ASSIGNMENT – 39

MACHINE LEARNING

Q10and Q15 are subjective answer type questions, Answer them in your own words briefly.

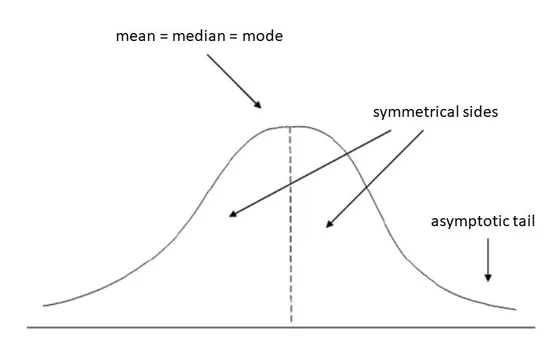
10. What do you understand by the term Normal Distribution?

The normal distribution is defined as continuous probability distribution that is symmetrical on both sides of the mean in such a way that the right side of the centre is a mirror image of the left side.

The area under the curve of normal distribution represents probability and the total area under the curve sums to one.

Most of the continuous data values in a normal distribution are cluster around the mean, and the further a value is from the mean.

For a perfect normal distribution the mean, median and mode will be of the same value, visually represented by the peak of the curve.



The normal distribution is the very important probability distribution in statistics because many continuous data in nature and psychology displays this bell-shaped curve when compiled and graphed.

For example, if we randomly sampled 100 individuals, we would expect to see a normal distribution frequency curve for many continuous variables, such as IQ, height, weight and blood pressure.

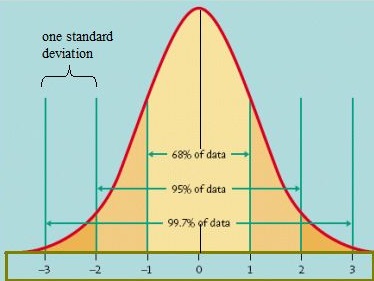
We can standardize the values (raw scores) of a normal distribution by converting them into z-scores. This procedure allows researchers to determine the proportion of the values that fall within a specified number of standard deviations from the mean (i.e., calculate the empirical rule).

Normal distributions become more apparent (i.e., perfect) the finer the level of measurement and the larger the sample from a population.

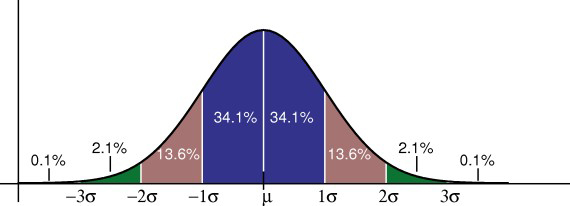
You can also calculate coefficients which tell us about the size of the distribution tails in relation to the bump in the middle of the bell curve. For example, Kolmogorov Smirnov and Shapiro-Wilk tests can be calculated using SPSS.

These tests compare your data to a normal distribution and provide a p-value, which if significant (p < .05) indicates your data is different to a normal distribution (thus, on this occasion we do not want a significant result and need a p-value higher than 0.05).

One way of figuring out how data are distributed is to plot them in a graph. If the data is evenly distributed, you may come up with a **bell curve**. A bell curve has a small percentage of the points on both tails and the bigger percentage on the inner part of the curve. In the **standard normal model**, about 5 percent of your data would fall into the “tails” (coloured darker orange in the image below) and 90 percent will be in between. For example, for test scores of students, the normal distribution would show 2.5 percent of students getting very low scores and 2.5 percent getting very high scores. The rest will be in the middle; not too high or too low. The shape of the standard normal distribution looks like this:



The [empirical rule](https://www.statisticshowto.com/probability-and-statistics/statistics-definitions/empirical-rule-2/) tells you what percentage of your data falls within a certain number of [standard deviations](https://www.statisticshowto.com/probability-and-statistics/standard-deviation/) from the [mean](https://www.statisticshowto.com/mean):  
• 68% of the data falls within one [standard deviation](https://www.statisticshowto.com/probability-and-statistics/standard-deviation/) of the [mean](https://www.statisticshowto.com/mean).  
• 95% of the data falls within two [standard deviations](https://www.statisticshowto.com/probability-and-statistics/standard-deviation/) of the [mean](https://www.statisticshowto.com/mean).  
• 99.7% of the data falls within three [standard deviations](https://www.statisticshowto.com/probability-and-statistics/standard-deviation/) of the [mean](https://www.statisticshowto.com/mean).



The standard deviation controls the spread of the distribution. A smaller standard deviation indicates that the data is tightly clustered around the [mean](https://www.statisticshowto.com/mean); the normal distribution will be taller. A larger standard deviation indicates that the data is spread out around the [mean](https://www.statisticshowto.com/mean); the normal distribution will be flatter and wider.

11. How do you handle missing data? What imputation techniques do you recommend?

[Handling](https://analyticsindiamag.com/get-started-preparing-data-machine-learning/) the missing values is one of the greatest challenges faced by analysts, because making the right decision on how to handle it generates robust data models. Let us look at different ways of imputing the missing values.

We will be using libraries in Python such as NumPy, Pandas and SciKit Learn to handle these values.

**1. Deleting Rows**

This method commonly used to handle the null values. Here, we either delete a particular row if it has a null value for a particular feature and a particular column if it has more than 70-75% of missing values. This method is advised only when there are enough samples in the data set. One has to make sure that after we have deleted the data, there is no addition of bias. Removing the data will lead to loss of information which will not give the expected results while predicting the output.

### Pros:

* Complete removal of data with missing values results in robust and highly accurate model
* Deleting a particular row or a column with no specific information is better, since it does not have a high weightage

### Cons:

* Loss of information and data
* Works poorly if the percentage of missing values is high (say 30%), compared to the whole dataset

## **2. Replacing with Mean/Median/Mode**

This strategy can be applied on a feature which has numeric data like the age of a person or the ticket fare. We can calculate the mean, median or mode of the feature and replace it with the missing values. This is an approximation which can add variance to the data set. But the loss of the data can be negated by this method which yields better results compared to removal of rows and columns. Replacing with the above three approximations are a statistical approach of handling the missing values. This method is also called as leaking the data while training. Another way is to approximate it with the deviation of neighbouring values. This works better if the data is linear.

Pros:

* This is a better approach when the data size is small
* It can prevent data loss which results in removal of the rows and columns

Cons:

* Imputing the approximations add variance and bias
* Works poorly compared to other multiple-imputations method

## **3. Assigning an Unique Category**

A categorical feature will have a definite number of possibilities, such as gender, for example. Since they have a definite number of classes, we can assign another class for the missing values. Here, the features Cabin and Embarked have missing values which can be replaced with a new category, say, U for ‘unknown’. This strategy will add more information into the dataset which will result in the change of variance. Since they are categorical, we need to find one hot encoding to convert it to a numeric form for the algorithm to understand it. Let us look at how it can be done in Python:

### Pros:

* Less possibilities with one extra category, resulting in low variance after one hot encoding — since it is categorical
* Negates the loss of data by adding an unique category

Cons:

* Adds less variance
* Adds another feature to the model while encoding, which may result in poor performance

## **4. Predicting the Missing Values**

Using the features which do not have missing values, we can predict the nulls with the help of a machine learning algorithm. This method may result in better accuracy, unless a missing value is expected to have a very high variance. We will be using linear [regression](https://analyticsindiamag.com/top-6-regression-algorithms-used-data-mining-applications-industry/) to replace the nulls in the feature ‘age’, using other available features. One can experiment with different algorithms and check which gives the best accuracy instead of sticking to a single algorithm.

Pros:

* Imputing the missing variable is an improvement as long as the bias from the same is smaller than the omitted variable bias
* Yields unbiased estimates of the model parameters

Cons:

* Bias also arises when an incomplete conditioning set is used for a categorical variable
* Considered only as a proxy for the true values

## **5. Using Algorithms Which Support Missing Values**

KNN is a machine learning algorithm which works on the principle of distance measure. This algorithm can be used when there are nulls present in the dataset. While the algorithm is applied, KNN considers the missing values by taking the majority of the K nearest values. In this particular dataset, taking into account the person’s age, sex, class etc, we will assume that people having same data for the above mentioned features will have the same kind of fare.

Unfortunately, the SciKit Learn library for the K – Nearest Neighbour algorithm in Python does not support the presence of the missing values.

Another algorithm which can be used here is Random Forest. This model produces a robust result because it works well on non-linear and the categorical data. It adapts to the data structure taking into consideration of the high variance or the bias, producing better results on large datasets.

### Pros:

* Does not require creation of a predictive model for each attribute with missing data in the dataset
* Correlation of the data is neglected

### Cons:

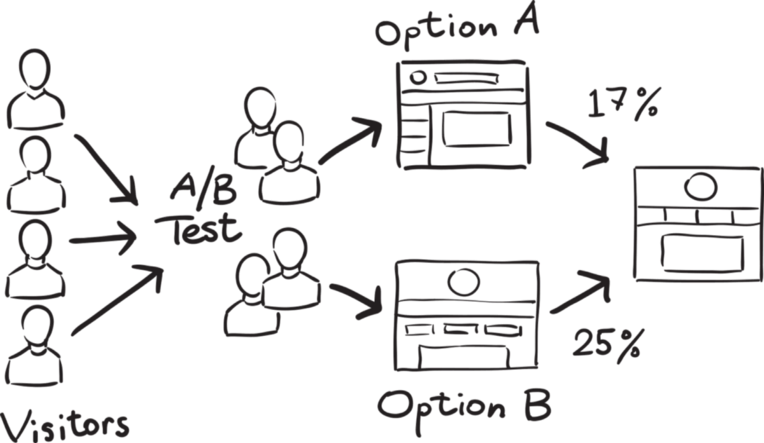
* Is a very time-consuming process and it can be critical in data mining where large databases are being extracted,
* Choice of distance functions can be Euclidean, Manhattan etc. which is do not yield a robust result.

12. What is A/B testing?

A/B testing is a basic randomized control experiment. It is a way to compare the two versions of a variable to find out which performs better in a controlled environment.

For instance, let’s say you own a company and want to increase the sales of your product. Here, either you can use random experiments, or you can apply scientific and statistical methods. A/B testing is one of the most prominent and widely used statistical tools.

In the above scenario, you may divide the products into two parts – A and B. Here A will remain unchanged while you make significant changes in B’s packaging. Now, on the basis of the response from customer groups who used A and B respectively, you try to decide which is performing better.



It is a hypothetical testing methodology for making decisions that estimate population parameters based on sample statistics. The**population** refers to all the customers buying your product, while the **sample** refers to the number of customers that participated in the test.

An example the logic and methodology behind the concept of A/B testing.

Let’s say there is an e-commerce company XYZ. It wants to make some changes in its newsletter format to increase the traffic on its website. It takes the original newsletter and marks it A and makes some changes in the language of A and calls it B. Both newsletters are otherwise the same in colour, headlines, and format.

A/B testing works best when testing incremental changes, such as UX changes, new features, ranking, and page load times. Here you may compare pre and post-modification results to decide whether the changes are working as desired or not.

A/B testing doesn’t work well when testing major changes, like new products, new branding, or completely new user experiences. In these cases, there may be effects that drive higher than normal engagement or emotional responses that may cause users to behave in a different manner.

To summarize, A/B testing is at least a 100-year-old statistical methodology but in its current form, it comes in the 1990s. Now it has become more eminent with the online environment and availability for big data. It is easier for companies to conduct the test and utilize the results for better user experience and performance.

There are many tools available for conducting A/B testing but being a data scientist, you must understand the factors working behind it. Also, you must be aware of the statistics in order to validate the test and prove its statistical significance.

13. Is mean imputation of missing data acceptable practice?

The process of replacing null values in a data collection with the data’s mean is known as mean imputation.

Mean imputation is typically considered terrible practice since it ignores feature correlation. Consider the following scenario: we have a table with age and fitness scores, and an eight-year-old has a missing fitness score. If we average the fitness scores of people between the ages of 15 and 80, the eighty-year-old will appear to have a significantly greater fitness level than he actually does.

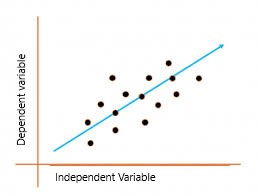
Second, mean imputation decreases the variance of our data while increasing bias. As a result of the reduced variance, the model is less accurate and the confidence interval is narrower.

14. What is linear regression in statistics?

Linear Regression is a supervised learning algorithm in machine learning that supports finding the linearcorrelation among variables. The result or output of the regression problem is a real or continuous value.

## **Simple Linear Regression**

Linear regression is a quiet and the simplest statistical regression method used for predictive analysis in machine learning. Linear regression shows the linear relationship between the independent(predictor) variable i.e. X-axis and the dependent(output) variable i.e. Y-axis, called linear regression*.*If there is a single input variable **X**(dependent variable), such linear regression is called simple linear regression.



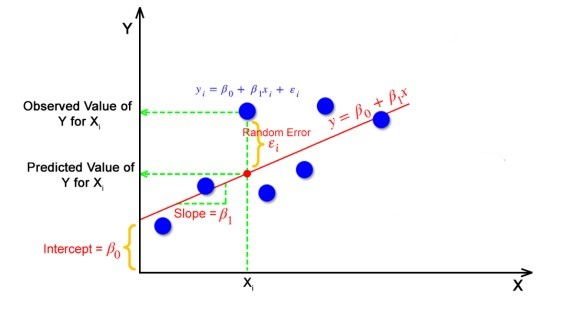
The above graph presents the linear relationship between the output(y) variable and predictor(X) variables.  The blue line is referred to as the*best fit* straight line. Based on the given data points, we attempt to plot a line that fits the points the best.

*T*o calculate best-fit line linear regression uses a traditional slope-intercept form which is given below,

**Yi= β0 + β1Xi**

where Yi = Dependent variable,  **β0** = constant/Intercept, **β1** = Slope/Intercept, **Xi** = Independent variable.

This algorithm explains the linear relationship between the dependent(output) variable y and the independent(predictor) variable X using a straight line  Y= B0 + B1 X.



But how the linear regression finds out which is the best fit line?

The goal of the linear regression algorithm is to get the **best values for B0 and B1** to find the best fit line. The best fit line is a line that has the least error which means the error between predicted values and actual values should be minimum.

**Random Error (Residuals)**

In regression, the difference between the observed value of the dependent variable(**yi**) and the predicted value(predicted) is called the residuals.

**εi=**  **ypredicted** –   **yi**

**where y predicted =   B0 + B1 Xi**

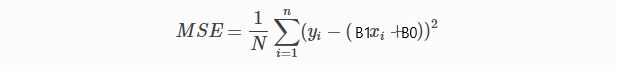
In simple terms, the best fit line is a line that fits the given scatter plot in the best way. Mathematically, the best fit line is obtained by minimizing the Residual Sum of Squares (RSS).

## Cost Function for Linear Regression

The [cost function](https://www.analyticsvidhya.com/blog/2021/03/data-science-101-introduction-to-cost-function/) helps to work out the optimal values for B0 and B1, which provides the best fit line for the data points.

In Linear Regression, generally **Mean Squared Error (MSE)** cost function is used, which is the average of squared error that occurred between the **predicted** and **yi**.

We calculate MSE using simple linear equation y=mx+b:

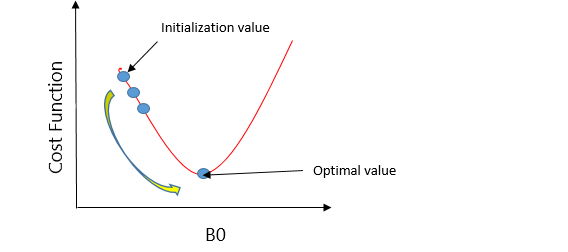


Using the MSE function, we’ll update the values of B0 and B1 such that the MSE value settles at the minima.  These parameters can be determined using the gradient descent method such that the value for the cost function is minimum.

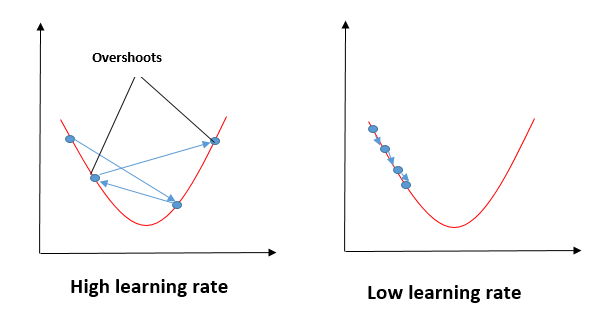
**Gradient Descent for Linear Regression**

Gradient Descent is one of the optimization algorithms that optimize the cost function (objective function) to reach the optimal minimal solution. To find the optimum solution we need to reduce the cost function (MSE) for all data points. This is done by updating the values of B0 and B1 iteratively until we get an optimal solution.

A regression model optimizes the gradient descent algorithm to update the coefficients of the line by reducing the cost function by randomly selecting coefficient values and then iteratively updating the values to reach the minimum cost function.



Let’s take an example to understand this. Imagine a U-shaped pit. And you are standing at the uppermost point in the pit, and your motive is to reach the bottom of the pit. Suppose there is a treasure at the bottom of the pit, and you can only take a discrete number of steps to reach the bottom. If you opted to take one step at a time, you would get to the bottom of the pit in the end but, this would take a longer time. If you decide to take larger steps each time, you may achieve the bottom sooner but, there’s a probability that you could overshoot the bottom of the pit and not even near the bottom. In the gradient descent algorithm, the number of steps you’re taking can be considered as the learning rate, and this decides how fast the algorithm converges to the minima.



We need to minimize the cost function J. One of the ways to achieve this is to apply the batch gradient descent algorithm. In batch gradient descent, the values are updated in each iteration. (Last two equations shows the updating of values)

The partial derivates are the gradients, and they are used to update the values of B0 and B1. Alpha is the learning rate.

## Evaluation Metrics for Linear Regression

The strength of any linear regression model can be assessed using various evaluation metrics. These evaluation metrics usually provide a measure of how well the observed outputs are being generated by the model.

The most used metrics are,

1. Coefficient of Determination or R-Squared (R2)
2. Root Mean Squared Error (RSME) and Residual Standard Error (RSE)

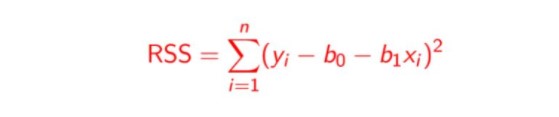
### 1. Coefficient of Determination or R-Squared (R2)

R-Squared is a number that explains the amount of variation that is explained/captured by the developed model. It always ranges between 0 & 1 . Overall, the higher the value of R-squared, the better the model fits the data.

Mathematically it can be represented as,

                                           R2 = 1 – ( RSS/TSS )

* Residual sum of Squares (RSS) is defined as the sum of squares of the residual for each data point in the plot/data. It is the measure of the difference between the expected and the actual observed output.

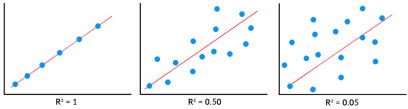


* Total Sum of Squares (TSS) is defined as the sum of errors of the data points from the mean of the response variable. Mathematically TSS is,

Total Sum of Squares

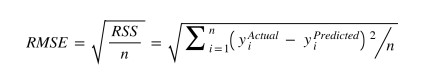
where y hat is the mean of the sample data points.

The significance of R-squared is shown by the following figures,

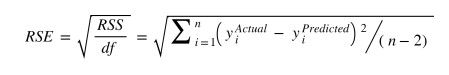


### 2. Root Mean Squared Error

The Root Mean Squared Error is the square root of the variance of the residuals. It specifies the absolute fit of the model to the data i.e. how close the observed data points are to the predicted values. Mathematically it can be represented as,



To make this estimate unbiased, one has to divide the sum of the squared residuals by the degrees of freedom rather than the total number of data points in the model. This term is then called the Residual Standard Error(RSE). Mathematically it can be represented as,

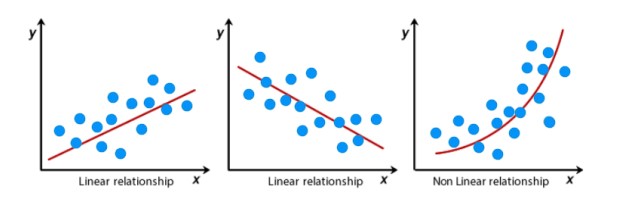


R-squared is a better measure than RSME. Because the value of Root Mean Squared Error depends on the units of the variables (i.e. it is not a normalized measure), it can change with the change in the unit of the variables.

## **Assumptions of Linear Regression**

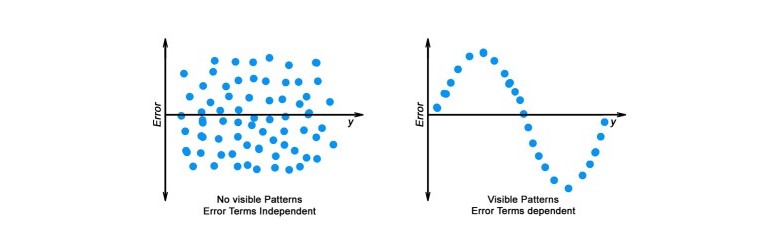
Regression is a parametric approach, which means that it makes assumptions about the data for the purpose of analysis. For successful regression analysis, it’s essential to validate the following assumptions.

1. **Linearity of residuals**: There needs to be a linear relationship between the dependent variable and independent variable(s).

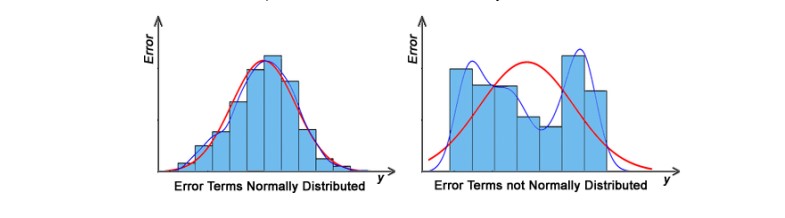


*2.*Independence of residuals: The error terms should not be dependent on one another (like in time-series data wherein the next value is dependent on the previous one). There should be no correlation between the residual terms. The absence of this phenomenon is known as Autocorrelation.

There should not be any visible patterns in the error terms.

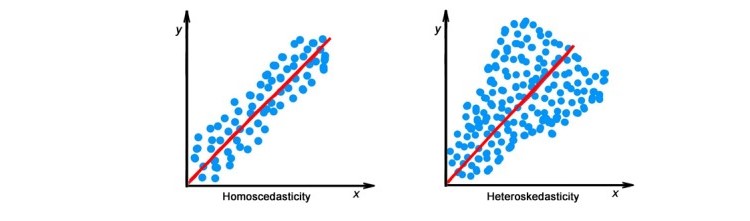


**3. Normal distribution of residuals:** The mean of residuals should follow a normal distribution with a mean equal to zero or close to zero. This is done in order to check whether the selected line is actually the line of best fit or not.  
If the error terms are non-normally distributed, suggests that there are a few unusual data points that must be studied closely to make a better model.



**4.** The equal variance of residuals: The error terms must have constant variance. This phenomenon is known as Homoscedasticity.

The presence of non-constant variance in the error terms is referred to as Heteroscedasticity. Generally, non-constant variance arises in the presence of outliers or extreme leverage values.



## **Hypothesis in Linear Regression**

Once you have fitted a straight line on the data, you need to ask, “Is this straight line a significant fit for the data?” Or “Is the beta coefficient explain the variance in the data plotted?” And here comes the idea of hypothesis testing on the beta coefficient. The Null and Alternate hypotheses in this case are:

H0: B1 = 0

HA: B1 **≠** 0

To test this hypothesis, we use a **t-test,**test statistics for the beta coefficient is given by,

### Assessing the model fit

Some other parameters to assess a model are:

1. **t statistic:** It is used to determine the p-value and hence, helps in determining whether the coefficient is significant or not
2. **F statistic**: It is used to assess whether the overall model fit is significant or not. Generally, the higher the value of the F-statistic, the more significant a model turns out to be.

## **Multiple Linear Regression**

Multiple linear regression is a technique to understand the relationship between a *single*dependent variable and *multiple*independent variables.

The formulation for multiple linear regression is also similar to simple linear regression with

the small change that instead of having one beta variable, you will now have betas for all the variables used. The formula is given as:

 Y = B0 + B1X1 + B2X2 + … + BpXp + **ε**

## **Considerations of Multiple Linear Regression**

All the four assumptions made for Simple Linear Regression still hold true for Multiple Linear Regression along with a few new additional assumptions.

1. **Overfitting**: When more and more variables are added to a model, the model may become far too complex and usually ends up memorizing all the data points in the training set. This phenomenon is known as the overfitting of a model. This usually leads to high training accuracy and very low-test accuracy.
2. **Multicollinearity**: It is the phenomenon where a model with several independent variables, may have some variables interrelated.
3. **Feature Selection:** With more variables present, selecting the optimal set of predictors from the pool of given features (many of which might be redundant) becomes an important task for building a relevant and better model.

## **Multicollinearity**

As multicollinearity makes it difficult to find out which variable is actually contributing towards the prediction of the response variable, it leads one to conclude incorrectly, the effects of a variable on the target variable.  Though it does not affect the precision of the predictions, it is essential to properly detect and deal with the multicollinearity present in the model, as random removal of any of these correlated variables from the model causes the coefficient values to swing wildly and even change signs.

Multicollinearity can be detected using the following methods.

1. **Pairwise Correlations:** Checking the pairwise correlations between different pairs of independent variables can throw useful insights in detecting multicollinearity.
2. **Variance Inflation Factor (VIF):** Pairwise correlations may not always be useful as it is possible that just one variable might not be able to completely explain some other variable but some of the variables combined could be ready to do this. Thus, to check these sorts of relations between variables, one can use VIF. VIF basically explains the relationship of one independent variable with all the other independent variables. VIF is given by,

multicollinearity VIF

where i refers to the ith variable which is being represented as a linear combination of the rest of the independent variables.

The common heuristic followed for the VIF values is if VIF > 10 then the value is definitely high and it should be dropped. And if the VIF=5 then it may be valid but should be inspected first. If VIF < 5, then it is considered a good vif value.

## **Overfitting and Underfitting in Linear Regression**

There have always been situations where a model performs well on training data but not on the test data. While training models on a dataset, overfitting, and underfitting are the most common problems faced by people.

Before understanding overfitting and underfitting one must know about bias and variance.

**Bias**

Bias is a measure to determine how accurate is the model likely to be on future unseen data. Complex models, assuming there is enough training data available, can do predictions accurately. Whereas the models that are too naive, are very likely to perform badly with respect to predictions. Simply, Bias is errors made by training data.

Generally, linear algorithms have a high bias which makes them fast to learn and easier to understand but in general, are less flexible. Implying lower predictive performance on complex problems that fail to meet the expected outcomes.

**Variance:**

Variance is the sensitivity of the model towards training data, that is it quantifies how much the model will react when input data is changed.

Ideally, the model shouldn’t change too much from one training dataset to the next training data, which will mean that the algorithm is good at picking out the hidden underlying patterns between the inputs and the output variables.

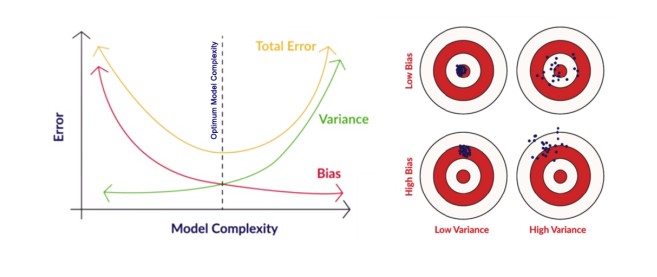
Ideally, a model should have lower variance which means that the model doesn’t change drastically after changing the training data(it is generalizable). Having higher variance will make a model change drastically even on a small change in the training dataset.

Let’s understand what is a bias-variance trade-off is.

## Bias Variance Trade-off

The aim of any supervised machine learning algorithm is to achieve low bias and low variance as it is more robust. So that the algorithm should achieve better performance.

There is no escape from the relationship between bias and variance in machine learning.



There is an inverse relationship between bias and variance,

* An increase in bias will decrease the variance.
* An increase in the variance will decrease the bias.

There is a trade-off that plays between these two concepts and the algorithms must find a balance between bias and variance.

As a matter of fact, one cannot calculate the real bias and variance error terms because we do not know the actual underlying target function.

Now coming to the overfitting and underfitting.

## **Overfitting**

When a model learns each and every pattern and noise in the data to such extent that it affects the performance of the model on the unseen future dataset, it is referred to as overfitting. The model fits the data so well that it interprets noise as patterns in the data.

When a model has low bias and higher variance it ends up memorizing the data and causing overfitting. Overfitting causes the model to become specific rather than generic. This usually leads to high training accuracy and very low-test accuracy.

Detecting overfitting is useful, but it doesn’t solve the actual problem. There are several ways to prevent overfitting, which are stated below:

* Cross-validation
* If the training data is too small to train add more relevant and clean data.
* If the training data is too large, do some feature selection and remove unnecessary features.
* Regularization

## **Underfitting**

Underfitting is not often discussed as often as overfitting is discussed. When the model fails to learn from the training dataset and is also not able to generalize the test dataset, is referred to as underfitting. This type of problem can be very easily detected by the performance metrics.

When a model has high bias and low variance it ends up not generalizing the data and causing underfitting. It is unable to find the hidden underlying patterns from the data. This usually leads to low training accuracy and very low-test accuracy. The ways to prevent underfitting are stated below,

* Increase the model complexity
* Increase the number of features in the training data
* Remove noise from the data.

15. What are the various branches of statistics?

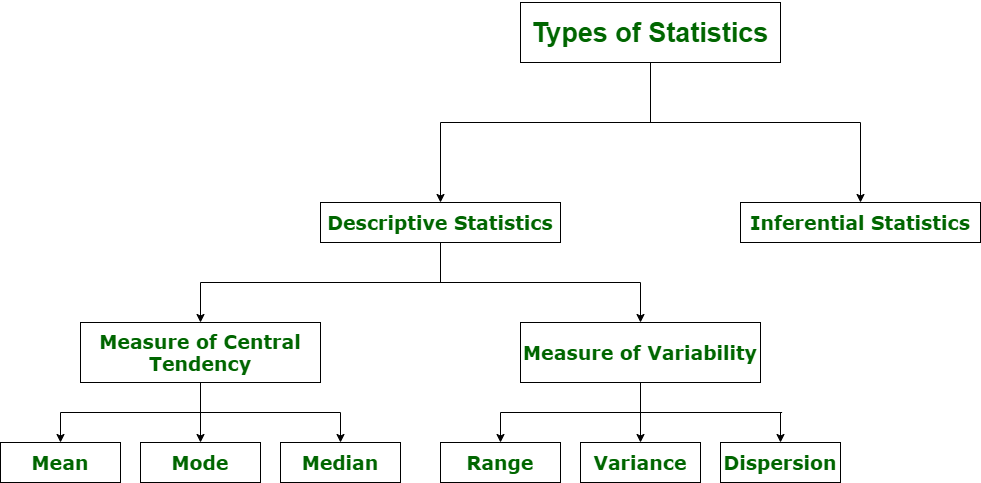
**Statistics** simply means numerical data, and is field of math that generally deals with collection of data, tabulation, and interpretation of numerical data. It is actually a form of mathematical analysis that uses different quantitative models to produce a set of experimental data or studies of real life. It is an area of applied mathematics concern with data collection analysis, interpretation, and presentation. Statistics deals with how data can be used to solve complex problems. Some people consider statistics to be a distinct mathematical science rather than a branch of mathematics.

Statistics makes work easy and simple and provides a clear and clean picture of work you do on a regular basis.

**Basic terminology of Statistics**

* **Population –**  
  It is actually a collection of set of individuals or objects or events whose properties are to be analyzed.
* **Sample –**  
  It is the subset of a population.

**Types of Statistics**



**1. Descriptive Statistics**

Descriptive statistics uses data that provides a description of the population either through numerical calculation or graph or table. It provides a graphical summary of data. It is simply used for summarizing objects, etc. There are two categories in this as following below.

**(a). Measure of central tendency**

Measure of central tendency is also known as summary statistics that is used to represents the centre point or a particular value of a data set or sample set.  
In statistics, there are three common measures of central tendency as shown below:

* + **(i) Mean**
  + It is measure of average of all value in a sample set.
  + **(ii) Median**
  + It is measure of central value of a sample set.
  + **(iii) Mode**
  + It is value most frequently arrived in sample set. The value repeated most of time in central set is actually mode.

**(b). Measure of Variability**

Measure of Variability is also known as measure of dispersion and used to describe variability in a sample or population. In statistics, there are three common measures of variability as shown below:

**(i) Range**

It is given measure of how to spread apart values in sample set or data set.

Range = Maximum value - Minimum value

**(ii) Variance**

It simply describes how much a random variable defers from expected value and it is also computed as square of deviation.

**S2=** ∑ni=1 [(xi - ͞x)2 ÷ n]

In this formula, **n** represent total data points, **͞x** represent mean of data points and **xi** represent individual data points.

**(iii) Dispersion**

It is measure of dispersion of set of data from its mean.

**σ=** √ (1÷n) ∑ni=1 (xi - μ)2

1. **Inferential Statistics**

Inferential Statistics makes inference and prediction about population based on a sample of data taken from population. It generalizes a large dataset and applies probabilities to draw a conclusion. It is simply used for explaining meaning of descriptive stats. It is simply used to analyse, interpret result, and draw conclusion. Inferential Statistics is mainly related to and associated with hypothesis testing whose main target is to reject null hypothesis.

Hypothesis testing is a type of inferential procedure that takes help of sample data to evaluate and assess credibility of a hypothesis about a population. Inferential statistics are generally used to determine how strong relationship is within sample. But it is very difficult to obtain a population list and draw a random sample.

Inferential statistics can be done with help of various steps as given below:

1. Obtain and start with a theory.
2. Generate a research hypothesis.
3. Operationalize or use variables
4. Identify or find out population to which we can apply study material.
5. Generate or form a null hypothesis for these population.
6. Collect and gather a sample of children from population and simply run study.
7. Then, perform all tests of statistical to clarify if obtained characteristics of sample are sufficiently different from what would be expected under null hypothesis so that we can be able to find and reject null hypothesis.

**Types of inferential statistics**

Various types of inferential statistics are used widely nowadays and are very easy to interpret. These are given below:

* One sample test of difference/One sample hypothesis test
* Confidence Interval
* Contingency Tables and Chi-Square Statistic
* T-test or Anova
* Pearson Correlation
* Bi-variate Regression
* Multi-variate Regression