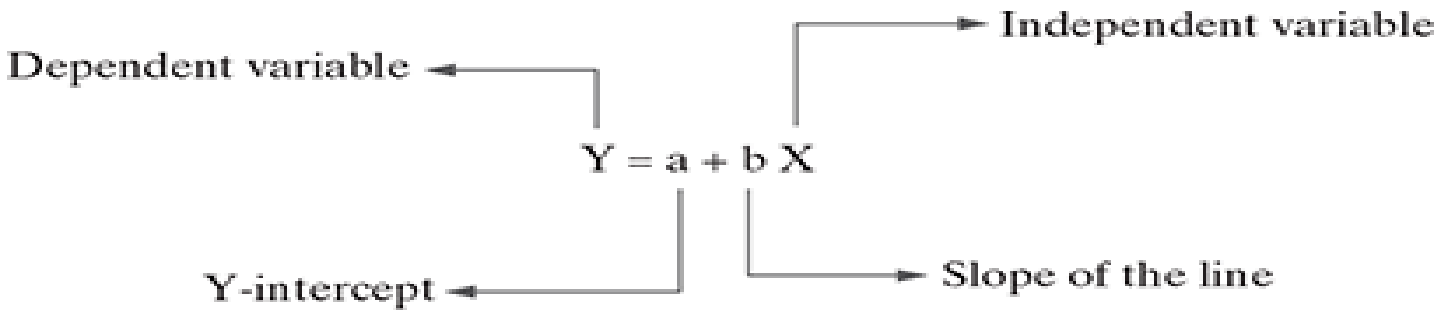
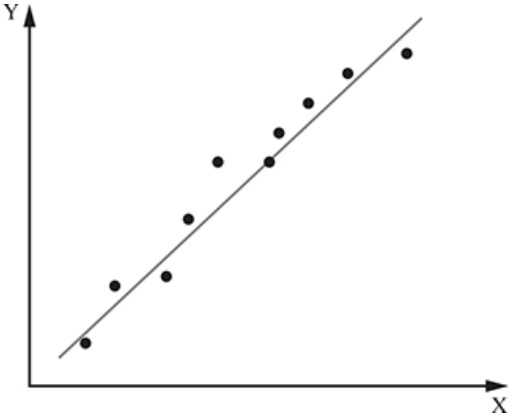
Chapter – 8

A study material for the students of GLS University
Compiled by Dr. Krupa Mehta

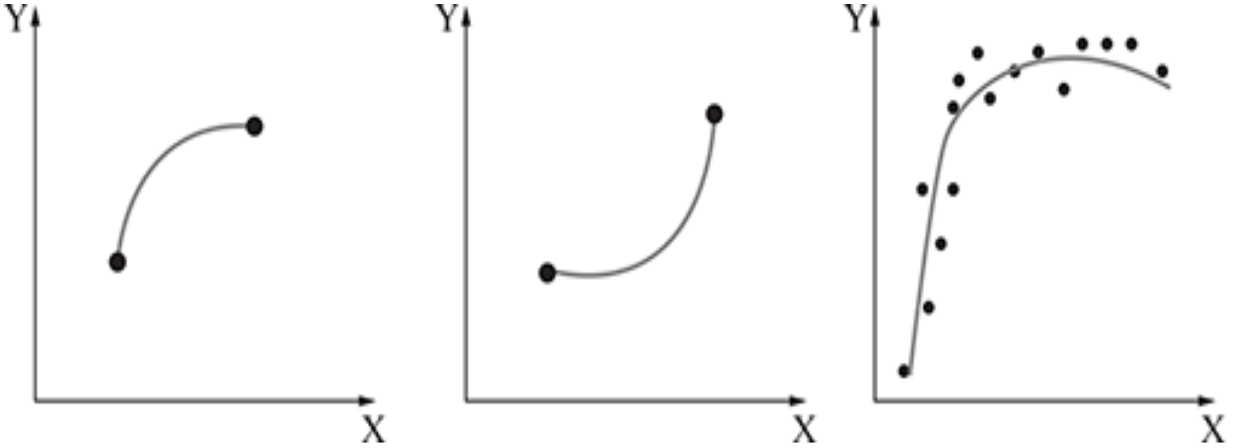
Simple Linear Regression



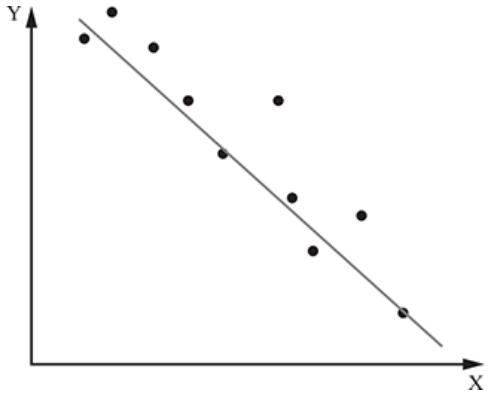
Simple Linear Regression



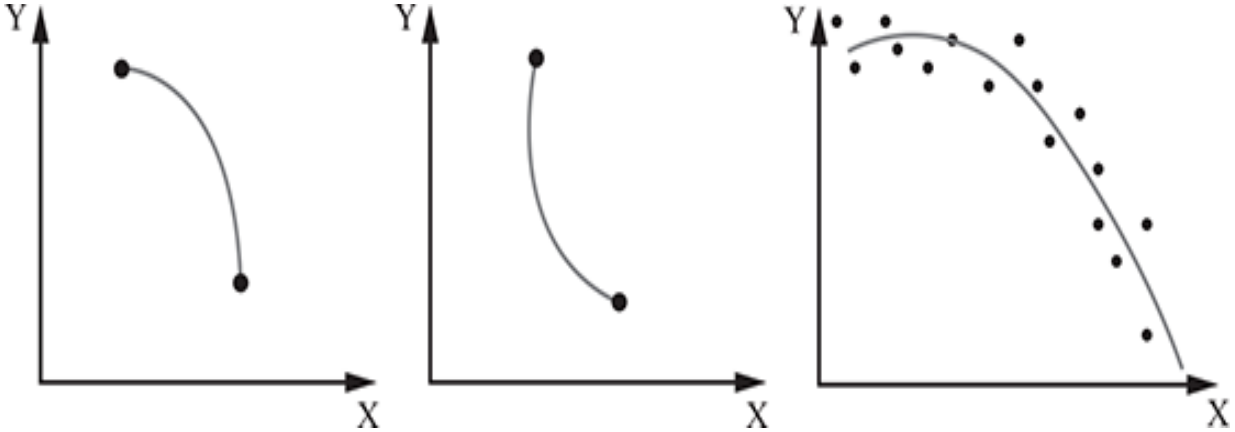
Linear positive slope



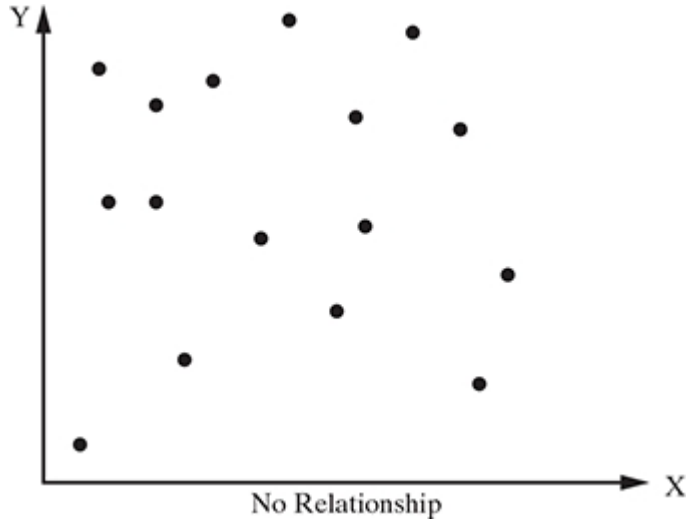
Curve linear positive slope



Linear negative slope



Simple Linear Regression



Error in simple regression

$$Y = (a + bX) + \varepsilon$$

Simple Linear Regression

- Assumptions in Regression Analysis
  - The dependent variable (Y) can be calculated / predicated as a linear function of a specific set of independent variables (X's) plus an error term (ε).
  - The number of observations (n) is greater than the number of parameters (k) to be estimated, i.e. n > k.
  - Relationships determined by regression are only relationships of association based on the data set and not necessarily of cause and effect of the defined class.
  - Regression line can be valid only over a limited range of data. If the line is extended (outside the range of extrapolation), it may only lead to wrong predictions.
  - If the business conditions change and the business assumptions underlying the regression model are no longer valid, then the past dataset will no longer be able to predict future trends.
  - Variance is the same for all values of X (homoskedasticity).
  - The error term (ε) is normally distributed. This also means that the mean of the error (ε) has an expected value of 0.
  - The values of the error (ε) are independent and are not related to any values of X. This means that there are no relationships between a particular X, Y that are related to another specific value of X, Y.

Logistic Regression

- Classification + Regression
- Binary prediction

Logistic Regression

- Assumptions in logistic regression
  - There exists a linear relationship between logit function and independent variables
  - The dependent variable Y must be categorical (1/0) and take binary value, e.g. if pass then Y = 1; else Y = 0
  - The data meets the ‘iid’ criterion, i.e. the error terms, ε, are independent from one another and identically distributed
  - The error term follows a binomial distribution [n, p]
    - n = # of records in the data
    - p = probability of success (pass, responder)

Reference

- Machine Learning by Saikat Dutt, Subramanian Chandramouli, Amit Kumar Das published by Pearson