Regression Analysis

answers

x	y
-1	0
0	1
1	2
2	3
3	4

$y = x + 1$

$= -1 + 1$

$= 0$ ✓

$y = 0 + 1$

$= 1$ ✓

$y = x - 1$

$= -1 - 1$

$= -2$ ✗

$y = x + 2$

$= -1 + 2$

$= 1$ ✗

$y = mx + c$

$y = x + 1$

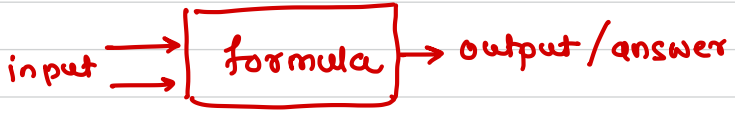
$m = 1$, $c = 1$

β estimator (coefficients)

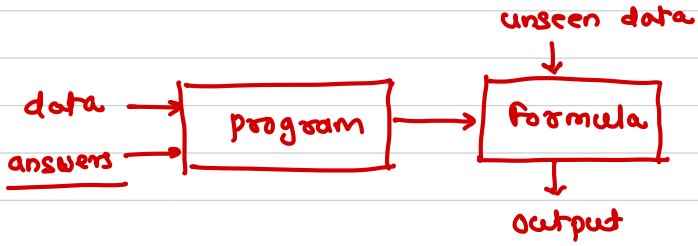
linear regression

$y = \beta_0 + \beta_1 x_1$

traditional



machine learning



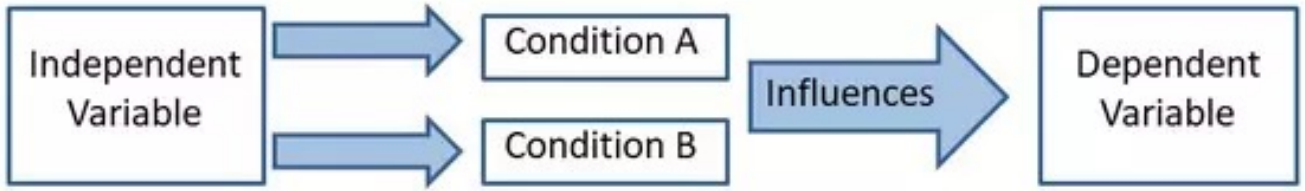
Covariance vs Correlation

Covariance	Correlation
It shows the extent to which two variables are dependent on each other	It measures the strength of two variables considering other conditions are constant
Range is $-\infty \; to \; +\infty$	Range is -1 to +1
It is affected by scale of variables	It is not affected by scale
It has definite units as it is derived by multiplying two numbers and they units	It is unit less

Terminology

- **Independent variable**
  - A variable that represents a quantity that is being manipulated in an experiment
  - Represents input
  - Also known as regressors in a statistical context.
  - x is often the variable used to represent the independent variable in an equation

- **Dependent variable**
  - A quantity whose value *depends* on how the independent variable is manipulated
  - Represents output
  - y is often the variable used to represent the independent variable in an equation



What is regression analysis

- Regression is a basic and commonly used type of predictive analysis
- The dictionary meaning of the word Regression is 'Stepping back' or 'Going back'
- Set of statistical processes for estimating the relationships between a dependent variable and one or more independent variables
- It attempts to establish the functional relationship between the variables and thereby provide a mechanism for prediction or forecasting
- The overall idea of regression is to examine two things
  - does a set of predictor variables do a good job in predicting an outcome (dependent) variable?
  - Which variables in particular are significant predictors of the outcome variable, and in what way do they– indicated by the magnitude and sign of the beta estimates–impact the outcome variable?
- These regression estimates are used to explain the relationship between one dependent variable and one or more independent variables

Applications of regression analysis

- It helps in the formulation and determination of functional relationship between two or more variables
- It helps in establishing a cause and effect relationship between two variables in economics and business research
- It helps in predicting and estimating the value of dependent variable as price production sales etc
- It helps to measure the variability or spread of values of a dependent variable with respect to the regression line
- In the field of business regression is widely used by businessmen in
  - Predicting future production
  - Investment analysis
  - Forecasting on sales etc.

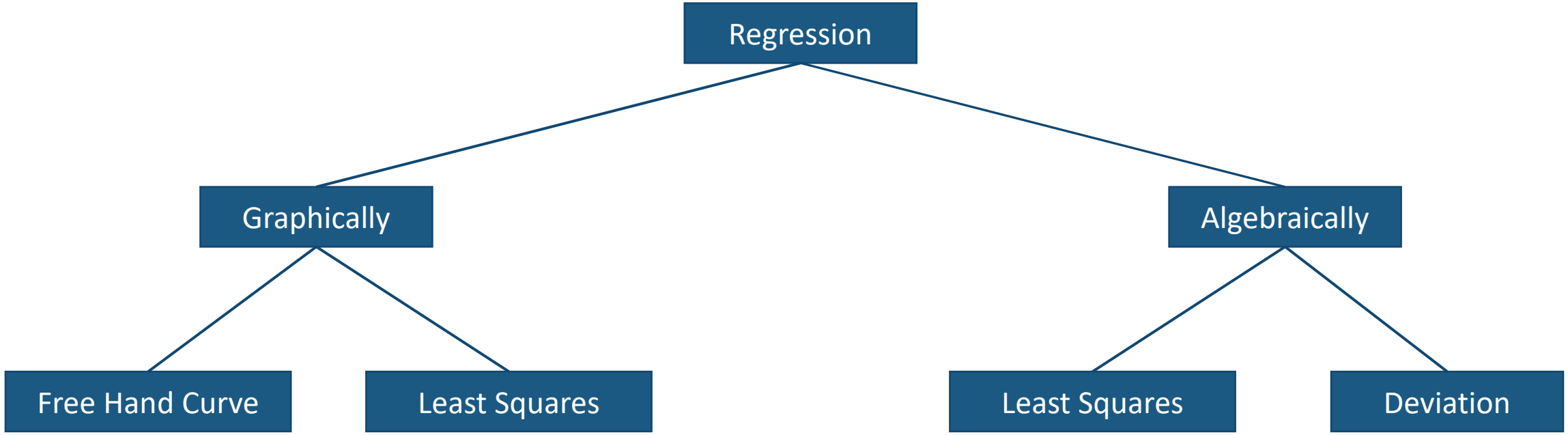
Where can it be used ?

- In the advertising business, an application delivering targeted advertisements
- In e-commerce, a batch application filtering customers to make more relevant commercial offers or an online app recommending products to buy on the basis of ephemeral data such as navigation records
- In the credit or insurance business, an application selecting whether to proceed with online inquiries from users, basing its judgment on their credit rating and past relationship with the company

Correlation vs Regression

Correlation	Regression
'Correlation' as the name says it determines the interconnection or a co-relationship between the variables.	'Regression' explains how an independent variable is numerically associated with the dependent variable.
In Correlation, both the independent and dependent values have no difference.	However, in Regression, both the dependent and independent variable are different.
The primary objective of Correlation is, to find out a quantitative/numerical value expressing the association between the values.	When it comes to regression, its primary intent is, to reckon the values of a haphazard variable based on the values of the fixed variable.
Correlation stipulates the degree to which both of the variables can move together.	However, regression specifies the effect of the change in the unit, in the known variable(p) on the evaluated variable (q).
Correlation helps to constitute the connection between the two variables.	Regression helps in estimating a variable's value based on another given value.

Methods of studying regression



Regression Equation

Adv	Sales
100	500
90	400
80	450
95	510
150	??

Sales depends on advertisement
Y depends on X [Y on X]

$(Y - \bar{Y}) = b_{yx} (X - \bar{X})$

$$b_{yx} = \frac{n \sum XY - \sum X \sum Y}{n \sum X^2 - (\sum X)^2}$$

$$b_{yx} = r \frac{\sigma y}{\sigma x}$$

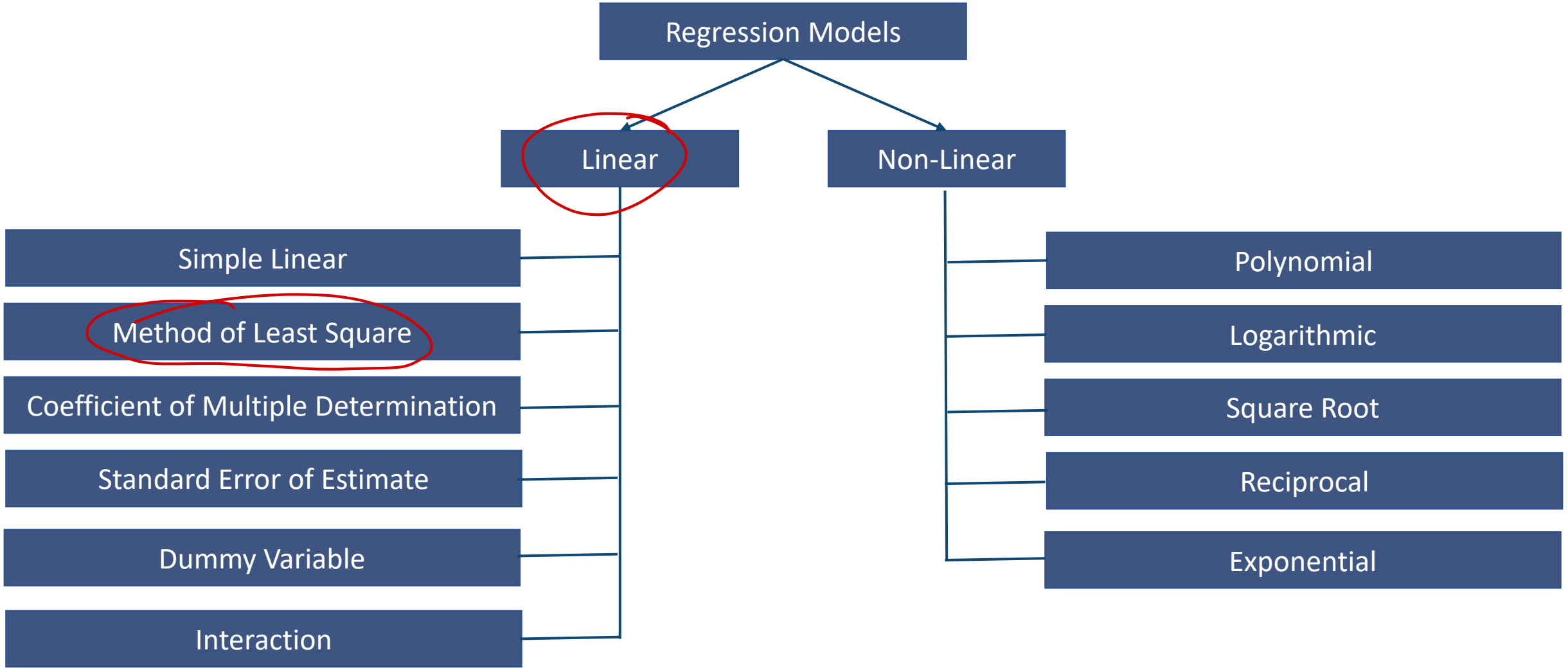
$$b_{yx} = \frac{cov(x, y)}{(\sigma x)^2}$$

Regression Equation

X	Y
3	11
4	12
8	9
7	3
2	5

1. What likely to be the value of Y if X = 10
2. What likely to be the value of X if Y = 10

Regression Models



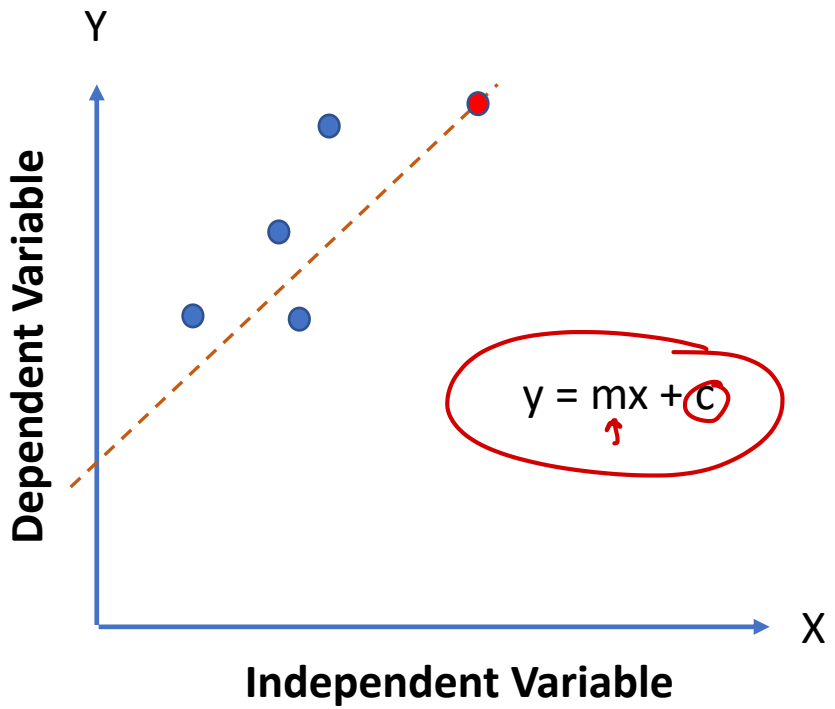
Regression Types

- Linear Regression
- Polynomial Regression
- Logistic Regression
- Ridge Regression
- Lasso Regression
- Elastic Net Regression
- Support Vector Regression
- Quantile Regression
- Principle Component Regression
- Partial Least Square Regression
- Ordinal Regression
- Poisson Regression
- Negative Binomial Regression
- Cox Regression

Linear Regression

Overview

- The data in Linear Regression is modelled using a straight line
- It is used with continuous variable
- It gives a future value as an output
- To calculate accuracy following methods are used
  - R-squared
  - Adjusted R-squared

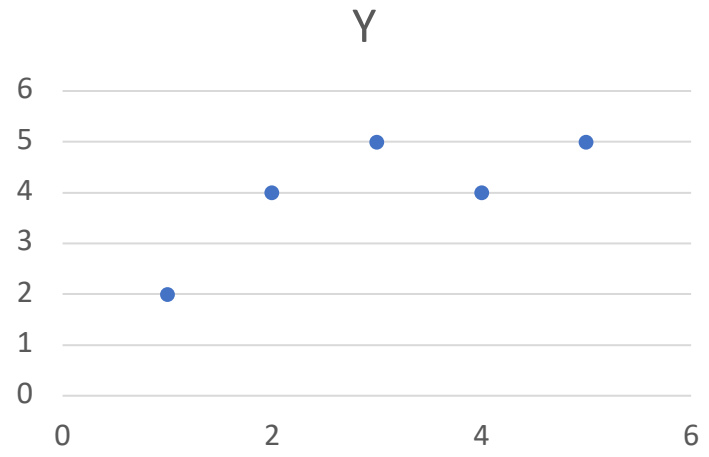


Least Square Method

- A form of mathematical regression analysis used to determine the line of best fit for a set of data, providing a visual demonstration of the relationship between the data points
- Each point of data represents the relationship between a known independent variable and an unknown dependent variable
- The least squares method provides the overall rationale for the placement of the line of best fit among the data points being studied
- It aims to create a straight line that minimizes the sum of the squares of the errors that are generated by the results of the associated equations, such as the squared residuals resulting from differences in the observed value, and the value anticipated, based on that model
- It begins with a set of data points to be plotted on an x- and y-axis graph
- An analyst using the least squares method will generate a line of best fit that explains the potential relationship between independent and dependent variables.

Least Square Method

X	Y
1	2
2	4
3	5
4	4
5	5



X	Y	$(X - \bar{X})$	$(Y - \bar{Y})$	$(X - \bar{X})^2$	$(X - \bar{X})(Y - \bar{Y})$
1	2				
2	4				
3	5				
4	4				
5	5				

$$m = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sum (X - \bar{X})^2}$$

R-squared

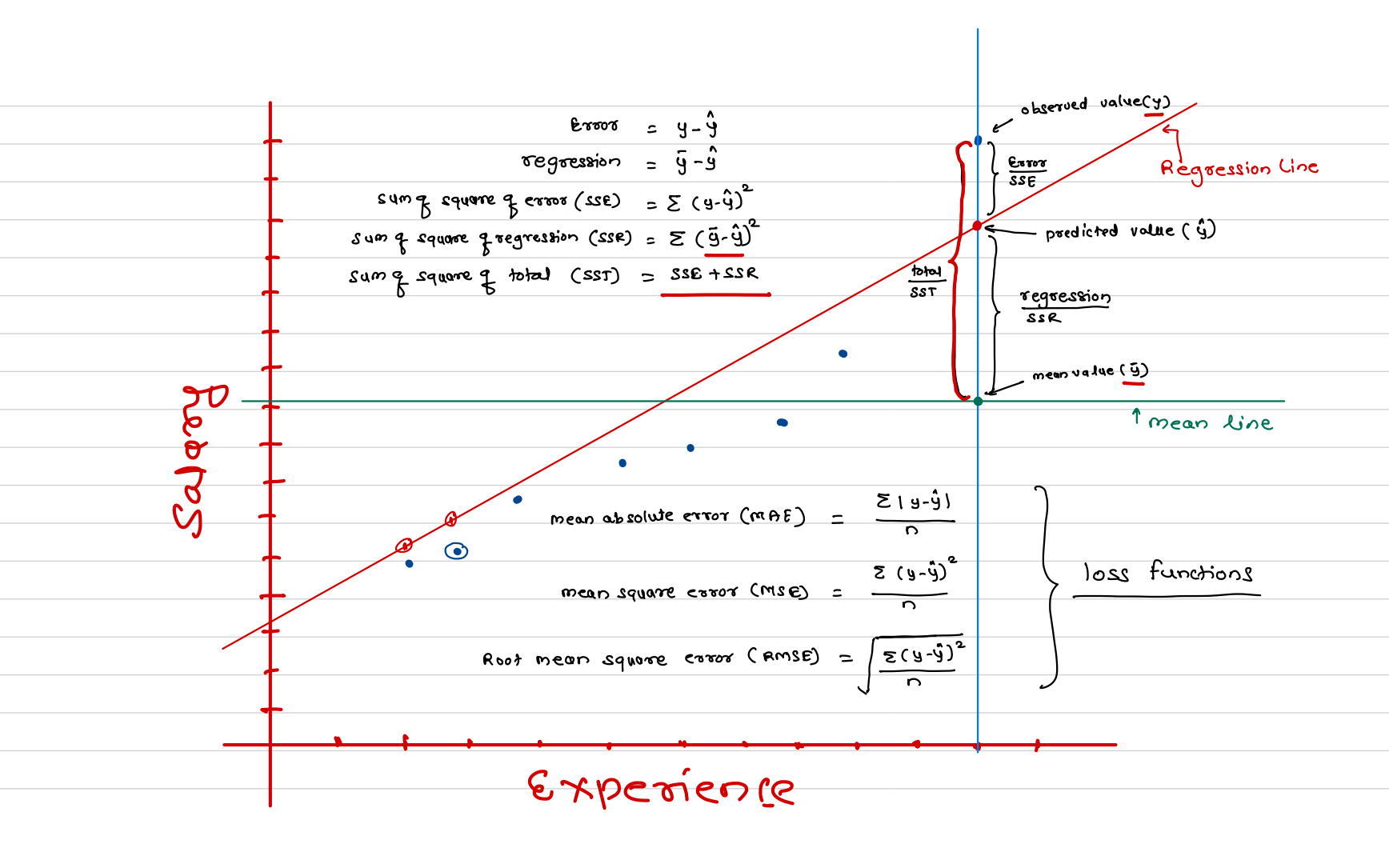
- R-squared value is a statistical measure of how close the data are to the fitted regression line
- It is also known as coefficient of determination or coefficient of multiple determination

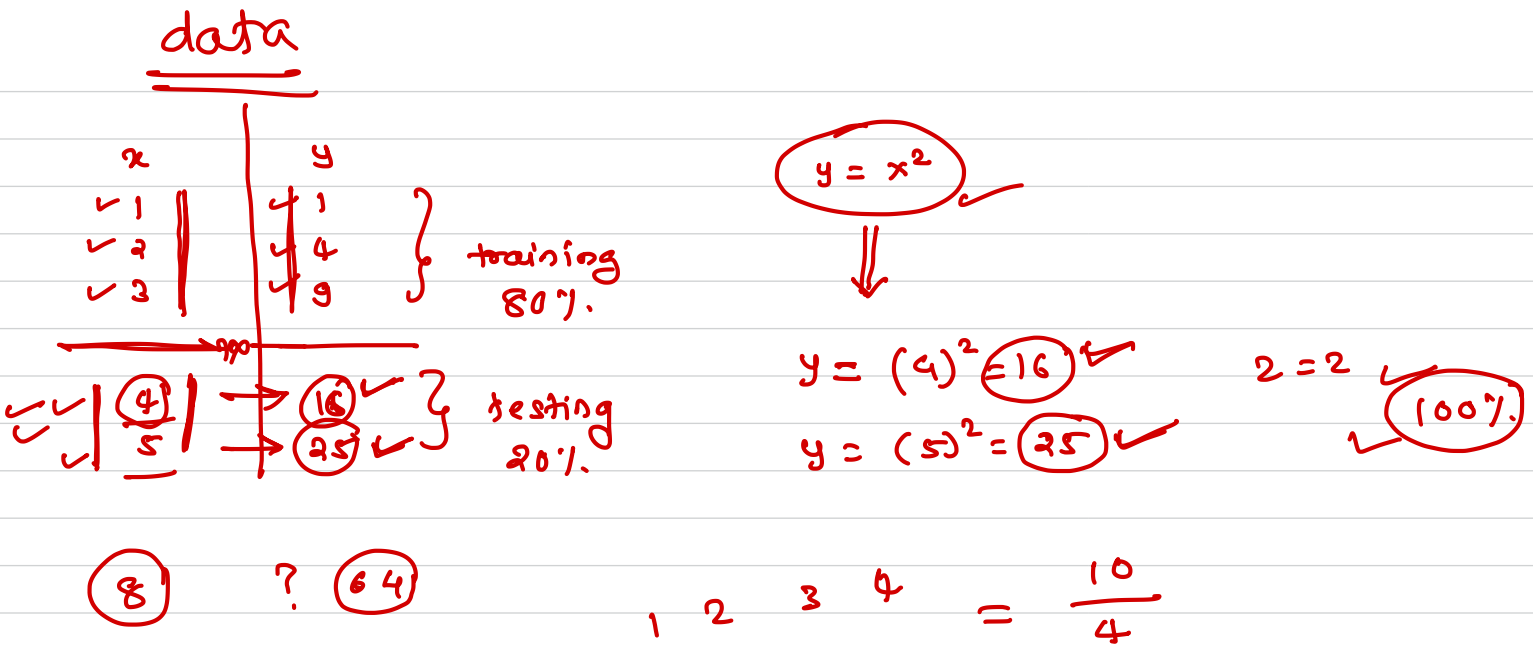
0 ⟷ 1

0.91 = R^2

$$R^2 = \frac{\sum (\hat{Y} - \bar{Y})^2}{\sum (Y - \bar{Y})^2}$$

higher the value of R^2, better is the model


Error = $y - \hat{y}$
regression = $\bar{y} - \hat{y}$
sum of square of error (SSE) = $\Sigma (y-\hat{y})^2$
sum of square of regression (SSR) = $\Sigma (\bar{y}-\hat{y})^2$
sum of square of total (SST) = SSE + SSR
observed value(y)
Regression Line
Error SSE
predicted value ($\hat{y}$)
total SST
regression SSR
mean value ($\bar{y}$)
mean line
Salary
Experience
mean absolute error (MAE) = $\frac{\Sigma |y-\hat{y}|}{n}$
mean square error (MSE) = $\frac{\Sigma (y-\hat{y})^2}{n}$
Root mean square error (RMSE) = $\sqrt{\frac{\Sigma (y-\hat{y})^2}{n}}$
loss functions


data
x
y
1
2
3
1
4
9
training
80%.
4
5
16
25
testing
20%.
8
?
64
$y = x^2$
$y = (4)^2 = 16$
$y = (5)^2 = 25$
2 = 2
100%.
1 2 3 4 = 10/4
= 0

data

Exp	Salary
1.1	15000
1.5	17000
2.5	25000
3	32000

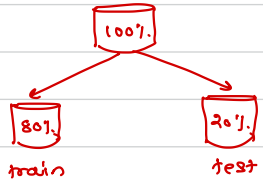
⇒

X

Exp
1.1
1.5
2.5
3

y

Salary
15000
17000
25000
32000



train_test_split(X , y , train_size = 0.8)

X_train (80%)

Exp
1.1
2.5
3

X_test (20%)

Exp
1.5

y_train (80%)

Salary
15000
25000
32000

y_test (20%)

Salary
17000