

Data Analysis for Biologists
Professor Biplab Bose
Department of Biosciences and Bioengineering
Mehta Family School of Data Science and Artificial Intelligence
Indian Institute of Technology, Guwahati
Lecture – 32
Nonlinear Regression

Hello everyone, welcome back. In this lecture, we will learn nonlinear regression. We have earlier learned about linear regression, when the predictor either you may have one predictor and one variable dependent variable; or you may have a situation where you have multiple predictors and one variable, depending upon the situation we have to use simple linear regression or multiple linear regression. We have learned those two things.

But life is not so simple. So, many a time although even though you wish that the data should have a linear relationship between variables, many a time the data will have nonlinear relationship.

(Refer Slide Time: 01:11)



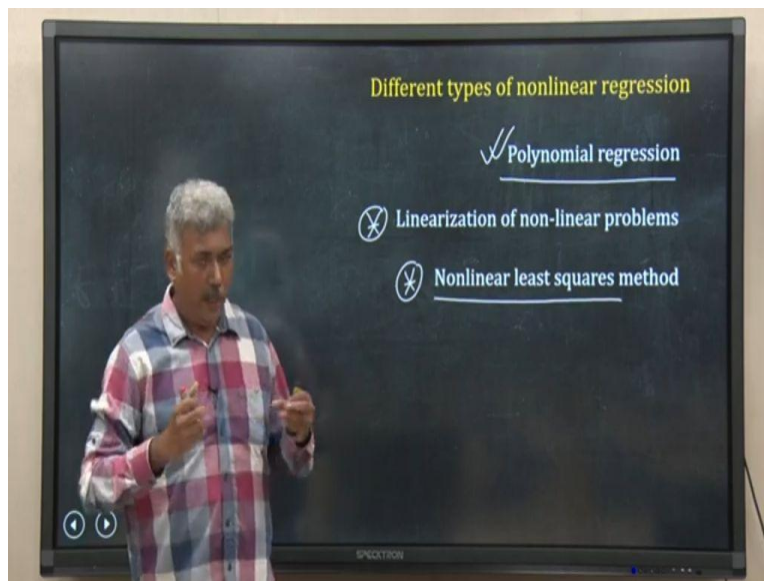
For example, take this data. I have just cooked up some data, maybe this one is your MTT data. What you have done? You have drug concentration on the horizontal axis and percentage viability of cells. Suppose you are assaying drug toxicity or something like that. So, the drug, the viability of the cells is on the vertical axis and you have plotted the data; the pink points are the

data point. So, you can easily see if you are doing cell biology experiments with the drug assays, you must be getting this type of data very frequently.

So, you can easily see we have a nonlinear decay type relationship between these two variables. So, linear regression cannot help us; we have to fit some nonlinear model. Whereas, in this case maybe this horizontal axis is time and you are measuring some property, cellular property or a property of your organism. And that maybe actually varying something like this undulating one.

Again, you need a nonlinear model. Now, for linear system we have simple linear regression. One single method if you learn that, you can actually use it universally; but for nonlinear system that is not true. There are a large number of algorithms or method to fit a nonlinear model to your experimental data. I will not discuss all of them; I will discuss only three of those.

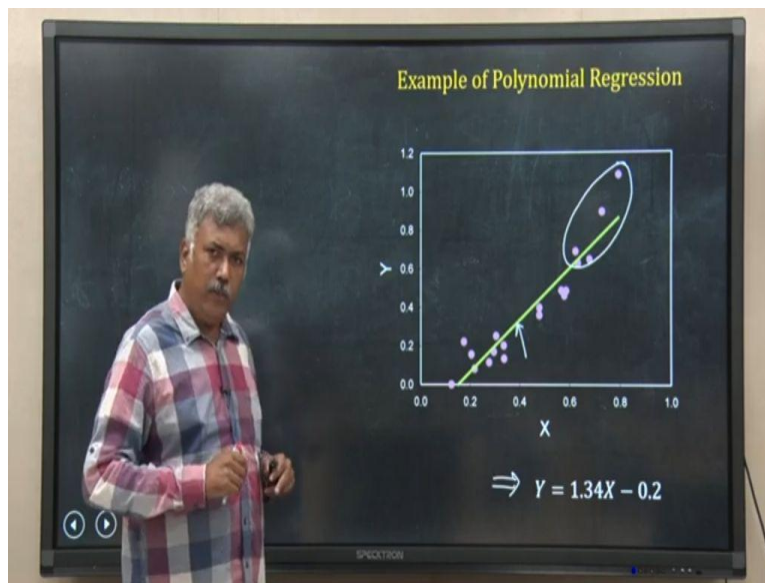
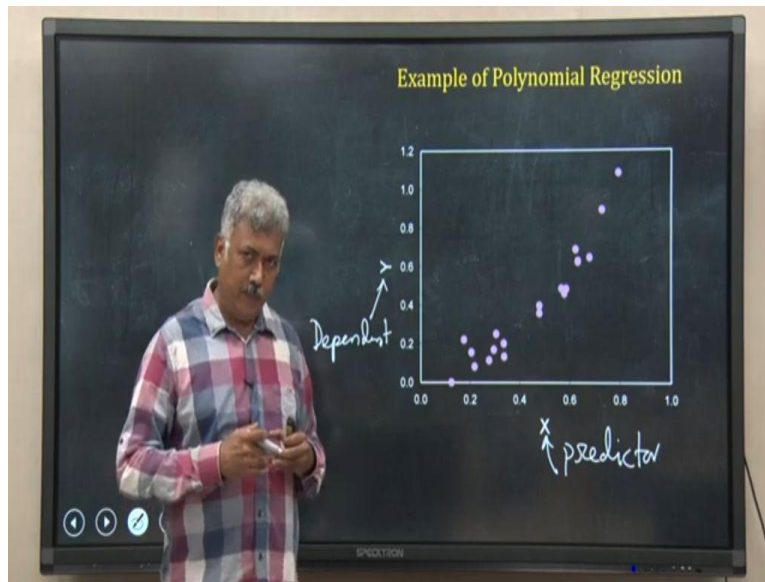
(Refer Slide Time: 02:49)

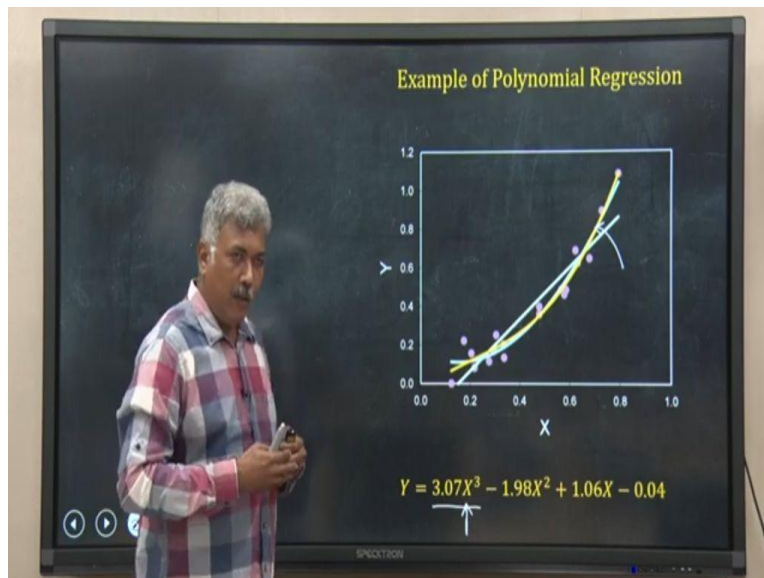
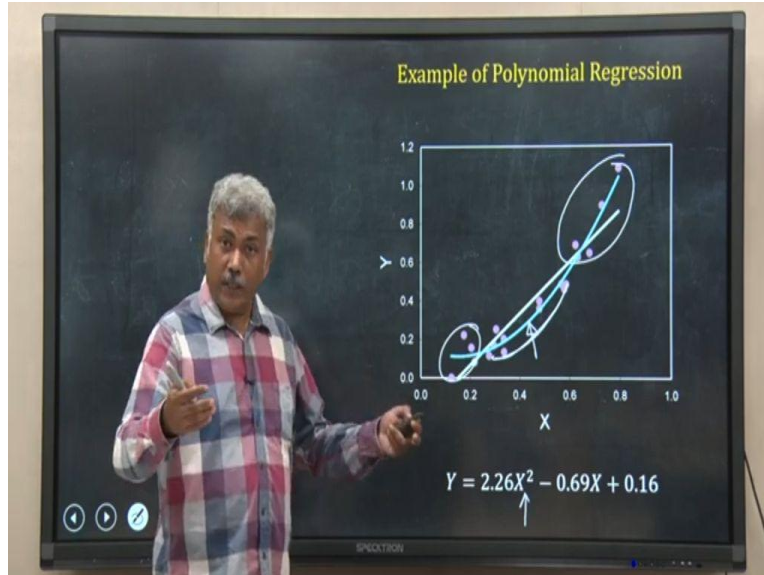


One method is that what you do? You linearize your nonlinear problem. That means you have a nonlinear equation, you convert it in some way into a linear equation, and then you use linear regression technique; we will discuss that. Another method which is called a nonlinear least square, I will also discuss that in this lecture. But, I will discuss more in detail about one method, which is called polynomial regression.

In fact, I will start this lecture with polynomial regression. And we will go into details of that how to do that; so let us start with polynomial regression.

(Refer Slide Time: 03:30)





$$1. Y = 1.34X - 0.2$$

$$2. Y = 2.26X^2 - 0.69X + 0.16$$

$$3. Y = 3.07X^3 - 1.98X^2 + 1.06X - 0.04$$

Suppose this is my data. X is a predictor or independent variable, and y is a dependent variable, dependent variable or response variable. Now, looking at the data you can easily see this data has some curvature, it is going up slightly. So, that means this is not a linear system; that means these two X and Y has some nonlinear relation between them. So, I want to do nonlinear regression and I will use polynomial regression method.

So, first thing I did just to show you what I have done I have just a simple linear regression, and I have got this equation, Y equal to $1.34x$ minus 0.2 . And that line is this green line; it has decently fitted except this part, where the data has taken a curvature, nonlinear curvature. Next what I did on the same data, I have used polynomial regression.

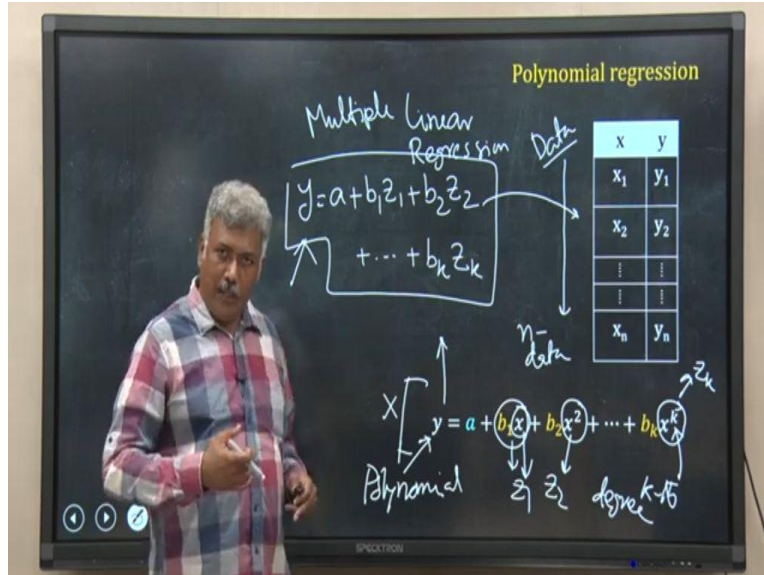
I will discuss how to do polynomial regression in few minutes; I have done polynomial regression for a quadratic equation. Why do I say it is quadratic equation? Because you can see you have x square term; and the result of that regression is Y equal to $2.26 X$ square minus $0.69X$ plus 0.16 , and that line is this blue line. Now, you can see that this regression has given me a better model.

This nonlinear line, this quadratic line blue line has a much better fit than a normal straight line that I have fitted earlier to my data. It is fitting both at this end as well as at this end and also capturing this curved area also. So, that means a quadratic model is much better for this data than a linear model. I have moved further, if I have done up to x square, why not x cube; and that is where I have the result for a cubic one.

So, here it is a cubic equation that is why the highest degree is X to the power three; and I have got this relationship by a polynomial regression and that curve is this yellow curve. And you can see it has fitted nicely to my data. And in fact, I cooked up this data for demonstration using considering a cubic equation.

And that is why actually this cubic polynomial regression is working so well; so, that is how you actually step by step can fit a polynomial. Remember here it is cubic equation, so, it is a polynomial of degree 3, to your data. So now, let me move and show how I can do this regression; and it is very easy to perform in R if you learn it properly.

(Refer Slide Time: 06:33)



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

$$y = a + b_1 x + b_2 x^2 + \dots + b_k x^k$$

So what is polynomial regression? I have data for x and y; so, I have n data points. It started x1, x2 up to xn. Correspondingly, the response variable or the dependent variables are y1, y2 and yn; and I want to fit it to this polynomial. I have written the generalized form, its a polynomial of kth degree; so the highest degree is k, so it is kth degree polynomial.

So, the degree is kth degree. Now, the equation is, the polynomial equation is a plus b1 into x, b1 is a coefficient that I have to calculate by regression, plus b2 into x square up to bk into x to the power k. Now, it may look a bit scary that I have so many terms with higher degrees, but it turned out that if we judiciously use the concept of linear regression, we can actually convert this linear regression problem.

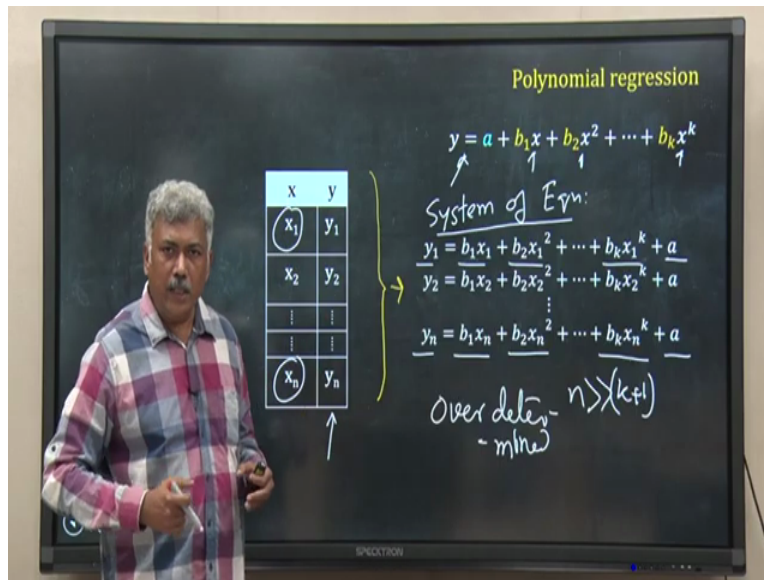
Let me explain why. So, imagine this b1 x, I consider as not b1 x, consider x as a variable z1; whereas, you consider x square as a new variable z2, whereas you consider x to the power k as zk. In this way, x to the power 3 will be considered as z3 something like that. Then, I can write

this equation, this polynomial, the original polynomial as y equal to a plus b_1z^1 plus b_2z^2 up to keep on adding b_kz^k .

Now, forget the original polynomial, forget about it, just look at this new equation that I have created for my polynomial by simply replacing the variables. x to the power k is replaced by a new variable z^k . Now, if you look at this, this is nothing but a problem; now I have to fit this one to this data. And this is nothing but a problem of simple multiple linear regression, isn't it?

I have k independent variable and they are linearly connected with the dependent variable y . So, I have to use simple multiple linear regression method and we have learned that earlier. And in fact, in R you can do the linear model for multiple linear regression; and we can use the same linear model function now for this polynomial regression also. Let us look into the linear algebra of doing multiple linear regression on this polynomial thing.

(Refer Slide Time: 09:49)



$$y = a + b_1x + b_2x^2 + \dots + b_kx^k$$

$$y_1 = b_1x_1 + b_2x_1^2 + \dots + b_kx_1^k + a$$

$$y_2 = b_1x_2 + b_2x_2^2 + \dots + b_kx_2^k + a$$

:

$$y_n = b_1x_n + b_2x_n^2 + \dots + b_kx_n^k + a$$

So, what I have to do? This is my target polynomial. I have this data, using this data I will write a system of equations. That is what we have done for simple linear regression and multiple linear regression. In this case, its just like multiple linear regression; I have k regressors, so it starts with x, x square up to x to the power k. So, that means I will take the value of x1.

So, the first equation will be y1 equal to b1x1 plus b2x1 square up to bk x1 to the power k plus a. Similarly, for the nth data, it will be yn equal to b1 xn plus b2 xn square, plus bk into xn to the power k plus a. So, you have this system of equation and obviously, just like any other linear regression problem n must be much bigger than k plus 1.

So, this will be a over determined system; so it will be over determined system, so you do not have a unique solution. That is why you have to do regression and we can do the regression least

squares method using the linear algebra approach. So, what I will do? I will convert this system of equation into a linear algebra problem.

(Refer Slide Time: 11:25)

Polynomial regression

$$y = a + b_1x + b_2x^2 + \dots + b_kx^k$$

$$\left. \begin{aligned} y_1 &= b_1x_1 + \dots + b_kx_1^k + a \\ y_2 &= b_1x_2 + \dots + b_kx_2^k + a \\ &\vdots \\ y_n &= b_1x_n + \dots + b_kx_n^k + a \end{aligned} \right\} \rightarrow$$

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} x_1 & x_1^2 & \dots & x_1^k & 1 \\ x_2 & x_2^2 & \dots & x_2^k & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x_n & x_n^2 & \dots & x_n^k & 1 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_k \\ a \end{bmatrix}$$

$(n \times 1)$ $(n \times (k+1))$ $((k+1) \times 1)$

Polynomial regression

$$y = a + b_1x + b_2x^2 + \dots + b_kx^k$$

$$\left. \begin{aligned} y_1 &= b_1x_1 + b_2x_1^2 + \dots + b_kx_1^k + a \\ y_2 &= b_1x_2 + b_2x_2^2 + \dots + b_kx_2^k + a \\ &\vdots \\ y_n &= b_1x_n + b_2x_n^2 + \dots + b_kx_n^k + a \end{aligned} \right\} \rightarrow$$

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} x_1 & x_1^2 & \dots & x_1^k & 1 \\ x_2 & x_2^2 & \dots & x_2^k & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x_n & x_n^2 & \dots & x_n^k & 1 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_k \\ a \end{bmatrix}$$

$\text{known} \quad \text{known} \quad \text{unknown}$

$Y = XB$

$$\begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{bmatrix} =$$

$$\begin{bmatrix} x_1 & x_1^2 & \dots & x_1^k & 1 \\ x_2 & x_2^2 & \dots & x_2^k & 1 \\ \dots & \vdots & \dots & \vdots & 1 \\ \dots & \vdots & \dots & \vdots & 1 \end{bmatrix}$$

Polynomial regression

$$y = a + b_1x + b_2x^2 + \dots + b_kx^k$$

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad X = \begin{bmatrix} x_1 & x_1^2 & \dots & x_1^k & 1 \\ x_2 & x_2^2 & \dots & x_2^k & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x_n & x_n^2 & \dots & x_n^k & 1 \end{bmatrix} \quad B = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_k \\ a \end{bmatrix}$$

Solution of this system of equations with minimum error:

$$\rightarrow B = (X^T X)^{-1} X^T Y$$

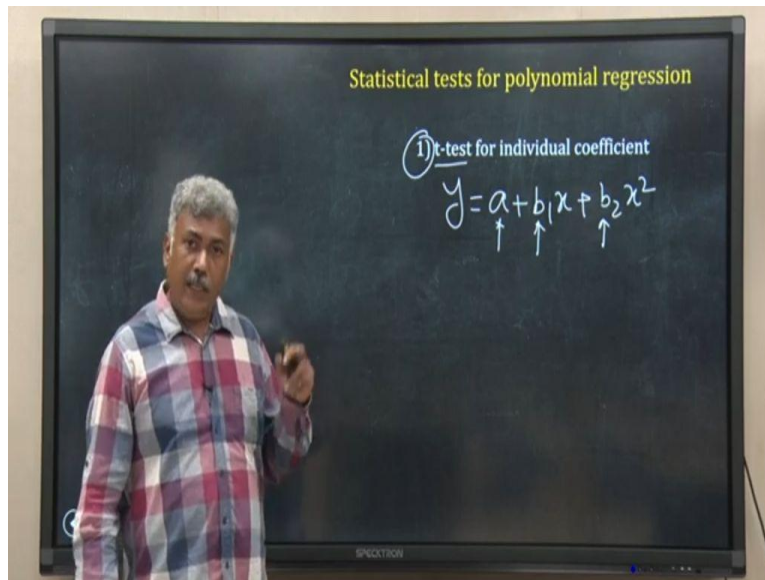
$$B = (X^T X)^{-1} X^T Y$$

And we know from the least square principle that the solution for B, the optimum solution for B will be equal to, B equal to inverse of X transpose X into X transpose into Y. So, that is how we have done in case of linear regression also, and we have to do the same here for polynomial regression also. That is why the linear model function in R will work for this polynomial regression because that can work using this simple principle.

(Refer Slide Time: 13:57)

Statistical tests for polynomial regression

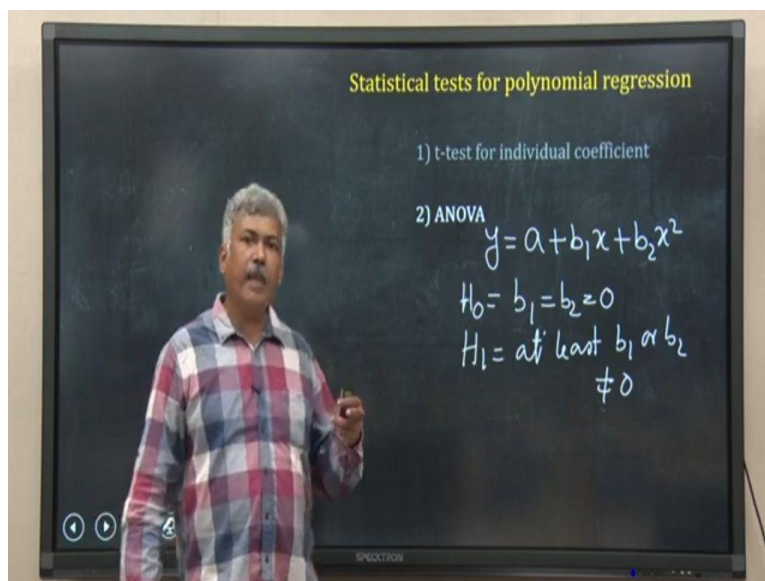
1) t-test for individual coefficient

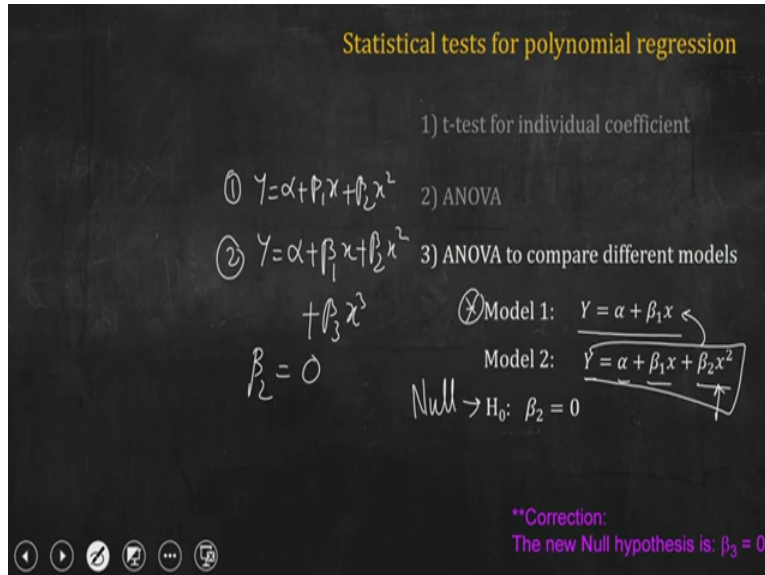
$$y = a + b_1x + b_2x^2$$


Statistical tests for polynomial regression

1) t-test for individual coefficient

2) ANOVA

$$y = a + b_1x + b_2x^2$$
$$H_0 = b_1 = b_2 = 0$$
$$H_1 = \text{at least } b_1 \text{ or } b_2 \neq 0$$




1. t-test

$$y = a + b_1 x + b_2 x^2$$

2. ANOVA

$$y = a + b_1 x + b_2 x^2$$

$$H_0 : b_1 = b_2 = 0$$

$$H_1 : \text{At least } b_1 \text{ or } b_2 \neq 0$$

3. ANOVA to capture different models

$$\text{Model 1: } Y = \alpha + \beta_1 x$$

$$\text{Model 2: } Y = \alpha + \beta_1 x + \beta_2 x^2$$

$$H_0 : \beta_2 = 0$$

Now, suppose I have done multiple linear, used this concept of multiple linear regression to do a polynomial regression essentially, I am done with the polynomial regression. Now, I want to check the quality of the regression. In case of linear regression, we have used some statistical test; if you remember we did t-test, we have done ANOVA; all those technique can be used here.

For example, let us start with t-test. I can test the significance of each of the coefficient separately using t-test. For example, take if I have done a regression for a quadratic equation, y equal to a plus b1x plus b2x square. Then, I can use t-test to check the (coeffi) significance of a,

b_1 and b_2 . That is what you have done in case of linear regression also, simple linear regression also.

Now, in case of multiple linear regression, if you remember we have learned ANOVA; here also, I can do ANOVA. So, what I will do in case of ANOVA for this problem? Suppose I am doing with a quadratic equation, I am using polynomial regression. So, y equal to a plus b_1x plus b_2x^2 , so I can make a null hypothesis H_0 is equal to b_1 equal to b_2 equal to 0. And obviously, the alternate hypothesis will be that at least b_1 or b_2 is not equal to 0, and then you can do ANOVA.

There is another type of ANOVA very useful when we are doing polynomial regression. Because remember when I am doing polynomial regression as I have shown in the example, I started with the first fitting the linear equation, simple linear regression of y equal to bx plus a . Then, what I tried? I have tried a quadratic equation, then I have tried cubic equation. I could have gone up to x to the power 4, x to power 5.

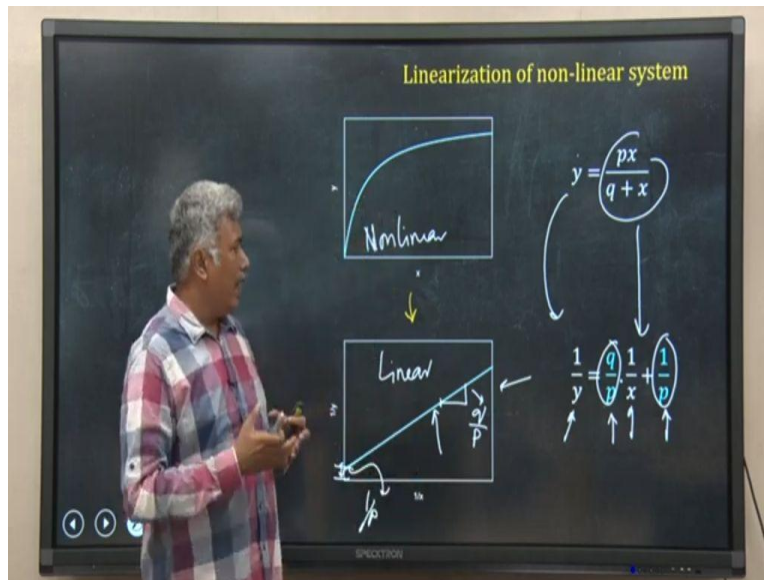
But, adding these extra term, extra term with higher degrees, we are helping in regression. Are they helping in making the right model? I can check by a comparative ANOVA. What I can do? Suppose the first I create a model, which is simple linear one, Y equal to α plus $\beta_1 x$. Then, I add another extra term and make it Y equal to α plus $b_1 x$ plus $b_2 x^2$.

Now, my question is, is adding this extra term $\beta_2 x^2$ is actually helping me in the model and I will use ANOVA for that. What I will do? In this case I will have the null hypothesis the H_0 is equal to β_2 equal to 0. And obviously, the alternate hypothesis will be β_2 is not equal to 0. Now, you calculate the F statistics, check the probability of that and the if the probability is such that you can actually reject null hypothesis; that means, this model is better than this model.

Now, once you have done up to quadratic, now you can create a new model. So, now what you can do? You can compare between y α plus $\beta_1 x$ plus $\beta_2 x^2$; and y equal to α plus $\beta_1 x$ plus $\beta_2 x^2$ plus $\beta_3 x^3$. So, this is model 1, this is model 2 now; you can again create a new hypothesis. What is the new hypothesis that β_2 is 0 that is my new null hypothesis. And I can check that new null hypothesis using ANOVA.

So, in this way step by step, I can actually use ANOVA to check whether adding a new polynomial term in my polynomial regression is actually helping or not. So, that way you can choose when to stop; should I stop at the cubic equation, or I should go up to x to the power 4 or higher degree terms in my polynomial or not. Now, we have done with polynomial regression; now I will move into two other type of regression, which we will learn in this in this lecture for nonlinear system.

(Refer Slide Time: 18:44)



$$y = \frac{px}{q+x}$$

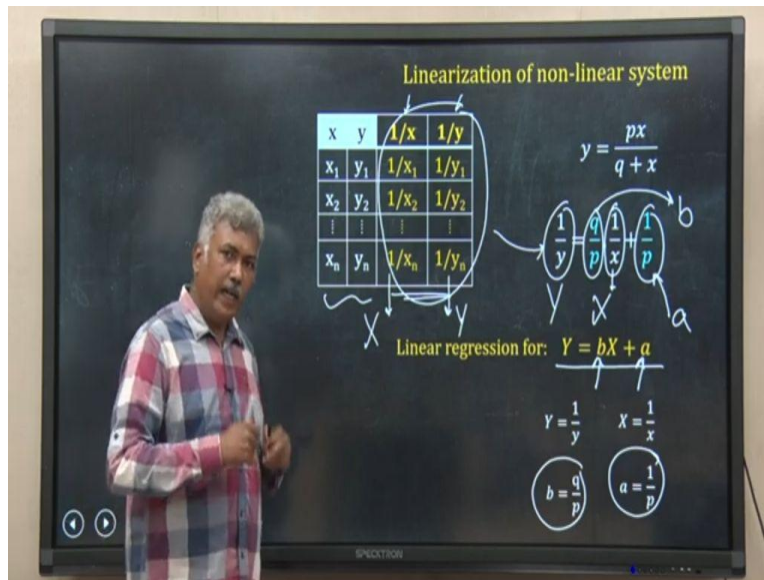
$$\frac{1}{y} = \frac{q}{p} \cdot \frac{1}{x} + \frac{1}{p}$$

The first one is linearization of a nonlinear system. What I have shown here? I have a equation something, this is Michaelis-Menten type equation if you can recognize. y equal to p into x divided by q plus x ; x is the independent variable or predictor, y is the dependent variable or the response variable, p and q are constant. So, it will look like if you plot this function it will look like this, a typical enzymatic reaction diagram that you have seen in your biochemistry textbook.

So, now, if I have to fit data to this type of nonlinear function, I can use a very simple trick. What is that? See, I will use some mathematical technique to simply convert this nonlinear system into linear one. What I will do? I will just invert. What I will do? I will invert y to 1 by y ; and so if I invert y , I have to invert this side also and I will get this one. So, you simply do arithmetic what you will get, algebra what you will get?

1 by y is equal to p divided by q into 1 by x plus 1 by p . And now you plot that one you see you have a straight line. So, I have converted a nonlinear system, this is nonlinear to a linear system; slope of this straight line is q by p and the intercept is 1 by p . So, now I can use simple linear regression to calculate the value of p by q and 1 by p . And if I can get q by p and 1 by p , then obviously I can calculate the value of p and q individually. So, let me show you.

(Refer Slide Time: 20:49)



x	y	1/x	1/y
x1	y1	1/x1	1/y1
x2	y2	1/x2	1/y2
:	:	:	:
xn	yn	1/xn	1/yn

$$y = \frac{px}{q+x}$$

$$Y = bX + a$$

$$Y = \frac{1}{y} \quad X = \frac{1}{x}$$

$$b = \frac{q}{p} \quad a = \frac{1}{p}$$

So, suppose this is the data, x versus y data and I have n data points x1, x2 up to xn; and correspondingly I have y1, y2 up to yn and I want to fit this data to this equation. So, what I have done earlier in the slide that I have actually done a inversion; so, I will invert it to make a linearized equation. So, that is why I have to invert both y and x data also; so, that is what I have done; I have taken 1 by x and 1 by y.

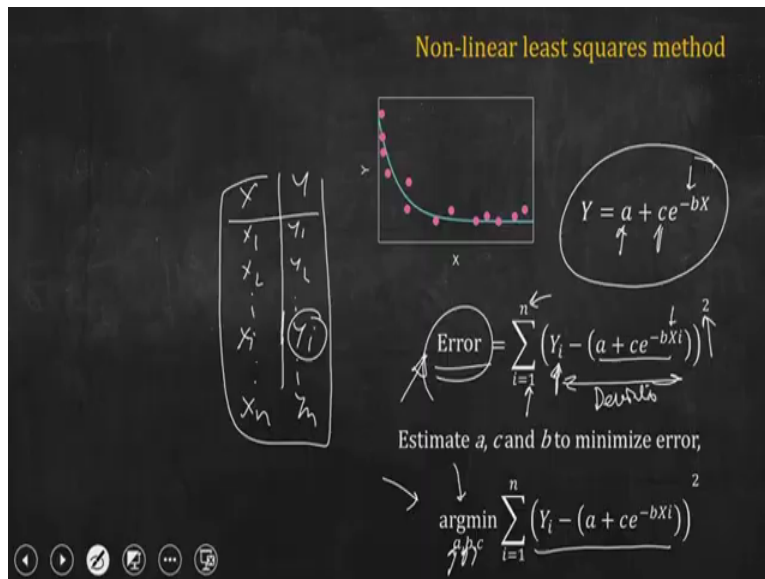
So, now I have to fit these data to this equation 1 by y equal to q by p into 1 by x plus 1 by p. Now, you can imagine I can consider this as capital Y, this as capital X, this one q by p I can say

b; and 1 by p, I can say it is a. Then, what do I get? I get Y equal to b into capital X plus a; and what is Y? This column is my capital Y and this column is my capital X.

So, I have a problem of simple linear regression and I will use a simple linear regression tool, whichever tool you are using. If you are using R, use the linear model function and fit the data, not the original data; but, this transformed data 1 by x, 1 by y data, calculate the value of b and a. And once you have calculated the b and a using this relationship, you calculate the value of p and q, and you are done.

Now, I have given one example. You can actually cook up lots of common functions that we face in biology in nonlinear functions, which can be actually linearized this way from nonlinear to linear; and you can simply use the linear regression method to fit the data. Now, let us look into another approach.

(Refer Slide Time: 22:57)



$$Y = a + ce^{-bX}$$

$$\text{Error} = \sum_{i=1}^n (Y_i - (a + ce^{-bX_i}))^2$$

$$\operatorname{argmin}_{a,b,c} \sum_{i=1}^n (Y_i - (a + ce^{-bX_i}))^2$$

This approach is called nonlinear least square method. What we are doing here? Again suppose this is a MTT data. So, I have drug concentration on this axis, percentage viability on the vertical

axis; and I want to fit an exponential decay equation, Y equal to $a + c \cdot e^{-bx}$; so, I have to calculate three coefficients a , c and b . So, it is a nonlinear least squares method, that means we will still use the least squares approach.

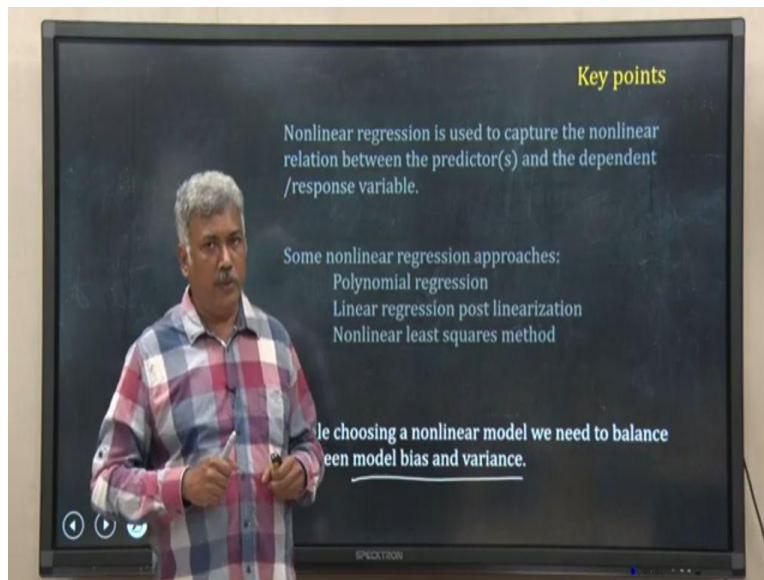
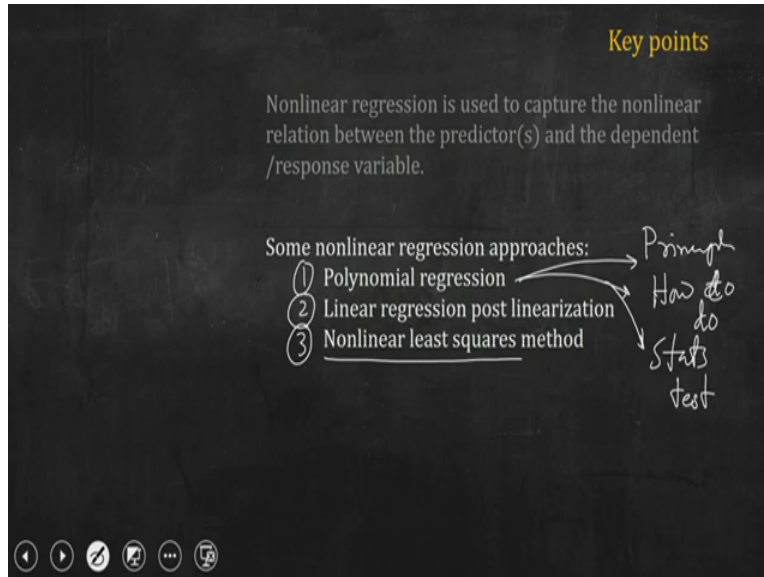
That means, we have to have some parameter which will be error term and we want to reduce that error. So, how should I define the error in this case, it's the least square method. So, I will define the error as we have defined in case of other least square method. Y_i , Y_i is an individual data; I have x versus y data. So, I have $x_1, y_1, x_2, y_2, \dots, x_i$ correspondingly y_i and then up to x_n, y_n . So, this Y_i a particular data Y_i is equal to $a + c \cdot e^{-bx_i}$, the corresponding X_i .

So, that is the deviation, this subtraction is the deviation; deviation of the predicted value from the real value. You take the square of that and sum it for all the data points from 1 to n ; so, this is nothing but sum of square error. So, what do you have to do now, you have an optimization problem; you have to calculate the value of a , c and b , such that for the given data set, this error is minimized.

And that is what I have written here; and that is what you do in any other least square method. You want to minimize this sum of square error such using by varying the value of a , b and c . Now, in case of simple linear regression, we know the method; I will not go in detail of the method used for this type of problem, because the method may vary depending upon the nonlinear equation that we are trying to fit.

There are multiple algorithms, but all of them eventually will try to minimize this error as I have written here. So, this is called nonlinear least square method. So, we are done for this lecture, let me jot down what we have learned in this lecture.

(Refer Slide Time: 24:46)



In this lecture, we have learned about nonlinear regression. There will be cases frequently where the predictor variables one or more has a nonlinear relation with the dependent or the response variable. In those cases, I have to use nonlinear regression. There are several approaches to do nonlinear regression; I have only discussed three of them.

One is polynomial regression that I have discussed in details; also I have a linear regression after linearization of the nonlinear system. That is very simple if you can linearize the system and the third one that I have discussed, I have discussed about nonlinear least square method. Now, in polynomial regression, what although I have not written here, what you have to understand

clearly the principle, how we are doing it; then how we are doing that and we also understand the statistical test that you do.

So, you should focus on these three things for polynomial regression, as we have discussed in the lecture. That is all for this video. Thank you for learning with me today.