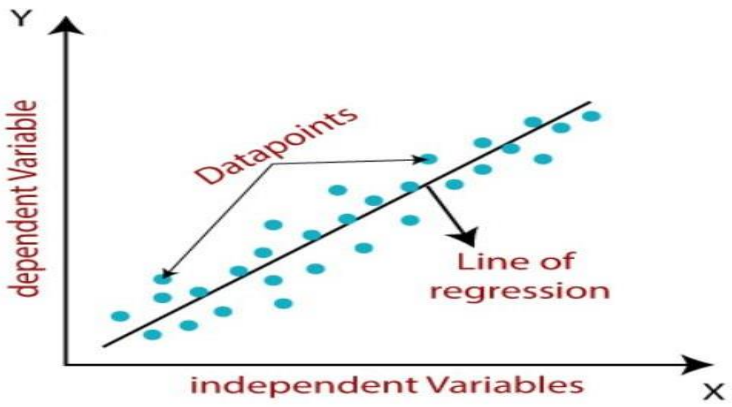
Linear Regression

- Linear regression is a statistical regression method which is used for predictive analysis.
- It is one of the very simple and easy algorithms which works on regression and shows the relationship between the continuous variables.
- It is used for solving the regression problem in machine learning.
- Linear regression shows the linear relationship between the independent variable (X-axis) and the dependent variable (Y-axis), hence called linear regression.
- If there is only one input variable (x), then such linear regression is called **simple linear regression**. And if there is more than one input variable, then such linear regression is called **multiple linear regression**.
- The relationship between variables in the linear regression model can be explained using the below image. Here we are predicting the salary of an employee on the basis of **the year of experience**.



$$y= a_0+a_1x+ \varepsilon$$

Here,

Y= Dependent Variable (Target Variable)
X= Independent Variable (predictor Variable)
a0= intercept of the line (Gives an additional degree of freedom)
a1 = Linear regression coefficient (scale factor to each input value).
ε = random error

The values for x and y variables are training datasets for Linear Regression model representation.

Types of Linear Regression

- **Simple Linear Regression:**

  If a single independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Simple Linear Regression.

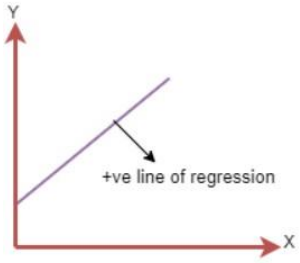
- **Multiple Linear regression:**

  If more than one independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Multiple Linear Regression.

Linear Regression Line

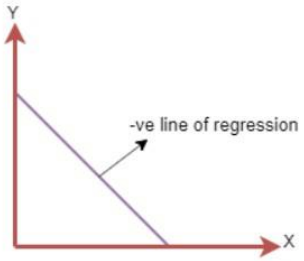
A linear line showing the relationship between the dependent and independent variables is called a **regression line**. A regression line can show two types of relationship:

- **Positive Linear Relationship:**
  If the dependent variable increases on the Y-axis and independent variable increases on X-axis, then such a relationship is termed as a Positive linear relationship.



The line equation will be: $Y= a_0+a_1x$

- **Negative Linear Relationship:**
  If the dependent variable decreases on the Y-axis and independent variable increases on the X-axis, then such a relationship is called a negative linear relationship.



The line of equation will be: $Y= -a_0+a_1x$

Assumption:

1. Linearity:

Assumption: The relationship between the independent variables and the dependent variable is linear.

Explanation This means that the change in the dependent variable (response) is proportional to the change in each independent variable (predictor) while holding

other variables constant.

Detection: Scatter plots of the variables can reveal linear or non-linear patterns. Additionally, residual plots can help identify deviations from linearity.

2. Independence:

Assumption: Observations in the dataset are independent of each other.

Explanation: The value of one observation should not be influenced by or dependent on the values of other observations.

Detection: Typically, independence is assumed in observational studies. In experimental studies, randomization helps ensure independence.

3. Homoscedasticity (Constant Variance):

Assumption: The variance of the error terms is constant across all levels of the independent variables.

Explanation: This implies that the spread of the residuals (errors) should be consistent across the range of predictor variables.

Detection: Residual plots (plot of residuals against predicted values) can help identify heteroscedasticity, where the spread of residuals varies with the predicted values.

4. Normality:

Assumption: The error terms (residuals) are normally distributed.

Explanation: This assumption implies that the distribution of residuals follows a normal (Gaussian) distribution with a mean of zero.

Detection: Histograms or QQ-plots of residuals can be used to assess normality. Shapiro-Wilk test or Kolmogorov-Smirnov test can also be employed for formal testing.

5. No Multicollinearity:

Assumption: Independent variables are not highly correlated with each other.

Explanation: Multicollinearity occurs when predictor variables are highly correlated, making it difficult to estimate individual effects accurately.

Detection: Correlation matrices or Variance Inflation Factor (VIF) can help identify multicollinearity. VIF values greater than 5 or 10 are often considered indicative of

multicollinearity.

These assumptions are crucial for the validity and reliability of the linear regression model. Violations of these assumptions may lead to biased estimates, inefficient predictions, or erroneous conclusions. Therefore, it's essential to assess these assumptions thoroughly before interpreting the results of a linear regression analysis.

Model Performance:

The Goodness of fit determines how the line of regression fits the set of observations. The process of finding the best model out of various models is called **optimization**. It can be achieved by below method:

1. R-squared method:

- R-squared is a statistical method that determines the goodness of fit.
- It measures the strength of the relationship between the dependent and independent variables on a scale of 0-100%.
- The high value of R-square determines the less difference between the predicted values and actual values and hence represents a good model.
- It is also called a **coefficient of determination,** or **coefficient of multiple determination** for multiple regression.
- It can be calculated from the below formula:

$$\text{R-squared} = \frac{\text{Explained variation}}{\text{Total Variation}}$$

Applications:

1.Predictive Modeling:

- Linear regression is widely used for predictive modeling in various fields such as finance, marketing, and healthcare.
- It can predict outcomes based on historical data, such as predicting sales based on advertising spending or forecasting stock prices based on historical trends.

2. Understanding Relationships:

- Linear regression helps in understanding the relationships between variables.
- It can identify the strength and direction of relationships between

independent and dependent variables, aiding in hypothesis testing and theory development.

3. Causal Inference:

- Linear regression can be used to assess the impact of independent variables on the dependent variable.
- It helps in identifying causal relationships, such as determining the effect of education level on income or analyzing the impact of environmental factors on health outcomes.

4. Model Interpretability:

- Linear regression models are relatively simple and interpretable compared to more complex models.
- Coefficients in linear regression provide insights into the magnitude and direction of the relationship between variables, making it easier to explain to stakeholders.

5. Variable Selection:

- Linear regression can aid in variable selection by identifying the most influential predictors.
- Techniques such as stepwise regression or regularization methods help in selecting the most relevant variables for the model.

Limitations:

1. Assumption of Linearity:

- Linear regression assumes a linear relationship between independent and dependent variables. If the relationship is non-linear, linear regression may

provide inaccurate predictions or biased estimates.

2. Sensitive to Outliers:

- Linear regression is sensitive to outliers in the data, which can significantly impact model performance.
- Outliers can distort the estimation of coefficients and affect the fit of the model.

3. Assumption of Independence:

- Linear regression assumes that observations in the dataset are independent of each other.
- Violation of this assumption, such as in time series data or spatial data, can lead to biased estimates and incorrect inferences.

4. Multicollinearity:

- Multicollinearity, where independent variables are highly correlated with each other, can affect the stability and interpretability of the regression coefficients.
- It can lead to inflated standard errors and difficulty in interpreting the individual effects of predictors.

5. Limited to Linear Relationships:

- Linear regression is limited to modeling linear relationships between variables. It may not capture complex relationships, such as interactions or non-linear associations, without transformation or additional modeling techniques.

6. Overfitting:

- In the case of multiple linear regression with a large number of predictors, there is a risk of overfitting the model to the training data, leading to poor generalization to new data.

Understanding these applications and limitations is essential for effectively applying linear regression in practice and interpreting the results accurately. It's important to consider these factors and choose appropriate modeling techniques based on the specific characteristics of the data and research objectives.