Multivariate Procedures with R

Prof. Shalabh

Department of Mathematics and Statistics

IIT Kanpur

Week – 08

Lecture – 38

Multiple Linear Regression Analysis: Estimation of Parameters

Hello friend, welcome to the course Multivariate Procedure with R. You can recall that in the last lecture we had understood an overview of the regression analysis. And I had given you some basic fundamentals like when a model is called as a linear model, what are the different steps in regression analysis. And I tried my best to give you an overview that how this analysis is conducted. And as I mentioned that there are many more steps which we will not be covering in this topic, but surely I would recommend you that you try to follow a book also, but certainly I can promise you that I will try to cover those many topics in this lecture that will help you in understanding all the lectures in the regression analysis which have not been covered in the book. So, now what are we going to do in this lecture? You know whenever we are trying to think about a statistical modeling, what are we trying to do and what do we expect? Basically if you try to do the computation manually it will take a very long time.

So, definitely we are going to depend on the software, but many times what people do that they will simply try to feed the data there, they will try to input the data and they will put some click click click and then they will say okay the regression analysis has been done. But when they try to look at the software they cannot understand what are these values trying to say. And as we have understood now that software is our obedient servant, it will do exactly what we have programmed or what we have asked, but how to interpret it, how to understand it this is our job. So, now in this lecture I am going to cover the topics which are produced in the output of a software when we try to do the or multiple linear regression analysis.

If you understand that how these values have been obtained and what they are trying to interpret, then it will be easier for you to understand the outcome of the software and

based on that you can take a wise decision about the statistical modeling. Well I am not going to give you here the proof of all those results, but surely I will give you the mathematical expressions which are obtained after a new mathematical and statistical analysis. So, let us try to begin this lecture and we try to understand different types of basic definitions and basic concepts in the multiple linear regression analysis. Okay, so now in this lecture we are going to talk about the estimation of parameters. Well this I will try to show you that whenever you are talking about the statistical modeling that is equivalent to saying that you want to estimate the parameters.

So, now how to get it done and what are the different other aspects related to the estimation of parameters we are going to now understand. So, suppose an experiment is conducted a small a number of times and the data is obtained here like this right. Suppose we have here k independent variable or k explanatory variable x1, x2, … xk which are given certain values. Suppose x1 is given the value x11, x2 is given the value x12, xk is given the value x1k and based on that we try to observe the value of the response variable which in this case is indicated by here y1. So, y1 is corresponding to the observation number 1 which is obtained by giving suitable inputs to the variable x1, x2, … xk.

Observation number	Response	Explanatory variables
	y	$X_1 \quad X_2 \quad \cdots \quad X_k$
1	y_1	$x_{11} \quad x_{12} \quad \cdots \quad x_{1k}$
2	y_2	$x_{21} \quad x_{22} \quad \cdots \quad x_{2k}$
$\vdots$	$\vdots$	$\vdots \quad \vdots \quad \ddots \quad \vdots$
n	y_n	$x_{n1} \quad x_{n2} \quad \cdots \quad x_{nk}$

And now this experiment is repeated and we try to take another set of values of independent variables on x1, x2, xk as x21, x22, x2k and we try to obtain the value of here it is observation on the y or second observation on the response which is indicated by here say observation number 2. And similarly we try to repeat this experiment a small a number of time and we choose say different values of explanatory variables xn1, xn2, xnk and based on that we try to observe the value of yn which is our nth observation. So, now you can assume linear relationship between this y and x1, x2, xk and we have discussed in the last lecture that how are we going to represent a linear relationship between y and x1, x2, … xk using the parameters beta1, beta2, beta k which are called as regression coefficient and it is giving an element of random error that is epsilon. Now the general relationship between y and x1, x2, xk is now written here as say y is equal to beta1 x1 + beta2 x2 + beta k, xk + epsilon. So, now if you try to understand I will try to explain you the same thing through a same example that we introduced in the last lecture.

In the last lecture we had considered the marks of 20 students and they were indicated by y and they took 200 and 50 marks and then these marks were supposed to be dependent on x1, x2, x3 where x1 is indicating the number of hours officially in a week, number of assignments submitted in a month and number of hours of play in a week and this is how we had got this data. So, you can see here these are the y1, y2, yn this is n equal to here 20. So, this is my here y1, this is my here y2, this is y3 up to here, this is my here y20, this is my here x11, this is my here x12 and x15 is my here x13 right. Similarly, if you try to see here this is my here x21, this is here x22 up to here x23 and so on. So, this is how I am going to.

So, I am trying to say that now I have conducted the experiment 20 times where I have 3 explanatory variables x1, x2 and x3 and then I have given different sets of values to this x1, x2 and x3. For example, if I give choose the value x1 equal to 34, x2 equal to 3 and x3 equal to 15 then I get here the first observation y1 as 180. Similarly, and I try to take here the third set of values here x1 equal to 15, x2 equal to 3 and x3 equal to 11 then I get the third observation y3 which is 118. So, this is how I have conducted this experiment right. I assume that each set of this observation is also going to follow the same multiple linear regression model.

For example, if you try to see here this is our assumed relationship between y and x1, x2, xk. So, I believe that each set of observation like as here between y and x1, x2, xk they are going to follow the same relationship right. So, if I try to say here in very simple words suppose if I take here this x11, x12, x1k and here y1. So, they are going to follow the or they are going to satisfy this relationship y equal to beta1 x1 + beta2 x2 + beta k xk. So, these relationships I can write for small n number of observation here like this.

$$
\begin{aligned}
y_1 &= \beta_1 x_{11} + \beta_2 x_{12} + \ldots + \beta_k x_{1k} + \varepsilon_1 \\
y_2 &= \beta_1 x_{21} + \beta_2 x_{22} + \ldots + \beta_k x_{2k} + \varepsilon_2 \\
&\vdots \qquad\qquad\qquad\qquad \vdots \\
y_n &= \beta_1 x_{n1} + \beta_2 x_{n2} + \ldots + \beta_k x_{nk} + \varepsilon_n .
\end{aligned}
$$

For example, for the observation 1 we have here y1 is equal to beta11 x11 + beta12 x12 up to here beta k x1k + epsilon1. And similarly for the nth observation we have here yn is equal to beta1 xn1 + beta2 xn2 up to beta k xnk this is my here observation number. Now these n equations can be written in a matrix format. I believe that this much you have done in your how you can write the linear equations in the form of a vectors and matrices. For example, if I try to give you here a very simple example suppose if I write down here y is equal to 2 x1 + 3 x2 and if I try to write down here y is equal to 4 x1 + 5 x2 right.

So, I can write down there as say y1 y2 is equal to 2 3 4 5 and here x1 and x2 like this right. And we generally you have seen in mathematics we try to denoted by a very popular symbol y equal to Ax right where y is this vector of y1 y2 x is a vector of x1 x2 and A is a matrix of 2 by 2 matrix of the element 2 3 4 5. Similarly, I am trying to write down here these y1 y2 yn as here in the form of a vector which is indicated by here y right. Similarly, I try to combine here all these x1 x2 x1 k x21 x22 x2 k up to here xn1 xn2 xnk in this matrix here like this right. So, this is here n cross k matrix which is indicated by here capital X.

And then I try to combine all these beta1 beta2 … beta k like this here in the form of here a vector right k cross 1 vector which is here like this. This is indicated by here beta and all these epsilon1 epsilon2 epsilon n they can be combined in a n cross 1 vector like here this which is indicated by here epsilon right. So, you can see here all this is small n number of equations which are indicating the n tuples of observations can be represented in the form of a equation y equal to x beta + epsilon right. And if you try to understand from the earlier example what are you trying to do here, then for example, for the first set of observation here like this it is x11 x12 and here x13 and this is y here y1 equal to 180. So, I am trying to write down here y1 is equal to 180 is the equation model beta 0 + this 34 times beta1, then this 3 times and then this 15 times beta3 + epsilon1.

Yes, I have introduced here the concept of beta 0 because I will try to explain you in the next slide right. And similarly for the second set of observation I am saying that this is the second set of observation and I am saying that now this y is equal to is equal to 12 into beta1 1 + 1 into beta2 + 13 into beta3 + epsilon2 and beta 0 here is the intercept term. And similarly if you try to go for the last observation say y20. So, y20 is equal to 197 is equal to beta0 + 34 beta1 + 1 into beta2 + 19 into beta3 and + epsilon 20. So, now you can see here y has been expressed as a vector of 20 by 1 order and x has been expressed as a 20 by 3 matrix of observations on the explanatory variables x1, x2, x3 where each of the explanatory variable has 20 observations right.

So, now if you try to understand this model on which we are going to now work this is here y equal to x beta + epsilon where y is a y n cross 1 vector of the observation on the response variable y1, y2, yn and x here is a matrix of order and yk where n observations on each of the k explanatory variables are obtained and beta is a k cross n vector of the regression coefficient beta1, beta2, beta k and epsilon is a n cross 1 vector of the random error component epsilon1, epsilon2, epsilon n right. Now, in many situations you would like to have an intercept term in the model right. What do you mean by intercept term? For example, if I write to like you can calculate the model say 1 beta 0 + beta1 x +

epsilon and say another model is my here y star is equal to beta1 x star + epsilon star right. So, what is the difference between the two models say let me try to write down here beta1 star. So, if I assume that expected value of epsilon or expected value of epsilon star is 0, if I try to take expected value of y this is equal to beta 0 + beta1 x from here and if you try to see when x is equal to 0 then average value of y is equal to beta 0 right.

$$X = \begin{pmatrix} x_{11} & x_{12} \cdots x_{1k} \\ x_{21} & x_{22} \cdots x_{2k} \\ \vdots & \vdots \ddots \vdots \\ x_{n1} & x_{n2} \cdots x_{nk} \end{pmatrix},$$

$\beta = (\beta_1, \beta_2, ..., \beta_k)'$ **is a** ***k*** **x 1 vector of regression coefficients and**

$\varepsilon = (\varepsilon_1, \varepsilon_2, ..., \varepsilon_n)'$ **is a** ***n*** **x 1 vector of random error components**

But if I try to take here the expected value of y star is equal to beta1 star x star + expected value of epsilon star which is equal to 0. So, if I take here x star is equal to 0 then average value of y star is also coming out to be 0. So, now there are two types of situations in real life sometimes when you put all the values of independent variables to be 0 even then there will be some outcome for y. For example, if you try to take the example of a crop in case if you do not put any additional fertilizer additional irrigation etcetera in the field and if you just throw some stick in the field then after sometime you will get some crop in general conditions right. So, in that case what will happen that if all the independent variables takes value 0 that means no additional fertilizer no additional irrigation etcetera then even then you are getting some output.

So, in such a cases I would like to have a model which has got interceptor. But on the other hand if you try to measure the light of a bulb and the current. So, in case if you make the input current to be equal to 0 that means you simply switch off the light then the current becomes 0 and the light also become 0. So, in such a cases when independent variable current takes value 0 then the outcome y is 0. So, in such a condition I would like to have a model which is a not having the interceptor.

So, interceptor like this one y star is equal to beta1 star x star + epsilon like this one right. So, the role of interceptor comes in those situation where the output is not going to be 0 when the values of all the input variables are 0 right or in a way in case if you take all the input variables to be equal to 0 then if your output becomes 0 then you try to consider a model without interceptor. But in case if the output does not become 0 then you try to consider a model with interceptor right. So, now if you understand it mathematically your this x matrix here is like this where your first column is x11, x21, xn1. Suppose if I try to make this in the first column to be equal to 1 like as here right and then remaining is your here the values of independent variables.

So, obviously when I am trying to take the first column to be 1 and if I want to retain only the k numbers of explanatory variables in the model then instead of x1, x2, … xk, I will be having x1, x2, here x k-1 right. But anyway the moral of the story is that if you want to have an interceptor in the model just try to consider the first column in the x matrix to be 1 1 1 1 that is all right. So, in but in this case you have to be careful that there will be only k - 1 explanatory variables in the model. This I am emphasizing because later on you will see when we are trying to conduct the test of hypothesis then the degrees of freedom will be depending on the number of explanatory variables in the model and in that case you will have to understand whether there are k variables in the explanatory model or k - 1 explanatory variables in the model. So, surely if you try to say as you see here the first independent variable here let me call it here as x star this is taking here value 1 1 1 1 1, but it is also a explanatory variable right.

So, this is what I meant when I try to take a model with or without intercept term. So, some basic assumptions are needed in the regression modeling in order to estimate the parameters and in order to explore the further statistical properties of the estimators of the regression coefficient right and they are helpful when we try to draw the statistical inference. So, in the model y equal to x beta + epsilon we try to make here certain assumptions. So, first assumption here is expected value of that means on an average the random errors are going to be 0. So, whenever we are trying to observe some data then some of the data may have error in the positive direction some of the data may have error in the negative direction.

So, if you try to take the average of all the errors that will come out to be very close to 0 or say exactly 0 ideally. Then we try to assume that the covariance matrix of the random error components is a diagonal matrix and whose all the elements are sigma square. That means, I am trying to see here epsilon1 epsilon2 epsilon n they are iid, identically and independently distributed independently and identically distributed. This means what? This means the one is going to I am trying to make that each of this epsilon i has got variance sigma square for i goes from 1 to n and covariance between epsilon i and epsilon j is equal to 0 for all i not equal to j i goes from 1 to n and j goes from 1 to n. So, this is what I mean if you try to see here this is the covariance matrix like sigma square sigma square sigma square and all other elements on the off and off diagonal matrix are 0.

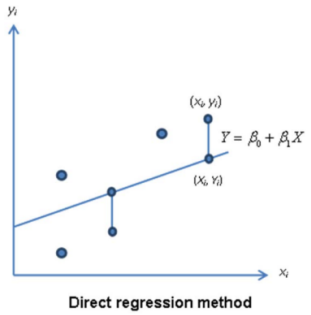
So, the fourth assumption I make here is that the rank of the x matrices is equal to k that is x is a full column rank matrix. What does this mean? This is trying to indicate that all x1, x2, … xk are linearly independent or in some way I am trying to say there does not exist any exact relationship among x1, x2, … xk right. And this condition will also be

needed when we will try to use of the matrix x transpose x right. So, because in that case it has to be a non-similar matrix. So, if x is a full column length matrix then this x transpose x will be a non-similar matrix, but anyway that I will show you later on.

The fourth assumption here is that x is a non-stochastic matrix that means x is non-random. So, well in many cases x can also be random, but anyway for our analysis we are assuming it to be non-random because it will help us when we try to take different type of expectations then the statistical calculation will become simpler and easy to understand. And the meaning of this statement is that if you try to conduct the again possibly you will get the similar value. The fifth assumption is that epsilons are following a multivariate normal with mean vector 0 and covariance matrix sigma x square in. Now, you understand what is the meaning of multivariate normal right and you know that it will be required when we try to conduct the test of hypothesis and confidence interval estimation.

There are two more assumptions which are required to study the last sample properties of the estimator that we are going to find. So, one of them is like limit n tending to infinity x transpose x upon n is equal to delta and this exists and it is a non-stochastic and non-similar matrix. Remember one thing do not get confused with this delta distance right if you remember that in that test of hypothesis we had indicated delta square to be the difference right. But it is not like this right it is only a delta some matrix which has got the finite elements. And we also assume that that limit tending to n tending to infinity x transpose epsilon upon n is equal to 0 right.

So, these two assumptions are going to help us when we try to prove the consistency property of our estimator that we are going to find right. Now, we come to the aspect of estimation of parameters. The question comes here why it is needed. Let me try to give you here a very simple example to explain you. You all have studied the equation y equal to m x + c which is the equation of a straight line like this right.



Direct regression method

And if you try to recall it is like this actually this part here is c and this here is the slope

parameter which is indicated by here m. So, m is going to indicate the slope of the line and c is going to indicate the intercept term. If I try to divide the components of this equation into two parts x and y and another here is m and c. Now, I ask you question number 1. The question number 1 is that if you know x and y then do you have the complete information about the line.

For example, if I say x equal to 2 and y equal to 3 then my this will become here p is equal to 2m + c. My question number 2 is that if I know the value of m and c, if the values of m and c are known. Suppose for example, I say here m equal to 2 and c is equal to 3. So, my equation becomes here y is equal to twice of x + here 3. Now, in the question number 1 and question number 2 what do you understand.

In my opinion if x and y are known then we are not getting any more information from this one no other information about the line. Means I cannot know anything except the information that x takes value 2 and y takes value 3. But on the other hand if I try to take here the second example and if I know the value of m and c then my equation becomes here y equal to 2x + 3 and then I have here the complete information about the line. What does this mean? That I know what is the slope, what is the intercept m and for any value of x, I can find out the value of y. So, if you try to see here this m and c they are playing a very important role in finding the distance line right.

So, I can say that m and c are the parameters of this line. And now you have seen that if somehow you know the value of the parameters then you know the whole line. Now, question to you all is that can you extend this concept to the multiple linear regression model. You are trying to write down your model here y equal to say x1 beta1 + x2 beta2 + up to here xk beta k + epsilon.

Now you want to know this model. So, if somehow you know the value of beta1 beta2 beta k and yeah this obviously epsilon also follows a normal 0 sigma x square. So, sigma x square is also unknown. So, in case if you somehow know the values of beta1 beta2 beta k and sigma square do you believe that you have the complete information about the line? The answer is yes and that is what we mean by the statistical modelling. So, in very simple word I can say that when you say that I want to know this model this is equivalent to saying that you want to know the values of the parameters of the model. Different models may have different parameters, but somehow if you know the values of those parameters then the whole model is known to you and after that you can use it in different applications.

So, the next question comes here how to know these parameters? So, definitely more simple option is that you try to collect the data on your variables and then try to find out the value of these parameters. And now you have understood some time back I had given you a lecture on estimation of parameters. So, we had talked about the maximum likelihood estimation and, but definitely I had told you that there are different types of estimation which are available and among them yeah maximum likelihood estimation is quite popular. So, here in this now lecture I am going to use here two methods. One is the principle of least square and another is the method of maximum likelihood estimation to estimate my parameters beta1, beta2, beta k and sigma square.

And when we try to use the principle of least square then there are different types of methods which are available in the based on the principle which is behind the least square. So, among them we are going to use here the ordinary or direct least square estimation which is based on minimizing the random errors in the vertical direction and I am going to use here the method of maximum likelihood. Well, I am not going to give you here that much mathematical detail, but I will give you only the some broad steps and then I would try to show you how you can find them out right. So, first I try to consider here the direct least square estimation or the ordinary least square estimation. So, let me try to give you this idea in the case of a simple linear regression model where I have only one variable x and my simple linear regression model is y equal to beta 0 + beta1 x + epsilon.

So, if you try to see here when you try to have say small n sets of data on say here x i and here y i they can be plotted here like this right. So, definitely ideally you expect that all the points should be lying exactly on the line, but in practice it is very difficult because of the involvement of the random error right. So, if you now try to find out here a line in a set of this which is passing through with the maximum number of points and now you can see here there is some error which is happening in each and every observation. For example, if this point this is observed value, but this point is expected to lie somewhere here on this line right. So, if I try to suppose this is my here x i y i which we are observing.

So, this is suppose here capital Xi and capital Yi right. So, all these points which are lying on this line they can be modeled by y is equal to beta0 + beta1 X right and you can see that this error is due to the random variation and different aspect yeah. Later on I will try to interpret this error in a different way I can tell you here the name, but later on I will try to explain you what is this, but similarly if you try to assume some points are below the line also. So, these are also error and different observations have means error in either

upward direction from the line or in the downward direction from the line right. So, what we try to do we try to take the sum of squares of this random variation like this one, this one etcetera and then we try to minimize them and you have understood earlier when we did the graphic that that we had used the command plot in our software.

So, that you can make such a scatter diagram and you can also find out to draw there a line which is passing through with the maximum number of points that we already have done right. So, this is how you can do it. So, the direct least square estimation involves the reminding the sum of squares of difference between the observation and the line in the scatter diagram right. So, if you try to that once you try to suppose I can say here now if you try to look here in this picture suppose I can see here this is my epsilon1, this is my epsilon2, this is my epsilon3, this is my epsilon 4, this is my epsilon 5 because this is the difference between the observed and the values which will be obtained from the model. So, if I try to write down here epsilon1 square + epsilon2 square up to here epsilon 5 square and then this will be equal to simply here y i - beta 0 - beta1 x i whole square and then I try to minimize this sum of squares and I try to find out the value of beta 0 and beta1 right.

So, this is actually done in the equal of least square. We try to minimize here in this case by using the principle of maxima and minima that we try to differentiate partially differentiate this equation with respect to beta 0 and beta1 put them equal to 0 and find and solve them and they are going to give us the value of beta0 and beta1 which are going to minimize the sum of squares right. So, the same thing is done in the case of multiple linear regression model here like this y equal to x beta + epsilon and we try to minimize summation epsilon i square which i goes from 1 to n which can be written here as epsilon transpose n epsilon this is equal to y - x beta transpose y - x beta and we try to minimize this quantity with respect to here beta substitute equal to 0 or null vector and then we try to solve it. So, once you try to solve it then you get here the value of here beta hat which is the value of the beta which is based on the observations on x1, x2, … xk and y. So, this value I am going to indicate by here b and this is here like this. So, the ordinary least square estimator of beta is obtained here like this x transpose x whole inverse x transpose y right.

I am not giving you here the that you can find in all the books on the regression analysis which is very common. Let me try to see here at the structure of this one here you are trying to find out x transpose x whole inverse. So, in order to do it to find it out you need the condition that x transpose x, x is a non-singular matrix and for that the rank of x has to be a full column rank matrix k that was the reason that I had assumed here this assumption. So, this rank of x equal to k is going to ensure that this x transpose x whole

inverse into x transpose y can be computed and this is called the ordinary least square estimator of beta and if you try to see here this is a k cross 1 vector. And similarly if you try to find out the estimator of sigma square then this is obtained here like this sigma square hat which is equal to 1 upon n - k y - x b transpose y - x b.

I am not going to give you all the details, but I will just give you here the guidelines. I had given you the idea that right. So, if you try to see here I am trying to consider here the summation epsilon i square and suppose the value of beta which are going to minimize this sum of a square are indicated by b1, b2, … bk. So, this sum of squared deviation can be written here say y - x beta transpose y - x beta for a given value of y and x and yeah means mathematically the minimum will always exist because this function s beta is a real valued convex and differentiable function. If you come to a few here I am just giving you the step here as beta can be expanded like this then we try to differentiate as beta with respect to beta which is here like this and then if you try to see the second order partial derivative is at least non-negative definite at the value of beta which we are going to get from this equation is really going to minimize the function s beta.

So, try to obtain here the normal equation by partially differentiating the s beta with beta then we get here these and if and from here we try to obtain here b is equal to x transpose x whole inverse x transpose y which is a ordinary least square estimator of beta. So, this is how you can obtain the estimator. Now, you will see that once you know this value. So, this b here is a function of here only x and y and x and y are known and hence this b is known b is also known to us right. So now you can see here in the model y equal to x beta + epsilon this beta was a known value.

Now, we have obtained the value of beta on the basis of given values of x and y which are indicated by here b. So, beta had equal to here b and now since the parameter beta1 beta2 beta k are known sigma square can also be estimated this function this estimator. So, now I can say that we have obtained the multiple linear regression model. So, now I try to give you some smaller definition which are helpful when we are trying to do the regression modelling. And now onwards we are going to simply assume that x is a full column rank matrix unless until I when x is not the full column rank matrix then it gives rise to the problem of multi collinearity, but I am not going to discuss it here right.

So, now you have obtained the value of beta as beta hat is equal to b. Now your model was equal to here x beta + epsilon. So, now your fitted model is obtained just by replacing the value of beta in the model which is here y equal to x b right. So, this fitted line or the fitted model or fitted linear regression model is obtained by y is equal to b into

x. Now if you try to give me some particular values of here x right then you can obtain the right. So, if I say here suppose x is something then you can obtain here the value of y and suppose that way I try to denote here say x into b.

So, the fitted value of for a given value of x are obtained by here y hat is equal to x b. And now if you try to substitute here the value of say here b which is x transpose x whole inverse x transpose y then here we have got a specific matrix this one x into x transpose x whole inverse into x transpose this is indicated by here h. And so this expression becomes here h y and this matrix h is termed as hat matrix. The hat matrix has a has an important role when we try to analyze the statistical properties of different estimators in the multiple linear regression analysis. For example, this hat matrix is a symmetric this is idempotent that if you try to multiply h and h the outcome will only be h and if you try to find out the trace of h.

So, the trace of h here is if you try to remember the rule trace of a b is equal to trace of b a. So, we try to this as a and this here as here b and I try to bring this x transpose x on this side. So, this becomes a trace of x transpose x into x transpose x whole inverse which is trace of identity matrix of order k which is here k. So, you will see that when you try to understand the the topics in the regression analysis then you will need these properties.

Ok, now I give you one more concept which is about the videos. What is the difference between the observed and fitted values of steady variables? And if you try to recall this picture which I had just shown you earlier I had taken this difference here as say epsilon. But now what I am trying to say that you try to obtain the value of y from say here y the fitted model y equal to y hat is equal to suppose x b and this is here the value which you are trying to observe say here y. So, if you try to see this difference which I have indicated in red colour this looks very similar to what we had interpreted as random error in the earlier picture. So, I am not saying at all that the random variable can be estimated. Please do not misunderstand or misquote me that I have told that we are going to estimate the random variable, but if you try to see here both this picture look like same.

So, what we try to see here that the difference between the y and the fitted value of y may give us some idea about the random errors and you will see that many diagnostic tests are based on this concept. So, if you try to take this difference and like we cut this y - y hat. So, y hat is equal to x b and we also have seen that y hat is equal to h y. So, if you try to see here this is here i - h into y. So, I try to indicate this i - h as here h bar right and so this E which is a residual vector can be indicated by h bar y.

Well all these things either this is b fitted value residual vector they will be obtained in the R software. So, it is important for you to first understand what they are trying to explain you when you look at the software outcome and then this h bar also has some nice properties that h bar is symmetric like h, h bar is also an idempotent matrix like as here h right and trace of h bar is equal to trace of i n - trace of h. So, this is here trace of identity matrix of order n. So, this is here n and trace of h we already have obtained by k. So, these statistical properties I am explaining you because they will help you later on when you try to understand the further topics in the regression analysis ok.

So, this was about the already used square estimation. Now, I come to maximum likelihood estimation right. So, when we try to consider here the model y equal to x beta + epsilon. Now, we are additionally going to assume that the errors are normally distributed and they are identically and independently distributed. Ah Remember one thing what is the difference between ordinary least square estimation and maximum likelihood estimation. In the ordinary least square estimation for estimating beta or sigma square we do not need any assumption on the distribution of the random error for example, normal or anything else.

But when we are trying to consider the maximum likelihood estimation then we have to assume some distribution of the random errors because maximum likelihood estimation depends on the likelihood function and likelihood function is obtained as a joint probability function of the random variables. So, that is why it is important for us to now assume that epsilon is following a multivariate normal with between vector 0 and covariance matrix sigma square i. Well, this assumption was made as one of the assumption of the multiple linear regression model, but we are going to use it here. And after that we are going to use when we are trying to conduct the test of hypothesis and confidence interval estimation. In that case yes in case if you are trying to use the least square estimator then you have to make a distinction that ok.

You have obtained the least square estimation without assuming any distribution for the random errors, but when you will be conducting the confidence interval estimation and test of hypothesis then you will be requiring that assumption. So, this is the basic difference which you always have to keep in mind. So, we know that the normal density functions form is like this. So, the normal PDF of epsilon i is obtained here by 1 upon sigma root 2 pi exponential of - 1 upon 2 sigma square into epsilon i square for i goes from 1 to n and all of them are independent. So, I can find out the likelihood function as the product of this individual PDFs which is obtained here like this.

And then summation epsilon i square I can write down here as epsilon transpose epsilon which is written here as say y - x beta transpose into y - x beta. Now, if you try to take here the log and if you try to partially differentiate it with respect to beta vector and a scalar sigma square you get here these two normal equation. If you try to recall we have done a similar exercise when we had obtained the maximum likelihood estimation in case of univariate as well as multivariate normal. Here we found the likelihood estimators for mean vector mu and covariance matrix sigma in case of multivariate normal distribution and we had obtained the maximum likelihood estimator of mu and sigma square in case of univariate.

But here this sigma is like here sigma square i. So, that is why if I try to estimate only sigma square that will give me the complete estimator of the sigma hat also. So, that is why I am trying to differentiate it with sigma square and if you try to solve them you will get here the mean estimator. The maximum likelihood estimator of beta is indicated by here beta tilde which is here x transpose x whole inverse x transpose y which is same as OLS Ordinary Least Square Estimator or the estimator that we obtained by direct regression. And but the maximum likelihood estimator of sigma square is indicated by here sigma tilde square it is here like this 1 upon n into y - x beta tilde transpose into y - x beta tilde, but you can see here the difference here is with the divisor. Say the factor is 1 upon n where the divisor is n and in the case of a variance integral estimator of sigma square you can see here it was 1 upon n - x square.

So, that is the basic difference between this ordinary least square estimator and maximum likelihood estimator. The ordinary least square estimator of sigma square is an unbiased estimator whereas the MLE of sigma square is a biased estimator of that is the mean difference. So, OLSE and MLE of beta they are the same, but OLSE and MLE of sigma square are different that is what you have to keep in mind. And you can I am just trying to some steps so that you can verify it later on that if you try to find out the second order partial derivatives with respect to beta, sigma square and beta and sigma square here and then try to find out the Bordered-Hesian matrix and it will ensure that whatever value you have obtained for beta equal to beta and sigma square equal to sigma square in that we are going to maximize the likelihood function. How we try to consider some properties of the least square estimator? They are discussed for the information that OLSE of beta is an unbiased estimator of beta that is the expected value of B is equal to beta.

If you try to find out the covariance matrix of B this will come out to be here sigma square into x transpose x whole inverse. And now if you want to know this value then it

is dependent on sigma square and sigma square here is unknown. So, in practice what you can do if you want to know its value then you can estimate sigma square and then unbiased estimator of sigma square is obtained here. It is obtained as sigma hat square is equal to 1 upon n - k y - x B transpose into y - x B and this is indicated by here s square. And once you have obtained this thing this value of sigma square can be substituted here and we find the estimator of the covariance matrix here like this sigma square hat into x transpose x whole inverse.

And you will see that I would try to show you all these things in the outcome of the software. So, it is important for me to at least give you here this idea that what are these value and how do they look like. Similarly, if you try to find out the covariance matrix of y hat that is the fitted this will come out to be here sigma square into x transpose x whole inverse x transpose which is sigma square h. And once again here this sigma square is unknown in real data sets. So, you try to estimate it by sigma square hat and you try to replace it here and so the covariance matrix of y hat which is the covariance matrix of the fitted value values it is obtained here by replacing sigma square here and it is coming out here sigma square hat h.

And yeah one question comes here what about the values of this model that we got. So, just to convince you we have a Gauss Markov theorem which explains or which mathematically proves that the ordinary least square estimator is the best linear unbiased estimator of beta that it has got the minimum variance in the class of linear and unbiased estimator. Beside this thing I can also tell you I am not going to prove it that the ordinary least square estimator is a consistent estimator of beta and this is sigma square hat which is indicated by also a consistent estimator of this sigma square. So, now we come to an end to this lecture and we can see here that was a pretty lengthy lecture, but if I try to break it into two parts then I will have to include many many things. So, my advice that you have an option that you can watch this video in parts, but if I try to break it then it will be difficult for me to continue and it will also be difficult for you to maintain the continuity. So, that is why I decided that I should give all these basic concepts in one lecture because now after that I will be going to different aspects where I am going to use these concepts.

And then you will see that in that when I am trying to conduct this analysis in the R software that you will have only their numerical values, but there I will try to show you that this is the value which is called as for example fitted value or ordinary least square estimator etcetera. So, now it is your turn that if you want to clearly understand the lectures which I am going to give now in the future lectures if you want to understand the content it is very important for you that you understand these basic definitions. In case if

you cannot remember the expression no issue, but at least you must know what they are trying to indicate how they had been obtained because in the software they will simply try to give you some numerical values which are based on the data set of X and Y, but what are they going to explain you this is our job. And I would say that okay I have not given here many proofs, many mathematical derivations, but I tried my best to strike a balance so that those participants who have a good background in mathematics and statistics they can they should not feel lonely. But definitely for those who are interested only in the application they may or may not be interested in the mathematical details, but definitely as I said in the last lecture that this regression analysis has many more topics which are available in the books and which I am not covering here, but I will try to give you the sufficient background so that if you want to understand those topics you should not face any problem, but definitely you have to study you have to work hard.

So, you try to revise this concept and come prepared for the next lecture where I will try to take up some more topics. So, we try to practice it and I will see you in the next lecture till then goodbye.