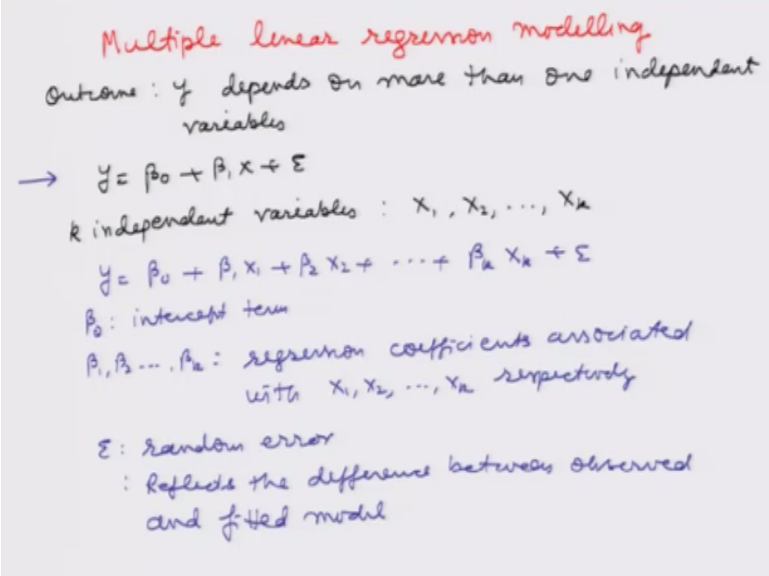
Regression Analysis and Forecasting
Prof. Shalabh
Department of Mathematics and Statistics
Indian Institute of Technology-Kanpur

Lecture – 11
Multiple Linear Regression Model

Welcome to the lecture today we are going to start with a new topic which is multiple linear regression modelling.

(Refer Slide Time: 00:20)



If you recall we started with the simple linear regression model, where we consider the situation where the outcome is going to be depended only on one independent variable now we are going to extent it. In practice this situation is more realistic, the outcomes usually depends on more than one factors or more than one variables, so we are going to consider here a situation where the outcome is going to depend on more than one independent variables.

The situation is the following that in the case of simple linear regression modelling we have developed many concepts and I have tried to explain you there utility and their interpretation, the same concept, the same interpretation will be brought forward in the case multiple linear regression modelling, so it is my request that before you start with multiple linear regression model it is very important that you are clear about all the concepts of the simple linear regression model.

Here we believe that the outcome which we had denoted as y this depend on more than one independent variables and earlier we had discussed the simple linear regression model that

was beta0+beta1x+epsilon, now here we assume that there are more than one independent variables and suppose there are k independent variables, and we denote them by here x1 x 2 up to here xk.

So the same model which we have considered in the case of simple linear regression model this can be extended to the case when there are more one independent variables, and this can be written as see y =beta 0 +beta 1 x1 + beta2 x2 up to here say beta k x k+ epsilon. Now about the interpretation means earlier we had said that this beta0 is the intercept term and this remains the same here.

And we say now that beta1, beta2, this beta k they are the regression coefficients associated with x1, x2, xk respectively, so essentially this beta j is the regression coefficient associated with jth explanatory variable xj, and epsilon because of the same thing as our random error. Now in this case the role of random errors becomes quite important when we are dealing with the real life situation.

The first step in doing a regression modelling is to identify what are my independent variables or what are variables, which is going to affect the outcome why? When we try to do so sometimes it is possible to obtain the observations on those variables and sometimes it becomes difficult to obtain the observations on the independent variable. For example if I take a variable like taste or intelligence.

It is difficult to obtain the numerical values on the variables like taste or intelligence. The intelligence is usually measured by IQ scores, but we are against a sort of indirect major of intelligence, similarly there are some variables which may not be very important or they may have a very small affect on the outcome y based on that some time we try to consider them or sometime they don't consider them.

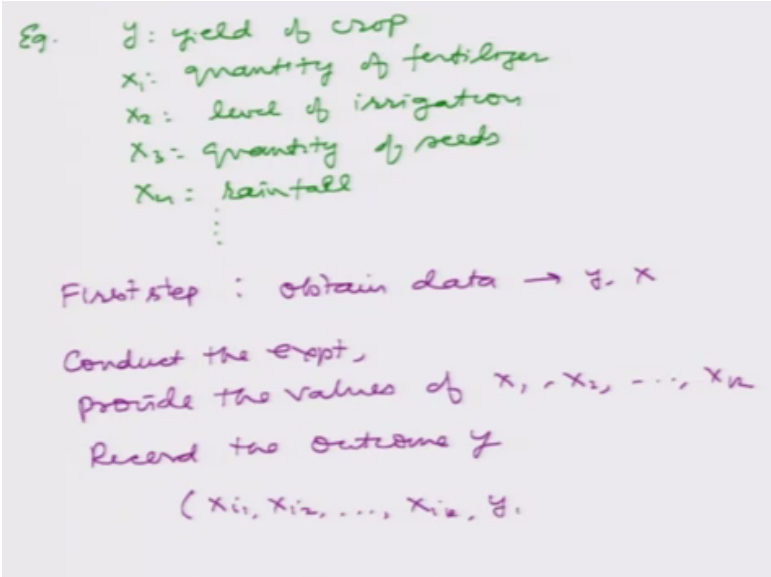
On the other hand in case if the number of explanatory variable become very, very large the situation become more critical and in that case we would try retain only the important variables which are trying to affect the outcome y. There will be many, many situations which are beyond our control and epsilon denotes the joint affect of all such factor which is beyond our control.

So epsilon is like a basket in which we try to put all those things which are beyond our

control. This epsilon essentially depicts or it reflects the difference between observed and fitted model and this area goes exactly on the same lines what we had done in the case of simple linear regression model. The situation in which we can use this multiple linear regression model are many.

First of all I try to extent the same example which I had considered in the case of simple linear regression model in the case of simple regression model I had taken an example of yield of a crop

(Refer Slide Time: 07:29)



where I denoted y as the yield of crop and we had taken x as the quantity of fertilizer, but do you really thing that the yield of a crop depends only on the quantity of fertilizer, but it depends on several other factors so now we have an opportunity to incorporate all those important factor which are affecting the yield of a crop, for example the first factor I can write down x1, which is my quantity of fertilizer similarly x2 can be level of irrigation.

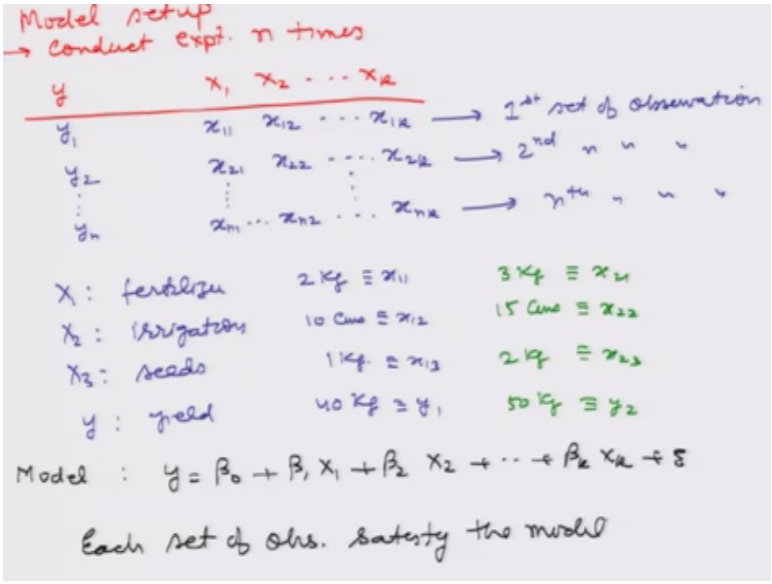
Third thing can be the quantity of seeds, x four can be rain fall and similarly you can identify some more important factors which are affecting the yield of a crop. Now under this things we have to now develop a multiple linear regression model, you may recall regarding the case of simple linear regression model the first step what we had defined for a linear regression model is to obtain data and in case of simple linear regression model we had obtain the data on y and x.

We conducted an experiment, we provided the value of x and then we had observed the values of y and this experiment was repeated n times. The same has to extended here also that

we have to conduct the experiment, provide the values of x1, x2, xk and then record the outcome y, so if you try to see earlier we had set of observation like xi, yi but now we are going to have a set observation ray which is something like xi one, xi two and up to here say the xik and yi, and i goes from here one to n in case if try to repeat the observations n times.

Earlier we had a assumed that all the observations they will also follow the same model, and in the case of simple linear regression model we had the model y = beta0, + beta1 x +epsilon and we assume that all observations xi, yi they are going to follow the same model and they will satisfy yi =beta 0 + beta 1xi + epsilon i. We have to extent the same definition in the case of a multiple linear regression model, so let us first set up our model.

(Refer Slide Time: 11:02)



So we consider the model set up it as simple as that conduct experiment n times, and we are going to obtain here the values of y and x1, x2 up to here xk this is how we are going to obtain our values suppose I conduct the experiment and I give x1 a value say x1, x2 a value x 1 2and xk a value x 1k and based on this values we try to observe the outcome y and we denote it here as y one, so this is our first set of observation.

Similarly we try to obtain the second set of observation that we try to give the value x1 as a x2 1 x 2 the value x 22 and say xk the value x2 k and we obtain the observation y two, so this gives us the second set of observation and we continue with this thing and finally we obtain the nth set of observation by giving x1, x2, xk the values xn 1, xn2, xnk and we observe the value here yn so this is the nth set of observation.

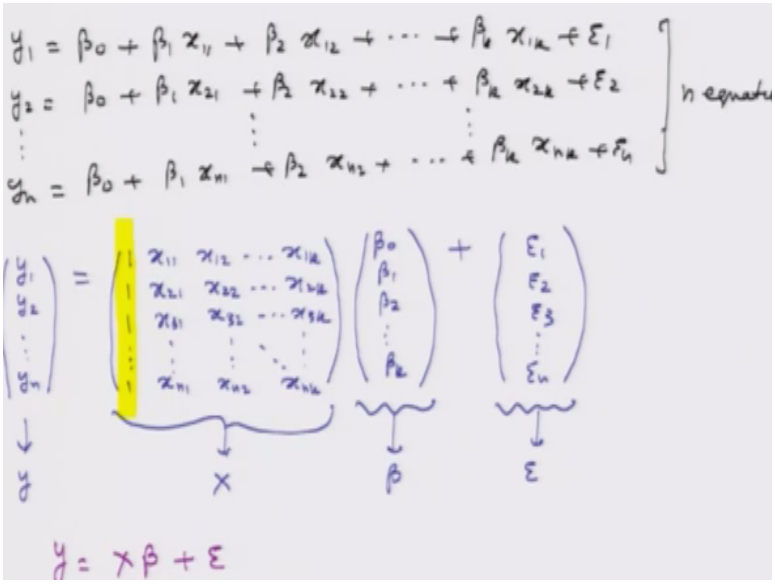
What is this actually mean for example if I try to take the same example of yield of crop, so

for example if I say x1 is my quantity of fertilizer, and see here x2 is my irrigation level and x three is my suppose seeds, what we try to do here suppose I try to give two kilogram of fertilizer and say ten centimetre of irrigation suppose I use one kilogram of seeds and based on that we try to observe the yield and we get suppose here forty kilogram of yield.

This is why x1 1 this is my x1 2 this my x1 3 and this is my y one and similarly I can repeat this experiments and I can take say I use three kilogram of fertilizer say this 15 centimetres of irrigation two kilogram of yields and based on that we observe suppose fifty kilogram of yield so this will be denoted here as a x2 1, this will be x 2 2, this will be x2 3 and this will be y2.

Similarly we try to repeat this experiment n times and we obtain n sets of observation, so now if you see we have here a model, which y = beta0+ beta1 x1 + beta2x 2 +up to here beta k x k plus epsilon, now we assume that that each set of observation satisfy this model.

(Refer Slide Time: 15:12)



This means I can express for the first observation I can write that y1 = beta0 + beta1 x1 beta 2 x 1 2 + beta k x 1k + epsilon1. Similarly for the second observation I can write down the model as a beta0 + beta1 x2 1 + beta2 x 2 2 + beta k x2 k + epsilon2 and so on for the nth observation I can write down y n= beta0 + beta1 x n1 + beta2 x n 2+ beta k x n k + epsilon n.

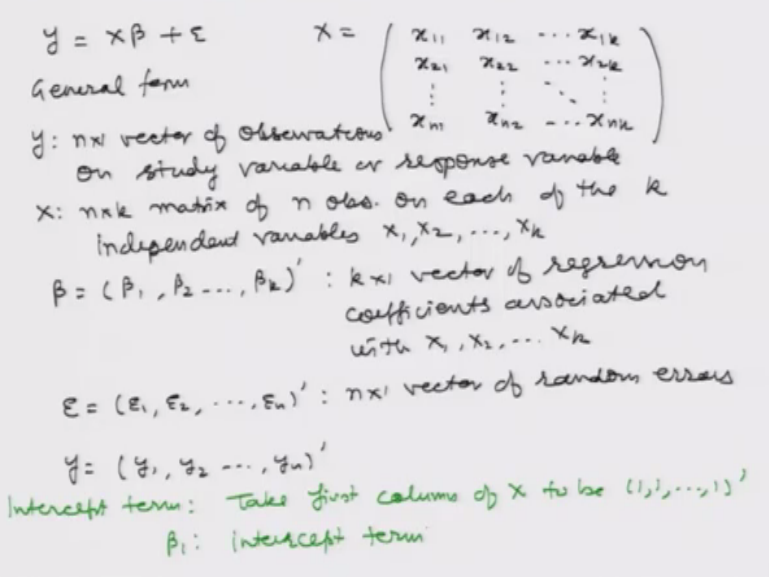
So essentially if you see here we have got here n equations, now these n equations can be expressed in the form of a vectors and matrix, so we can write down this n equations as follows, let us try define here vector of y1 y2 y n and this is equal to so we define here one matrix and here we define here a vector beta0 beta1 beta2 up to here say here beta k and

based on that the first row of this matrix will be one x1 1x one2 up to here x one k.

The second row will1x 21 x 22 x 2 k and similarly the third row will be x 31 x 32 up to here x 3 k and this will continue up to here x n1, x n2 up to here x n k and+ epsilon one epsilon2 epsilon three up to here epsilon n. So now I can denote this vector as y and this matrix here as x this vector here has beta and this vector here as epsilon, so i can write down the entire model as a here y = x beta+epsilon.

Now we try to observe here that this first column is here only 1,1,1,1,1 this is indicating the intercept term this can be made a little bit more general

(Refer Slide Time: 18:41)



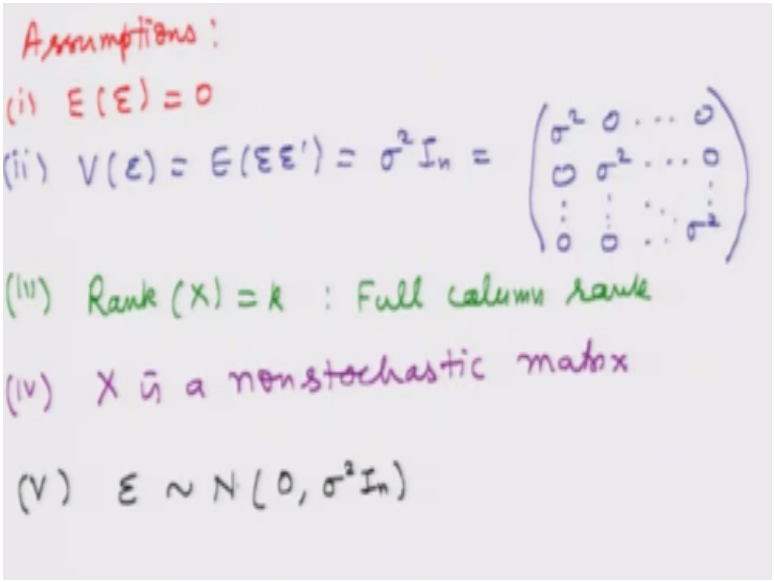
That I can write my model in journal has se here y is equal to x beta+ epsilon and where I can say that x is going to be something like x11 x 1 2 x 1 k, x 2 1 x 22 x 2 k, x n 1 x n 2 x n k and in case if I want to consider the intercept term in the model then the first column of the x matrix has to be made 1,1,1,1 and in case if i don't need an intercept term in the model this x matrix will remain as such.

So this is a very general form, in which we assume that y is say n cross one vector of observation on study variable or let me call say response variable some time x is a n cross k matrix of n observations on each of the k independent variables x1 x2 x k beta is going to be something like beta1, beta2 and beta k, this is going to be a k cross one vector of regression coefficients associated with x1 x2 x k and epsilon here is as usual epsilon1, epsilon2, epsilon n which is n cross one vector of random errors.

For the sake of completeness I can also write here y as a y1, y2, y n transpose, now the question is that in case if I want to have intercept term in the model then what I have to do take first column of x matrix to be 11 say 1 and then correspondingly this beta1 will become the intercept term, so now onwards we will start with the model y=x beta+epsilon and we will not bother whether there is an intercept term or not.

In case if I wanted the intercept term I simply have to write the first column of x matrix to be 1,1,1,1 otherwise I will simply continue with the s matrix as the matrix of the observation obtain on the explanatory variable.

(Refer Slide Time: 22:50)



You may now recall that in case of simple linear regression model we had made certain assumption about the model the similar assumptions we are going to make for the multiple linear regression model, so if you remember the first assumption what we had made was that expected value of epsilon i is 0, now in case of multiple linear regression model we do not have one epsilon i but we have a vector of epsilon i so I can assume that expected value of epsilon=null vector.

The interpretation part of this thing that we already had discussed in the case of simple linear regression model, the second assumption is about the variance covariance matrix, so we assume that the variance covariance matrix of epsilon which is the same as expected value epsilon, epsilon prime this we assume is sigma square i n, so it is something like this it will look like this the diagonal elements.

They are going to denote the variances of epsilon1 epsilon2 epsilon n and the half diagonal

elements they are going to denote the covariance between epsilon i and epsilon j, which are zero. Again, this is the same assumption that all epsilon i's are ours identically and independently distributed, so we can see here from this matrix that we are assuming that all epsilon1, epsilon2

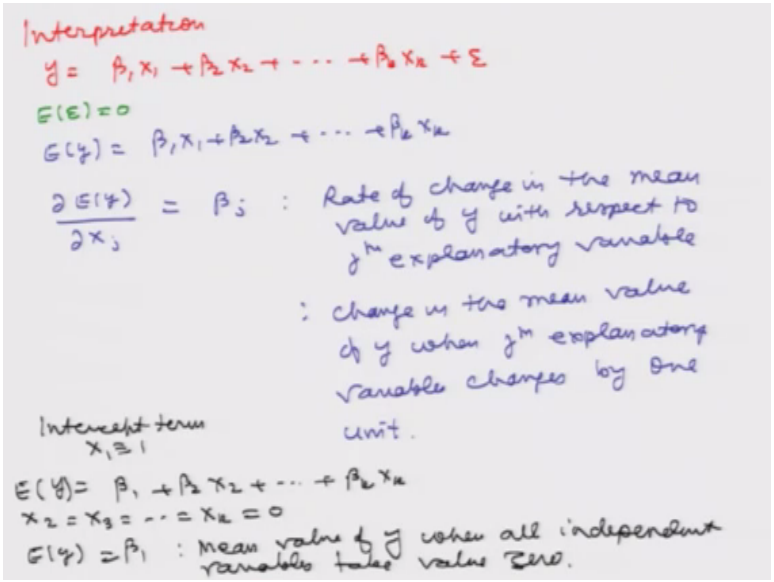
Epsilon n they are having the same variance sigma square and they are mutually independent of each other. The third assumption which we are going to make here is that rank of x matrix is going to be the k, and remember k is the number of independent variable so essentially we assume that this is a full column rank, the advantage of making this assumption will be clear to you in the next lecture when we go for the estimation of parameters.

The next assumption we make is that x is a non-stochastic matrix you may recall that similar assumption was also made in case of simple linear regression model where we assume that x is a fixed quantity, it is an non-stochastic random variable, so similarly here we are trying to make it more general we have now not one variable but more than one variable so we trying to extent the same assumption of the simple linear regression model to a more general case for all the k independent variables.

The last assumption what we make here that epsilon are following multivariate normal distribution with null vector and covariance matrix sigma square i n. This assumption is a gain similar to the assumption what we made in the case of simple linear regression model there we assume that epsilon i's are following our normal distribution a univariate normal distribution with mean 0 and variant sigma square.

Now we are trying o extent it for all epsilon1 epsilon2 epsilon n, again I would like to emphasize that the utility of normal distribution comes into picture when we consider the maximum likelihood estimation of the parameter or when we go for the test of hypothesis and confidence interval estimation.

(Refer Slide Time: 26:54)



Next we come on the aspect of interpretation of these regression parameters, so we have considered here a model y= beta1 x1 + beta2 x2 + beta k x k +epsilon and now we have assumed that expected value of epsilon=a null vector, so I can write down expected value of y to be here beta1 x1 + beta2 x2 + beta k x k and now.

Here itself you can see the utility of assuming that x k are non-stochastic another advantage of assuming that x one x two x k are non-stochastic is that the outcome of the experiment will not be dependent on the values of x1 x2 x k. So if somebody is conducting an experiment in city number one and somebody collecting the observation in city number two and somebody else is collecting the observation in city.

Number three then whatever the analysis we are going to obtain on the basis of collected set of data that is not going to dependent on the city number one, city number two or say city number three right but that will be valid for everyone. Now based on this if I try to find out the partial derivative of expected value of y with respect to here certain variable x j this comes out to be beta j.

So you can see here that beta j is nothing but the rate of change in the mean value of y with respect to jth explanatory variable. So this essentially denotes the change in the mean value of y when jth explanatory variable changes by one unit, and if you try to recall this is the similar interpretation as in the case of simple linear regression model so whatever interpretation I had given to beta1 in case of simple regression model that is now extended to beta1 beta2 beta k.

In case if you say what is the interpretation of having an intercept term in the model so in

case if I try consider here a intercept term so I simply have to take here all value of x1 to be one, in this case, the model will become expected value of y beta1 + beta2 x2 plus beta k x k. Right, so if try to take all x2, x3 and all other values of x2, x3, x k to be 0 then expected value of y becomes nothing but beta1.

So in this case also the intercept term will denote the mean value of y when all independent variables take value 0 and again this is the same interpretation that we had given in the case of simple linear regression model there I consider only one variable x to be 0, now am saying that all x1, x2, xk they are going to take the value 0. So we have completed here the description of the model, in the next lecture we will consider the estimation of the model parameters, till then good bye.