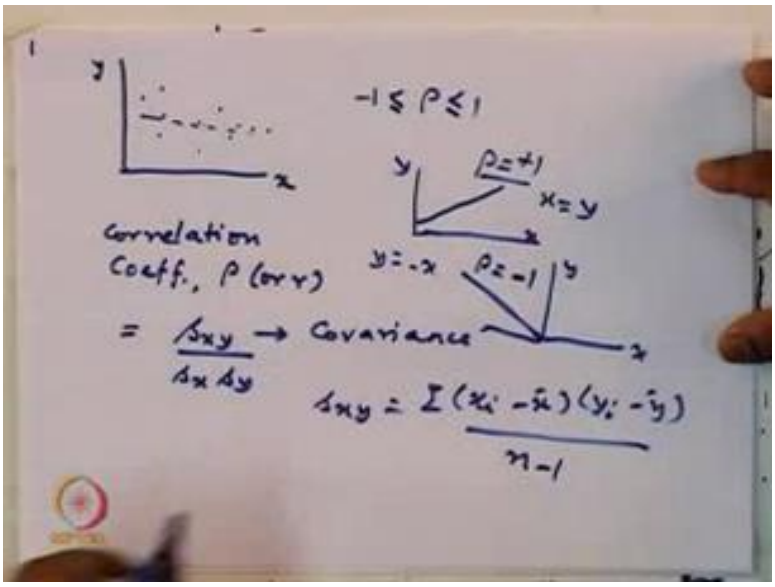
Introduction to Biostatistics
Prof. Shamik Sen
Department of Bioscience and Bioengineering
Indian Institute of Technology, Bombay

Lecture – 13
Interpolation and extrapolation

Hello and welcome to today's lecture. Today we will briefly review, what we had you know discussed in greater detail in last two classes which was correlation and regression and discuss about some aspects of fitting data. So, just to summarize what we had done in the few last few lectures.

(Refer Slide Time: 00:41)

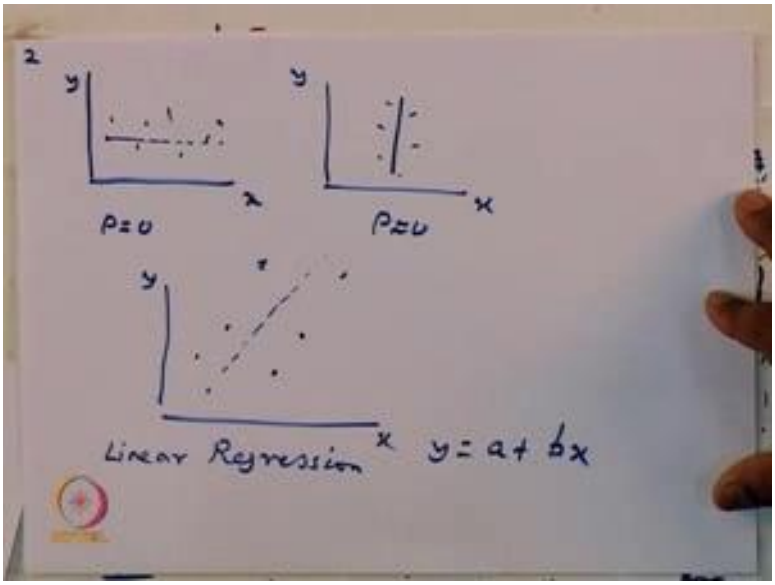


We had discussed how to plot bi variant data x and y and then that from the plots. Can you get some estimate of, what is the extent of correlation between these two variables x and y? So, in this particular case, is it we can draw a curve something like this, we can draw a curve which is horizontal so on and so forth.

So, to quantitatively determine the extent of correlation between two variables, x and y so, the correlation coefficient, we compute the correlation coefficient. So, the correlation coefficient is written by either rho or r and it is. So, and it is defined by s xy by s x times s y where, s xy refers to the covariance and defined by s xy is equal to summation of x i minus x bar into y i minus x bar whole divided by n minus 1. And we also discussed that the correlation coefficient is actually bounded between minus 1 and 1. So, for two

variables which are perfectly correlated; for example, if we take x equal to y in this case my correlation coefficient is going to be plus 1, but since if I take x equal to minus y, this is my x, is my y; x equal to minus y then you have this particular value. So, minus x is S. So, this is y is equal to minus x. So, in this particular case my correlation coefficient is going to be minus 1.

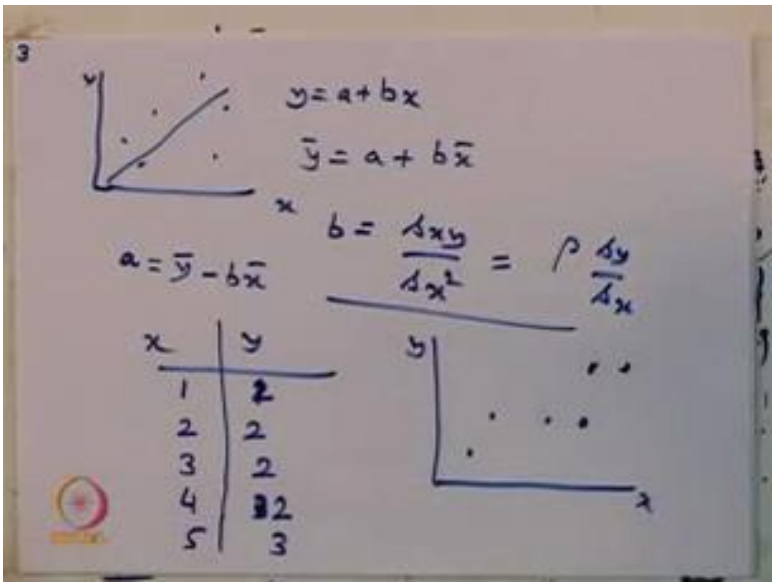
(Refer Slide Time: 02:34)



So; however, if you are data points are all scattered, as in this case let us say. When you compute the correlation coefficient you might have. So, for a line which is perfectly horizontal your correlation coefficient will come to either perfectly horizontal or let us say if it is perfectly vertical like this, in both these cases by correlation coefficient will approximately turn out to be close to 0. Now one of the immediate follow ups of correlation is regression.

So, if you have a set of points x and y like this, can you estimate y given one value of x or vice versa? So, here the objective is to approximate these data points by some curve may be linear as is you know as seems from the trend from these particular points and can we in case of linear regression, I want to approximate this data by some equation y a plus b x.

(Refer Slide Time: 03:50)

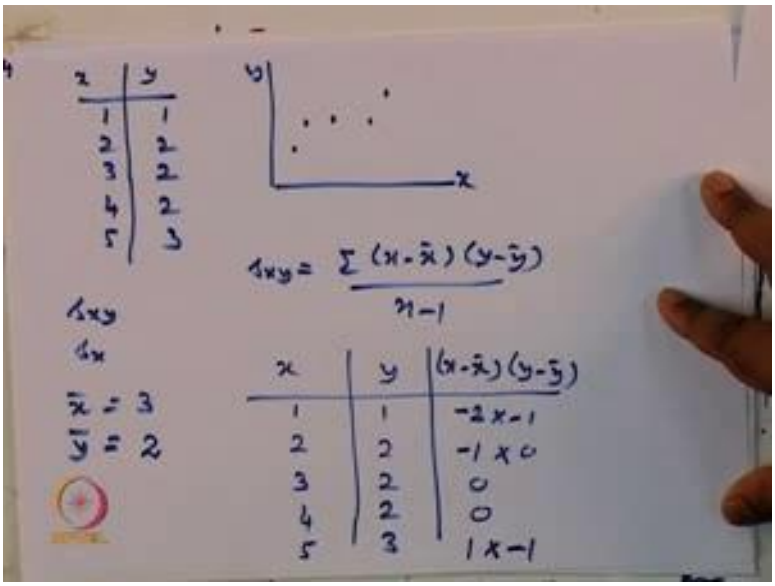


So, for doing linear regression; so once again you have x and y which are plotted like this and I want to approximate by this particular curve which is y is equal to a plus b x. So, I had derived in few you know two or three more back lectures that you can find out the value of a. So, you have from this equation I get one equation which is y bar is equal to a plus b x bar and the other equation which I get is b is s xy by s x square and this will give me a value, which I can also write it as rho times s y by s x.

This is the value of b and once I calculate the value of a, a is nothing but y bar minus b x bar. So, this is how you calculate a and b. So, let us take one example once more and do one more thing. So, let us say the first step is of course, to find out the values of a and b respectively, but the next step is to ask, is this equation a good enough fit for this particular experimental data?

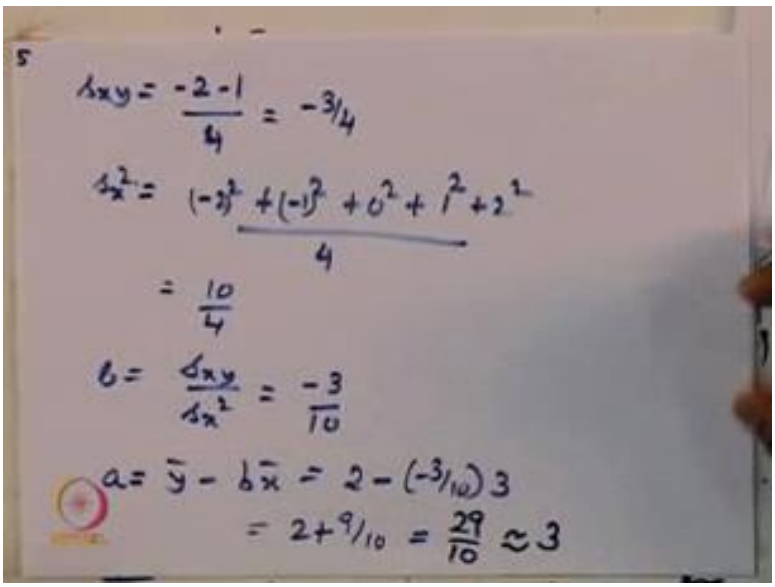
So, let us do one particular example. So, let us take the following values x is 1 2. So, let us let me you know make a data like this, this is my x, and this is my y. So, x is let us say 1 1, then you have 2 2, 3 2, 4 3. So, what you can clearly see. So, if I can actually add one more point or maybe, what I do is; 4 2 and 5 3, so my data points look something like this. So, my points are once more.

(Refer Slide Time: 05:43)



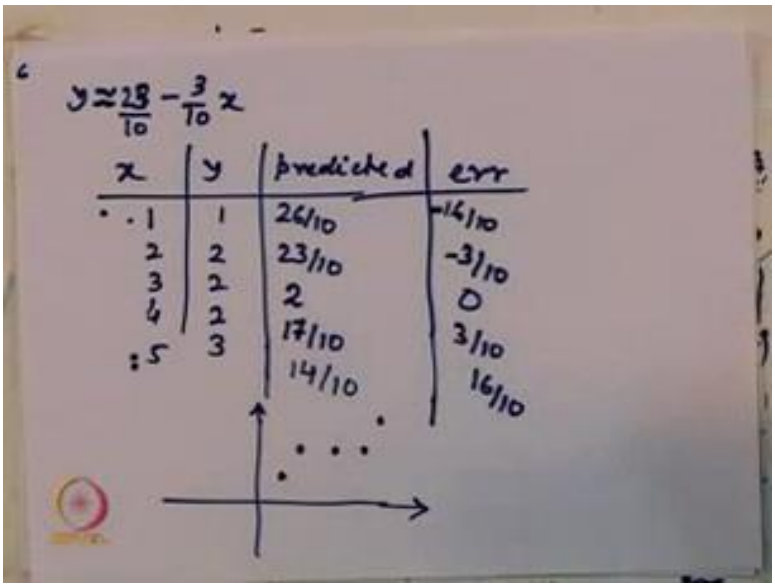
X and y; 1 1, 2 2, 3 2, 4 2, 5 3; so I know. So, I have to calculate s xy and s x right to in order to be able to compute b. So, for these particular values of x, I can calculate. So, s xy is given by summation x minus x bar into y minus y bar whole divided by n minus 1. So, x bar comes out to be 3, y bar comes out to be 4 1 6 10, y bar comes out to be 2; 1 2 3 4 5, 1 2 2 2 3. In this case x bar is 3 is minus 2 into minus 1, in case of 2; is to minus, minus 1 into 2 minus 2 0, this 3 is also will give you a value of 0, 4 will also give you a value of 0. In case of 5, you have 1 into y minus 1 bar minus 1.

(Refer Slide Time: 07:28)



So, in this particular case I can then find out the value of s xy is going to be minus 2, minus 1 divided by n minus 1 is 4 is equal to minus 3 4th and s x is s x square is going to be x minus x bar whole square, x bar is minus 2 square plus 2 minus 1 square plus 3 0 square plus 4 minus 1 square plus 2 square by 4. So, it is going to be 4 plus 4, 10 by 4. So, b turns out to be s xy by s x square is equal to minus 3 by 10 and a comes out to be y bar minus b times x bar is equal to 2 minus of minus 3 by 10 into x bar is 3 is equal to 2 plus 9 by 10 is equal to 3 plus 9 by 10 is turns out to be 29 by 10, this is approximately equal to 3.

(Refer Slide Time: 08:56)

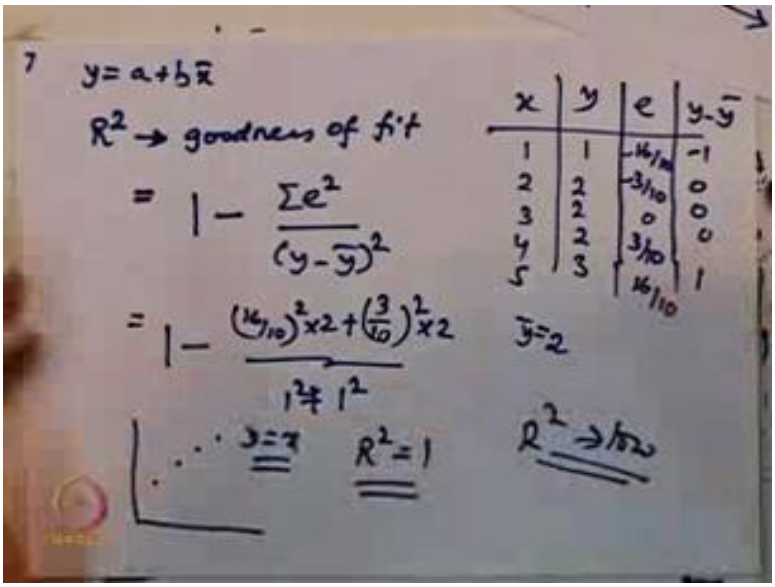


So, my equation is y is equal to I am approximately 3 minus 3 by 10 into x. Now I want to find out, so for each of the values. So, let us say 1 2 3 4 5, 1 2 2 2 3, predicted value is, so I can actually maybe I should write 29 by 10 x equal to 1, 29 minus 26 by 10, 2 6 23 by 10, 43, so 9 2 4 is 22 5 (Refer Time: 10:04) 4 into 3 12 and 5 15 14 by 10. So, I can clearly see that for these two values it is doing a reasonable job, for 1 and 5 it is really doing a very bad job and the reason is obvious. If I were to draw the points again, see these are your points and the line it has been drawn is roughly y is equal to 3 (Refer Time: 10:40) plus P, so your line.

So, this is doing a very bad job for these two extreme points. So, I can compute the error which is the difference between actual value of y and this deviation. So, this comes to 16 by 10, this comes to minus 3 by 10, this comes to 0, this comes to 3 by 10 and this comes

to 16 by 10. This is minus 16 by 10. So, the way to estimate the goodness of fit of this data is to compute is. So, what you typically have is, if you do it in excel or origin or any of the standard (Refer Time: 11:29) softwares.

(Refer Slide Time: 11:31)



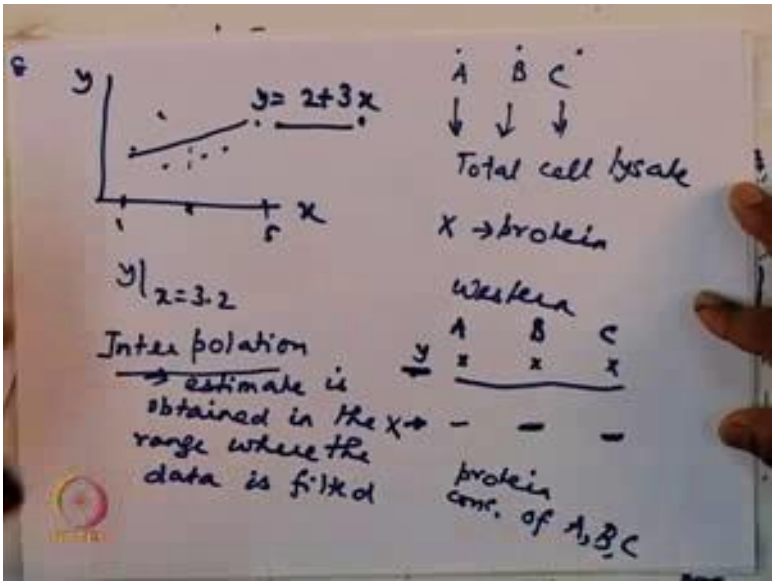
So, in additional to this equation that they write, they also write a value of R square. This R square is a measure of goodness of fit. So, this R square is given by the expression R square is equal to 1 minus summation of e square by y minus y bar whole square. So, we have calculated the errors we have. So, if I do the particular values, we have error and I want to plot y minus y bar. So, I have 1 2 3 4 5, y is 1 2 2 2 3 error predicted. So, error is minus 16 by 10 minus 3 by 10, 0 3 by 10 and 16 by 10.

And y bar y minus y bar. So, y bar I know is equal to 2. So, y minus y bar is minus 1 0 0 0 and 1. So, my goodness of fit in this case is equal to 16 by 10 whole square into 2 plus 3 by 10 whole square into 2 and y minus y bar is 1 square plus 1 square. So, this is a significantly large number 16 by 10 here. So, I will get a value of R square which is reasonably low. In other words this is not a good fit. So, you can compute the R square value. So, for whenever you do the calculation you should also compute the R square value and then come to a conclusion if you are if your line is a good representative of a data.

So, if for a very good curve, so let us say for if we have a curve which exactly point are fit in to the point y is equal to x. So, in this case you will get a value of R square is equal

to 1. So, in best case situation, if you have a perfect fit R square, should return your value of 1 versus if it is a bad value you should get a very R square value which is very low. So, this tells you that this is how we compute the R square value or the goodness of fit of the data. Now one of the obvious usages of doing linear regression is to use it to estimate a value of y given a value of x.

(Refer Slide Time: 14:07)



Let us say we have this particular. So, we had this data based on which we fit this line and we want to know what is the value of x. So, let us say this is 1 and this is 5 and I want. So, this is my x is my y and this is a particular equation, let us say 2 plus 3 x.

So, I want to know the value of x, I want to know the value of y given a value of x. Let us say, I want to know the value of y at x equal to 3.2. So, what I do is, I simply plug in the value of x and get the corresponding value of y. So, this is called interpolation. Interpolation means that the point you are trying to get an estimate. So, estimate is obtained in the range where the data is fitted. So, because the range over which I was fitting my curve was between 1 and 5. So, when I put an x equal to 3.2 values, it falls within this range. So, this process is called interpolation and this so let me give you one you know very common use of interpolation in biology.

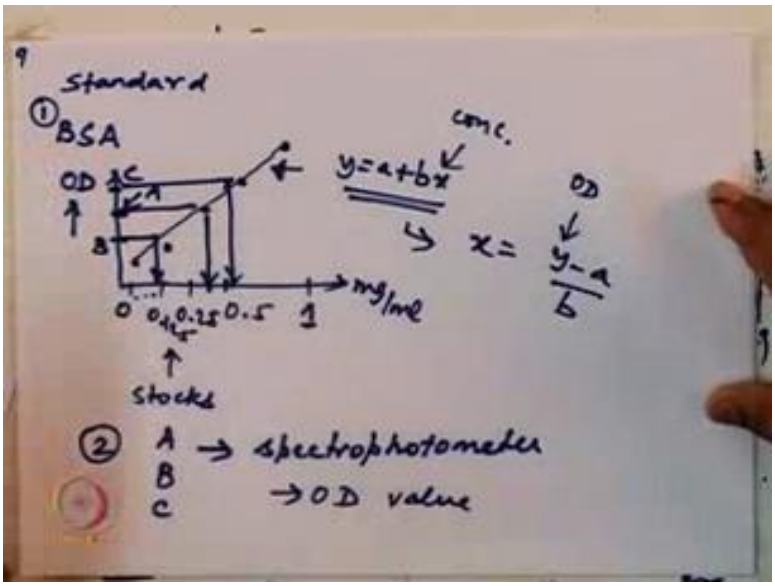
So for example, if you want to do, let us say a western right. You have three samples; A, B, C from these three samples you have collected the total cell lysate which is the content of all the proteins present in these cells cultured under let us say cells which have

been cultured under conditions A, B, C. So, and I want to ask the question how does the expression of protein X? So, X is a protein, how do, how does the expression of protein X vary as a function of the conditions A, B, C?

So, what I need to do is as a first step; if I want to run a western. So, in western, what I do? I want to load proteins of so, I want to load total proteins y let us say of each sample A, B, C and then get some bands in electrophoresis using page and I want to develop these bands and ask the question how is the expression profiles. So, this is the band you get corresponding to y and I want to know what is the, how is y variant? So, one of the critical steps is to ensure that is to so, this is your protein X band and you want to know how X is varying across different conditions. So, these are your conditions A, B and C. So, in order to make sure that you are comparing X as a fraction of the same equal amount, you want to make sure that the same amount of protein y is loaded across all the conditions.

So, which means that you want to know, what is the protein concentration of each of the samples A, B and C? So, the question to begin with you want to estimate the protein concentrations of A, B and C. So, how do you do it? So, what you do? You use something called a standard.

(Refer Slide Time: 17:34)



Let us say standard and typically people use BSA or Bovine Serum Albumin. What you do is you prepare stocks of various concentrations. Let us say 0.5 so on and so forth. So,

you use your protein samples of known concentrations. So, if you know that the highest concentration of any of your samples A, B, C is going to be less than 1 mg per meal. What you do is you prepare protein stocks. So, these are all stocks of varying concentrations, you prepare this stocks and you get the OD measure from a spectrophotometer.

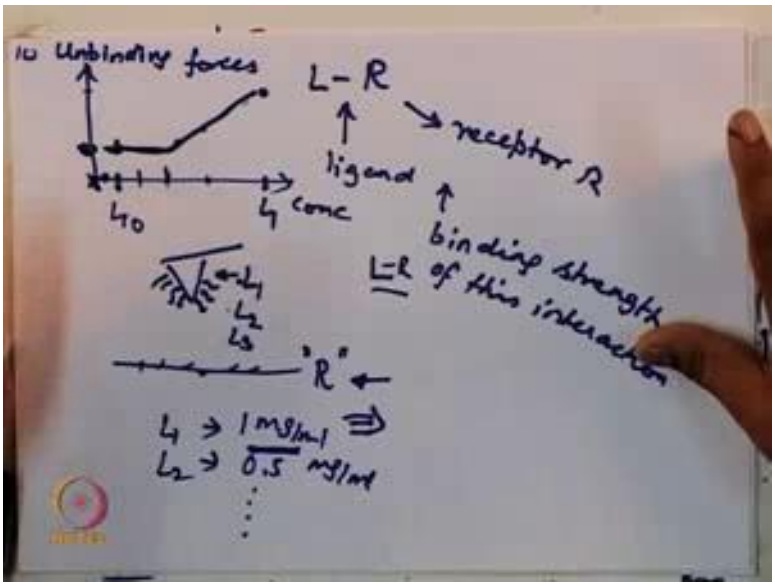
So, what you will have are these points, which will give you these data points and to this you would fit either a line, you would fit a line to get what is the estimate of your sample. So, to get an estimate of what is y is equal to a plus b x is what you will obtain. You will obtain the equation corresponding for these data points which are prepared with stocks of various concentrations. So, this is your step 1.

In the next step, what you do is let us say you take your protein samples A, B, C and you use the spectrophotometer to take the OD value. So, in the background of this what you have is let us say you have the y value. So, let us say this is for sample A, this is for sample B and this is for sample C. These are your OD values corresponding to sample A and sample B and sample C. So, how we do backtrack and find out the concentration? So, if you know the equation you can invert the equation to find x is nothing but y minus a by b. So, because you know the OD value which is your y. So, y is your OD value and x is your concentration.

So, given the OD value, you backtrack using this equation to find out the exact concentration of A, the exact concentration of B and the exact concentration of C. So, this is how it is widely used for interpolating data and to finding out how much protein you should use. So, that in order to assess that across condition A, B, C what is the expression changes in the expression level of protein A? So, this is the way procedure of how you make use of interpolation to find out a given value.

Now, interpolation is reasonably straightforward. If your data points do fit a very clean line then, the fit is nice and your protein estimates are good. What is more difficult is to extrapolate. What is extrapolation?

(Refer Slide Time: 20:33)



Let us say you have you are proving the binding interaction between a ligand L and its receptor R. So, when L binds to R and you want to find out the binding strength of this interaction. You want to probe the binding strength of this interaction. So, experimentally one of the (Refer Time: 21:05) used. So, if you want to measure actually in terms of forces, the binding strength between L and R. So, what you want to know is how much force is required to tear up out a bond which is formed when L and R, so when L binds to R. So, you can use any experimental assail, let us say for example, atomic force microscopy is one such assail which can be used to prove this.

So, the critical problem here; so you want to estimate the binding strength of one molecule, but it is nearly impossible to estimate the you know the binding strength of one molecule because this measure of let us say 10 Pico Newton force is way below the resolution limit of your system. So, what you do is let us say you take a surface and you put down known concentrations of your receptor R. You put known concentration of R and then what you do is you take a tip and you vary the concentration, you functionalize this E F M tip with your protein L and you vary the concentration of L.

So, essentially you do the experiments between L 1, L 2, L 3 so on and so forth. So, let us say, I is begin with 1 mg per ml concentration. L 1 is 1 mg per ml concentration, L 2 is 0.5 mg per ml concentration and so forth and so on and so forth and I keep reducing the concentrations and let us say. So, I have no way of knowing exactly how much, how

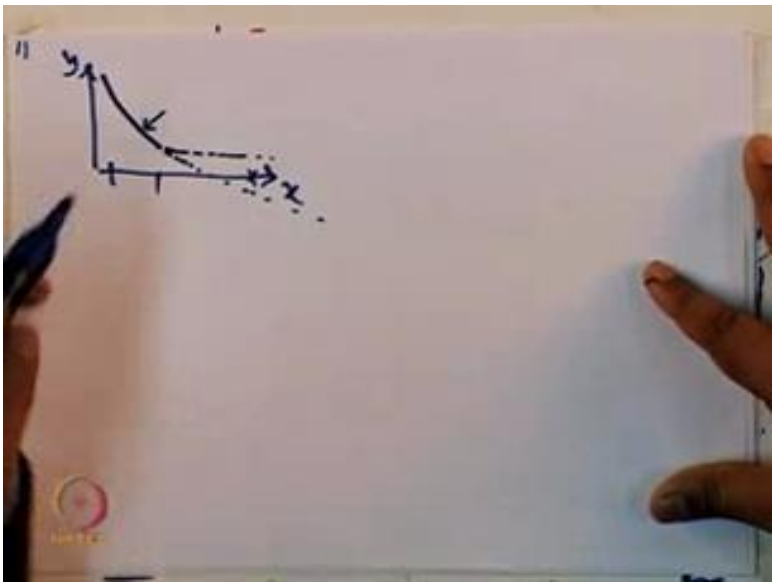
many bonds have been formed when this L 1 number of ligands bind to this particular things because you are operating at a macro scale where you only have control over this concentration which is 1 mg per ml. It does not you do not know exactly how many molecules of this particular ligand at that on the tip. So, it is an estimate, it is an approximant you know the scaling, but you do not know the exact value.

So, how will you go ahead and do it? So, let us say I plot the unbinding force. So, this is my concentration axis and I plot my unbinding forces, that I measured using an EFM. So, at a very high concentration, logic would dictate if at L 1 which is my highest concentration I get this particular force, that an L 2 which is at half concentration. On an average I should see a lower force and. So, this curve, this curve should keep on reducing should reduce and then for some concentration for below a certain concentration I begin to see a plateau. So, I begin to see a plateau. So, if I begin to see a plateau then, I know and let us say this is my lowest concentration which was L 10.

So, till L 10 I begin to see a plateau. So, what this means is, in this range of concentration it is likely that only 1 bond is forming because if I reduce the concentration further, I only hit a value of 0 and nothing. Below this concentration I only hit 0. So, this means that I am operating very close to either 1 bond or 2 bonds. So, using this if I know at this concentration, I am getting a plateau, I know that below this. So, this is a roughly mimetic of the binding strength of a single bond. So, this is how I can make use of extrapolation, even though I do not have a way of controlling the interaction between a single you know single bond or forming up a single bond, but by doing these experiments and by extrapolating it might be possible to operate at a level such that this force can be estimated.

So, this is extrapolation, but they are of course lot of fear. So, the problem with extrapolation is you are still making so.

(Refer Slide Time: 25:17)



Let us take an example, imagine my curve is something like this. Now and this is my x this is my y. This is the regime within which I have collected my data. So, and I want to know the value of y at this particular value of x. So, do I assume that the curve is doing something like this? Or the curve is actually saturating beyond this point. So, this is what makes the process of extrapolation somewhat more complicated and many times you might actually get lot of errors, as if you pre assume the nature of the function.

So, the challenge is really to esteem use the appropriate function to fit your data. So, that the when you extrapolate this particular function to get the estimate of y for a given value of x, you are it is a reasonable estimate. With that I conclude my lecture for today and I hope you have gotten an idea of how we can make use of linear regression to fit data and to make use of it to interpolate and extrapolate. So, if your data is you know if have data is very erroneous then, if your data has wide scattered; in this case extrapolating can give you large errors. Interpolation is reasonably ok still, but still as I have drawn in this case, if I get a line like this at this point you know you see the amount of deviation from this line. So, you might still be, the estimate might still be very erroneous.

With that I thank you for your attention, and I look forward to next class.