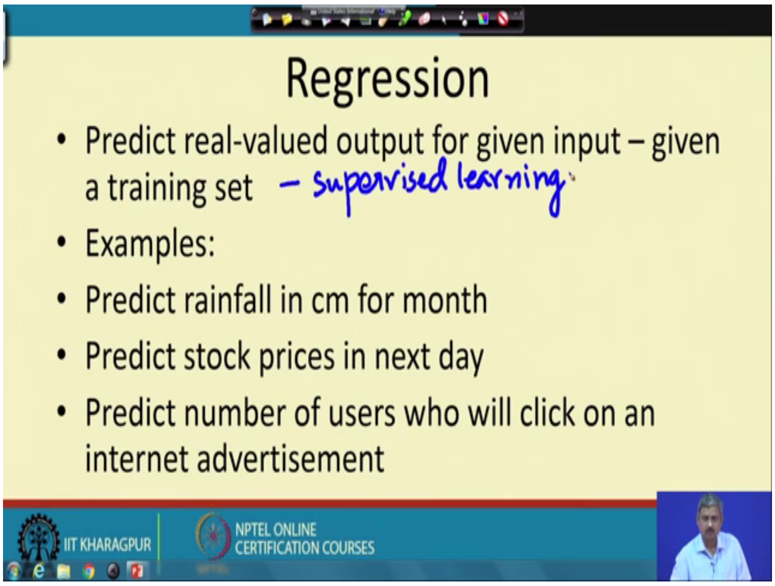
Data Mining
Prof. Pabitra Mitra
Department of Computer Science & Engineering
Indian Institute of Technology, Kharagpur

Lecture - 37
Regression – I

We will now discuss another Data Mining problem known as Regression Analysis. So, in this week, the 8th week, we will discuss Regression Analysis along with few other techniques that are essential for Data Mining.

(Refer Slide Time: 00:40)

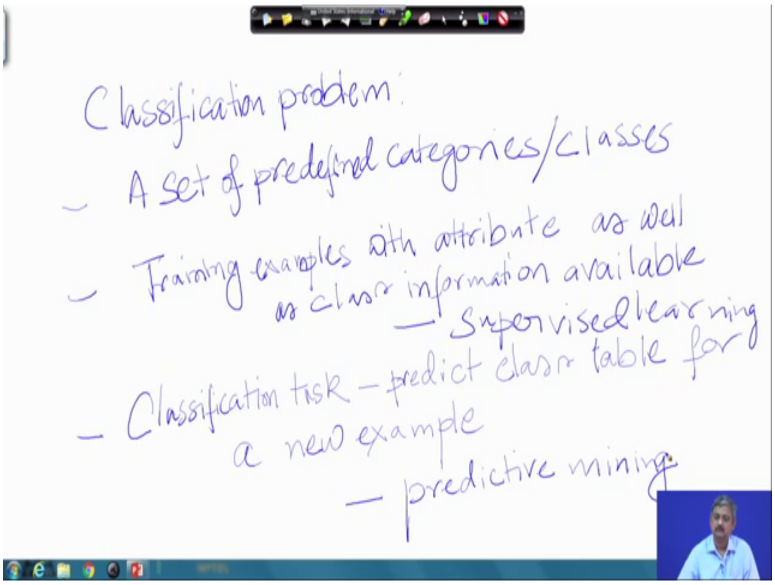


Regression

- Predict real-valued output for given input – given a training set – supervised learning
- Examples:
- Predict rainfall in cm for month
- Predict stock prices in next day
- Predict number of users who will click on an internet advertisement

So, we had previously studied mainly two important Data Mining class, besides the association rule mining. One is the classification, where there are some predefined set of classes. For example, a spam or a non-spam; for example, a news article falling in a category say sports, politics, local. So, there are defined categories and you have a training example, a set of training example, where there is a supervised output; that means, for a every example, what should be the category, it is mentioned.

(Refer Slide Time: 01:45)

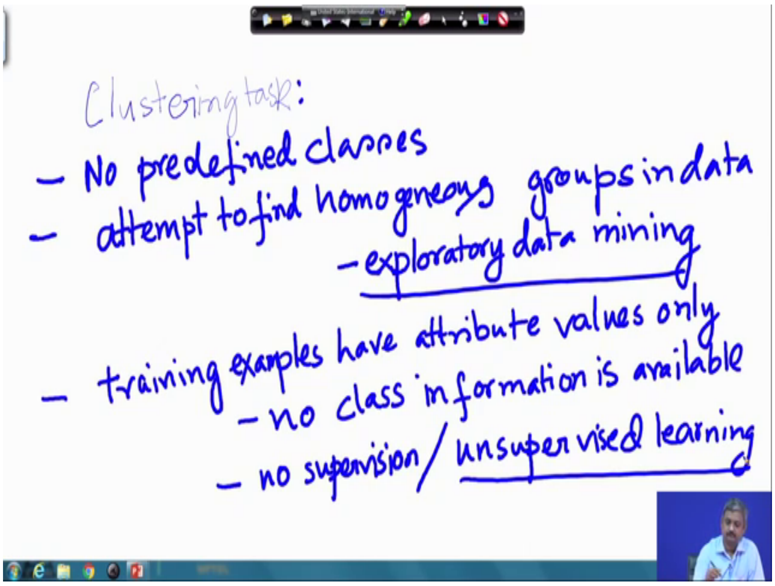


Classification problem:

- A set of predefined categories/classes
- Training examples with attribute as well as class information available
  - Supervised learning
- Classification task – predict class table for a new example
  - predictive mining

So, if I quickly write down, a set of predefined categories or; you have a training set where not only the attributes of the examples but also the class levels are available you have to predict the class level for a new example. So, you have examples of spam and non-spam email, you have to predict a new email whether it is spam or not a spam. So there are two aspects, one is there is this class level information along with 10 attributes which is called Supervised learning. Second is, you are predicting for an unknown example, why it is predictive mining ok. So, these are the two important problem.

(Refer Slide Time: 04:27)

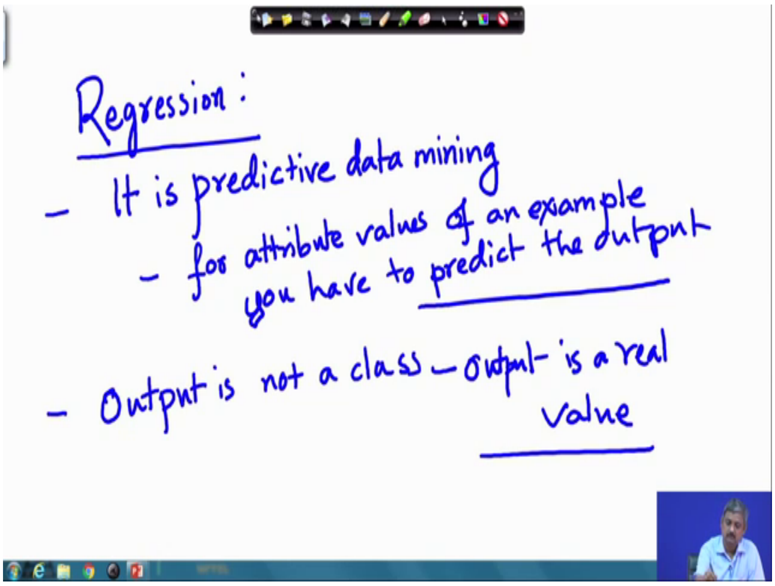


Clustering task:

- No predefined classes
- attempt to find homogeneous groups in data
  - exploratory data mining
- training examples have attribute values only
  - no class information is available
  - no supervision/ unsupervised learning

Similarly, we also saw another task. We also saw the Clustering task where no predefined classes. You just find to try to find the homogenous groups in that data. So, it is that is why called exploratory data mining and training examples, they have attribute values only; no class information is available ok. So, nobody tells that for this example, this should be the class. So, there is no supervision. So, it is unsupervised learning. That is exploratory and unsupervised.

(Refer Slide Time: 07:02)



Regression :

- It is predictive data mining
  - for attribute values of an example you have to predict the output
- Output is not a class – output is a real value

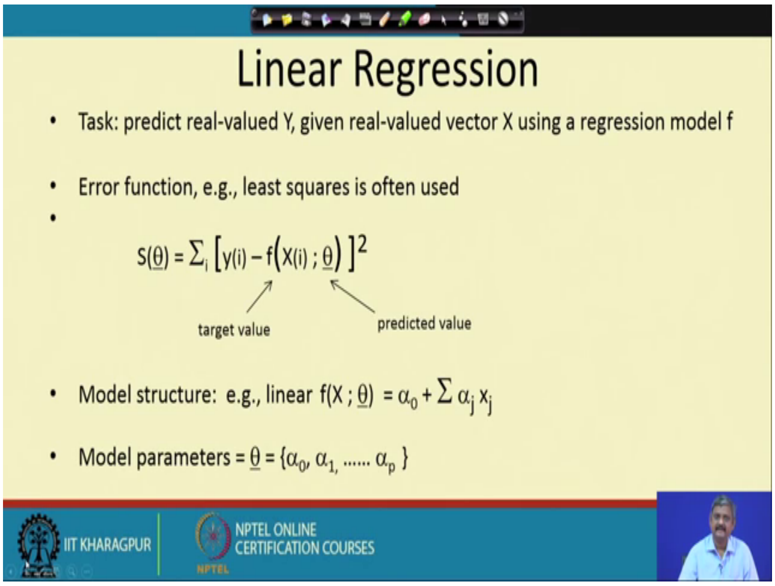
Now, we will look at another problem, Regression. It is predictive data mining; that means values of an example. You predict output but output is not a class; it is not a class. This, if it is a class, we can probably represent them by replace them by in tiers. Say spam is 0; non-spam is 1 ok, fraud is 0; non-fraud is 1, fraud is minus 1, non fraud is we can represent them by a, but here it is not a class. So, output is no longer an integer, output is a real value ok.

So, regression is predictive mining where output is real and you still have a training set. So, it is supervised also, learning also ok. So, some examples, may be you have gone to weather dot com and you want to predict the rainfall let us say in centimetre for the next one month, some real value say 230 centimetre or something. Or for example, you want to predict the value of a stock. In the next step that is also a real value whereas, if you predict if the stock goes up or goes down that is a 2 class classification problem ok.

Another example, many people study is the number of users who will or the fraction of users who will click on certain internet advertisement.

So these are all examples of Regression.

(Refer Slide Time: 10:30)



Linear Regression

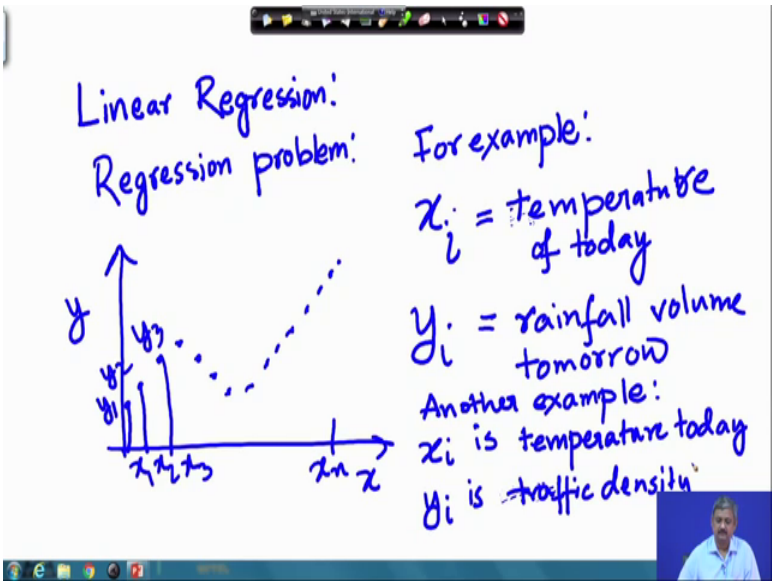
- Task: predict real-valued Y, given real-valued vector X using a regression model f
- Error function, e.g., least squares is often used
-

$$S(\underline{\theta}) = \Sigma_i \left[\underbrace{y(i)}_{\text{target value}} - \underbrace{f(x(i)\,;\underline{\theta})}_{\text{predicted value}} \right]^2$$

- Model structure: e.g., linear $f(X\,;\underline{\theta}) = \alpha_0 + \Sigma\, \alpha_j x_j$
- Model parameters = $\underline{\theta} = \{\alpha_0, \alpha_1, \ldots\ldots \alpha_p\}$

One of the most common technique of Regression is what is called a Linear Equation ok.

(Refer Slide Time: 10:55)



Linear Regression:

Regression problem:

For example:

x_i = temperature of today

y_i = rainfall volume tomorrow

Another example:

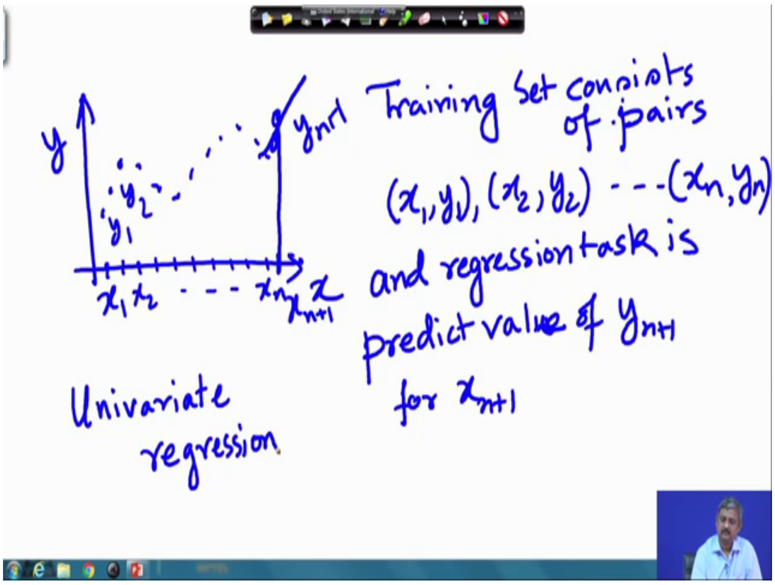
x_i is temperature today

y_i is traffic density

So, let me explain what it is. So, what is given to me is, at a set of values of x, x 1, x 2, x 3, x 4 up to say x n, I have the corresponding output y 1, y 2. So, x i say for example, say is the temperature today of certain place and y i can be volume tomorrow, ok; rainfall volume tomorrow all right. So, there can be other also. Let me take another example. So, x i is let us say, say how does traffic density, number of vehicles in a road depend on temperature?

If it is very hot people do not go out, if this very cold people do not go out. So, can I predict the traffic density as a function of the temperature of the day and that is a Regression problem ok.

(Refer Slide Time: 14:57)



Training Set consists of pairs

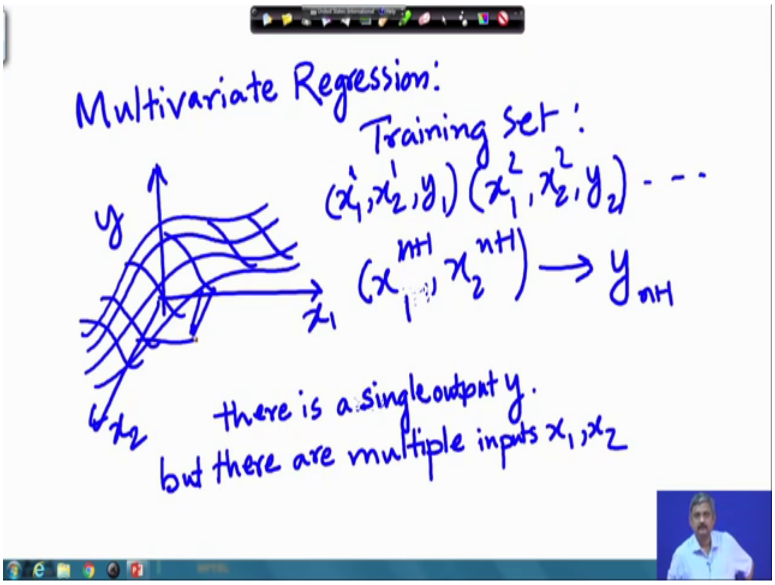
$(x_1,y_1),(x_2,y_2)---(x_n,y_n)$

and regression task is predict value of y_{n+1} for x_{n+1}

Univariate regression

So, in general, Training Set consists of y n. Given a new point x n plus 1, what should be the value of y n plus 1? Ok, that you have to predict, all right. So, may be you extend the curve like these and see what value it takes, like that you do ok. So, this particular example when the x has only a single value x 1, x 2 is called a Univariate regression.

(Refer Slide Time: 17:10)



Multivariate Regression:

Training set:

$(x_1^1, x_2^1, y_1)\ (x_1^2, x_2^2, y_2)\ \ldots$

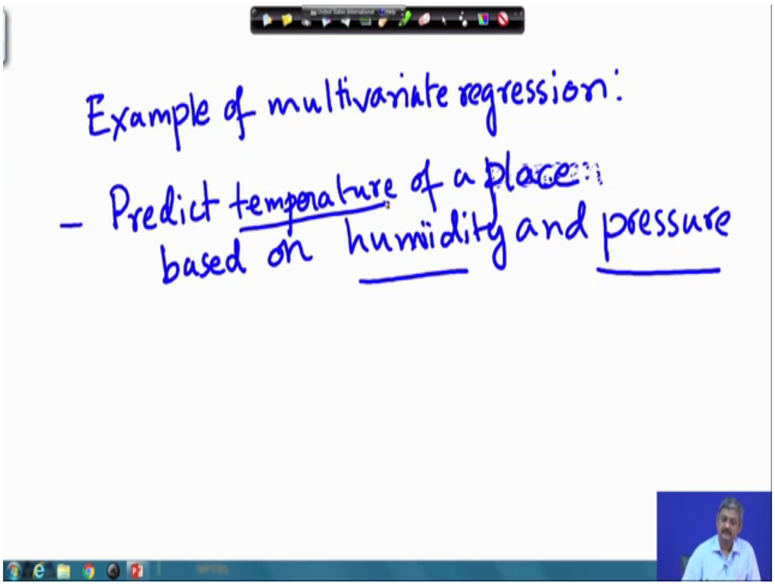
$(x_1^{n+1}, x_2^{n+1}) \rightarrow y_{n+1}$

there is a single output y.

but there are multiple inputs x_1, x_2

We can have, we can have Multi value. There is a x 2 ok. So, here it what will be and so, on. And for a new point sorry 1, you have to predict is Multi value.

(Refer Slide Time: 19:44)

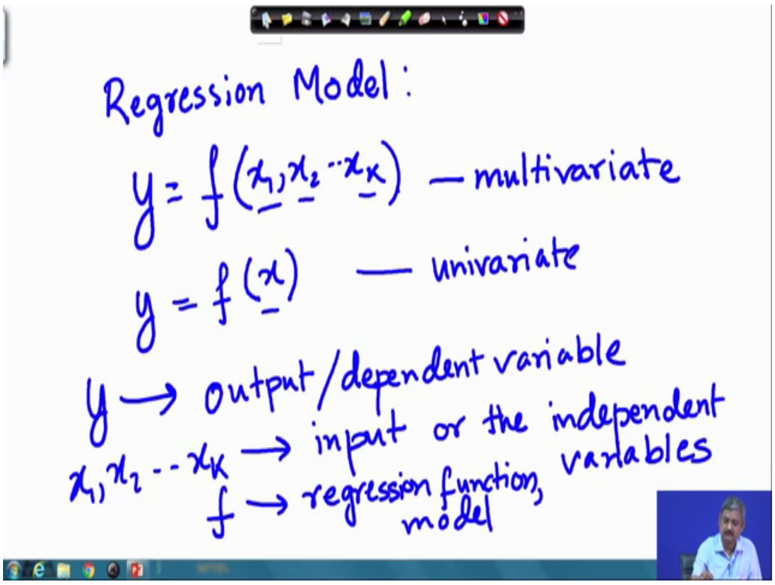


Example of multivariate regression:

- Predict temperature of a place: based on humidity and pressure

So, basically one point here, you know you have to predict it. For 2, given x 1, x 2, you have to predict 1 ok. So, let us see some example of Multivariate regression. Place based on pressure, based on humidity and pressure, you predict the temperature place. So, there are two inputs, bivariate and one output. Actually, in general, there can be multiple outputs also.

(Refer Slide Time: 21:21)



Regression Model:

$y = f(\underline{x_1}, \underline{x_2} \cdots \underline{x_K})$ — multivariate

$y = f(\underline{x})$ — univariate

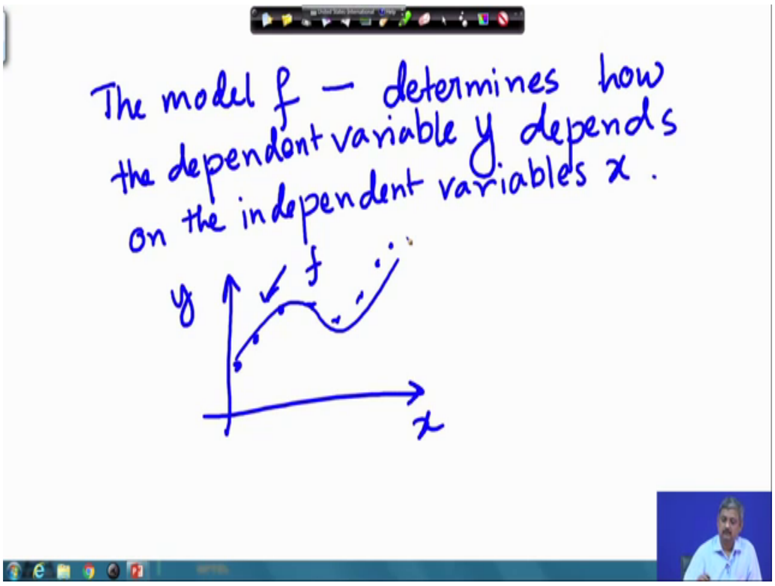
$y \rightarrow$ output/dependent variable

$x_1, x_2 \cdots x_K \rightarrow$ input or the independent variables

$f \rightarrow$ regression function, model

So let us write down what is called a Regression Model. Regression Model is nothing but expressing y as a function of the input x ok. As a function, if it is univariate. So, these are the attributes and this is attribute, for multivariate case multiple attribute, univariate one attribute; y is called the output or the dependent variable; whereas, input or the independent variables and f is called the function or the regression model ok.

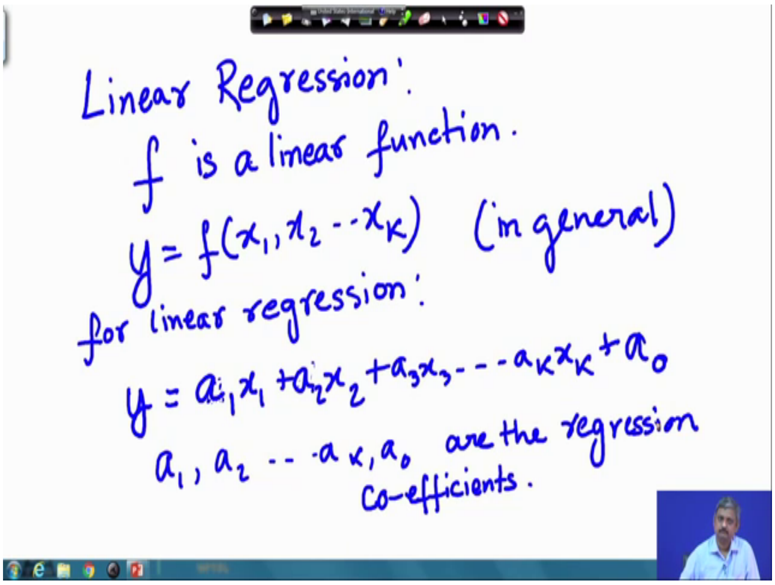
(Refer Slide Time: 23:42)



The model f — determines how the dependent variable y depends on the independent variables x.

So, the idea is that dependent variables or idea is, f determines how ok. That is the idea.

(Refer Slide Time: 24:54)



Linear Regression:

f is a linear function.

$y = f(x_1, x_2 \cdots x_K)$ (in general)

for linear regression:

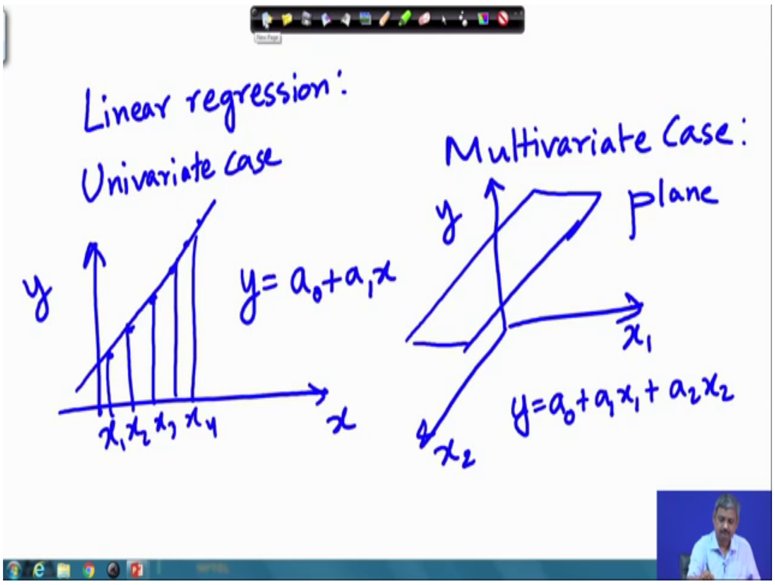
$y = a_1x_1 + a_2x_2 + a_3x_3 \cdots a_Kx_K + a_0$

$a_1, a_2 \cdots a_K, a_0$ are the regression Co-efficients.

So you basically, f is this function, f is this function.

So, what is Linear Regression? When f is a linear function ok; so, for linear regression, some constant alpha 1, alpha 2 or not alpha; let me write the mouse all right, write them as a 1, a 2. So, f is this linear function, f is this linear function are the regression are the called the regression Co-efficient; a 1, a 2, a k, a naught, they are called the regression Co-efficient ok.

(Refer Slide Time: 27:18)



Linear regression:

Univariate case

$y = a_0 + a_1x$

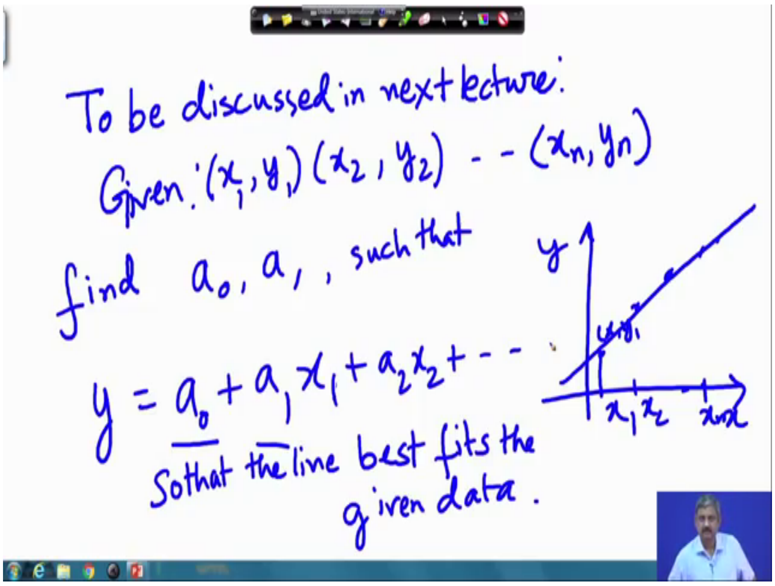
Multivariate Case:

plane

$y = a_0 + a_1x_1 + a_2x_2$

So, just f is of this particular form, a 1, a 2 a simple ok. So, in the univariate case, defining values if they are given, then a 1 x, straight line; a naught is the shift and a 1 is the slope. In bivariate, it is a plane.

(Refer Slide Time: 29:15)



To be discussed in next lecture:

Given: $(x_1, y_1)\,(x_2, y_2)\,--\,(x_n, y_n)$

find a_0, a_1, such that

$y = \underline{a_0} + a_1 x_1 + a_2 x_2 + ---$

So that the line best fits the given data.

So that is linear equation as you can expect. If we are given some regression points say x 1, y 1; x 2, y 2; given some points, a naught such that so, basic. So, you find values of a naught and a 1, so that the given data it matches; best fits the given data that we will find out. If we have higher dimension, we have more terms. So, the all these terms you have to find out ok.

So, that is the Regression design problem; so that we will do in our next lecture, how to find that.

Thank you.