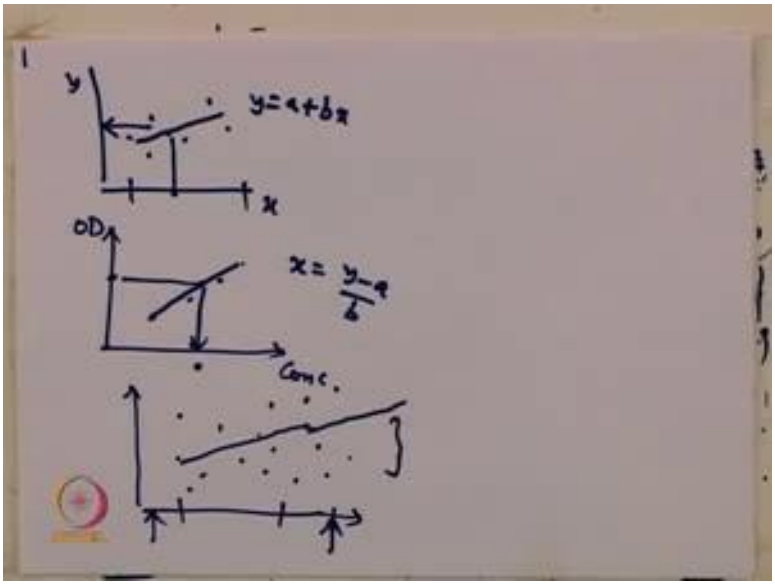
Introduction to Biostatistics
Prof. Shamik Sen
Department of Bioscience and Bioengineering
Indian Institute of Technology, Bombay

Lecture – 14
Nonlinear data fitting

Hello and welcome to today's lecture. We will briefly recap what we discussed in last class, which was you know regression linear regression and its usage in extrapolating and interpolating and extrapolating the data.

(Refer Slide Time: 00:34)



So, as we saw that interpolation refers to finding an estimate within the range within which you have fitted the data. So, if these are your data points and this is the line that you know you have found out by linear regression right. So, if this case for a given value of you know you might want to know, what is the value of y given x; in that case you plug in the value of x and use this expression to get the value of y, but you can also do the reverse thing where you are in, we discussed in last class where you can use linear regression to find out the protein concentration of your sample.

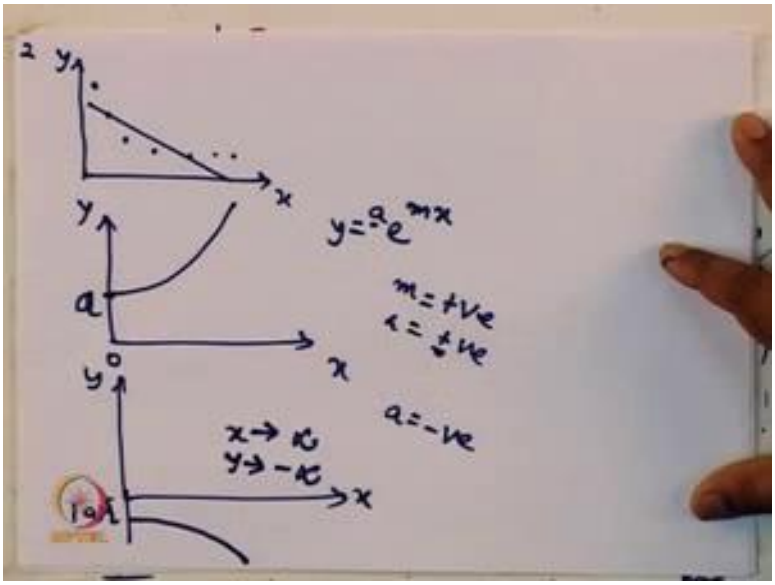
In which case you create a standard curve you fit a line and so in that case, so in case of protein concentration estimates. So, your x axis is concentration, y axis is OD value which would take with the spectrophotometer. So, and then for a given sample you find

out what is this OD value and then invert this particular equation in the form x is equal to y minus A by b to determine, what is your protein concentration?

So, this shows how you can make use of interpolation both to find out the value of y and to find out the value of x as the case may be depending on whether you know the x or the y value for one particular measurement. So, extrapolation refers. So, if you have this particular data and I also this is your range and I want to estimate a value which is beyond this range either here or here. So, this is what extrapolation is about. So, if your curve is well fit then extrapolation will not give you too much of erroneous results; however, if there is lot of scatter in your data, let us just say I make this scatter this bad.

In that case, so as you go further you know if your scatter was here. So, your extent of error will keep on increasing as per this equation. So, in interpolation whatever you use a maximum error that sets the bound of how much error you will accumulate when you interpolate, but in your extrapolation step your error can be much higher or much lower depending on the case under consideration. So, these are of course, for linear curves. Let us take few cases of non linear curves; more often than not you will have curves.

(Refer Slide Time: 03:01)



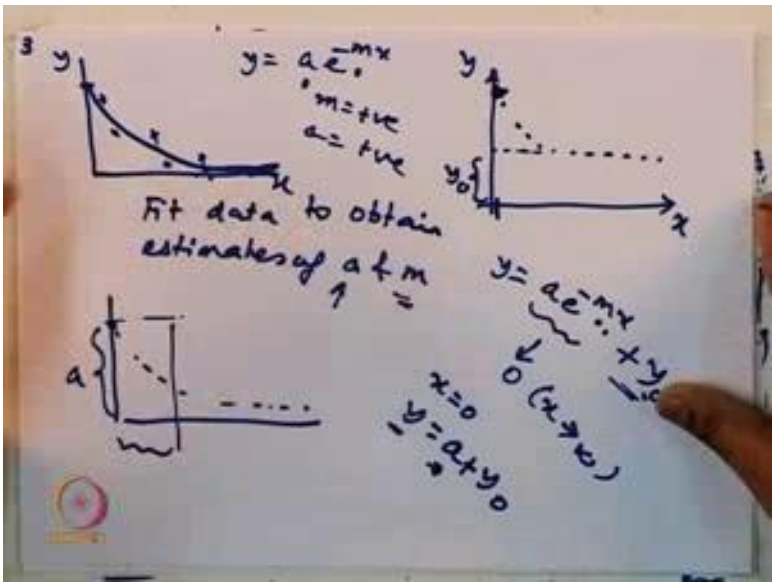
So, let us take an example where, this is x this is y and your data points look something like this. Of course, it does not make any sense to fit it with a line because you know right away. If you were to do a line fit you will get something like this which does not

make any sense, but these x you know these points give you the feeling that there is something which decays and then plateaus off.

So, this can be used this is where it is important to know how different functions behave. So, let us take very you know simple cases well let us assume that the underlying phenomena as such that you are expecting an exponential decay. So, we had briefly decided in the class. We have this briefly decided in your initial lectures about plotting that how does a curve, y is equal to e to the power m x look like. So, if let us say m is positive or we can write y is A to the power m x. If m is positive and let us say y is positive also let us assume y is equal to positive also, y is also positive. So, your curve at x equal to 0 e to the power m x will always return your value of 1. So, you know that the curve starts from A, A times 1. So, at x equal to 0 your, it starts from A and then it increases exponentially. So, your curve should look something like this.

Now, for the same curve if, A is negative let us say. If let us say A is equal to negative. So, then the curve will actually look like this. So, you see the value of A will dictate what is the nature of the curve. So, in this case when A is negative this is your magnitude of A or mod A, this is just a mod A is same as A. So, in this case your curve will you know go to x tend as x tends to infinity, y tends to minus infinity because of A is negative. So, of course, we have these curves where when m is positive, but let us say the more interesting thing for many biological phenomena is when m is negative.

(Refer Slide Time: 05:25)



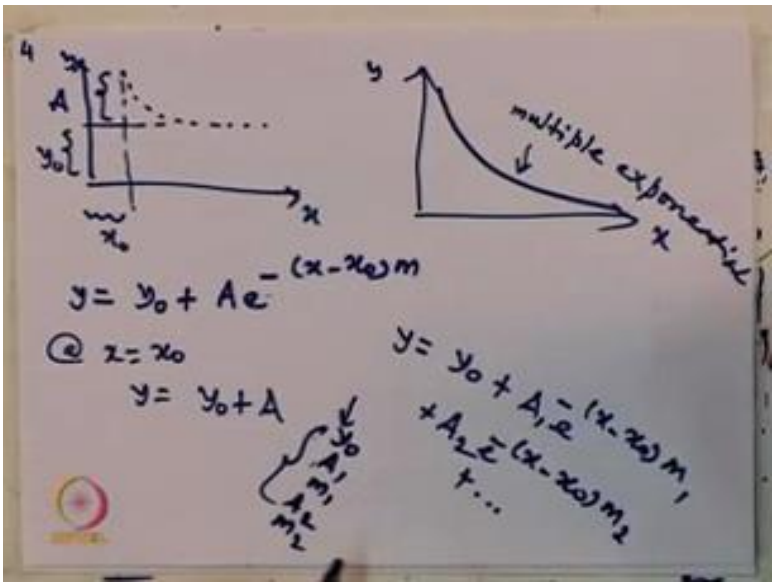
In that case, your curve should look something. So, let us say we are plotting y is equal to A e to the power minus m x. So, I put the minus so I can keep m as positive and A as positive. So, I know that whatever as at x equal to 0 it starts from y particular A and then as x tends to infinity it goes tends to 0. So, your curve should look something like this. So, I have it should basically meet at x tends to infinity. So, it should not hit 0. So, it is obvious that for this particular kind of curve. So, for example, if your data was like this;

So, if you knew, if this is the nature of your data then, you can fit an exponential and with two parameters. So, in this case you need to fit the data to obtain estimates of A and m. Now estimate of A is reasonably easy to obtain, all you have to do is you look at your data and if it starts from one particular value. So, if this is your data, so this particular value of A. So, whatever is your maximum value of x, maximum value of y this itself should be A and then m is what you will have to estimate. As a rule of thumb, you know that within where you have most of the decay your m roughly is of that nature.

Now, let us say how if this curve, but will change. So, let us say it starts from a given value and your data plateaus off to some particular value. So, it does not go to 0 and x tends to infinity, but it plateaus off to this particular value. In this case, you will have to modify your equation to fit by writing y is equal to A e to the power minus m x plus some y naught. As you observe that this y naught, so as x tends to infinity this component will return your value of 0, when x tends to infinity.

So, y naught is the plateau value. So, from your data you know how to estimate y naught which is nothing but this particular value y naught. So, this is the plateau value is going to give you the value of y naught and you know at low values of x, whatever is your value of y right. So, at x equal to 0, y is equal to A plus y naught. So, whatever is your value at this particular point you know. So, if you know that, let us say at x equal to this point you know the value of y you can find out what is an estimate of A and then that leaves us with this particular constant m to be determined. So, as m increases your curve should look sharper and sharper.

(Refer Slide Time: 08:30)



Now, let us say. So, we had if there was a vertical shift in y axis you modified the equation from y is equal to A e to the power minus m x to y is equal to y naught plus A to the power minus m x. Let us say your data is something like this. So, what you see is the data this first starting point does not start from x equal to 0, but it is offset. So, in this case let us say, let this offset be x naught and this offset be y naught. So, this kind of a data you will have to fit using a function y is equal to y naught plus A e to the power minus x minus x naught into m.

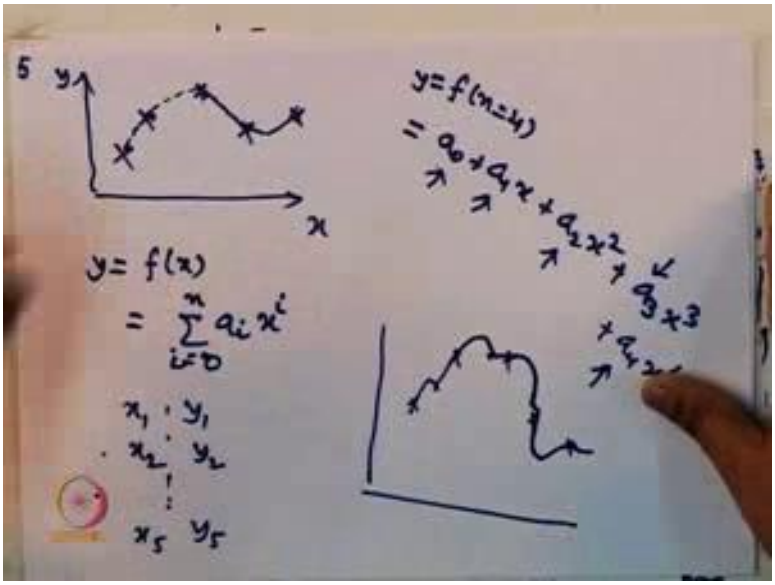
So, y x minus x naught you see that at x equal to x naught which is your starting value, this will return your value of 1 right; x minus x naught is 0. So, e to the power minus this is will be return a value of 1. So, at x equal to x naught y is going to be y naught plus A. So, this is your estimate of A and this is your estimate of y naught. So, accordingly you see that depending on how your function shifts either in x or in y axis you need to appropriately modify the functional form of the equation with which you are fitting the data.

Let us take another example, where the data has this long tail. As opposed to showing a very quick saturation, it saturates in a much you know slower manner. In that case suggesting, that this curve has multiple exponentials, so in this kind of fitting this kind of curve you might have to modify this particular equation, but y naught plus A 1 e to the

power minus x minus x naught into m plus A 2 e to the power minus x minus x naught into m, so m 2 so on and so forth.

But one of the dangers of increasing the number of parameters to fit is that there. So, you are what you are essentially doing is multi dimensional optimization. You have various multiple variables. So, for example, in this particular case you have y naught A 1 m 1 A 2 two m 2. So, you have five variables to fit, why now it might be easy to understand, but all these other variables are not need not be as straightforward. So, increasing the number of variables you have to optimize for four particular parameters, which is difficult in higher dimensional space. So, it is best that you use the minimum number of fitting parameters to get a good fit of your population.

(Refer Slide Time: 11:37)



So, what is often done, if you have lot of points? So, let us say if this is my x y axis and I had these five points. Now looking at the way I have plotted, I might want to connect my point something like this, tending to give me a feeling that these points are actually part of some sinusoidal curve. In that case, I can use a sinusoidal curve or in the most generic case if I want to get a unique curve which passes through five points, I can use with a four dimensional polynomial equation.

So, in the polynomial equation y is of the form summation a i x to the power I, i is equal to 0 to n. So, for let us say, if I have five points to fit; that means, I have at x 1 the value is y 1, at x 2 the value of y 2 so on and so forth up to x 5, the value is y 5. If I want to fit

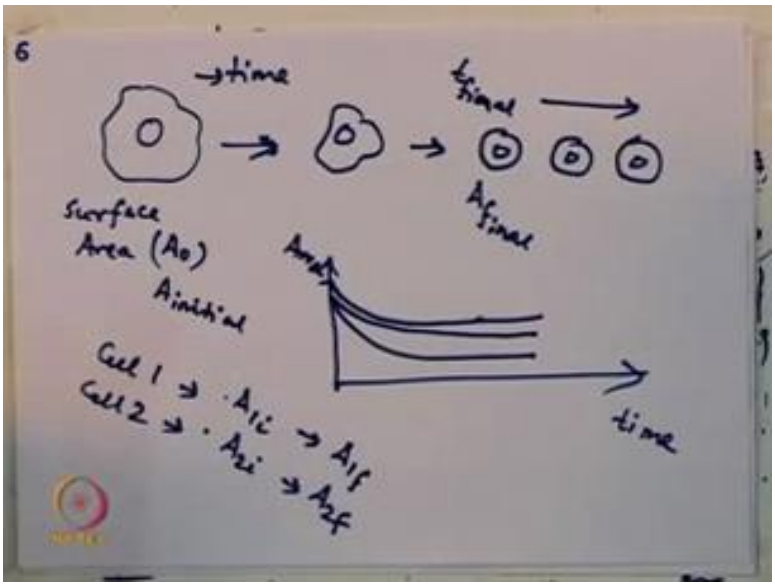
this particular data right, I have five equations. So, for a four dimensional polynomial, so if y is equal to f of up to n equal to 4 then, I can write y using the form a 0 plus a 1 x plus a 2 x square plus a 3 x cube plus a 4 x 4th.

So, I see that there are five fitting (Refer Time: 13:04) a naught, a 1, a 2, a 3 and a 4 and I have five equations. So, plugging the values of x 1, I will get y 1 is equal to a naught plus a 1 x 1 plus a 2 x 1 square plus a 3 x 1 cube plus a four x one 4th. So, you can have five equations in five unknowns which can be uniquely solved. So, that would mean that for every number of variables you can fit a higher dimensional, you know you can fit a polynomial such that you get a good fit, but it is still not advisable for the simple reason that if you fit a higher dimensional polynomial.

So, many times you will get between these two also you might see the fitting which looks like this. You have these you know a higher dimension polynomials will give might give you these traps and these are just an artifact of the fitting itself and the equation. So, this is not advisable. So, higher dimension polynomials are not advisable in most cases. You can use up to cubic because you know that the function is gives a very well defined form, but for different combinations of these parameters many times the polynomial is such that you will get very complicated shapes which does not have any bearing on the physical phenomena that is underlying this particular process.

So, I will give you another example, a biological example where, the nature of the fitting can change depending on what you are measuring or what you are quantifying.

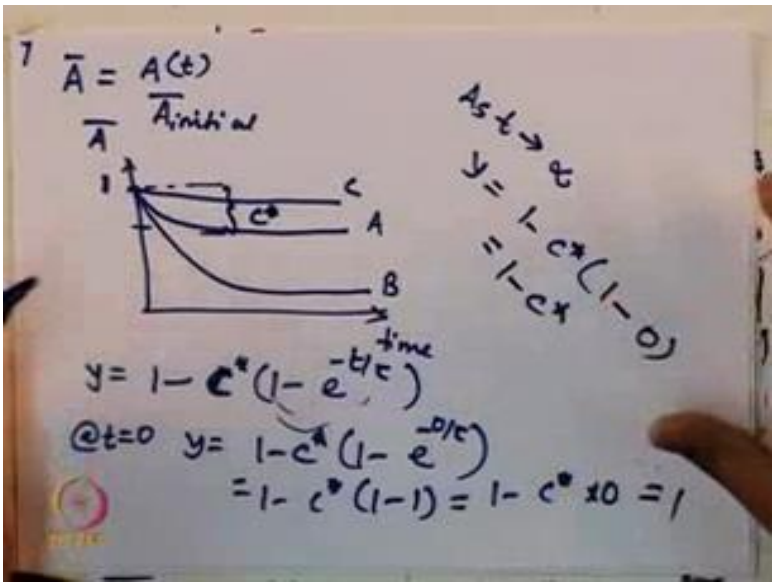
(Refer Slide Time: 14:38)



So, imagine I track a cell, this is a biological cell and as a function of time. It goes from this round you know this shape and I can calculate its surface area. Let us say it, is surface area at this time point is A naught and then it is rounding up because of some reason and it ends up with a very rounded shape and then once it is rounded beyond that time point it remains rounded. So, if I call this A f or A final, I can also write A 0 or A initial.

So, after t final, this is constant. So, if I were to plot this particular area as a function of time. So, area as a function of time will look something like this. So, this is my time axis, this is my area axis. Now let us say, I want to fit this with what some equation, now because now this when you do this statistical measure let us say for cell 1, for cell one it went from A 1 initial to A 1 final. For cell 2, it went from A 2 initials to A 2 final. So, this tells you. So, there is no point in plotting all for different cells. So, for one particular cell it might be like this, for another particular cell it might be like this, for another particular cell it might be like this. So, I need to account for the fact that initial cell area of a different cells is different so; that means, that there is it does not make sense to plot the area itself, but rather the normalized area.

(Refer Slide Time: 16:43)



So, if I plot the normalized area. So, let us let me define A bar equal to A of time t by A of initial. So, in this case my curve will always start from 1 and plateau off to one particular value. So, this is my time axis and this is my A bar axis. So, what i can? So, this at least the way I have plotted it, it gives that this somehow mimics an exponential decay. So, I can plot this particular function with let us say the equation y. So, it plateaus off so; that means, that there is some drop and there is a characteristic time constant beyond which the area does not change. So, there is a time constant. So, as opposed to this particular case let us say for condition A, for condition B I might have a situation like this and for condition C, this is my condition C.

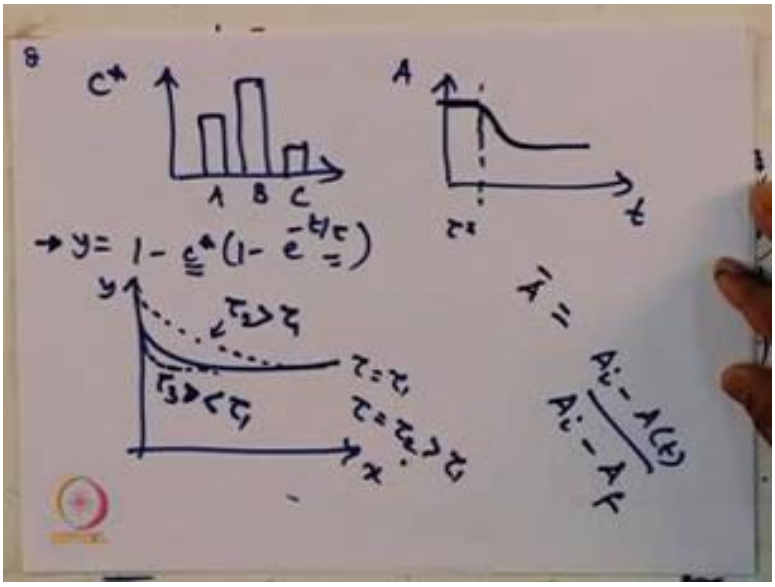
So, I want to have a way of quantitatively finding out matrix so that I can compare what these conditions A, B and C are doing to my data. So, let us say if I use this particular expression 1 minus C star into 1 minus e to the power minus t by tau. So, this is the equation that I have used 1 minus C star into 1 minus e to the power minus t minus tau. What is the basis of my using this particular equation? So, let us see what is happening to this particular equation when t is equal to 0.

So, when t is equal to 0, this equation returns me a value of 0. So, C star multiplied by 0, then why should give me a value of 1? So, at t equal to 0 y is equal to 1 minus C star into 1 minus e to the power minus 0 by tau. So, is equal to 1 minus C star into 1 minus 1 is equal to 1 minus C star into 0 equal to 1. This equation is giving me a value of 1 and as

so let me also support when t tends to infinity. When t tends to infinity, exponential anything times t to the power of infinity would give me a value of 0. So, as t tends to infinity y will give me a value 1 minus C star into 1 minus 0 which is nothing but 1 minus C star.

So, what it tells me? So, 1 minus C star is this, which would mean that in the original case this is what is C star. So, this is what is measured by C star and this is 1 minus C star is the final plateau. So, what I can compare across different situations A, B, C.

(Refer Slide Time: 19:46)



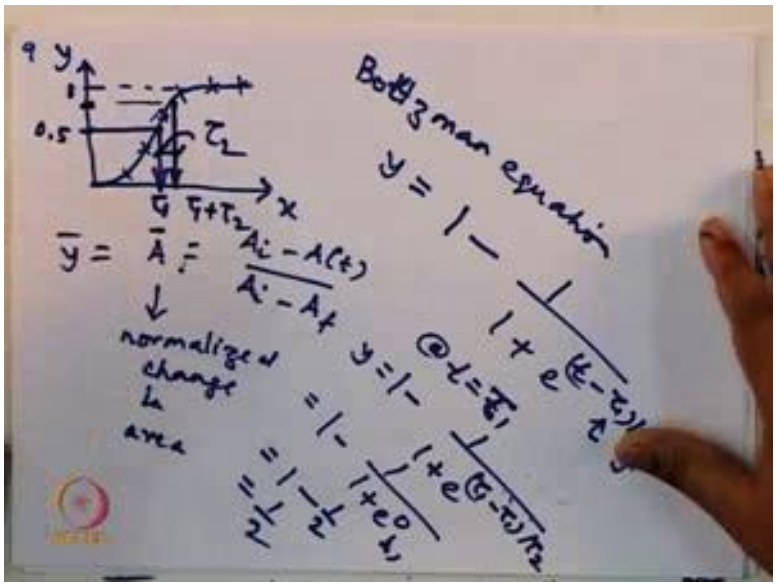
Is I can find out the value of C star and let us just say I can have a situation. So, this is C star for condition A it is like this, for condition B it is like this, for condition C it is like this. So, this is A, condition A, condition B, condition C. I can also do so in the previous case. So, if you, I can also find out the characteristic time constant.

So, I have this particular equation right, y is equal to 1 minus C star into 1 minus e to the power minus t by tau. So, let us say this is for a particular value of tau. So, this is my x this is my y, this is for one particular value of tau 1, tau is equal to tau 1. So, for if tau equal to tau 2 is much is greater than tau 1. Then, I know that this process should get delayed. So, for the same value of C, if I have to have another condition where, tau is only changing C is not changing, the curve should look something like this. So, what it means is it is taking a much larger time to relax at for this to drop.

So, in this case I have tau 2 greater than tau 1. Similarly, if I have something which is super fast then the curve should look like this. So, this is tau 3 greater than much less than tau 1. So, this is how you can come you know, you can find out these two. So, in this particular case; they are two well defined time constants C star and tau and this is how it plots.

Now, this particular equation has a caveat. So, what you see is this change begins right away. Now imagine you have a situation from initial area. So, imagine have you how you have a situation where from initial area if I were to just plots the A as a function of t. I have actually my curve looks something like this. So, which means at for a certain time point tau star there was no change in area and then it began to drop. So, this equation cannot take care of this time tau star into this calculation. This tau only captures for the time constant when there is a change in area. So, how can we incorporate this into the equation? So, in this case maybe instead of tracking in the area itself A, we can track the change in area.

(Refer Slide Time: 22:42)



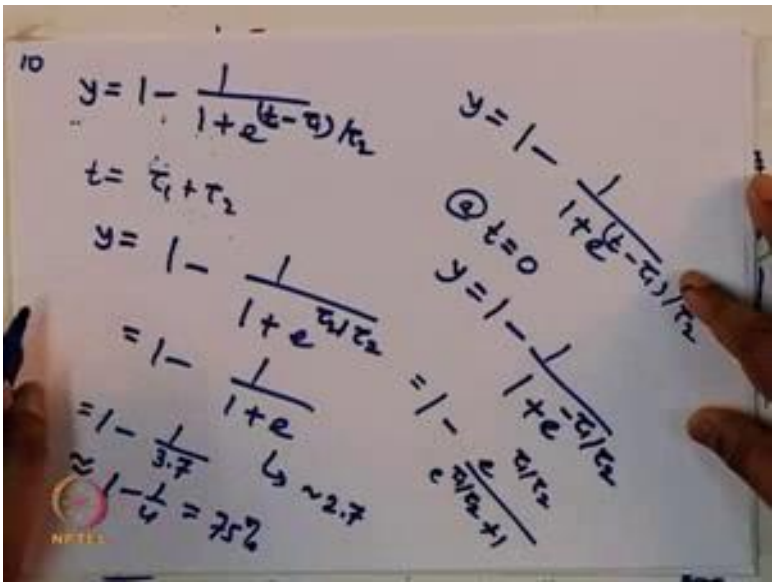
So, I can define this normalized parameter A i A bar which is A i minus A of t by A i minus A f. So, in this particular case, the curve will look something like this. This is y, this is x, so once again why do we need to do this normalization? Once again, why do we need to do this normalization? This is because for cell 1, again for cell 1 you have might start from an initial area A 1, for cell 2 you might start from initial A 2 so on and so

forth. So, this normalization will make it independent of the starting area. So, in this way what I can do is, if I plot this particular equation. So, if I plot this data in this particular way. So, this was your experimental data let us say, I can plot it using this particular expression y bar is equal to, so I have plotted A bar, I have defined A bar is now it is normalized, change in area and A bar is defined by A i minus A at time t by A i minus A final.

So, this curve is called a sigmoidal curve. Now one of the equations which is commonly used for sigmoidal curve is a Boltzmann equation. So, here the Boltzmann equation uses the equation y is equal to 1 minus 1ne plus e to the power t minus tau 1 by tau 2. So, here also, so as opposed to a term C star. So, in the earlier fit, you had two constants C star and tau of which C star does not have any units, tau should have units of time. So, these two units have, so this is unitless and this is units of time, but now when I plot the normalize change in area, I have a curve where I cannot come up with two time constants; tau 1 and tau 2 respectively. Let me think, from this equation where exactly what is tau 1 and tau 2 mean? So, let us see what is the value when t is equal to tau 1? When at t equal to tau 1, what is the value of? So, at t is equal to tau 1, what is the value of y? So, y becomes 1 minus 1 by 1 plus e to the power tau 1 minus tau 1 by tau 2 equal to 1 minus 1 by 1 plus e to the power 0.

So, which is e to the power 0 is 1. So, it is becomes 1 minus half equal to half. So, tau 1 is essential because this value is 1, tau 1 is that value where this normalize change in area is equal to half. So, you have tau 1 as a very distinct time constant, which says when this relaxation is half done. What happens at t equal to tau 2?

(Refer Slide Time: 25:56)

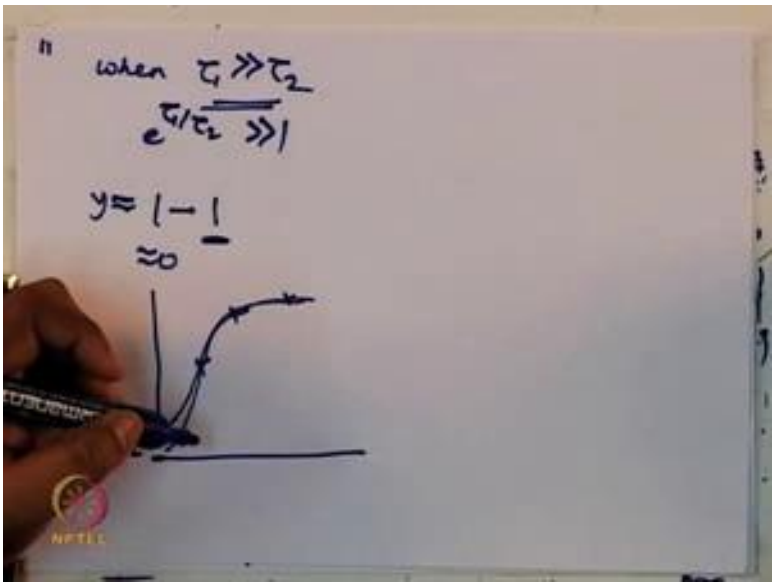


So, we have once again the equation is y is equal to 1 minus 1 by 1 plus e to the power t minus tau 1 by tau 2. What is the importance of this parameter tau 2?

Let us see, what happens when t is equal to tau 1 plus tau 2. When t is equal to tau 1 plus tau 2, what you see is the value of y becomes 1 minus 1 by 1 plus e to the power tau 2 by tau 2 is equal to 1 minus 1 by 1 plus e. So, e is roughly 2.7. So, this is roughly 1 minus 1 by 3.7, if I approximately equal to 1 minus 1 4th is equal to 75 percent. So, at t equal to tau 1 plus tau 2, you roughly reach slightly more than 75 percent. So, once again in this curve, so tau 2 is somewhere here, this is my 0.75 mark and this is my tau 1 plus tau 2; that means, this difference is my tau 2.

So, I have two time constants; tau 1 is the time takes for half of the relaxation process to be done and tau 2 is another time at which tau 1 plus tau 2, you have nearly 75 percent of the relaxation which is done. Now is this fit very good? If you look at the equation very closely, t minus tau 1 by tau 2, so what happens at t equal to 0? At t equal to 0, my y value is 1 minus 1 by 1 plus e to the power minus tau 1 by tau 2. So, y is 1 minus so is 1 minus e to the power tau 1 by tau 2 by e to the power tau 1 by tau 2 plus 1, so 1 minus e to the power tau 1 by tau 2 plus 1.

(Refer Slide Time: 28:10)



So, let us say, when e to the power tau 1 by tau 2. So, when tau 1 is much greater than tau 2 then, e to the power tau 1 by tau 2 will be much greater than 1. In that case my y is simply 1 minus. So, I can write approximately 1 minus 1 is close to 0, but if this is not true, then you accumulate an error because as from the experimental observations you know that at t equal to 0, you must have a normalized change in area is equal to 0. So, depending on the values of tau 1 and tau 2, so when your experimental data is like this let us say- you might in your fit. So, let us say this was the experimental data, in your fit you might actually get an equation like this.

So this is an error because it predicts that at t equal to 0 also you start from some normalized change in area. So, this is one of the caveats of this particular fit. So, to complete what we are discussed so far. We discussed the different aspects of things to consider when you do fitting of non linear curves, how would you change the type of equation you used to fit your data depending on the experimental data.

So, with that I thank you for your attention, and I will meet again in next lecture.

Thank you.