Business Analytics for Management Decision
Prof. Rudra P Pradhan
Vinod Gupta School of Management
Indian Institute of Technology, Kharagpur

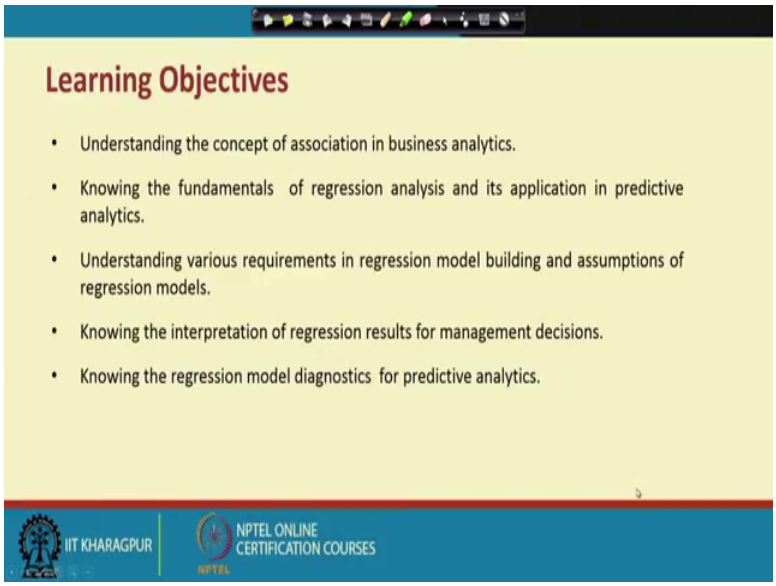
Lecture – 26
Predictive Analytics

Hello everybody, this is Rudra Pradhan here and welcome you all to B M D lecture series. Today, we will continue the new component called as predictive analytics. In the predictive analytics, we have three parts; predictive analytics one, predictive analytics two, and predictive analytics three, and in the last couple of lectures we have already discussed various issues, relating to business analytics that too we have discussed descriptive analytics, and inferential analytics one, and inferential analytics two, and all these analytics, the descript analytics and inferential analytics that is the part 1 and part 2 are very useful or in fact, it is a more or less compulsory for this particular, you know predictive analytics without knowing the descriptive analytics and inferential analytics is very difficult to understand the concept relative analytics.

So, my suggestion is you must have actually enough knowledge or enough exposure, on the basis of descriptive analytics and the inferential analytics and then we will come down to predictive analytics and. In fact, in the business analytics, predictive analytics tool is a kind of you know key component. So, you know through the, through these predictive analytics, we can go for you know prescriptive analytics and for the predictive analytics we need to have a connection with the descriptive analytics and inferential analytics.

So, now in this particular you know predictive analytics, a particularly the part one, part two, and part three. We will discuss various issues, you know business related problems and dealing with, you know the kind of, you know prediction and forecastings and then we are in a position to knows certain, you know the decision making process through which you can solve some of the business problems and let us see what are these, you know concepts behind predictive analytics and how we will proceed and what are the techniques through which you can solve the business problems and coming to, coming for a kind of you know management decisions and the objective of this particular.

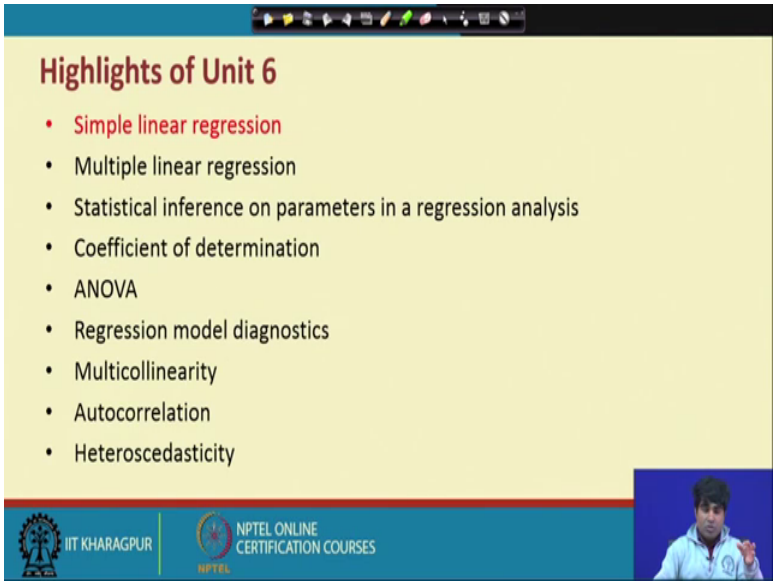
(Refer Slide Time: 02:33)



You know unit is like this, understanding the concept of association in business analytics and knowing the fundamentals of regression analysis and it is application infective analytics, then understanding various requirements in regression model building and assumptions of regression models and knowing the interpretation of regression results for management decisions, knowing the regression model diagnostics for predictive analytics.

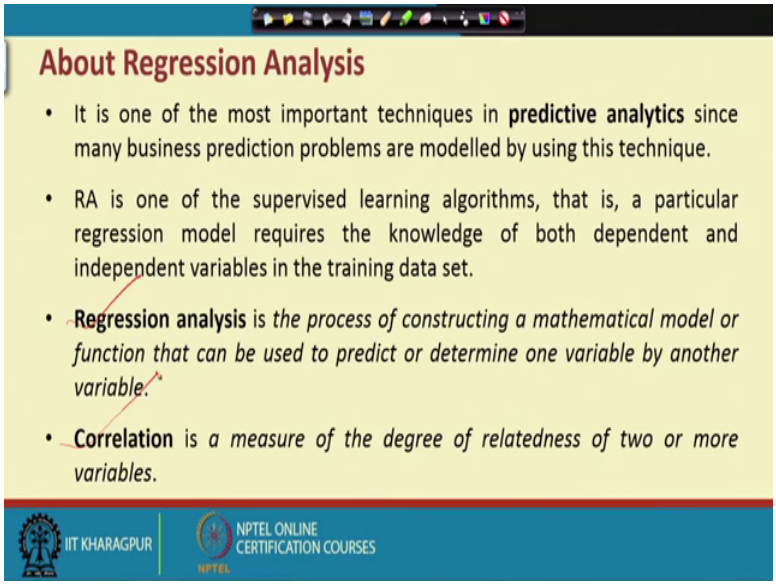
So, these are the broad objectives, for this you know unit.

(Refer Slide Time: 03:02)



And the unit highlights are like this. So, we will start with you know simple regression structure, then we will connect with multiple regression and then is will be dealing with statistical inference, on parameters connecting to regression analysis, then we will you know coefficient of determination analysis of variance, then recall a address regression model with various diagnostics issues like multicollinearity autocorrelation and heteroscedasticity besides, we will be dealing with several other, you know predictive analytics tools. So, in this particular units, we will be dealing all these components.

(Refer Slide Time: 03:47)



So, let us start with you know, the simple regression analysis. So, in the predictive analytics regression analysis is one of the most important techniques and since many business prediction problems model by you know using this particular, you know technique and regression analysis is one of the supervising. You know learning algorithms that is a you know that is a particular regression model, requires the knowledge of both dependent, independent variable in the training data set.

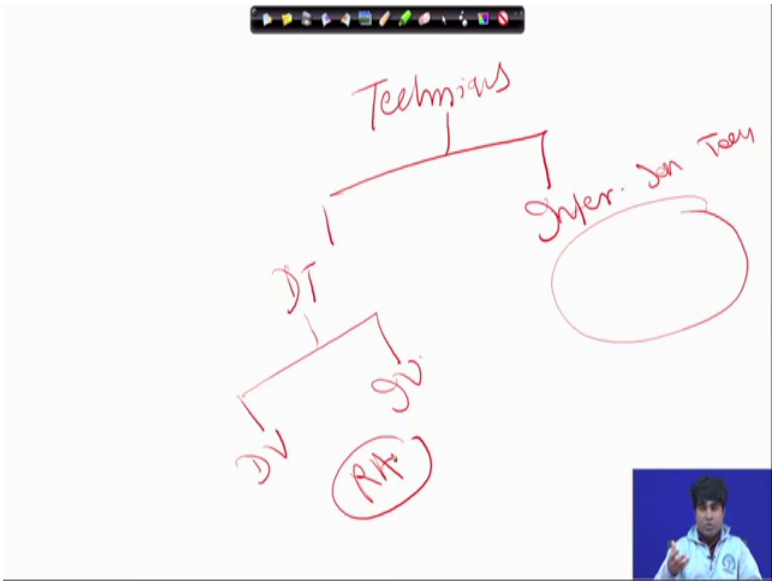
So, in the couple of you know, we expect, we have discussed the concept of you know associations and in that case, we have discussed the issues about the covariance and correlation and here we are dealing with the concept called as you know regressions and. In fact, we have already discuss the concept collagen correlation and a before correlations, the association technique is called as you know covariance and then it will be followed by correlations and what covariance and correlations can give you the kind

of you know degree of association between two and more variables; however, it will not give you the kind of you know, the causality part, which one is the cause and which one is the effect and after knowing the cause and effect then we are in a position to predict as per the kind of you know problem requirement.

So, regression analysis is a kind of you know what we call as you know dependence technique, through which is some of the prediction can be done, with the help of the kind of you know independent variables in. In fact, in the last couple of lectures, we have discussed about the understanding about the variables and there you know we have classified variables into two parts; dependent variable and independent variables.

So, as a result so, we have a kind of, you know different clusters of you know about analytic tools, where we need to know the dependent structure and independent structure, as you know that will help you to predict the kind of you know things. In fact, the kind of you know structure is a (Refer Time: 06:00). So, the business analytics particularly for predictive analytics and that to the so far, as you know association structure is concerns. So, we have actually two groups of you know techniques.

(Refer Slide Time: 06:16)



So, basically the techniques can be divided into techniques, can be divided into two parts, one is called as you know dependence technique and the another is called as you know interdependence techniques. So, in the case of interdependence techniques, we need to have a classification about the dependent variables and the independent variable,

in the case of interdependence technique, all most of the variables are you know seems to be interdependent to each others, under you know dependence technique. So, one particular tool is a nothing, but called as you know regression analysis. So, we will be dealing with now, the regression analysis, but you know in the analytics baskets, particularly the kind of you know association techniques.

We have a plenty of you know techniques, but one of the most important technique, which will be very useful for most of the business related problem is a nothing, but collagen, you know regression analysis.

So, we will be you know dealing this particular concept and this lecture will give you some kind of snspshot about the concept of you know, regression analysis. How to start the particular process? How to connect with the problems and what are the process through which you can, you know predict, the kind of you know business requirement and then accordingly we can, you know come for, you know some kind of you know management decisions.

Let us see, how is the particular structures and then we will follow up a particular problem. So, you know. So, that we can, you know we can address the business problems as per the requirement and typically in this lectures. I will highlight the bivariate structure of you know regression modeling then I will connect with his some problems and then I will take, you know to the, you know data analysis that is a spreadsheet to highlight, how the particular, you know prediction can be happened. If you know, if you have a problems and you know relating to that problems, you know you may have actually data.

So, having problems and having data. So, you know, you can use actually kind of in a spreadsheet and then you can do the kind of interpretive analytics and that too through regression analysis. This is what the summary of these particular lectures, let us see how is the kind of you know concept. So, this is how the background about the, you know regression structure. So, it is the process of constructing a mathematical model or a function that can be used to predict or to determine one variable by another variables, that is what the structure of dependence technique.

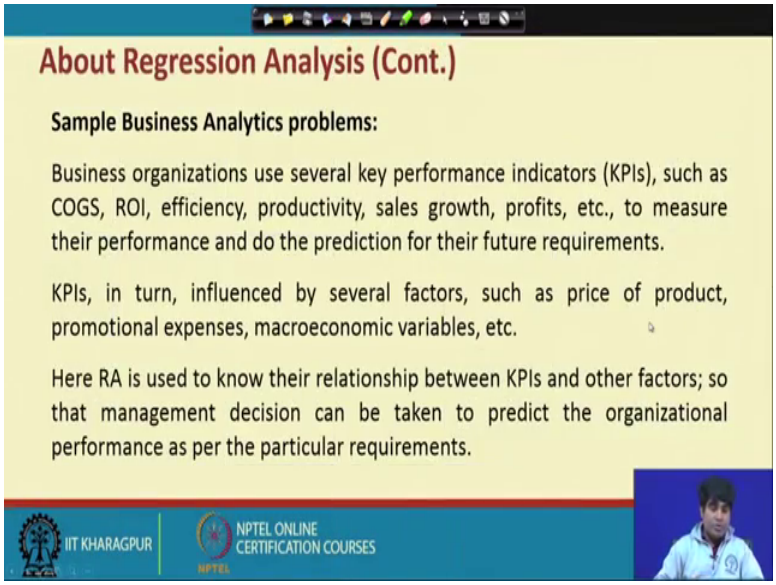
So that means, here is we are, you know trying to predicting to the usual structure collagen dependent variable and that to with the help of you know independent variables

correlation is a measure of the degree of you know, associations of two and more variables and a compared to regression analysis. So, you know this is not. So, you means correlation is not. So, effective to predict, you know useful for you know kind of in a prediction, but regression is the useful technique, which can be very useful for kind of you know prediction.

So, when we are dealing with you know regression, you know problems relating to regression analysis by default, correlation will be there in, you know by the process right. So, let us see how is this concept and how correlations and regression can work together and. In fact, in the bivariate analysis particularly by create regression. So, you know the summary of the regression output can be represented through also correlation, you know analysis. So, I will show you the typical structures. So, so the; that means, typically the understanding of regressions, you know highly depends upon the understanding of the correlation, if you do not understand the correlation, then you may not in a position to understand regression.

So, by the way so for the business requirement for you know that making management decisions. So, you need to know these techniques and then connect the problems as per the particular, you know requirement. So, before we proceed actually.

(Refer Slide Time: 10:39)

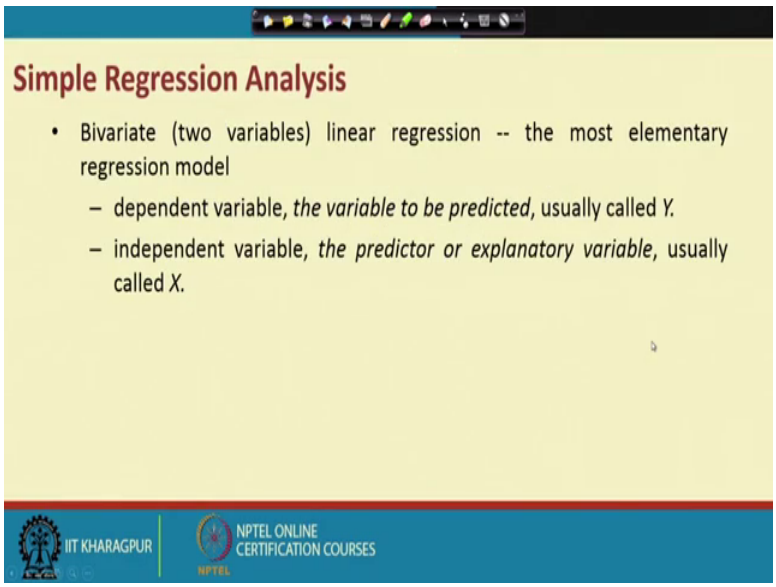


So, here is the kind of you know requirement, the requirement is a. So, it is a kind of you know so, let us you know, I will give you, know kind of you know connection. So, what

are the areas through, which you can actually apply business analytics, it means that to this you know simple regression, you know analysis. So, you know in many business organizations, there is a you know the performance measures, which you called as you know K P is key performance indicators. So, like you know cost of goods and services return on an investment efficiency productivity, a cells growth profits and these are the indicators through which you can measure the performance of a particular. You know company and we can do the, you know prediction as per the companies requirement or organizations requirement in this case regression. You know in the kind, you know prediction of you know organizational performance or you know industry performance. So, we have to use these key indicators that is the K P is and in fact, most of the instances.

So, K P is are you know influenced by several factors, these are the factors, which you stated as you know independent variables. So, what we like to know or the regression will be help you to know, you know how these independent variable are linked with the K P I, that is the dependent variable indicators or predict, you know predictive indicators. So, now, once you know the kind of, you know structure, the kind of you know strength, then we are in a position to predict, the kind of you know business performance or you know organizational performance. So, you know this is a simple, you know examples, where regression can be very effectively used for some kind of you know requirement of management decisions.

(Refer Slide Time: 12:25)

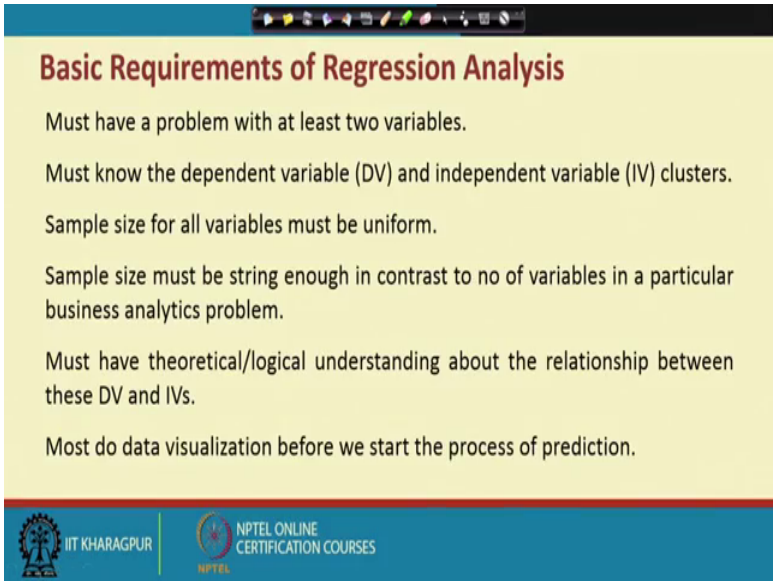


So, now I will take you to the kind of, you know structure of the regression analysis. So, we will be starting with you know a simple one then we will go; that means, technically regression can be a structured into three different formats, the bivariate that is what the structure called as a simple one, then trivariate and then we try to go for you know some kind of you know structure called as a multivariate, under multivariate.

So, we will have a some kind of you know multiple regressions, then actually there is a multivariate, where there may be many dependent variables along with many dependent variables. So, now, these are all the kind of you know complexity, but we need to know the simple one first, then we move slightly, you know slowly towards the complexity, but more or less if you are in a simple understanding is very clear, then you can follow up the complex one and simple in the simple regression analysis.

So, you know it will be dealing with two variables and that two, one is the dependent variables and the another one is the independent variable. It is the game between a predicted and predictors or explend, you know explanatory. So, this is the kind of you know structures, through which you have to discuss.

(Refer Slide Time: 13:39)



And before we start the process for any, you know any kind of you know problems, we need actually management decision and that to use of you know regression. So, regression has some kind of you know requirements or you know in this particular, you know predictive tools, in this particular predictive tool cannot be used, you know each

and every problems, it will be apply in a kind of you know environment, where you know minimum requirement, you know it should be, you know it filled up.

So, these are the minimum requirements you must have, before you go for some kind of, you know, using the regression analysis for you know, kind of you know business prediction and a first things must have a problem with at least two variables and if it is only one variable, then this technique cannot be use, then second one must know the dependent variable and independent variables, you know and then sample size for all variables must be uniform.

For instance, for the simple regressions, you need two variables and if there are two variables, then corresponding to these two variables, you must have a sample point and that sample point must be uniform; that means, technically if X is a variable have been 10 points, then Y must have a variable 10 points. If one is having 10 points, another is having 15 points or vice versa.

Then you may not be in a position to around the regressions, sometimes there may be the case that you know, you have 10 observation or 15 observation, then through some intrapolation and extrapolation, you have to, you know fill up the gap and then bring the uniformity, after that you can apply the regression and again a sample size must be strong enough to contrast, to a number of variables in a particular business analytics problem and they must have theoretical and logical understanding about the relationship between these dependent variable and independent variable and then you must have, you know go for, you know data visualization before you start the process of predictions.

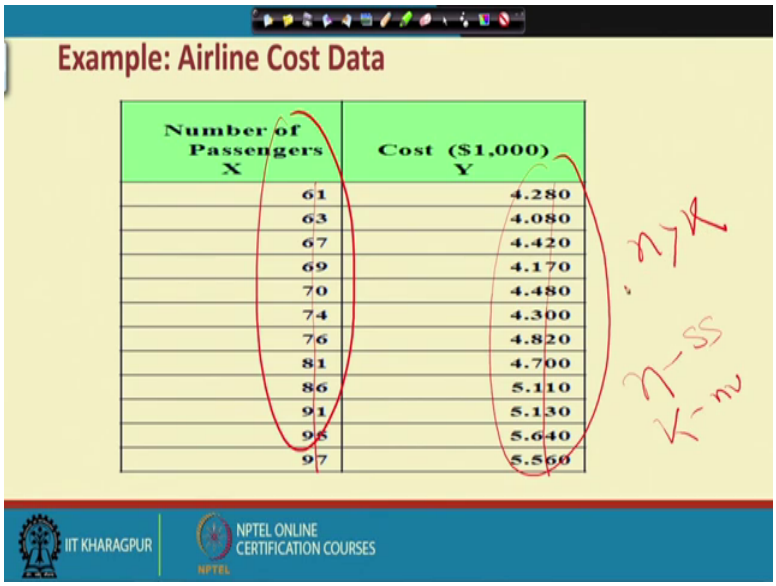
So; that means, this there you have to set a hypothesis. They you know with respect to the problems that through the you know theoretical hypothesis will be like this that you know you have to predict these cells or you have to predict the profit or you have to predict the return on investment, then subject to identification of the independent variables.

It means the factor which can affect these key performance indicators and in that case how they can be actually modeled means sometimes you know a model, means it is a kind of no functional relationship between the dependent variable and independent variable, but here is until unless you know the particular, you know structures that is the

visualization part then you may not, you know build the model perfectly. For instance, the functional relationship depends upon the kind of visualization of the data.

So, there may be chance that the model can be linear one. The model can be non-linear one. Let us say, if the visualization will give you, kind of you know perfect path through which you can go for, you know perfect predictions. So, this is what the basic requirements of the regression analysis.

(Refer Slide Time: 16:48)



And with this we can move to this particular process and then you know though I have theoretically highlighted, what are the areas through which you can apply the kind of you know regression analysis and that thorough the kind of you know predictive analytics and here the sighted example. This is the live example of you know a kind of, you know airline cost data and here is, there are two variables, the first variable is the number of passengers that is represented by X and the second variable is the cost represented by Y. So, mathematically Y is usually recognized as a dependent variable and X is recognized as a independent variable, but there is no hard and fast rules, what they are will start with the process, that you know. Y is every time dependent variable and X is the independent variable.

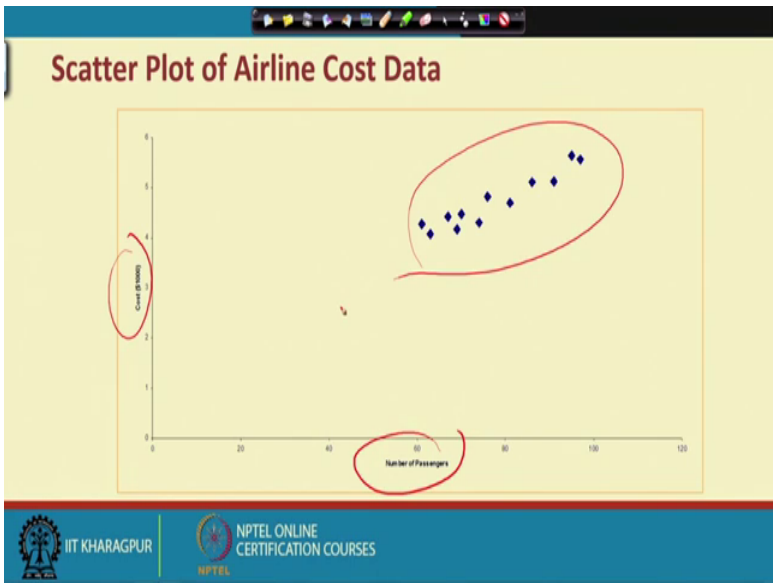
So; that means, see here actually, we have actually series of data points here. So, this is, this is actually the you know independent variable informations and this is the dependent

variable informations. What I have already mentioned that you know, you must have a uniform data points for both X and Y.

So, you have actually whatever data points are here at the same data points are also here so; that means, it is actually the first hand requirement is satisfied and the second hand requirement is here, the sample size should be greater enough to you know variable size, since there are two variables and we have actually you know more than two observations so; that means, you know the necessary condition is satisfied, but actually you know I mean, see in the kind of you know predictive kind of, you know structures. So, we will represented by the sample information, say n and the K I S the number of variables.

So, this is actually sample size and this is the number of variables and usually the structure is n greater than to K, but that is the necessary condition, but the sufficient condition depends upon the number of variables and the kind of you know particular requirement of you know data points usually, we have a small sample and large samples and we have already discussed in the case of you know inferential analytics and that part will it take care here, while dealing with the problems and the kind of you know predictions. So, let us see the game. How we can proceed, you know furthers to know these particular, you know tools and to predict this so; that means, in this problem our job is to predict the cost subject to number of passengers.

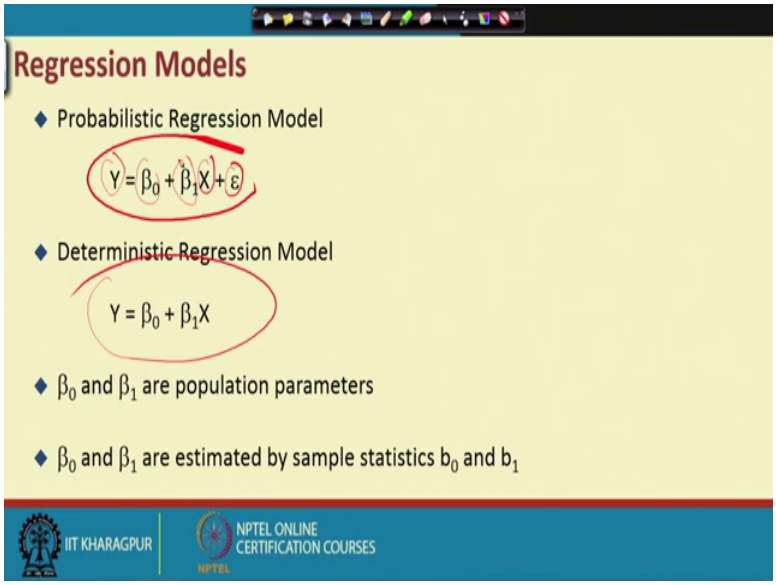
(Refer Slide Time: 19:11)



So, now how will do for the kind of you know, in how we can use the predictive analytics that to the use of in regression analysis. So, here in as per the requirement, we will go for data visualizations and in this case and. So, here the X axis is number of passengers and the Y axis is the cost and the plotting is coming like this and accordingly. So, you can actually get to know that you know there is some kind of, you know linkage or you know the kind of in a moment.

So, having the kind of in a slight hint from the data visualization and then we have actually theoretical knowledge that you know cost and passengers. So, that is actually customers are having, actually you know correlated and then we like to know how passengers can? You know help; they help you to predict the cost structures. So, actually you know it in the theoretically, that you know increased the passengers may increase the cost, but you know, we like to know or the quantitative pictures or the quantitative amount or quantitative, you know numbers through which you can do the perfect prediction. So, that is how the regression analysis can be applied and can be used for this particular, you know quantifications.

(Refer Slide Time: 20:29)



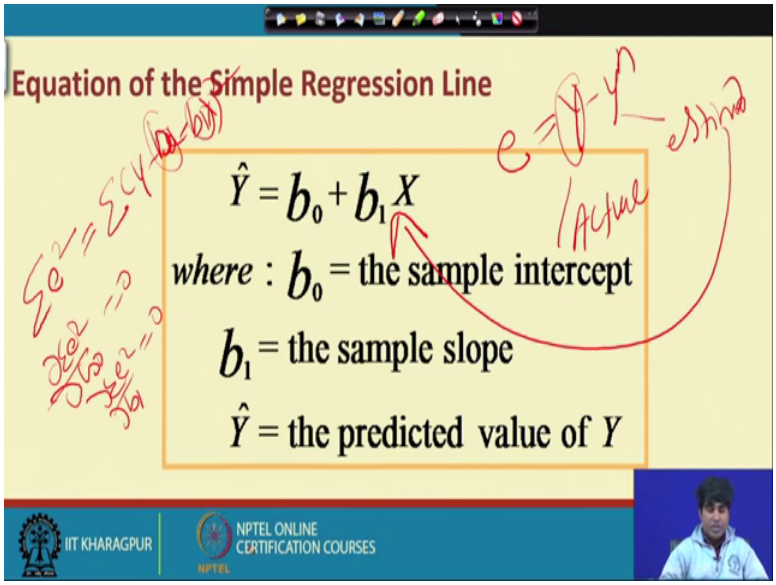
So, now moving forward so, we see here is that you know the basic regression models can be written like this. So, we are every times, we are dealing with the sample problem and then give the comment about the populations.

So, the same structure will, we can have here. So, having this sample data, you can actually build a models and then is you can predict for the, you know population that is the; that means, it is a kind of you know generalization and. So, it can be a probabilistic form and it can be the deterministic forms in that case of you know probabilistic. So, the the structure error term would be there in the system and in the case of deterministic. So, there will be no error terms and.

So, here is y is the dependent variable X is the independent variable and beta 0, beta 1 are the parameters and that need to that the population parameter and that need to be actually estimated for the kind of you know predictive requirements and beta 0, beta 1 are estimated by the sample statistics and followed by you know small b 0 and small b 1.

So, having this sample information between these two variables, we have to estimate the particular, you know equation and then get the estimated outputs and that is what the regression output and on the basis of regression output, we can go for some kind of, you know predictions.

(Refer Slide Time: 21:51)



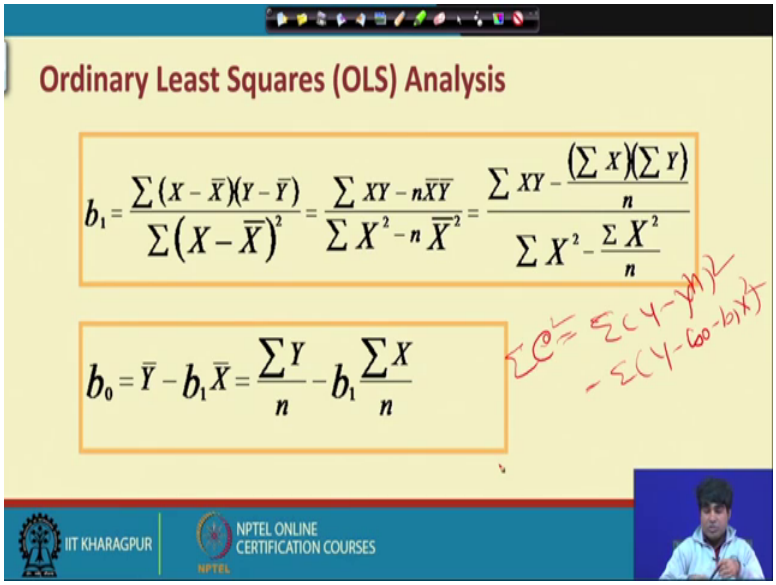
So, let us see how we can move further and this is the usual structure of an estimated equation. So, we will be dealing with you previously. So, this is the equation, which you are, you know going to estimate with the sample informations and then we will follow. Here is the, you know estimations after the estimation of the error terms by default will be removed and the final equation will be Y, estimated will be beta 0 plus beta 1 X, that

is actually sample intercept and the slope, you know slope coefficient and then we have to predict the, you know Y corresponding to the particular, you know X availability.

So, now so the, first hand requirement or regression will be, you know you are supposed to know the tool. How did you know parameters can be obtained from the sample data? So, these are the indicators through which you can do the predictions so. In fact, we have already discussed, you know demand predictions, you know self predictions in the first unity of this particular, you know lecture series, but now, we are going in depth, actually get to know how these are all coming, you know all together.

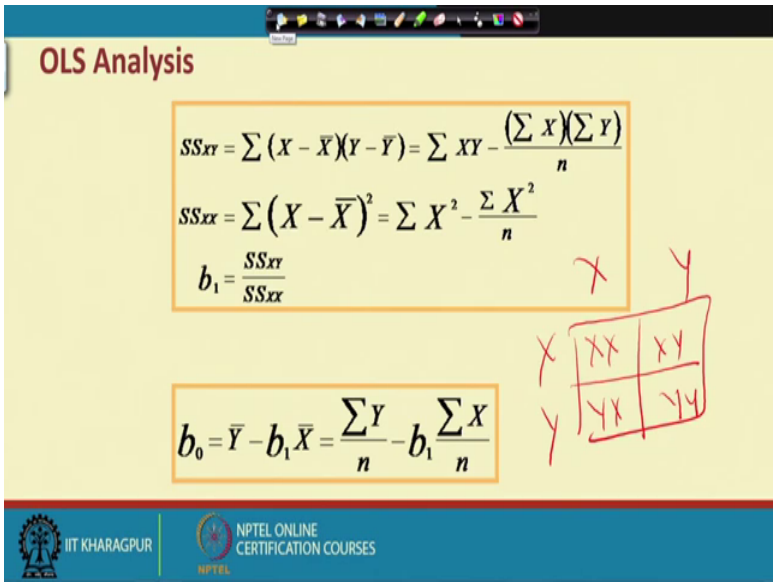
So, this regression will help you to know the, in depth understanding and the kind of in a inference exact, inference as well the particular, you know requirement. So, now, having you know, knowing the particular, you know structure. So, we can actually move further and then.

(Refer Slide Time: 23:23)



So, now we actually the particular process is the. So, this is actually the kind of you know this is the model, estimated model.

(Refer Slide Time: 23:28)



And then what we need to do actually, you know to get these parameters, we like to find out the error terms which is the difference between the actual values of the dependent variable and the estimated values of the dependent variables so; that means, technically. So, we have actually we need to calculate to E that is error component, which is the difference between Y minus Y hat.

So, Y is the actual value and the Y hat is the estimated value and estimated value will be derived from these particular equations and then the difference will help you, you to get the kind of an error component. Now, what will you do actually? So, we have actually some of the tools like you know one of the tool is ordinary least squares. So, we like to apply ordinarily least squares. To minimize this error, sum squares in, you know the way you will minimize the error, sum square then by default you will get the values of these parameters.

So; that means, technically what we like to do. So, we will we will square the both the sides and then we will take summation. So, technically summation E square equal to summation Y minus you know b 0 minus b 1 X and that 2 squares and then and the you know what will you do; since our requirement is b 0 and b 1. So, we have to minimize with respect to this summation E square by d beta 0 d b 0 equal to 0 and d summation E squares by d E b 1s equal to 0.

So, now, once you simplify then you know you will get, you know two different equations depending upon these two, you know variables or two parameters and this the solving, these twos you know equation with the help of, you know data by default, you will get the values of these parameters. So, in the next slides I will let you know, how these are all coming.

So, this is how the once, you simplify the particular structure then ultimately you will be get, you know beta b 1 coefficient, that is actually covariance of X Y by variance of X and this is on the basis of the previous. You know simplification that is with respect to d summation is square by d beta 0 b 0 and d summation E square by d d 1s, then once you simplify then the first hand value, you will get this 1s, that is the actually bit b 1 coefficient. Now, the second one is the, you know b 0 equal to Y bar minus b 1 X bar.

So, technically so you may find summation Y equal to means technically. So, what will you do? So, let me give you the idea, actually it is a, it is nothing, but actually E equal to oh summation E square equal to summation Y minus Y hat whole square that is nothing, but summation Y minus b 0 minus b 1 X whole squares. So, now, if we simplify then and you know, then you will get two different equations.

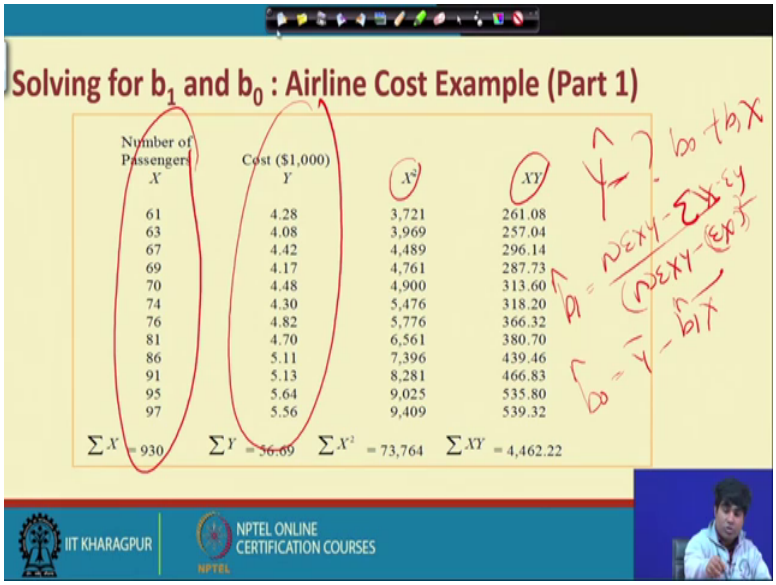
So, in the first equations will follow the, you know structure, then the second equation will follow the b 1 structure. So, the then you know ah after doing all these things. So, automatically you will get the beta zero coefficient and beta one coefficients. So, i am not going in the derivations, but ultimately once you simplify with respect to beta zero b zero and b one then by default you will get you know b one coefficient and you know ah b zero coefficients.

So, now b one coefficient and b zero coefficient will help you to predict the particularly you know y value subject to availability of you know x value. So, let us see here. So, how is the particularly in a structure then ah accordingly. So, if we will again in a simplify. So, it will be coming actually b one equal to s s x y sum of square that is actually covariance; that means, technically.

So, the structure will be like this. So, we have actually x y and then if you prepare the covariance matrix. So, it will be. So, this is x x and this is x y this is y x and this is actually y y. So, now, with the help of this covariance matrix we can we can in a position to calculate these values and accordingly you can also get the values of the b zero

coefficients. So, in the next slides i will let you know how these are all coming in a kind of you know a systematic process. So, let us see the kind of you know structure ok.

(Refer Slide Time: 28:23)

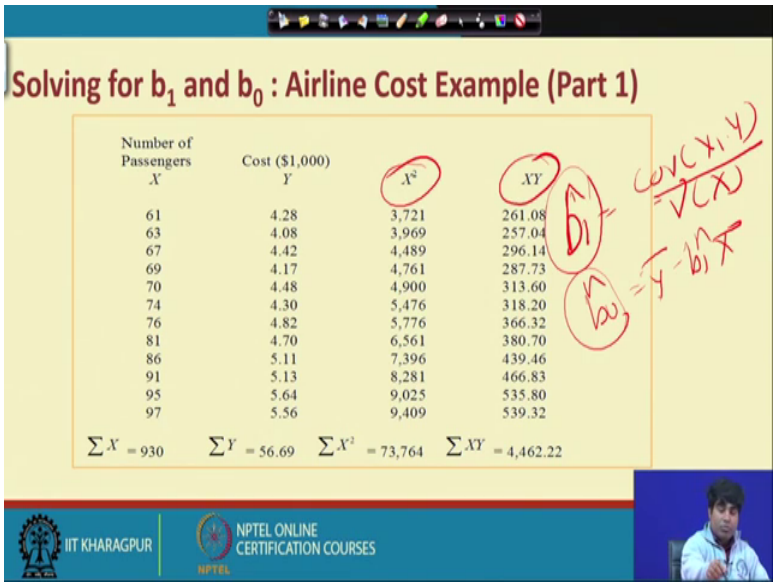


So, let us see here. So, this is actually as far our you know problem structure. So, we have actually x information and we have a Y informations and our job is to find out, you know Y predictions right.

So, which is nothing, but actually b 0 plus b 1 X and that two number of passengers and; that means, technically a cost is a function of actually the number of passengers. So, now, as per our requirement. So, you need to have actually. So, fast you know b 1. So, that is actually the n summation X Y minus sum X into some Y divided by n summation X Y minus sum X, you know some X whole square.

So, this is actually you know b 1 coefficient and b 0 equal to Y bar minus b b 1 X bars right, b 1cap and these are all actually estimators and so; that means, technically covariance of you know X by variance of X. So, now, for that actually so, you need actually XS square and then X Y. So, after knowing this so, automatically you can, you know, get the kind of you know values. So, it is technically and summation X Y minus sum of X sum of X square. So, this will be the kind of you know structuring. So, let me give you another way that mean copy.

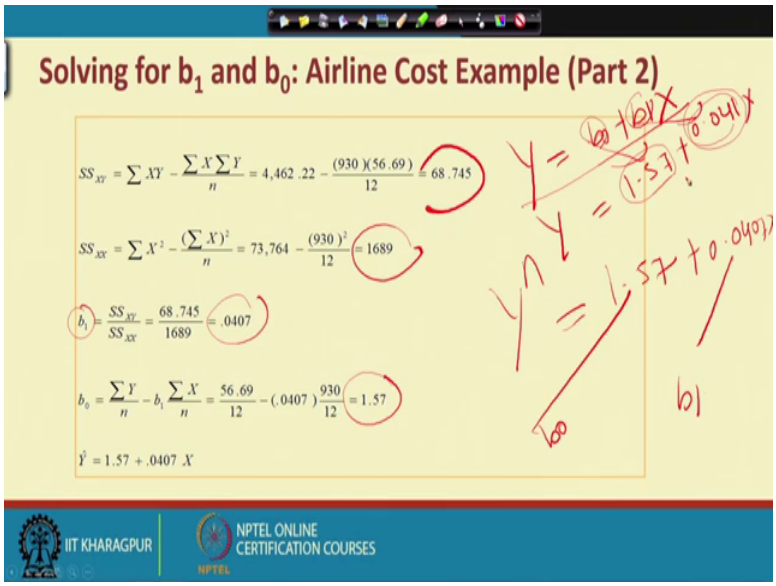
(Refer Slide Time: 29:55)



So, this is actually the beta one, coefficient that is covariance of X Y by variance of X and then beta 0 equal to Y bar minus b b 1 hat X bar right. So, this is also cap and and the first hand requirement is covariance. So, for that you need actually X Y and then variance of X, you need actually XS squares.

So, now, having this information so, first you find out the b 1 value and with the help of b 1 value you can find out the b 0 value so; that means, actually this is Y bar so; that means, it is the sum of Y divided by sample observation, that is the n and similarly, X bar that is summation X Y X by n that is the X bar of this, you know X series. So, after knowing the b 1, then you can easily get the b 0. So, that is how the the model requirement. So, once you have these model requirements, then you can move for you know, estimation or you know, kind of you know prediction.

(Refer Slide Time: 31:00)



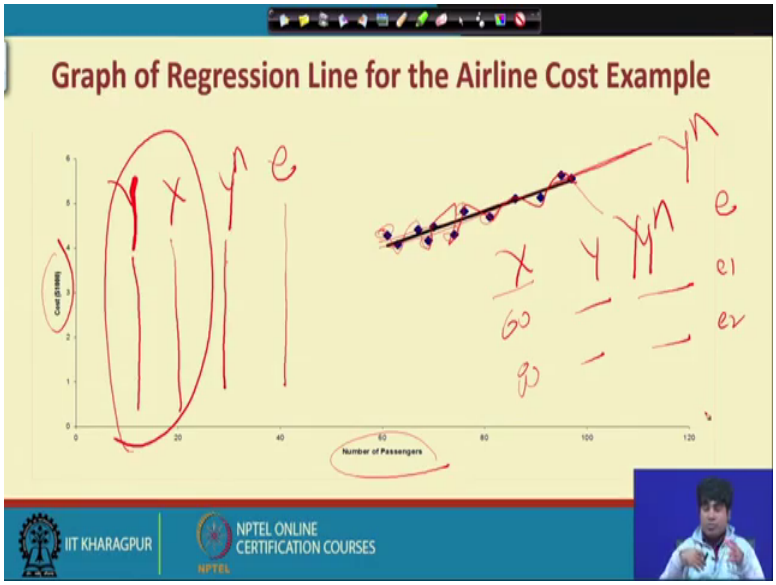
And. So, this is how the kind of you know simplification results. So, on the basis of you know. So, we need actually the lower part of these values. So, sum of X sum of Y sum of X square, sum of X Y and then you go to, this you know structures. So, here is actually the requirement. So, S S X Y, that is variance of X Y covariance of X Y and that is how it is coming here, actually 68.745, then this is variance of X and by default your b 1 coefficient will be covariance of X Y by variance of X and it is coming actually 0.041 right and b 0 or coefficient is coming, actually Y bar minus b 1 X bar and that is coming actually 1.57s, as a result your estimated equation will be Y cap equal to 1.5 and plus 0.0407 X right.

So, this is how the kind of you know estimated equation and this is what the b 0 and this is what the b 1s right. So, now, this earlier you know, earlier functional form is b 0 plus b 1 X. So, now, the model will be Y equal to 1.57 plus 0.041 X right. You know so; that means,. So, by default b 0 is this much and b 1 is this much. So, this is these are all unknown parameters. Now, these are all known parameters and that is obtained with the help of you know sample information, the past availability of you know data or the kind of you know the, you know problem structure.

So, once you are connecting with data and the models. So, the estimated output will give you the kind of in on non values, on the basis of these non values, you are in a position to go for some kind of you know prediction that to in this case. We are predicting the airline

cost subject to the end, you know number of passengers and in the case of you know further, you know about this particular technique.

(Refer Slide Time: 33:15)



So, we what we can do actually so, we like to check how is the particular, you know modeling right so; that means, technically. So, we have actually here is Y and we have actually X, then we will get Y, Yth victed and the E. So, technically if it will go to the excel spreadsheet. So, the original problem boundary will be like this and then with the help of this, you know estimation. So, we can get the Y head series and then Y minus Y hat series, you to get the E series right.

So, now, what is happening here? So, number of passenger here with is in actually X axis and the cost is in Y axis that is what we need actually prediction and as a result. So, corresponding to a number of passenger one point to another point. So, we have actually cost structure from one point to another point.

So, as a result so, the kind of you know these points are you know actual points, all did you know, all these points are actual. You know these are all called as a actual points, that is with respect to number of passengers and number you know and the cost structures, but the line, straight line is nothing, but called as a predicted line and now, if we join all these points, then the curve will be coming like this and this is curve, is nothing, but college Y actual and the straight line that is the predicted line is nothing, but college a Y head. Now, you know for every point, for every you know passengers, let us

say 60. So, for you know, when X equal to 60. For instance, X Y and Y Y head right. So, now, for X equal to 60.

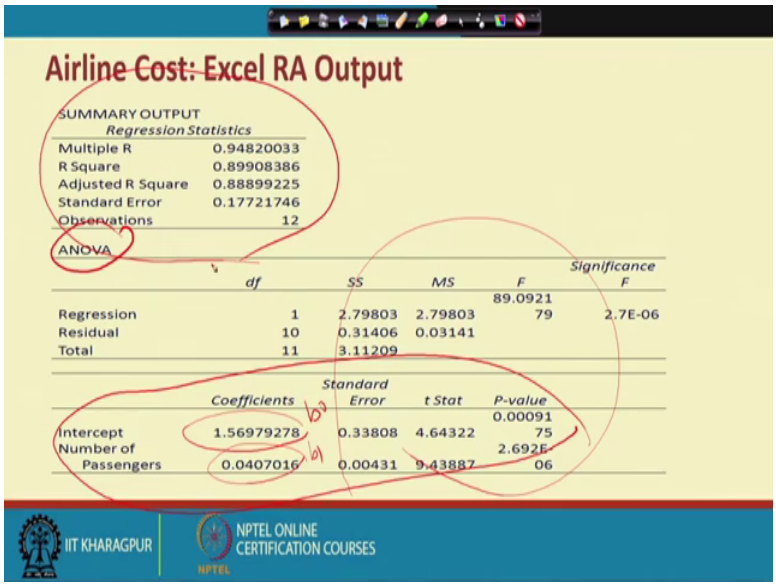
So, you have a Y information and we have Y head information similarly, when it is 80. So, you have Y information and Y head information. Then now, the difference between the two will give you the error terms right. Well, you know like E 1 and E 2. So, what we like to check finally, that you know how is the position of this particular predictor lines usually, the regression line, this called as regression line and otherwise it is called as average line. So, the line by default of course, by you know mathematical process, we are getting this particular line, but you know.

So, far as a validation is concerned, so, the line can be, the line looks like you know 45 degree angle. So, it can be, it can be further you know ups further, you know downs, but the thing is that you know. So, whether this line is perfect or not. So, we have a lots of you know checks, you know on the basis of these checks you can find out, to the kind of you know or you can justify that you know this line is perfect for the kind of no prediction or the kind of a requirement. So, one of the one of the cross check is nothing, but you know the error sum should be equal to 0. So, if you find this sum error is coming exactly 0.

So; that means, actually when you draw the printed line then some of the actual points will go up and some of the points, you know actual points will go down. So, then what is the you, how you to judge that. You know this line will be best fitted line. So; that means, the sum of the upper points must be equal to some of the lower points corresponding to the estimated, you know that is the regression line, the estimated regression line.

So, if that is the case, then you know we can say that you know this model can be immense. There is a reliability or in there is a kind of you know accuracy through which you can do the predictions of course, there are lots of other diagnostic checks, but this is the first hand kind of you know, you know inspection through which you can do the kind of, in a predictions as per the kind of in a problem requirement.

(Refer Slide Time: 37:06)



And accordingly so, we can move further and. So, this is actually the kind of you know overall regression output. So, on the basis of you know the Y information and X informations.

So, you can actually find out the entire regression output and the regression output will help you, know for, you know the Y prediction, corresponding to you know acceptability or you know X informations. So, in the next lectures, I will take this problem to the spreadsheet and then I will show you how these outputs, regression outputs are actually coming and now in this, in this output, you defined you know there are plenty of informations and there is a box.

Here is, like you know anova and which you have already discussed and then the, these are all collagen or regression outputs, that is you know estimated on the basis of Y and X so. In fact, we, you know till now we only know, how to get the coefficients of you know b 0 and coefficients of you know b 1s right, but in this particular, you know process, you we have actually lots of, you know other informations, you know why these informations are there and how they are, you know how. You know these are all coming in the regression. What is the utility of these informations and how this information can be helpful for this predictive analytics? That we will discuss in details, in the next lecture. So, with this we will stop here.

Thank you very much have a nice time.