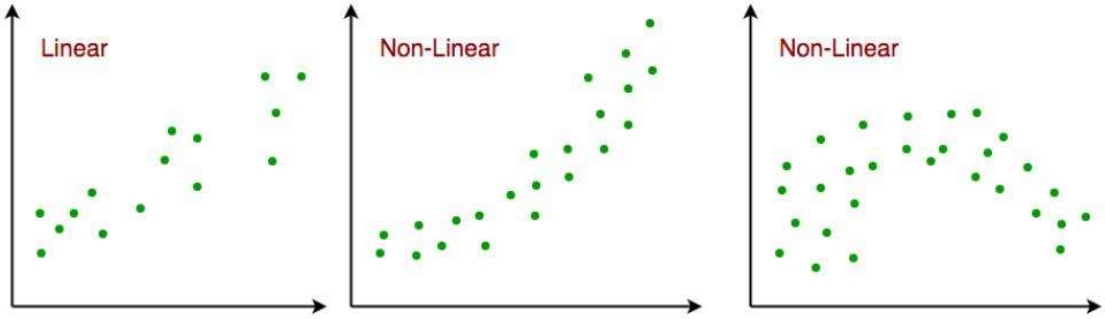
Linear regression comes with a set of assumptions that must be met for the results to be valid and reliable. These assumptions play a crucial role in the interpretation of the regression analysis. The key assumptions of linear regression are as follows:

1. Linearity

The fundamental assumption of linear regression is that there exists a **linear relationship** between the independent variables and the dependent variable. This means that the change in the mean of the dependent variable is proportional to a change in the independent variable. It is essential to check for linearity by examining scatter plots and ensuring that the relationship follows a straight line.



Linear Relationship between Variables

Concept: *The relationship between the dependent variable Y and the independent variable(s) X is represented as a linear combination*.

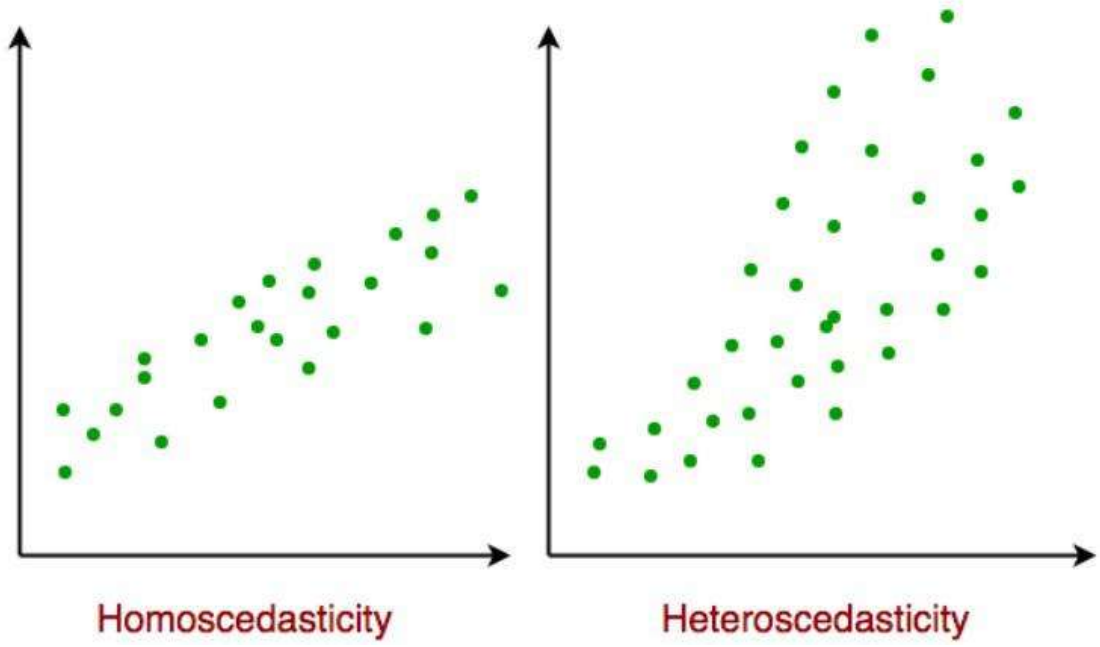
2. Independence

Another critical assumption is the **independence of observations**. Each data point should be independent of others, meaning that the value of the dependent variable for one observation should not be influenced by the values of the independent variables for other observations. This assumption is often violated in time series data or repeated measures studies, requiring special attention and techniques.

Concept: *Each observation is independent of others.*

3. Homoscedasticity

Homoscedasticity refers to the constant variance of errors across all levels of the independent variable. In other words, the spread of residuals should be consistent throughout the range of predictor values. Heteroscedasticity, where the variance of errors is not constant, can lead to inefficient parameter estimates and affect the reliability of hypothesis tests. Residual plots are commonly used to assess homoscedasticity.



Homoscedasticity vs Heteroscedasticity

Concept: *The variance of errors is constant across all levels of the independent variable(s).*

4. Normality of Residuals

The assumption of **normality** pertains to the distribution of the residuals, which should ideally be normally distributed. While normality is not critical for large sample sizes due to the **Central Limit Theorem**, deviations from normality in small samples may impact the validity of statistical inferences. Techniques like the Shapiro-Wilk test or normal probability plots can be employed to assess normality.

Concept: *The residuals follow a normal distribution.*

5. No Perfect Multicollinearity

Multicollinearity occurs when two or more independent variables in the regression model are highly correlated. This can create issues in estimating the individual contributions of each variable to the dependent variable. **Variance Inflation Factor (VIF)** and **correlation matrices** are common tools to detect multicollinearity and corrective measures may involve excluding variables or combining them.

Concept: *The independent variables are not perfectly correlated.*