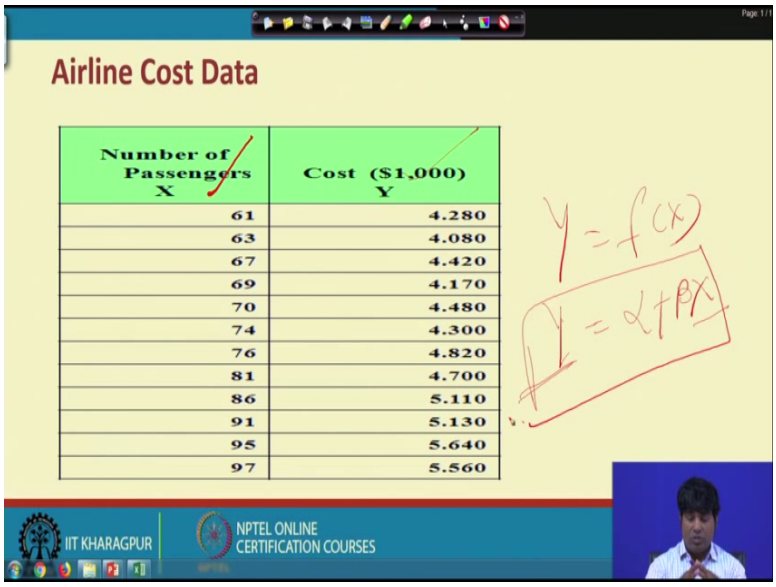
Engineering Econometrics
Prof. Rudra P. Pradhan
Vinod Gupta School of Management
Indian Institute of Technology, Kharagpur

Lecture – 17
Linear Regression Modelling (Contd.)

Hello everybody, this is Rudra Pradhan here. Welcome to Engineering Econometrics. Today we will continue with Linear Regression Modelling. And in the last lecture, we have discussed the linear regression modelling with respect to the requirements, the kind of you know structure or the kind of you know applications.

(Refer Slide Time: 00:47)



So, to continue this particular aspect, so, we have to proceed for the kind of you know discussion, you know. Let us start with the something called as you know Airline Cost Data and that is a transportation engineering problem, where the objective is to predict the cost with respect to couple of variables. But since we are discussing about the simple regression modelling,so we will we will integrate cost with the one particular variables in the transportation sectors that is the number of passengers.

In the transportation engineering, we have a couple of transportation infrastructure starting with the roads, railways airports, ports etcetera. And this is this particular problem related to airline industry and here one of the basic objective is how to reduce the cost and as a result the basic structure is used to find out the determinants of you

know airline cost. And from the theory or the kind of you know overall understanding the first factor which can influence the airline cost is the number of passengers.

So; obviously, our modelling will be is to determine or to predict the cost, airline cost with respect to number of passengers in you know airline. In fact, this is not the only variable which can determine the airline cost, there are couple of other variables which can equally it determine the airline cost. And a number of passengers is the prime factor which can in influence the airline cost you know a significantly.

So, as a results so, the first you know work is to connect a airline cost with number of passengers, then we can extend this models and that too we will discuss same models in the context of multiple regression modelling where we will determine the cost with respect to number of variables including number of passengers. Now the idea is here whether you know number of passengers you know can influence the airline cost.

So, we have a dataset and these are the typically known as data structure and this is what the number of passengers and this is what the cost per you know 1000 dollar and these are all you know passengers numbers. And we like to check whether the number of passengers can influence the cost factor and if it is the case, what is the kind of you know nature? And in that whether there is a negative relationship or positive relationship and then we have to check whether the particular relationship is statistically significant or not.

And a corresponding to this objective and the kind of you know transportation engineering problem and we need to apply engineering econometric tools that to bivariate regression modeling or means in other words it is a simple regression modeling where cost is declared as a dependent variable and number of passengers is declared as a independent variables.

And since it is a 2 variable case and the model will be very simple, we start with the model like this Y equal to functions of a X, where since we are discussing with respect to linear regression modeling. By default, a the you know the model can be written as a Y equal to alpha plus beta X, where Y is declared as a airline cost and X is declared as a number of passengers in the airline and alpha and beta are the parameters.

And this is what you know mathematical form of the models. From the theory or the kind of understanding, we like to you know build a model cost with respect to passengers and that too there is a linear relationship. And we are hoping the particular you know variable will have a positive impact on cost; that means, technically if a number of passengers are high, cost will be high or number of passengers will be high, cost can low.

So; that means, in the first instance we have now clear idea whether increasing the number of passengers will reduce the cost or in you know increasing the a number of passengers will be reduce the cost. So, the thing is that you know, we like to integrate with the help of you know bivariate regression modelling and then you check the kind of you know relation so, check the kind of relationship. So, for the management principle is concerned increase the numbers that is with respect to passengers, then the cost can decline.

If the passengers numbers are low, then cost should be high because there are lots of you know fixed cost involvement. As a result if the number of passengers will be lower one then by default cost will be high and the moment passengers will be high, then the cost will decline considerably.

So, means a theoretically we have this kind of in our knowledge, but so, for as a mathematics is concerned or a modelling is concerned, we like to you know state that you know there is a negative relationship or we can start if there is a positive relationship. On the basis of the data, we can check whether the particular you know relationship is a negative or positive and that too whether it is a theoretically supporting or there is something wrong.

Because, there is a high chance that you know the relationship will be negative. So, by default the indication is that you know increasing the number of passengers reduce the cost and vice versa. But, in reality this may not actually the case because we are missing some of the other variables to determine the cost. So, the nature maybe coming positive you know probably due to you know not involvement of this particular you know other important variables.

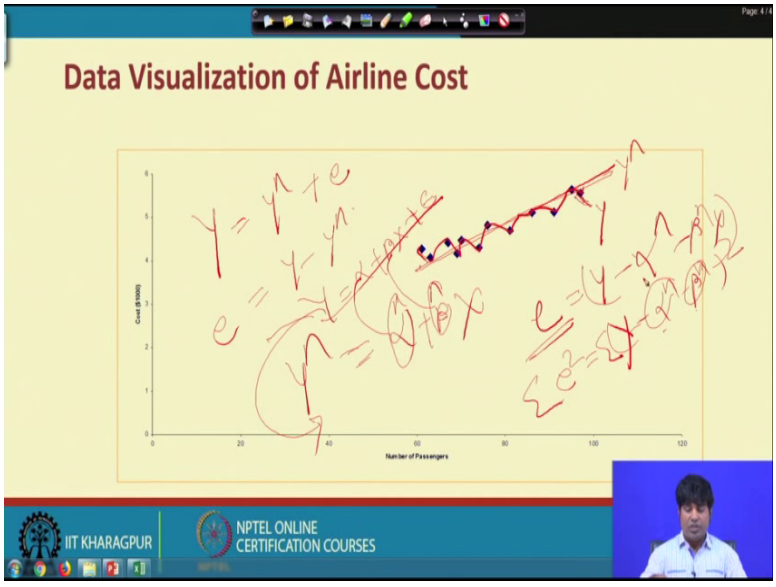
So; obviously, we will check and you know find out the kind of you know solution. So, the first hand of this particular process is to you know build a mathematical models which can establish the relationship between airline cost and number of passengers. As a

result we have read we have already identify a particular model linear model, not necessarily that they are always in linear in nature. But there may be some kind of non-linearity and for that we need to have a data visualization in the. On the basis of data visualization, we can get to know whether the particular relationship is a linear one or non-linear one.

But, since a since we are discussing linear regression modelling, let us start with the assumption that they are linearly related. And according with the mathematical model which we have to introduce here is Y equal to alpha plus beta X, where alpha is a parameter, beta is a parameter, Y is a dependent variable and X is a independent variable. And the model itself you know interprets that you know even passengers are 0, still there is a airline cost. Because, cost usually means here we are representing the cost that involves that involves total cost which includes variable cost and fixed cost.

So, variable cost depends upon the number of passengers involved in the process and fixed cost even that does not you know depend on the kind of an airline passengers that is a often X becomes 0 Y cannot 0. So, Y represented by alpha so; that means, what is the particular level of alpha through which the particular you know problems can be analyzed or the kind of you know predicted? And further we will we will go for you know detailed discussions.

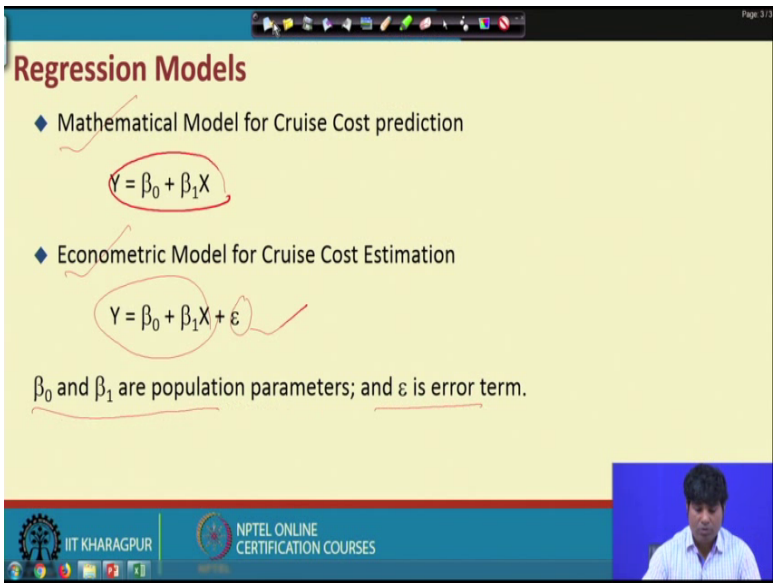
(Refer Slide Time: 09:40)



Let us see how is the kind of you know structure. So, first end requirement is the data visualization and if you plot these 2 number of passengers corresponding to cost. So, this X axis we can put number of passenger and Y axis put you know cost factors. Then we define the plottings are plottings are here like you know this is what the plottings and this is what the boundary and more or less they are actually if you will connect the flow will be coming like this and somehow it is actually is slightly linear not perfect way non-linear.

So, as a result we can actually mean regress the particular cost with respect to airline passengers and that too with the help of linear regression modelling. So, the data visualization is also supportive. So, since data is supportive for the linear regression modelling we can proceed for the modelling or the kind of you know; estimations. So, let us see how is the kind of you know, process.

(Refer Slide Time: 10:51)



So, this is how the mathematical kind of you know structure, here we have 2 different model, first one is the called as the mathematical model, the second one is called as the econometric model. The first models brings that Y equal to beta 0 plus beta 1 X and a second one is Y equal to beta 0 plus beta 1 X plus epsilon. And in the first case, particularly for mathematical model; there is a relationship between Y and X and that too with involvement of 2 parameters beta 0 and beta 1 and that too without the involvement of error terms that is for the particular you know case here.

But, in the context of in the context of econometric model, so, in addition to Y equal to beta 0 plus beta 1 X we have a epsilon and for the that for that you know we need to clarify and in the econometric model. And in the statistical model we usually introduce the error terms with the justification that nothing is exact so; that means, technically X is not the only variable which can influence Y and that is the first and kind of you know clarifications.

So; obviously, the question is you what is the need of you know introducing the error term. Because, the thing is that the in the process of statistical investigation or economic econometric investigation we assume that there is something you know wrong or something lacking. And if you say that is perfectly you know and there is nothing lacking or something wrong, then there is no rule of you know statistics and econometrics. In fact, if it is a perfect still we can use statistic and econometrics in that case we will find the error will be completely 0.

So, whether there is kind of you know error or not error. So, on the process of modelling, you can get the clue and accordingly we can go for you know establishing this particular you know relationships and then go ahead with the kind of you know requirement. So, we have 2 different models here, mathematical models which is without the involvement of error and econometric model which is actually involvement of error term. And I like to point out to why there is a need of an error terms every time when we do the econometrics analysis and that too regression modelling.

So, in the first hand as a researcher or as a kind of analyst, so, you cannot even assure whether the particular you know set of all the variables are clearly identified. And you have no idea how these variables are actually regretted whether there is a multivariate kind of a relationships or simple kind of a relationship and whether there is only one variable which can influence the dependent variables. So, whether there are multiple variables which are involved in the process of you know regression dependent variable.

The thing is that, at a particular point of time you are not sure how many variables can influence the cost here airline cost here. Even if you theoretically you identify you know for 5 variables still you will not sure that these are the only variable which can influence the dependent variable. So, as a result, there is something wrong. So, that particular wrong can be captured through the error term, this is the first justification which we can

you know bring here to justify that there is an involvement of error term in the econometric modelling.

And likewise there are couple of you know reasons you can point out for that you know to justify the involvement of error term in any kind of you know econometric modelling and or in a regression modelling. So, for instance, there are certain factors which cannot be identify there are certain factor can be identify and cannot be quantified, cannot be captured and sometimes there are human error sometimes there is a kind of you know external factors which is not in your control.

So, likewise there are you know couple of factors or couple of indications through which we are in a position to justify that you know there is a need of you know involving error term in the process of modelling that too establishing the relationship between dependent variable and independent variable. The movement we are establishing the relationship between dependent variable and independent variable that too Y with respect to X, where Y is the only variable and X may be 1, X may be many. So, in both the cases so, there is a need of you know involvement of an error terms.

In fact, when you go for multivariate, where there are more number of Y, more number of X. The situation maybe some maybe more drastic and more complex, but in the meantime for the bivariate structures when we are integrating Y with a only X then; obviously, there is a serious doubt that you know how do you decide that this is the only factor which can influence Y. Again when you go for multiple regression; let us say, instead of single X you can have a you know 3 independent variables or 4 independent variable. But, still there is again doubt or there is a not sure that you know these are the only variable which can influence the dependent variable.

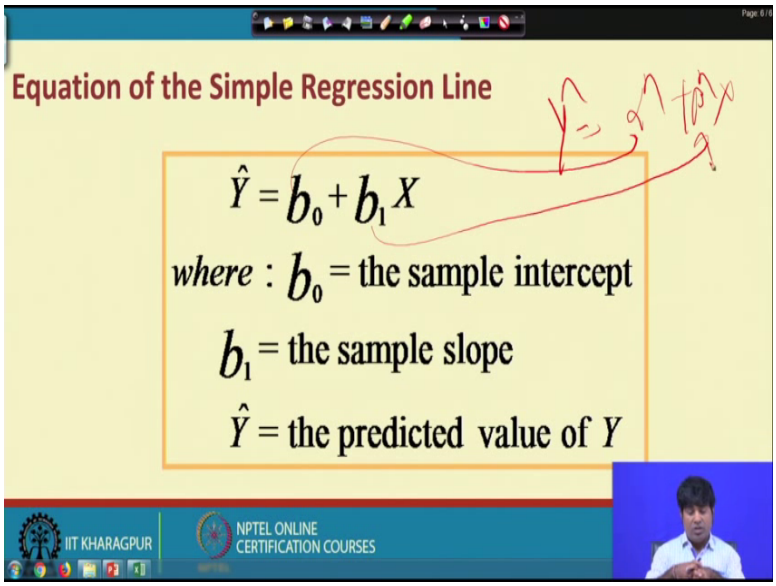
So; that means, a in totals we are in a you know position to justify that you know there is involvement of error term in the process of econometric modeling that too establishing the relationship between the Y and X, that too the relationship between dependent variable and independent variable. So, the difference between the 2 model mathematical model and econometric model is that in the mathematical model there is a no involvement of error terms and in the case of econometric modeling there is a involvement of error term.

And the process of econometric modeling is to check what is the level of error terms, whether it is a 0 or not 0 if not 0 in what extent it will affect the systems or the kind of you know modelling. So, the whole idea of econometrics analysis or whole process of econometrics analysis is to how to minimize the error terms and so that the model will be very perfect and the accuracy will be come into the picture.

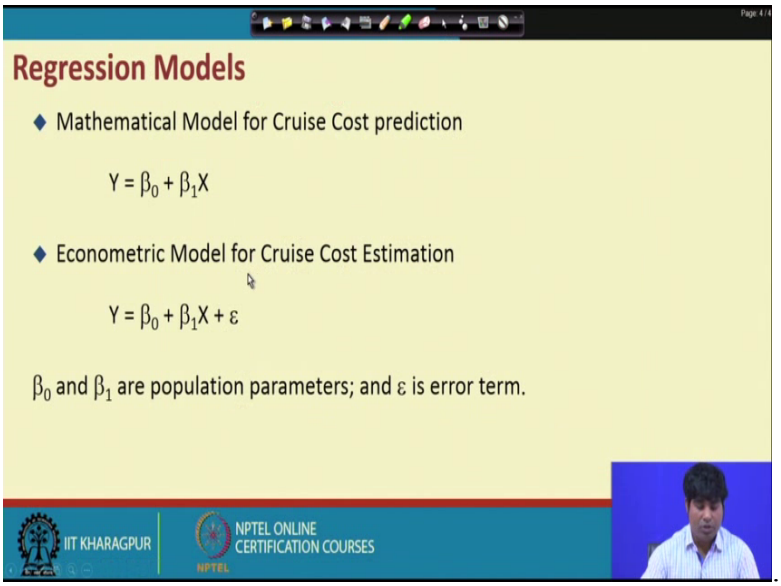
So, the whole job of the econometrics and the engineering econometrics process is to minimize the error term with respect to these parameters or the kind of you know variables involved in the process.

So, here on the basis of this econometric modelling, so, we have 2 parameters beta 0, beta 1 and then a epsilon is the error term, on the basis of that we will we will do the processing and in a first hand processing the compared to econometric modelling.

(Refer Slide Time: 18:20)



(Refer Slide Time: 18:25)



So, the thing is that you know since the model is not perfect, there is some kind of you know volatile is there. So, we need to find out within the particular volatility what should be the perfect line or predictor line or sometimes called as the best fitted line or line of the best fit.

So, for that we need a estimated equations; that means, technically the job is like this. So, this is how the kind of you know, data visualization. So, in the first hand; so, if you join all these points so, the point will be coming like this ok. So, this is how you join all these points, then we will get actually this is how the curve so; that means, there is a slight way you know volatility and as a result the prediction accuracy or forecasting accuracy is not so, significant.

So, what is the role of regression modeling is that to find out a line which can be you know very easy for a researcher or analyst to predict the dependent variable with respect to independent variable and for that I particularly we can draw line like this. So, this is what which called as you know predicted line. So, we call as a Y estimated and this particular you know the volatile a kind of a line which called as actual Y.

So, technically the actual Y is equal to estimated Y plus error terms. So, as a result error term is equal to Y minus Y predicted so; that means, the actual behavior Y is not in a line. So, there is a ups and downs, but the predicted line will be will be in a line, that is what is called as a line of the best fit. Since it is a linear one so; obviously, the particular line

will be you know like a straight line and the line should be in between all these points, that because it justifies the average relationship between dependent variable and independent variables.

So, theoretically we are drawing the line like this, but actually it can be derived mathematically so; that means, technically once you know the parameters value. So, we can get to know, what are these points corresponding to this particular; you know estimated variable and then, you can draw the means you generate the points and draw the lines. So, we defined the particular line will be straight line and that will be within the in between all these points.

So, technically Y head equal to that is estimated equation which is nothing, but alpha plus beta X. So, Y head equal to alpha head and beta head. So, when we write actually first hand equation Y equal to alpha plus beta X plus epsilon that is the error term. So, this is econometric modelling in the first hand, but in the process of estimations we will get a estimated line which is called as a Y cap and that too which is equal to alpha cap plus beta cap into X. Alpha cap is the estimated value of alpha and beta cap is the estimated value of beta.

So, now the moment when we transfer Y equal to alpha plus beta X plus epsilon to Y head equal to alpha head plus beta head X. So, error will be actually minimized so; that means, technically we need to minimize the error term. The way we minimize the error term by default we can get the values of you know alpha cap and beta cap. That means, technically your error equation equal to Y equal to equal to Y minus Y hat and that is Y cap that Y equal to alpha cap minus beta cap X ok.

So; that means, we need to minimize this e and that is how we have to optimize the particular process means technically we if you go for optimization then the single e cannot be optimized. So, what we will do? We will we will go for you know squaring the particular you know error term that is called as an error mean scale. And then we will minimize the error mean depending upon the particular you know you know a situation or the kind of you know equations.

So; that means, technically so, what will you do mathematically? So, we can square both the sides. So, then it becomes a square equal to Y minus Y minus alpha head minus beta head X squares and then we take summation both the sides. And now we have to

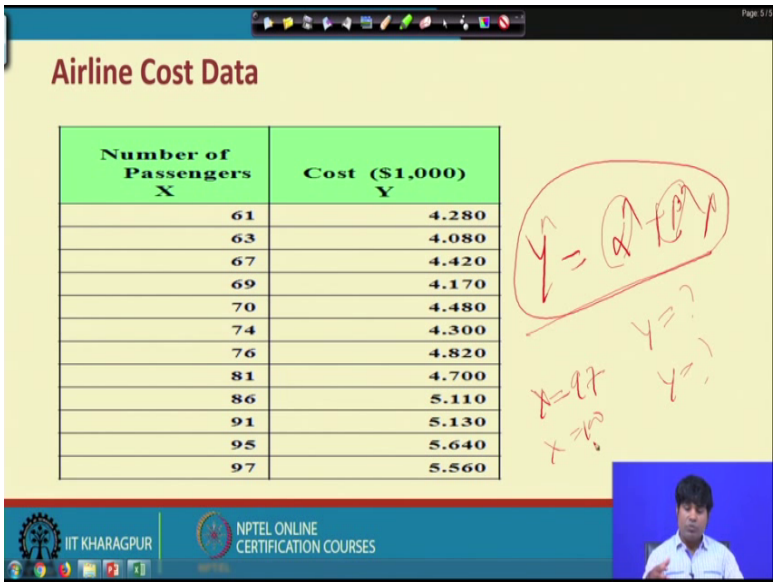
optimize or minimize the errors on squares with respect to alpha head and beta head. Because, these are the only 2 parameters which we need to know because previously they are unknown.

And the process of this minimization will bring you bring you the alpha head value and beta head value; that means, initially we have the equation where alpha head alpha and beta are unknown now, with the process of you know estimation we can get to know this you know values of alpha and beta.

As a result, now alpha and beta are known to you. So, the moment you know alpha and alpha head and beta head, then it is easy to predict the particular you know Y corresponding to the involvement of you know X. Because, in the in the process of you know data as we defined. So, this is what is the actually X data so; that means, technically against 61 passenger, the cost factor is 4.28 against 63 4.08, likewise the last component will be when there is a 97 passengers the cost factor is 5.56.

So, now likewise there is a kind of an increase number of passengers 61 to 63, 67, 69, 77, 74 then 91, 95, 97. So, likewise let us say if it is 99 then what should be the cost factor? So, that is how the simple understanding of this particular problem. So, now, once you know you know a particular you know equation then it is very easy to predict the kind of you know for instance. The line is equal to Y equal to alpha head plus beta head X.

(Refer Slide Time: 25:13)



So, now for the next you know cost factors depending upon the number of passengers. So, here it is last number is 97, now if we assume that if X equal to say X equal to say 97 or X equal to hundred then what should be the X value? Ok what should be the Y value? So, now, by the process of regression modelling we will get a equation like this where alpha head and beta head are known to you and for that we are minimizing the error sums here and then getting the value of alpha head and beta head.

So, now since alpha head and beta heads are known may be equal alpha head equal to 2 and beta head equal to 3 and; that means, technically the equation will be Y head equal to 2 plus 3 X. So, now, putting X equal to 97 so, you can get Y head value and again putting X equal to 100 you can get the Y head value. So; that means, it is very easy to predict when the values of you know alpha and beta are known to you. So, now, our first task is to know, how these values are you now means this values can be obtained ok.

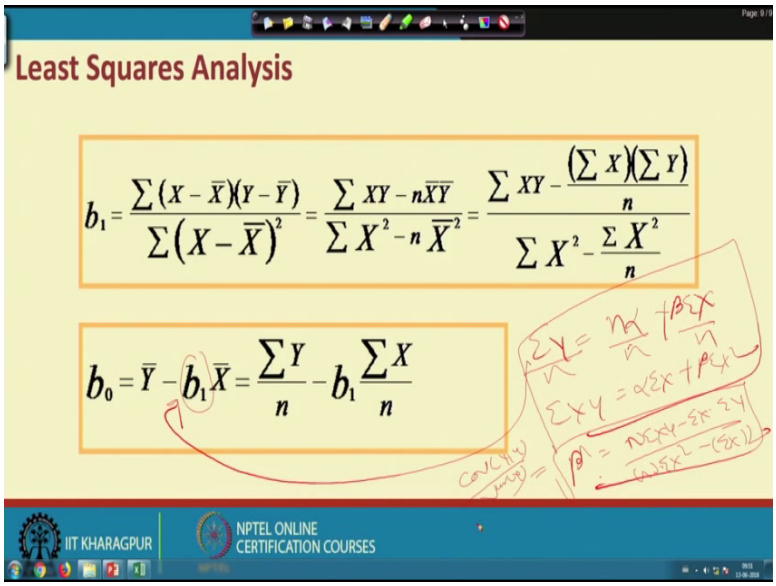
So; that means, technically we like to know the procedure through which you know alpha value and beta value can be calculated. The moment you calculate the alpha value and beta value, then you can easily predict a predict the particular you know cost subject to number of you know passengers right whether it is increasing decreasing there is you know such issue. But, ultimately you will get a kind of you know path through which you can predict the particular you know cost depending upon the number of you know passengers involvement.

So, now, to do that so, we can proceed for the particular you know processing and this is what the modelling. And then let us assume that this is what the estimated equations. So, we assume that the estimated equations which we have already obtained is called as a Y Y cap equal to alpha head plus beta head X which I have already highlighted in the previous slide.

So, now here b 0 is nothing, but called as a alpha head and beta b one is nothing, but called as a you know beta head. So; that means, technically beta 0 b 0 is a kind of you know intercept line of the intercept and b 1 is the slope and Y head is the predicted value of Y and then now we like to know how b 0 and b1 to be calculated. Of course, b 0 is nothing, but alpha head that is the estimated value of alpha and b1 equal to estimated value of beta that is beta cap.

So, what is the typical you know structure or the kind of you know formula through which you can calculate the b 0 and b1 that too alpha head and beta head? So, for that the process is to you know minimize the errors sum squares and the moment you will minimize the error sums square you can easily get the values of this parameters. So, ultimately the process we like to follow is like this.

(Refer Slide Time: 28:35)

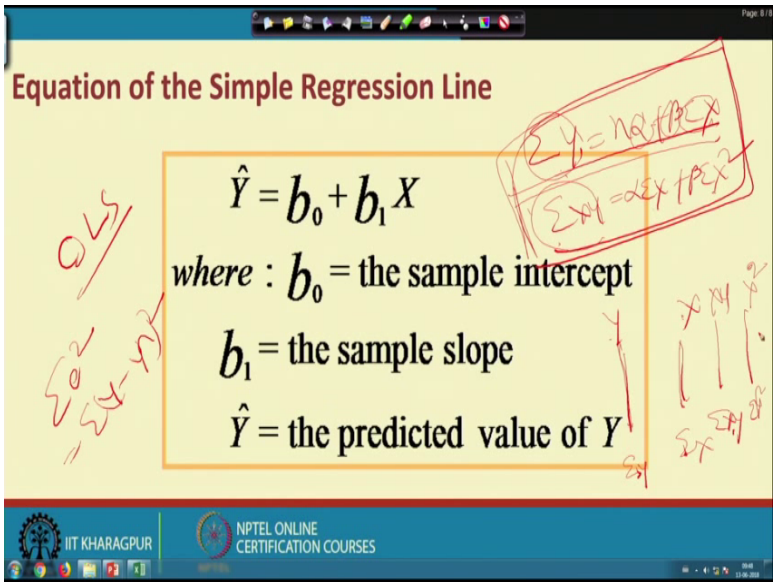


Usually, the process will be you know process will be followed by some of the tools or you know techniques that are called as you know ordinary least squares, like you know generalized least square, weighted least squares, maximum likelihood estimators. So,

these are the different techniques are available to help the process that too minimize the error some squares and to get the values of these parameter.

But, usually in the regression modelling or econometric modelling the starting procedure to have the estimated values of this parameter by deploying the technique called as a OLS, that is called as a ordinary least square methods.the process of ordinary least square is a to minimize the error sum.

(Refer Slide Time: 29:28)



So; that means, here we use a technique called as an ordinary least squares. So, here the process of you know OLS is to minimize the error sum square summation square which is nothing, but summation Y minus Y cap whole squares and that too we have to minimize with respect to the unknown parameters that is alpha and beta. And the moment we will you know minimize then the process will give you 2 different equations all together.

So, the equation will be like this summation Y equal to summation Y equal to n alpha plus beta summation X and summation X Y equal to alpha summation X plus beta summation X square. So, now, these are the 2 equations which will get from the process from the process of you know OLS, OLS that is through you know minimizing the error sum square.

And since we are minimizing with respect to alpha and beta, so, there are 2 different procedures, first is the necessary condition and the sufficient condition. The necessary condition is a, differentiate the particular equation with respect to alpha and differentiate the particular equation with respect to beta. The moment we differentiate the particular equation with respect to alpha and to and then to simplify the process we will get the first equation which is nothing, but summation Y equal to n alpha plus beta summation X.

And again when you differentiate with respect to beta and then simplify the process. You will get a equation called as a summation X Y equal to a summation alpha summation, X plus beta summation X square. So, now, you know data structures we have Y value and we have X value ok. So, now, if you got to this modeling so, we need to find out alpha value and beta value. And by the by the way in the dataset we have a Y information and X information. So, that too we need summation Y since it is a Y series we can easily get summation Y.

And since again X series you can easily get a summation X just to add these numbers with respect to Y then you will get summation Y and again you add all the X points then you will get summation X. And for first equation the unknown component is summation Y which we can derive from the Y series and summation Y which is derived from the X series and n is the sample size number of sample points. And then in the second equations the requirement is a summation X Y.

And for that Y and X we can create a variable here X Y just to multiple Y with respect to X then we will get a series here and again take this sum we will get summation X Y. Then, summation X is already here. And we need again summation X square since X is a series here. So, we can create another series here X square by squaring X item, then you generate the points and finally, you will get summation X square here.

So, ultimately with the data set you can easily get all these items and then put these items in the particular you know in this 2 particular equation and simplify you will get the values of you know alpha and beta. And once you get the alpha values in the in this process and beta values in this process. So, alpha can be treated as a beta 0 and beta will be treated as a you know b 1.

So; that means, previous alpha and beta are unknown now this estimated beta alpha and beta are known to you and which is which are named as you know b 0 and b1 and that

that these are all derived from the particular you know sample. So, that is the process through which you can get the estimated equations or estimated line and that too the values of the parameter.

And now with the with the help of these 2 equation, if we will simplify then you will get you know b 0 like this and b 1 like this for instance the equation structure will be equation structure will be like this.summation Y equal to n alpha plus beta summation X and summation X Y equal to alpha summation X plus beta summation X square.

So, now you can divide n both the sides then if you simplify then this will give you to this equation and as a result it will give you solution to alpha component and now forfor beta component that is b 1. So, you have to simplify the entire equation and that will give you to value like this. So, it is beta estimate equal to n summation; X Y minus sum X into sum Y divide by n summation X square minus sum X square.

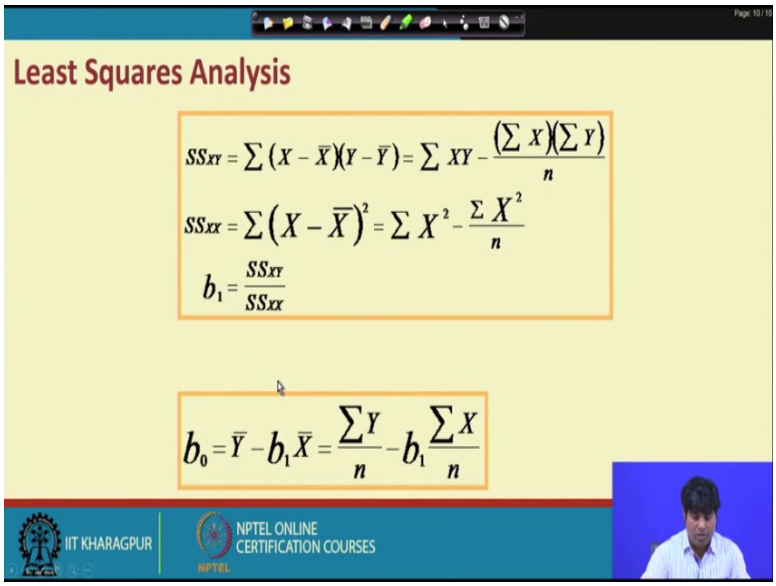
So, this is actually the kind of you know beta coefficient, the process is to first find out beta coefficients by this particular formula. It is something called as you know covariance of X Y by variance of X.

So; that means, beta coefficient for a bivariate model is nothing, but a covariance of X and Y divided by variance of X and then you calculate first beta coefficients and after getting the beta coefficient you can put here in the first equation and you will get the alpha coefficient. So; that means, this is a mathematical procedure through which you can get alpha coefficient and beta coefficient to predict the dependent variable that is Y with respect to independent variable X.

Of course software will help you to get this values very easily, but this is what the a theoretical understanding or technical understanding how to get or how to process the particular you know you know modelling. So, that the alpha values can be obtained and beta values can be obtained in the process of you know OLS estimation in that too you know determining the cost or predicting the cost with respect to airline passengers ah.

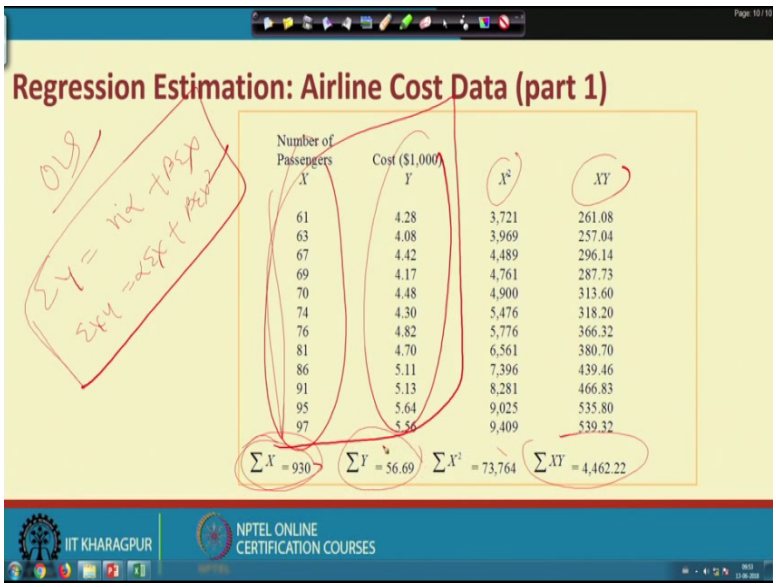
So, these are the process through which you can actually obtain the values of the unknown parameters.

(Refer Slide Time: 36:35)



And once you get the values of the unknown parameters we can easily read the kind of you know a Y and this is how simple procedure what I have already mentioned you know beta coefficient is nothing, but covariance of X Y by variance of X. And this is nothing, but actually Y bar minus b beta head X bar since beta head is obtained here. So, you can just simplify and get the answer.

(Refer Slide Time: 36:57)



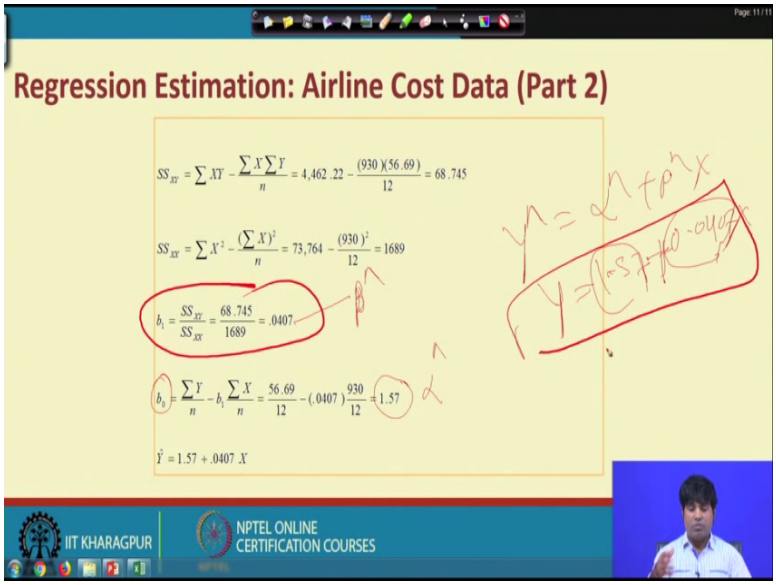
And for that this is what you know procedures. So, we have X information and we have a Y information. The way which I have already highlighted you can easily get summation

X Y because, your our standard equation equal to summation Y equal to nn alpha plus beta summation X and summation X Y equal to alpha summation X plus beta summation X square.

So, now, from X and Y this is our original data structure then with the process you can get X square series you can get X Y and that too as per the requirement of these 2 equation and which we will derive from the process of OLS that too minimizing the error sum square. And to obtain the values of the alpha and beta and now creating X square you can get summation X square, creating XY you can get summation XY and with the availability of X you can get easily summation X and with the availability of Y you can easily get summation Y.

Now, put all these value in this 2 equation and then simplify you will get the values of these parameters. In the next slide it will give you the indication here.

(Refer Slide Time: 38:11)



So, now, after putting all these things, so, you will get a beta one first that is the covariance of X Y you know covariance of X Y by variance of you know X. So, that is for the particular you know beta coefficient which is called as you know b 1 and this is alpha coefficient which estimated alpha coefficient which called as b 0 which is nothing, but Y bar minus beta head X bar that is b1 X bar and as a result you will get this values. So, this is technically b e called as beta coefficient and this is technically called as a alpha coefficient.

So; obviously, our estimated equation Y head equal to alpha head plus beta head X. So, now, alpha head is coming here 1.57 and plus beta coefficient is coming 0.0407 that too X. So, this is how the estimated equation and this is estimated equation or estimated line and for that you know you can now predict the Y with respect to X. Because, alpha is now known and beta is now known of course, they are constant whatever values of you know you know X and that too predict the Y.

So; that means, actually X will change one point to another point, but alpha intercept will b e constant and beta intercept will be constant for this particular you know sample. Because, it is now the average values of the values of this particular series that represent the parameter and then you can predict X with Y. So, this very easy to process and then it will help you lot to predict you know some of the engineering problems like you know the transportation cost or something like that.

So, the so, ones you understand is clear and the process you will be acquainted then you can actually easily workout. And then it will help you lot to examine the kind of you know things and to decide what should be the weighted cost subject to the passengers, increase of passengers, decrease.

And after getting the predicted lines you can do the prediction in this problems that too airline cost subject to airline passengers. But, before you go for you know prediction about the airline cost with respect to airline passenger, the model need to be checked for factor, this is the estimated models. Now, you can easily go ahead with the predictions.

But, there may be some of the requirements means there are some of the errors which you need to check and clarify before you go for the predictions. So, in the next class we will discuss all these obstacles. Before you must start the prediction you should you know check all these obstacles. And you know give the green signal that yes this model is perfect to predict and you can go ahead with the predictions as per the particular you know engineering require engineering problems requirement. With this we will stop here.

Thank you very much.