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

Lecture – 12
Descriptive Econometrics (Contd.)

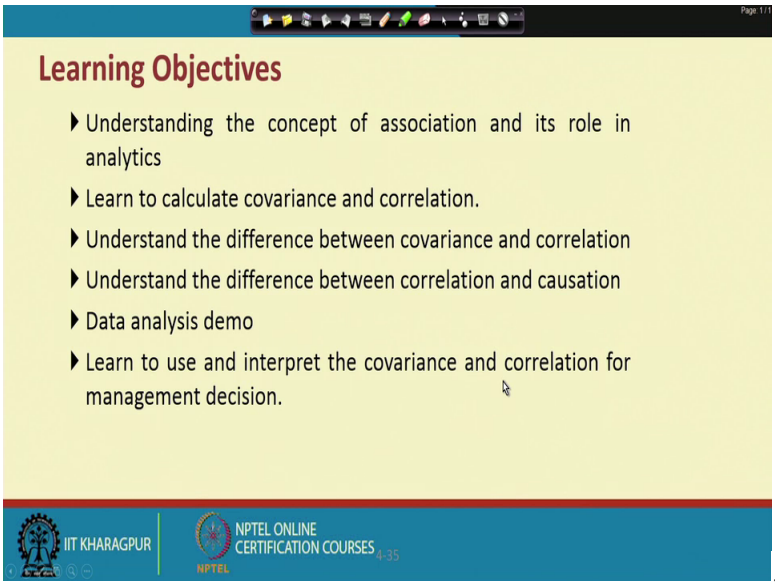
Hello everybody. This is Rudra Pradhan here. Welcome to Engineering Econometrics. Today, we will continue with Descriptive Econometrics. And, in the last lectures we have discussed about this particular component. And, specifically we have discussed the issues like central tendency dispersion and a kind of you know spreads of the particular you know data series or you know distributions.

So, these are all you know collagen of basics and some sometimes called as you know univariate statistics and that will help a lot to analyze certain problems. And against that is the kind of you know, typical requirement of a hard core kind of you know engineering econometrics.

So, when like to solve a complex problem. And, this descriptive econometrics that to the kind of you know central tendency dispersion, then spreadness all these things are you know very much requires, until unless you know details about the data structure and the kind of you know variable natures. You may not in a position to pick up good models to analyze any kind of you know engineering problems.

So, that is a it is a mandatory to know certain things in the basics, that is what we can call as you know basic statistics or descriptive econometrics, and something called as you know unvariate statistics, or univariate econometrics to analyze the kind of you know data. And, that to know something more about the data before we proceed for some kind of you know, complex empirical investigation you know process. And the idea about this particular you know topic is like this.

(Refer Slide Time: 02:12)



Ah. Since, we have already discussed the kind of you know unvariate structures, then slightly move into bivariate structures where, we have certain you know a tools through which we like to know the measurement or we like to know the relationship between 2 or more variables.

In fact, the accurately you know engineering econometrics we will start with you know more number of variables, maybe at least with you know 2 variables or you can say 3 variables or more than 3 variables, but with a single variables then you know hardcore econometrics may not be applied. But, when we solve a particular engineering problems, we assume that there are you know various variables which can be a which can be used or which can be very helpful to analyze the engineering problem, but in that contest every a variables corresponding to the you know data availability, we need to check independently and through some kind of you know relationship measurements and then we proceed for some kind of you know hardcore a you know engineering investigations.

For instance, in this particular le lectures we specifically highlight and the bivariate a structures like you know covariance and correlation. And, that really add value to the hardcore engineering econometrics.

So, the basic idea about this particular lecture easy to know the concept of association and it is role in you know analytics or some kind of in econometrics, learn to calculate covariance and correlation, to understand the difference between covariance, and

correlations and understand the difference between correlation and causation. And, then we the really solve some of the problems with the help of you know data; this is like we have discussed earlier. So, whether it is a univariate structure or bivariate structure or multivariate structures or whatever may be the kind of problem.

So, we have 2 different setups. The first setup is the input structure and then the second set of is the output structures, corresponding to a particular engineering problem. So, we have an input clusters where variables are identified named properly and we must have a data. And, then the second stage of you know processes the engineering economics output and for that these inputs are very essential, then what we are supposed to do?

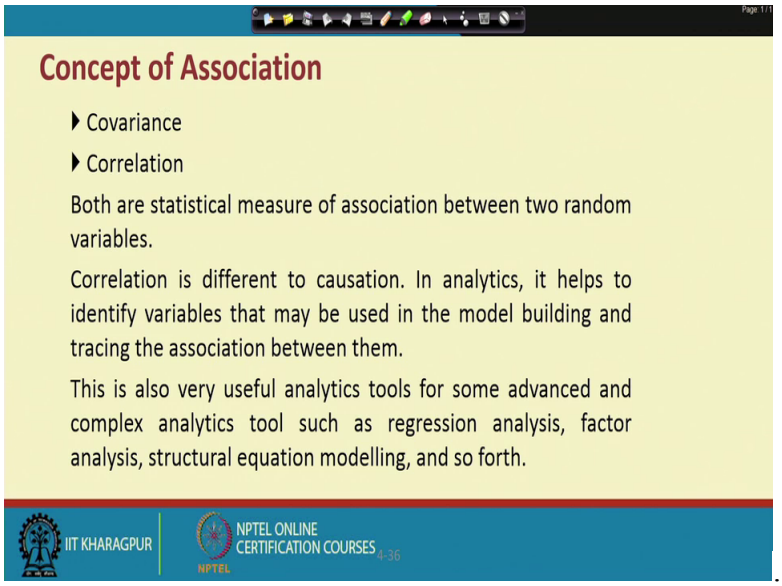
So, we are supposed to pick up certain engineering econometrics tools maybe you know manually can calculate or through some software's. So, the idea behind this software's or the manual process to transfer this inputs to some kind of you know engineering econometrics output. And, for that we have to apply some of the z engineering econometrics tools. And, it may be univariate structure, it may be bivariate structure, or it may be a multivariate structure, but in this particular lectures we specifically highlighting the basics only.

So; that means, for any kind of you know hardcore engineering econometrics problem. So, the minimum requirements or the basic need or basic check is the descriptive statistic and the kind of you know correlation matrix, which we will discuss in this particular lecture. Of course, the descriptive statistic the kind of you know a central tendency, dispersion and spreadness, all these things are discussed in the last lecture, but in this case we specifically highlight the association structure, because the association structure will take you to a path of you know hardcore engineering econometrics investigations.

So, with the help of data, so we like to know the details about the descriptive statistics, then the correlation statistics and after that we can proceed for the kind of you know actual econometrics investigations. So, when you go for actual econometrics investigation. So, we must have a clear cut you know details. So, starting with you know variables I think the data availability and then and the kind of you know the basic descriptives or univariate statistic corresponding to the is variables and that minus corresponding to this data.

So, once we have all these details then you will think about the particular process, how a beautiful models or how a accurate model can be used or can be applied to analyzed the particular you know engineering problems? So, the starting process of this particular lecture is the covariance issue and a correlation issue.

(Refer Slide Time: 07:05)



So, what is about this correlation and covariance? Let me a little bit you know brief about this. A correlation is not actually something called as you know causations; it simply represents the association between 2 variables. And so; that means, technically when you have a 2 variable or 2 variables.

So, we may have association or you may not have association. So, our job is to check whether there is a kind of you know association or no association. So, again so, when we go for you know harcore engineering econometrics investigation. So, one of the basic techniques we frequently use is called as you know regression modeling, which I have already highlighted you know very specifically in the first couple of lectures. So, that means, the entire engineering converting problems more or less you know connected to a regression modeling.

So, here what is our job you know, you have to just check the basics before you go to the regression modeling. For instance, one of the requirement of regression modeling is you know the classification of dependent variable and independent variable and that too

when there is a kind of you know independent variable clusters, there should not be any further relationship you know among them.

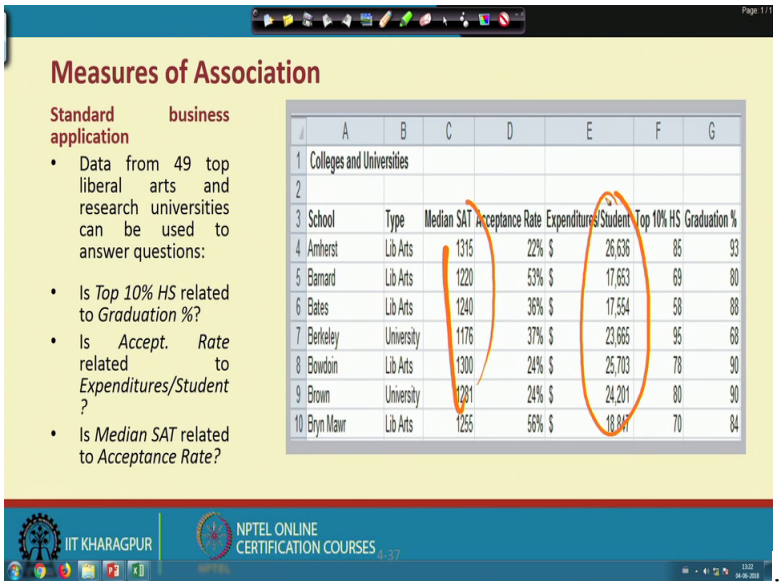
So, now is that is the requirement of you know regression modeling, but you know in reality the things are you know different. So, there may be a it is a that you know there is a association among these independent variables. So, that is how? So, descriptive econometrics will let you know, whether you are in the correct track or you know something you know right way to proceed for the kind of you know harcore investigation. And, the kind of you know solution for the particular you know engineering econometrics problem.

So, it is simply starts with you know it starts with you know covariance and correlation; covariance is simply you know you know the degree of association between 2 variables. And, that to it is not a univariate measurement technique, well correlation is a unitless measurement technique and it is a standardized component. So, what we like to check here? Whether there is a relationship between 2 variables, it is it may be independent 2 independent or independent to dependent.

So, technically for covariance and correlation the issue of dependent and independent you know is not there at all, but it will give you a kind of you know path for the regression modeling. So, that is how we need actually the entire correlation matrix with respect to all variables simultaneously.

So, it truly it you know by default it will give you a clear cut pictures, you know before you start the harcore you know engineering in econometrics investigations.

(Refer Slide Time: 09:58)



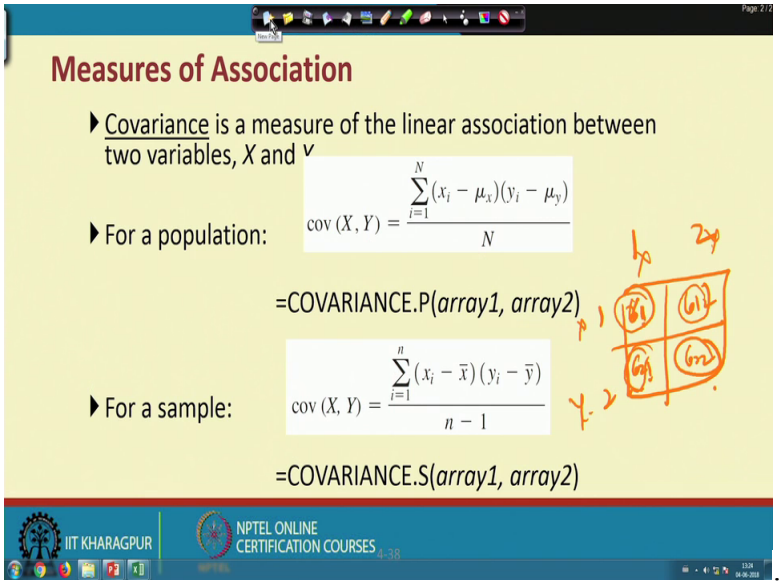
So, now the process is like this, you know let us say these are all you know different scores. And, we like to know whether you know this course are you know correlated to each other's. For instance so, putting money you know putting money in a particular you know investment process that is called as you know let us say expenditure. And, this particular example is related to students scoring and in the for instance this is actually students scoring, and then the kind of an expenditure you know invested towards you know particular students.

So, then you know usually the perception is that you know, if you put more money towards you know students for their you know preparation or something like that then; obviously, they will score well. So, also theoreticall;y so, there is a link between putting money towards the students and the student performance.

So, now to justify this particular you know theoretical evidence, so, we like to empirically test and one of such a technique which you can use is called as you know association technique. And that we can either use covariance or you can use correlations. So, of course, correlation is better technique than covariance, but ultimately the requirement is to check whether there is an association or not. So, co in fact, covariance is actually one of the component in correlations. So, just you know having covariance and divide by a variance of these 2 variables will give you the correlation coefficient.

So, ultimately so, there is you know big difference between correlation and co you know covariance. So, so what will they do here? So, we like to just check how is this particular you know case? And, then will they really analyze the requirement.

(Refer Slide Time: 11:45)



So, this is just typical formula through which you can actually calculate. So, it is a it is a kind of you know degree of association between 2 variables. In the last lecture we have discussed the concept called as you know standard deviations and variance. So, when you are you know a relating one variable with you know same means variable 2 variable same variable, then it is called as you know variance or if you know mathematical, if you know calculate the square root of that variant, then it is called as a standard deviation.

So, now when you are you know connecting variable upon same variables then it is called as a variance. So, then variable variant variable with you know different variable, then when you are you know con you know connecting then it is called as you know covariance. So, that means, technically. So, this is x minus x bar into y minus y bar. So, when you are actually just having x minus x bar into again x minus x bar then it is called as a variance. Otherwise it is simply called as you know covariance. So, for instance for a 2 variable let us say variable 1, variable 1, and variable 2.

So, then we can have a kind of you know matrix. So, 1 2 and again 1 2 so; that means, technically we have actually it took correlation matrix. So, this is what actually let us say a you know sigma 1 1 and this is sigma 2 2, then this is sigma 1 2 and sigma 2 1. So; that

means, this is what actually called as you know variance? If you know this is also variance; that means, square root of variance is standard deviation so, variable upon variables.

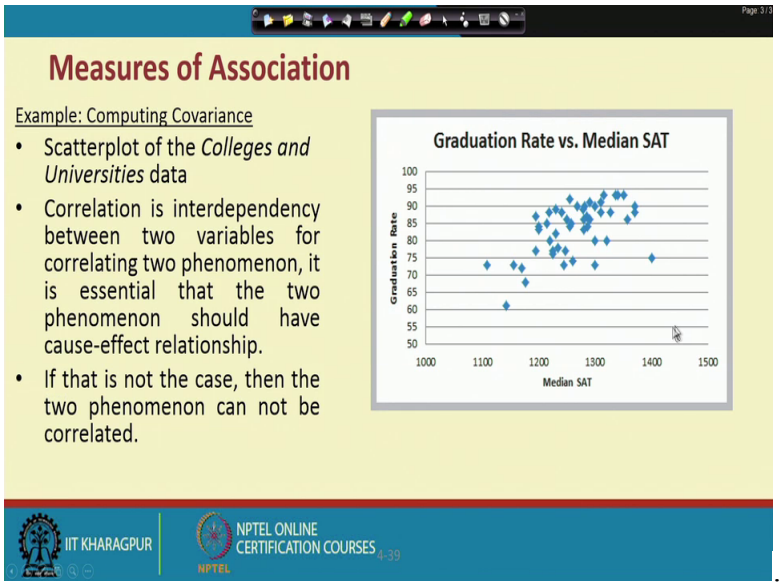
So, let us say this is x and this is x and this is y and this is y. So, x upon x this is stand you know standard deviations or you know called as variance. And, then y upon y again it is a standard deviation and that is the variance. And, this is

a x upon y and this is y upon x this is called as you know covariance. So, that is the only difference. So, it is kind of you know a variable upon same variable that is the indication is the variance, and variable with you know other variable that is called as you know covariance.

So, likewise you know you can calculate co correlation. So, co correlation is just you know you know covariance divided by a square root of you know variance of both the variables. And then it will give you some kind of you know weight vector which can describe the detail about the relationship between these 2 variables.

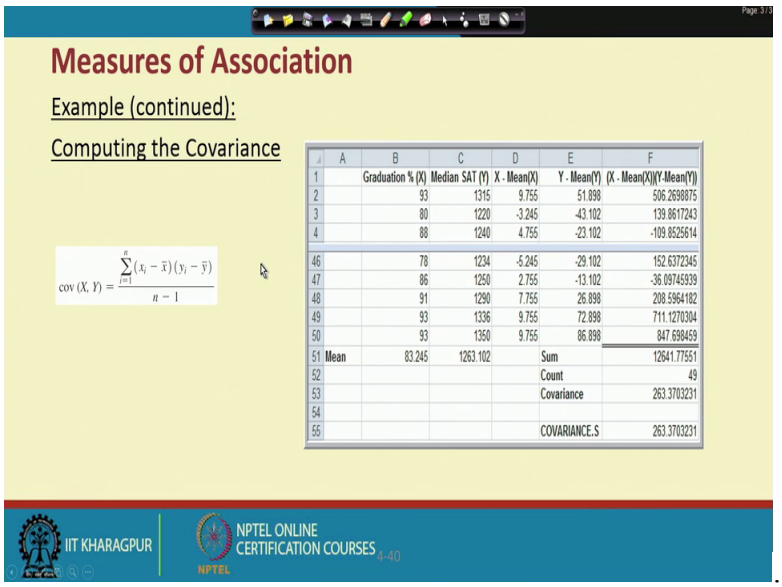
So, now corresponding to this you know yes there is a population structure and sample structure. So, in fact, we have actually separate lectures to highlight the issues about the sampling. So, that times I will clarify what is the big deal between you know sample and population. So, usually most of the empirical process is the sample specific, we just pick up the sample and then test accordingly. Then on the basis of you know inference and the kind of you know results from the particular samples we generalize the particular, you know engineering problems or we will go for general policy issues or something like that.

(Refer Slide Time: 15:06)



So, this is how the kind of process. So, this is a plotting and sometimes gives you some kind of you know idea about whether there is a kind of you know association or not.

(Refer Slide Time: 15:15)



But, ultimately so, we need to quantify. It and covariance correlation is the you know measurement techniques, which can give you a quantitative you know output that can you know, address, that will be you know better way to you know address the particular associations, you know otherwise graphically showing something else. So, you may not

actually properly, you know interpret so, because it will not give you some kind of you know unique kind of you know quantification.

But, the beauty of the correlation is you know it will be ranging from minus 1 to plus 1. And, if you transfer into percentage for then it will be actually 0 to 1 both positively and negatively correlation can be negative correlation can be positive. And, if correlation coefficient is coming positive, then you know the variables are positively related to each other. And, if the coefficient is in now becoming negative then the variables is negative, let it to each other.

So, the 3 specific you know kind of an extremes are minus 1 0 and 1. So, when the correlation coefficient is coming minus 1 so; that means, technically t represents that you know, there is a strong negative association between the variables. And, when you will get you know plus 1, then there is you know perfect positive relationship between the 2. And when you get exactly 0; that means, there is no association between the 2. And so; that means, technically in the first instance you have a 2 different classifications. So, no association and association, if no association the correlation coefficient will be 0, and if is the associate, if there is a association then either it will be minus 1 or 1 or in between minus 1 to 1 not exactly 0.

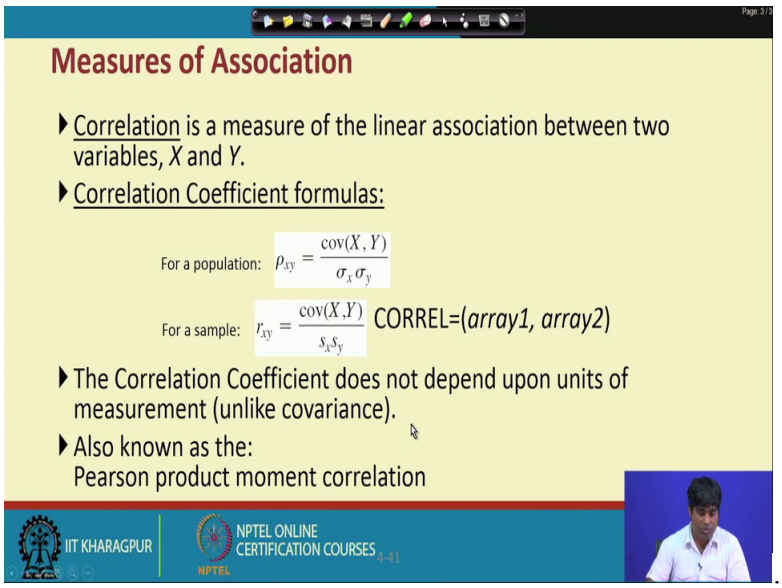
So, having minus 1, then perfectly negative representation having plus 1 perfectly positive association and if in between let us say minus 0.9 or plus 0.9 minus 0.8 something like that. So, that means, say it is it can indicates that you know highly correlated or you know more significant associations or more in a considerable association or if it is close to 0, then it is a less associate some something like that. So; that means, one way classification is an association no association and that to the quantification is a correlation coefficient r equal to 0 and correlation coefficient r not equal to 0, against when the ca means in the case of you know r not equal to 0.

So, we have minus 1 this is one extremes, another is actually you can put actually 0.5 that is the middle. And, anything less than 0.5 is called as in association. And, anything above then 0.5 is called as you know high association. And, close to ones means very high association and close to 0 means very low associations may be positive may be negative.

So, it is not a big deals, but you know our job is to know whether there is a kind of you know positive association or negative association or no association. Of course, you know research levels or empirical kind of you know testing. The negative inference or the kind of you know positive inference you have a different kind of you know policy implication. That is why having just plotting you may not actually a perfectly you know visualize or you can just perfectly can guess, that you know there is a positive associations or you know there is a strong negative association. That is why there is need of you know some kind of you know quantification.

Of course, visualization is a good trick, but sometimes it may not to give you a kind of you know clear cut path to address the problem very perfectly. That is why some kind of you know quantitative measures we need to use to you know represent the particular effect in a much better way.

(Refer Slide Time: 19:20)



So, this is how the particular structure? So, this is how they see you know formula through which you can actually calculate. This is the correlation coefficient formula and that to that to between the 2 variables. And, this is the kind of you know population case and this is a kind of you know sample case.

Of course, in the in the first instance you may not confuse the population correlation coefficient, sample correlation coefficient just you know just you know from the beginning, you can start with the structure that you know you have to the at least 2

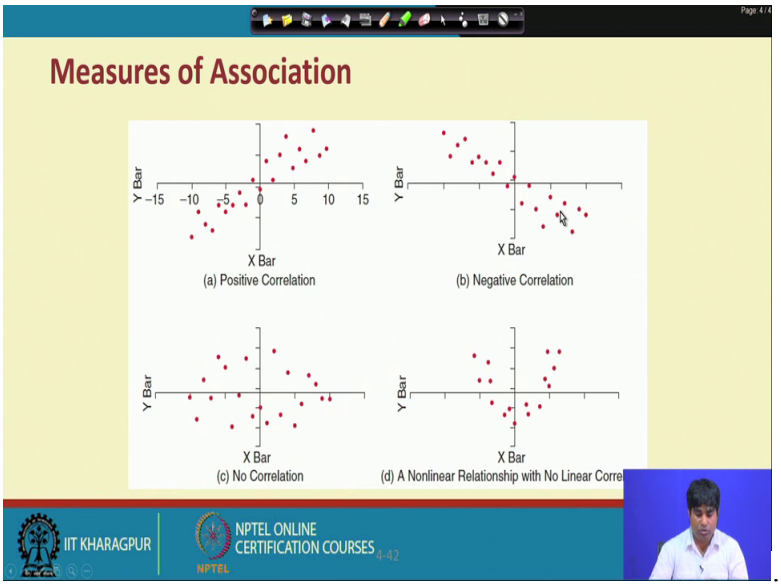
variables. And corresponding to each variable you have a sample points and all the sample points must be uniform, that is one of the basic requirement and if it is actually that is called as what you know balanced sampling.

If it is unbalanced sampling, then you need to prepare a balanced sampling by removing the other items in a particular variables or you know fill up the missing observations of a particular variable. Then finally, you must have a uniform observation corresponding to both the variables, then you are in a position to calculate the correlation coefficient. And, this is one of the basic requirement of you know regression modeling or you know engineering econometrics, until unless you know the relationship you should not start in regress. That means, technically before the start you know harcore engineering econometrics investigations by using regression modeling.

So, you must have theoretical evidence, whether there is a kind of you know association between the 2 variables, and there is also some kind of you know statistical evidence whether there is a kind of you know; that means, we start with the theoretical evidence, and some kind of you know statistical evidence and then will go for you know in depth; in depth kind of you know investigations by using regression modeling to know the details about the particular you know relationship. Who is the cause? Who is deeper? And, how is the particular then you know relationship and whether the particular relationship is affected or you know to be affected by other variables?

So, these are the issues we which we you know you know, we usually actually look for forward and that to the requirement of you know engineering econometrics. We typically, you know give tress on engineering econometrics solve all this you know complexity issues. So, that is how we need to know the basics first and then we will go for the kind of you know harcore you know engineering econometrics problem. And, then specifically so, this is how the particular you know structure?

(Refer Slide Time: 22:01)



So, now, this is how this snapshots. So, how they are you know related to each other? And, if you could see here, you know this is a positive correlation, because it is going in one direction, this is a negative correlation it is also going opposite directions. And, the process which we are you know discussing here, there may be 2 divisions. It is a kind of in only a linear associations and there is something called as you know non-linear association.

That means, if the kind of you know plotting between these 2 variables are going in a kind of you know one a similar kind of you know directions may be positively may be negatively, then this is called a general linear structure for instance the figure 1 and figure 2 is the classic examples. And figure 3 so, the plotting's or the common points are actually a very different. Not exactly for a particular you know path as a result simple declaration is there is you know association. And, that too we are just detecting through visualization, but still if you can use correlation coefficient and get to know.

So, in this particular case for instance in the third case. So, this case so, if you apply correlation coefficient and then; obviously, the value will be close to 0, but in this case the value will not be 0, but the relationship is coming non-linear. So; that means, you know correlation coefficient can be linear and can be you know non-linear. So, that means, the typical structure is the association is the linear 1 and the association is the

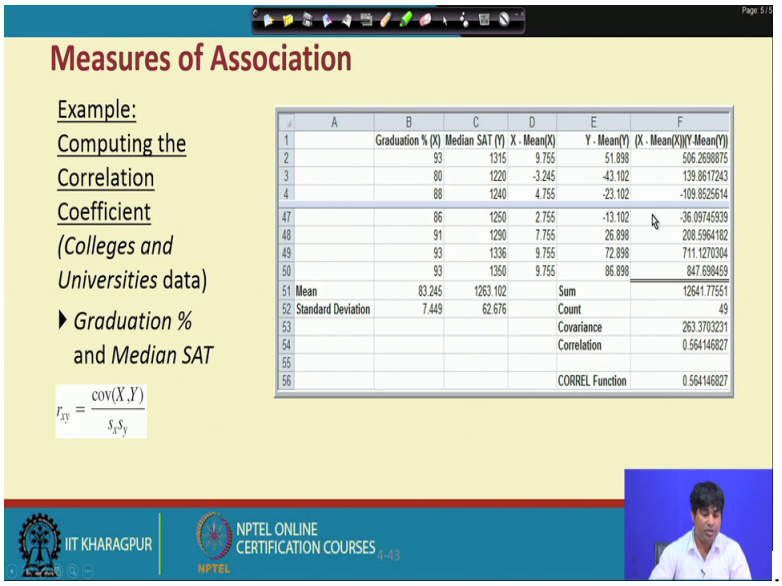
non-linear 1. In other words the relationship is linear 1 and the relationship is non-linear 1. So, then it depends time to time situation to situation problem to problem.

So, today the today the 2 variables are linearly related to each other simultaneously going, but because of certain other factors influence or something like you know prospectors and pull factors. So, they may be related, but not in a linear platforms. They will go for they really relate each other in a kind of you know non-linear platforms. So, our major deal is not whether linear or non-linear our major deal is whether there is an association or no association. And, whether there positively linked or negatively links. Of course, in the case of you know non-linearity.

So, some instances it may be positive association some instances, it is a negative association which is not there in that case of you know linear structures. In the linear structures the straight straightforward answer is a negative ones or positive one, but in the case of non-linear some extent it will be positive and some dimension it may be negative. So, that is what the complexity is all about? So, that means, real life problem will be like that only.

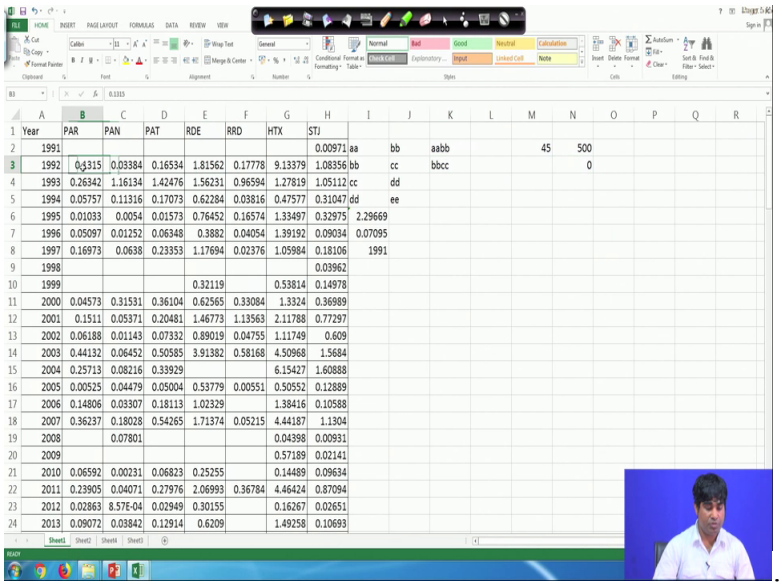
So, you will not get actually very perfect signal in the first instance. So, that is how we need actually in depth investigation to find out the actual fact and reality to address the engineering problems as for the particular you know requirement. And for that so, the engineering econometrics and that to particular regression modeling is very essential with your to which we are not in a position to find out the perfect kind of you know requirement for the engineering problem.

(Refer Slide Time: 25:26)



So,; obviously, so, this is how the kind of you know structure? And, then ultimately so, this is how the excel spreadsheet? And, in the excel spreadsheet is just you will the kind of you know structure and you will get the results very easily for instance.

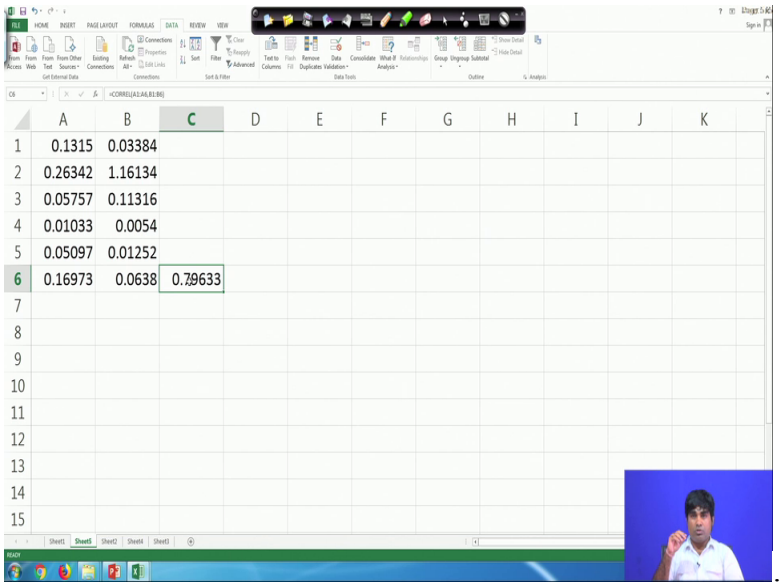
(Refer Slide Time: 25:39)



Let us say this is the excel spreadsheet and we will just take few variables for instance, let me take you know few data point here. This is a big data and I will take you to new excel sheet and then you paste it here ok.

So, this is this paste you will do the pasting here and just I am just elaborating and ok.

(Refer Slide Time: 25:59)



So, this is how the variables and you can say that this is the variable 1 and this is variable 2. And, we have 6 data points for x ones and we have 6 data points also for x 2. So; that means, we have a balanced sampling and for well, if the first hand requirement is a balanced sampling, then we can easily use covariance or correlation, but if the data points are not balanced.

So, then we really first at least the balanced you know balance data and; that means, either you know remove the extra points or you can fill up the missing observations. Whatever which you can do, but ultimately we need balanced sampling to do the particular analysis. Ultimately having these inputs; so, we need actually let us say correlation.

So, what I have already mentioned, we have a couple of software. And, in the last unit we specifically highlighted the beauty of the excel spreadsheet and within the excel spreadsheet; we have a data analysis package. And, then we have actually a excel start. And, excel start is a very harcore package, which can solve almost all problems of you know engineering econometrics, but in the mean times. So, we are just using simple excels and to report the kind of you know correlation.

So, that means, for that we just you know put anywhere else you know equal to signs and go to this particular you know functions and find out to report the correlation coefficient ok. So, what did we do?. So, so, otherwise you can go to the data and data analysis package. So, here's we defined you know correlations ok. So, if you put you know correlations ok. So, this will be showing that there will be a correlation coefficient ok. Just click here and you indicate the variable response here and close the loop and then you just put enter yes this is coming.

So, now so, you know so, this is how the indication clear cut indication? So, that means, what we have given the indication that you know, we are just putting the correlation a you know formula. So, just you are right you know put equal to signs and give the command about the correlation. So, it will ask you or to give you know x 1 range and x 2 range. So, obviously, when we are putting the x 1 range and x 2 range. So, it rally it really is you know check the balance sampling. And, then by default it will be processed through the particular you know formula and you will get the correlation coefficient.

So, this is a simple way to calculate, but you know in the advance software's like you know except start or s p s s e p o is micro pet cats rats saddams, whatever software you can use. So, you just you know give the indications that you know variable unvariable at tools. So, automatically it will be read the range and check you know balancing, whether it is a balanced data or unbalanced data. Then a report the correlation coefficient as per the particular you know requirement.

Then ultimately, this will give you some kind of you know you know some kind of you know clarity to know, whether there is a kind of you know association or non-noise association. For instance, in this case the better is coming 0.79 6; that means, it is a eighty percent association. So, what I have already mentioned? So, the correlation coefficient will be in between minus 1 to 1 and in the first instance the division is the association and no association, if it is like that then quantification would be r will be equal to 0 and r not equal to 0. If r not equal to 0 then the game is on. So, where we like to know more about the relationship whether it is a positive or negative a high or low, something like that and against there is issue of you know whatever correlation coefficient, you have negative positive whether it is statistically significant or not.

Till now we have not started this particular investigation process. We are just you know reporting the descriptive statistic or by various statistics to ignore the effects, about these variables corresponding to the reported data.

But, ultimately the particular statistic need to be tested again empirically a through some process, that is called as a inferential econometrics where we need to convert these statistic with you know testing procedures, that is what called as a hypothesis testing. So, hypothesis testing again is a kind of you know step by step process, where the idea about the problem is whatever parameters value or you know test statistic value, e which you have receives from the particular variable also the particular problems corresponding to the data. So, that need to be you know checked and the reporting must be requires, whether it is empirically in a valid or statistically valid that is what we use the term called as a statistical significance.

So, at the econometrics you know angles or the level of testing followed by hypothesis testing. So, you know we take care of the issues like you know one percent 5 percent and 10 percent. And, that means, you know the issue is like that you know we have to formulate the null hypothesis and the alternative hypothesis with the indenting that you know the particular test statistic is you know 0 or not 0. Then ultimately from the data we will get the test statistic value and it is connect with you not tabulated values. Then, we will check whether you know calculated value through the data is overtaking the tabulated value with a particular probability level of significance. Econometrically we have to we have to check or we have to you know compare with a 1 percent, 5 percent, and 10 percent. Anything we are rejecting at 1 percent it is always best, if not then it will come down to 5 percent, and then will come down to 10 percent.

And if they do the 10 percent therefore, if we are not in a position to reach it, then the declaration is that you know yes there is a correlation correlations, but the correlation coefficient is not you know strong enough or not statistically significant.

So, if something is not statistically significant then the empirical validations or empirical evidence is not so, supportive. So, that is why we try to try to analyze in such a way. So, we should have some kind of an output you know, such that we can empirically you know have evidence. And, in the first instance you may not have and that is how the game is all about?

So, it is a very interesting game. So, because the things are you know very continuous and you know structuring restructuring, it is the kind of you know iterative process. Sometimes, you can increase the sample size you know decrease the sample size and you know add variables remove variables, then means you will continuously do and find out to some hint and as per the particular you know requirement. You know the thing is that you know before we enter to this you know games you are theoretically convinced that you know there is a relationship and there is a you know logic behind it and there is a theory behind it.

So, now once the empirical results are not coming as per the log logical evidence or you know theoretical evidence. So, you must be very careful why not coming? So, that is how we have to do the iterative process, till you get the particularly an answer yes. You know even after that you are not getting the answer, then you know the theory a theory we need to recommend that you know there is something wrong about this theory and there is something wrong about this particular technology. So, you know that is how the challenge part or the kind of you know complexity part and where, the engineering econometrics rule is actually absolutely you know highly and as a means you can say high requirement.

So, ultimately in these particular lectures, we have gone through the particular component called as a covariance and correlations which can specifically highlight the issues about the relationship between 2 variables, which is one of the basic requirement of regression modeling and that to engineering econometrics and that to address a particular you know engineering problem.

So, whatever we have discussed right now this is called as a simple correlation structure. And, it has also complex you know issue and complexity issue, you know starts with you know more number of variables and the significance issues and the kind of you know regression requirement all these details will be there. So, we will discuss these details in the next lectures in the meantime, I will stop here.

Thank you very much have a nice day.