Linear regressor

What is Linear Regression?

- Linear regression is a type of **supervised machine learning** algorithm that computes the linear relationship between the dependent variable and one or more independent features by fitting a linear equation to observed data.
- Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable.
- When there is only one independent feature, it is known as `Simple Linear Regression`, and when there are more than one feature, it is known as `Multiple Linear Regression`.
- Similarly, when there is only one dependent variable, it is considered `Univariate Linear Regression`, while when there are more than one dependent variables, it is known as `Multivariate Regression``.`

Importance of Linear Regression

Linear regression's strength lies in its interpretability. The model's equation reveals clear coefficients that explain how each independent variable affects the dependent variable, leading to a better grasp of the underlying dynamics. Its simplicity is advantageous, making linear regression transparent, easy to use, and a fundamental concept for more complex algorithms.

Beyond its predictive capabilities, linear regression serves as the foundation for many advanced models. Techniques such as regularization and support vector machines are inspired by linear regression, broadening its applicability. Moreover, linear regression plays a crucial role in assumption testing, allowing researchers to validate essential assumptions about the data.

Key assumptions of effective linear regression

Assumptions to be considered for success with linear-regression analysis:

- **For each variable**: Consider the number of valid cases, mean and standard deviation.

- **For each model**: Consider regression coefficients, correlation matrix, part and partial correlations, multiple R, R2, adjusted R2, change in R2, standard error of the estimate, analysis-of-variance table, predicted values and residuals. Also, consider 95-percent-confidence intervals for each regression coefficient, variance-covariance matrix, variance inflation factor, tolerance, Durbin-Watson test, distance measures (Mahalanobis, Cook and leverage values), DfBeta, DfFit, prediction intervals and case-wise diagnostic information.
- **Plots**: Consider scatterplots, partial plots, histograms and normal probability plots.
- **Data**: Dependent and independent variables should be quantitative. Categorical variables, such as religion, major field of study or region of residence, need to be recoded to binary (dummy) variables or other types of contrast variables.
- **Other assumptions**: For each value of the independent variable, the distribution of the dependent variable must be normal. The variance of the distribution of the dependent variable should be constant for all values of the independent variable. The relationship between the dependent variable and each independent variable should be linear and all observations should be independent.

Types of Linear Regression

There are two main types of linear regression:

1. **Simple Linear Regression**
2. **Multiple Linear Regression**

1. Simple Linear Regression

This is the simplest form of linear regression, and it involves only one independent variable and one dependent variable. The equation for simple linear regression is:

$y=\beta 0+\beta 1Xy=\beta 0+\beta 1X$

where:

- Y is the dependent variable
- X is the independent variable

- β0 is the intercept
- β1 is the slope

2. Multiple Linear Regression

This involves more than one independent variable and one dependent variable. The equation for multiple linear regression is:

$y=\beta 0+\beta 1X+\beta 2X+.........\beta nX$$y=\beta 0+\beta 1X+\beta 2X+.........\beta nX$

where:

- Y is the dependent variable
- X1, X2, ..., Xp are the independent variables
- β0 is the intercept
- β1, β2, ..., βn are the slopes

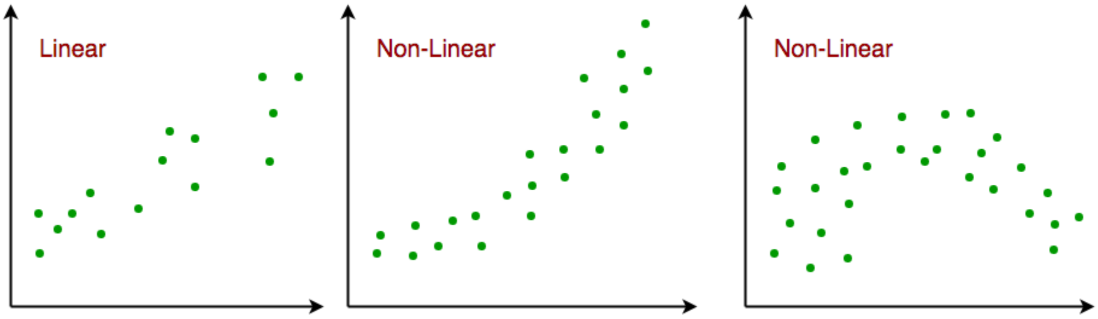
The goal of the algorithm is to find the best Fit Line equation that can predict the values based on the independent variables.

In regression set of records are present with X and Y values and these values are used to learn a function so if you want to predict Y from an unknown X this learned function can be used. In regression we have to find the value of Y, So, a function is required that predicts continuous Y in the case of regression given X as independent features.

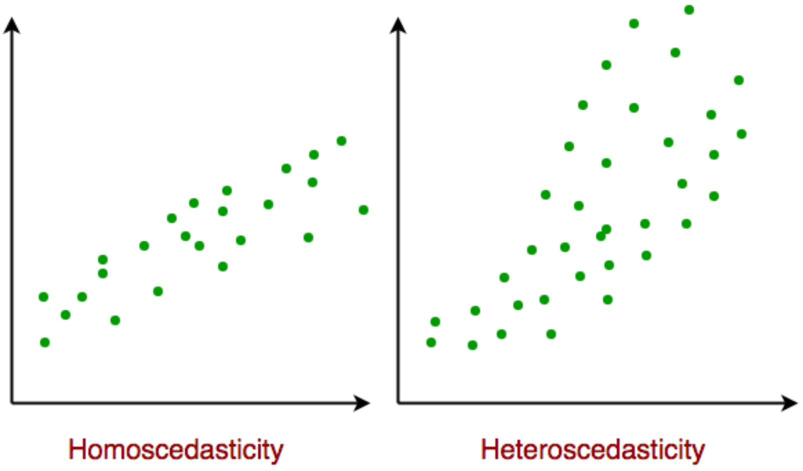
Assumptions of Simple Linear Regression

Linear regression is a powerful tool for understanding and predicting the behavior of a variable, however, it needs to meet a few conditions in order to be accurate and dependable solutions.

1. **Linearity**: The independent and dependent variables have a linear relationship with one another. This implies that changes in the dependent variable follow those in the independent variable(s) in a linear fashion. This means that there should be a straight line that can be drawn through the data points. If the relationship is not linear, then linear regression will not be an accurate model.



2. **Independence**: The observations in the dataset are independent of each other. This means that the value of the dependent variable for one observation does not depend on the value of the dependent variable for another observation. If the observations are not independent, then linear regression will not be an accurate model.
3. **Homoscedasticity**: Across all levels of the independent variable(s), the variance of the errors is constant. This indicates that the amount of the independent variable(s) has no impact on the variance of the errors. If the variance of the residuals is not constant, then linear regression will not be an accurate model.



4. **Normality**: The residuals should be normally distributed. This means that the residuals should follow a bell-shaped curve. If the residuals are not normally distributed, then linear regression will not be an accurate model.

Assumptions of Multiple Linear Regression

For Multiple Linear Regression, all four of the assumptions from Simple Linear Regression apply. In addition to this, below are few more:

1. **No multicollinearity**: There is no high correlation between the independent variables. This indicates that there is little or no correlation between the independent variables. Multicollinearity occurs when two or more independent variables are highly correlated with each other, which can make it difficult to determine the individual effect of each variable on the dependent variable. If there is multicollinearity, then multiple linear regression will not be an accurate model.
2. **Additivity:** The model assumes that the effect of changes in a predictor variable on the response variable is consistent regardless of the values of the other variables. This assumption implies that there is no interaction between variables in their effects on the dependent variable.
3. **Feature Selection:** In multiple linear regression, it is essential to carefully select the independent variables that will be included in the model. Including irrelevant or redundant variables may lead to overfitting and complicate the interpretation of the model.
4. **Overfitting:** Overfitting occurs when the model fits the training data too closely, capturing noise or random fluctuations that do not represent the true underlying relationship between variables. This can lead to poor generalization performance on new, unseen data.

Examples of linear regression

Certainly! Here are five examples of linear regression applications with brief explanations:

1. **Predicting Sales**: Linear regression can be used to predict sales based on factors like advertising expenditure, seasonality, and previous sales data. This helps businesses plan their marketing strategies and manage inventory more effectively.
2. **Forecasting Stock Prices**: In finance, linear regression is used to forecast stock prices based on historical price data, trading volume, and other relevant factors. This information is valuable for investors and financial analysts.
3. **Healthcare Outcome Analysis**: Linear regression can analyze the impact of a new drug or treatment on patient outcomes, helping healthcare providers

make informed decisions about patient care and treatment strategies.

4. **Environmental Impact Assessment**: Linear regression can predict pollution levels based on factors like meteorological data, industrial activity, and traffic patterns. This information is crucial for environmental impact assessments and policy-making.
5. **Education Performance Prediction**: In education, linear regression can predict student performance based on factors like attendance, study hours, and socioeconomic background. This information can help educators identify at-risk students and provide targeted support.