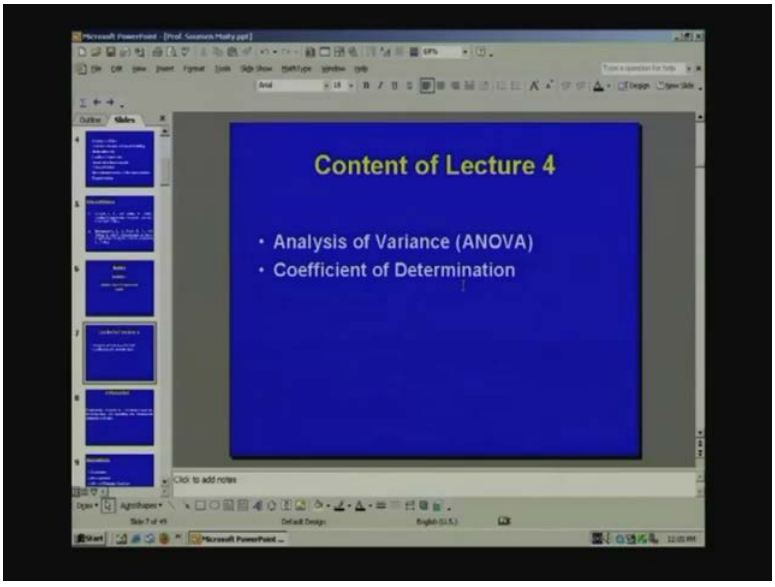
Regression Analysis
Prof. Soumen Maity
Department of Mathematics
Indian Institute of Technology, Kharagpur

Lecture - 4
Simple Linear Regression (Contd.)

Hi this is my 4th lecture on Simple Linear Regression, till now I mean given a set of observation, we know how to a beat Simple Linear Regression model to the data and once the module as been fitted then, we need to determine the goodness of the feet. So, and also we need to test, the statistical significance of the regression and coefficients.

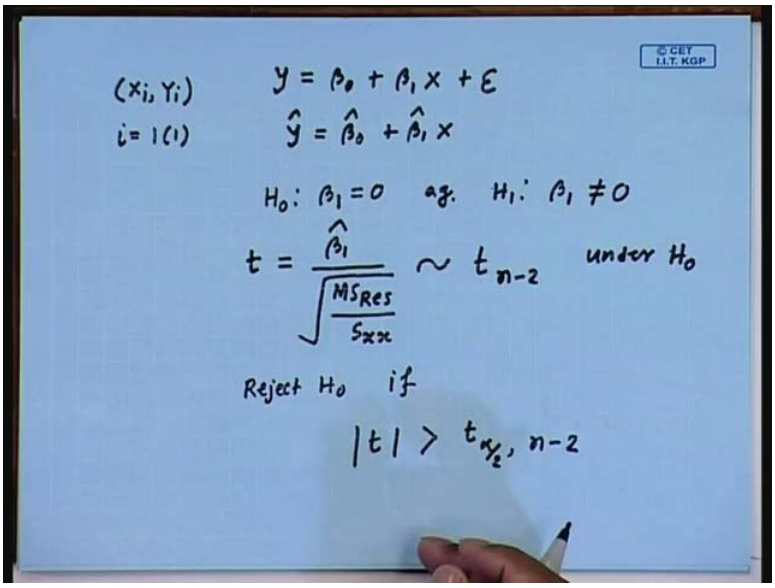
Well so, one will do, this on like we in the last lecture, we have test it hypothesis is not, which is beta 1 equal to 0, against the alternative hypothesis that, beta 1 is not equal to 0. And we have used the test statistics t to test this hypothesis; so another way to approach this problem is the analysis of variance, so today we basically talking about ANOVA.

(Refer Slide Time: 02:28)



So, the content of lecture 4 is analysis of variance, in abbreviation it is ANOVA and also will talking about the coefficient of determination.

(Refer Slide Time: 02:53)

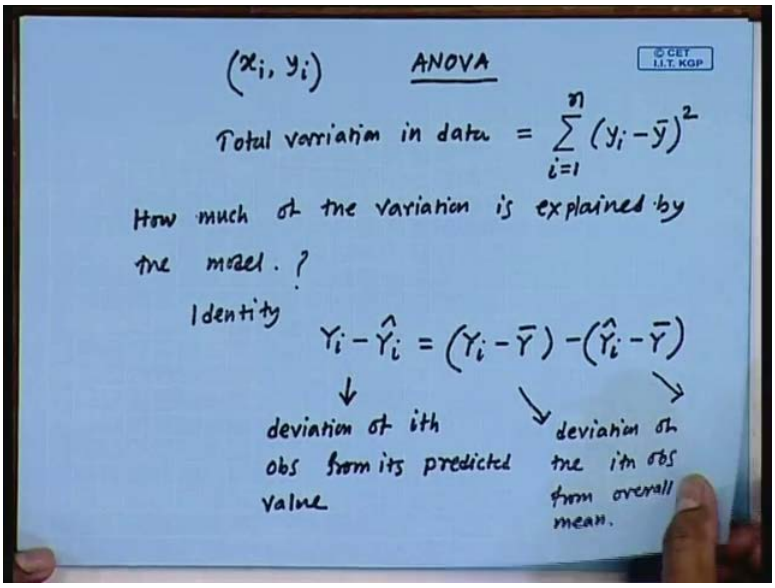


Well so, our module is Y equal to beta naught plus beta 1 X plus epsilon, so given a set of data; say for example, X i Y i, for i equal to 1 to end, we know how to fit a simple linear regression module this to data that is Y hat equal to beta naught hat plus beta 1 hat X. And once the module has been consecrated, it is important to conform the goodness of feet and statistical significance of the regression coefficient.

So, we have pasted the hypothesis H naught, which is equal to beta 1 equal to 0, against the alternative hypothesis H 1, which says that beta 1 is not equal to 0. And in lecture 3, we have used the test statistic t, which is equal to beta 1 hat by M S residual by S x x root over, which follows t distribution with degree of freedom in minus 2 and this is under H naught. So, we have used this test statistic, 2 test the hypothesis H naught and we reject H naught, if is t value is greater than, t alpha by 2 n minus 2 at the level of significance, here is equal to alpha.

So, now, another approach to solve this problem is called ANOVA technique, it is analysis of variance.

(Refer Slide Time: 06:18)

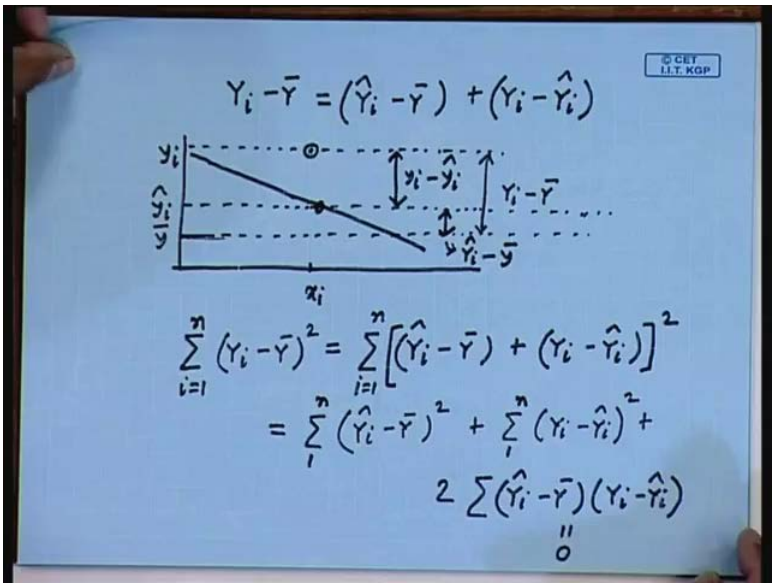


So, here given the set of data x i y I, so I am talking about ANOVA now, the various in response variable, the total variation data. So, the variation data, I mean the variance in responsible variable, which is equal to summation y i minus y bar square, i equal to 1 to n.

So, this y i minus y bar, is basically the deviation of the i th observation from the total, from the over all mean. Well, now the question, this is the total variance in responsible variable. Now, how much of this variability is explained by the regression variable and how much of this variation is left on explained; well, the question is how much of the variation is explained by the model or regressive variable. Now, constant identity Y i minus Y i hat, which is equal to Yi minus Y bar minus Y i hat minus Y bar.

So, this one is basically the deviation of i th observation, from it is predicted value and this quantity is deviation of the i th observation, from over all mean, this is the deviation of the i th observation from overall mean. Similarly, this quantity is you know, this is the deviation of the i th predicted value, from the overall mean.

(Refer Slide Time: 10:49)



Now, this identity can be written as Y i minus Y bar, which is equal to Y i hat minus Y bar plus Y i minus Y i hat. So, the basically the significance of this identity is that, this is the deviation of the i th observation from overall mean. And this is the deviation of the i th observation i th fitted observation, fitted value of i th observation from overall mean and this is the residual basically, this is e i. So, how much of the variation, I mean, I am talking about the i th observation, so how much deviation of i th observation from over all mean explained by the model and this is the portion, which is remain unexplained.

Now, let me just draw figures, suppose given set of observation x i y i, I have the fitted model, this is my fitted model and my i th observation is here, this is my i th observation. So, basically this is x i and this height is y i, now this is y i, this is y i hat, because this point is basically x i y i hat and suppose, the overall mean of the responsible variable or of the data O i is Y bar, which is this well, now you see that, this is the distance is Y i minus Y bar, so this is the deviation of i th observation from the overall mean.

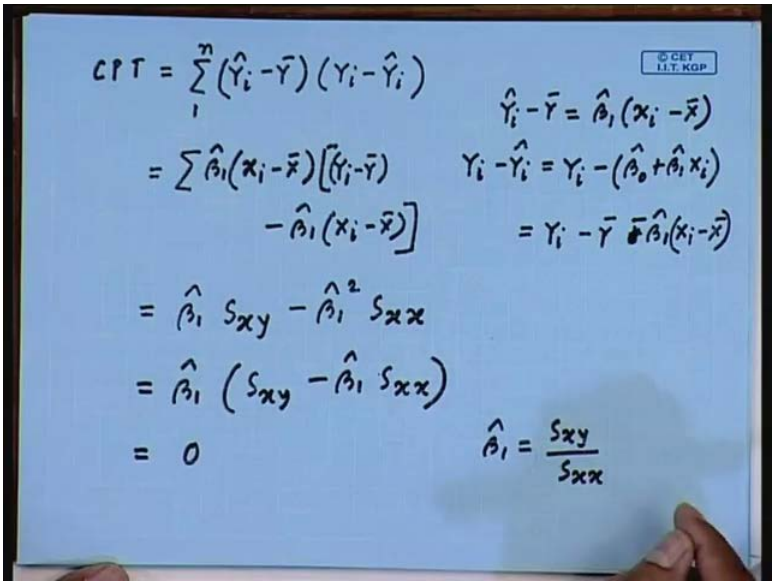
Now part of the deviation is explained by the regression model and this distance basically, this distance is basically, it is y i hat minus y bar and this portion is y i minus y i hat well. So, the total deviation is this much and part of this deviation is explained by the regression model and remaining portion is the unexplained fault. Now, if we, if the square both side of this equation and sum from 1 to n, then we get summation Y i minus

Y bar whole square is equal to Y i hat minus Y bar plus Y i minus Y i hat square from i equal to 1 to n i equal to 1 to n.

This is basically, the variation in the responsible variable or variance in the data, now you want split this variation into several parts, basically 2 parts the part, which is explain by the regression variable and the part the variance, which is not explained by the by the regresses variable. So, you want to the part, which is not explained by the regression variable is basically S S residual and we want to minimize the part, you want minimize is S S residual.

You want the model to this as that, it can explained the variation in observation, I mean most of the part is explained by the model that is what, we want well. This is equal to summation Y i hat minus Y bar whole square plus summation Y i minus Y i hat, whole square plus 2 times 1 to n 1 to n plus 2 times Y i hat minus Y bar into Y i minus Y bar right and I am going to prove that, this cross product on is equal to 0 well.

(Refer Slide Time: 17:52)

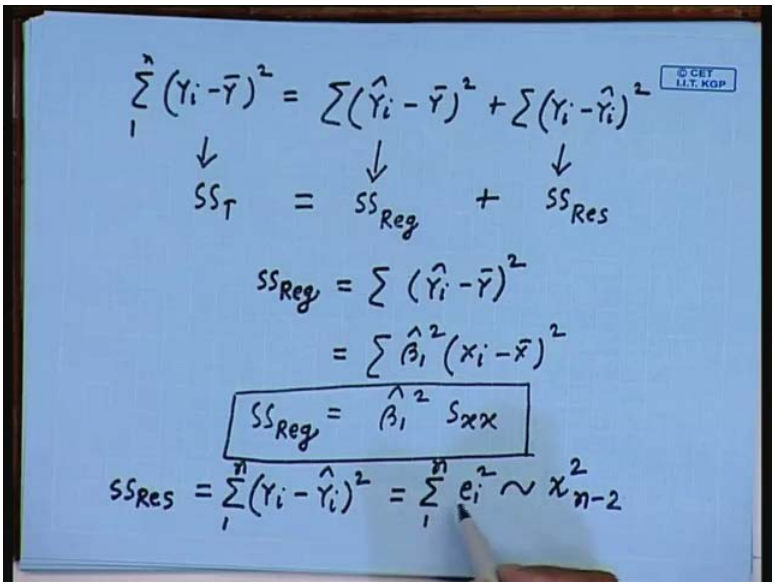


So, cross product on T is equal T summations Y i hat minus Y bar, now we can check that, Y i hat minus Y bar this is nothing but, beta 1 hat into x i minus x bar and similarly Y i minus Y i hat is basically, Y i minus beta naught hat plus beta 1 hat x i. So, just replacing Y i hat by the fitted value and now replace beta naught that is basically Y i minus Y bar plus beta 1 hat x bar, so I can write this is equal to x i minus x bar. So, if I

now replace 2 quantity here, i equal 1 to n, I will get some over beta 1 hat x i minus x bar into Y i minus Y bar.

So, here it is minus, minus beta 1 hat x i minus x bar, basically it is beta 1 hat S x y minus beta 1 hat square is S x x, this is the notation for this summation, which is equal to beta 1 hat S x y minus beta 1 hat is S x x. And this quantity is equal to 0, this is equal to 0, because we know that, beta 1 hat is equal to is x y by is S x x, so what we proves that the cross product on is equal to 0.

(Refer Slide Time: 21:56)



So, we are left to it, then summation Y i minus Y bar whole square is equal to summation Y i hat minus Y bar whole square plus summation Yi minus Y i hat whole square. So, this quantity is denoted by S S T, so basically, it is a total sum of square and this quantity is denoted by S S regression; that means, sum of square due to regression and this is the portion, this called S S residual. So, what we have is that, S S total is equal to S S regression plus S S residual, while this is the splitting of total sum of square into 2 parts.

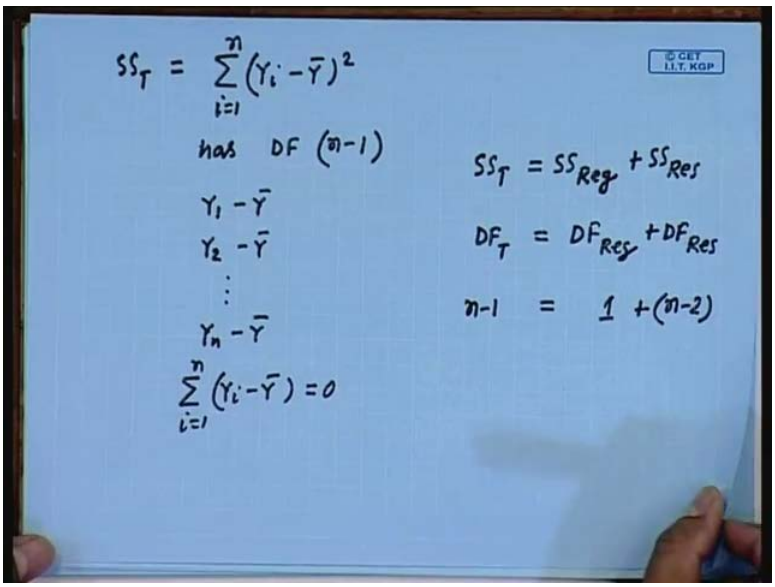
The total variation in Y is spitted into 2 parts, the variation the regression and the variation residual sum of square that means, variance which is not explained by the regression variable well. Now, we have been proved that, now what is S S residual S S sorry, S S regression is equal to summation Y i hat minus y bar whole square and this

quantity is nothing but, just now I mean, we have proved that, this is this quantity is equal to beta 1 hat X i minus X bar.

So, square here square here, summation and this is going to be beta 1 hat square S x x, so S S residual sorry, S S regression is this quantity, now S S residual, which is equal to summation Y i minus Y i hat whole square 1 to n, this is nothing but, the i th residual. So, i can write this own as summation e i square 1 to n and you have proved that, this quantity is I mean, this follows chi square distribution with degree of freedom is not aim, it is the degree of freedom is n minus 2, because you know all the e i is there not independent.

We know that residual stay satisfy the constant that summation e i is equal to 0 and the constant is summation e i X i is equal to 0. So, because of the due to this 2 constant all the e i is not independent, you know out of n e i is you can choose n minus 2, e i is independently and the remaining 2 have to be choose in such a way that, they satisfy this constant. So, be losing the 2 degree of freedom, because of these 2 constant, because e i satisfies to constant that is why, S S residual here, it follow chi square in minus anyway, I have proved this thing for also.

(Refer Slide Time: 27:21)



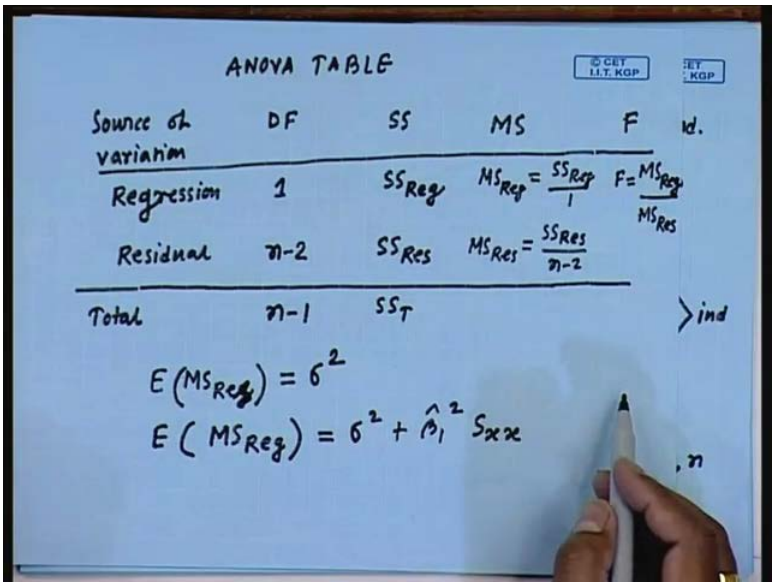
Now, S S T some of square, due to total sum of square, so S S T is equal to sum Y i minus Y bar whole square i equal to 1 to n and this has the degree of freedom n minus 1, because of the fact that, you know you have Y you on minus Y bar Y 2 minus Y bar and

Y n minus Y bar. And this satisfy the condition that, of course this always show that, some Y i minus Y bar is equal to 0, i equal to 1 to n. So, for this reason, you know out of in quantities n minus 1 can be choose in independently and the n th 1 has to be choose in such a way that, this constant is satisfied.

So, that is why, total sum of square has n minus 1 degree of freedom and also you know, you came to another S S T is equal to S S regression plus S S residual, the variation, which as been explained by model. And this is the variation, which as not explained by the regression variable, if this as degree of freedom and also degree of freedom as the addict properties, so degree of freedom total is equal to degree of freedom of regression plus degree of freedom of residual.

So, this quantity is n minus 1 and we know that, S S residual as degree of freedom n minus 2 and then, the degree of freedom of S S regression is equal to 1.

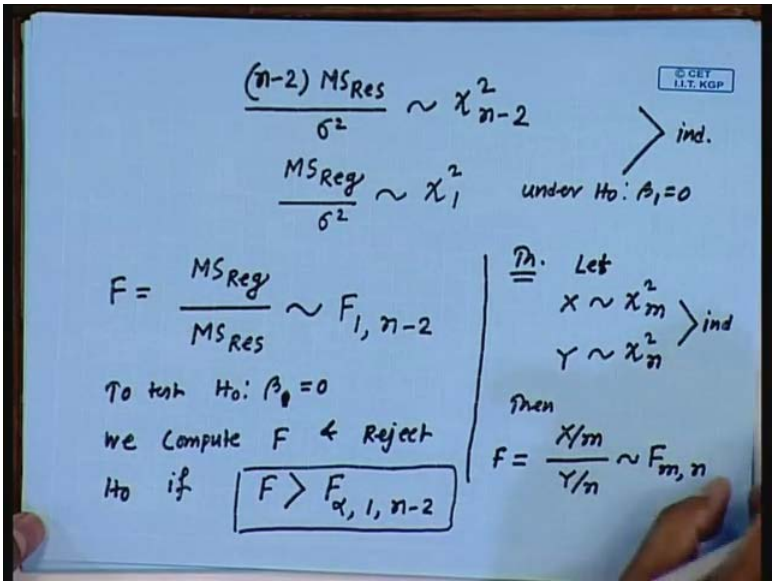
(Refer Slide Time: 30:20)



Now we make the ANOVA table, will source of variation degree of freedom sum of square well, let me write here, total sum of square and the source of variations are regression and the residual. So, this has degree of freedom 1, the residual has degree of freedom n minus 2 and this has degree of freedom n minus 1, this one denoted by we know, what is this quantity, this denoted by S S regression, this is called S S residual and this is called S S T.

Now M S mean square, which is obtained by dividing the S S by degree of freedom, so here it is M S regression is equal to S S regression by 1 and similarly, M S residual is equal to S S residual by degree of freedom that is n minus 2 yes. Now, we already know that expected value of M S regression is equal to sigma square, this we have proved before and it can be proved that, expected value of sorry, this is M S residual. It can be proved that M S regression is equal to sigma square plus beta 1 hat square S xx right.

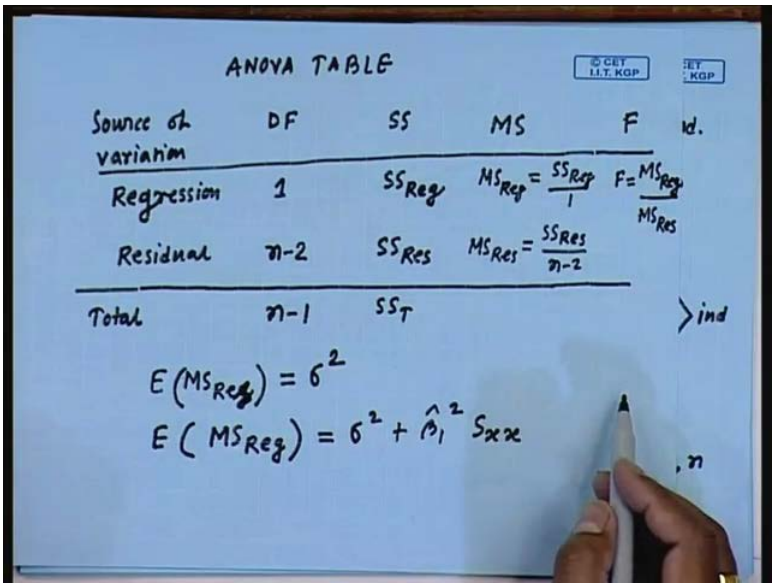
(Refer Slide Time: 33:56)



And also, we know that n minus 2 M S residual by sigma square, this follows chi square with degree of freedom n minus 2 and it can be proved that, M S regression by sigma square, this follows, chi square with degree of freedom 1 that is under H naught, there is beta equal to 0. And these 2 quantities are this 2, these are basically this is function of random variables Y i Y, so and they are independent.

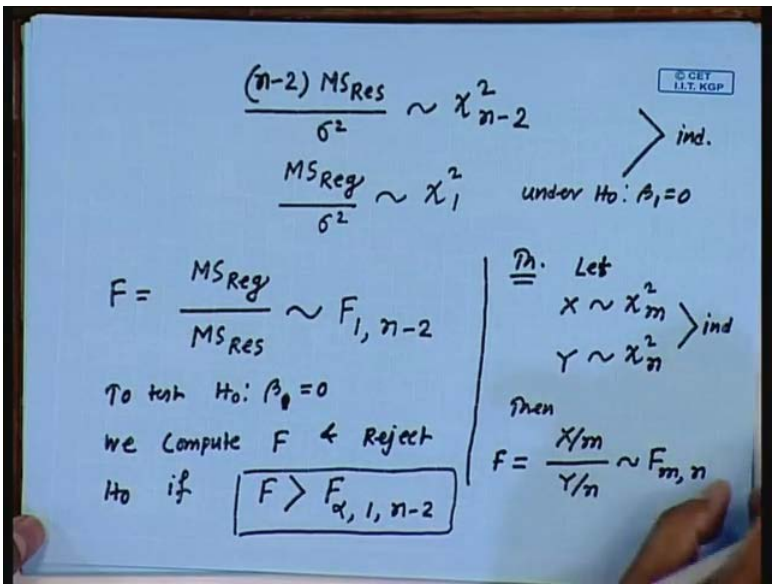
Now, a statistical theorem, we have theorem says that, let X follows chi square m and Y follows chi square n and they are independent, then X by m by Y by n, this follows F with degree of freedom m and n, this quantity is divided by this. Basically, the residual of 2, chi square they follow, if distribution from there, from this theorem, we can say that, we can defined now, for 1 F for ANOVA table, it F is equal to M S regression by M S residual, this follows F with degree of freedom 1 n minus 2.

(Refer Slide Time: 36:54)



And obviously, looking at this 2 expected value, so what we going to do is that ANOVA table next, we are going to compute, F this F is equal to M S regression by M S residual and looking at their expected value, it is initially clear that. If F is large, then it likely that, this beta 1 is not equal to 0, if beta 1 is equal to 0, then this residual going to be close to 1.

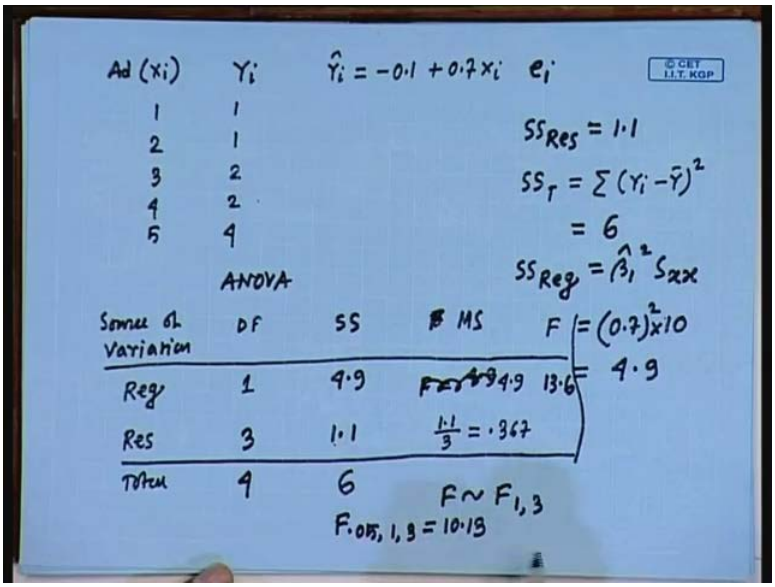
(Refer Slide Time: 38:04)



So, from there, we can say that to test the hypothesis H naught beta 1 equal to 0, we compute F and reject not, if F is greater than, F alpha with degree of freedom 1 n minus

2. This is another way to another approach to test the hypothesis is beta 1 equal to 0, I mean this the same test, we can do using that, t distribution also basically, those 2 or same, I am going to proved that 1. First let me a give a 1 example for this ANOVA table well, I am going to constant the same example.

(Refer Slide Time: 39:37)



Cost on advertisement that is X i and this is the sales amount Y i and we have the data 1 1 2 1 3 2 4 2 and 5 4 and we know the fitted model is Y i hat is equal to minus 0.1 plus 0.7 X i. And found here, we can you know before also, we compute it, e i is and we know that, S S residual for this problem is for this data is equal to 1.1 and what you need to compute is that, we need to compute. What is the total variation in the data S S T, which is equal to summation Y i minus Y bar whole square, you can take that Y bar is equal to 5 here and it is not difficult to check that, this is equal to 6.

And also we know that, S S regression, which is equal to beta 1 hat squares S x x, we know that beta 1 is 0.7. So, this is equal to 0.7 whole square and 1 can check that S x x, basically we call to 10. So, this quantity is equal to 4.9 and here is my ANOVA table for this problem, ANOVA the source of variation, this is regression residual total degree of freedom, for this 1 is equal to 1 well, the total degree of freedom here is n minus 1 and n is equal to 5.

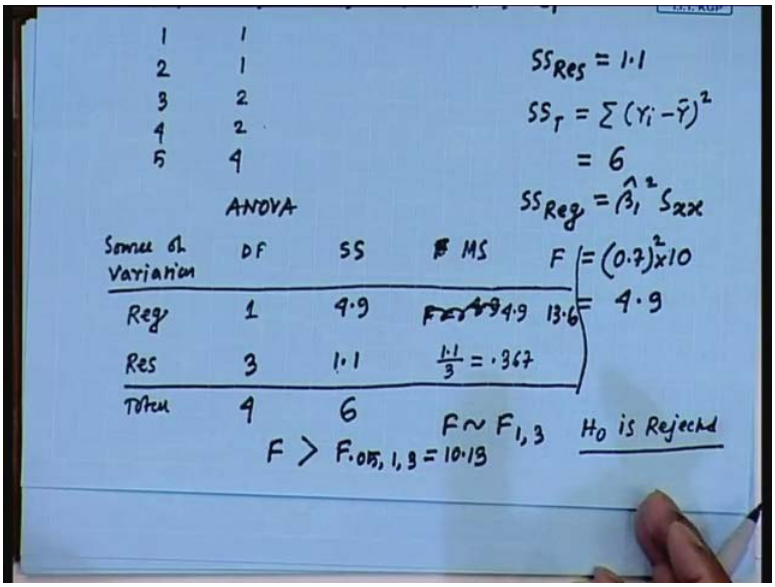
So, total degree of freedom is 4 and the degree of freedom residual is n minus 2 that is equal to 3 and hence the degree of freedom, for the regression is equal to 1 and the S S

values are 4.9 1.1 and the total variation is equal to 6. So, here you can see that, for this problem the repeated model is really good, because the total variation in Y is 6 and most of the part of this variation has been explained by the regression.

So, out of 64.9 as been this is the part of the variation, which has been explained by the regression model, so most of the part has been explained the regression model and the portion, which is not explained by the regression model is 1.1 well. So, F value here is equal to 4.9 sorry, sorry, first step we need to compute M S value, the M S value is S S by degree of freedom. So, this is going to be 4.9 and M S residual 1.1 by 3, which is going to be 0.367 and the F value, is basically M S regression by M S residual, so 4.9 by 0.367, which is going to be 13.6.

And now you know, this F follows, this F here that follows, F distribution with degree of freedom 1 3. And you find the value of F 0.0513, which is going to be equal to 10 point 13, so you can take this value from the statistical table and now you check, you can see that our computed F value, which is 13.6 is larger than the tabulated value.

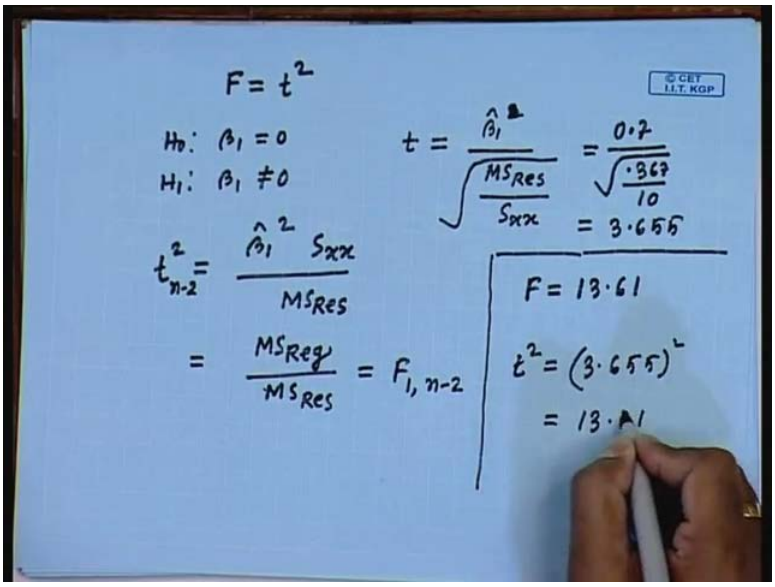
(Refer Slide Time: 45:48)



So, we can conclude that, which is I mean, this computed value is larger than F 0.051 with degree of freedom 1 and 3. So, we conclude that, we can reject H naught is rejected at the level of significance of this test is equal to alpha is equal to 0.05, so basically we have got the same result, using the t test. Now, I am going to prove that, this 2 test are

basically, you know this 2 whether you using F test or t test is does not mater in the case of simple linear regression, basically this 2 test are same well.

(Refer Slide Time: 47:04)



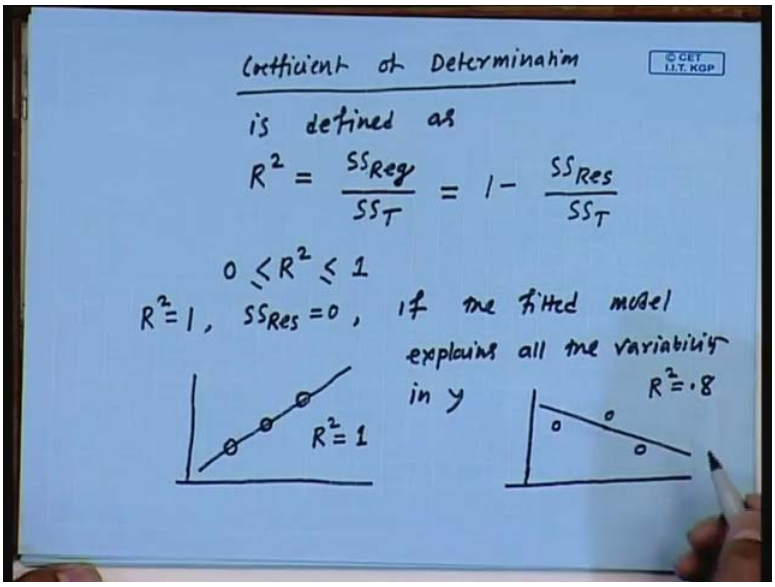
Let me prove that, F is nothing but, t square, I mean t square valuable well, so to test this hypothesis beta 1 is equal to 0, against the alternative the hypothesis beta 1 not equal to 0. Either you can go for t test, which is that t is equal to beta 1 hat square sorry, beta 1 hat by M S residual by S x x, this is done t statistic to test, this hypothesis, now we compute t square, which is going to be beta 1 hat square into S x x by M S residual right. And this quantity is well, this quantity is nothing but, M S regression and the trigonometry M S residual, which is nothing but, the F test.

So, the value of the distribution with degree of freedom 1 n minus 2 is same as the value of that distribution of the degree n minus 2, now just for you can check this 1, you know previous example, what we got is that for testing this hypothesis, we got F equal to 13.61. Now, if you go for the t statistic to test this hypothesis, we have this is beta 1 is equal to beta 1 add is equal to 0.07 and M S residual is 0.367 by S x x is equal t 10.

And this is going to be equal to 3.655 and you can check that, this 3, I mean that t square, which is basically equal to 3.655 is equal to 13.61, which is equal to F. So, whether you use the t statistic to test, this hypothesis or this is the F ANOVA approach that is the F statistic to test this hypothesis, there are basically same. And basically same for simply linear regression model once, we talking about multiple linear regression than, we need

to follow the ANOVA approach only. So, next I will be talking about the coefficient of determination well.

(Refer Slide Time: 51:02)



So, what is this that, this is denoted by is this is defined by defined as at square, which is basically ratio of S S regression by S S total well, you know several approach, to evaluate the performance of fitted model. So, this is 1 parameter, which can be, which is used to evaluate the performance of the repeated model, here this quantity is nothing but, 1 minus S S residual by S S T.

And you know, that is S S T is equal to S S regression plus S S residual, so that is why and obviously, the range for R square is going to be from 0 2 1, it can be at most 1, if the S S residual equal to 0. So, this R square, it is basically determines that, the proportion of variability that has been explained by the by regression model, because R square is equal to S S regression by S S total, so residual give you the proportion of variability that has been, you know explained by the regression model well.

Let me just give, this is the very important parameter, when R square is going to be equal to 1, so R square is going to be equal to 1, if S S residual equal to 0, so S S residual is equal to 0. And this will happened, if the predict model explained, this will happen, if the fitted model explains all the variability in Y right, that means, there is no parts, which remain unexplained by the model. And the example of this one is this is the case basically, suppose you have the data like this then, your fitted model is going to be this.

So here R square is going to be equal to 1 and this case, it is entity clear that S S residual is going to be equal to 0, while if suppose in some example R square is equal to 0.8 may be, this is the case. These are the 3 data point and you fitted this model and entity ability, you know is the R square is going to be very high.

And what we say here is that, what is the meaning of R square is equal to 0.8 is that, approximately 80 percent of the total variability in Y that has been explained by the regression model and then remaining 20 percent of the variability in Y remain unexplained by the model. So, the higher the value of the regress R square value, the better model is, so that is all for today and next class will be talking about, talking more about R square.

Thank you very much.