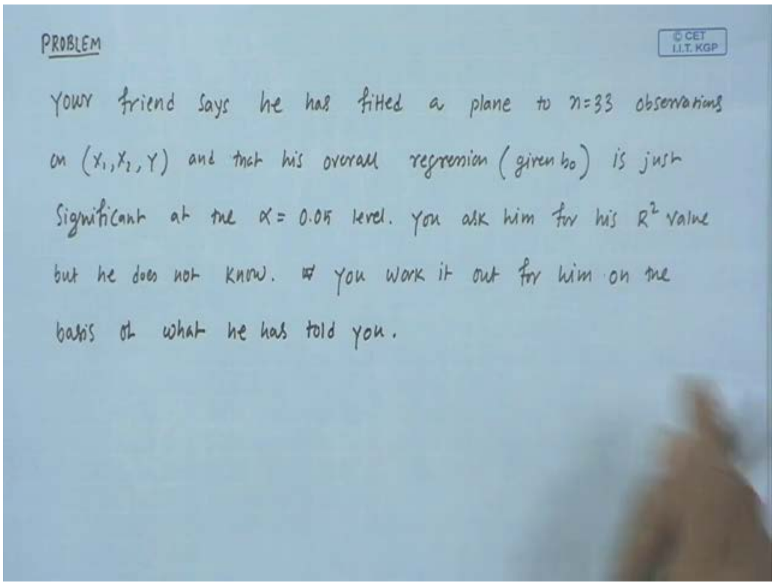
Regression Analysis
Prof. Soumen Maity
Department of Mathematics
Indian institute of Technology, Kharagpur

Lecture - 39
Tutorial IV

Hi, so this is my fourth tutorial and today we will be solving some problems involving coefficient of determination, and also you know module fitting with auto correlated, auto correlated errors.

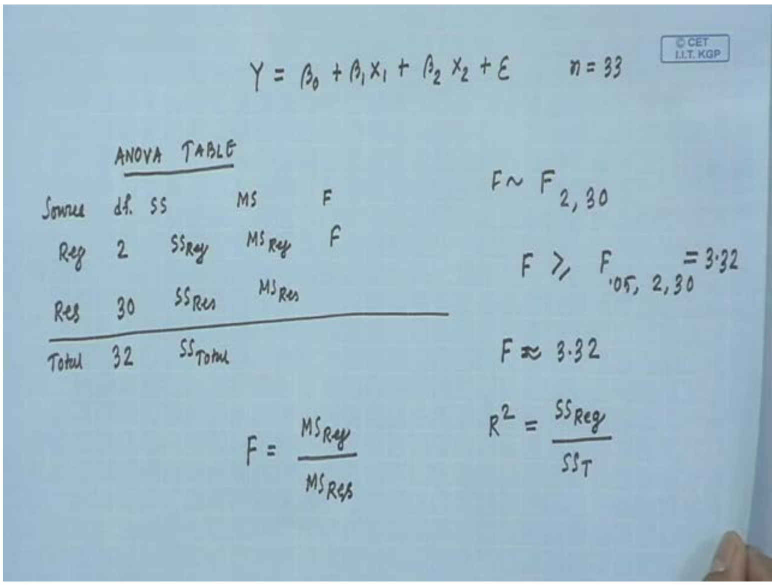
(Refer Slide Time: 00:41)



So, here is the first problem, the problem says that your friend says he has fitted a plane to 33 observations on x 1, x 2 and y and his overall regression is just significant at 0.05 level of significance. That means your friend has fitted a multiple linear regression model with 2 regressions x 1 and x 2 and 1 response variable and his test is significant that means the fitted model is significant at 5 percent level of significance. Now, you ask him for his R square value that is the coefficient of determination, but he does not know, you work out for him on the basis of what he has told you.

So, the information given to you is that he has fitted multiple linear regression models with two regressions x 1 and x 2 and his feet his significant at 0.05 levels, but he does not know the R square value where R square is the coefficient of determination. It is sort of measure the proportion of variability that is explained by the model. Now, you have to work out this R square value from whatever he has told to you.

(Refer Slide Time: 02:53)

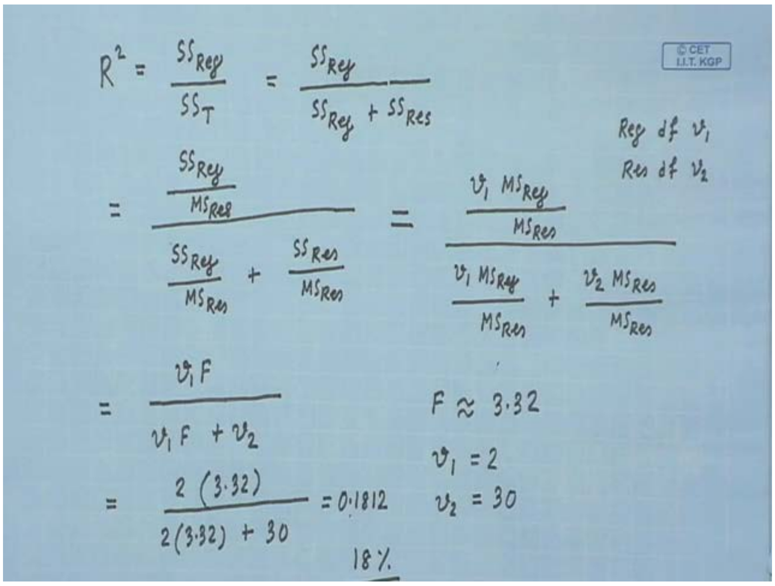


So, what you know is that your friend has fitted a model like this y equal to beta not plus beta1 x 1 plus beta 2 x 2 plus epsilon and this model has been fitted for n equal to 33 observations. So, from this information what you can do is you can construct ANOVAs table for this, here is the ANOVAs table having source sum of square S S, degree of freedom and M S and the F statistic. So, the total degree of freedom is 32 because there are already 33 observations and then the regression degree of freedom is 2 because there are 3 parameters.

The residual degree of freedom is 30 and, of course your friend has you know S S regression, S S residual and S S total and hence the M S regression, M S residual and F value. So, what you know is that this F is significant, so this F follows F distribution with degree of freedom 2, 30 and his test is just significant. So, that means the observed F value is just greater than or equal to the F value 0.05 at the level with degree of freedom 2, 30 which is equal to 3.32.

So, you can assume that the observed F value is close to 3.32 and from, here you have to compute the R square value, the R square value is S S regression by S S total which is the proportion of variability in Y about mean that is explained by the model. But, what you know that F which is equal to M S regression by M S residual, so you know the F value their degree of freedom. But, of course you do not know separately the M S regression value and M S residual value from F you have to compute R square. So, this is the problem we will see whether this R square can be written in terms of F.

(Refer Slide Time: 07:27)



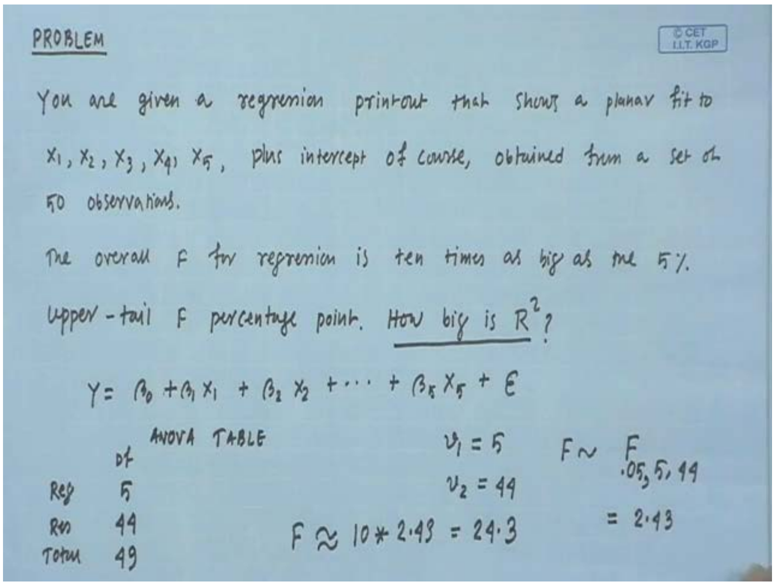
So, that the R square which is equal to S S regression by S S total that I can write as S S regression and S S total is equal to S S regression plus S S residual. I want to express this one in terms of F, so I can write this as S S regression by M S residual, so I will divide both the numerator and the denominator by M S residual, So, this is S S regression by M S residual plus S S residual by M S residual, so this one is equal to v 1 M S regression by M S residual, where v 1 is the regression degree of freedom.

So, regression degree of freedom is v 1 and similarly, here also I can write that as v 1 M S regression by M S residual plus v 2 M S residual by M S residual, where residual has degree of freedom v 2. So, now this one is equal to M S regression by M S residual which is nothing but F, so this can be written as v 1 F by v 1 F plus v2. Now, we can compute R square, so this is how you know we can express R square in terms of F. So, we know that F is close to 3.32, and also we know that our v 1 is the regression degree of freedom that is equal to 2 in our case and v 2, which is the residual degree of freedom that is equal to 30.

So, R square can be, now written as 2 in to 3.32 by 2 in to 3.32 plus 30 this is equal to 0.1812 which means you know what R square is. The R square is the proportion of variability in Y about mean that is explained by the model, so you can see, here only 18 percent of total variability has been explained by the model.

So, R square is this means only 18 percent of total variability has been explained by the model, so you can see that from this example R square is very good parameter to measure how good feet is the test is significant. So, the module is significant according to the global app test, but only 18 percent of the total variability has been explained by the model this is quite low, so we will consider another problem of similar type.

(Refer Slide Time: 13:05)

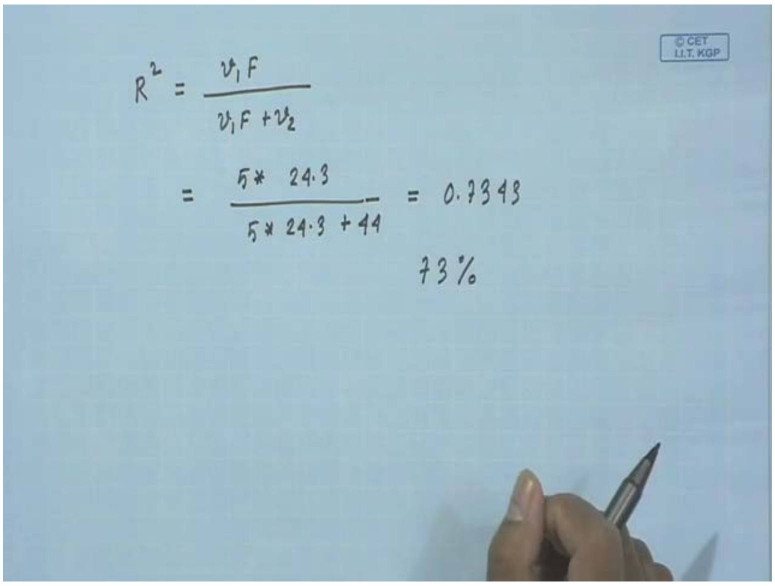


So, this problem says that you are given a regression printout that shows a planar fit to x 1, x 2, x 3, x 4 and x 5 plus intercept of course obtained from a set of 50 observations. The overall F for regression is 10 times as big as the 5 percent upper tail F percentage point.

So, you have to compute how big R square is, so here you are concentrating on a module involving 5 regressions, so Y equal to beta nod plus beta1 x 1 plus beta 2 x 2 plus beta 5 x 5 plus epsilon and this module is fitted on 50 observations. So, we can quickly have ANOVAs table, so that the total degree of freedom is, of course 49 and regression degree of freedom is 5 as there are total 6 parameters and the residual degree of freedom is then 44. So, what you know is that we know my v 1 is equal to 5, here v 2 is equal to 44 and it says that the observed F value is 10 times the tabulated F value.

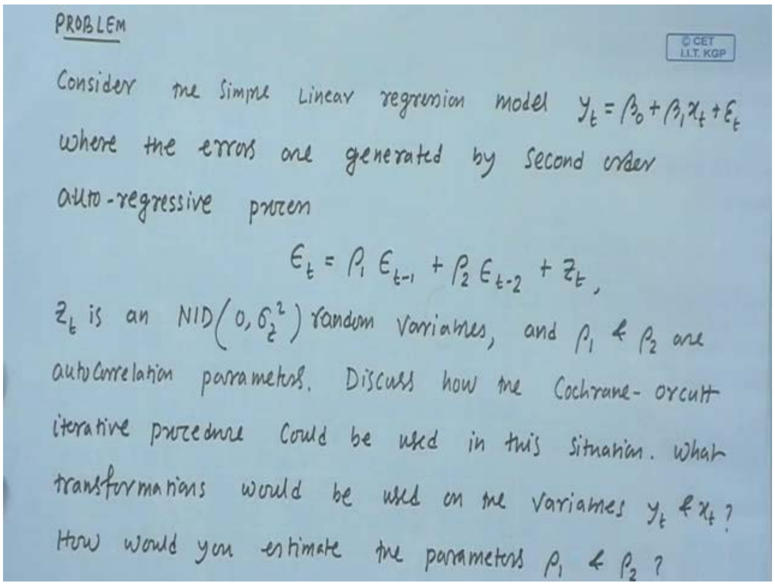
So, for this test, here F has degree of freedom F 5, 44 and you have to find the tabulated value for this one at 0.05, and you can check that this value is equal to 2.43. So, what were given is that your observed F value is 10 times bigger than this one, so your observed F value is then equal to 10 times of this tabulated value that is 2.43, which is equal to 24.3 and you know v 1, v 2 you know F. So, you can compute R square, now problem is that how big is R square.

(Refer Slide Time: 17:02)



So, you know the formula that R square is equal to, R square is equal to v 1 F by v 1 F plus v 2. So, here your v 1 is 5 and F is 24.3, so 5 into 24.3 plus v 2 that is the residual degree of freedom that is 44. So, here 0.7343 that means 73 percent of the total variability in the response variable has been explained by the fitted module which is quite good and next we will be concentrating one problem from regression models with auto correlated errors.

(Refer Slide Time: 18:17)

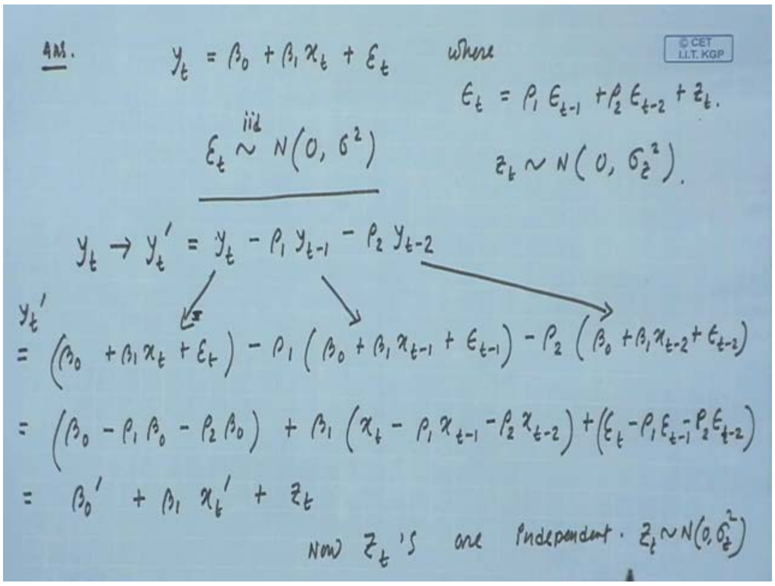


So, here is the problem, consider the simple linear regression model y t equal to beta nod plus beta 1 x t plus epsilon t where the errors are generated by second order auto regressive process. So, I hope that you can recall, so here you can see this observations are collected sequentially in time. So, they are the time series data basically that is why it is denoted by t, here and in case of time series data we know that this epsilon t, the error term they are not independent they are basically correlated. Here, it is given that the errors are, second order errors are having second order auto regressive relation, so epsilon t is equal to rho 1 epsilon t minus 1 plus rho 2 epsilon t minus 2 plus z t.

So, this z t is independent with mean 0 and variance sigma z square, here rho 1 and rho 2 are called auto correlation parameters. So, the problem is you know you have to discuss how Cochrane Orcutt iterative process could be used in this situation if you can recall we talked about how to feed the regression parameters beta nod and beta 1 in case of first order auto regressive error. So, here instead of first order auto regressive error we have second order auto regressive error.

So, this is quite straight forward problem, so it says that what transformation would be used on the variables y t and x t and how would you estimate the parameters rho 1 and rho 2. Well, so this is the problem how do you estimate this parameter beta nod and beta 1 had beta 1 and beta nod in the presence of second order auto regressive error.

(Refer Slide Time: 21:13)



So, let me start the solution for this problem, so we need to fit this model y t is equal to beta nod plus beta 1 x t plus epsilon t where this epsilon t is equal to rho 1 this is second order auto regressive epsilon t minus 1 plus rho 2 epsilon t minus 2 plus z t.

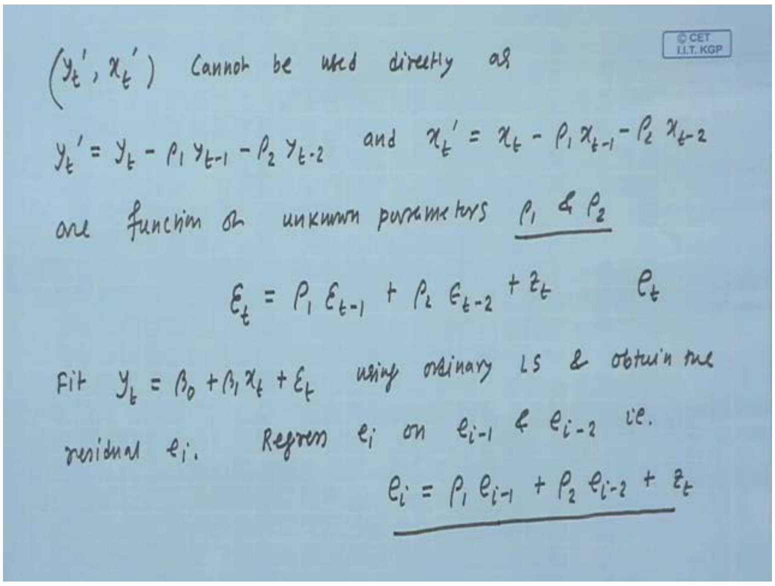
This z t follows normal distribution with mean 0 and variance sigma z square, so usually you know we need if it is a simple regressive model we assume that epsilon t follows normal distribution with 0 mean and constant variance. If it ended also independent if this is true then we can estimate beta nod and beta 1 using the ordinary least square technique. But, here it is given that this error term they are not independent they are correlated and they follow the second order auto regressive process. So, in this situation how to fit this model, so what you do is that we transform this response variable y t to y t dash which is equal to y t minus rho 1 y t minus 1 minus rho 2 y t minus 2.

So, this is equal to so this is this is equal to, so y t prime is equal to beta nod plus beta 1 x t plus epsilon t minus rho 1 beta nod plus beta 1 x t minus 1 plus epsilon t minus 1 minus rho 2 beta nod plus beta 1 x t minus 2 plus epsilon t minus 2. So, you understood, so this is my y t, this is rho 1 y t minus 1 and then rho 2 y t minus 2 right. So, this can be written as beta mod minus rho 1 beta nod minus rho 2 beta nod plus beta 1 x t minus rho 1 x t minus 1 minus rho 2 x t minus 2 plus epsilon t minus rho 1 epsilon t minus 1 minus rho 2 epsilon t minus 2 right epsilon t minus 2 and this one is equal to z t. So this term is equal to z t, so I can rewrite this one as say beta nod dash plus beta 1 call this as x t prime.

So, what you do is you transform y t to y t prime and, similarly x t to x t prime, so this is my x t prime plus z t. Now, in this transform module we have transformed y t to y t prime where y t prime is given, here similarly we have transformed x t to x t prime and as a consequence the error term has been transformed to z t this z t is, now independent. Now, z t are independent, so and also you know that z t follows z t follows normal 0 sigma z square, so the advantage of this one is that.

Now, we have the model like y t prime equal to beta nod dash plus beta 1 x t prime plus z t and, here the error term is error terms follow normal distribution with mean 0 and variance sigma square and they are also independent. Now, you are in a position to apply your ordinary least square technique to estimate the regression coefficients beta nod prime and beta 1, but the problem is that you know this y t.

(Refer Slide Time: 27:23)



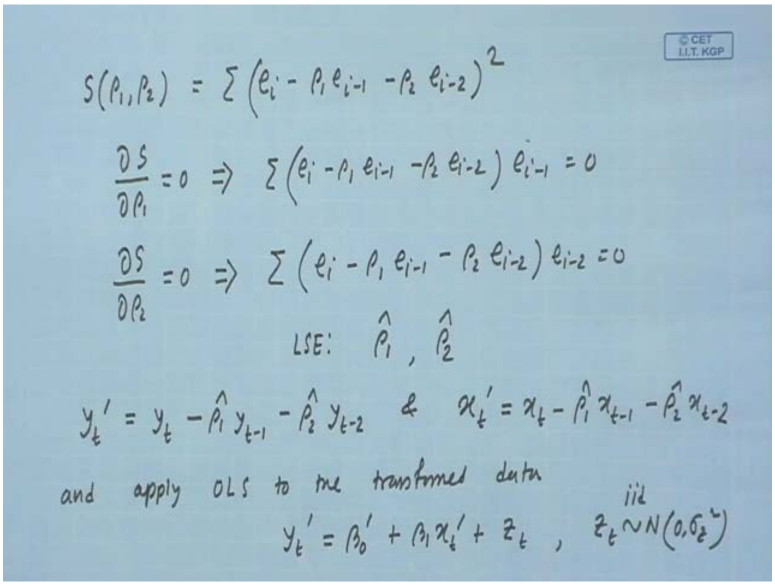
You see, here this y t prime that involves rho 1 and rho 2 similarly, x t prime involves rho 1 and rho 2, so we cannot use this transformation unless we know the value of rho 1 and rho 2. So, y t prime and x t prime cannot be used directly as y t prime which is equal to y t minus rho 1, y t minus 1 minus rho 2 y t minus 2 and x t prime which is equal to x t minus rho 1, x t minus 1 minus rho 2 x t minus 2 are function of unknown parameters rho 1 and rho 2.

So, unless you know see you are given the data x t and y t you do not know rho 1 rho 2 unless you do not, unless you know the value of rho 1 rho 2 how do you use this transformation. So, we need to estimate them and we know that these are that auto correlation parameter what is this rho 1 and rho 2 are auto correlation parameter. For this second order auto regressive process rho 1 epsilon t minus 1 plus rho 2 epsilon t minus 2 plus z t, so we need to estimate this rho 1 and rho 2.

So, how to do that you know that the e t the observed value of e t epsilon t are t h residual e t, so what you do is that we fit this model y t equal to beta nod plus beta 1 x t plus epsilon t using ordinary least square technique and obtain the residual e i. So, we will fit this model we do not ignoring the fact that they are auto correlated, so you fit this model and once you have fitted the model you can compute the residuals.

Then you regress e i on e i minus 1 and e i minus 2 that is you fit model like e i is equal to rho 1 e i minus 1 plus rho 2 e i minus 2 plus some error z t. So, you know epsilon i, so you can fit this model, so this is nothing but multiple linear regression model with two regressions, so how do you estimate the parameter rho 1 and rho 2.

(Refer Slide Time: 31:34)



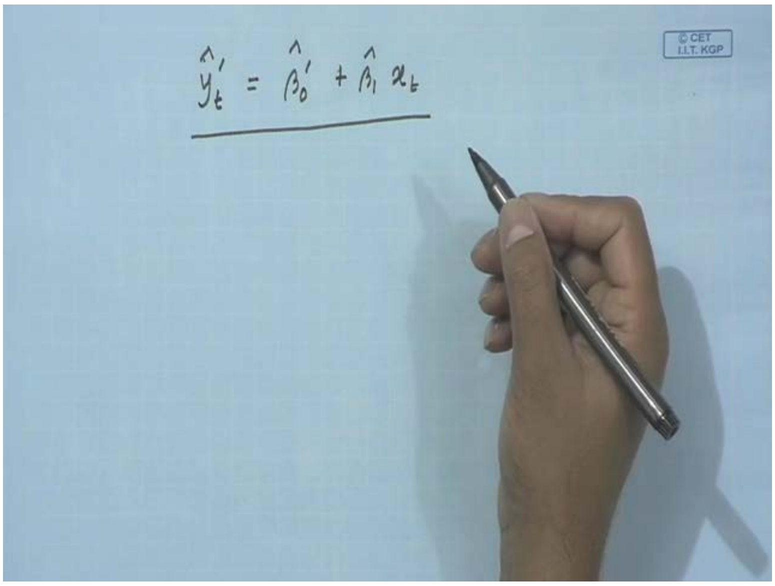
You compute the least square function S rho 1 rho 2 which is equal to e i minus rho 1 e i minus 1 minus rho 2 e i minus 2 and you minimize this. So, you estimate rho 1 and rho 2 in such a way that this least square function is minimized and what you do is that you just differentiate this S with respect to rho 1 that equal to 2 will give you one normal equation.

Similarly, S with respect to rho 2 is equal to 0 this will give you another normal equations let me write down those things. So, e i minus rho 1 e i minus 1 rho 2 e i minus 2 in to e minus 1 this is equal to 0 this is the first normal equation and the second normal equation is e i minus rho 1 e i minus 1 minus rho 2 e i minus 2 in to e i minus 2 equal to 0.

So, you have two normal equations and two unknowns, so you can estimate you can find rho 1 and rho 2 call them rho 1 hat and rho 2 hat, these are the least squared estimate of rho 1 and rho 2. So, once you have these estimated values, now you can use this rho 1 hat and rho 2 hat to get the transformed values y t prime is equal to y t minus rho 1 hat y t minus 1 minus rho 2 hat y t minus 2. Similarly, you get x t prime which is equal to x t minus rho 1 hat x t minus 1 minus rho 2 hat x t minus 2, now you can obtain y t prime and x t prime.

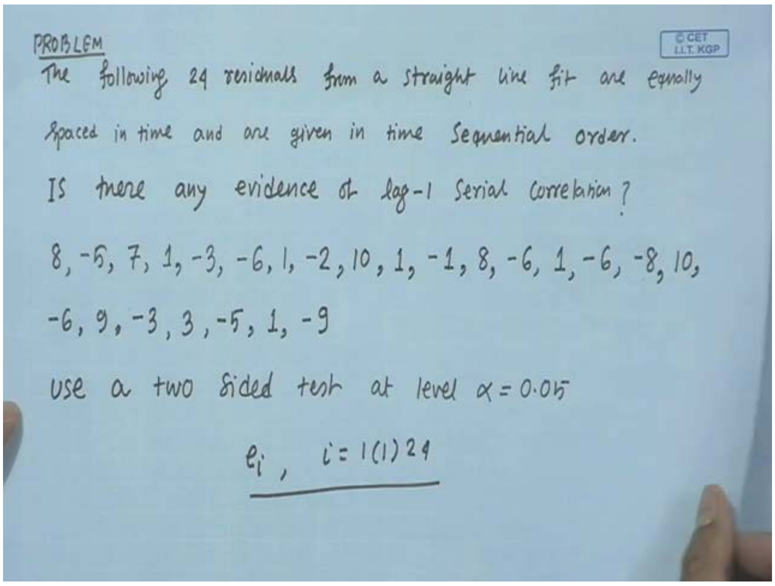
So, these are the transformed data and now you are in position to apply ordinary least square to the transformed data y t prime equal to beta nod prime plus beta 1 x t prime plus z t. Here, for the transformed data you know this z t the error term from transformed data this follows normal 0 sigma z square. So, normal with mean 0 and constant variance sigma z square and they are independent. So, you are in position to apply ordinary least square technique and you apply ordinary least square technique.

(Refer Slide Time: 35:17)



Here, to get the estimated value and the final fitted observe, fitted model is y t prime hat is equal to beta nod prime hat plus beta 1 hat x t. So, this is how you can fit the model in the presence of second order auto regressive error, so this is called Cochrane Orcutt procedure and in a module called regression model with auto correlated errors. We talked about the same technique for auto, for auto regressive errors with first order auto regressive order we solved the same problem.

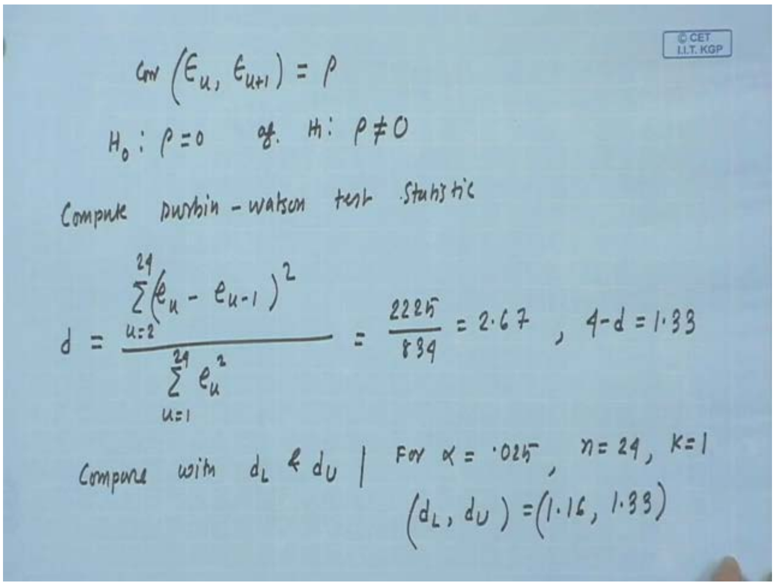
(Refer Slide Time: 36:24)



When there exists first order auto regressive error and just now we solved for second order auto regressive order auto regressive error. So, let me consider one more problem this is also you know to check whether there exists lag 1 auto correlation or serial correlation it says that the following 24 residuals. So, these are the residuals from a straight line fit are equally spaced in time and are given in time sequential order. That means these residuals are obtained from time series data and they are equally spaced, I mean the times are equally spaced.

So, the question is there any evidence of lag 1 serial correlation for these 24 residuals, so it says that you use a two sided test at level alpha equal to 0.05. So, you are given e i for i equal to 1 to 24 and how do you test that there exists lag 1 serial correlation or not we know that for testing lag 1 serial correlation we need to go for Durbin Watson test.

(Refer Slide Time: 38:16)

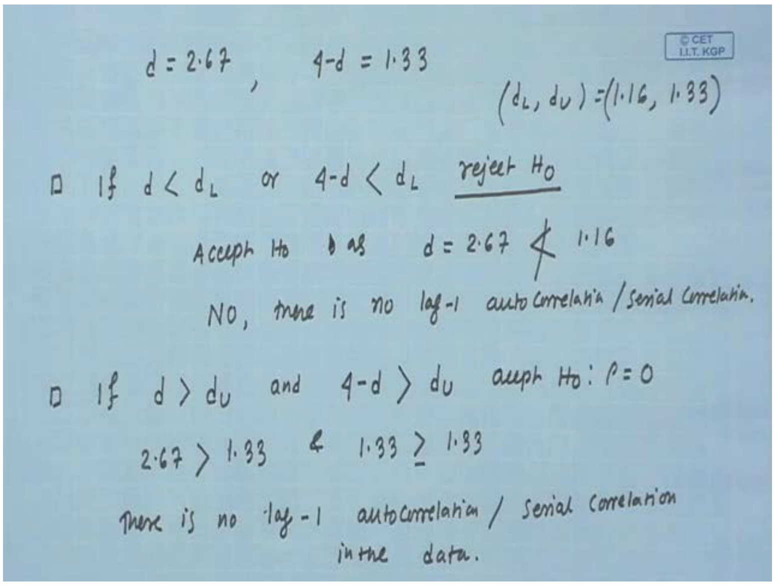


So, what we will do is that, so what is lag 1 correlation is that the correlation between epsilon u and epsilon u plus 1 this correlation is equal to rho which is not equal to 0 if there exists lag 1 auto correlation, so what you have to test is that. So, here you can see the errors are one step apart, so we need to test that h nod rho equal to 0 against the h 1 that rho is not equal to 0. So, h nod says that there is no lag 1 auto correlations in the residuals h nod says that, and h 1 says there exists lag 1 auto correlation.

So, what you do is that we compute Durbin Watson test statistic what is that that is d equal to summered e u minus e u minus 1 square u is from 2 to 24, here by e u square for u equal to 1 to 24 and that you can check that you are given e i S. You are given e u for u equal to 1 to 24 we can compute this one this is 2225 by 834 which is equal to 2.67 well, so my d is 2.67 and we know that this d is d the range of d is from 0 to 4 and it is symmetric about 2. We compute 4 minus d also 4 minus d is equal to 1.33 and what you do is that for testing this two sided alternative, we compare this d value compare with d L and d U and this value will get from the d table.

So, for alpha equal to 0.025 because it is two sided test that is why I am taking alpha equal to 0.25, n equal to 24 and k equal to 1 because it is a straight line fit, so there is only one regression in the module that is why k is equal to 1. You can check that your d your d L and d U is equal to 1.16 and 1.33, now we need to think about the critical values for this one.

(Refer Slide Time: 41:49)



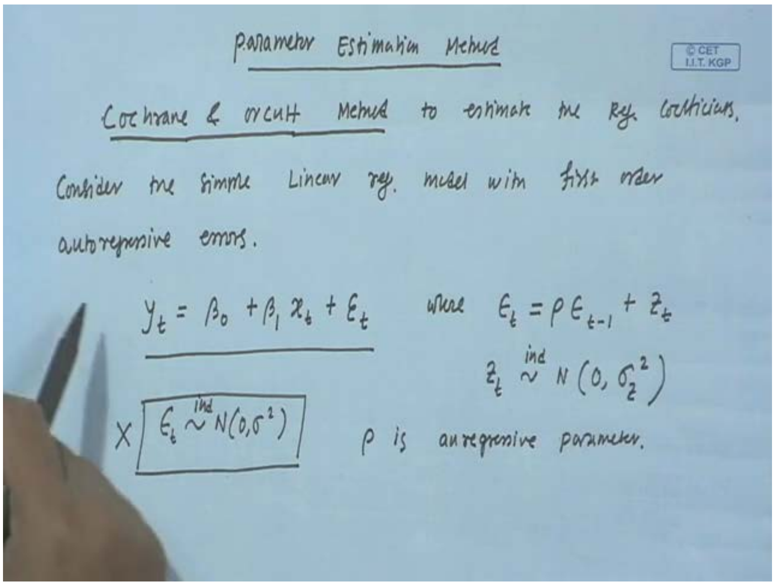
So, what we have is that we have we know that my d is equal to 2.67 my 4 minus d is equal to 1.33 and my d l and d U is equal to 1.16, 1.33, so if d is less than d L or 4 minus d is less than d L you reject h nod. That means if the Durbin Watson test statistic is small then you reject h nod, rejecting h nod means there is no auto correlation, let us see what is this d value d is 2.67 which is not equal to d L. Similarly, 4 minus d which is equal to 1.33 which is not strictly, sorry which is again not less than d L value which is 1.66, so here d value is large.

So, we reject h nod rejecting h nod means well, so d is not true, so this is not true we are not going to reject h nod we accept h nod because as my d which is equal to 2.67 which is not equal to d L which is not less than equal to 1.16, so this is not true. So, I will accept h nod accepting h nod means no there is there is no lag 1 auto correlation or serial correlation, so this is the first one and also you can do. What you can do is that you check with this one if, of course you will get the same result if d is greater than d U and 4 minus d is greater than d U then it says that you accept h nod which is equal to rho is equal to 0 is it.

So, this is the test in terms of d U value d upper value, so d is greater than d U yes d is 2.67 which is greater than d U 1.33 and also my 4 minus d which is equal to 1.33 is greater than or equal to d U that is 1.33, so both are true. So, we accept h nod, accepting h nod means there is, there is no, there is no lag 1 auto correlation or serial correlation in the in the data. So, here we have used Darwin Watson test, to test whether there exists you know serial auto correlation lag 1 auto correlation or not.

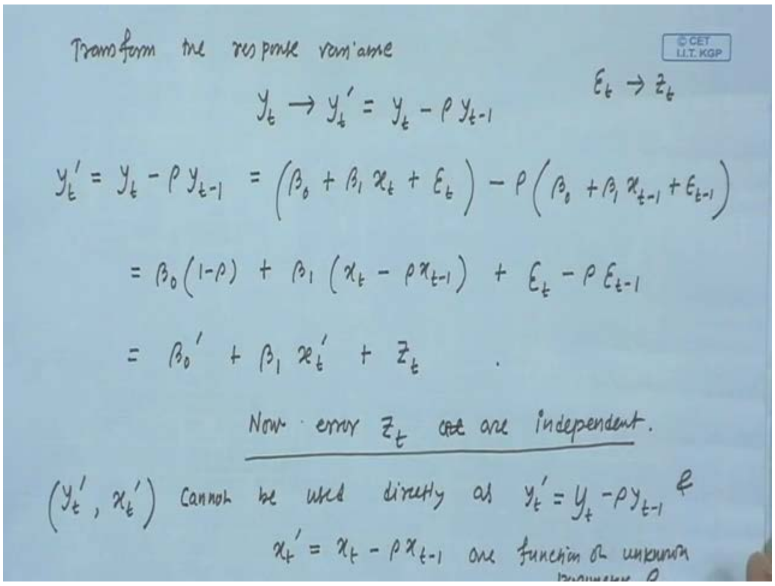
So, when the d value that is the Darwin Watson test statistic whether when that is small that means there exist auto correlation. Here, you can see that the d value is not small, smaller than the d lower value that is why, finally the conclusion is that there is no lag 1 auto correlation in the in the data. Well, so we talked about this problem and still we have you know sometimes.

(Refer Slide Time: 48:39)



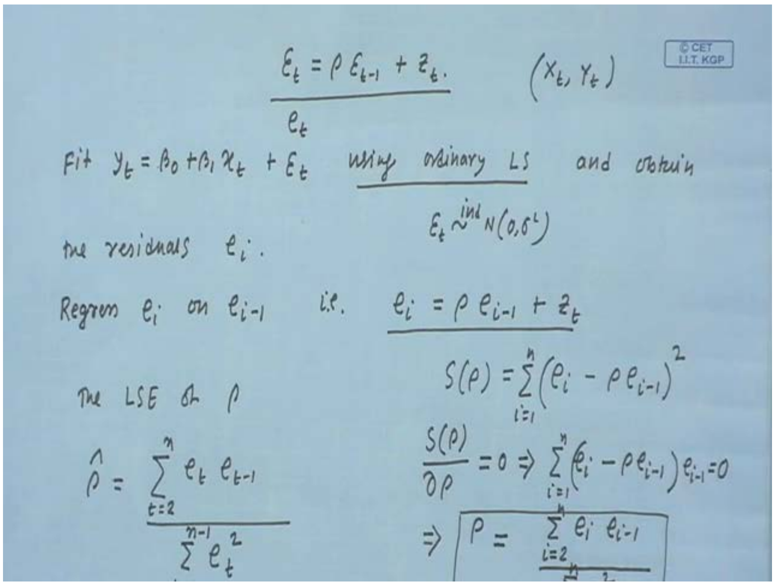
So, we can sort of recall the parameter estimation technique in case of in existence of first order auto regressive errors. So, this is what the Cochrane Orcutt method you know and this says that you know how to estimate this parameter beta nod and beta 1 when there exists first order auto correlation or among the errors well and this epsilon t they are not independent 0 sigma square they are having the first order regressive error.

(Refer Slide Time: 49:44)



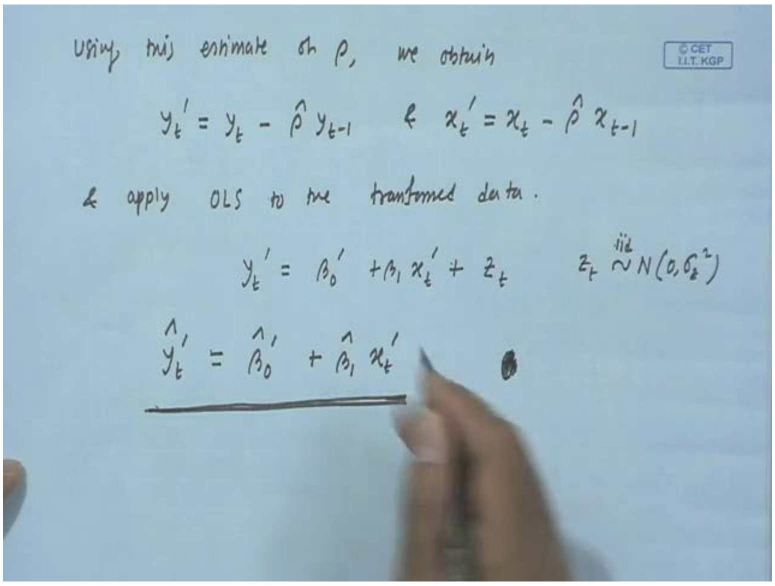
What we did there is that we transform y t to y t prime which is equal to y t minus rho y t minus 1 and you can finally check, you can finally check that this y t prime is equal to beta nod prime plus beta 1 x t prime plus z t and here z t is equal to epsilon t minus rho epsilon t minus 1. We know that this z t this transform error they are independent and that is why you know you can ordinary least square technique to this transformed data. But, the problem again is that you know this x t prime and y t prime they involve some unknown parameter rho.

(Refer Slide Time: 51:00)



We talked about how to estimate those rho I mean the similar technique, I mean just repeating the things here. So, you fit the module using the ordinary least square technique, forget ignore the thing that assumption is not true. Here, and fitting the ordinary least square technique you will get the residuals e i and then you regress e i on e i minus 1 right and you fit this module that is e i is equal to rho e i minus 1 plus z t. The least square estimate of rho is this one you can check that you know you have to estimate rho in such a way that this is this is minimum, so the rho value is equal to this.

(Refer Slide Time: 52:04)



Using this rho, now you can transform the data y t to y t prime, similarly x t to x t prime and now you know that this transform data or the transform module has error z t, which is normal 0 sigma square and they are independent. So, you can apply ordinary least square technique to this transformed data and you can happily know estimate beta nod hat and beta 1 hat. So, this is sort of just you know we solved the same problem for second order auto regressive error. Today well, so that is all for today again in the next tutorial class we will be solving some randomly selected problems.

Thank you very much.