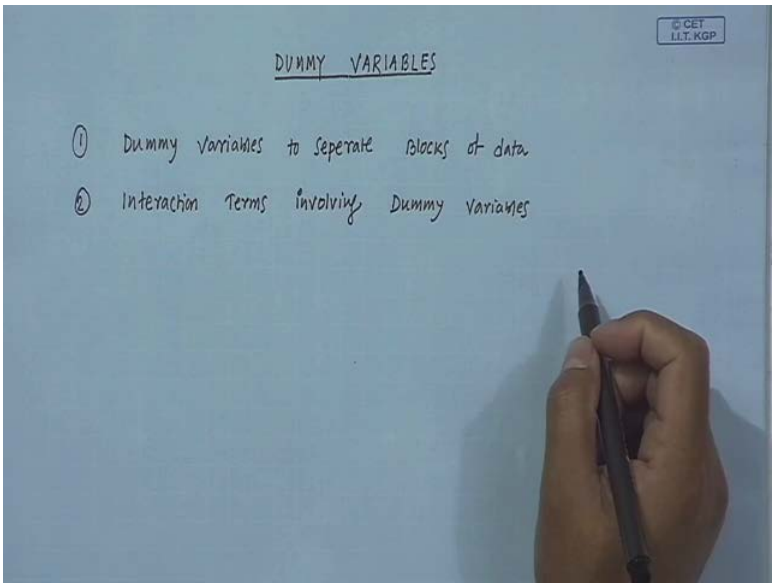
Regression Analysis
Prof. Soumen Maity
Department of Mathematics
Indian Institute of Technology, Kharagpur

Lecture No. - 24
Dummy Variables

Hi, today we will start a new model called dummy variables and

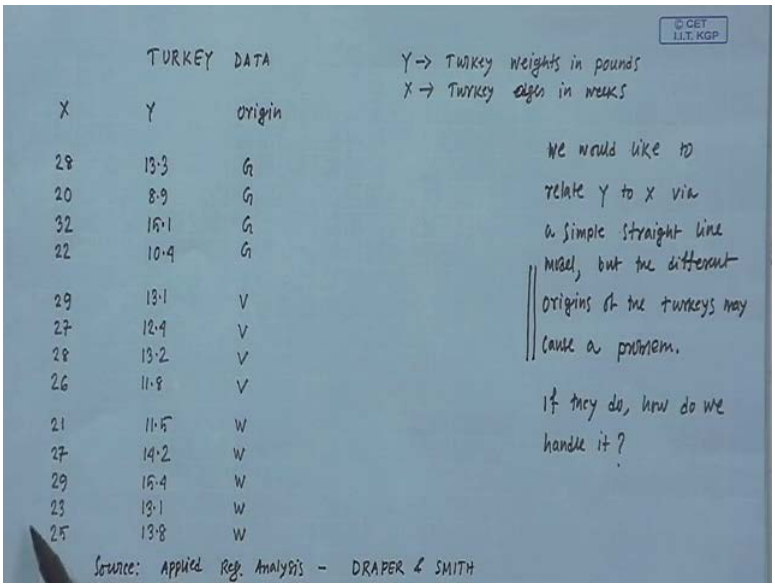
(Refer Slide Time: 00:28)



Here is, content of this module: Dummy variables to separate blocks of data and interaction terms involving dummy variables ok. Let me explain what is the objective of this module. What happen in the regression analysis? In most of the cases, we use quantitative variables and this variables have well defined a scale of measurement. For example, the variables you considered here like: weight, age, height or temperature, pressure or may be income, expenditure all this things. But, occasionally it is necessary to use some qualitative variables like, the variable like you know employment status whether the i-th person is employed or un employed or sex may be whether male or female. It could be marital status also, it could be origin also like whether the observation is from city say from Calcutta or from Mumbai or from Delhi.

So, these are called sort of qualitative variable and the objective of this module is to use how to incorporate this qualitative information in regression analysis.

(Refer Slide Time: 02:59)



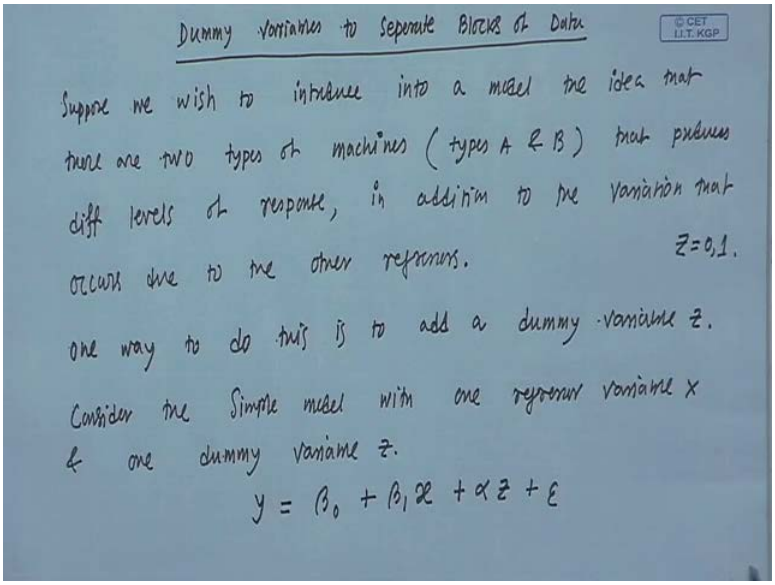
I mean well, let me just give a examples to illustrate qualitative information in the regression model. Let me consider turkey data. Well, so here this response variable Y, it stands for the turkey weights in pound and the regression variable X, it stands for the turkey age in weeks. Well, and here you can see that the 1st 4 observation, are from city called Georgia and the next 4 observation are from Virginia and the last 5 observation are from Wisconsin. And of course, I should mentioned that this data is from a book called applied regression analysis by Draper Smith. Well so, what you want is that so, we have a response variable, we have the regression variable. We want to fit relationship between straight line fit may be for example, between Y and X. So, what it says that we would like to relate Y to X via a straight line, simple straight line model. But, the different origin of the turkey may cause a problem.

Let me, explain this part. You know what is the problem here is, that there might be significant difference in response level, for different origin and that information we need to incorporate that part. I mean we cannot if you, if equal you can fit a model for this block fast and then you can fit a simple straight line model for this block. And finally, another simple linear regression model for the last block. But, that you do not want. I mean, you want to fit a single straight, single model between X and Y and also there will be some problem with that. I mean if you consider all the data together and fit a single model there could be some problem, will explain that part. So, what you want is that we want to fit a single model and at the same time you want to incorporate the qualitative

information we have. Like you know, they are from 3 origin, 3 different origins and there could be significant different in the response level. So, we need to incorporate that part. So, it is a nice idea here.

You know, how to incorporate that information that there are from different origin then will fit a single model to handle this situation ok. So, let me talk about the model we are going to fit.

(Refer Slide Time: 07:10)



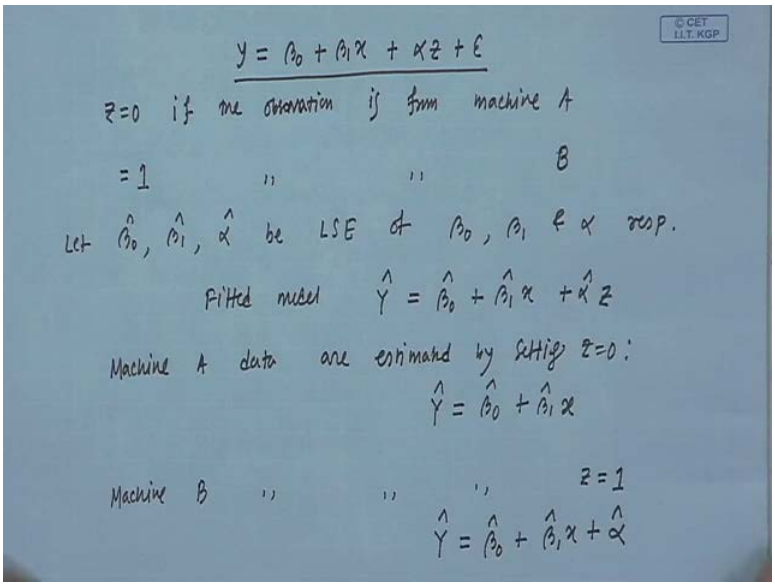
So, dummy variables to separate blocks of data. So, block means this could be different origin. This could be for different employment status. One set of data for employed person, one set of data for unemployed person, one set of data for male and one set of data for female. This is what we mean by block. And, there could be you know could be significant difference between the response level between, it is a male and female ok. So, you have to incorporate that part also in the model, by using dummy variables ok

So suppose, we wish to introduce into a model the idea that there are two types of machines say types A and B that produces different level of response, in addition to the variations that occurs due to the other regressor ok. So, what I mean by this is that there could be two blocks of data: one for machine A and the other for machine B. And, there could be significant difference in their response level, in addition to the variation due to the other regressor variables.

So, how to incorporate the information that one set of data is machine B the production of the machine A and the other set of data is production from machine B ok. So, one way to do this is to add a dummy variable say, call it Z ok. So now, will introduce a model involving a dummy variable. So, consider the simple model with one regressor variable X and one dummy variable so, Z. And, I should mention that this dummy variable can take value either 0 or 1 ok.

So, here is the model involving one regressor variable and one dummy variable. So, the response Y is equal to beta naught plus beta 1 x which is simple linear regressor model involving one regressor. And then, I will add a dummy variable Z here the co efficient alpha plus epsilon. So, the model is will write down the model again here.

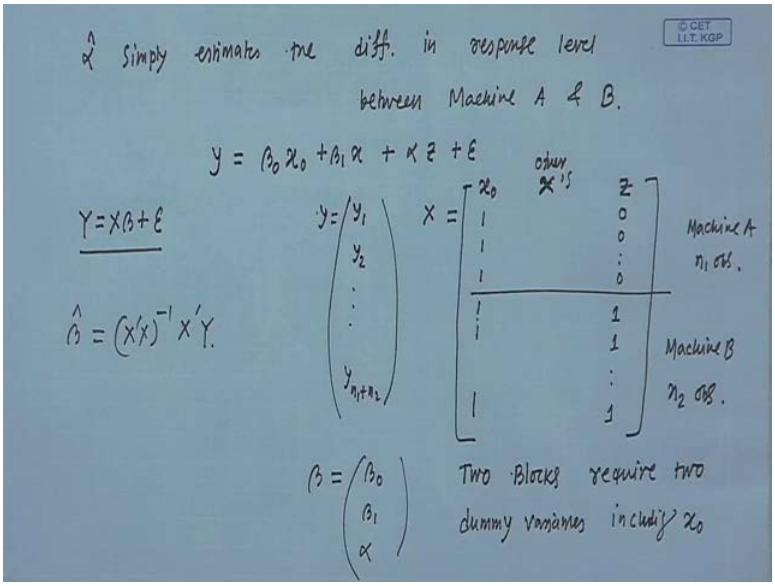
(Refer Slide Time: 12:49)



So, Y equal to beta naught plus beta 1 X plus alpha Z plus epsilon. So, this is the model. So, here Z equal to 0 if the observation is from machine A and it is 1 if the observation is from machine B. Well, so the model for block A data or machine A is basically Y equal to beta naught plus beta 1 X plus epsilon because for machine A it is Z equal to 0. And, for machine B the model is beta naught plus beta 1 X plus alpha Z plus epsilon ok. Well but, we considering a single model to fit all the observations together. Well, I will explain it again later on. Let, beta naught hat, beta 1 hat, alpha hat be least square estimate of beta naught, beta 1 and alpha respectively. Then, the fitted model is model is, Y hat equal to beta naught hat plus beta 1 hat at X plus alpha hat Z ok.

So now, machine A data are estimated by setting Z equal to 0. So, the fitted model for the machine A is basically Y hat equal to beta naught hat plus beta 1 hat X and machine B data are estimated by setting Z equal to 1 ok. So, the fitted model for machine B is Y hat equal to beta naught hat plus beta 1 hat x plus alpha hat ok. So, you can see that the fitted model for machine A and the fitted model for machine B, both are straight line and both of them are having the same slope beta 1 hat only. So, this is called intercept. Only the different is machine A has different intercept than machine B. So here, the intercept is beta naught hat where as for machine B the intercept is beta naught hat plus alpha hat ok.

(Refer Slide Time: 17:09)



So, what this alpha hat does is that, this alpha hat simply estimates the difference in response level between machine A and B ok. Well now, let me explain now, this is beautiful concept you know, you need to think it little bit to understand it more. Now, let me explain the technical part like so, you are given the model Y equal to beta naught plus beta 1 X plus alpha Z plus epsilon. And you need, I said that ok let beta naught hat and beta 1 hat and alpha hat are lea st square estimates of these parameters. Now, I explain how to do it. So, once you add a new variable, its multiple linear regressor model type of things. So here, you can write this as Y equal to X beta plus epsilon. So, here the X matrix X is equal to, what I will do is that I will write X naught here. So, X naught is sort of 1 for all observation ok. So, there is no harm in putting X 1 here.

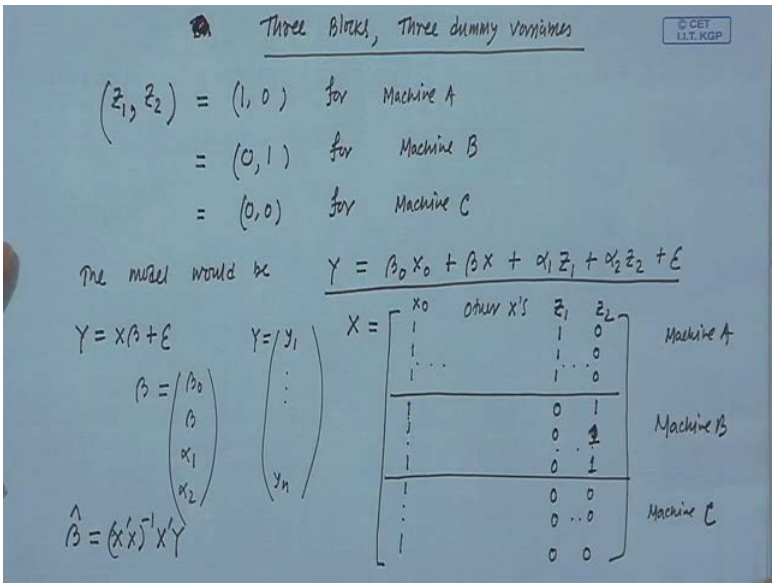
So, this is the column corresponds to X 1 and they are all equal to 1 and here is the regressor variable X or there could be several regressors. So, I should write other regressor and then Z the dummy variable. The dummy variable is 0 for machine A. So, this part is for machine A. Suppose, there are n 1 observation and dummy variable value Z equal to 1 for machine B. So, this is the X matrix. This is for machine B. And suppose, there are n 2 observation so, this is what the X matrix is.

So, you know what is X. Y is simply the response observation, response variable y 1 y 2 up to y n 1 plus n 2, I should write. And, what is beta? Beta is a vector of course, beta is equal to beta naught, beta 1 and alpha. So, this can be written as multiple linear regression model. Y equal to X beta plus epsilon by X is this matrix. Y is this matrix, this vectors and beta is this vector of the regression coefficient. And, you know how to estimate beta for linear regression model. So, beta hat is nothing but X prime X inverse X prime Y ok. This is how we estimate beta naught hat, beta 1 hat and alpha hat.

So here, I want to say that you have two blocks of observation and you need basically one dummy variable. But, I said to dummy variables takes the value 0 and 1. I mean in that since you can say this X naught is also dummy variable because it already takes the value 1. So, I will write that two sets of data or two blocks of data requires, two dummy variables including X naught. I mean I am considering X naught is also a dummy variables. So, for two sets of data we need two dummy variables: one is Z and the other dummy variable is X naught. So, X naught is equal to 1 always and Z is equal to 0 for the 1st block or block A or machine A. And, Z equal to 1 for the 2nd block or machine B.

Well, now suppose, instead of two sets of data you have three sets of data like in the example of turkey data, there are three sets of data. So, how to fit a single model for a turkey data? Because, there you have three sets of data, I mean instead of three sets could be, are any number of sets I mean any number blocks say r blocks then how to handle? How to fit a single regression model involving dummy variable to incorporate the qualitative information that they are from different block? Let me 1st talk about three blocks and then you know we will talk about n blocks in general.

(Refer Slide Time: 24:43)



So here, let me talk about how to handle two blocks, sorry I said three blocks, two blocks we are done, three blocks. And, how many dummy variables, let me say three dummy variables. So, how to do with three dummy variables? I mean three dummy variables means two real dummy variables say Z 1 and Z 2 and one dummy variable is of course, X naught ok. So, how to incorporate the information that there are three set of data and that you have to incorporate using two dummy variables Z 1 and Z 2. So, you cannot put in a Z value is equal to: 0, 1 and 2, for 3 blocks not like that because the dummy variable takes the value 0 and 1 ok. So, here is the idea.

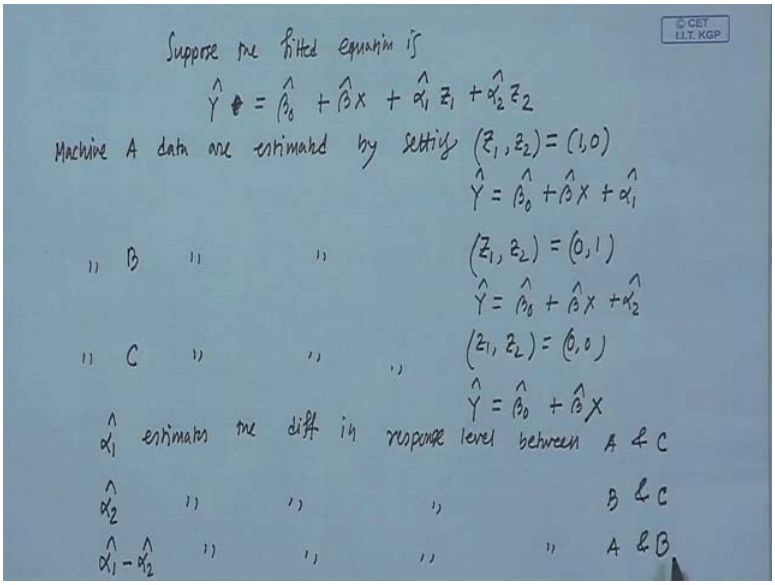
You have, two dummy variables say I mean we are going to use two dummy variables Z 1 and Z 2. So, Z 1 and Z 2 is equal to 1, 0 for; let me write machine A or block A and this is equal to 0, 1 for machine B and this is equal to 0, 0 for machine C ok. So, the model would be here. The model would be Y equal to beta naught, let me put X naught right now, x naught. So, beta naught plus beta X which is the regressor variable, plus alpha 1 Z 1 plus alpha 2 Z 2 plus epsilon ok. So, here we have, you know we have three dummy variables basically including X naught Z 1 and Z 2. And, let me explain again that you know similarly, like previous case this can be written as in matrix form Y equal to X beta plus epsilon ok.

So here, what is the X matrix here? So, X is equal to, so 1st X naught. X naught is all equal to 1 and then other X is I mean here instead of one regressor, there could be several

regressor. It does not matter you put the matrix here and then Z 1, Z 2 so Z 1, Z 2. Z 1 is 1 and Z 2 is 0 for the 1st block for the machine 1. So suppose, these are the data for machine 1 and then for block 2, Z 1 is equal to 0 and Z 2 equal to 1. So, 0 1, 0 1, 0 1. So, this is for machine A this is for machine B and for the block C, Z 1 is equal to 0 and Z 2 equal to 0. So, this is how you get the X matrix and this is for machine C. And, you know what is this Y. Y is the vector of observations, response variable Y 1 up to, let me write Y n only and beta is equal to beta naught beta alpha 1, alpha 2 ok.

So, you are all set to fit a multiple linear regression model. Because you know what is X, you know what is beta, you know what is Y everything is given here. So, you know the list square estimate beta hat is equal to X prime X inverse X prime Y. So, you can compute: beta naught hat, beta hat, alpha 1 hat and alpha 2 hat by solving this one ok.

(Refer Slide Time: 31:22)



So, suppose the fitted equation is Y hat equal to beta naught hat plus beta hat X plus alpha 1 hat Z 1 plus alpha 2 hat Z 2. So, given three sets of data we have fitted a simple regression model and see for machine A or block A data are estimated by setting Z 1, Z 2 equal to 1, 0 right. So, the fitted equation or yeah so fitted equation for block A is Y hat is equal to beta naught hat plus beta hat X plus alpha 1 hat. Similarly, machine B data are estimated by setting Z 1, Z 2 equal to 0, 1.

So, Y hat is equal to beta naught hat plus beta hat X plus alpha 2 hat and machine C data are estimated by setting Z 1, Z 2 equal to 0, 0. So, Y hat is equal to beta naught hat plus

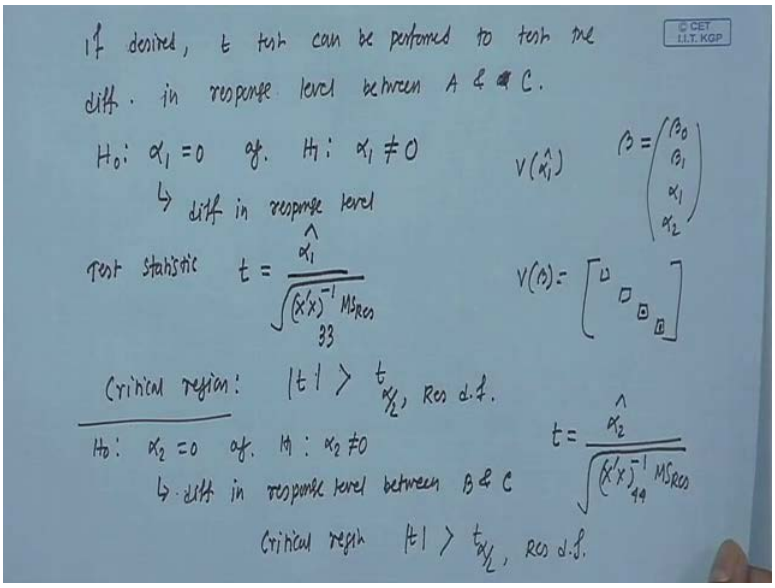
beta hat X. Well, so we have three sets of data and what we are doing is that we are fitting a single linear regression model involving dummy variable to all the data. I mean including block: A, B and C.

And finally, you can see that there are three different fitted equations. One for block 1 which is different from the fitted model for block 2 and which is again different from the fitted model for block 3. So, we get three different equations for three blocks and they are having the same slope but different intercept. And, I mean this is what we want, I mean to fit a single model involving all the data and we won't to fit a separate model for separate set of data ok, separate block, separate blocks of data ok. Now, this alpha 1 hat it estimate the difference in response level between C and A. Similarly, the alpha 2 hat is I mean this estimates the difference in response level between B and C.

So, let me write that alpha 1 hat estimates the difference in response level between A and C. Alpha 2 hat estimates the difference in response level between B and C. And then, how you estimate the difference in response level between A and B? So, the different between in response level A and B can be estimated by alpha 1 hat minus alpha 2 hat. So, alpha 1 hat minus alpha 2 hat, this estimate the difference response level in between A and B. So, if this estimate value is large then you can say that, say alpha 1 hat is large then you can say that, there is a significant different in response level between A and C. I mean again it is hard to say what I mean by for large alpha hat.

So, we need to go for statistical test to test the significant of alpha 1 hat. So, what you do is that you will test whether alpha 1 hat, the hypothesis whether alpha 1 hat is equal to 0 against the hypothesis is alpha 1 hat is not equal to 0. So, if the hypotheses is reject alternative hypothesis alpha 1 hat is not equal to 0 is accepted that means there is significant difference in response level between A and C ok. And, we know how to test that.

(Refer Slide Time: 39:17)



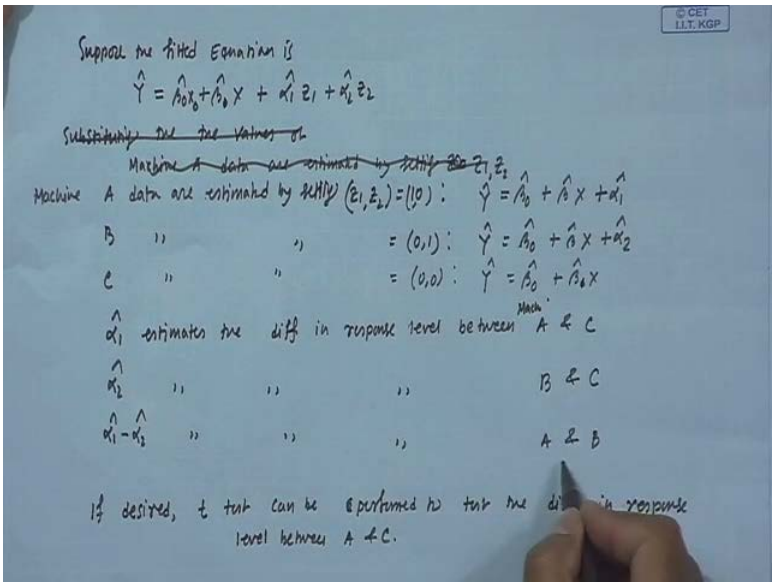
So, it says that if desired t test can be performed to test the difference in response level between A and C. So, we are formally going to test that whether the hypotheses H naught which is equal to alpha 1 equal to 0 against the alternative hypotheses H 1 which says that alpha 1 is not equal to 0. So, this alpha 1 is basically the difference in response level. So, I hope you can understood what you mean by this response level. So, if you recall the turkey data example, so suppose one set of data is from, there we have one set of data from Wisconsin and then another set of data from Georgia and the response is turkey weight. So, by this different in response level between turkey I mean between say, Georgia and Wisconsin. What I mean that is that whether the difference in turkey owlet, significant difference in turkey weight which are originated from Georgia here from the turkey which are originated from Wisconsin.

So, that is what I mean the different in response level between two sets of data and for the turkey data it is whether the yet are significant different of two different origin. So, to test this hypotheses, we know how to test this hypotheses. We go for test statistic am sure that you recall from my 1st model and 2nd model simple linear regression model or multiple linear regression model. The test statistic for testing hypothesis is that t equal to alpha 1 hat by X prime X inverse M S residual ok. I mean here basically I am looking here for the variance of alpha 1 hat and if you recall the beta matrix there, so this is beta naught, beta 1, alpha 1, alpha 2. So, in this matrix the 3rd diagonal element is the variance of alpha 1. So, this is I will put 3 3. 3 3 means it is a 3-rd.

Suppose, this is variance co variance matrix for beta and then here you have the data, the diagonal elements. So, this element is the variance of, so this quantity is the variance of alpha 1 hat. So, I denoted by 3 3 and the critical region is that you reject hypotheses if t is significantly large. Large means it modules value of t is greater than t tabulated value, alpha by 2 at the alpha level of significance and the degree of freedom is residual degree of freedom ok. I hope that you can ((Refer Time: 44:11)) you remember all this things. This is how we test a alpha 1 so, this is what we are testing for the significant different in response level between A and C. Similarly, we can test for H naught: alpha 2 equal to 0 against H 1, alpha 2 not equal to 0. And, this one is, this one stands for difference in response level between B and C so, to check whether there is the significant different in response level between that data from B and from data in C ok.

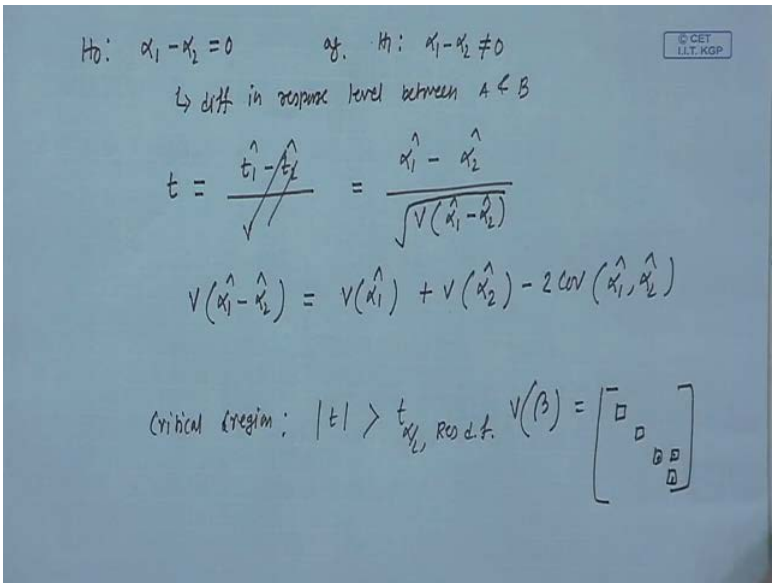
So, the same test statistic so, t is equal to alpha 2 hat by X prime X inverse and M S residual and this is the 4th parameter regression co efficient. So, the variance of alpha 2 hat is basically, the 4th element at the position of the 4th diagonal elements and here is the t statistic. And similarly, region is same, critical region is t greater than t alpha by 2 and residual degree of freedom for this distinction. And, so we are left with how to check the difference between difference in response level between A and B.

(Refer Slide Time: 46:27)



So, I told before that alpha 1 hat minus alpha 2 hat estimated the different in response level between A and B.

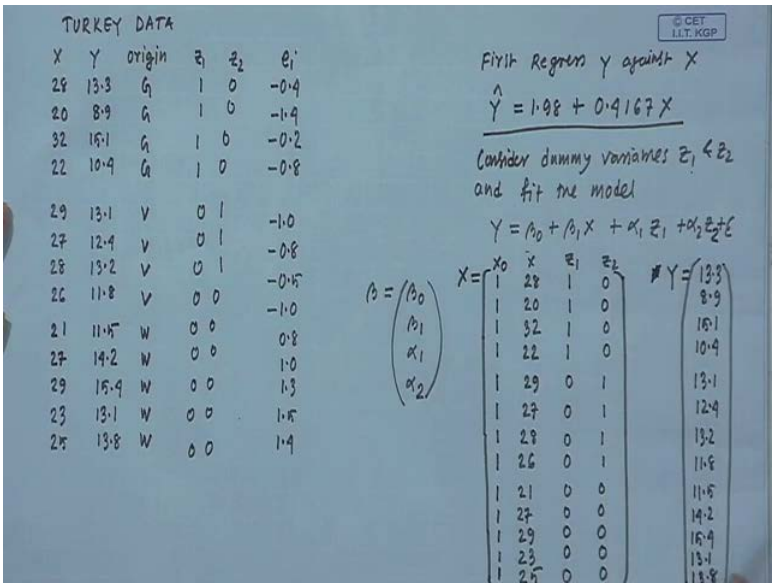
(Refer Slide Time: 46:40)



So, what will do here will check H naught whether alpha 1 minus alpha 2 is equal to 0. So, this is the difference in response level between A and B. So, if this is accepted that means there is no difference in significant difference in the response level between A and B against the alternative hypotheses H 1, which is alpha 1 minus alpha 2 is not equal to 0. So, here also we will use t statistics. That is t equal to alpha 1 hat minus alpha 2 hat by variance of alpha 1 hat minus alpha 2 hat square root. And, this quantity you can compute, because variance of alpha 1 hat minus alpha 2 hat is equal to variance of alpha 1 hat. So, you know how to compute it plus variance of alpha 2 hat minus twice covariance of alpha 1 hat alpha 2 hat.

So, if I write down variance coefficient of beta, you know so, this variance of alpha 1 hat is this element, variance of alpha 2 hat is this element and the covariance between alpha 1 alpha 2 hat is this one. Yeah I think so. So, you can compute this because you know this one and then its same you know this follows. The critical region is that modules of t is greater than t alpha by 2 residual degree of freedom. So, if this t is observed t is greater than the tabulated t you reject null hypotheses and you can say that there is a significant different between response level of A and B. Well, still you know I hope that everything is clear but I would like to illustrate the same thing for the turkey data.

(Refer Slide Time: 49:41)

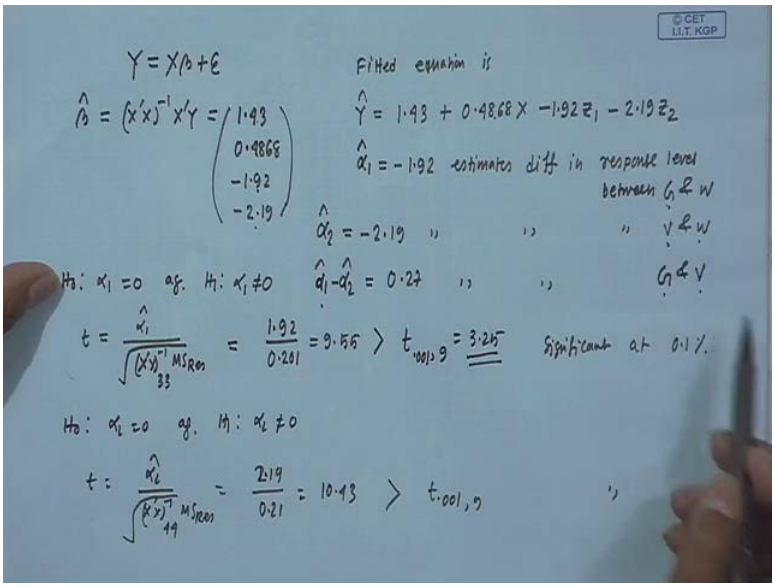


So, we have this data. I explained that this is the age in weeks and this is widths in pounds and the origin different origin. And, what you can do is that suppose, you do not know the dummy variable technique and you want to find relationship between X and Y. So, 1st regress Y against X and here is the fitted model. Now, what we do is that once we have the fitted model ignoring the dummy variable, you compute the residual. So, we can see there is something wrong residual wrong in the sense that for this block of data residuals are all negative and for this block of data the residuals are again all negative. And perhaps, these all smaller than overall compared to this and for the last block all positive.

So, this sort of indicates there is a significant difference in the response level between this three sets of data. So, that implies that you need to consider dummy variable say Z 1 and Z 2 and fit the model ok. So, constant of dummy variable Z 1 and Z 2. Here is the model I explained just now and to clear any sort of doubt I just wrote down the X matrix. So, here you see the X matrix. This one is for the dummy variable X naught which is always equal to 1 and here you have only one regressor. So, put the data here, what the corresponds to X only. So, if you have X 1, X 2 you put X 1, X 2 here. And you have the dummy variables Z 1 and Z 2. So, for the 1st block it is all 1 0, 1 0, 1 0, 1 0. For the 2nd block it is 0 1, 0 1, 0 1, 0 1 and for the last block it is all 0 0, 0 0, 0 0, 0 0, 0 0.

So, here is the X matrix. So, you know what is the X matrix and here is the Y vector. Observation of response variable is same as this one and you have four parameters here: beta naught, bet 1, alpha 1, alpha 2. So, you are all set with to fit the multiple linear regression model

(Refer Slide Time: 53:10)



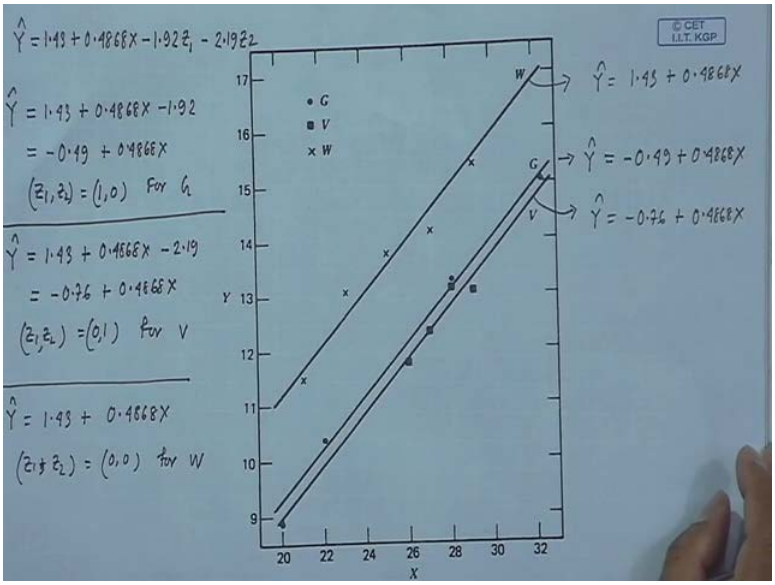
Y equal to X beta plus epsilon. So, here you do all this calculations. Here is the estimated parameters. Now, you see, so the fitted equation is this one. This is beta naught, this is beta 1, alpha 1, alpha 2. So, this is the fitted equation. As I told that alpha 1 estimates difference in response level between G and W that is block A and block C perhaps. And, alpha 2 hat estimates different in response level between V and W that is, B and C. And, alpha 1 hat minus alpha 2 hat that is equal to 0.27 that estimates different in response level between G and V. So, you can, see looking at this data you cannot say whether this quantity or this alpha 1 modules value is significantly big or not. So, you can go for formal test. So, you test for alpha 1 equal to 0 against H 1 is alpha 1 is not equal to 0.

So, you go for the t statistic and here is the t value and the degree of freedom is 9 here. I hope you can recall, why it is 9. Because there are 13 observation total. And since, so 13 observations means 13 residuals at there and since there are 4 regression co efficient. There will be 4 constraint on the residual. So, you cannot choose all the residual independently. There are some restriction, there are four restriction. So, you can choose, you have the freedom of choosing 9 residuals and then the remaining 4 have to be chosen

in such way that all this 4 restriction satisfied. So, that is why the degree of freedom is 9. And, tabulated value is this much which is much smaller to this observed value.

So, this concludes that alpha 1 is equal to 0 is rejected and that means this is accepted, which means there is the significant difference in response level between the turkey from Georgia and Wisconsin. Similarly, you test for alpha 2 and to test the difference between this V and W and this test will give you difference in response level between G and V.

(Refer Slide Time: 56:33)



And here, you can see that basically you know if you put, so this is the fitted model and if you put Z 1, Z 2 equal to 0 1, here is the fitted model for Georgia. If you put 0, 1 in this model you get, this is for Virginia and this is for, if you put 0, 0 here you get this model for Wisconsin. And you plot them, getting ultimately you are getting three different fits and their parallel. But, there have different intercept and if you see them carefully you can identify, you know the difference in response level between the turkey originated from three different origins. So, we have to stop now.

Thank you for your attention.