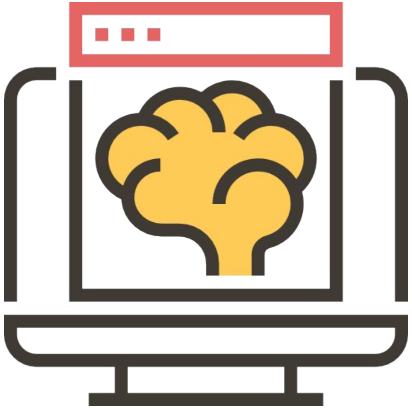


Supervised Learning Regression

Agenda

- Business Problem
- Introduction To Linear regression
- Working Of Linear Regression
- Model Evaluation Of Regression
- Assumptions Of Regression

Business Problem

Business Problem : Predict Salary to be given to new recruits.

It is important for companies to develop models that accurately forecast salary quotient for new employees.

These model estimates can be used to predict salary that can assist HR to set the various components of salary, depending on the number of years of experiences, domain knowledge, location and etc.

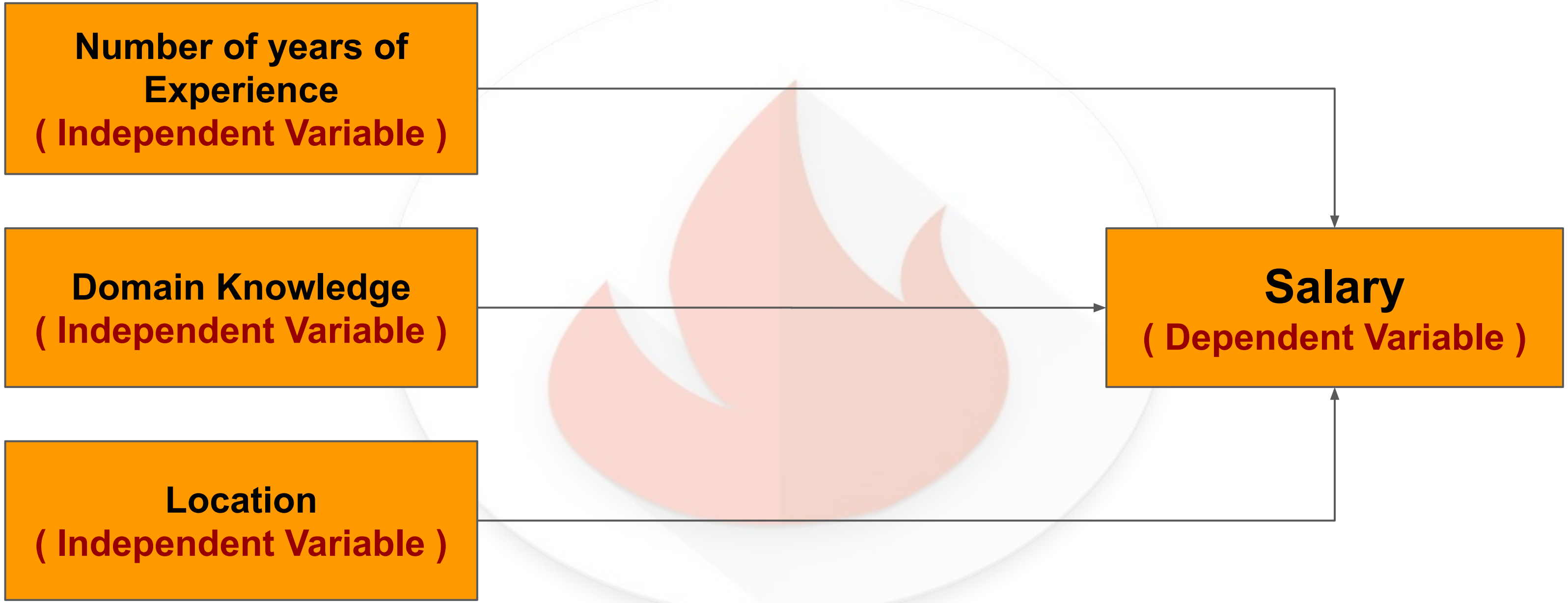
Dependent Variable

- The variable we want to predict or explain.
- It is as well referred to as Target variable, Response variable or Dependent Variable.
- It is denoted by Y.
- In previous example Salary was our dependent variable.

Independent Variable

- The variables used to explain the dependent variable.
- It is as well referred to as Predictor variable, or Independent variable.
- It is denoted by X.
- In previous example Number of Years of Experience, Domain Knowledge, and Location was our independent variable.

Variables that may contribute to predict salary

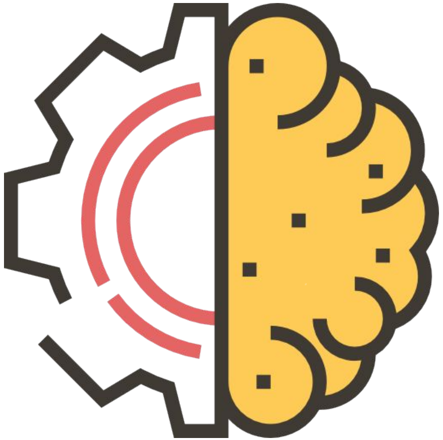


Number of years of Experience
(Independent Variable)

Domain Knowledge
(Independent Variable)

Location
(Independent Variable)

Salary
(Dependent Variable)

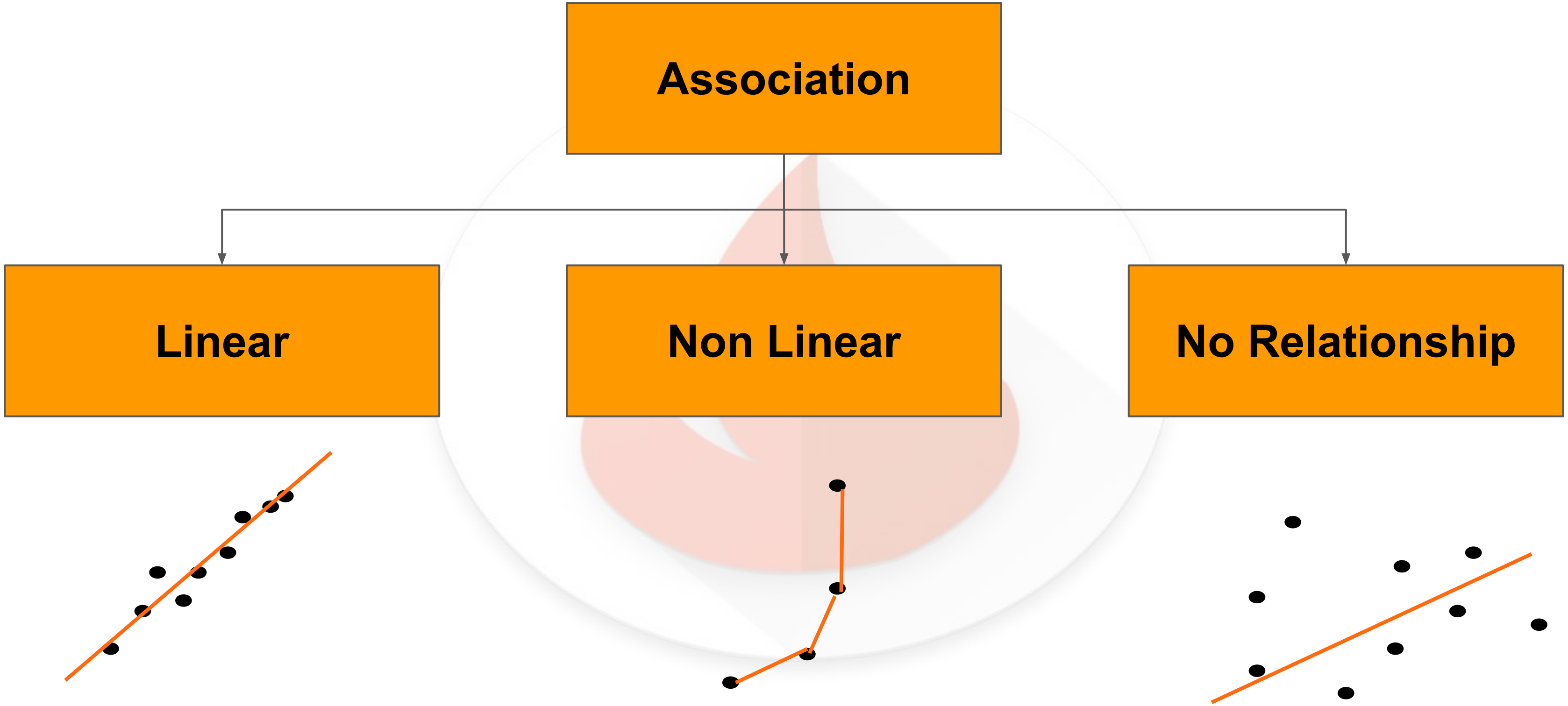


Linear Regression

What is Regression ?

- Regression analysis is a set of statistical processes for estimating the relationships between a dependent variable and one or more independent variables.
- Regression is used to examine the relationship between one dependent and independent variable.
- It is supervised technique.

Types of Association



Association

Linear

Non Linear

No Relationship

Linear Regression

The core idea is to obtain a line that best fits the data. The **best fit line** is the one for which total prediction error (all data points) are as small as possible. Error is the distance between the point to the regression line.

Simple Linear Regression

Simple linear regression is an approach for predicting a quantitative response using a single feature (or "predictor" or "input variable"). It takes the following form called as **Linear Regression Equation:**

$$Y = \beta 0 + \beta 1 * X + \varepsilon$$

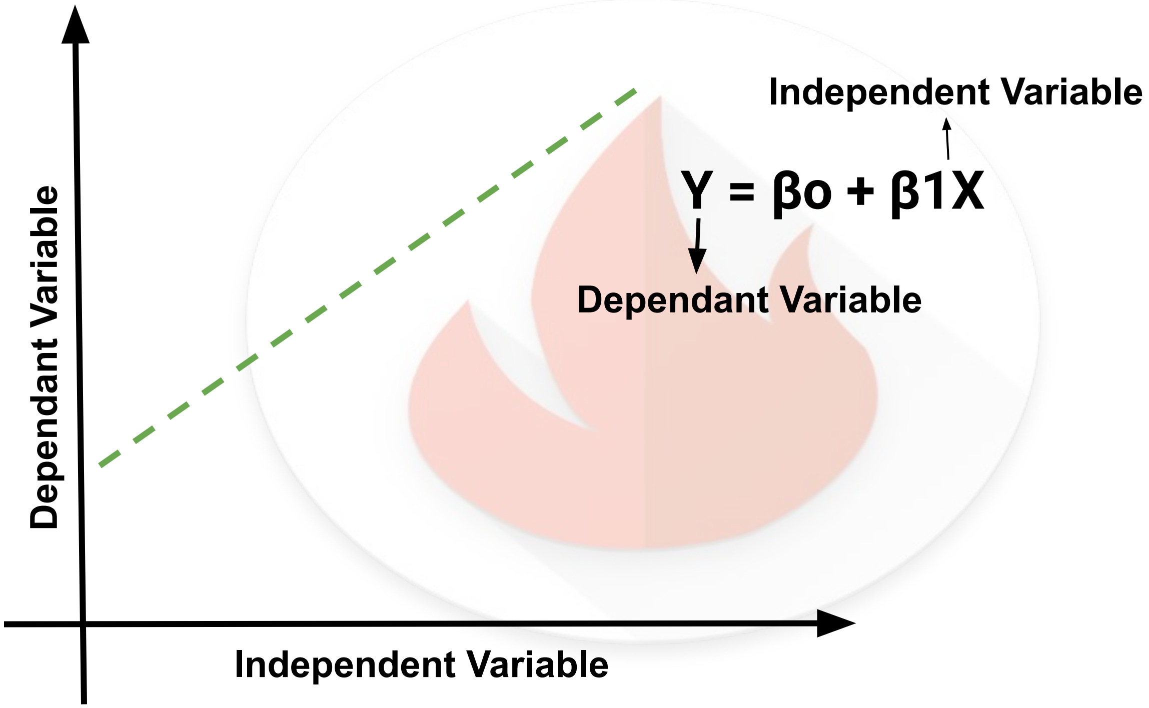
- Y is the response
- X is the feature
- $\beta 0$ is the intercept
- $\beta 1$ is the Multiplier / Unit change Together $\beta 0$ and $\beta 1$ are called the **model coefficients**

Simple Linear Regression

$$Y = \beta0 + \beta1 * X + \varepsilon$$

- The linear equation assigns one scale factor to each input value or column called a coefficient and represented by the capital **Greek** letter **Beta (B).**
- In higher dimensions when we have more than one input (x), the line is called a plane or a hyperplane.

Linear Regression



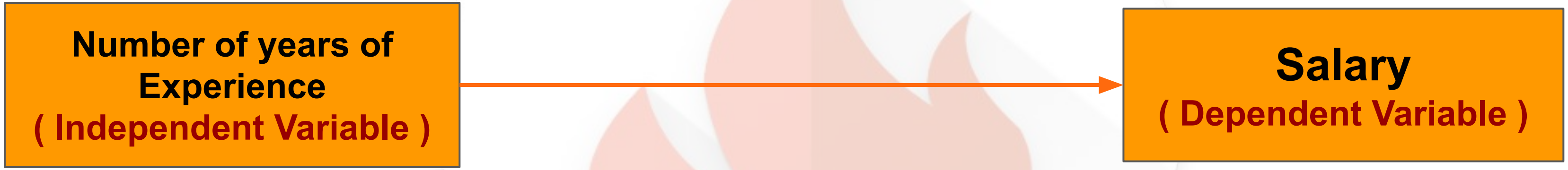
Independent Variable

$Y = \beta o + \beta 1X$

Dependant Variable

Linear Regression

Let us consider impact of a single variable for now.



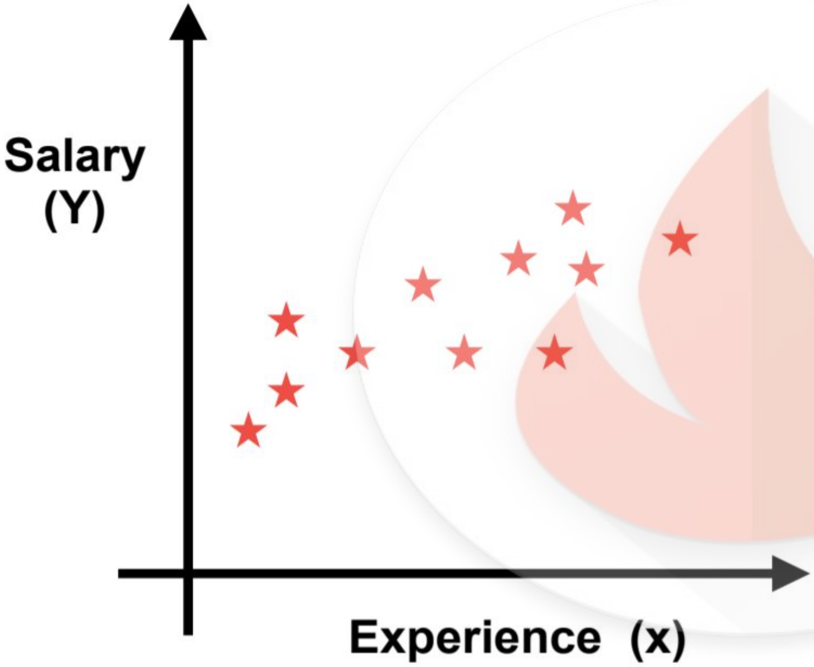
Number of years of Experience
(Independent Variable)

Salary
(Dependent Variable)

Linear Regression

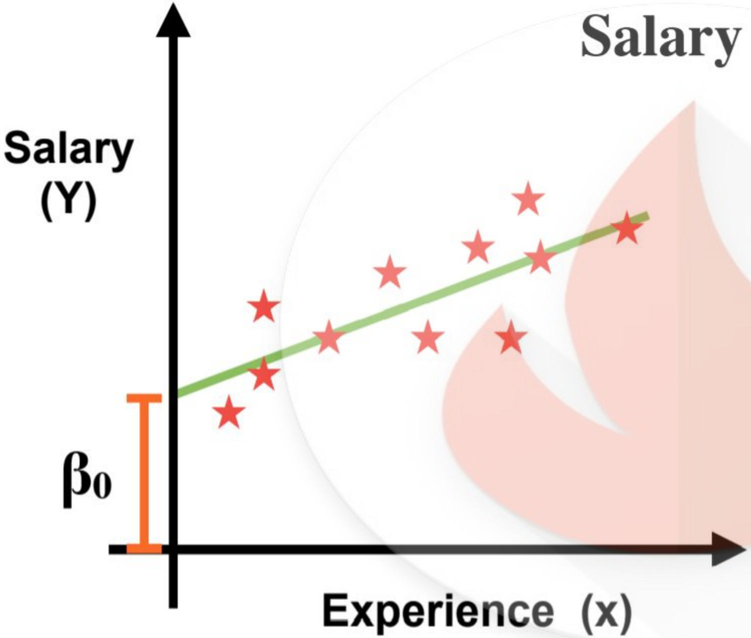
Experience (x)	Salary (Y)
1	10000
2	15000
3	25000
4	35000
5	46000

Linear Regression



$$Y = \beta_0 + \beta_1 * x$$

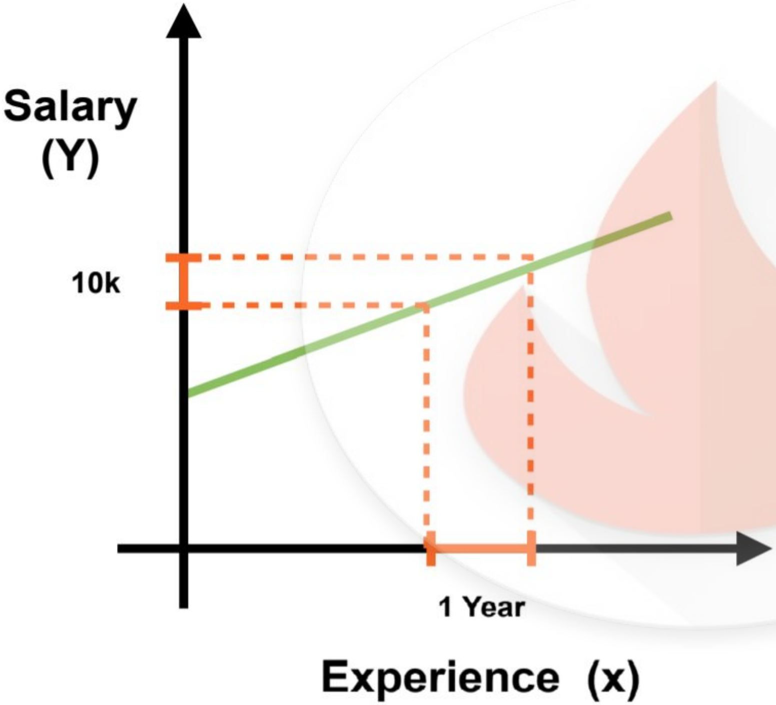
Linear Regression



$$\text{Salary} = \beta_0 + \beta_1 * \text{Experience}$$

$$\beta_0 = \$\ 30K$$

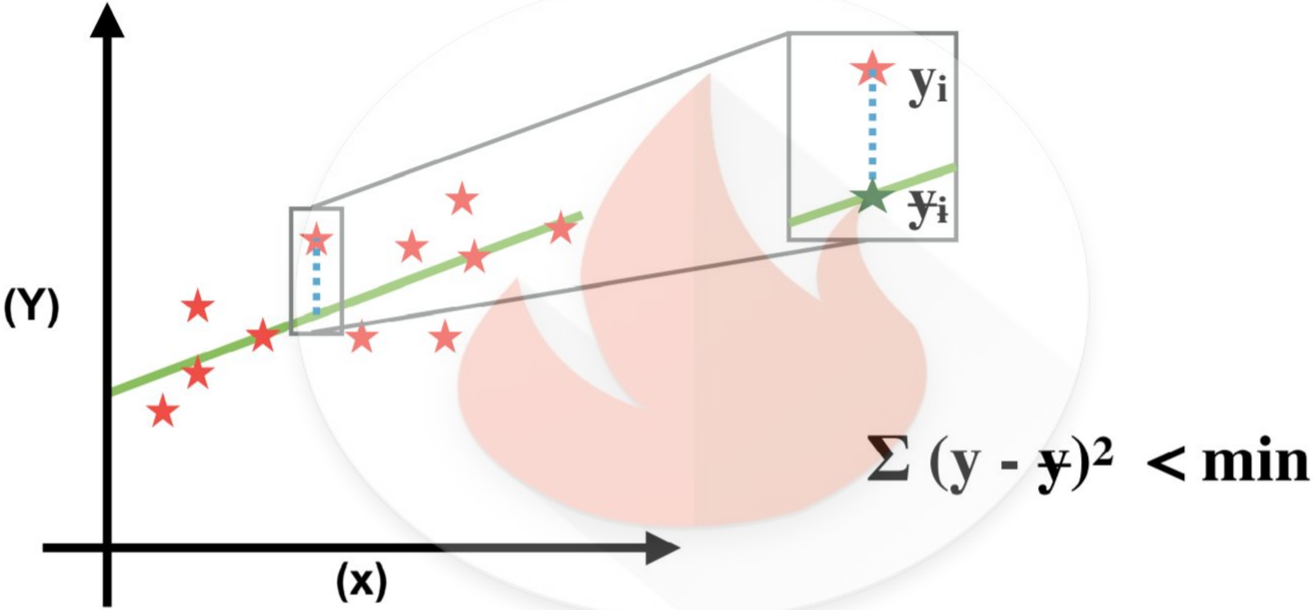
Linear Regression



$$Y = \beta_0 + \beta_1 * x$$

β_1 -Unit Change in x / Multiplier

Linear Regression



Linear Regression

Linear regression line using example,

Salary = β0 + β1 * (No. of years of Experience) + Ɛ

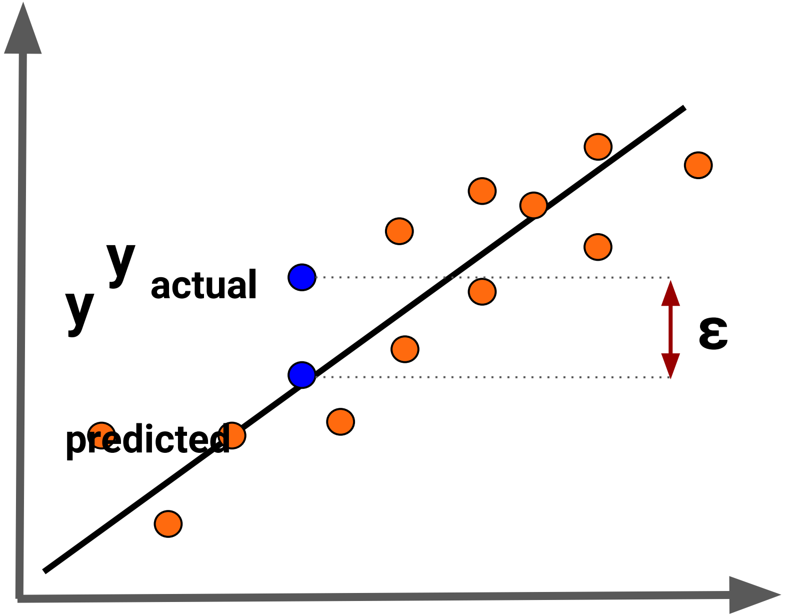
- y = set of values taken by dependent variable, Salary
- x = set of values taken by independent variable, No. of years of Exp.
- β0 = Salary value where the best fit line cuts the Y - axis (Salary)
- β1 = beta coefficient for Number of years of Experience
- ε = random error component

What is error term (Ɛ) ?

- Error term is the distance between observed value and the predicted value by the regression line.
- It is also called as residual.
- In our example,

Error term = Actual Salary - Predicted Salary for each observation

What is error term (Ɛ) ?



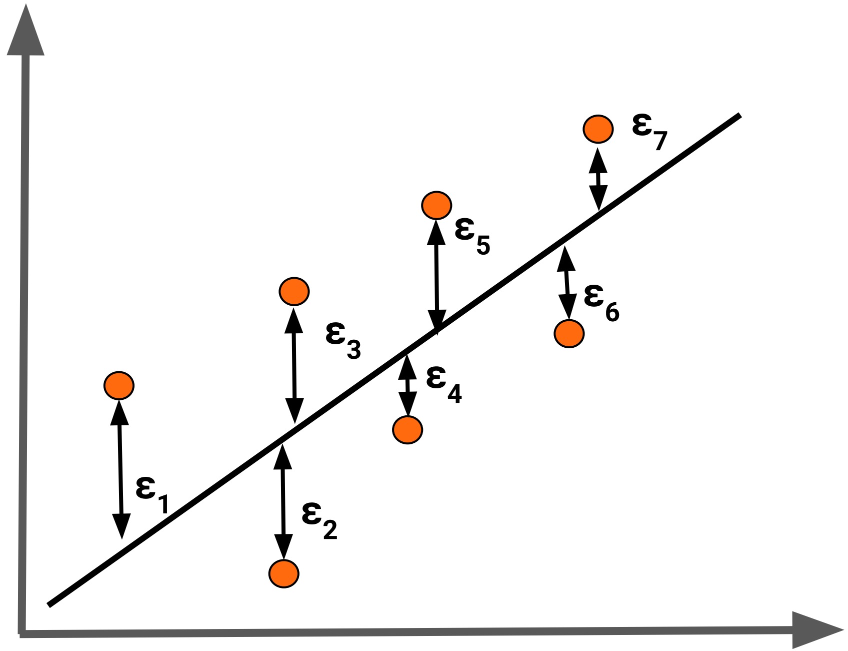
Equation of regression line is given by,

$$Y = \beta 0 + \beta 1 * x + \varepsilon$$

$$\therefore \ \varepsilon = y - (\beta 0 + \beta 1 * x)$$

$$\therefore \ \varepsilon = y_{\text{actual}} - y_{\text{predicted}}$$

What is error term (Ɛ) ?



To compute overall error,

$$\varepsilon_i = y_{\text{actual}} - y_{\text{predicted}}$$

Squared error,

$$\varepsilon_i^2 = (y_{\text{actual}} - y_{\text{predicted}})^2$$

Sum of squared errors = $\sum \varepsilon_i^2$

What is Best Fit Line

- The ordinary least square method is used to find the best fit line.
- This method aims to minimize the sum of squared error, using optimized value of β0 and β1.
- Min Error - Min $\sum_{i=0}^{n} = (y_i - \beta_i x_i)^2$

Linear Regression

Estimating(Learning) Model Coefficients : Generally speaking, coefficients are estimated using the least squares criterion, which means we find the line (mathematically) which minimises the sum of squared residuals (or "sum of squared errors")

Linear Regression Line : While doing linear regression our objective is to fit a line through the distribution which is nearest to most of the points. Hence reducing the distance (error term) of data points from the fitted line.

Interpretation of β coefficients

- β1 gives the amount of change in target / response variable per unit change in predictor variable.
- β0 is the y intercept which means when X=0, Y is β0.
- Depending on whether β's take a positive value k or - k the response variable increases or decreases respectively by k units for every one unit increment in a predictor variable, keeping all other predictor variables constant.

Find Model Coefficients

$Y = \beta o + \beta 1X$

$\beta 1 = cov\ (X,Y) / var\ (X)$

$\beta 0 = \bar{Y} - \beta 1 * \bar{x}$

Where,

$\bar{Y}$ - Mean of Y

$\bar{X}$ - Mean of X

Types of Linear Regression

1. Simple Linear Regression
2. Ordinary Least Squares
3. Gradient Descent
4. Regularization
    a. Lasso Regression
    b. Ridge Regression

Types of Linear Regression

1. **Simple Linear Regression:-**
   a. With simple linear regression when we have a single input, we can use statistics to estimate the coefficients.
   b. This requires that you calculate statistical properties from the data such as means, standard deviations, correlations and covariance.

Types of Linear Regression

2. Ordinary Least Squares:-

a. When we have more than one input we can use Ordinary Least Squares to estimate the values of the coefficients.

b. The Ordinary Least Squares procedure seeks to minimize the sum of the squared residuals.

c. This approach treats the data as a matrix and uses linear algebra operations to estimate the optimal values for the coefficients.

Types of Linear Regression

3. Gradient Descent:-

a. When there are one or more inputs you can use a process of optimizing the values of the coefficients by iteratively minimizing the error of the model on your training data.

b. The sum of the squared errors are calculated for each pair of input and output values.

c. A learning rate is used as a scale factor and the coefficients are updated in the direction towards minimizing the error.

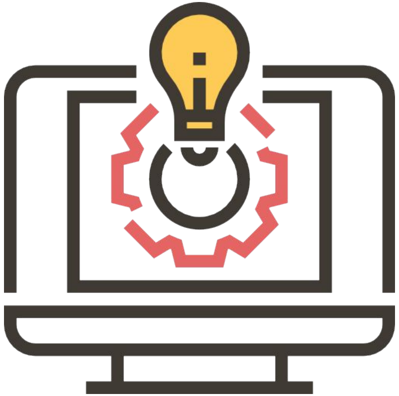
Types of Linear Regression

4. Regularization:-

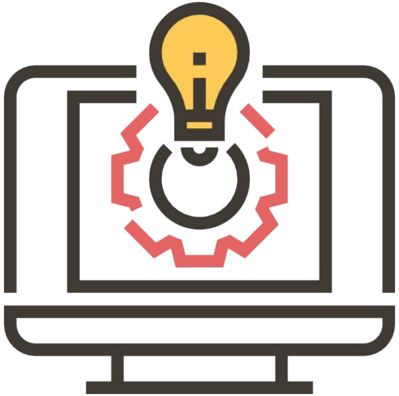
a. These seek to both minimize the sum of the squared error of the model on the training data (using ordinary least squares) but also to reduce the complexity of the model (like the number or absolute size of the sum of all coefficients in the model).

Types of Linear Regression

A. **Two Popular Methods**
   a. **Lasso Regression -**
      i. where Ordinary Least Squares is modified to also minimize the absolute sum of the coefficients (called L1 regularization).
   b. **Ridge Regression -**
      i. where Ordinary Least Squares is modified to also minimize the squared absolute sum of the coefficients (called L2 regularization).



THE END



Model Evaluation

Assumption for Evaluation

Assume that we have a collection of paired data containing the sample point (x , y), that

- $\hat{y}$ is the predicted value of y,
- $\bar{Y}$ is the Mean of the sample y-values.

Sum of Squares of Error

- The sum of squares of error (SSE) is the sum of squared differences between observed response variable and its predicted value.
- SSE is the measure of variability in the response variable remaining after considering the effect of dependent variable.
- It is the unexplained variation.
- Also known as Error Sum of Square (ESS) SSE =$\Sigma (y - \hat{y})^2$

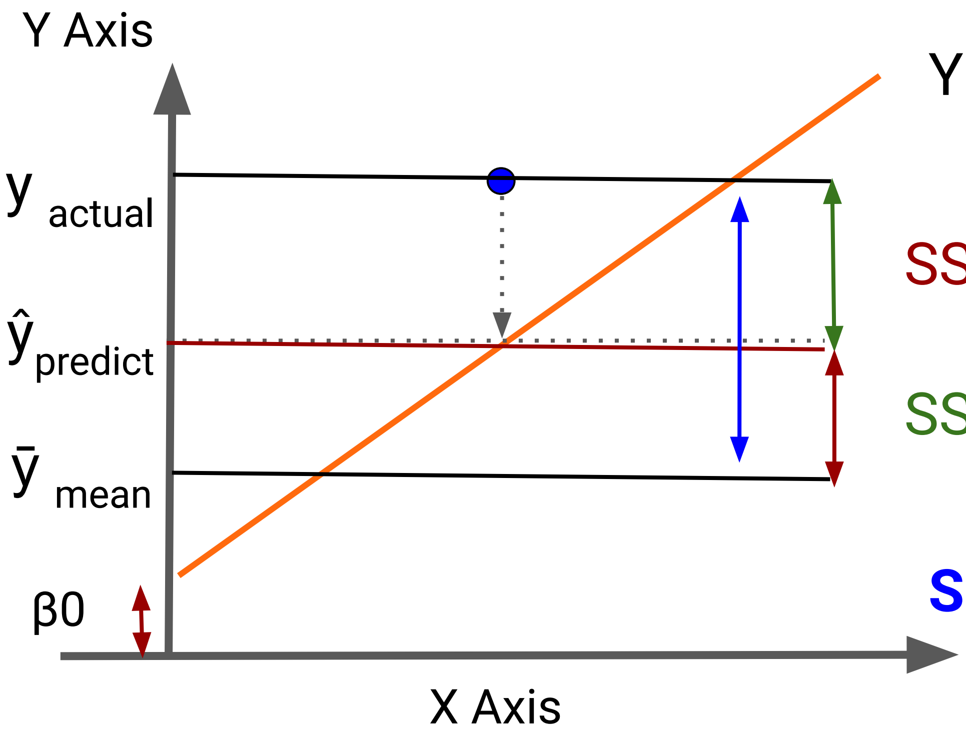
Sum of Squares Regression

- The sum of squares regression (SSR) is the sum of squared differences between the predicted value and the mean of the response variable.
- SSR is the measure of variability in the response variable considering the effect of dependent variable.
- It is the explained variation.
- Also known as Regression Sum of Square (RSS) = $\boldsymbol{\Sigma(\hat{y} - \bar{y})}^2$

Sum of Squares Total

- The sum of squares total (SST) is the sum of squared differences between the observation and its mean.
- It can be seen as the total variation of the response variable about its mean value.
- SST is the measure of variability in the response variable without considering the effect of dependent variable.
- Also known as Total Sum of Square (TSS) = $\boldsymbol{\Sigma(y - \bar{y})^2}$

Anatomy of Regression Errors



$SSR = \Sigma(\hat{y} - \bar{y})^2$ | Explained

$SSE = \Sigma\ (\ y - \hat{y}\ )\ ^2$ | Unexplained

$SST = \Sigma(y - \bar{y})^2$ | Total

Total Variation

Total variation = Explained variation + Unexplained variation

SST = SSR + SSE

$$\Sigma(y - \bar{y})^2 = \Sigma(\hat{y} - \bar{y})^2 + \Sigma (y - \hat{y}) ^2$$

Evaluating the Algorithm

Coefficient of Determination (R Square) It suggests the proportion of variation in Y which can be explained with the independent variables. Mathematically:

R2 = SSR/SST

or R2 = Explained variation / Total variation

In other words, it explains the proportion of variation in the dependent variable that is explained by the independent variables.

Evaluating the Algorithm

Coefficient of Determination (R Square)

- R2 ranges from 0 to 1.
- R2 of 0 means that the dependent variable cannot be predicted from the independent variable
- R2 of 1 means the dependent variable can be predicted without error from the independent variable
- If the value of R2 is 0.912 then this suggests that 91.2% of the variation in Y can be explained with the help of given explanatory variables in that model. In other words, **it explains the proportion of variation in the dependent variable that is explained by the independent variables.**

Mean Absolute Error (MAE)

- MAE is robust to outliers

$$\mathbf{MAE = 1 / n \sum_{i=1}^{n} | y - \hat{y} |}$$

Mean Square Error (MSE)

- Squaring of error terms handles the negative values of error and also emphasizes larger errors.

$$MSE = 1 / n \sum_{i=1}^{n} (y - \hat{y})^2$$

Root Mean Square Srror (RMSE)

- Lower the value of RMSE, better is the fit of regression line.

$$\mathrm{RMSE} = \surd\left(1 / n \sum_{i=1}^{n} (y - \hat{y})^{2} \right)$$

ASSUMPTIONS OF LINEAR REGRESSION

ASSUMPTIONS

- Linear relationship
- Multivariate normality
- No or little multicollinearity
- No autocorrelation
- Homoscedasticity

ASSUMPTIONS

- **Linear relationship** : First, linear regression needs the relationship between the independent and dependent variables to be linear.
- It is also important to check for outliers since linear regression is sensitive to outlier effects. The linearity assumption can best be tested with scatter plots.
- It is important to check this assumption because if you fit a linear model to a non-linear one, the regression algorithm would fail to capture the trend.

ASSUMPTIONS

- **Linear relationship :**
- **What to do if linear relationship assumption isn’t met**
  - You can apply nonlinear transformations to the independent and dependent variables.
  - You can add another feature to the model.
    - For example, if the plot of x’ vs. y’ has a parabolic shape, then it might be possible to add x2 as an additional feature in the model.

ASSUMPTIONS

- **Multivariate normality** : Secondly, the linear regression analysis requires all variables to be multivariate normal. Means data should be normally distributed.
- If residuals are non-normally distributed, the estimation may become too wide or narrow.
- Ways to Check Normal Distribution
  - Distribution Plots
  - Q-Q Plots - “**quantile-quantile**”

ASSUMPTIONS

- **Homoscedasticity** (also known as homogeneity of variance.):
- It describes a situation in which the error term (that is, the “noise” or random disturbance in the relationship between the independent variables and the dependent variable) is the same across all values of the independent variables.
- This means the residuals are equal across the regression line.
- Simply put, residuals should have **constant variance**. If this condition is not followed, it is known as **“heteroscedasticity”**.

ASSUMPTIONS

- **Homoscedasticity** (also known as homogeneity of variance.):
- The plot will show a funnel-shaped pattern if **heteroscedasticity** exists.
- **How to Test if Homoscedasticity Assumption is met?**
    - Heteroscedasticity can also be computed using the statistical approach. They are as following:

The Breush – Pegan Test: It determines whether the variance of the residuals from regression depends on the values of the independent variables. If it is so then, heteroscedasticity is present.

ASSUMPTIONS

White Test: White test determines if the variance of the residuals in a regression analysis model is fixed or constant.

Methods to handle Heteroscedasticity:- There are two ways to handle the Heteroscedasticity.

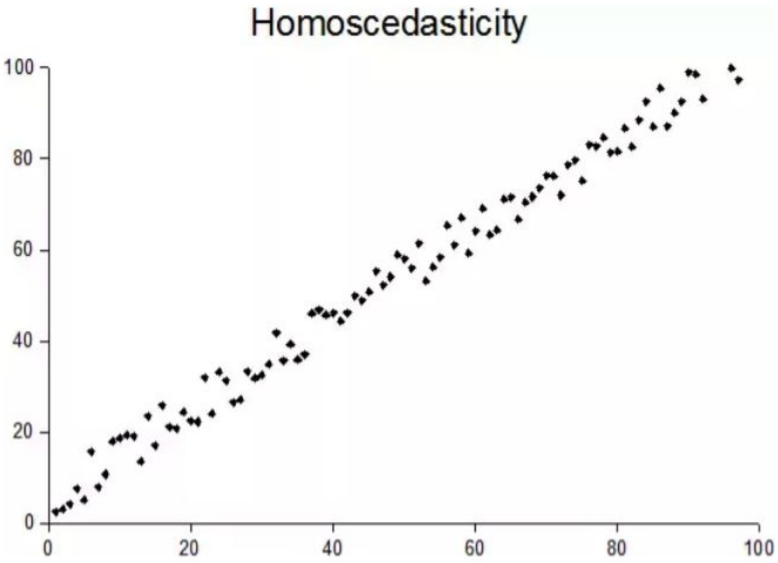
1. **Transform the Dependent Variables:-** We can transform the dependent variables to avoid heteroskedasticity. The most commonly used transformation is taking the **log** of dependent variables.

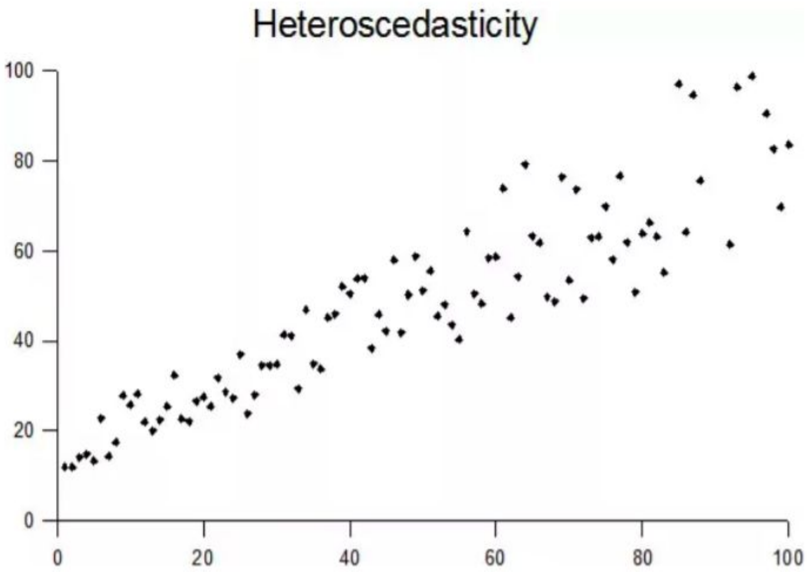
ASSUMPTIONS

1. **Transform the Dependent Variables:-** We can transform the dependent variables to avoid heteroskedasticity. The most commonly used transformation is taking the **log** of dependent variables.

- Using the **log of the target** variable helps to reduce the heteroskedasticity.

ASSUMPTIONS





ASSUMPTIONS

- **No autocorrelation** : The linear regression analysis requires that there is little or no autocorrelation in the data. Autocorrelation occurs when the residuals are not independent from each other.
- In other words when the value of y(x+1) is independent from the value of y(x).

ASSUMPTIONS

- **No autocorrelation :**
- **How to Test Autocorrelation Assumption is met?**
  - This is a plot of residuals vs. time.
  - Usually, most of the residual autocorrelations should fall within the **95%** confidence intervals around zero.
  - Which are located at about **+/- 2** -over the square root of N, where N is the dataset’s size.

ASSUMPTIONS

- It can also be checked using the Durbin-Watson test.
- Durbin-Watson test statistics can be implemented using statsmodels.durbin_watson() method.

$$\Sigma^{T}_{t=2}((e_t - e_{t-1})^2) / \Sigma^{T}_{t=1}e_t^{\ 2}$$

- If the value of **durbin_watson = 2**, it implies no autocorrelation
- If the value of durbin_watson lies between **0 and 2**, it implies **positive** autocorrelation.

ASSUMPTIONS

- If the value of durbin_watson lies between **2 and 4**, it implies **negative** autocorrelation.
- **Methods to Handle Autocorrelation**
    - Include the dummy variables in the data.
    - Predicted Generalized Least Squares.
    - Include a linear sequence, if the residuals showing a consistent increment or decrement in pattern.

ASSUMPTIONS

- **Multicollinearity** : It refers to a situation where a number of independent variables in a multiple regression model are closely correlated to one another.
- Multicollinearity generally occurs when there are high correlations between **two or more** predictor variables. In other words, one predictor variable can be used to predict the other. This creates redundant information, skewing the results in a regression model.

ASSUMPTIONS

- If you drop one correlated variable from the model, its predicted regression coefficients will change. It can lead to wrong conclusions and poor **performance of our model**.
- An easy way to detect multicollinearity is to calculate correlation coefficients for all pairs of predictor variables. If the correlation coefficient, r, is exactly **+1 or -1**, this is called perfect multicollinearity.
- If r is close to or exactly **-1 or +1**, one of the variables should be removed from the model if at all possible.

ASSUMPTIONS

- **How to Test Multicollinearity:-**
- We can test multicollinearity by using the following approaches.
  - Correlation Matrix
  - Tolerance Variance
  - Variance Inflation Factor

ASSUMPTIONS

- **Correlation Matrix:-**
  - Correlation represents the changes between the two variables.
  - While calculating Pearson's Bivariate Correlation matrix, it is recommended that the correlation coefficient among all independent variables should be less than 1.

ASSUMPTIONS

- **Tolerance Variance:-**
  - Tolerance helps us to determine the effect of one independent variable on all other independent variables.
  - Mathematically, it can be defined as $T = 1 - R^2$,
  - where R2 is computed by regressing the independent variable of concern onto the remaining independent variables.
  - If the value of T is **less than** 0.01, i.e., $T<0.01$, then your data has multicollinearity.

ASSUMPTIONS

- **Variance Inflation Factor:-**
  - VIF approach chooses each feature and regresses it against the remaining features.
  - It is calculated by using the given formula,

$$\mathbf{VIF = 1 / 1 \ - R^2}$$

  - If VIF **value <=4**, it implies no multicollinearity
  - If VIF **value>=10**, it implies significant multicollinearity

ASSUMPTIONS

- **Methods to handle Multicollinearity:-**
  - You can **drop** one of those features which are **highly correlated** in the given data.
  - Derive a **new feature** from collinear features and drop these features (used for making new features).

Multiple Linear Regression

If we have more than one independent variable then we will use Multiple Linear Regression.

It takes the following equation:

$$\mathbf{Y = \beta 0 + \beta 1 * X1, + \beta 2 * X2 + \beta 3 * X3 +..}$$

What does each term represent?

- Y is the response
- X1, X2, X3 are features
- β0 is the intercept
- β1, β2, β3 are different coefficients for X1, X2, X3 respectively

Thank you