Multivariate Procedures with R

Prof. Shalabh

Department of Mathematics and Statistics

IIT Kanpur

Week – 09

Lecture – 40

Test of Hypothesis and Confidence Interval Estimation on Individual Regression Coefficients

Hello friend, welcome to the course multivariate procedure with R. So, you can recall that in the last two lectures we have discussed about the multiple linear regression model and we have learned that how we can obtain it on the basis of given set of data in the R software. We also understood that what is the correspondence between the theory what we have done in the second last lecture and the software output what we have done in the last lecture. And there are different types of interpretations to this regression coefficients which are helpful in getting a good statistical model. So, whenever you are trying to think about finding a statistical model using the multiple linear regression model, what is the first step? The first step is to identify the input variable, the independent variable or the explanatory variable which are going to affect the outcome. Now some of the independent variable, some of the explanatory variable may have very high impact on the values of response or the y and they are playing a very significant role in explaining the variation in the values of y.

But there may be some variable which may not be playing a very important role in the senses that their contribution in explaining the variability in y is very very less. For example, if I try to take the example which I considered in the last lecture the marks of the student. If some student says that okay that the student could not get the good marks because the price of the patrols were quite high. In first say reaction do you agree with this statement possibly not, but the argument is from the student side is okay because the price of the patrol were very high.

So, it was difficult for the student to come to the classes or the classroom every day and the distance between the classroom and the hostel may be just say this not more than 500

meters and because of which I could not come to classes. So my marks and the examination are low. Are you really going to accept these types of argument? I am sure I will not accept it, but the question comes how are you going to justify it on the basis of some scientific approach. So, when we are trying to consider such variables in the multiple linear regression model then those variables which are contributing more which are important they are considered in the model and those variable which are not important in the sense that they are not contributing significantly they are not considered, but the question is how to identify whether a particular regression or a particular input variable is really going to contribute significantly or not. So, this can be achieved by testing of hypothesis and confidence interval estimation and you will see that we will try to conduct the test of hypothesis on the individual regression coefficient on more than one regression coefficients and we will get different type of information.

And based on that we can do a variable selection also. And you have to keep in mind that suppose we start with suppose 8 independent variables in my model and suppose I conduct my regression analysis and suppose I come to know that okay 3 variables are not important. So, in the next step I have to remove them from my dataset and I have to refit the statistical model. So, this is how the process is a recursive process and you come to a finally come to a model which is good or say well fitted for the given dataset. Well, I would like to address here before I move forward that I am not really talking of the variable selection techniques like as forward selection, backward elimination or a combination of them here.

But my modest objective is that to show you how you can conduct that test of hypothesis on the regression coefficients, how you can construct the confidence interval estimation on the regression coefficient and what type of information you can obtain from there. So, I will try to explain you here both the things, the concept and their application in the R software also. And we already have discussed about what is test of hypothesis and how do we conduct the test of hypothesis in the R software. So, similar type of outcomes will also be here. The only thing is this, yes, now instead of here so called here mu, your population mean, I will have here beta0, beta1, beta2, etc.

And there is one more parameter, sigma square. So earlier when we conducted that test of hypothesis, I had not explained you that how are you going to conduct a test of hypothesis for the variant. So that I will be taking in the setup of multiple linear regression model, but that will be valid for any setup also. So this is how I want to produce, proceed in the next couple of lectures. So let us try to begin this lecture and try to understand how are you going to conduct the test of hypothesis, confidence interval

estimation and their implementation in the R software and how are you going to interpret them.

So let us begin our lecture. So now in this lecture, in the setup of multiple linear regression model, we are going to talk about the test of hypothesis and confidence interval estimation on individual regression coefficient. Why individual? Because sometime I will try to give you the test of hypothesis where you would like to tell the equality of more than two regression coefficient that will be analysis of variance in the setup of multiple linear regression model. So there are several important questions which can be answered through that test of hypothesis concerning the regression coefficient. For example, one important question is what is the overall adequacy of the model? Which specific explanatory variable seems to be important, etc.

And in order to answer such questions, we develop here the test of hypothesis in the setup of multiple linear regression model for the individual regression coefficients. So firstly let me try to introduce here the null hypothesis H0: beta j = 0 and we want to conduct test of hypothesis for the alternative hypothesis H1: beta j $\neq$ 0. Well, after having detailed information or knowledge about that test of hypothesis, you know that if you want to consider here any other type of alternative greater than, less than that you can also consider. So I would like to consider the test of hypothesis for individual regression coefficient and for all the regression coefficient including beta0, beta1, beta2, beta3, etc. But before that how are you going to take the conclusion? How are you going to interpret the outcome of the test of hypothesis? So if you understand the rule is like this.

If H0 is accepted, my H0 here is remember beta j = 0. If this is accepted, it implies that the explanatory variable Xj can be ignored, can be deleted from the model and it is not contributing significantly in explaining the variation in the values of Y in the model. Because if you try to see what is the meaning of beta j? Beta j is the partial derivative of expected value of Y with respect to Hj. This is here beta j. So this is trying to indicate the average change in the value of Y when there is a unit change in the value of Xj, right.

That is the rate of change in the average value of Y with respect to Xj. So you are trying to say here that beta j = 0 is accepted and beta j is the value of the parameter in the population. So in the population beta j is 0 that means the rate of change in the average value of Y with respect to Xj is very small and ideally 0. That means Xj is not contributing in the model, right. So in this case suppose if you start or you begin with the model Y = X1 beta + X2 beta2 + X3 beta3 + epsilon and suppose you try to conduct the hypothesis H0: beta1 = 0, H0: beta2 = 0 and H0: beta3 = 0.

And suppose H0 beta2 = 0 is accepted, right. So now the conclusion is going to be like this. Now beta2 = 0 in the population and so beta2 is not significant and we need to revise our model this one as X1 beta1 + X2 into 0 that is beta2 = 0 + X3 beta3 + epsilon and my model will become only here X1 beta1 + X3 beta3 + epsilon. Now when you conducted the regression analysis for this model with X1, X2, X3 you had suppose beta1 hat, beta2 hat, beta3 hat, right. Based on that you have to conduct that test of hypothesis.

But now this model is changed and your new model now here is like this Y = X1 beta1 + X3 beta3. Now if you try to obtain here the regression estimators there will be something like beta1 star hat, beta say this here beta hat, beta3 star hat and so on and these are going to be different from beta1 hat, beta2 hat, beta3 hat or at least beta1 hat and beta3 hat, right. So this is how we try to choose the important variables in the model using the concept of Test of hypothesis, right. So now let me try to give you some fundamentals about the Test of hypothesis about the individual regression coefficients beta j's. So as you can recall that when we conducted the Test of hypothesis in the case of univariate normal population where we assume that like X1, X2, Xn be a random sample from a normal population normal mu sigma square and we had conducted that Test of hypothesis for H0: mu = 0 using the t statistics when sigma square is unknown, right.

That same concept is being used here also. If you try to recall your statistics was H0 if I see here H0 is mu = mu0 then now your statistics become here X bar - mu0 divided by S square upon here n, right. So if you try to see here now I am trying to replace this X bar beta hat, mu0 is my here 0. So X bar become beta j hat, mu0 become 0 and now I have here beta hat j - 0 and divided by so what is here the square root of SU square by n if you remember that was a standard error of this was standard error of X bar. So now I can replace it by here standard error of beta hat j.

So this will also be a t statistics and that will be following a t distribution with certain degrees of freedom. So how to obtain it? This is the job which we try to learn in our undergraduate program and postgraduate program in statistics but here I am going to give you here the final conclusion. So if you want to test here this hypothesis H0: beta j = 0. So suppose beta j has been estimated by ordinary least square estimator bj, yeah that can be MLE also because MLE is also the same as beta j tilde or beta j tilde like this, whatever you want to call. So, OLS and MLE of beta are the same.

So you try to estimate it by here bj - 0 and divided by standard error of bj, right. How to obtain the standard error of bj? If you try to recall we had done that the covariance matrix

of beta hat is sigma square X transpose X whole inverse and because sigma square is unknown, so sigma square is estimated on the basis of given sample of beta by here say sigma square hat and then we try to obtain the estimate of covariance matrix V hat sigma square hat into X transpose X whole inverse. So now if you try to say here that this sigma square hat into here some matrix here, C matrix where your C is say X transpose X whole inverse. It will be something like here C11, C22, Ckk in the diagonal and say Cij and Cji in the off diagonal elements. So I can, so we know that in the covariance matrix the jth diagonal element gives the variance of the jth regression coefficient.

So the standard error of bj can be obtained as the, from the jth diagonal element of this C with sigma square hat and by taking the positive square root. So the standard error of ordinary least square estimated bj is obtained here as the square root of sigma square hat Cjj where Cjj is the jth diagonal elements of, element of X transpose X whole inverse which is corresponding to bj. Well, you are not going to compute this thing. The R software will give you this value but it is important for you to understand that what R is going to tell you how it has been obtained. So that is why it is important for us to understand for this basic concept.

Now how are you going to take the decision? The decision rule is if you are going with the software then reject H0 against H1 at alpha level of significance if p value is less than alpha. That is our old classical rule. And if you are trying to take the help of the T tables then you have to reject H0 at alpha level of significance whenever absolute value of T is greater than the tabulated value from the T table T alpha by t (n - k – 1). Why this here n - k - 1? Because this T statistics here, this follows T distribution with n - k - 1 degrees of freedom under H0.

$$t = \frac{b_j}{se(b_j)} \sim t(n-k-1) \text{ under } H_0 \quad , \quad se(b_j) = \sqrt{\hat{\sigma}^2 C_{jj}}$$

That is when H0 is true. So this test of what we have done here H0: beta j = 0 is only a partial or marginal test because bj depends not only on xj but on all other variables also. Because if you try to see means b was X transpose X whole inverse X transpose Y and then b1, b2, bk they are not independent. So the value of bj should not be misunderstood that it depends only on the jth set of values but it is depending on all other values also. So this test gives us the information about the contribution of xj given the other explanatory variable in the model. So mainly the rule is that if that we are interested in that test of hypothesis H0 beta j = 0, if this hypothesis is accepted then we remove the corresponding variable from the model and we refit the model and we try to obtain other things.

Now let me try to take here the same example that I took earlier and I will try to conduct the test of hypothesis first on the same set of data. So we have the same data set where we had obtained the marks of 20 students and which are here given here like this and these marks are obtained on the marks of students out of 250 and we believe that these marks are dependent on 3 variables say x1 which is the number of hours in a week the students studies and then here and then the second thing here is that it is trying to say here x2 which is the number of assignments submitted in a month and x3 is the third variable which is the number of hours of play in a week time. So it is like that therefore, the student number 1 the student 1 has studied for 34 hours in a week the student has submitted 3 assignments in a month and the student has paid for 15 hours in a week and the student has got handed 30 marks and similarly all the we have the data for the 20 students. So now in order to get the information about the test of hypothesis and another aspects we try to use the command summary. You can recall that summary command has been used in different context when you try to use this command this summary with a data vector then you get the information about the means like minimum, maximum, first quartile, second quartile etc.

Similarly if you try to use this summary command with the alum command that is the linear model command which is used to find out the fitted linear regression model then we also get the results about that test of hypothesis along with some other results which I will try to take up now one by one. So your command to find out the linear model was lm then formula and then data etc. Now we are simply going to find out the summary of this command so summary of this expression and this will give us some output and we have to understand how are we going to understand it how are we going to interpret it. So I try to take here this model yi = beta0 + beta1 xi 1 + beta2 xi 2 + beta3 xi 3 + epsilon i and I have just stored the data on these three variables say x1, x2, x3 and y in the data vector here x1, x2, x3 and here y. Now I already had demonstrated the outcome of the command lm inside the parenthesis yi ~ x1 + x2 + x3 in the last lecture where we had obtained the linear regression model only.

Now if you I am using the summary() command, summary all in lower case alphabets with this lm() command and you will get here this type of output. You can see here this is here residuals and about residuals you are getting here the output what we had obtained with the summary command over a data vector. The minimum value of the residuals, maximum value of the residuals, first quartile, median that is the second quartile and then the third quartile of the values of residuals. If you recall that in the last lecture we had taken the same example where we had got 20 values of residuals. So this outcome here is

trying to find out the minimum, maximum, first, second and third quartiles of the 20 values of residuals.

After this if you try to see here briefly I will tell you then I will move forward that try to observe this here, this block. I will try to take one at a time but you have to understand what it is trying to tell you. First here is intercept term, then here x1, then x2, then x3. So intercept term is something like whose associated coefficient was beta0 and x1 had a regression coefficient beta1, x2 has a regression coefficient beta2 and x3 has a regression coefficient say here beta3. And then we have here first column which is here estimate.

So this is the value of here b0, then the value here is b1, then the third value here is b2 and the fourth value here is of the b3 that we already obtained in the last lecture. And based on that I can say that my model here is 8.76945 x1 + 1.99675 times x2 + 3.91840 x3 + 6.10603 x4 and so on. After this you have here these are the standard errors of b0, b1, b2, b3, then we have the t value, t value are the t statistics corresponding to H0: beta j = 0 and then we have here p values. Then we have here some here significance code, then here residual standard error on 16 degrees of freedom, multiple r squared, adjusted r squared, f statistics, etcetera and p value. So this part we are going to discuss later. But today we are going to talk about this part here.

So let us try to take it one by one. So this is the command. But I think before I move further let me try to show you this command so that you are confident that whatever I am telling you that is really going to happen and then I can concentrate on my outcome. So this is I have entered my data and then I am just trying to execute here summary command here like this. You can see here this is the same screenshot which I have shown you here. So now let me come back to my slide and try to take it here one by one how are you going to understand it.

So you can see here we have taken here 20 observations. So n is here 20. And what about the number of independent variable, number of explanatory variable? Are they only x1, x2, x3? No. You also have taken this intercept term. So as I explained you in the beginning when I introduced the multiple linear regression model that if you want to have the intercept term in your model then you can take the first column of the X matrix to be 1, 1, 1, 1, 1, so on and so but this will also be another explanatory variable which takes always value the 1.

So you have here one explanatory variable whose value is 1 and there are remaining

three explanatory variables x1, x2, x3. So this is how we are going to understand it. So I will take here k = here 4 and there they are corresponding to beta0, beta1, beta2 and beta3. So this is the, that is why we have here intercept term for this beta0, x1, x2 and x3 for here beta1, beta2, beta3 and this is the part in your, this screenshot which I am going to consider here.

You can see here this part. So now let me try to take here this one by one. So now if I try to take here this hypothesis S0 beta j = 0 versus H1 beta j is ≠ 0. So now my j will take value here 0, 1, 2, 3. And the t statistics will be bj upon standard error of bj and which follows a t distribution with n - k - 1 degrees of freedom under H0 and standard error of bj is obtained here that the positive square root of sigma square hat Cjj where Cjj is the j diagonal element of X transpose X whole inverse. So now in this output, let us try to concentrate or try to understand one by one that what are they trying to indicate.

So firstly let me try to take here the first row. This is about intercept term you can see here. So if you try to see here what is this value? This is the value of b0. And what is here? This is standard error. This is the standard error of b0. What is here this value, next value? This is the t value corresponding to H0 beta0 = 0 and it is here its t value.

And then here in the fourth column this is the p value corresponding to this H0: beta0 = 0. So if you try to see here I have explained here my null hypothesis is H0 beta0 = 0 and my alternative here is H1: beta0 ≠ 0. Now I try to create here a t statistics which was the value of b0 divided by standard error of b0. So this you can see here this is here the value of b0 and this is here the value of standard error. And if I try to calculate it manually this will come out to be 8.509 and this is given here you can see here, right. And it has got a t distribution with 20 - 4 - 1 which is t15, t distribution with 15 degrees of freedom under H0. And now here the p value here is like this which is given here. You can see here like this. So now if you try to choose here alpha = 0.05 then p value is 2.47 into 10 power of -7 something like 0.0000 like this and here 247. So you can see that it is very close to 0. So p value is less than alpha. So I can conclude that reject H0: beta0 = 0 at alpha level of significance.

So that means beta0 is ≠ 0. And so that means the concerned explanatory variable is important and it is contributing. That means intercept term beta0 is important and we need to keep it in the model. And if you remember I can I had already explained you that if a student is studying from some online courses not going to the class, solving the questions in the assignment correctly but not submitting it, not going to play but going to watch a film or involve in some could extra curricular activities etcetera etcetera. Do you

think the students will get 0 marks? Certainly not. Similarly, in the yield of a crop if there is no additional fertilizer, there is no additional irrigation etcetera, do you think the crop is going to be 0? No.

Under the normal condition there will be some crop if you try to put some seed in the soil that is the natural fertility of the soil. So that is why this is indicating what is happening in the real life. So this is what we mean by this test of hypothesis. So now beta0 is going to remain in the model and we are going to consider the model with intercept term.

Now I try to take here the next row that is about x1. So that is the same thing. This is this one in the first column under the estimate this is the value of here b1, the second value here is the standard error of b1, then it is here the value of here t statistics corresponding to H0: beta1 = 0 and this is here the p value like this. So if you try to consider here that test of hypothesis, the null hypothesis will be H0: beta1 = 0, alternative will be beta1 is $\neq 0$ and the t statistics here this b1 value upon standard error of b1, this b1 value will be coming from here 1.99675, then standard error of b1 that will be coming from the standard error of this here b1 which is 0.03228 and if you try to calculate it, this will come out to be 61.850 which is mentioned here and this is again going to follow a t distribution with 15 degrees of freedom under H0 and its p value, this is given here 2 into 10 power of -16 which is very close to 0 and it is smaller than the value of the level of significance alpha which is considered here as a 0.05, 5 percent level of significance. So once again I can conclude that since p value is less than alpha, so reject H0: beta1 = 0 at alpha level of significance, hence x1 is an important variable.

So number of hours of study, they are going to affect the marks significantly that is our conclusion now. And similarly, the same exercise you can do here for the row corresponding to x2 because x2 is indicating about the b2, so my null hypothesis here is H0: beta2 = 0, alternative here is H1: beta2 $\neq 0$ and the concerned t statistics is here like this t is equal to b2 upon standard error of b2, this b2 value is obtained from here, the standard error is obtained from here and t statistics is obtained from here which has got the t distribution with 15 degrees of freedom and the p value here, it is here 7.9 into 10 power of - 14 very close to 0, this is obtained from here and which is less than alpha which is = 0.05. So I can say once again that H0 beta2 = 0 is rejected at alpha level of significance and hence x2 is an important variable, right. So if you try to recall what was your x2, you can see here, x2 was your number of assignments submitted per week and we agree from our experience that if some student is submitting the assignments in the class, the student will always get a good marks. Now similarly I come to the variable here

x3, the same exercise what I told you here that in this case my H0 is beta3 = 0, H1 is beta3 ≠ 0.

Similarly I can compute my t statistics as b3 upon standard error of b3, b3 is here, standard error is here and this is here that p value and which is again the t distribution with 15 degrees of freedom and the p value is here 2 into 10 power of -16 which is less than alpha = 0.05. Hence once again I can conclude that H0 beta3 = 0 is rejected at alpha level of significance and hence x3 is important variable. So x3 was your number of hours a student plays in a week.

So that means yeah playing is important. So you are trying to conclude that if a student is also playing, then it is going to enhance the marks, but yeah I do not mean to say that if the student only plays and does not study, the student will get always a good marks. I am not trying to say that thing. But it is an important variable which is contributing in determining the marks of the student. That is what I am trying to say here. So now after this I come to another aspect that is confidence interval estimation.

So you can recall that we already have discussed this concept in the case of normal distribution where we had constructed the confidence interval for mu under two cases when sigma square is known and when sigma square is unknown. So here also you can do, but now you can see that in this case we have assumed that sigma square is unknown and we are trying to estimate it from the given set of data. So if you try to recall what we have done that in order to find out the confidence interval, we had taken the statistics x bar - mu divided by s by root n and then we had put it between - t alpha by 2n -1 t alpha by 2n-1 and we have said that okay the probability of that this t statistics is lying between these two critical values on a t distribution here like in this. It is going to be 1 - alpha and then we had solved this equality and we had obtained the value of mu and the lower and upper limits of the confidence interval.

The same process I am going to do here for beta j. And you also know that there is a close relationship between the test of hypothesis outcome and confidence interval estimation. So the same thing I try to do here and I try to construct here the confidence interval for this beta0, beta1, beta2, beta3 and so or say in general beta1 say this beta individual regression coefficient. So here now once again I have to assume that epsilons are identically and independently distributed following the normal contribution with mean 0 and variance sigma square in the model y = x beta + epsilon and so I can say here that y will also be following a multivariate normal distribution with mean vector x beta and covariance matrix sigma square i. And since this ordinary least square estimator is a

linear function of y and y is normally distributed, so I can say that ordinary least square estimator that is B vector will also be following a multivariate normal distribution with mean vector beta and covariance matrix sigma square x transpose x whole inverse. So now using this thing I can say that the individual values suppose Bj will be following a univariate normal distribution with mean beta j and covariance sigma square Cjj where Cjj is the jth diagonal element of x transpose x whole inverse but I just explained you in the beginning.

So now you have this statistics tj = bj - beta j divided by square root of sigma square hat Cjj which is the standard error of bj and this is going to follow a t distribution with n - k degrees of freedom under H0 and yes we are going to estimate the value of sigma square s sigma square hat as sum of square due to residual I can write down as sum of squares due to residual. Okay, I have not introduced this term here but then I had introduced this thing in the case of analysis of variance if you remember in the case of testing the equality of more than two means population mean from different populations and I am going to explain you this concept once again when we try to conduct the analysis of variance but here I can say simply this sum of square due to residual is simply that you already had obtained the residual vector e whose elements were e1, e2, … en from there we try to obtain it and this expression comes out to be here like this Y transpose Y - B transpose X transpose Y and this sigma square hat is SS that is sum of square due to residual divided by n - k the degrees of freedom.

So this is the expression by which we will try to compute this value of sigma square here right but anyway software is going to do it so now if you try to see here this statistics which is following the t distribution this is lying between the left hand side and right hand side of the t distribution like as here t alpha by 2n - k right and the probability of certain event is 1 - alpha if you simply try to solve it you will get here that beta j is lying between this limit and this limit and so 101 - alpha percent confidence interval for beta j is here like as beta bj - t alpha by 2n - k standard error of bj and bj + t alpha by 2n - k into standard error of bj right. But anyway software will give us this information but it is important for you to understand what software is trying to give us. So in our software the command is confint to obtain the confidence interval of the regression coefficient in the when we have more than one variables in the model right.

So and there are different choice of the parameters which have to be mentioned inside the parenthesis but we are going to use here only couple of them like as object, parameters, parm and then level of significance is given as confidence level that is 1 – alpha. So if you want to have here alpha = 0.05 then you have to mention this level here as say 1 - alpha and this confint is based on the class here lm which is the used to find out the linear models. So this is how we are going to use it confint object parameter level. So

object will is something like we have to specify the model which we want to fit and base and whose parameters are estimated and their confidence interval has to be obtained right and then parm this will specify that which parameter are to be given confidence interval either a vector number or a vector of names right and the label is the confidence level as I explained you that is 1 - level of significance.

So now if I try to see here you this command if I try to show you here confint now the object is lm() inside the parenthesis y ~ X1 + X2 + X3 and level confidence level is 1 - 0.05 1 - alpha. So you will get here this type of outcome. Now it is very easy for you to understand it.

You can see here this is giving you here 2.5 percent 97.5 percent this means what here if you want to plot it then the area on the left hand side and the area on the right hand side of the distribution is alpha by 2 alpha by 2 this is here alpha by 2 this alpha by 2. So this is 2.5 percent and this is here from here to here this is 97.5 percent area right. So this area will be 95 percent this is what 2.5 and 97.5 are indicating. Now for the intercept term beta0 the lower confidence limit is 6.584563 that is the lower limit actually this column is giving the value of lower confidence limit and this column is giving the values of say upper confidence limits right.

So the first row is trying to say that the lower confidence limit of intercept term is 6.58 and the upper limit is 10.954344 right. And similarly the next row it is about X1 so this is about the confidence interval for beta1 which is 1.928308 which is the lower limit of the confidence interval and the upper limit of confidence interval is 2.065184 right. So if you try to see here this is the confidence interval for the intercept term whose value is coming out of here like this, this one and this one. And similarly if you try to see here this is the confidence interval for the beta1 and whose values are obtained here like this, this and this.

And similarly the remaining two values are for the X2 and X3 here so this is for here beta3, this is for here sorry here beta2 and beta3. So this is the lower confidence limit and this is the upper confidence limit right. So these are obtained as like this the confidence interval for beta2 is obtained here like this whose lower limit is obtained here like this and upper limit is obtained here like this and for beta3 the confidence interval is the same like this one whose lower limit is obtained here and upper limit is obtained here right. So you can see here it is not a very difficult thing and this is here the screenshot and let me try to show you these operations on the R software also. So you can see here that like this

and yeah means so I already have this values of X1, X2, X3 and here Y so I simply have to say here conf int and you can see here this is here like this.

So I can see here that for example if you try to look at the values of X1, I can see here that there are 95 percent chances that the value of beta1 in the population will be lying between 1.92 and 2.06. Similarly for this X 3 there are 95 percent chances that the value of beta3 will be lying between 5.95 and 6.25 right and if you try to change here this level of confidence actually if you try to make it here 0.99 it is very simple you can see here that this confidence interval are changed right. And now if you try to suppose if I try to take here only here this X1 and X2 suppose if I fit only two variables here you can see here this confidence interval is now different than this confidence interval. These values are actually changed right. So this is what I meant that if you drop a variable from the model and refit the model the value of the regression coefficient as well as the confidence interval are going to be changed right.

So now with this discussion on the test of hypothesis and confidence interval estimation in the regression coefficient we come to an end to this lecture. And you can see that it was not a very difficult thing. Why? You already have understood what is test of hypothesis, what is confidence interval estimation. So I did not have to explain you but I used those things in the context of linear regression model and now you have seen that they are going to give you different type of information about the data which data wants to speak but the data is unable to speak so you are understanding the symbolic language of data and you are trying to get the information although you have not seen the whole population but still you are getting an information that how the marks of the students are dependent on these three variables in the population not in the sample. Whatever conclusion you have taken in the test of hypothesis or confidence interval estimation they are going to be valid for a they are valid in the whole population not only in this sample and for that you have yeah there will be some uncertainty so you have used the concept of level of significance or the confidence level.

You are saying that okay there are 95% chances that the value of say beta1, beta2 etc will lie in these two intervals or between these lower and upper limits. Yes certainly we are not God. We are trying to approximate a very complicated process through some mathematical procedures so and we are unable to obtain the exact data also because if I say that if you ask any student that how many hours you have studied in a week or how many hours you have played in a week it is very rare that they will be able to give you the correct information. So this information is also approximate so its impact is going to be indicated in the why also through this random variation. So you can see that here this is a multi parameter setup this is a multivariate setup and we have obtained a multiple linear

regression model and then there are many more questions whether the model is good or bad how to judge it etc that we are going to take up in the next lecture.

So now in this lecture your job is to take different type of dataset and try to compute these values from the R software and you can create your own artificial data set also. Try to take some values of beta1, beta2, beta3 and then try to generate some values of x1, x2, x3 and try to check suppose for example if you try to take beta3 to be very close to 0 like 0.00001 and then try to get your values of y by adding some random errors into it you know how to generate the normal random errors. Then you try to do this analysis once again on this dataset and try to see that you had taken the value of beta3 in the original model to be very close to 0 is it really getting reflected through that custom hypothesis and it will happen in case if your process is correct. It is something like this if somebody has a body temperature high temperature and if you put our fingers on the forehead if the forehead is quite warm then if you try to put a thermometer inside the mouth then the thermometer will also show the high temperature.

The only thing is this by putting the fingers it is very difficult for us to tell whether that temperature is 101 degree or say 101.5 degree or say 102 degrees. But temperature will give us the correct information so that is what exactly we are trying to do it. Your statistical tools are working just like a thermometer and your experience to handle the data to work with the statistical tool is working like as if you are trying to put your fingers on the forehead to measure the temperature. So you try to take some example and try to read from the books also and I will see you in the next lecture with more topics till then goodbye. Thank you.