Fundamentals of Artificial Intelligence

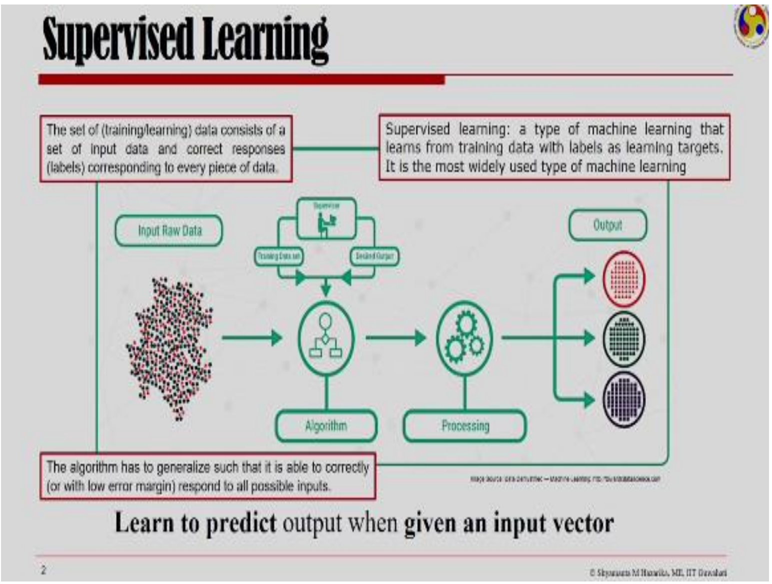
Prof. Shyamanta M. Hazarika
Department of Mechanical Engineering
Indian Institute of Technology – Guwahati

Lecture - 29
Linear Regression

Welcome to Fundamentals of Artificial Intelligence. We continue our discussion on machine learning with focus on supervised learning. In the last lecture, we had seen learning from observations. In particular, we saw learning a decision tree which is about getting to a decision, given a number of variables. We particularly saw how for a given problem of deciding whether you wait in a restaurant or not, we could derive the definition for the predicate we will wait for the problem from 12 examples of the different variables involved. Today our focus is on prediction. Prediction involving numerical data, say we want to predict what would be the salary of a recent graduate based on his cumulative grade point average. This we do within supervised learning using a process called Regression.

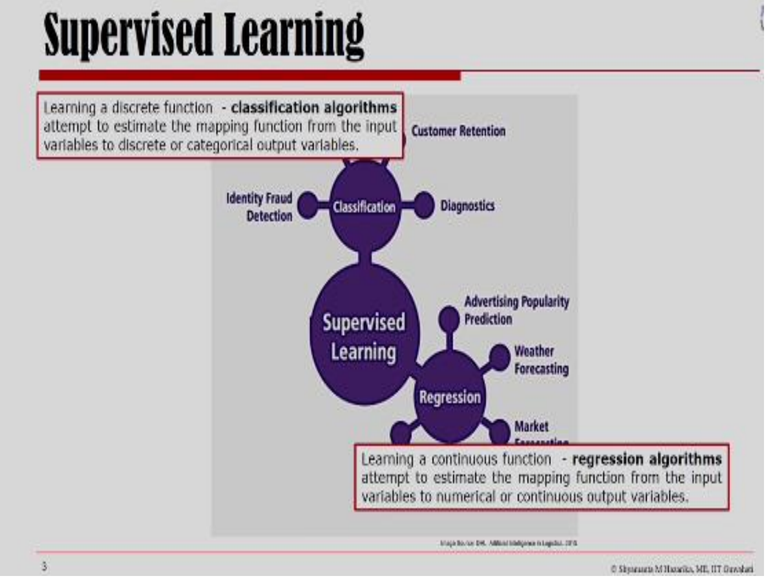
We would start our discussion by quickly reviewing what is supervised learning and look at regression vis-a-vis supervised learning.

(Refer Slide Time: 02:19)



So supervised learning as we have discussed in our previous lecture is about learning to predict an output when we are given an input vector and this is learning from training data with labels as learning targets. Supervised learning is about getting to the levels for the output based on a given set of labels for the inputs. Now the algorithm has to generalize such that it is able to correctly or with very low error margin respond to all possible inputs.

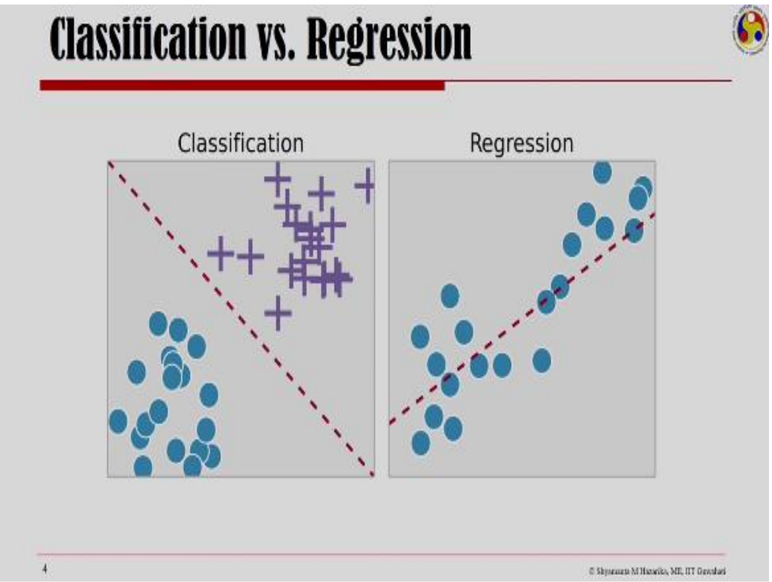
(Refer Slide Time: 03:05)



As we have discussed previously supervised learning can be categorized as classification and regression. Classification is about getting to the categories of the output variables or it is about learning a discrete function which is referred to as the classification algorithm to estimate the mapping function from the input variables to discrete or categorical output variables.

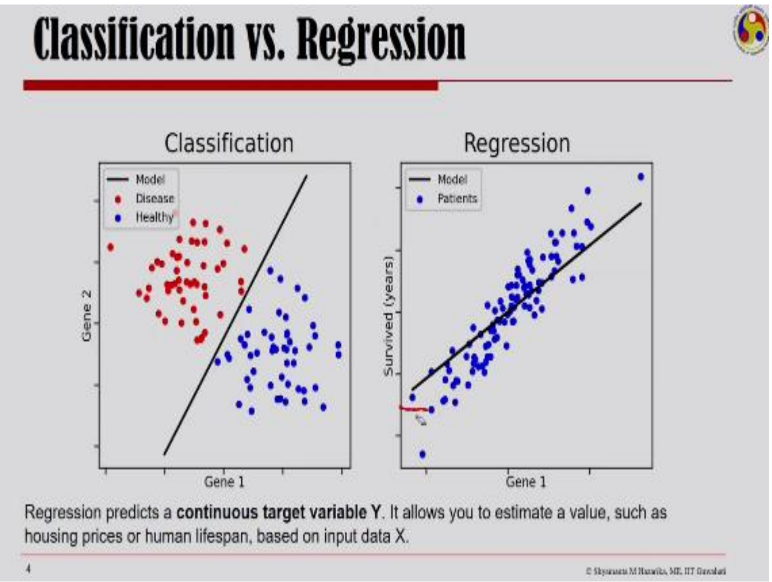
Whereas on the other hand regression is about learning a continuous function and is an attempt to estimate a mapping function from the input variables to numerical or continuous output variables and this is what is our focus today. We want to look at how given a variable x which is called the explanatory variable which is the independent variable I get to predict y the dependent variable and this for numerical or continuous variables.

(Refer Slide Time: 04:21)



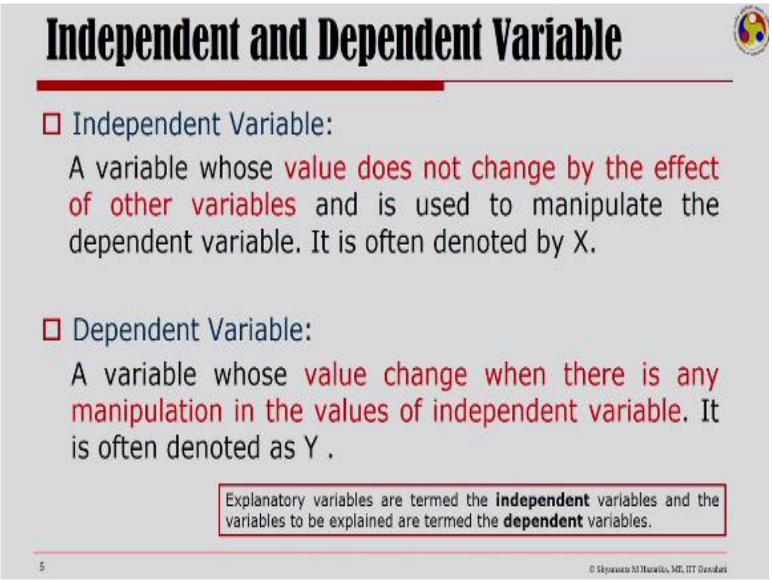
Now classification vis-a-vis regression let us take a minute to understand the difference, if I have two types of outputs here the circle and the cross as shown in your screen. A classification algorithm would try to draw a boundary between them whereas if I am talking of regression I have the same type of output which are circles, but given the variable X for that I would love to know what is the Y.

(Refer Slide Time: 05:05)



So let us say I have two genes Gene 1 and Gene 2 which are either healthy or disease genes and then when I am trying to understand a classification model my aim is to draw a boundary between the two types of genes which is either disease or healthy, whereas in regression I am trying to get a model where when I am told about a particular gene I would love to have certain characteristics of that gene reflected such as for this example shown here, I would love to know from a particular gene what would be the survival of a particular patient and then what I have now here is a mapping from a variable which is gene to a variable which is the survival in number of years. So regression predicts a continuous target variable Y and it allows one to estimate a value this could be as mundane as housing price or as interesting as human lifespan based on an input data X and this is what would be our focus today.

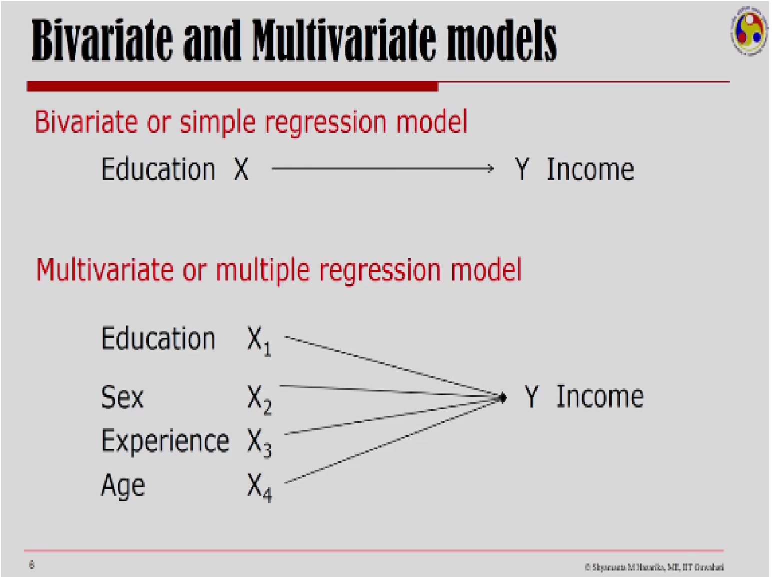
(Refer Slide Time: 06:42)



When we are talking of regression we are talking of two variables one the independent variable, an independent variable is one whose value does not change by the effect of other variables and is used to manipulate the dependent variable. And independent variable is often denoted by X and then in a regression problem we have a dependent variable a variable whose value change when there is any manipulation in the values of the independent variable.

A dependent variable is often denoted as Y. Now the explanatory variables are termed the independent variable and the variables that are to be explained are termed as the dependent variable. In the gene and human lifespan example that I was showing you in the previous slide the gene is what explains, so the explanatory variable or the independent variable is the gene and the human lifespan is what is being explained and this is the dependent variable.

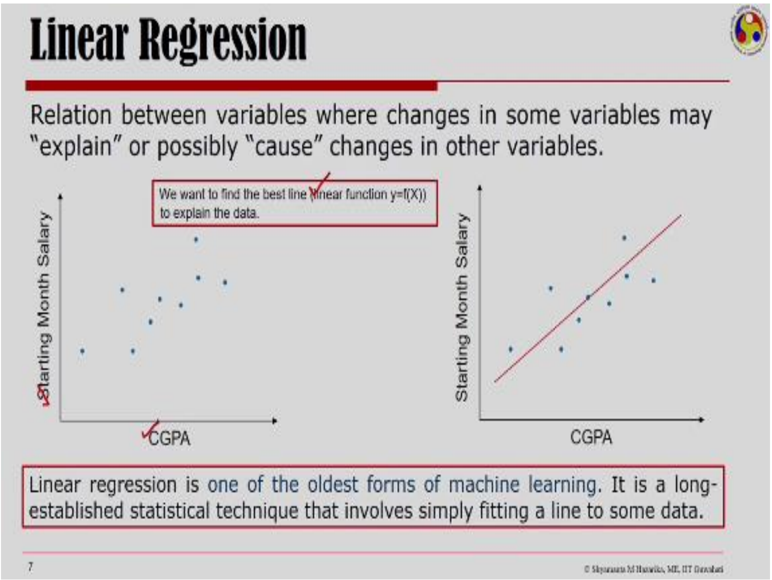
(Refer Slide Time: 08:16)



Now when we are talking of regression we could have what is referred to as Bivariate or simple regression and Multivariate regression. Bivariate or simple regression model is about one explanatory variable mapping to a dependent variable. So here I have shown education X which is the independent variable able to somehow give a prediction of income Y that is a Bivariate or simple regression.

In Multivariate or multiple regression I am looking for more than one explanatory variable for example income could be explained by education, sex, experience, age. So here we have a number of variables X1, X2, X3, X4 which all try to explain Y the dependent variable such a regression is called Multivariate or multiple regressions.

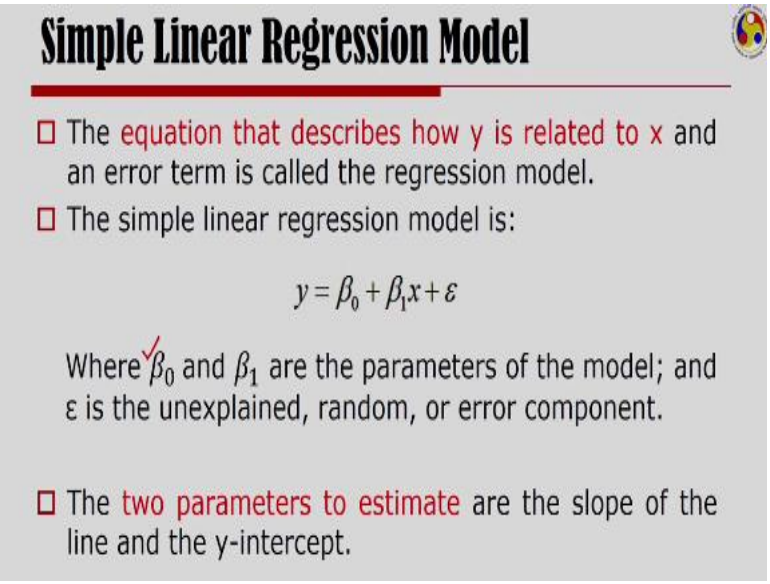
(Refer Slide Time: 09:30)



Now we will today look at a very simple model of regression referred to as Linear Regression. The relationship between variables where changes in some variable may explain or possibly cause changes in other variables could be explained by a linear function. Now coming back to this example of what could be the starting monthly salary of a graduate based on his cumulative grade point average.

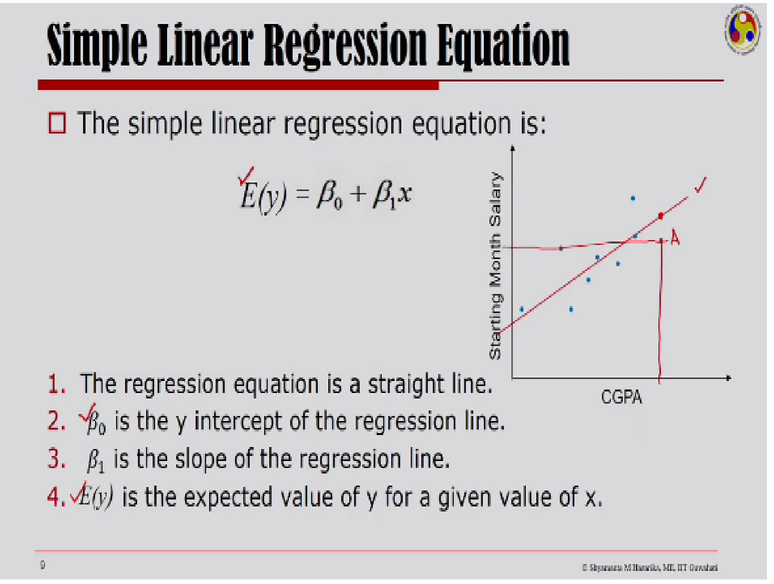
We could have data as shown on your screen here where we have the cumulative grade point average as our independent variable X and the starting monthly salary as our dependent variable Y. Now when I am trying to get a regression model one that is a linear model I want to find the best line the linear function y = f of X to explain the data and that line is the regression model. Now the Linear Regression is one of the oldest forms of machine learning. It is a long-established statistical technique that involves simply fitting a line to some data but let us look at, how do you get to the best fit.

(Refer Slide Time: 11:14)



Now the equation that describes how y is related to x and an error term is called the regression model and here it is called Linear Regression model because the equation is linear. The simple linear regression model is y = beta 0 + beta 1 x + epsilon where beta 0 and beta1 are the parameters of the model and epsilon is the unexplained random or error component. The two parameters beta 0 and beta1 are to be estimated and actually if you look at this equation more closely these are the slope and the y-intercept of the line.

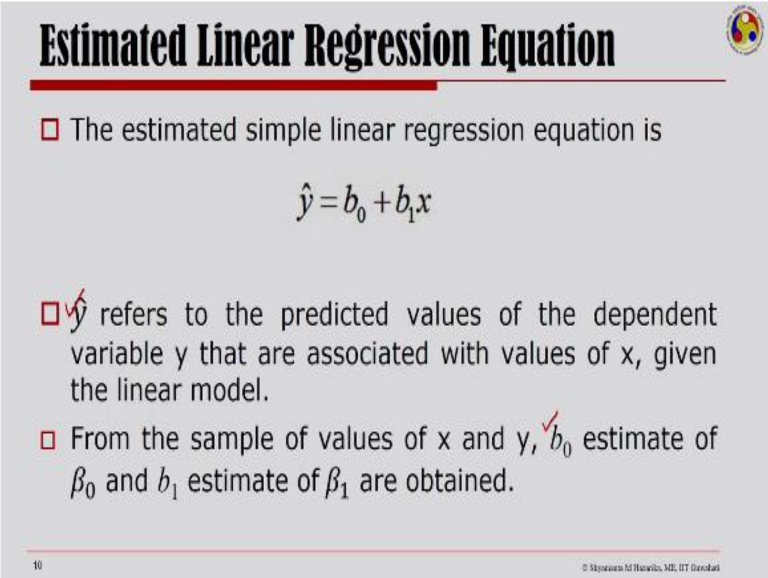
(Refer Slide Time: 12:16)



So here is the data that is about the CGPA versus the starting monthly salary and the simple Linear Regression equation is about the expected value of y given as beta 0 + beta 1 x where the regression equation is a straight line and beta 0 is the intercept of the regression line. This value is beta 0 and beta 1 is the slope of the regression line. E is the expected value of y for a given value of x.

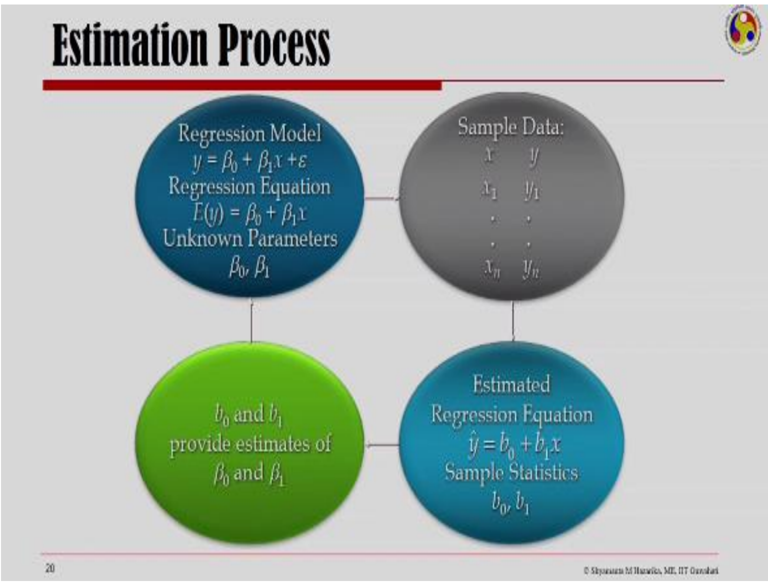
Now we would make this clear in a short while, but even then we should realize that when I am fitting this line here the value of y for a given value of x is different from the real value. For example, as shown this point A that I have here this is the x value and correspondingly I will have a value somewhere here which would be the real y value whereas when I am using a simple linear regression equation the y value for the particular x is somewhere here and that y is the expected value of y for the given x.

(Refer Slide Time: 14:09)



Now the estimated simple linear regression equation is given as y hat = b0 + b1 x. Y hat refers to the predicted value of the dependent variable y that are associated with values of x given the linear model and from the sample of values of x y that we started with b0 gets to an estimate of beta 0 and b1 estimate beta 1.

(Refer Slide Time: 14:50)

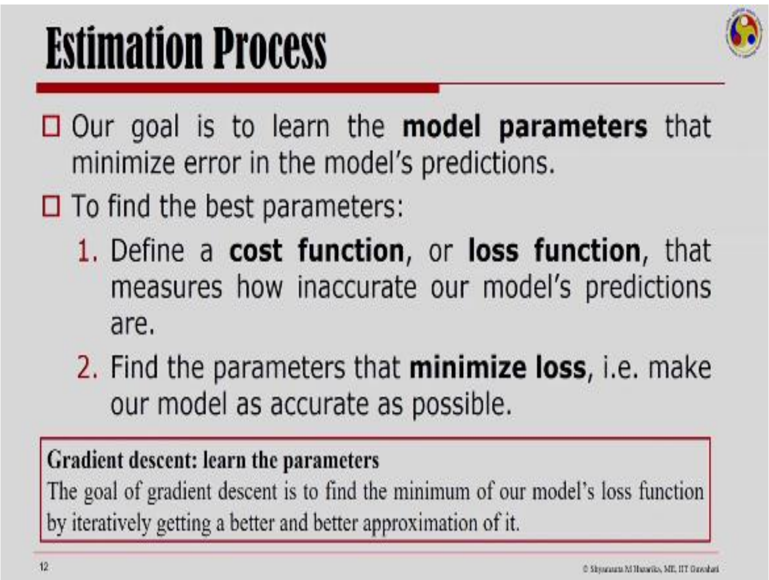


u

Now let us look at what we mean by the estimation process. So we have the regression model beta 0 beta 1 + epsilon and we have the regression equation beta 0 + beta 1 x and we have the unknown parameters which are beta 0 and beta1. Once we have the regression model and the sample data are given to us we would love to arrive at the estimated regression equation which is y hat = b0 + b1 x and the sample statistics of b 0 and b1.

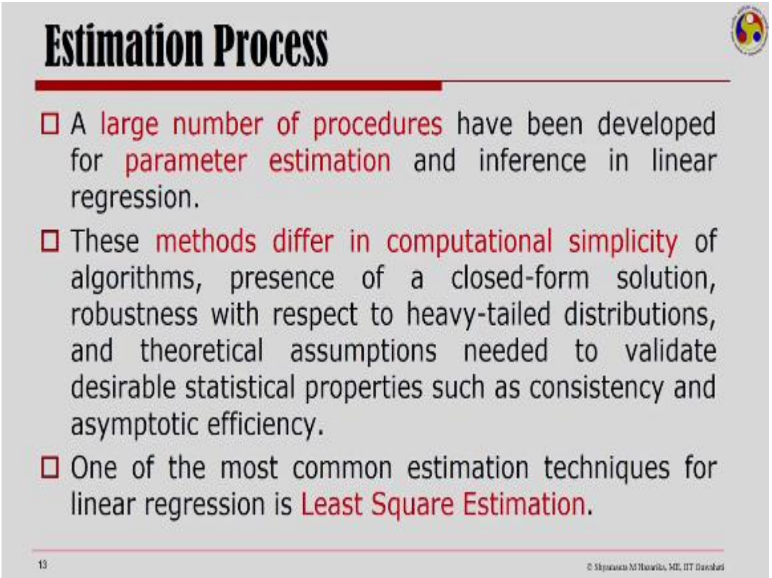
B0 and b1 actually provides estimate of beta 0 and beta 1. Our estimation process therefore runs like this. We have the regression model together with the regression equation, the sample data is known to us from that we estimate the regression equation and the sample statistics and the sample statistics b0 and b1 provide estimates of beta 0 and beta1.

(Refer Slide Time: 16:16)



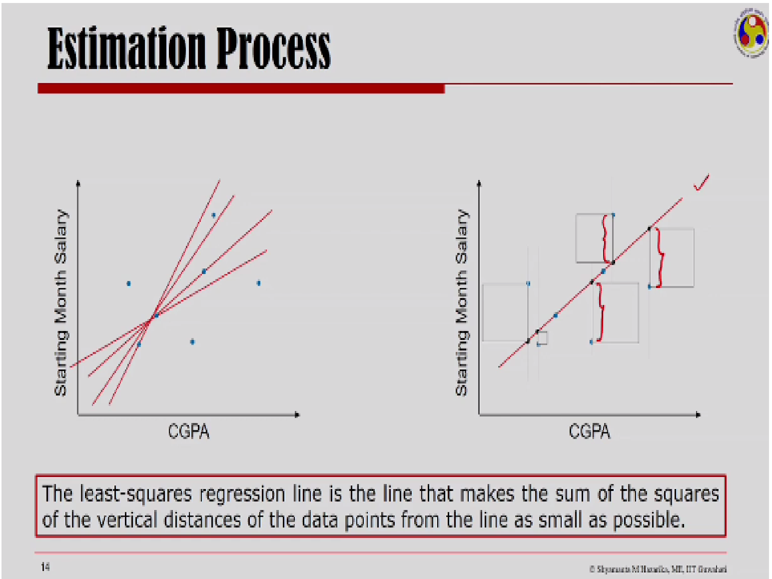
Our goal is to learn the model parameters that minimize error in the models prediction. To find the best parameters we define a cost function or loss function that measures how inaccurate our models predictions are and then the idea is to find the parameters that minimizes loss that is make our model as accurate as possible. Now there are different ways to do this. One of the most popular is about gradient descent to learn the parameters. The goal of gradient descent is to find the minimum of our models loss function by iteratively getting a better and better approximation of it.

(Refer Slide Time: 17:11)



A large number of procedures have been developed for parameter estimation and inference in linear regression. Now all of these methods differ in computational simplicity of algorithms, presence of closed-form solution, robustness with respect to heavy-tailed distribution and theoretical assumptions needed to validate certain desirable statistical properties. One of the most common estimation techniques for linear regression is the technique of Least Square Estimation which we will look at now.

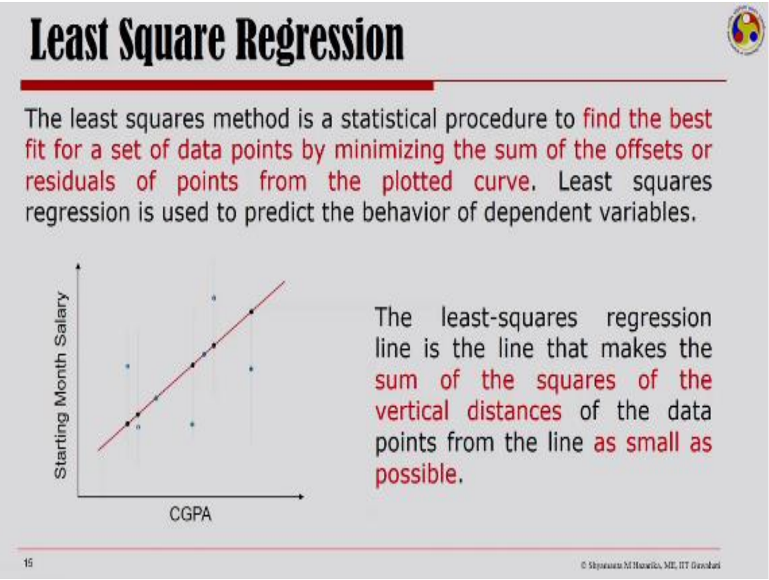
(Refer Slide Time: 17:53)



So the least square regression is about getting a line that would make the sum of the squares of the vertical distances of the data points from the lines as small as possible. Let us see what we mean by that. So given the data distribution as shown on your screen we could have a number of lines that could pass through them. Now we would try to get to the line where the vertical distances of data points from the line that I am trying to fit and the point themselves.

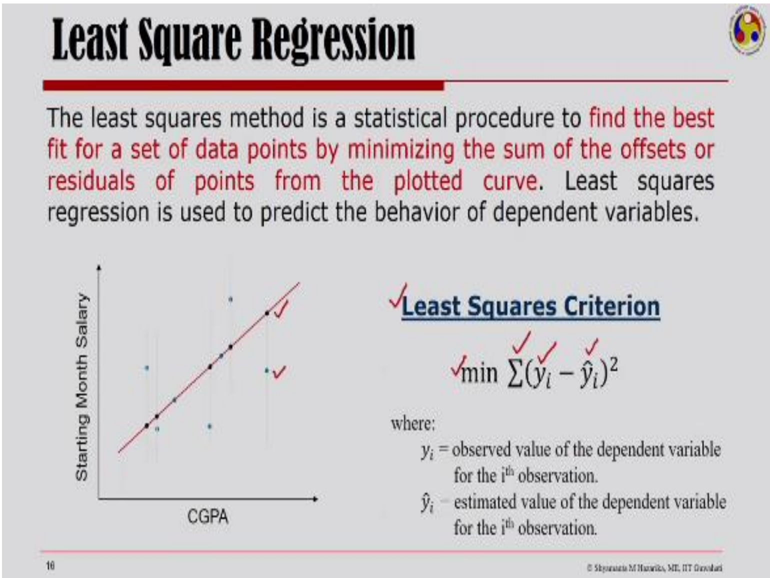
I try to figure them out first. So here is the vertical distance of the first point, the vertical distance of the second point, third point is on the line, fourth point here is the vertical distance so and so forth I find the vertical distances for each of these points. So I find these vertical distances of the points from the line that I am trying to fit in and in order to minimize error what I would do is I would take the squares of this. So I will take the squares of this and the idea is to get to the line that makes the sum of these squares the minimum.

(Refer Slide Time: 19:38)



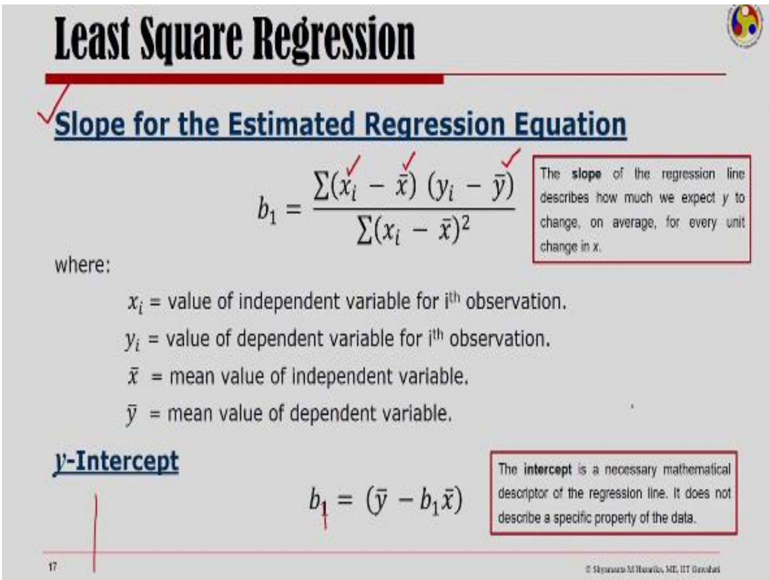
So the least square regression is a statistical procedure of finding the best fit for a set of data points by minimizing the sum of the offsets or residuals from the plotted curve and these vertical distances are what is the offset or residual and I want to minimize the squares of this. So the least square regression line would be the line that makes the sum of the squares of the vertical distances of the data points from the line as small as possible.

(Refer Slide Time: 20:25)



So the least square criteria is given as taking the observed value of the dependent variable for the ith observation and the estimated value of the dependent variable of the ith observation. As already explained the observed value is this whereas the estimated value is here. So you take the square of that and you sum it up for all the points and then what we want to do is get to the minimum of such a sum that is called the Least Squares Criterion.

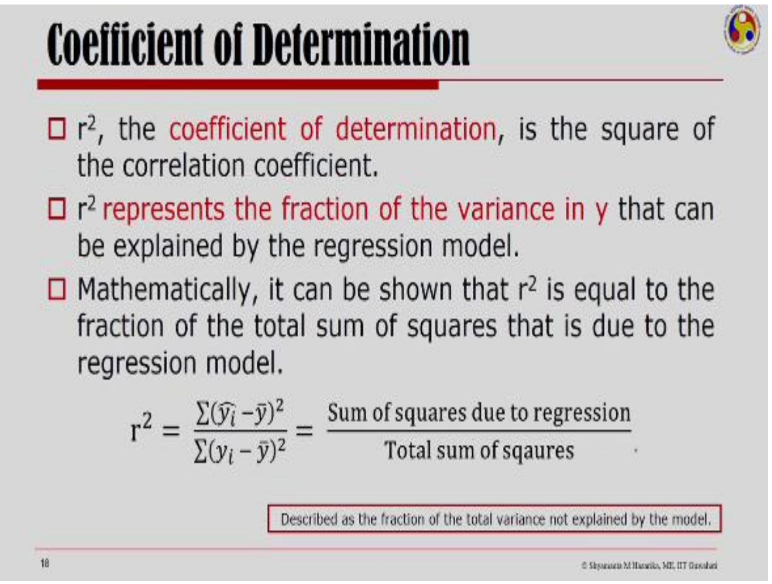
(Refer Slide Time: 21:15)



And from the least square criteria we finally arrive at the slope for the estimated regression equation which is b1 and it is based on summing up the value of the independent variables ith observation and the mean value of the independent variable and we have here the value of the dependent variable of the ith observation and its mean value divided by the difference between the independent and the mean square of that sum of squares of that.

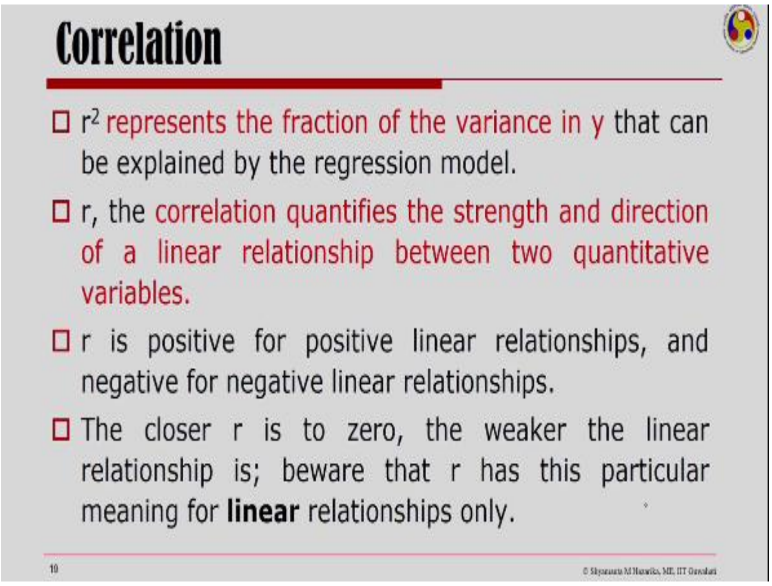
So slope for the estimated regression equation actually describes how much we expect y to change on average for every unit change in x and then we have the intercept which is the mean value of the dependent variable minus the slope into the mean variable of the independent variable. Now the intercept is a necessary mathematical descriptor of the regression line. It does not describe a specific property of the data.

(Refer Slide Time: 22:45)



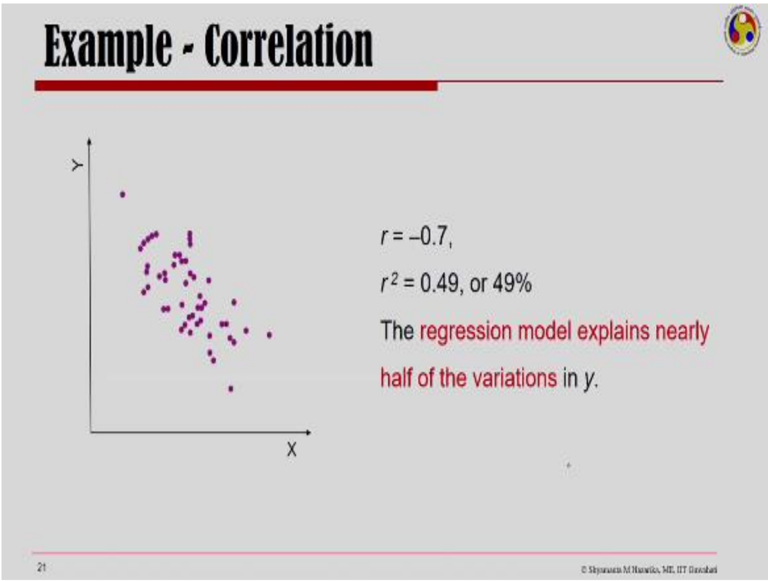
The coefficient of determination is the square of the correlation coefficient and it represents the fraction of the variance in y that can be explained by the regression model. Mathematically, it can be shown that r square the coefficient of determination is equal to the fraction of the total sum of squares that is due to the regression model. So r square is the ratio of the sum of squares due to regression by the total sum of squares and is described as the fraction of the total variance not explained by the model.

(Refer Slide Time: 23:32)



The coefficient of determination represents the fraction of variance in y that can be explained by the regression model. Correlation where as quantifies the strength and direction of a linear relationship between the two quantitative variables, r is positive for positive linear relationships and it is negative for negative linear relationships. The closure correlation is to zero the weaker the linear relationship is, but then r has this particular meaning only for linear relationships.

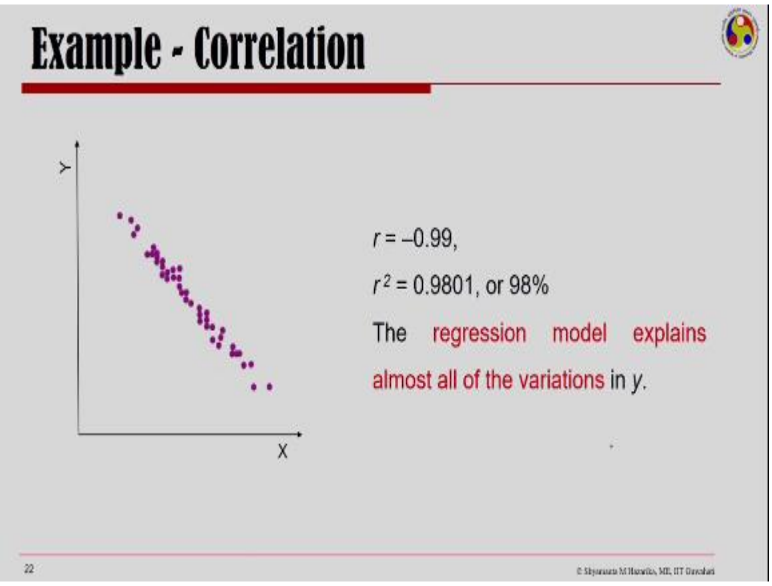
(Refer Slide Time: 24:16)



Now let us look at a couple of examples and try to understand the correlation vis-a-vis the coefficient of determination. So here on your screen is some data that has correlation of 0.3 and the coefficient of determination is 0.09. So what this means is that the regression model explains not even 10% of the variations in y. Here is another example where correlation is 0.7

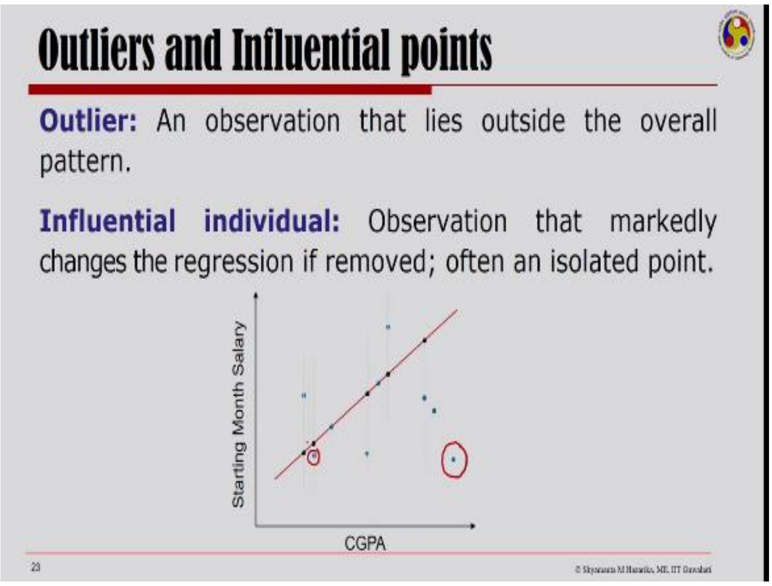
the coefficient of determination is 0.49 that is 49% what it means is the regression model could explain nearly half of the variations in y.

(Refer Slide Time: 25:14)



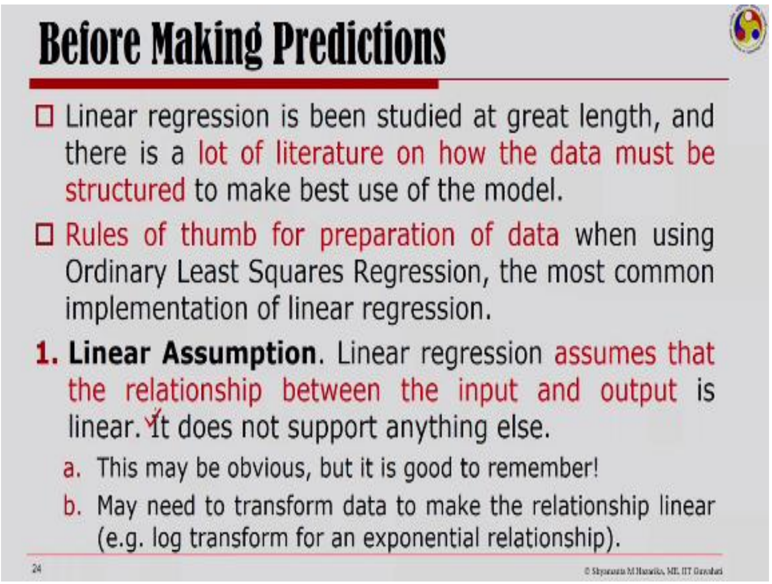
And as we move to more linear relationships between x and y we have correlation here of 0.99 and the coefficient of determination is almost 98%. The regression model explains almost all of the variations in y.

(Refer Slide Time: 25:38)



Other thing that needs to be taken care of when we are talking of linear regression is about outliers and influential points. An outlier is an observation that lies outside the overall pattern. So in this example here, this point that I have here is an outlier and on the other hand influential individual is an observation that could markedly change the regression if removed and often it is an isolated point. Like it could be a point here which is very near to my model isolated and this could be one that if I do not consider could markedly change the regression model.

(Refer Slide Time: 26:31)

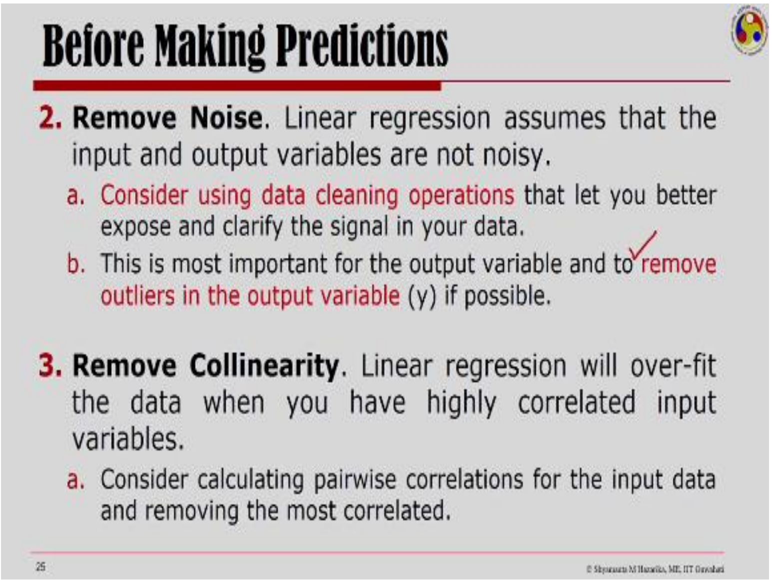


Now how do one make prediction using such regression models we would focus in the next couple of slides and the most important thing is about making certain preparations of the data before you use this data for making predictions. So linear regression has been studied at length and there is a lot of literature on how the data must be structured to make the best use of the regression model.

Rules of thumb for preparation of data when using this Ordinary Least Square Regression we have defined here is what we will highlight now. Now we should realize that the Ordinary Least Square Regression that we have highlighted here is one of the most common implementation of linear regression and the rules of thumb for preparation of data include one the very first being an assumption of linearity between x and y.

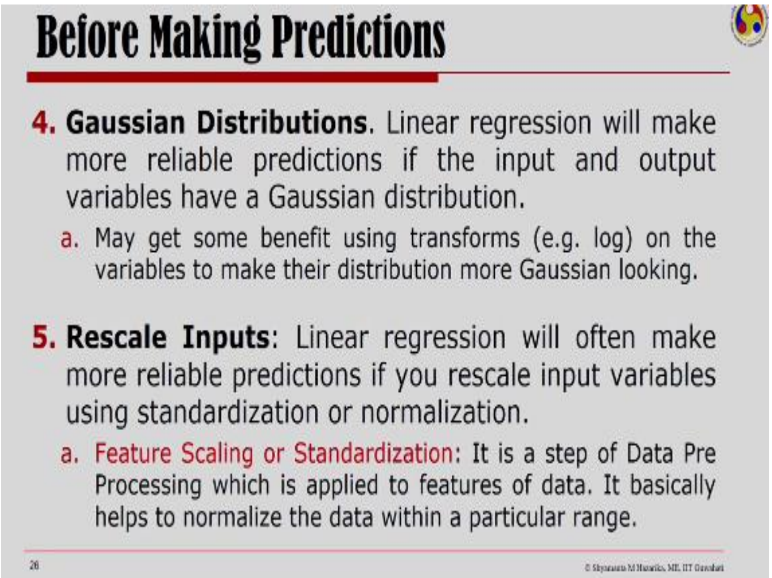
So linear regression assumes that the relationship between the input and the output is linear it does not support anything else. Now this may be obvious, but then it is good to remember. A point that needs to be taken note of here is that when I want to use linear regression for certain data I may need to transform the data to make the relationship linear. For example, I could take log transform for an exponential relationship and then use linear regression over it.

(Refer Slide Time: 28:22)



The second is about removing noise so linear regression assumes that the input and output variables are not noisy. So we could consider using certain data cleaning operations that expose and clarify the signal in our data. This is most important for output variable and about removing outliers in the output variable if possible. The third rule of thumb is about removing collinearity. So linear regression will over-fit the data when you have highly correlated input variables. So the idea is to consider calculating pairwise correlations for the input data and removing the most correlated before you apply linear regression on it.

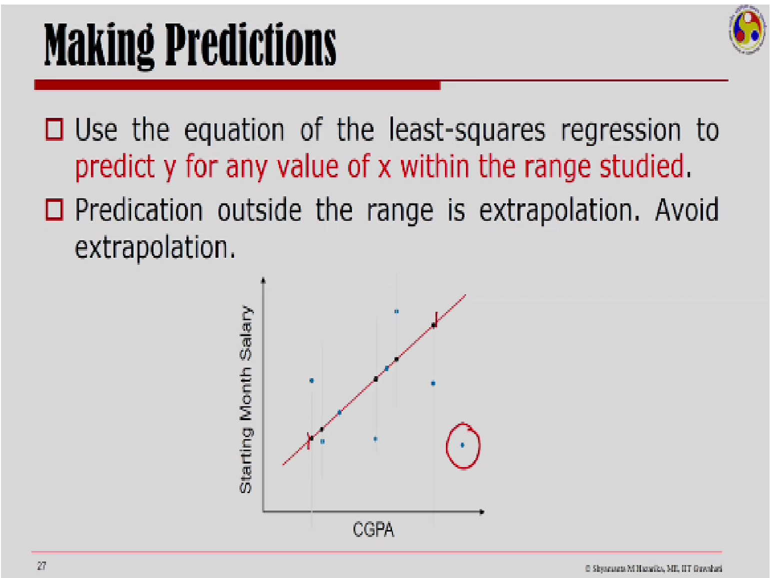
(Refer Slide Time: 29:19)



One most important point on linear regression is that linear regression will make reliable predictions if the input and output variables have a Gaussian distribution. Now in order to explain that one may get some benefit using transforms on the variables to make their distribution more Gaussian looking. Finally, linear regression will often make more reliable predictions if you rescale the input variables using standardization or normalization.

Now features, scaling or standardization is a step of data pre-processing which is applied to features of data across many algorithms in machine learning. It basically helps to normalize the data within a particular range and this is to be done for the data before we apply linear regression on it.

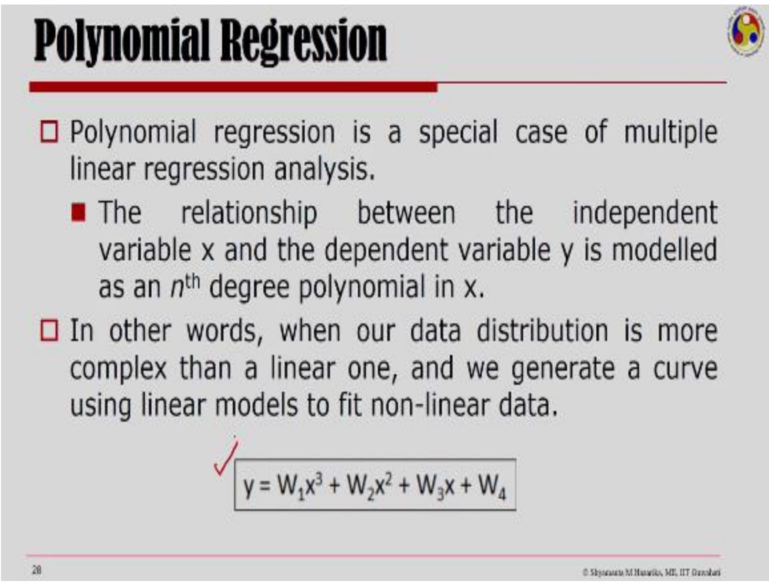
(Refer Slide Time: 30:25)



So once we have a data prepared according to this rule of thumbs and then we have had our equation of the least square regression model. Now we can use the equation of the least square regression to predict y for any value of x, but then this must be well within the range studied. So, one point to note is that even if I am talking of fitting a line here, I am not talking of extrapolation.

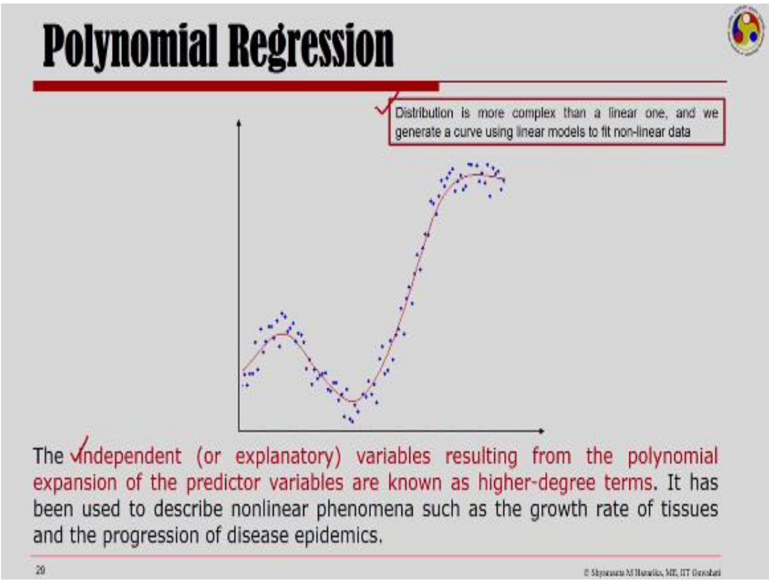
So predictions if I do outside the range that have been studied is extrapolation and we need to avoid extrapolation. So we use the equation of the least square regression to only predict values for this example possibly between here and here, because as we have already discussed we will be removing this as an outlier and therefore we will restrict our predictions only between these two range.

(Refer Slide Time: 31:29)



Now a quick note on a very special case of multiple linear regression analysis called polynomial regression. So, polynomial regression is a special case of multiple linear regression analysis. Here the relationship between input variable x and the dependent variable y is modeled as an nth degree polynomial in x, so, in other words when our data distribution is more complex than a linear one and we generate a curve using linear models to fit non linear data. So this is the polynomial regression model where we can see that we have used a cubic polynomial in x.

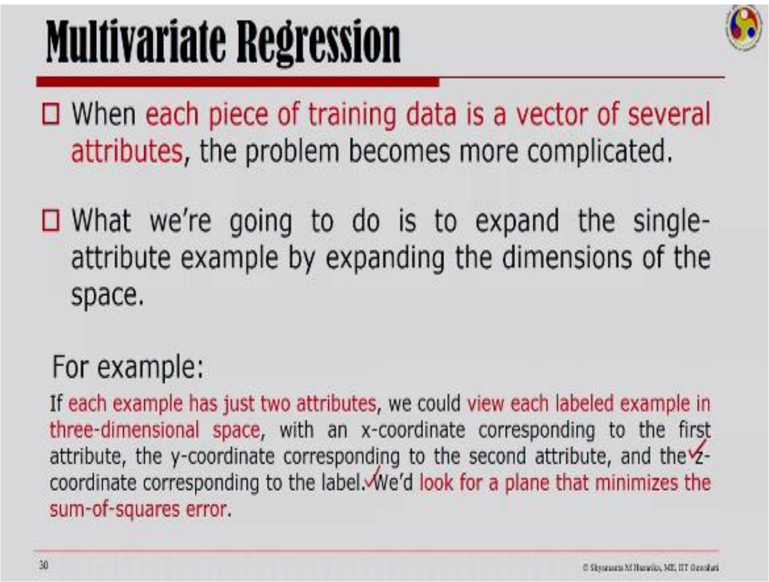
(Refer Slide Time: 32:27)



Now polynomial regression we will use if the distribution is more complex than a linear one and here the independent variable or the explanatory variable is resulting from the polynomial expansion of the predictor variable and these are known as higher degree terms.

Polynomial regression has been used to describe nonlinear phenomena such as growth rate of tissues and progression of disease epidemics.

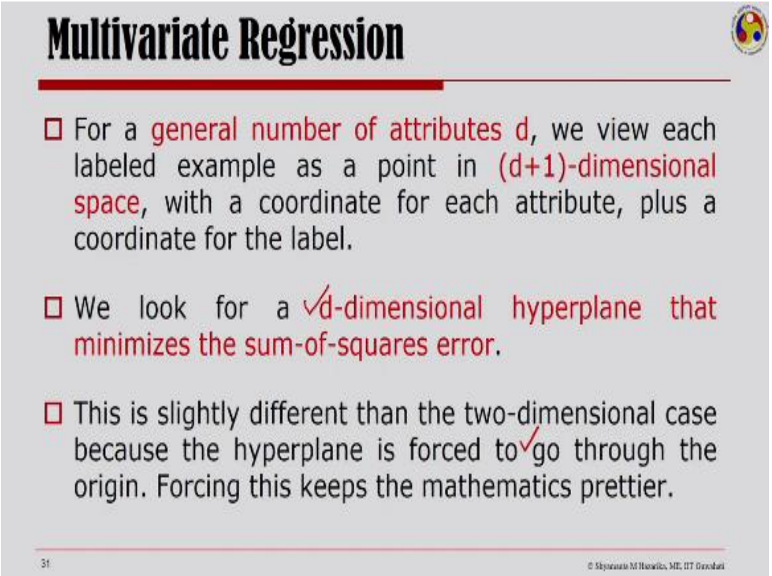
(Refer Slide Time: 33:04)



Now let us quickly focus on multivariate regression. So far we were talking of Bivariate or simple regression however when each piece of training data is a vector of several attributes, the problem becomes more complicated. What we are going to do is expand this single attribute example by expanding the dimensions of the space. For example, if each example now has two attributes, we would then view each label example in a 3 dimensional space with an x coordinate corresponding to the first attribute, the y coordinate corresponding to the second attribute and then the z coordinate would be the corresponding label. Now recall that when we were doing Bivariate or simple regression we were talking of the two dimensional plane where we had the independent variable x and the dependent variable y.

And then we were looking for a line to fit in to explain these data points. Here instead of a line we would look for a plane that would minimize the sum of squares error

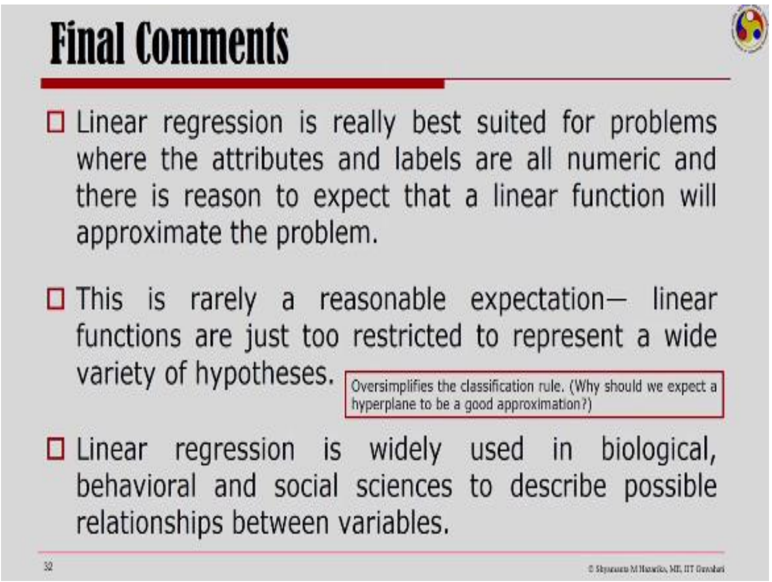
(Refer Slide Time: 34:40)



For a general number of attributes this could be expanded and we could look at a d + 1 dimensional space with a coordinate for each attribute plus an additional coordinate for the label and then we would be talking of a hyperplane, a d-dimensional hyperplane that would minimize the sum of squares error to explain the data. Now this is slightly different from the two dimensional case.

Because when I am talking of multivariate regression with a general number of attributes I assume that the hyperplane go through the origin, whereas when I was trying to fit in a line in linear regression in the two dimensional case I did not have such a constraint. So basically forcing this keeps the mathematics prettier and this is how we do multivariate regression.

(Refer Slide Time: 35:52)



Now linear regression is what we have discussed today and is really best suited for problems where attributes and labels are all numeric and there is enough reason to expect that a linear function will approximate the problem. Now linear functions actually are just too restricted to represent a wide variety of hypotheses and therefore this is rarely a reasonable expectation. This actually in a way oversimplifies why, should we expect a hyperplane to be a good approximation and one that we force to go through the origin.

Nevertheless, linear regression is widely used in biological, behavioral and social science to describe possible relationship between variables. So we have looked at Linear Regression today and we would take up a classification algorithm called the support vector machines in our next lecture. Thank you.