Data Science and Artificial Intelligence

Machine Learning

Regression

Lecture No. 06

By- SIDDHARTH SABHARWAL SIR

Recap of Previous Lecture

Topic — R^2

Topic — TSS/RSS/MSE/RMSE

Topic — Homework.

Topic

Topic

Topics to be Covered

Topic — Hat matrix

Topic — What is Outlier and it's effect on LR

Topic — Assumption in LR

Topic

Topic

Optimistic
+ve thoughts.

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

What is Coeff of Determination

$$R^2 = 1 - \frac{RSS}{TSS}$$

$R^2 = 0 \longrightarrow$ underfit, bekar model

$R^2 = 1 \longrightarrow$ overfit

$\longrightarrow$ goodness of fit.

What is MSE and RMSE

- $\frac{RSS}{N}$
- $\sqrt{\frac{RSS}{N}}$

Analysis $\Rightarrow \mathcal{L} = \sum_{i=1}^{N} (y_i - y_i')^2$

$$\sum_{i=1}^{N} \left(y_i - (\underbrace{m x_i + c})\right)^2$$

min L

$$\Rightarrow \frac{\partial L}{\partial m} = 0 \quad \frac{\partial L}{\partial c} = 0$$

$$\Rightarrow [X^T X]\,\beta = (X^T Y)$$

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

Known

$$(X^T X)\beta = (X^T Y)$$

(we are finding β)

$$\boxed{X^T X \beta = X^T Y}$$

model

$y = mx + c$

$y = 3x + 2$

data

x	y	$\hat{y}$
1	3	5
4	6	14
-2	7	-4
-3	9	-7

What is Hat Matrix

- X matrix $\rightarrow \begin{bmatrix} 1 & x_1^1 & x_1^2 & \cdots & x_1^D \\ 1 & x_2^1 & x_2^2 & \cdots & x_2^D \\ 1 & x_3^1 & x_3^2 & \cdots & \\ \vdots & & & & \\ 1 & x_N^1 & x_N^2 & \cdots & x_N^D \end{bmatrix}$

- Y matrix $\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}$

  actual Values

- B matrix $= \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_D \end{bmatrix}$

$$Q(X\beta) = ?$$

$$\begin{bmatrix} 1 & x_1^1 & \cdots & x_1^D \\ 1 & x_2^1 & & x_2^D \\ 1 & x_3^1 & & \vdots \\ \vdots & & & \\ 1 & x_N^1 & & x_N^D \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_D \end{bmatrix}$$

Model: $y = \beta_0 + \beta_1 x^1 + \beta_2 x^2 - - - \beta_D x^D$

$Q(X\beta) = ?$

$$\begin{bmatrix} 1 & x_1^1 & \cdots & x_1^D \\ 1 & x_2^1 & & x_2^D \\ 1 & x_3^1 & & \vdots \\ \vdots & & & \vdots \\ 1 & x_N^1 & & x_N^D \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_D \end{bmatrix} \rightarrow \begin{bmatrix} \beta_0 + \beta_1 x_1^1 + \beta_2 x_1^2 - - - \beta_D x_1^D \\ \beta_0 + \beta_1 x_2^1 + \beta_2 x_2^2 - - - \beta_D x_2^D \\ \vdots \\ \beta_0 + \beta_1 x_N^1 + \beta_2 x_N^2 - - - \beta_D x_N^D \end{bmatrix}$$

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

not much Imp

What is Hat Matrix

So we can say that

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

$$\Rightarrow X\left[(X^TX)^{-1}(X^TY)\right] = \hat{Y}$$

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

So $\underbrace{\left(X(X^TX)^{-1}X^T\right)}_{\Downarrow}Y = \hat{Y}$

Hat matrix

What is Hat Matrix

Linear Regression

$$MSE \Rightarrow \frac{.5^2 + .8^2 + 1.1^2}{3} \Rightarrow .7$$

A simple linear regression model is given by:

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

• $y = \beta_1 x + \beta_0 = .5x + 1$

Given the following dataset:

x	y	$\hat{y}$	Error
1	2	1.5	.5
2	2.8	2.0	.8
3	3.6	2.5	1.1

If the initial model parameters are $\beta_0 = 1$ and $\beta_1 = 0.5$, compute the Mean Squared Error (MSE).

$\beta_0 = 1 \quad \beta_1 = .5$

.7

$$X = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix}$$

$$Y = \begin{bmatrix} 2 \\ 2.8 \\ 3.6 \end{bmatrix}$$

$$\hat{Y} = X\beta = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix} \begin{bmatrix} 1 \\ .5 \end{bmatrix}$$

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \begin{bmatrix} 1 \\ .5 \end{bmatrix}$$

$$\hat{Y} = \begin{bmatrix} 1.5 \\ 2 \\ 2.5 \end{bmatrix}$$

Linear Regression

In a linear regression model, the cost function is:

2Mark

$$J(\beta_0, \beta_1) = \frac{1}{2m}\sum_{i=1}^{m}(y_i - (\beta_0 + \beta_1 x_i))^2 = \frac{1}{2\times 2}\sum_{i=1}^{2}(y_i - \beta_0 - \beta_1 x_i)^2$$

Given the dataset:

$$\partial J/\partial\beta_1 = \frac{1}{4}\, 2 \sum_{i=1}^{2} -(y_i - \beta_0 - \beta_1 x_i)\, x_i$$

x	y
1	3
2	5

If $\beta_0 = 1$ and $\beta_1 = 1$, compute the update for β_1 after one step of gradient descent with a learning rate $\alpha = 0.1$.

$$\boxed{\beta_1^{new} = \beta_1^{old} - \alpha \left.\frac{\partial J}{\partial \beta_1}\right|_{@\beta_1^{old}\beta_0^{old}}}$$

(A) 1.5

(B) 1.25 1.25

(C) 1.3

(D) 1.4

$$= \frac{1}{2\times 2}\sum_{i=1}^{2}(y_i - \beta_0 - \beta_1 x_i)^2$$

- $$\partial J/\partial \beta_1 = \frac{1}{4}\, 2 \sum_{i=1}^{2} -(y_i - \beta_0 - \beta_1 x_i)\, x_i \Big|_{@\, \beta_0 = \beta_1 = 1}$$

$$= \frac{-1}{2}\left[(y_1 - 1 - x_1)x_1 + (y_2 - 1 - x_2)x_2\right]$$

$$= \frac{-1}{2}\left[(3-1-1)1 + (5-1-2)2\right] \Rightarrow \frac{-1}{2}\left[1+4\right] = -5/2$$

$$\boxed{\begin{aligned} \beta_1^{new} &= \beta_1^{old} - \alpha \frac{\partial J}{\partial \beta_1}\Big|_{@\, \beta_1^{old}\, \beta_0^{old}} \\ &= 1 - \cdot 1\,(-2\cdot 5) \Rightarrow 1\cdot 25 \end{aligned}}$$

Given the following actual and predicted values:

y_{actual}	$y_{predicted}$
3	2.8
5	4.9
7	6.5

If the mean of actual y values is 5, what is the R^2 score?

(A) 0.94

(B) 0.96

(C) 0.98

(D) 0.92

$$RSS = \sum_{i=1}^{3}(y_i - y_i')^2$$

$$= .2^2 + .1^2 + .5^2$$

$$= .3$$

$$TSS = \sum_{i=1}^{3}(y_i - \bar{y})^2$$

$$= 2^2 + 0 + 2^2 \Rightarrow 8$$

$$R^2 = 1 - \frac{RSS}{TSS} \Rightarrow 1 - .3/8 \Rightarrow .9625$$

Consider a linear regression model where the feature x is scaled by a factor of 10 (i.e., $x' = x/10$). How does the estimated coefficient β_1' of the new model compare to the original β_1?

(A) $\beta_1' = \beta_1$

(B) $\beta_1' = 10\beta_1$ ✓

(C) $\beta_1' = \beta_1/10$

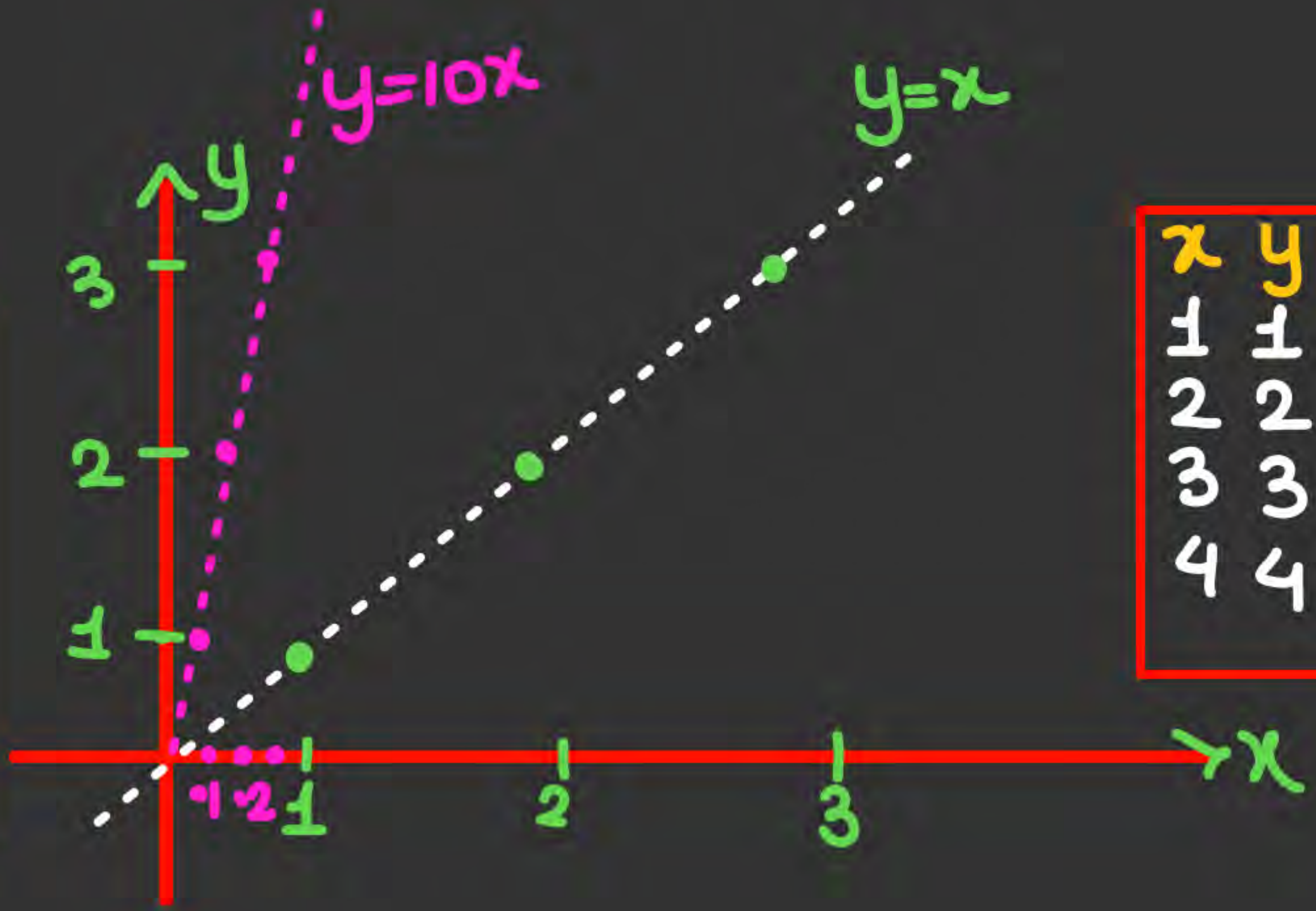
(D) No change in coefficients

$y = \beta_1 x + \beta_0 \Rightarrow$ ex. $y = 3x + 5 = 30\frac{x}{10} + 5$

x	y
x_1	y_1
x_2	y_2
⋮	⋮

x	y
$\frac{x_1}{10}$	y_1
$\frac{x_2}{10}$	y_2
⋮	⋮

x	y
.1	1
.2	2
.3	3
.4	4



x	y
1	1
2	2
3	3
4	4

Simple LR

1D LR

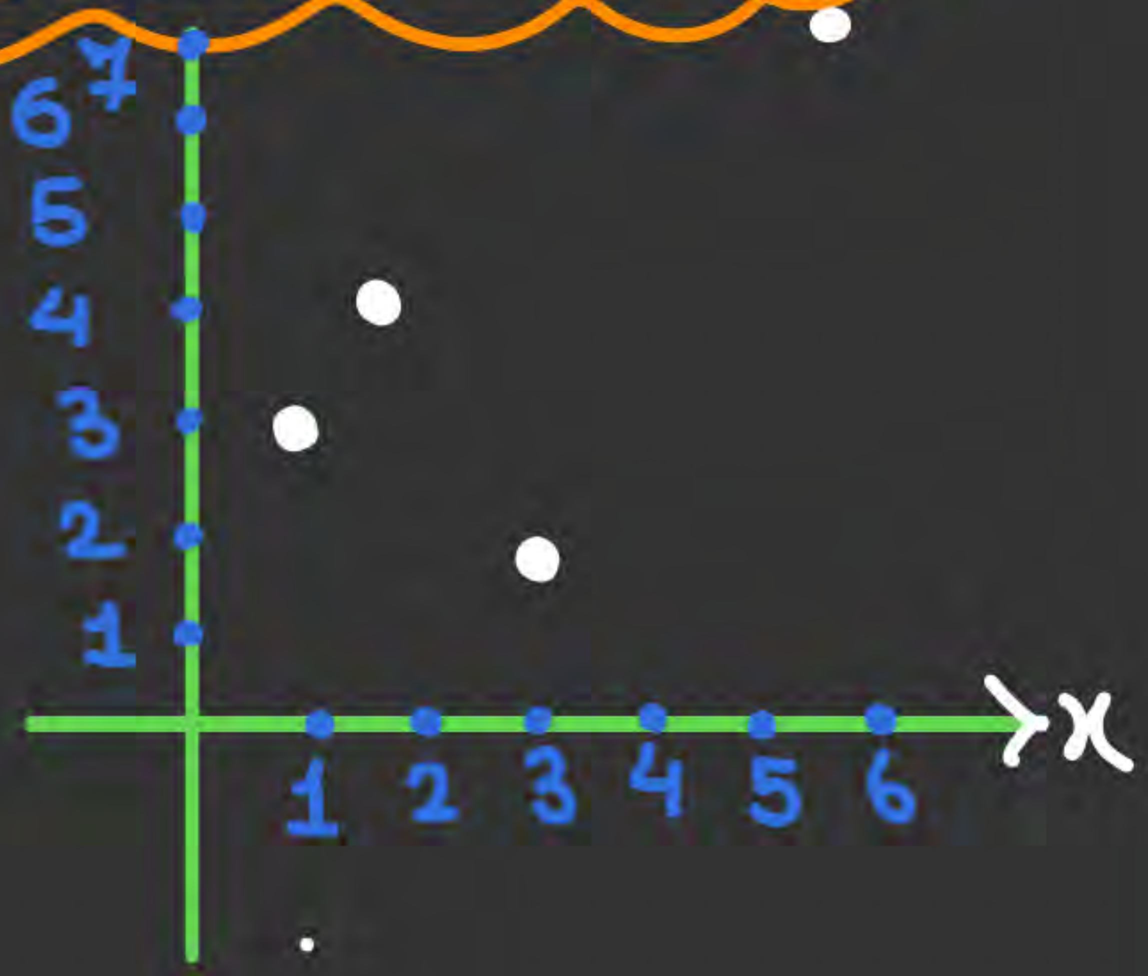
$y = mx + c$

Multiple LR

When data has more than 1D

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

What is a Centred data



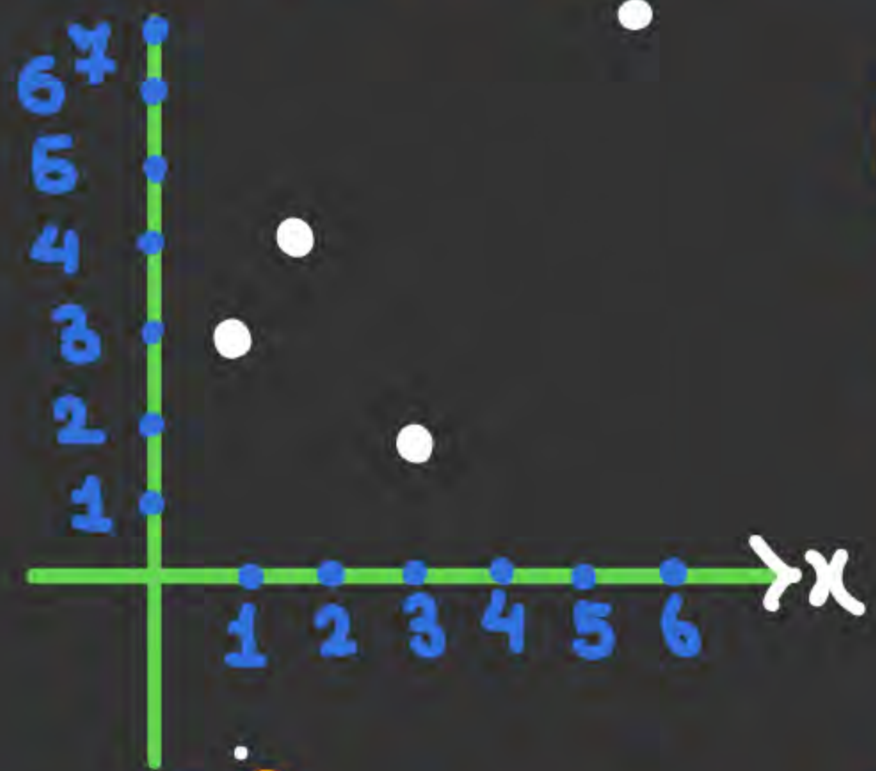
1D data

x	y
1	3
2	4
3	2
6	7

x_c	y_c
-2	-1
-1	0
0	-2
3	3

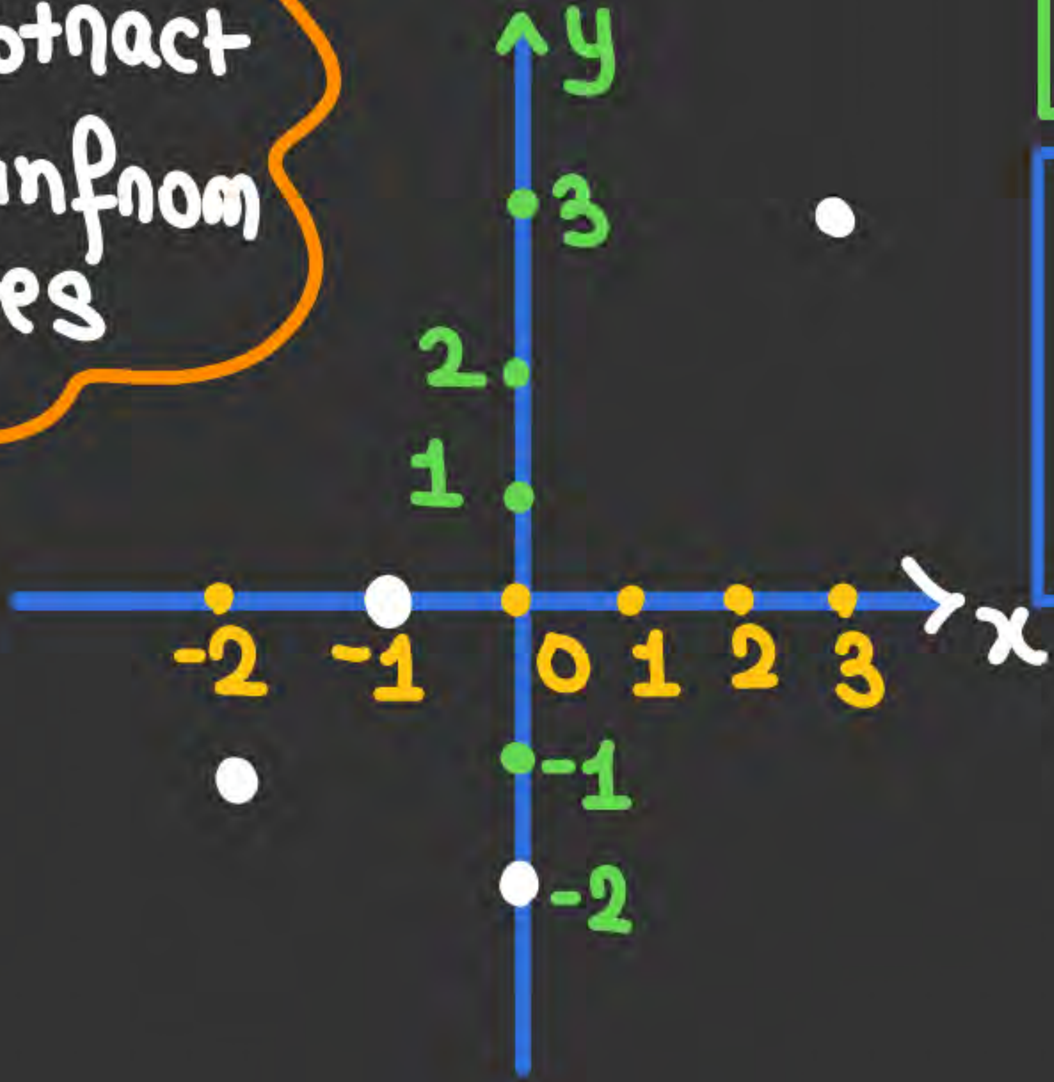
Now find mean of all columns of data and then subtract the mean from all values

What is a Centred data



Now find mean of all columns of data and then subtract the mean from all values

* After centering of data the data is spread all around origin



1D data

x	y
1	3
2	4
3	2
6	7

x_c	y_c
-2	-1
-1	0
0	-2
3	3

* After centering mean value of all column = 0

Question 5: In a simple linear regression analysis, if the mean of the dependent variable (Y) is 50, and the slope coefficient (a) is 3, what is the mean of the predictor variable (X)

a) Cannot be determined without the value of the intercept

b) Cannot be determined without the value of the intercept

c)Can be determined without the value of the intercept

$\bar{y} = 50$ $\bar{x} = ?$

$m = 3$

$y = mx + c$

$\bar{y} = m\bar{x} + c$

$\bar{y} = 3\bar{x} + c$

$50 = 3\bar{x} + c$

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 → To find goodness of fit of LR model.

C. To calculate the p-value of the regression

D. To identify outliers in the dataset

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

A. The correlation between predictor variables
B. $\boxed{\text{The percentage of variance in the dependent variable explained by the model}}$ ✓
C. The significance of the intercept term
D. The number of predictor variables in the model

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

↳ 1D ↳ more than 1D

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.

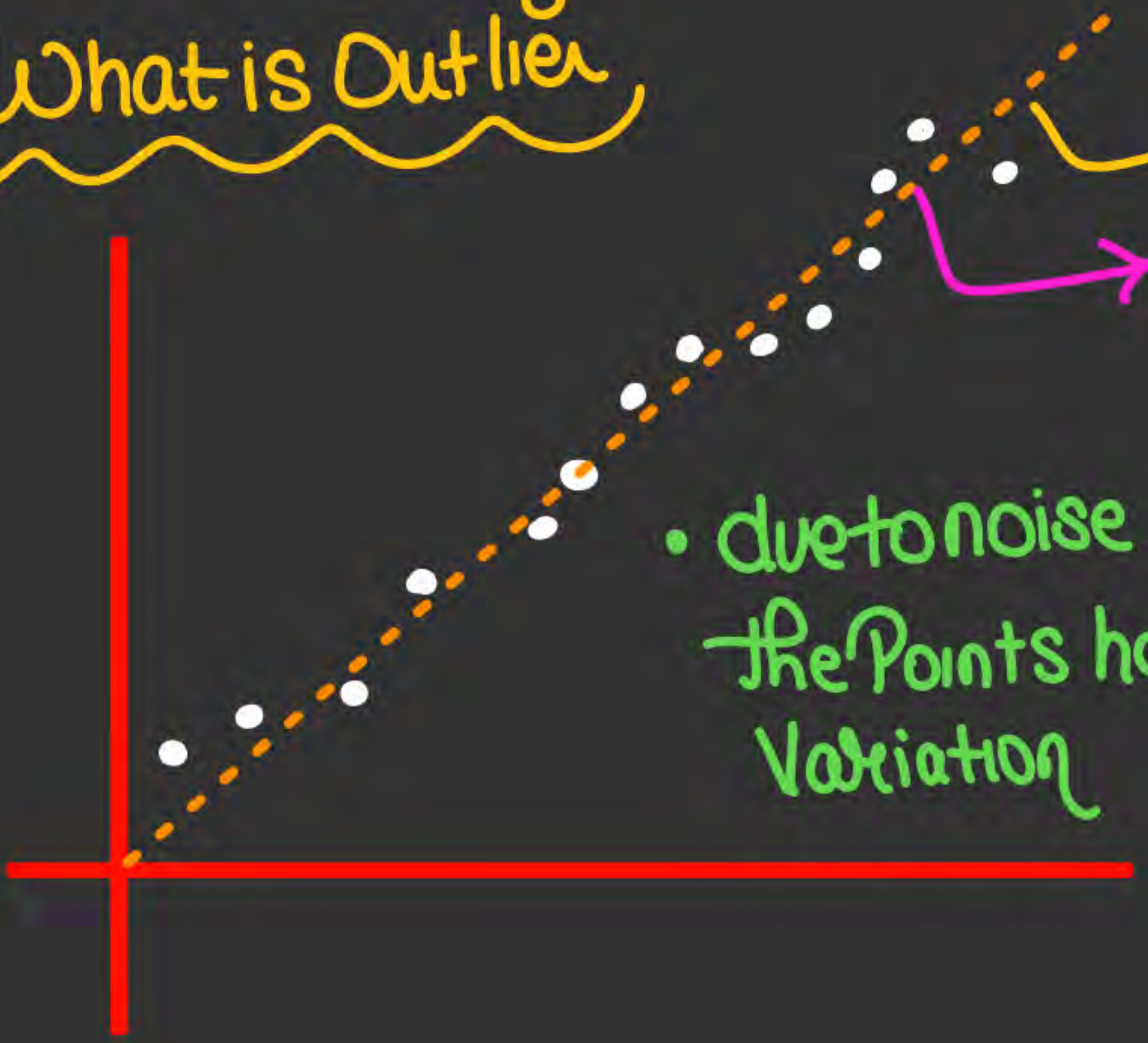
C) Simple linear regression is used for classification, while multiple linear regression is used for prediction.

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

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?

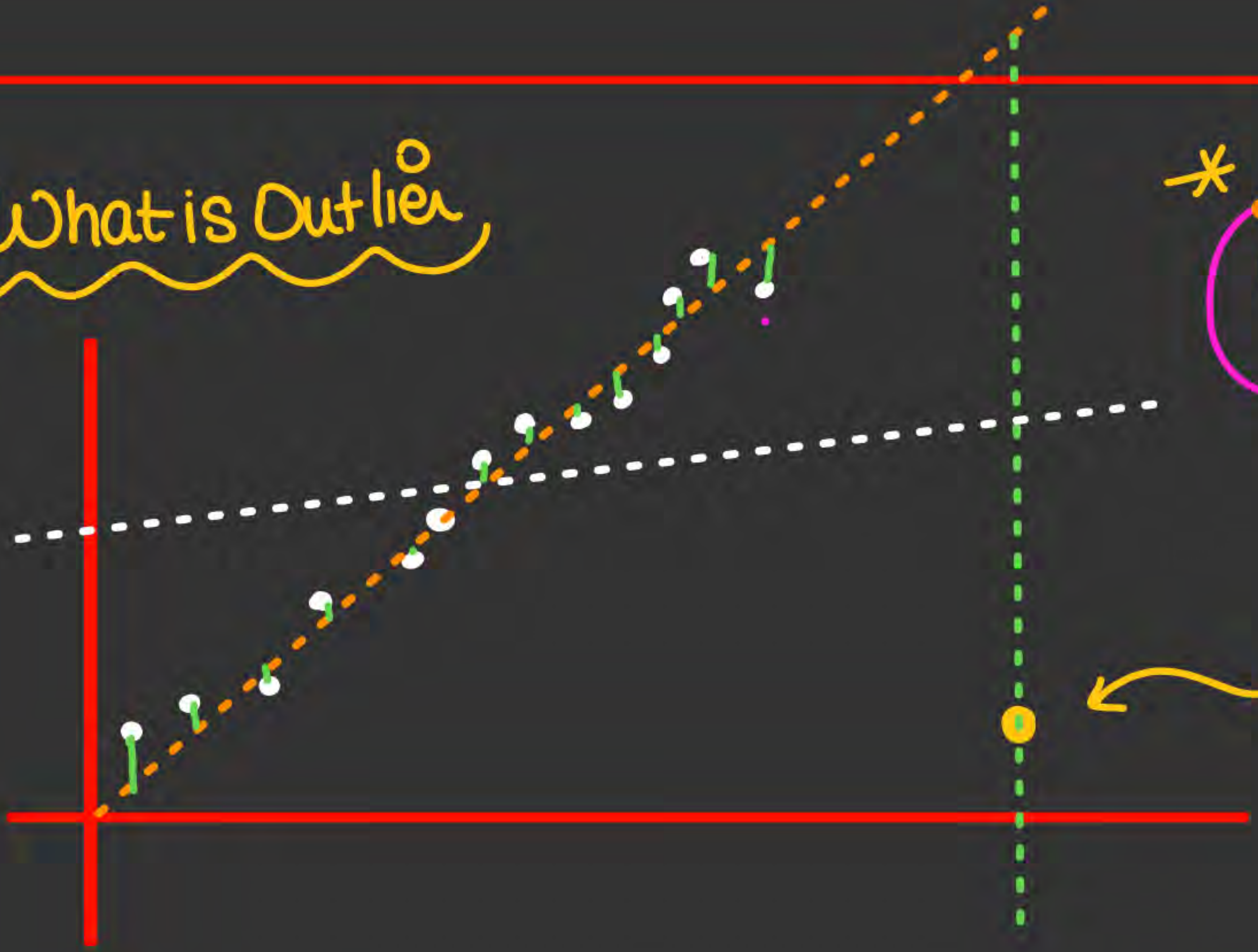
A) Mean absolute error (MAE)
B) R-squared (R^2)
C) Standard error
D) Confidence interval

What is Outlier



- due to noise in data the Points have small Variation

What is Outlier



Very large noise and this is Outlier

* LR is effected by Outlier = ??

$$\min \sum_{i=1}^{N} (y_i - y_i')^2$$

So LR $\Rightarrow$ Fail in Case of outlier

Outlier create a poor model

SKip

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.

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

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.

D) It is not used in multiple 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

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 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

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.

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

A) It represents the slope of the regression line.

B) It is used to control for multicollinearity.

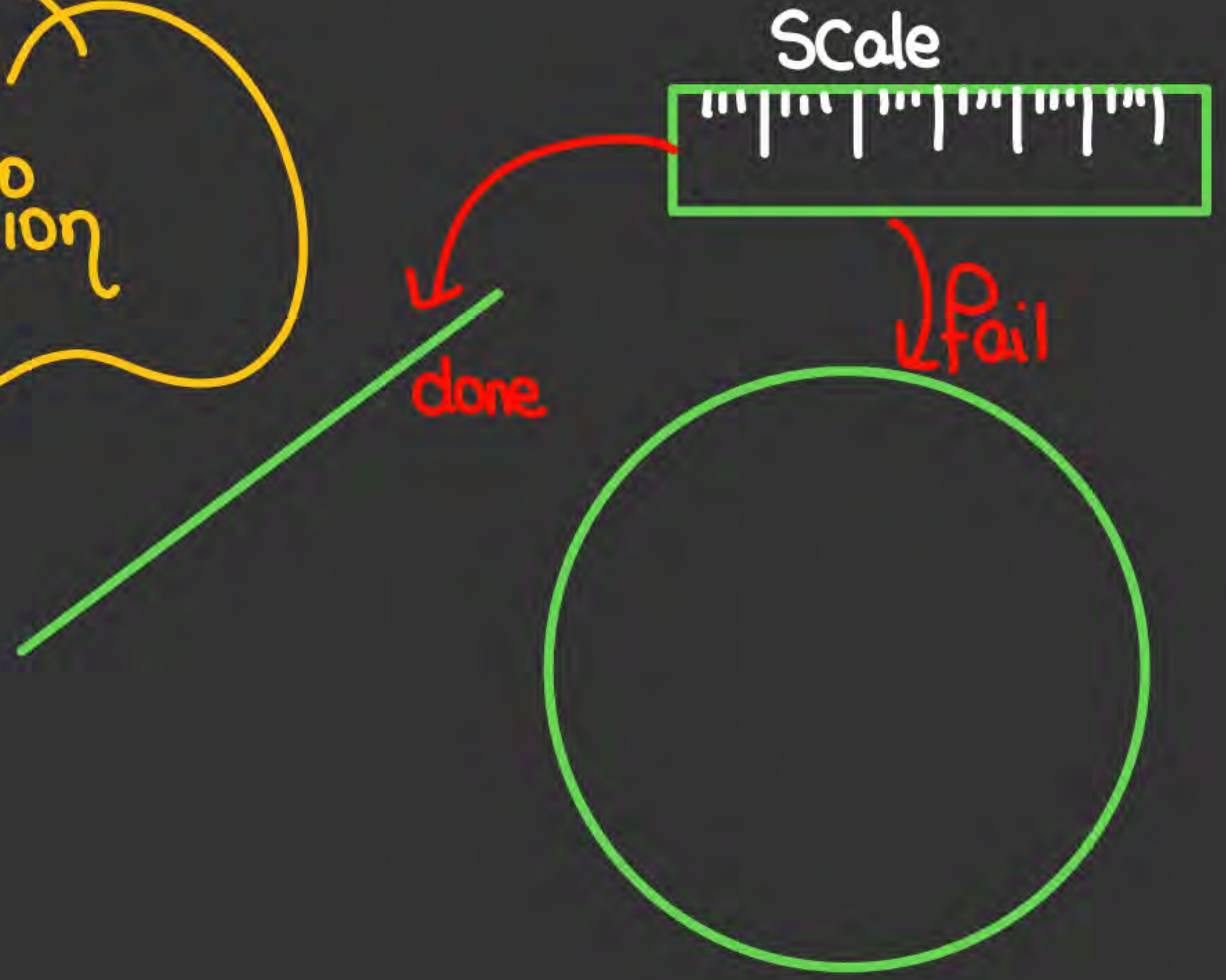
C) It represents the expected value of the dependent variable when all independent variables are zero.

D) It is not used in multiple linear regression.

What is an outlier and how it effect the Linear Regression

done

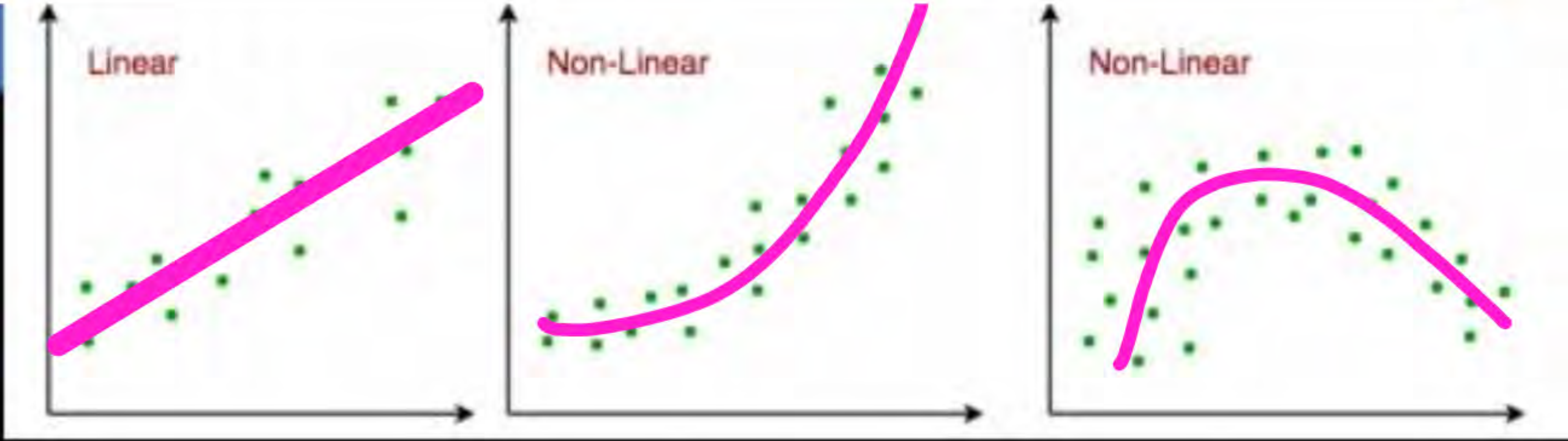
Assumption in Linear Regression



Assumptions in Linear Regression

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

1. Linearity: while applying LR we assume that y and x independent & dependent variable have linear Relationship.



Important Property of LR → Multiple Dimension

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

* Because model is created from data

β_1 : Coeff of 1st Dim
β_2 : " " 2nd "
β_3 : " " 3rd "
so on

- So these coefficient show contribution of each dimension in model/data.

2^{nd} Assumption in LR $\Rightarrow$ The dimensions / attributes / features Shd be independent of each other.

data

Temp	Usage of Ac's	Electricity Bill
inc	inc	inc
dec	dec	dec

• dependent features

- If features of data are dependent on each other then it is said that data has multicollinearity.

- For LR we need No multicollinearity.

Assumptions in Linear Regression

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.

THANK - YOU