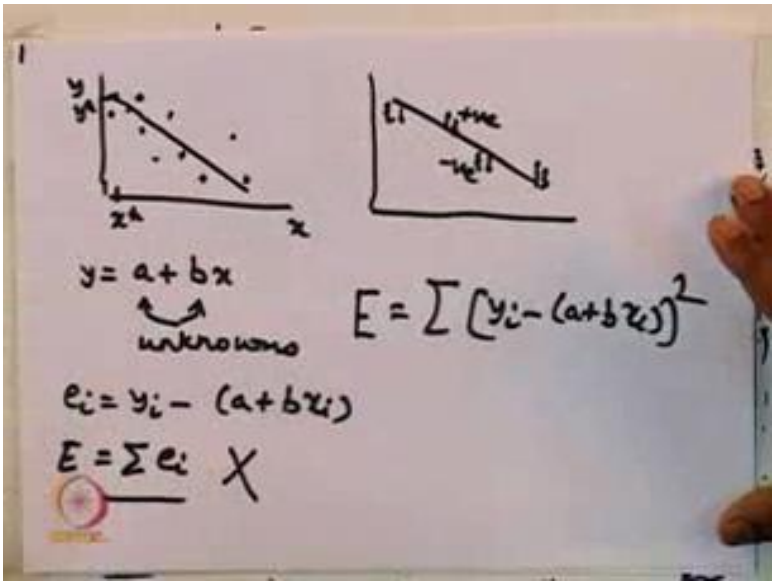
Introduction to Biostatistics
Prof. Shamik Sen
Department of Bioscience and Bioengineering
Indian Institute of Technology, Bombay

Lecture – 12
Correlation and Regression Part-II

Hello and welcome to today's lecture. We will as always we will start with briefly recapping, what we discussed in last class. We had been mostly discussing about correlation and regression.

(Refer Slide Time: 00:33)



In case of bivariant data; so, when you have x and y and you want to see that if you have some friend, can you find out a line, which gives a way of interpolating or extrapolating to determine y at some unknown location let us say. I can find out given this particular x star, I want to know; what is the value of y star. So, the way to do it is from the trend itself. You want to find out this equation of the line right. So, what we do? So, in this particular case, when it gives us the impression that there is some linear trend; we fit these data points using a line which is y is equal to a plus b x.

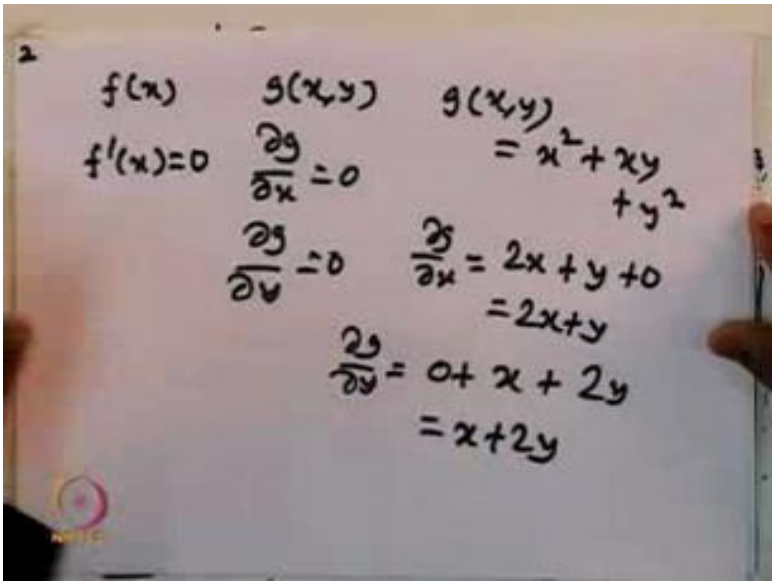
So, clearly this line has equation of line has two unknowns. These are your two unknowns and you want to find out, what is the value of a and b? If you had taken, if your points were two; if you had only two points, then a and b will be uniquely determined, but when you have multiple number of points, then there are infinite

solutions to a choice of a and b and that is what brings us to linear regression. How do we go about finding out the equation of a line which we think is a best representative of this data? So, one of the strategies which is followed is to minimize. So, if this is the line that I am drawing and these are our points, I minimize the deviation, I minimize the sum of these deviations from this line.

So, in other words and I had previously discussed that if it is deviations, so deviations are simply y minus a plus b x i. So, this is you can say, E i equal to y i minus a plus b x i. Now if you simply do total E, as summation E I, this will underestimate the true error because there are some errors which are positive. So, when y is greater than a plus b x i. So, here you have positive error, here you have negative error. So, these will cancel each other, hence this is not the approach to take.

So, what you minimize is E is equal to summation of y i minus a plus b x i whole square. Square means you are minimizing the net deviation without taking into consideration whether it is positive deviation or negative deviation and we had briefly discussed that as opposed to a function of a single variable where, you take the f prime or the.

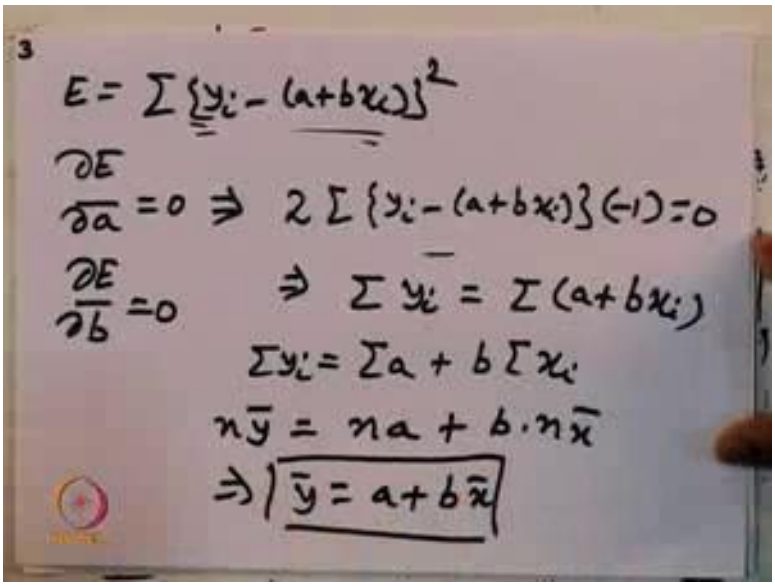
(Refer Slide Time: 03:22)



So, if you have f of x. So, finding out what value this is you know minimum you set f prime of x equal to 0. So, when you have g of x comma y, what you do is, you set g del x equal to 0 and del g del y is equal to 0.

So, once again if g x i is let us say equal to x square plus x y plus y square then, del g del x is going to be 2 x plus y because if you are taking the derivative assume y is constant plus 0 is equal to 2 x plus y and del g del y will be is equal to 0 plus x plus 2 y, x plus 2 y. So, you said both of these equal to 0.

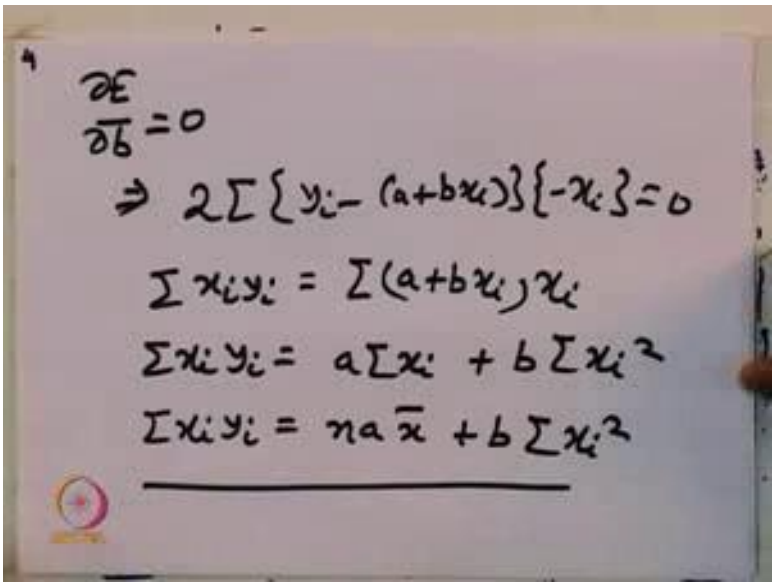
(Refer Slide Time: 04:18)



So, in our case the equation is, E is equal to summation y i minus a plus b x i whole square right. So, I will set del E del a is equal to 0 and del E del b is equal to 0 because these are the unknown variables, this is what we want to find out. So, if I use this particular expression, I can right down 2 minus 1 equal to 0. So, I essentially take the derivative with respect to this entire term, this is of the form z square. So, when I take a derivative of z square I get a form of 2 z. So, 2 and this is the z and then of this if I take a derivative with respect to a. So, del y i del a is 0, del minus a del a is equal to minus 1 and del b x i del a is 0. So, I get a value of minus 1. So, this equation basically will give you the expression summation y i is equal to summation of a plus b x i.

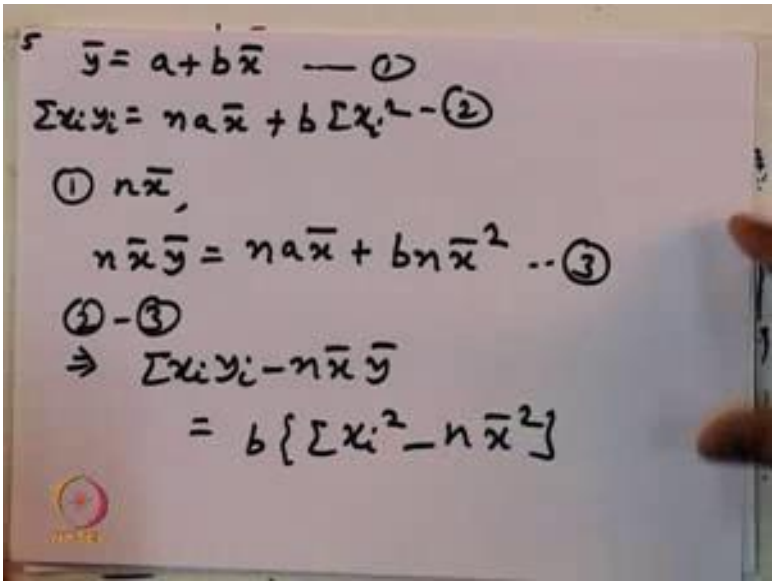
So, I can write summation y i is equal to summation a plus b summation x i. So, summation y i is basically n times y bar summation a is n times a and summation b is n times x bar. So, this gives me the equation y bar is equal to a plus b x bar, this gives you the equation y bar is equal to a plus b x bar, I can then.

(Refer Slide Time: 06:00)



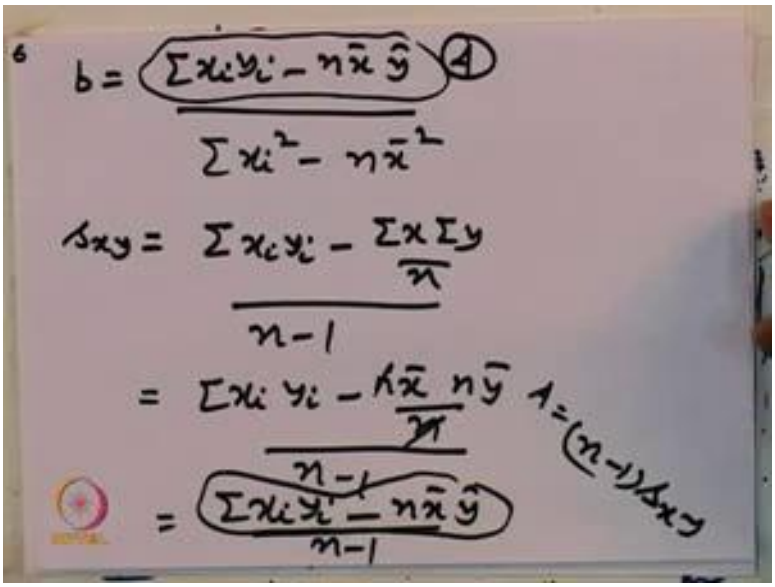
So, let us do the other expression, del E del b equal to 0 would imply 2 summation first part remains the same is only thing is now I have to multiply with this taken derivative of this with respect to b which is nothing but minus x i. So, I can simplify it I can get rid of 2 n minus. I can write it as summation x i y i is equal to summation of a plus b x i into x i. So, I can simplify it further, is equal to a summation of x i plus b summation of x i square. So, I can further simplify it, x y is equal to n a x bar plus b summation x i square.

(Refer Slide Time: 07:09)



So I have two equations. So, let me write down the final equations, I have y bar is equal to a plus b x bar and summation x i y i is equal to n a x bar plus b summation x i square. So, let us say this is my equation 1, this is my equation 2. I have two equations in two unknowns which are a and b respectively. Now can I eliminate 1? So, I can multiply first equation by n x bar and let us see what we get? So, I have n x bar y bar is equal to n a x bar plus b n x bar square. So, let us say this is equation 3. So, if I deduct then, 2 minus 3 would imply summation x i y i minus n x bar y bar is equal to b summation x i square minus n x bar square. So, I can uniquely determine b, the expression for b becomes.

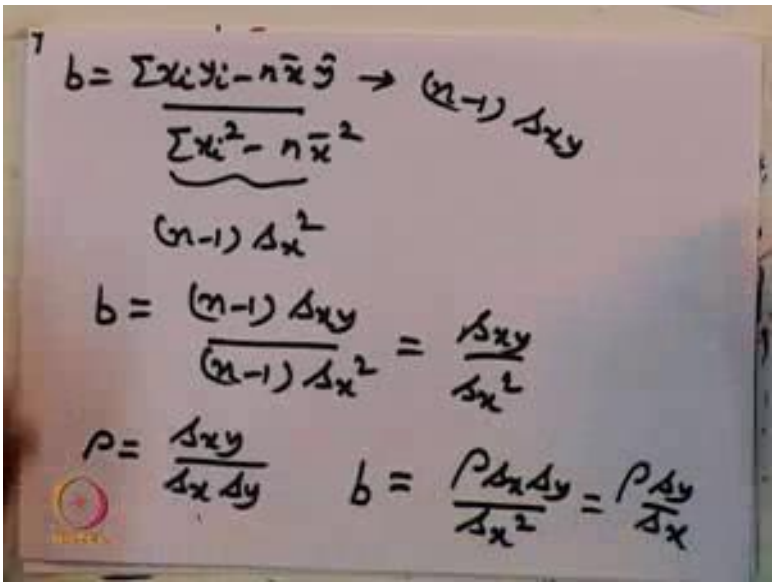
(Refer Slide Time: 08:19)



So, b becomes summation x i y i minus n x bar y bar y summation x i square minus n x bar square. Now does it, can I compare it with some of the definitions of covariance of correlation coefficient? Let me write down the expression for covariance that we determined s xy is equal to summation x i y i minus summation x summation y by n whole divided by n minus 1.

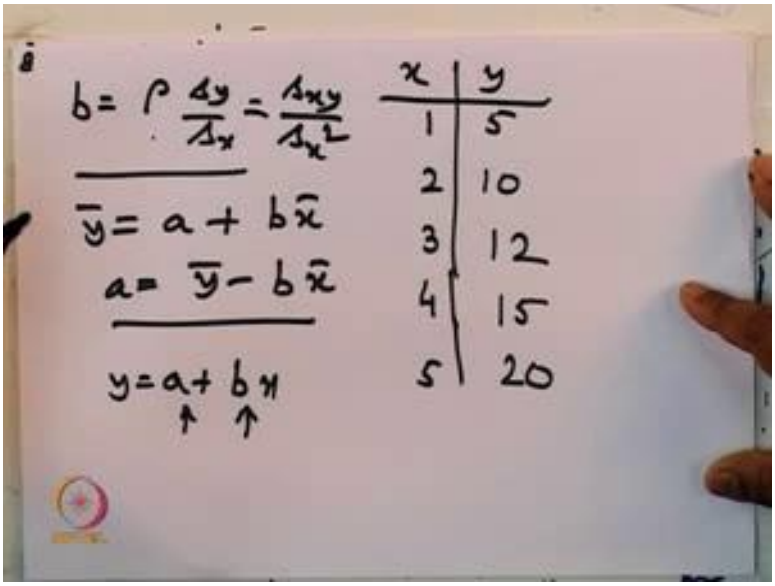
So you can clearly see. So, I can rewrite this equation. So, I can write summation x as n x bar, summation y as n y bar and by n by n minus 1. So, I can cancel each other out. So, I simply get, x i y i minus n x bar y bar whole divided by n minus 1. In other words you see that this whole term is nothing but this whole term here. So, you can clearly get an idea that this is about. So, this term is nothing but n minus 1 into s xy. So, if I call it term A. So, my A is nothing but n minus 1 to s xy.

(Refer Slide Time: 09:51)



So, let us see what is the zero? So, again I will write down the expression for b, this summation x i y i minus n x bar y bar by summation x i square minus n x bar square. So, what is this? This is nothing but n minus 1 into s x square, this n minus 1 into s x square. So, I can and this found was n minus 1 into s xy. So, my b is nothing but n minus 1 into s xy by n minus 1 into s x square is equal to s xy by s x square. Now I know that my correlation coefficient is defined by s xy divided by s x times s y. So, my b can be determined to be rho, so s xy I can write, rho s x s y by s x square. So, then b is nothing but rho into s y by s x.

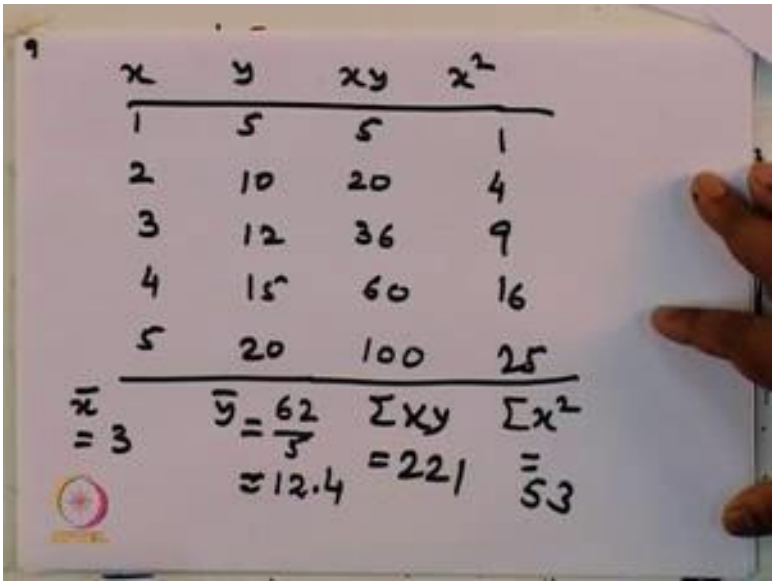
(Refer Slide Time: 11:01)



So I can determine my b is equal to rho times s y by s x. So, b is the coefficient, the second coefficient I have determined and I have the other equation y bar is equal to a plus b x bar. So, a is nothing but y bar minus b x bar. So, I can determine once I know b I can determine what is a. So, let us work out a specific case and see what value we get for x and y. So, let us take a very simple example 15. So, these are our four values or let me make it 5, 5 is 20. So, this is my x and y data and I want to find out y is equal to a plus b x and what are our values a and b respectively. So, what I need to do? So, as per this equation b is rho times s y by s x.

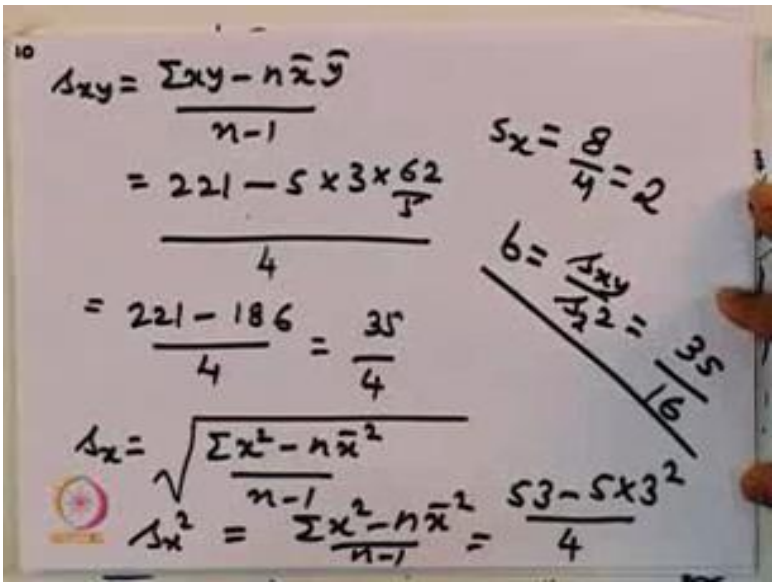
So, I are or I can also write s x square s xy by s x square. So, I need to find out s xy and s x square respectively. So, I will write down these values again.

(Refer Slide Time: 12:27)



So, I have x, I have y; 1 2 3 4 5; 5 10 12 15 20. So, for calculating s xy, I need x y, I also need x square. So, my x bar is equal to 3, y bar is equal to 25 (Refer Time: 12:56) 50 60, 62 by 5 is approximately. So, it is exactly 12.4. Let me calculate x y, it is 5, it is 20, it is 36, it is 60, this is 100. So, x y become 80 180 (Refer Time: 13:24). So summation xy becomes 221, let me just cross check again 180 221, x square is 1 4 9 16 25. So, summation x square is equal to 6 40 53.

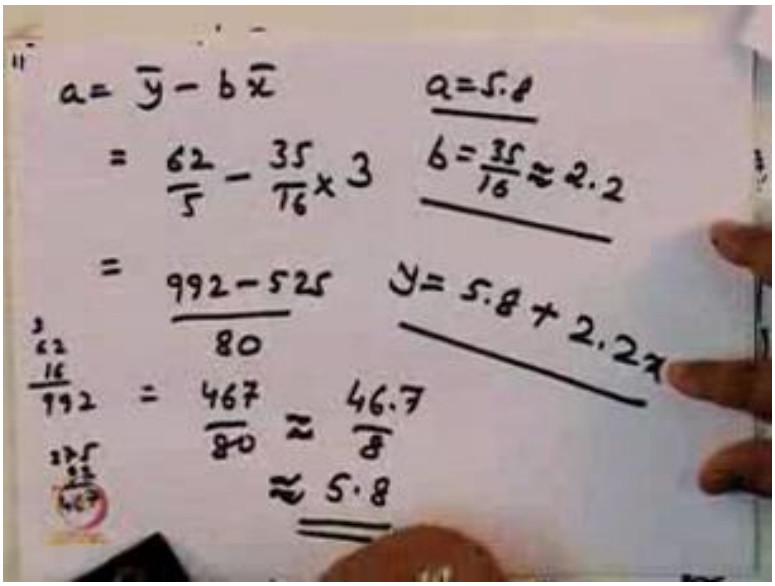
(Refer Slide Time: 14:07)



So, this is summation xy. So, I can find out the expression for s xy. So, s xy is equal to summation xy minus n x bar y bar by n minus 1 equal to. So, summation xy is 221 minus n is 5, x bar is 3 into y bar is 62 by 5 whole divided by n minus 1 is 4 equal to 221 minus 5 (Refer Time: 14:34) 62 into 3, 186 by 4; 221 is 35 by 4 and s x is equal to square root of. So, s x is equal to square root of summation x square minus n x bar square by n minus 1.

So, s x square will be summation x square minus n x bar square by n minus 1, summation x square we found out was 53 minus n is 5 into 3 square by 4. So, I can find the value of s x is equal to, so 9, 45; 8 by 4 equal to 2. So, s x is equal to 2, s xy is equal to 35 by 4 my b comes out to be s xy by s x square is equal to, so 35 by 4 into 4, 60. So, b comes out to be 35 by 60 and a.

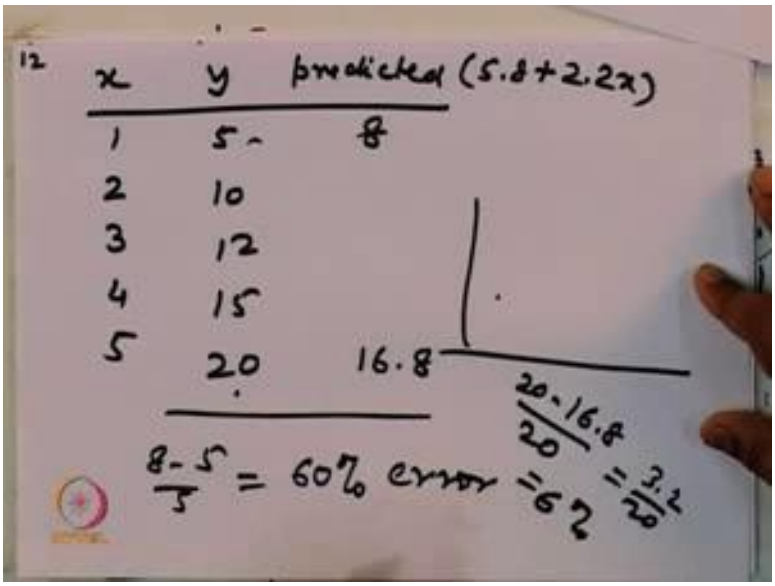
(Refer Slide Time: 16:00)



a comes out to be y bar minus b times x bar is equal to y bar is 62 by 5 minus b is 35 by 16 into x bar is 3. So, I cannot cancel in anything, (Refer Time: 16:18) 80; 62 into 16, so (Refer Time: 16:23) also 99. So, 992 minus (Refer Time: 16:30) 467 by 80 is approximately to 46.7 equal to 46 by 7. Let us say, we approximate it equal to 8 into 46.7, 8.

So, I say a is 5.8 and b. So, a becomes 5.8 and b becomes 35 by 16, 2 is a 32, 30 approximately 2.2. So, your final expression becomes 5.8 plus 2.2 x. So, if we substitute x is equal to, so we have the value of x is 1. So, x is 1 I see.

(Refer Slide Time: 17:43)

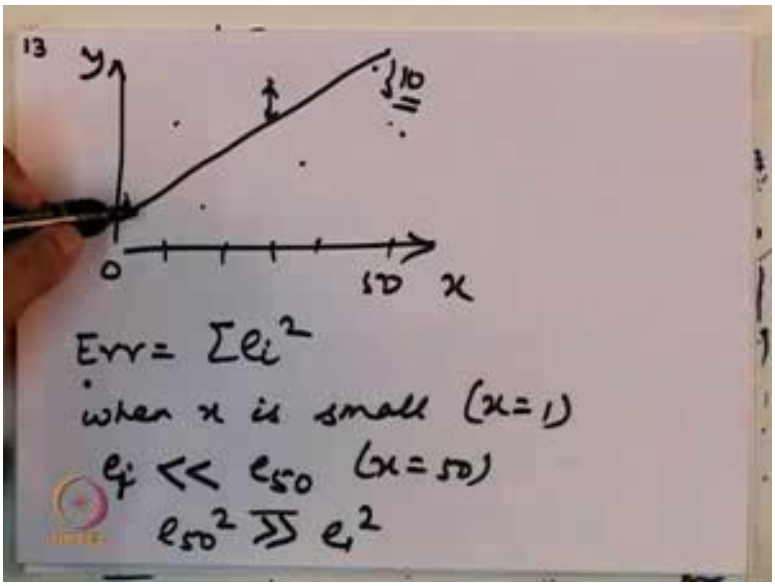


So, let me again recalculate all our values; x y and then predicted. I had 1 2 3 4 5, my y values are 5 10 12 15 20. So, my predicted equation is 5.8; 5.8 plus 2.2 into x.

So, what you see here x equal to 1 is 8, x equal to 5 is 11 is 16.8. So, what you see here is in the way our points are. So, you are if you see the percentage of error. So, this raises one important point, if you see the percentage of error. So, here you are off by 8 minus 5 by 5 is equal to 60 percent error. In here your error is 16.8, so roughly 20 minus 16.8 by 20 is equal to 3.2 by 20, which is only 6 percent.

So, this is this brings us at, you know it raises a very important point; when you mark these errors, as y increases.

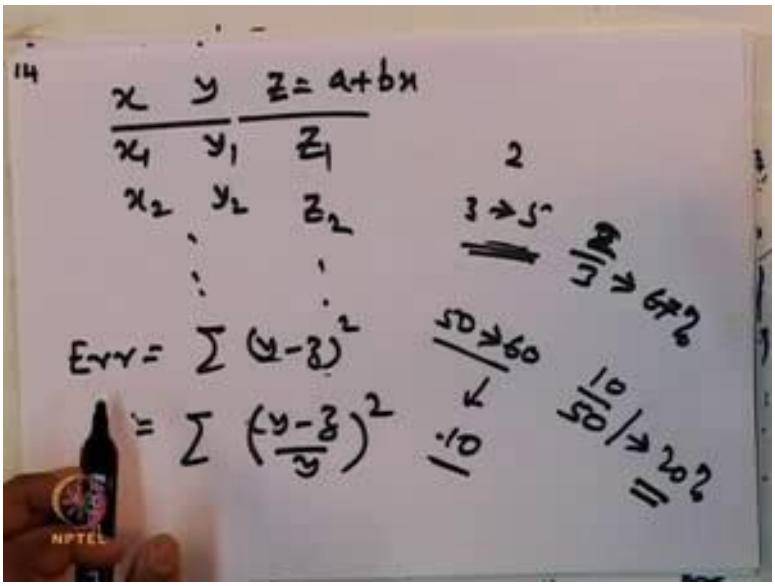
(Refer Slide Time: 19:11)



So, what you see? So, let us say in our particular case my x varies from 0 to 50. So, when I and you know let us say these are my points, when I am here this error is actually insignificant. So, if you count the magnitude of the error right. So, my error, total error I define as summation e i square. So, when x is low, when x is small say let us say x is equal to 1. So, my magnitude of error e 1 or e 1 is going to be way insignificant compared to let us say e 50, that is when x equal to 50. So, here instead of 3, you are getting a value of let us say 5 or 6.

So, your error is 2 or 3, but instead of 50, you are getting a value of 60 or 70. So, that error the proportion of error is so, this is 10 let us say, this is off by 10, but this off by 10 has a much significant contribution. So, when you write your error expression. So, e 50 square is going to be significantly greater than e 1 square. So, this would tell you that for higher values your estimate is going to be better than for lower values. So, how do you tackle this problem? One way to tackling this problem is actually. So, one way to tackle this problem is.

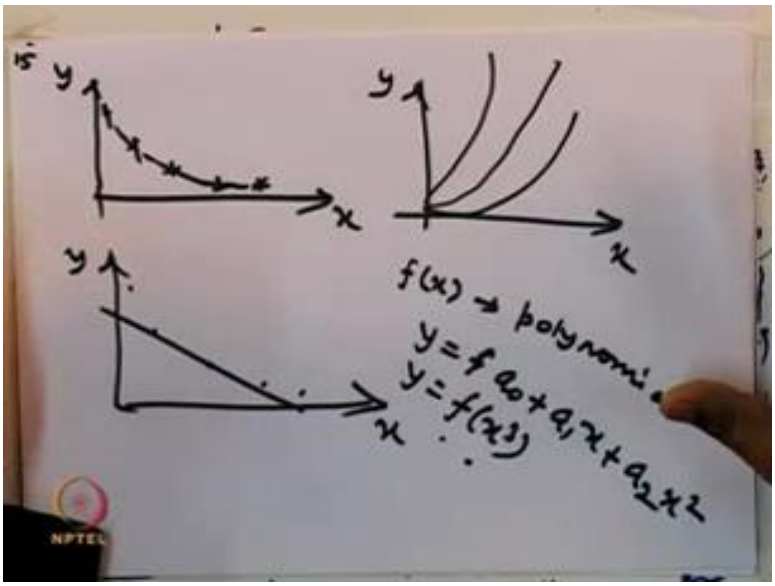
(Refer Slide Time: 20:51)



Let us say you have x y, x 1 y 1, x 2 y 2. So, you calculate your let us say z is equal to a plus b x, you define z 1 z 2 z 3 so on and so forth. So, instead of defining your error as simply summation y minus z whole square, you can define it as summation of y minus z normalized to y whole square. So, what this will ensure is it normalizes the magnitudes of the errors as well. So, between 3 it going up to 5, you have a huge amount of error, but when you normalize by 5 by 3 you get a much lesser. Similarly when you go from 50 to 60 right, your actual error is 10, but 10 normalized by 50 will give you a value which is comparable to this value. So, you here your error is 2 by 3. So, this is 67 percent and here, it is 20 percent.

So, earlier you are comparing this jump 2 with this jump 10. So, in that case your data will be much better fit in lateral portion because this error is getting a much greater weightage in this expression. So, this is one way to normalize. So, of course, in this case you cannot do it by hand, but you have to depend on programming to write down the appropriate code.

(Refer Slide Time: 22:29)

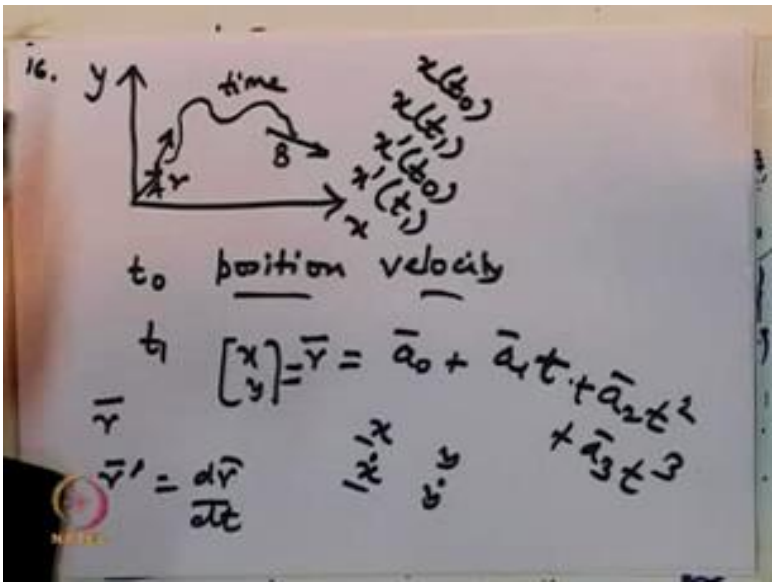


So, one more thing I wanted to point out. So, of course, let us say in the best case situation your data has some linear trend and it is easier to in a fit a linear curve, but in the more generic case let us just say which you will see is very. Let us say you have x and y which are related as follows. So, I am actually drawing the curve and not the exact points, your points you can draw those points on top of each other, but what you see is the curve is not just a linear line.

So, if you fit this data with a linear line so, your curve will look something like this, but this does not capture the essence of the phenomena you are trying to study. So, it is not always beneficial to just use a line to fit your data. So, you take a look at your data and see what information it conveys. So, for example, right if your data is like this right, you can clearly see that there is a non linear increase in y with x. What kind of a function should you use? One way is to use a polynomial right, you can choose a polynomial.

Now, depending on how fast this curve is rising, you can choose y is equal to f of x square. So, let us say you can have a naught plus a 1 x plus a 2 x square or you have a function of x cubed. So, you can write function of x cube also, but the nature of the rise how fast the rise is will dictate what kind of a polynomial you will use. Now let us take one more example.

(Refer Slide Time: 24:16)



So, let us say you are designing a robot which goes from two points. So, this is my x y. So, two points and at each point, it is going in this particular plane from point A to point B at different times and you want. So, you know at this particular point, it has a certain velocity here and at this point it has its velocity in a different direction right. So, what you, so this is the x y plane, but this is the actually A is a function of. So, this is let us say a trajectory as a function of time right. So, this is time axis, as a function of time you are moving along x y axis.

So, you are given two pieces of information at time t naught and time t 1, you are given the position and the velocity. So, it tells you that there are four conditions being prescribed. You are given the position and the positioned way. So, if I were to represent this as my r. So, I know what is r vector? And what is r dot or r prime, which is equal to d r by d t which is a vector, this is the direction. So, what kind of a function can I choose? So, there are four unknowns right, let just say for each trajectory let us say if I were to assume a trajectory as a naught.

Let us say I assume this particular function. Now I write them as a naught a 1 a 2 a 3 because it is vector. Your r position stands for x position and y position so, but what you see is. So, what you are provided at every time is either the x or the x dot; similarly the y or the y dot. So, you have two conditions right. So, from this depending on the number of variables you can choose. So, in this particular case you can actually find out the exact

solution if you know the position and you know position at two different points and the speeds at two different points. So, I can use x at t 0, x at t 1, x at x prime at t 0 and x prime at t 1 to find out. So, I can put four equations and if you have four independent equations in four unknowns you can uniquely determine. So, that gives you an idea of what kind of function you can use to fit it.

So, linear is only one particular case within the generic case you can you need to fit with different kind of functions. So, that brings our lecture today to a close. We discussed about regression, worked out some examples of how you would find out regression and we showed that using one particular example, we showed that the way how we calculate the error which is the square of the deviations. Then for low values of x your deviation are small, hence it is it gets less represented in the overall definition or overall computation of the error. So, in our way of eradicating it is to normalize this error with respect to the value that you are doing. So, then you relatively give equal weightage to each value.

With that I thank you for your attention and I look forward to meet you again in next class.

Thank you.