Linear Regression

OVERVIEW & PURPOSE

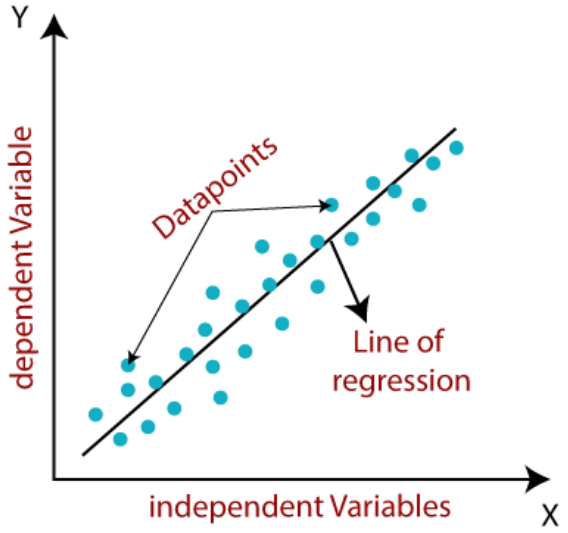
In this session, participants will learn the widely used techniques of linear regression in supervised learning to predict the desired outcome

OBJECTIVE

- Data Preparation: To preprocess the data to make it ready for use in a machine learning model. This can include steps like feature engineering, normalization, and splitting the data into training and testing sets.

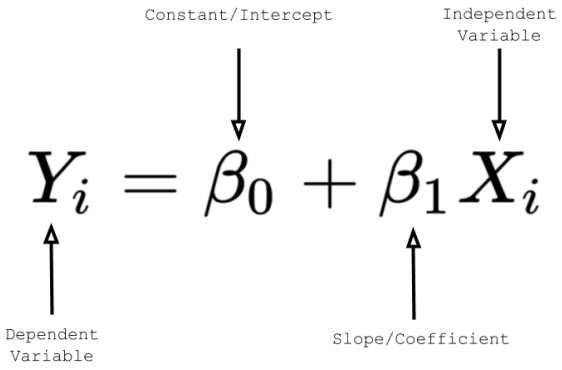
What is Linear Regression?

Linear regression is a statistical method and a type of supervised learning algorithm used in machine learning and data science. It attempts to model the relationship between two or more variables by fitting a linear equation to the observed data.



The simplest form of linear regression (simple linear regression) involves two variables: an independent variable X and a dependent variable Y.

- For simple linear regression, the relationship is represented as:



- For multiple linear regression, when there are more than one independent variables, the equation becomes:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_i X_i$$

Y : Dependent variable
β_0 : Intercept
β_i : Slope for X_i
X = Independent variable

Why is it Used?

Predictive Analysis: It's a useful tool for predicting a quantitative response.
Understanding Relationships: It helps to understand the strength and direction (positive/negative) of the relationship between the dependent and independent variables.
Simple and Interpretable: The results of linear regression are easy to understand and interpret in real-world scenarios.
Foundation for Other Methods: Many other methods, including logistic regression and ANOVA, are extensions of linear regression.

Important Steps in Linear Regression:

Data Collection: Gathering relevant data that will help in building the model.

Data Cleaning: Removing outliers, handling missing values, and ensuring the quality of data.

Exploratory Data Analysis (EDA): Understanding the distribution of data, checking for multicollinearity, and visualizing relationships between variables.

Feature Selection/Engineering: Picking the relevant features which contribute most to the output. Engineering new features if necessary.

Model Fitting: Using training data to train the linear regression model. This involves determining the coefficients

Validation: Using a validation set (or cross-validation) to check how well the model performs on unseen data.

Assumption Checking: Linear regression makes several assumptions like linearity, independence of errors, homoscedasticity (constant variance of errors), and normality of errors. These need to be checked and if violated, necessary transformations or other methods should be applied.

Prediction: Using the model to predict the dependent variable for new data.

Model Evaluation: Using metrics like Mean Absolute Error (MAE), Mean Squared Error (MSE), or R-squared to evaluate the performance of the model.

Iterate: Depending on the performance and business requirements, one might need to iterate over some of the above steps to improve the model.

Deployment: If the model's performance is satisfactory, it can be deployed in a real-world scenario to make predictions on new data.

Reading Material:

- https://www.geeksforgeeks.org/ml-linear-regression/
- https://www.w3schools.com/python/python_ml_linear_regression.asp
- https://www.analyticsvidhya.com/blog/2021/05/all-you-need-to-know-about-your-first-machine-learning-model-linear-regression/

Resources:

- https://www.youtube.com/watch?v=8jazNUpO3lQ
- https://www.youtube.com/watch?v=zUQr6HAAKp4
- https://www.youtube.com/watch?v=lzGKRSvs5HM
- https://www.youtube.com/watch?v=UZPfbG0jNec