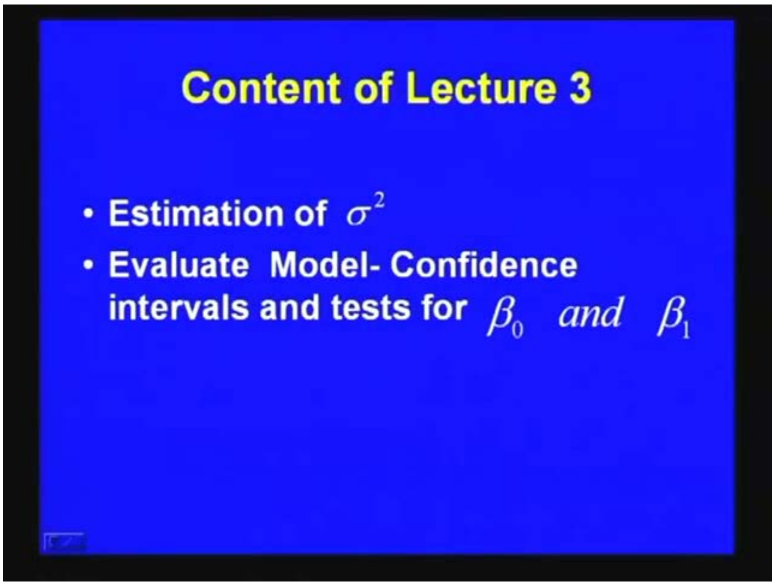
Regression Analysis
Prof. Soumen Maity
Department of Mathematics
Indian Institute of Technology, Kharagpur

Lecture - 3
Simple Linear Regression (Contd.)

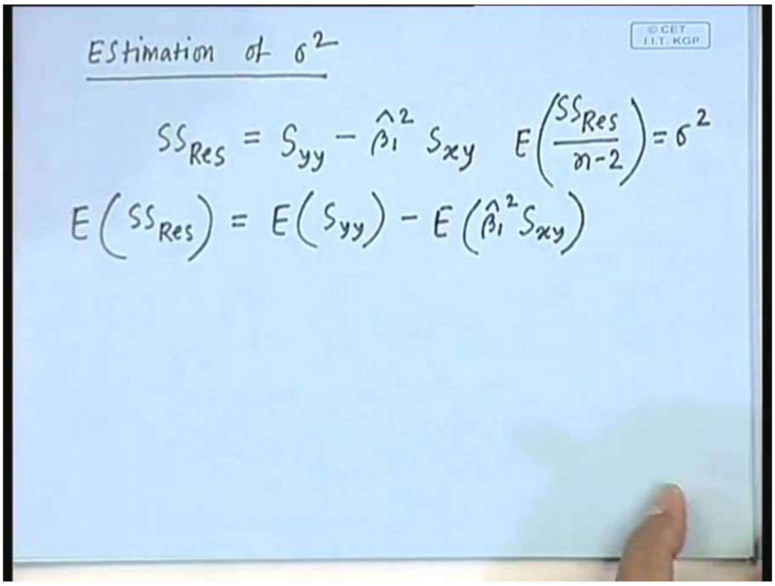
Hi, this is my third lecture in Simple Linear Regression. In the first lecture, we have learned how to fit simple linear regression model to a data set, which fits the data best. That means, we have learned how to fit a simple linear regression model using the least square technique. In the second lecture, we have learned the statistical property of the regression coefficient that is, beta naught hat and beta 1 hat. And we have observed that, both the beta naught hat and beta 1 hat, they are unbiased estimator of beta naught and beta 1 respectively. And we computed the variance of beta naught hat and beta 1 hat and we found that, both the variance of beta naught hat and beta 1 hat, they involve sigma square. So, sigma square is the population variance, which is unknown, so what we need to do is that, we need to estimate the population variance sigma square.

(Refer Slide Time: 02:48)



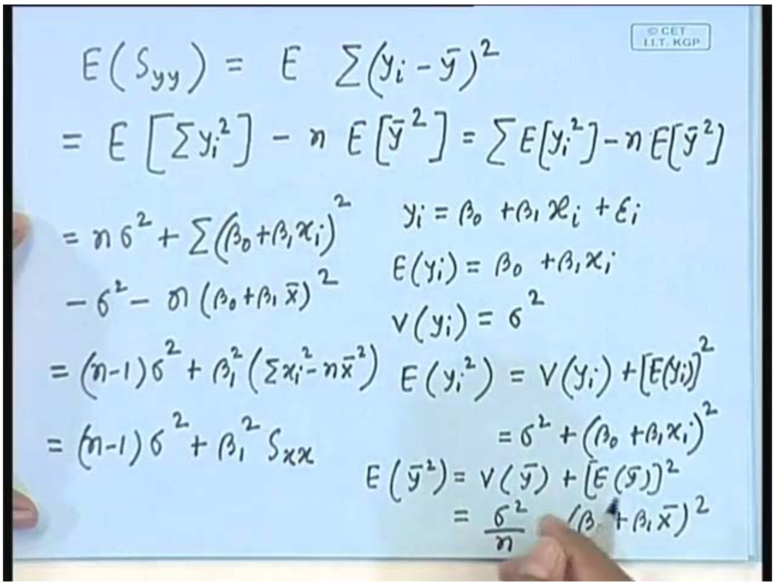
So, here is the content of today's lecture, we are going to estimate the population variance, sigma square. So, we will give an unbiased estimated of sigma square and next, we evaluate the performance of the fitted model. So, we will talk about the confidence intervals and tests for the regression coefficients, beta naught and bets 1.

(Refer Slide Time: 03:32)



So, first we talk about the estimation of sigma square, the estimation of sigma square is obtained from S S residual. And in lecture 2, we have proved that, S S residual, this can be written in the form S y y minus beta 1 hat square S x y. Now, our ultimate aim is to prove that, S S residual by n minus 2, this is an unbiased estimator of sigma square. So, what you need to do is that, we will find the expected value of residual sum of square. So, expected value of residual sum of square, S S residual is equal to expectation of S y y minus expectation of beta 1 hat square S x y. So, first let me find the value of expected value of S y y.

(Refer Slide Time: 05:52)



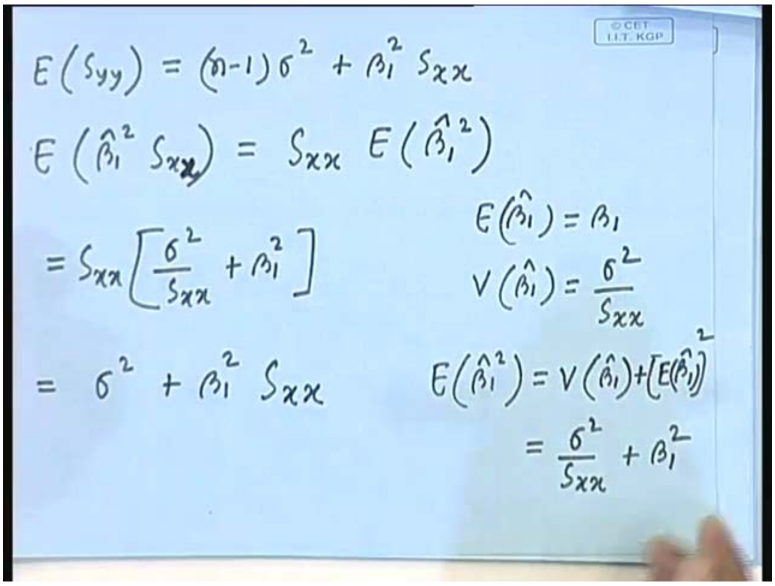
So, expected value of S y y, which is equal to, expectation of S y y is nothing but summation y i minus y bar whole square and this can be written as, expectation of summation y i square minus n time expectation of y bar square. Now, again this one is equal to summation of expectation of y i square minus n times expectation of y bar square. So, what is expectation, let me recall the model y i, y i equal to beta naught plus beta 1 x i plus epsilon i.

And we assume that, expected value of epsilon i is equal 0, so expected value of y i equal to beta naught plus beta 1 x i, x i is not a random variable. And we also know, we also assume that, the variance of epsilon i is equal to sigma square, so the variance of y i is also equal to sigma square. Now, expected value of y i square is equal to variance of y i plus expectation of y i whole square, this is from the definition of the variance. Now, the variance of y i is equal to sigma square and expectation of y i is this quantity, so this is equal to beta naught plus beta 1 x i whole square.

And similarly, we can find out the expected value of y bar square, it can be proved that, expected value of y bar square is equal to, of course this is equal to variance of y bar plus expectation of y bar whole square. So, variance of y bar is equal to sigma square by n and expectation of y bar is equal to beta naught by beta 1 plus x bar, this is equal to beta naught plus beta 1 x bar whole square.

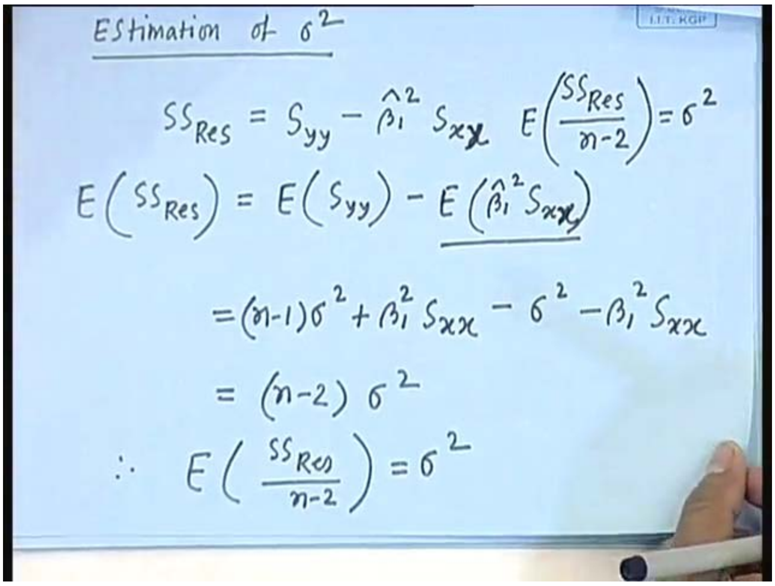
Now, basically what I will do is that, I will plug these values here, so expected value of S y y is equal to this thing, summation over i, so n sigma square plus summation beta naught plus beta 1 x i whole square minus n times expectation of y bar square, which is this quantity, so n times this term. So, this is basically sigma square minus n times beta naught plus beta 1 x bar whole square. So, little bit algebra, we will prove that, this is nothing but n minus 1 sigma square plus beta 1 square summation of x i square minus n x bar whole square. And this is nothing but n minus 1 sigma square plus beta 1 square S x x, this is the notation for this term, summation x i bar whole square.

(Refer Slide Time: 12:56)



So, what we proved that, expected value of S y y is equal to n minus 1 sigma square plus beta 1 square S x x.

(Refer Slide Time: 13:20)

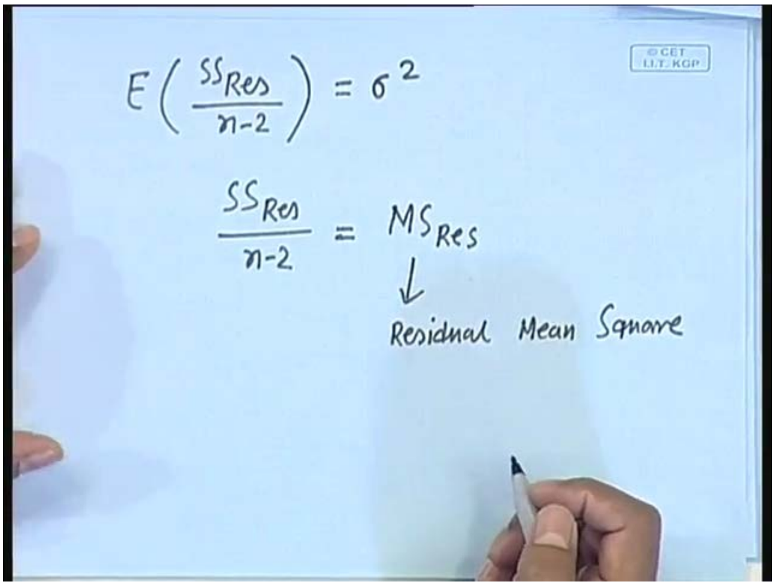


Now, see that, you want to compute the expected value of residual sum of square, so this involves the expected value of S y y and the expected value of beta 1 hat square S x y. Next we will compute the expected value of this one, so this is equal to next expectation of beta 1 hat square S x y, is equal to, I think I did a mistake here, this is not x y, this is x x, so this is not x y, this is x x. So, this one is equal to S x x, expected value of beta 1 square and we know that, expected value of beta 1 hat is equal to beta 1, this is an unbiased estimator and the variance of beta 1 hat is equal to sigma square S x x.

So, expected value of beta 1 hat square is equal to variance of beta 1 hat plus expected value of beta 1 hat whole square. We know both the values here, this is equal to sigma square by S x x and this one is equal to beta 1 square. So, this thing is equal to S x x and expected value of beta 1 hat square is equal to sigma square S x x plus beta 1 square, so this is going to be equal to sigma square plus beta 1 square S x x.

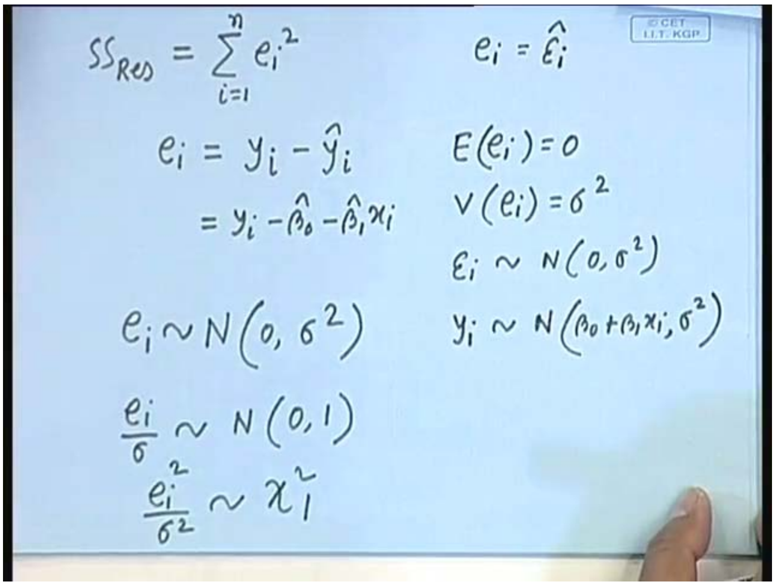
Now, just we need to plug these two values here, expected value of residual sum of square is equal to, we proved that, this one is equal to n minus 1 sigma square plus beta 1 square S x x and this one is equal to minus sigma square minus beta 1 square S x x and this one is nothing but n minus 2 sigma square. So, what we proved is that, expected value of S S residual, residual sum of square by n minus 2 is equal to sigma square that means, residual sum of square by n minus 2 is an unbiased estimator of sigma square.

(Refer Slide Time: 17:41)



So, what we proved is that, expected value of S S residual by n minus 2 is equal to sigma square and this S S residual by n minus 2, this is also denoted by M S residual and this is called residual mean square. This is ultimately found an unbiased estimator of sigma square, which is M S residual, next we will talk about the distribution of M S residual.

(Refer Slide Time: 18:59)

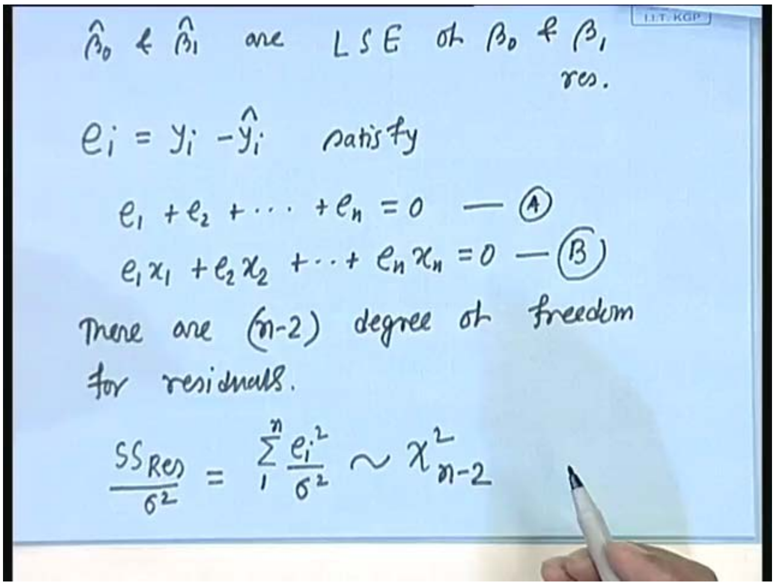


So, what is S S residual, residual sum of square is nothing but summation e i square, i is from 1 to n. So, this e i is nothing but the i th residual, this is the difference between the absorbed response value and the predicted response value. So, it can be proved that, expected value of e i is equal to 0, so you can prove it and also it can be proved that, the variance of e i is equal to sigma square. See, this e i is nothing but the estimate of i th error term, the variance of e i, this can be proved sigma square.

And also we have assumed that, epsilon i follows normal zero sigma square and they are independent, which implies that, the observation y i, they are also normal with mean beta naught plus beta 1 x i and variance sigma square. Now see, this e i is, it is linear combination of y i, this y i hat is nothing but beta naught hat plus beta 1 hat x i. And both we proved that, beta 1 hat is linear combination of the observations and also beta naught hat is also a linear combination of the observation.

So, the whole thing this is e i, is a linear combination of the observations, which implies that, e i follows normal. Because, e i is a linear combination of normal variables and the linear combination of normal variables is also normal. So, that is why, e i follows normal with mean 0 and the variance is equal to sigma square and from here, we can say that, e i by sigma, this follows standard normal (0,1). And also since e i by sigma follow standard normal, we can say, e i square by sigma square, this follows chi square 1. And S S residual is, basically it is a sum of e i square, i equal to 1 to n, but the distribution of S S residual is naught chi square n, because all the e i are not independent, this e i satisfies the some constant.

(Refer Slide Time: 23:22)

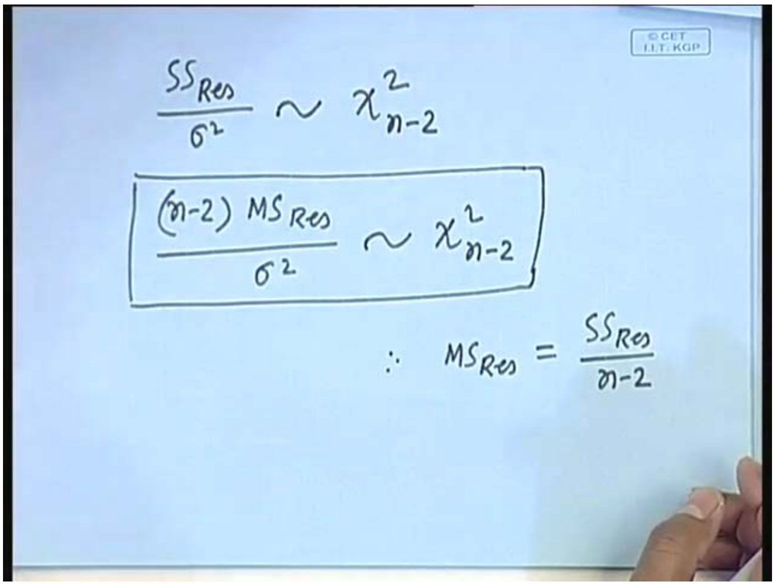


We know that, beta naught hat and beta 1 hat are least square estimator of beta naught and beta 1 respectively. And this e i, which is equal to y i minus y i hat, they satisfy the constant, that e 1 plus e 2, e n is equal to 0. That is, this is what we proved before also, the sum of the residuals is equal to 0 and also it satisfies, this basically the first normal equation, which residuals satisfy. And the second normal equation is summation e i x i is equal to 0 that is, e 1 x 1 plus e 2 x 2 plus e n x n is equal to 0.

So, what I want to prove here is that, this S S residual by sigma square, which is equal to summation e i square by sigma square. This does not follow chi square n, it follows chi square with degree of freedom n minus 2, because there are n minus 2 degree of freedom for the residuals. All the e i are not independent, the first I mean, you can choose n minus 2 residuals independently and then the remaining two residuals have to be chosen in such a way that, they satisfy the condition that, summation e i is equal to 0 and summation e i x i is equal to 0.

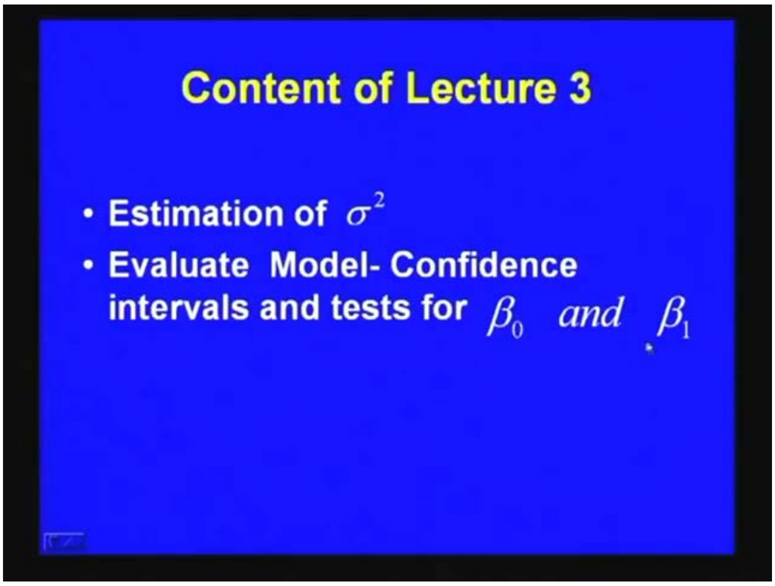
So, you have the freedom of choosing n minus 2 residuals or n minus 2 e i independently and the remaining two residuals must be chosen in such a way that, they satisfy these two conditions. That is why, the distribution of S S residual by sigma square follows chi square with degree of freedom n minus 2, it is not chi square n. If all the e i are, if there is no constrain on e i then this could follow chi square n, but since we have two constrain on the residuals, it does not follow chi square n, it follows chi square n minus 2. So, I repeat just that, we have the freedom of choosing n minus 2 residuals independently and the remaining two residuals have to be chosen in such a way that, they satisfy these two conditions like summation e i equal to 0 and summation e i x i equal to 0.

(Refer Slide Time: 28:10)



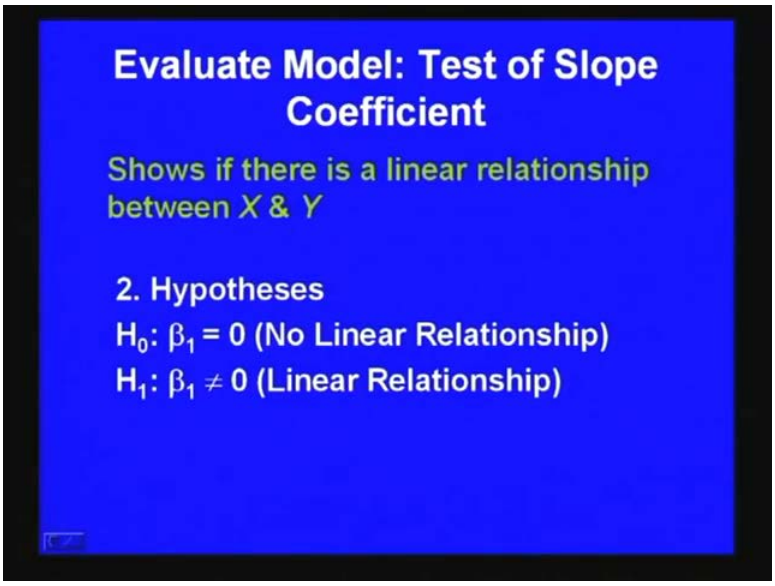
What we proved is that, summation S S residual by sigma square follows chi square n minus 2, which is equivalent to say that, n minus 2 M S is residual by sigma square follows chi square n minus 2, because M S residual is nothing but S S residual by n minus 2. So, this is the result we proved and this is very much useful in testing of hypothesis.

(Refer Slide Time: 29:09)



So, next we move to the evaluation of model, so what we learnt in the 1 st lecture is that, given the state of data or given the state of observations, we learned how to estimate the regression coefficient. That means, we have learned, how to fit regression model to the data, so once the linear model has been fitted, the next job is to confirm the goodness of the fit. So, what we will do is that, we are going to test the significance of the regression coefficients, beta naught and beta 1.

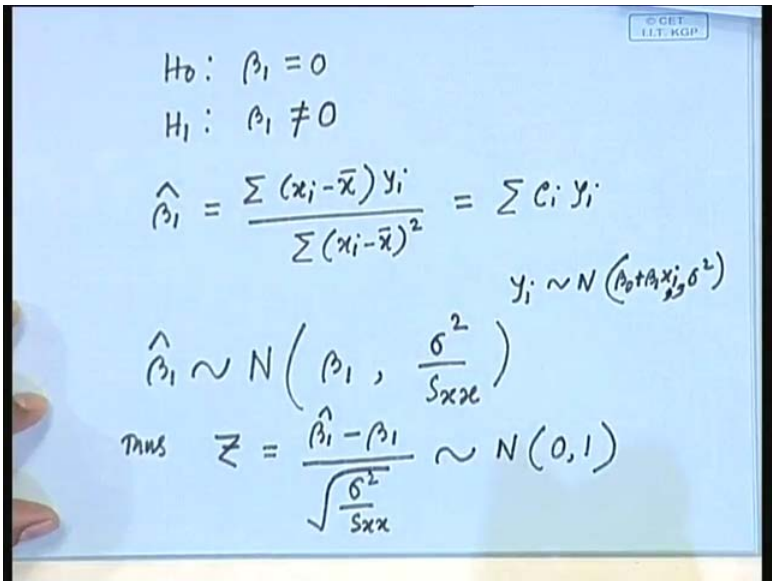
(Refer Slide Time: 30:28)



First we will test the hypothesis that, beta 1 is equal to 0, this is null hypotheses against the alternate hypothesis that, beta 1 not equal to 0. So, what is the significance of this null hypothesis, beta 1 is equal to 0, if beta 1 is equal to 0 then the model will become Y equal to beta naught plus epsilon that means, there is no linear relationship between the variables X and Y. So, if H naught is accepted then we conclude that, there is no linear relationship between the regressor variable and response variable.

And we say that, X is of little value in explaining the variation in Y, whatever be the value of X, we can estimate the regressor variable y i by y bar. If H naught is rejected that means, if the alternative hypothesis, the two sided alternate hypothesis beta 1 is accepted. That means, it says that, there is a linear relationship between the regressor variable X and the response variable Y and X is of value in explaining the variation in Y, this is the significance of the alternative hypothesis, so how to perform this test, we have to compute the value of the test statistics.

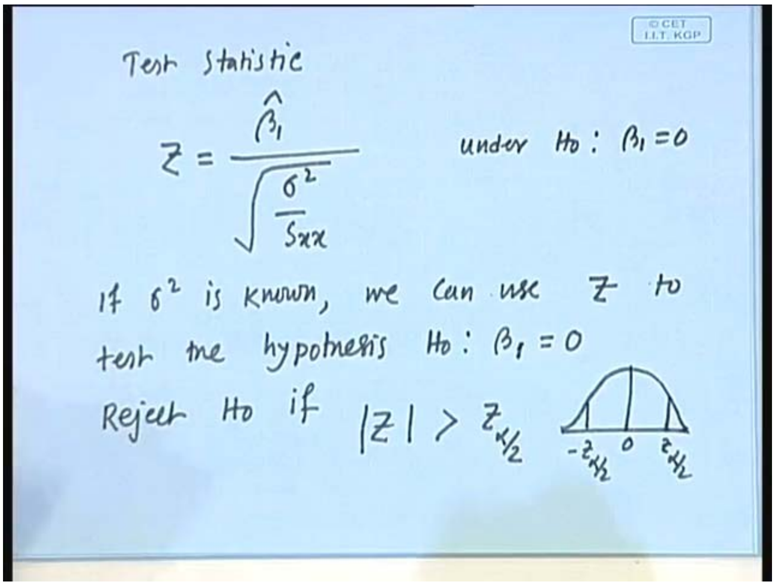
(Refer Slide Time: 33:07)



And here, to test the hypothesis, H naught which is equal to beta 1, is equal to 0, against the alternative hypotheses H 1, beta 1 not equal to 0. The test statistics here is beta 1 hat, which is an estimator of beta 1 and before we have prove that, beta 1 hat which is equal to summation x i minus x bar into y i by summation x i minus x bar whole square, this can be written as summation C i y i. That means, beta 1 hat is linear combination of the observation y i and you know that, y i follows normal distribution with some mean and variance sigma square and the mean is beta naught plus beta 1 x i.

So here, beta 1 hat is linear combination of normal variables, which implies that, beta 1 hat is also normal with mean. We know the mean of beta 1 hat, beta 1 hat is an unbiased estimator, the mean of beta one hat is beta 1 and the variance of beta 1 hat is we proved before that is, sigma square S x x. So, we need the sampling, this is called the sampling distribution of beta 1 hat, to find the critical value for the testing of hypothesis. Now, from here, we can say that, the z equal to say, beta 1 hat minus beta 1 by sigma square S x x, we know that, this follows normal (0,1) standard normal.

(Refer Slide Time: 36:30)

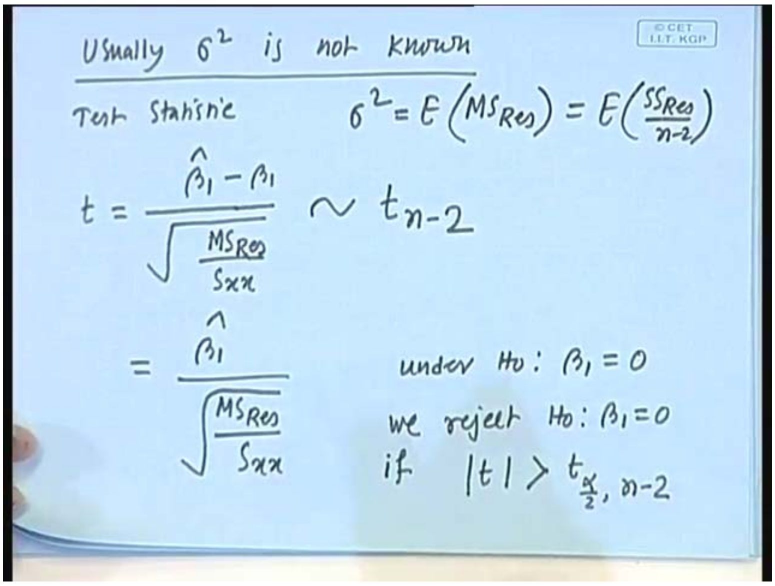


And the test statistic is equal to beta 1 hat by sigma square S x x, this is z, z equal to this under H naught, because under H naught, beta 1 is equal to 0. And this is the test statistic to test the hypothesis that, H naught beta 1 equal to 0 against the alternative hypothesis that,.beta 1 is not equal to 0. Now, see usually these variance I mean, sigma square is not known, the population variance is not known. If sigma square is known, we can use z to test the hypothesis H naught beta 1 equal to 0.

And we reject the critical region here, we reject H naught if z is greater than z alpha by 2, so we reject the null hypothesis at alpha level of significance, if z is greater than z alpha by 2. So, this z alpha by 2 is nothing but the upper alpha by 2 percentage point of standard normal distribution. So, this is the point, z alpha by 2 and this is the point minus z alpha by 2, this is the period of standard normal distribution. So, intuitively I mean, we reject the null hypothesis, if beta 1 is not close to 0.

And formally I mean, we found the test statistic value z and we checked, if the z is greater than z alpha by 2 then we reject the null hypothesis, otherwise we accept the null hypotheses.

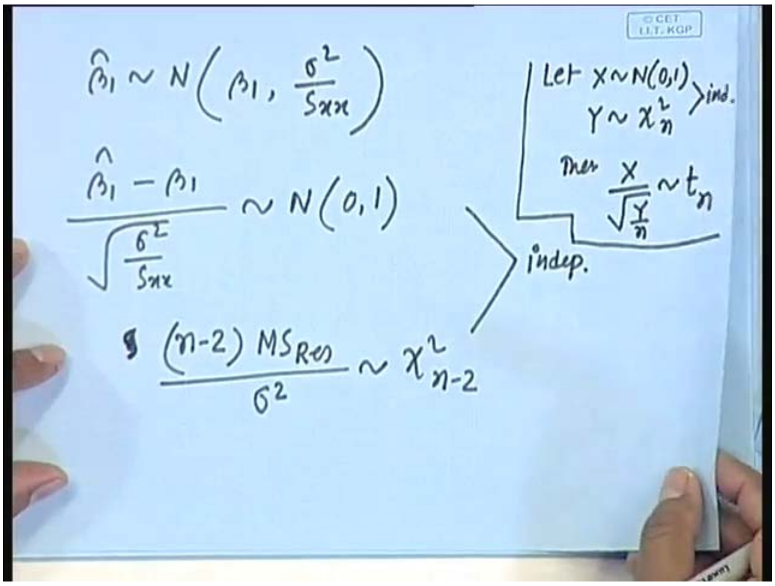
(Refer Slide Time: 39:53)



But, I mean, usually sigma square is not known, this is the practical case I mean, we cannot assume that, sigma square is known. And we know that, unbiased estimate of sigma square is M S residual, because we proved that, expectation of M S residual, which is equal to expectation of S S residual by n minus 2, this is equal to sigma square. So, what you do is that, that is, the statistic to test the hypothesis beta 1 is equal to 0, this was of the form, beta 1 hat by root over of sigma square by S x x.

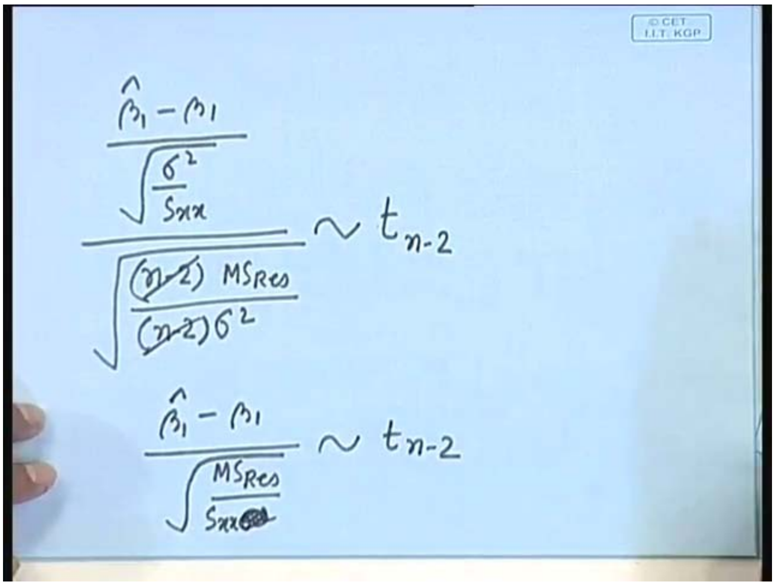
But, sigma square is not known here I mean, this is the more practical case that sigma square is not known. So, what we do is that, we just replace this sigma square by it is unbiased estimator, M S residual and we call it t and this is the test statistic say, beta 1 here. This is the test statistic to test a given hypothesis, now this does not follow normal, we need to find the distribution of this one.

(Refer Slide Time: 41:49)



What we need to find is that, beta 1 hat follows normal with mean beta 1 and variance sigma square by S x x. And so from here, we can say that, beta 1 hat minus beta 1 by sigma square S x x root, this follows standard normal, normal (0,1) this is one. And also we know that, n minus 2 M S residual that means, S is residual by sigma square, this follow chi square n minus 2 and it can be proved that, these two are independent. Now, there is one very standard result in sampling distribution, let X follows normal (0,1) and Y follows chi square n. And they are independent then X by root over of Y by n, this follows t distribution with the degree of freedom n. This is a very standard result, I am expecting that, you know sampling distribution well and now, we are going to make use of this result here. So, this one follows standard normal and this one follows chi square n minus 2.

(Refer Slide Time: 44:12)

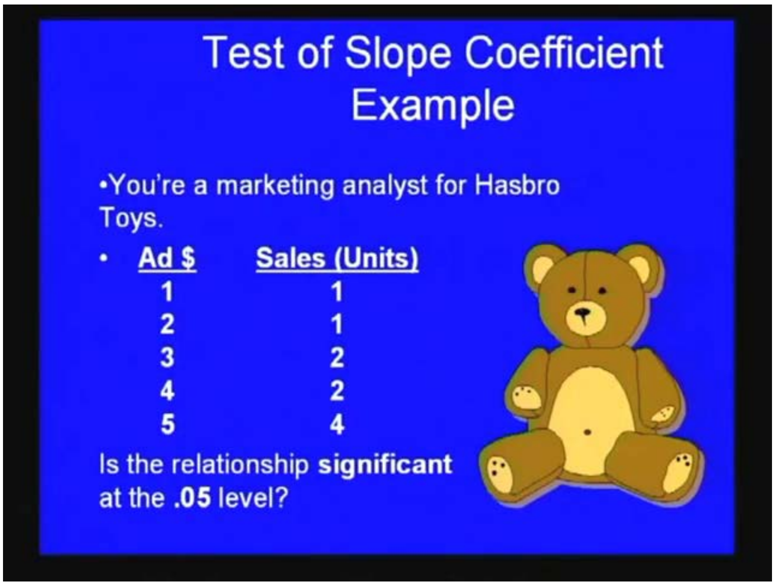


And from here, I can say that, beta 1 hat minus beta 1 by sigma square S x x by n minus 2 M S residual, this is my X, this is my Y. So, Y by n minus 2, root of this thing, this follows t with n minus 2 degree of freedom. I am just making use of this result, this is my X, this is my Y and Y follows chi square n minus 2, instead of n and x follows standard normal. So, X by root of Y by n minus 2, this follows t n minus 2 and from here, we get that, beta 1 hat minus beta 1 by, this will cancel out, by M S residual by S x x, this follows t n minus 2, we obtain this from here only.

So, when we are talking about testing the hypothesis, beta 1 equal to 0 against the alternative hypothesis, beta 1 not equal to 0. And we are considering the case that, sigma square is not known, so basically what I did here is that, I replaced the sigma square by S S residual. And we proved that, this follows t n minus 2 and this t is equal to beta 1 hat by M S residual by S x x under H naught. Under H naught, beta o1e is equal to 0, I am expecting that, you know testing of hypothesis well and now, we know the distribution of test statistics.

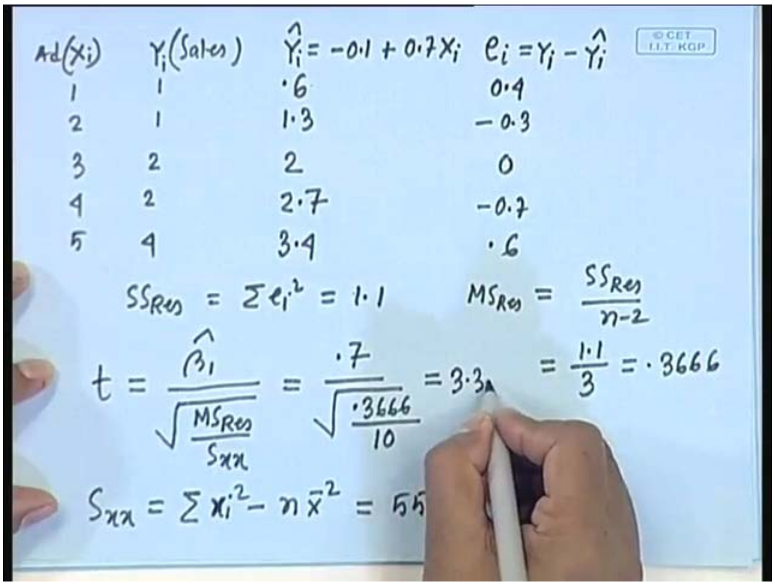
And this is a two sided test, so we reject the null hypothesis, so we reject H naught, beta 1 equal to 0. If this t value is greater than t alpha by 2 with degree of freedom n minus 2, this t alpha by 2, n minus 2 is the upper alpha by 2 percentage point of t distribution with n minus 2 degree of freedom, so this is the critical region. And if the t value is greater than this one, the mode of t value is greater than this one and tehn, we are going to reject the null hypothesis, beta 1 equal to 0. And we conclude that, there is a linear that, rejecting the null hypothesis means, accepting the alternative hypothesis that, beta 1 not equal to 0. And we conclude that, there is a linear relationship between the regressor variable and the response variable.

(Refer Slide Time: 48:34)



Now, let us recall the toy example and these are the cost, the money spent on advertizing and this is the sales amount and what you want to check, before we already have fitted relationship between X and Y. And now, you want to check, whether the relationship is significant at 0.05 level of significance that is, alpha equal to 0.05 that is, type one error, so let me do this one.

(Refer Slide Time: 49:30)



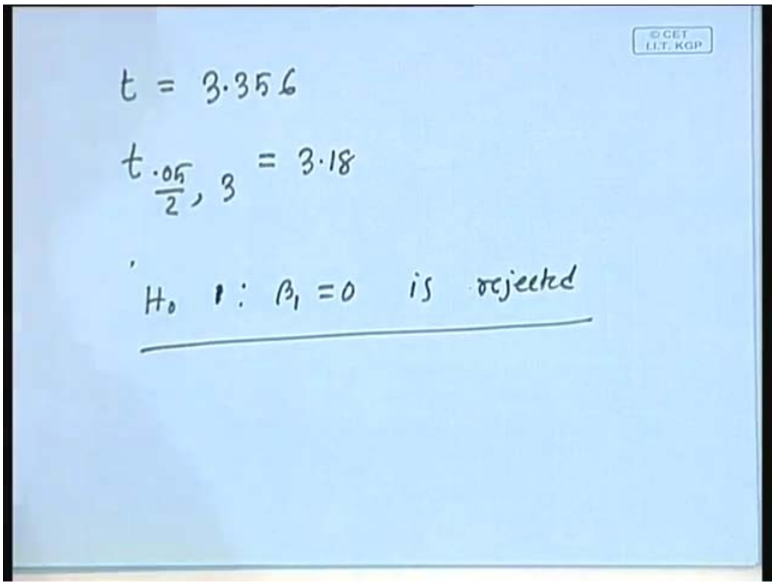
Here X is equal to the cost, the money spent on advertisement that is, the regressor variable. The values for this one is 1 2 3 4 5 and Y is the sales amount and the values are 1 1 2 2 4. And using the least square technique, we already have estimated the regression coefficient and the fitted model is Y hat equal to minus 0.1 plus 0.7 X, so beta 1 is 0.7 and beta naught equal to minus 0.1, so this is the fitted model. Now, corresponds to X equal to 1, the predicted value of the response variable is 0.6, you just put X equal to 1 here, that gives that Y 1 hat for the predicted value is 0.6, corresponds to X equal to 1.

Similarly, for X equal to 2 the predicted value is 1.3, for X equal 3 the predicted response value is equal to 2, for X equal to 4 Y hat is equal to 2.7, for X equal to 5 Y hat is equal to 3.4. So, this is Y 1 hat Y 2 hat Y 3 hat Y 4 hat Y 5 hat, now let me compute e, the residuals e i, we will call them i, so e i equal to Y i minus Y i hat that is, the observed response and the predicted response, so the values are 0.4, minus 0.3, 0, minus 0.7, 0.6. So, from here, we can compute S S residual, which is equal to summation e i square, this can be proved that, from here you can compute this, this is equal to 1.1.

And here, M S residual is equal to, you know M S residual is S S residual by n minus 2, here n is equal to 5, we have 5 observations. So, n minus 2 equal to 3, so this gives, this equal to 1.1 by 3, which is equal to 0.3666. Let me compute the test statistics value t, t equal to beta 1 hat by root over of M S residual by S x x. So, we know what is the value of M S residual, beta 1 hat equal to 0.7 and the value of M S residual is 0.3666. What is

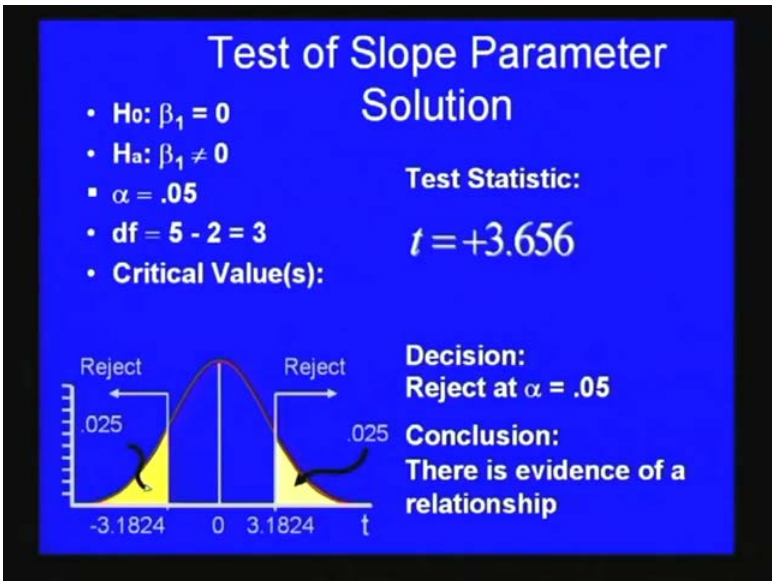
S x x, S x x is equal to summation x i square minus n X bar square, so it is not difficult to compute from here, you are given the x i's values. So, this is equal to 55 minus n equal to 5 and X bar equal to 3 square, which is going to be 10 and the t value, ultimate t value is going to be 3.356.

(Refer Slide Time: 54:27)



So, the t value is equal to, the absorbed t value is 3.356 and from the table, you find the value of t alpha by 2. Here, alpha is equal to 0.05 by 2 and degree of freedom is n minus 2, n is 5, so this is equal to 3, this is going to be 3.18. So, since t is greater than this quantity, we conclude that, H naught is that is, beta 1 is equal to 0, is rejected. That means, we say that, there is a linear relationship between the regressor variable and the response variable.

(Refer Slide Time: 55:22)



Here is the summary, we are trying to test this hypothesis, we have the given observation and the level of significance is 0.05, degree of freedom is 5 minus 2 equal to 3. We already computed the test statistic value, which is equal to 3.656 and the critical value is 3.1824, which is nothing but t 0.025 with degree of freedom 2. And since this 3.6 is in the critical region, it lies in the critical observed value, is in the critical region, we are going to reject the null hypothesis at the alpha level of significance. And conclude that, there is evidence of linear relationship between the response variable and the regressor variable and now, we can stop for today.

Thank you.