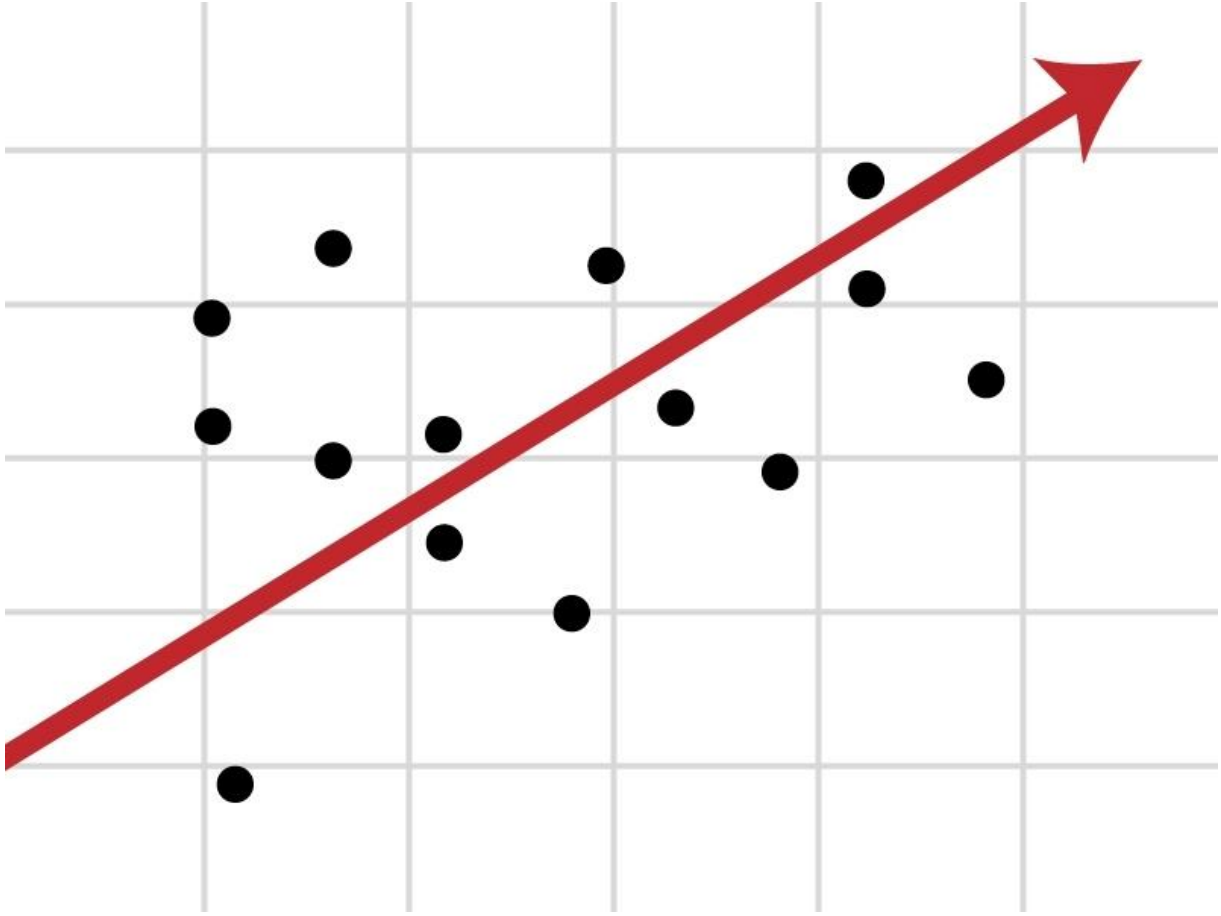
Regression Techniques



Regression

- Regression is a supervised learning technique that supports finding the correlation among variables. A regression problem is when the output variable is a real or continuous value.

- In Regression, we plot a graph between the variables which best fit the given data points. The machine learning model can deliver predictions regarding the data.

- In naïve words, ***"Regression shows a line or curve that passes through all the data points on a target-predictor graph in such a way that the vertical distance between the data points and the regression line is minimum."*** It is used principally for prediction, forecasting, time series modeling, and determining the causal-effect relationship between variables.

- **"Regression"** is a generic term for statistical methods that attempt to fit a model to data, in order to quantify the relationship between the dependent (outcome) variable and the predictor (independent) variable(s).

- **Dependent Variable:** The variable that we are trying to explain or predict is called the dependent variable.

- **Independent Variables:** The variable that is used to explain or predict the response variable is called the independent variable.

Regression Analysis

- Regression analysis is a form of predictive modelling technique which investigates the relationship between a **dependent** (target) and **independent variable (s)** (predictor).

- The regression analysis approach allows us to accurately establish which elements are most important, which factors may be ignored, and how these factors interact.

- Regression analysis is an important tool for modelling and analyzing data. Here, we fit a curve / line to the data points, in such a manner that the differences between the distances of data points from the curve or line is minimized.

- This technique is used for forecasting, time series modelling and finding the causal effect relationship between the variables. For example, relationship between rash driving and number of road accidents by a driver is best studied through regression.

Why Do We Use Regression Analysis?

- Regression analysis indicates the **significant relationships** between dependent variable and independent variable.

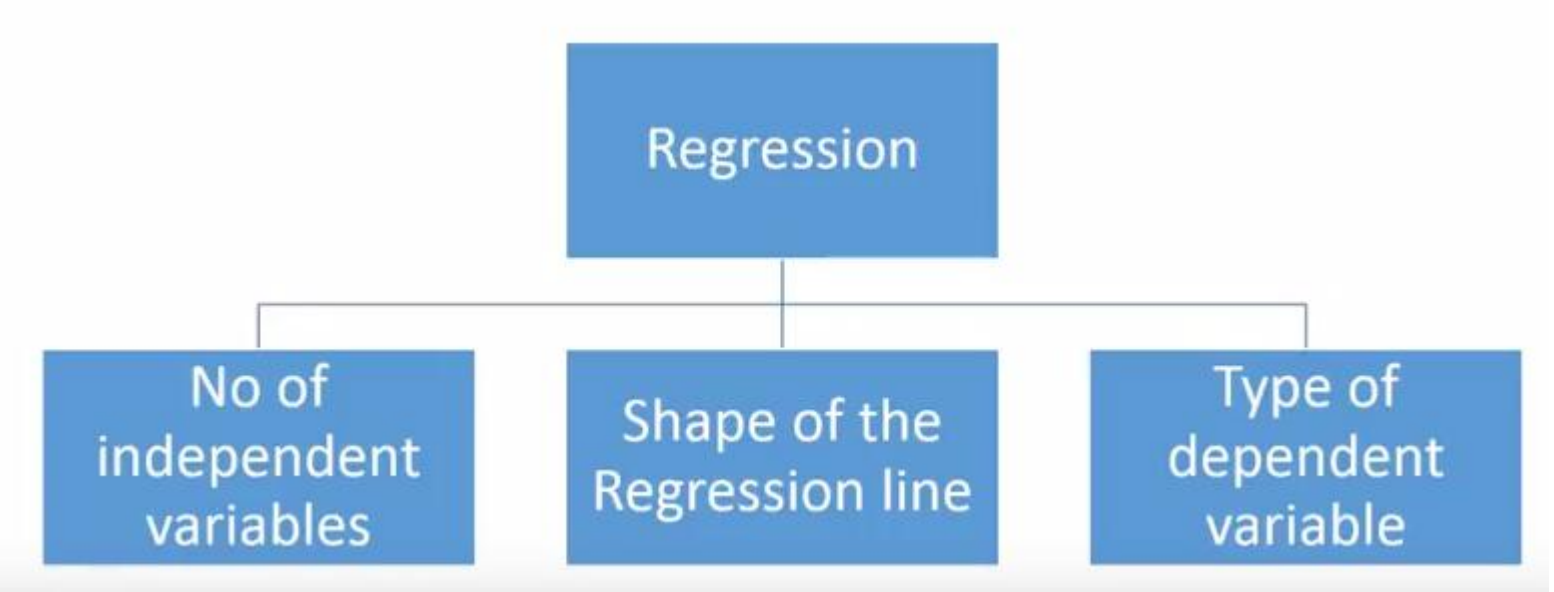
- Regression analysis indicates the **strength of impact** of multiple independent variables on a dependent variable.

- Regression analysis also allows us to compare the effects of variables measured on different scales, such as the effect of price changes and the number of promotional activities.

- These benefits help market researchers / data analysts / data scientists to eliminate and evaluate the best set of variables to be used for building predictive models.

Types of Regression Techniques

There are various kinds of regression techniques available to make predictions. These techniques are mostly driven by three metrics (number of independent variables, type of dependent variables and shape of regression line).



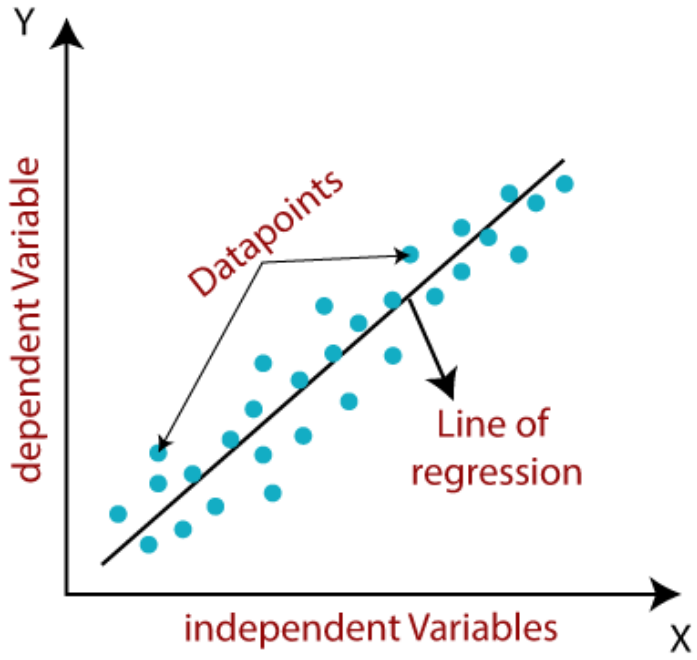
Linear Regression

- **Linear Regression** is a **supervised learning Technique**. It performs a **regression task**.

- Linear regression is a technique that statisticians use to describe the relationship between **dependent variable** and **independent variables**.

- **Simple linear regression.** Regression between **one dependent** variable and a ***single*** **independent** variable.

- **Multiple regression.** Regression between **one dependent variable** and ***two or more*** **independent** variables.

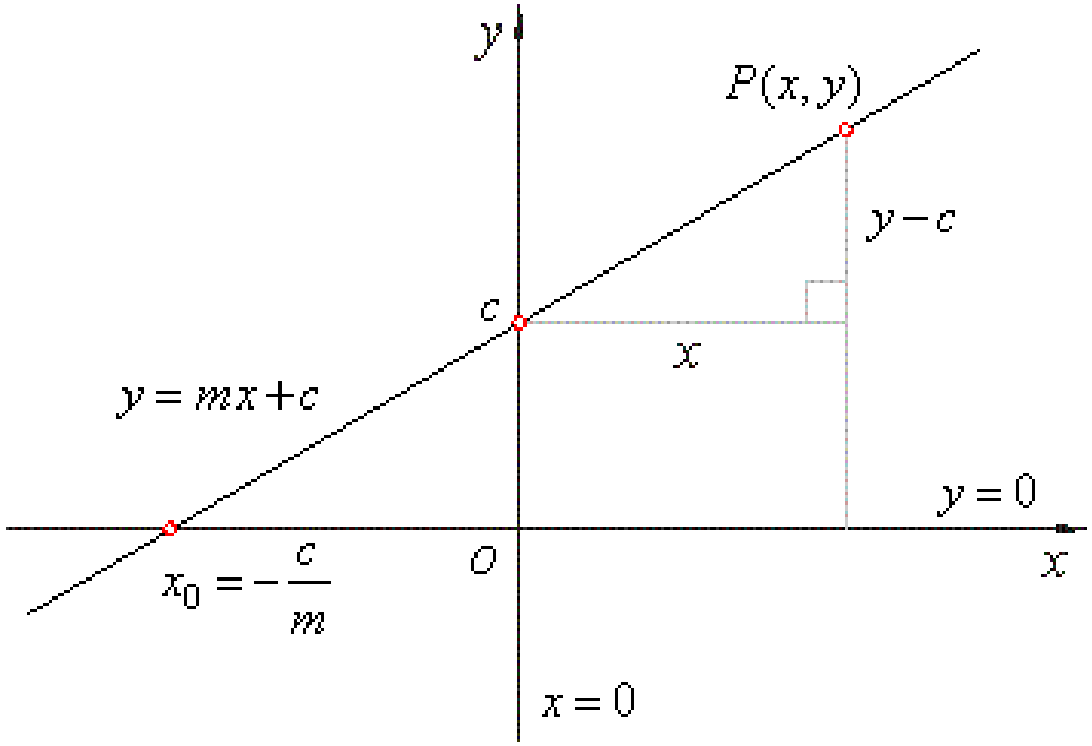


Linear Regression

linear regression is a linear approach to modelling the relationship between a dependent variable and one or more independent variables. Let **X** be the independent variable and **Y** be the dependent variable. We will define a linear relationship between these two variables as follows:

$$Y = mX + c$$

In **Y = mX + c, m** is the slope of the line and **c** is the y intercept. Today we will use this equation to train our model with a given dataset and predict the value of **Y** for any given value of **X**. Our challenge is to determine the value of **m** and **c**, such that the line corresponding to those values is the best fitting line or gives the minimum error.



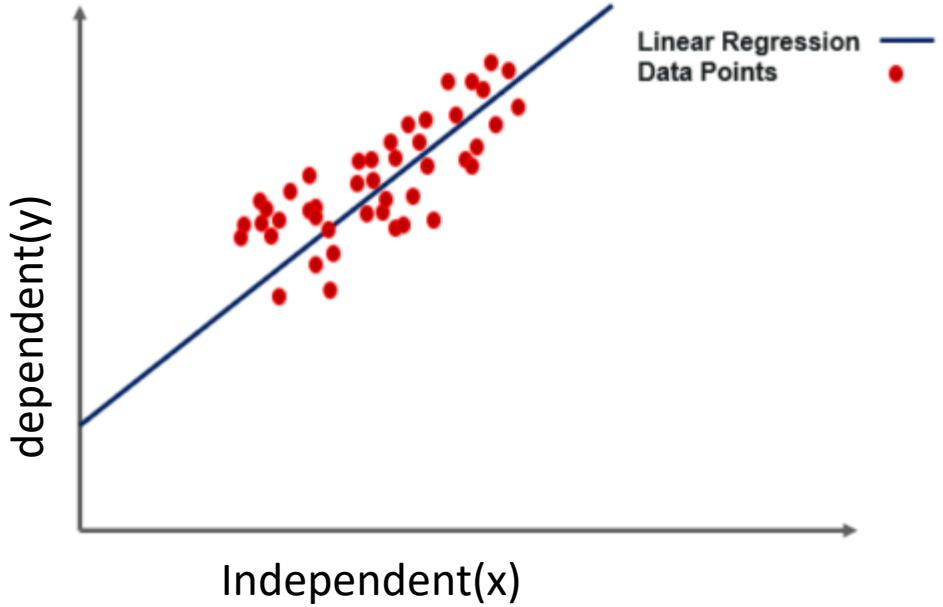
Simple Linear Regression

If you know something about X and this knowledge helps you to predict about y

In this equation:

y=mX+c

Y – Dependent Variable
m – Slope
X – Independent variable
c – Intercept



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**(independent variable), such linear regression is called ***simple linear regression***.

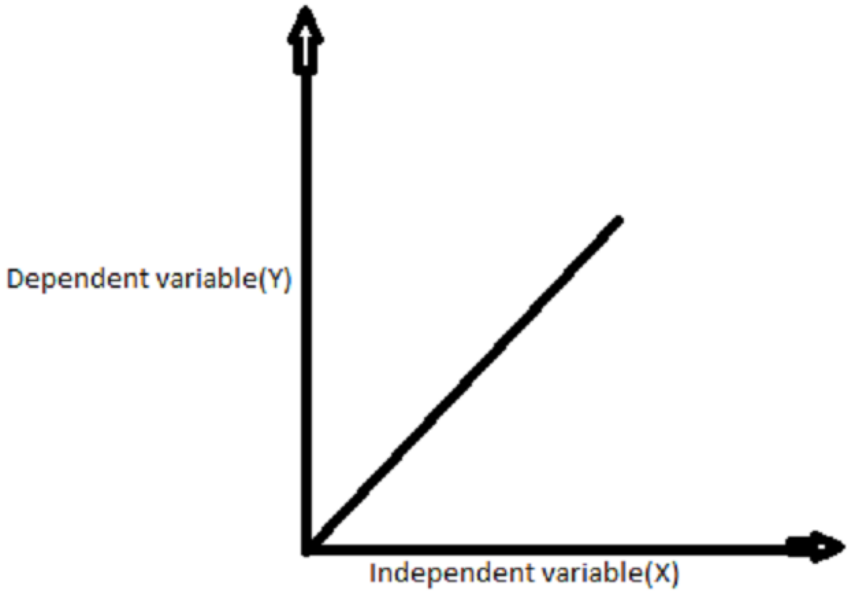
These coefficients m and c are derived based on **minimizing the sum of squared difference of distance between data points and regression line.**

Look at the example. Here we have identified the best fit line having linear equation **y=0.2811x+13.9**. Now using this equation, we can find the weight, knowing the height of a person.

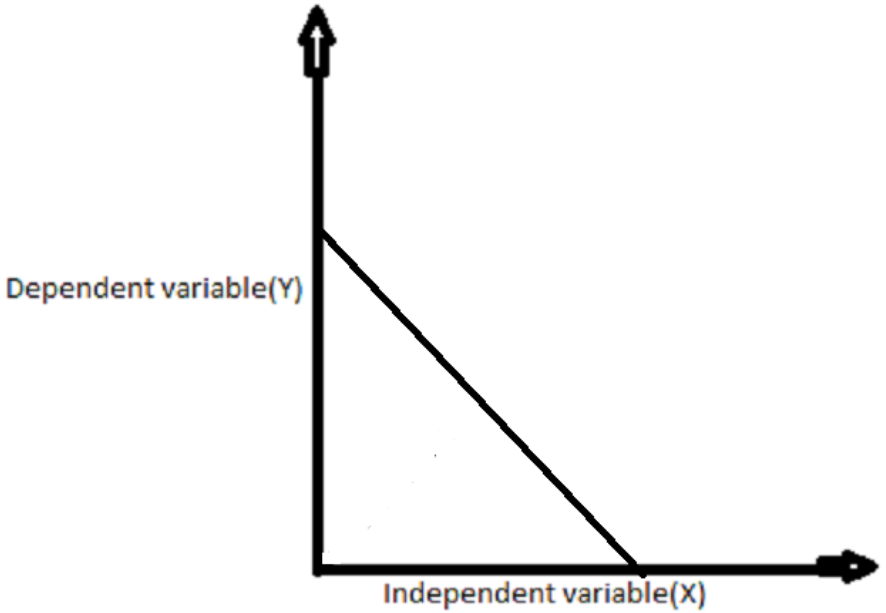
Slope of 2 means every change of 2 units in X will yield 2 units change in Y

Types of Simple Linear Regression

Positive Linear Regression– If the value of the dependent variable increases with the increase of the independent variable, then the slope of the graph is positive; such Regression is said to be Positive Linear Regression.



Negative Linear Regression-
If the value of the dependent variable decreases with the increase in the value of the independent variable, then such Regression is said to be negative linear Regression.



Simple Linear Regression Example

	Student	x_i	y_i	$(x_i - x)$	$(y_i - y)$	$(x_i - x)^2$	$(y_i - y)^2$	$(x_i - x)(y_i - y)$
	1	95	85	17	8	289	64	136
	2	85	95	7	18	49	324	126
	3	80	70	2	-7	4	49	-14
	4	70	65	-8	-12	64	144	96
	5	60	70	-18	-7	324	49	126
Sum		390	385			730	630	470
Mean		78	77					

The regression equation is a linear equation of the form: $\hat{y} = b_0 + b_1 x$ To conduct a regression analysis, we need to solve for b_0 and b_1. Computations are shown below.

$b_1 = \Sigma [(x_i - \bar{x})(y_i - \bar{y})] / \Sigma [(x_i - \bar{x})^2]$ $b_1 = 470/730 = 0.644$	$b_0 = \bar{y} - b_1 * \bar{x}$ $b_0 = 77 - (0.644)(78) = 26.768$

Therefore, the regression equation is: $\hat{y} = 26.768 + 0.644x$.

Simple Linear Regression Analysis

Best Fit line?

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 helps to work out the optimal values for B_0 and B_1, 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 $\mathbf{y_{predicted}}$ and $\mathbf{y_i}$.

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

$$MSE = \frac{1}{N}\sum_{i=1}^{n}(y_i - (B1x_i + B0))^2$$

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,
Coefficient of Determination or R-Squared (R2)
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,
$R^2 = 1 - (RSS/TSS)$

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.

$$RSS = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$$

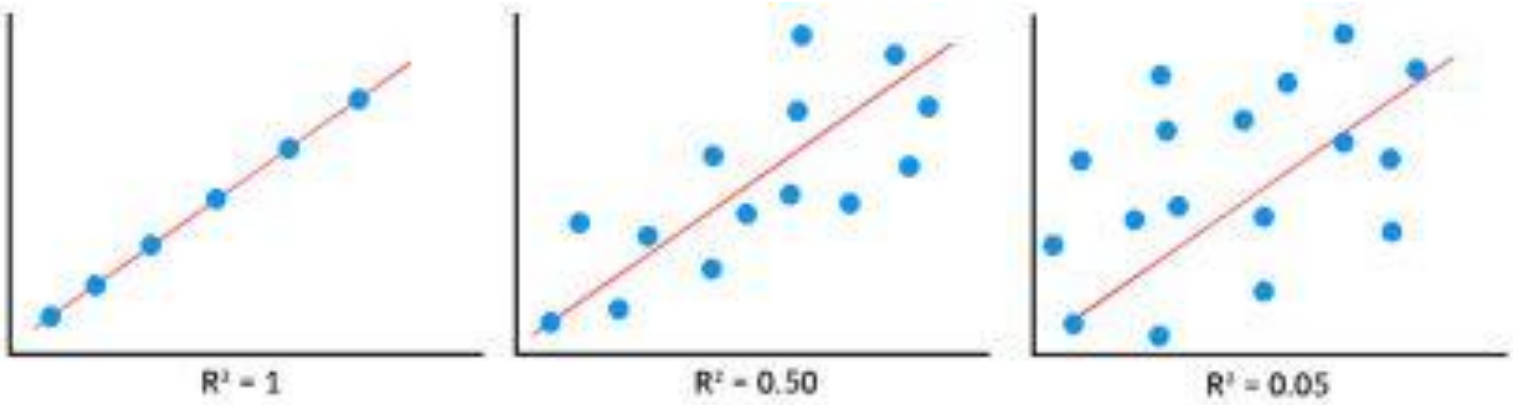
Evaluation Metrics for Linear Regression

•**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,

$$TSS = \sum (y_i - \bar{y}_i)^2$$

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

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



Evaluation Metrics for Linear Regression

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,

$$RMSE = \sqrt{\frac{RSS}{n}} = \sqrt{\sum_{i=1}^{n} \left(y_i^{Actual} - y_i^{Predicted}\right)^2 / n}$$

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,

$$RSE = \sqrt{\frac{RSS}{df}} = \sqrt{\sum_{i=1}^{n} \left(y_i^{Actual} - y_i^{Predicted}\right)^2 / (n-2)}$$

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.

Multiple Linear Regression

- Multiple or multivariate linear regression is a case of linear regression with two or more independent variables.
- There is a linear relationship between the dependent variables and the independent variables.
- The independent variables are not too highly correlated with each other.
- Observations are selected independently and randomly from the population.
- Residuals should be normally distributed with a mean of 0 and variance σ.

Multiple Linear Regression

- If there are just two independent variables, the estimated regression function is $f(x_1, x_2) = b_0 + b_1x_1 + b_2x_2$. It represents a regression plane in a **three-dimensional space**.

- The goal of regression is to determine the values of the weights b_0, b_1, and b_2 such that this plane is as close as possible to the actual responses and yield the minimal SSR.

- The case of more than two independent variables is similar, but more general. The estimated regression function is $f(x_1, \ldots, x_r) = b_0 + b_1x_1 + \cdots + b_rx_r$, and there are $r + 1$ weights to be determined when the number of inputs is r.

Polynomial Regression

- **Polynomial Regression** is a form of linear regression in which the relationship between the independent variable x and dependent variable y is modeled as an *nth* degree polynomial.

- Polynomial regression fits a nonlinear relationship between the value of x and the corresponding conditional mean of y, denoted E(y |x).

- In addition to linear terms like b_1x_1, your regression function f can include non-linear terms such as $b_2{x_1}^2$, $b_3{x_1}^3$, or even $b_4x_1x_2$, $b_5{x_1}^2x_2$, and so on.

- The simplest example of polynomial regression has a single independent variable, and the estimated regression function is a polynomial of degree 2: $f(x) = b_0 + b_1x + b_2x^2$.

- Now, remember that you want to calculate b_0, b_1, and b_2, which minimize SSR. These are your unknowns!

- In the case of two variables and the polynomial of degree 2, the regression function has this form: $f(x_1, x_2) = b_0 + b_1x_1 + b_2x_2 + b_3{x_1}^2 + b_4x_1x_2 + b_5{x_2}^2$.

Polynomial Regression

Uses of Polynomial Regression: These are basically used to define or describe non-linear phenomenon such as:

- Growth rate of tissues.
- Progression of disease epidemics
- Distribution of carbon isotopes in lake sediments

Advantages of using Polynomial Regression:

- Broad range of function can be fit under it.
- Polynomial basically fits wide range of curvature.
- Polynomial provides the best approximation of the relationship between dependent and independent variable.

Disadvantages of using Polynomial Regression

- These are too sensitive to the outliers.
- The presence of one or two outliers in the data can seriously affect the results of a nonlinear analysis.
- In addition there are unfortunately fewer model validation tools for the detection of outliers in nonlinear regression than there are for linear regression.

Underfitting

- **Underfitting** occurs when a model can't accurately capture the dependencies among data, usually as a consequence of its own simplicity. It often yields a low R^2 with known data and bad generalization capabilities when applied with new data.

- Techniques to reduce under fitting :
  1. Increase model complexity
  2. Increase number of features, performing feature engineering
  3. Remove noise from the data.
  4. Increase the number of epochs or increase the duration of training to get better results.

Overfitting

- **Overfitting** happens when a model learns both dependencies among data and **random fluctuations**. In other words, a model learns the existing data too well.

- Complex models, which have many features or terms, are often prone to overfitting. When applied to known data, such models usually yield high R^2. However, they often don't generalize well and have significantly lower R^2 when used with new data.

- The model then learns not only the relationships among data but also the noise in the dataset.

- Overfitted models tend to have good performance with the data used to fit them (the training data), but they behave poorly with unseen data (or test data, which is data not used to fit the model).

Overfitting

- **Overfitting usually occurs with complex models**.

- **Regularization** is the process of adding information in order to solve an ill-posed problem or to prevent over fitting.

- **Regularization** normally tries to reduce or penalize the complexity of the model. Regularization techniques mostly tend to penalize large coefficients b_0, b_1, …, b_r:

- Two regularization techniques:

1. L1 regularization
2. L2 regularization

Bias & Variance

- **Bias** –Bias occurs when an algorithm has limited flexibility to learn the true signal from the dataset

- Bias is an error from erroneous assumptions in the learning algorithm. **High bias** can cause an algorithm to miss the relevant relations between features and target outputs **(underfitting).**

- **Variance** – If you train your data on training data and obtain a very low error, upon changing the data and then training the same previous model you experience high error, this is variance.

Bias & Variance

Low Bias — High Variance:

- ***A low bias and high variance problem is overfitting****. Different data sets are depicting insights given their respective dataset. Hence, the models will predict differently. However, if average the results, we will have a pretty accurate prediction.*

High Bias — Low Variance:

- *The predictions will be similar to one another but on average, they are inaccurate*

Regularization

- A regression model that uses L1 regularization technique is called ***Lasso Regression*** and model which uses L2 is called ***Ridge Regression***
- *The key difference between these two is the penalty term.*
- **Ridge regression** adds "***squared magnitude***" of coefficient as penalty term to the loss function. Here the *highlighted* part represents L2 regularization element.

$$\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

Here, if lambda is zero then you can imagine we get back OLS. However, if lambda is very large then it will add too much weight and it will lead to under-fitting. Having said that it's important how lambda is chosen. This technique works very well to avoid over-fitting issue.

Regularization

- **Lasso Regression** (**Least Absolute Shrinkage and Selection Operator**) adds "***absolute value of magnitude***" of coefficient as penalty term to the loss function.

Again, if *lambda* is zero then we will get back OLS whereas very large value will make coefficients zero hence it will under-fit.

$$\sum_{i=1}^{n}(Y_i - \sum_{j=1}^{p} X_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

The **key difference** between these techniques is that Lasso shrinks the less important feature's coefficient to zero thus, removing some feature altogether. So, this works well for **feature selection** in case we have a huge number of features.

THANK YOU