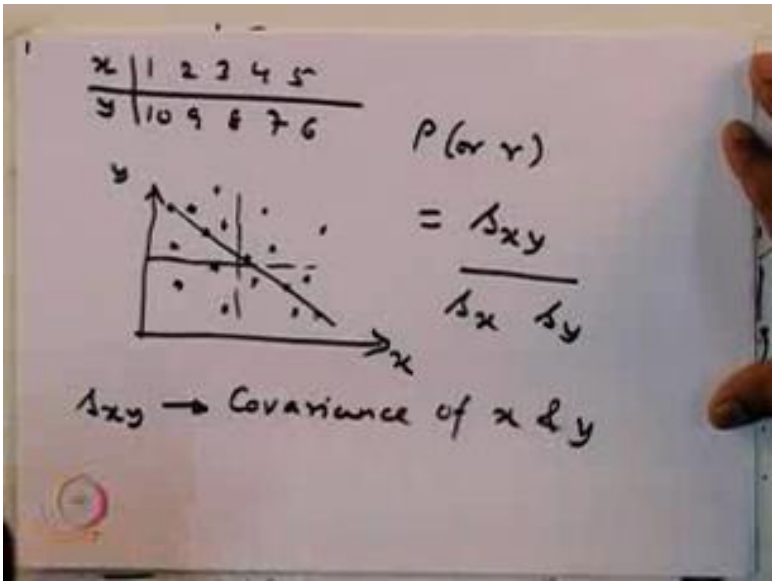
Introduction to Biostatistics
Prof. Shamik Sen
Department of Bioscience and Bioengineering
Indian Institute of Technology, Bombay

Lecture – 11
Correlation and Regression

Hello and welcome to today's lecture. In the last lecture, we had a brief recap of R and how you can enter vectors in R and do basic calculations or basic operations with vectors and also we saw how you can use R to calculate descriptive metrics including me in media and standard deviation so on and so forth. Towards the end of last lecture, we started discussing about bivariate data, and how you can say something about correlation between 2 different variables.

(Refer Slide Time: 00:53)

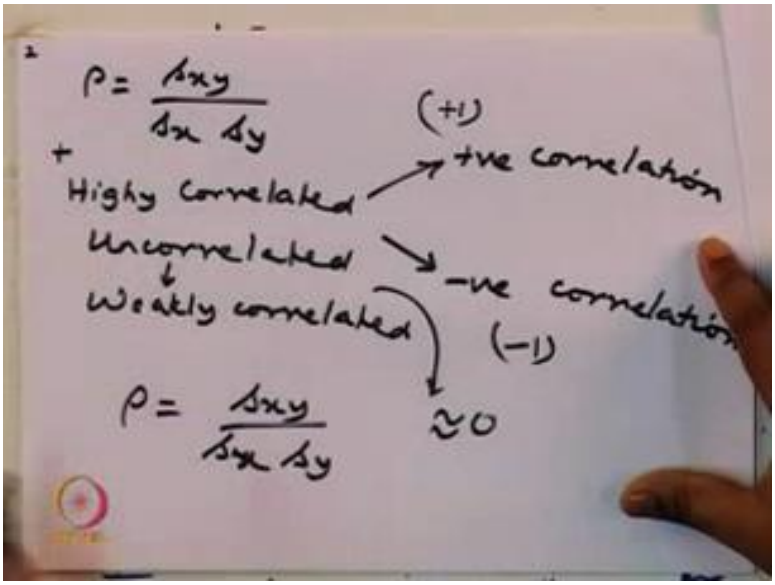


So, let us take a look at what you mean by bivariate data. Let us say we have 2 variables x and y and we can have various values let us say 1, 2, 3, 4, 5, 10, 9, 8, 7, 6, let us say this is one example of x and y. So, if you plot this data in the x and y plot. So, what I have is something which. So, this is 1 and 10, 2, 1, 9. So, you have these 5 points; let me make the points darker. So, these 5 points, you get the feeling that there exists and correlation or an inverse correlation that is with increase in x there is a decrease in y and how do you and let us just say. So, this is a very good data you can clearly see that there

is a clear line which passes through all these points, but in the mode generic case let us say you have points.

This is more likely to be a data now in this case what kind of a line will you draw will you line a draw will you draw a line which is like this will you draw a line which is like this will you draw a line which is like this. In order to go ahead with these kinds of third processes what we do is we find out something called correlation coefficient and then based on that we go ahead to a concept called regression. So, let us revisit what we did about correlation coefficients. So, we define the correlation coefficient rho or R and this is given by the expression is equal to s xy by s x times s y. So, s xy refers to the covariance of x and y. So, if you see the symmetry.

(Refer Slide Time: 02:59)

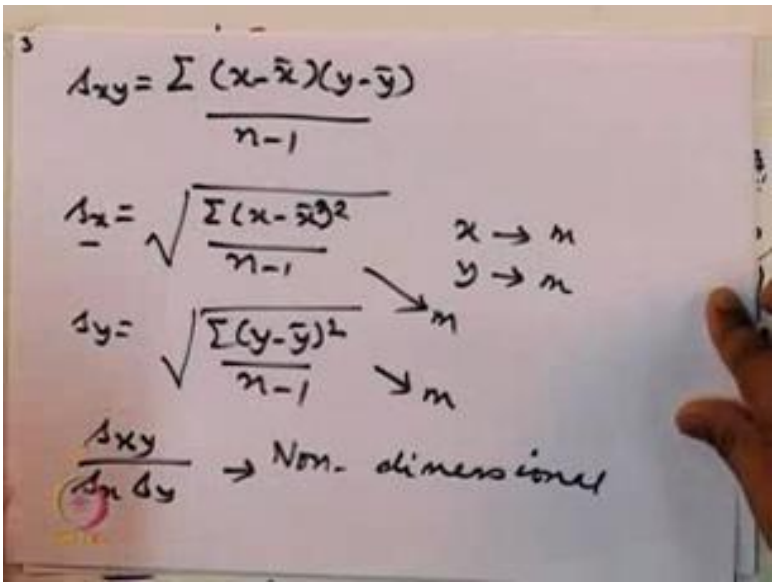


So, if you have s x, so we know, so, we have R or rho defined as s xy by s x comma s y s x product s y. Now rho is a is a measure of correlation. So, what would ideally with the values of flow; you can say something as highly correlated or uncorrelated or weakly correlated uncorrelated or weakly correlated and within high correlated you can have positive correlation or negative correlation. So, in the example we showed in the first slide. So, this is an example the original data set is an example of negative correlation where increase in x leads to a decrease in y.

So, if we see the way rho is defined; rho is defined as s xy by s x by s y. So, ideally we should have when it is highly correlated, we should give a value which is positive. So,

highly positively correlated should give me a value which is closer to plus 1 negative correlation should give me a value which is closer to minus 1 and uncorrelated should give me a value approximately equal to 0. So, if we see our definition of rho and ask, how do we define the contra the covariance?

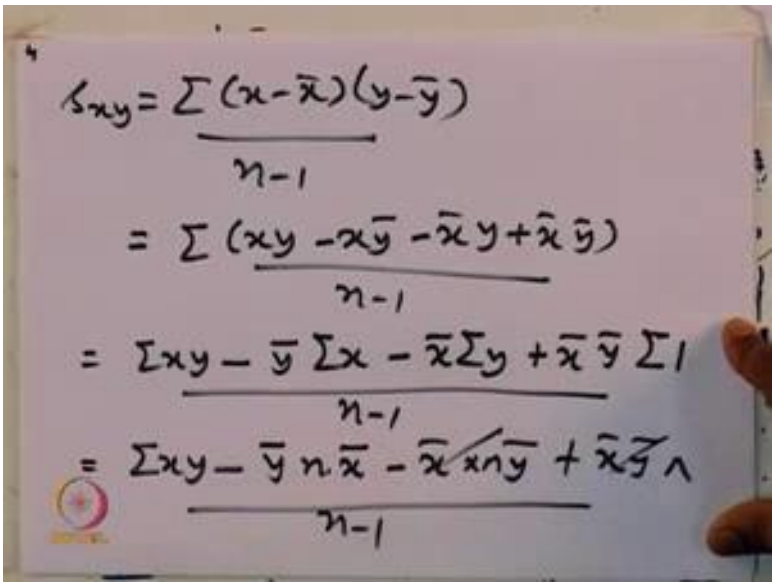
(Refer Slide Time: 04:38)



So, covariance is defined as. So, s xy is defined as summation of x minus x bar times y minus y bar whole divided by n minus 1. So, what you clearly see missing is the square root. So, if I write down the corresponding expression of s x i have square root of x minus x bar square whole square by n minus 1. So, in the definition of covariance this square root is missing and the reason is obvious. So, you are multiplying s x by s y which has square root y minus y bar whole square by n minus 1.

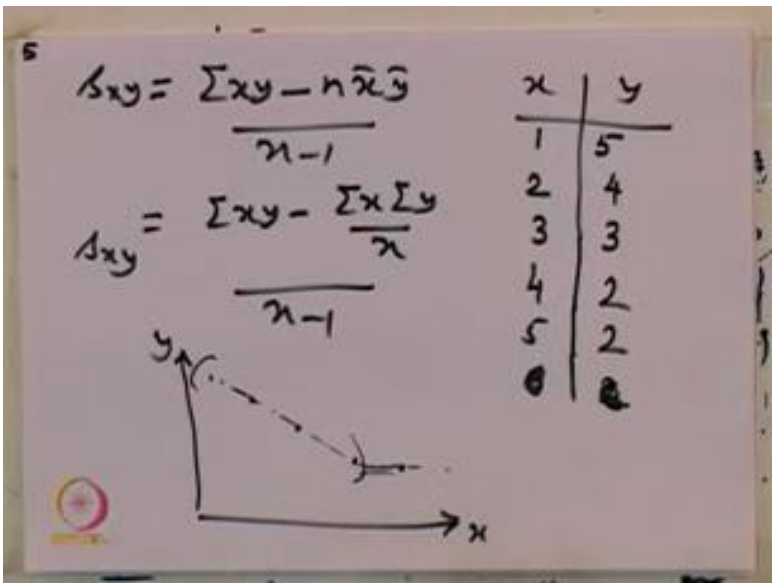
So, if x has units of meters let us say then and why has units of meters let us say, then x y. So, this will give me a value of meters this will also give me a value of meters and this; the top portion should I also give me a value of meter square so that s xy by s x by s xy, so, s xy by s x times s y should be non dimensional and this is why you have this the root does not exist in s xy. So, it is possible we can simplify s xy.

(Refer Slide Time: 06:01)



We can simplify s xy by writing summation of x minus x bar into y minus y bar by n minus 1. So, I can simplify this to summation of x y minus x by bar minus x bar y plus x bar y bar whole by n minus 1 summation xy minus I can take y bar out summation x minus I can take x bar out summation y plus x bar y bar summation 1 whole by n minus 1. So, this I can write as summation xy. So, y bar is. So, I can have y bar and summation x is to n into x bar; similarly x bar into n y bar plus x bar y bar into n by n minus 1. So, this and this cancel out and what you are left with.

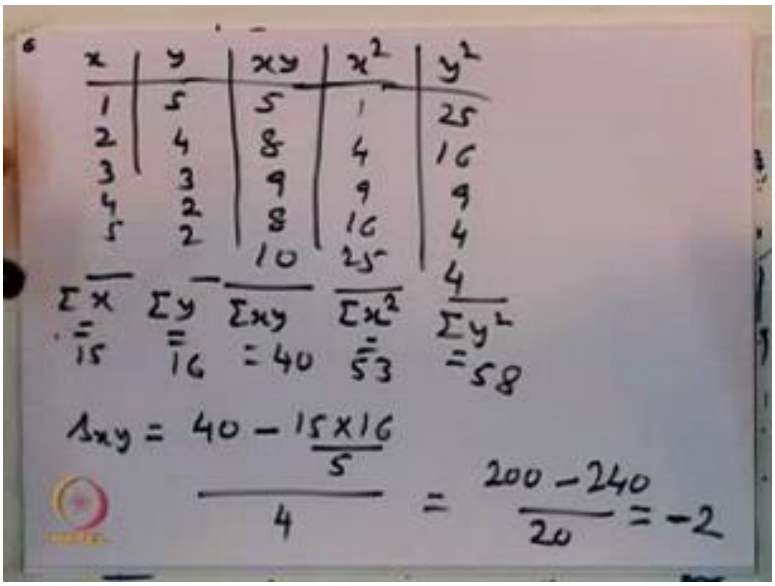
(Refer Slide Time: 07:06)



What you are left with is the expression s xy is equal to summation of x y minus n x bar y bar by n minus 1. So, I can rewrite this equation as; so, summation xy minus summation x summation y by n whole divided by n minus 1. So, this is a simplified expression for calculating s xy. Of course, when you are doing these calculations in a laptop or in a program then you can use the original formula to calculate the final expression, but when you are doing it by hand then this is a easier way to calculate. So, let us take one sample example and work it out to find out what is the expression for s xy. So, let us take x and y.

So, let us say, so you have the following values of x and y. So, so let me remove the last one for simplicity, let me leave the last value you have 5 values of x and 5 values of y. So, I want to calculate the value of s x, s xy and I want to calculate the value of correlation coefficient. So, first before doing the calculation let us see how the values look. So, 5, 4, 3, 2 and then 2, you have these points which are coming down, but then there is a flattening beyond this point. So, had I plotted 6 and 2 then it would have also some lied somewhere here. So, let us see. So, what I can clearly see is towards the initial part there is a very clear you know negative correlation and towards the latter side it is kind of flattish. So, let us calculate the exact expression for s xy.

(Refer Slide Time: 09:09)

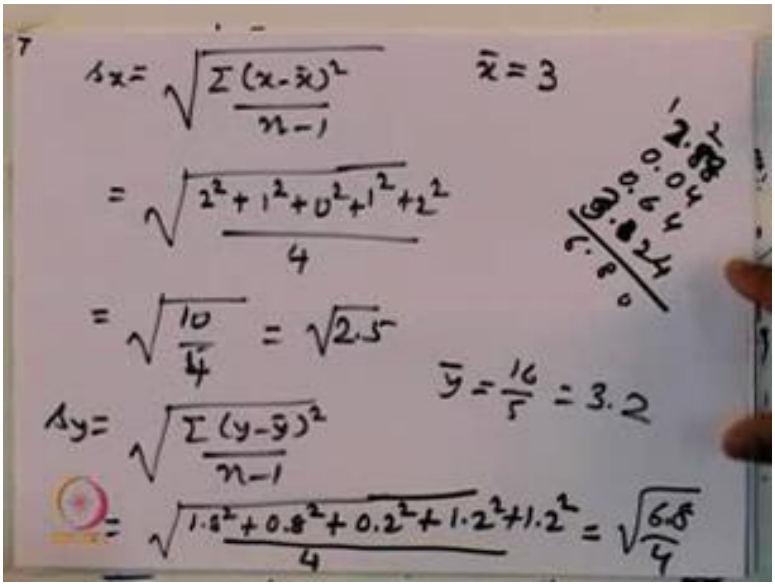


I will again write down the values of x 1, 2, 3, 4, 5 x is y is 5, 4, 3, 2, 2 I can write x y i can write x square and y square I can calculate these values 9, 8, 10, 1, 4, 9, 16, 25, 25,

16, 9, 4, 4. So s, summation of x y is equal to 27 plus 13, 40 summation of s x square equal to 5, 6, 40 and 31, 53, summation of y square equal to 9 and 17 and 33, 58, summation of x of course, y is equal to 16. So, my s xy is given by summation xy which is equal to forty minus summation x is 15 into summation y 16 by n which is 5 whole divided by 4.

So, you see a value. So, this I can simplify 15 square is to 25, 240 by 20 40 minus 2. So, s xy is giving me a value of minus 2.

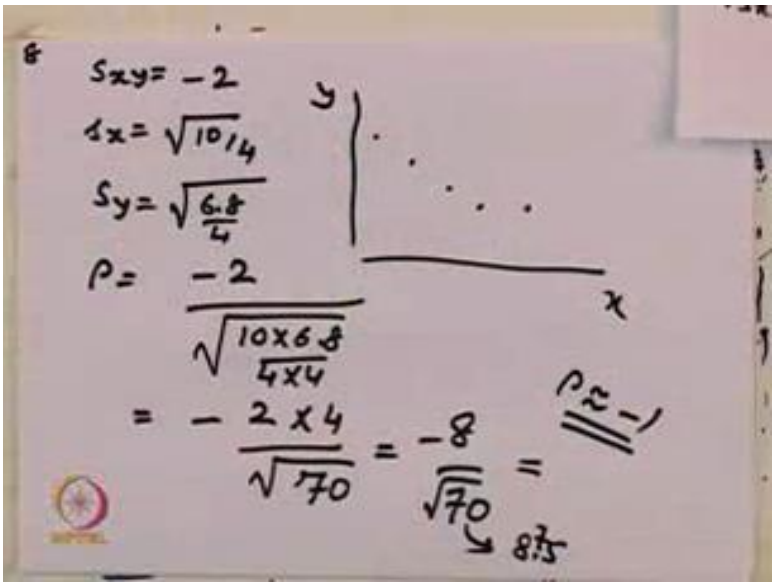
(Refer Slide Time: 11:22)



I can go on to calculate the value of s x, s x is equal to square root of summation of x minus x bar whole square by n minus 1 is equal to root of; so, in this case x bar, so x bar is equal to 3. So, 2 square plus 1 square plus 0 square plus 1 square plus 2 square whole divided n minus 1 is 4 into square root of 4 plus 4 [FL] 10 by 4 equal to root of 2.5. I can calculate the value of s y is equal to root of y minus y bar whole square by n minus 1 y bar is equal to 5 plus 16 by 5 is equal to 3.2.

So, s y 1.8 whole square, let us take a different pen 1 plus 8 whole square plus 44; 0.8 whole square 3.2 whole square plus 2; 1.2 whole square plus 1.2 whole square whole divided by 4 roughly 3.24. So, I can calculate this as. So, 2.4, 4.44, this 2.88 plus 0.04 0.64 0.324, sorry, 3.2, 6.8, so, root of 6.8 by 4 is square root of 1.7.

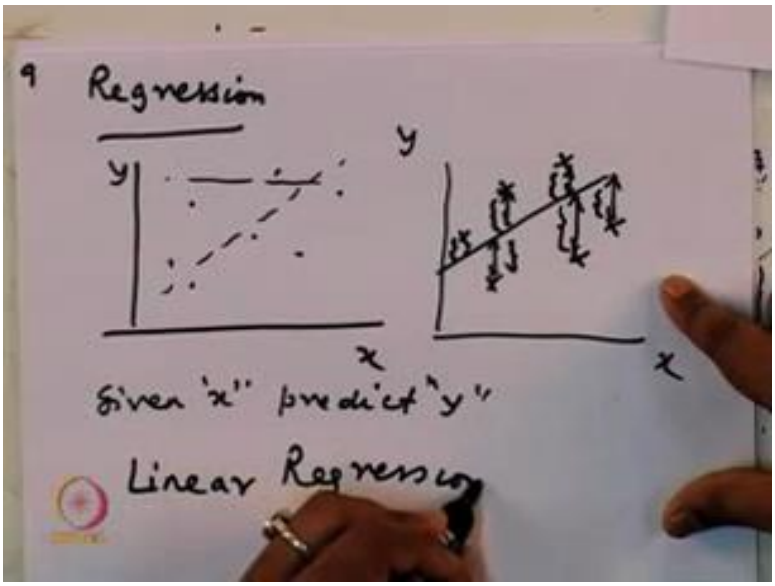
(Refer Slide Time: 13:34)



So I can write s xy is equal to minus 2 s x equal to square root of 10 by 4 s y is equal to square root of 6.8 by 4. So, my rho is equal to minus 2 by square root of 10 into 6.8 by 4 into 4 which is a minus 2 into I can have the 4 out by root of 70 equal to minus 8 by root 70 which gives me a slightly negative value which is some; this I can approximate as 8.5 let us say. So, I have the; it has to be greater than 8.75, I will get a value rho which is almost close to minus 1 and this is obvious from the data because the data looked something like.

But what you also this is my x axis? This is my y axis. So, what you clearly see is this correlation coefficient is negative; that means, that as you increase x y decreases now based on this can I come up with a way such can I come up with a way to predict that given a value of x what will be the value of y and that is what we do using regression analysis.

(Refer Slide Time: 15:09)

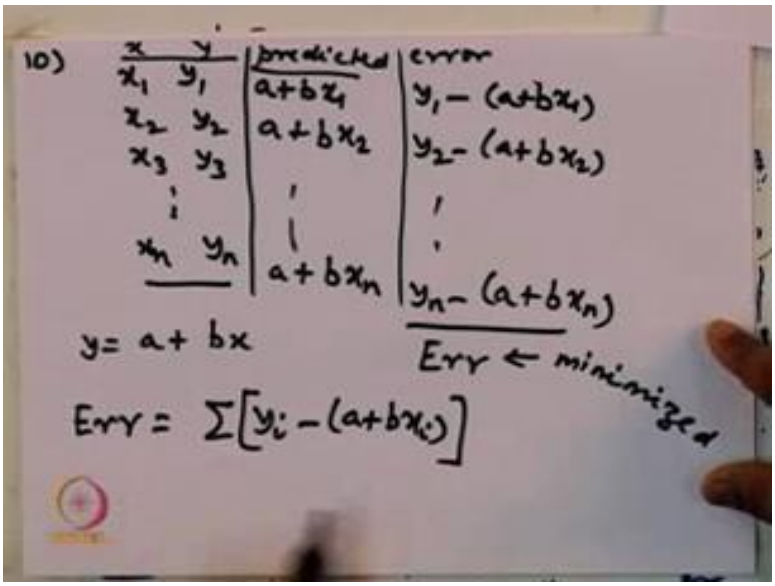


So, in regression what you do is let us say you have a set of data points x and y imagine these are your data points and you want to be able to predict such that given x predict why now if let us say one way the way I have drawn this data I can see that if I draw a line which is roughly like this then this might be a good way of predicting the value of y and I have intentionally drawn a line which somehow crosses through the middle of all these points I could have also drawn the points like this, but in this case what you can clearly see while this line will predict um y quite well in this regime of x, but in this regime of x there will be a huge disparity.

So, what is the method to madness what is done in linear regression is what you try to minimize see if you have x you have y and you have these points you draw a line such that the errors which accumulate for each of these points away from this line. So, for each point you can calculate how far is the estimate from the actual value? So, these are your actual values.

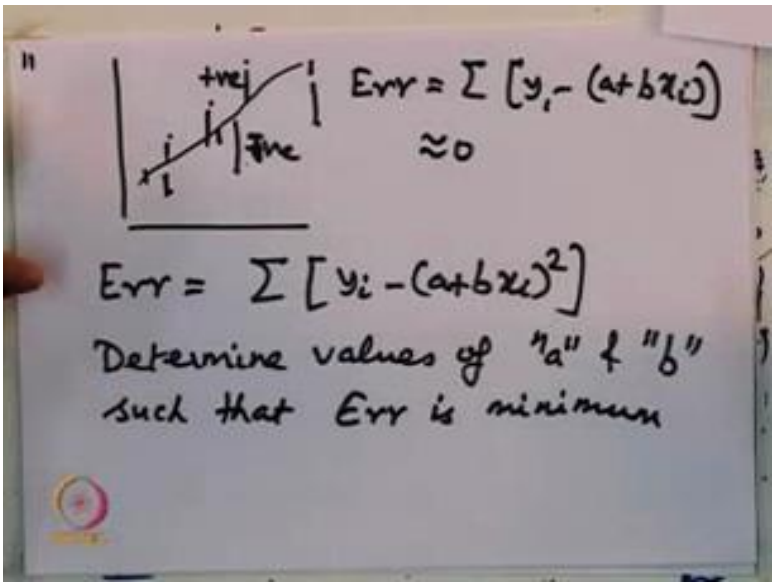
So these are your actual values and these are you know this is the predicted line. So, what you want to know is can I draw a line such that the sum of these numbers or these errors which accumulate from the prediction is minimized and that is what is convene called in lean? What is done in linear regression? So, let me reframe the problem statement let me reframe the problem statement.

(Refer Slide Time: 17:22)



So, you have, let us say x 1 y 1 x 2 y 2 x 3 y 3 n values x n. So, you have x n y n given in values can you draw line y is given by a plus b x. So, if; So, these are my x and y values and my predicted values if I were to draw this particular line as the estimate of the values I can write that at this particular value it is a plus b x 1 a plus b x 2 a plus b x n. So, what is the error that I accumulate from my prediction y 1 minus a plus b x 1 y 2 minus a plus b x 2 y n minus a plus b x n. Now let us say if I say that I want to assume I total error is what I want to minimize this has to be minimized. So now, if I were to define error as just summation of y i minus a plus b x i would this be a good metric and the answer to that is no because given your points.

(Refer Slide Time: 19:04)

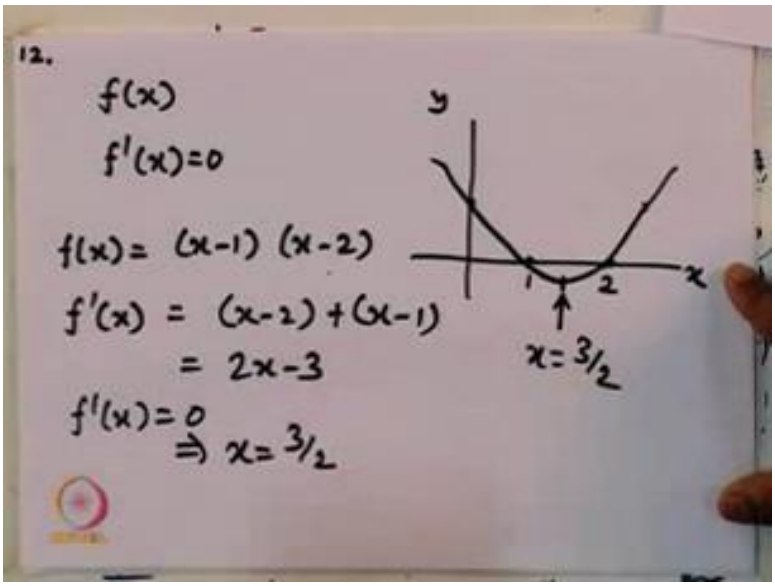


So, given your points, if you are giving your points and you were to draw this particular line. So, you can clearly see that in some cases this error will be positive in some other cases. So, this is negative, this is positive sorry, the other way around this is positive this is no, this is negative, this is positive. So, some errors will be positive and some errors will be negative.

So, this function error is equal to summation of y i minus a plus b x i if you actually do it, you might get a value of error which is close to 0, but this is not what we are minimizing, we are minimizing the sum of these; each of these deviations from the actual value. So, what we need to minimize is this particular function error given by. So, I take the square which means that what I count both whether it is positive or negative when I take the square it is something which contributes to the error. So, I want to minimize this function how do I go about minimizing this function.

So, there and you want to go at the end of the exercise what you want to do is shown to determine values of a and b such that error is minimum. So, this is what you want to minimize the error how do you do it. So, if you recollect. So, in class 12 standard calculus what you did was.

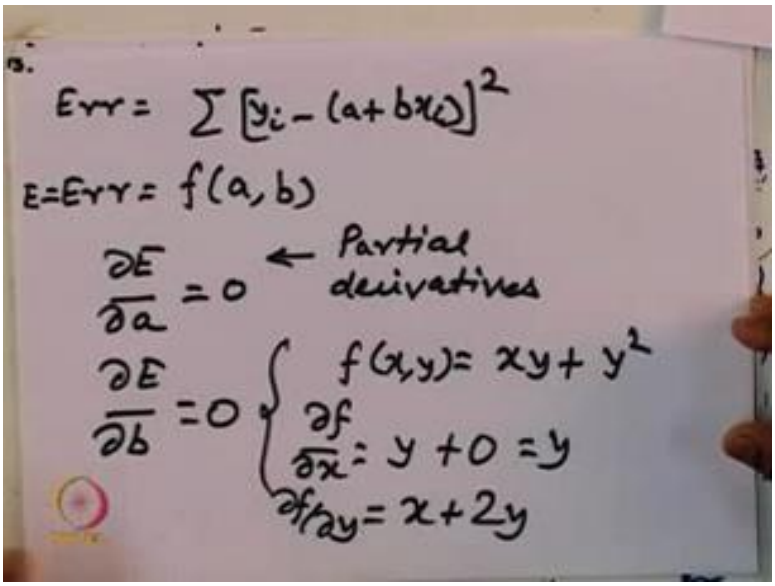
(Refer Slide Time: 21:01)



If you had a function f of x in order to minimize in order to minimize you said f prime of x is equal to 0. So, let us take a sample example let us say f of x is equal to x minus 1 into x minus 2, what is the value of f prime x is equal to x minus 2 plus x minus 1 is equal to 2 x minus 3 if I said f prime of x is equal to 0 then I get x is equal to 3 by 2. Now if I actually plot this whole function. Since x this is y let us say this is 1 and this is 2 at x equal to one the function is 0 at x equal to 2. So, let us say this is my value x is my value 2 at x equal to 0 if I put x equal to 0 then it is minus 1 into minus 2 which is something which is positive at x is equal to 3 for example, again it is positive. So, this tells me that the function will look something like. So, it is nothing but a parabola.

So with a minimum this is where I am calculating the minimum. So, f prime f prime x is equal to 0 means x is equal to 3 by two. So, this is x equal to 3 by 2 point.

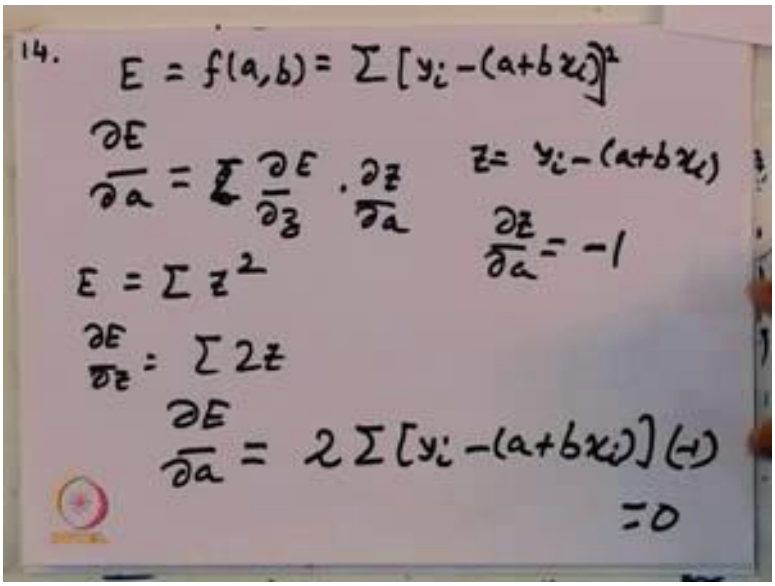
(Refer Slide Time: 22:36)



But let us say, in our case we want to minimize this particular function. So, in our case we want to minimize this particular error is equal to summation. So, my error is actually a function of a comma b. So, if you have more than one variable then the way you minimize is by instead of taking a regular differential. So, let us say e is equal to error which is function of a comma b. So, what I write I write down 2 equations del e of del a is equal to 0 and del e of del b is equal to 0. So, these are called partial derivatives these are called partial derivatives. So, let us say for example, if f of x y is equal to x y let us say plus y square. So, del f del x is derivative with respect to x assuming y is constant. So, my del f del x will return me a value of y my del f del y will return me a value of x plus 2 y.

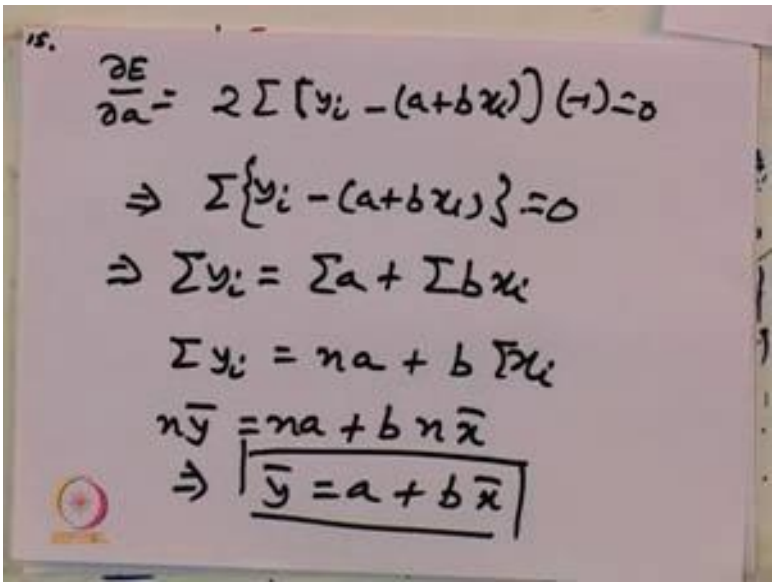
If this is how; when you take the derivative partial derivative with respect to x you assume that y is constant. So, of course, if y was the constant when you take this derivative it is simply y, and because y square is treated as a constant del f del x or del x of y square is equal to 0. So, this is actually y plus 0 is equal to y. Similarly I can take del f del y is equal to x plus 2 y is what I get. So now, coming back to the problem I have 2 equations my error is equal to f of a comma b.

(Refer Slide Time: 24:28)



So, let me find out the partial derivative possessive to a. So, you can let us say if I what to say is z is equal to y i minus a plus b x i. So, I can write this as is equal to summation of. So, if I do this particular z then I can take the derivative. So, my error is of the form summation of z I square z square. So, I can write down my del e del z. So, this is nothing but it is nothing of the form. So, del e del z is of the form summation 2 z. So, and del z del a. So, del z del a is of the form minus 1 because he treat everything else as constant and you take the partial derivative with respect to a. So, my del e del a will give me an expression of twice summation of y i minus a plus b x i into minus 1 and this is equal to 0.

(Refer Slide Time: 26:16)



So, I have del e del a is equal to 2 summation of y i minus a plus b x i whole into minus 1 equal to 0 which I can simplify, but simply imply summation of y i minus a plus b x i is equal to 0. So, this implies summation y i is equal to summation a plus summation be x i. So, I can. So, this n times a plus b summation x i. So, this is nothing but n times y bar, I can write n a plus b times n times x bar. So, this would give me the expression y bar is equal to a plus b x bar. So, what we have determined.

So, 2 things we have discussed we have worked out the case of correlation and shown how to calculate the correlation coefficient and then we wanted to go to the next step where we wanted to show can you find out a line which is which can be used to predict the value of y given a value of x and as part of that what we use we use basics of calculus and partial derivatives. So, we determine an error function and we say that we want to minimize the sum of all the deviations of the values of y i from what is predicted from that line and this is what we want to minimize and this is the first equation that we minimize when we when we write del e del a equal to 0.

So, with that I will I stop here, I will continue in next lecture where we derive the next equation for regression and see how regression can be made use of to predict a line and predict a value of y based on the line estimate.

Thank you for your attention.