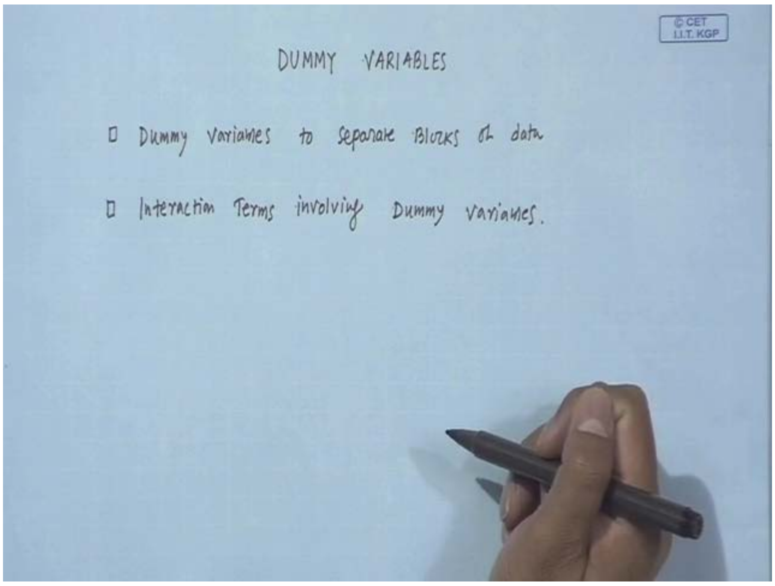
Regression Analysis
Prof. Soumen Mailty
Department of Mathematics
Indian Institute of Technology, Kharagpur

Lecture - 25
Dummy Variables

(Refer Slide Time: 00:38)

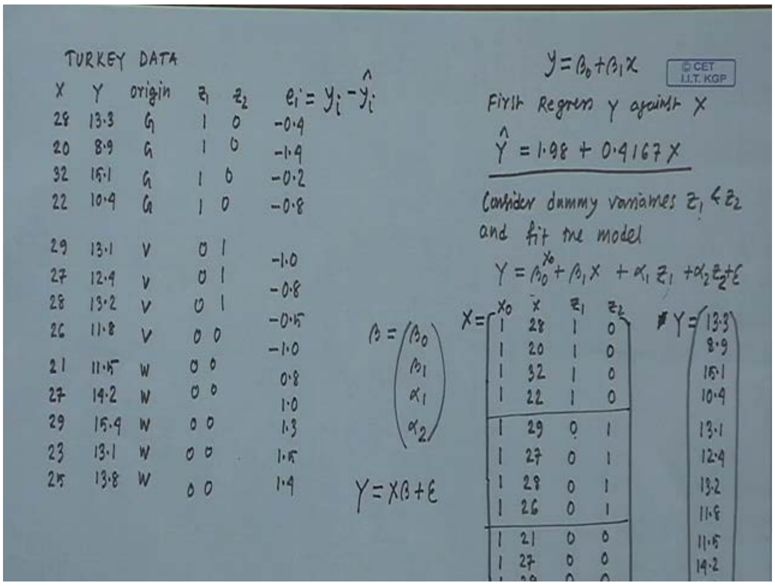


Hi, so, this is my second lecture on dummy variables and here is the content of this module, dummy variables to separate blocks of data and interaction terms involving dummy variables. Let me explain the objective of this modules once more. The variables used in regression analysis are, usually qualitative and this variables have well defined scale of measurements. Example of qualitative variables are like temperature, pressure, income, expenditure ok and occasionally you know, we need to use some qualitative variables in regression analysis.

Say, I am considering the qualitative variable symmetrical status and my response variable is say, per month expenditure and the regression variable is income per month. So, I am interested to find relationship between, expenditure and income. But, there could be you know, sort of difference in a significant difference in response level. Here, response is expenditure. So, there could be in significant difference in expenditure amount, between a set of married people and in the set of unmarried people. So, we cannot really no plug both data set, together and find the relationship between expenditure and income.

So, we cannot put them together and fit a single straight line model, if there is significant difference in response level. So, let me once more explain the turkey data, we considered in the last class and using this example, I will say, what I mean by significant difference in the response level, ok.

(Refer Slide Time: 03:29)



So, here is the data, we considered in the last previous class and here the Y is turkey weight in pounds and X is edge in weeks, ok. So, we have 13 observation here, we have sort of 3 sets of data, I mean 3 blocks you can say. So, this is this 4 data are originated from Georgia and similarly, this fourth data, I mean this 4 data from Virginia and this 5 data's data are from is Wisconsin. Now, here is the response variable is the weight of turkey in, in pounds. So, what you mean by, whether you can see or fit a simple straight line module between X and Y here. We have 3 sets of data right, well so, first you try, you start with simple regression model between X and Y and you fit the model.

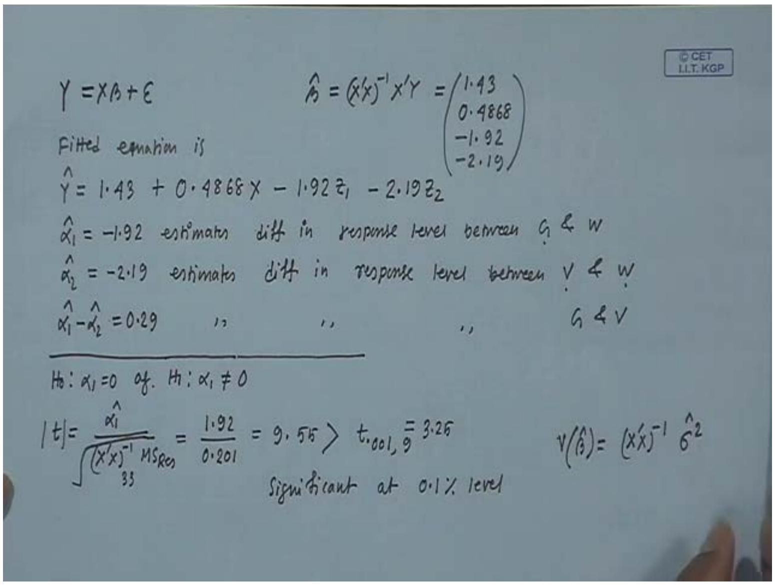
So, this is the fitted model and you know how to fit this model is nothing but, the fitting Y equal to beta naught plus beta 1 X and here is the model. So, once you get the fitted, the model you can find the residuals. Now, we can see the residuals for the first block are all negative residuals. For the second block, are all negative and here the residuals for the third block all positive, but, it means is that, the response value, for the third block or the third set of data are, because this residual is nothing. But, this is y i observed observation minus the fitted. So, this is positive, means the fitted response value of this block are smaller than the original data and similarly, here it is just opposite.

So, that means there is this sort of this residual, if we plot this residual graphically. So, that indicates that, so, this residual indicates that, there is significant difference in the response level. So, we cannot go for simple straight line model between, regressor and response variable ignoring the origin, ignoring the origin means ignoring the qualitative information we have. That they are from, 3 different origins. So, here is the use of dummy variables. So, dummy variable is used to incorporate qualitative information in the regression analysis.

Well so, I already talked about this is the. S o, we need to fit model involving in 2 dummy variables, because we have 3 sets of data. So, we know about it already. So, here is the model, we want to fit. The model is y equal to beta naught plus beta 1 X, plus alpha 1 Z 1, plus alpha 2 Z 2 epsilon. So, this is the first dummy variable, this is the second dummy variable and then the whole model, you know this can be treated as multiply linear regression model and it can be written as, Y equal to x beta plus epsilon. So, this is the matrix rotation, where Y is the response here and beta is the coefficient vector. So, this is beta naught 1 alpha 1 and alpha 2 and X is the coefficient matrix.

So, here we can put x naught, x naught is 1 for all the observations. And here is the regression variable and this column corresponds to, this regression variable here and Z 1 and Z 2 are dummy variable. So, for the first block Z 1 is equal to 1 and Z 2 equal to 0. For the second block Z 1 is equal to 0 and Z 2 is equal to 1 and for the third block Z 1 is equal to Z 2 equal to 0. So, you have the X matrix, you have the Y vector, you have Z vector. So, you know how to find beta hat.

(Refer Slide Time: 09:19)



So, here is the estimated regression coefficients, you know x and y. So, you can find it and the fitted equation is this one, ok. So, I told this one also before that this alpha 1 hat, estimates difference in response level, between the first block and the last block and alpha 2 hat, estimates difference in response level, between second block and the last block. Whereas alpha 1 hat, alpha 2 hat, estimate difference in response level between the first block and the second block, ok.

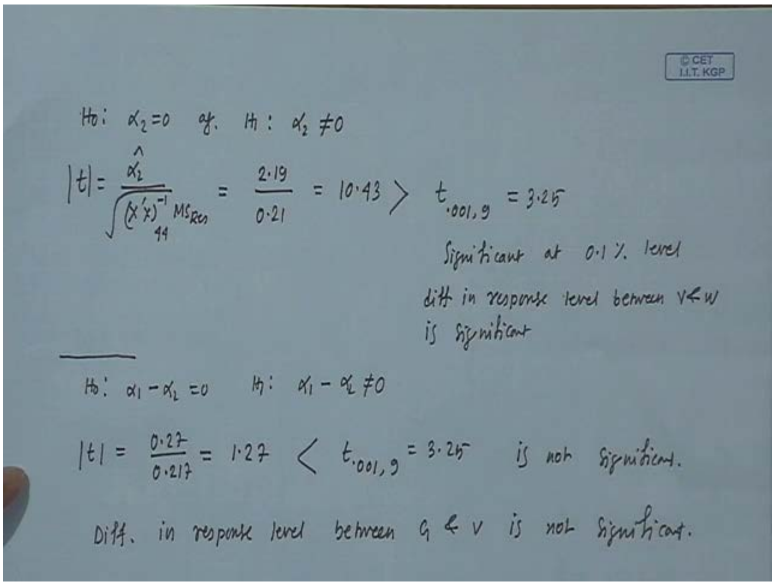
Well so, I go back to the data again (Refer Slide Time: 03:29) looking at this residual you know, is very clear that. There is a significant difference, between G and W and difference in response level is estimated by alpha 1 hat. While so, there is a significant difference between second block and third block, that means V and W, which is estimated by alpha 2 hat and intuitively appears me that, there is no significant difference between the first block and the second block that means G and V. Which is estimated by alpha 1 minus alpha 1 hat minus alpha 2 hat. But, you know we cannot say whether this is significant or not just looking at the value.

So, we need to test the hypothesis is not alpha 1 equal to 0 against alpha 1 not equal to 0 and we know that this can be tested using the test statistic this one. Where, X prime X inverse 3 3 denotes, the third diagonal element of X prime X inverse and this M S residual is nothing but, sigma square hat. Because, we know, that this is nothing but, variance, coefficient matrix for beta hat well ok. So, you can see the observed value is

9.55, which is bigger than that tabulated value and it as degree of freedom 9. I told why it is 9, in the last class. So, that means testing the significant at 0.1 percent level. What is the meaning of that? The testing significant means, this is rejected, this is accepted and the meaning of this one is that.

There is a significant difference in response level between, Georgia and Wisconsin, ok. In response level, that means in terms of turkey weight. Now, let us take whether, there is a significant difference between, Virginia and Wisconsin.

(Refer Slide Time: 13:04)



So, for that we need to test alpha 2. So, alpha 2 is equal to 0, against alpha 2 is not equal to 0, same test statistic. So, here only it will it replaced by 4 4 the 4-th diagonal element in the variance covariant matrix and you can see the observed value is 10.43 and the tabulated value is 3.25.

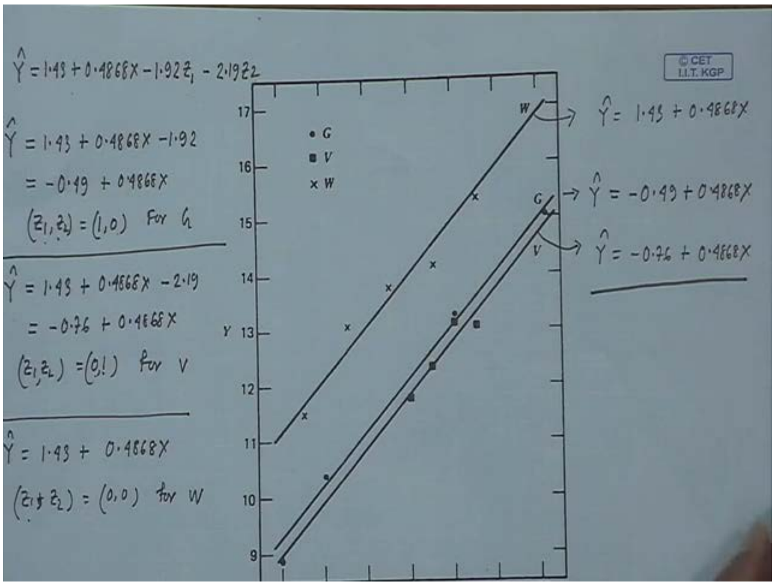
So, that means the test is significant. In other words, there is significant difference in response level between, V and W. (Refer Slide Time: 03:29) So, as I told you know, intuitively it appears to me that, there is no significant difference between, G and V, which is estimated by alpha 1 hat minus and alpha 2 hat. Let us check whether there is the significant difference between these 2 or not.

So, for that we need to test the hypothesis alpha 1 minus alpha 2 is equal to 0 against. That alternative hypothesis is alpha 1 minus alpha 2 is not equal to 0. And you can check that the object t value, is this one is nothing but, alpha 1 hat minus alpha 2 hat and this one is the variance of alpha 1 hat minus alpha 2 hat. So, they can check that this one is 0.217. So, the observed value is 0.127, which is less than the tabulated value at 0.1 percent level of significance and so, this says that test is not significant. That means, the difference in response level between G and V is not significant, ok.

So, if you see that, well what we have done till now. Is that we have given 3 sets of data and we have fitted a model involving 2 regress able variables. Sorry, 1 regress able variable and 2 dummy variables. And then we have tested the, we have tested whether there is a significant difference in response level between different sets. So, what we have observed that there is a significant difference between first set of data and third set of data that is Georgia and Wisconsin.

There is significant between second set of data and third set of data. That means Virginia and Wisconsin. But, there is no significant deference between, first set and second set. But, since there are significant differences between is 2 pairs of set. We cannot go for, simple straight line model; ignoring the origins. That means, ignoring the qualitative information, we have with us. So, we need to go for a model involving dummy variable, ok.

(Refer Slide Time: 17:03)

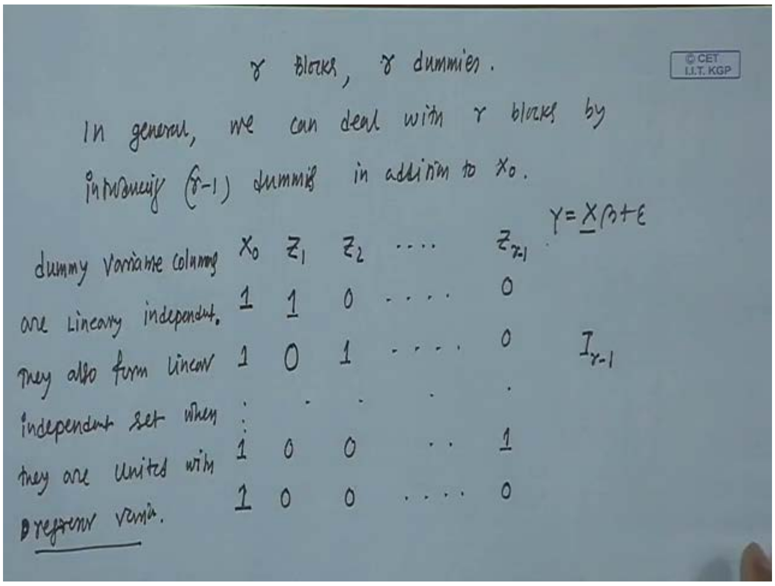


So, here is just, the graphical representation. So, this is the fitted model, we have seen before and then the response in the first set or in Georgia can be estimated by using this fitted equation. What we done here is that, you put Z 1 equal to 1 and Z 2 equal to 0, in the general model to get the fitted equation of the first set. To get the fitted equation for the second set, you put Z 1 equal to 0 and Z 2 equal to 1.

So this is the fitted equation for second set and this one is fitted equation for the third set. Were you put Z 1 equal to Z 2 equal to 0 and you can see the graph of this 3 straight lines fits and you must have noticed that, this 3 fits they have the same slope. They are basically the same model, with different inter set ok. Now, we will talk about, we will talk in this class only whether, we can go for model. I mean whether we can go for the same straight line model, I mean, or straight line model with the same slope all set of data. That might not be true for all set of data. But, here it seems ok.

So, we will talked about the general case we can think of different straight lines for different sets of data.

(Refer Slide Time: 19:09)



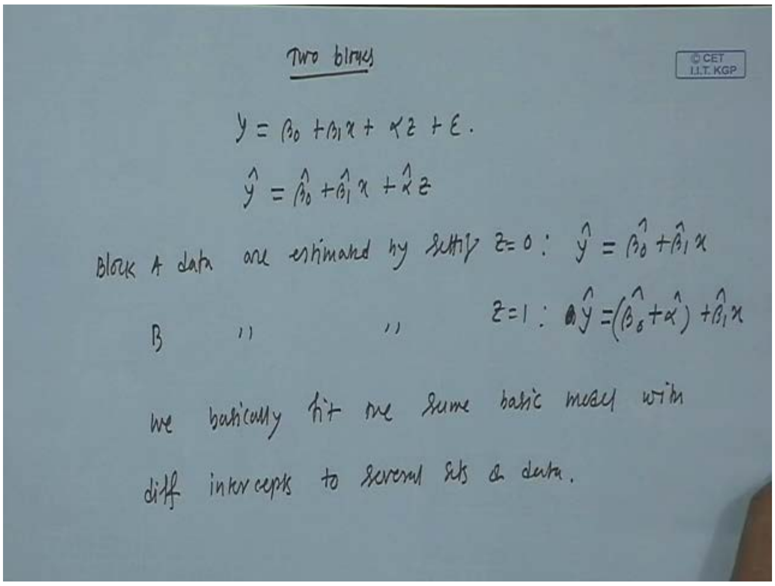
So, before that I will talk about the general case like suppose, you have r blocks, instead of 3 blocks, you have r blocks. And then, how many dummy variables you need? You need let me say its r dummies. So, it says that in general, in general we can deal with r blocks, by introducing r minus 1 dummies, in addition to X naught. So, how do you assign the value for first block, second block and the rth block?

So, suppose the dummy variables are Z 1, Z 2 and Z r minus 1. And for the first block, what we do is that? We put Z 1 equal to 1, Z 2 equal to 0 and all of them are equal to 0. For the second block, we put Z 1 equal to 0, Z 2 equal to 1 and this and for the r-th block. We put 0 0 0 and Z r minus 1 equal to I. So this is nothing but, you know identity matrix of order r minus 1 and for the final block, we put 0 0 and all 0. Now, so, here we have assign value for r minus 1 dummy variables. Now, if we include the dummy variable X naught, which is equal to 1, for all blocks.

So, this is what, this is how you know we assign a value for r blocks and we can deal r blocks with r dummy variables. So, what is special about you know this assignment, it says that dummy, you must have noticed that dummy variables columns are linearly independent. So, to assign the value, you can think about the other assignment value here. But, the constant, the condition is that, all this column, have to be independent and they also form a linear independent set, when they are united with regressor variable.

So, perhaps, you understood why this is true. Because, you have to finally, write the whole thing in matrix notation. Suppose there are r dummy variables. So, you can write in matrix notation Y equal to X beta plus epsilon and then this coefficient matrix S, has to be a full rank, right. So, that is why, they are this dummy variables columns are independent among themselves, not only that is true. They are also linearly independent, with the column for the regressor. So, you have to put regressor variable here also right, ok.

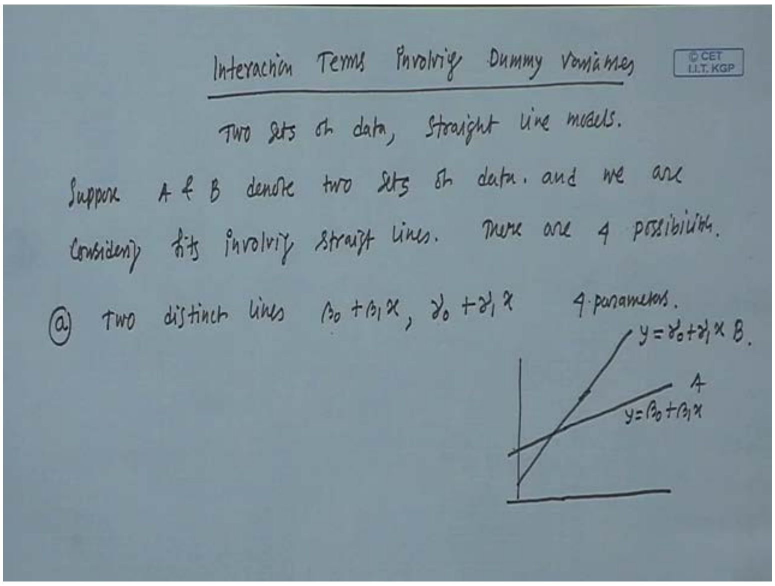
(Refer Slide Time: 24:22)



Well so, for 2 blocks, or 3 blocks you know what model we fit. For 2 blocks, we fit the model y equal to beta naught plus beta 1 x, plus alpha Z, plus epsilon and suppose the fitted model is y hat equal to beta naught, plus beta 1 hat x plus alpha hat Z. Then block A data, are estimated by setting Z equal to 0 right. That means for block A, data estimated and the fitted equation is y hat is equal to beta naught hat, plus beta 1 hat x and block B data are estimated by setting Z equal to 1. That is, that is the fitted equation is y hat, is equal to beat naught hat, plus alpha hat, plus beta 1 hat x.

So, what we are doing is that, we basically fit the same basic model. Basic model with different intercepts right, to several sets of data. So, whatever we have learnt till now is that basically finally, you understood that. We are setting the same basic model, with the same slope, with different intercept for different set of data. But, this might not be true always, I mean different set require or may require different straight line fit, ok. So, next we will talk about a model which involves interaction terms, involving a dummy variable.

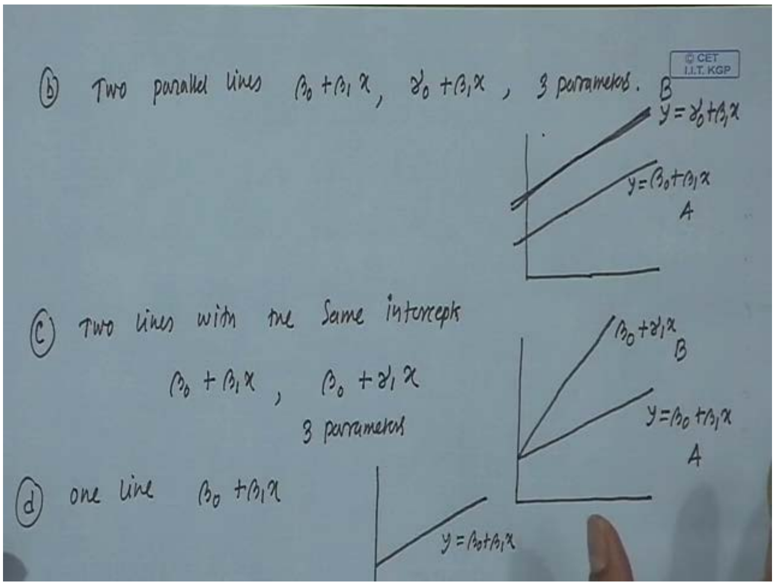
(Refer Slide Time: 27:04)



So, interaction terms, involving dummy variables. So, we will be talking about general case here. Suppose, you have 2 sets of data and we are thinking of straight line models. So, we have 2 sets of data and we are planning to fit straight model for both the sets. But, here we will be talking about general case, you know, 2 sets might have 2 different straight lines. That means in the sense that, they might have different slopes, they might have different intercept and all this things, ok.

Suppose, A and B denote two sets of data. And we have considering, fits involving straight lines. So, there are 4 possibilities. The first possibility is that, 2 distinct lines, 2 distinct lines one is beta naught, plus beta 1 x, for the first set A and then, gamma naught plus gamma 1 x for the second set. So, here this involves 4 parameters, 1, 2, 3 and 4. And here I am talking about the case like, you have 2 sets of data and we are fitting 2 different line for 2 sets, ok. So, suppose this is the line, beta naught plus beta 1 x for set A and suppose this is the line, this is the fit for, y equal to this, y equal to gamma naught plus gamma 1 x, for set B. So, this is the first case.

(Refer Slide Time: 30:53)

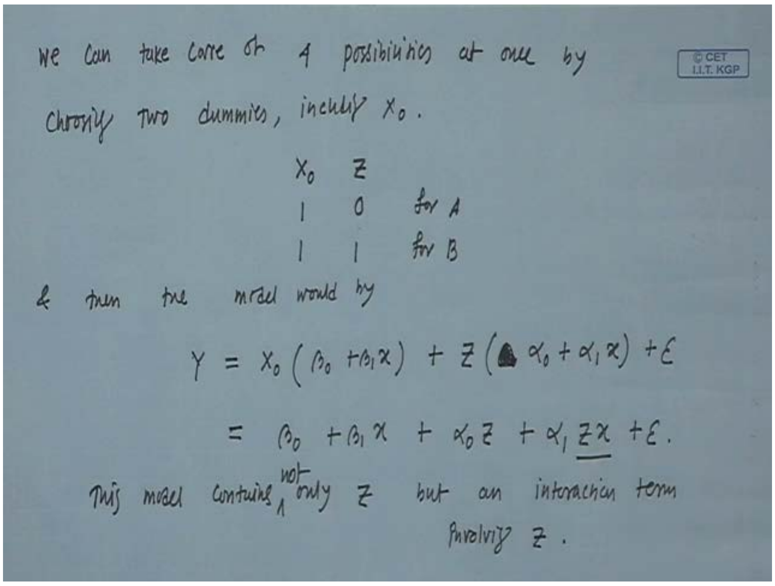


Now, the second case is, case B is that two parallel lines, ok. So, the first line is beta naught, plus beta 1 x, for set A and the second plan is gamma naught, plus beta 1 x. They have the same slope. So, here you need to estimate 3 parameters, ok and here I am talking about this situation. Suppose, this is my line, beta naught plus beta 1 x, this is for the y equal to this. This is for the first set A and I am fitting another line, with the same slope, these 2 are parallel, y equal to gamma naught plus beta 1 x, for set B, right. So, this is the second possibility. The third possibility is two lines, with the same intercept, ok.

First line is say beta naught, plus beta 1 x. This is for the set A and the second one is beta naught, same intercept beta naught, plus gamma 1 x. The slope could be different, so, this is the possibility, third possibility. So, here this is my say, y equal to beta naught, plus beta 1 x and the other line has same intercept but, different slope. So, this is beta naught, plus gamma 1 x. For this is for block A. This is for block B, y equal to this and the 4-th possibility is, is one line ok. So, let me write here only, one line. So, we are fitting in the same line, or just one straight line it fit for both sets of data.

So, here I did not say how many parameters 1, 2, 3. So, 3 parameters here and 1 line and the line is beta naught, plus beta 1 x. This is the line we are fitting for both the sets A and B and here is the graph. So, this is the line, y equal to beta naught, plus beta 1 x. Well, so, we talked about all the 4 possible situations, for two sets of data involving straight line fit. Now, we will be looking for a model involving dummy variable. A single model which can take care of all this 4 possibilities.

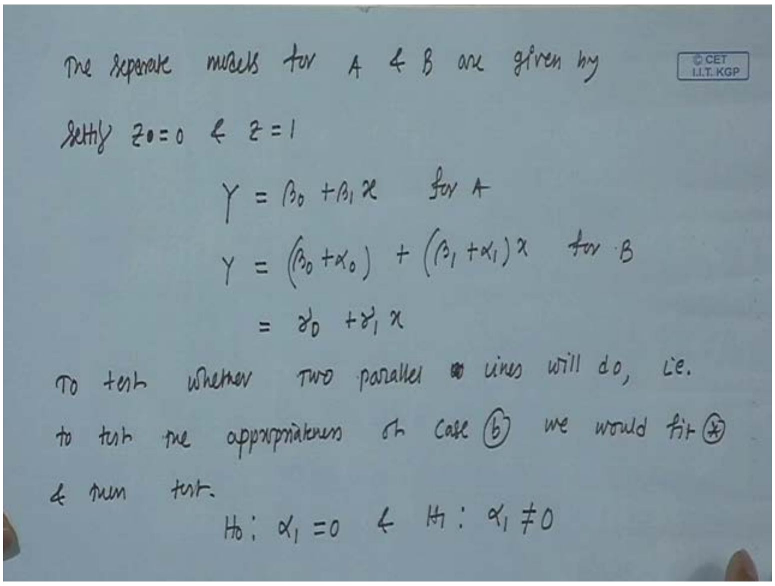
(Refer Slide Time: 35:04)



So, here is the model. So, we can take care of 4 possibilities. At once by choosing two dummies, including X naught. So, basically one dummy and then X naught. So, X naught and Z and Z equal to 0, for block A and Z equal to 1, for block B. And as usual in the X naught is always equal to 1. And then, the model would be Y equal to X naught, beta naught plus, beta 1 x. This is for the first block. Plus Z alpha naught, plus alpha 1 x, plus epsilon, ok and this can be written as now, in anyway X naught is equal to 1. So, I can write this X equal to beta naught, plus beta 1 x, plus alpha naught Z, plus alpha 1 Z x, plus epsilon.

So, this is the model, which involving one dummy variable, which can take care of all the 4 possibilities. Let me explain it. So, here one observation you can do that, this model contained. So, this model contains, not only Z but, an interaction term. Here is the interaction term Z X, involving Z, ok. So, this is the final model and I mean, this is the model, we have decided. And now, if you put Z equal to 0, you will get the model for A and by putting Z equal to 1, you will get the separate model for B.

(Refer Slide Time: 39:19)

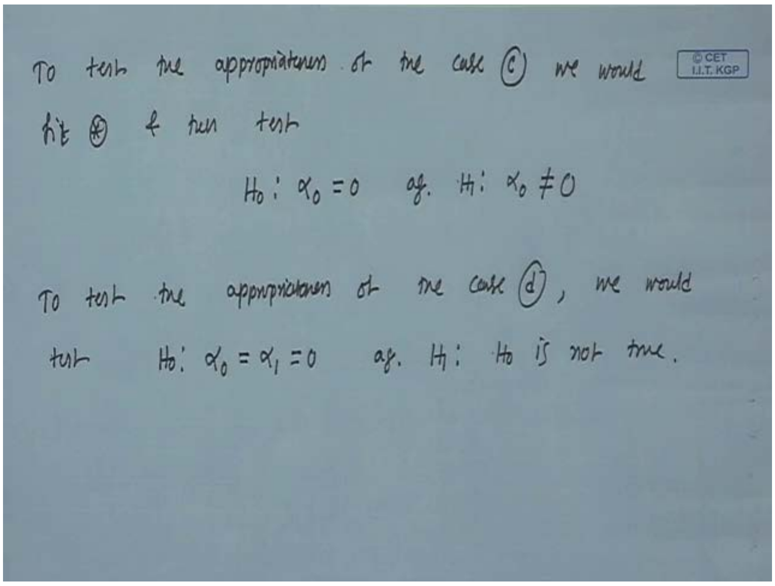


So, the separate models for A and B are given by setting Z equal to 0 and Z equal to 1. So, if you put Z equal to 0, in that model you will get, Y equal to beta naught, plus beta 1 x. That is the model for A and Y equal to beta naught, plus alpha naught, plus beta 1, plus alpha 1, x for B. So, which is you know nothing but, gamma naught plus, gamma 1 x, model for B. Well now so, what given a two sets of, or 2 data sets you can go for this general model and then you fit the model. And then you test whether, you need two separate now, you can go for 2 separate testing.

We really need different straight line fit for two sets, or we can go for a parallel lines. Whether you can go for single line, or you can go for two lines, with the same intercepts. All this 4 possibilities we can check now. To test, whether two parallel lines will do, that is to test the appropriateness of case b. We would fit the model, let me call this model star (Refer Slide Time: 35:04) we would, fit this model first; you consider the general model you fit this model first. So, this is general model you fit this model first and then you test whether you can go for, I mean whether 2 parallel lines are enough or not. So, first we would fit star and then and then, test this hypothesis H naught.

Well so, H naught which hypothesis you need to test? You need to test whether they are parallel. That means you have to test whether alpha 1 is equal to 0. You test H naught, that alpha 1 is equal to 0, against the alternative hypothesis, alpha 1 not equal to 0. So, I hope that you know how to test this one, ok. I will talk about some more examples later on. So, this is the test you need to do. To test whether, you can go for second case. That means, whether you can go for two parallel lines, right. And if this is rejected, this null hypothesis is rejected, that means you cannot go for two parallels lines. So, we have to fit two different lines, I mean we have to go for some other cases.

(Refer Slide Time: 43:48)



Now, to test the appropriateness of the case c, what is that? c is two lines with the same intercept, ok. To test the appropriateness of case c. We would first fit star the general model and then test this hypothesis, H naught. So, what hypothesis we have to check, is that (Refer Slide Time: 39:19) whether they have the same intercept but, different slope.

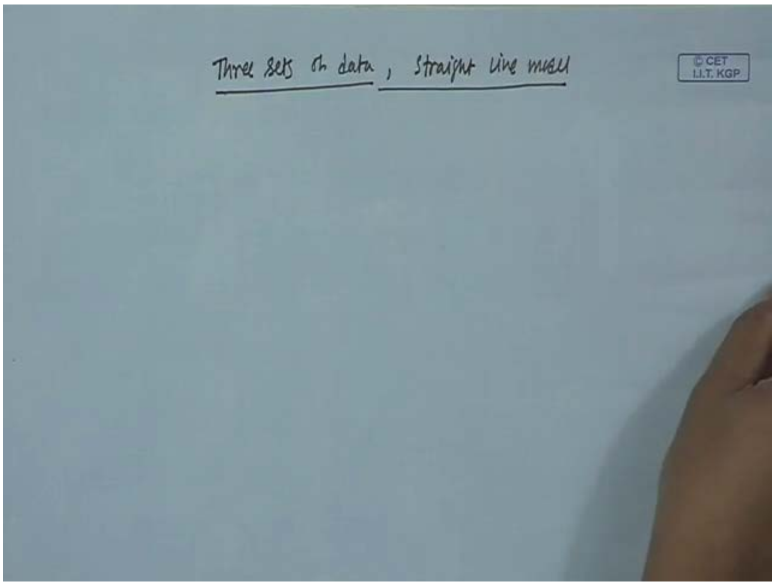
So, you have to check, if you test whether alpha naught equal to 0, or against the alternative hypothesis that alpha naught, is not equal to 0. So, to test the appropriateness of case c. You have to test the hypothesis alpha naught equal to 0, against the alternative hypothesis H 1 alpha naught, not equal to 0, ok. And finally, to test the appropriateness of the case d, that means you go for the identical model. We would test, what we have to test is that whether we can go for identical model. That means you have to test whether, alpha 1 is equal to alpha 2, is equal to 0, ok. So then both the models are same.

So, you have to test the hypothesis. You test H naught that alpha naught, equal to alpha 1, equal to 0, against the alternative hypothesis H 1 that h naught is not true, ok. Well so, let me summarize this part. So, here now, we are given two sets of data and we are trying to fit a general model involving 2 dummy variables. Such that we can cover, so, we talked about model which can cover all the 4 possibilities and then, you fit this model fast. For your given data sets, you fit the model general model involving the dummy variable and then you go for several testing like, whether you test the appropriateness of case b, if that is rejected.

Suppose, if that is, I mean rejected means, H naught is rejected. That means what is that? (Refer Slide Time: 39:19) H naught is rejected. So, H naught is rejected, that means; the lines are not parallel, ok. So, you cannot go for two parallel lines, for two sets of data. Next you check the appropriateness of case c, here. If your null hypothesis is rejected. That means you cannot go for two straight line model, having the same intercept and also if the last case is also rejected. That means, you cannot fit two, I mean you cannot fit same straight line for both the data. So, you fit the general model and then you go for testing to test the appropriateness to case b, case c and case d.

If, all of them are rejected, then you go for two separate, I mean two distinct straight lines fits for two sets of data, ok. So, this is what, this is how, I mean we have generalized the case here.

(Refer Slide Time: 49:46)



Now, let me talk about instead of two sets, if we have say 3 sets of data and also trying for the same straight line model. Then 3 sets of data say a b c and we are going for, straight line model, ok. Perhaps you know I should not start this one, because I do not have time today. So I will talk about now, how to fit a general model for three sets of data? Instead of two sets, you have three sets of data. Now, how to, what is the general model for that? Which perhaps covered all the possibilities; ok. So, that we will be talking in the next class.

Thank you very much.