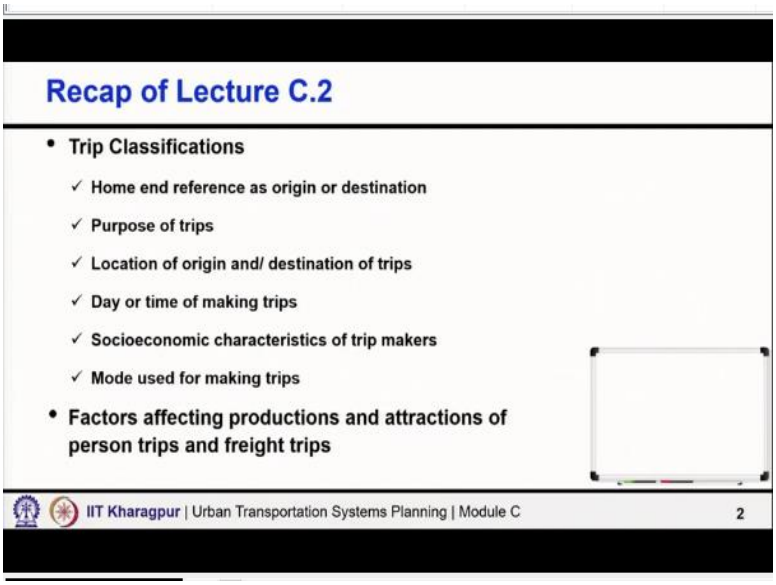
Urban Transportation Systems Planning
Prof. Bhargab Maitra
Department of Civil Engineering
Indian Institute of Technology-Kharagpur

Lecture-13
Modelling Approaches and Step-Wise Approach of Multiple Regression Analysis

Hello friends, welcome to module C, lecture 3. In this lecture, we shall discuss about various modeling approaches initially and then we shall talk about step-wise approach for multiple regression analysis, which are used for modeling trip generation.

(Refer Slide Time: 00:40)



What we discussed in lecture 2 a quick recap, we discussed about various classifications of trip, various ways of looking at the trips, say based on human reference as origin or destination. So, primarily classifying trips as home based and non home based trip. Then based on purpose of trips, work trip, business trip, shopping, recreational trips and so on so, forth. Then, based on the location of origin and destination of trips.

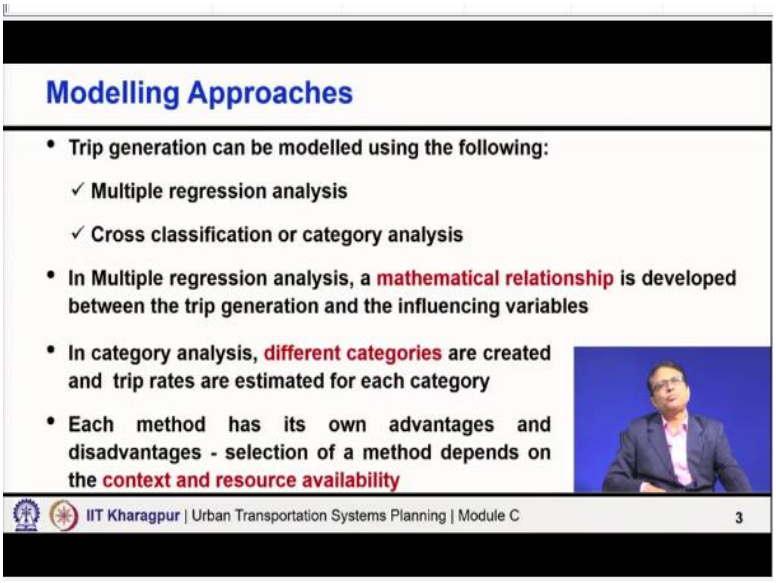
So, with reference to study area internal, external and then finally going about internal-internal, internal-external, external-internal and external-external and also discuss that why such classifications may be significant or maybe important. Then classifying trips based on when the trips are happening day or night time of making the trips, day or time of picking trips. Say peak hour trip or pick period trip, then OPIC period trips.

Then weekday trip, weekend trips like that, then also classifying trip based on socio economic characteristics of trip pickers, say for example, based on income of household or personnel income and so on so forth. Also, we discussed that how we can look at the trips based on mode which are used for trip picking, say car trips, trips made by using taxi, using public transport, using non motorized mode.

And the whole idea was to understand that how so many ways we can look at the trips? And then use a proper classification or most suitable classification for a given context of the work. What is my objective of doing this transportation planning studies? And then keeping that objective in mind, what could be the most appropriate classification of trips?

Then, we also discussed about various factors which are affecting productions and attractions of person trips as well as freight trips. So, production, attraction, person trip production attraction, also freight trip production attractions we discussed.

(Refer Slide Time: 03:24)



Now, with this background, we shall now discuss about the modeling approaches for trip generation. So, we are all set with that our basic information and basic understanding and now we shall enter into the modeling approaches. There are 2 broad approaches which are used for trip generation modeling, one is called multiple regression analysis. The other is called cross classification or category analysis.

What we do in multiple regression analysis are this approach, it is mathematical relationship that is what we try to build or develop between the trip generation and the influencing

variables. So, you remember in the lecture 2, we talked about various variables or factors which may influence person trip productions. Say, if we are making a trip production model using multiple regression analysis.

Then we are trying to develop a mathematical relations where y may be the number of trips that are produced from a zone or by household and expressing this y as a function of x where x may include independent variables such as maybe income of household, car ownership, household size and maybe other relevant variables. So, the basic idea here is to develop a mathematical relationship between trip generation and the influencing variables.

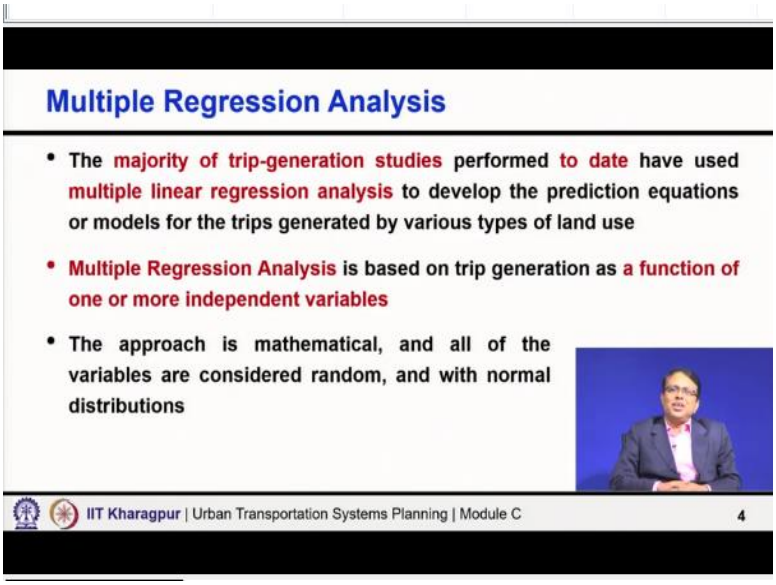
In cross classification or category analysis, what we do we create different categories of households, if it is production, if not accordingly based on the land use or other considerations, we create different categories, but altogether we create different categories and then we estimate the trip rates per unit how many trips are being made. So, we calculate the trip rates for each category.

These are the 2 different methods. Then, different categories may be based on income. So, different income household how that triplets are changing that we try to plot and maybe see the trend and then accordingly for different households from that trip rate, we use pick up the values of appropriate tip rate and use it for a given category for forecast or for predicting the number of trips in the future.

These are the basic 2 approaches. Both approaches have their advantages and disadvantages which of course, we shall discuss towards the end of this module not now, because once you learn both approaches then you will be in a better position to understand their merits and demerits. But, as it happens, every in every context in this world, whenever there are alternative methods, alternative techniques, each method or each technique has its own advantages and disadvantages.

So, here also multiple regression analysis and cross classification or category analysis both have their own advantages and disadvantages. So, what method to select that depends on the context and also the kind of resource which are available for the work. Again, you will appreciate it better when we discuss more.

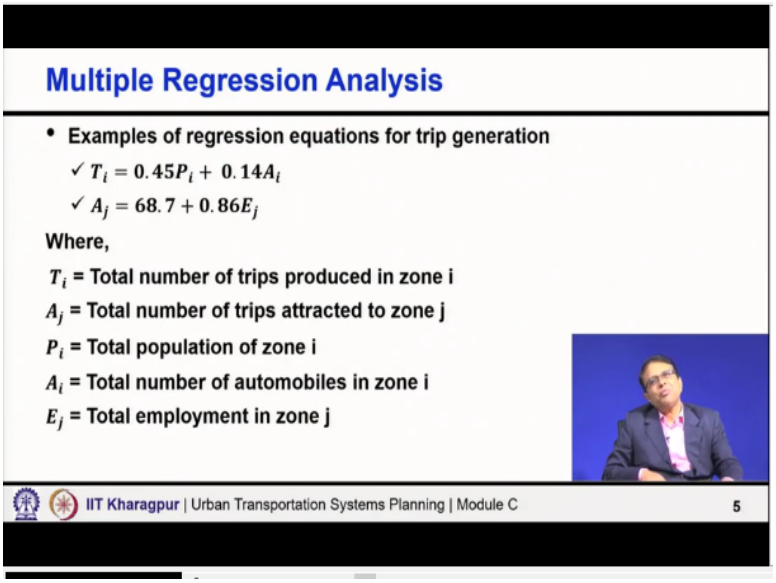
(Refer Slide Time: 07:34)



Now, first coming to the multiple regression analysis. Now, majority of the trip generation studies they performed actually they used multiple regression analysis to develop prediction equations or models equations are also models for trips generated by various types of land use. So, large number of studies actually used till date the regression based model. Multiple regression analysis is based on trip generation as a function of one or more independent variables.

So, we know how many number of trips are produced per household is a function of the factors which influence the household trip production and that is expressed using a mathematical relationship. In this case, it is a multiple regression model. The approaches mathematical and all the variables are considered as random variables and assumed to also follow normal distributions. That is some basic assumptions which are being made.

(Refer Slide Time: 09:01)



Let us give 1 or 2 example. Let us take this example T i = 0.45 P i + 0.14 Ai. What is T i? T i is total number of trips produced in zone i. So, when we are telling that how much produced and getting attracted to and from zones then it is basically zonal based model. That means traffic analysis zone is my unit for analysis. So, it is a zonal model we are trying to predict the number of trips produced from zone i as a function of what population yes number of trips logically depend on the population production.

And also the number of automobiles in that zone. Yes, automobile also is an indicator of income one can use, one can use automobile ownership or number of automobiles that are available in the zones, since it is a zonal model, number of automobile ownership available is really used as a logical variable. So, you can see such kind of mathematical equation we can develop to express number of trips produced the zone as a function of the population in that zone and as a function of total number of automobiles in that zone.

The variables are indicative, you may use many other variables, some cases some variables may be influencing the results, some cases some variables may not really indicate adequate influence based on the statistical test. So, what variable to take and what variable to finally retain these are all the things we will discuss, but similar kind of equation we would like to develop.

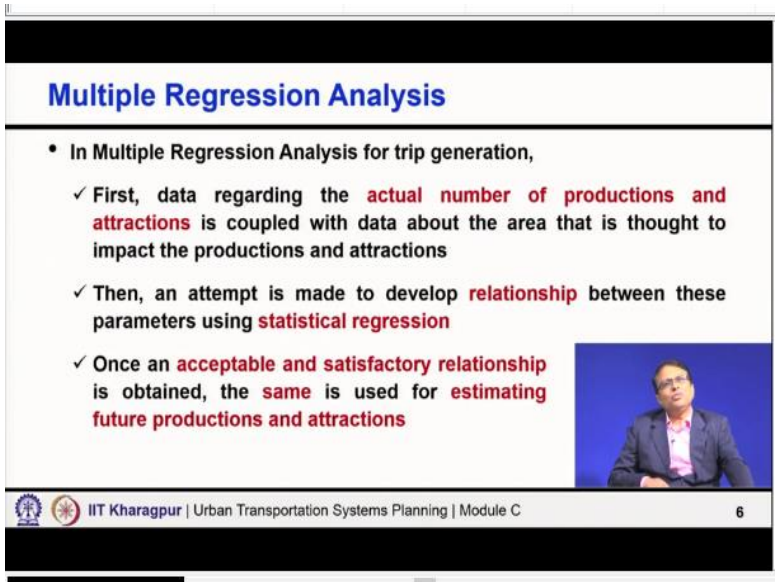
Similarly, attraction is expressed as a function of the total employment in a zone, if there are more employment in a zone more trips are expected to get attracted to that zone. So, again very logical and what is the 68.7 you can say? That is the constant. So, often a regression

model comes with a constant, we shall again discuss further about even about this aspect also the value of the constant.

The constant is what in a way it is the unexplained component. But, so, you have to look at this value, this value should not be too large. That means, the unexplained component yes, it is quite logical and quite practical that once you develop model through this use of this limited number of variables, one may not be able to explain the variation 100%. So, there may be some unexplained component**.**

That is okay, that is not anything seriously wrong also, but the value should not be too high. So, my unexplained components should not dominate my model, that one has to ensure. So, this is the kind of simple models we want to develop.

(Refer Slide Time: 12:34)



Now, how we go then, once you have decided that, we will probably go for multiple regression analysis, what we do? First data regarding the actual number of productions and attractions, these are the 2 things we want to model, we want to model the number of trips which are likely to get produced and number of trips which are likely to get attracted, these 2 things we want to model.

So, we need data about these 2 number of productions and attractions and what we do, then these data are coupled with the data about the area, that area what we think could impact this productions and attractions. So, number of productions we try to relate it with the data what

we think are likely to influence production, we try to build the relationship. Similarly, the trips attracted also we know we quantify.

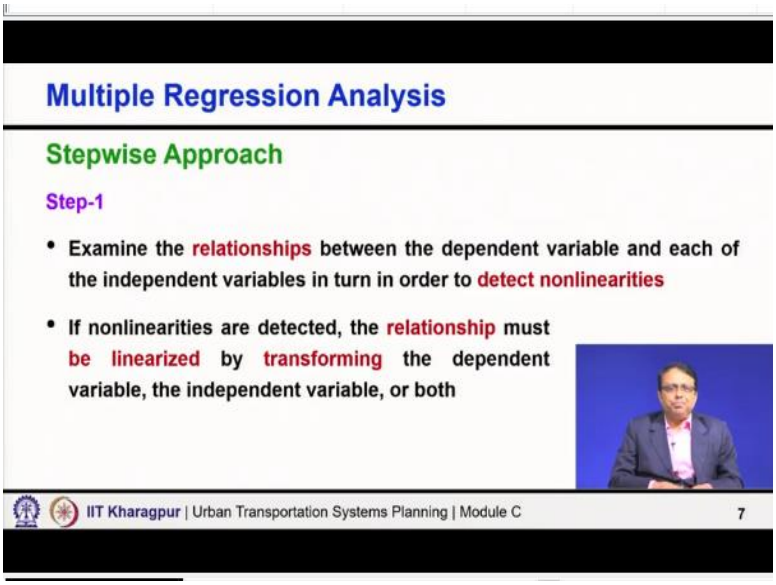
And then try to relate it with the variables or the data that we think would probably impact or influence these trip attractions. So, first we get the data about y about various X, X are independent variable, Y are dependent variables, dependent variables are productions and attractions, independent variables depending on whatever we discussed earlier several factors we say which might influence productions and attractions.

Those factors would be logically taken in terms of independent variables, then what we are trying to do? Then we are trying to make or develop the relationship between these parameters factors and Y, independent and dependent variables, we are trying to build the relationship. How we can say estimator Y given the values of X? That is we are trying to do using statistical regression.

That once we are successful to develop an acceptable and satisfactory relationship we are using the same for the future because whole modeling any model we developing the model for the future, base year we have the data for x, base year we have the data for Y. So, base year we develop the model, develop the relationship, but developing it for the future. Because future we want to know Y given X, that is the objective.

So, once we have developed a satisfactory relationship acceptable and satisfactory relationship then we use this one for estimating future productions and future attractions So, that is what it is.

(Refer Slide Time: 15:58)



Then, let us go to stepwise approach, how step 1, step 2, step 3, how we really build this equation or develop this equation? What we do first is the step 1. Examine the relationship between the dependent variable and each of the independent variable in turn, in order to detect nonlinearities. I shall explain everything very clearly and try to understand. You have identified a few logical variables which are likely to influence or explain the variation in that production among different zones.

Some zones are producing more, some zones are producing less, some zones are producing moderate and you know how different genes are producing different number because some of the x values are different in different zones. So, that we are trying to explain this variation of Y with the help of X, we have identified those and we have collected the data for all Y and all X what we want.

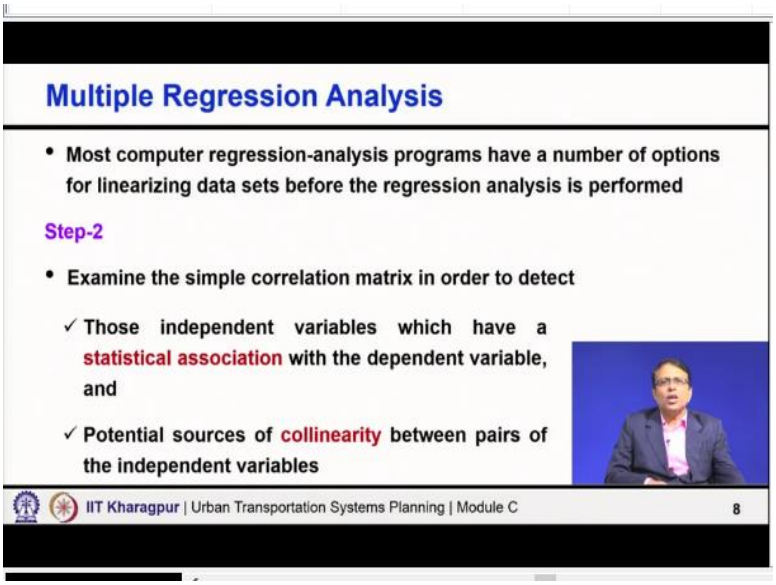
Then the first step I am saying plot Y versus X, 1 X at a time just look at the scatter. So, you say number of trips and maybe number of vehicles or number of population. So, you have say 100 zones, you have. So, you have 100 values of trip production, 100 values of population in each zones, put population versus productions. Just look at the scatter. Do you see apparently that the relationship is nonlinear?

Just from visual observation itself, you can see the pattern. Of course, there will be scattered, there will be variation, but you can still see a trend. Does it apparently show any distinct trend that it is not linear? It appears to be nonlinear. If you find that it is nonlinear, if you do

not detect fine, take it linear. If you find that no it is likely to be nonlinear, then you try to transform the assumed x or assumed y or both.

In such a manner that the relationship becomes linear. That is the first thing because what we are doing? We are doing linear regression. So, the relationship is supposed to be linear. So, if it is nonlinear and if we can detect that, then we can transform either X or Y or both suitably to make the relationship linear. There are many functions available.

(Refer Slide Time: 19:47)



Most computer regression analysis program have number of options for linearizing this data sets before regression analysis is performed. Say for example you can take log just to transfer that. So, I may take instead of X I will take log x as my x now. So, I will not consider x but if I find once I take log X then it transforms or some other thing, it makes linearize then that variable that transform variable now, I will consider an X.

And then that variable will be used for building a developing regression model. So, we are making the relationship between Y and X linear, that is the first step. Second examine the simple correlation matrix. Who will give me the correlation matrix? It is just a matter of giving a command in excel, you select the data set and you can give a command, it can give you the correlation matrix.

So, what is the correlation? Y 2 X 1 Y 2 X 2 Y 2 X 3 Y 2 X 4, it will give X 1 to X 2 X 1 to x 3 x 1 to X 4 X 2 to X 1 X 2 to X 3 X 2 to X 4. So, in between independent variable and dependent variable. So, each X versus this Y what is the correlation? What is the statistical

association? And then in between also independent variables in between individually X 1 X 2 X 2 X 3 X 1 X 1 X 4 X 2 X 4 X 3 X 4 what is the correlation?

So, this correlation matrix you can easily get computer can give you, but computer cannot give you a decision. So, how to use that to make a decision, that is what we are going to discuss now. Computer will give you, I am not going to discuss how you get the correlation matrix because that in different software's computer program can give you even as I said the excel also you have the database, it can give you the correlation matrix, that is not a big issue.

But computer will not be able to give you a decision. So, you have to make a decision as a planner or as a modeller when you are building a model. So, what do you do with that correlation matrix? Two things we check. First those independent variables whichever statistical association with the dependent variable, I want those variables, I just thought that X 1 X 2 X 3.

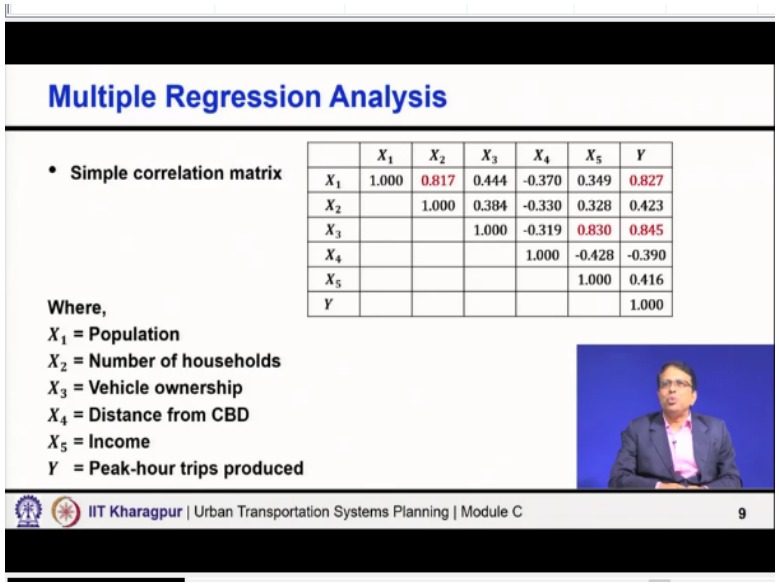
All these variables would probably help me to explain the variation of y. I just thought but the given data, the correlation matrix will indicate are they really going to if I consider all these variables, each of these variables are they going to really help me to estimate my Y? So, we want a statistical association between X and Y. If there is a statistical association between X and Y, then I can express y as a function of X.

So, maybe out of 5 variables you can take you may find only X 1 X 3 and X 5 are having statistical association with Y, but X 2 and X 4 no very weak. So, we need to identify those variables which have statistical association with the dependent variable. Number 2, we want to identify also potentials sources of collinearity between pairs of independent variables.

What is called collinearity between independent variables? That means, 2 variables X 1 and X 2 would be set collinear when their correlation matrix will indicate a high value. That means, there is again a statistical association between X 1 and X 2. So, if X 1 and X 2 two independent variables have high statistical association then they are not independent variable, I can express X 1 as a function of X 2 or X 2 as a function of X 1.

So, then they are actually collinear variable. So, I want variables which have high statistical association, but I do not want variables independent variables which have collinearity .So, both can be detected with the simple correlation matrix.

(Refer Slide Time: 25:41)



	X_1	X_2	X_3	X_4	X_5	Y
X_1	1.000	0.817	0.444	-0.370	0.349	0.827
X_2		1.000	0.384	-0.330	0.328	0.423
X_3			1.000	-0.319	0.830	0.845
X_4				1.000	-0.428	-0.390
X_5					1.000	0.416
Y						1.000

So, here let us take an example, I have taken a simple correlation matrix here the values are shown here, if we take 0.7 as a threshold for example, nobody has said that you have to take only 0.7 and higher or 0.7 as a threshold value. In this example I have taken. So, then what do you find here X 1 has got a very high statistical association with Y, X 3 has a high statistical association with Y, because both cases values are more than 0.8.

But not X 1 X 2, X 4 and X 5, they do not have that strong association with y. So, they independently these X 2 X 4 and X 5 are actually weak, but X 1 and X 3 are quite strong. So, I can obviously make a relationship y is a function of X 1, y is also a function of x 3, both these models are worth investing further likely to work, likely to explain the variation of Y, because of this strong association. So, that is the first one.

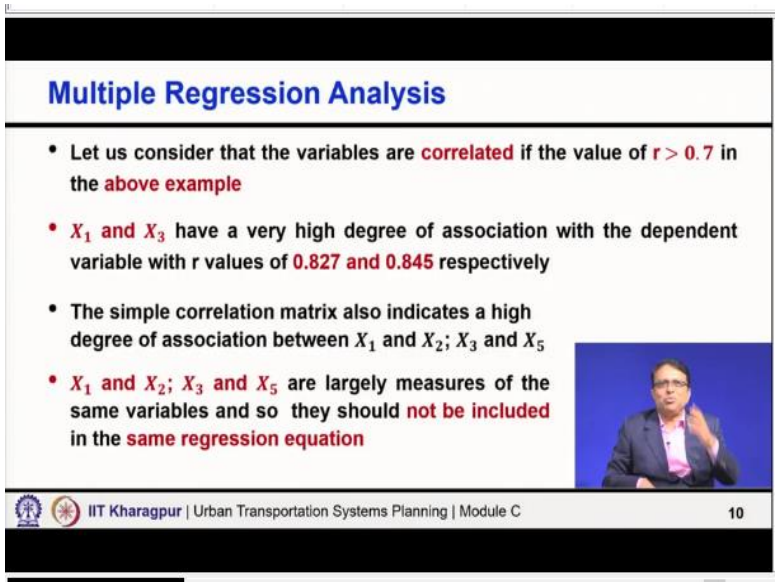
That is the first one what I said earlier, first thing those independent variables which have a statistical association with the dependent variables, that we have checked. The second is potential sources of collinearity between the pairs of independent variables, look at this thing, here also some values are red which are more than 0.7. For example, only and I have highlighted these values as red.

So, that you can easily identify, that is the purpose. So, you can see here X 1 and X 2, the statistical association again is very high 0.8. So, X 1 and X 2 are collinear variable that means they are not really independent variable, if there is so, high statistical association then you can express X 1 as a function of X 2 or you can express X 2 as a function of X 1. So, then in one model, I cannot use X 1 and X 2 both because they are not independent.

My fundamental assumption of linear correlations are the independent variables, they are independent variables here X 1 and X 2 are not independent, they are collinear. Similarly, you can see X 3 and X 5 they are again collinear. So, I can express X 3 as a function of X 5. I can express X 5 as a function of X 3. So, again these 2 are not independent variable. So, I cannot use both of them in the same equation or same model.

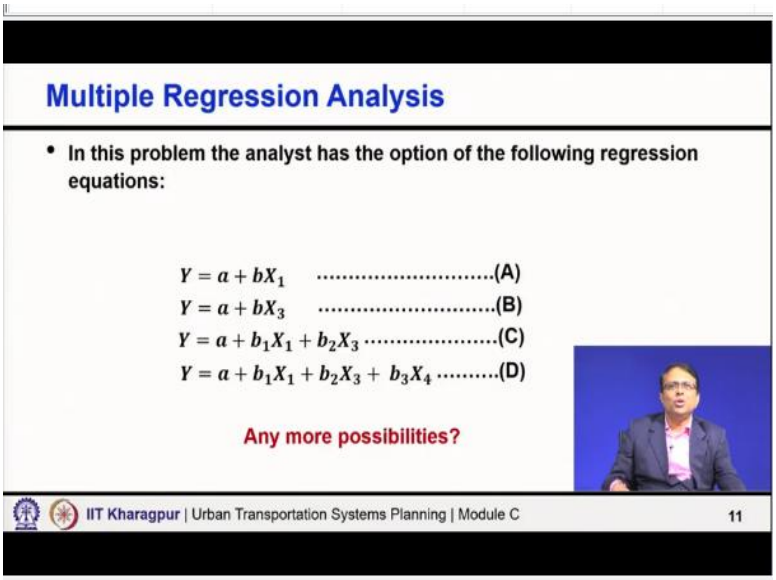
So, what I find here? So, I know we have independent variable has strong association with the dependent variable. So, I would like to include them when I am trying to explain the variation of y in my different trial models and the collinear variables should not be included together in the same model. So, with those, that is what is explained here.

(Refer Slide Time: 29:44)



I have said let us consider 0.7 and above are value as correlated, so you can see X 1 and X 3 have got very high degree of association with the dependent variable, which I have shown here, and then we said here that X 1 X 2 And X 3 X 5 cannot be used together because of their collinear X 3 and X 5 are collinear, X 1 and X 2 are also collinear.

(Refer Slide Time: 30:22)



So, with this then I select a few equations. I select this is what, this is also you remember I discussed about model specification, model calibration, validation and forecasting, this is model specification remember that, we are specifying the model, we are decided that maybe zonal based we will do or household well, and these are going to be my specification of the model.

So, in this example, I have taken y = a + b x 1 as one model. Again go back to this thing we know x 1 and y the statistical association is very strong the value is 0.827. So, it is worth trying to explain the variation of y with the help of x 1, that is what we have done, Y = a + b X 1, we also have similar model Y = a + b X 3 why because X 3 and Y are also having strong association you can say 0.845 in this hypothetical example.

So, these 2 we have done, but then we have not tried similar model Y as a function of X 2 alone or X 4 alone or X 5 alone it does not make sense because the correlation coefficients are lower. So, the association is not that strong individually. So, we do not want to try them alone. But does it mean that they cannot be considered? Yes, they can be considered, I can always include it we tried Y as a function of X 1 only.

I can also try another model to bring along with X 1 some other variables to see that if I include it may be individually they are not strong, but if I included along with other strong variables in the same model, can they make the overall model even stronger, but then carefully observe here we take one I will not take X 2, because X 1 and X 2 are collinear variable.

So, we take x 1 not X 2, but we take X 1 I can take X 3 yes X 1 and X 3 their correlation is 0.44. So, not so, high. X 4 also I can take, X 5 also I can take along with X 1. Similarly, for along with X 2 I can take X 3 along with X 3 I can take X 2 because X 3 is the more stronger variable. So, anyhow have a model only with X 3. So, I can also include X 2 along with that no issue.

So, this is our guidance which one makes sense and which are the alternatives worth investing. So, you can see here similarly, I have selected Y = a + b X 1, Y = a + b X 3 the two very strong variables which have got very strong correlation with Y independently. Then with X 1 I have added X 3, X 1 X 3 are not correlated. So, I could add both are strong. So, I start even the combination also very strong.

And they could be added because X 1 and X 3 are not so, highly correlated and along with that I also added X 4. Now X 4 X 1 not collinear, X 4 X 3 again not collinear and X 1 X 3 also not collinear. So, they all can be there in the same model. Now, any more possibilities? Yes there could be. Similarly you can try like only X 3 and X 4, you can try and maybe even bringing X 5 also.

One can bring X 5 because X 5 only cannot go with X 3, but with X 1 X 5 can also go. I have no problem even putting X 1 X 2 X 5, but you have to look at the correlations values and things judiciously that what are the combinations I should really take. Here is the guidance, look at X versus Y and think which are the stronger one and then can we take the combination of stronger variables.

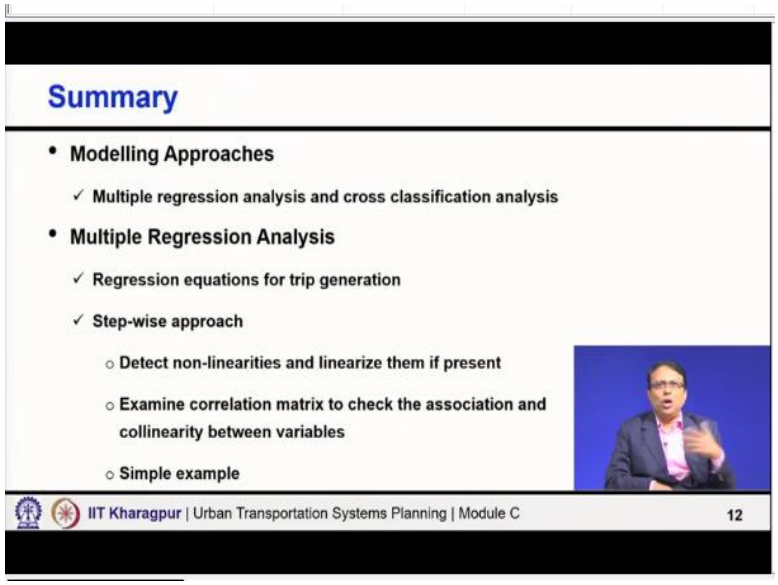
Suppose, here luckily we found X 1 X 3 individually strong association with Y but not collinear. So, I could add them, but if they would have been collinear then X 1 X 3 I would not be able to add together. So, I have taken these 4 models for further investigation too. So, but yes you could theoretically take more combinations, here there are opportunities even for more combination.

Because I take X 1 X 3 independently then X 1 X 3, then X 1 X 3 X 4, X 5 not with X 3, X 5 could have been taken then with X 1, along with X 1 I could have taken X 5. X 1 X 4 and X 5

that could have been also, then X 2 X 3 and X 4 without taking X 1. But logically you see I would still prefer to take because when X 1 has got such a strong association with Y.

X 3 has got a strong association of Y and X 1 X 3 and not collinear then I must take a combination of the stronger variables and then keep on adding if I want to add more variables, that was the reason why I have not selected more. Theoretically you can select but logically if you say not just theoretically any possible combination I will start taking, here you have to apply your logic. So that is what we selected.

(Refer Slide Time: 37:27)



So, what we discussed here, we discussed briefly the 2 modeling approaches multiple regression analysis and cross classification analysis. And then in multiple regressions, we introduced what we trying to do and the stepwise approach we started discussing or all steps are not over till now. But we said first try to detect nonlinearities, if you find nonlinearities, then try to linearize them using proper transformation of variable.

Then examine the correlation matrix. We want stronger association between each independent variable and dependent variable that they are stronger association we want. But between 2 independent variables, we do not want stronger association because that will mean they are collinear, they are not independent. And if they are collinear, then they are not even independent and cannot be used together in the same model.

So and then with that example, tried to show how logically based on these considerations, you can actually select logical and right combinations of model specifications for investigating further, we shall continue in the next class. Thank you so much.