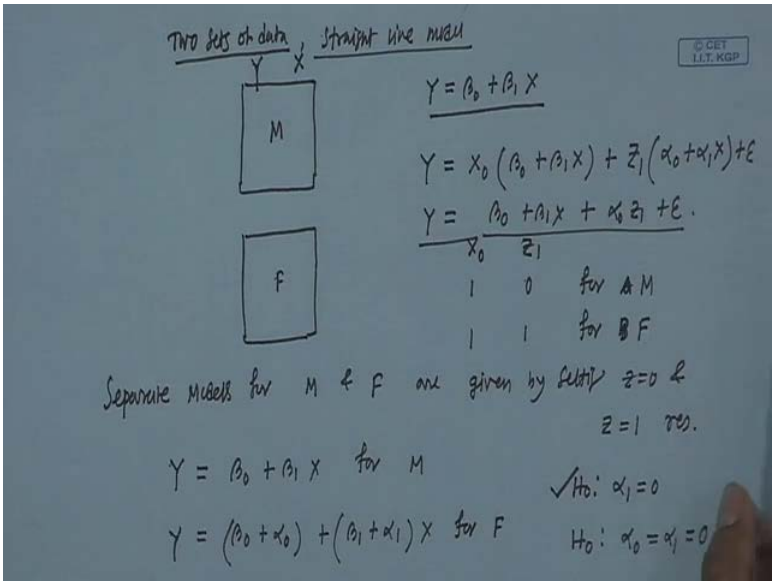
Regression Analysis
Prof. Soumen Maity
Department of Mathematics
Indian Institute of Technology, Kharagpur

Lecture - 26
Dummy Variables (Contd.)

Hi, this is my third lecture on Dummy Variables. The variables used in a regression analysis are generally qualitative, but occasionally we need to it is necessary to use qualitative variables, and the dummy variables are used to incorporate the qualitative information in regression analysis.

(Refer Slide Time: 01:11)



Let me give an example the use of dummy variable, suppose you have two sets of data well, I forgot to say that you know this the example of qualitative information's or qualitative variables are like a employment status or sex whether male or female or on the origin from where the data taken, whether it is city a or city b or city c. And if there is a significant difference in response level between two sets of data, it is not advisable to use a simple linear regression model contrasting all the data together.

So, here is an example suppose you have two sets of data, this is say for male and this for female, and the data are on two variables say the response variable Y, and the regression variable X. So, this could be like you know expenditure per month, and this is the income per month, and you have to two sets of data, so it is not advisable to fit a model

like Y equal to beta naught plus beta 1 X for all the data together, because if there is a significant difference in the response between the male data and the female data, this is not recommended.

So, what we need to do is that, we need to incorporate the qualitative information that one set of data is for male, and the other set of data is for female. And the model which is recommended here, using one dummy variable is Y equal to X naught beta naught plus beta 1 X plus Z 1 alpha naught plus alpha 1 X plus epsilon. So, of course, you know here we are using two dummy variable, including X naught and their value X naught and Z 1 is 1 0 for the first set A and 1 1 for the second set B I said I mean for male and for the female.

So, the separate models for male and female are given by setting Z equal 0, and Z 0 to 1 respectively. So, the model we recommend is that Y equal to beta naught plus beta 1 X for male, and beta naught plus alpha naught plus beta 1 plus alpha 1 into X for female. So, if you know that there you know if the data are given in two sets, you fit this model which is equivalent to Y equal to beta naught plus beta 1 X for male, and Y equal to beta naught plus alpha naught plus beta 1 plus alpha 1 into X for female.

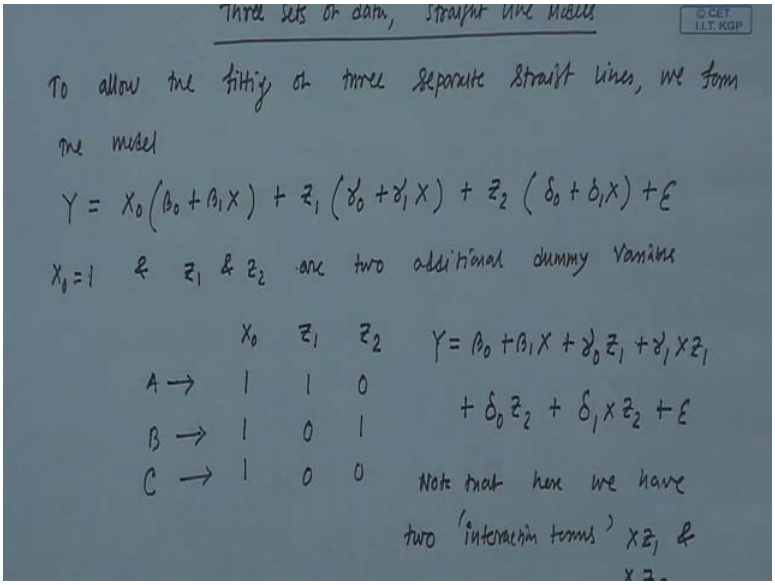
So, you fit this model and then by testing the hypothesis like say H naught equal to say alpha 1 equal to 0. So, if you test the hypothesis that whether alpha 1 is equal to 0 or not, that is basically about testing the appropriateness that whether two parallel straight line can be fitted for two sets of data. Similarly, you test the hypothesis H naught which is equal to say alpha naught equal to alpha 1 equal to 0.

So, this hypothesis is testing this hypothesis is equivalent to your testing the appropriateness of the fact that, whether the same model can be fitted for both sets of data or not. And the other one is that, you test this hypothesis H naught which is say alpha naught equal to 0, so this one testing this hypothesis means, you are testing whether two straight line with the same inter shaped can be fitted for two sets of data.

Now, if you see that all of them are rejected then you can go for this model, I mean this is your final model which feeds the rate of data best. And if you see that say for example, this one is accepted for example, H naught which is alpha 1 equal to 0; that means, this indicates that if this hypothesis is accepted. That means, you can go for a two parallel line, you can fit two parallel straight line for two sets of data, but with different intercept.

That means, in that case if this is accepted you can go for the model Y equal to say beta naught plus beta 1 X plus alpha naught Z 1 plus epsilon. So, this is the model you can fit for the given set of data, so this is about two sets of data and straight line model.

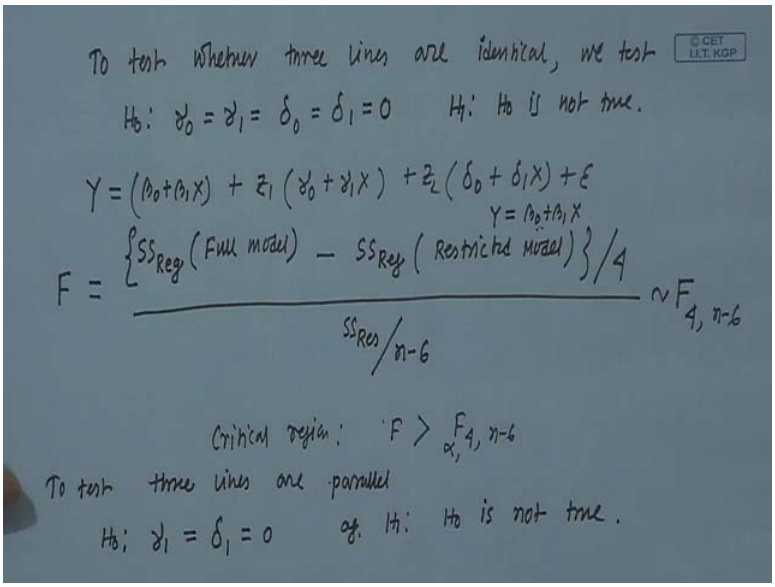
(Refer Slide Time: 09:54)



Now, we will go for three sets of data and straight line models, so to allow the fitting of three separate straight lines, we form the model Y equal to X naught beta naught plus beta 1 X plus Z 1 gamma naught plus gamma 1 X plus Z 2 delta naught plus delta 1 X plus epsilon. So, this is the model for three sets of data, and we are trying to fit straight line model for each set and here of course, X naught is equal to 1 is dummy variable, and Z 1 and Z 2 are two additional dummy variables X naught Z 1 Z 2 1 1 1 1 0 0 1 and 0 0 this is for the first set, this is for set A, this is for set B and this is for set C.

And this can be rewritten as Y equal to beta naught plus beta 1 X plus gamma naught Z 1 plus gamma 1 X Z 1 plus delta naught Z 2 plus delta 1 X Z 2 plus epsilon. So, note that here we have two interaction terms involving dummy variable, two interaction terms, so X Z 1 and X Z 2. So, this is the general model which sort of cover all possibilities of fitting three straight line for three sets of data, now we can test some hypothesis, like a since we are given three sets of data whether we can go for a three parallel line for three sets or whether we can fit one single straight line model for all the three sets. So, those I mean you can form the appropriate testing null hypothesis to test all this things.

(Refer Slide Time: 15:05)



So, we will talk about two testing here, so first one is to test whether three lines are identical, we test the following hypothesis I need to write down the model may be once more. So, test whether three lines are identical; that means, we have to test whether gamma naught is equal to gamma 1 is equal to delta naught is equal to delta 1, so let me write the model once more. So, Y equal to beta naught plus beta 1 X plus Z 1 gamma naught plus gamma 1 X plus Z 2 delta naught plus delta 1 X plus epsilon.

So, this is the model now we want to test whether you can go for a single straight line model, with the three lines identical to test this hypothesis we need to test this one, that gamma naught is equal to gamma 1 is equal to delta naught is equal to delta 1 is equal to 0. Against the alternative hypothesis that H naught is not true, I hope that you know how to test this sort of hypothesis for a model like this because this one is nothing but multiple linear regression model right.

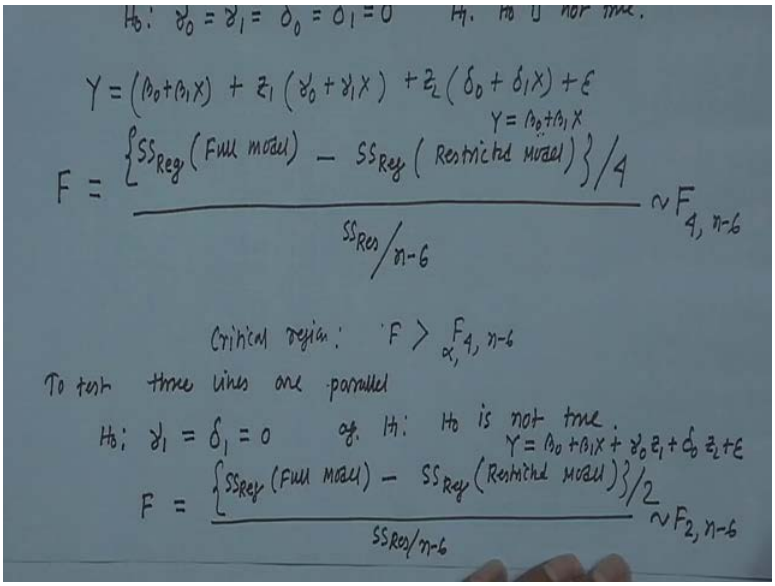
So, you have sort of three regressors you can consider this dummy variable as a regressor variable of course, it involves some interaction term that does not matter. So, you I hope that you can recall the extra some square techniques, so the to test this hypothesis we use the test statistic F which is equal to the SS the regression for the full model, you know what I mean by this you compute SS regression for the full model, this is the full model minus this is regression for the restricted model.

So, what is the restricted model here the model under the null hypothesis, so the restricted model is Y equal to beta naught plus beta 1 X. So, that is what the null hypothesis suggest that you go for this model; that means, the null hypothesis suggest that, you go for a single straight line fit for three sets of data. So, you compute the SS regression for the restricted model, and you divide this quantity, so this was has degree of freedom this is the full model. So, it has degree of freedom 6 and this one has degree of freedom 2.

So, 6 minus 2 it is you divide this quantity by 4 and this by this is residual by residual degree of freedom that is n minus 6. So, this follows F distribution with the degree of freedom 4 n minus 6, and then the critical region of course, critical region is a you reject this null hypothesis, if this F value is a greater than F 4 n minus 6 at some level of significance alpha. So, what is this numerator is basically, it is basically the contribution of this parameters or the associated regression variable to explain the variability in Y.

Similarly you can say to test whether the three lines are parallel; that means, whether you can fit three parallel line for three sets of data, to test hypothesis you what you have to test, you test the null hypothesis H naught that gamma 1 is equal to delta 1 is equal to 0 against the alternative hypothesis that H naught is not true.

(Refer Slide Time: 21:50)



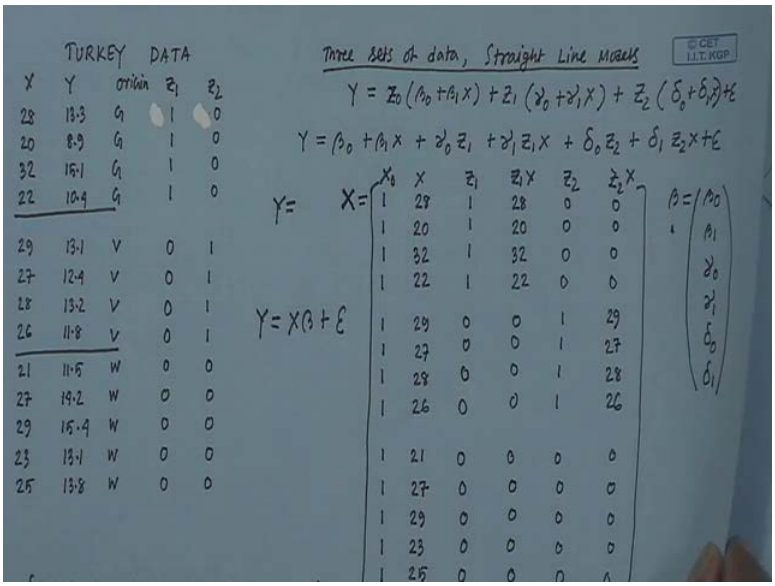
And this one also you know same technique, you can test this non hypothesis by using the F statistics F equal to SS regression for the full model minus SS regression for the

restricted model. What is the restricted model that is the model under the null hypothesis, so here the restricted model is basically you just put lambda is lambda 1 equal to 0 and delta 1 equal to 0 in this equation. So, that restricted model is Y equal to beta naught plus beta 1 X plus lambda naught Z 1 plus delta naught Z 2 plus epsilon.

So, this is the restricted model, so you see what is the SS regression under this restricted model, and this has of course, degree of freedom six and this has degree of freedom 1, 2, 3, 4. So, you divide this by 2, 6 minus 4 by SS residual by degree of freedom m minus 6, so this follows F to n minus 6 and of course, the critical region is you reject the null hypothesis (Refer Time: 23:29) if F is greater than F alpha 2 n minus 6.

Now, what I will do is that, so I talked about now I am at this moment I am talking about suppose you are giving three sets of data, and I am trying to fit three straight line model for three sets of data. And we have just now we talked about you know, whether we can go for a three parallel line or whether you can go for a single straight line model for all three sets of data, and we learn how to test those possibilities. So, what I will do is that I will explain this if we can recall now, we had data called turkey data and there, we had a three sets of data. And I will try to explain whatever I talked just now by using the turkey data.

(Refer Slide Time: 24:44)



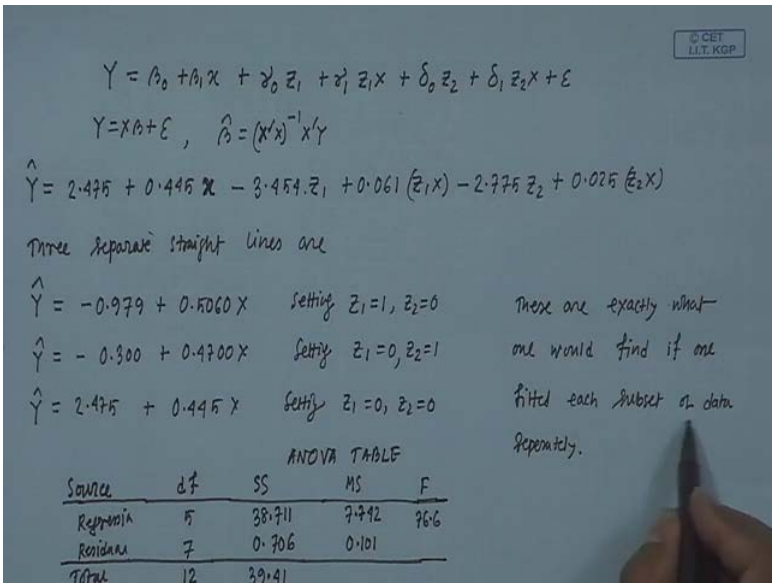
So, here is the turkey data we had and Y is weight in pounds X is H in wigs, and they are some three different origin well. So, we have three sets of data and we are trying to fit

straight line model for three sets, so the model you have to consider is that is a general model is this 1 Z naught into Z naught or you can call X naught also no problem, Z naught into beta naught plus beta 1 X plus Z 1 into gamma naught plus gamma 1 X plus Z 2 into delta naught plus delta 1 X plus epsilon.

So, this can be rewritten in this way now I will just explain you know how to fit this how to estimate these regression coefficients, here you have six regression coefficients. So, here is the parameter vector beta, this involves all beta naught beta 1 gamma naught gamma 1 delta naught delta 1. And here is the coefficient matrix, so the first one is stands for say you can put here Z naught also, so this is either Z naught or X naught you put. And this X is for the regressor variable X here, and this X is same as the X here.

Now, Z 1 you know the dummy variable scheme for two sets, so this is the scheme, so Z 1 is this column, Z 1 into Z you just multiply Z 1 with X and you will get the column associated with Z 1 X, Z 2 you know this is the scheme for the dummy variable for two sets of data. And once you have Z 2 you can compute Z 2 X, so this is what you know I am just you know explaining how to estimate this parameters. So, you have the coefficient matrix X, you know this is the parameter you have to parameter vector that you have to estimate, you know what is Y vector this is the y vector. And I am sure that you understand that you know this X and this X is different, this is the matrix coefficient matrix and this is a regression variable well. So, you can write this multiple linear regression model in this form Y equal to X beta plus epsilon.

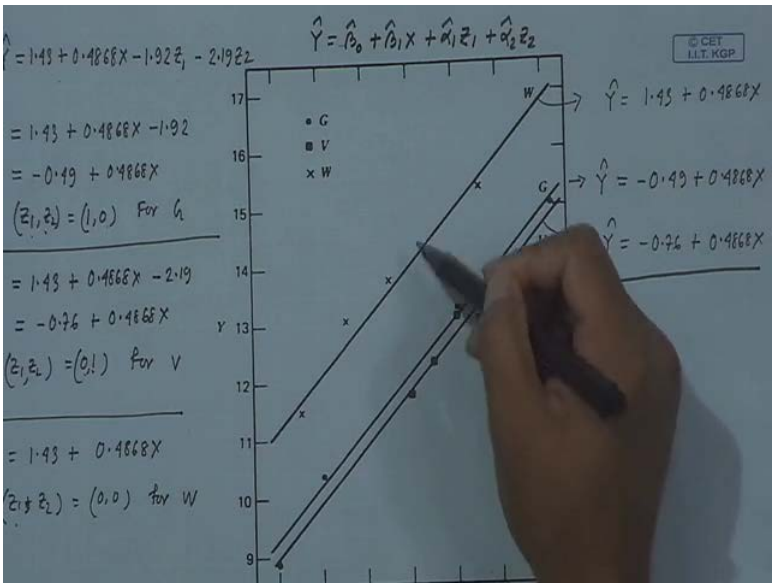
(Refer Slide Time: 28:08)



And you can see, you know that you know for multiple linear regression model beta hat is X prime X inverse X prime Y. So, you know what is X you know the X matrix you know Y, so you can compute beta hat, and this are the estimate of the regression coefficients. So, this is the estimated I mean this is the fit basically, now three separates straights lines for first block is this one, so this one is obtained by setting Z 1 is equal to 1, and Z 2 equal to 0 in this fit.

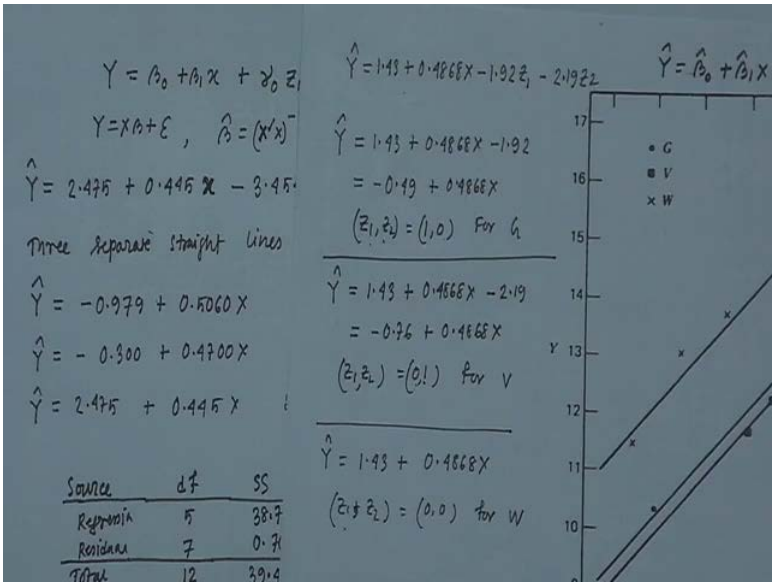
Similarly for the second block, the fit for the second block is obtained by setting Z 1 is equal to 0, and Z 2 equal to 1 in this equation here, in the fitted model here. And the fit for third block is obtained by setting Z 1 equal to 0 and Z 2 equal to 0, and this 3 fits it says that this are exactly what, one would find if one fitted each subset of data separately. So, if you go for say separate fitting for separate sets, then we will get exactly the these fits.

(Refer Slide Time: 30:08)



And now let me compare this one this is what we got before, this is without using the interaction term. So, this is the model we considered here before, this model did not this is basically here we consider that all three lines are parallel and that means the slope for all three fits is a beta 1 hat, and you can see here we got this is fit for the first set this is the fit for the second model for the second block.

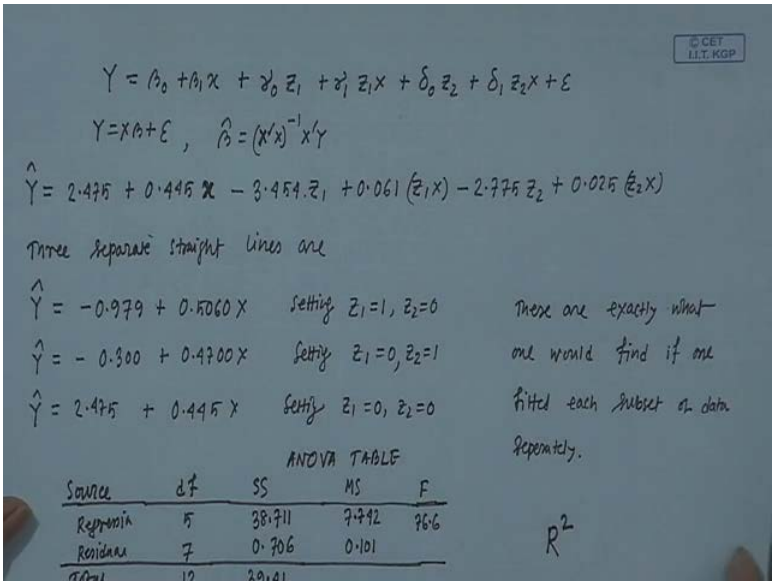
(Refer Slide Time: 30:52)



Now, you can compare that you know this are not different too much I mean the fit you obtained here, and the feet we had before they are not very different. But; that means,

what I am trying to say is that, so we have fitted the generalize model and this the fit we got here are not, so different from the fit we had before considering that, considering three parallel straight line for three sets of data. Now, what we need to do is that we need to statically test, whether this three set of data for turkey data, they really require three different straight line model or three parallel straight line are for the turkey data that we can do statically by testing suitable null hypothesis right.

(Refer Slide Time: 32:02)

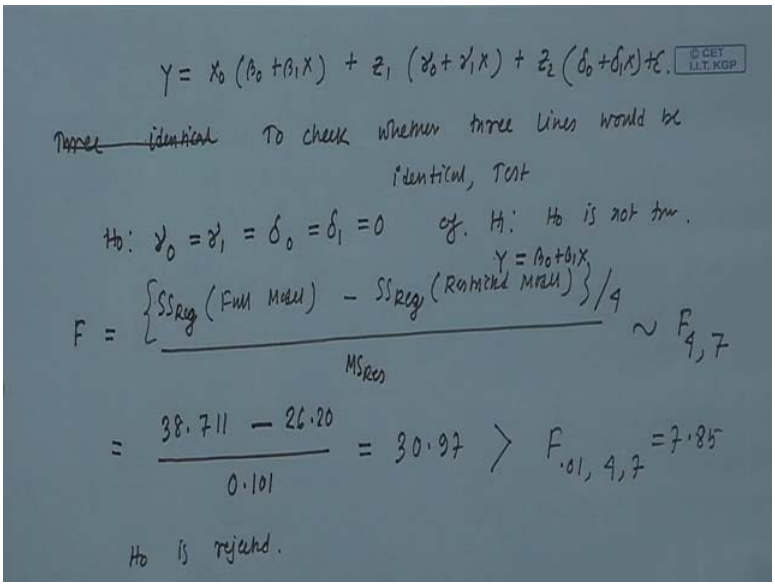


So, before doing that I will just write and here is the ANOVA table for the turkey data, we had you know this is the total SS T and the degree of freedom is 12 because we had 13 observations. And this is the SS regression, and SS residual is this one, and SS residual degree of freedom is safe and because you know that here in the model we have 6 parameters, and we have total 13 residual, 6 parameters means there are 6 restriction on the residual.

So, 7 residuals can be selected you have the freedom of selecting 7 residual and the remaining 6 residuals have to be chosen in such a way that, they follow those 6 restrictions. So, that is why the residual degree of freedom is equal to 7, and then the regression coefficient regression degree of freedom of course, it is 5 and this is the part of variability which is explained by the model, you know this is the total variability in the response variable.

This is the variability which is explained by this model I mean here, if you can recall the parameter R square is very high I mean most of the viabilities explained by this model, and here is the F value. So, to check now what I wanted to do say is that, we have observed that these three straights lines are very similar to the straight line, we obtained using by considering three parallel line fit before, now we need test that formally well.

(Refer Slide Time: 34:29)



So, our model is the model we are fitting here is Y equal to X naught beta naught plus beta 1 X plus Z 1 gamma naught plus gamma 1 X plus Z 2 delta naught plus delta 1 X plus epsilon. Now, first we check whether three lines could be identical, so three I mean while let me write it to check a whether three lines would be identical, your tester the hypothesis that H naught gamma naught equal to gamma 1 equal to delta naught equal to delta 1 equal to 0, against H 1 that H naught is not true.

So, we have the fitted model and now we are trying to check whether we really need such a general model or we can just fit a single straight line for all three sets of data. So, that we will get the answer from by testing this hypothesis, so how do we test this hypothesis, the F statistics is the SS regression for the full model. And SS regression for the restricted model and by the degree of freedom here it is 4 and by MS residual, so this SS residual by residual degree of freedom.

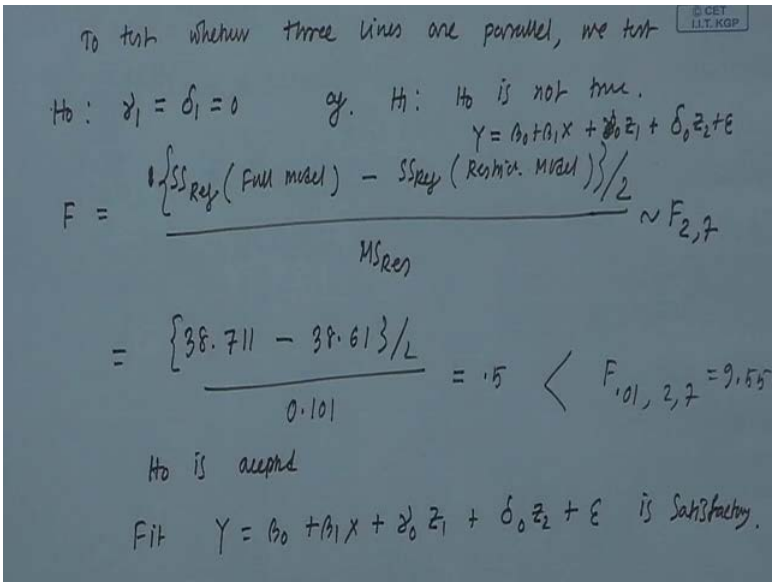
So, now if you recall this ANOVA table here the SS regression is considering for the full model is 38.711. And of course, you have to compute the SS regression for the restricted

model, and the MS residual is 0.101 and you can check that for the restricted model, what are the restricted model. Restricted model is Y equal to beta naught plus beta 1 X, so just fit this model for the given data and you, see how much of the variability is explained by this model that is what the SS regression for this model, you can check that that is 26.20.

So, this one is 30.97 and this follows F 4 the residual degree of freedom is 7, so this one is greater than F 0.0147 which is equal to 7.83. So, you refer the tabulated value for F statistics, so; that means, this is significant, significant means we cannot go for I mean H naught rejected. So, this test suggest that for the turkey model you fit a general straight line fit, and then you are testing whether this three blocks of data can be fitted by a single straight line.

And the test here implies that or you know the test says that, you cannot go for a single straight line model, you have to go something else. So, fitting a single model for turkey data is rejected, next what we will do is that we will test a whether we can go for parallel lines.

(Refer Slide Time: 39:37)



To test whether three lines are parallel to test this we test null hypothesis H naught, which is a gamma 1 is equal to delta 1 equal to 0 against the alternative hypothesis that H naught is not true. So, similarly we use the F statistics here and that is nothing but SS regression for the full model, and then SS regression for the restricted model right, and

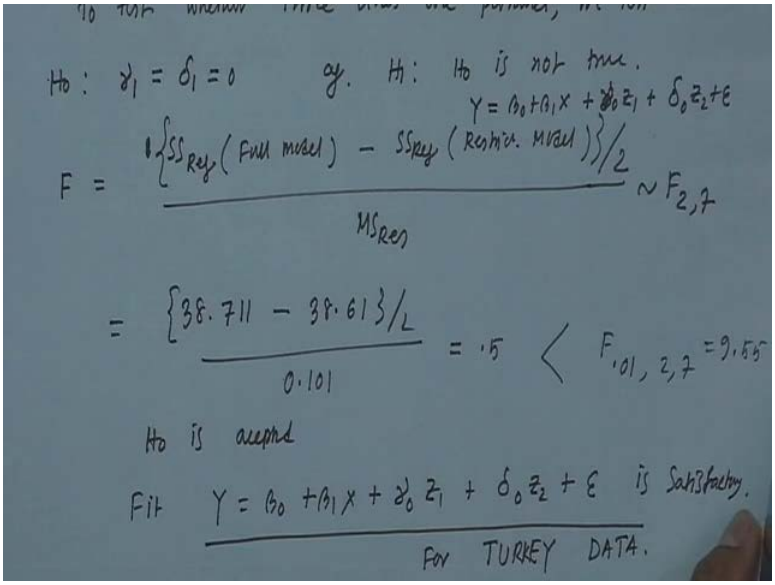
the full model has 6 parameters and the restricted model has 4 parameters. So, 6 minus 4 is 2 you divide it by 2 and by MS residual, this follows F 2, 7.

Now, we know that for the full model the SS regression is 38.711 you have to find this one, you know what is the model here? So, that is 38.61 let me write at least the model, so Y equal to beta naught plus beta 1 X plus delta naught Z 1 plus Sorry gamma 1 gamma and delta naught delta naught Z 2 plus epsilon. So, this is the restricted model, so you split this model you find the SS regression due to this model that is 38.6 hardly any difference.

So, this one by 2 by MS residual that is 0.101 this is equal to 0.5, which is less than F 0.0127 that is 9.55 this is the tabulated value. So, this says that the test is not significant; that means, we accept the null hypothesis, so here H naught is accepted, so the fit is the fit we talked about the fit Y equal to beta naught plus beta 1 X plus gamma naught Z 1 plus delta naught Z 2 plus epsilon is satisfactory well.

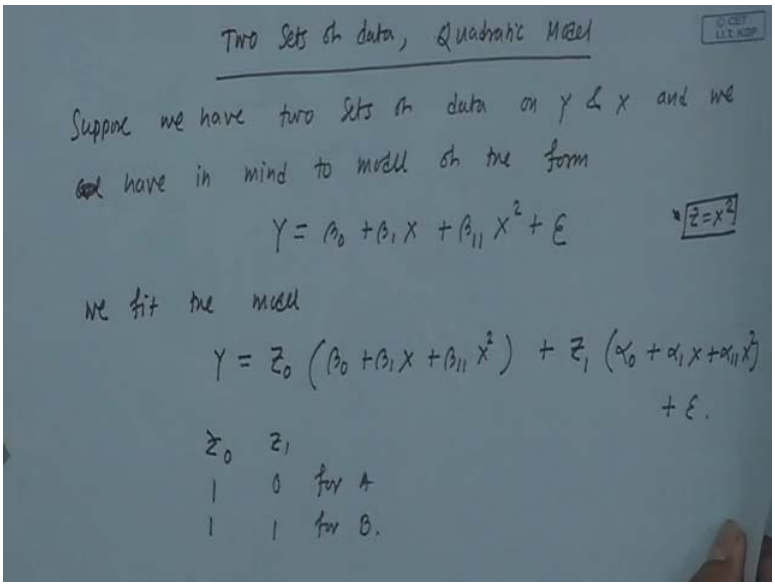
So, let me conclude this part, so for the turkey data involves three sets of observations and what we did here is that, first you fit a general model involving three dummy variables. If it is straight line generalize model, involving three dummy variables and then you test whether this three lines can be parallel or not. So, form the test here we observe that, the hypothesis accepted; that means, we can go for three parallel line for the turkey data.

(Refer Slide Time: 43:55)



And based on this test finally, we conclude that this model is satisfactory for turkey data well. So, this is about three sets of data straight lines fits.

(Refer Slide Time: 44:28)



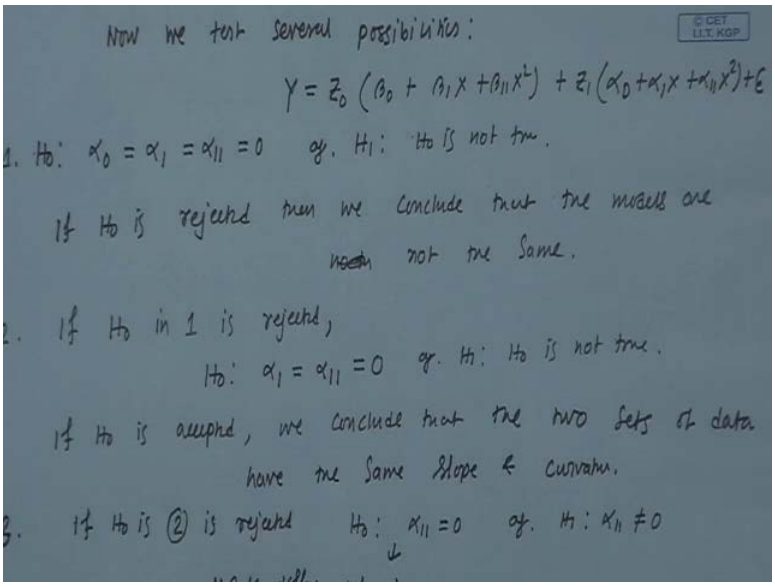
Now, we will go for two sets of data quadratic model, suppose we have two sets of data on Y and X. And we have in mind a model of the form Y equal to beta naught plus beta 1 X plus beta 1 1 X square plus epsilon, so we have two sets of data, and we are thinking of fitting a quadratic model for this one. We did not talk about this sort of quadratic model before, this are called polynomial fitting, but let me just say that this not difficult.

Because, you just consider X square as say call it Z equal to Z square, then this is nothing but a multiple linear regression model that is all for the time being, but we will be talking about polynomial regression later on. So, here we have two sets of data and we are planning to fit quadratic model, so we fit the model involving two dummy variables that is Y equal to say let me call Z naught beta naught plus beta 1 X plus beta 1 1 X square plus Z 1 alpha naught plus alpha 1 X plus alpha 1 1 X square plus epsilon.

So, this is the same technique we used for linear fitting also straight line fit, instead of straight line we are replacing this by quadratic fit here. So, here the dummy variable scheme is Z naught Z 1 this is always 1 and this is 0 for A and 1 for B, now what we want is that, we want to test several possibilities. Like, whether if we need two different quadratic model for two sets of data or we can go for identical quadratic fit, for both the

sets or something else. So, we will be talking about two three possibilities and we will talk how to test them.

(Refer Slide Time: 48:34)



Now, we test several possibilities first of all let me test let me write down the model once more, Y equal to Z naught beta naught plus beta 1 X plus beta 1 1 X square plus Z 1 plus alpha naught plus alpha 1 X plus alpha 1 1 X square plus epsilon, so this is the model. Now, to test whether we can go for the same quadratic fit for both the data or not, we will test hypothesis H naught is a say alpha naught equal to alpha 1 equal to alpha 1 1 equal to 0, against H 1 that H naught is not true.

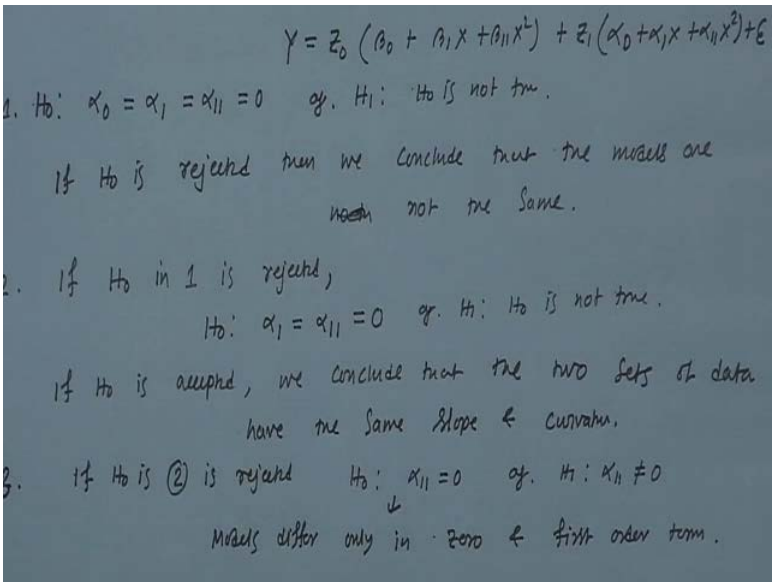
So, you know how to test this hypothesis by using extra some of square technique, so of course, you understood by this null hypothesis that, if the null hypothesis is accepted; that means, you can go for single quadratic fit. So, if H naught is rejected then we conclude that the models are not the same for two sets of data, so this is 1, 2, so if H naught in 1 is a rejected. That means, we cannot go for the same quadratic fit for both sets, now we will check several other possibilities.

If this is rejected, then you test this hypothesis H naught that is alpha 1 equal to alpha 1 1 equal to 0, against the H 1 that H naught is not true. What does it means is that well the two quadratic models are different, but they have different intercept, but the null hypothesis say that, they have the same slope and curvature. So, it says that if H naught

is accepted here, then we conclude that the two sets of data have the same slope and the curvature.

So, this is how you know we can test many hypothesis and may cases let me do one more is that, if H naught in 2 is rejected. Then you can now go for you test whether alpha 1 1 equal to 0, against H 1 that alpha 1 1 is not equal to 0; that means, null hypothesis is that their slope and curvature are not the same. Now, test whether the curvature is same for both the fits or not.

(Refer Slide Time: 54: 05)



So, this says that the null hypothesis said that the model defers, the models defer only in 0 and first ordered term. That means the curvature is same, but they have different intercept and different slope, so that is what you know we can test a several hypothesis once you have a model. So, let me conclude this module dummy variable, so here the dummy variable in regression analysis is used to incorporate, qualitative information in the data. Here, we understood that if you have say three blocks, then you need three dummy variable to fit a general model.

So, what is recommended is that you know if you have say three sets of data, you should not go for a single straight line fit for all the blocks together. You fit a general model, involving dummy variables and then you go for several testing, like whether you can go for a identical straight line fit for all the data or you can go for parallel straight line fits for different blocks like that.

And depending on the result of your hypothesis testing, you chose the final model, so I hope you understood the use of dummy variable in regression analysis, they are used to incorporate a such qualitative information available with the observation. And that is all for today.

Thank you very much.