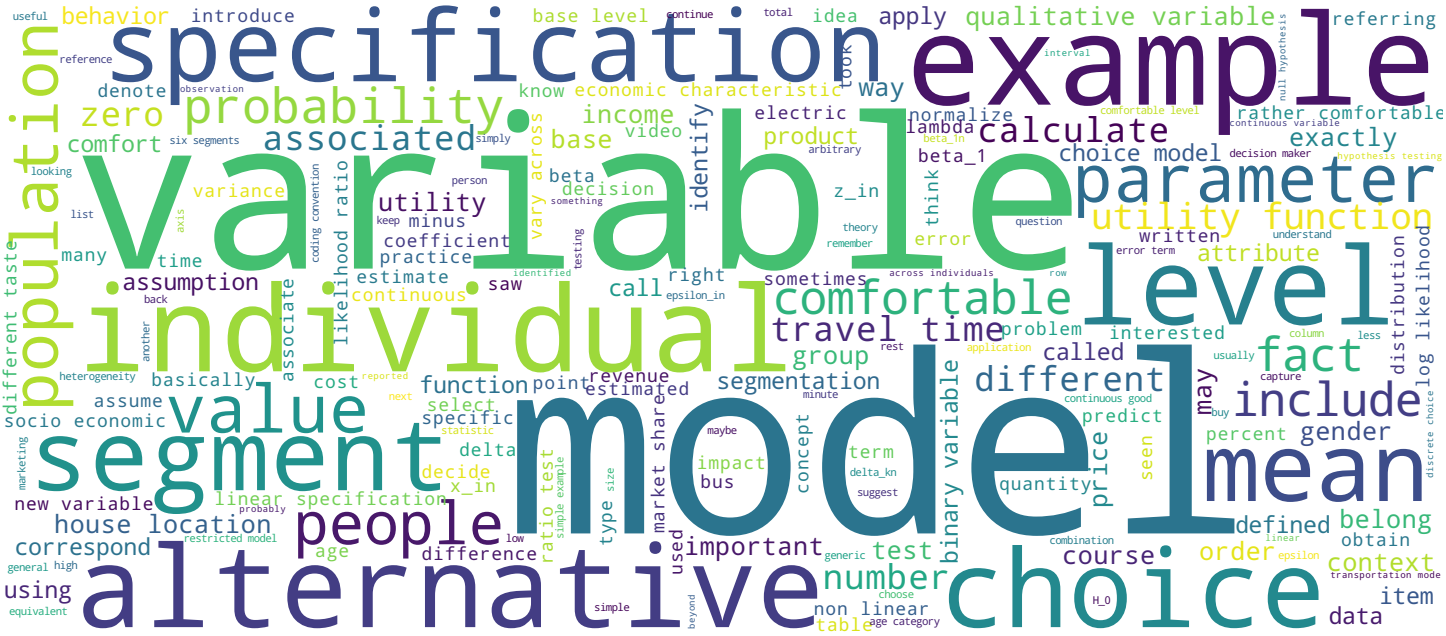
Choice with multiple alternatives

Specification of the deterministic part

Michel Bierlaire

Introduction to choice models

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE



Search MOOC

Video

EPFL

Qualitative explanatory variables

Hello. In this video, we will now look into the qualitative variables. We would like to model these qualitative variables and include them in the choice model. And we will see how to do that.

Notes

Summary

0m 04s

Qualitative attributes

Examples

- Level of comfort for the train
- Reliability of the bus
- Color
- Shape
- etc...

In many circumstances, they influence the choice a lot. So, in the context of transportation mode choice, for example, we can think of the comfort of the train, we can think about the reliability of the bus. If you are interested in marketing, if you want to know how people choose one item or the next, it may depend on the shape, or the color of the box. And so on. So all these concepts are qualitative. But they are very important. So we need to be able to include them into the model.

Notes

Summary

0m 16s

Modeling

Identify all possible levels of the variable

- Very comfortable,
- Comfortable,
- Rather comfortable,
- Not comfortable.

Select a base level

- Very comfortable,
- Comfortable,
- Rather comfortable,
- Not comfortable.

In order to see how to model a qualitative variable, let's take the example of comfort. So we consider the variable "comfort". The first thing to do is to identify all possible levels that this variable can take. So we consider it as a discrete variable. In this example, we can consider four levels: very comfortable, comfortable, rather comfortable and not comfortable. Once we have identified these levels, we need to select one level which is considered as the reference. We call it the base level. In this example, let's consider "very comfortable" as the base level.

Notes

Summary

0m 48s

Modeling

Introduce a 0/1 attribute for all levels except the base case

- z_c for comfortable
- z_{rc} for rather comfortable
- z_{nc} for not comfortable

	z_c	z_{rc}	z_{nc}
very comfortable	0	0	0
comfortable	1	0	0
rather comfortable	0	1	0
not comfortable	0	0	1

If a qualitative attribute has K levels, we introduce $K-1$ binary variables (0/1) in the model

Now we need to make the mathematical modeling of this variable. To do this, we introduce an attribute, which is a binary variable, binary meaning that it can take the value zero or one, for all the levels, except for the base case. So in this case, in the example of the comfort, we had four levels, we took "very comfortable" as the base case, so we include z_c for "comfortable", z_rc for "rather comfortable" and z_nc for "not comfortable". And we associate with these variables a coding convention. This coding convention is reported in the table on the right. When all the variables are zero, it means that it corresponds to the base case "very comfortable". And when one of these variables is one, then it means that the qualitative variable is associated with the corresponding level. So if z_c is one, it means that we are referring to "comfortable". If it is z_rc, it means that we are referring to "rather comfortable". And if it is z_nc, it means that we are referring to "not comfortable". So in general, if a qualitative attribute has K levels, then we introduce K-1 binary variables in the model.

Notes

Summary

1m 28s

Modeling

Utility function

$$V_{in} = \beta_c z_c + \beta_{rc} z_{rc} + \beta_{nc} z_{nc} + \cdots$$

Note

The choice of the base level is arbitrary.

After we have defined these binary variables, we can include them in the utility function exactly as we did with the quantitative variables. So in this example, the utility function will be defined as beta_c times z_c, plus beta_rc times z_rc, plus beta_nc times z_nc. Note that the choice of the base level is arbitrary. And I let you do as an exercise the comparison of a model where you normalize the "very comfortable" level, with another model where you normalize for example the "comfortable" level as a base case. They are equivalent. And this can be verified easily.

Notes

Summary

2m 49s

Qualitative characteristics

Examples

- Sex
- Education
- Professional status
- etc.

In terms of socio-economic characteristics, there are also plenty of them who are qualitative: sex, education, professional status, and so on, are qualitative variables. So we need to be able to model them in the same way as we did for the attributes.

Notes

Summary

3m 31s

Modeling heterogeneity

Behavioral assumption

- Individuals have different taste parameters.
- The difference is explained by a qualitative socio-economic characteristic.

$$V_{in} = \beta_{1n} z_{in} + \cdots$$

where

$$\beta_{1n} = \beta_{1n}(\text{education}_n).$$

So, as we did before, the beta coefficient now would vary across individuals. In the continuous case, we took the example that beta_1 would vary with income, which was a continuous variable. Here we would like to model the fact that beta one may vary across education levels, for example. The level of education. How do we do that?

Notes

Summary

3m 48s

Modeling heterogeneity

Segmentation

- Assume that there are K levels for the qualitative variable (e.g. education).
- They characterize K segments in the population.
- Define

$$\delta_{kn} = \begin{cases} 1 & \text{if individual } n \text{ is associated with level } k \\ 0 & \text{otherwise} \end{cases}$$

- Introduce a parameter β_1^k for each level and define

$$\beta_{1n} = \sum_{k=1}^{K} \beta_1^k \delta_{kn}$$

This is called segmentation. So we assume that there are K levels for the qualitative variable (so, let's say education, we have like three levels of education), and each level is associated with a segment in the population. And what is important, is that, for each individual, we need to identify which segment they belong to. So segmentation is a partition of the population. Each individual should belong to exactly one segment. Based on this segmentation, we can now define binary variables, that I denote here by delta_kn, which is equal to one if individual n belongs to segment k, and zero otherwise. And now, what we do, is that we introduce a different coefficient, a different parameter, beta_1k, for each level, for each segment in the population. And now, the beta_1n that we would like to include in our specification, can be written as the sum over all levels of beta_1k times delta_kn. And because each individual belongs to exactly one segment, this sum will include only one non zero term. So basically, this sum will pick the beta which corresponds to the segment, and include it in the utility function.

Notes

Summary

4m 10s

Modeling heterogeneity

Segmentation

$$V_{in} = \beta_{1n} z_{in} + \cdots = \sum_{k=1}^{K} \beta_1^k \delta_{kn} z_{in} + \cdots = \sum_{k=1}^{K} \beta_1^k x_{ink} + \cdots$$

where

$$x_{ink} = \delta_{kn} z_{in}$$

So we have the first specification beta_1n z_in. We replace beta_1n by the sum that we have defined. An we obtain this specification. And like we did before, for the quantitative attributes, we interpret this specification as the definition of a new variable. So we define a new variable x_ink which is the product of delta_kn and z_in. So, basically x_ink will be equal to z_in for individuals who belong to the segment k, and will be zero for all other individuals. And, therefore, this segmentation can be written as a classical linear-in-parameters specification.

Notes

Summary

5m 26s

Segmentation with several variables

Example

- Gender (M,F)
- House location (metro, suburb, perimeter areas)
- 6 segments: (M, m), (M, s), (M, p), (F, m), (F, s), (F, p).

Clearly, you can generalize this with more than one variable. So let's take this example where I consider that people behave differently depending on their gender, and depending on their house location. So we have two levels for the gender, and we have in this example three levels for house location: the metropolitan area, the suburbs, and the perimeter areas. The combination of gender and house location identifies six segments in the population, which correspond to the six possible combinations of gender and house location: two genders, three house locations equal six segments. And the rest is exactly the same as before.

Notes

Summary

6m 14s

Segmentation

Specification

$$\beta_{M,m} TT_{M,m} + \beta_{M,s} TT_{M,s} + \beta_{M,p} TT_{M,p} + \\ \beta_{F,m} TT_{F,m} + \beta_{F,s} TT_{F,s} + \beta_{F,p} TT_{F,p} +$$

$TT_i = TT$ if indiv. belongs to segment i, and 0 otherwise

Remarks

- For a given individual, exactly one of these terms is non zero.
- The number of segments grows exponentially with the number of variables.

We define this variable, let's say travel time in this example, for each of the segments. And this is the product of the travel time and the delta, the delta that identifies the membership of a segment. So each of these TT will be zero for people who are not in the corresponding segment, and the value of travel time for those who are in the segment. So this the exact same specification as we saw before. The only difference is that we have segmented the population based on two variables instead of one. Well, of course, you can continue. You can base your segmentation on three, four, five variables. The problem is that then, the number of segments will grow exponentially with the number of variables. And, at some point, it will be too high. You will have too many parameters. And, probably, you will not have enough data in each of the segments in order to be able to identify the coefficients. So it may be useful to find another specification that captures the heterogeneity in a different way.

Notes

Summary

6m 58s

Segmentation

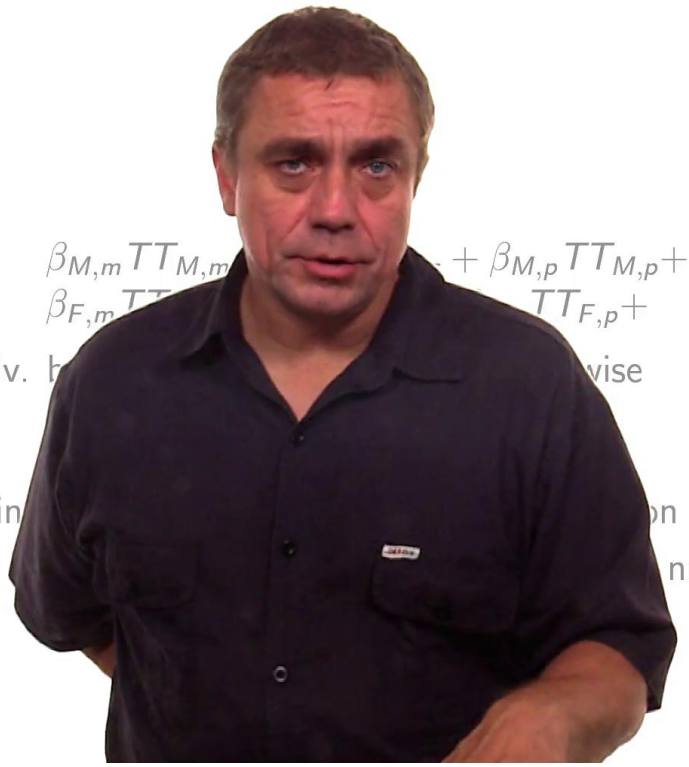
Specification

$\beta_{M,m} TT_{M,m}$ $+ \beta_{M,p} TT_{M,p}+$
$\beta_{F,m} TT$ $TT_{F,p}+$

$TT_i = TT$ if indiv. h vise

Remarks

- For a given in n zero.
- The number number of variables.



At this point, we have investigated a lot of specifications that are linear in parameters. We were able to model quite complex phenomena: the fact that a variable can be generic or specific, and the fact that people have different tastes. So we have seen our to segment the population, and to model the fact that different people, with different types of socio-economic characteristics, have different taste parameters. This linear specification is extremely powerful. But, sometimes, we need to go beyond. And we need to calculate a model with a non linear specification. This is what we will see in the next video. But before we do that, I suggest that you do some practice quiz to be more familiar with this specification.

Notes

Summary

8m 04s