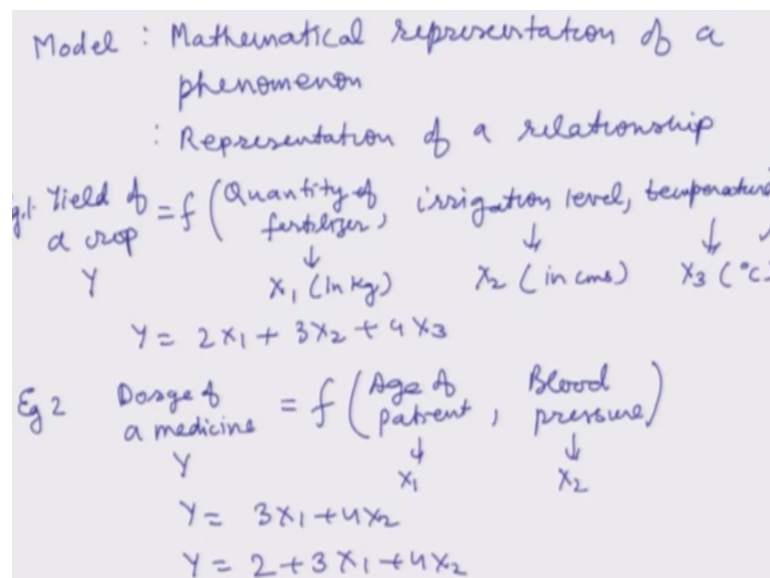


**Regression Analysis and Forecasting**  
**Prof. Shalabh**  
**Department of Mathematics and Statistics**  
**Indian Institute of Technology – Kanpur**

**Lecture - 01**  
**Basic Fundamental Concepts Of Modelling**

Welcome to the first lecture on regression analysis and forecasting in this lecture we are going to learn the basic fundamentals there are very small topics, but these are the topics which create the background to learn the statistical tools relate to the regression analyses later on. So in this lecture we are just going to learn the small topic, the basic fundamentals which will help us later on.

**(Refer Slide Time: 00:43)**



Model : Mathematical representation of a phenomenon  
: Representation of a relationship

Eg 1. Yield of a crop  $= f$  (Quantity of fertilizer, irrigation level, temperature)  
 $Y$                        $X_1$  (In kg)                       $X_2$  (in cms)                       $X_3$  ( $^{\circ}C$ )  
 $Y = 2X_1 + 3X_2 + 4X_3$

Eg 2. Dosage of a medicine  $= f$  (Age of patient, Blood pressure)  
 $Y$                        $X_1$                        $X_2$   
 $Y = 3X_1 + 4X_2$   
 $Y = 2 + 3X_1 + 4X_2$

So, well if I ask you what is a model that is my very basic question to all of you? You see the model is nothing this is only a mathematical representation of a phenomenon. What you mean by this? This is nothing but a mathematical representation of phenomenon. What does this mean? For example when you try to conduct an experiment you know there is a process which is happening and in that experiment there are two types of variables.

One variable which is an output variable and there are some variables which are causing that output. So now in case if you try to find out the relationship between the input and output variables that will be called as a model. So essentially model is nothing, this is simply the representation of a relationship. Now let me take a very simple example to explain you, for example if I sale a yield of a crop, yield of a crop depends on several factors.

Like quantity of fertilizers, quality of fertilizers, irrigation, rainfall, temperature and so on. For example let me take some variables like quantity of fertilizer, irrigation level, and temperature and so on. Now in case if I try to find out a mathematical function relationship between these three variables quantity of fertilizer, irrigation level and temperature then this will be called as mathematical relationship between yield of crop.

And this three variables, suppose if I denoted by Y suppose quantity of fertilizer this is suppose denoted by some variable  $x_1$  and supposed this is measured in kilogram this irrigation level. Let us try to denote here as a  $x_2$  and because this is measured in centimetres and temperature let us try to denote by some other variable  $x_3$  and supposed this is measured in degree centigrade.

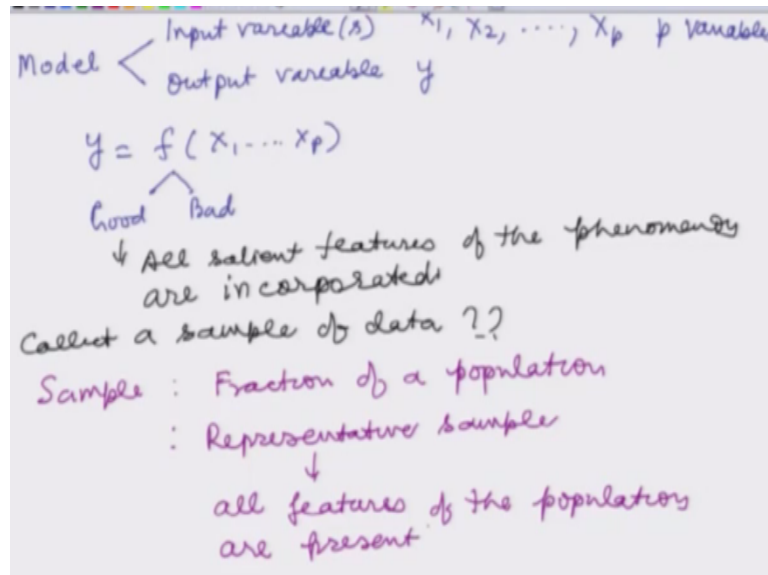
So one possible model can be  $y=2x_1+3x_2+4x_3$ , this is a mathematical relationship between the output which is a yield of a crop and 3 input variables which are your quantity of fertilizer, irrigation level and temperature. Similarly if I try to take another example suppose, if I say dose of a medicine. Dose of a medicine mean, whether a patient has to be given half tablet, quarter tablet or a full tablet or 2 tablets.

That depends on several factor and one of the important factor is age of patient. Second thing can be his health conditions which can be determined by different variables like his blood pressure, blood sugar or say some other variable. So let us try to assume for a while that a dose of a medicine depends on the age of the patient and blood pressure, and when we try to find out the functional relationship or a mathematical relationship between the dose of a medicine with respect to age of patient as well as blood pressure.

Then this will be a sort of mathematical model. For example if I try to denoted just dose of medicine by  $y$ , age of patient by  $x_1$  and blood pressure  $x_2$  with some appropriate units. For example if I try to write three  $x_1+4x_2$  or that can also being something like  $2+3x_1+4x_2$ , so these are simply mathematical equation, which are trying to represent a relationship between the input variables and say output variable.

So they are essential a model. Now once you try to see in this model there are 2 types of component, one is input variable or they can be more than one variables and there is an output variable.

(Refer Slide Time: 05:18)



So we will try to denote the output variable by  $y$  and the input variable suppose if I say there are  $p$  variable so we can denote them by  $x_1, x_2, x_p$ , so these are here  $p$ ,  $p$  variables which are affecting the output  $y$ . What is your objective? The objective is that we want to find out a relationships between  $x_1, x_2, x_p$  and  $y$  in a mathematical frame work, so now the question is this what is this model representing?

This model is actually representing some phenomenon, for example if I take that same example of yield of a crop. The yield of crop that is happening, that has been controlled by irrigation level, temperature, quality of fertilizer, quantity of fertilizer and so on, we have no control over it and there is some mathematical relationship which existing in the nature the problem is that we don't know direct relationship.

And our objective is that if somehow I come to know about this relationship possibly that will help us for the better future, and better planning. So now there can be two things suppose I say that this is my model, so now this model can be a good model or this can be a bad model. What you really understand by a good model and a bad model, in a very simple sense I can say a good model will be a model, which incorporates all salient features of the phenomenon.

For example, if I say that in the model for yield of a crop, for example in this case if I try to take the same thing here that yield of the crop, this depends on quantity of fertilizer, irrigation level and temperature. Now suppose for a while you say well I try to find out a relationship where yield of the crop depends on some factor and we say we are not considering this quantity of fertilizer.

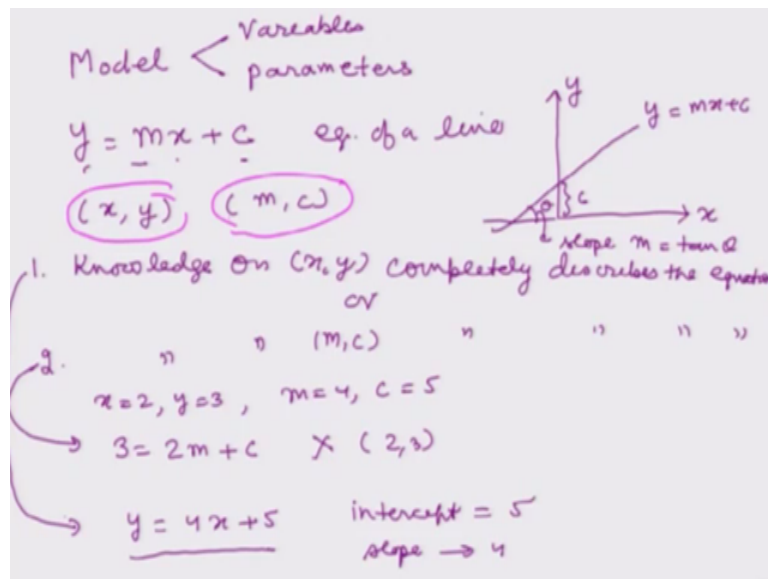
Do you really think that is it going to be good model? Not really, because we all know, we are the social scientist and we know that the yield of a crop depends on the quantity of fertilizer up to certain extent. I can say here that in case of a good model I would like to have that all salient features of the phenomenon are incorporated. The next question comes well I would like have a good model no doubt about it, but then how to obtain it.

So for that we try to collect a sample of data. What is sample that is a basic question, sample here is nothing but this the fraction of a population. A small fraction of the population but these needs to have one very important statistics that we assume that sample is representative sample. This is essentially a representative sample in the sense that whatever are the features have level in the population they are present in my sample.

Representative means that all features of the population are present, and now as a statistician how do we start, we try to collect a data, we try to conduct the experiment or we try to observe the phenomenon and we try to collect the data. Now one thing which what you have to keep in mind which is very important is that as a statistician we have no right to change the process.

Whatever is the natural mode of happening of a phenomenon or an experimental process that would continue, so now what you have in your hand before you start, you simply have a small sample of data. The data can be of twenty observation, thirty observation, hundred observation or two hundred observations depending on the situation. So now let us try to try to consider another aspect of a model, a model has two components once is variables and another is parameters.

**(Refer Slide Time: 11:13)**



Now what is the difference between a variable and a parameter let me try to explain it in a very simple and layman's language you all have learned this equation in your possibly class twelve or class ten something like this  $y=mx+c$  what is this, this is a equation of a line. Now there are four components here  $x, y$  and  $c$ , so we are going to try to group them into two groups  $x, y$  and say here  $m$  and  $c$ .

We all know that this equation if I try to draw on a graph, if this is something, something like a line this I know this my  $x$  axis and this my  $y$  axis and this is my line  $y=mx+c$  and this distance is  $c$  and this is the slope which is represented by here  $m=\tan \theta$ . Now let us try to consider this line and now I ask you one simple question or if I write two statements.

Let me try to explain which of the statement is correct. First statement is knowledge on  $x$  and  $y$  completely describes the equation or the second sentence is this knowledge on  $m$  and  $c$  completely describes the equation, because at this moment we are considering a simple equation, so the second sentence is that knowledge on say  $m$  and  $c$  completely describes the equation, now what you think which one is correct.

Let me take simple example suppose if I say  $x=2$  and  $y=3$  and suppose  $m=4$  and  $c=5$ , right. Now once I try to consider here this statement number one, so knowledge on  $x, y$  is known to us, so I can write this model  $3=2m+c$  and now in case if I try use here this thing the second statement then I can write down here  $y=4x+5$ .

Now, I arrived these two equations now do you know or can you say that which of the equations is completely describing the line, well I can see here this is not describing the line because these are the values of  $m$  and  $c$  which are unknown to us and this is only telling us that there is one pair of point this is two and three that's all. On the other hand if you try to consider this equation, this is telling us everything about the line.

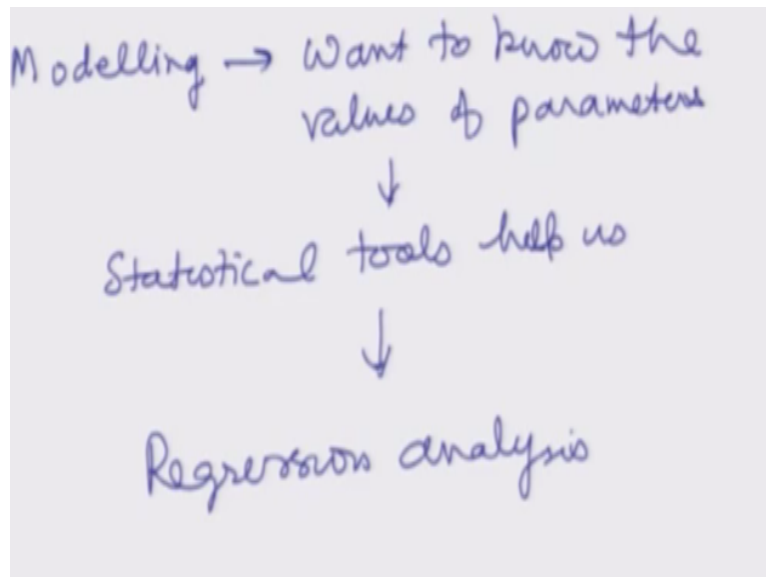
This is giving us complete information about the line. It is telling that intercept term is, intercept is equal to 5 and the slope can be controlled by the value four, that  $4 = \tan \theta$ , because the slope is  $\tan^{-1} 4$ . So now once I know the value of  $m$  and  $c$ , I can always find that for a given value of  $x$  what will be the value of  $y$  or for given value of  $y$  what will be the value of  $x$  and all information about this line is completely known to us.

So broadly speaking here I have a two sets of values  $x$  and  $y$  and say  $m$  and  $c$ , which I am marking here  $x$  and  $y$  and say here  $m$  and  $c$ , and we observe that by knowing the value of  $m$  and  $c$  I can determine the complete line, where as just by knowing the value of  $x$  and  $y$ , I cannot know the entire line, so in a layman's language I can say that  $x$  and  $y$  are my variable and  $m$  and  $c$  are the parameters.

So parameter are essentially the component of a model, which determine the entire equation or in our language parameters are the component of an equation which describes the complete the model, so whenever I say that I want to do modelling that is very simple sentence that I just want to know the values of my parameters that's all. So you have heard many people say okay he want to do the modelling and modelling is this and modelling is that.

But now you can say that modelling is nothing, but just knowing the values of the parameter now but then what is hurdle in obtaining a model. The hurdle is that parameter values are not known to us. The question is how to know them once I know the parameter value the entire model is known to us. So in a statistics I would say that whenever I say that I want to do a modelling this means what?

**(Refer Slide Time: 17:16)**



It simply means I want to know the values of parameters. So question is that how do we know it, they are unknown to us they exist in that nature, that is okay, but we have no idea, so in this case the statistical tools help us. Now there are different types of statistical techniques, which can help us in finding out the values of these parameters and among those techniques we are going to use here the technique of regression analysis.

Right and this regression analysis tool will help us to find out the values of the parameters on the basis of a given set of data. So there are many statistical tool which helps in knowing the value of this parameters and from those techniques we are going to use here the regression analysis technique. This regression analysis technique as an advantage that it helps us in finding out the unknown values of the parameters on the basis of a given set of data

And this data is obtain from some experimental setting from some experiment or from some survey and so on. So here we stop and we will continue in the next lecture, thank you.