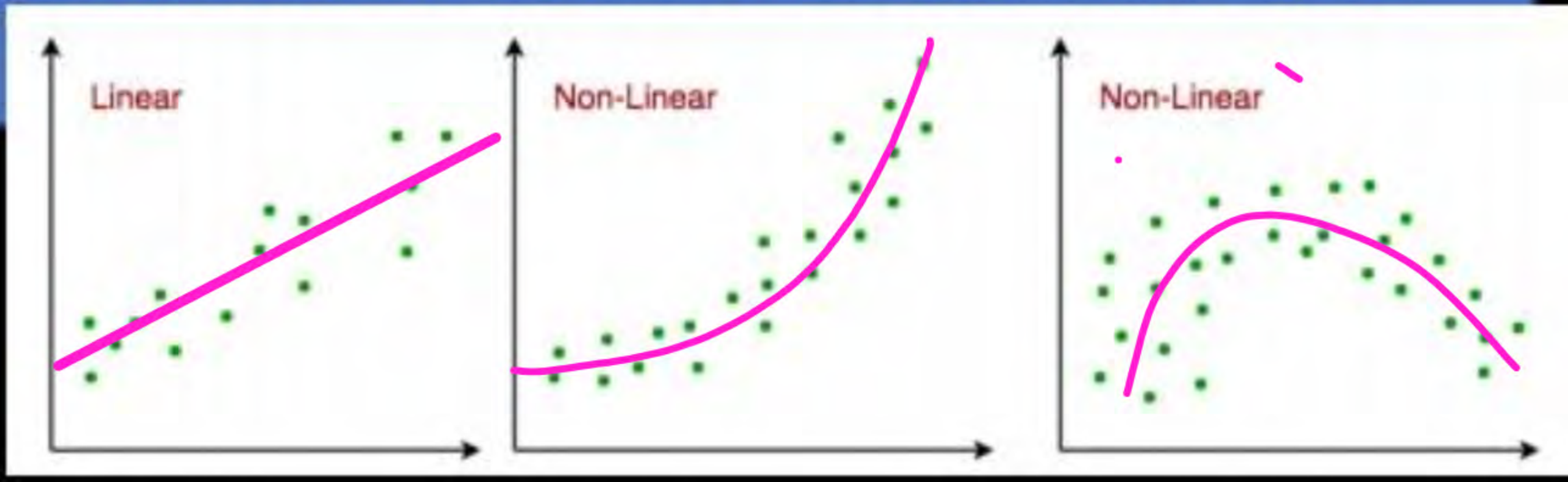
1. we assume that actual pattern of data is linear

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 variable follow those in the independent variable(s) in a linear fashion. This means that there should be a straight line that can be drawn through the data points. If the relationship is not linear, then linear regression will not be an accurate model.



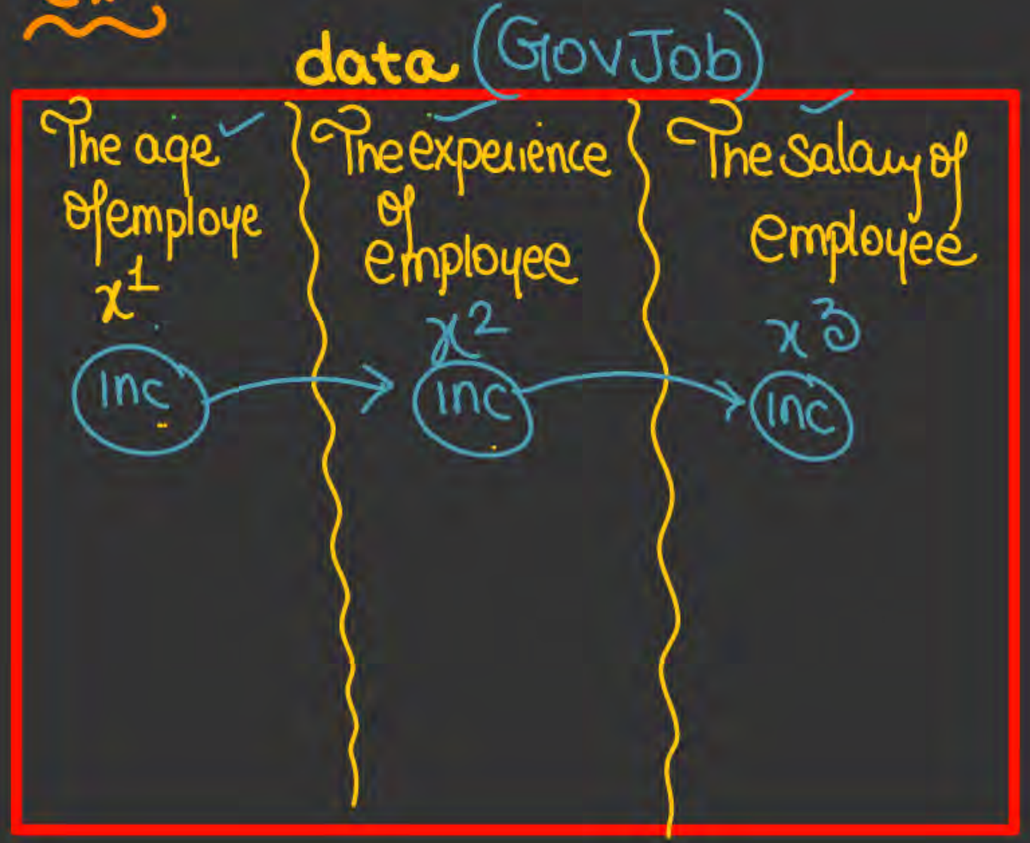
2) In L-R the model ⇛

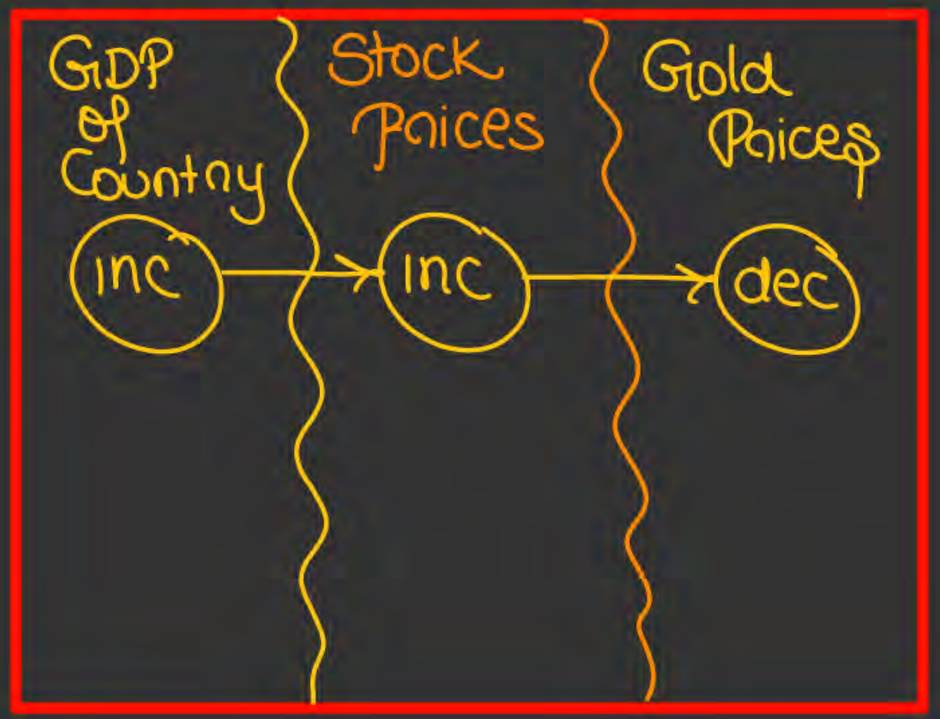
$$\left(y = \beta_0 + \beta_1 x^1 + \beta_2 x^2 + \beta_3 x^3 - - - - - -\right)$$

⇛ So from this eq, ⇛ if x^1 ⇛ changes by 1 unit then y changes by β^1

So we can say that $\beta_1, \beta_2, \beta_3$ - - - Shows contribution of 1st, 2nd, 3rd dimension when these dimension changes by 1 unit.

ex


data (GovJob)
The age of employe
x^1
The experience of employee
x^2
The salary of employee
x^3
inc
inc
inc

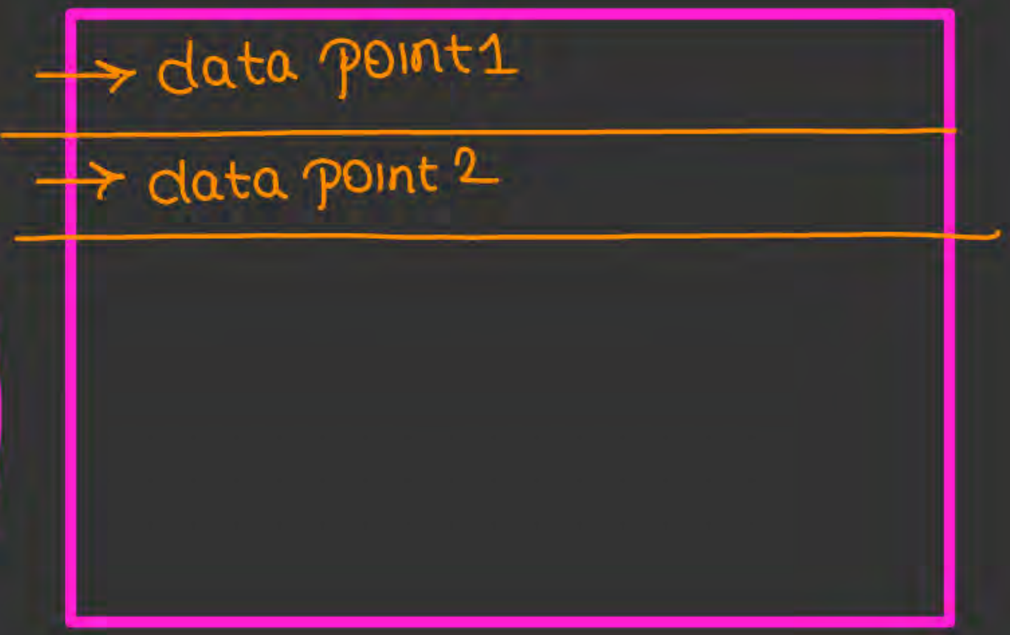

GDP of Country
Stock Prices
Gold Prices
inc
inc
dec

- So in L.R we need that the dimensions in the data shd be independent of each other.
- If data dimensions are dependent on each other then we say that data has multicollinearity ⇐
  - So L.R cannot work as desined in case of multicollinearity.

- also we assume that the data points are also independent

dependent ??

There shd be no Relationship b/w the value of one data point with any other data point.



Linear Regression

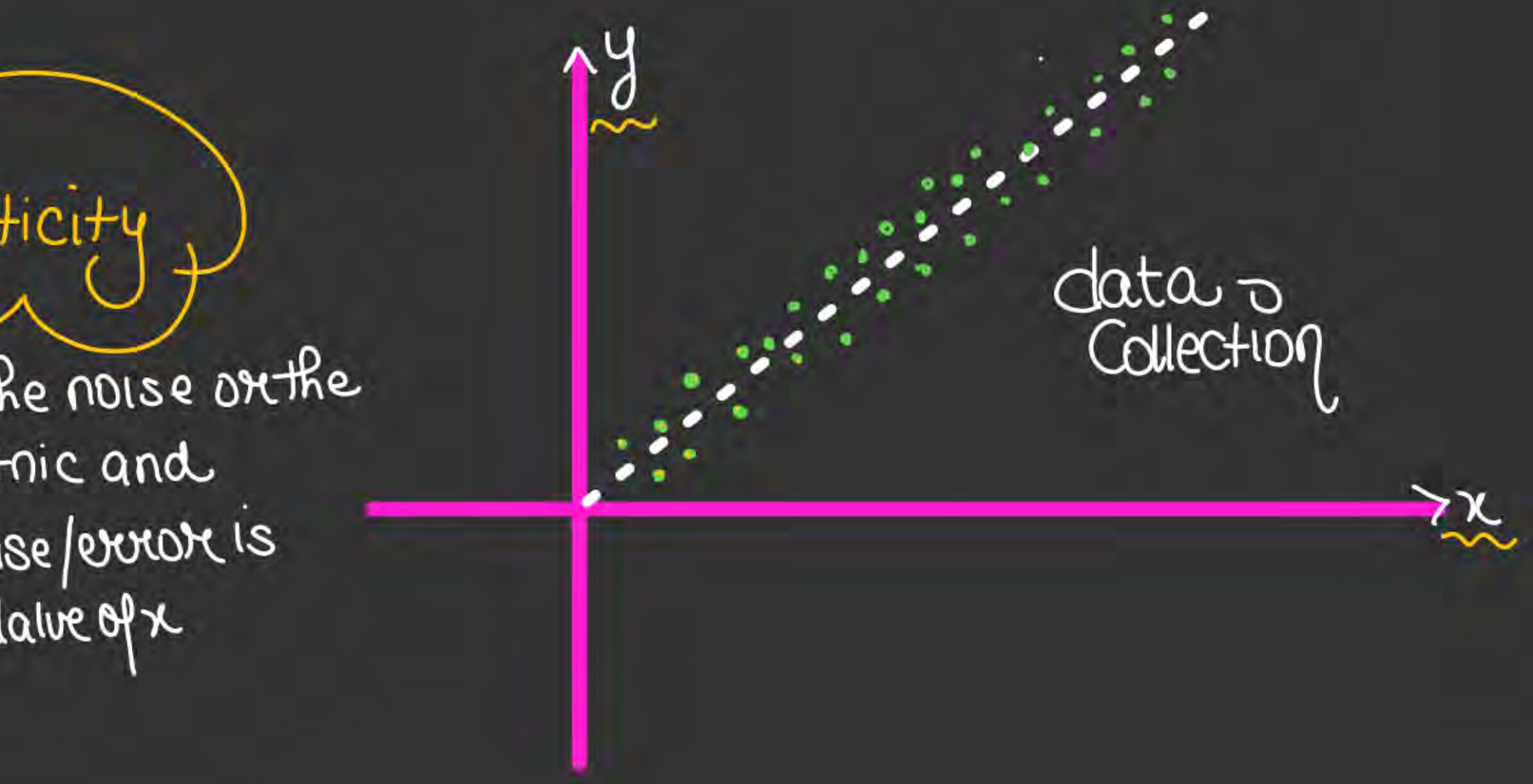
Assumptions in Linear Regression

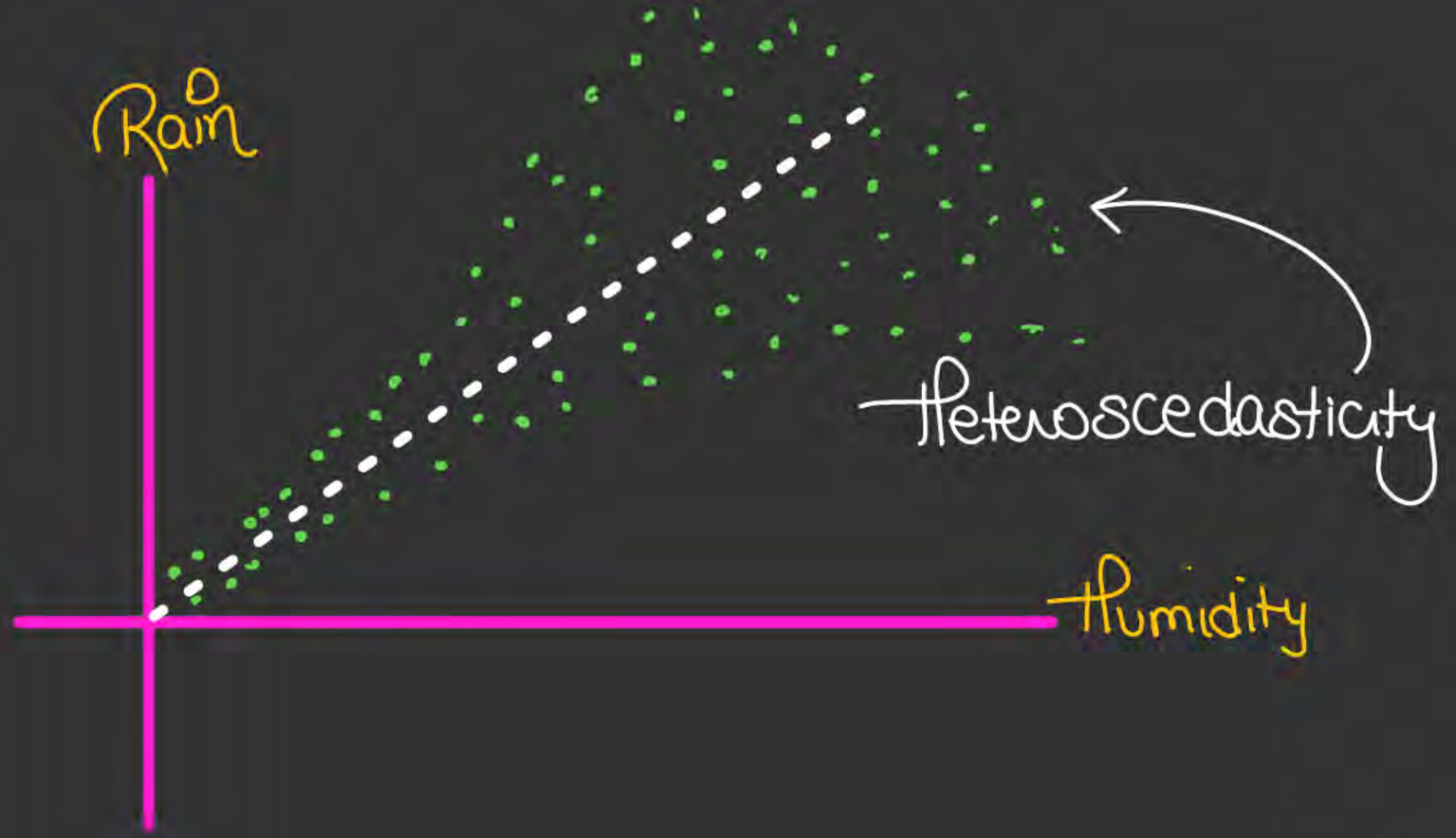
datapoints shd be independent of each other.

2. Independence: The observations in the dataset are independent of each other. This means that the value of the dependent variable for one observation does not depend on the value of the dependent variable for another observation. If the observations are not independent, then linear regression will not be an accurate model.

3. Homoscedasticity

- we assume that the noise or the error is symmetric and the amount of noise/error is independent of value of x



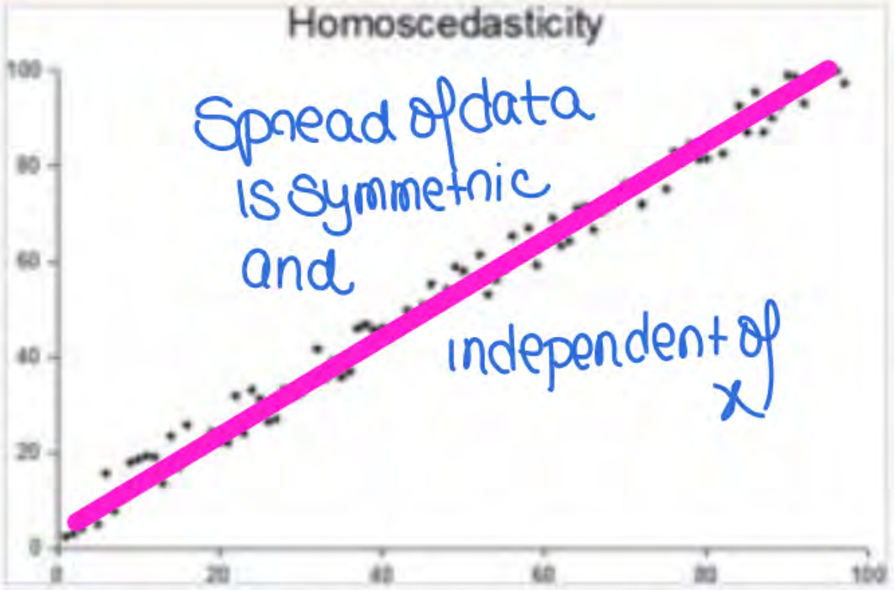

Rain
Heteroscedasticity
Humidity

So L-R will work when data has homoscedasticity.

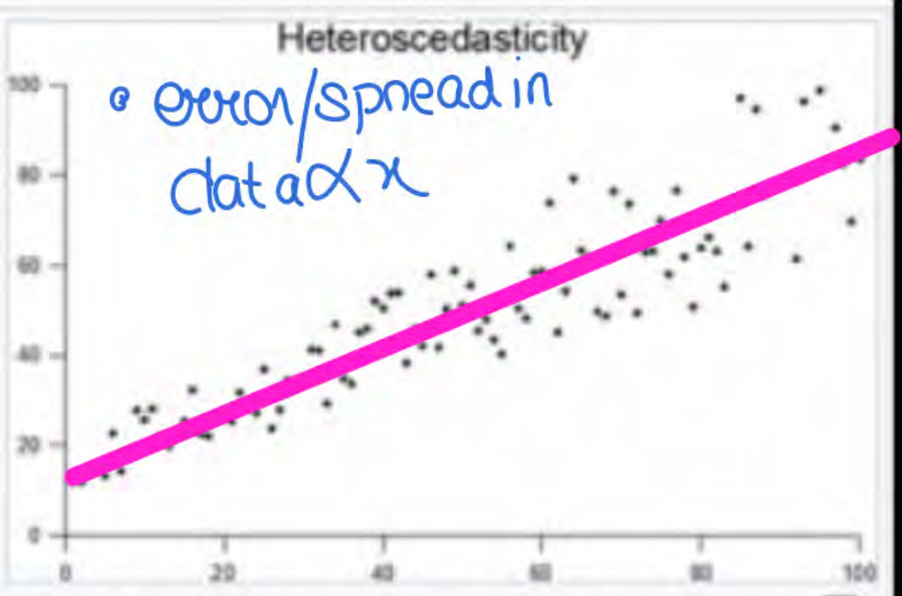
Linear Regression

Assumptions in Linear Regression

3. 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.

Linear Regression

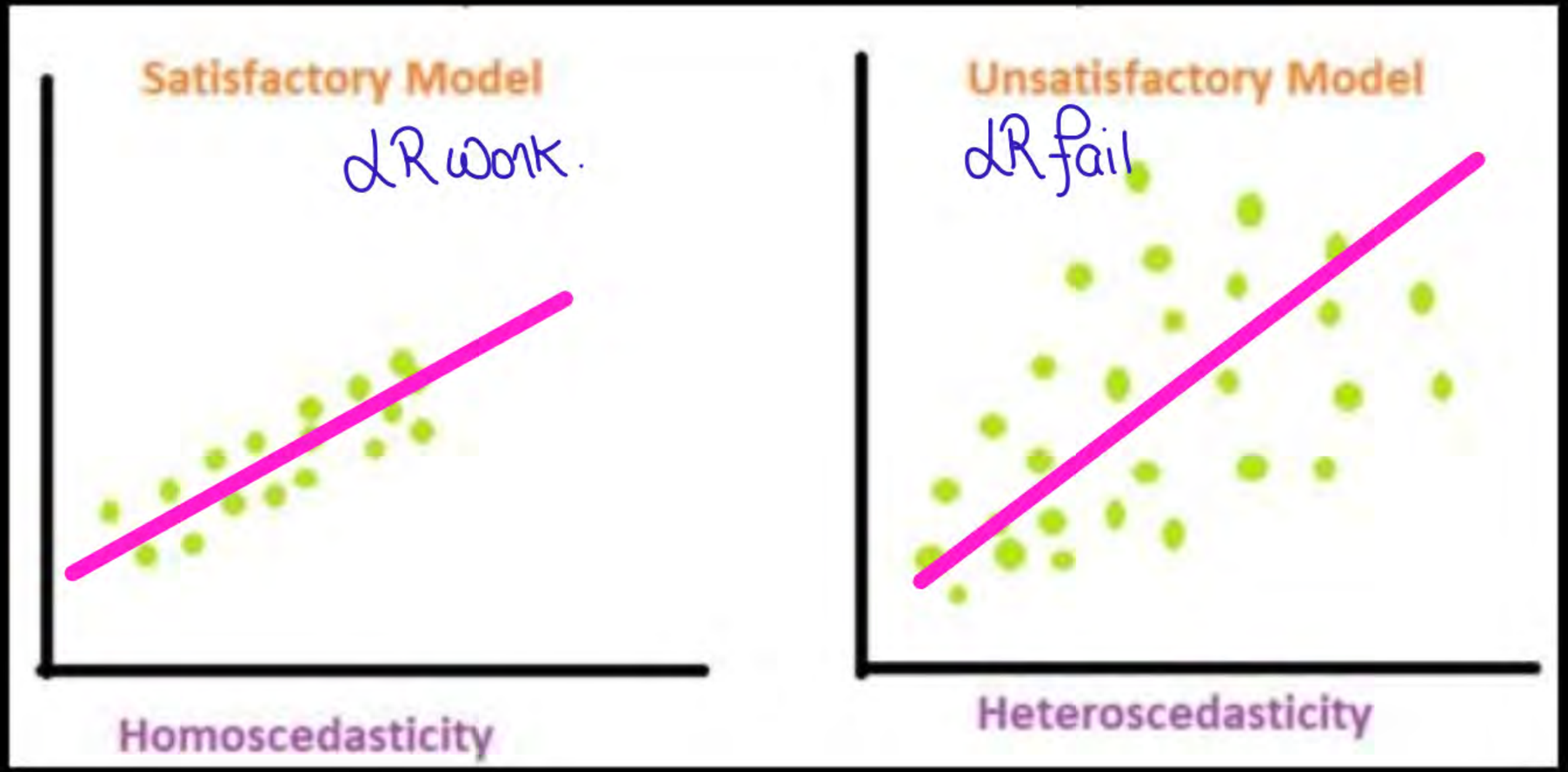
Assumptions in Linear Regression

3. we assume that data in L-R has Homoscedasticity.

In Heteroscedasticity ⇒ due to huge noise and outlier the L-R will fail.

Linear Regression

Assumptions in Linear Regression



Linear Regression

Assumptions in Linear Regression

already told in 2 ⇒ independence of dimension

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.

Linear Regression

Assumptions in Linear Regression

Detecting Multicollinearity includes two techniques:

- Correlation Matrix: Examining the correlation matrix among the independent variables is a common way to detect multicollinearity. High correlations (close to 1 or -1) indicate potential multicollinearity.
- VIF (Variance Inflation Factor): VIF is a measure that quantifies how much the variance of an estimated regression coefficient increases if your predictors are correlated. A high VIF (typically above 10) suggests multicollinearity.

Correlation

↓

Correlation b/w any two RV's ⇒ i.e dependence b/w two RV's

Correlation Coef ⇒ $\rho_{X^1X^2} = \frac{Cov(X^1, X^2)}{\sigma_{X1}\,\sigma_{X2}}$

$$Cov(X^1, X^2) = \frac{1}{N-1}\sum_{i=1}^{N}(X_i^1 - \bar{X}^1)(X_i^2 - \bar{X}^2)$$

$$\rho_{X'X'} \Rightarrow \frac{Cov(X', X')}{\sigma_{X1}\cdot\sigma_{X'}} \Rightarrow \frac{Var(X')}{Var(X')} = 1$$

- Value of Correlation Coef ⇒ -1 to 1 , if this is zero ⇒ the RV's are independent of each other

  → if this is 1, -1 ⇒ then RV's are highly dependent on each other.

How to find multicollineauty in data ⇒ → Correlation Coeff → VIF

* So if we have 3 dimensions in data then to check Correlation b/w dimensions we create a matrix

$\rho_{X^2X'} = \rho_{X'X^2}$

⇒ So if data has no multicollineauty then Values in the Correlation matrix shd be close to zero.

ρ_{xx}	X^1	X^2	X^3
X'	$\rho_{X'X'} = 1$	$\rho_{X'X^2}$	$\rho_{X'X^3}$
X^2	$\rho_{X^2X'}$	$\rho_{X^2X^2} = 1$	$\rho_{X^2X^3}$
X^3	$\rho_{X^3X'}$	$\rho_{X^3X^2}$	$\rho_{X^3X^3} = 1$

Linear Regression

Assumptions in Linear Regression

Correlation between two variables :

$$\text{Correlation} = \rho = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y}$$

done

Linear Regression

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. $J(w) = \frac{1}{10}\Sigma_i^5 (y^{(i)} - w_1 x^{(i)} - w_0)^2$ where the constants $x^{(i)}$, $y^{(i)}$ are provided in the table below

$$= \frac{1}{10} \sum_{i=1}^{5} (y^i - w_1 x^i - w_0)^2$$

$w_0, w_1 \geq 0, 0$

i	$x^{(i)}$	$y^{(i)}$
1	0	1.4822
2	0.25	1.8165
3	0.50	1.9171
4	0.75	2.3930
5	1.00	2.5826

Dataset

7) The dimension of w is _____.

Linear Regression

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 2 iterations of gradient descent will be _____.

Find w_0 after 2 iteration

Linear Regression

What is Multicollinearity

- 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

- **Example of multicollinearity :**

example

Linear Regression

What is Multicollinearity

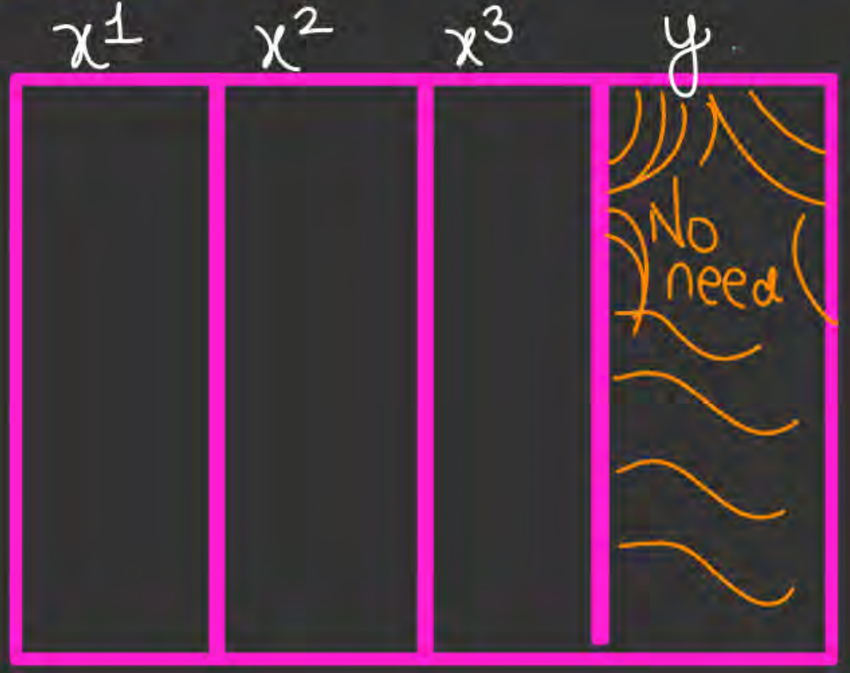
- Why this is a problem ?
- Because in regression we are looking at how the independent variables are individually effecting the output label.

done

So why we need VIF (Variance Inflation factor)

- ~~*~~ in Corelation matrix we can see dependence of one dimension on other individually

- * To see whole data together we use VIF



- Here we want to check linear dependence of one dimension on others

if $x^1 = a + bx^2 + cx^3$

So taking $x^1 \approx$ label now perform L-R

$x^1 = a + bx^2 + cx^3$

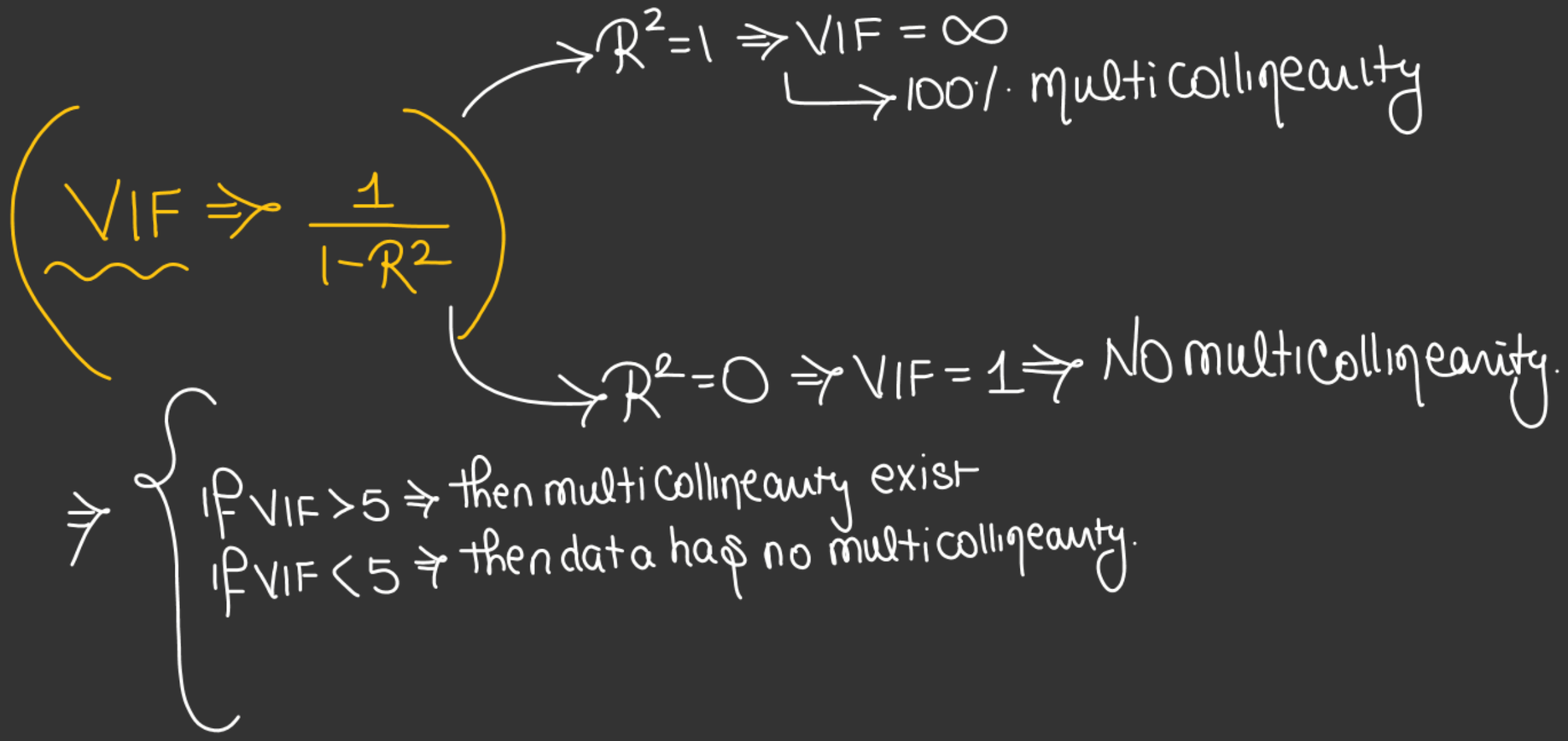
⇓

model

Now R^2 of this model $(R^2 = 1)$ → model = Pattern of data → 100% dependent data

If $R^2 \to 0$ the model is not able to understand pattern of x^1.

So independent data


VIF ⇒ 1/(1−R²)
R²=1 ⇒ VIF = ∞
100% multicollinearity
R²=0 ⇒ VIF = 1 ⇒ No multicollinearity.
⇒
if VIF > 5 ⇒ then multicollinearity exist
if VIF < 5 ⇒ then data has no multicollinearity.

Linear Regression

What is Multicollinearity

- **How to solve the problem of multicollinearity ?**

Dimension ≯ features

*By feature selection method we remove the highly dependent dimensions

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.

Linear Regression

How do we measure Multicollinearity?

- 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?

Formula and Calculation of VIF

The formula for VIF is:

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

∞ = M·C exist

1 ≠ M·C donot exist.

where:

R_i^2 = [illegible] coefficient of determination for regressing the ith independent variable on the remaining ones

Linear Regression

Advantage & Disadvantage of Linear Regression

Advantages	Disadvantages
Linear Regression is simple to implement and easier to interpret the output coefficients.	On the other hand in linear regression technique outliers can have huge effects on the regression and boundaries are linear in this technique.
When you know the relationship between the independent and dependent variable have a linear relationship, this algorithm is the best to use because of it's less complexity compared to other algorithms.	Diversely, linear regression assumes a linear relationship between dependent and independent variables. That means it assumes that there is a straight-line relationship between them. It assumes independence between attributes.
Linear Regression is susceptible to over-fitting but it can be avoided using some dimensionality reduction techniques, regularization (L1 and L2) techniques and cross-validation.	But then linear regression also looks at a relationship between the mean of the dependent variables and the independent variables. Just as the mean is not a complete description of a single variable, linear regression is not a complete description of relationships among variables.

Linear Regression

Why Linear Regression is Important

The interpretability of linear regression is a notable strength.
The model's equation provides clear coefficients that elucidate the impact of each independent variable on the dependent variable, facilitating a deeper understanding of the underlying dynamics.
Its simplicity is a virtue, as linear regression is transparent, easy to implement, and serves as a foundational concept for more complex algorithms.

Linear Regression

Space and Time Complexity of Linear Regression

Assumptions:

n = number of training examples, **m** = number of features, **n'** = number of support vectors,
k = number of neighbors, **k'** = number of trees

- **Linear Regression**
  - Train Time Complexity=O(n*m^2 + m^3)
  - Test Time Complexity=O(m)
  - Space Complexity = O(m)

Question 12: In simple linear regression, which variable is considered the independent variable?

H.W.

A. The variable being predicted
B. The response variable
C. The predictor variable
D. There is no independent variable in simple linear regression

Question 19: If the R-squared value in simple linear regression is 0.75, what does it indicate?

A. A strong linear relationship between the variables
B. A weak linear relationship between the variables
C. No linear relationship between the variables
D. The model is overfitting

Question 20: Which of the following statements is true regarding the residual plot in simple linear regression?

A. Residuals should exhibit a clear linear pattern.
B. Residuals should be randomly scattered around the horizontal line.
C. Residuals should be negatively correlated with the predictor variable.
D. Residuals should have a positive correlation with the dependent variable.

5.FOr a give N independent input variables (X1,X2... Xn) and dependent (target) variable Y a linear regression is fitted for the best fit line using least square error on this data. The correlation coefficient for one of it's variable(Say X1) with Y is -0.97. Which of the following is true for X1? ?

- A) Relation between the X1 and Y is weak
- B) Relation between the X1 and Y is strong
- C) Relation between the X1 and Y is neutral
- D) Correlation does not imply relationship

6.Given below characteristics which of the following option is the correct for Pearson correlation between V1 and V2? If you are given the two variables V1 and V2 and they are following below two characteristics. 1. If V1 increases then V2 also increases 2. If V1 decreases then V2 behavior is unknown ?

- A) Pearson correlation will be close to 1
- B) Pearson correlation will be close to -1
- C) Pearson correlation will be close to 0
- D) None of these

1) A regression analysis is inappropriate when;
   a) you have two variables that are measured on an interval or ratio scale.
   b) you want to make predictions for one variable based on information about another variable.
   c) the pattern of data points forms a reasonably straight line.
   d) there is heteroscedasticity in the scatter plot.

2) In regression analysis, the variable that is being predicted is;

a) the independent variable

b) the dependent variable

c) usually denoted by x

d) usually denoted by r

3) In the regression equation $y = b_o + b_1x$, b_o is the;

a) slope of the line

b) independent variable

c) y intercept

d) coefficient of determination

6) Least square method calculates the best-fitting line for the observed data by minimizing the sum of the squares of the _____ deviations.

a) **Vertical**
b) Horizontal
c) Both of these
d) None of these

7) Which one is the least square method formula;

a) $\min \sum(y_i - \hat{y}_i)^2$

b) $\min \sum(\hat{y}_i - y_i)$

c) **$\min \sum(y_i - \hat{y}_i)^2$**

d) $\min \sum(y_i - \hat{y}_i)$

13) Below you are given a summary of the output from a simple linear regression analysis from a sample of size 15, SSR=100, SST = 152. The coefficient of determination is;

a) 0.5200

b) 0.6579

c) 0.8111

d) 1.52

10) **A residual is defined as**

a) The difference between the actual Y values and the mean of Y.

b) The difference between the actual Y values and the predicted Y values.

c) The predicted value of Y for the average X value.

d) The square root of the slope.

11) If the regression equation is equal to $y=23.6-54.2x$, then 23.6 is the _____ while -54.2 is the _____ of the regression line.

a) Slope, intercept

b) Slope, regression coefficient

c) Intercept, slope

d) Radius, intercept

Q8. Suppose we have N independent variables (X1, X2... Xn) and Y's dependent variable.

Now Imagine that you are applying linear regression by fitting the best-fit line using the least square error on this data. You found that the correlation coefficient for one of its variables (Say X1) with Y is -0.95.

Which of the following is true for X1?

A) Relation between the X1 and Y is weak

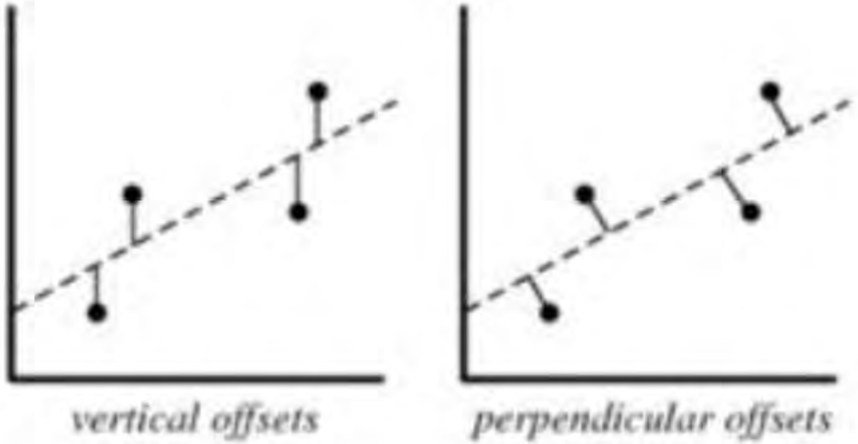
B) Relation between the X1 and Y is strong

C) Relation between the X1 and Y is neutral

D) Correlation can't judge the relationship

Solution: (B)

Q11. Suppose the horizontal axis is an independent variable and the vertical axis is a dependent variable. Which of the following offsets do we use in linear regression's least square line fit?



B) Perpendicular offset

C) Both, depending on the situation

D) None of above

Q12. True- False: Overfitting is more likely when you have a huge amount of data to train.

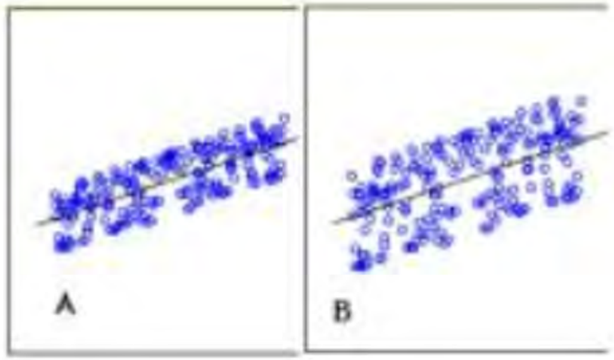
A) TRUE

B) FALSE

Solution: (B)

Q14. Which of the following statement is true about the sum of residuals of A and B?

Below graphs show two fitted regression lines (A & B) on randomly generated data. Now, I want to find the sum of residuals in both cases, A and B.



A) A has a higher sum of residuals than B

B) A has a lower sum of residual than B

C) Both have the same sum of residuals

D) None of these

Q18. Which of the following statement is true about outliers in Linear regression?

A) Linear regression is sensitive to outliers

B) Linear regression is not sensitive to outliers

C) Can't say

D) None of these

Q19. Suppose you plotted a scatter plot between the residuals and predicted values in linear regression and found a relationship between them. Which of the following conclusion do you make about this situation?

A) Since there is a relationship means our model is not good

B) Since there is a relationship means our model is good

C) Can't say

D) None of these

Suppose that you have a dataset D1 and you design a linear model of degree 3 polynomial and find that the training and testing error is "0" or, in other words, it perfectly fits the data.

Q20. What will happen when you fit a degree 4 polynomial in linear regression?

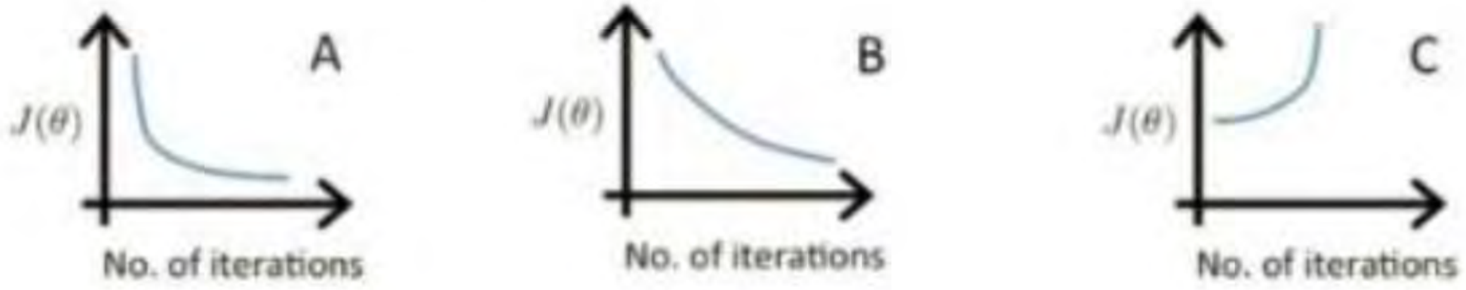
A) There is a high chance that degree 4 polynomial will overfit the data

B) There is a high chance that degree 4 polynomial will underfit the data

C) Can't say

D) None of these

Below are three graphs, A, B, and C, between the cost function and the number of iterations, I1, I2, and I3, respectively.



Q23. Suppose I1, I2, and I3 are the three learning rates for A, B, and C, respectively. Which of the following is true about I1,I2, and I3?

A) $I2 < I1 < I3$

B) $I1 > I2 > I3$

C) $I1 = I2 = I3$

D) None of these

Intercept

$$SE(b_0) = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n-2}} * \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}}$$

Slope

$$SE(b_1) = \sqrt{\frac{\frac{\sum_{i=1}^{n} e_i^2}{n-2}}{\sum_{i=1}^{n} (x_i - \bar{x})^2}}$$

THANK - YOU