Covariance and Correlation

Terminology

- **Univariate**
  - This type of data consists of only one variable
  - It does not deal with causes or relationships
  - the main purpose of the analysis is to describe the data and find patterns that exist within it

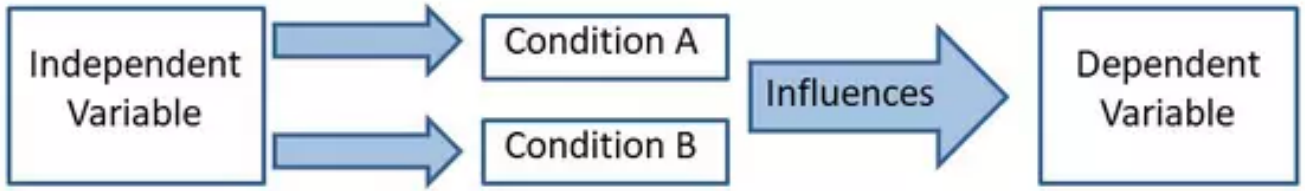
- **Bivariate**
  - This type of data involves two different variables
  - The analysis of this type of data deals with causes and relationships
  - the analysis is done to find out the relationship among the two variables

- **Multivariate**
  - When the data involves three or more variables
  - It is similar to bivariate but contains more than one dependent variable

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 variables**
  - 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



Covariance

- A measure of the relationship between two random variables
- The metric evaluates how much – to what extent – the variables change together
- A positive covariance would indicate a positive linear relationship between the variables
- A negative covariance would indicate the opposite

$$cov(x,y) = \frac{\sum(X_i - \mu)(Y_i - \mu)}{N}$$

$$cov(x,y) = \frac{\sum(X_i - \overline{x})(Y_i - \overline{y})}{n}$$

- Where
  - X_i – the values of the X-variable
  - Y_i – the values of the Y-variable
  - $\overline{x}$ – the mean (average) of the X-variable
  - $\overline{y}$ – the mean (average) of the Y-variable
  - n – the number of the data points in sample
  - N – the number of the data points in Population

Correlation

- Measures the strength of the relationship between variables
- Correlation is the scaled measure of covariance
- It is dimensionless: the correlation coefficient is always a pure value and not measured in any units

$$\rho(X,Y) = \frac{cov(X,Y)}{\sigma_x \sigma_y}$$

- Where
  - ρ(X,Y) – the correlation between the variables X and Y
  - cov(X,Y) – the covariance between the variables X and Y
  - σ_X – the standard deviation of the X-variable
  - σ_Y – the standard deviation of the Y-variable

Correlation coefficient

- Following are the ways to calculate correlation coefficient
  - Karl Pearson's coefficient of correlation
  - Spearman's Rank correlation
  - Scatter diagram
  - Coefficient of concurrent duration

- Correlation (r)
  - -1 <= r <= 1
  - r = 1 (perfect correlation)
  - r = -1 (perfect negative correlation)
  - r > 0 (positive correlation)
  - r < 0 (negative correlation)
  - r = 0 (no correlation)

Karl Pearson's coefficient of correlation

- Also known product moment coefficient of correlation

$$r = \frac{N\sum XY - \sum X \sum Y}{\sqrt{N\sum X^2 - (\sum X)^2}\sqrt{N\sum Y^2 - (\sum Y)^2}}$$

Spearman's Rank correlation

- Find the correlation using the rank

$$r = 1 - \frac{6 \sum d^2}{n(n^2 - 1)}$$

$$r = 1 - \frac{6 \left[\sum d^2 + \frac{\sum (m^3 - m)}{12}\right]}{n(n^2 - 1)}$$

Covariance vs Correlation

- Both primarily assess the relationship between variables
- The closest analogy to the relationship between them is the relationship between the variance and standard deviation
- Covariance measures the total variation of two random variables from their expected values while correlation measures the strength of the relationship between variables
- Using covariance, we can only gauge the direction of the relationship while correlation is the scaled measure of covariance

Regression Analysis

What is regression analysis

- Linear 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

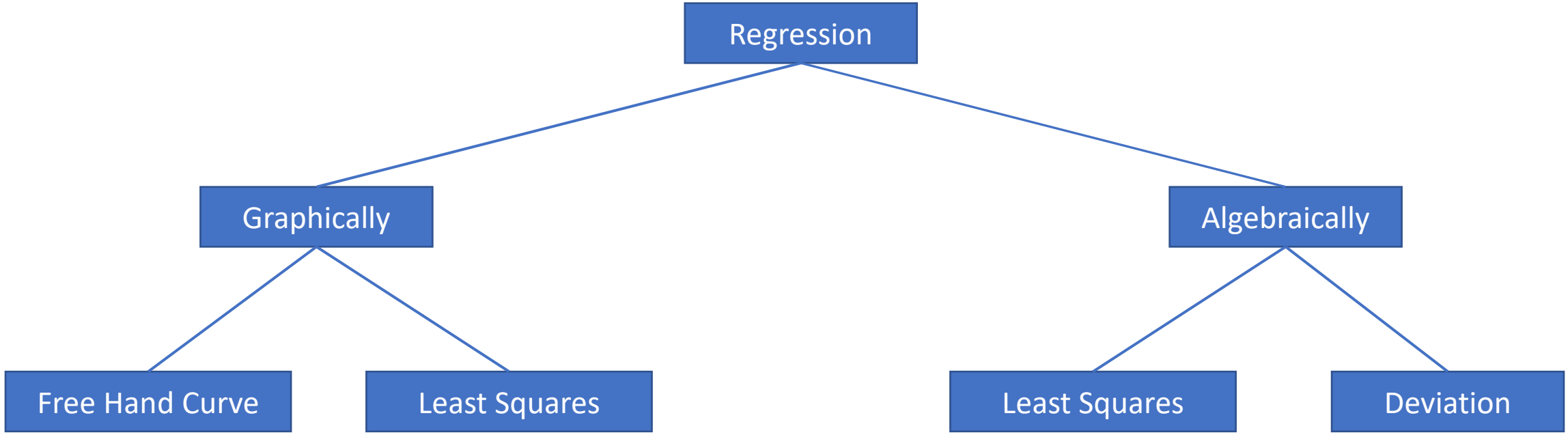
Correlation vs Regression

- Correlation is a statistical measure which determines co-relationship or association of two variables while Regression describes how an independent variable is numerically related to the dependent variable
- Correlation is used to represent linear relationship between two variables while regression is used to fit a best line and estimate one variable on the basis of another variable.
- Cov(x, y) = Cov(y, x)
- but reg(x,y) != reg(y, x)

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.

Methods of studying regression



Types of regression

- 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

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.

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

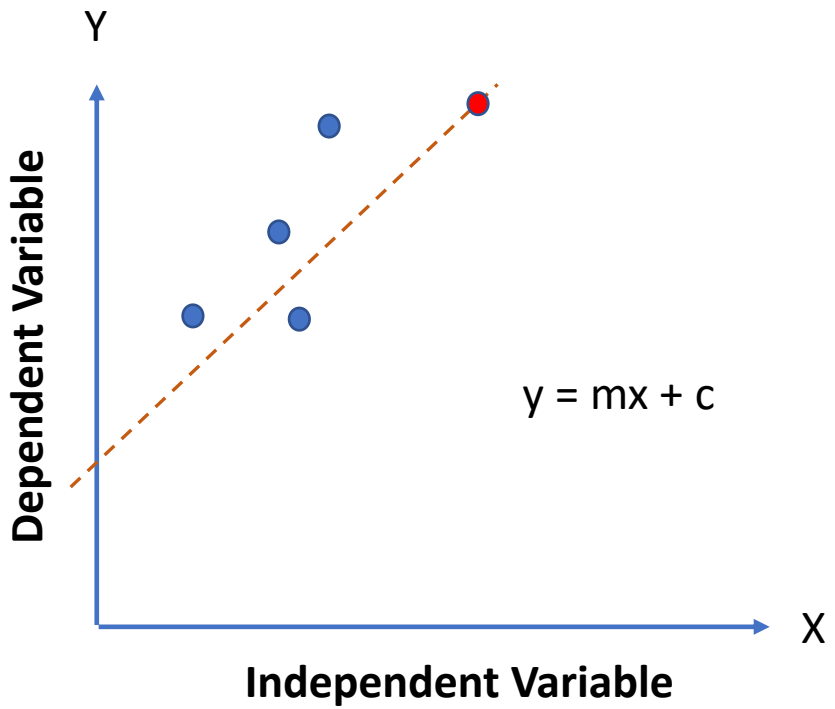
X	Y
3	11
4	12
8	9
7	3
2	5

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

Linear Regression

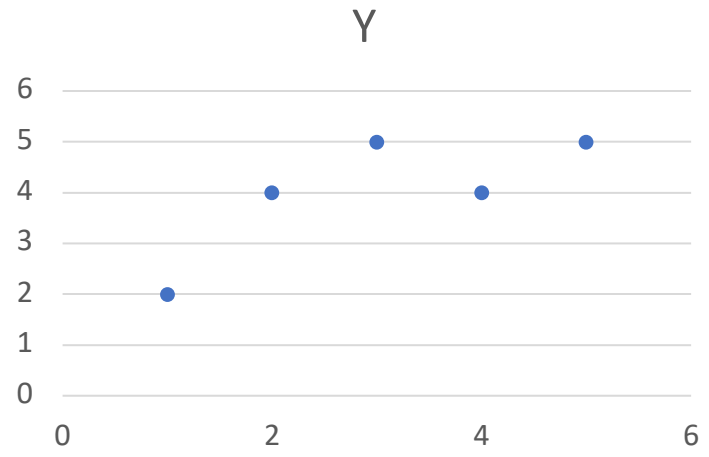
Overview

- 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

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

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