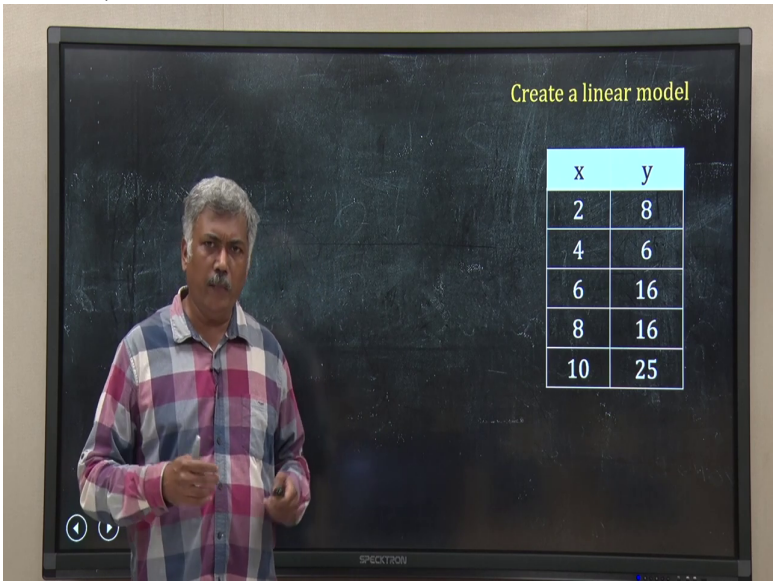
Data Analysis for Biologists
Professor Biplab Bose
Department of Biosciences & Bioengineering
Mehta Family School of Data Science & Artificial Intelligence
Indian Institute of Technology, Guwahati
Lecture – 27
Linear Regression - I

(Refer Slide Time: 00:47)

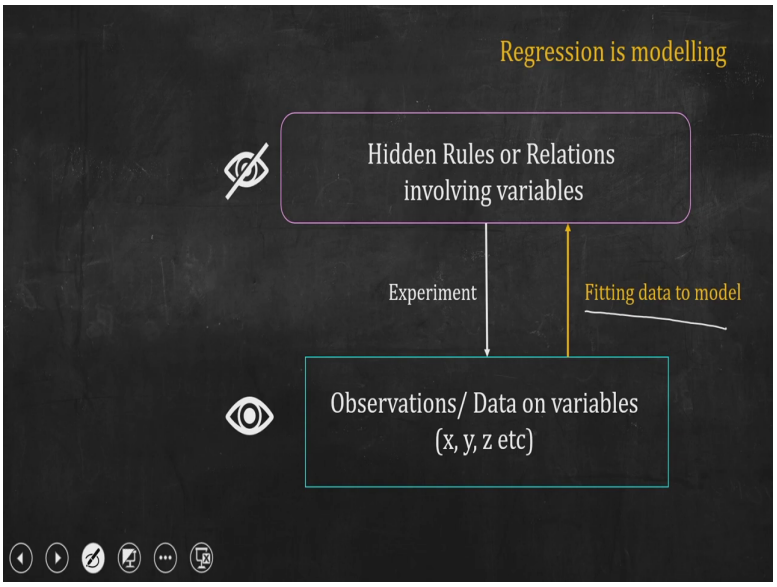


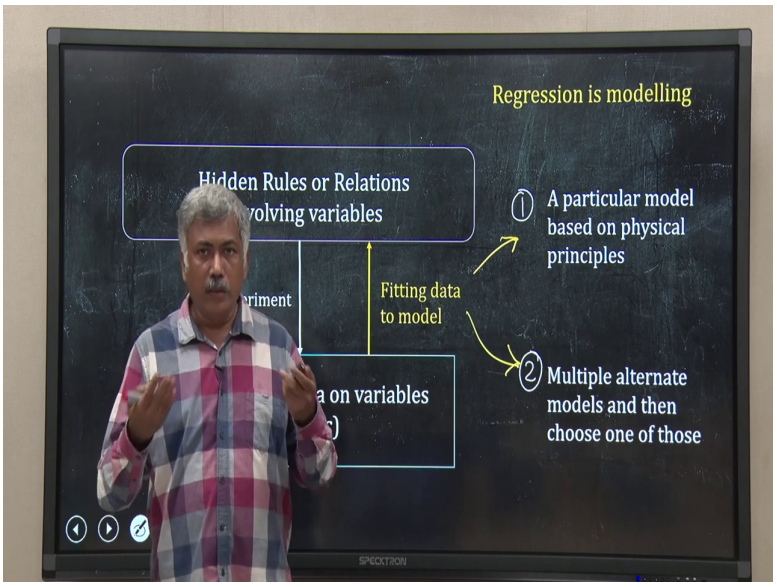
X	Y
2	8
4	6
6	16
8	16
10	25

Hello everyone. Welcome back, in this lecture we will learn linear regression. In fact, we will have two lectures on linear regression, this is the first one. So, let us start with the problem, suppose I have a data and I have shown a data set here, 5 data points, for two variable x and y. And I want to create a linear model out of this data set.

Now, you must be wondering, why I am talking about modelling. This is supposed to be a lecture on regression, data fitting and you must have done data fitting like this in using excel or some other thing, in your experiment, in your lab. Let me explain, why I am talking this problem of regression as a modelling problem. In fact, linear regression in general regression is a technique of modelling. What do I mean by that?

(Refer Slide Time: 1:21)





See, when we do experiments, we measure something. We observe something and we get a data. So, for example we have three variable that we may be measuring. Suppose, some blood parameter in some animal experiment you are measuring that is your observation. Now, for some treatment or some other thing these parameters are changing and you have measured that, those changes you have measured. Now, those changes are happening because there must be some relationship between these variables. Somehow these variables must be connected to what you are doing with the animal.

So, what you are checking in the experiment. So, those relationships are unknown to you, those are hidden to you. So, you have hidden rules or relations involving variables. You do not know them, you cannot see them by the experiment, you see the effect of those things.

Now, what we are trying to do, what we are trying to do, once we got the data we try to create a model by fitting the data to model, so that I can understand these hidden rules or relations between variables. That is what we do in science. Now, what type of model I should fit? Now, that you have two options broadly, one is you take a particular model which is based upon physical principles that you believe is correct in this case.

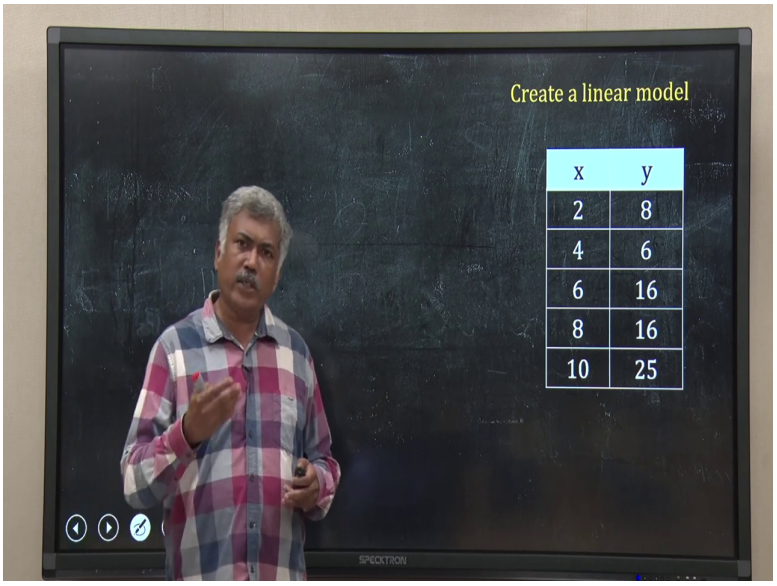
This is my first case, a particular model based on physical principles. What do I mean by that? Suppose, you have measured the absorbance of different concentration solutions of DNA. So, you have two things, one variable is concentration, the other variable that you have measured is the absorbance. So, you have absorbance versus concentration data.

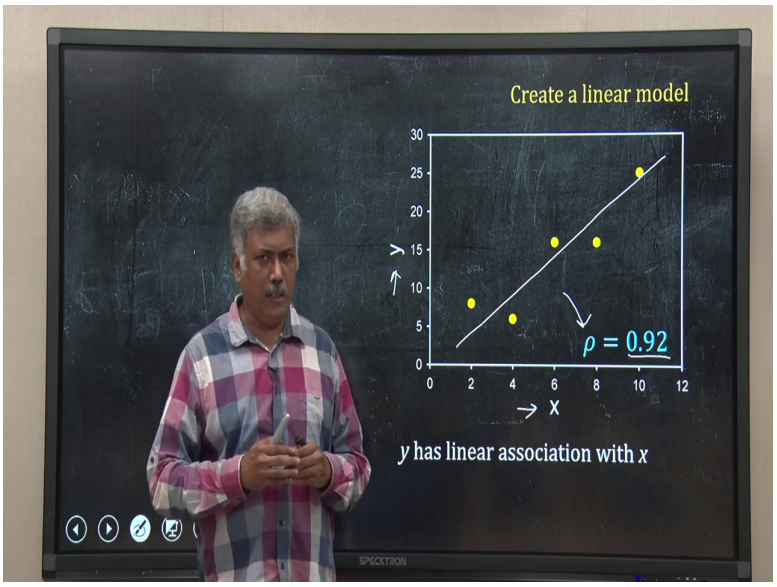
Now, from the basic chemistry we know physical chemistry we know that absorbance follows Beer Lambert law with certain assumptions obviously. So, you know that Beer Lambert Law is a linear thing. Absorbance is linearly related to concentration. So, in this case the physical principle is defined by the Beer Lambert Law and you want to fit absorbance with concentration by a linear model. That is a straight line.

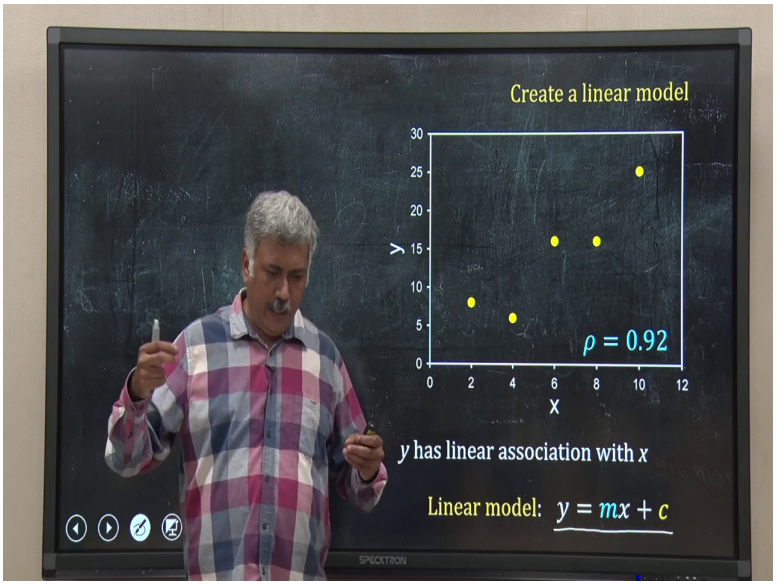
So, this is simple. So, in the other case you may have a situation where you may have to assume multiple alternate models and then you fit your data to all of them and then eventually you choose one depending upon some mathematical characterization and other characteristic judiciously you choose one of the those model which you have fitted to your own data.

So, eventually whatever you are doing? Out of these two more techniques whenever you are doing regression you are actually fitting your data to a model or more than one models to identify the hidden relationship which is hidden, which you cannot see, you want to identify those hidden rules and relations.

(Refer Slide Time: 4:21)







$Linear\ model\colon\ y\ =\ mx\ +\ c$

So, let us go back to our data and find out the hidden linear relationship between x and y. Now, the first question you should ask, why do I believe that there is a linear model behind this data? Why there is a linear relation between x and y and that is why every time when you think of regression, you should think about it. So, just look at it, the first thing what I did, I looked at the data, I plotted it.

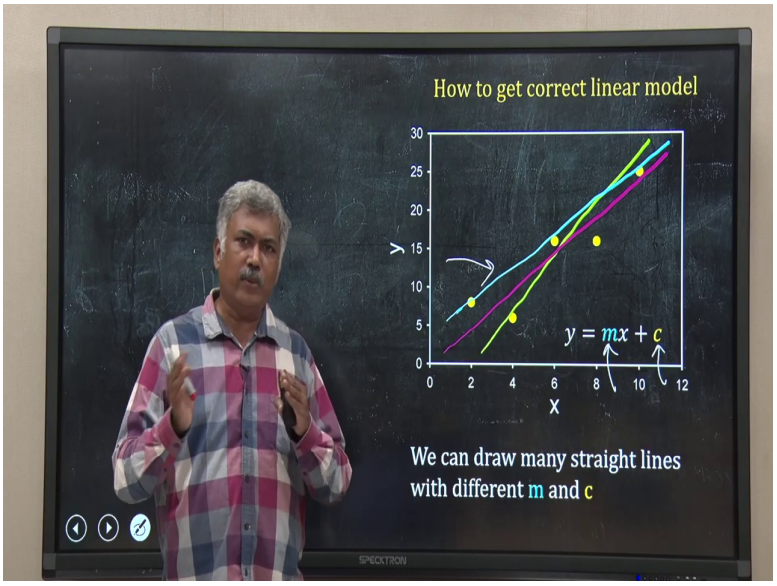
That is the best approach to start with. So, I have x in the horizontal axis, y in the vertical axis and I have five yellow data points. And looking at it without thinking of something complicated, undulating line, I can simply think maybe possibly, I can fit these to a straight line like this.

The second diagnostic thing I have done, I have calculated the Pearson correlation coefficient. If you remember, in last lecture we discussed Pearson correlation coefficient gives you the strength of linear association between two variable in my data and that came out to be 0.92. It is quite reasonable in linear strength. So, that means my belief that I will be able to fit a linear model to this data set. It is quite a good belief. So, I want to fit this data to a straight line, a linear model is a straight line. So, I want to fit this data to the equation, y equal to mx plus c.

Now, here is a problem, as you know from your school level geometry idea that if I give you a straight line that can be defined by two points. If you give me just two point, I can draw a straight line through that. If you have more than two points then you are in trouble. If they are

all in the same straight line, that is fine. But if they are not on the same straight line then how can you fit all these three points to one single straight line? You cannot.

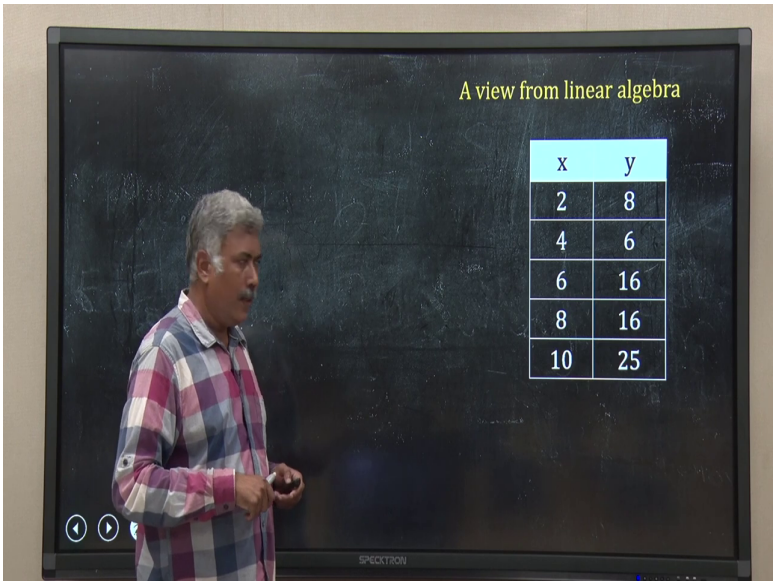
(Refer Slide Time: 6:14)



The same problem is here. I have five data point and I can draw in finite numbers of different straight line with different values of m and c. None of them will go through all the points. They will always leave some of them. But I can have infinite this type of straight line, which one should I choose?

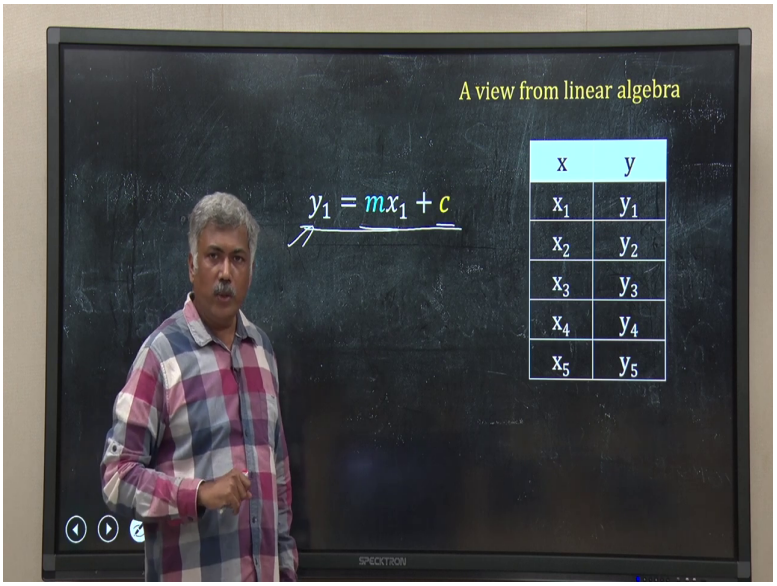
So, there is no straight answer to that, how should I choose the right one out of this infinite number of straight line each having different value of m and c. So, we have to sort this out. This problem can be looked from another angle also, from a linear algebra perspective. If you remember, in one lecture on linear algebra, we have studied a system of linear equations.

(Refer Slide Time: 7:03)



A view from linear algebra

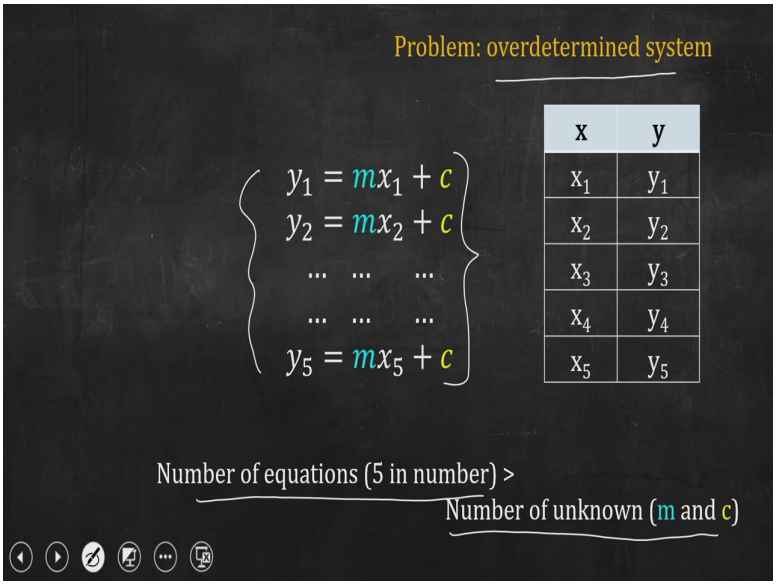
x	y
x_1	y_1
x_2	y_2
x_3	y_3
x_4	y_4
x_5	y_5



A view from linear algebra

$$\text{5 Eqn}\begin{cases} y_1 = mx_1 + c \\ y_2 = mx_2 + c \\ \cdots \quad \cdots \quad \cdots \\ \cdots \quad \cdots \quad \cdots \\ y_5 = mx_5 + c \end{cases}$$

x	y
x_1	y_1
x_2	y_2
x_3	y_3
x_4	y_4
x_5	y_5



x	y
x1	y1
x2	y2
x3	y3
x4	y4
x5	y5

y1 =
mx1 + c
y2 =
mx2 + c
y3 =
mx3 + c
y4 =
mx4 + c
y5 =
mx5 + c

$$Number\ of\ Equations\ (5\ in\ number) >\ \ Number\ of\ unknown\ (m\ and\ c)$$

So, let me pose this problem from that perspective. So, I have this data, 2, 4, 6, 8 something like that. This numerical thing is difficult to handle when I write equations, so let me replace them with x1, x2, y1, y2 something like that up to x5 and y5. Now, take this one, the first row of data x1, y1. When x is x1, y is y1.

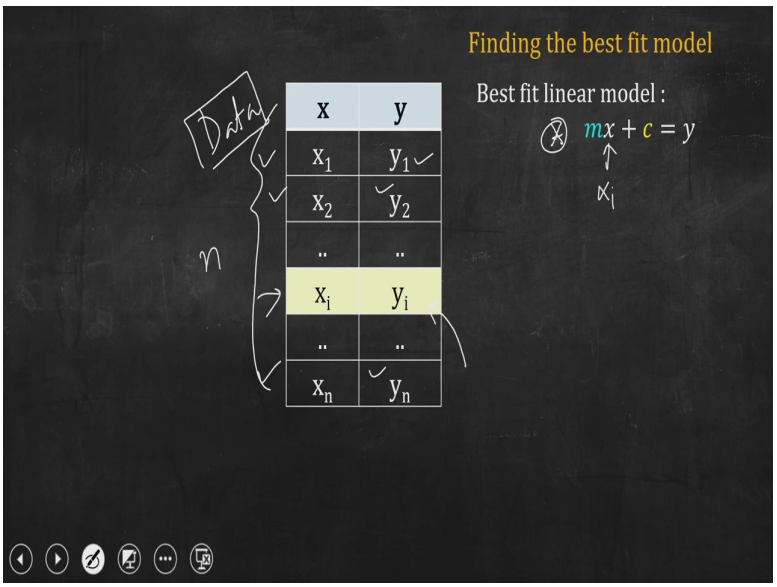
Now, if I consider that there is a linear relationship of the form y equal to mx plus c, then for this first data I can write a equation like this, m into x1 plus c must be equal to y1 because the data says when x is x1, y is y1 and if there is a linear relationship then, this equation must be valid. I do not know what is m and c but this equation must be valid. Now, for each of this data, each of this data I can do that. From 1 to 5.

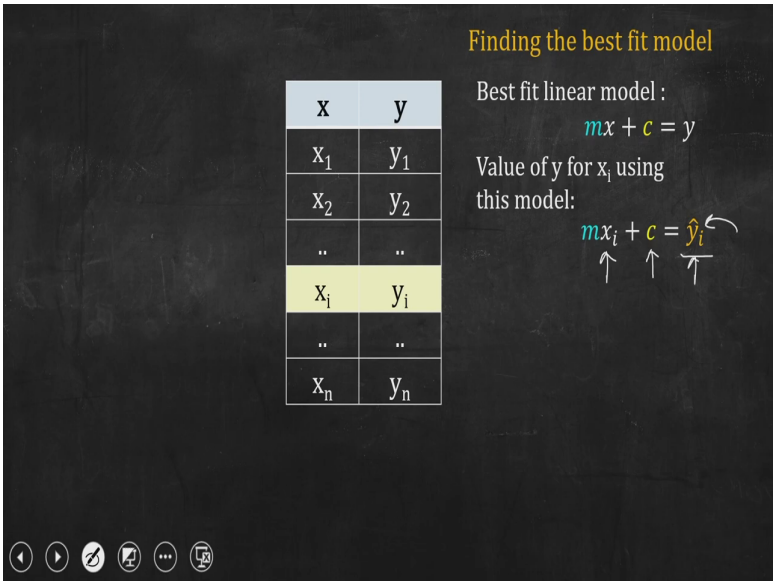
So, in this way how many equations you have got. You have 5 equations, how many unknowns you have? You have two unknowns, m and c are unknown or variable you can vary those value of m and c to find out the correct one. So, we have land up in a situation where the number of equation that is 5 in this case is much bigger than number of unknown that is 2, m and c.

So, if you remember our lecture on linear equations from the linear algebra lecture series, in this course, this is a situation which what we call over determined system. So, this system is a over determined system and for over determined system, there is no unique solution. So, these five equations they do not have any unique solution. So, I do not have a unique value of m and c which will satisfy all these five data points.

So, what will be the way forward? If you remember that linear equation lecture, we discussed that in that lecture that in this type of situation, when I do not have any unique solution, what we do? We convert this problem into optimization problem. We do not have a unique solution but we can find an optimum solution. So, what we do? We put an extra constraint and try to optimize something. We want to find something a middle ground we want to find.

(Refer Slide Time: 9:41)





x	y
x1	y1
x2	y2
..	..
xi	yi
..	..
xn	yn

$$mx + c = y$$

$$mx_i + c = \hat{y_i}$$

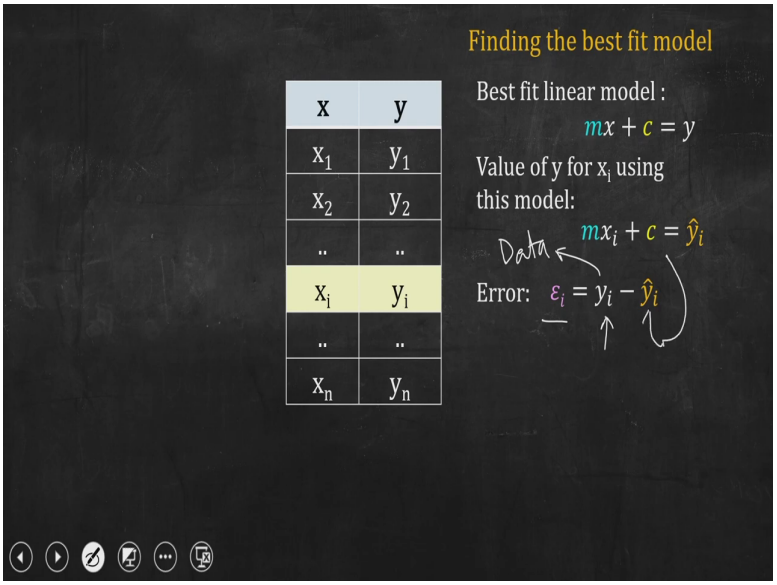
So, how will do that? Let us see, so this is my data and I have left that 5 data point issue, I have generalized it. I have started from x1, y1, x2, y2 up to xn, yn. So, I have n data points. In pairs, and my model is mx plus c equal to y. Let us just flip this one, the equation so that it

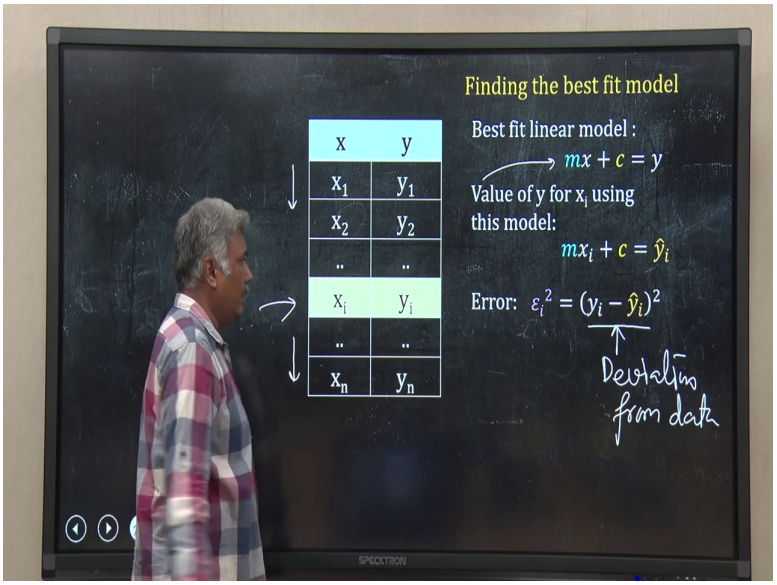
will be much easier for me to write. So, mx plus c, is equal to y this is my linear model, I want to fit the data to this.

So, let us take this, ith row xi and put that value of xi in place of x. Suppose, I know m and c, I have assumed suppose, so I know m and c, so multiply xi with m then add c to that and you will get a value of y. Now, for most cases as you can easily understand, this y will not be same as this y1 which is present in the data, it may be different.

So, let me write it down, so I have put the value of xi and added the value of c that I have assumed and I have got a value of y which I will mark as yi hat because that may not be same as the yi which is in the data. So, as I said my yi hat is coming by putting xi into the equation and it is not equal to possibly most of the time, is not equal to the original yi that means there is a difference between yi and yi hat. So, that is my error.

(Refer Slide Time: 11:24)





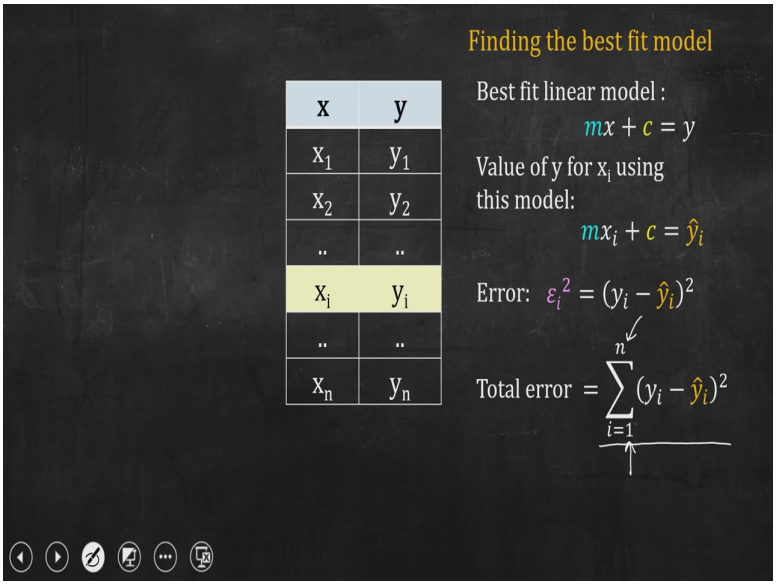
$$Error: \epsilon_i = y_i - \hat{y_i}$$

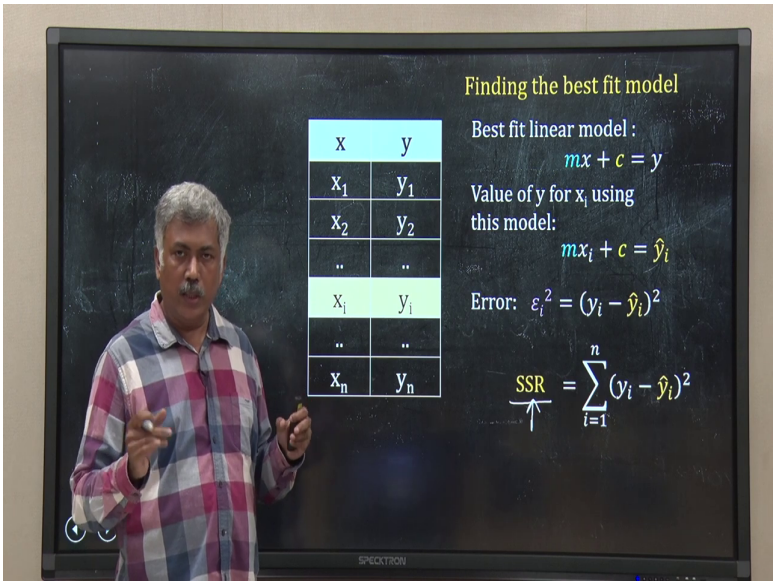
So, I define that as error eta i, so what is that? The error eta i is yi that is coming from the data. And yi hat which is coming from the equation. You find the difference between that, that is the error. Now, this error can be positive and negative because in some cases yi hat can be bigger than yi. So, in that case the error eta will be negative. If yi hat is smaller than yi, it will be positive. So, if you have that type of situation, it is better to square the error term so that we can get rid of the negative terms from my calculation.

So, that is why I will define error not by eta by eta square and that will be nothing but simply the deviation, this is nothing but the deviation. Deviation from data, so yi hat is deviated from yi, the data point and you take the deviation square. So, that is my error, error for what? Error for this point when I have taken that point and use this equation with a particular value of m

and c. Now, I can do this whole thing for all the data points starting from 1 up to n. So, I will have n error squares, n number of error squares. So, I will add all those errors together. That will be my total error.

(Refer Slide Time: 13:00)





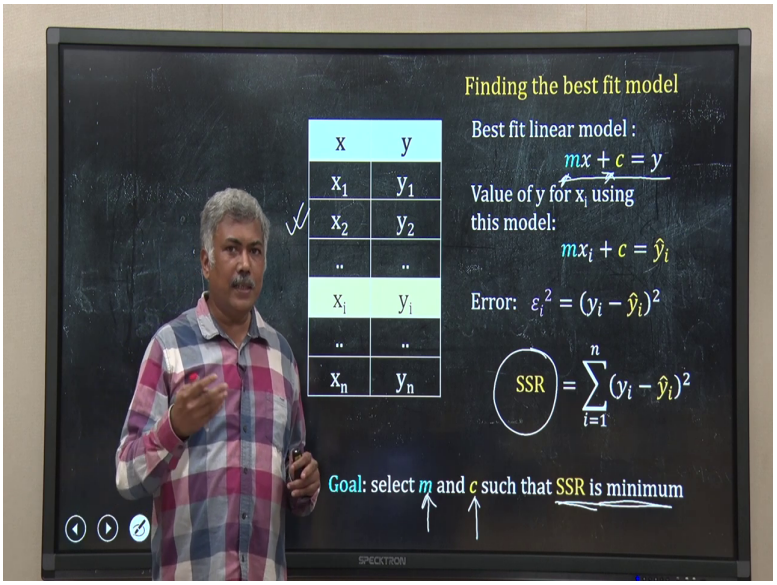
$$Error: \epsilon_i^2 = \left(y_i - \hat{y_i}\right)^2$$

$$SSR = \sum_{i=1}^{n} \left(y_i - \hat{y_i}\right)^2$$

So, that is how what I do? I add all of those, errors, the summation sign shows that it start from one, the first one first data set point to the last one, nth one and I get the total error. This error in some time in mathematics books will be called SSR, Sum of Square of Residuals. And what it shows, it shows the overall, I mean total error in the calculation. Now, what is

my goal in this problem? Remember, I want to find the optimum solution because I do not have a unique solution. So, I want to find a middle ground, I want to find the optimum straight line which will fit to this data that means it will have the least error.

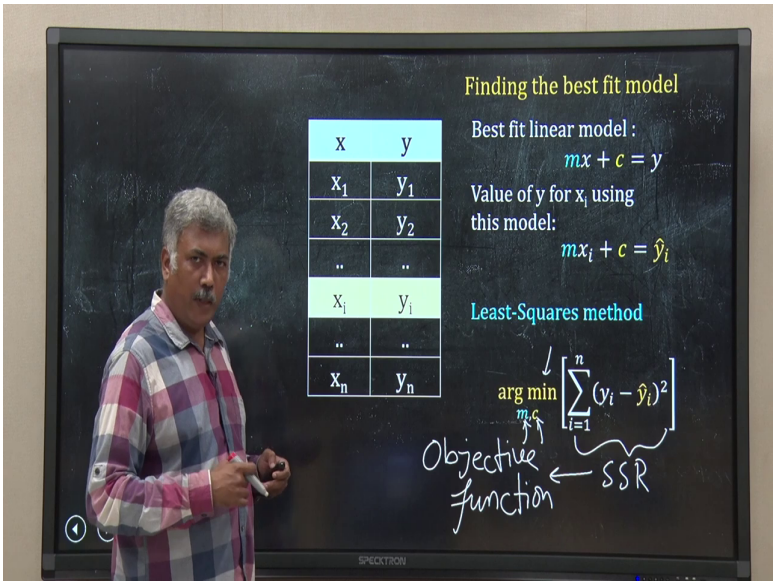
(Refer Slide Time: 13:50)



$$SSR = \sum_{i=1}^{n} \left(y_i - \hat{y_i}\right)^2$$

So, the goal for my whole exercise is that choose a particular value of m and c in such a way that this error SSR is minimum. So, that is what I have written here, the goal for me is to find the value of m and c such that for this particular given data set, the SSR will be minimum, if I use this equation mx plus c equal to y. So, that is the goal. You are trying to minimize the SSR. What is the SSR? SSR is a square deviation term.

(Refer Slide Time: 14:27)

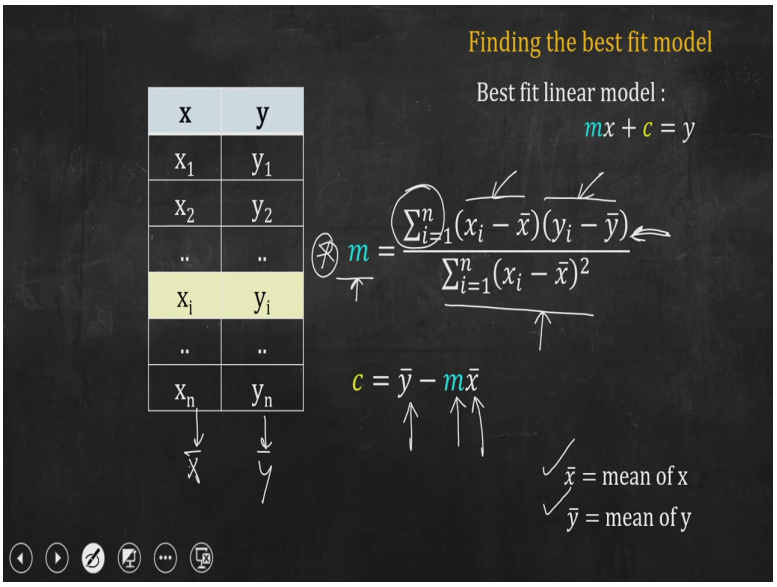


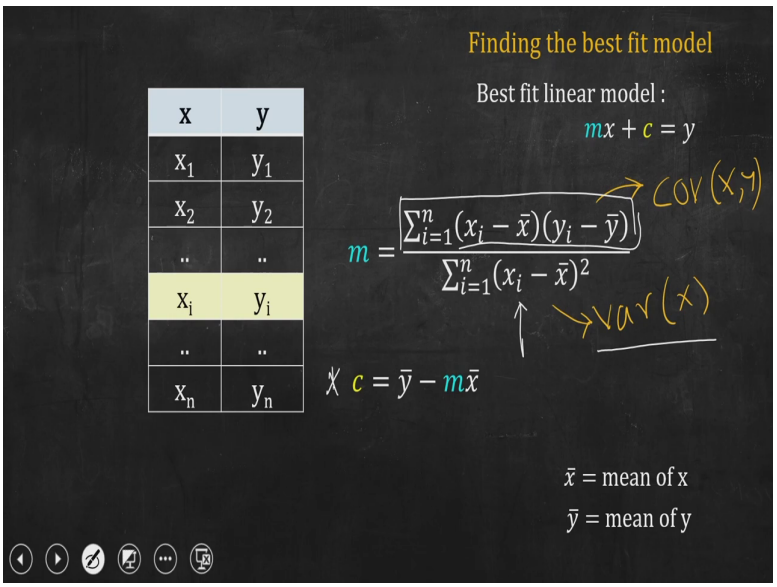
$$\arg arg \left[\sum_{i=1}^{n} \left(y_i - \hat{y_i} \right)^2 \right]$$

So, that is why this method is called Least-squares method and you must have heard this over Least-square many a times in many other places. So, this is what we are doing in least squares method. We have the SSR here, this is the SSR, the total error and in data analysis language we will call that is our objective function. Our purpose is to minimize this objective function using some algorithm, some method by doing what?

By choosing a particular m and c. We have to choose m and c, so this objective function which is the total error in my calculation becomes minimum. So, that is what is done in linear regression using the least squares method. Now, how should I minimize this objective function? There are two way, one is to use calculus because if you remember school days calculus, we can minimize maximize some function using differentiation. The other one is by linear algebra, I will not go in details of those mathematical thing.

(Refer Slide Time: 15:50)





$$m = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}$$

$$c = \bar{y} - m\bar{x}$$

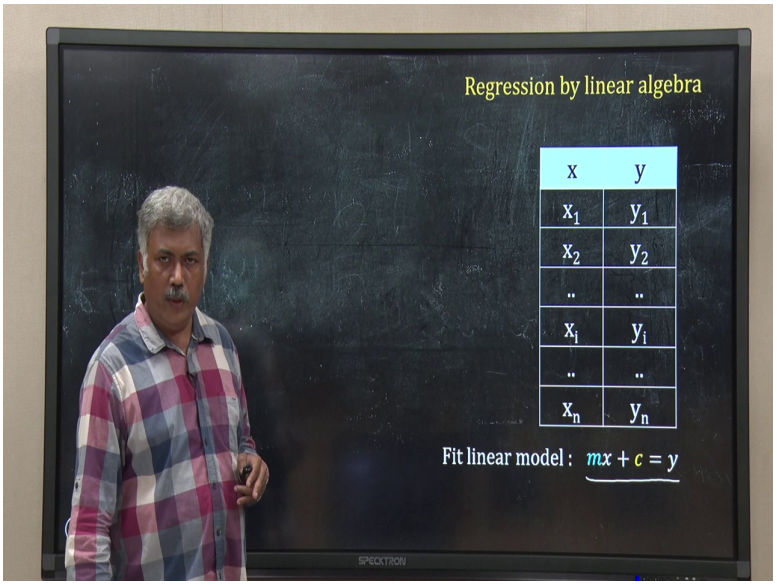
What we can show is that eventually if you do the optimization, that is the minimization of this objective function, you will get a relationship for m and c. So, m for my linear model y equal to mx plus c, m turns out to have the relation that it will be equal to xi minus x bar into yi minus y bar and summation of that for all the n data point. What is x bar and y bar? x bar is the mean of x, y bar is the mean of y.

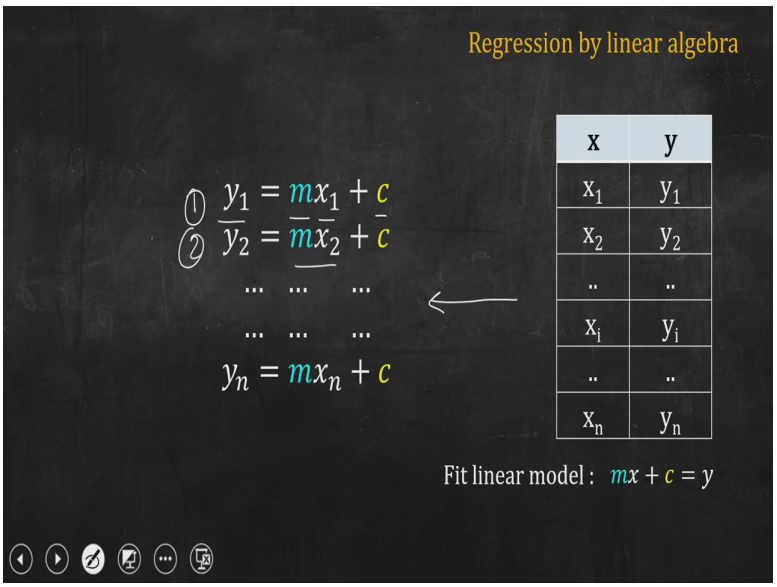
So, here for y you calculate the y bar. For all these data in this first column, you calculate the mean, so x bar. And then divide this numerator thing, this numerator thing by a denominator term which is xi minus x bar whole square and this is the sum for all the data point from 1 to n. This is m and once you have got m, you can easily calculate c, where c will be equal to y bar mean of y minus the m which you just now calculated using this relationship into the mean of x. So, using this formula, this relationship for a given data set, I can calculate m and c which is the solution which gives me the minimum SSR, minimum sum of square residual error.

Now, this one if you just pay attention a bit to these terms in the denominator and numerator. What you will realize is that, this thing, this upper term the numerator is nothing but covariance between x and y in your sample. Whereas the denominator, if you remember in one of our lecture we have discussed about sample variance and sample covariance.

So, the denominator is nothing but the variance of x, sample variance of x. So, what I can say that the m as per the least square method is the ratio between covariance between two variable divided by the covariance of the x. And once you have got the m obviously you can calculate c using this relationship. So, let us briefly look into how linear algebra transform this problem.

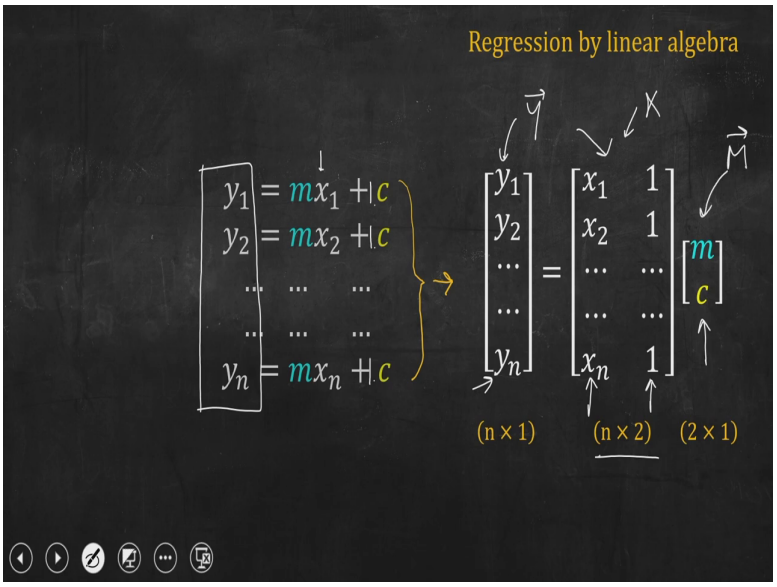
(Refer Slide Time: 18:19)

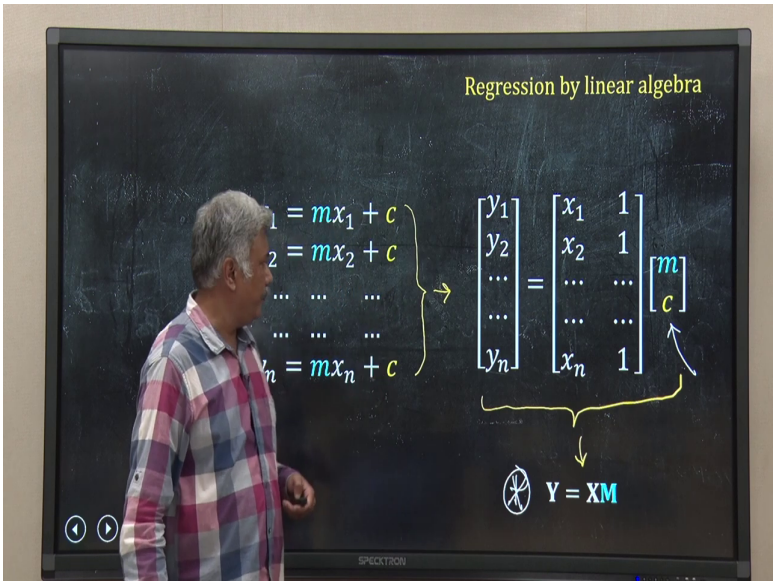




So, I have the data, same data set, n data points and I want to fit it to y equal to mx plus c. So, what I will do, I will convert this data set into a system of linear equation because we are fitting it to a linear model. So, I will convert this data into a system of linear equation. So, the first equation is y1 is equal to m into x1 plus c, the second equation is y2 equal to mx 2 plus c and that way I have the nth equation. So, I have n simultaneous linear equation, each having the same value of m and c, which we do not know, we want to calculate.

(Refer Slide Time: 19:06)





$$\begin{vmatrix} y1 \\ y2 \\ .. \\ .. \\ yn \end{vmatrix} = \begin{vmatrix} x1 & 1 \\ x2 & 1 \\ .. & 1 \\ .. & 1 \\ xn & 1 \end{vmatrix} \times \begin{vmatrix} m \\ c \end{vmatrix}$$

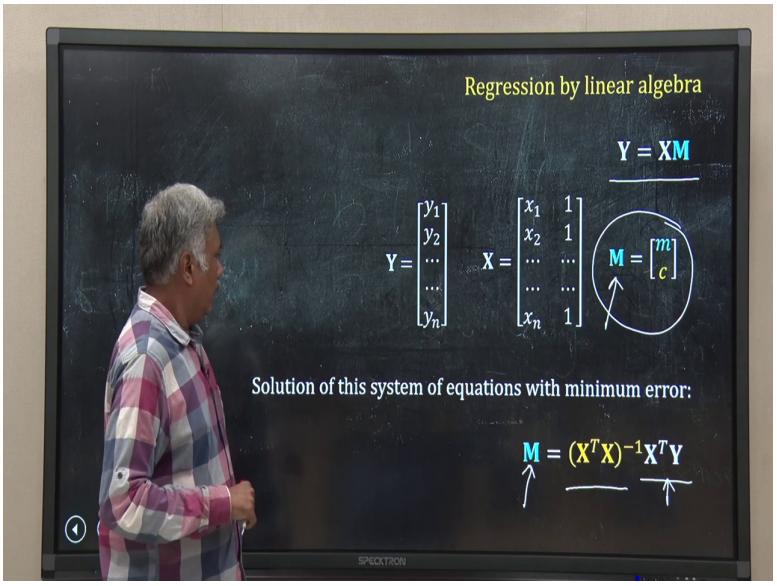
$$(n \times 1) = (n \times 2) \times (2 \times 1)$$

$$Y = XM$$

So, I will convert this system of equation in terms of vectors and matrices. So, by taking this y's, I get a n by 1 vector whereas by taking this x and remember c is actually 1 into c, 1 into c, so taking this x values and 1, I get a matrix which is n into 2 matrix, n into 2 matrix where the first column of the matrix is all the x and the second column is just 1. And then I multiply that with a vector m of m and c.

That dimension is 2 into 1. Now, this one, this vector we'll call y bar, this we can call x and this will call m bar suppose. So, that is what I have written here. So, I have converted the data into a system of equation in vector and matrix format and I get y equal to y vector is equal to x matrix into m vector, m vector is having this m and c which are unknown to me.

(Refer Slide Time: 20:15)

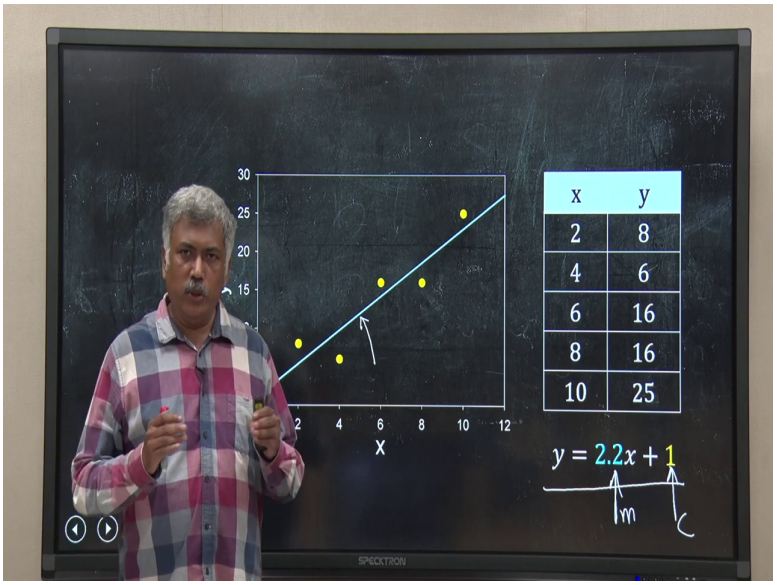


$$M = \left(X^T X\right)^{-1} X^T Y$$

So, what do I have? I have a system of equation of the format y equal to x into m and I have to calculate this m which has m small m and c, which is the intercept and the slope of the line. So, that using the concept of projection which I will not go in detail, how to do the calculation here, but using the concept of projection, what we have learnt in earlier in vectors, one can show that this m vector will be equal to x transpose x inverse, so you take the inverse of x transpose x and multiply that by x transpose and y.

So, you will get the m vector. So, as you can see here x and y matrix and vectors are coming from the data and then what you are doing? You are doing simply, some inversion, transposition and then multiplication and those are very easy to do using a computer. And that is why most of the time, if you are using a computer we use this method to calculate the value of m and c, using this linear algebra technique.

(Refer Slide Time: 21:25)

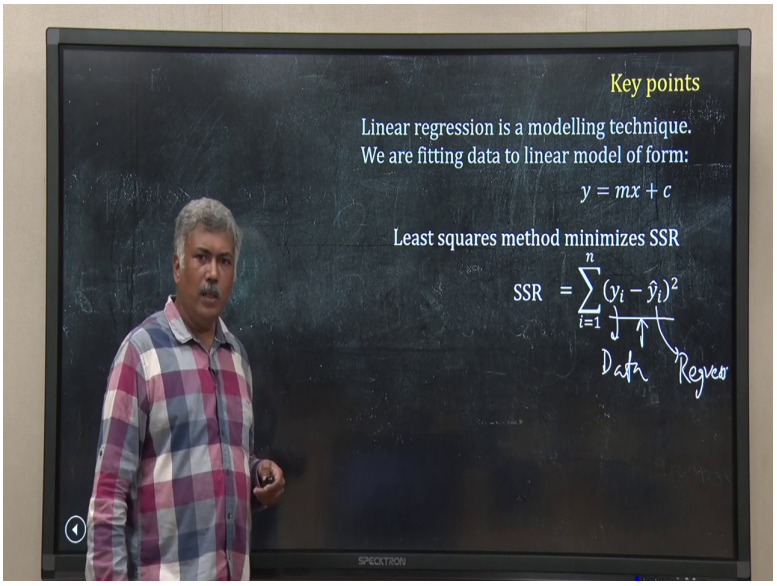


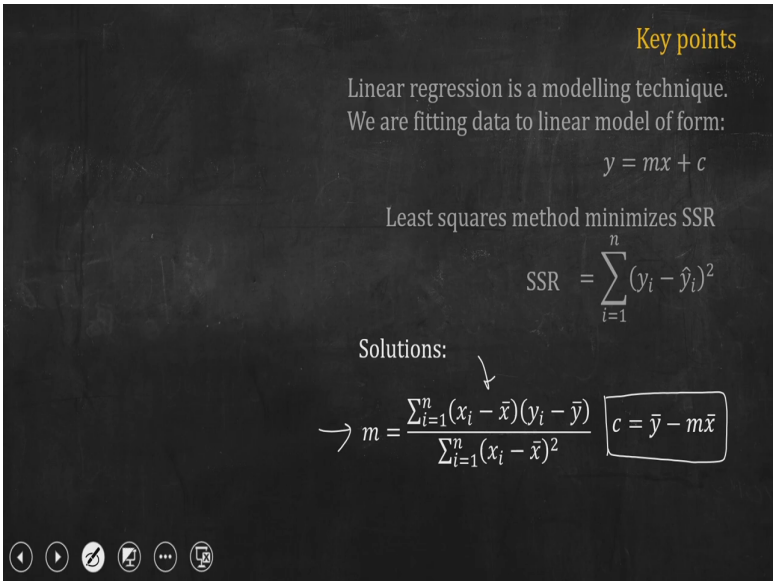
x	y
2	8
4	6
6	16
8	16
10	25

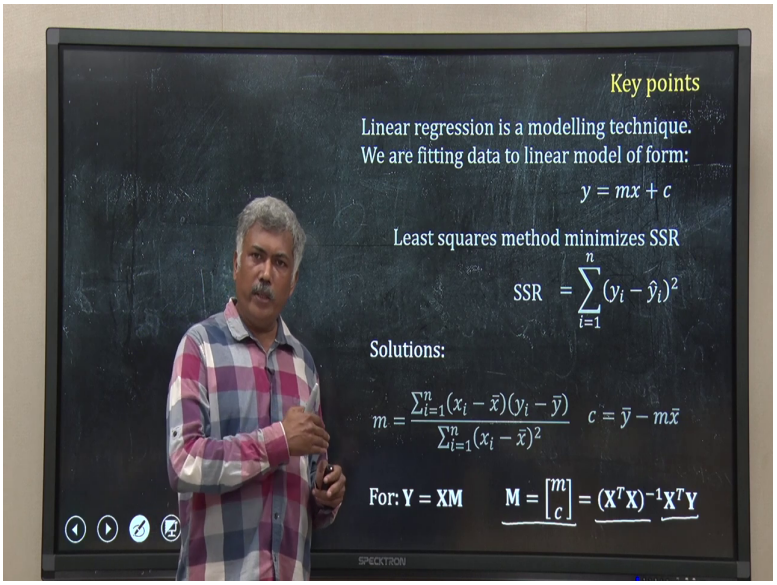
$$y = 2.2x + 1$$

So, far so good. We have discussed the method but let us go back to our original problem, I have the data and I have used r to perform the linear regression which has used the same principle of least square method using linear algebra and we have got the solution. So, what I have got? My m turn out to be 2.2 and c is 1 and I have shown that, I have plotted that line here. So, that is y equal to 2.2 into x plus 1. So, that is all for learning linear regression.

(Refer Slide Time: 22:01)







$$y = mx + c$$

$$SSR = \sum_{i=1}^{n} \left(y_i - \hat{y_i}\right)^2$$

$$m = \frac{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2}$$

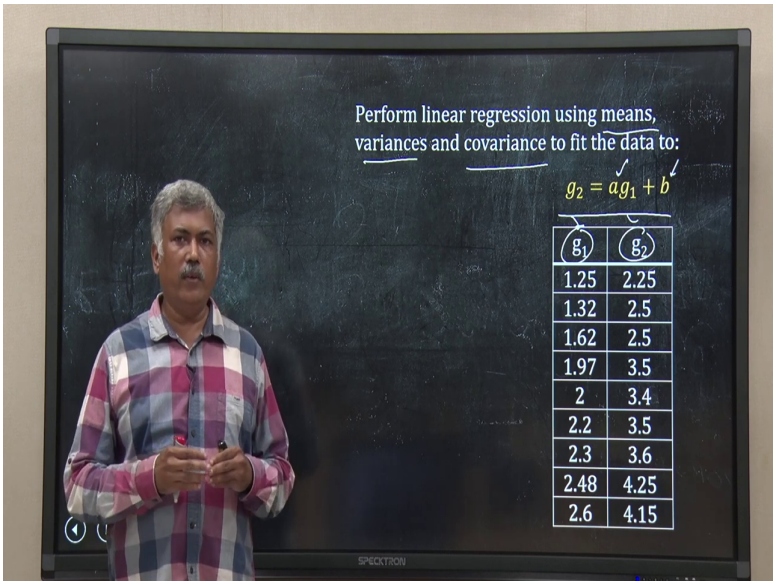
$$For: Y = XM$$

$$M = [m\ c] = \left(X^T X\right)^{-1} X^T Y$$

So, let me jot down the key point. The first thing that we have to remember is that linear regression is just like a modelling technique. What we are doing, we are fitting our data to a linear model of the form y equal to mx plus c and to perform linear regression, we use the method called least squares method where we are minimizing sum of square residual error and that SSR is nothing but yi minus yi hat where yi is coming from data and yi is the regressed value that means what I have got by regression and summing over all the data points.

So, we are minimizing SSR and if you do the minimization as it is optimization problem, you will get the relationship, the solutions are m is given by this relationship and c is given by this relationship. You can use these relations. Otherwise, if you use linear algebra you get the same solution but in a different format. Here you get m, the m and c vector will be equal to the inverse of x transpose x into x transpose into y. So, that is all for this first lecture on linear regression. We will continue to the next lecture, where I will discuss about some diagnostic check for our regression that we have just done now.

(Refer Slide Time: 23:36)



$$g_2 = ag_1 + b$$

g1	g2
1.25	2.25
1.32	2.5
1.62	2.5
1.97	3.5
2	3.4
2.2	3.5
2.3	3.6
2.48	4.25
2.6	4.15

But before I leave you, I will leave you with a problem to solve. In this case, I do not recommend you to use r to solve this problem. We will have a separate lecture where we will learn how to use r to do perform linear regression but for this data set, you can use a different approach. What I want you to do, for this particular data set where I have data for two genes and I believe there is a linear relation between g2 and g1.

You find the linear model that means you fit this data to a straight line, linear equation between g1 and g2 using means, means of these two data, variances and covariance. You can do it by paper and pen, you can do it on excel, you can use r to calculate mean and variance and covariance and then think about it which formula you will use to calculate the value of a and b. I hope you will be able to do this job yourself and let us meet again for the next lecture on linear regression where we will do some diagnostic check for our linear regression model. Till then, happy learning.