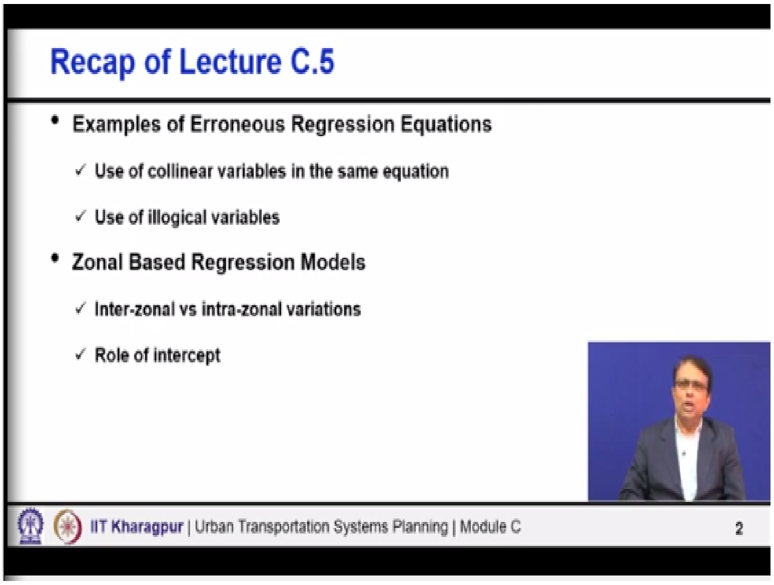
Urban Transportation Systems Planning
Prof. Bhargab Maitra
Department of Civil Engineering
Indian Institute of Technology-Kharagpur

Lecture-16
Zonal and Household Based Regression Models

Welcome to module C, lecture 6, module C was on trip generation.

(Refer Slide Time: 00:22)



Recap of Lecture C.5

- Examples of Erroneous Regression Equations
  - ✓ Use of collinear variables in the same equation
  - ✓ Use of illogical variables
- Zonal Based Regression Models
  - ✓ Inter-zonal vs intra-zonal variations
  - ✓ Role of intercept

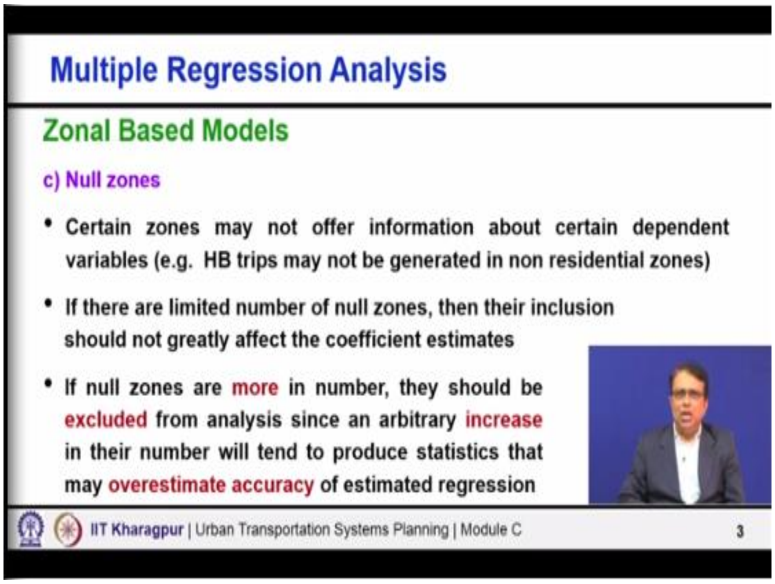
IIT Kharagpur | Urban Transportation Systems Planning | Module C
2

In previous lectures particularly in lecture 5, we gave you several examples of erroneous regression equations to tell you what kind of problems may occur or what kind of mistakes people normally do while developing regression models. And then we started discussing some of the key characteristics of zonal based regression models. And then the two specific aspects we discussed one is inter-zonal versus intra-zonal variation.

Regression models, we actually prefer or actually we want more of inter-zonal variations. But as I said sometimes problem may happen, because one can find from the data that inter-zonal variation is relatively lesser and it is a significant intra-zonal, variations. Then the second thing what we discussed was about the role of intercept, the constant and we said that constant represents basically unexplained component.

And therefore, as far as possible, we should try to make these constants minimum and impossible even to make it 0. If that does not distort the overall model, if there is a significant distortion in the model because we are forcing the intercept to be 0, then we return those constant. In continuation to these various characteristics of zonal base regression model. So, with these two issues, which were discussed in lecture 5, we will continue and discuss a few other aspects related to zonal based regression model.

(Refer Slide Time: 02:34)



Multiple Regression Analysis

Zonal Based Models

c) Null zones

- Certain zones may not offer information about certain dependent variables (e.g. HB trips may not be generated in non residential zones)
- If there are limited number of null zones, then their inclusion should not greatly affect the coefficient estimates
- If null zones are more in number, they should be excluded from analysis since an arbitrary increase in their number will tend to produce statistics that may overestimate accuracy of estimated regression

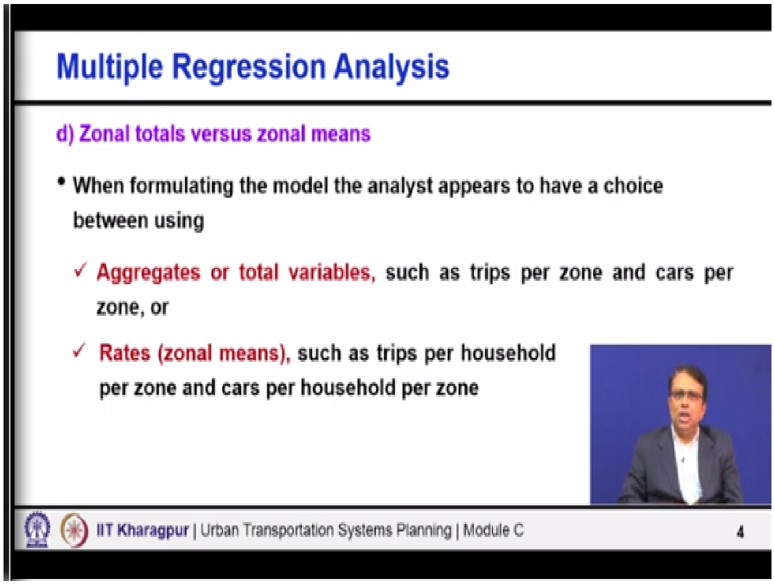
IIT Kharagpur | Urban Transportation Systems Planning | Module C 3

So, the next point is about the null zones, now what is null zones? When you develop models, certain zones may not offer information about certain dependent variable. Say for example, house or home, based trips may not be generated in non residential zones. So, if you have purely nonresidential zones, then the trip productions maybe 0. Now question is, so this kind of zones where you know, which do not offer information about certain dependent variables maybe called as null zones.

Now if there are a limited number of null zones, say for example, a few sale values in the dependent variables are 0, then it may be fine because their inclusion may not significantly affect the coefficient estimates, so you may not really bother much. But if there are more such null zones in the overall database then there will be many zeros in the dependent variable sales.

So, what will happen that will actually may overestimate the accuracy of the estimated regression. So, that in that case one has to really think and probably one has to think that to exclude the null zones from the database, if the numbers are significant. If there are a small in number then there is no issue.

(Refer Slide Time: 04:24)



Multiple Regression Analysis

d) Zonal totals versus zonal means

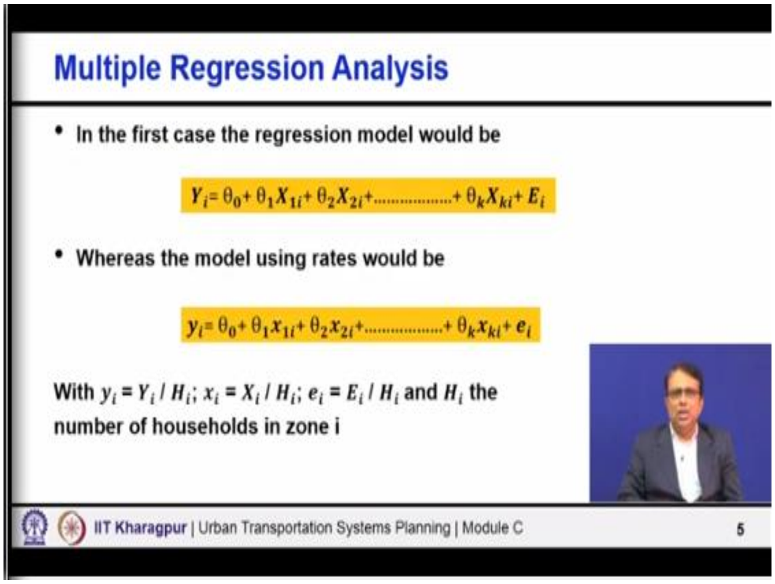
- When formulating the model the analyst appears to have a choice between using
  - ✓ Aggregates or total variables, such as trips per zone and cars per zone, or
  - ✓ Rates (zonal means), such as trips per household per zone and cars per household per zone

IIT Kharagpur | Urban Transportation Systems Planning | Module C
4

Also, the other part next point is a very, very interesting feature of zonal based multiple regression is use of zonal totals versus zonal means as variables. I can use say for example, total variable, say maybe total car in the zone or I can use car per whatever it is per household. Then it becomes of course a household model or you can rate per household also we can use.

So, then maybe vehicles, number of vehicles per household. So, one is basically total vehicles or total cars that zone wise that maybe a variable. And in second case in zone what is the rate, that means what is the general means? How many vehicles per household? Or similarly one can say that total trips produced per zone, another case trips per household in that zone. So, the first case we are saying we are using zonal total or aggregate or total variable. And in the second case we are using rate or zonal means, now what are the implications?

(Refer Slide Time: 06:09)



Multiple Regression Analysis

- In the first case the regression model would be

$$Y_i = \theta_0 + \theta_1 X_{1i} + \theta_2 X_{2i} + \dots + \theta_k X_{ki} + E_i$$

- Whereas the model using rates would be

$$y_i = \theta_0 + \theta_1 x_{1i} + \theta_2 x_{2i} + \dots + \theta_k x_{ki} + e_i$$

With $y_i = Y_i / H_i$; $x_i = X_i / H_i$; $e_i = E_i / H_i$ and H_i the number of households in zone i

IIT Kharagpur | Urban Transportation Systems Planning | Module C 5

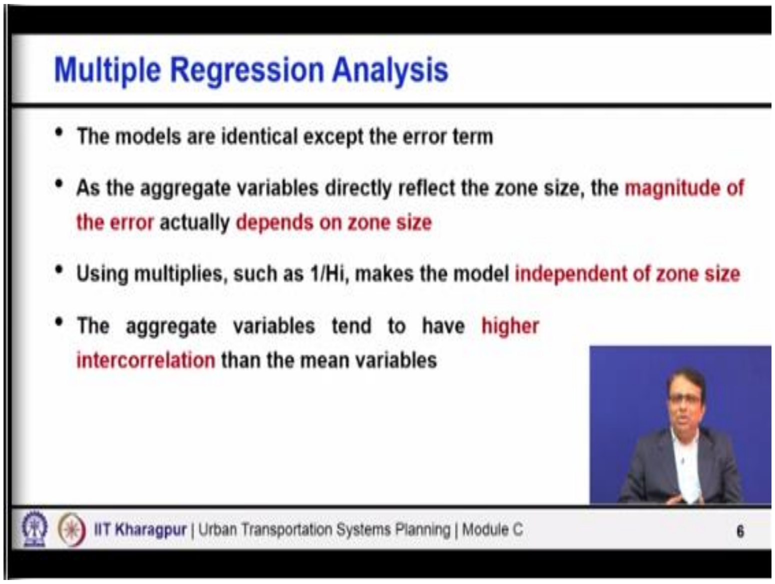
Before we talk about the implications, let us see that mathematically then what does it means? If we are using zonal total maybe we are representing the dependent variable as capital Y. And if we are using zonal mean then let us say we are using the dependent variable as small y. So, either capital Y or small y for each zone, so it is Y i. Then you know that theta 0, theta 1, theta 2 these are the coefficients.

And let us say zonal total the variables are capital X i 1, capital X i 2, capital X i 3 like that. And when we are using zonal means, they are small x i 1, small x i 2 and so on, I have written it here. Now obviously if you say small y i = capital Y i divided by number of households H i in that zone, zone i, similarly small x i = capital X i divided by H i number of households. And I have written also one error term because obviously we are estimating, so there will be an error.

So, that error term first case aggregate case model, we are saying capital E i and in the second case we are saying it as small e i. Now look at these two equations very carefully and tell me are they identical, they may not be identical, what is then the fundamental difference? Fundamental difference is actually in the error term distribution. The distribution of small e i and capital E i is not same, why they are not same? Can you think?

The answer is H i is also variable, suppose H i if number of households in each zone say hypothetically is equal, then what is H i? H i is no more variable, it is like a constant, every value, every sale you divide by one number, so that is not a problem. But in this case, number of households in zone 1, zone 2, zone 3 and like that for all n zones or k zones. For all n number of zones may not be same, they are unlikely to be different. So, actually H i is also a variable, so the error term distribution will be different.

(Refer Slide Time: 09:31)



Multiple Regression Analysis

- The models are identical except the error term
- As the aggregate variables directly reflect the zone size, the magnitude of the error actually depends on zone size
- Using multiplies, such as 1/Hi, makes the model independent of zone size
- The aggregate variables tend to have higher intercorrelation than the mean variables

IIT Kharagpur | Urban Transportation Systems Planning | Module C 6

So, I have written exactly for this reason, that the models are identical except the error term. As the aggregate variables directly reflect the zone size when you are using aggregate variables, the magnitude of the error actually depends on the zone size. Because how, what will be your variable values or inputs, that will depend on how big the zones are. So, the bigger zones will obviously be expected to have higher value, smaller zones will have lesser value.

So, the somewhere the error term in the estimate also actually is getting influenced by the zone size. Now second case, what we are doing in the second case? We are actually dividing it by or multiplying it by 1 by H i or dividing it by H i, number of households to make it rate. Once you do that then the values the bigger zones bigger value, smaller zone smaller value the size of the zone does not really matter, so this model is almost independent of zone size.

I said almost independent because the size of the zone is not coming directly. So, that is why the error term distributions are also different. So, the first case you have to remember that if you are using aggregate variable, the magnitude of the error actually depends on the zone size. And once you are using a multiplier 1 by H i that means using the variables at rate rather than the total then you are making the model independent of zone size.

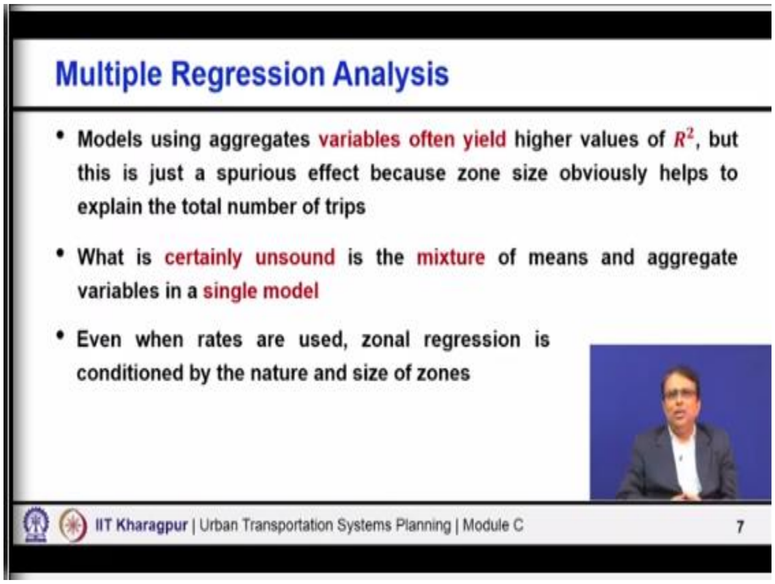
Now in the same way, what is also found? That aggregate variables when you are using in the models, they tend to have higher correlation than the mean variables. Now this is again logical if you think it sounds logical, why? Now every variable the bigger the zone size, every variable will be bigger, population also will be more, number of cars also will be more. So, every variable for smaller zones generally smaller, larger zone generally bigger.

So, when you have more population, probably if everything else is same, you will generally find the more, number of cars also. Of course, that is not that always they are perfectly every variable has to be collinear, you will not get that as well. But obviously, because this zone size is going inside the aggregate variables tend to have higher inter correlation than the mean variable.

The chances of such correlations from the data will be really lesser when you are using mean variable or the rate. So, this is also another point what you have to bear in mind and you have to be careful, that you may get simply that two variables are collinear simply because you have used actually aggregate total. So, obviously somewhere at the back end, the zone size is actually controlling everything.

So, these are very interesting aspects one has to bear in mind and that is why just getting something out of the data computer will tell you correlation value, computer will tell you R square value but the interpretation is very, very important. One more interesting thing I would like to say about this zonal base model is the next one.

(Refer Slide Time: 13:56)



Multiple Regression Analysis

- Models using aggregates variables often yield higher values of R^2, but this is just a spurious effect because zone size obviously helps to explain the total number of trips
- What is certainly unsound is the mixture of means and aggregate variables in a single model
- Even when rates are used, zonal regression is conditioned by the nature and size of zones

IIT Kharagpur | Urban Transportation Systems Planning | Module C
7

Models using aggregate variables often yield higher R square as two independent variables are likely to be more correlated the bigger the independent variable values, the bigger will be the dependent variable values as well. Because everywhere that the whole thing is getting influenced by the size of the zone. So, when you develop a model, you may apparently get a higher r square value.

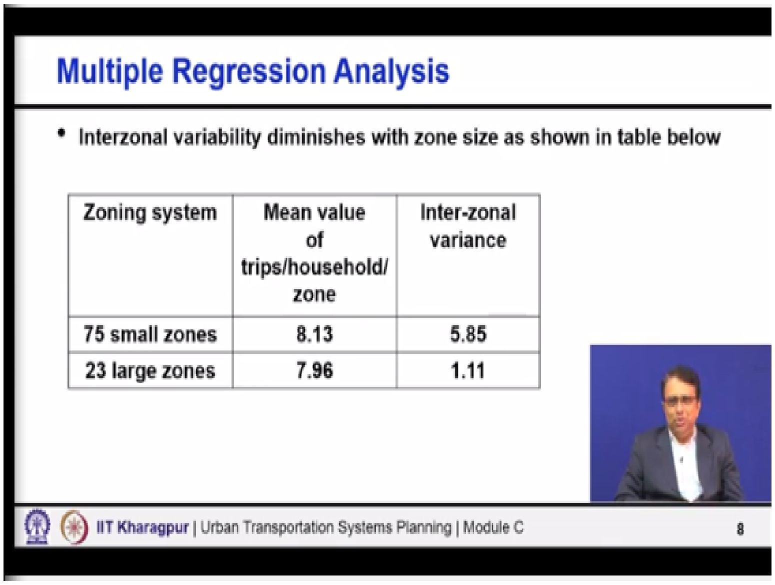
But really, that does not mean that your model using zonal total your because just simply R square is higher, you are getting a superior model than a model what you will get probably using that zonal means or the rates? Remember that the zone size again is playing a factor here. And actually, giving that it is because of that reason you are actually getting a higher R square value.

So, even what I am trying to say you have to be very careful when you are developing model understand all these aspects. And that is the reason why do not try to conclude only based on R square that my model gives better R square than his model, so my model is superior, that is too simple the conclusion**.** Inside if you have used zonal total and somebody has used zonal means.

There is no wonder that the model with zonal total will give you higher r square. What is really certainly unsound, that what happens if we use a mixture of means and aggregate variables in the single model? Normally I have also not come across such models either people use zonal means or zonal aggregate. The last point is very interesting, where we are saying that even when the rates are used zonal regression is still conditioned by the nature and size of the zone.

This probably we will find that what say I told in the previous slide and say telling now are contradictory statement. Actually, not contradictory in that sense, when we said that, yes when you were using a multiplier 1 by H i, you are largely trying to make the model independent of zone boundary. Because you were using the rate, so that is definitely true. But even then, what we are trying to say that zonal regression is still conditioned by the nature and size of zone, why we are seeing this? Look at this example, in the next slide.

(Refer Slide Time: 17:08)



Multiple Regression Analysis

- Interzonal variability diminishes with zone size as shown in table below

Zoning system	Mean value of trips/household/ zone	Inter-zonal variance
75 small zones	8.13	5.85
23 large zones	7.96	1.11

IIT Kharagpur | Urban Transportation Systems Planning | Module C

8

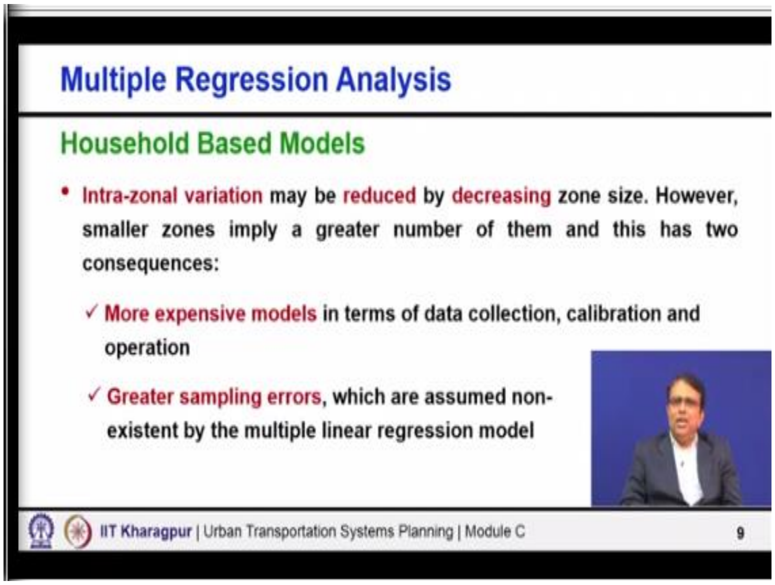
In the same study area, when 75 small zones values, that means the study area was divided into 75 small zones then the mean value of trips per household per zone came out to be 8.13. And the inter-zonal variance came as high as 5.85, the same study area when we divided into 23 large zones to represent 23 large zones to represent the study area. Then the mean value comes at 7.96 and inter-zonal variance came only as 1.11.

So, does it not indicate that even when using the zonal means, remember that our value where I am indicating here, they are mean value of trips per household per zone mean value. So, this mean value is also getting influenced by the zone size, when we are using small number of zone or large number of zones, the values are not same. More particularly the inter-zonal variation is going to be very different when you use 23 large zone you get inter-zonal and variation as only 1.11.

Whereas whenever you use 75 small zones, your inter-zonal variance came as 5.85, it increased, increased significantly and what we want? We actually want higher inter-zonal variants, so our model will be good if the inter-zonal variance is high. So, that shows that overall regression, even when you are using mean or mean value of trips per household are essentially rates, then also the regression is conditioned by the zone size.

So, that is what I made the statement even when rates are used, zonal regression is still conditioned by the nature and the size of the zones.

(Refer Slide Time: 19:53)



Multiple Regression Analysis

Household Based Models

- Intra-zonal variation may be reduced by decreasing zone size. However, smaller zones imply a greater number of them and this has two consequences:
  - ✓ More expensive models in terms of data collection, calibration and operation
  - ✓ Greater sampling errors, which are assumed non-existent by the multiple linear regression model

IIT Kharagpur | Urban Transportation Systems Planning | Module C 9

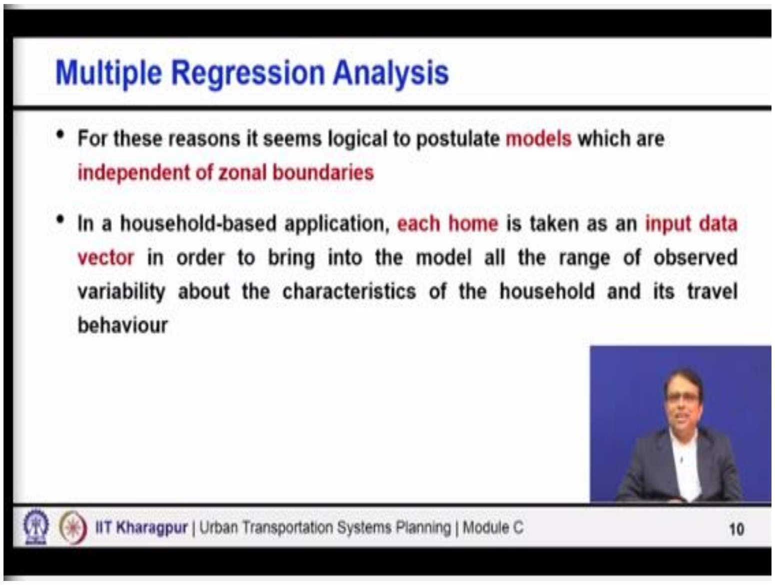
So, what it tells you that? If we use more, number of zones then we can increase the inter-zonal variation or variants. And in fact, we want, you remember that when I talked about several characteristics of zonal distribution, I said that intra-zonal versus inter-zonal variation, we want inter-zonal variations to be more and intra-zonal variation to be less. So, if I am going for smaller number of zones, then my inter-zonal variation is increasing, so what it says.

That go for smaller zone and obviously then same area if you are trying to make zone smaller. That means, you are actually going for more, number of zones and that will definitely help you to increase your inter-zonal variance and reduce intra-zonal variance. But then there are certain implications, what are those implications? First of all, you will end up having more expensive models in terms of data collection, calibration and operation.

Naturally if I am considering 23 zones or 25 zones and I am considering 75 zones, my data collection is effort is not going to be same, when you take small zones then you need much more data to represent all these zones in the overall study area in a proper way. So, your data collection, your reviewing subsequent OD data collection, all efforts will be magnified like anything.

The second is you will likely to face greater sampling error, when you increase more number of zones. This is also it goes like as expensive model, you go more, spend more to collect the data, spend more time, more money, overall, more research and you are also likely to face greater sampling error when the so many zones you are trying to consider.

(Refer Slide Time: 22:15)



Multiple Regression Analysis

- For these reasons it seems logical to postulate models which are independent of zonal boundaries
- In a household-based application, each home is taken as an input data vector in order to bring into the model all the range of observed variability about the characteristics of the household and its travel behaviour

IIT Kharagpur | Urban Transportation Systems Planning | Module C

10

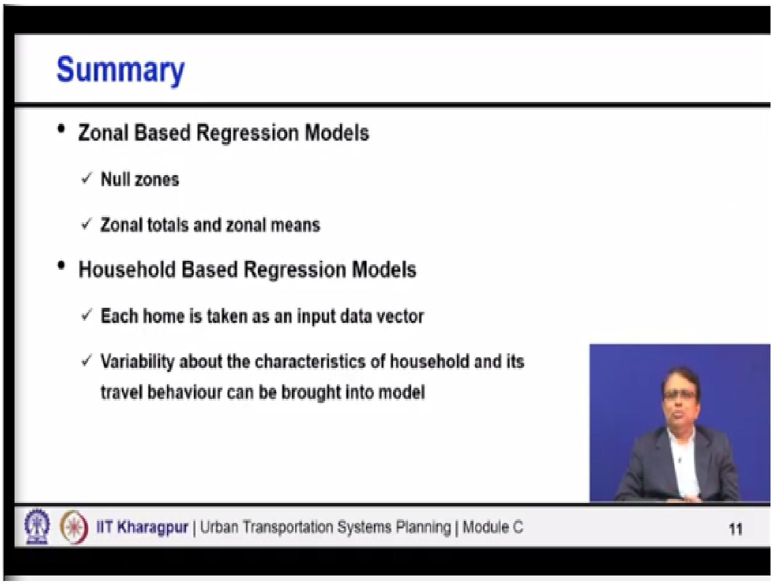
So, for these two reasons, what appeared to be logical? That we make models, which are independent of zone boundaries, can we totally eliminate that zone effect? How to do that? So, if you start reducing the zone further where we end up finally, what is the smallest unit? Household, so can we just make the model then household based model rather than zonal base model? We started with zonal base model, we told you so many characteristics, so many things.

And then reducing the intra-zonal variability by increasing the number of zones, implications also we told what logically then sound comes out? That probably we can go for household based models to eliminate the effect of zone size in our model. And also, to basically absorb or basically eliminate to some extent the impact of increase in the cost, resource, sampling errors and so, all other issues.

So, in a household based application, each home is taken as an input data vector in order to bring into the model all the range of observed variability about the characteristics of the household and it is travel behaviour. So, you take all possible characteristics income, household size, car ownership and further additional variables. And then try to explain how the household trip making is happening?

And once you do that that in a zone then how many households are there of what characteristics. And accordingly, from the household based models, ultimately I will still come back and make my zonal estimates. So, end of the day I will still go back and try to say zone 1 trip production is so much, zone 2 trip production is so much and like that.

(Refer Slide Time: 24:43)



Summary

- Zonal Based Regression Models
  - ✓ Null zones
  - ✓ Zonal totals and zonal means
- Household Based Regression Models
  - ✓ Each home is taken as an input data vector
  - ✓ Variability about the characteristics of household and its travel behaviour can be brought into model

IIT Kharagpur | Urban Transportation Systems Planning | Module C 11

Now, if we have to then summarize this whole thing, remember that zonal based regression models are widely used. You have also typically, so many variables say residential density, say for example value of land. This kind of variables you can use only when you are using zonal based model. But zonal base model has certain aspects which one has to understand, one is zonal total versus zonal means, then use of null zones, type.

Then intra-zonal variability versus inter-zonal variability, then the value of intercept and then what we found. If we try to reduce the zone size or make more number of zone then obviously that helps us to increase the inter-zonal variations and reduce intra-zonal variations which is welcome. But we ended up developing models which are more expensive in terms of data collection, calibration, operation, and also greater sampling error.

So, then and still the zone boundary influence remains in some form or other, so then what is really can be done? Probably as an alternative you can use zonal base, instead of zonal base model you can go for household base models. And in that each household is taken as a unit or an input data vector and then we try to take all different characteristics of household to explain or to relate household trip making behaviour to the characteristics of household. So, with this I close today's lecture, thank you so much.