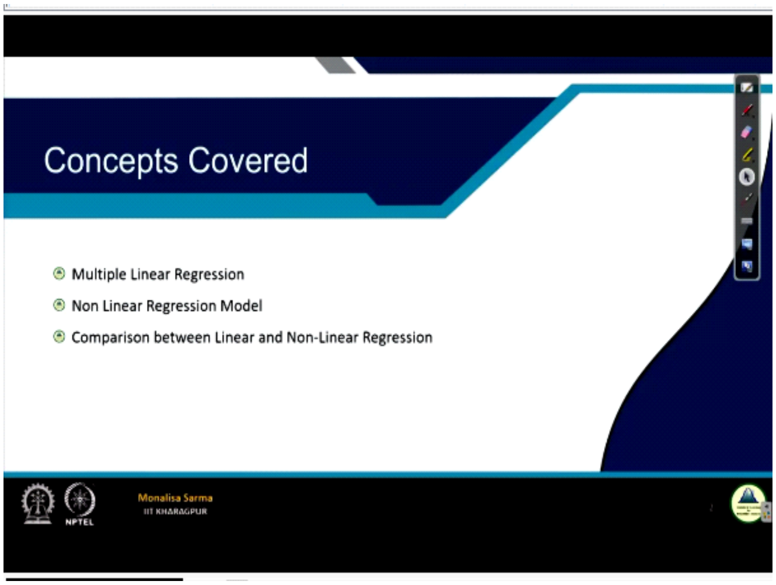
Statistical Learning for Reliability Analysis
Prof. Monalisa Sarma
Subir Chowdhury School of Quality and Reliability
Indian Institute of Technology-Kharagpur

Lecture-42
Regression Analysis (Part-III)

Hello guys, so in continuation of our discussion in relationship analysis, basically we were discussing linear regression. In my last lecture I discussed linear regression.

(Refer Slide Time: 00:38)

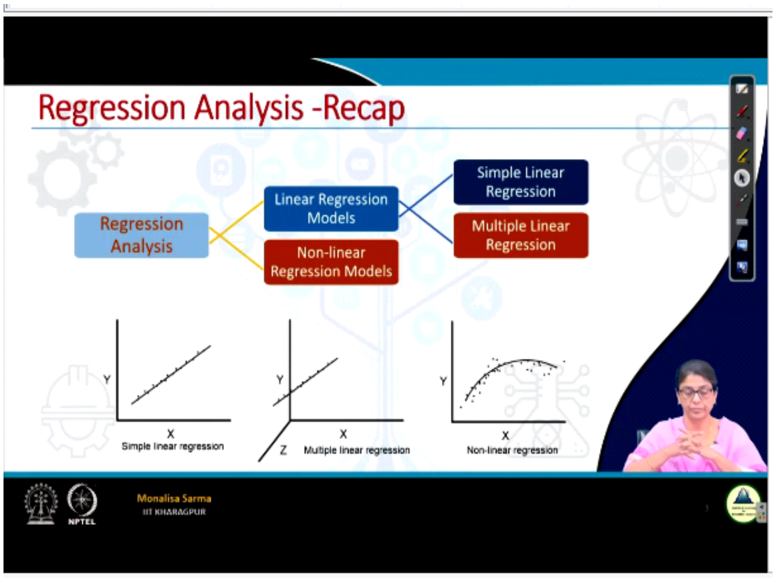


Concepts Covered

- Multiple Linear Regression
- Non Linear Regression Model
- Comparison between Linear and Non-Linear Regression

So, today's lecture I will discuss multiple linear regression, non-linear regression model and then I will try to see a compression between this linear and non-linear regression model. So, when I talk of multiple linear regression means there will be more than one factor; one factor means which factor more than one independent factor. Like a very good example is the same example I want to give basically which I have given is the price of the house, it not only depends on the square feet it also depends on the age of the house, there are two factors, square feet as well as age.

(Refer Slide Time: 01:10)



Regression Analysis -Recap

Regression Analysis

Linear Regression Models

Non-linear Regression Models

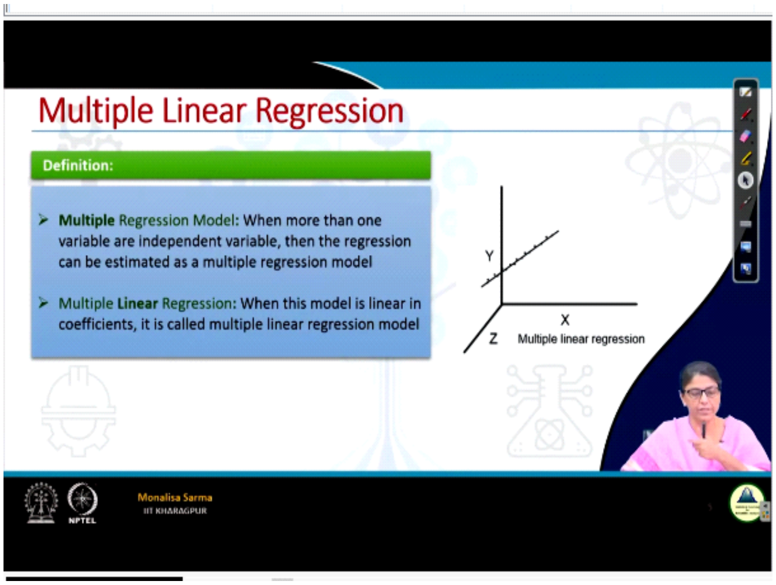
Simple Linear Regression

Multiple Linear Regression

So, that is basically multiple linear regression, if the regression line is a straight line then it is a multiple linear regression. If the regression line is not a straight line it is a curve then it is a multiple non-linear regression. So, this figure already we have seen in my last lecture also. So, just a quick recap. So, this is the simple linear regression, simply it features straight line y = mx.

So, mx + c or y = α + β x whichever you take. So, now this is the multiple linear regressions where if it is a linear regression but more than one factor, so we will consider a straight line in a multi-dimensional plane, it is a non-linear regression where basically it does not graph a straight line but maybe a curved line, it may be a parabola, it may be hypo anything like.

(Refer Slide Time: 02:10)



Multiple Linear Regression

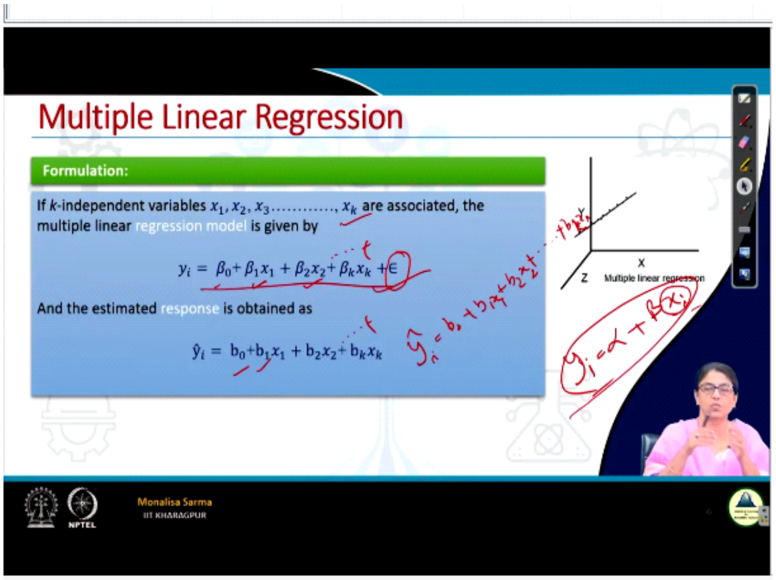
Definition:

- **Multiple Regression Model:** When more than one variable are independent variable, then the regression can be estimated as a multiple regression model
- **Multiple Linear Regression:** When this model is linear in coefficients, it is called multiple linear regression model

So, when more than one variable are independent variable, I am just talking dependent variable is one only. So, when more than one variable are independent variable then the regression must be estimated as multiple regression model. So, more than one independent variable, so like in this example in the price of the house there can be some more other example we can think of anyway forget it.

One example is sufficient; you got the point basically that is the most important thing. So, now this multiple linear regression, when this model is linear in its coefficient it is called a multiple linear regression model, when they have more than one factors and the model is linear then we call it a multiple linear regression.

(Refer Slide Time: 03:04)



Multiple Linear Regression

Formulation:

If k-independent variables $x_1, x_2, x_3 \ldots\ldots\ldots\ldots, x_k$ are associated, the multiple linear regression model is given by

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_k x_k + \epsilon$$

And the estimated response is obtained as

$$\hat{y}_i = b_0 + b_1 x_1 + b_2 x_2 + b_k x_k$$

$\hat{y}_i = b_0 + b_1 x_1 + b_2 x_2 + \dots + b_k x_k$

$y_i = \alpha + \beta x_i$

Monalisa Sarma
IIT KHARAGPUR

So, now multiple means there can be more than one factor, more than one independent variable. So, let us consider k independent variable x 1, x 2 up to x k. Suppose these are associated, now the multiple linear regression model is this. Earlier what was my regression model is y i = α + β x i. But at time using i is a subscript there may be n data and x i data, for n x I am getting y and y i data, y data.

So, that is why this subscript i. So, this is earlier my first simple linear regression this was my model, y i = α + β x i. Now there are more than one independent variable; earlier just one independent variable that is x. Now there are more than one x independent variable. Let us take there are k independent variable. So, if there are k independent variable x 1 to x k. So, the regression model will be in this form β 0 that is the intercept β 1 x 1, β 2 x 2 **+,** so it will be dot dot dot + β k x k.

There is a slight mistake here; it is β 2 x 2 + dot dot + β k x k + ε. Why again this ε is there because this is not a deterministic equation. As we have mentioned why it is not a deterministic equation because of this random component that is why it is making it y i as a random, we cannot deterministically say that for particular value of x 1 x 2 up to x k we can knowing the coefficient will be able to tell the value of y i.

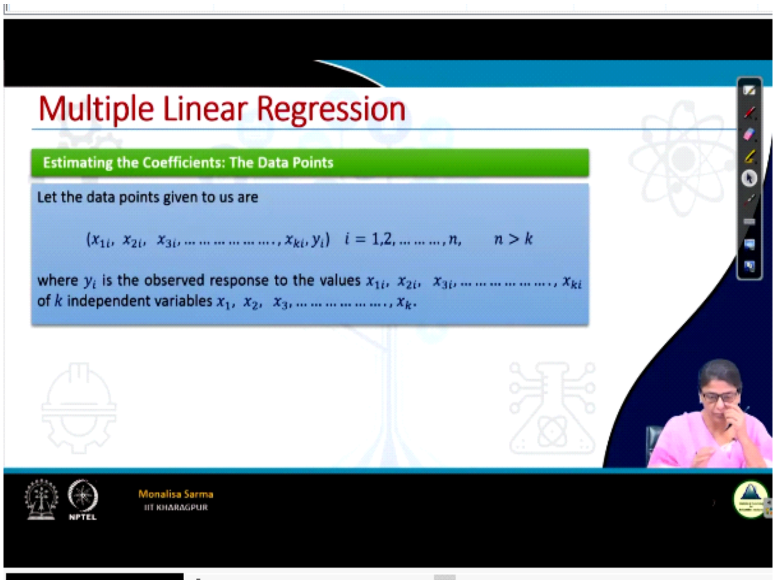
No we cannot deterministically say there is always a randomness to it and now we need as same as what we have done for linear regression. So, we need to estimate the value of the coefficient, here the coefficient is β 0 β 1 β 2 β 3 β 4 up to β k. So, we have to estimate this value of this coefficient. Now how we have to estimate the same method least square method what we have used in the linear regression method, for that first what we will do? We will first assume some values for this coefficient.

Now since there are k variables, when there are, k variables how many coefficients are there? K variables then there are total k + 1 coefficient, there are one variable, there are two coefficient α and β. So, there are k variables, so there are total k + 1 coefficient. So, what is the k + 1 coefficient β 0 β 1 β 2 up to β k. So, now suppose we have assumed as value for β 0 as b 0, β 1 as b 1.

So, plus this equation I am just writing one equation both the equations are wrong basically y i cap = β 0 + β 1 x + β 2 x and β 1 x 1 β 2 x 2 + dot, dot + β k x k. So, here basically dot, dot means it goes from x to x l, it is not only it is just showing as if x 1 x 2 x k that is wrong. So, this is my if I have estimated the value for β 0 β 1 β 2 I have estimated as to be b 0 b 1 b 2.

So, if that is the case, this is my predicted value y i cap is my predicted value. If my y i cap is the predicted value, I need to estimate this parameter I will be using the least square method only. Now using this same method again I will have to find out the residual. So, what is the residual for each y i value. So, how do I find out a residual? Same method y i - y cap. So, this residual may be positive, it may be negative. So, that I get a proper residual, I will square it, so, sum of all this residual square.

(Refer Slide Time: 07:24)



Multiple Linear Regression

Estimating the Coefficients: The Data Points

Let the data points given to us are

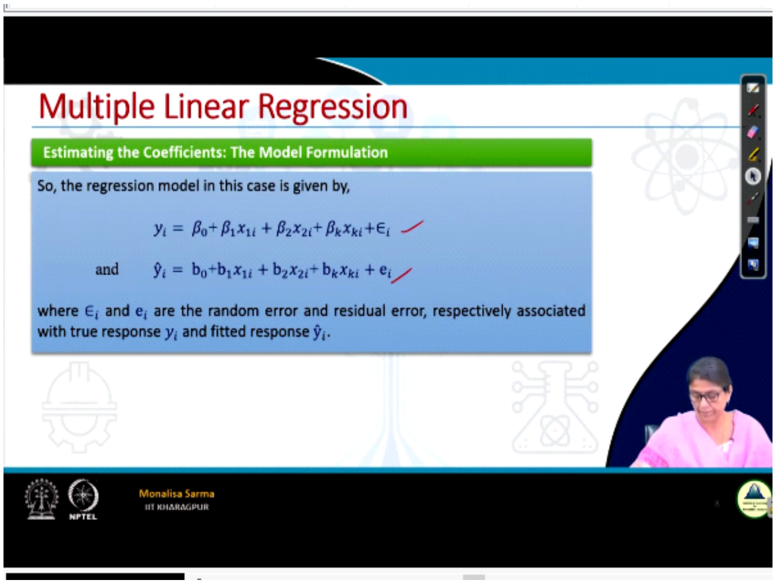
$$(x_{1i},\ x_{2i},\ x_{3i}, \dots\dots\dots\dots\dots\dots, x_{ki}, y_i) \quad i = 1,2, \dots\dots\dots, n, \qquad n > k$$

where y_i is the observed response to the values $x_{1i},\ x_{2i},\ x_{3i}, \dots\dots\dots\dots\dots\dots, x_{ki}$ of k independent variables $x_1,\ x_2,\ x_3, \dots\dots\dots\dots\dots\dots, x_k$.

Monalisa Sarma
IIT KHARAGPUR

So, say and here let the data points to be used are these are my data points, for y i where i = 1 to n, so y i is the observed response to the values this of k independent variable. This is if you have understood the previous one simple linear equation, multiple linear results exactly the same just that instead of one variable I am using here more than one variable that is all everything remains same.

(Refer Slide Time: 07:55)



Multiple Linear Regression

Estimating the Coefficients: The Model Formulation

So, the regression model in this case is given by,

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_k x_{ki} + \epsilon_i$$

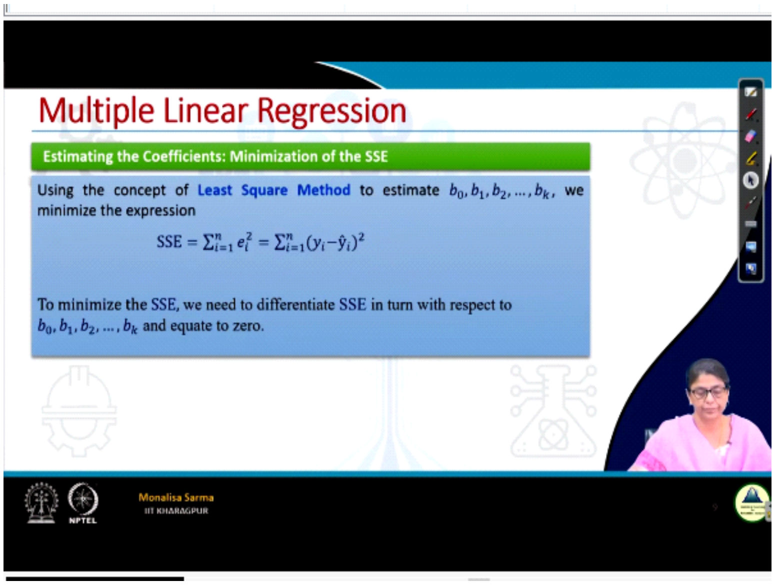
and $$\hat{y}_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + b_k x_{ki} + e_i$$

where ϵ_i and e_i are the random error and residual error, respectively associated with true response y_i and fitted response $\hat{y}_i$.

Monalisa Sarma
IIT KHARAGPUR

So, the regression model in this case is this is the ε i is a random component. Now y cap also and this is e i; where ε i is a random error and e i is the residual error. Residual error means the predicted value and minus the actual value is what I get is that that is the residual error. So, where ε i and e i the random error, residual error respectively associated with a two response y i and a fitted response y i cap.

(Refer Slide Time: 08:29)



Multiple Linear Regression

Estimating the Coefficients: Minimization of the SSE

Using the concept of Least Square Method to estimate $b_0, b_1, b_2, \dots, b_k$, we minimize the expression

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

To minimize the SSE, we need to differentiate SSE in turn with respect to $b_0, b_1, b_2, \dots, b_k$ and equate to zero.

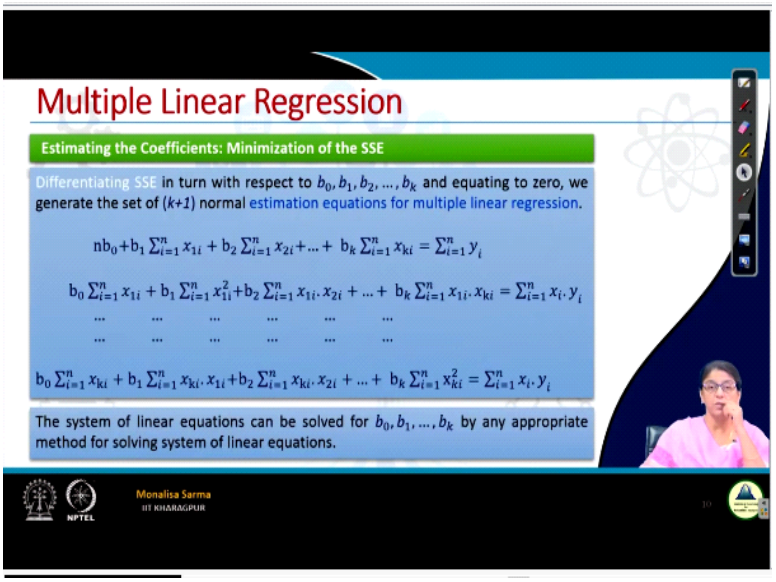
So, using the same concept of least square method, I will use the same to estimate this b. So, what is my SSE? SSE is nothing but the residual error square, sum of the all the residual error square. So, what is the sum of this residual error square is nothing but this y i - y cap whole square summation up from i = 1 to n, i = 1 to n means all the data points what I have taken.

To minimize SSE we need to differentiate SSE in terms with respect to b 1, b 2, b k = 0. Now we have to find out the value of the parameters b 1, b 2, b 3 whatever we have assumed we have to find out the values of b 0, b 1 up to b k. So, how to find out the we have to find out the value means now basically our SSE should be minimum or fitted line should be such that the SSE should be minimum.

So, the same concept how do I find a value for which value the function is minimum, so how do I that to find out the for which value of x the function is minimum basically I do the differentiation of that first order differentiation and then equate it to 0. Similarly, whatever we have done it for simple linear regression I will differentiate SSE now here earlier I have differentiated SSE for partial differentiated SSE with respect to α and with respect to β.

Now I will differentiate SSE with respect to b 0, b 1, b 2 up to b k and then equate all this to 0 and then once I equate all this to 0, so total how many variables I need to find out?

(Refer Slide Time: 10:00)



Multiple Linear Regression

Estimating the Coefficients: Minimization of the SSE

Differentiating SSE in turn with respect to $b_0, b_1, b_2, \dots, b_k$ and equating to zero, we generate the set of (k+1) normal estimation equations for multiple linear regression.

$$\mathrm{nb}_0+\mathrm{b}_1\sum_{i=1}^{n} x_{1i}+\mathrm{b}_2\sum_{i=1}^{n} x_{2i}+\dots+\mathrm{b}_k\sum_{i=1}^{n} x_{\mathrm{k}i}=\sum_{i=1}^{n} y_i$$

$$\mathrm{b}_0\sum_{i=1}^{n} x_{1i}+\mathrm{b}_1\sum_{i=1}^{n} x_{1\mathrm{i}}^2+\mathrm{b}_2\sum_{i=1}^{n} x_{1i}.x_{2i}+\dots+\mathrm{b}_k\sum_{i=1}^{n} x_{1i}.x_{\mathrm{k}i}=\sum_{i=1}^{n} x_i.y_i$$

$$\begin{matrix} \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots \end{matrix}$$

$$\mathrm{b}_0\sum_{i=1}^{n} x_{\mathrm{k}i}+\mathrm{b}_1\sum_{i=1}^{n} x_{\mathrm{k}i}.x_{1i}+\mathrm{b}_2\sum_{i=1}^{n} x_{\mathrm{k}i}.x_{2i}+\dots+\mathrm{b}_k\sum_{i=1}^{n} \mathrm{x}_{ki}^2=\sum_{i=1}^{n} x_i.y_i$$

The system of linear equations can be solved for $b_0, b_1, \dots, b_k$ by any appropriate method for solving system of linear equations.

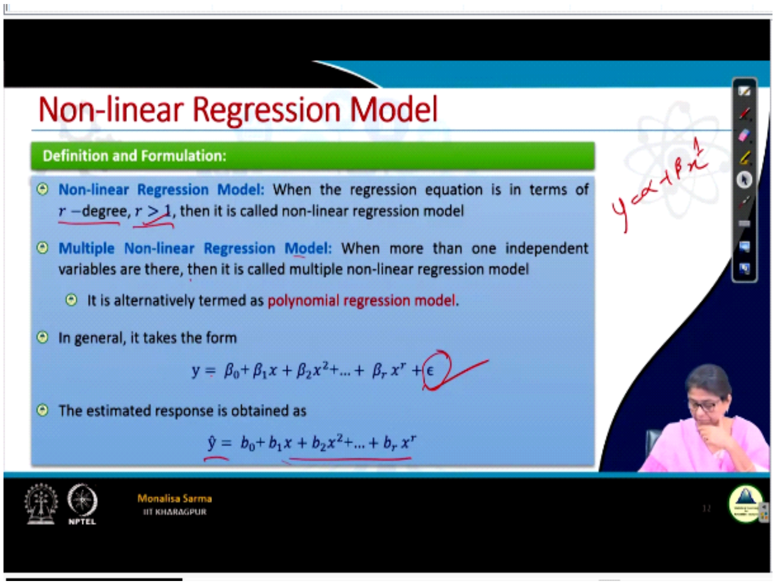
Monalisa Sarma
IIT KHARAGPUR

K + 1 variables. So, differentiating SSE in terms with respect to and equating to 0 we generate total k + 1 estimation equation. I will get total k + 1 because a SSE I will differentiate with respect to all these parameters. How many parameters are there? K = 1 parameters, when I differentiate SSE with respect to all these k = 1 parameters I will get to a set of k + 1 equations.

And I will put that equations to 0, equalize it to 0, so that I can get the value of the parameters b 1, b 2. So, now this is a system of linear equations. The system of linear equation you can solve by any method. So, you can solve it by any method and then by solving that and you can find out the values of the b 0, b 1, b 2 which basically, once we find out the b 0, b 1, b 2 then we got the regression model. Basically the mathematical model that gives us the relationship between all the x to the y, x 1, x 2, x k, 2 y.

So, that is nothing the same thing whatever we learn for simple linear regression same thing it is for multiple linear equation. Similarly for multiple linear equations also we can do for quality of it for that we can again find out the r ^{2} which I am not discussing it here.

(Refer Slide Time: 11:26)



Non-linear Regression Model

Definition and Formulation:

- **Non-linear Regression Model:** When the regression equation is in terms of r −degree, $r > 1$, then it is called non-linear regression model
- **Multiple Non-linear Regression Model:** When more than one independent variables are there, then it is called multiple non-linear regression model
  - It is alternatively termed as polynomial regression model.
- In general, it takes the form

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_r x^r + \epsilon$$

- The estimated response is obtained as

$$\hat{y} = b_0 + b_1 x + b_2 x^2 + \ldots + b_r x^r$$

$y = \alpha + \beta x^1$

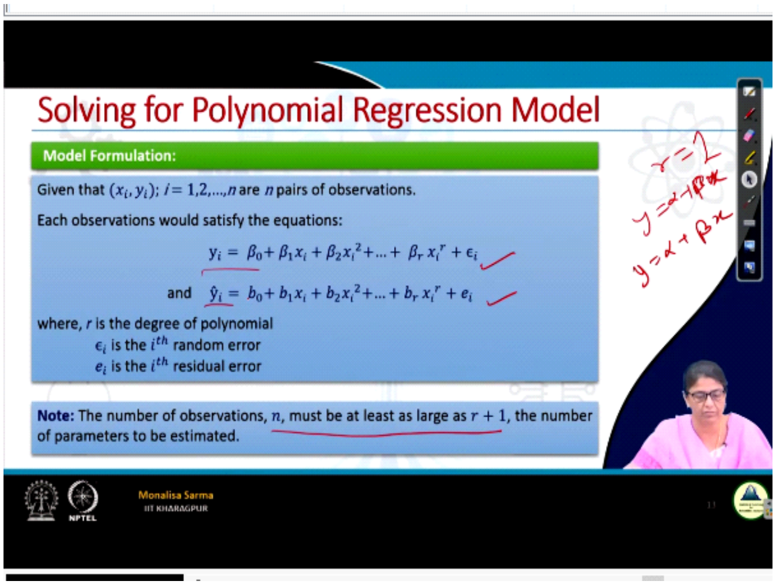
Monalisa Sarma
IIT KHARAGPUR

So, now non-linear regression model: So, non-linear when we called it is a non-linear regression model when the regression equation is in terms of r-degree, where r is greater than 1. Earlier y = α + β x, degree of x is what? Degree of x is 1. So, this is linear, we get a straight line. Now when the; regression equation is terms of r-degree, where r is greater than 1 then it is called non-linear regression model. So, x may be x ^{2}, x ^{3}, x 4 whatever it is.

Now it is simple non-linear. Now again multiple non-linear, so multiple non-linear when more than one independent variables are there then it is called multiple non-linear regression model, it is also termed as polynomial regression model. So, when there are more than one independent variable as well as degree is also more than one then we call it as a multiple non-linear regression model or we also call it a polynomial regression model.

So, in general it takes this form b 0, b 1 x + β 2 x ^{2} + + up to β r x r + ε. This ε is a random component that will always be there. Similarly the estimated response same way square methods, so, this is y cap we have estimated b 0 as the value of β 0, b 1 is the value of b 1 and β 1 b 2 is the value of β 2. We have estimated this value then we try to find out the residual, how is the residual, summation of all residual square will get it SSE. So, that is y i - y cap.

(Refer Slide Time: 13:02)



Solving for Polynomial Regression Model

Model Formulation:

Given that (x_i, y_i); $i = 1,2,\ldots,n$ are n pairs of observations.

Each observations would satisfy the equations:

$$\underline{y_i} = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \ldots + \beta_r x_i^r + \epsilon_i \checkmark$$

and $\underline{\hat{y}_i} = b_0 + b_1 x_i + b_2 x_i^2 + \ldots + b_r x_i^r + e_i$ ✓

where, r is the degree of polynomial

ϵ_i is the i^{th} random error

e_i is the i^{th} residual error

Note: The number of observations, n, must be at least as large as $r + 1$, the number of parameters to be estimated.

$r = 1$

$y = \alpha + \beta x$

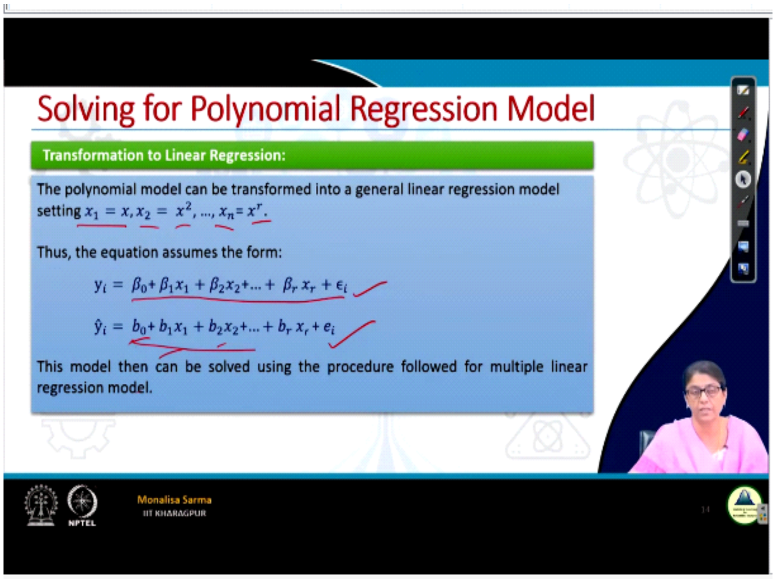
$y = \alpha + \beta x$

Monalisa Sarma
IIT KHARAGPUR

So, this is the y i, this is y i cap, so now here the number of observation n must be at least as large as r + 1. The number of parameters to be estimated, so, how many parameters we have to estimate? Total r + 1 parameters, when my r = 1 y = α + β x, here my r is 1, when r is 1 how many parameters I have to estimate? Two parameters is not it.

So, total how many equations I got? Two equations, so number of observations that n must be at least as large as r + 1. So, total I have to take number of observation means total data set I have to take at least as large as r + 1 the number of parameters to be estimated then only I will be able to estimate all the parameters. So, this is y i y i cap, same nothing difference here.

(Refer Slide Time: 14:16)



Solving for Polynomial Regression Model

Transformation to Linear Regression:

The polynomial model can be transformed into a general linear regression model setting $x_1 = x, x_2 = x^2, \dots, x_n = x^r$.

Thus, the equation assumes the form:

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_r x_r + \epsilon_i$$

$$\hat{y}_i = b_0 + b_1 x_1 + b_2 x_2 + \dots + b_r x_r + e_i$$

This model then can be solved using the procedure followed for multiple linear regression model.

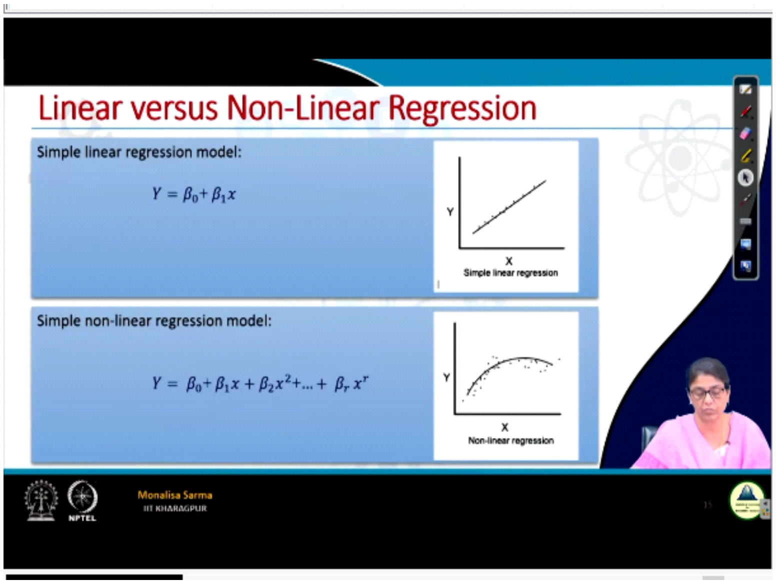
So, now this is in a polynomial form, non-linear form, is not it? Very much in a non-linear form. Nowadays nonlinear form it is difficult to solve it I can convert this non-linear to linear. How I can convert this nonlinear to linear? Say if I write x 1 = to x, x 2 = x ^{2}, x 1 = x r. I have used different variables x 1, x 2, x 3 are different variables to indicate the different powers of x.

X 1 = x, x 2 = x ^{2} then x n = x r = x to the power basically. So, initially this was my equation. So, this equations this b 0 + this equation β 0 + β 1 x i β 2 x i ^{2}. So, this equation I can write it in this form in a linear form. Similarly my y cap will be in this form. So, this model can be solved using the procedure followed by multiple linear. But with the same procedure what we have followed for multiple linear regression models because here we were getting a polynomial expression where our degree is more than 1.

So, if I want to estimate the parameters then it is becoming complicated. So, I want to solve it using the same method, the least square method estimate. So, that what I have done? This equations from linear non-linear I have made it linear, just that what I have done? I have used more number of variables, different variables I have used to indicate the different degrees. So, this expression is same as what I have used in my multiple linear regressions the same expression this too is not it? So, how I have solved it? I will find out the SSE then basically what is SSE is the summation of the residual square.

And then to find out the different values of this parameter what I will do? I will differentiate SSE with respect to all this parameter β 0 b and β 2 and all. Then based on that I will get the value for this and once I get the value for this I can find out the regression line. That is the regression model.

(Refer Slide Time: 16:33)



Linear versus Non-Linear Regression

Simple linear regression model:

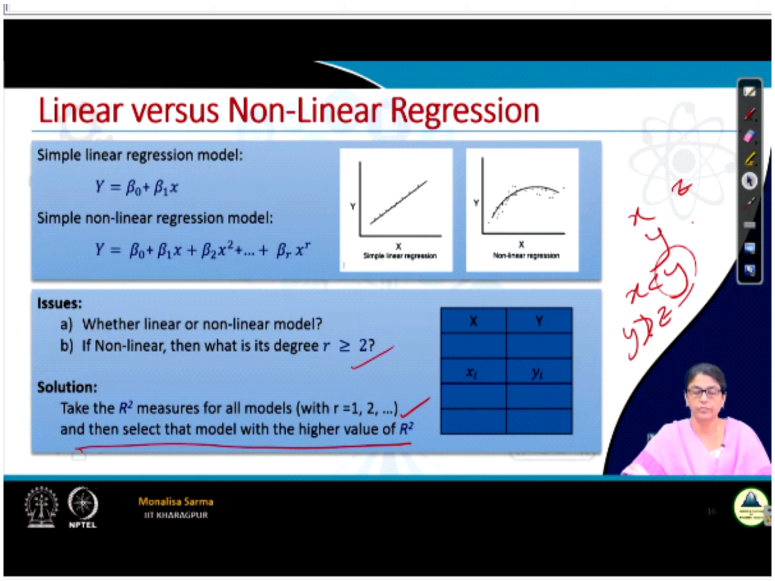
$$Y = \beta_0 + \beta_1 x$$

Simple non-linear regression model:

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_r x^r$$

Clear; same thing just that things is getting a bit more complicated that is all the theory behind everything is same, the procedure is same, theory is same, everything is same. Just the calculation is becoming a bit more intensive. So, this is my linear regression model, this is my non-linear regression model.

(Refer Slide Time: 16:55)



Linear versus Non-Linear Regression

Simple linear regression model:

$$Y = \beta_0 + \beta_1 x$$

Simple non-linear regression model:

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_r x^r$$

Issues:

a) Whether linear or non-linear model?
b) If Non-linear, then what is its degree $r \geq 2$?

Solution:

Take the R^2 measures for all models (with r =1, 2, …) and then select that model with the higher value of R^2

X	Y
x_i	y_i

Now the thing is that given a set of data x and y let us assume one parameter, one factor only. It is not multi-normal, this is just one what to say here we are using multiple nonlinear regression model, multiple non-linear models. So, let us not consider multiple, just consider it is just linear only, single only simple linear. So, here just one factor it is not multiple, it may be linear, it may be non-linear, but it is not multiple.

Suppose so and this is the case. Now the issue is given the different values of x and its corresponding y values are given, now which model we will try to fit? Whether we will try to fit linear or non-linear model where this y is map thing x is mapped to y based on a linear model or based on a non-linear model. So, if non-linear then what is his degree? Its degree is 2 or 3 or 4 what it is.

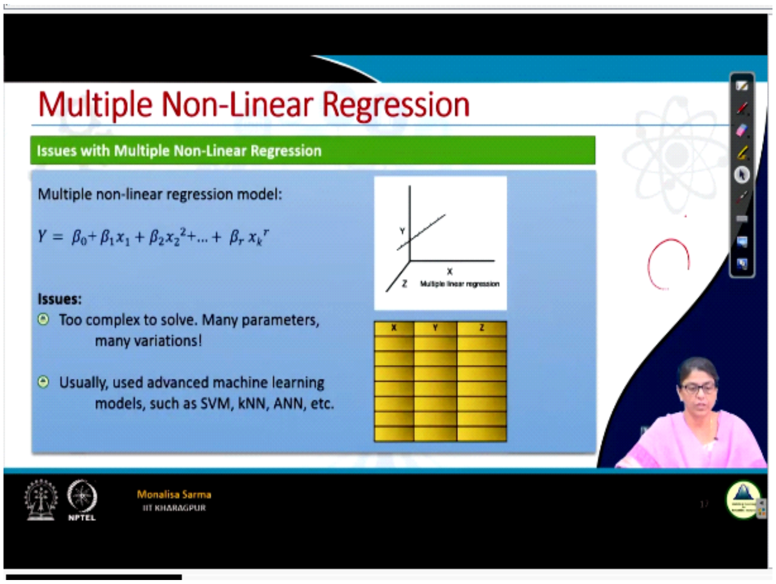
So, first what we will do? So, solution to this is because given this value of x and y we really do not know first it is really very difficult just if I try to plot the scatter plot also, from the scatter plot also visually sometimes very difficult to know whether it is the linear, it will fit a linear regression line or it will fit a non-linear regression line. So, if moreover even if you come to know that it will fit a non-linear regression line then what will be the degrees?

There will be 2 or 3 or 4 or what, so in that case what we will do? First we will take the r^2 measures for all the models. First we will try to find out the simple linear regression and will find out the r^2 model. That will be degree 1, then again I will find out for degrees equals to 2, that is the non-linear I will again find out the r^2 value.

Similarly I will again take $r = 3$ and I will find r^2 value and then select the model with the higher value of r^2. By taking a simple linear regression model if I get a particular r^2 value say x then I try to fit in a non-linear regression model where degree is 2, I get a value suppose y. Now my x is less than y. Then I am sure that it is not a linear regression model, it definitely it will fit a non-linear integration model.

Then I will again try it for $r = 3$ nonlinear fine. But let me just check for $r = 3$. For $r = 3$ I got z, however y is greater than z, then I need not to continue further then I know my regression is it is a non-linear regression model whose degree is 2. That model will fit best for the given set of data action y.

(Refer Slide Time: 19:59)



Multiple Non-Linear Regression

Issues with Multiple Non-Linear Regression

Multiple non-linear regression model:

$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2^2 + \ldots + \beta_r x_k^r$

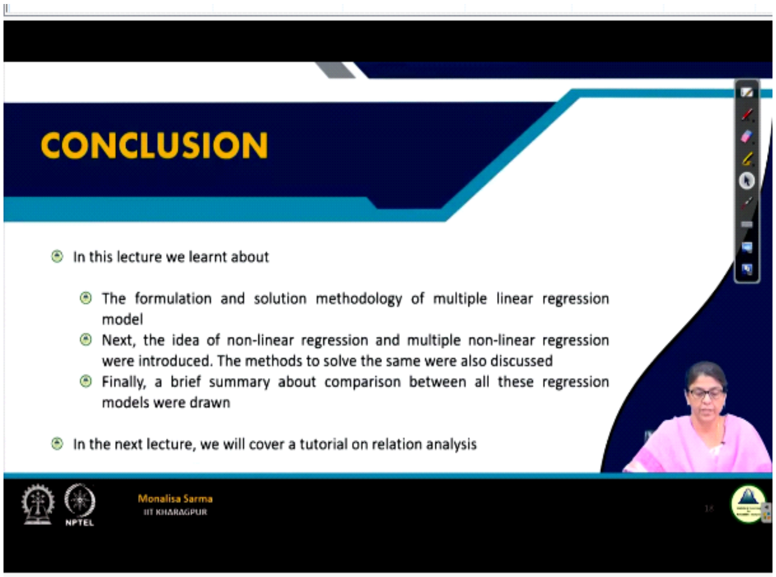
Issues:

- Too complex to solve. Many parameters, many variations!
- Usually, used advanced machine learning models, such as SVM, kNN, ANN, etc.

So, issues with multiple nonlinear regression model. So, what is the issue with multiple nonlinear? Too complex to solve, many parameters, many variations, so all this when we have multiple nonlinear even if we are choosing simple nonlinear regression model then also it is quite complex. But for multiple non-linear regression model it is too complex to solve we have many variations.

That is really not possible to solve using simple statistical matter. So, we use advanced machine learning models such as SVM, KNN, ANN etcetera which we will be discussing in coming few lectures.

(Refer Slide Time: 20:36)



CONCLUSION

- In this lecture we learnt about
  - The formulation and solution methodology of multiple linear regression model
  - Next, the idea of non-linear regression and multiple non-linear regression were introduced. The methods to solve the same were also discussed
  - Finally, a brief summary about comparison between all these regression models were drawn
- In the next lecture, we will cover a tutorial on relation analysis

So, now to conclude this lecture; in this lecture we learn about the formulation and solution methods of multiple linear regression model. Next idea of non-linear regression and multiple non-linear equations were introduced, the methods to solve the same that also we have discussed and then we have given a brief summary about the completion between all these regression models.

So, in the next lecture basically we will cover a tutorial on the relation analysis, by relation analysis means I will cover a tutorial on correlation analysis and regression analysis still what we have learned till, auto regression is still not covered, logistic regulation is still not covered, till whatever we have learned. So, correlation analysis and regression analysis everything put together will take a tutorial in my next lecture. Thank you, guys.

(Refer Slide Time: 21:22)

REFERENCES

- The Elements of Statistical Learning, Data Mining, Inference, and Prediction (2^{nd} Edn.), Trevor Hastie, Robert Tibshirani, Jerome Friedman, Springer, 2014.

Springer Series in Statistics
Trevor Hastie
Robert Tibshirani
Jerome Friedman
The Elements of Statistical Learning
Data Mining, Inference, and Prediction
Springer
NPTEL

Monalisa Sarma
IIT KHARAGPUR

So, this is the reference, thank you.