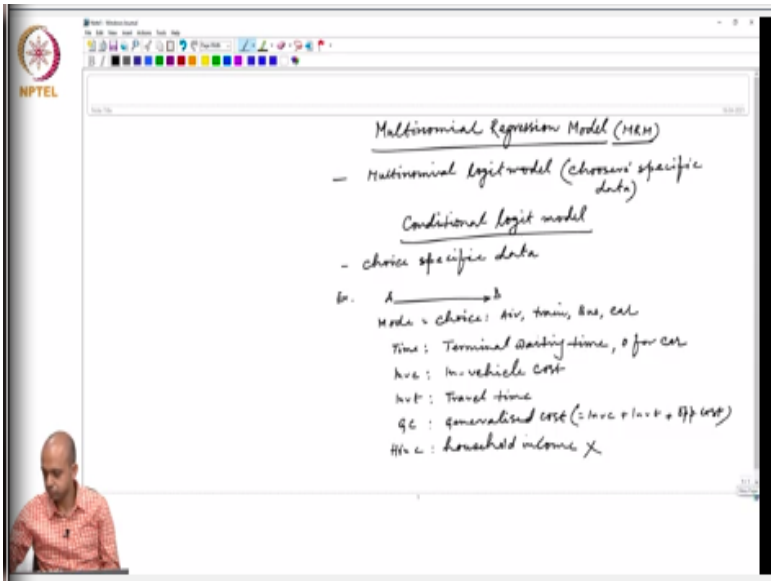
Applied Econometrics
Prof. Sabuj Kumar Mandal
Department of Humanities and Social Sciences
Indian Institute of Technology-Madras

Lecture-37
Multinomial Regression Model-Part IV

In our discussion on qualitative response model, we were discussing multinomial regression model. That means the dependent variable, it takes more than 2 values, the respondents or individuals are facing more than 2 options from which to select. And in our last class we were discussing about multinomial logit model.

(Refer Slide Time: 00:49)



So, that means this is multinomial regression MRM, we were talking about multinomial regression model or in-short this is MRM. And in the last class we talked about multinomial logit model which is basically a chooser's specific data. So, what does this chooser's specific data mean? That means the example what we are discussing in our previous class that was a college choice problem.

It means a candidate after getting his or her +2 degree is deciding whether to go for a 2-year college or 4-years college or no college at all. And that college choice problem we are hypothesizing that whether an individual will take no college, 2-year college or 4-year college that basically depends

on chooser's specific data that means individual specific data. Like the individual's family income, family size, individual's average grade up to +2 so on and so forth.

All these individual specific data will determine whether the individual will go for a 2-year college, 4-year college or no college at all. Now today what we are going to discuss is another interesting model which is called conditional logit model. So, this is conditional logit model and here in this conditional logit model this is called choice specific data. For example, let us take an example let us say there are 3 modes of travel available from A to B.

What is the choice you have? You can travel by air, you can travel by train, you can travel by bus or by car, these are the 4 modes that means these are the 4 options available, so 4 choices are there out of which individual will select 1. But that does not depend on individual specific data like the multinomial logit model we discussed earlier. Rather whether the individual will choose air, train, bus, or car that depends on the characteristics which are related to air, train, bus and car.

What are those? The choice specific variable is time, what is the terminal waiting time? So, here time means terminal waiting time. For example, when you go to airport what is the waiting time? When you go to railway station you have waiting time and you go to the bus stop you have a waiting time, but we assume for car there is no waiting time, terminal waiting time which is 0 for car. Then we have INVC, one variable called INVC which is in vehicle cost.

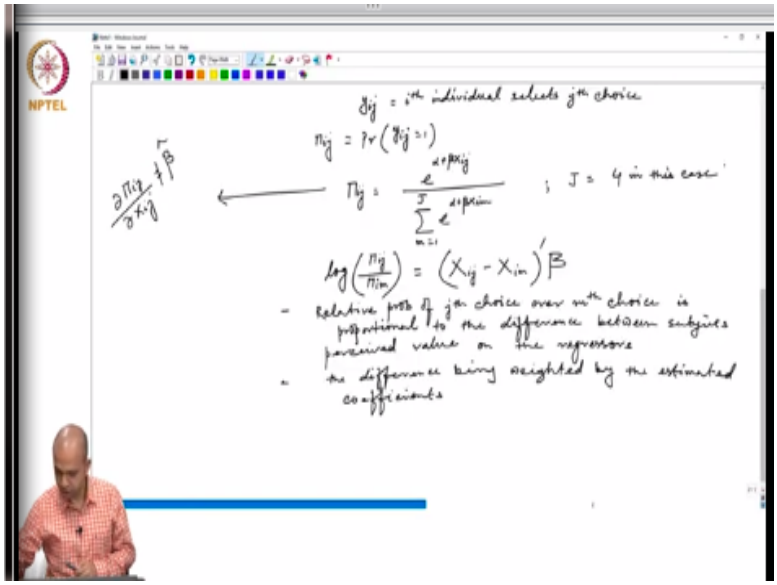
What is the cost of that mode of travel? If you hire a car or if you buy an air ticket, if you buy a train ticket, everywhere there is a cost involved this is called in vehicle cost. Then in vehicle time INVT which is different from terminal waiting time, this is called travel time basically. And then you have let us say GC generalized cost which is defined as your in-vehicle cost INVC + INVT + opportunity cost.

So, that means cost does not mean only in monetary cost, you have your time, money, and energy, so everything including it is called generalized cost. And let us say that in our dataset we have HINC which is called household income. Now since this is a choice specific data, in conditional logit model we have to clearly keep in mind that this type of individual specific data we cannot

include. We cannot include this type of individual specific data, why this is so because individual specific data remain constant across these different modes of travel.

Here I have different modes of travel and it does not vary, we are hypothesizing that it depends only one choice specific data. So, since household income or individual's specific data are even fixed across different modes of travel in conditional logit model we cannot include any individual specific data. We have to include only the choice specific data. What is the terminal waiting time, in vehicle cost, travel time so on and so forth?

(Refer Slide Time: 08:29)



And then the econometric model let us say once again y_{ij} is basically indicates that the ith individual selects jth choice. So, it indicates ith individual selects jth choice.

$$\Pi_{ij} = Pr(y_{ij} = 1)$$

And probability y_{ij} = 1 is basically indicated Π_{ij}, the probability at which ith individual selects jth choice that is nothing butΠ_{ij} like the previous example. Then what is your Π_{ij}? $\Pi_{ij}$is basically

$$\Pi_{ij} = \frac{e^{\alpha+\beta x_i j}}{\sum_{m=1}^{j} e^{\alpha+\beta x_i m}}$$

Sorry this is x i m where am runs from 1 to J, J is the total number of choices available, where J equals to let us say 1, 2, 3, 4 in this particular case. You have air, train, bus and car so J = 1, 2, 3,

4 in this case or rather I will say that J equals to simply 4 in this case J = 4, we have 4 options available. Now one thing we have to keep in mind that here unlike the multinomial logit model there is no subscript attached with α and β but here the subscript are attached with x, x is the explanatory variable.

So, when I am saying x_{ij} basically it indicates what is the value of the explanatory variable for that particular choice. So, x_{ij} for example may indicate travel time for the jth mode of choice or may indicate in vehicle cost for the jth choice or maybe terminal waiting time for the jth choice. So, since this x varies across choice that is why x_{ij}, it does not there is no subscript or attached on α and β that is the point what we wanted to make.

And then unlike the multinomial logit model what we have to do? We have to take the relative probability

$$log\left(\frac{\Pi_{ij}}{\Pi_{im}}\right) = \left(x_{ij} - x_{im}\right)'\beta$$

Here x is a vector of regressors, β is also a vector because you have many regressors here. So, that means from this expression log of pi ij by pi im equals to this the way we have written one thing is very clear.

That the relative probability of what I can write that relative probability of jth choice over mth choice is proportional to subject's value on the proportional to the difference $\left(x_{ij} - x_{im}\right)$ to the difference between subjects perceived value on the regressors. Who are the subjects here? Subjects are the individuals who are basically selecting the modes of choice and each subject he or she has a perceived value on travel time, travel cost so on and so forth.

When the individual is selecting jth choice, so relative probability of selecting the jth mode of travel over the base category mth mode of travel is basically the individual will calculate what is the difference in travel time between jth choice and mth choice. Secondly the individual will also calculate what is the difference in travel cost of jth mode of travel and mth mode of travel? So, this relative probability is always proportional to the difference.

Difference between subjects perceived value on the regressors**.** And the difference is being weighted by the estimated coefficient. Each individual they have a perceived value of this explanatory variable. Each individual is thinking if I take the air mode of travel what would be my travel time? What is my travel cost? What is my terminal waiting time? And the individual is comparing these values for the base category mode of travel let us say car.

The difference between air and car, train and car, bus and car and bus and car will determine the relative probability of a particular mode over the base category that is what we wanted to say. Now what we will do? We will take an example; this particular example is based on a hypothetical data on 4 modes of travel that the individuals are selecting. And that 4 modes of travel how the individuals are selecting based on the travels or mood specific feature. **(Video Starts: 18:27)**

So, we will take one dataset. So, if you look at these are the choices individuals are making. And look at this there are mode of travel air, air, train, bus and car, 4 modes of travel available and then these are the explanatory variable, what is the terminal, waiting time then in vehicle cost, then travel time, then travel cost. So, travel cost is basically different from in vehicle cost, this travel cost actually includes total that means this is basically in vehicle cost plus some opportunity cost.

And this is individual specific data income and this is party size. Party size means when you are traveling then from the same family how many members are traveling? It is assumed that if the size of the travel is more far from a particular family than the family will take basically the least cost mode of travel. So, we will take this data, total 840 observations are there we will copy this and then we will put it in stata. In stata how will you do that?

You have to go to stata editor and then I will paste it here and then stata will always ask you do you want to treat the first row as variable names? We have to say yes, so this is now copied in our dataset. So, now in conditional logit model what we first do? We will first specify the data as this is conditional logit and what is the command for that? You have to put a specific command cmset id and mode.

So, it shows that command cmset is unrecognized, so you have to install ssc, install cmset. So, now if you put cmset id and mode, it will take some time, since this is not connected with the internet we are not able to install the cmset but if you have a licensed version then we can actually work with the cmset command, not a problem. We have another command which is without specifying this we can basically put this command c logit, c is for conditional, c logit.

And notice your dependent variable; dependent variable is basically your choice. Then you have terminal time, you have in vehicle cost, you have travel time, travel cost and then you have to include what is your group variable here. So, you have to say group, what is your group? Group is mode. So, this is how you can actually you can actually estimate the multinomial logit model.

Now as I told you earlier then estimating the model is very simple but the problem here is how will you interpret the coefficient. Look at here terminal time is negative but significant, what is the meaning of this? So, let us put a let me see whether in this version we can, here we are not able to select the base category also. So, if we do not specify any base here then stata will automatically take the last category as the base.

Here what is our last category? 1, 2, 3, 4, fourth category which is car that is basically is taken as the base category. And then after running this regression let us look at the sign of each and every variable, all the variables are highly significant here. So, what we can do? We can just now try to explain, let us say terminal time which is negative. So, that means if terminal time that means if waiting time at the terminal for a particular category is increasing.

Then probability of selecting that mode of transport compared to the base category goes down, there is negative relationship. But once again you have to clearly keep in mind that we cannot take this as marginal effect. Because always we have to go back and see what is the model we are estimating? This is the model, so our model dependent variable is relative probability $log\left(\frac{\Pi_{ij}}{\Pi_{im}}\right)$.

So, that means we cannot say that if travel time increases by 1 unit the relative probability of selecting jth mode of transport compared to the mth mode of transport which is the base category

goes down by 0.11 you need. Because that is in conditional in any type of qualitative response model thus estimates they do not give you a direct measure of marginal effect.

Because if you differentiate this, this is our function

$$\Pi_{ij} = \frac{e^{\alpha+\beta x_i j}}{\sum_{m=1}^{j} e^{\alpha+\beta x_i m}}$$

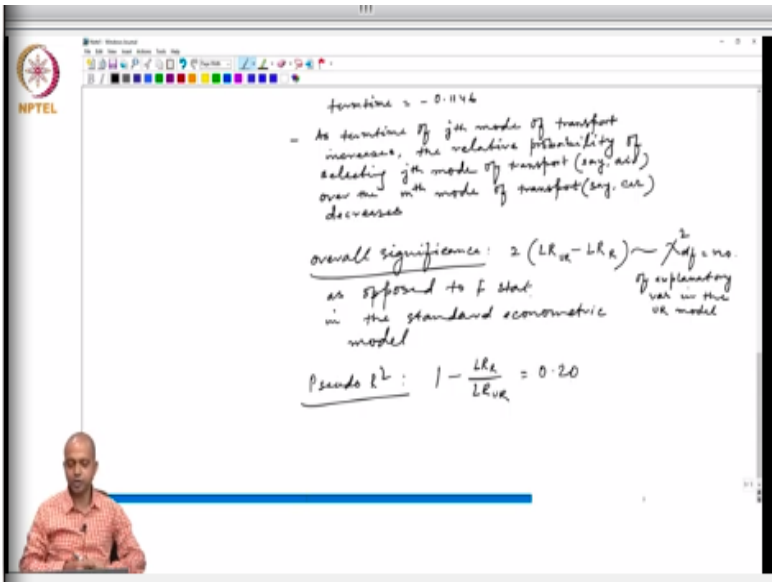
So, what basically if you want

$$\frac{d\Pi_{ij}}{dx_{ij}} \neq \hat{\beta}$$

This is clear from this expression itself that is why we said that in qualitative response model we will never take this β as a direct measure of marginal effect. So, we will try to interpret the results only in terms of whether there is a negative relationship or positive relationship between the log odds ratio.

That means log of the relative probability of selecting jth mode of transport over mth mode of transport that is what we will take. So, that is why when we get this type of result what we will interpret? The interpretation is that as terminal time increases, so as terminal time increases for a particular mode of transport so terminal time is - -0.11, so how will you interpret this coefficient?

(Refer Slide Time: 28:38)



So, terminal time = -0.1146, so what is the interpretation? The interpretation would be as term time of jth mode of transport increases the say air relative probability of selecting jth mode of transport over the mth mode of transport say mth mode of transport is let us say car decreases, that is how we will interpret. As the terminal time of jth mode of transport increases the relative probability of selecting jth mode of transport over the mth mode of transport decreases.

Similarly, if we go back as in vehicle cost increases then the relative probability of jth mode of transport over the base category mth mode of transport also decreases. As the travel time of jth mode of transport increases relative probability of selecting the jth mode of transport over the base category let us say which is car decreases. So, while the first 3 variables all are having expected sign the fourth one is little problematic to explain because this is based on hypothetical data.

The travel cost says that means which is the in-vehicle cost plus your opportunity cost, as opportunity cost of a particular mode of transport increases. Well, that means when I am selecting air, I am calculating the opportunity cost of what I am sacrificing to take that air mode of transport, opportunity cost is calculated with reference to the base category. So, when a passenger is taking air mode of transport, passenger is thinking what is the opportunity cost of traveling air in terms of car?

Probably I will save some amount of money; probably I have to wait less amount of time in the airport so on and so forth. So, as opportunity cost increases, opportunity cost of a particular mode of transport increases probabilities of selecting that mode of transport should not actually increase. Travel costs should also come up with a negative sign but since this is based on hypothetical data that is also coming at as positive.

Now as I said in this command when I am putting c logit command to estimate the conditional logit model this is the c logit command as you know. The c logit after that your dependent variable then all independent variables and then you have to make the group id, what is the group id? That is mode actually, you have different mode, so stata will decide that means when I am saying conditional logit model what are the options?

Option is defined in terms of different mode of transport. So, here stata will automatically consider the last mode as the base category but if we could estimate the model using that cmset command unfortunately which is not available in this version. So, cmset will allow us to select the base category according to our own requirement. According to our own requirement you can select category 1 as base, category 2 as base, category 3, 4, everything you can select as your base category.

So, this is all about the interpretation of the coefficient and other measures are again same like LR chi square which is basically defined as, how do you define LR chi square?

$$LR\ Stat = 2(L1 - L0)$$

So, what is the unrestricted, is the final one, -361 and the restricted one is basically -394. So, this minus this into 2 is this value. Restricted means when you have no explanatory variable included in your model, unrestricted means when you have all the variables included in on your model.

So, that means basically we are trying to understand the overall significance of the model. Overall significance of the model is always given by LR chi square, so that also we should mention somewhere. **(Video Ends: 35:18)** That overall significance generally is given by F statistic, let me

$$LR\ Stat = 2(L1 - L0)$$

restricted as opposed to this will follow chi square distribution with degrees of freedom equals to number of explanatory variable in the unrestricted model that is LR chi square.

So, if you take this, so this would become 2 into what is the LR unrestricted? LR unrestricted is basically -361.41 minus and another minus will come +394 I think, 394.61. So, if you do that then what you will get? This value 181, so this is how you have to basically calculate this. So, I am not writing this value, so let us say this is the chi square value, you have to get the overall significance of the model by this as opposed to F stat in the standard econometric model. So, overall significance is given by LR chi square.

And then you have also another measure called pseudo R^2, pseudo R^2 is 0.20. So, pseudo R^2 which is again divided by

$$R^2 = 1 - \frac{LR_R}{LR_{UR}}$$

That is given by the pseudo R^2 which is 0.02 in this particular case. So, this thing we have to keep in mind. So, now based on the result what you can see that the LR chi square value is, what is the corresponding probability? It is 0.000.

So, that means which is highly significant, so model is overall significant. And that is true also see if you look at the individual variables all these explanatory variables are significant at 1 percent level. So, when variables are significant at 1 percent level individually there is no point in thinking that they would be insignificant at the model, overall. So, this is how we have to basically select the command to estimate the model.

And lastly you have to put the MFX command for getting, even the MFX command or the margins if I put margins, even the margin command is also not working properly in this version of stata but it is available in the higher version of this. Actually, the marginal effect is little complicated in this model. So, up to this if we can understand that itself is enough for our requirement.

We need to basically understand how to estimate the model and interpret the coefficient and then checking the overall significance of the model and the goodness of it measure. In short I will once again repeat that this is a conditional logit model which is basically the chooser's choice specific data, so that means no individual specific data is included here. So, if we think that an individual will select which mode of transport that depends on the trans mode specific information travel time, cost so on and so forth.

Then this is the appropriate model. And if we think the individual will select the mode of transport based on the individual specific data then basically multinomial logit model is the correct option. But in reality you might be thinking actually whether the individual will take bus, train, air or car that basically depends on mood specific as well as individual specific data. Because we all know if we go for air then travel time is very less, I can reach there on the same day.

But just because I can reach their same day if an individual does not have enough income to afford for that mode of transport, then it is basically meaningless. So, individual will consider the individual specific data, age, income, education as well as the choice specific data travel time, comfort, waiting time, hygiene so on and so forth then actually decide. So, that means a real model.

A real-life scenario should include actually chooser's specific as well as choice specific information which is basically the nested multinomial logit model or mixed multinomial logit model, wherein we will include both chooser's specific as well as choice specific information which we will discuss next. **(Video Ends: 43:21)**