

Essentials of Data Science with R Software - 2
Sampling Theory and Linear Regression Analysis
Prof. Shalabh
Department of Mathematics and Statistics
Indian Institute of Technology Kanpur

Linear Regression Analysis
Lecture - 39
Introduction to Linear Models and Regression
Introduction and Basic Concepts

Hello, friends welcome to the course Essentials of Data Science with R Software 2, where we are trying to learn the topics of Sampling Theory and Linear Regression Analysis. From this lecture we are going to begin our new module on linear regression analysis, and we are going to talk about the basic fundamentals of linear models and regression in this lecture.

Linear regression analysis that is a very important topic, when you try to do any type of statistical modeling, modeling that is a very fancy word. And the linear regression models, they provide a very good way out to model the things in real life and as far as data sciences is related is concerned. You are always interested in finding out a model, nothing more than that, your ultimate goal in most of the cases is just to find out a model.

But, this model can be good, model can be bad, when the model is obtained through some logical procedure through some scientific procedure or not. These are the things which are practically decides, whether the obtained model is good or bad. And if you do not have a proper knowledge of statistics, your model cannot be dependable and it is not only dependency.

Even if you are trying to use some statistical tool and you are not getting the outcomes, which you are expecting from the real life. There can be many many violation, there can be different types of violation. So, you need to learn the basic background and the philosophy of the statistical tool, which are used in the regression analysis.

So, that you can diagnose that, if there are such some minor issues possibly you can take care. And if you do not know all these things, and you do not have a proper knowledge of regression analysis, even one simple problem can destroy the entire statistical model.

So, that is the motivation which you need to have to learn the linear regression analysis. So, in this lecture I will try to give you a background, what is linear model, what is regression analysis and what are the different types of terminologies which are used in this? And why should we do it? Why do we do it this regression analysis? Why this is called as regression? That is the first question I will ask you.

(Refer Slide Time: 03:22)

Basics

Linear models play a central part in modern statistical methods.

These models are able to approximate a large amount of metric data structures in their entire range of definition or at least piecewise.

2

So, all these small things I will try to address in this lecture, ok. So, let us try to begin our lecture with this slide. So, whenever we come to the aspects of modeling, there can be 2 types of model: linear models and non-linear models. What are those things? I will try to explain them in the forthcoming slides.

But with my experience I can share with you that, the linear model play a very important role, play a central part in the modern statistical methods. And these models are able to approximate a large amount of metric data structure in their entire range of definition, or at least piecewise. Suppose, there is a complicated process, whose curve is going like this this, this, this, this, this, this like this.

You cannot make a 1 model that will explain the entire phenomena. So, in these cases we can have 1 model for this region, one model for this region, one model for this region, another model for this region and so on. So, this is what I mean to say, when I say at

least piecewise. Your ultimate objective is to have a model, which is trying to explain the phenomena, which you are trying to consider, ok.

(Refer Slide Time: 04:50)

Basics: Example

Suppose the marks of students depend upon several factors, e.g., number of hours per week of study, number of assignments submitted per month, number of hours of play per week etc.

Suppose we want to find out the relationship between marks of students and number of hours per week of study, number of assignments submitted per month, number of hours of play per week in terms of a mathematical equation.

3

So, let me take a simple example to explain you the concept behind the models and regression model. We all know that the marks of the students in any examination they depend on several factor, they depend on numbers of hours of per week this they have a study, number of assignments they have submitted per month, number of hours they have played per week, number of hours they have spent in the library etc. etc.

So, we know that these are the factor which are going to affect the performance of the students, and hence the marks of the students. And suppose we want to find out a relationship between these variable, that is the relationship between the marks of the students and a group of variable which is affecting it, that is number of hours of per week number of study.

Number of assignments submitted per month, number of hours of play per week and so on, right. And we are interested in finding out a mathematical model. Why mathematical model? If you remember in the beginning itself, I have explained you that, if your conclusions are based on the rules of mathematics, then this word will believe on them.

It is just like if as a teacher if I say those students who are wearing a blue shirt they will be getting a grade A, those who are wearing a green shirt they will be getting a grade B.

This type of criteria nobody will accept in grading of the students. Rather if I say, ok, we will have some number of quizzes, examination, based on that you will obtain some marks on the topics, which have been taught and then based on your marks.

You will be graded means anybody who is getting 80 percent or more marks he or she will get a grade A. And say anybody who is getting a marks between say 60 and 80 percent, they may those student might be getting say grade B and so on. So, this criterion will always be acceptable why? Because there is an involvement of mathematics into it.

(Refer Slide Time: 07:10)

Basics: Example

Observations on 20 students are collected

Let

y : Marks of students

X_1 : Number of hours per week of study,

X_2 : Number of assignments submitted per month,

X_3 : Number of hours of play per week

Student no.	y	X_1	X_2	X_3
1	180	34	3	15
2	116	12	1	13
3	118	15	3	11
4	139	33	1	10
5	195	31	5	17
6	152	24	1	15
7	218	40	5	18
8	170	31	5	13
9	179	21	2	20
10	210	37	3	19
11	178	29	4	16
12	104	15	1	10
13	145	17	1	16
14	203	38	5	16
15	163	17	1	19
16	216	36	3	20
17	106	13	1	11
18	216	39	5	18
19	191	36	5	15
20	197	34	1	19

So, let me try to take some example, the same example. But I try to collect some data, because I do not know what is the equation and as we have discussed in the beginning itself, that in statistics whenever we do not know anything and we want to know about some phenomena, the most simple option for us is to collect some data.

So, first of all we have a faith, we have a belief, that these marks of the students are going to be affected by the variables, like number of hours per week which is denoted by say X_1 , number of say assignment submitted per month X_2 . And number of hours of play per week say X_3 and there can be some more also but, I am taking only here 3 and then marks of the students are denoted by Y , ok.

(Refer Slide Time: 08:13)

Basics: Example

How to find f ?

X_1 : Number of hours per week of study,
 X_2 : Number of assignments submitted per month,
 X_3 : Number of hours of play per week

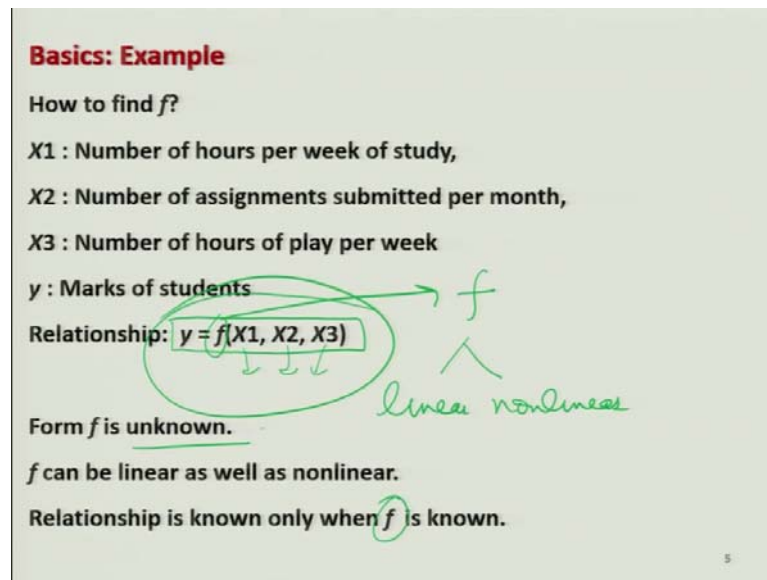
y : Marks of students

Relationship: $y = f(X_1, X_2, X_3)$

Form f is unknown.

f can be linear as well as nonlinear.

Relationship is known only when f is known.



Now, if you try to see, there is going to be a relationship like this one. That marks are a function of say number of hours per week of study X_1 , number of assignments submitted per month X_2 and number of hours of play per week X_3 . So, the question here is this, we have the values of y , X_1 , X_2 , X_3 for some number of students, but we do not know, what is here the form of relationship f . Form of f is unknown to us and this form can be linear or this can be non-linear.

And this relationship can be known, only when f is known to us. And nobody is coming from sky to tell us, that a what is the f . But, we have to look into the data and we have to learn that how to retrieve the information, which is concerned, which is contained inside the data on X_1 , X_2 , X_3 . And what this data is trying to inform us about the form of the equation.

(Refer Slide Time: 09:47)

Basics:


2 types of variables

- Input variables or independent variables
- Output variables or dependent variables.

Objective:

To determine a relationship between dependent and independent variables which describes the phenomenon/process in the best possible way.

This is a model.]



6

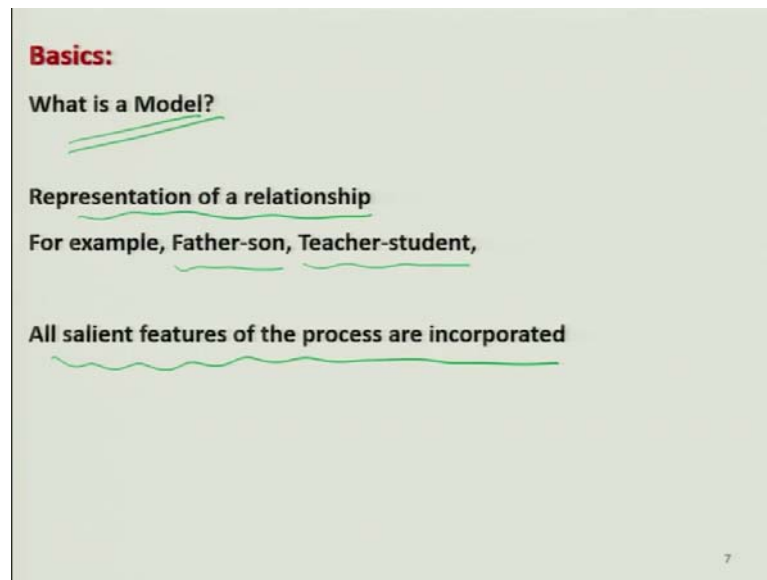
So, in this process you have understood that, now there are two types of variables; one say input variables or they are they can be called as independent variables, and another is output variable or variables or they are also called as dependent variables. So, what is the relationship between the two? Suppose, if I say I have here one variable marks and say number of hours of study.

So, you can see here very clearly that, if a student is studies more then he will get he or she will get more marks. But, in case if the person is getting more marks, then how many hours he should or she should study that is very difficult to say in general. So, then, that means number of marks are affected by the number of hours of study, but not vice versa.

So, the number of hours of study can be treated as input variable, and the marks can be treated as output variable. Well, there can be situation when we try to consider more than one output variables also. We are not going to consider here in this course, but it is important for you to know that, yes this is possible, ok.

So, now based on this data set, we have an objective that, we would like to find a relationship between the dependent and independent variables, which describes the phenomena or the process in the best possible way. Now, what is best possible way, that we have to find and whatever we will obtain, this will be called as model.

(Refer Slide Time: 11:47)



So, now the next question is what is a model? This is actually representation of a relationship and why do we talk only of the mathematical relationship. Why cannot there be a relationship between father and son, teacher and student? For example, you have heard many time, that many people say that, ok, I want to be like my father or I want to be like my teacher and so on.

So, what is really happening? They are their role model; that means, whatever they are trying to do, they will try to copy the same thing and they will try to follow possibly the same model values, ethical values like them. So, in a simple language they want to become like their father or teacher. So, this means all salient feature of the process are incorporated.

For example, if I become just like my father; that means, whatever are the qualities and the salient features of my father, they are also inside me and then people might say, ok, Shalabh you look just like your father, that is the model.

(Refer Slide Time: 12:56)

Basics:
Statistician's role:

- No right to change or alter the process.
- works only on the basis of a small sample.
(Sample : A small fraction of population)
- Sample is expected to have all features of the population.

Now, whenever you are trying to make a model, you should know your limitations. As a statistician or as a data scientist, we have no, right to change or alter the process, whatever is happening that will continue to happen.

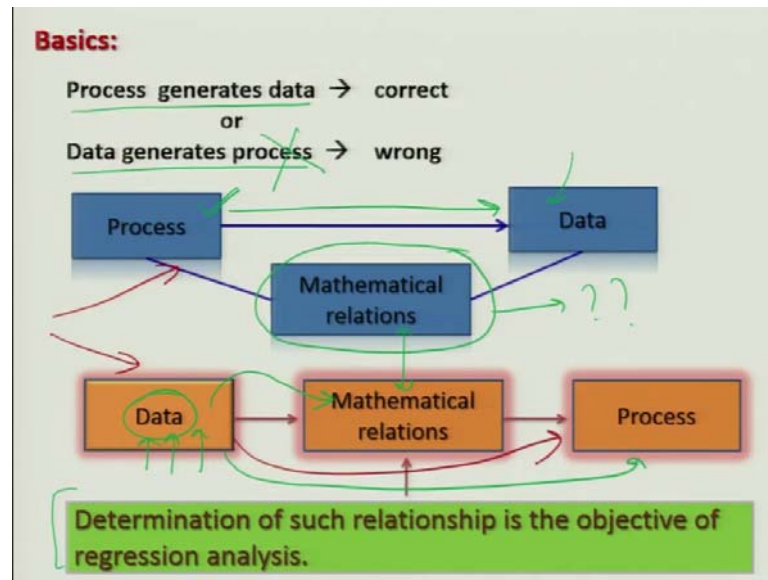
The way the yield is coming, the way the crops are going you cannot change their way. But whatever has been done based on that, whatever way the crop has adopted to grow, we simply have to observe and the biggest challenge is that, we have to work only on the basis of a small sample.

But, the results will be valid for a very large population. For example, if somebody creates a model for the efficiency of a medicine, then that model will be valid all over the world. Have you ever heard that medicine is more effective on an American person, but less effective on European person? Medicine to control the body temperature or fever can control the body temperature, for an American up to 8 hours and for European only up to 4 hours.

These things should not happen, these are wrong conclusions in general unless and until the medicine is affected by the geographical condition or weather conditions, right. So, whatever we have to conclude, whatever inferences we have to draw, they have to be for the entire population. Now, what is sample? What is population? Now, you have understood that why we have studied the sampling theory, right.

The sampling theory will help you and guide you to get a representative sample. Why representative sample? Because the sample is expected to have all features of the population and how to get it done?

(Refer Slide Time: 15:10)



Sampling theory will help you in this job. Now, I ask you a simple question. I give you two statements and please let me know which one is the correct one, first statement is say process generates data or the second statement is data generate process.

Which one is correct? Obviously, the second statement is wrong. Whenever there is a process, the data will be generated. So, if I try to depict this phenomena through this graph, then I am saying that this is my here process and this is my here data. And this process is trying to generate the data and whatever is the process, we believe that can be represented by some mathematical relations.

But, my problem is that, we do not know this mathematical relation. And never objective is this that, somehow if I can find out this mathematical relation, then I can control the entire process. So, what we do? We go in the reverse direction. We try to collect the data and then we try to observe the process. And obviously, this mathematical relationship will remain the same.

So, now based on this data, we try to find out the mathematical relationships. And the determination of such mathematical relationship is the objective of regression analysis.

This is what exactly we do when we say statistical model. Finding out a statistical model is like you have got a data. Based on that data you give me a mathematical equation and the form of the equation should be like which is copying the real process.

So, if I try to handle that mathematical relationship, then we will get the outcome just like as if they are happening in real life, and this is your forecasting prediction, ok.

(Refer Slide Time: 17:59)

Linear models:

Suppose the outcome of any process is denoted by a random variable y , called as dependent (or study) variable, depends on k independent (or explanatory), variables denoted by X_1, X_2, \dots, X_k

Suppose the behaviour of y can be explained by a relationship given by

$$y = f(X_1, X_2, \dots, X_k, \beta_1, \beta_2, \dots, \beta_k) + \varepsilon$$

Handwritten notes on the slide include:

- Red circles around X_1, X_2, \dots, X_k and $\beta_1, \beta_2, \dots, \beta_k$ with arrows pointing to the word "fixed".
- A red circle around ε with an arrow pointing to "random variata".
- A green circle around y with an arrow pointing to "y: random variable".
- A handwritten equation $y = mx + c$ with (x, y) and (m, c) in boxes, and (m, c) labeled as "parameters".
- A handwritten example: $\bar{x}, y = 2 \rightarrow 2 = m + c$ with "NO" written next to it.
- A handwritten example: $m = 1, c = 2, y = x + 2$.

Now, let me try to explain you in a different way. Now, whenever you are trying to work in a real life, right, then whenever you are trying to conduct an experiment. Do you think the outcome of that experiment is always fixed? Suppose, I ask you that you takes some flower pots, put some soil into the flower pot try to put some seeds into it and say after a month try to see, how many flowers have bloomed on each and every plant.

And what is the height of that plant. Do you think that is it going to be a constant value? Certainly, not there will be some difference, some plants may have say 10 to 12 flowers, some plants may have 5 to 5, 6, 7 flowers the height of the plants might be different.

So, now my question is, when you have taken the same soil, you have taken the similar pots, you have taken the same seed, you have put the same quantity of seeds in the soil, then why this difference is coming? There is something which is happening, but whatever is happening whatever is the reason that is not in my control. There can be some reasons which we do not know and which are beyond our control.

So, whenever we are trying to conduct any real life experiment, this situation will occur again and again. Whenever as a student we go to the examination, we prepare our best and all the students are being taught by the same teacher inside the classroom, same teaching material has been delivered to them. But, even the same student will get different marks in different examination, sometime low, sometime high.

Do you think that any of the students intentionally tries to get a lower mark? No. But something happens, there can be factor 1, there can be factor 2 or there can be some factor, which we do not know. So, whenever we are trying to work in a real life, there will always be some random variation which will be always present in the observations.

And whenever we are trying to formulate a mathematical relationship, we cannot formulate it without the showing the presence of random factor. So, whenever we are trying to consider a mathematical relationship, the mathematics does not understand randomness, but statistics understand the randomness. I can take a very simple example, suppose if you try to find out, how much time do you take in going from your home to your college?

The distance is the same, route is the same, but every time you take a different time. Why this is happening? Is it possible that each and every time you get you take exactly 20 minutes? That may be 17 minutes, sometime 18, sometime 19, sometime 21, sometime 22; that will be close to 20 that is acceptable.

So, now this is the basic difference between a mathematical relationship and a statistical relationship. The statistical relationship will take care of the random variation in the data. And here we are interested in finding a statistical relationship, but definitely without the help of mathematics, we cannot do it.

So, now I am in going to introduce the sort of formal relationship between input and output variables, but I will also add the randomness into it. And in practice if you try to see, you have two types of variable, input and output. The input variables are in your control for example, you know that in every flower pot, you have put say 20 seeds, they are in that is in your control.

You know that in every flower pot, you have put only 2 kg of soil; you know that in every flower pot every day you have given 100 ml of water. So, these things are actually fixed and they are in your control. So, the so in general usually we consider the input variable to be fixed, there cannot be random variation.

Well, there can be random variation in some cases for example, if somebody is using the irrigation in a big field, then you do not know that how much water, how much exactly the water is going to which part of the field. But in general, there will be not much different, the variability will be very small and for all practical purposes we can take it fix.

So, unless and until we believe that there is a lot of variation in the input variable also, we will assume that is in this course that my input variables are fixed. So, now the question is why this variation is coming? The variation is coming because, there are certain factors which are not in our control and this variation is affecting the output values.

So, that is why in this relationship, we assume that the output variable is a random variable and the input variables are fixed. That is our basic idea in linear regression analysis. Well, at this moment I would make a note that, it is also possible to consider the input variable as random, but we are not considering that aspect here, ok.

So, let us try to come back to our slides. So, suppose there is some process, and the outcome of the process is denoted by a random variable, say y and this will be called as dependent or study variable. And this study variable or dependent variable suppose depends on a small k number of independent, or say explanatory variable.

So, this dependent, independent, study variable, explanatory variable they are the different nomenclatures, which have been used in in regression analysis, but they are the same thing. So, suppose we denote the independent or input variable by capital X_1, X_2, \dots, X_k .

So, these are the k different variable for example, this can be the number of hours of study, X_2 can be the number of hours of playing, and X_k can be the number of

assignments submitted and so on. So, the behavior or the relationship between y and X_1, X_2, \dots, X_k . This can be controlled by some coefficients and random error.

So, what we try to do here? That we try to write down a mathematical function or a statistical function like this, $y = f(X_1, X_2, \dots, X_k, \beta_1, \beta_2, \dots, \beta_k) + \varepsilon$. What are these $\beta_1, \beta_2, \dots, \beta_k$?

Let me try to take a very simple example, here you remember that you have done. You know this equation y equal to $mx + c$. This is a line, a straight line and the components of this line, I can divide in 2 components say x and y and here m and c , right. So, now if I ask you that, suppose you know the value of x and y , suppose I can say that x is equal to 1, y equal to 2.

So, this relationship becomes 2 is equal to $m + c$. So, do you know anything about this line, y equal to $m x + c$ from this data? No, but on the other hand if I say here suppose m is equal to 1 and c is equal to 2, then my line becomes $y = x + 2$. So, now I have the complete information about this line.

So, you can see here out of these 2 sets of values, m and c are the set of values, which if you know then you know all the properties of this line. So, they are termed as parameter. So, I am trying to introduce here the parameters which are denoted by $\beta_1, \beta_2, \dots, \beta_k$ and these parameters are associated with this with these independent variables. β_1 is related to X_1 , β_2 is related to X_2 and so on, right.

And this ε here, this is going to denote the random variation. And we also assume that all this $\beta_1, \beta_2, \beta_k$ are fixed. But just to tell you that it is also possible that β_1 to β_2, β_k can also be random, but here at least in this course we are assuming them to be fixed and they will play a very important role.

So, ε is the random variable and because of the involvement of ε , this y becomes a random variable. Why? Because X 's are fixed, X_1, X_2, \dots, X_k are fixed, $\beta_1, \beta_2, \dots, \beta_k$ are fixed. So, whatever is this part this is going to be fixed? So, this is the component here ε which is controlling the behavior of y , and it is transforming y into a random variable, right.

(Refer Slide Time: 28:38)

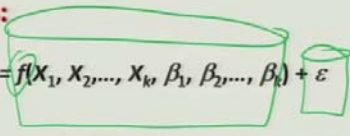
Linear models:

$$y = f(X_1, X_2, \dots, X_k, \beta_1, \beta_2, \dots, \beta_k) + \varepsilon$$

where f is some well defined function and $\beta_1, \beta_2, \dots, \beta_k$ are the parameters which characterize the role and contribution of X_1, X_2, \dots, X_k , respectively.

The term ε is the random error. It reflects the stochastic nature of the relationship between y and X_1, X_2, \dots, X_k .

It indicates that such a relationship is not exact in nature.



And here f is some well defined function and you will see later on that $\beta_1, \beta_2, \dots, \beta_k$ are the parameter, which will characterize the role and contribution of X_1, X_2, \dots, X_k also. And as I said this term ε is the random error, actually this random error reflects the stochastic nature of the relationship between y and X_1, X_2, \dots, X_k . And there can be many many factors in any experiment which are beyond our control.

So, what we try to do? All those factors which we cannot control. We try to put them inside the basket called as ε . And all those factor which we can control, they are put in this basket of X_1, X_2, \dots, X_k , right. And the presence of ε indicates that the relationship between y and X_1, X_2, \dots, X_k is not exact in nature that is the stochastic that is random.

(Refer Slide Time: 29:54)

Linear models: Mathematical and statistical models

When $\varepsilon = 0$, then the relationship is called the mathematical model.

When $\varepsilon \neq 0$, then the relationship is called the statistical model.

The term "model" is broadly used to represent any phenomenon in a mathematical frame work.

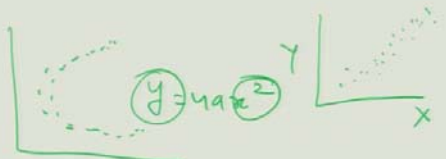
12

And obviously, means if you try to take ε is equal to 0, then the relationship becomes mathematical relationship and the corresponding model is called as mathematical model. And when this ε is not equal to 0, then the relationship is called as a statistical model. Because, there is a involvement of randomness. So, the term model is broadly used to represent any phenomenon in a mathematical framework.

(Refer Slide Time: 30:24)

Linear and nonlinear models:

A model or relationship is termed as linear if it is linear in parameters $\beta_1, \beta_2, \dots, \beta_k$
and
nonlinear, if it is not linear in parameters.



13

Now, the question is under what type of condition you will say that a model is linear or non-linear? Here, I would like to have your attention, that many times people try to look on the relationship between x and y in a graphic way and based on that they try to decide whether the model has to be linear or non-linear.

This is not always correct, because as long as the statistics is involved, we define or we term a model or a relationship to be linear, if it is linear in parameters. And this relationship is termed as non-linear, if it is not linear in parameters. So, I am just showing you here you may get confused.

Suppose if I try to take here a relationship like parabolic, right. Something like $y = 4ax^2$ type of thing; by looking at the relationship of y and x^2 you may think that, this is going to be a non-linear model. But statistically this is a linear model. How to convert it into linear model and how to handle it? That is what we are going to learn in this course.

(Refer Slide Time: 31:57)

Linear and nonlinear models:

In other words, if all the partial derivatives of y (or $E(y)$) with respect to each of the parameters $\beta_1, \beta_2, \dots, \beta_k$ are independent of the parameters, then the model is called as a linear model.

If any of the partial derivatives of y (or $E(y)$) with respect to any of the $\beta_1, \beta_2, \dots, \beta_k$ is not independent of the parameters, the model is called as nonlinear.

Note that the linearity or non-linearity of the model is not described by the linearity or nonlinearity of explanatory variables in the model.

14

So, now in order to remove this confusion, we have a simple criteria to check, whether the model is linear or non-linear. So, we say if all the partial derivatives of y or say expected value of y with respect to each of the parameters $\beta_1, \beta_2, \dots, \beta_k$ are independent of the parameters.

Then the model is called as linear model and if not, then the model is called as non-linear model. And remember the linearity or nonlinearity of the model is not described by the linearity or non-linearity of explanatory variables in the model. So, if the explanatory variable is entering as say X_1^2 or say X_1^3 that will not determine whether the model is linear or not.

(Refer Slide Time: 32:50)

Linear and nonlinear models:

For example

$$E(y) = \beta_1 X_1^2 + \beta_2 \sqrt{X_2} + \beta_3 \log X_3 + \varepsilon \quad \rightarrow E(\varepsilon) = 0$$

is a linear model because assuming $E(\varepsilon) = 0$,

$$\frac{\partial E(y)}{\partial \beta_i}, (i = 1, 2, 3)$$

are independent of the parameters $\beta_1, \beta_2, \beta_3$.

linear model

$$\frac{\partial E(y)}{\partial \beta_1} = X_1^2 \rightarrow \text{ind of } \beta_1 \text{'s}$$

$$\frac{\partial E(y)}{\partial \beta_2} = \sqrt{X_2} \rightarrow \text{"}$$

$$\frac{\partial E(y)}{\partial \beta_3} = \log X_3 \rightarrow \text{"}$$

15

So, let me take here some example and try to show, that how you can decide whether a model is a statistically linear or not or a model is linear in statistical sense or not.

Suppose I take a model here $y = \beta_1 X_1^2 + \beta_2 \sqrt{X_2} + \beta_3 \log X_3 + \varepsilon$. . We assume that the expected value ε is 0; that means, it simply means that if you try to take some observation, in some observation the random error will be in with positive sign, in some observation the random error will be with some negative sign.

So, when you try to take their average, the average will be 0. So, this is what we mean by saying expected value ε is equal to 0 and when we try to go for further analysis we will make this assumption also. So, I try to take a respected value of here y on both the sides. So, since this is constant. So, this will remain as such and this expected value of ε will become 0.

Now, if you try to take the partial derivative of expected value of y with respect to β_1 , β_2 as well as β_3 . You can find out that these are independent of the parameters. For example, say the partial derivative of expected value of y with respect to β_1 , that is simply here X_1^2 .

So, this is independent of β s. Partial derivative of expected value of y with respect to β_2 is $\sqrt{X_2}$, this is again independent of β s and partial derivative of expected value of y with respect to β_3 . This is again $\log X_3$. So, this is again independent of β . So, this model is a linear model.

Now, think if I had given you this model in general, and if I ask you that whether this is a linear or non-linear, you will say this is a non-linear model and possibly that is the wrong answer, ok.

(Refer Slide Time: 34:53)

Linear and nonlinear models:

On the other hand,

$$y = \beta_1^2 X_1 + \beta_2 X_2 + \beta_3 \log X + \varepsilon$$

is a **nonlinear model** because \rightarrow N.L. model

$\frac{\partial E(y)}{\partial \beta_i}$, ($i = 1, 2, 3$) depends on $\beta_1, \beta_2, \beta_3$.

but $\frac{\partial E(y)}{\partial \beta_2}$ and $\frac{\partial E(y)}{\partial \beta_3}$ are independent of any of the β_1, β_2 , or β_3 .

Handwritten notes on the slide:

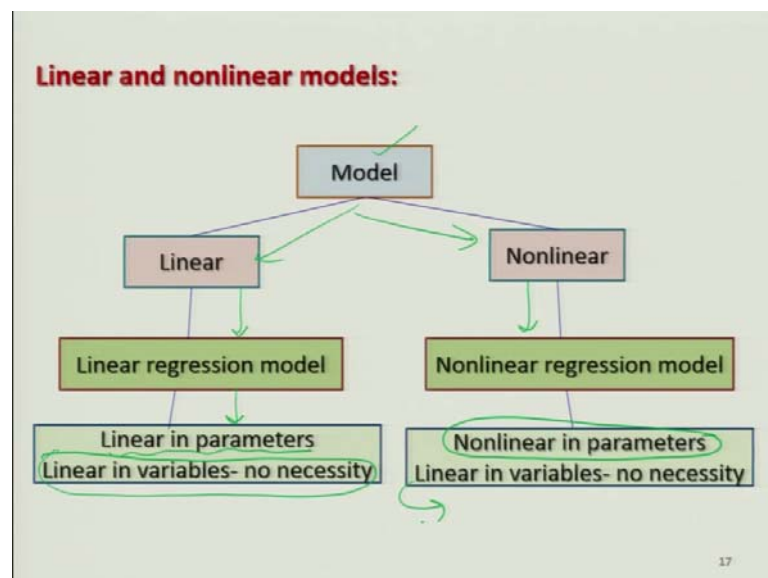
- $\frac{\partial E(y)}{\partial \beta_1} = 2\beta_1 X_1$ (with β_1 circled and β_1 written below)
- $\frac{\partial E(y)}{\partial \beta_2} = X_2$
- $\frac{\partial E(y)}{\partial \beta_3} = \log X$

Similarly, if I try to take here one more example, Suppose I try to take here a model like this one, $y = \beta_1 X_1^2 + \beta_2 \sqrt{X_2} + \beta_3 \log X_3 + \varepsilon$, then under the same assumption you can see here, that the partial derivative of expected value of a y with respect to β_1 , β_2 , β_3 they are not independent of the parameter, right.

You can see here, the partial derivative of expected value of y with respect to β_1 this is twice of $\beta_1 X_1$. So, this is a function of β_1 . So, this model is a non-linear model and even if one of the partial derivative is a function of β_1 ; that means, the entire model will be non-linear, right.

Because, here if you try to see in this case, if I try to take the partial derivative of expected value of y with respect to β_2 , this is independent of parameter. And similarly with respect to β_3 also the partial derivative of expected value of y is simply $\log X$. This is also independent of the parameter, right, ok.

(Refer Slide Time: 36:13)

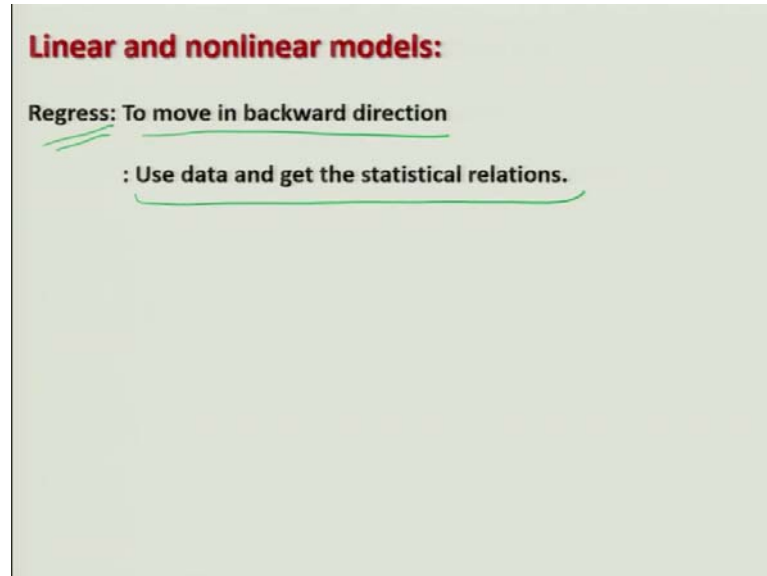


So, now when I come to the aspect of linear and non-linear model, So, I have now explained you how are you going to decide whether the model is linear or non-linear. So, when we are trying to consider the modeling then we have 2 options, the model can be linear or the model can be non-linear.

So, when we are trying to consider the linear model, based on that we will have linear regression analysis or we will have a linear regression model, and when we are trying to consider a non-linear model, based on that we will have non-linear regression model. And in case of linear regression model, the model is going to be linear in parameters and the linearity in variables is not necessarily needed.

And similarly in the case of non-linear regression model, the model will be non-linear with respect to the parameters and the linearity in variables is not a necessity, ok.

(Refer Slide Time: 37:11)



So, now I come to another aspect that why do you call it as regression analysis? The dictionary meaning of regress is to move in the backward direction. So, if you try to see here in this graph, I had shown you that you are going in the reverse direction. The correct process is this one, that the process generate the data, but you are moving in the reverse direction. That you are trying to generate the data and then you are trying to determine the process.

So, since you are going in the opposite direction. So, that is why this is called as regression and analysis. So, by looking at the data, we try to find out the statistical relationships.

(Refer Slide Time: 38:07)

Linear regression model: Example

Consider a simple example to understand the meaning of "regression".

Suppose the marks of students (y) depend upon

- number of hours per week of study (X_1),
- number of assignments submitted per month (X_2), and
- number of hours of play per week (X_3).

We want to find out the relationship between y and X_1, X_2 and X_3 as

The slide contains handwritten mathematical derivations. At the top, it shows a simple linear regression equation: $y = m_1x + c$, with a circled m_1 and a circled x . Below this, it shows a multiple linear regression equation: $y = m_1x_1 + m_2x_2 + m_3x_3 + c$, with circled m_1, m_2, m_3 and circled x_1, x_2, x_3 . Below that, it shows the standard form: $y = X_1\beta_1 + X_2\beta_2 + X_3\beta_3 + \varepsilon$, with circled X_1, X_2, X_3 and circled $\beta_1, \beta_2, \beta_3$. To the right of this equation, there are handwritten notes: $\alpha, \beta, \gamma, \delta \dots$ and x, y, z .

So, that is the reason it is called as regression analysis. So, now we will consider here a linear regression model. Because we are aiming to learn here is the linear regression analysis.

So, we consider the same example where we have said that the marks of the students are depending on number of hours per week that is X_1 , number of assignments submitted per month say X_2 and number of hours of play X_3 . So, if you remember I had given you this equation y equal to $mx + c$.

So, although I will give you in more detail, but at this moment my idea is to explain you that, how the things are happening. So, if you look at this relationship there is 1 y and 1 x , but now once I say suppose I have more than one x . So, one option is this I can extend it to $y = m_1 X_1 + m_2 X_2 + m_3 X_3 + c$.

And one simple note that in statistics we always try to denote the random variables by alphabets like x, y, z and all the parameters, they are denoted by the Greek letters like $\alpha, \beta, \gamma, \delta$ etc. So, all this m_1, m_2, m_3 etc. or whatever is here m , that is actually in a statistical language. That will be denoted by some Greek letter and we have decided to denote by β .

So, the same model which was written only for one variable, that is extended that is extended to three variables and this can be written as something like $y = X_1 \beta_1 + X_2 \beta_2 + X_3 \beta_3 + \varepsilon$ and if you want to take care of your c , I can write here one more parameter. Well, anyway these things are at a very introductory level, later on we will see how this things may happen.

(Refer Slide Time: 40:08)

Linear regression model: Example

There exist the true values of β_1, β_2 and β_3 in nature but are unknown to the experimenter.

Some values on y are recorded by providing different values to X_1, X_2 and X_3 . For example,

Student no.	y	X_1	X_2	X_3
1	180	34	3	15
2	116	12	1	13
3	118	15	3	11
4	139	33	1	10

There exists some relationship between y and X_1, X_2, X_3 which gives rise to a systematically behaved data on y, X_1, X_2 and X_3 .

$y = f(X_1, X_2, X_3)$

Such relationship is unknown to the experimenter.

So, my objective is this there are some values of $\beta_1, \beta_2, \beta_3$ β are which are existing in nature, but they are not known to us. So, what we try to do? We try to conduct the experiment. We try to give an examination to a student and suppose he gets he or she gets 180 marks, then we ask the values of X_1, X_2, X_3 and their responses are recorded like this. Then we try to repeat this and we try to collect the same observations on y, X_1, X_2 , and X_3 .

So, definitely now there exist some relationship between y and say here X_1, X_2, X_3 , but this relationship is not known to us. But in case if I try to look at this data, then it is possible that by looking at this data, this can give rise to a systematically behaved data. The behavior there can be some hidden behavior, which is systematic. But my problem is this, this relationship is not known to us not known to the experimenter.

(Refer Slide Time: 41:28)

Linear regression model: Example

To determine the model, we move in the backward direction in the sense that the collected data is used to determine the unknown parameters of the model.

In this sense such an approach is termed as regression analysis.

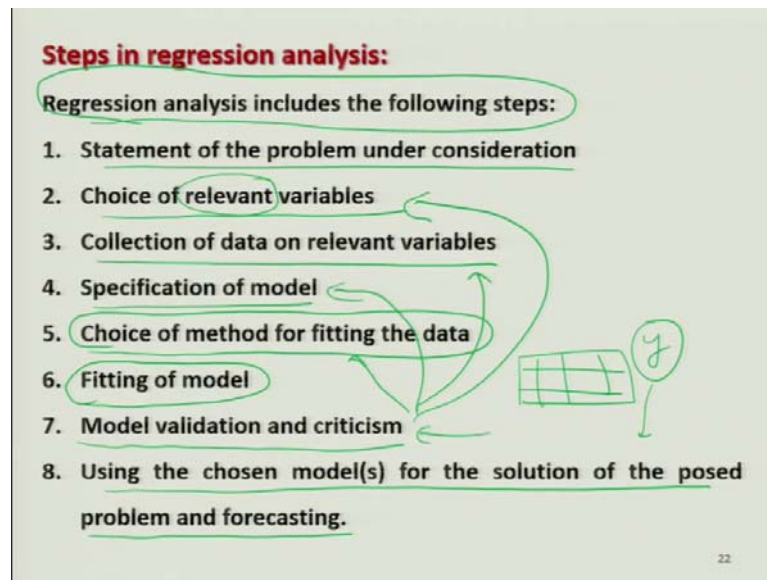
The theory and fundamentals of linear models lay the foundation for developing the tools for regression analysis that are based on valid statistical theory and concepts.

21

And in order to determine this relationship or the model, we move in the backward direction, in the sense that we collect the data and the collected data is used to determine the unknown parameters of the model, and that is why in this sense such an approach is termed as regression analysis.

And many times you will see in the undergraduate and post graduate levels, we have two types of courses one course is on the linear models and say another course is on the regression analysis. So, actually the theory and foundation of linear models lay the foundation for developing the tools for regression analysis. That are based on some valid statistical theory and concepts.

(Refer Slide Time: 42:18)



So, that is the relationship between the two although they are more or less similar. So, now just to give you a quick idea that, how this regression analysis is done, right. So, the first step is this, we have to write the statement of the problem, that what we really want to know.

Suppose, if I want to know the performance of the student, then I have to collect the data on relevant variable. And suppose if I am interested in some agricultural experiment, where I have to record the yield of the crop and the variables, which are affecting the crop like as quantity of fertilizer, temperature, irrigation etc..

So, we have to choose the relevant variable. What is called relevant variable? How it can be done? That is a statistical issue, that we will try to discuss. But at this moment you can believe or you can imagine, you will get easily convinced, that we need to data on we need to have data on relevant variables, right; means some variable, some irrelevant variable should not be added in the model.

Now, once you have decided this thing, then we have to collect the data on those relevant variable. And then we have to start with a possible specification of a model. For example, if we are working in the linear regression analysis we will always assume that we have a specification of a linear model. And if you are going for non-linear, then you

have to assume some linear; some non-linear model and then you have to work and see how it goes.

Once you have specified the model, then you have to choose that how are you going to estimate the parameters? Or in the language of regression analysis, how you are going to fit the data? How you are going to obtain the fitted model? There are different estimation techniques in the statistics like as principle of least square method of maximum likelihood method of movement's etc. etc.

So, you have to first choose which of the method is more appropriate under the given conditions. Once you decide for those methods, then you have to fit the model, you have to use some statistical techniques for the fitting of models. Once you have obtained the model on the basis of the given set of data, you have to validate it and you have to you have to criticize it. That means, try to see whatever model you have obtained, is this giving you the similar type of data in the real life also?

For example, you already have collected some data set and based on that you have some outcome also y . Now, you try to take some input variable and try to see what is your here y and whether this y is close enough to the real value that you are observing in practice. And if yes, that means the model is good, otherwise you have to revise your model. And actually I must admit that getting a good model is not usually achieved in one step.

So, we come here surely, there are high chances that there will be some problems in the model. So, we try to go back once again and we try to see whether we have chosen the relevant variable or not? Or whether we have collected the data properly or we have specified the model correctly or have we chosen the correct method of estimation. And you have to just look into different aspects not one time not two time, but several times.

But at the end, after you have taken care of all the problems, you will have a very good model and which will be depicting as if the real process is being controlled by the mathematical model, right. And once you have the good fitted model, this model is used for the solution of the posed problem and making forecasting prediction etc. etc. many many things happen, right.

So, now I stop here, my idea in this lecture was to just to give you a broad overview of the regression analysis. What is this, what are we going to do, what are the different complications. But I would try to address one thing. This regression analysis is not a small topic, this is a vast topic there are many many aspects which have to be handled. Since, I have limited time in this course,

So, it may nearly it may not be possible for me to cover all the topics. But my objective is that I want to give you the sufficient material, I want to give you the sufficient correct fundamentals, which will help you in taking out the fear of learning regression analysis yourself from your heart. I will try to put you on the, right track, I will try to develop a sort of statistical thinking, how to think about the linear regression model and.

once you do this much, after that I believe that if you try to take up some good book, it should not be a very big problem for you to continue further and learn the regression analysis in more detail. That is my objective remember, so I would say you try to look into some regression analysis book, try to look into the first chapter which is usually the introduction. They are trying to give different types of very simple example, in very simple language to give you an overview of the problem to motivate the students.

So, you study and I will see you in the next lecture, till then good bye.