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 - 45
Multiple Linear Regression Analysis
Basic Concepts

Hello 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. In this module on the linear regression analysis we are going to begin with a new chapter today on Multiple Linear Regression Analysis.

I am sure that you all of be getting irritated with me when I was trying to teach you simple linear regression model because every time I will say that ok when we will do it in multiple linear regression model right ok. So, well there are now many questions which are cropping up and we would like to obtain the answers of those questions one by one.

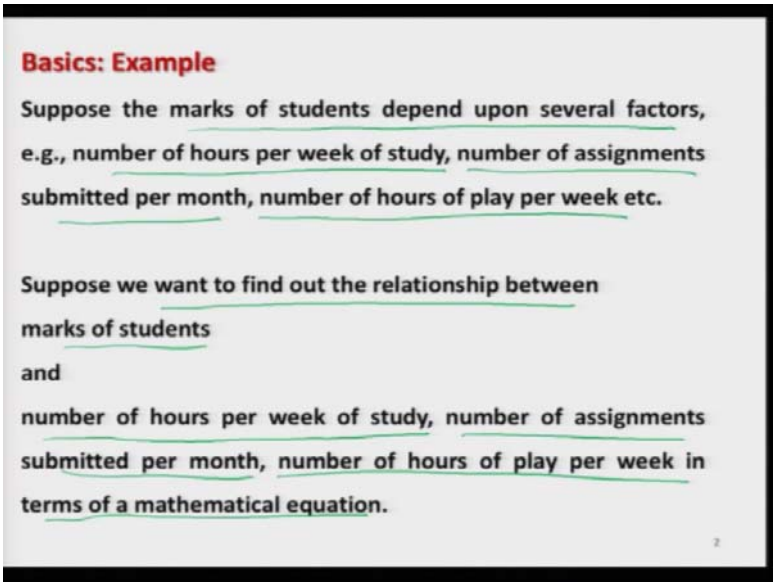
First thing I can assure you this is one of the most useful tool in statistical modelling. Even whatever you will think I will say just make it double then whatever you think then I will say again make it double, that will be the importance of this topic.

So, the first question comes what is multiple linear regression model? So, what I am going to do, I am going to do this lecture almost on the same lines as I did in the case of simple linear regression model. First I will try to take the basic concepts and now I will be using the concept from simple linear regression model and I will try to extend it to the multiple linear regression model. And I will to try to take those small issues and in every case I will try to show you the statistical part, computational part as well as interpretation.

And I will try to take the same data set which I took earlier, but I will be just adding here one more variable so that you will have an opportunity to compare the results whatever we are going to obtain here in multiple linear regression model and whatever we have obtained in the simple linear regression model.

You can compare that how the things are changing right. For example, we had obtained the regression coefficients β_0 and β_1 by ordinary least square estimation numerically. So, now, you can see that now what will happen if some more data is added there will be many more things. So, let us begin and try to learn it one by one, ok. So, let us start, right.

(Refer Slide Time: 03:01)



So, the first question comes over here that, what is multiple linear regression model? So, let me try to take one example. Earlier we had considered an example where we discussed that what are the different variables which affect the marks of a student? So, we had considered say just one variable that the number of hours of study per week..

But do you think as a student that is it the only variable beside this thing number of hours you have spent in the library; that means, how many hours you have studied books how many assignments you have attended, how many hours you have played, how many times you have attended the class all these factors you know they are going to affect the marks.

So, now first question is how to take care of these variables? How should I obtain a model which is a function of all such variable? Up to now my model was a function of only one variable, but now we will have a general model which will depend on all these factors all these variables.

Some of the variables are going to be very important good variable which are contributing in the modelling some variable may not be good for example, if I say do you think that the price of the petrol will really influence the marks of the students think about it.

Yeah, some people may argue yes I have a scooter by which I come to my college and if the price of the petrol increases then it becomes difficult to come to the college and that is why I have missed the classes well these arguments are not really acceptable. So that means, there will be some variable which may have some indirect relation with the output, but they may not be that important.

So, that will be another challenge for us that how to identify those variables. And now if I tell you that whatever test of hypothesis you have used that will be used in identifying this thing will you believe on me certainly not, but I will prove it. This is what I said in the beginning that all the concepts of simple linear regression model will be used in multiple linear regression model to make bigger conclusions.

So, now what I am assuming here suppose I want to make a model of number of marks of the students which are depending on more than one variable. Similarly, if you try to take another example from the this agriculture, the crop of a field the amount of crop of a field depends on many factors quantity of rainfall, quantity of fertilizer, quality of soil, different chemicals compounds in the soil, temperature and so on.

So, you would like to have a more realistic model where the outcome that is the yield of the crop is dependent on all such variables. Similarly, if you come to a medical experiment the effect of a medicine is measured on patients, this depends on several factors age of the patient, gender of the patient, height of the patient, weight of the patient, blood pressure of the patient, blood sugar of the patient and there can be many parameters.

So, now we are going to start a more realistic modelling, where the dependent variable is going to be depend dependent on more than one explanatory variable and when we have this type of setup this is called as multiple linear regression model ok.

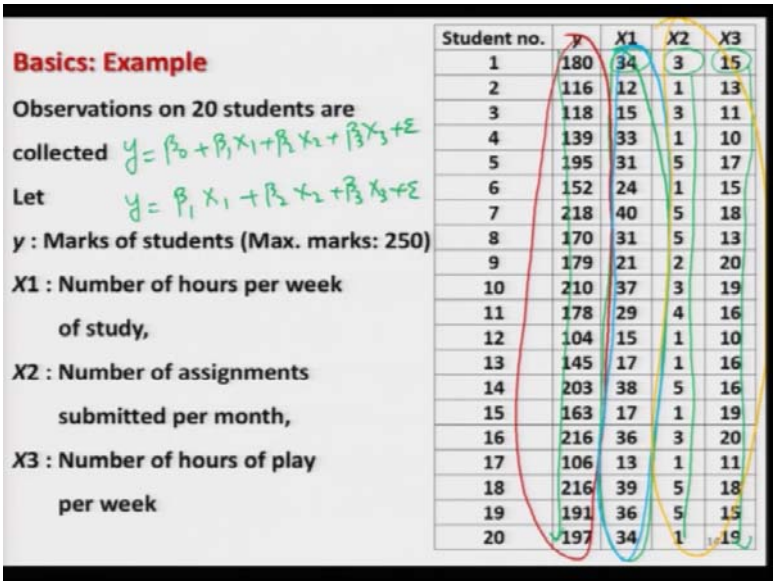
So, now, in order to move forward we consider the same example that we considered earlier that we had obtained the marks of the students, which we know that depend they

depend on several factors several variables number of hours of study per week, number of assignments submitted per month, number of hours of play per week and so on there are many other things.

So, now, just for the sake of simplicity because I have a limitation on my slide I have a limited space to type. So, I will be considering only here three such variables, but there can be more such variable and they can be dealt exactly on thus on those lines.

So, suppose we want to find out the relationship between marks of students and all this variable number of hours per week of study, number of assignments submitted per month, and number of hours of play per week. And I want this relationship in terms of a mathematical equation now you know; that means, I want to find out a fitted model.

(Refer Slide Time: 08:16)



Student no.	y	X1	X2	X3
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

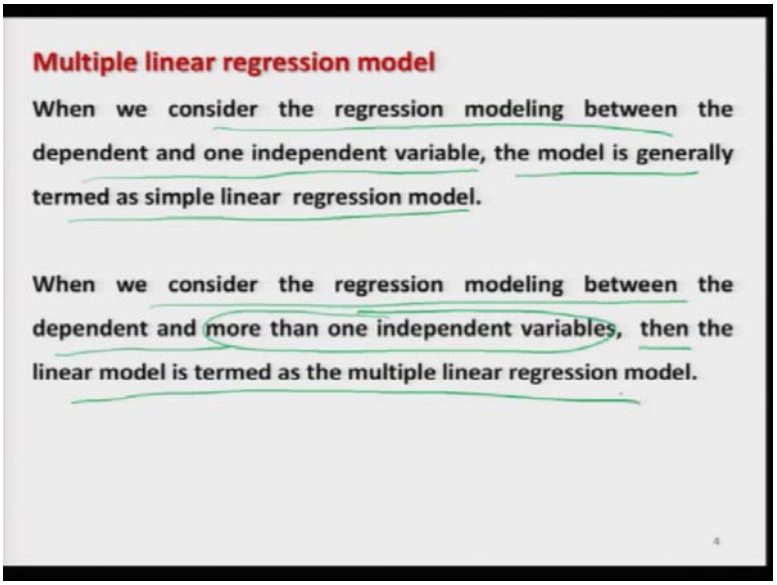
So, now this is our data set have a look because this data set is going to be used again and again here and when I try to implement all these things in the R software. So, we have obtained the data on 20 students which are numbered as 1 to 20.

And then we have obtained the data on the marks of student which are denoted by here y then data on the number of hours per week of study which is denoted by here X_1 which is here in the table it is in the 3rd column. Then similarly we are denoting number of assignments submitted per week by X_2 and number of hours of play per week this is X_3. So, and this data here is obtained in the table.

So, the interpretation of the data goes like this suppose I consider student number 1 the student number 1 has got 180 marks out of 250 and this student has studied 34 hours per week and the student has submitted 3 assignments per month and the student has played 15 hours per week.

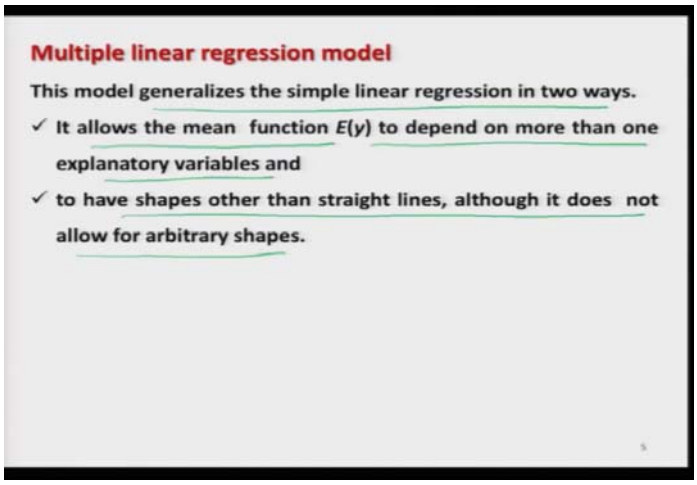
Similarly, the student number 2 the student has got 115 marks out of 250 and the student has studied for 12 hours every week submitted three 1 assignment in a month and played 13 hours in a week and so on this data is collected for all the 20 students ok. So I believe now that X_1 , X_2 , X_3 they affect the values of y.

(Refer Slide Time: 09:56)



So, when we try to consider the regression modelling between the dependent and only one independent variable then we have learnt that this model is generally called as simple linear regression model. Now, when there are more than one independent variable; that means, when we consider the regression modelling between the dependent and more than one independent variables then the linear model is termed as the multiple linear regression model, ok.

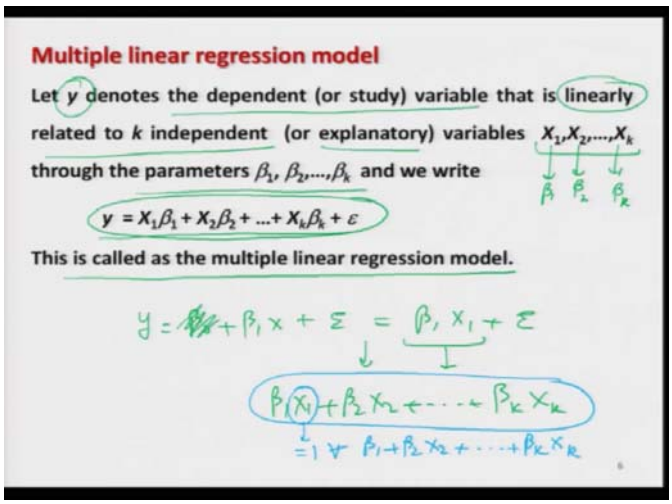
(Refer Slide Time: 10:28)



So, actually this multiple linear regression model generalizes the simple linear regression model in two ways. First it allows the mean function which is the expected value of y to depend on more than one explanatory variable's, earlier $E(y) = \beta_0 + \beta_1 X$ so there was only one X now there will be more than one X's. And this can also entertain the different types of shapes for example, straight line although it does not allow for any arbitrary shapes right..

There are different types of things which I will try to show you in the forthcoming slide that it is possible that if a relationship which is looking non-linear in the scatter diagram may be considered or may be transformed into a linear relationship and which will look at least linear on the scatter plot.

(Refer Slide Time: 11:27)



Ok. So, now, we will use the similar notation that we use in the case of simple linear regression model here this small y is going to denote the dependent or steady variable. And we assume that this dependent variable is linearly related to k independent variables or k explanatory variables.

In earlier we had denoted the one explanatory variable as X, but now we have here X_1 , X_2 ,…, X_k . And earlier the regression coefficient or the regression parameter associated with X was β_1 in the case of simple linear regression model, but now we have here X_1 , X_2 ,…, X_k so I can do one thing I can associate one β with each of this variable. So, X_1 will have the coefficient β_1 regression coefficient X_2 will have regression coefficient β_2 and X_k will have regression coefficient β_k..

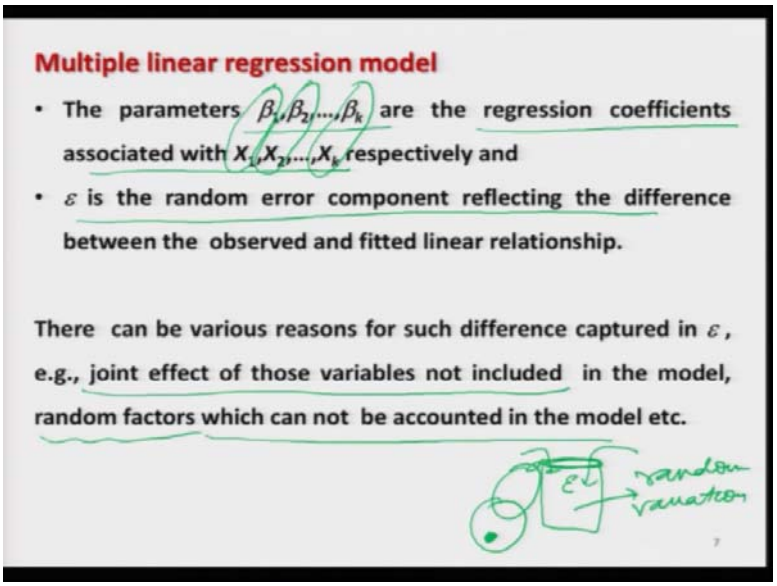
So, now, we have here small k number of parameters β_1, β_2,…, β_k and we can extend the line what we had considered earlier for example, we had considered $y = \beta_0 + \beta_1 X + \varepsilon$ right. For a while I am not writing here β_0 because I will show you that β_0 can be entertained in a much simpler way in case of multiple linear regression model, but suppose this is my model in general say $\beta_1 X_1 + \varepsilon$ that was the simple linear regression model.

Now, I am trying to extend it to k variable so this $\beta_1 X_1$ can be extended to $\beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$ right. So, this is what I am writing here $y = X_1 \beta_1 + X_2 \beta_2$ + up to $X_k \beta_k + \varepsilon$, ε you know that is the same random error component as we have discussed in the case of simple linear regression model and this model is called as multiple linear regression model ok that is a very general form right.

And now if you are confuse that why I am not writing here the intercept term. Let me show you suppose I take look here I try to take this X_1 to be which takes value 1 for all the values and so now, this will become here something like $\beta_1 + \beta_2 X_2$ + up to here $\beta_k X_k$.

So; that means, if you want to have then have an intercept term in the model then you simply try to take the first variable which always takes the value 1 that is all. Do not worry I will try to show you in more detail in the further slides.

(Refer Slide Time: 14:32)



But before we try to understand what are these things. So, now, we have understood that in the case of multiple linear regression model the parameters are β_1 β_2 β_k which are called as regression coefficients.

Regression coefficients associated with X_1 , X_2 ,…, X_k respectively. That means β_1 is related to X_1 β_2 is related to X_2 and β_k is related to X_k . And ε is the random error component which is reflecting the difference between the observed and fitted linear relationship that was our basic fundamental concept.

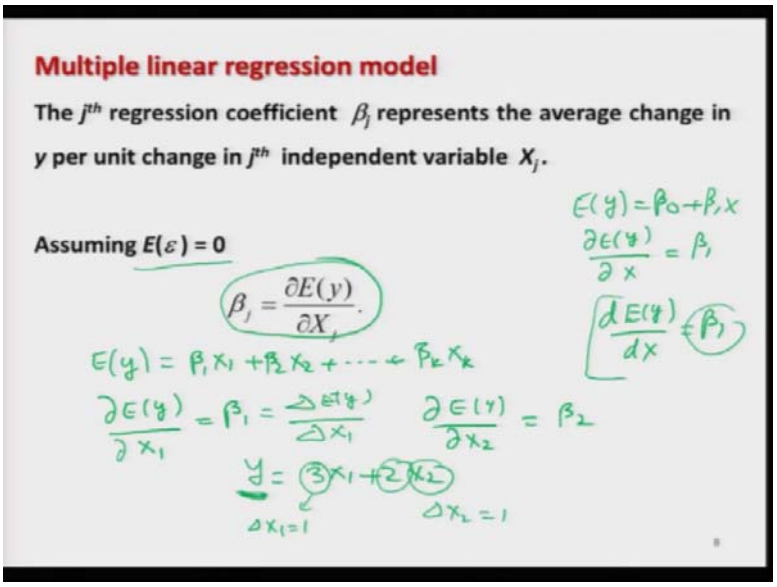
But now, in this case this can be generalized and there can be various reason which can contribute in this ε For example, there are we are considering here more than one variable so, there can be a joint effect of those variable which we are unable to measure. So, that is why we have not included in the model there are certain random factors which cannot be accounted.

For example if I ask you a simple question what is the temperature in your room you will say some numbers say 28 degrees centigrade, but tell me one thing that do you think that at every point inside your room the temperature is 28? For example, the location where you have a fan or air conditioner or a cooler around that the temperature is going to be less and if you move away from these gadgets the temperature will increase.

So, now you cannot say that the temperature inside the room is uniformly 28, but what are you saying it is close to 28 there is some difference in the observed and the true values. So, all such things which you cannot entertain just try to create here a basket name it ε and whatever you cannot observe, whatever you cannot entertain just try to put all the effects inside this basket and close it. And after closing it just make here a big lock this is here the hole of the key lock the key and throw away the key.

So, now you do not know what is happening inside ε you cannot observe, what you can observe there is only some random variation that is all, but it does not mean that if you know that there is one variable there is some there are some variables which are very important which are going to affect the model, but still since you do not want to collect the observation you are throwing them in the ε no that is not correct.

(Refer Slide Time: 17:28)



All those things where which are random where you cannot handle because of any reason those effects are put in the ε random error or disturbance term. Now we come to the aspect of interpretation of this regression coefficient.

So, this interpretation goes exactly on the same line as we have does as we have discussed it in the case of simple linear regression model. If you remember we had taken the model $E(y) = \beta_0 + \beta_1 X$ and then we had differentiated say expected value of y with respect to X and this was coming out to be β_1.

Actually there we consider the exact differential so, this was β_1 and it was trying to give us the rate of change in the average value of y when there is a unit change in the value of X. So, this was the interpretation for β_1.

Now, if you try to extend this concept on the platform of multiple linear regression model then you have here more than one β_j. If you assume that expected value ε is equal to 0 then I can write $E(y) = \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$.

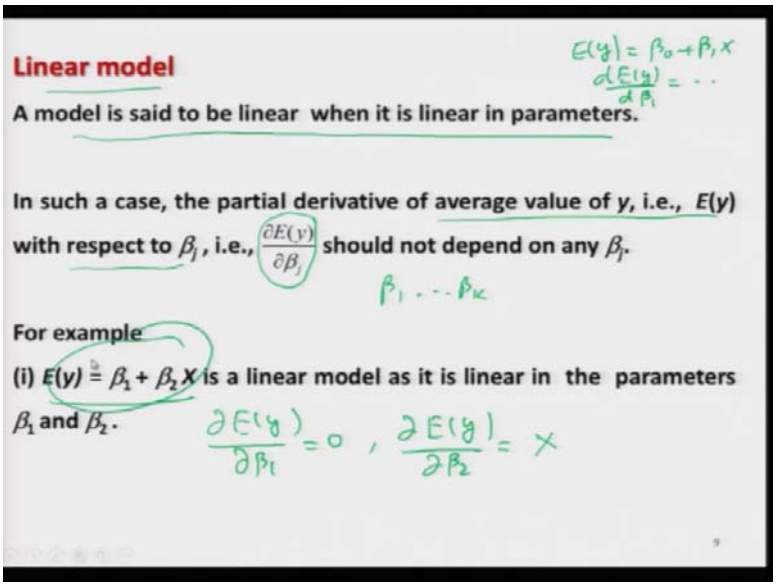
And now, if you try to partially differentiate E(y) with respect to say any variable say let me take it here X_1 this will come out to be β_1. So, now, you can see here that β_1 is nothing but the change in the average value of y when there is a unit change in the value of X_1 . So, this is the change right.

Similarly, if you try to say here a partial derivative of E(y) with respect to X_2 this is say here β_2. So, in general I can write down that β_j = partial derivative of expected value of y with respect to X_j.

So, this is going to give us the rate of change in the value in the average value of y when there is a unit change in the value of X_j, right. So, if you have a model like $y = 3 X_1 + X_2$ for example, suppose this is the fitted model suppose you obtain it. So, this is in the indicating that the average value of y will increase by 3 units if there is a unit change in the value of X_1 .

And similarly if there is a unit change in the value of X_2 then the average change in the value of y will be 2 units. So, this is how we interpret the value of the regression coefficient and the same interpretation will go when you try to estimate the value of betas on the basis of given sample of data right.

(Refer Slide Time: 20:42)



So, now we come to another definition linear model. So, now, if you remember in case of simple linear regression model we had discussed that a model is said to be linear when it is linear with respect to the parameters we had not considered the aspect of linearity with respect to the variables.

Now, the same concept is going to be extended in the case of multiple linear regression model also. And if you remember what we had done we had considered the model the simple linear regression model like expected value of y is equal to say here $\beta_0 + \beta_1 X$ and then we had said you try to take out the derivative of expected value of y with respect to here say here β_0 or say β_1 and when you try to see here whether this term comes out to be independent of the parameters or not based on that you can do it right.

So, the same concept is being extended here and we simply try to say that in such a case you try to take the partial derivative of the expected value of y with respect to each β_j means earlier you had only two parameters β_0 and β_1, but now you have here k parameters $\beta_1, \beta_2, \ldots, \beta_k$.

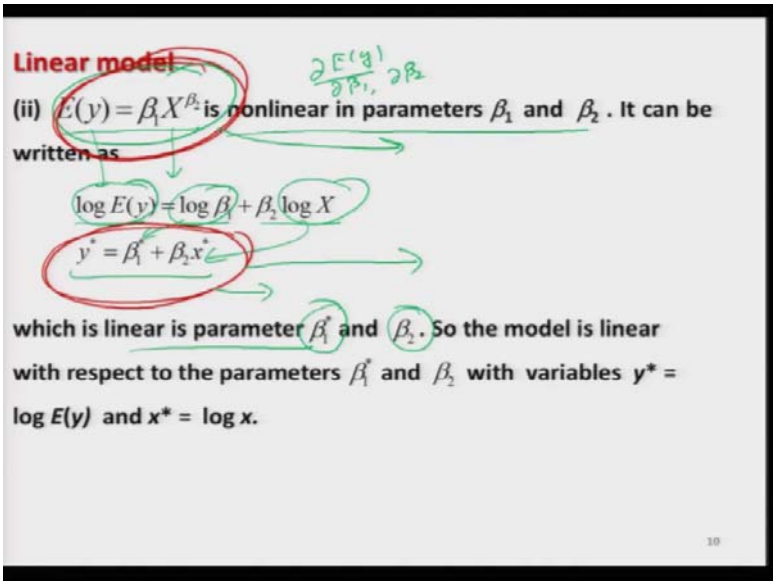
So, try to find out the partial derivative of the average value of y that is respective value of y with respect to β_j and you have to check whether it depends on the parameters or not. So, the model is said to be linear if all the partial derivatives with respect to each of the parameter are not depending on any of the $\beta_1, \beta_2, \ldots, \beta_k$. So, if there are more than

one variables the condition is this every partial derivative should be independent of the parameters

If this happened that out of there, suppose two partial derivatives are independent of parameters and the remaining one is dependent on the parameter then even the model will be termed as non-linear, ok. So, let me try to take here some examples and try to explain you this concept. So, you see this is our old friend simple linear regression model. So, expected value of y is now written as $\beta_1 + \beta_2$ X. So, you can see here that partial derivative of expected value of y with respect to β_1 is 0.

This is so this is not dependent on say any β and partial derivative of with respect of expected value of y with respect to β_2 is simply here X which is independent of β_1 and β_2. So, this model here is a linear model because it is linear in parameters.

(Refer Slide Time: 23:44)



Now, I try to take some more example which are little bit complicated suppose if I try to take the model expected value of y = β_1 $x^{\beta 2}$. So, if you try to find out here the partial derivative of E(y) with respect to β_1 and with respect to β_2 you will see that they are dependent on the parameter. So, this model is non-linear in parameters β_1, β_2.

But there is one option we can use some transformation and possibly we can linearize it. So, if you try to take here the log on both the sides, you can write down expected value of y here as say log of expected value of y and then β_1 will become log of $\beta_1 + \beta_2 \log X$.

And let me denote this log of expected value of y as say here y* and this log of β_1 as β_1* and log of X here x here x*. So, now, if you try to differentiate this function with respect to β_1 and β_2 then it will come out to be independent of the parameters. So, if you try to differentiate y* with respect to β_1* and β_2 then they will be independent of the parameter.

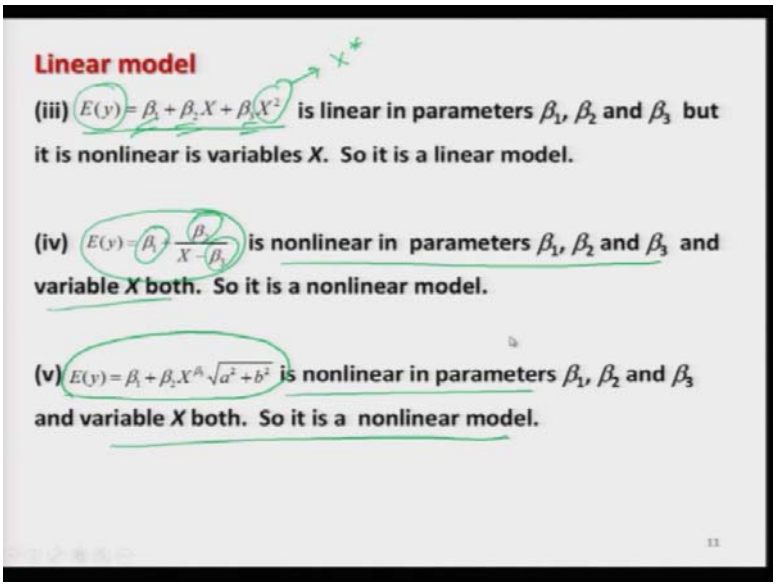
But here you have to be careful your original model is this one and your transform model is this one. So, be careful in stating the correct conclusion that which is that this model I mean the original model is non-linear whereas, the transform model $y^* = \beta_1^* + \beta_2 x^*$ is linear in parameters β_1* and β_2 not in β_1 and β_2, right.

So, these are now two different models in one case the parameters are β_1, β_2 and in other case the parameters are log of β_1 and β_2 which are indicators as β_1* and β_2.

So, sometime you will see that people try to when you are trying to deal with more complicated models where the mathematical forms is very complicated then you try to make some such transformation and you try to work with this transformed model y. Because whatever tools we are going to develop they are essentially for linear model.

So, once you are trying to deal with non-linear model you will have to develop the tool according to the need and requirement of the model, which is more difficult. So, people try to first attempt that if a non-linear model can be converted into a linear model. So, that all the tools whatever have been developed for the linear model and which are easily available in books in software they can be directly used in the analysis.

(Refer Slide Time: 27:06)



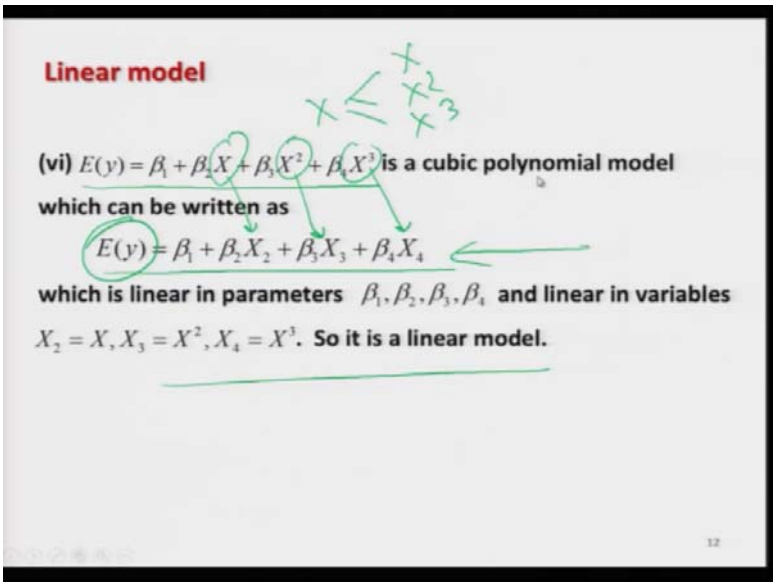
Now, let me take here one more model third one, expected value of y = $\beta_1 + \beta_2 X + \beta_3 X^2$ Now, if somebody looks it appears that this is a sort of second order parabolic model so, this is non-linear. Because its scatter diagram will not come out to be linear, but if you try to differentiate E(y) with respect to β_1 β_2 β_3 you obtain the partial derivatives then they will come out to be independent of the parameters.

Because here I can do one thing this X^2 can be written as say here say X*. So, this is again a linear model, but if you try to take this type of model in the 4th case $E(y) = \beta_1 + \beta_2 /(X - \beta_3)$ you can see that the partial derivatives of expected value of y with respect to β_1 , β_2 and β_3 they are not going to be independent of the parameters. So, this model is non-linear in parameters.

As well as if you want to make it more clear you can say that it is non-linear in variable X also, but that does not affect the definition of the linear linearity of a model. Similarly, if you try to take here one more example which I am taking here in the case 5th expected value of y is $\beta_1 + \beta_2 X^{\beta 3}$ square root of a square + b square they are some constant right some known value.

So, this again you can if you try to partially differentiate E(y) with respect to β_1 and β_1, β_2 and β three this will come out to be dependent on parameter. So, this is again a non-linear model ok.

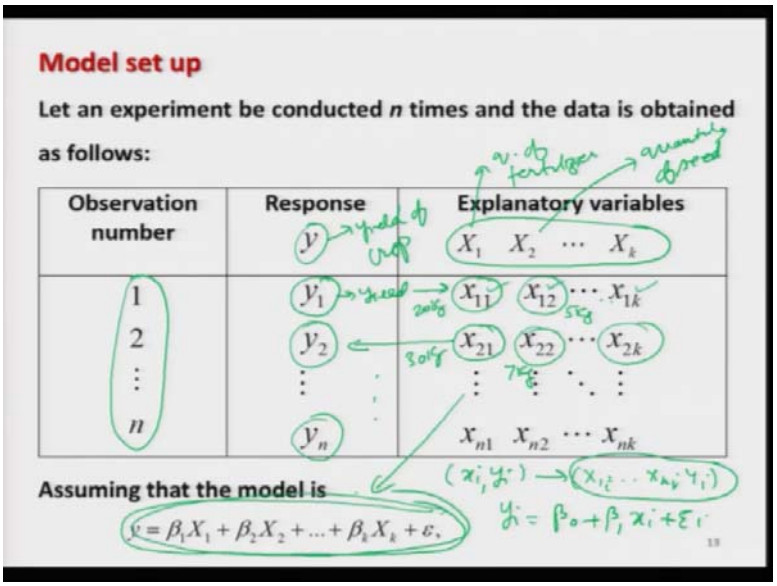
(Refer Slide Time: 28:58)



So and then the last example I try to take which is a polynomial model. Which is something like $E(y) = \beta_1 + \beta_2 X + \beta_3 X^2 + \beta_4 X^3$. So, you can see here the variable here is the same x. And now it has here it is appearing as different variable X^2 and X^3.

So, this type of model are called as cubic polynomial model which can be written here like this something expected value of $y = \beta_1$ and X can be represented as X_2 X^2 can be represented as X_3 and X^3 can be represented as X^4. So, now this model looks like a linear model. And if you want to verify it just take the partial derivative of expected value of y with respect to β_1 , β_2 , β_3 and β_4 and you will see that they will come out to be independent of the parameter hence this is also a linear model.

(Refer Slide Time: 30:06)



So, now you have learned that how to formally decide whether a model is linear or non-linear, ok. Next we try to understand the basic setup that how are you going to obtain the response and how the observations are going to be indicated in the model?

So, suppose you conduct the experiment and suppose you obtain a small n number of observations which I have denoted here by 1 2 up to n. And now you are considering here k independent variable k explanatory variable so these are my X_1 , X_2 ,…, X_k .

So, how are you going to conduct the experiment? Suppose this X_1 here is quantity of fertilizer and X_2 is suppose here quantity of say seeds and suppose y is your here yield of a crop. So, now, what you will do? You will take a plot you will put some quantity of seeds say maybe 5 kg and then you will try to put some fertilizer say 20 kg and similarly you will try to take other values of the variable right. X_1, X_2 ,…, X_k they can be different variable affecting the yield of the crop..

And then after sometime you will obtain the yield of the crop, which is denoted as here y_1 . And this yield is dependent on these values this yield has the value of this yield y_1 is obtained using the values of X_1 , X_2 ,…, X_k as x_{11}, x_{12},…, x_{1k}. Similarly, you will then you will try to take another choice for this X_1 , X_2 ,…, X_k for example, now you try to repeat the experiment and if you try to take say 30 kg of fertilizer you try to take say 7 kg of seeds. And similarly you try to choose another value of X_k and you conduct your

experiment and after sometime you obtain the value of yield of the crop which is here y_2 .

And similarly, you can repeat this experiment say small a number of times. So, now, you have this type of data. So, if you try to remember earlier I had taken the paired observation say (x i, y i) now this has been extended and now you have data on X_1 , X_2 ,…, X_k as well as here y and then you will have here the i th observation here like this which is going to be a tuples not a pure observation right.

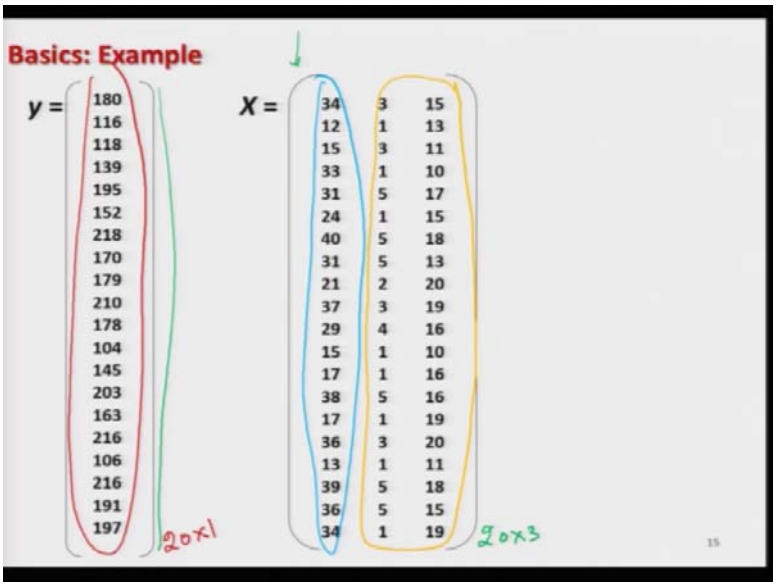
So, this one set of observation will consist will constitute one value of the observations. And now since, we are assuming the model to be here like this $y = \beta_1 X_1 + \beta_2 X_2 + \beta_k X_k + \varepsilon$ So, similar to the concept that we introduce in the case of simple linear regression model that every set of observation will satisfy the simple linear regression model $\beta_0 + \beta_1 x_i + \varepsilon_i$ this concept can be extended and these sets of observation will satisfy this relationship this one this multiple linear regression model right.

For example in this case if you try to take the same example here which we consider now, it is dependent on 3 variables. So, I can write down my model here say here for example, $\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$ and $+ \varepsilon$ right.

And what are these things this 34, 3, 15 these are the values of X_1 , X_2 , X_3 . So, in case if you try to make your life more simpler let me try to take the same model which you have considered just for a while just for explaining it I am trying to take the model without intercept term. So, that I can introduce the notation. But later on we will consider the intercept term in the model because that is obvious that if a student is not studying if he is not submitting or the student is not playing even then the student will get some marks ok.

So, now, if you try to see in this case these are the values of X_1 similarly these are the values of X_2 and these are the values of here X_3 and these are the values of here y, right now see how I try to arrange them.

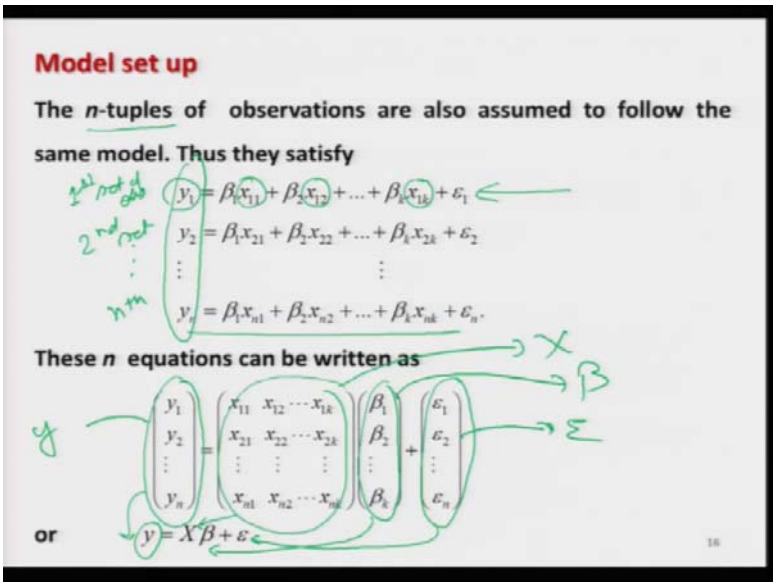
(Refer Slide Time: 35:04)



You can see here I have created here a vector and a matrix, all these observations on y you can see here you can see here it is 180, 116 and you can see here this is the same data set I will try to make it in red colour this whole data set is being copied here. But now it is in the form of a vector this is a vector of order 20 by 1. And similarly, if you try to see here the variable X_1 has these values which I am making here in blue colour these values have been entered here. And similarly the values of here see here X 2 and X_3 they are copied here.

So, now, you can see here what is happening that capital X is denoting a matrix of order 20 by 3. And now, here I can tell you actually if you want the intercept term to be here just try to insert here one column with all the values 1, but anyway I will try to show you at later stage. So, this is how we are going to collect the data and this is how we are going to indicate it in terms of our symbolic notations.

(Refer Slide Time: 36:36)



So, now we assume that the observations which are in the form of n tuples they are going to satisfy the same linear model. So, the for the 1st set of observation we have the multiple linear regression model in which the value of y is y_1 the value of x_1 is x 11 the value of x 2 is x 12 and the value of x_k is x 1k and they are going to satisfy the same linear regression model.

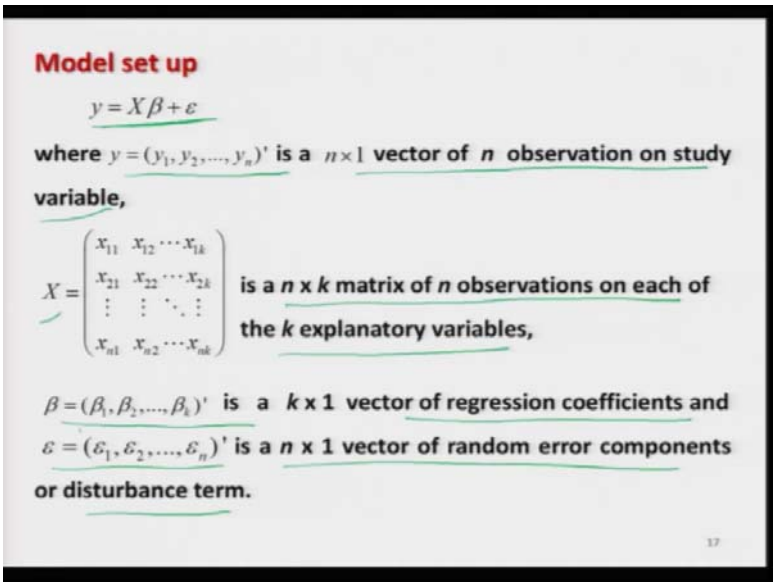
And similarly, the second set of observation up to here n th set of observation they are going to follow the same model. And now, I try to write down these equations in terms of vectors and matrices, you know means from your mathematics knowledge that I can write these linear equations in the form of vectors and matrices.

So, these observation y_1 , y_2,…, y_n they can be written here in the form of a vector. Similarly, all these observations on X_1 , X_2,…, X_k they are written in the form of this matrix all the regression coefficients they are written in the form of β_1, β_2,…, β_k and all the random error component they are given here as say ε_1, ε_2,…, ε_n.

So, in general what I can do that I can rewrite all this y_1 , y_2,…, y_n, I can denote them in a vector small y and all these observations on capital on this explanatory variable capital X_1 I have capital X 2 capital X_k the observations are denoted by the small letters small x all these observations are going to be indicated by this matrix X and all the regression parameters β_1, β_2,…, β_k which are inside this vector they are going to be denoted by β.

And whatever are your random errors that this vector is going to indicated by ε So, this here is y this here is X this here is β and this vector here is ε So, I can write down this model as y = Xβ + ε.

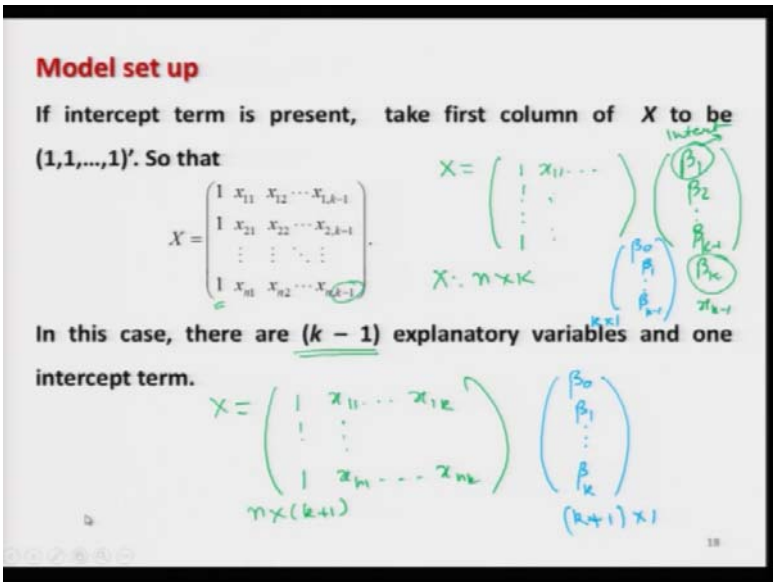
(Refer Slide Time: 38:55)



And you will see here this for y = X β + ε is a standard form in which we try to write down the multiple linear regression model. So, now I will say briefly explain you what is here y, X β and ε So, y here is a n x 1 vector of observations on the study variable..

Capital X is the n x k matrix of n observations on each of the k explanatory variable β is a k x 1 vector of regression coefficients and ε is a n x 1 vector of random error components or this is called as disturbance term. So, this is going to be our multiple linear regression model on which we are going to work.

(Refer Slide Time: 39:39)



So, now I try to show you that if you want to include an intercept term in your model then how you can do it? So, there are two options the first option is this that you try to introduce the first column of your matrix X as 1 or I can say whatever are your values here in the first column they will become 1 and all other values will come over there as such.

But here in this case you have to be little bit careful that how are you going to count the number of explanatory variables? Means if you are trying to say that X is going to your here n cross k matrix then the number of columns are k. Since now, first column here is 1. So, now, the number of explanatory variable will be here k minus 1, although the order of the X will remain as n cross k. So, in this case the number of explanatory variables will become k minus 1.

Sometime people try to simply use the earlier matrix up to k variable up to just like here this is the same metric that we have used earlier and they try to add here one more column. In this case the order of the matrix will become n x (k + 1).

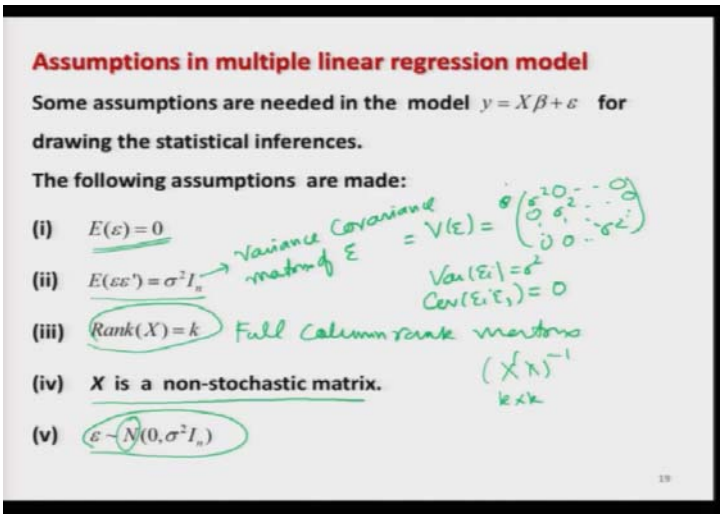
So, I am giving you here an advice that later on you will see at different places we will need the number of observation as well as we need the number of explanatory variable and sometime we will need the order of the matrix X. So, you have to be careful how are you going to do it.

And once you try to do it do it here then in this case your β_1 you had the regression coefficient vector $\beta_1, \beta_2, \ldots, \beta_k$ now, it will become here β_k minus 1, but it will remain as here say here β_k.

So, now, this β_k will be corresponding to the $x_{(k-1)}$ variable and this β_1 will become your intercept term right or better notation will be something like if I introduce the intercept term as β_0 then it will remain as β_0, β_1 up to here $\beta_{(k-1)}$ so, that this will remain as a k cross 1 vector.

And in this option if you want to have this one then your regression coefficient vector will become $\beta_0, \beta_1, \ldots, \beta_k$. So, now, this will become here k + 1 cross 1 vector. So, this is the point where you have to be extremely cautious because this is the thing which is going to help you in finding out the degrees of freedom so just be careful.

(Refer Slide Time: 42:30)



Now, I try to make certain basic assumptions in this multiple linear regression model which are needed to draw statistical inferences. So, we assume that expected value of ε is 0; that means, the mean of the ε in the population is 0 and the expected value of ε epsilon prime is $\sigma^2 I_n$. So, this is indicating the variance covariance matrix of ε.

So, that will be suppose we try to denote it by V of ε So, it is trying to say it is something like this that on the diagonal elements we have σ^2 and all other elements are 0. So, this is

indicating that the variance of ε_i is σ^2 and all those ε_i ε_j they are independent so the covariance between them is 0.

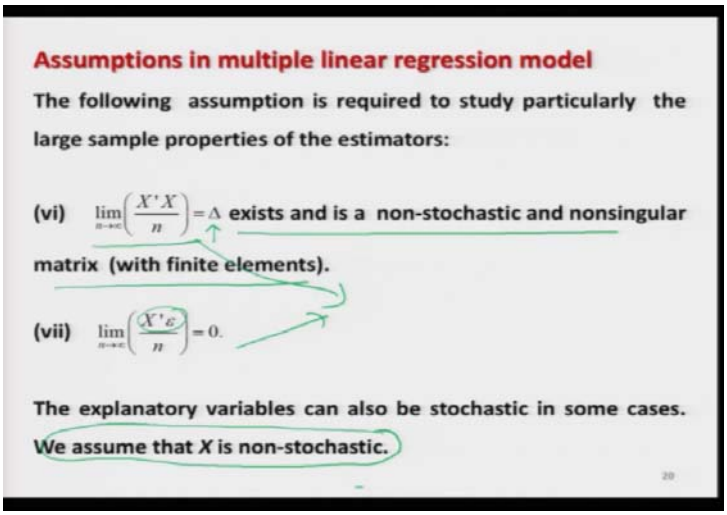
Third basic assumption we make is rank of the matrix X = k; that means, this is a full column rank matrix. Later on I will show you that why this assumption is needed because you can assume that if I have to invert a matrix X transpose X which is of order k by k then you need this matrix X to be of full column rank otherwise you cannot find the unique inverse of the matrix X'X.

Fourth assumption we make is X is a non stochastic matrix; that means, all the independent variable all the explanatory variable X_1 , X 2, …, X_k they are fixed. So, that is an extension of the assumption that we made in case of simple linear regression model also and you have seen that this assumption helps us in taking various types of expectation, but in practice it might be possible that sometime the X's are random variable they are stochastic. So, corresponding to those thing we have random regressors models also.

And finally, we assume that ε are following a multivariate normal distribution with mean vector 0 and covariance matrix $\sigma^2 I_n$. So, this is again an extension of the earlier assumption what we made in the case of simple linear regression model. And here I would like to repeat once again that the assumption of normality will be needed when you are trying to consider the confidence interval estimation and testing of hypothesis and also in the case of maximum likelihood estimation.

In case if you are using the ordinary least square estimation then the normality assumption is not needed as long as you are estimating the parameters, but when you, but when you are trying to construct the confidence intervals and the test of hypothesis then you will need this assumption ok.

(Refer Slide Time: 45:31)



Beside, those things we will try to make two more assumptions although it might not be possible for me to show you the outcome of these assumptions because we are not going to take all the topics in the regression analysis. So, and these two assumptions are required to study the large sample properties of the estimator.

So, as we had discussed earlier that whenever you are trying to find out the exact finite sample properties of estimator sometime the algebra becomes very complicated and then we try to approximate their properties, right. And in order to do it we try to use the last sample asymptotic approximation theory.

And for those things you need this assumption you simply need the assumption that as limit n is going to infinity X transpose X upon n is a matrix which is indicated by capital delta this matrix exists and it is a non stochastic and non singular matrix having some finite elements.

And we also assume that as limit n goes to infinity the matrix X' and ε vector. So, this quantity x transpose ε divided by n this goes to zero. So, now, we are simply going to assume in all the lectures what we are doing here that X will remain as a non stochastic matrix.

And this is very important for you to understand that we will always assume X to be non stochastic, because sometime if you go to the literature of regression analysis or

econometrics these two assumptions what I told you here they are also used when X's are random ok.

So, now, we come to the end of this lecture. So, I have tried my best to give you an overview of the multiple linear regression model. And I have tried my best to give you an idea that how the model will look like and how the observations will look like.

These assumptions what I explain you, you have to just understand them and you will see that when I progress further I will try to show you at different stages that these assumptions are going to help you in different ways. Well, I would like to address one thing more that, suppose these assumptions are not satisfied then what will happen? Then in that case different violations of these assumptions will lead to different type of problem..

For example, if your εs are not independent they are correlated that will lead to the problem of autocorrelation. If the variances are not the same for εs they are not like what we have assumed to be σ^2 they are σ_1 square σ_2 square σ n^2 if that is not satisfied then it will lead to the problem of heteroscedasticity.

We have assumed that rank of X = k; that means, all X_1 , X_2,…, X_k are linearly independent if that assumption is violated then it leads to the problem of multi collinearity. And if you assume that your y_1 , y_2,…, y_n they are not independent they are correlated possibly that can be a case of time series analysis.

And similarly, we have a list of problems which may come when these assumptions are not fulfilled and there are different types of solution to handle those problems and that is exactly what we try to do when we try to teach a full course on regression analysis and econometrics also. But, here in this course definitely we do not have that much of time to understand all those things and our objective is different we are trying to deal with the data sciences. So, my objective and aim is that I want to show you that how these things can be used in the framework of data sciences.

So, I hope you will excuse me for not explaining those things, but definitely those things are possible to learn books are there you can read them yourself you can attend the

lectures in your college universities etc. So, you try to learn them you practice you try to revise this concept and I will see you in the next lecture till then good bye.