Linear Regression

Agenda

- Prerequisite :
  - Correlation
  - Covariance
- Introduction to Linear Regression
- Least Squares Regression Line
- Assumptions of Linear Regression
- Simple Linear Regression
- Multiple Linear Regression
- Selecting Best Regression Model

Pearson Correlation Coefficient

Linear Regression Pre Requisite : Correlation

- Are two variables related?
  - Does one increase as the other increases?
    - e. g. sales and promotions
  - Does one decrease as the other increases?
    - e. g. experience and salary
  - How can we get a numerical measure of the degree of relationship?

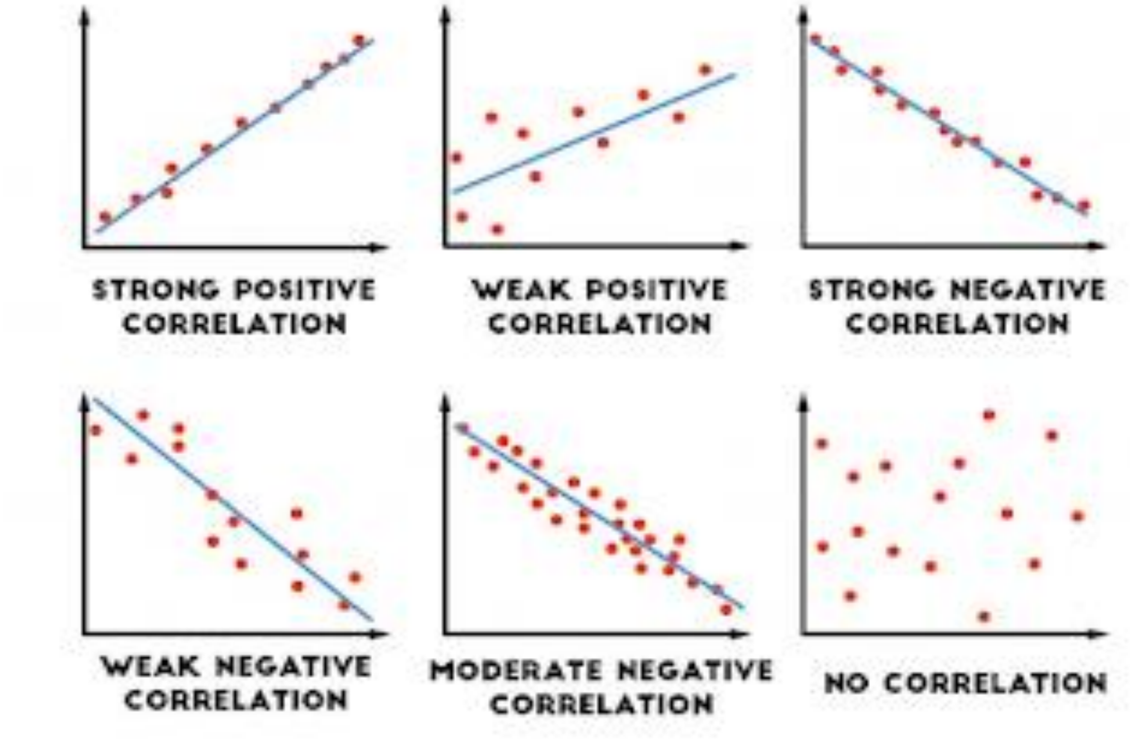
Scatterplots

- Scatter diagram or scatter gram.
- Graphically depicts the relationship between two variables in two dimensional space.

Correlation : **Pearson Product-Moment Correlation**

- It is a measure of relationship between two variables
- Measured with a correlation coefficient
- Most popularly seen correlation coefficient: **Pearson Product-Moment Correlation**

Types of Correlation



STRONG POSITIVE CORRELATION

WEAK POSITIVE CORRELATION

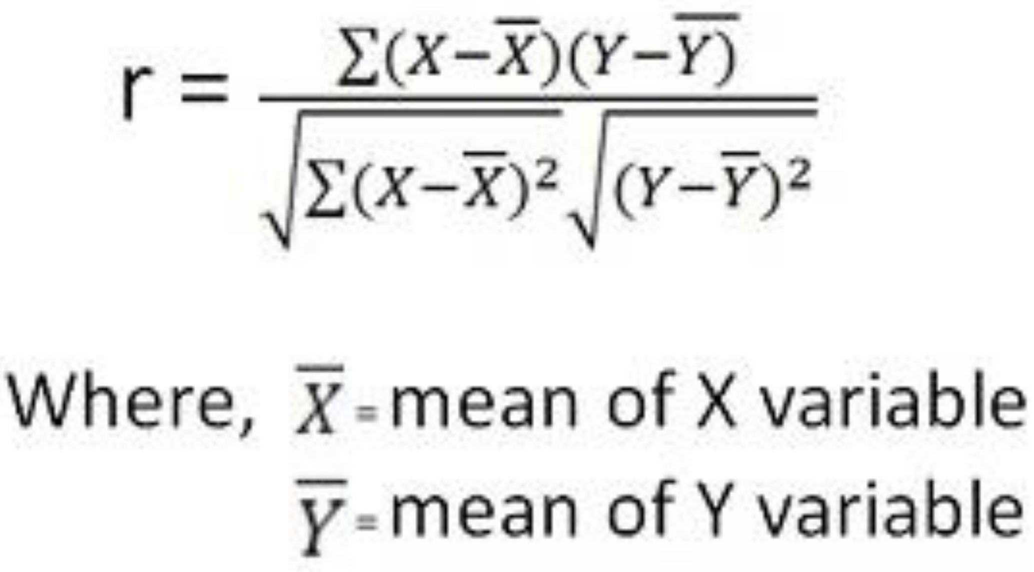
STRONG NEGATIVE CORRELATION

WEAK NEGATIVE CORRELATION

MODERATE NEGATIVE CORRELATION

NO CORRELATION

Pearson Correlation - Formula

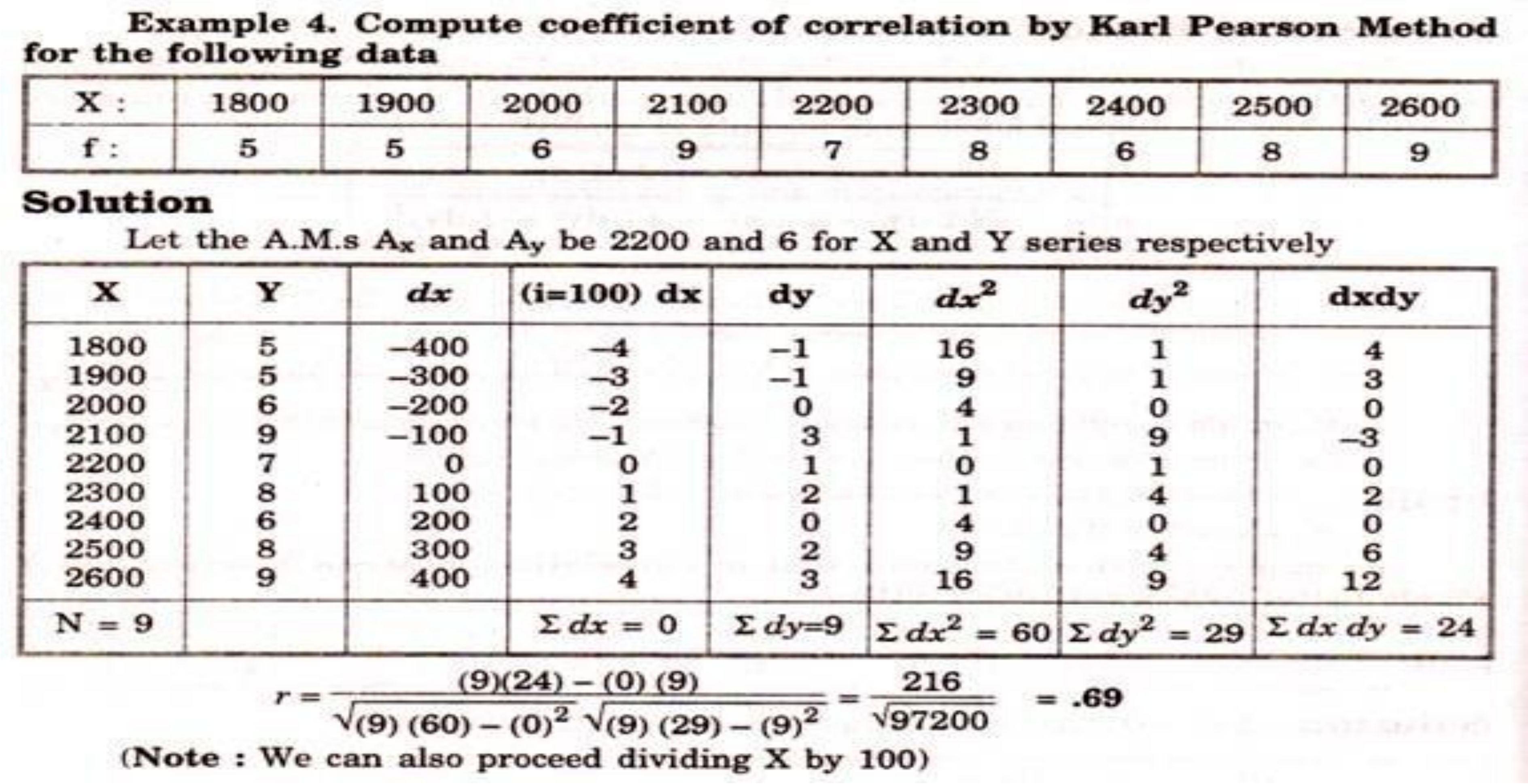


$$r = \frac{\Sigma(X-\overline{X})(Y-\overline{Y})}{\sqrt{\Sigma(X-\overline{X})^2}\sqrt{(Y-\overline{Y})^2}}$$

Where, $\overline{X}$ = mean of X variable

$\overline{Y}$ = mean of Y variable

Pearson Correlation - Calculation



Example 4. Compute coefficient of correlation by Karl Pearson Method for the following data

X :	1800	1900	2000	2100	2200	2300	2400	2500	2600
f :	5	5	6	9	7	8	6	8	9

Solution

Let the A.M.s A_x and A_y be 2200 and 6 for X and Y series respectively

X	Y	*dx*	(i=100) dx	dy	dx^2	dy^2	dxdy
1800	5	–400	–4	–1	16	1	4
1900	5	–300	–3	–1	9	1	3
2000	6	–200	–2	0	4	0	0
2100	9	–100	–1	3	1	9	–3
2200	7	0	0	1	0	1	0
2300	8	100	1	2	1	4	2
2400	6	200	2	0	4	0	0
2500	8	300	3	2	9	4	6
2600	9	400	4	3	16	9	12
N = 9			$\Sigma\, dx = 0$	$\Sigma\, dy{=}9$	$\Sigma\, dx^2 = 60$	$\Sigma\, dy^2 = 29$	$\Sigma\, dx\, dy = 24$

$$r = \frac{(9)(24) - (0)\,(9)}{\sqrt{(9)\,(60) - (0)^2}\,\sqrt{(9)\,(29) - (9)^2}} = \frac{216}{\sqrt{97200}} = .69$$

(Note : We can also proceed dividing X by 100)

Pearson Correlation - Merits

- This method indicates the presence or absence of correlation between two variables and gives the exact degree of their correlation.
- In this method, we can also ascertain the direction of the correlation; positive, or negative.
- This method has many algebraic properties for which the calculation of co-efficient of correlation, and other related factors, are made easy.

Pearson Correlation - Demerits

- It is more difficult to calculate than other methods of calculations.
- It is much affected by the values of the extreme items.
- It is based on a many assumptions, such as: linear relationship, cause and effect relationship etc. which may not always hold good.

Covariance

Definition

- In mathematics and statistics, covariance is a measure of the relationship between two random variables. The metric evaluates how much – to what extent – the variables change together. In other words, it is essentially a measure of the variance between two variables. However, the metric does not assess the dependency between variables.

Covariance : Formula

$$Var_X = \frac{\Sigma(X-\overline{X})^2}{N-1} = \frac{\Sigma(X-\overline{X})(X-\overline{X})}{N-1}$$

- Sample : n-1
- Population : n

$$\text{Cov}\,(X,Y) = \frac{\Sigma(X_i-\overline{X})(Y_j-\overline{Y})}{n}$$

X_i – the values of the X-variable

Y_j – the values of the Y-variable

$\bar{X}$ – the mean (average) of the X-variable

$\bar{Y}$ – the mean (average) of the Y-variable

n – the number of data points

Covariance : Calculations

Introduction to Linear Regression

What is Regression?

- Regression is a statistical method; Regression is a technique that predicts the value of variable **'Y'** based on the values of variable **'X'**.
- In simple terms, Regression helps to find the relation between one **dependent variable** (usually denoted by Y) and one or more **independent variables** (usually denoted by X).

What is Linear Regression?

It is an analysis that can be modeling the relationship between dependent and one (or) more response in independent variable. Linear regression is a type of supervised algorithm

What is Linear Regression?

- (or) In simple terms, linear regression is a method of finding the best straight-line fitting to the given data, i.e., finding the best linear relationship between the independent and dependent variables.

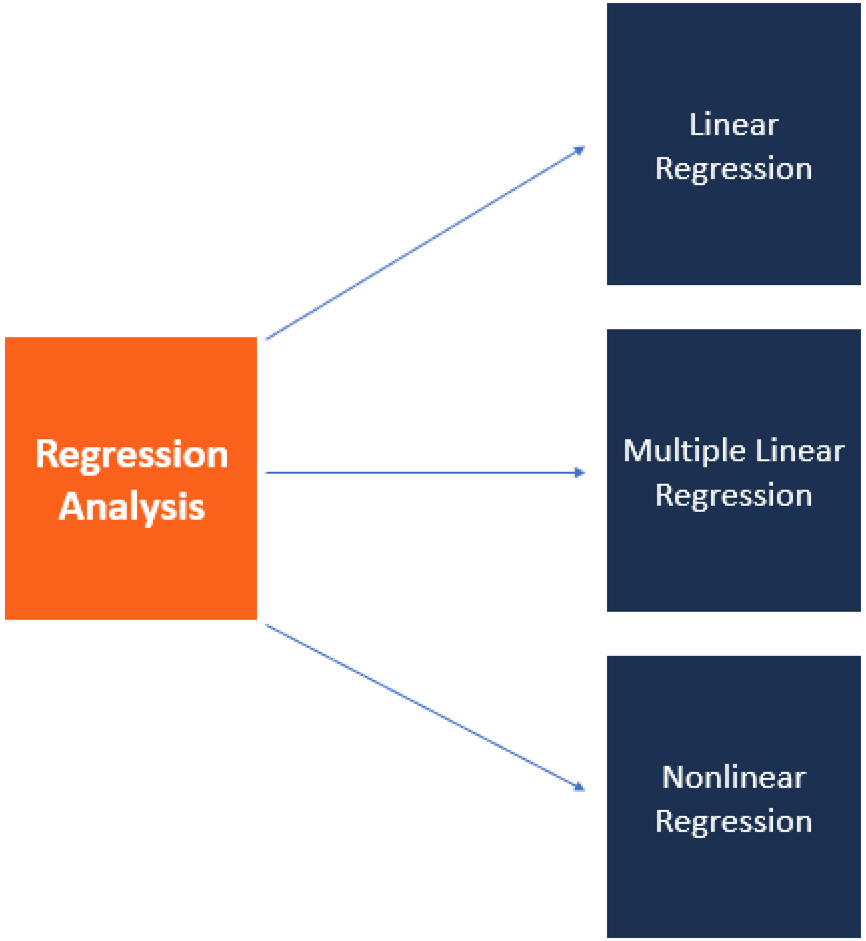
What is Linear Regression?

- Y - the variables you are predicting
  - i.e. dependent variable
- X - the variables you are using to predict
  - i.e. independent variable
- - your predictions (also known as Y′)

What is Linear Regression?

- A technique we use to predict the most likely score on one variable from those on another variable
- Uses the *nature of the relationship* (i.e. correlation) between two variables to *enhance* your prediction

Types of Regression Analysis

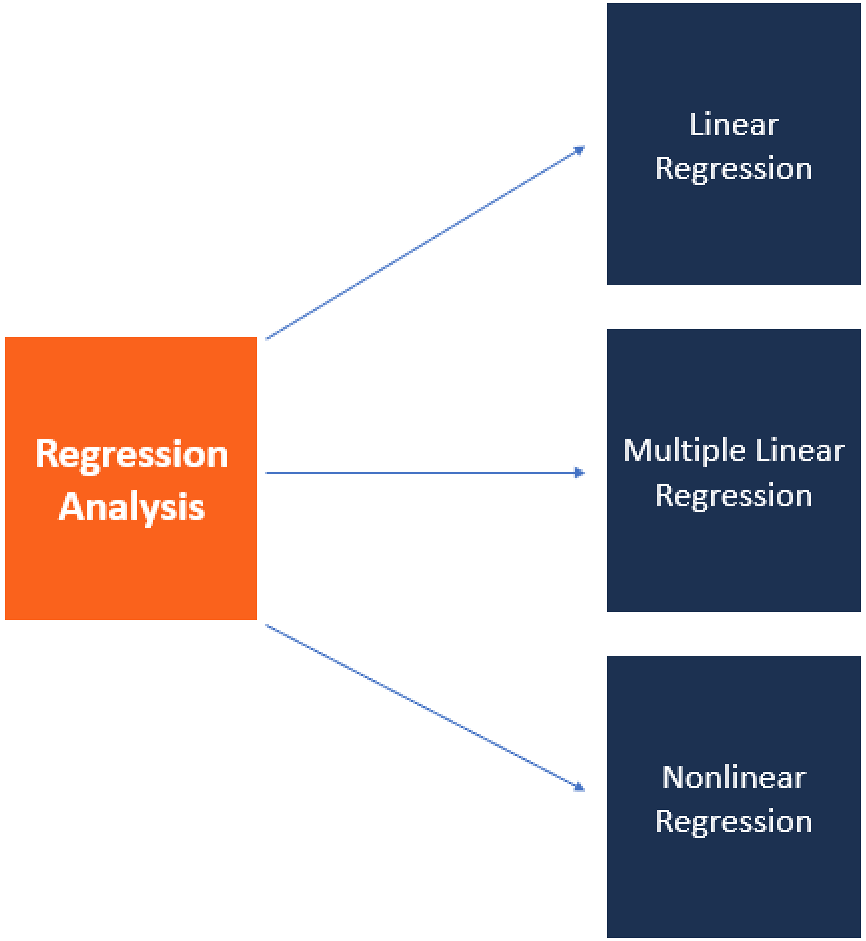


There are two basic types of regression: Simple linear regression and **Multiple linear regression**.

The general form of each type of regression is:

Simple linear regression: $Y = a + bX + u$

Types of Regression Analysis



Regression Analysis

Linear Regression

Multiple Linear Regression

Nonlinear Regression

Multiple linear regression:

$Y = a + b_1X_1 {}^{+} b_2X_2 + b_3X_3 + \ldots + b_tX_t + u$

Y = the variable that you are trying to predict (**dependent variable**).

X = the variable that you are using to predict Y (**independent variable**).

a = the intercept.

b = the slope.

u = the regression residual.

Assumptions of Linear Regression

Assumptions in a linear regression model

There are five assumptions associated with a linear regression model:

- **Linearity**: The relationship between X and the mean of Y is linear.
- **Homoscedasticity & Heteroskedasticity**: The variance of residual is the same for any value of X.
- **Normality**: For any fixed value of X, Y is normally distributed.
- **Multicollinearity:** It is also known simple as multiple regression; Multicollinearity is a state of very high intercorrelations or inter-associations among the independent variables.
- **Independence**: Observations are independent of each other and there are no hidden relationships among observations.

Assumption 1: Linearity

- **linearity** and **additivity** of the relationship between dependent and independent variables.
- The expected value of dependent variable is a **straight-line** function of each independent variable, holding the others fixed.
- The effects of different independent variables on the expected value of the dependent variable are additive.
- We Should have linear relationship between **ex:(y,x1),(y,x2),(y,x3).**
- No-Correlation between the independent variables.

i.e., $X1 \neq X2 \neq X3$

Assumption 1: Linearity

How to diagnose:

- nonlinearity is usually most evident in a plot of **observed versus predicted values** or a plot of **residuals (versus) predicted values**, which are a part of standard regression output.
- The points should be symmetrically distributed around a diagonal line in the former plot or around horizontal line in the latter plot, with a roughly constant variance.

How to fix:

- consider applying a nonlinear transformation to the dependent and/or independent variables if you can think of a **transformation** that seems appropriate.

Assumption 2: Homoscedasticity

- If the variance of the errors is increasing over time, confidence intervals for out-of-sample predictions will tend to be unrealistically narrow.
- **Heteroscedasticity** may also have the effect of giving too much weight to a small subset of the data (namely the subset where the error variance was largest) when estimating coefficients.

Assumption 2: Homoscedasticity

- The behaviour of error increasing (or) decreasing exponentially is called "**Heteroscedasticity**" in case of outliers exist in the data.
- The errors should distribute normally(values are between predicted lines distributed equally) is called "**Homoscedasticity**" .

How to fix:

If the dependent variable is strictly positive and if the residual-versus-predicted plot shows that the size of the errors is proportional to the size of the predictions (i.e., if the errors seem consistent in percentage rather than absolute terms), a **log transformation** applied to the dependent variable may be appropriate.

Assumption 3: Normality

- **Y is normally distributed for any value of X**.
- The data should follow a **normal distribution**. Assumptions for linear regression denotes a mean zero error or residual term. The normality are typically mode normality assumptions for the residuals.

Remediation for Normality:

- Remove outliers
- Take transformations(log, x^2 , $\sqrt{x}$)
- Fit other distributions and use non-parametric test(T-test)

  => for Normality.

Assumption 4: Multicollinearity

- **Multiple Linear Regression(MLR)** is also known simple as **Multiple Regression** and we can't have one multi-collinearity in the data. One to many(**Multicollinearity**)

$$\mathbf{x1} \cong x2+x3$$

$$\mathbf{x2} \cong x1+x3$$

$$\mathbf{x3} \cong x1+x2$$

- Multicollinearity is a state of very high intercorrelations or inter-associations among the independent variables. It is therefore a type of disturbance in the data, and if present in the data the statistical inferences made about the data may not be reliable.

Assumption 4: Multicollinearity

There are certain reasons why multicollinearity occurs:

- It is caused by an inaccurate use of dummy variables.
- It is caused by the inclusion of a variable which is computed from other variables in the data set.
- Multicollinearity can also result from the repetition of the same kind of variable.
- Generally, occurs when the variables are highly correlated to each other.

Assumption 4: Multicollinearity

How to diagnose:

- **Correlation matrix –** when computing the matrix of Pearson's Bivariate Correlation among all independent variables the correlation coefficients need to be smaller than 1.

- **Tolerance –** the tolerance measures the influence of one independent variable on all other independent variables; the tolerance is calculated with an initial linear regression analysis. Tolerance is defined as $T = 1 – R^2$ for these first step regression analysis. With $T < 0.1$ there might be multicollinearity in the data and with $T < 0.01$ there certainly **is.**

Assumption 4: Multicollinearity

- **Variance Inflation Factor (VIF) –** the variance inflation factor of the linear regression is defined as VIF = 1/T. With VIF > 10 there is an indication that multicollinearity may be present; with VIF > 100 there is certainly multicollinearity among the variables.
- **If multicollinearity is found in the data, centering the data (that is deducting the mean of the variable from each score) might help to solve the problem. However, the simplest way to address the problem is to remove independent variables with high VIF values.**

Assumption 5: Independence of errors.

- This assumes that the errors of the response variables are uncorrelated with each other. (Actual statistical independence is a stronger condition than mere lack of correlation and is often not needed, although it can be exploited if it is known to hold.)

Assumption 5: Independence of errors.

- Some methods (e.g., generalized least squares) are capable of handling correlated errors, although they typically require significantly more data unless some sort of regularization is used to bias the model towards assuming uncorrelated errors. Bayesian linear regression is a general way of handling this issue.

Simple Linear Regression – In Depth Study

Multiple Linear Regression – In Depth Study

Total Error:

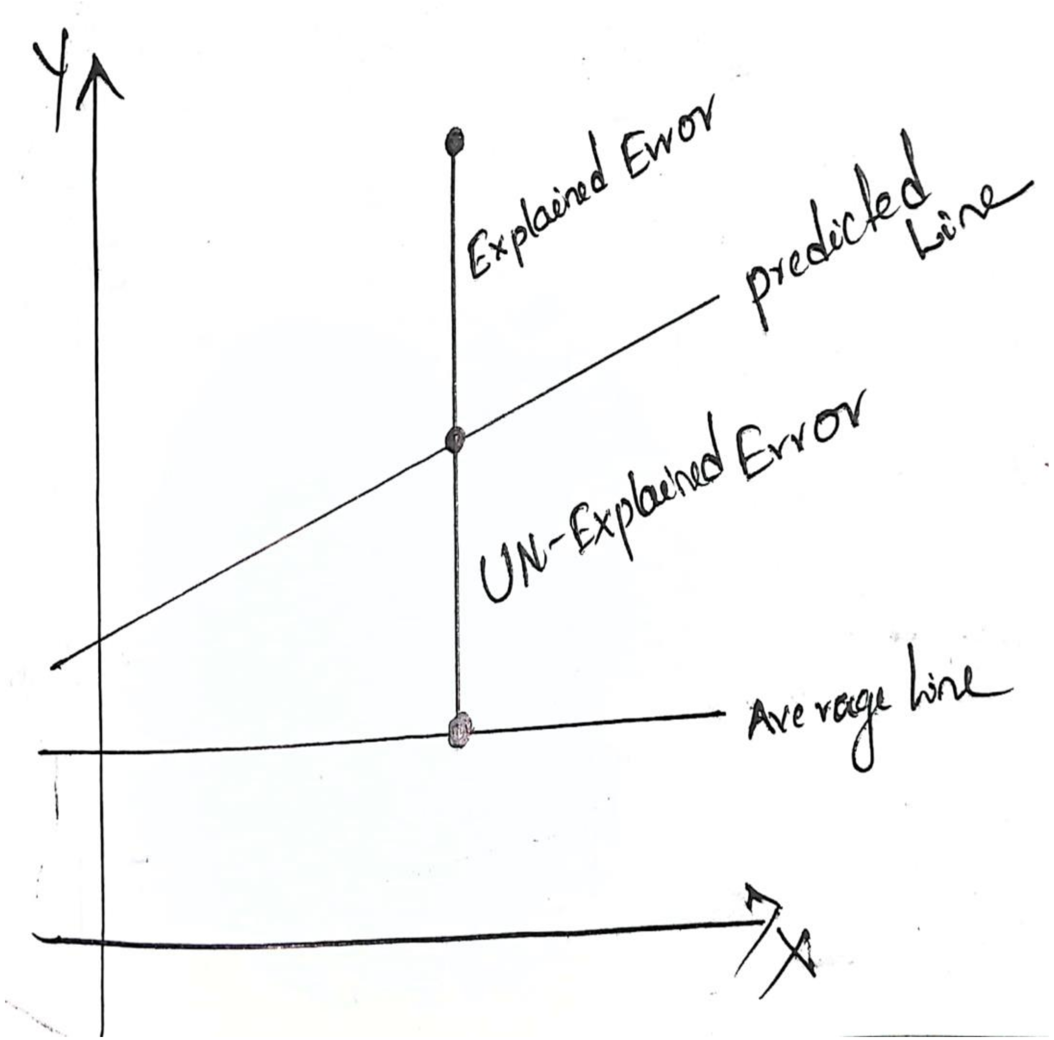
loss function=[actual-predicted]

Mean absolute error

$$MAE = \frac{\sum_{i=1}^{n} |y_i - y_i^p|}{n}$$

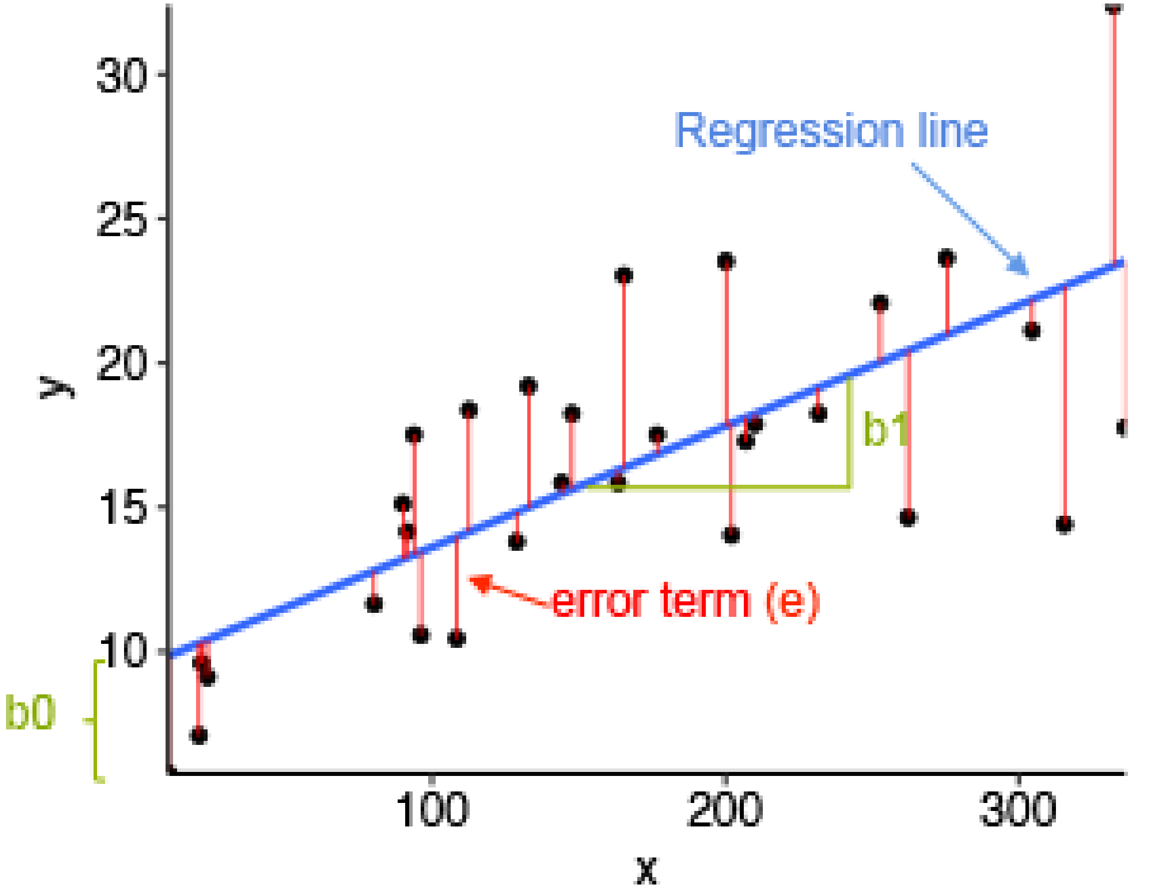
regression line

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2$$



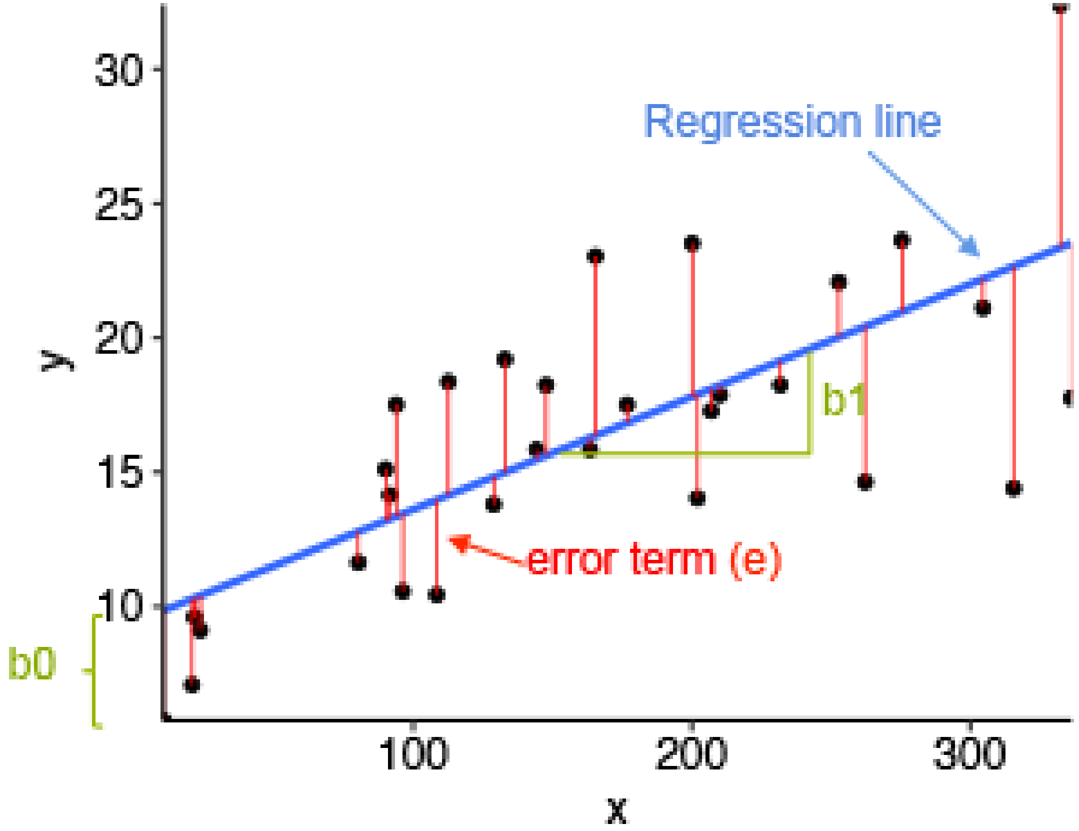
Total Error(Regression line Representation):

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}$$



Total Error(Regression line Representation):

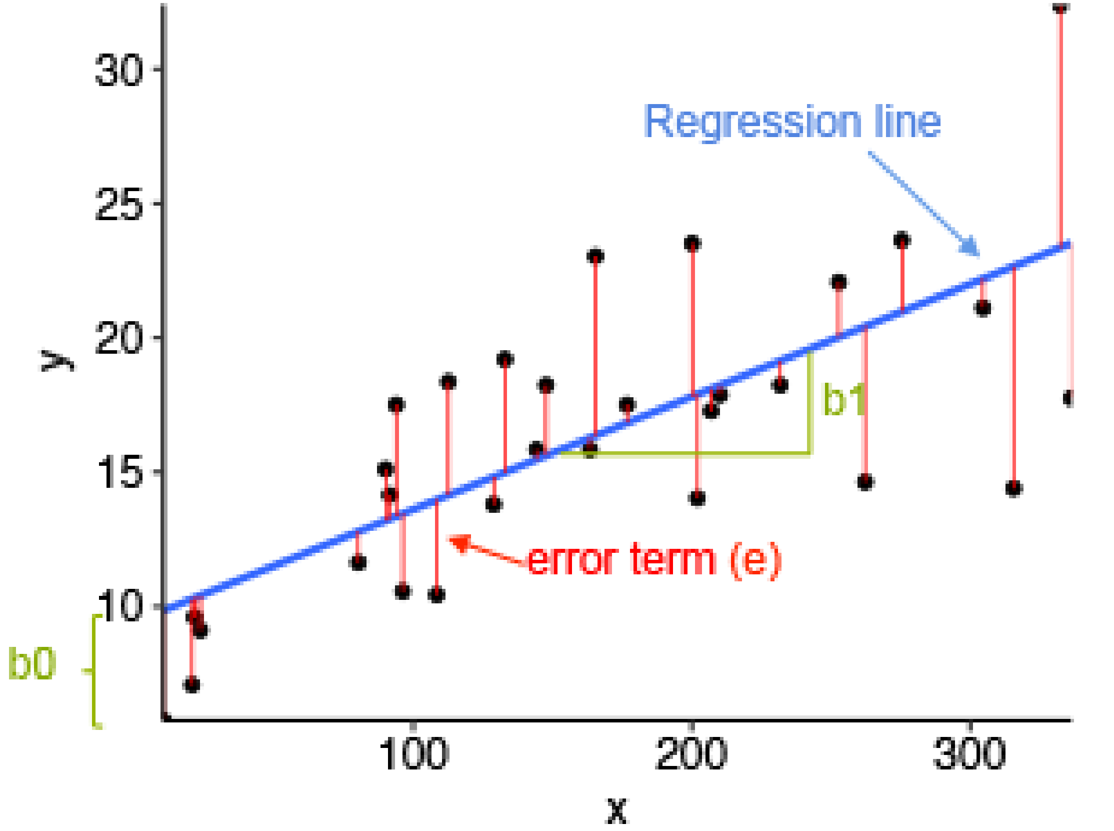
$$\frac{1}{N}\sum_{t=1}^{N}\frac{ABS(Actual_t - Forecast_t)}{Actual_t} * 100\%$$



==
==============================

Total Error(Regression line Representation):

$$MSE = \frac{1}{n}\Sigma\underbrace{\left(y - \widehat{y}\right)^2}_{\text{The square of the difference between actual and predicted}}$$



R2 And Adjusted R2: Concept of Errors

Sum of Squared Error (SSE)

$$SSE = \Sigma (Y_i - \hat{Y}_i)^2$$

Sum of Squared Resuduals (or) Regressor

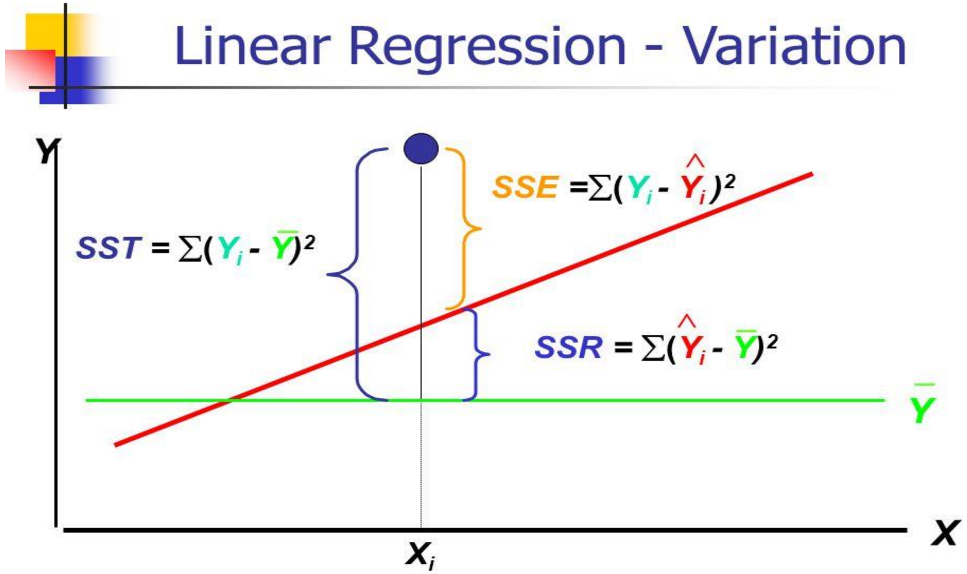
$$SSR = \Sigma (\hat{Y}_i - \bar{Y})^2$$

Sum of Squares of Total (SST)

$$SST = SSR + SSE$$

$$SST = \Sigma (Y_i - \hat{Y})^2 + \Sigma (\hat{Y}_i - \bar{Y})^2$$

$$\boxed{SST = \Sigma (Y_i - \bar{Y})^2}$$



Linear Regression - Variation

Y
$SSE = \Sigma(Y_i - \hat{Y}_i)^2$
$SST = \Sigma(Y_i - \bar{Y})^2$
$SSR = \Sigma(\hat{Y}_i - \bar{Y})^2$
$\bar{Y}$
X_i
X

Sum Squared Regression Error

$$R^2 = 1 - \frac{SS_{Regression}}{SS_{Total}}$$

Sum Squared Total Error

R-Squared (or) Coefficient of Determination

[R-Squared(R2) =(SSR/SST)=(1-(SSE/SST)]

R^2 = 1-(Unexplained variation/Total Variation)

===

==============================

R-Squared (or) Coefficient of Determination

- **R2 is to finding the line of best fit.** The R-squared values range from 0 to 1 and are commonly started as percentages from 0% to 100%

same (or) increase

==

=================================

Adjusted R2: Concept of Errors

R2 shows

Adjusted R2

$$R^2_{adj} = 1 - \left[\frac{(1 - R^2)(n - 1)}{n - k - 1} \right]$$

where:

- **N** is the number of points in your data sample.
- **K** is the number of independent regressors, i.e., the number of variables in your model, **excluding the constant.**

F-Statistics:

$F - ratio \rightarrow$

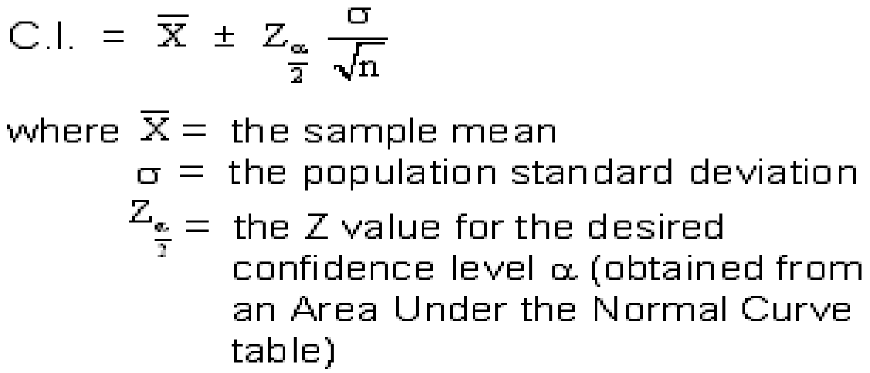
$$F = \frac{MSR}{MSE} = \frac{\frac{SSR}{df_{MSR}}}{\frac{SSE}{df_{MSE}}}$$

$degree\ of\ freedom \rightarrow$

$$df_{MSR} = p$$

$$df_{MSE} = n - p - 1$$

Parameters Estimates and the Associated Statistical Tests:



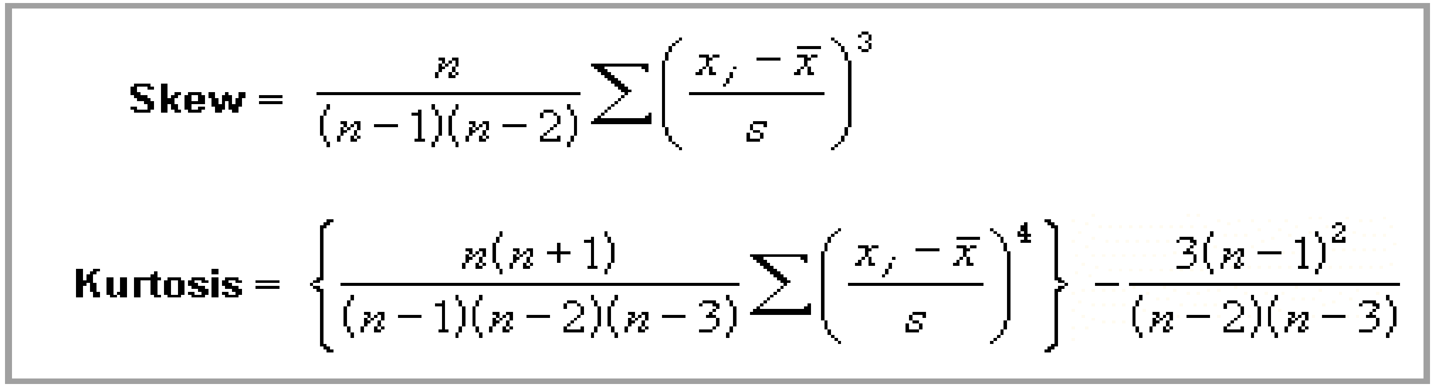
$$\text{C.I.} = \overline{X} \pm Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

where $\overline{X}$ = the sample mean

σ = the population standard deviation

$Z_{\frac{\alpha}{2}}$ = the Z value for the desired confidence level α (obtained from an Area Under the Normal Curve table)

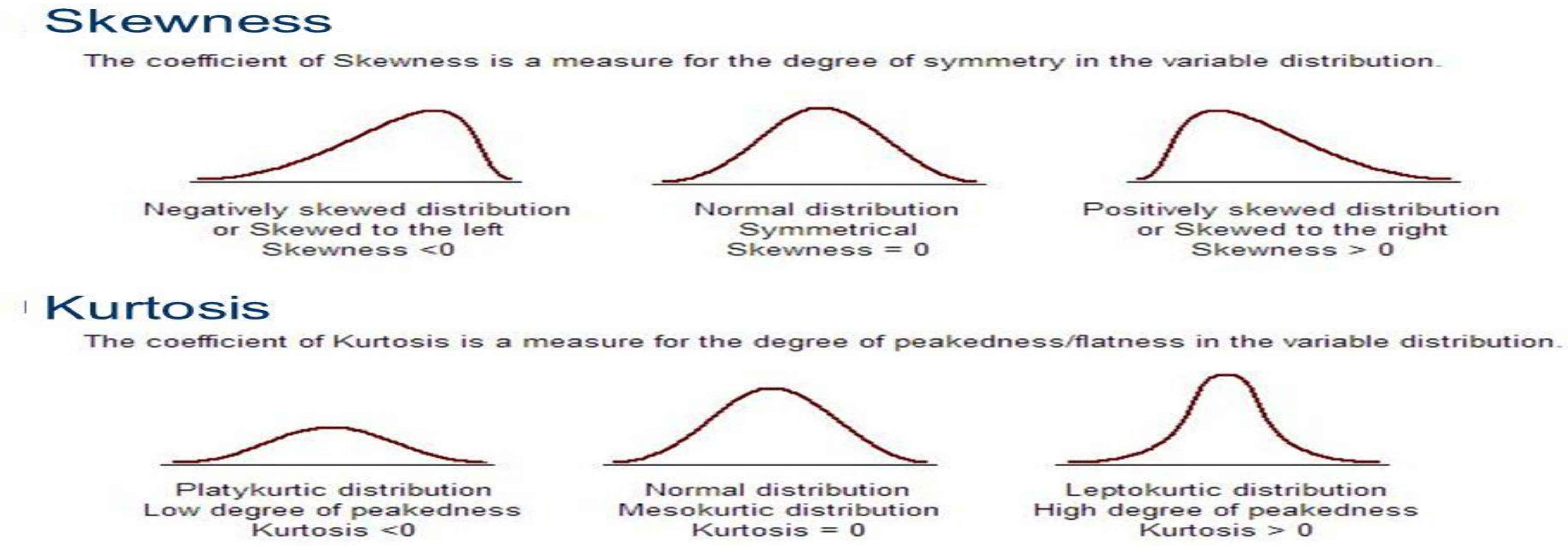
Residual Tests:



$$\text{Skew} = \frac{n}{(n-1)(n-2)} \sum \left(\frac{x_i - \bar{x}}{s} \right)^3$$

$$\text{Kurtosis} = \left\{ \frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum \left(\frac{x_i - \bar{x}}{s} \right)^4 \right\} - \frac{3(n-1)^2}{(n-2)(n-3)}$$

Residual Tests:



Skewness

The coefficient of Skewness is a measure for the degree of symmetry in the variable distribution.

Negatively skewed distribution
or Skewed to the left
Skewness <0

Normal distribution
Symmetrical
Skewness = 0

Positively skewed distribution
or Skewed to the right
Skewness > 0

Kurtosis

The coefficient of Kurtosis is a measure for the degree of peakedness/flatness in the variable distribution.

Platykurtic distribution
Low degree of peakedness
Kurtosis <0

Normal distribution
Mesokurtic distribution
Kurtosis = 0

Leptokurtic distribution
High degree of peakedness
Kurtosis > 0

Difference Between Correlation and Regression?

Basics of comparison	Correlation	Regression
Meaning	correlation is a statistical measure which determines co-relationship or association of two variables.	Regression describes how an independent variable is numerical related to a dependent variable.
Usage	To represent a linear relationship between two variables.	
Dependent and Independent Variables	No difference	Both variables are different.
Indicates	correlation coefficients indicates the extent to which two variables move together.	Regression indicates the impact of a unit change in the known variable(x) on the estimated variable(y).
Objective	to find the numerical value expressing the relationship between variables	To estimate values of random variables based on the values of fixed Variable.

Ridge, Lasso & Elastic net Regressions

What are Ridge, Lasso, Elastic net regressions?

- **Ridge, Lasso, Elastic net Regressions** are Regularization techniques that are used for solving overfitting problem.

- These techniques **penalize the magnitude of coefficients** of features along with **minimizing the error** between predicted and actual observations.

Regularization:

Overfitting: ML model performing well on training data but poorly on validation (test) data.

Regularization helps to solve overfitting problem.

Regularization solves this by adding a penalty term to the objective function and controls the model complexity using that penalty term.

Regularization:

Regularization is generally useful in the following situations:

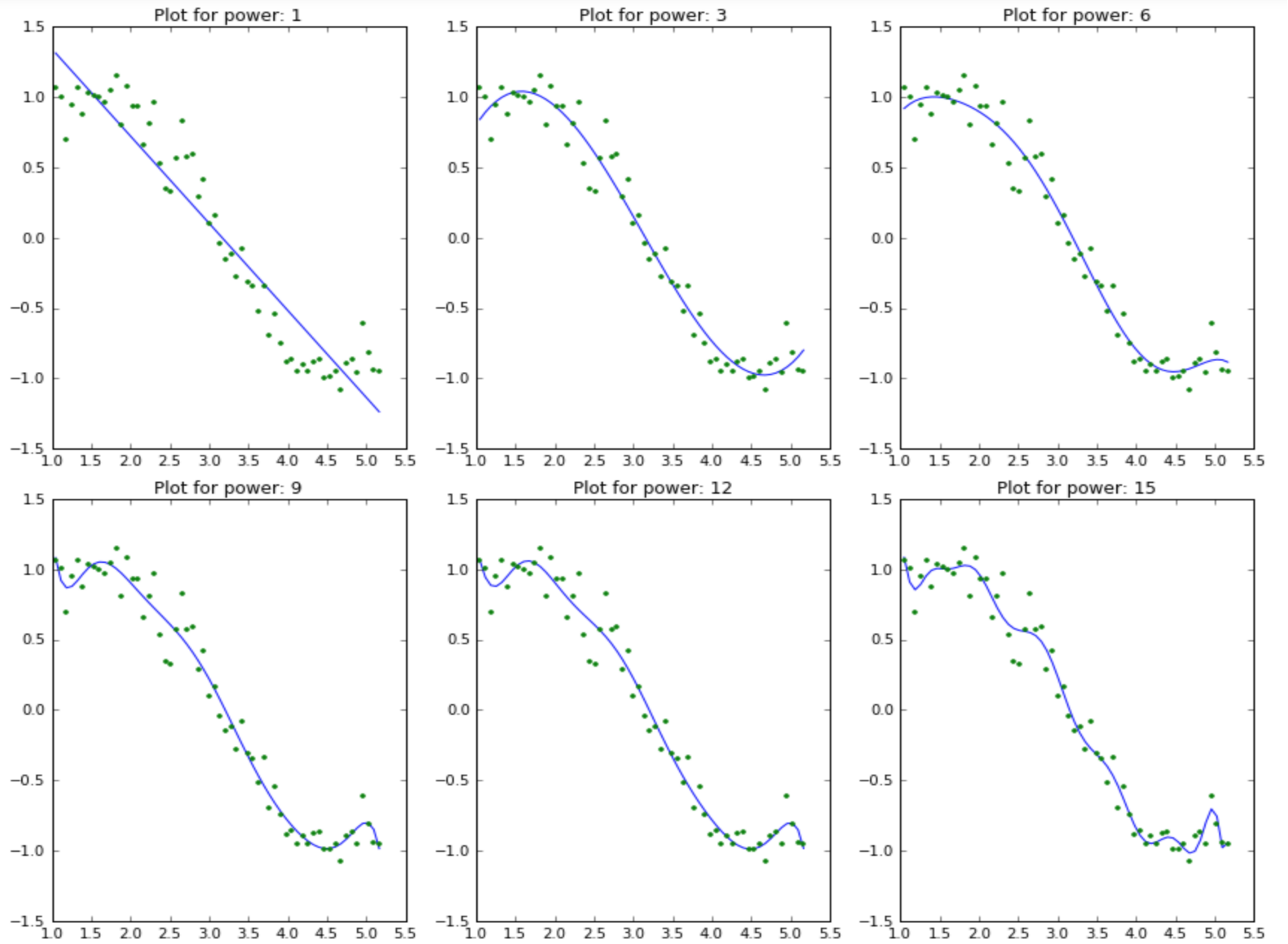
1.Large number of variables/features

2.Low ratio of number of observations to number of variables/features

3.High Multi-Collinearity

Model Complexity Vs Overfitting

- With increase in complexity, models tend to fit even smaller deviations in the training data set.
- This leads to overfitting.
- The **number of coefficients** increase exponentially with increase in model complexity



Ridge Regression:

Ridge regression imposes an additional shrinkage penalty to the ordinary least squares loss function to limit its squared *L2* norm **(L2 Regularization):**

Objective = RSS + α * (sum of square of coefficients)

Ridge Regression:

α can take various values:

α = 0:

The objective becomes same as simple linear regression.

We'll get the same coefficients as simple linear regression.

α = ∞:

The coefficients will be zero, since infinite weightage on square of coefficients, anything other than zero will make the objective infinite.

0 < α < ∞:

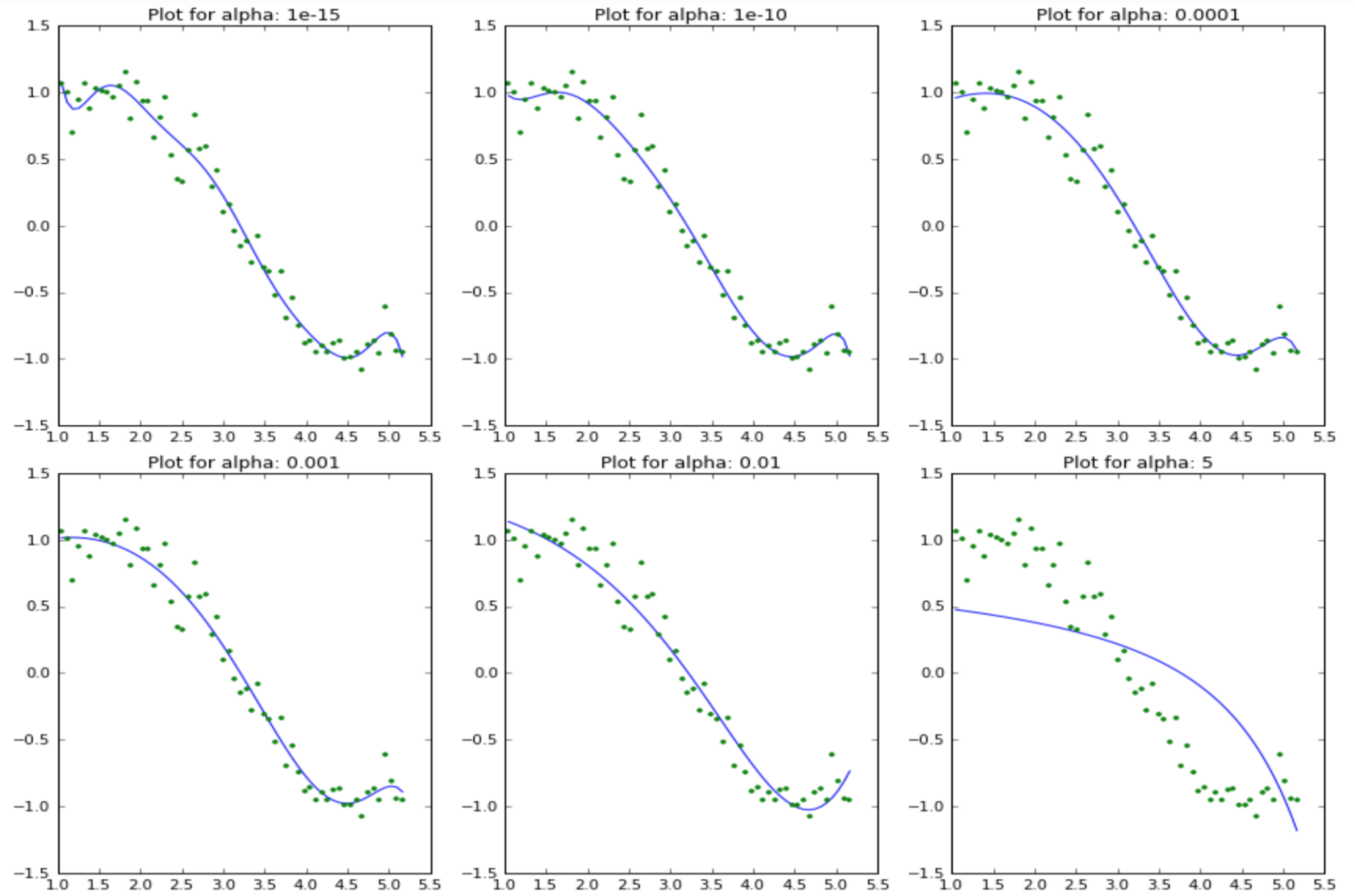
The magnitude of α will decide the weightage given to different parts of objective.
The coefficients will be somewhere between 0 and that of simple linear regression.

Alpha Vs Model Complexity:

ML Labs Pvt Ltd

The model complexity reduces as the value of alpha increases.

High alpha values can lead to significant underfitting.



Ridge Regression:

L2 Regularization

—This adds regularization terms in the model which are function of square of coefficients of parameters. Coefficient of parameters can approach to zero but never become zero.

—Good for multi-collinearity.

—Not Good for feature selection.

Lasso Regression:

Lasso Regression: Performs L1 regularization, i.e., adds penalty equivalent to absolute value of the magnitude of coefficients.

Objective = RSS + α * (sum of absolute value of coefficients)

Lasso Regression:

Here, α (alpha) is same as that of ridge and provides a trade-off between balancing RSS and magnitude of coefficients.

α = 0: Same coefficients as simple linear regression

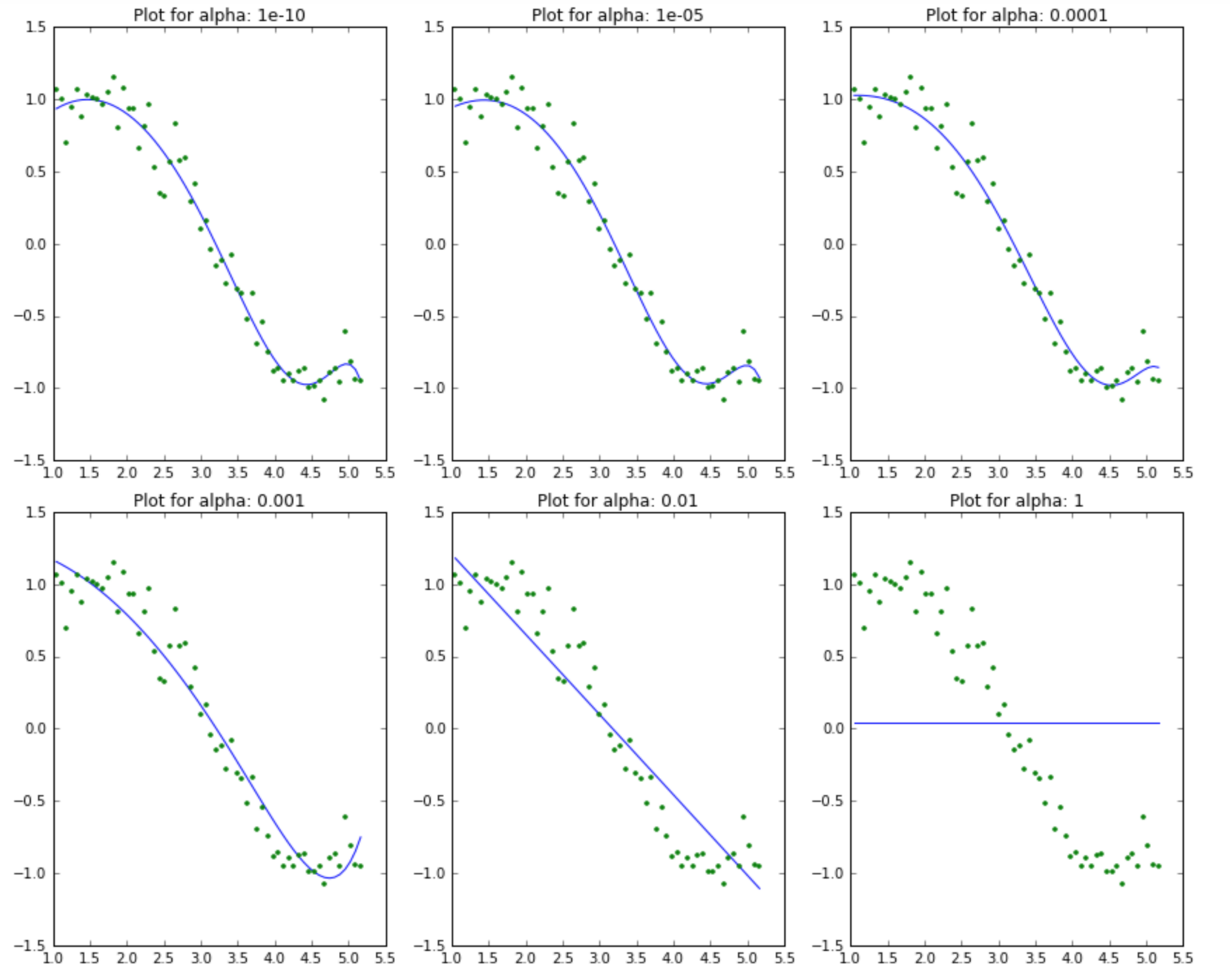
α = ∞: All coefficients zero

0 < α < ∞: coefficients between 0 and that of simple linear regression

Lasso Regression:

ML Labs Pvt Ltd

The model **complexity reduces** as the value of **alpha increases**



Lasso Regression:

L1 Regularization

— Adds regularization terms in the model that are function of **absolute value** of the coefficients of parameters.

— The coefficients of the parameters can be driven to zero as well during the regularization process. Hence this technique can be used for **feature selection.**

— Not good for grouped selection for highly correlated features.

Lasso Regression

At higher alphas, Residual Sum of Squares (RSS) will be higher.

For the same values of alpha, the coefficients of lasso regression are much smaller as compared to that of ridge regression .

For the same alpha, lasso has higher RSS (poorer fit) as compared to ridge regression.

Many of the lasso regression coefficients are zero even for very small values of alpha.

ElasticNet

- ElasticNet is a Regularization technique that **combines both Lasso and Ridge** into a single model with two penalty factors: one proportional to *L1* norm and the other to *L2* norm.

Objective = RSS + α * (sum of absolute value of coefficients)

+λ(sum of square of coefficients)

- ElasticNet model will be sparse like a pure Lasso, but with the same regularization ability as provided by Ridge.

ElasticNet

—This adds regularization terms in the model which are combination of both L1 and L2 regularization.

—Trades bias for variance reduction

—better prediction accuracy