Regression Analysis and Forecasting

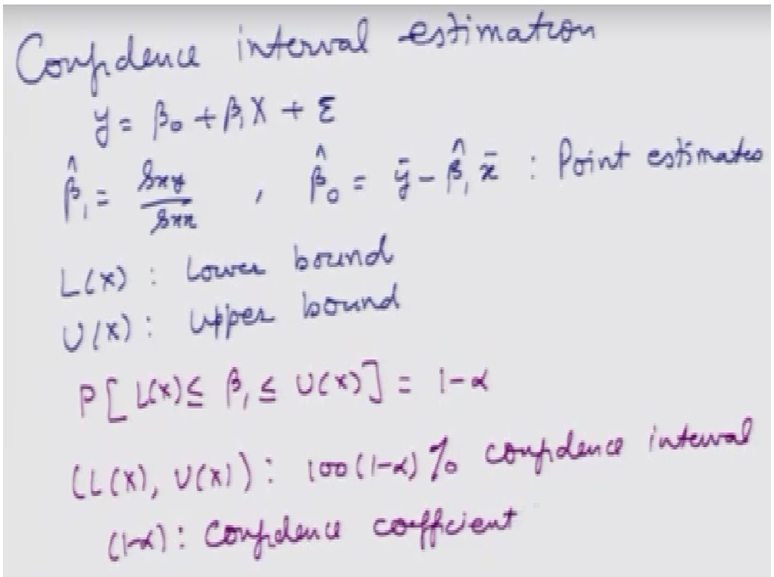
Prof. Shalabh
Department of Mathematics and Statistics
Indian Institute of Technology-Kanpur

Lecture – 09
Testing of Hypotheis and Confidence Interval Estimation in Simple Linear Regression Model (continued)

Welcome to the lecture you may kindly recall that in the earlier lecture we had discussed about the test of hypothesis for the slope parameter under 2 types of condition when sigma square is known and when sigma square is unknown. Now we are going to discuss about the confidence interval estimation for the slope parameter okay, so first of all the question arises what is a confidence interval?

If you say we had consider the model y=beta naught+beta1 x + epsilon, and we had estimated for example beta1 has s x y upon s x x and we had a estimate beta naught as y bar – beta1 hat xbar. These are actually our point estimates, point estimates means they are trying to estimate the value of the parameter at a particular point.

(Refer Slide Time 00:40)

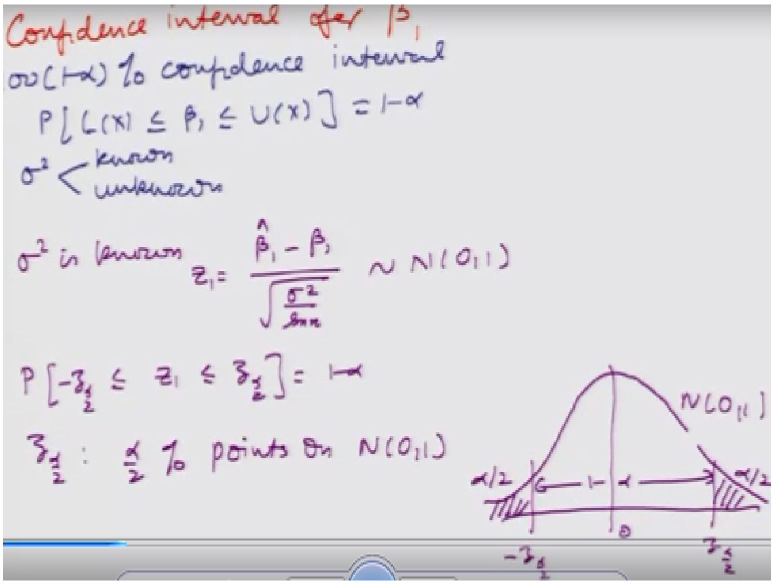


Similar to this we can also estimate these parameters in an interval 1 of the use of the confidence interval estimation is that, the interval estimates are given in the form of lower and upper bound. For example in the case of point estimate suppose I estimate beta1 say 2, then I would say 2 beta1 hat=2, but in the case confidence interval estimation, we will estimate beta 1 is lying between 1 and three with certain probability.

In the case of a confidence interval, we try to create 2 bounds Lx and Ux, Lx is the lower bound and Ux is the upper bound and our objective is to find out the Lx and Ux in such a way such that the probability that beta1 lies between Lx and Ux and these 2 bounds has to be found in such the way such that probability of such an event is 1- alpha. So this Lx and Ux called as hundred 1- alpha % confidence interval, and 1- alpha is called as confidence coefficient.

Confidence interval estimation and test of hypothesis they are very closely related, there is 1 to 1 correspondence between the 2, so many times when we are trying to deal with software the software usually give us the confidence interval, so by looking at the confidence interval we can also infer about the outcome of test of hypothesis. So first of all lets us try to construct the confidence interval for beta1.

(Refer Slide Time: 03:59)



So in the case of confidence interval estimation is hundred 1- alpha % confidence interval for beta1 is found in such a way such that probability that Lx less than = beta1, less than =x should be =1- alpha. In this case we will have 2 options, 1 option is sigma square is known and another option is sigma square is unknown.

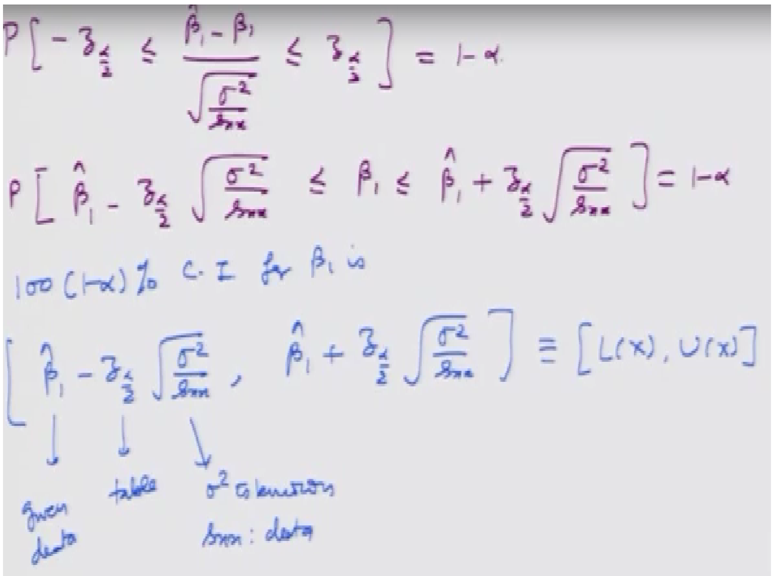
We had seen earlier that when sigma square is known then beta1 hat – beta1 over square root of sigma square sxx this follows a normal distribution with mean 0 and variance 1. So, what I can do here is the following that I can write down and this statistics suppose here z1, so I would say here z1 lies between -z alpha by 2 and +z alpha by 2 and the probability of such an event is 1- alpha

Where your z alpha by 2 are the alpha by 2% points on normal 01. It is something like this for example if I try to draw a curve this is normal 01, so if there are alpha by 2% point on the left tail and if alpha by 2 points are on the right tail and somewhere it is 0, and this reason is 1- alpha then the value of z on the x axis here this is -z alpha by 2 and here this is + alpha by 2.

These values can obtain from that tables and these are standard value which are well known and depending on whether alpha is 1% or say 5% and so on. So now look lets us try to continue further and we try to write down probability that -z alpha by 2 is less than or = beta1 hat – beta1 over square root of sigma square or s x x less than z alpha by 2 and this is 1- alpha.

So we try to solve it further and we can obtain that beta 1 hat - z alpha by 2 square root of sigma square over s x x is less than or = beta1 says beat1 hat+z alpha by 2square root of sigma square over s x x =1 - alpha.

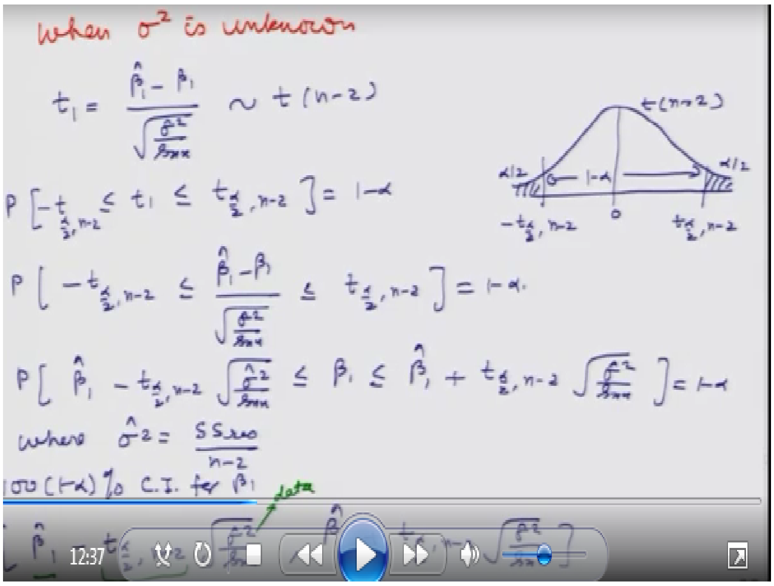
(Refer Slide Time: 06:46)



So finally can write down that hundred 1-alpha % confidence interval for beta1 is the interval that is beta1- z alpha by 2 sigma x square over sxx, 2 beta1 hat+ z alpha by 2 sigma x square over s x x. So you can see her that this beta1 can be obtain from the given data and z alpha by 2 can be obtain from the table and this can again sigma square is known and s x x can be obtain from data and the same thing over here on the right hand side interval.

So this is equivalent to something like we have obtain the Lx and Ux which are the function of the sample values x1 x2 xn and this interval can be obtain on the bases of given sample of data. So this about 1 hundred 1 - alpha percent confidence interval. Now before it giving the detail of is relation with test hypothesis let us try to construct the confidence interval when sigma square is unknown.

(Refer Slide Time: 09:01)



We have seen in the earlier case that when sigma square is unknown then in that case the statistics beta 1 hat - beta 1 over sigma square hat over s x x this follows a t distribution with n- 2 degree of freedom, and let us call this as a statistics t1. So now based on the statistics on the similar lines I can find out the confidence interval that the statistics t1 lies between - t alpha by 2n-2 and + t alpha by 2n-2 and the probability of such an event is 1- alpha.

So this is again equivalent to saying that if I have got t distribution like this 1 then this is here alpha by 2% point on the left hand tail and these are your alpha 2% point on the right tailed test and the mid area is 1-alpha and here somewhere it is 0 and corresponding to this thing the value on the x axis from the t table can be obtain as - t alpha by 2 and -2 degrees of freedom.

This will be your t alpha by 2 and -2 2 degrees of freedom and this is the probability density function of t distribution with L -2 degrees of freedom. Now lets us try to further solve it and we can write beta1 hat – beta1 over sigma square hat over s x x and t alpha by 2n- 2=1-alpha or this can be written as beta1 hat - t alpha by 2, n-2 sigma square hat over s x x less than or =beta1.

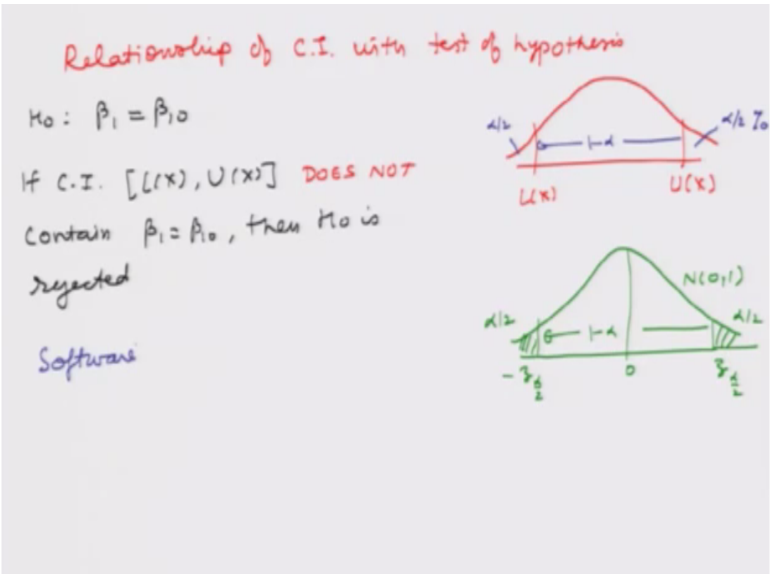
And beta1 hat + t alpha by 2 and –2 sigma hat square over s x x. This probability = - alpha and here we are going to obtain the value of sigma square hat by sum of his squares due to residuals divided by degrees freedom where the sum of the squares due to residual are obtain based on the fitted model using the ordinary least square estimation are maximum likelihood estimation.

So now I have obtain the hundred 1- alpha % confidence interval for beta1 when sigma square is unknown as interval like this and beta1 hat + t alpha by 2 and -2 and sigma hat square over s x x. So here again you can see her that this value of beta1 can be obtain from the given set of data and this sigma square hat can be obtain from the data s x x can be obtain from the data.

This t alpha by 2n-2 is the alpha by 2% points on the t distribution and its value can be obtain from the t table, so this interval can be obtain on the basis of given set of data. Now before going further let us try to understand what is the relationship of this confidence interval estimation with test of hypothesis? So let us try to discuss here relationship of confidence interval with test of hypothesis.

So you have seen that in the case of confidence interval what are we doing essentially we are trying to find out here 2 statistics Lx and Ux the lower bound and upper bound.

(Refer Slide Time: 12:59)



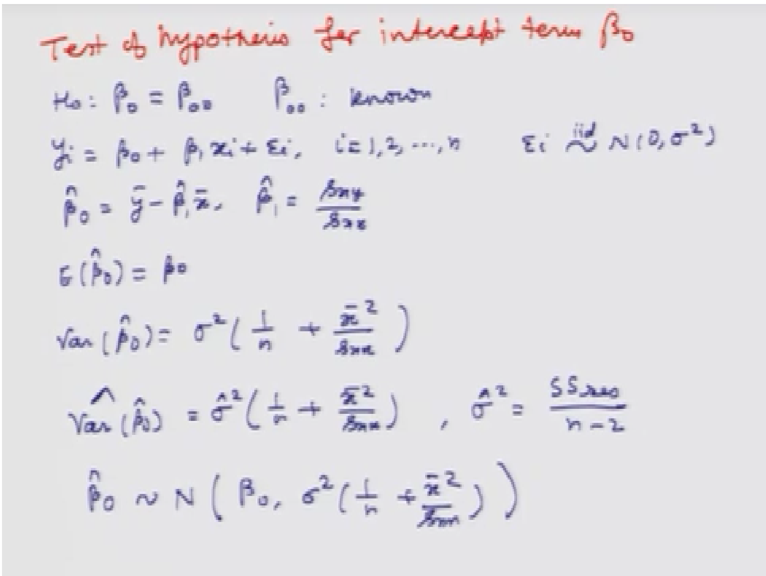
We believe that this area is 1- alpha % and this area is alpha by 2% and this area is alpha by 2%. Similar to this thing you can also recall that in the case of a test of hypothesis when we are considering the 2 sided test hypothesis for example if you try to recall the case of z statistics based on the normal 0 1 what you will trying to do our critical reason was lying on the left hand side on right hand side and this was alpha by 2.

This was alpha by 2 and the mid area was 1 - alpha with something like this on the normal 0 1 curve and this point was - L alpha by 2 and this point was + z alpha by 2 on the normal distribution. Right so now if I try to consider here the test of hypothesis that h1 beta1=beta1 naught. So the rule is very, very simple if confidence interval say.

This here Lx, Ux does not contain beta 1 equal to beta 1 naught then h naught is rejected, so that is relationship test of hypothesis and confidence interval estimation. What is the use of this thing, this is actually more useful when we trying to use the software sometime the software do not gives the outcome about the test of hypothesis rather they provide us the confidence interval in the form of lower limit and upper limit Lx and Ux.

So by looking at the confidence interval we can again a infer about the test of hypothesis. Now after this we try to consider the test of hypothesis for interceptor term beta0.

(**Refer Slide Time: 16**:**00**)



Once you have a understood that how we have constructed the test statistics in case of slope parameter beta1 it is not difficult to create the test statistics for the intercept beta0 and though all the testing procedure everything that remains the same.

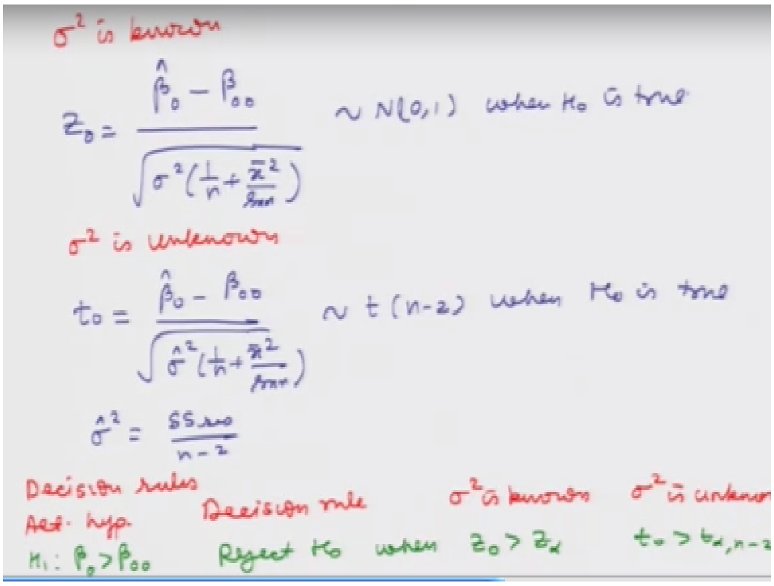
Exactly on the same line as we have done in the case of test of hypothesis for the slope parameter. Just for sake for understanding I will try to quickly give you the results that we are interested now here in testing the null hypothesis for the intercept term beta0=beta to not, not, where beta naught, not is some known value, it is some given value. We had considered the model if you remember beta0+ beta1 xi +epsilon.

We had observed n excess of data we had assume that epsilon i's are iid following normal distribution with mean 0 and variance sigma square and we had obtain the estimate of beta0 hat as y bar – beta1 hat xbar where beta1 hat was obtain has s x y over s x x.We had also demonstrated that expected value of beta0 t is beta0 so it was an unbiased estimator and variance of beta0 hat was obtain has sigma square 1 over n+ xbar x square over s x x.

Further we had obtain the estimate of its variance like by replacing sigma square by sigma square hat in the variance of beta naught hat were sigma square hat= some of square due to residual divided by n-2. We can show that this beat0 hat, this is a linear function of the normally distribute random variable wise.

So this also follows a normal distribution with mean beta0 and variance sigma square 1 over n + xbar square over s x x, so now using this result we can construct the test of hypothesis under 2 cases that when sigma square is known and when sigma square is unknown. So first we try to consider the case when sigma square is known.

(Refer Slide Time: 18:53)



So in this case we can write beat0 hat - beta0 divided by variance of beta hat naught which is equal to sigma x square over s x x and this follows a normal 0 1 distribution when h naught is true. So this is we call as a z statistics and by z naught and similarly when sigma square is unknown then exactly on the similar lines I can develop statistics beta0 hat – beta0 0 divided by sigma square hat1 over n+xbar square over s x x.

This follows a t distribution with n- 2 degrees of freedom when h naught is true and let us try denote this quantity as t naught, and her we try to estimate the sigma square hat again by ss res divided by n-2. Now what about the decision rule? Decision rules remains the same we will just try to write down briefly, so now depending on my alternative hypothesis I will try to write down my decision rule under 2 cases when sigma square is known and when sigma square is unknown.

So first case, I would say suppose the alternative hypothesis is h1 beta naught greater than beta naught, naught in this case the decision rule is to reject h naught when in case of sigma square is known the value of z0 is greater than z alpha.

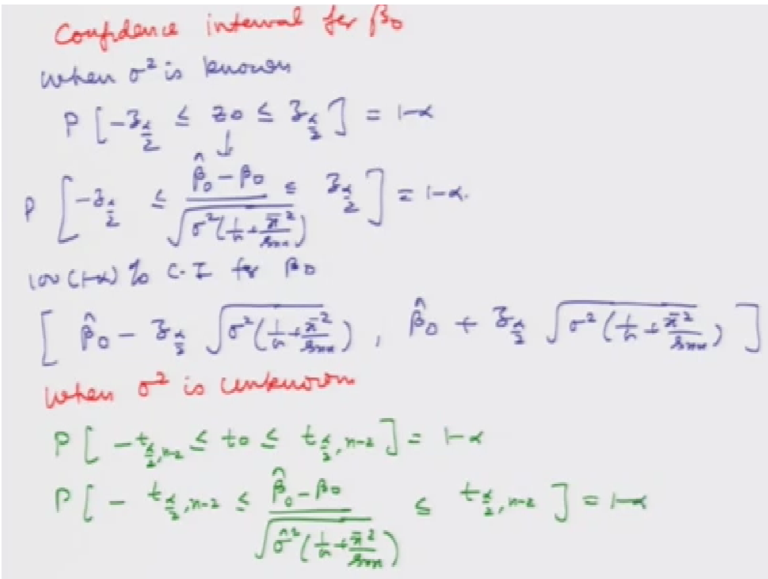
Where Z alpha is the alpha % point on the normal0 1 on the right hand side on the right tail of the normal0 1 distribution and in case if sigma square is unknown then the decision rules to reject h naught when t0 is greater than t alpha and - 2 where t alpha n- 2 are the alpha percent point on the t distribution with n-2 degrees of freedom.

Similarly if my hypothesis alternative hypothesis is h naught beta less than beta naught, then I would say again that the decision rule is to reject h naught whenever z0 is less than - z alpha in case sigma square is known and in case when sigma square is unknown then t0 is less than - t alpha n-2. On other hand if my h1 is beta naught is not equal to beta naught, naught.

Then in that case I will say reject h naught when mode of z naught is greater than z alpha by 2. In case sigma square is known and in case sigma square unknown then this is mode of t0 is greater than t alpha by 2 and -2 where t alpha by 2 and n-2 are the alpha by 2% points on the t distribution with n-2 degrees of freedom. So now after this we can also construct the confidence interval for beta naught.

This is again is not difficult and it goes exactly on the same lines as we have developed the confidence interval for the slope parameter beta naught.

(Refer Slide Time: 23:08)



I am going to use the same statistics what we have used in case of test of hypothesis the statistics z naught and t naught under the 2 cases when sigma square is known and when it is unknown. So, the first is when sigma square is known than in that case I can wrote down that the confidence interval will be obtain in such a way such that z0 lies between - z alpha by 2 and z alpha by 2 and the probability of such an event is 1 - alpha.

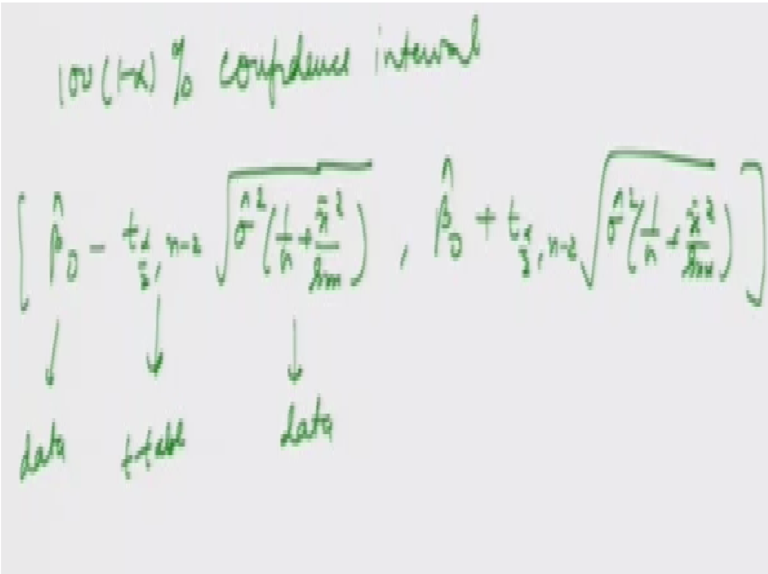
Now if you try to substitute this z0 which is beta hat naught – beta0 divided by sigma square 1 over n+ xbar x square over sxx this becomes - z alpha by 2 between z alpha 2 and the probability of such an event =1- alpha and if you try to solve it say beta0 then we get the hundred 1- alpha percent confidence interval for beta naught like beta hat naught - z alpha by 2 sigma square.

1 over n+xbar square over s x x and beta0 hat + z alpha by 2 sigma square 1 over n+ xbar square over s x x. Similarly in the case when sigma square is unknown to us we simply have to replace sigma square by sigma square hat and in that case we can write the probability that t zero lies between - t alpha by 2n-2 and t alpha by 2n- 2 and the probability of such an even t is1- alpha.

After substituting the value of t naught which is beta naught - beta naught over sigma square root of sigma square hat1 upon n xbar square over s x x t alpha by 2n-2=1-alpha. We can obtain the hundred 1- alpha % confidence interval for beta naught like beta hat not - t alpha by 2 and – 2 square root of sigma square hat1 over n+ xbar square over s x x to beta0 hat + t alpha by 2n - 2 sigma square hat1 over n+xbar x square over s x x.

So now again you can see that this is the value that can be obtain from the data this is the value, that can be obtain from the data. This is the value which is obtained from the tables of the t probabilities that is t table and based on an given set of data this confidence interval can also be obtained. Now we consider the test of hypothesis for sigma square and we would like her to test the null hypothesis sigma square = sigma naught square.Where sigma naught square is some known value and it is specified.

(Refer Slide Time: 27:23)

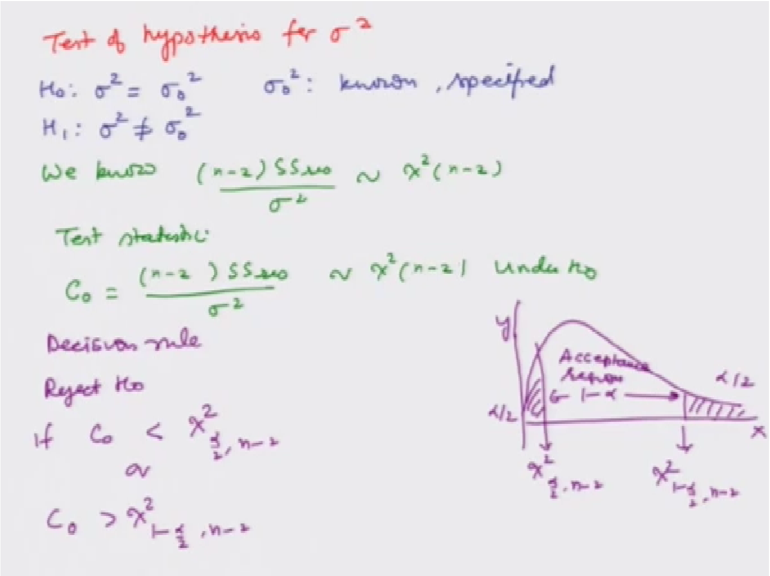


The alternative hypothesis in this case is going to be sigma square naught = sigma naught square. 1 question arises here under what type of condition we are going to use this type of test of hypothesis. We have seen that we have constructed various test statistics for conducting different type of test of hypothesis under 2 type of condition, first condition is sigma square is assumed to be known.

And second condition is sigma square is estimated from the given sample and it is unknown. So in the situation when we assume that sigma square is known then the value of sigma square is obtained by the experimenter either from his experience from some similar kind of experiments conducted in the past or from somewhere. So before assuming that the value, which he is going to use is correct or not he would like to conduct the test of hypothesis.

About the value of sigma square which he is going to assume to be known and then to use in the concern test of hypothesis. So in order to construct the test statistics in such a situation we are going to use the result where we know that n - 2 ss res upon sigma square follows a chi-square distribution with n-2degrees of freedom. So in this case the test statistic is defined as n-2, ss res upon sigma square and this follows a chi-square distribution with n - 2 degrees of freedom under h naught.

(Refer Slide Time: 29:33)



So let us try to denote this statistics as c naught and in this case the decision rule is obtained using the chi-square distribution, so we know that the chi-square distribution is not a symmetric distribution, so this is our x axis this is y axis and somewhere here there is alpha by 2 region, critical region and here on this side, this is alpha by 2 critical region and in between this is the 1- alpha region which is here acceptance region.

The critical value which are obtained from the tables of chi-square probability they will be here like this chi-square at alpha y 2 level with n-2 degrees of freedom and this value is going to be chi-square with 1 - alpha by 2 and -2 degrees of freedom. So we can see here that in this case that we reject the null hypothesis h naught if c naught is less than chi-square alpha by 2 and -2 or c naught is greater than chi-square 1- alpha by 2 and - 2.

So this will help us in testing the hypothesis about the sigma square. Now we can also construct the confidence interval for sigma square, so now we consider the confidence interval for sigma square, so in order to construct the confidence interval for sigma square we are going to use the result which we had discussed earlier that sum of a square due to residual divided by sigma square follows a chi-square distribution with n - 2 degrees of freedom.

So using this result I can find out the confidence interval in the following way that ss res divided by sigma square this is going to lie on the chi-square with alpha by 2% point and n- 2 degrees of freedom with chi-square point 1- alpha by 2% and n-2 degrees of freedom and the probability of such an event is 1- alpha. You have to just keep in mind.

That chi-square is not a symmetric distribution like normal distribution or t distribution so that is why this distribution may look like this and in this case we case say that this is our alpha by 2% and on this side this is alpha by 2%, so if I say that this is our chi-square alpha by 2 with n-2 degrees of freedom then this point become chi-square with 1 - alpha by 2n- 2 degrees of freedom and this mid part is 1- alpha.

This is the probability density function of chi-square with n - 2 degrees of freedom, so now if you try to solve it this become simply here ss res divided by chi-square with 1 - alpha by 2n-2 less than or= sigma square ss res or chi-square with alpha by 2 and n -2degrees of freedom 1- =1- alpha, so hundred 1- alpha percent.

Confidence interval for sigma square is now simply ss residual divided by chi-square 1- alpha by 2n-2 degrees of freedom and ss residual divided by chi-square alpha by 2n-2. So these 2 values of chi-square they can be obtained from the tables of probabilities for chi-square. So they are available sum of a square due to residual can be obtained on the basis of given set of data.

So this hundred 1- alpha % confidence interval can be obtained on the basis of given set of data. Now we have completed the test of hypothesis and confidence interval estimation for all the model parameters namely beta0 that is the intercept term beta1, which is the slope parameter and sigma square. Once we have estimated all the parameters equivalently we have done the modeling and we have obtained a statistical model.

So the next turn I will try to take simple example and using a software, I will try to analyze the data my main objective will be that whatever we have done in the case of simple linear regression model we will try to demonstrate how those values have been obtain from the software and how they can be interpreted for getting a statistical model till then good bye.