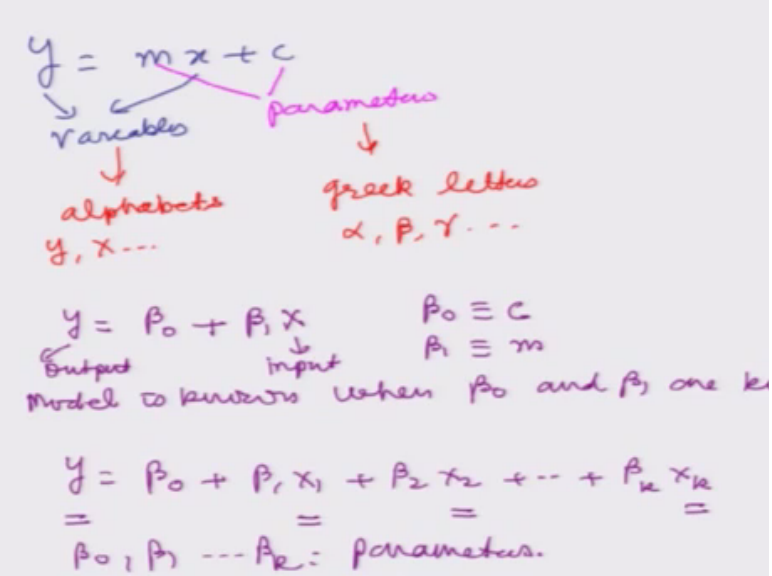
Regression Analysis and Forecasting
Prof. Shalabh
Department Of Mathematics and Statistics
Indian Institute of Technology – Kanpur

Lecture - 02
Regression Model - A Statistical Tool

Welcome to lecture 2 as in the lecture 1 we had discussed some basic fundamental concepts about the modeling so we will be continue to learn some more basic fundamentals about the regression modeling, so in the last lecture you had seen that I had taken an example of a simple line like as y=mx+c.

(Refer Slide Time: **00:34)**



Here I denoted that x and y are your variables and whereas m and c they are the parameters. In statistics we have a understanding or there is a tradition that usually we try to express these variables by some alphabets or by some Latin letters and whereas parameters are expressed in Greek letters, so you will see that we will try to express the variables something like y, x and so on.

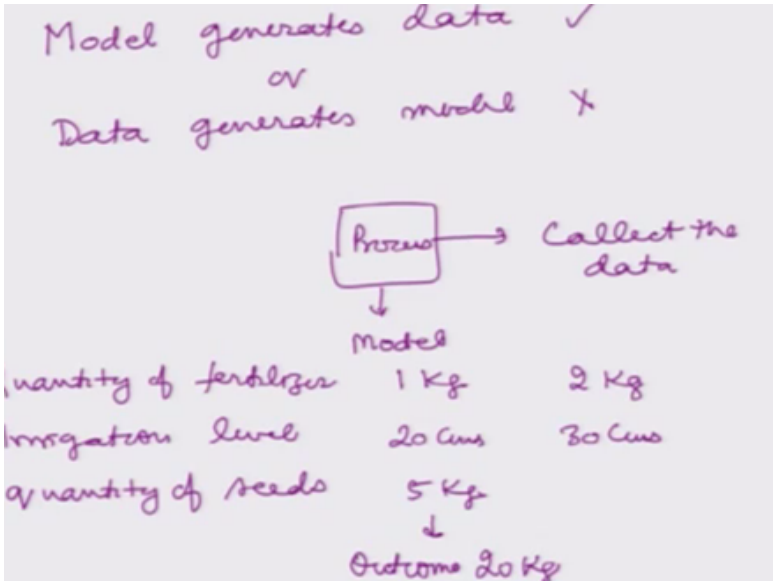
Whereas the parameters are usually denoted by alpha, beta, gamma, and so on, this is our common language of statistics. So now incase if I try to translate now this model y=mx+c in the language of a statistics then I can write down the variable as y, this intercept terms c this can be written as beta0 and this m can be written has beta1 and X. So if you try to see beta 0 has the same role as of c and beta1 as the same role as of m and x.

And y remains the same. So now once I say that the model is known when beta0 and beta1 are known at least in this case, this is a very elementary situation and you will see later on, well because this is too simple for us, but at this moment this is very important to understand the basic concepts using this small examples. So if you try to see here, here I am saying that y is my output variable and this x is my input variable.

So this is very simple situation, well in practice there are usually more than one input variables which affect the outcome variable. So now if you try to see it is very natural to extent this simple equation to a more general situation why? because in practice usually there are more than one independent variables which affect the outcome. So I can extent this equation as a y= beta0 + beta1x1+ beta 2x2 and + beta k.

X k and we assumed here that there are not 1 but, there are k input variables x1, x2, x k which are affecting the output variable here y. So I can say that this beta0, beta1 up to beta k they are the parameters of the model.

(Refer Slide Time: **03:48)**



Now I try to address another question, so now we come to another simple concept to understand one more basic fundamental. If I try to write down here two sentences model generates data or data generates model, which one correct in the Lecture one we had discussed that model is simply a mathematical relationship to express a phenomenon which is happening in the nature and as a statistician we had no authority and we don't change the phenomenon.

We simply observe the phenomenon and we tried to present a model, so obviously the first sentence this is correct that the model generates the data and second sentence that the data generates a model this is not correct. But what is now really happening, we want to know the parameters of a model. The parameters are unknown to us, so the question is how should we know them? Once I know the parameters then the entire model is completely known to us.

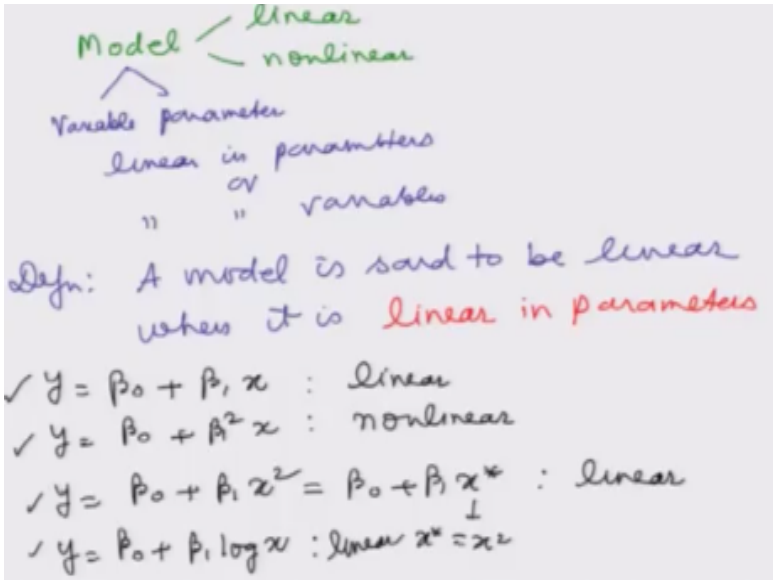
So I can say here that for example if I try to make a simple diagram that this is here somewhere this is the process, this process is happening and there is a model, which is describing this process and when we try to observe the process we try to collect the data. What do understand by this sentence, for example if I take a simple example of agriculture, suppose we decide that there are three variables say this year quantity of fertilizer, irrigation level and quantity of seeds.

These are the 3 variable we consider and we assume that they are going to affect the yield of a crop suppose I say that I try to use one kilogram of fertilizer and I put 20 centimeters of irrigation and I use 5 kilogram of yield and based on that I try to observe some outcome of the yield say suppose I get 20 kilogram of yield. Now I try to repeat this experiment again with some different values.

For example now I say am using here two kilogram of fertilizer and I try to use here thirty centimeters of irrigation and suppose I use seven kilogram of yield and suppose now at this time I get 30 kilograms of yield. So, we try to conduct this type of experiment and then we try to collect the data and now my objective is that based on this set of data I want to know the model.

So here you are saying that the model is generating the data, but now what are we doing, we are trying to collect this type of observations and we are trying to find out a mathematical relationship between outcome, which is a yield of a crop and these three variables. So, now if you try to observe, what are we doing? we are going in the opposite direction. The model generates the data, but we are trying to observe the data and we are trying to find out the model.

(Refer Slide Time: 08:13)



So do not you think that we are going in the opposite direction, answer is, yes and this is the advantage of regression analysis when I say what is the meaning of regression? Or more general if I say what is the meaning of regress? This means to move in the backward direction, means what why are moving in the backward direction, what I am trying to say model generates the data that is acceptable that is the correct sentence, but now we are going in the backward direction.

We are trying to observe the data and then we are trying to find out the model that is why this is called a regression analysis. We go in the opposite direction, we collect the data and from that set of data, we try to infer about the phenomenon and then we try to express it in the mathematical format and whatever mathematical format we get that is essentially a model. So now if I say another aspect about model.

There can be 2 types of model, one is mathematical model, and others are statistical model. But, what is the difference between that 2, 1 of the important difference between this mathematical model and a statistical model is that they are usually exact in nature and a statistical model they take care of random variation, and the problem with this random variation is that this variation is not in our control.

Let me take a very simple example to explain the difference between the interpretation of a mathematical model and a statistical model. So let me try to consider a simple equation $y=3x+2$ and suppose I say that y is my yield of a crop say in kilograms and x is my quantity

of fertilizer and that is also in kilograms. Now suppose if I try to put here some values of x and y. If I try to put here x=1, I will get here y=3 in 2.1+2=5.

Similarly If I try to put x=2 I get y=8 and so on and suppose if I try to put x1000 then I get y=3002. So now what is the mathematical interpretation? The mathematical interpretation says that once we are trying to use 1 kg of fertilizer then we will get exactly 5 kg of yield, more deviation from 5, exactly 5.00000 and so on.

Similarly if try to use here 2 kilogram of yield, then I would get exactly 8 kilogram of fertilizer. In case if I try to keep on increasing the quantity of fertilizer say for example 1000 kilogram then I will get exactly 3002 kilograms of yield and from the mathematical point of view I can say that y is an increasing function of x, but is that really correct. Now let us try to see the statistical interpretation.

We know from our common experience that whenever we try to take particular size plot in which we try to put suppose 1 kg of fertilizer there is no 100% that I will exactly 5 kilogram of yield. But possibly what we mean by saying 5 kg of yield is that if I try to repeat the experiment several time this means if try to take the same size of plot and almost under similar condition.

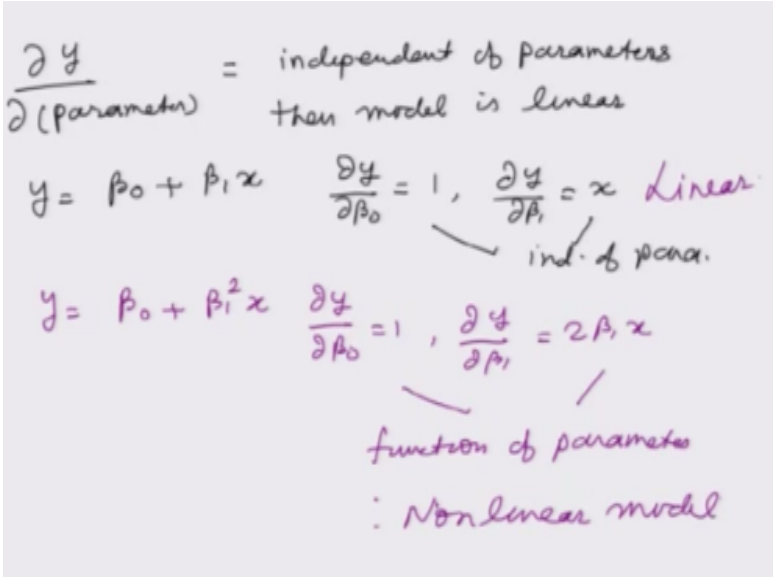
If I try to put suppose 1 kg of fertilizer every time then some time I will get a yield of 5 Kg some time 5.5 Kg sometime 4.5 Kg sometime 5.2 Kg sometime 4.8 Kg and so on. Once I try to take the average of all these values that will be close 5 kg. So, similarly if I say from the statistics point of view if I try to use 2 kilogram of fertilizer.

Then on a average I will get here 8 kilogram of yield. But here comes the main difference the mathematics says if you keep on increasing the quantity of fertilizer the yield is going to continuously increase. But now the statistics comes into picture and we use our common sense and we try to observe the phenomenon and we say, yes the quantity of yield is going to increase as the quantity of fertilizer increases.

But this happens only up to a certain extent, and if you try to increase the fertilizer beyond a certain limit the entire crop will burned up. So, that is the main difference between a

statistical interpretation of the data and the mathematical interpretation of the data. Please note that I am not criticizing either statistics or say mathematics but I am trying to take a very simple example and I am trying to explain that what is the difference between a statistical model and a mathematical model.

(Refer Slide Time: 15:03)



Both of them have their own importance. Now we try to take up another concept in the model. So model can be linear or this can be nonlinear and you also have seen that model has 2 components one is variable and another is parameter.So now the question is when you would call a model to be a linear, when it is linear in parameters or linear in variables, that is the question.

Here I would say this is the basic definition, a model is said to be linear when it is linear in parameters and this is very important definition. So now just to illustrate it let me try to take some examples suppose if I take a model only in with the one input variable something like beta0+ here beta1 x.

What do you think, here there are two parameters this beta0 and beta1 and both of them have got the power one, so this going to be a linear model. Whereas if I try to take it here suppose here beta0+ beta1 square x, so now you can see here although beta0 as got a power one but beta1 is a square quantity, so this is a nonlinear model. Another example if I try to write down beta0=beta1 say x square.

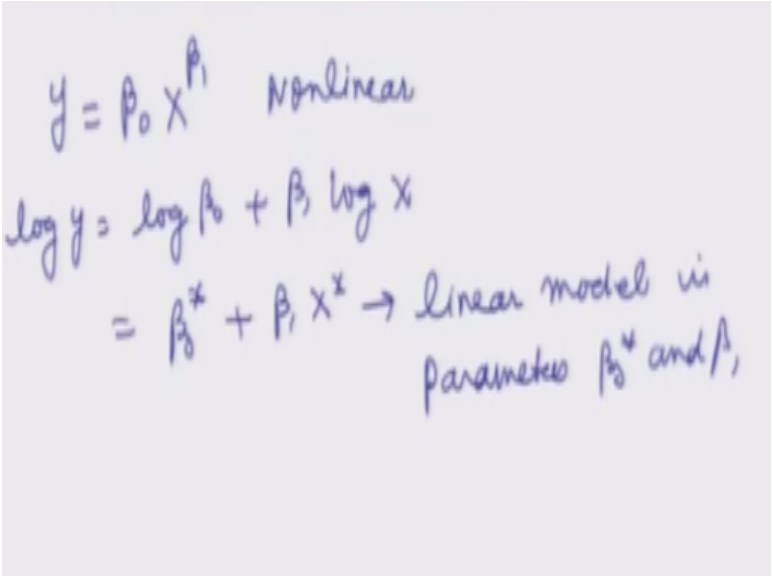
Then it is very simple I can rewrite this model as a beta0+ say beta1 say x star right, where a x star is, say this here x square so this is again going to be a linear model. So this is a basic definition that you have to keep always in mind. Another example I can also write beta0+ beta1, log of x this is also a linear model, so this is what you have to keep in mind that how do define the model to be linear or nonlinear.

Now our issue is this, we want to translate it into a proper definition, so if you try to observe over this or say these examples. So now we need to translate this definition into a proper way, looking at these four equations, If you try to observe, incase if I write the partial derivative of the output y with respect to all the parameters if this is independent of parameters then model is linear. For example if I try to write down y=beta0+ beta1 x.

You can see here that partial derivative of y with the respect to beta0 this is one and partial derivative of y with the respect to beta1 this is x and both are independent of parameters. Let me try to take some examples if I say y= beta0+beta1 square x, now in this case partial derivative of y with respect to beta0 is 1, but partial derivative of y with respect beta1.

This is twice of beta1 x and this is a function of parameters. In this case I can say this is a linear model whereas in this case I can see that this is a nonlinear model.

(Refer Slide Time: 19:56)



Now let me try to take another simple example and let me try to explain how a linear model can be converted into a nonlinear model, suppose if I try to consider a function like beta0 x

rest of the power of here beta1, obviously this is nonlinear, but in case if I try take it here log on both the side then this becomes log of beta0+beta1 times here log of x, so this I can write down here beta0 star and this I can write down here as x star.

Although I can see here that this is essentially a nonlinear model, so now we can see here that this becomes a linear model. But you have to keep in mind the parameters are now changed. Linear model in parameters, beta0 star and beta1 and now variables are also changed here you had x and now your input valuable is x star, so this one example where I have tried to show that in case if we have a nonlinear model.

Possibly that can be converted into a linear model and we can continue with our statistical tools which we try to develop for a linear model. In this lecture two we have tried to consider some other basic elementary concept related to the regression modeling and in the next lecture we will start with a simple linear regression model, till then good bye.