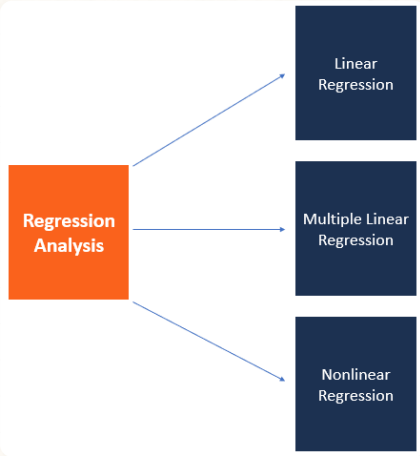
Regression Analysis

In stastical modelling, regression analysis is a set of stastical processes for estimating the relationships between a dependent variable (often called the "outcome variable") and one or more independent variables (often called predictors or features).

It can be utilized to assess the strength of the relationship between variables and for modeling the future relationship between them.



Regression Analysis – Linear model assumptions

Linear regression analysis is based on six fundamental assumptions:

1. The dependent and independent variables show a linear relationship between the slope and the intercept.
2. The independent variable is not random.
3. The value of the residual (error) is zero.
4. The value of the residual (error) is constant across all observations.
5. The value of the residual (error) is not correlated across all observations.
6. The residual (error) values follow the normal distribution.

Regression Analysis – Simple linear regression

Simple linear regression is a model that assesses the relationship between a dependent variable and an independent variable. The simple linear model is expressed using the following equation:

$$Y = a + bX + \epsilon$$

Where:

- **Y** – Dependent variable
- **X** – Independent (explanatory) variable
- **a** – Intercept
- **b** – Slope
- **ϵ** – Residual (error)

Regression Analysis – Multiple linear regression

Multiple linear regression analysis is essentially similar to the simple linear model, with the exception that multiple independent variables are used in the model. The mathematical representation of multiple linear regression is:

$$Y = a + bX_1 + cX_2 + dX_3 + \epsilon$$

Where:

- **Y** – Dependent variable
- **X_1, X_2, X_3** – Independent (explanatory) variables
- **a** – Intercept
- **b, c, d** – Slopes
- **ϵ** – Residual (error)

Multiple linear regression follows the same conditions as the simple linear model. However, since there are several independent variables in multiple linear analysis, there is another mandatory condition for the model:

- **Non-collinearity:** Independent variables should show a minimum of correlation with each other. If the independent variables are highly correlated with each other, it will be difficult to assess the true relationships between the dependent and independent variables.

Regression analysis is used in statr to find trends in data.

It will provide an equation for a graph so that we can make prediction about the data.

Correlation – It is a measure of the linear relationship beth two variables.

It is worth pointing out that regression analysis always involves the use of historical data in order to understand the relationship between two or more variables. Those who use regression analysis assume that whatever relationship existed in the past will continue to exist in the present or the future. Some observers refer to this as the issue of the lag between the past and the present and the future. Nonetheless, regression analysis is a popular forecasting and estimating technique. Although many users might find the mathematics involved quite difficult, the technique itself is relatively easy to use, especially when a model or template has previously been developed.

Terminologies related to regression analysis

1. Outliers

Suppose there is an observation in the dataset which is having a very high or very low value as compared to the other observations in the data, i.e. it does not belong to the population, such an observation is called an outlier. In simple words, it is extreme value. An outlier is a problem because many times it hampers the results we get.

2. Multicollinearity

When the independent variables are highly correlated to each other then the variables are said to be multicollinear. Many types of regression techniques assumes multicollinearity should not be present in the dataset. It is because it causes problems in ranking variables based on its importance. Or it makes job difficult in selecting the most important independent variable (factor).

3. Heteroscedasticity

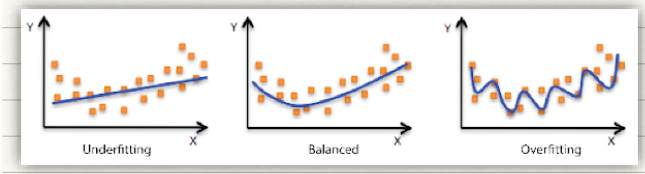
When dependent variable's variability is not equal across values of an independent variable, it is called heteroscedasticity. **Example -**As one's income increases, the variability of food consumption will increase. A poorer person will spend a rather constant amount by always eating inexpensive food; a wealthier person may occasionally buy inexpensive food and at other times eat expensive meals. Those with higher incomes display a greater variability of food consumption.

4. Underfitting and Overfitting

When we use unnecessary explanatory variables it might lead to overfitting. Overfitting means that our algorithm works well on the training set but is unable to perform better on the test sets. It is also known as problem of **high variance.**

When our algorithm works so poorly that it is unable to fit even training set well then it is said to **underfit the data.**It is also known as **problem of high bias.**

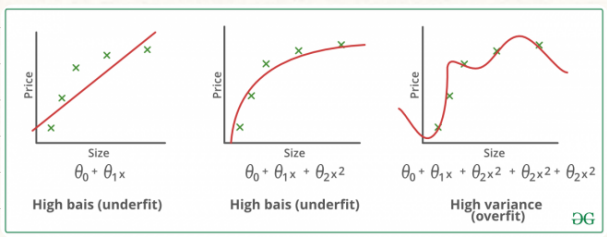
In the following diagram we can see that fitting a linear regression (straight line in fig 1) would underfit the data i.e. it will lead to large errors even in the training set. Using a polynomial fit in fig 2 is balanced i.e. such a fit can work on the training and test sets well, while in fig 3 the fit will lead to low errors in training set but it will not work well on the test set.

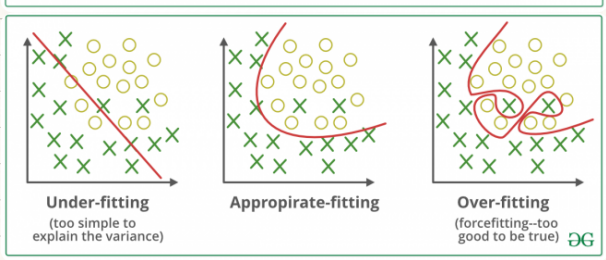


Underfitting occur when we have less data and we are trying to build a linear model using non-linear data.

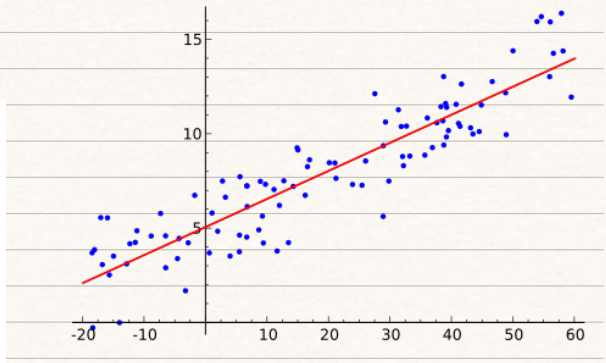
Overfitting occurs when we train our model with lots of data.

It starts learning from noise or inaccurate data entries in our datasets.





Linear Regression :-



Linear regression is one of the most basic **types of regression in machine learning**. The linear regression model consists of a predictor variable and a dependent variable related linearly to each other. In case the data involves more than one independent variable, then linear regression is called multiple linear regression models.

The below-given equation is used to denote the linear regression model:

y=mx+c+e

where m is the slope of the line, c is an intercept, and e represents the error in the model.

The best fit line is determined by varying the values of m and c. The predictor error is the difference between the observed values and the predicted value. The values of m and c get selected in such a way that it gives the minimum predictor error. It is important to note that a simple linear regression model is susceptible to outliers. Therefore, it should not be used in case of big size data.

> *When you have only 1 independent variable and 1 dependent variable, it is called simple linear regression.*
> *When you have more than 1 independent variable and 1 dependent variable, it is called Multiple linear regression.*

The equation of multiple linear regression is listed below -

$$y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_k X_k + \varepsilon$$

Here 'y' is the dependent variable to be estimated, and X are the independent variables and ε is the error term. βi's are the regression coefficients.

Assumptions of linear regression:

1. There must be a linear relation between independent and dependent variables.
2. There should not be any outliers present.
3. No heteroscedasticity
4. Sample observations should be independent.
5. Error terms should be normally distributed with mean 0 and constant variance.
6. Absence of multicollinearity and auto-correlation.

Important Points:

- There must be **linear relationship** between independent and dependent variables
- Multiple regression suffers from **multicollinearity, autocorrelation, heteroskedasticity**.
- Linear Regression is very sensitive to **Outliers**. It can terribly affect the regression line and eventually the forecasted values.
- Multicollinearity can increase the variance of the coefficient estimates and make the estimates very sensitive to minor changes in the model. The result is that the coefficient estimates are unstable
- In case of multiple independent variables, we can go with **forward selection, backward elimination** and **step wise approach** for selection of most significant independent variables.

Simple Linear Regression

$$y = b_0 + b_1 * x_1$$

Multiple Linear Regression

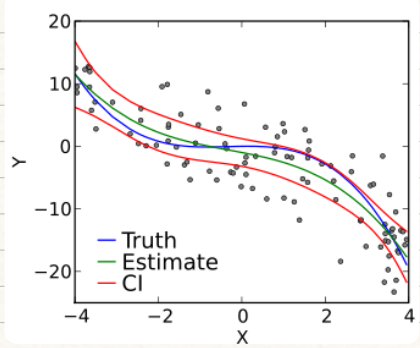
Dependent variable (DV) Independent variables (IVs)

$$y = b_0 + b_1 * x_1 + b_2 * x_2 + \ldots + b_n * x_n$$

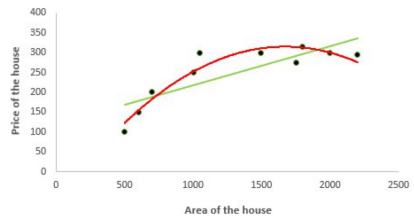
Polynomial Regression :-

Polynomial Regression is another one of the **types of regression analysis** techniques in machine learning, which is the same as Multiple Linear Regression with a little modification. In Polynomial Regression, the relationship between independent and dependent variables, that is X and Y, is denoted by the n-th degree.

It is a linear model as an estimator. Least Mean Squared Method is used in Polynomial Regression also. The best fit line in Polynomial Regression that passes through all the data points is not a straight line, but a curved line, which depends upon the power of X or value of n.



It is a technique to fit a nonlinear equation by taking polynomial functions of independent variable.
In the figure given below, you can see the red curve fits the data better than the green curve. Hence in the situations where the relation between the dependent and independent variable seems to be non-linear we can deploy **Polynomial Regression Models.**



Simple Linear Regression

$$y = b_0 + b_1 x_1$$

Multiple Linear Regression

$$y = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_n x_n$$

Polynomial Linear Regression

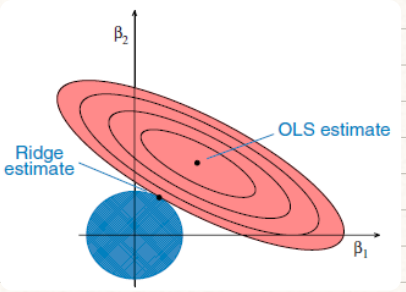
$$y = b_0 + b_1 x_1 + b_2 x_1^2 + \ldots + b_n x_1^n$$

Ridge Regression :-

This is another one of the **types of regression in machine learning** which is usually used when there is a high correlation between the independent variables. This is because, in the case of multi collinear data, the least square estimates give unbiased values. But, in case the collinearity is very high, there can be some bias value. Therefore, a bias matrix is introduced in the equation of Ridge Regression. This is a powerful regression method where the model is less susceptible to overfitting.

Below is the equation used to denote the Ridge Regression, where the introduction of λ (lambda) solves the problem of multicollinearity:

β = (X^{T}X + λ*I)^{-1}X^{T}y



Non-linear Regression - When the relationship between x and y is non-linear, we use non-linear regression.
We use non-linear function.