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

Lecture – 21
Linear Regression Modelling (Contd.)

Hello, everybody. This is Rudra Pradhan here. Welcome to Engineering Econometrics. Today, we will continue with Linear Regression Modelling and that too part – 2. In the part – 1 we have discussed about the simple regression modelling and that too bivariate engineering econometrics where we must have two variables; one is dependent variable, one is independent variable and then the objective is to predict dependent variable with independent variable and we have discussed how to estimate, how to obtain the estimated results and the types of you know test to validate the empirical results and the regression model and that too the line of the best fit to predict the dependent variable with respect to independent variables

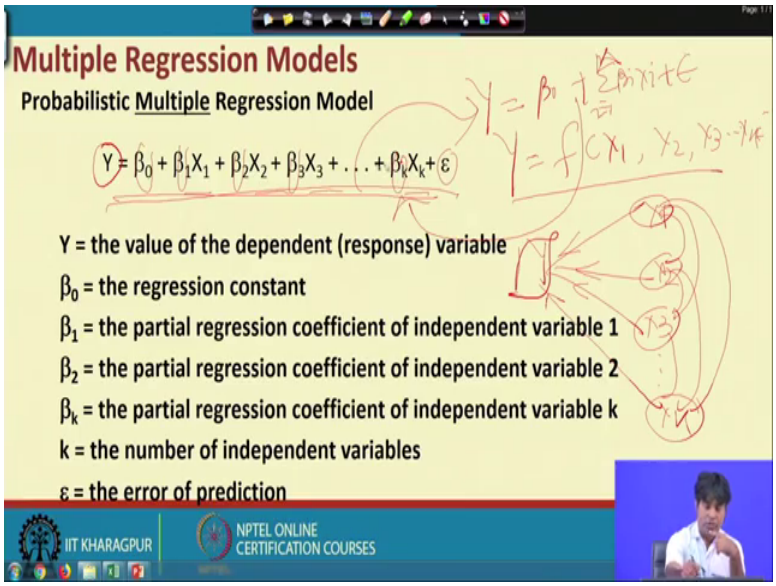
In fact, we have gone through specification test, goodness of fit test, out of sample prediction test, to predict the dependent variable with respect to independent variable and that too with a with a good fit model and the diagnostic the other the other part is the diagnostic test which we have not covered and after the multiple regression modelling we can touch up on that particular component. And, in this particular lecture we like to you know solve some of the engineering problems where we where the situation will be more than two variables that is a dependent variable with many independent variables.

So, here the rule is you know similar that too predict the dependent variable with respect to independent variables. In the part – 1, we try to predict the dependent variable with respect to single variable which is very simple very easy to process and easy to predict the kind of you know engineering requirement. But, in this case we have you know means we have this scenario where we have to predict the dependent variable, but instead of one dependent variable independent variable we have multiple independent variables and that is how the complexity will start. That means, when you add one after another independent variable the if the level of complexity will start increasing because when we add one after another independent variable one of the complexity will obvious that you know there may be some kind of you know linear relationship among the

independent variables that we need to check and to sort out the particular you know you know solutions before we go for predicting the predicting the y with respect to all the independent variables.

So, let us first see how is the multiple regression case and then we will go through the reliability you know you know reliability part of this estimated models where with the find all such problems will be there and the and these will be included in the diagnostic a kind of you know test and we like to know how these diagnostic test can be used and can be applied and to clean the estimated models before you do the predictions and the kind of you know forecasting's.

(Refer Slide Time: 03:54)



So, usually the regression the multiple regression modelling will be like this and here we have a we have a dependent variable that is called as a Y and series of independent variables starting with X 1, X 2, X 3 and X k and then by default there is a requirement of you know independent remains error terms the involvement of error term in the models you know is a essential which we have already justified in the part one, whether it is a bivariate case or trivariate case or multiple case or multivariate case. So, by default error term will be the mandatory component to do the processing for the prediction of Y with respect to independent variable.

And, here the structure will be like this so, in a implicit format Y equal to function of X 1, X 2, X 3 and so on up to X k and then there will be error terms. So, the idea is that you

know we like to check how these independent variables can influence the dependent variable. So, the path diagram will be like this. So, Y is the dependent variable which can be connected with the several independent variables. So, starting with X 1, X 2, X 3 and then it will be with you know X k. So, every times the impact will be unidirectional since it is a multivariate regressions.

In the case of multidirec[tional]- multivariate regression modelling the impact maybe bidirectional, but in this case the impact will be unidirectional every times we like to check how is the influence of X 1 on Y, X 2 on Y, X 3 on Y and likewise you can continue and then X k on Y and that means, every times we like to check individual impact means individual impact of a particular independent variable on dependent variable but, while you know connecting more number of independent variables to Y that too independently means individually then by default you know you must be you know means you must be sure that all these independent variables are technically independent.

So, that means, if you if you looking to this path diagram. So, there may be some relationship among these independent variables like X 1 X 2, X 1 X 3 the next one X k then X 2 X 3, X 2 X k, then X 3 X k. So, these are all possible you know correlation or association you know in the setup. Since it is in a kind of you know multivariate scenario where one dependent variable with many independent variables.

So, first of all; that means, technically first requirement is that all these independent variables should be technically independent, but in reality it is very difficult to obtain. So, our job is how do you know bring a kind of you know estimated model which is free from all these obstacles. If not then we have to minimize if the level best through which we can actually predict the dependent variable with these independent variables without any a kind of you know significant problems you know or a significant issue.

So, first of all so, we should you know ensure that all these independent variables are independent and they are all theoretically or logically you know must have influence on dependent variables. So, without any you know kind of you know knowledge or the kind of you know hint a particular independent which variable should not be you know involved in the estimation process, otherwise you know it will give you some kind of you know biased results or sometimes it may be called as you know spurious regression.

So, one means typically if you if you know compare strictly mathematical regression modelling then the econometric engineering econometric modelling. So, one of the one of the fundamental requirement is that you know you must have theoretically you know evidence or logical evidence before you start the process. Because, it is it is the concept of called as you know applied econometrics it is not completely mathematical econometrics where you can you know connect X 1 with a Y X 2 with Y without any kind of you know theoretical evidence because it is a mathematically we will established technique. So, even if you do not know the nature of those particular variables you can actually do the modellling.

But, since we are actually interested to solve some of the engineering problems; that means, technically we are using regression modelling and that too it is a kind of you know application towards the engineering problem means the whole objective is to work out the engineering problems through the help of you know engineering econometrics. So, we must have actually complete knowledge or complete evidence how these variables are connected each other before you start the processing, before you start the means do the predictions or in the kind of you know forecasting.

So, as usual the general framework of multiple regression is like this and you know this is implicit format and this is what the explicit format and again you can write the problem like this Y equal to beta 0 plus summations beta i X i; i equal to 1 i equal to 1 to n plus epsilon that is the error term and if you actually a make it you know separate then it will start with Y equal to beta 0 beta 1 X 1, beta 2 X 2, beta 3 X 3 up to beta k X k and since it is a k. So, this should be i equal to k.

So, as a result when i equal to 1, then it becomes beta 1 X 1 then when i equal to 2 it will be beta 2 X 2 and so on and continue ah, where i equal to k then the last term will be beta k X k. So, so ultimately this is the model. So, that means, this model can be transferred to this way or if you generalize this model this can be come into this way. So, depending upon the kind of you know requirement either you can put this way or you can put this way, ok. So, here the objective is to estimate these parameters that are beta 0, beta 1, beta 2, beta 3 up to beta k.

So, as usual like we discussed in the part – 1, so beta 0 is the intercept which represents that you know when X all X are you know 0, still Y has some kind of you know weight.

For instance, the previous problem which we have discussed is called as you know costing and determination of you know costing that too airline industry and where one of the independent variable is the number of passengers so; that means, when number of passengers will start increasing or decreasing the cost of the airline will start you know increasing and decreasing.

So, that means, number of passengers will affect the costing of the airline industry. So, when there is a change you know happening in the case of you know number of passengers so, the costing will also you know change accordingly. So, usually theoretical hint is that when passengers are high cost may decline when passenger will low cost may more. So, because it cannot actually when passenger will be low it may not you know possible to minimize the fixed cost, but when passengers are high we can actually very much optimize the fixed cost that is why. So, increase in passengers will have a significant negative impact on you know castings.

So, that means, increase in passengers will reduce the cost factor and decrease the passenger will increase the cost factors. So, that is how the case we have discussed. So, then if you put you know X equal to 0, where number of passengers and the cost factor will not actually 0, because it will take care by the intercept factor. So, that means, technically in the part – 1 we have discussed the model up to this point ok, plus of course, plus error terms. So, here this one is the airline passengers and this is the total airline cost.

So, now, when the passengers are 0, then still the airline cost is there because some fixed cost are always there whether passengers will be in the system or not. Because passenger is the is the ultimate customers for this particular you know industry when customer will in or customer will be out then the cost variation will start you know affecting, but still even if there is no of passengers in the system still there will be the cost because the initial setup involves you know large amount of the cost. So, as a results since initial investment is the large one so, to optimize the process and to go for you know cost effectiveness. So, you have to increase the passenger size, so that you know it will be it will reduce the cost factor as per the particular you know industry requirement.

So, now, what is happening in this case we add one after another variables to justify that you know same airline cost can be affected by the passengers the infrastructural

availability, manpower you know availability or something like that investment or advertising. So, many factors which can affect the airline costs. So, once you identify you know the particular independent variable then you can add into the system then it then it becomes you know simple to multiple one.

So, now mathematically this is for the model without knowing the means without connecting to any engineering problem. So, let us first know the mathematical concept behind this multiple regression modelling and then we will connect to a particular problems and to check how this you know these problems can be you know process and then what type of you know outcome will you obtain and how to do the predictions and how to do the forecasting and that too with the check of you know reliability and something like that.

So, here these coefficients beta 0 is the intercept as usual you know discussed in the case of you know bivariate regression modelling and then beta 1 is the slope coefficient for X 1 variables, beta 2 is the slope coefficient for X 2 independent variables, similarly beta k is the slope coefficient for the X k independent variables and k is the number of independent variables and epsilon is the error terms.

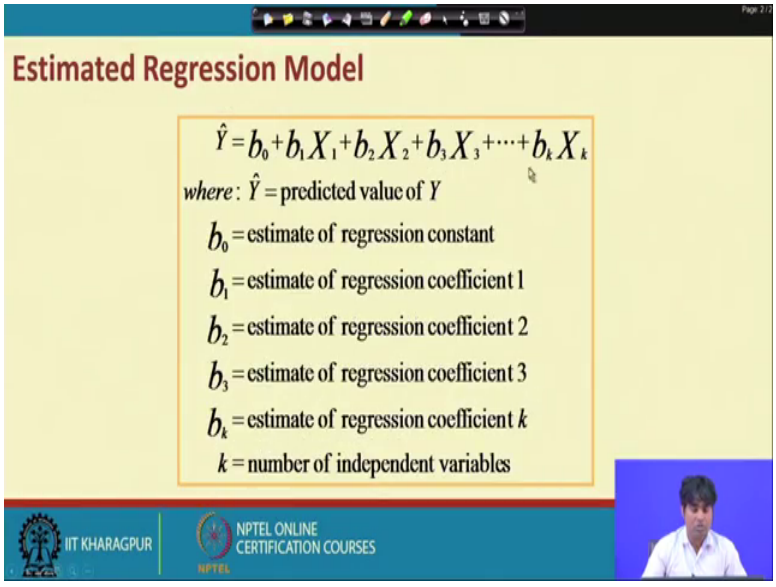
So, that means, technically we have you know couple of independent variables through which we can actually predict the dependent variables. So, as usual so, the predicted model will be Y head equal to beta 0 head, beta 1 head X 1, beta 2 head X 2, beta 3 head X 3 plus you know it will continue then the final term will be beta k head X k and then by default error term will be removed. So, that means, the original Y involves you know intercept and all these independent variables and then the predicted and the error term then the estimated model will be only intercept at the estimated parameters that too connect with the all these independent variables.

So, by default so, the difference between Y and Y cap that is Y estimated will give you the error weightage, like the previous case which we have discussed in the part one that too bivariate regression modelling or you know simple regression modeling. So, that means, technically whether the system involves 2 variables or 3 variables or 4 variables. So, the stepping we need to obtain the empirical model will be more or less same. So, you must be very careful how you to deal with the situation and you add one after another variables, but systematically it will be the derived of course, the complicity part

is that you know the relationship among the independent variables. Otherwise the process is more or less same whether it is a bivariate structure or trivariate structure and multivariate structure.

So, let us see how is this kind of you know environment and then how we can obtain the particular you know process.

(Refer Slide Time: 16:50)

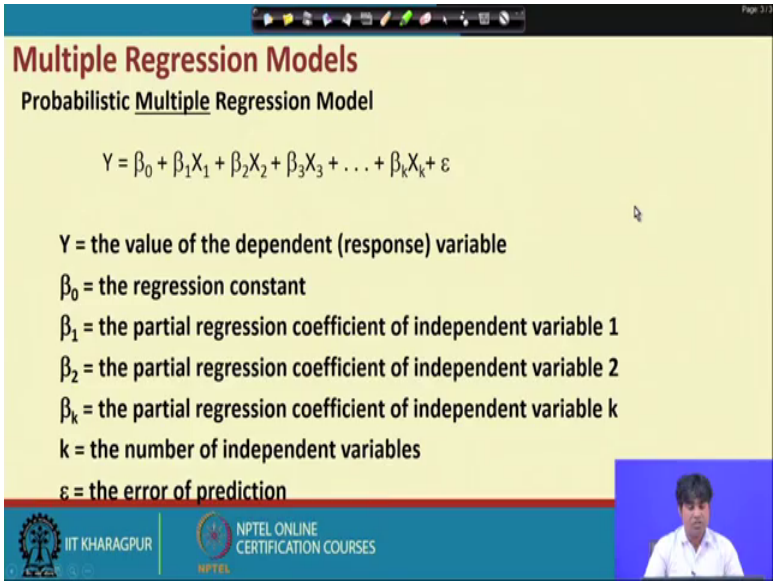


So, let us you know as usual I have mentioned. So, Y cap is the estimated models in default error term will be removed in the process. So, as usually know we have to apply the ordinarily square mechanism initially till unless we have a some something wrong and then the process of OLS is to minimize the error terms and the way we will minimize the error terms by default we will obtain the estimated parameters and this estimated parameters will help to predict Y with respect to all these independent variables.

So, by default so, beta 1 b 1 is the estimated value of b 1 is the estimated value of beta 1 and then b 2 is the estimated value of beta 2 similarly b 3 is the estimated value of you know beta 3 head and b k is the estimated value of you know beta k head. So, and b 0 is the estimated value of you know beta 0. So, likewise we will means we have all these estimated parameters and in the in comparison to first equations where all these parameters are unknown.

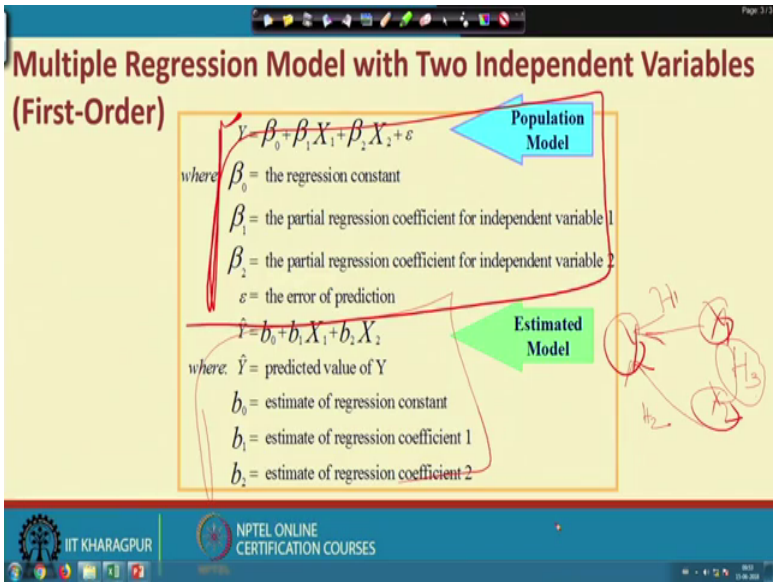
Now, with the help of OLS and then with the help of with the availability of you know data the way will you process all these estimated all these unknown parameters to be now known. And, once you know all these estimated parameters, then it is easy to predict Y with respect to all these independent variables. That is the usual procedure through which you can actually process the things get the estimated model and then go for the reliability check and then go for the predictions and forecasting as per the particular you know engineering problems requirement.

(Refer Slide Time: 18:45)



So, now let us see the first step is a the structure of the a multiple regression modelling and the second step we have the estimated model of the multiple regression modelling and the kind of you know understanding between the variables and the parameters and then the kind of you know involvement of you know error term.

(Refer Slide Time: 19:06)



So, now in order to simplify the particular process because you know when you talk about you know multiple regressions. So, we just generalize that you know it starts with the X 1, X 2 up to X k, but X k when I write X k then it is a kind of you know in a infinite sign because we do not know what is exactly k.

But, for a applications or you know when we solve when we like to solve a kind of an engineering problem. So, make sure that you know all these variables are clearly identified until unless you close the loop or you know identify all the variables you may not start the process. So, the moment you will start the process so, everything should be in a particular loop. So, the number of variables the sample size the kind of you know functional form everything you should be clear you know either you know through some kind of you know homework and testing checking, then at the end of the day your basket must be all kinds you know information's with you know removal of all obstacles. Then you can easily proceed and do the processing and have the estimated model and go for solving the engineering problems as per the particular you know requirement.

So, now these are all very simple to you know convince and to understand then process, but ultimately how is the kind of you know set up. Since beta k X k is the kind of you know infinite signal so, in order to understand the reality and the kind of you know structure for the engineering problem predictions and the kind if you know forecasting. So, let us bring a particular situation, where Y equal to beta 0 beta 1 X 1 and plus beta 2

X 2 that is question is called as you know that is particular structure is called as you know trivariant modelling.

In the in the trivariate modeling so, that is also multiple case. So, that means, in the first instance you can you can have two different you know structure or the kind of you know modelling. So, in the first case it is a simple bivariant and this second case it is the trivariant and you know kind of you know multiples. So, that means, so, the minimum requirement of one independent variables and the more than one independent variable it may be start with you know two independent or it may be twenty independent or it may be thirty independent. When the number of independent variable will start increasing then the then the complexity will start increasing because you have to address two different issues all together when you are dealing with you know multiple regression modelling.

If first requirement or first issue is you have to check the existence of linear relationship among the independent variables and again you have to be you have to be very careful about the sample size. For two variable case, then the sample size maybe small one, but when we add one after another variable then the sample size should start you know increasing accordingly. Otherwise keeping sample size remain constant if you add one after another independent variable, then you may not go for you know means you may not you know in a position to test the model and to validate the model before you start predicting and you know forecasting.

Because sometimes you may not obtain the parameters while you know, processing with a particular data because of the low sample size and large number of variables. In fact, in the first lecture of regression modelling I have given you some kind of you know instruction the particular requirement before you start or you know using regression modelling for the solution of a particular you know engineering problems. What I remember I means what I have discussed earlier that you know one of the fantastic requirement or you know typical requirement is your the that is the equation or game between n and k the sample size and the number of variables so, every time your sample size should be substantially greater than 2 k and ah; that means, and greater than 2 k

Compared to previous problem the airline cost determinations where we have we have analyzed the problem with respect to two variables, the airline cost along with you know

airline passengers and where we have we have started the problem or investigate the problems or we have connected the regression modelling with that you know air airline cost determinations within you know sample size of you know 2 l only because in that problems your sample size n equal to 2 l and k equal to 2, so, 2 l and 2. So, the there is you know such significant issue.

Of course, for econometric kind of you know principle or statistical kind of you know principle. So, we have two rules of you know sampling one is called a small sampling and another one is called as a large sampling. So, the division between small sampling and large sampling depends upon the size of the sample 30 to 32. So, when it is actually you know less than 30 or less than 32 so, it is called as you know small samples and more than 32 it is called as you know large samples.

But, the meaning of you know small sample large samples which can be very easy to understand, easy to pickup in the context of bivariate setup, but when we have more number of variables then technical for instance instead of 2 if you have 20 then the small sample large samples with the particular you know typical link like you know 30, 32 will not actually a convince you are actually not help you to solve the problem. So, so the simple understanding or simple structuring is that you know you must have enough sample size that too you must you must be in a position to optimize significantly with the equation between n and k.

So, that you know your degree of freedom should be substantially very high to validate the models as per the particular you know requirement. So, now, these are all the kinds of you know clarity and understanding before you start the processing whether it is a bivariator trivariator you know kind of you know multiple. So, in order to you know justify the kind of things so, we can restrict the model with you know two independent variables. So, when your model is a with you know two independent variables that that becomes called as you know multiple regression, but when we have actually only one independent variable that is called as you know simple regression modelling.

So, now here with two variables two independent variable it is called also as multiple regression modeling because one of the fantastic you know issue between bivariate and multiples; that means, simple regression and multiple regression is the number of

independent variables involvement and their relationships. That is most important you know typical difference between these two different case.

So, now since the typical issue is the relationship between independent variable, so, when there are two independent variables that relationship can be there. But, when more such independent variables then the then the number of relationship among the independent variable will start increasing and by default the complexity part will be start you know you know adding in the process of you know investigation, in the process of you know estimation as a result. So, the issue will be very complicated and until unless you solve these issues, you should not start the process and you cannot use the cut particular regression model for the requirement of you know predictions or you know forecasting of a particular you know engineering problem.

So, now so, we have the so now, the general models can be restricted to two independent variables for that for that. So, this is the first hand requirement, and then this is this by using a particular sample this is the estimated model with respect to independent variables. So, that means, it is typically Y X 1 and X 2 ok. So, obviously, X 1 has a influence on Y X 2 has a influence y. So, this is this is hypothesis 1 and this is hypothesis 2 and this can be additional hypothesis which can which you can create where H 3 and the H 3 indicates that you know there is no link between X 1 and X 2.

But, in reality there maybe some link, but as per the requirement of you know multiple regression that links should not be statistically significant. If that link is a completely 0, then you are absolutely imperfect you know path, but ah, but you know situation is something different. When you integrate something or solve a particular you know real life problem or any real engineering problem. So, you will find these variables are not really independent. So, they may have some kind of you know relationship and you try to you know you try to actually minimize that relationship. Or, sometimes you can two variables are you know same and you are introducing in the system then the problem the estimated model will have some kind of you know biasness.

So, that means, you know what is happening in reality there may be high chance that you know X 1 and X 2 are same, but the way you will introduce in the system are you know different. For instance, let us let us let us take an examples in a kind of you know financial environment, so, we have we have a kind of you know component called as

inflation and while you know predicting the kind of you know air lining airline cost or real estate problem, that means, housing price etcetera, so, inflation can be one of the factor. Of course, it is a macro econometric variables, but still it has some kind of you know link to the original problem and in that instance what will you do you can actually link the inflation factor and predict the housing price or airline cost or something like that.

So, now what is happening to this particular problems or the kind of my discussion is that you must have you know something like you know you know understanding that air airline costing means determining airline costing or housing price should be actually a clear a clear with all these you know independent variables clarity. For instance, if you use inflations and now how you have to actually measure the inflations. So, inflation rate is also still there.

And, sometimes what is happening, if inflation data is not available because these models or this process are exclusively depends upon the data to get these you know parameters value. And, when inflation data is not available then you can use the proxy variables, that is that is called as you know intermediate variable through which you can actually a obtain the actual variable that is inflation here and for that there are two different variables which can be used frequently to get the inflation factor that is called as you know consumer price index CPI and wholesale price index that is called as a WPI.

So, now both are actually you know used independently to calculate the inflation and that variable that is the inflation can be used as a used as a as an independent variable to predict the airline cost or to predict the housing price or to predict the engineering products or something like that. So, in that case if you are not having clear cut understanding or something you know lack of you know knowledge then there is a high chance that you may use CPI and WPI in the system because CPI WPi actually has a different kind of understanding different kind of you know information, but all will you give you the hint about the inflation.

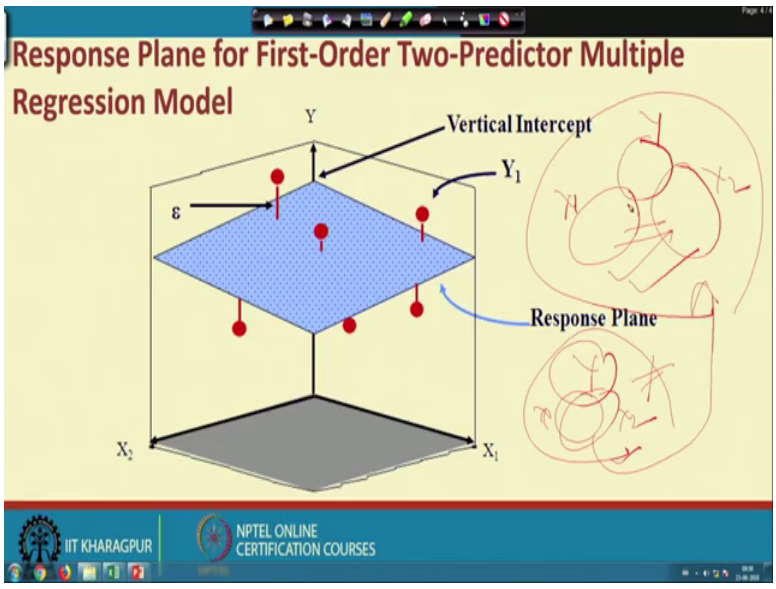
So, now instead of you know using inflations if it is not available, then you may use CPI either CPI or WPI. But, if you simultaneously use CPI and WPI and that too for you know naming X 1 and X 2 then there is high chance that you know X 1 and X 2 would be highly correlated. So, in the first instance you can drop a particular variable either CPI

or WPI and keep WPI and CPI you know as per the particular requirement and then go ahead with the prediction, then the problem by default will be solved.

So, that means, technically this kind of you know scenario so, you first understand the reality, understand the problem then you check whether these independents variables are something you know same or you know different. If theoretically they are different you can keep and then even theoretically they independent, but with the help of some kind of you know structuring and the process of integration these variables may have some kind of you know relationship and then you try to laid down that you know you know kind of relationship by using some kind of you know mathematical structuring or data structuring or something like that, then ultimately at the end of the day you have a clean environment to do the things as per the particular you know engineering problems requirement.

So, this is what the trivariate case that too restrict the model with two independent variables and then we will see how is the particular process. So, this is the actual scenario and this is what the estimated scenario here b 1 is the estimated coefficient of you know beta 1 and b 2 is the estimated coefficient of you know beta 2 that too for X 1 and X 2 impact respectively on Y dependent variable Y. So, this is what called as you know estimated model.

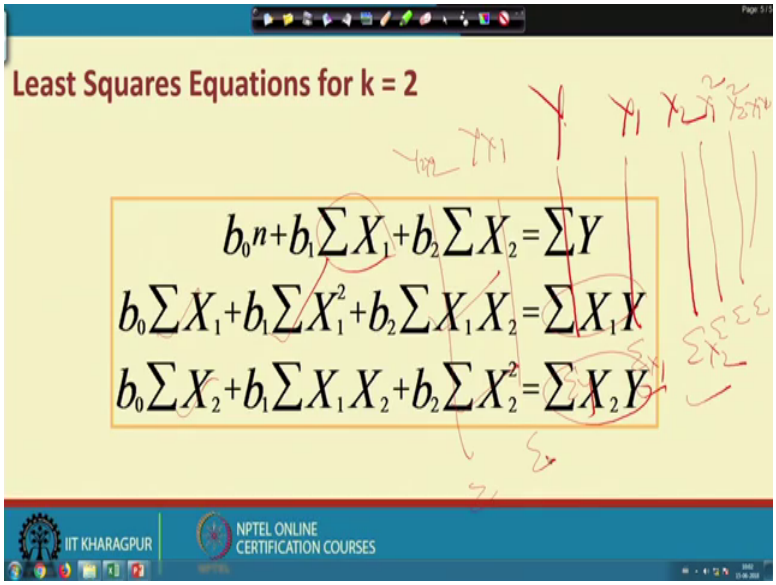
(Refer Slide Time: 33:19)



And, if you actually plot this case in a kind of you know you know plotting case that is the 3D, then this look will be like this and how this Y will be a reflecting within X 1 and X 2. Ultimately the flowchart will be like for a for a regression modeling, this is what you know visualization process. Ultimately the actual flow of model you can you know site mathematically or you know with a kind of you know flowchart, but in a reality it maybe something different. We can start with the concept like this; that means, this is what we need actually you see here. So, between X 1 and X 2 there is no relationship, but ultimately when you linking Y with X 1 and X 2 there may be chance that you know X 1 and X 2 are actually linked.

So, this is not actually valid, but this is this is actually valid so, but what is the actually reality whether you are in this track or this track. So, you need to check and then you know proceed accordingly unfortunately, if you are here then you should not proceed with the a prediction and estimation you know forecasting. So, you try to bring this case to this case, if not like this somehow little bit closer and then you go ahead with the kind of you know prediction and forecasting. This is what the real understanding and real need you must have before you start you know processing in the kind of you know things and then ultimately so, this is what the kind of you know structures.

(Refer Slide Time: 34:53)



Now, like you know the part one where we have discussed the link between Y and only X for instance if you will restrict the general generalized the model to two variable case

then it becomes Y equal to beta 0 plus beta 1 X 1 and then by using OLS where we like to minimize the error sum square to obtain the parameters like you know beta 0 and beta 1 for predicting Y with respect to X 1. So, we once we actually go through the optimization process by minimization of you know means that is to minimize the error sum squares. So, we will have two different equations that is with respect to beta 0 and beta 1 and with these two equation you can get the formula of you know beta 0 how to obtain from the data and then beta 1.

Similarly, in the case of you know trivariate where we have a dependent variable within two independent variables. So, then when we process that is you know the error terms which is equal to Y minus Y hat where Y hat is not actually we know beta 0 beta 1 X 1 it becomes you know you know beta 0 beta 1 X 1 beta 2 X 2. So, that means, there are two different component; so, ultimately beta 0 beta 1 and beta 2, so we have a three different equation.

So, now if you solve these three different equations, then you can get the formula of beta 0 beta 1 and beta 2 that is the estimated beta 0 estimated beta 1 and estimated beta 2. And, then with the help of the data if you do the processing that too in manually so, you can easily obtain these unknown parameters that too beta 0, beta 1 and beta 2. So, this is how the structural form of these equations.

So, with the help of so, for instance here spreadsheet or data will give you something like this a. For instance it is it will be like this Y equal to X 1 and X 2. So, you must have a data you must have a data you must have a data. So, now summation X 1; so, here you will get summation X 1 and summation X 2 here you will get summation X 2, then summation Y here you will get summation Y. So, then again summation X transformation X 1 you can have here, but summation X 1 square means. So, what will it do you can create a particular you know component again X 1 square that is squaring all these items then we can also squaring X 2 squares and then you can create X 1 and X 2.

So, by multiplying these two; so, you can have here sum you can have here sum you can have here. So, so, that means, you can get also summation X X 1 square summation X 1 and X 2 and summation X 2 square. Again, so, you can create another you know kind of you know column by integrating Y X 1 and Y X 2 to fill this particular component, then again you take this sum take this sum. So, ultimately from the original data set Y X 1 and

X 2. So, you can get summation X 1 summation X 2 summation Y summation X 1 square summation X 2 square summation X 1 X 2 summation X 1 Y summation X 2 Y.

And then finally, you simplify this three equations. Since there are three parameters and three equations you can easily get these values and then it becomes it becomes you know means you will come to a kind of you know position to have the estimated model to do the predictions and do the forecasting as per the particular you know engineering problems requirement how to do these processing and that too with three variables, we will discuss in the next lecture. With this, we will stop here.

Thank you very much. Have a nice day.