SNS COLLEGE OF ENGINEERING

Kurumbapalayam(Po), Coimbatore – 641 107
Accredited by NAAC-UGC with 'A' Grade
Approved by AICTE, Recognized by UGC & Affiliated to Anna University, Chennai

Department of Information Technology

19IT601 – Data Science and Analytics

III Year / VI Semester

Unit 3 – PREDICTIVE MODELING AND MACHINE LEARNING

Topic 1: Regression

Regression

Regression:

- Regression is a statistical technique to determine the linear relationship between two or more variables. Regression is primarily used for prediction.

- In its simplest (bivariate) form, regression shows the relationship between one independent variable (X) and a dependent variable (Y).

- Regression is a supervised learning technique which helps in finding the correlation between variables and enables us to predict the continuous output variable based on the one or more predictor variables.

- It is mainly used for prediction, forecasting, time series modeling, and determining the causal-effect relationship between variables.

- Regression shows a line or curve that passes through all the datapoints on target-predictor graph in such a way that the vertical distance between the datapoints and the regression line is minimum.

Regression

Types of Regression

- Linear Regression
- Polynomial Regression
- Multivariate Regression

Some examples of regression can be as:

- Prediction of rain using temperature and other factors
- Determining Market trends
- Prediction of road accidents due to rash driving.

Linear Regression

Linear Regression

- Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x).

- So, this regression technique finds out a linear relationship between x (input) and y(output). Hence, the name is Linear Regression.

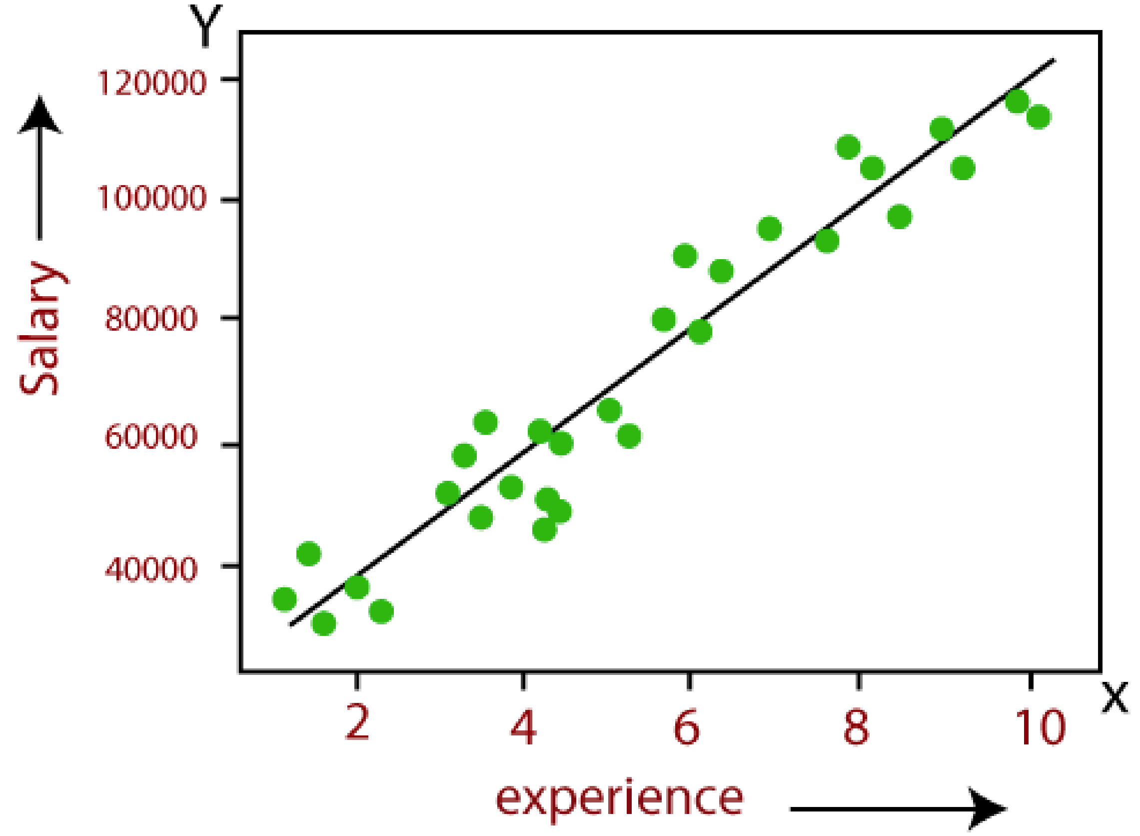
- It can be described as Y=f(x).

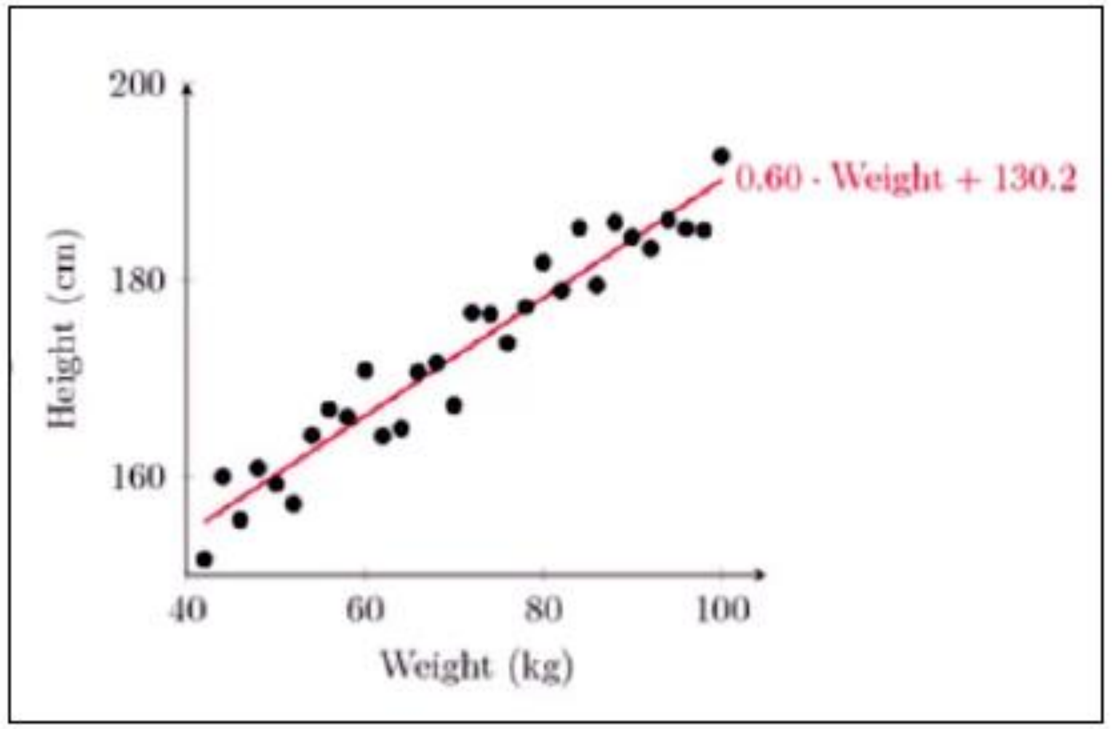
- It is fitting a straight line to a set of data points.

- Generally, a linear model makes a prediction by simply computing a weighted sum of the input features, plus a constant called the bias term (also called the intercept term).

- Bias is the difference between the average prediction of our model and the correct value which we are trying to predict.
- In this technique, the dependent variable is continuous, the independent variable(s) can be continuous or discrete, and the nature of the regression line is linear.

Linear Regression





The equation of a straight line is **y=mx+b**, where

- Y is dependent variable (Target variable)
- m is the slope (linear coefficients)
- b is the y-intercept (Bias)

Linear Regression

Calculating linear regression

The ordinary least squares technique - it tries to minimize the **squared error** between each point and the line, where the error is just the distance between each point and the line that you have.

Then we sum up all the squares of those errors.

The gradient descent technique –

Gradient descent is an optimization technique used to tune the coefficient and bias of a linear equation

Using the gradient descent technique can make sense when dealing with 3D data.

The co-efficient of determination or r-squared

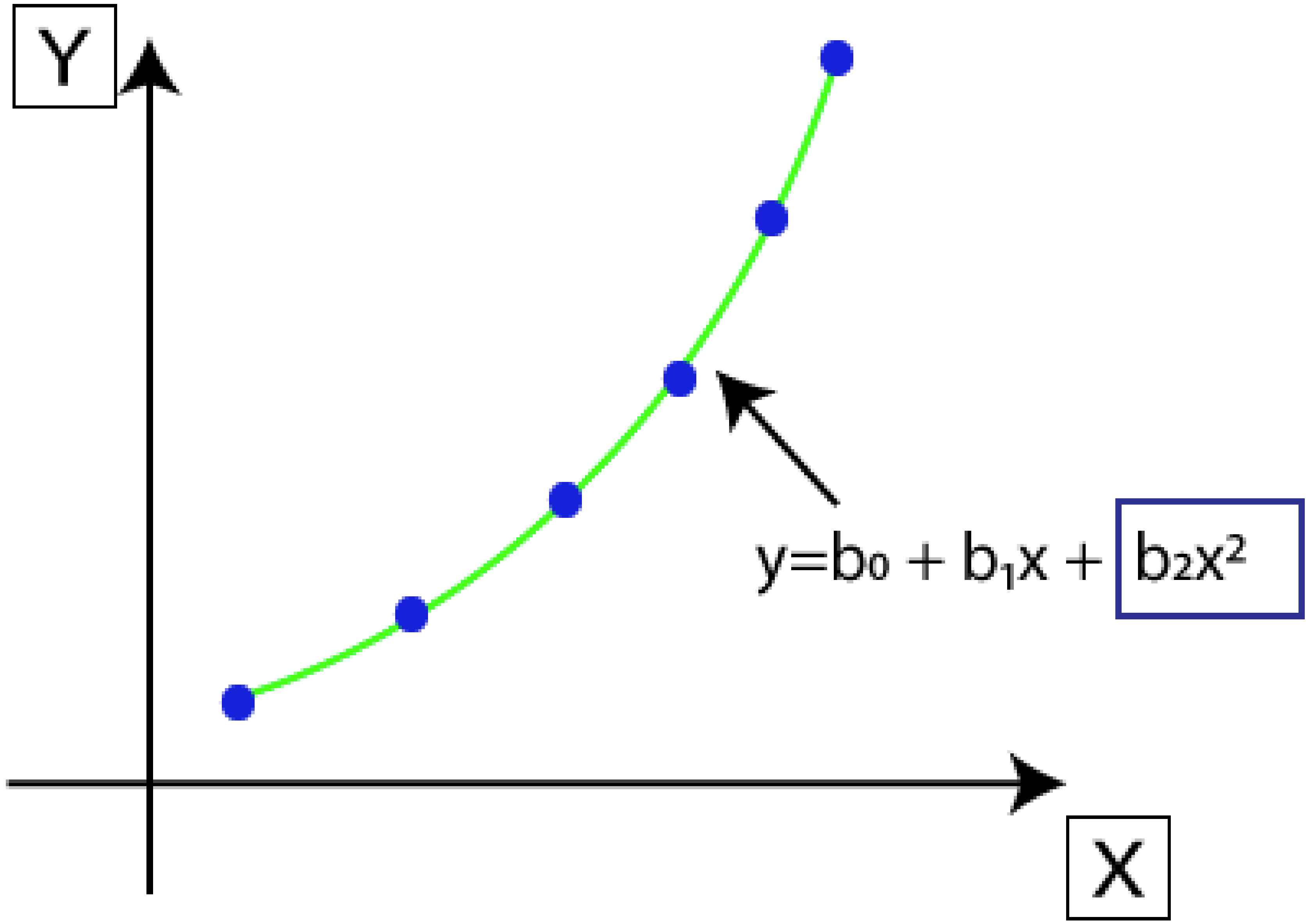
It is the fraction of the total variation in Y that is captured by your models

$$1.0 - \frac{sum\ of\ squared\ errors}{sum\ of\ squared\ variation\ from\ mean}$$

Polynomial Regression

- Polynomial Regression is a type of regression which models the non-linear dataset using a linear model
- It is used when data doesn't actually have a linear relationship, or maybe there's some sort of a curve to it.
- Suppose there is a dataset which consists of datapoints which are present in a non-linear fashion, so for such case, linear regression will not best fit to those datapoints.
- To cover such datapoints, we need Polynomial regression. In Polynomial regression, the original features are transformed into polynomial features of given degree and then modeled using a linear model.
- Which means the datapoints are best fitted using a polynomial line.
- The equation for polynomial regression **Y= b0+b1x+ b2x2+ b3x3+.....+ bnxn.**
- Here Y is the predicted/target output, b0, b1,... bn are the regression coefficients. x is our independent/input variable. The model is still linear as the coefficients are still linear with quadratic.

Polynomial Regression



Multivariate Regression

- Multivariate Regression is a supervised machine learning algorithm involving multiple data variables for analysis.
- Multivariate regression is an extension of multiple regression with one dependent variable and multiple independent variables.
- Based on the number of independent variables, we try to predict the output.

Here, the plane is the function that expresses y as a function of x and z. The linear regression equation can now be expressed as: **$y = m1.x + m2.z + c$**

- y is the dependent variable, that is, the variable that needs to be predicted.
- x is the first independent variable. It is the first input.
- m1 is the slope of x1. It lets us know the angle of the line (x).
- z is the second independent variable. It is the second input.
- m2 is the slope of z. It helps us to know the angle of the line (z).
- c is the intercept.

Below is the generalized equation for the multivariate regression model-

- **$y = \beta 0 + \beta 1.x1 + \beta 2.x2 + ... + \beta n.xn$**
- Where n represents the number of independent variables, $\beta 0 \sim \beta n$ represents the coefficients, and $x1 \sim xn$ is the independent variable.

Multilevel Models

- Multilevel modelling is a statistical model that is used to model the relationship between dependent data and independent data when there is a correlation between observations.
- These models are also known as hierarchical models, mixed effect models, nested data models or random coefficient models.
- Here, the individual observations are nested inside different groups. The observations within each group are correlated
- The concept of multi-level models is that some effects happen at various levels in the hierarchy.
- In Multi-level models there is a hierarchy of effects that influence each other at larger and larger scales.

Types

- Random Intercept Model
- Random Coefficient Model or Random Slopes and Intercepts Model

THANK YOU