

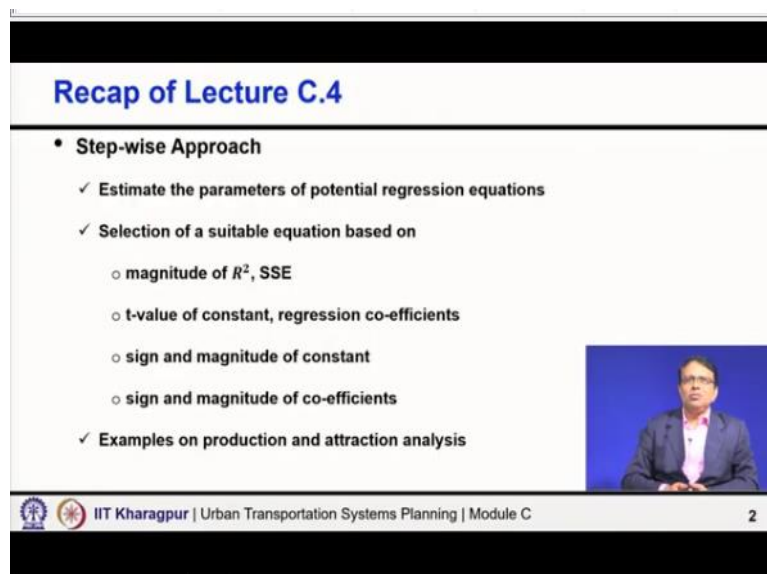
Urban Transportation Systems Planning
Prof. Bhargab Maitra
Department of Civil Engineering
Indian Institute of Technology-Kharagpur

Lecture-15

Examples, Common Mistakes and Zonal Based Models of Multiple Regression Analysis

Welcome to module C, lecture 5. In this lecture will give you some more examples of development of regression models and also highlight some of the common mistakes what people do and give you an introduction about the zonal based models of multiple regression analysis.

(Refer Slide Time: 00:40)



The slide is titled "Recap of Lecture C.4" and lists the following points:

- Step-wise Approach
 - ✓ Estimate the parameters of potential regression equations
 - ✓ Selection of a suitable equation based on
 - magnitude of R^2 , SSE
 - t-value of constant, regression co-efficients
 - sign and magnitude of constant
 - sign and magnitude of co-efficients
 - ✓ Examples on production and attraction analysis

A small video inset shows Prof. Bhargab Maitra speaking. The footer of the slide includes the IIT Kharagpur logo and the text "IIT Kharagpur | Urban Transportation Systems Planning | Module C" and the number "2".

In lecture 3 or lecture 4, we were basically talking about the step-wise approach, lecture 3 we said that, once you get the data, try to see if there are nonlinearities between dependent and independent variables, what you thought to include in your model and if you find nonlinearities then linearize them, then second is develop the correlation matrix see Y and X we want in dependent and dependent higher association.

So, wherever there is a stronger association, we want to select those independent variables because they are the stronger candidates, but also when we are including multiple independent variables, we want to make sure that they are not collinear. So, that checking was necessary. Then, that we explained with an example and showing the steps how to screen and how to select the model specification.

So, select a few models for further investigation, then we say when you do the calibration, how you get the coefficient estimates and after getting the coefficients estimate what all you check, you check the magnitude of R square, you check the t values because your estimated coefficients are to be statistically significantly different from 0, are statistically significant.

Then also you have to we said that you need to check the sign and magnitude of the constant, the sign has to be correct or logical and magnitude also has to be checked, we do not want very high constant or unexplained component, then also checking the sign and magnitude of the coefficient system. It is what we get. And then we took various examples, and also the last example, what I took in lecture 4 was an interesting examples where for 2 equations, all these aspects were getting satisfied and both models were fine.

So, I told also then, then, I would think about the application point of view, where it is easy to apply. So, obviously, if my one variable and two variable models, both are giving me the same results, I will prefer one variable model especially, if that one variable is an aggregate prediction, total employment. So, obviously, that would be preferred rather than predicting the category wise employment and then finally, the model does not give me any superior performance than the other one. So, that was another interesting application aspect, very practical and application aspect that was covered.

(Refer Slide Time: 03:52)

The slide is titled "Multiple Regression Analysis" and is labeled "Example-2". It displays a regression equation: "Zonal peak-hour work trips produced = 0.3036 (zonal households) + 0.5638 (zonal population) (R² = 0.92)". Below the equation, it asks "Should we accept this model?". A small video inset shows a man speaking. The footer includes the IIT Kharagpur logo and text: "IIT Kharagpur | Urban Transportation Systems Planning | Module C" and the number "3".

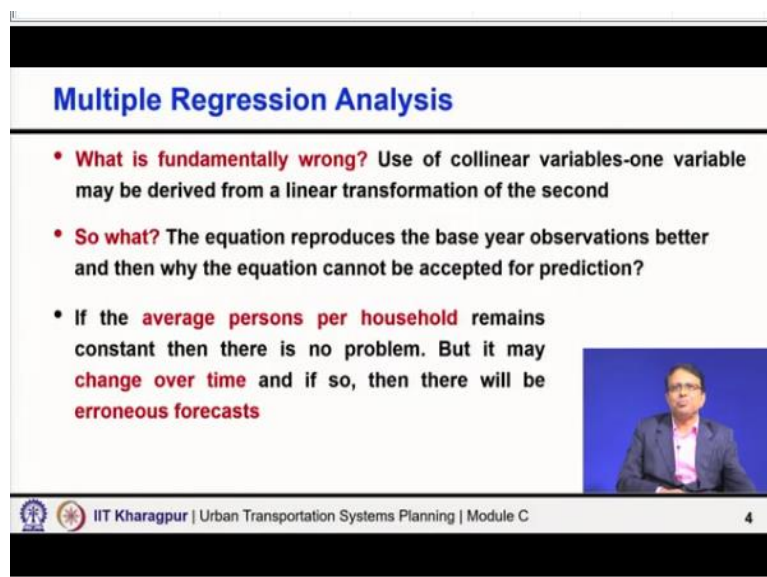
So, with this example, I am still interested to take 1 or 2 more examples to you and bring various considerations which are important for this model selection. Look at this example 2 zonal peak hour work trip produced is 0.3036 zonal households plus 0.5638 zonal population

and let us say R square is 0.92, just to ensure that, just to communicate that R square is very good.

So, which models we should then take up, which models should be taken up? Can you think what is the problem here? The problem is not with R square. The problem is not with sign but one obvious problem is there in this equation, that is useless collinear variable, we are using zonal population and zonal household, zonal population and zonal household is highly likely to be collinear.

What is zonal population? Zonal population is zonal household multiplied by the average number of persons per household, that gives you the zonal population. So, that way zonal household and zonal populations are likely to be highly collinear.

(Refer Slide Time: 05:33)



Multiple Regression Analysis

- **What is fundamentally wrong?** Use of collinear variables-one variable may be derived from a linear transformation of the second
- **So what?** The equation reproduces the base year observations better and then why the equation cannot be accepted for prediction?
- If the **average persons per household** remains constant then there is no problem. But it may **change over time** and if so, then there will be **erroneous forecasts**

IIT Kharagpur | Urban Transportation Systems Planning | Module C 4

So, what is fundamentally wrong here is the use of collinear variables and since they are collinear one variable may be derived from a linear transformation of the second, one variable independent variable is a linear function of the second independent variable. They are not utterly independent one variable is actually in a linear function of the other variable. So, they are not independent variables rather they are collinear variable.

So, they should not be used. Because I said this one repeatedly in my last lecture, it is lecture 4 that lecture 4 or lecture 3 probably that collinear variable should not be used. Anyhow, I told it probably in lecture 3 and lecture 4 both in different context. So, you should not use it.

But then somebody may be very stubborn and ask. So, what if I am getting a good R square value even with use of collinear variable, what is your problem?

What is the answer? So, the logic is his logic is that the equation produces the base year observations better than other models, then why the equation cannot be accepted for prediction. Ultimately, my objective is to predict Y, I am able to predict Y. So, why you have a problem even if I use collinear variables, I am getting my Y value estimates. The fundamental wrong thing why collinear variable should not be used, it is not that do not use collinear variable because sir somebody like me or any other person or your professors told that you should not use, try to understand why you should not use it.

I told you an example that, if they are collinear, one variable can be expressed as a linear function of the other variable. So, zonal population is equal to K into zonal households. So, whatever is the average number of households now in the base year that K average number of households have persons per household, that K value is inbuilt in the model, you do not see that, but that is there in that model.

Because your this 0.3036 and 0.5638. These coefficient estimates are based on that value of K in the base year, average number of persons per household. That number is there in this equation, you do not see it, but it is actually influencing that those coefficient estimates. So, if in future, if this average number of households changes, then you have a problem. If the number of households does not change, then you may still give you a reasonable result, but in all possibilities, different times there are so many dynamics in the whole system.

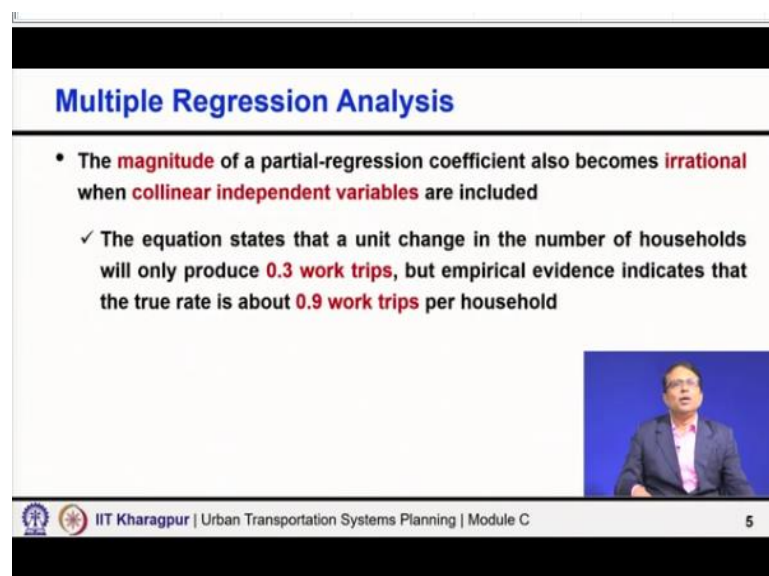
So, the number of households per person will change. And if it is changing, you will not notice normally you will not understand also and you have used collinear variables so you will get something because you will be very happy that my base year model was very good, but your predictions will be wrong. And most importantly, why I have taken this example just to tell you that in reality for every variable, it will not be so, obvious that you will be able to identify the collinearity and it will be led not so, simple like average number of persons per household.

Many cases, we may not be willing to understand that what is that constant or that relationship that is getting in built in sight. So, one thing is you must check, if there is

collinearity and you must not include collinear variables as independent variables in the same model and this is the reason for you, why you should not do it and remember, every example will not be so simple and straightforward, but the same message will remain something getting assumed or inbuilt in developing this model or equation.

And that may change over time and which you will not notice I will also not notice, but we will be very happy looking at the R square and everything use the model and our predictions will be ultimately wrong.

(Refer Slide Time: 11:38)



The slide is titled "Multiple Regression Analysis" in blue text. It contains two bullet points: the first states that the magnitude of a partial-regression coefficient becomes irrational when collinear independent variables are included; the second, marked with a checkmark, states that the equation predicts 0.3 work trips per household, while empirical evidence shows a true rate of about 0.9 work trips per household. A small video inset shows a man in a suit speaking. The footer includes the IIT Kharagpur logo and text: "IIT Kharagpur | Urban Transportation Systems Planning | Module C" and the number "5".

Also look at the thing why I said that you have to check the coefficient estimate value also. The magnitude of the partial regression coefficient also becomes irrational when the collinear independent variables are included. The equation shows go back to the equation check it once again the coefficient estimates are 0.3 into zonal household, plus 0.5 into zonal population. Actually empirical evidence shows that per household the trip rate is not low as low as 0.3 for work trips, it is actually nearly close to 1.

For that kind of data, which were actually used for this analysis, because of the shortage of time and other things I will not refer to every details or this may be even hypothetical. But 0.3 work trips in a typical urban context is not correct, it is actually low value, but the value got distorted because we use collinear variable the same thing getting explained some partly through the population coefficient and partly through the number of household coefficients. So, obviously, there will be distortion that also happens. So, when you use collinear variables.

(Refer Slide Time: 13:20)

Multiple Regression Analysis

Example-3

Other non-home based trip attracted = 0.485 (industrial & manufacturing employment) + 4.330 (retail and service employment) + 0.298 (population) + 4.624 (total employment) ($R^2 = 0.90$)

Should we accept this model?

IIT Kharagpur | Urban Transportation Systems Planning | Module C 6

Let us take one more example. Other non home based trip attracted equal to 0.485 into industrial plus manufacturing employment plus 4.330 into retail plus service employment plus 0.298 into population plus 4.62 for total employment, R square again is very high. So, should we accept this model yes or no? The answer is no. What is wrong if I look at this equation, I can find 2 things are fundamentally wrong here?


One is what we are trying to model trip attraction, other non home based trip attraction, non home based trip attraction is population a logical variable for that population is not a logical variable, how the non home based trip attraction depends on the population residential population in that area. It is not a logical variable, we may get the equation that the computer does not know what is the physical meaning of the data.

Any data you give it will try to fit with that data, but what is the logical meaning of that variable? So, the inclusion of population is not correct. The second thing, when we are considering the categories employment, industrial manufacturing employment and again retail and service employment, then again total employment. So, total employment again include industrial employment, manufacturing, employment, retail and service employment all are again included. So, again that is not acceptable.

(Refer Slide Time: 15:36)

Multiple Regression Analysis

- One must check whether there is a **causal basis** to the apparent dependency between variables
- The **relationship** between **travel demand** and the **intensity of land-based human activities** is quite direct and the validity of trip-generation equations may be assessed easily
- In the above equation, **population** is not a logical variable for trip attraction and **total employment along with categories of employment** are included which is a sort of repetition of the same variables in one regression equation



IIT Kharagpur | Urban Transportation Systems Planning | Module C 7

So, we should not accept this model, that is what I say one must check whether there is a causal basis to the apparent dependency of the variable. The relationship between travel demand and intensity of land based human activities is quite direct, as we say that attraction is whether the population is logical or not, this kind of thing is very, very direct relation and you can check the logical inclusion of the variable.


And the validity of trip generation equation needs to be assessed easily. In the above equation, population is not a logical variable for the trip attraction and total employment along with categories of employment are included, which is a sort of repetition of the 7 variable in one regression equation. So, again, somewhere there, somewhere, it is somewhere here and again, similar kinds of problems. So, that is again to be avoided.

(Refer Slide Time: 16:36)

Multiple Regression Analysis

Common Mistakes

- Use of the **coefficient of multiple determination** as the **only criterion** for the statistical validity of the regression model
- Inclusion of **collinear independent variables** in the same model
- Not checking the **logical** inclusion of variables in the model
- Not considering **model application** in future



IIT Kharagpur | Urban Transportation Systems Planning | Module C 8

Now, let us see what are the common mistakes people do when developing regression equations or carrying out multiple regression analysis. I have listed here 4 common mistakes, which I found over the years students are doing very frequently. One is use of coefficient of multiple determination or R square, as the only criteria for the statistical validity of regression models.

I am sure maybe some master students taking this course each of you will ultimately in your thesis will give one equation at least, whether you do experimental work, whether you do pavement engineering related work or planning or traffic. And in most cases, I find people report only R square and try to conclude based on the R square, which is wrong and I have given here number of examples in the last few lectures to convince you that R square is not everything.

This lecture, previous lecture, just to convince you that R square is not enough and not always a high R square means a good model, maybe we need a lower R square can give you a much better model, if all other aspects are really correct. Second, inclusion of collinear independent variables in the same model, we will check also simply include and get the results without checking if there is a collinearity.

Third not checking the logical inclusion of variables in the model. Students often do this mistake, they are so enthusiastic about developing equation that any data they get they include everything and try to fit a model without even bothering what is that data, what that column means, whether there is any logical dependency and not only for regression equation, when they do even machine learning, when they apply deep learning, they have a common tendency to do it.

Any data they give all the data they try to use. So, logical inclusion that is very important. Last but not the least, not considering model application in the future. A model is developed not for present condition, not for application in today's context. Today's context any modeling work, in today's context, we know independent variable, we know dependent variable we can go to the field measure everything we want for building the relation.

So, that is very important. For building relation, we need to know both X and Y, but application to be done only in the future. So, this model application must be considered

properly that when I apply, I must consider the applicability of the model. I give you one example where they said that 2 models nearly same performance, I selected the one which is simpler one, thinking the applicability.

But not that always I will take a simpler model. No, in most cases, you will find that the parameter richer models are generally superior that means, if you include more variables, your overall goodness of fit will improve, your R square will improve, your model performance will improve. If it is so by including more variables, if you get a superior model, which is likely to happen, most cases.

Because parameter richer models are generally superior, then please use more variables because we want to use a better model. But when it is all neck to neck, very similar, very similar, very similar, may be very slight better a square 0.01, 0.02 higher R square at that level 0.02 higher R square really does not mean anything. I should then focus on applicability of the model.

So, think of the application because any model you are doing and do not use very complex variable in the model which you cannot forecast accurately. Ultimately, please it is not only for this multiple trip generation modeling or so, anywhere any modeling you are doing you are using explanatory variables, please think that all these explanatory variables, unless you are able to forecast this explanation telling variables clearly and accurately reasonably accurately.

You need to forecast these independent variables first. If it is so, complicated that you cannot forecast these variables properly, then whatever relation you develop, maybe the relationship is very good, but if your input is not correct, your output also will not be correct. So, if your relationship is not correct, your output will not be correct, if your input is not correct, your output will again not be correct.

So, you need to strike a balance, an acceptable relationship, reasonably good relationship, reasonably good inputs will give you a reasonably good output. So, that should be kept in mind.


(Refer Slide Time: 23:39)

Multiple Regression Analysis

Zonal Based Models

a) Inter-zonal variations vs. Intra-zonal variations

- Models can only explain **variation in trip making behaviour** among zones
- Only be successful if inter-zonal variations adequately reflect the real reasons behind trip variability
- Zones should have a **homogeneous socioeconomic composition** to represent a range of conditions
- A major problem is that the main variations in person trip data sometimes occur at the intra-zonal level



IIT Kharagpur | Urban Transportation Systems Planning | Module C 9

Now, coming to as I said that regression model trip generation model we may do it at zonal basis, we may do it at household base also, household may be my units, when I describe the variables I said that many cases the household could be the unit. Let us see, most cases we go for zonal base model. That means, zonal population, zonal number of vehicles and all zonal characteristics may be residential in density, many logical variables you can use.

But there are certain considerations of zonal which model that you should be aware of. First international variation free service intra-zonal variations, what model is giving us, this zone number of trips are higher than that zone and if the income in this zone is higher than the income in that zone or the population density here is higher than the population density there. The variation of Y is getting explained by the variation of X where one zone to another zone.

Some zone it is higher, some zone it is lower. So, it is actually trying to say one zone to another zone to another zone how things are changing. So, basically inter-zonal variations is very important, many cases the data could be like intra-zonal. How I can explain this? Let us say maybe 2 hypothetical examples, one case let us say you have 10 different zones. In one zone if all households said their every case the family income is 20,000 or let us say 30,000.

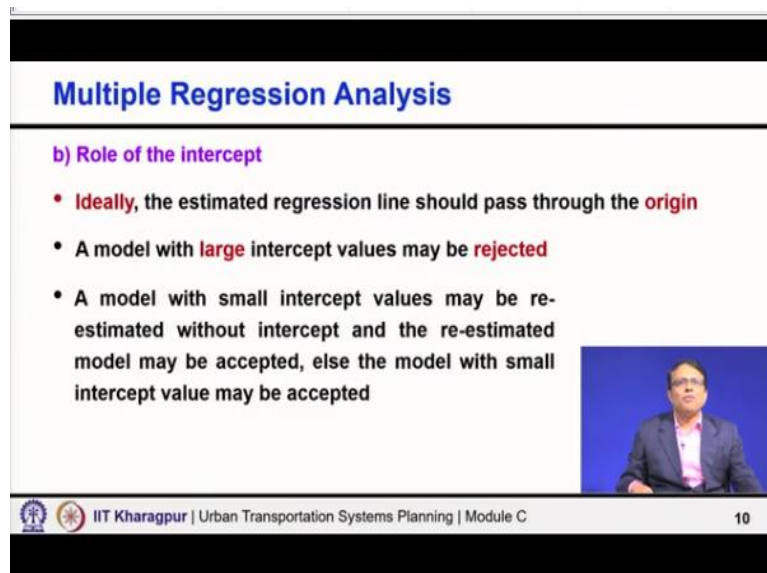
Another zone every household the income is 50,000, another zone every household income is one lakh like that, one zone to another zone to another zone, there is a distinct variation and within that zone, everybody is homogeneous, everybody is having some income, then you will actually get a very good model because one zone to another zone how really the income changes and then what is the impact on the trip making behavior or the trip generation.

But if it happens, here also you have people whose income is 1 lakh to 10,000 every zone u have 1 lakh to 10,000 but on an average some income here some income there. So, here you get 30,000, there you get 40,000, somewhere average is 38,000, somewhere 49,000, you get that variation, but variation from one zone to another zone to another zone inter-zone variation is relatively small, your actual variation is the intra-zone.

Within a zone the variance is much higher than the variance in the data from one zone to another zone. If that is so, your model is not going to be good. So, what we say model can only explain variation in the trip making behavior among zones. So, only be successful is the inter-zonal variations adequately reflect the real results of trip making behind trip variability and for that reason, it would be ideal if the zone become homogeneous as far as in every characteristics.

And what we sometimes find the major problem is the main variation in the personal trip data occurs at the intra-zone level, the variation is more within zones rather than in between zones.

(Refer Slide Time: 28:04)



Multiple Regression Analysis

b) Role of the intercept

- **Ideally**, the estimated regression line should pass through the **origin**
- A model with **large** intercept values may be **rejected**
- A model with **small** intercept values may be re-estimated without intercept and the re-estimated model may be accepted, else the model with small intercept value may be accepted

IIT Kharagpur | Urban Transportation Systems Planning | Module C 10

Second, before I come to this point, so, how we can sort out this problem? One possible is to reduce the zone size, if the zones becomes smaller it is likely to be more homogeneous. There are implications and other we will discuss further. The second consideration is the role of intercept, as I said ideally the estimated regression line theoretically should pass through the origin.

So, unexplained component intercept should be 0. So, that is theoretical. Is it possible to get that every model? No. Any data will have some source of variation and any practical data you will get some constant value, but what is important? If we are getting large intercept value that means, unexplained component is very significant if your values are in suppose a 10, 20,000 and if you are getting some 4000 as a constant I will consider it is very significant.

But if your support values are in 10, 20, 30, 40,000 and if you are getting 5100 as a constant, it is nothing. So, large intercept is a problem, then model may be rejected. But if there is a small intercept it is okay. And when there is a small intercept what can be done, one can try to re-estimate the model forcing the line to pass through the origin. That means when you develop model, you can develop it with constant and without constant also.

So first we develop without constant and we know that it is not very high value. So, we can retain it, that is one possibility. The other is we try to force it to pass through the origin. So, re-estimated without constant, there could be 2 consequences with that, once you do it, you may still find that maybe the coefficient estimates marginally has changed here and there, t values have marginally changed here and there.

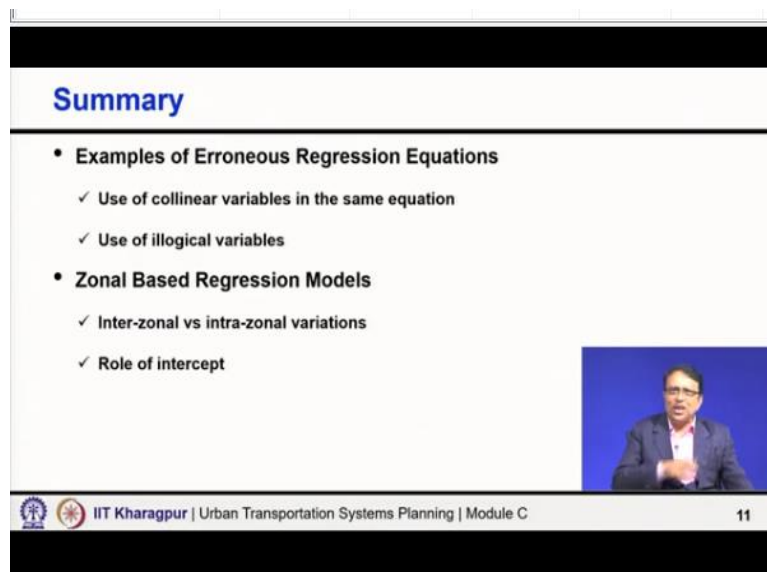
R square also has changed marginally here and there. But overall model sounds still logical, maybe I was getting 0.8 R square now I am getting 0.72. But all variables still a significant, all signs are still logical. If that is so, then use it, use that model. Even though you have a little lower marginally lower R square or so, but still use that model, use the model without any intercept.

But some cases you may find, we do not know now, which is a small value, which is a large value I said that if it is small, but what is the small value, it is all relative. So, by forcing it to pass through origin, if you do that, then you find the whole model is distorted instead of plus maybe something because minus some variable went insignificant, an R square drastically change from 0.8 to 0.4 or 0.5.

A visible change in the overall model and something went wrong totally, you may get both. If you get that, then you say fine, like enough is enough, I should go back let the constant be there and I go ahead with that model, but generally I would say when you are selecting model, even if you cannot force the model to pass through the origin.

When selecting model, we should really give attention and care to this fact that we want to select a model without compromising other fundamental things like sign should not be wrong or the statistically significantly, the coefficient estimate should be significant statistically, those things we cannot compromise but without compromising those we would obviously prefer a model where the intercept is smaller or lower as compared to others.

(Refer Slide Time: 33:00)



The slide is titled "Summary" in blue text. It contains two main bullet points, each with a sub-bullet. The first main bullet is "Examples of Erroneous Regression Equations" with sub-bullets "Use of collinear variables in the same equation" and "Use of illogical variables". The second main bullet is "Zonal Based Regression Models" with sub-bullets "Inter-zonal vs intra-zonal variations" and "Role of intercept". In the bottom right corner of the slide, there is a small video inset showing a man in a suit speaking. At the bottom of the slide, there is a footer with the IIT Kharagpur logo, the text "IIT Kharagpur | Urban Transportation Systems Planning | Module C", and the number "11".

So, that is brings the end. So, what we discussed here, we continued with the previous example, here also, we give some new example regarding the regression equations and we explained with one example, that how the collinear variables why it should not be used in the model and also said that why it is important to check the logical inclusion of the variables and then the 2 basic considerations of zonal base regression that we discussed.

One is inter-zonal variations versus intra-zonal and we say we want inter-zonal variations to be high, not the intra-zonal, but in many cases that may be a problem. And also we discussed about the role of intercept in the context of zonal based equation and then what are the possibilities and what are the different aspects we tried to discuss? So, with this, I close this lecture and thank you so much.