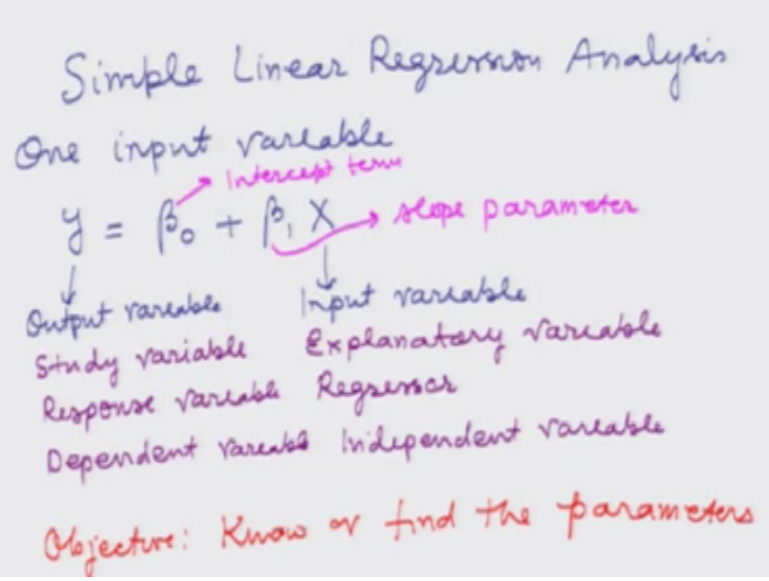
Regression Analysis and Forecasting
Prof. Shalabh
Department of Mathematics and Statistics
Indian Institute of Technology – Kanpur

Lecture – 03
Simple Linear Regression Analysis

Welcome to the lecture number 3 from this lecture we are going a start with basic fundamental of a linear regression modeling we are essentially going to a start with. the chapter.

(Refer Slide Time: 00:23)



Simple Linear Regression Analysis, well in practice the contents of this chapter or the basic concept which we are going to learn in this chapter may not really be helpful. More important chapter will be the next chapter that will be multiple linear regression analysis, but whatever concept we are going to learn here they are going to build up the basic fundamental for the next chapter.

The different between simple linear aggression analysis and multiple linear regression analysis is that, in the case of simple linear regression we are going to consider only one input variable whereas in the case of multiple linear regression analysis we are going to consider more than one input variables and in practice we know that any output is depend on more than one input variables.

So that will be a more realistic chapter, but whatever the concept we are going to discuss in the case of multiple linear regression model they are base on the concept that we are going to

learn in this simple linear regression model and as a instructor it is also easy for me to explain things when there is one variable, and I can use this one dimension and two dimension graphics and later on I can simply extended to a multiple case.

Here we are going to consider a situation where we consider only one input variable. So now we consider that the output variable y that is linearly related by this function beta0+beta1X. Now you can see here, this is the same model that a we had discussed in case of lecture number one and lecture number two okay, so in this case just for your information this is so called output variable.

This is how we have denoted it earlier and this was we had called earlier in an elementary language as input variable. Now we are going to talk about that in pure statistical language. In pure statics language this y and x they have got different names for example this y is called as study variable and in connection with the study variable this x is called as explanatory variable. This has got several other names also.

For example this y is also called as response variable and when we talk of response variable then we talk of x as regressors or say sometime regressor variable and so on and this is also called as dependent variable, y is called as dependent variable and in connection with dependent variable I can call this x as independent variable, so similarly there are some other names also that are popular in the literature.

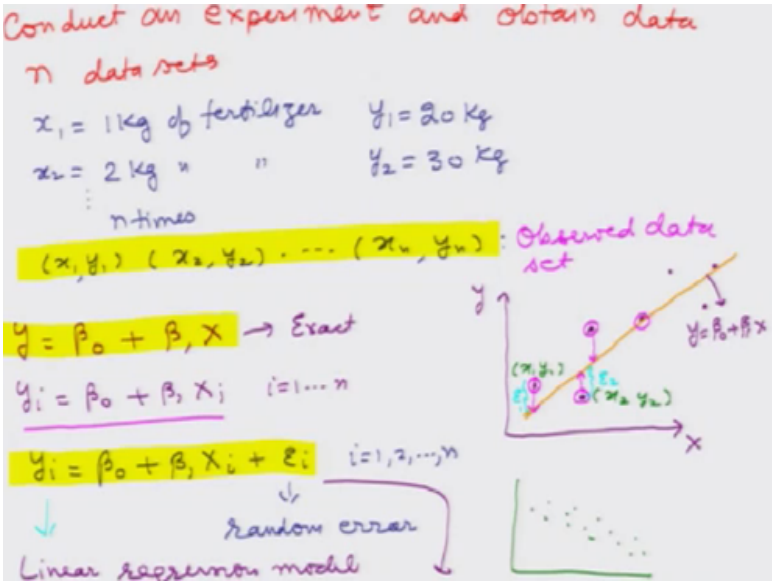
But these are some common names that we try to use for this y and x and just for your remembrance this beta0 we had taken this was our intercept term and this beta1 was the slow parameter if you try to recall in the lecture number two we had consider the linear equation y=mx+c which I had translated in terms of y+beta0+beta1 x, so this is my model.

The question is that how to know this model, we had discussed earlier that model is nothing, but a function between the output and input variable and when I say I want to find out the model that is equivalent to saying that I want to find out the parameters of the model. Once I say that I want to know the model, our objective is to know or find the parameters.Up to now we have a defined here two parameter one is beta0 and say another beta1.

Well they can be some other parameter also that we will see it as we move further. Now we are discussing here the technique of regression analysis, we had discussed that regression means to move in the backward direction, so what are we going to do here that we are going to a conduct an experiment or we are going to observe some data in some real life experiments, and finally we will have sample of data.

Now I have to find out the values of the parameter or equivalently I want to find out the model using this data, so let us try to start with this thing, suppose I say that now we are going to conduct an experiment and obtain data. Well, we are going to find some finite number of data. Let us say we want to obtain n data sets.

(Refer Slide Time: 06:39)



What is this mean? That means we are going to conduct the experiment and record the data n times this means what? For example if I consider a same example that we consider earlier that y is the yield of a crop and x is the quantity of fertilizer then for example I can take here that x1 that= 1 kg of fertilizer and I give it to the field and then I try to observe how much yield I get after sometime.

Suppose I get the yield y1= 20 kilogram, then next time I try to take here 2 kg of fertilizer and this will be the value of x2. and suppose I get 30 kg of outcome, 30 kg of yield so this going to be my y2 and so on I try to repeat it in times, so what will happen at the end I will have a data something like x1, y1, x2, y2 and so on say xn, yn.

So, this is my observed data set. Now what is your s your objective? Your objective is that using this data set I have to find out the value of the parameters of my model and that would be nothing but finding out a model. Now the next question is how to start it, well, we have assumed that there is a linear model between y and x that we have to keep in mind because we are going to consider here only the linear regression modeling.

So the first step is that I can plot this data on a 2dimensional plot something like this, this is my x axis, this my y axis and suppose I plot data and say x1, y1, say x2, y2, x3, y3, x4, y4, x5, y5, x6, y6 and so on. Now, one can see from this graph that the points are following a sort of linear trend right and one can see here that if somebody tries there is a trend like this one.

But this only a trend and now if I assume that this is my model for a while then I would like to know what is the equation of this line? So incase if you try to see let me mark here these values here something like this my x1, y1, this is my x2, y2 and so on. On other hand it is not always necessary that the trend between x and y is always increasing in fact there can be other possibility also that the observations are lying like this one

So once you get the data try to plot it and then try to see the trend in the data. Incase if the trend in the data is linear that give us first assurance yes, a linear model can be fitted to the given dataset or in simple words I can assume, yes the process can be describe by a linear equation. Now you if you try to see I simply assume that there is a model, the model is now a going to be something like y= beta0 + beta1 x this the model that we have considered.

I believe that whatever observation I have obtain here x1, y1, x2, y2, and xn and yn they are actually going to follow this model and this observation are generated from this model this means that these set of observation will satisfy this equation, beta0+ beta1 xi, i goes for 1 to n. But if you try to observe this graphic this two dimension plot is it really happening?

No, if I try to write down this orange line is something like y= beta0 + beta1X, and we are assuming that whatever are my points here they are lying on this line, so essentially what we are assuming that this point, this point is going lie somewhere here, this point is going to lie somewhere, this point is going to lie here and well this point is lying on the line and so on.

But now you can see that this model is not appropriate because that is not really describing the true phenomenon. The observation which I have got here x1, y1, x, y one, x n, y n they are indicating, yes the approximate relationship is here yi= beta0 + beta1 xi, but this not exact. So I can do one thing in order to make this model more realistic I can add here a term, for example I can rewrite it yi= beta0+beta1 xi and + some random error epsiloni, i goes from one to say here n.

What is this epsilon i, epsilon I is the random error involved in ith set of data random error.So when I observe the ith set up of observation xi, yi, this has got some error epsiloni. So I can assume here for example in the case of this first observation x1, y1, I can write that okay this difference whatever is going to be here this is trying to denote something like epsilon1.

Similarly when I try to go for this x2, y2, this difference is going to be something like epsilontwo and so on. So in some cases you can see here that this epsilonone this distance, this is the distance of x1, y1 from this line that we want to know, that we want to fit and this point x1, y1 is lying above the line whereas if you try to see in the case of x2, y2 this epsilon2 is the difference between x2, y2, and this line that you want to fit and this observation is lying under the line.

The same thing is happening with all other points, so I can say that some random errors are in the upward direction of the line and some random errors they are under the line, so I can assume that here that epsilonI has to be a random quantity, which takes some value that can be positive or that can be negative. So now I call this thing as a linear regression mode and if you try to see what is the difference between this model and this model.

Now if I say what is the different between these two model, so if you try to see this was a sort of exact relationship, but this is a this is going to be a, going to depict a statistical relationship, this means what I am not saying that all my observations are exactly going to follow the straight line, but they are very close to the straight line and there are certain random errors which are deviating the observations from the main line.

So now I can see her that this is, this orange line is going to depict the line that you want to fit on the basis of this given set of data say x1, y1, x2, y2, x n, y n and now our objective is this

how should I find this line, but before that we have to make some assumption for this random error. So what we try do here, we are going to assume here that the mean of epsilonI is 0.

In statistical language I can write down here say expected value of epsilon i=0 for all i goes from 1 to n, well if you do not know the meaning of this E, I can just explain you before I move further E is the expectation operator.

We define it in the following way if I say if there is a random variable, some random variable z, which is going to follow a probability density functions fz with a parameter theta, this is my probability density function where z is the random variable and theta is the parameter then we try to define say expected value of z integral over the range of z, fz, of dz. Incase if you are using a discrete random variable then in place of the probability density functions.

We try to use the probability mass function and this integral is replaced by the submission, similarly at the same point I can also tell you that how do we define the variance. We will denote it by Var, variance of z, this is defined as expected value of z - expected value of z whole square. So this is nothing but integral of z minus expected value of z whole square fz, dz and the integral is a over the range of random variable z, so this how we try to define this expectation and variant.

So now I can make two assumptions as I wrote here expected value of epsilon i to be 0 and variance of epsilon I to be sigma square. What does this mean? Once I write that expected value of epsilon i=0 that means we are trying to say that the mean of epsilon i is 0, this means that some observations they have got random error, which is positive and some observation they have got random error which is negative.

And once you to try find out its arithmetic mean there average value is going to be 0 and that is expected also from the precaution point of you that in a realistic experiment sometime the errors are positive and sometime the errors are negative, so it is reasonable to assume, that the average of those errors is going to be 0. But it is a random variable, epsilon i is a random variable so we need to describe its behavior by variance also.

For example if you try to see in this figure also this epsiloni sometimes they are lying in the upward direction sometimes they are lying in the downward back to correction and moreover

every observation will have a different type of random error with different amount of random error, so that is described by the quantity of sigma square. In simple words if say, if the sigma square is low, then I would say that my observations have got less variance.

And they are lying closer to the line and in case if this sigma square is high then I would say observations have got more scatteredness and observations are lying quite away from the line. At this moment, if I try to consider this model that y= beta0 + beta1 x+ here epsilon, so you can see her that I have assumed this to be random. Now we also assume something about y and x and let us see what their behavior.

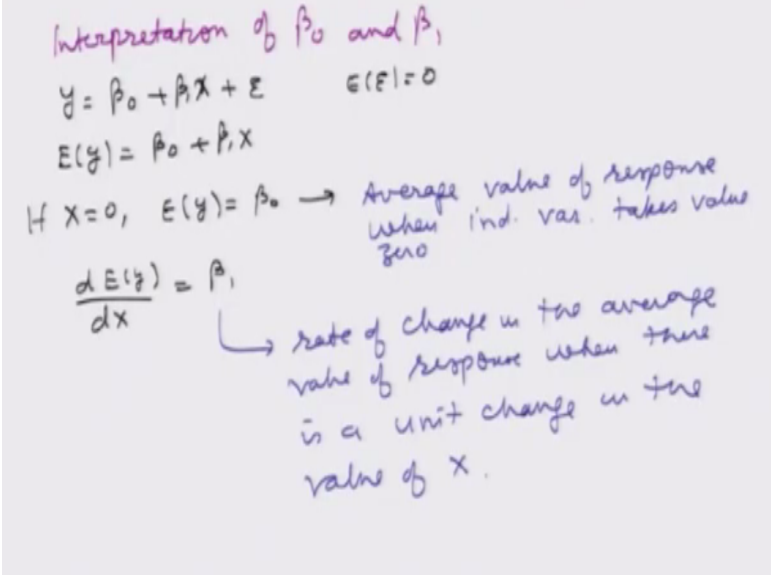
So we assume here that this x is also non-stochastic, non-stochastic means it is nonrandom in simple language. In practice, there can we situation where x can be random also, but here in this course we are going to assume that x will remain as fixed. Similarly going on the same concept I also assume that beta0 and beta1, they are my parameters and I will say that they are fixed but unknown.

In some cases this beta0 and beta1 can also be random, but we are not going to consider those situations in this course. So now since epsilonis random beta0, beta1 and say x they are not random, so this y also becomes random. After this if you try to see we also have assume that is expected value epsilon is 0 and variance of epsilon sigma square. So that means if I want to know my complete model I also need to know the value of sigma square.

So the sigma square also become a parameter of the model so now you can see here that we have got here three parameters, beta0, beta1 and sigma square. So once I say that I want to know my model that is equivalent to saying that I want to find out the values of beta0, beta1 and sigma square and how to find them just on the basics of sample of data x1, y1, x2, y2, x n, y n.

We also assume here that when we observe the observations x1, y1, x2, y2, x n, y n then the corresponding random errors epsilon1, epsilon2, epsilon n, they are IIDs, IIDs means they are identically distributed as well as they are independently distributed. So all epsilon1, epsilon2, epsilon n, they are mutually independent of each other and all epsilon1, epsilon2, epsilon n, they are coming from the same distribution.

(Refer Slide Time: 25:01)



Now let us try to understand what is the interpretation of beta0 and beta1.So if you try to look at this model y= beta0+ beta1, x+ epsilonwe had assumed that expect value of epsilon =0, so I can write expect value of beta of y=beta0+ beta1 x. So now if I say x 0 that independent variable takes the value 0 then expected value of y= beta0.

If I try to find out the first derivate of expect value of y with respect x then this comes out to be beta1 So now this beta0 and beta1 have some interpretation, beta0 is the average value of y or average value of response when independent variable takes value 0 and beta1 is the rate of change in the average response, average value of response when there is a unit change in the value of independent variable x.

So this is how we try to interpret the values of beta0 and beta1 in case of linear regression modeling, so beta0 is simply the average value of a y when x takes value0, and beta1 is the slope of the line, which is measured by the first derivative of average value of y with respect x.

Now we stop here and in the next lecture, I would explain you how to estimate this parameters using two techniques, one is ordinary least square estimation and another is maximum likelihood estimation, till then good bye.