**Engineering Econometrics**
**Prof. Rudra P. Pradhan**
**Vinod Gupta School of Management**
**Indian Institute of Technology, Kharagpur**

**Lecture - 33**
**Non-Linear Regression Modelling- Dummy-Variable Regression Modelling**

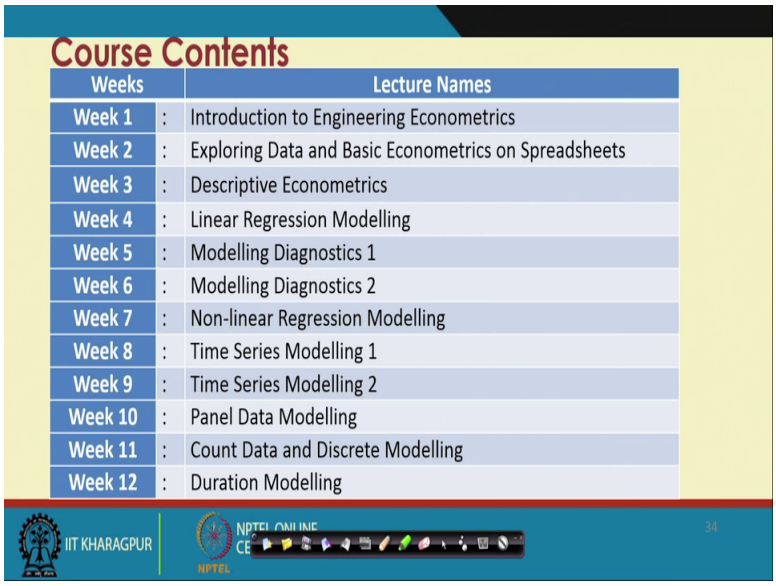
Hello everybody. This is Rudra Pradhan here. Welcome to Engineering Econometrics. Today we will start a new topic. The title of this discussion is Non-Linear Regression Modelling. We have discussed couple of you know models relating to linear regression modeling.

We started with you know simple regression modeling that too with a linearity structure, then we have discussed multiple regression modeling with more number of you know independent variables again that too with a linear structure and we have discussed or we have acquainted with the system that is related to estimation process, then the kind of you know reliability check starting with specification test goodness of fit test, then all kinds of you know diagnostics again starting with multicollinearity issue, autocorrelation issue, heteroskedasticity issue, then you know model misspecification.

So, technically we have gone through all these details so far as model building is concerned and then, the empirical testings. Here, we are connecting with an engineering theory and that too connecting with the data as per the variables requirement and then, we have gone through the linearity models and the kind of you know a reliability check process starting with you know specification to diagnostics and that too with the out of sample prediction test.

So, we have gone through all these details and then, we like to know you know how decision making process will be with respect to the theory, the empirical testings and to what the data tells about this particular you know problems and how the estimated outcomes you know behaves that together the estimated outcome which is derived through the empirical process that too as per this choice of the linearity model and the kind of know availability of data and the idea is that to have the estimated output and then, compare with the existing theory whether it is the, it is supportive or it is going against the theory. So, all kinds of you know discussion we have already done.

(Refer Slide Time: 03:09)




So, that means technically if you go through this you know contents, so we start with you know introductions to the basically about the Engineering Econometrics exploring data and knowing the use of you know spreadsheet and softwares to drilling to understand the data. Data visualization, then starting with descript econometrics and the linear regression modeling and then, modeling diagnostic one model like modeling diagnostic tools and then, we are in this you know structure what we called as you know non-linear regression modeling technically, whether the model I mean modeling structure is the linear one or the non-linear ones or whether it is a time series structures or panel data structure or it is kind of you know discrete structure.

So, every aspects of the process of empirical testing is more or less same and what is the kind of you know requirement you must have engineering theory and after understanding the engineering theory or problems, you have to transfer into you know models that we have to develop a model where all kinds of you know variables will be identified to analyze this particular theory and then, we like to gather information behind all these very variables relating to cross-sectional specification times, the specification, full specification, panel specification and then, we have gone through the kind of you know empirical process.

We are using model a you know model and data. We will have estimated output and then, we will go through all kinds of you know check process and you know finally declaring

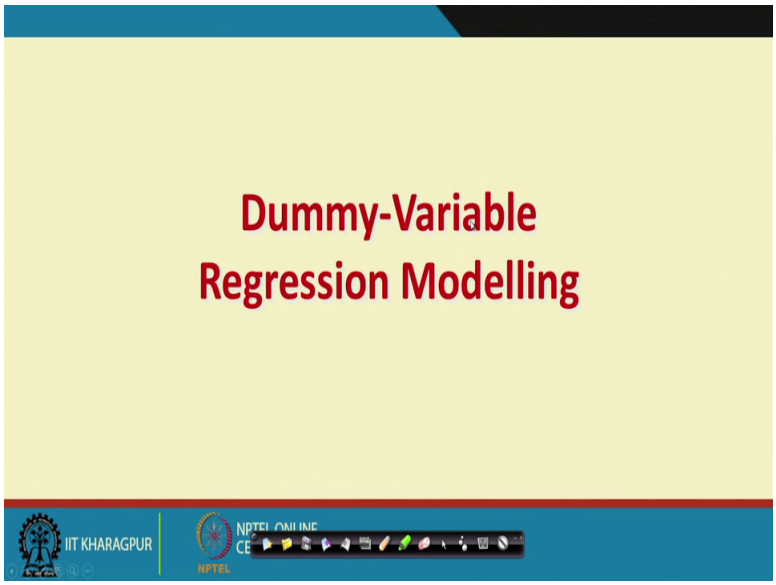
that the modeling is good enough to do the kind of you know for castings or as per the decent you know management decision making process is concerned.

So, this means here what we like to do, we follow the similar kind of you know structure, but here the framework will be a non-linear setup. Of course, we have gone through slightly non-linear setup earlier by highlighting different functional form and that too in the case of you know misspecification inspection and here we will follow the same route and then, we like to elaborate the modeling structure in a kind of you know elaborate way by using various non-linear in the non-linearity structure and what we can you know do here, we start with actually you know a simple one, then you know you know integrating with you know complex one and to understand you know you know structure of the non-linearity, we can start with a dummy modeling first, because dummy modeling has a kind of you know cluster of you know means it is a kind of you know mixture setup, where we have a linearity structure and as well as a non-linearity structure.

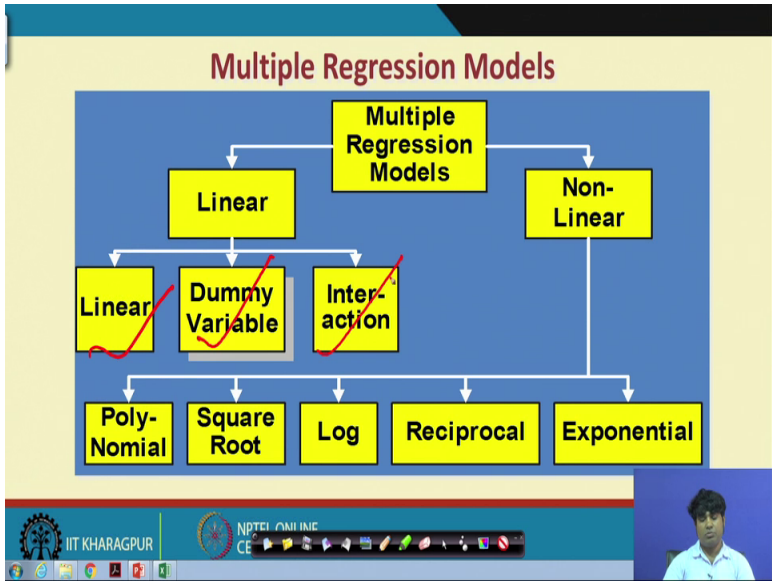
The model itself has given you know giving you a different kind of you know you know structure, where you can actually understand the non-linear structure and the similar kind of you know problem can be compared with the linearity structure with the various you know restructuring process. Let us start with first slightly about the dummy modeling, then we will move to the various functionality or you know various forms of you know non-linear regression modeling.

So, let us see what is all about the particular you know process. So, the process will be, so let me start the kind of you know, ok. So, the process will be here is the dummy, dummy variable regression modelling, then we will go to the other forms of you know non-linear models, we can actually go through details. What I can say the classification of you know multiple regression modeling that too linear versus non-linear and you know we with respect to various forms.

(Refer Slide Time: 07:06)




(Refer Slide Time: 07:12)




In the first instance, the classification will be like this linear and non-linear and then, in the linear setup, we have in fact a simple linear structure which you have already discussed then by using the dummy variable structures and then, there is a structure of interactive effect and that too it is with respect to linear set up by using simply y with x, then y with you know dummy variable set d and then, the interactions may be y with the x d and d x the kind the item d x will be the interactive effect. For instance, predicting you know cricket score. Then, you know the linear setup may be with respect to number of batsmans in a scoring contribution to the total score and dummy variables will be huge

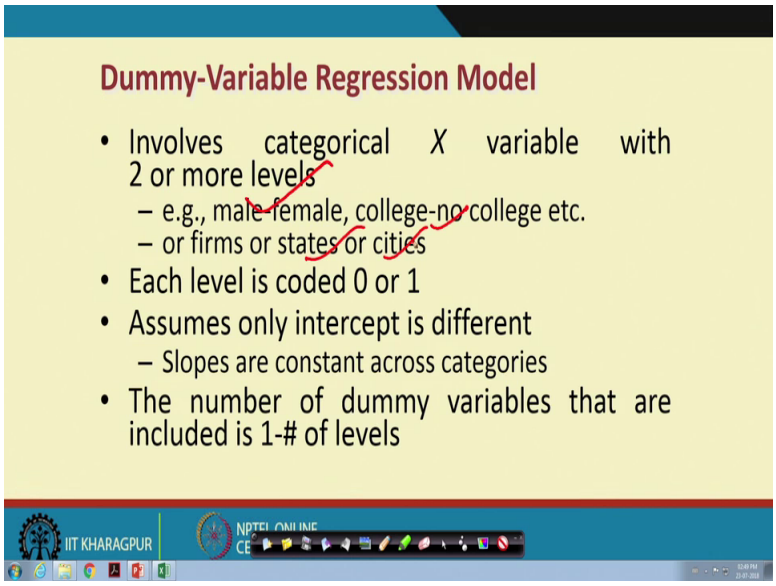
in the sense you know whether there is a kind of you know no ball, you know wide or not, no wide like that and then interaction effect that is with respect to both the variables batsman score and the kind of you know no wide of sums that will be the interactive or in fact, you know you can cause you know use a you know partnership contribution to total score can be interactive effect.

So, this will be you know you know means the problem will be like that only and in the non-linear set of, so we have various forms starting with the polynomial regression in the form of square root log, logarithmic format, reciprocal from art and exponential format again little bit we have discussed earlier while you know addressing the specification that with respect to typical functional form.

So, the requirement may be non-linear and if we start with the linearity and go ahead with the estimation process and forecasting process, so we will find you know some kind of you know specification bias because of you know choosing the wrong functional form. So, that is why we have slightly discussed all these you know other forms of you know linear, you know modeling that too you in a kind of you know non-linear format.

So, let us move with this is that you know how we can do the kind of you know structuring. So, what we can do here, we they start little bit you know discussing the concept of you know dummy modeling, then we really discussed the concept of you know non-linear more modeling that too with respect to the use of you know dummy structure and then, non dumbest rupture that with you know respect to various functional form.

(Refer Slide Time: 10:15)




By the way 1st question is you know first understanding must with respect to dummy. A dummy is a kind of you know proxy variables or it is a kind of you know we can in a use like you know or you can call as you know categorical variables which initially may be in the form of you know qualitative in natures and then, through some kind of you know kind of you know structuring and restructuring, we specify the model in a quantitative format.

Ultimately, we need actually quantitative informations about the variable to do the you know estimation that to why we get by obtaining and obtaining the regression put through a particular software and for that if you put you know categorical or qualitative variables like let us say gender, then the answer will be simply male and females. If we write male and female in the excel sheet, then software cannot run even manually. You cannot actually do the processing because ultimately we need actually some kind of you know mathematical operation like you know summations and then, you know you know cross product summation, dot product summations in this kind of you know situations. So, variables in information should be actually numeric in nature.

That means, you know it should be in a quantitative format, so that we can do the operation and finally, have a kind of you know summation to know the values of the parameters and then, go through the kind of you know reliability test or diagnostic test

Etcetera and that is how first understanding is the kind of you know dummy understanding and technically.

So, we have 2 variables all together in the regression framework that two dependent variables say y and independent variable say X. So, in the X sides the variables you know maybe dummy, that means maybe you know one dummy or two dummy, dummy which actually other independent variables, dummy without other independent variables and again the dependent variable can be also categorical and domain nature. For instance, you know yes no type of things you know for instance people having bank account or depends upon means if the problem will be what are the factors having actually you know bank account of a household. So it is mostly because it means the problem will be like you know to understand the financial inclusion plan.

So, one of the question you know variables information will be whether people have account or not. Not having account and for that what are the variable responsible and in that case, here you know dependent variables definitely will be categorical and if people have a bank account, then we can say yes and people have no bank account, they can go for you know no. So, then the further transformation to do the processing or testing, the yes situation can be represented as 1 and by default no situation will be represented by 0 or vice versa can also be true.

So, I saw some of the examples we have written here is like you know male, female and you know like you know or various forms, various states, various cities and religions and then, the kind of you know or you know different qualifications like you know BTech, MTech or some kind of you know literate, illiterate kind of you know things every times.

So, the first you know requirement is you know you properly name this particular you know categorical variables, understand the particular categorical variable and then, you specify the coding and then, you transfer the qualitative information into coding that too representing the same variables with the help of you know quantitative informations. So, now if it is the question of you know gender, then the classification will be male, female.

So, if male will be coded as 0, a female may be coded as 1 and vice versa can also be true. Similarly, you know Indian, non Indians, then this can be classified with you know 0, 1 again. So, every time the kind of you know transformations you can do right either with you know 0, 1 or 1, 2, 3 something like that. For instance, if it is a gender, then one

way to go for that you know classification is 0, 1 or 1, 2 and if it is say religion, then depending upon the number of religions and that too involves with you know a particular sample. So, you can put you know different kind of you know coding 1, 2, 3, 4 depending upon the number of religions which you have in the data set and then, we will do the empirical processing in order to check you know whether religion having an impact with any dependent variable.

If it is the region have been part and what religion you know what kind of you know impact with respect to different religions, so how we will do all these things. So, we can get to know from this you know modeling and that too in this lecture assume only intercept is different and slopes are constant across the categories and the double, the number of dummy variables that you know included you know in this particular you know say you know genders can be one's, but depending upon the problems and the requirement, it can be more than that.

For instance, in a particular you know problems you may ask for you know religion impact, gender impact, literacy impact. So, these are all actually can be used or can restructure the process with you know different domains. So, every day dummy has a different kind of you know specification and representation for gender, then the dummy description would be different or it can be used like that 0, 1 or you know 1, 2 something in that format and then, for religions depending upon this particular you know number of religions, you can code accordingly and you like to know whether the religion is having impact and if there is impact which religion having in a high impact or which religion having actually low impact.

So, all these details can be you know analyzed through the dummy modeling and then, interactive effective it can be huge. For instance, if you have two dummies say gender and you know you know you know say religions, then the interactive effect will be in a particular religions how the gender impact will be there. So, that means D1, D2 and with you know let us say D1 is Hindu, then under Hindu some you know sample we like to check how many are male and female. Similarly let us say another religion Muslim.

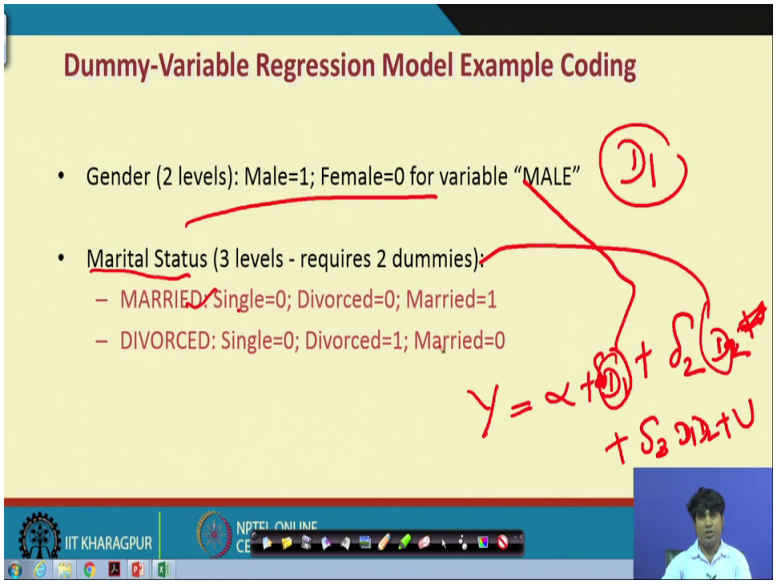
So, you have to check the gender impact against female. So, the 1st dummy will be religion and first of all you identify how many religions are there in the sample, then for every religions, we have 1 dummy and then, again for that religion the specification will

be d1 another kind of you know dummy with the specification of you know male, female.

So, then D1 and D2 can be clubbed. So, this will be interactive effect. So, that means the interpretation will be the impact of you know Hindu male and Hindu female. So, that is the interactive effect through which you can actually regress again with you know dependent variables. So, that means dummy variable like have you know this kind of you know advantage to create a non-linear structure. So, when we are you know you know addressing the interactive effect like you know whether you know let us say income is a dependent variable and then, whether you know Hindu male has a high contribution or Hindu female has a high contributions, again whether Muslim male has a high contribution or you know Muslim female has a high contribution towards you know income.

So, if these are the things we can actually easily check and that too in this kind of you know situations. So, we can use the interactive kind of you know structure. The moment will be an integrate, interactive effect to the dependent variable, then by default that model can be one form of you know non-linear regression modeling. That is a simple way of you know starting the non-linear regression modeling that too with the use of you know dummy variables. So, the idea is exactly like this.

(Refer Slide Time: 19:40)



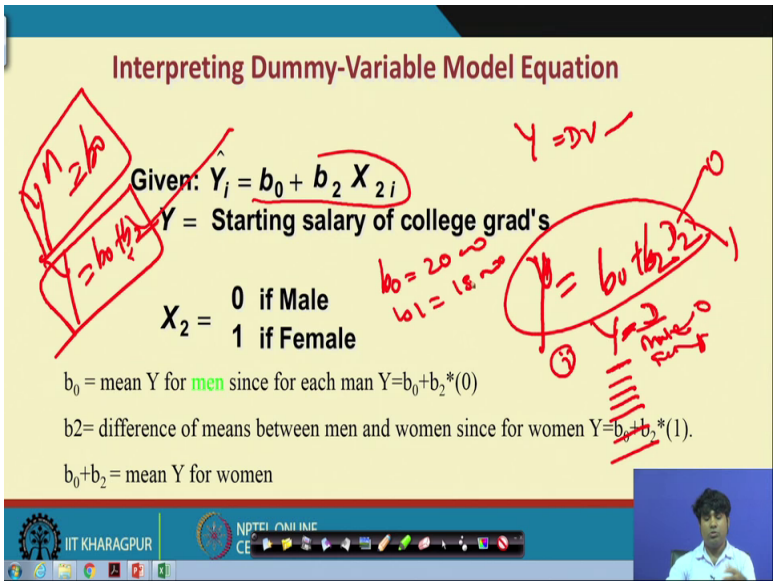
So, let us see oh you know one after another example. So, like you know it is a religion and gender against another way of you know addressing this particular non-linear modeling is you know gender with you know you know marital status. So, that means here a gender having actually two levels of you know classifications; male and female. So, for that you know say what we can say we have actually male one and female one. So, in this case, so male can be represented as 1 and female can be represented as 0. So, it can be treated as you know D1 say this is the dummy variable D 1s and your models may be equal to Y equal to say alpha of loss you know D 1s delta 1 D 1s, ok. Delta 1 D 1s plus delta 2 D 2 plus error terms, ok.

So, this will be the kind of you know models and this first D 1s will be represented as a gender and second D 2 represented as a marital status and then, we can create 3rd dummy. So, let us say delta 3, then D1 D2 that is the interactive effect and then, the error term here. We can remove this error term here, then this could be a non-linear model. So, now in this case you have a marital status. So, married thing is you know that could be single and divorce and then, you know married.

So, we can put actually 0, 1, 2 or simply called as you know whether actually you know say married or you know say unmarried or singles we can say, then in that case the classification will be like that. So, again what is happening you know say that means the kind of you know you are staying alone. So, you will be staying alone when you are not married and if you are married and with you know divorce, then this will be one case.

So, now we have to see two situations if you married, then you stay together. This is one level and if you are in a singles either you are not married or you know say divorce, so just again staying you know single. So, with you know in that case it will be like you family, the 2 members here the family with 1 member. So, family with a 2 members and whether male and female, so again single with male or females. So, there then you will add this will be kind of you know interactive effect. So, likewise you know various forms of you know an interactive effect can be you know used and can be checked. So, it is all actually once you know the simple structure of you know modeling.

(Refer Slide Time: 22:34)




So, introducing dummy and bringing the interactive effect that too representing the non-linear regression modeling, it is very easy because you know just to add a particular you know variables and then with the specification of dummy and for that you know theoretical, it should be you know connected or it should be well addressed, otherwise there is a high chance of you know misspecification and without objective, a specific objective and the kind of you know specific requirement you should not use the dummy and that too transfer this simple model into the kind of you know non-linear models.

In order to understand you know you know a little bit more you know a elaborative way, let us have an example here. So, this is a kind of you know estimated model here and here the estimated model will be Y, Y that equal to b 0 plus b 2 X 2 and in fact, here the Y is the dependent variables which is actually is a salary of a kind of you know called as you know a student and then, what are the factors which can be responsible for that and there may be many factors which can be responsible for that, but we are interested here to know the actually gender impact.

So, then you simply actually start with a dependent variables Y that is actually salary of you know all these you know students those who are you know in job after the college and then, we are checking whether actually gender has an impact. So, that means which particular group is having high average income after the college education and they are in the job. So, for that you know you simply represent the model like this Y equal to and

then, you know say b 0 plus b 2 or we can either put X 2 or you can simply put actually D2 that represents the dummy and here the data structure will be Y with you know say cross sectional data with you know different you know students you know salary after the colleges educations and then, we like to just ask you know the candidate whether you are in job and if so, what is your you know salary.

So, you represent the salary level here and in the same times, you just you know call it whether the particular you know sample or the particular respondent is male or female. If it is male, so you write in male and if it is a female, then you can write female and male can be recorded into the 0 and by default female can be coded as 1. So, then we can regress Y with D. So, initially you have one variables and then, finally with the gender as you know gender can be added a particular variable, then we can study the impact and again in order to understand further, so this is the you know understanding further means is the kind of you know interactive effect.

So, for that what you can do? We add another dummy variable, but in the mean time let us see here how is the particular structure. So, when we start with you know dependent variable with a simple dummy that to male and female, the model transformation will be having two parts. So now since it is a question of male and female, so if male is coding as 0 by default female will be coding to 1. So, in this case D2 will be ranging from 0 to 1.

So, after having the estimated model, you just put you know h 2 equal to 0 and X 2 equal to 1. In one instance when a putting X2 equal to 0, say then the model will be restricted to simply Y hat equal to b 0 only ok, then by default when you are putting X 2 equal to 1 that is the female representation, then in that case X 2, X 2 will be 1, then Y will be b0 plus b2, b2 because b2 into 1 is you know the result will be b2.

So, now Y hat equal to b 0 which is where the case you know male representation, that means the average income it will b0 and in the case of you know female, the average income will be average income will be b0 plus b2. So, that is the kind of you know female income and what is happening here is the difference is actually b2 and that means, technically let us say in the estimation process we are having b0 say let us say 20000, right and let us say b1 equal to say 15000. So, that means technically Y hat that is you know average salary equal to 20000 and that too for he may know a male candidates

and then, female candidate case the average salary will be b0 plus b2. That means, technically 20000 plus 18000. So, domestic means a 15000 here, it should be 35000.

So, now after you know the kind of you know calculation, you can say that you know the female average salary will be higher than the male salary where the difference will be 15000. Now, in this you know kind of an hypothetical examples, so female salary is much higher than male salary, but in reality the situation may be opposite.

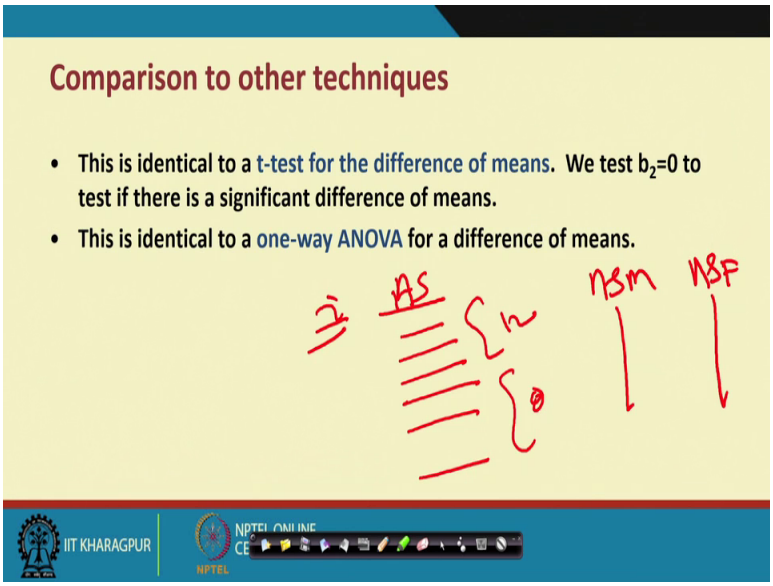
If you take the actual data and you know do the kind of you know estimation process and in that case suppose you know actual data or actual model saying that you know male income should be higher even if this model is correct. So, in that case the particular coefficient may be coming negative. So, if the coefficient will be coming negative, then by default you know if Y hat equal to b0 that is where you know average salary of male group is 20000 and in other case the female income will be 20000 minus 15000, that means you know 5000. So, you know whether actually you know male contribution or male salary will be high or females average salary will be high.

So, that will not explicitly depend upon the, you know and the kind of you know classification which we have made here 0 for male and 1 female. So, if you change the order actually you know 1 for male and 0 for female, so again the theory or the kind of you know empirical testing will not actually deviate too much.

So, actually if it is a you know male is high, then by default if you start with 0 for male and 1 female, then actually b0 or whatever b0 value, we will get that may be positive, but you know the you know female component that is for b1 will be negative and if you change the order and if the theory says that you know male average salary will be high, then by default you know coefficients both b0 and b1 will be positive. So, in that case female average salary will be 20000 and male salary, average salary will be 20000 plus 15000 that will be 35000.

So, likewise you know different ways of you know kind of you know interpretation. So, with this you know we can actually have a kind of you know structure, but the same structure which you can actually go for you know comparison by other techniques. So, that means you know what do we have done is, we have actually clubbed the samples.

(Refer Slide Time: 30:37)




For instance, we have here average salary you know and this is one specifications. So, 1st you are just collecting the data of you know 20 you know candidates out of which you know let us say 8 would be female and you know 12 will be male. So, now the use of dummy means using dummy. So, we club the data and then, dependent variable will be represented as a salary and then, whether male has a high impact or female has high impact or whether the difference between male female will be positive or negative.

So, that would be you know a user in analyzed through you know introducing dummy. So, here dummy will be qualitative and can be represented male female depending upon the samples and then, regress and we will find out the answers which you have already discussed right here and another way of you know checking the particular you know difference is it through a simple one way ANOVA.

So, in that case the entire sample can be divided into 2 parts me. So, that means average salary of male and average salary of you know females. So, first 12 data will be here and then, that I will be here so you can go ahead with the you know ANOVA test here where ANOVA is using one way ANOVA, not necessarily the data requirement is a balanced data, but maybe unbalanced data or if you require you can actually do the balancing by removing the additional data points or by bootstrapping you can you know normalize the data and you have the balanced data and that balance data can be used to you know do

the comparisons and say whether the difference between male female will be significant one or not significant ones.

Suppose there is no difference at all, then by default in the case of you know say in this case of say a dummy modeling if there is no difference, then the beta 2 coefficient will be becoming 0. So, if that is the case, then by default the male female you know salary structure ever self-structure to be more or less same. So, if there is a difference, then beta 2 will not beta 2 means coefficient of beta 2 will not be 0. If not 0, it may be positive, it may be negative.

If it were if it is positive, then it has a you know you know male will be high salary than female or female has a high salary than male and if there is no difference, the coefficient will be simply 0 and that we can find out here also through ANOVA. So that means, a dummy variable is an interesting you know structure or technique through which you can identify similar in or you know get similar kind of definite results or similar kind of you know inference depending upon the kind of you know engineering problems in most of the engineering kind of you know structures.

So, we have a kind of you know gender issue or something like that and that can be addressed through this particular dummy modeling and again for adding to the non-linearity, you can add another dummy variables which you have already discussed like you know a religion and genders and then, marital status with you know genders. So, these are the different levels of you know starting with linearity and transferring into the non-linearity as per the requirement and by the way, we can get to know you know flow of you know non-linearity.

So, this is one level of you know addressing the non-linearity structure by the use of you know W molding. So, we have more extension on dummy modeling and we have also different types of you know non-linear regression molding which we will discuss in the next class and with this, we will stop here.

Thank you very much.