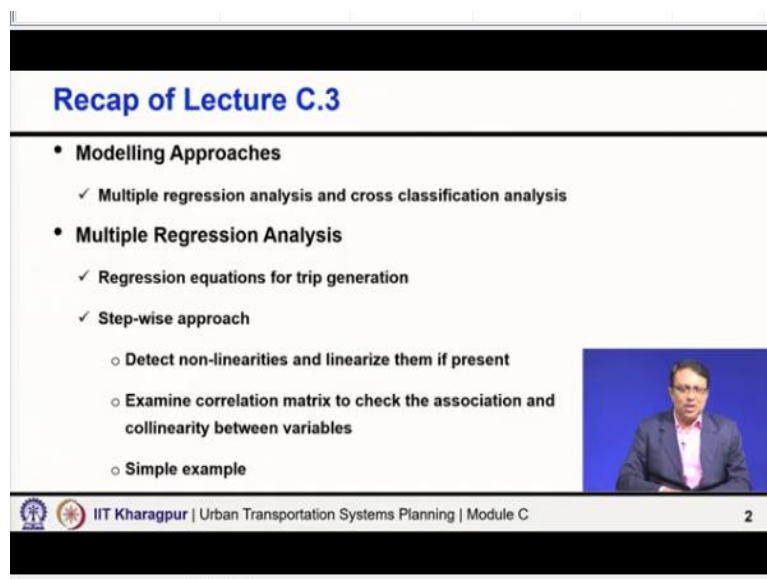


Urban Transportation Systems Planning
Prof. Bhargab Maitra
Department of Civil Engineering
Indian Institute of Technology-Kharagpur

Lecture-14
Step-Wise Approach of Multiple Regression Analysis and Examples

Welcome to module C lecture 4. In this lecture, we shall continue our discussion about the step-wise approach of multiple regression analysis and shall take a few more examples.

(Refer Slide Time: 00:32)



The slide is titled "Recap of Lecture C.3" and contains a bulleted list of topics. The first bullet is "Modelling Approaches" with a sub-bullet "Multiple regression analysis and cross classification analysis". The second bullet is "Multiple Regression Analysis" with sub-bullets: "Regression equations for trip generation", "Step-wise approach" (which includes "Detect non-linearities and linearize them if present", "Examine correlation matrix to check the association and collinearity between variables", and "Simple example"). A small video inset shows a man in a suit. The footer includes the IIT Kharagpur logo and text: "IIT Kharagpur | Urban Transportation Systems Planning | Module C" and the number "2".

In the last lecture initially, we talked about 2 approaches of doing that trip generation analysis, the multiple regression analysis and also the cross classification or category analysis. Then we started more discussion about regression analysis, how to develop regression equations and what we try to do in regression model and then, when to talk about the steps 1 by 1.

We said the first step is try to detect nonlinearities and if we find nonlinearities between X and Y between any independent and dependent variable, then we pick suitable transformation of variables to make the relationship linear. Then second is develop the correlation matrix which any computer program can give and then try to investigate if the correlation each independent variable and dependent variable have high statistical association, in fact, we want that.

So, if we get strong statistical association there are strong candidate. So, that means, for modeling Y expressing Y we can use those X values, the weak 1 all cannot be used. And also you wanted to see if some of the independent variables are really collinear. So, each pair of variable what is the correlation coefficient that we wanted to check.

And if the correlation coefficient between 2 variable is high, that simply indicates that they are collinear. Collinear means they are not independent variable. So, by default our assumption is that all X are independent variable, but if 2 variables are collinear, then they are not independent, because 1 variable can be expressed as a function of the other variable. So, we do not want and we cannot include collinear variables in the same model.

Then, we said that given a hypothetical example of a correlation matrix, how we can logically select the models? And there I said that, we first tried to find out who are our strongest player.

(Refer Slide Time: 03:28)


Multiple Regression Analysis

- Simple correlation matrix

	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

Where,

- X_1 = Population
- X_2 = Number of households
- X_3 = Vehicle ownership
- X_4 = Distance from CBD
- X_5 = Income
- Y = Peak-hour trips produced



IIT Kharagpur | Urban Transportation Systems Planning | Module C
3

In that example, if you go you will find that X 1 and Y is a potential specification, is a potential combination because X 1 and Y very strong association. Similarly, X 3 and Y strong association so, you also took that we took 1 model as $Y = a + b X_1$, we took another model $Y = a + b X_3$ fine, then logically if 2 are so strong can we put them together then our model also will be very strong?

Yes, we should try to do that. So, we tried to put X 1 and X 3 together and luckily in this case X 1 and X 3 are not collinear So, there is no problem in putting them together. So, you put

them together then we found X_4 is also not X_1 , X_4 not collinear X_3 X_4 also not collinear. So, we could easily add also X_4 and develop another model and try to see if that gives us better goodness of fit.

What was left out theoretically some more combinations? Yes, but you will know that when X_1 and X_2 are too strong players we try with X_1 alone we try X_2 alone X_1 and X_3 alone then X_1 and X_3 combination then also want to include X_4 , but X_5 we cannot include with X_3 because X_3 and X_5 are collinear. So, then I could theoretically try another model where only I have X_2 X_4 and X_5 , omitting X_3 .

But then whom you are putting in the sideline you are putting a very strong variable X_3 and bringing X_5 in due of have that one can try, but most cases I will say based on my experience of modeling that you are unlikely to probably get that as a finally, the best model. But of course, there is no harm you can always try a few more combinations.

(Refer Slide Time: 06:20)

Multiple Regression Analysis

- In this problem the analyst has the option of the following regression equations:

$$Y = a + bX_1 \dots\dots\dots(A)$$

$$Y = a + bX_3 \dots\dots\dots(B)$$

$$Y = a + b_1X_1 + b_2X_3 \dots\dots\dots(C)$$

$$Y = a + b_1X_1 + b_2X_3 + b_3X_4 \dots\dots\dots(D)$$

Any more possibilities?

IIT Kharagpur | Urban Transportation Systems Planning | Module C 4


So, with this 4, let us take this 4 or if you wish maybe let us take for the sake of argument another 2 or 3 variables models also we could consider fine, we could consider that nobody stops we can explore in whenever you will actually develop a model you can explore even more possibilities, but here let us take for this example, we go ahead now with this 4 model specification. So, what is the further step after specification, it is the calibration. So, let us go for that.



(Refer Slide Time: 07:00)

Multiple Regression Analysis

Step-3

- Estimate the parameters of each of the potential regression equations and investigate the following
 - ✓ What is the magnitude of R^2 ?
 - ✓ Do the partial regression coefficients have the correct sign and are their magnitudes reasonable?
 - ✓ Are the partial regression coefficients statistically significant?
 - ✓ Is the magnitude of 'a' reasonable?





 IIT Kharagpur | Urban Transportation Systems Planning | Module C
 5

So, we now go for calibration. So, we estimate the parameters of each of the potential equations that again computer will give you, excel can give you, many other software, which are commonly used can give you. So, I will again not tell you how to estimate the equation, but that most computer program will able to do it, estimation technique and other things you will learn. I will again discuss something which the computer will not be able to do.

That you have to do as a planner or as a modeler, decision making, again based on the values given. So, how to then select a model? You fit that thing you decide the model specification you have a set of data. So, in each case, the coefficient estimates, go to the previous focus on first case A and B, second case A and B, third case A B 1 B 2 and the 4 th case A b 1 b 2 and b 3.

All this coefficient estimates will be available to you. Now decision has to be taken by you which model you are going to finally accept. So, what we all need to check? We need to actually check maybe these 4 aspects, first what is the magnitude of the R square, that we need to check. Second, we need to check do the partial regression coefficients have the correct sign very important logical sign.

If I know if X increases Y should also increase. Say for example, if a type of employment increases the number of trips attracted should increase cannot decrease. Everything else remains if it cannot decrease. So, population increases number of trip production should increase it cannot decrease. So, that logical sign correct sign and second are their magnitudes reasonable? Magnitude sometimes get distorted.

So, if you have used collinear variable without checking if collinear variables are inserted in a model, the coefficients get distorted because it is getting distributed in 2 variables. Same thing getting distributed in 2 variables because they are collinear. So, this is again and that you should check the are you getting a logical value per household, trip rate you know wherever you were working Indian cities.

This kind of cities, medium midsize cities or big metros are these in what range they vary normally. So, I should not get something totally upside down a completely different value. It is not possible. So, that you may not know the exact value of course, none of us, we may not know, but we should give value sounds logically explainable, you should be able to explain it logically. So, the correct sign and logical magnitude.

Third are the personal regression coefficients statistically significant. Take any model specification, give any data set, it will give you some coefficients estimate. But does it mean we are going to use that? No, we need to check whether these coefficient estimates are statistically significant? What does it mean? You get a value. But the value alone does not tell you whether they are statistically significant why?

Because of value of the coefficient depends on the unit of the variable. Say for example, just I am giving you an example, if I take rupee as my unit, not directly related to this, but still a good example, let us say rupee, I express my cost in rupee I many give a big coefficient into rupee rupees by variable let us say X. So, big coefficient value a into X, you take the same rupee in express it in 1000s, the coefficient value will be small.

You will express rupees in crore or million, billion you may get a few decibels also a very small value. So, how big or small that value that does not tell you anything, what is important is that coefficient estimate statistically significantly different from 0, if the coefficient estimate is statistically significantly different from 0, then it is actually contributing to explain you why.

And if it is not statistically significantly different from 0, then that coefficient and 0 statistically are not different anyway. So, inclusion of that does not make sense, whatever remaining thing we interpret that variable you have included forcibly it does not contribute, it

is a garbage. So, we must check that. Fourth is the magnitude of a reasonable. What is a? I hinted in my previous lecture, a is the unexplained component constant, it is quite okay if you get a component which is unexplained component there is nothing wrong.

But if you find that my constant value is a very high value, that means I am not able to with my combination of variables, I am not able to explain the variation of Y adequately, lot of portion is unexplained. So, given everything else same, if I am getting 2 alternative model, everything else is same. But one case I have a high a value another case a lower value, obviously will prefer the **low** model with a low value of a, low unexplained value or constant.

So, all these have to be checked, most students, they forget all this, only select high R square, sir I got high R square means my model is very good. No, you have to actually check all these aspects. And these are extremely, extremely important. And the computer program will not give you a decision. It is you who have to take a call and then select and tell which model is the best for the given data or for the given context. Let us take it example, to move forward.

(Refer Slide Time: 15:01)

Multiple Regression Analysis

- The four equations were fitted to the data with the following results:

Model (A):
 $Y = 894.5 + 0.66X_1$
 • Standard error of estimate (SSE) = 570.4
 • R^2 coefficient of determination = 0.68
 • t value of regression coefficient = 9.08

Model (B):
 $Y = 444.39 + 1.045X_3$
 • Standard error of estimate (SSE) = 542.93
 • R^2 coefficient of multiple determination = 0.74
 • t value of regression coefficient = 9.75

Legend:
 X_1 = Population
 X_2 = Number of households
 X_3 = Vehicle ownership
 X_4 = Distance from CBD
 X_5 = Income
 Y = Peak-hour trips produced

IIT Kharagpur | Urban Transportation Systems Planning | Module C 6

The 4 models what we selected, let us assume that for a given data when we fit the model, we get this results, to carry forward our discussion. The model models we selected for the given data, we put those models and whatever computer generated are given here, look at this equation $Y = 894.5 + 0.6 X_1$ a $Y = 444.39 + 1.045 X_3$.

(Refer Slide Time: 15:47)

Multiple Regression Analysis

$Y = 180.4 + 0.38X_1 + 0.65X_3$ (C)


- Standard error of estimate (SSE) = 429
- R^2 coefficient of determination = 0.85
- t value of regression coefficients = 4.88, 5.49


$Y = 957.55 + 0.35X_1 + 0.55X_3 - 32.65X_4$ (D)

- Standard error of estimate (SSE) = 420.87
- R^2 coefficient of multiple determination = 0.92
- t value of regression coefficient = 4.56, 4.26, -1.27

(Note: t-value depends on confidence level (in the example, it is 95%) and degrees of freedom (depends on number of variables considered in an equation). If estimated value > tabulated value, it is significant)

X_1 = Population
 X_2 = Number of households
 X_3 = Vehicle ownership
 X_4 = Distance from CBD
 X_5 = Income
 Y = Peak-hour trips produced





IIT Kharagpur | Urban Transportation Systems Planning | Module C

7

Then $Y = 180.4 + 0.38 X_1 + 0.65 X_3$ and the last one $Y = 957.55 + 0.35 X_1 + 0.55 X_3 - 32.65 X_4$. And along with that, I have given standard error of estimate R square value and T value. So, all are given. Standard error of estimate also you can check. Obviously, even given everything else same all other parameter, we would actually selective model where the SSE is lower, standard error of estimate is lower.

So, now the decision to be taken, which model among these 4 should be accepted. Let us look at this based on R square 0.68 0.74 for b 0.85 for C and 0.92 for D. So, in terms of R square, the best one is model D. But then, let us check other points. As I said R square is not everything, R square is one aspect which should be checked. And from the point of view of R square D is my selection. But can I go ahead with that?

Let us see, first thing immediately I can see in this slide, we need to check if the coefficient estimates are statistically significant, I can clearly see the t value of X_4 in model D is 1.27 which is not statistically significant, even at 90% confidence level. Of course, I have indicated here, the t value depends on the confidence level. Normally we take 95% confidence level.

Some cases, you can even take 90% confidence level that is also fine. And planning purpose we normally do not expect this 0.92, we have shown because this is an hypothetical example I have taken the freedom to put a high R square value but it is much high R square value unlikely to be obtained in any planning studies even a lower square is fine. But here what you find this t value 1.27 it is not statistically significant.

I have indicated here if you want to calculate t value, you have to see the confidence level you have to see what is the degrees of freedom, which depends on the number of variables considered in the equation. And then what is the estimated t value after calibration of the model that estimated t value has to be greater than the tabulated t value. So, generally I would say even in 90% confidence level it should be around 1.64.

But I find it is even lesser than that, around 1.6, 1.64 around that it should vary expected to be around that, but here it is 1.27. So, 1.27 is definitely not statistically significantly different from 0 at 90% confidence level. So, this is the forced inclusion does not help. So, how can even if I get high R square it does not make sense. Second, let us look at the sign, sign as come as negative, what is X 4?

X 4 is the distance from CBD yes, it is logical, distance from CBD, the more the distance from CBD trip generation will reduce, we are trying to say peak hour trip reduce, peak hour trip produced will reduce with the distance. So, with population it should be positive, with number of households it should be positive, with vehicle ownership it should be positive, but with income it should be positive, but with the distance from CBD, it is likely to be negative because distance increases.

So, number of trips production will reduce. So, it is negative sign is logical. So, what I found here the t value coefficient estimates is problematic. So, maybe I will not use this model. I know other models 0.85 t values are fine, signs are also fine. If I go back all cases t values are fine, the signs are also fine, but look at this, the unexplained component A B and C. Unexplained component here is the highest 900 which are which service nearly half of that.

And compared to B A is again further reduced, even less than half of that. So, if I have to select I would probably reject A because of the high unexplained component unusually high. Now, between B and C so A is rejected because of the high everything else is fine. So, I have to compare it based on other parameters. So, I found that A is basically high unexplained component or constant as compared to other 2. So, we will reject.

Then B and C also A has got the least R square value, that is also one result. So, the model is not good. So, the R square is also getting reflecting that. So, if I have got a model with 0.85

or 0.74 and then with a higher unexplained component and with a lower R square value, why should I accept it. So, obviously, A gets eliminated. Now, I have B and C, B and C sine fine coefficient values logical, statistically significant estimates both cases it is statistically significant.

Here it is 444.39 and there it is 180.4. But obviously, I also find that the R square is more. So, if I have not missed any point, as I am looking it now here in the screen, then the choice should be actually model C. That means I am using population and vehicle ownership. So, vehicle ownership data or vehicle data. So that probably we should take a final call and we get a better model here 0.85 is the R square as compared to 0.74. So, obviously, and standard error of estimate is also low. Yes, it is lower than even what we got from B and R square is also improved.

(Refer Slide Time: 25:22)

Multiple Regression Analysis

Examples

Example-1

• For trip attraction analysis, following two alternative models are developed:

$Y = 61.5 + 0.93X_1$

• SSE= 208.4; $R^2 = 0.90$; t value of coefficient = 42

$Y = 25.8 + 0.89X_2 + 1.29X_3$

• SSE= 199.4; $R^2 = 0.91$; t values of coefficients = 51, 17

Which model should be selected?

X_1 = Total employment
 X_2 = Manufacturing employment
 X_3 = Retail and service employment
 Y = Peak-hour trips attracted


IIT Kharagpur | Urban Transportation Systems Planning | Module C 8



So, we select that.

(Refer Slide Time: 25:23)

Multiple Regression Analysis

- Standard errors of estimate of both the models are almost close
- R^2 values of both the models are very high and close
- The two models are statistically significant
- Magnitudes and signs of the co-efficients of both the models are reasonable
- There is no big difference between the constant values of both the models
- In this case, a **simple model with a single variable is chosen**





 IIT Kharagpur | Urban Transportation Systems Planning | Module C
 9

Now, let us take another example. These 2 equations are given, these are again hypothetical examples, because I quickly want to convey to you a few other things. Let us take a model we are trying to do now peak hour trip attraction, one case I gave you $61.5 + 0.93 X 1$ is total employment and here I get standard error 208 R square 0.9 and t value 42, 42 is very high.

So, obviously, it is statistically significant. There is no doubt about that. And second equation, I get $25.8 + 0.89 X 2 + 1.29 X 3$ R square is 0.91, it is also very high even 0.01 value higher than the previous one and t values are 51 and 17 that is fine. Now, one thing I should tell and you should know probably the t value we are only worried whether it is statistically significant or not.

Now, once it is statistically significant, whether the R value is 10 or higher value is 50. That is we are not going to consider that, because it does not make sense to consider. Both cases it is statistically significant. So, we only want to make sure that it is statistically significant. So 51 and 17 both R square values are very high, both are acceptable at 95% level. So, which model to be selected, both are very competitive model.

You see, it is very interesting example, because both models are competitive, both are good, both cases R square are same, statistically significant coefficient estimate and if I say peak hour trips attracted number of trips getting attracted in peak hour 60, 162 is higher than obviously 25 or 26 values, but 60 is also not a high value, in a zone how many trips are getting attracted 50, 60 is unexplained number is nothing actually.

Both are low value, not high. So, which one to select? Here look at this part, very competitive, highly comparable models. So, anyone you can select, but then finally, if I have to select one I will go now, beyond all checks what I have said to you so far, because all those checks are giving me a tie, but I have to do something to break the tie. So, I bring another consideration now.

You see what is X_1 , total unemployment, what is X_2 and X_3 at the 2 categories of employment? Here first one 1 variable, second one 2 variables. So, normally 2 variables are more parameter richer than 1 variable. That model will be more parameters richer, but by adding more parameter or variable, I am not able to get a significantly better model. Because 0.90 and 0.91 R square no difference, 61 and 25 constant values again no difference.

So, there is actually no difference. So, then why I consider more variables in the model, because it not just developing the model now, but you are developing the model for the future. So, you have to predict it, apply it for the future. And if I am using parameter richer model then I have to do forecasts of these each variables, accurately why I should additionally force?


Because if I say you predict 1 variable and you predict 5 variables and that to you look at this, total employment prediction is likely to be much easier than the predictions of category wise employment. Like prediction of population for India, and prediction of population for a state, these are this prediction of population or forecast of population for a district, where we are going to be more where the error is likely to be more, think carefully.

So, the same spirit we are bringing, because there is no significant gain by putting more variables. And looking at the variables that first case it is the total employment. The second case, there are 2 different categories of employment. If I am not getting any significantly better model, then I must use or accept a model, which I can apply easily in the future. So, I am now thinking beyond all the checks that in the future I have to apply, I have to predict my or forecast my independent variable. So which one will be easy for me to apply? So, based on that, I will select the first model.

(Refer Slide Time: 32:04)

Multiple Regression Analysis

- Standard errors of estimate of both the models are almost close
- R^2 values of both the models are very high and close
- The two models are statistically significant
- Magnitudes and signs of the co-efficients of both the models are reasonable
- There is no big difference between the constant values of both the models
- In this case, a **simple model with a single variable** is chosen



IIT Kharagpur | Urban Transportation Systems Planning | Module C 9


So with this that is what I explained here, the standard errors of estimate of both the models are almost close. R square values about the models are very high and close, 2 models, all variables are statistically significant, magnitude and sine of the coefficients of both models are reasonable. And altogether there is no big difference between the constant values of both the models.

So, in this case, simple model with a single variable is chosen, but if I would have got by inclusion of more variable a better model significantly better my R square jumps from 0.8 to 0.9 or 0.7 to 0.78 or 75 I would have used that parameter richer model.

(Refer Slide Time: 32:56)

Summary

- 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



IIT Kharagpur | Urban Transportation Systems Planning | Module C 10

So, that is how we discussed here stepwise approach, once you have the estimated parameters, then how to select the model based on magnitude of R square, standard error,

looking at the t value or statistical significance of the coefficient estimates, looking at the sign magnitude of the constant and the sign and magnitude of the coefficients and then also give some examples to explain you, different aspects.

And once you have a tie of all these, every case it is given good from all this perspective, then what to think, then think of the applicability. So, with this I will close here, we will continue in the next lecture. Thank you so much.