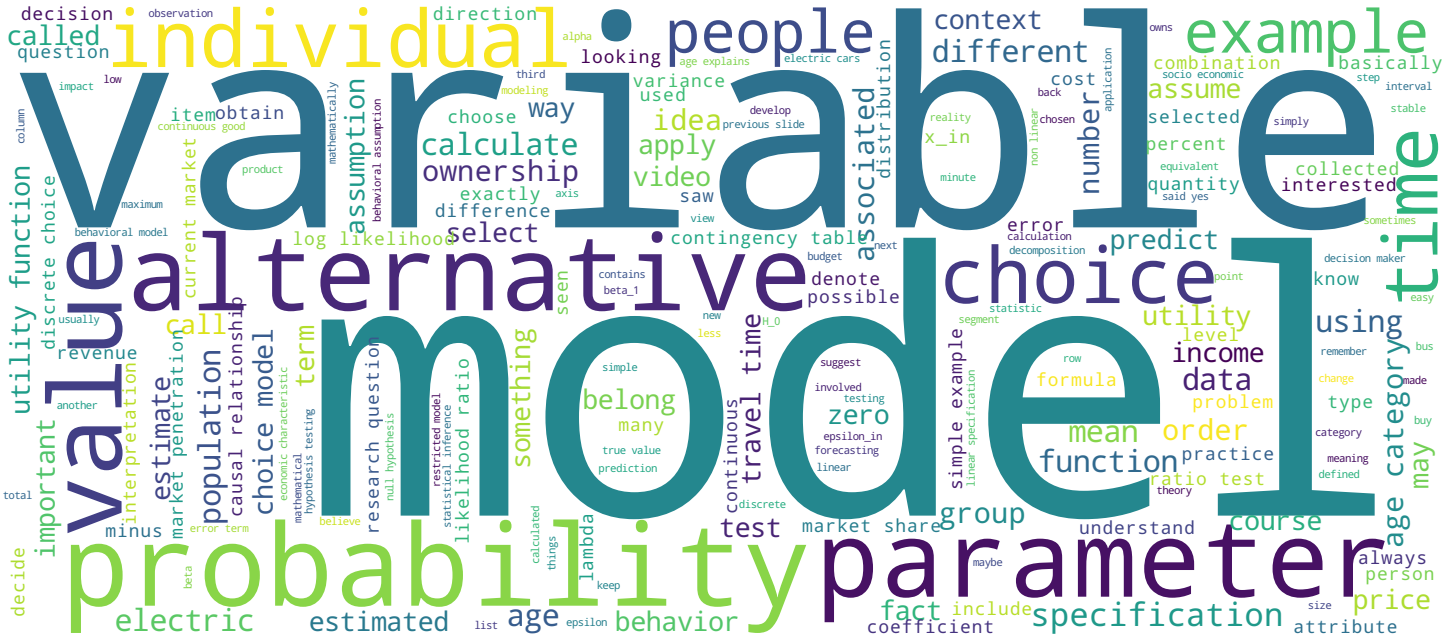
Introduction to behavior modeling

Simple example

Michel Bierlaire

Introduction to choice models

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE



Search MOOC

Video

EPFL

Simple example: developing a model

We are analyzing a simple example in order to understand the steps that are necessary to develop a discrete choice model. And we saw before that using the responses to a questionnaire, and using a contingency table, we were able to respond to the first research question, which was: "what is the current market..." "... penetration of electric cars in our population." But the second research question, which is about prediction, what will be the future market penetration, requires a model. Here, we mean a mathematical model. A model is a simplification of the reality that allows us to manipulate it and to make calculations about it. That's what we will do in this video.

Notes

Summary

0m 04s

Model

Variables

- i: status of electric car ownership (yes or no)
- k: age category (20–39, 40–64 or 65+)

First, we need to define variables. In this example, there are two variables that are involved. The first variable is the status of electric car ownership. This variable, that we will denote by i, can take two values. Yes or no. The second variable characterizes the age category. This variable, denoted by k, can take three values: either 20-39, 40-64 or 65+.

Notes

Summary

0m 50s

Model

Decomposition

$$P(i,k) = P(i|k)P(k) = P(k|i)P(i)$$

Interpretation

- $P(i|k)$: age explains electric car ownership
- $P(k|i)$: electric car ownership explains age

So what we get from the data is the combination of the value of i and the combination of the value of k. Our model would like to predict that. What is the probability that a given person in the population has this configuration i and k, meaning a given car ownership status, and a given age category. This probability can be decomposed. And actually, using the definition of probability, we can decompose it in two different ways. We can say that the probability of i and k is the probability of i given k, times the probability of k. Or, we can do it in the other direction. The probability of k given i, times the probability of i. So if these two decompositions are equivalent from a mathematical point of view, they are not the same in terms of interpretation. And, actually, and this is important, because we will always make the link between the mathematical formulation and the behavioral interpretation. So if you look at the interpretation of the two conditional models, P(i|k) and P(k|i), we have that the first basically explains the car ownership as a function of age. So age explains the electric car ownership. While the second one is exactly the opposite. The electric car ownership explains the age.

Notes

Summary

1m 24s

Model

Decomposition

$$P(i,k) = P(i|k)P(k) = P(k|i)P(i)$$

Interpretation

- $P(i|k)$: age explains electric car ownership
- $P(k|i)$: electric car ownership explains age

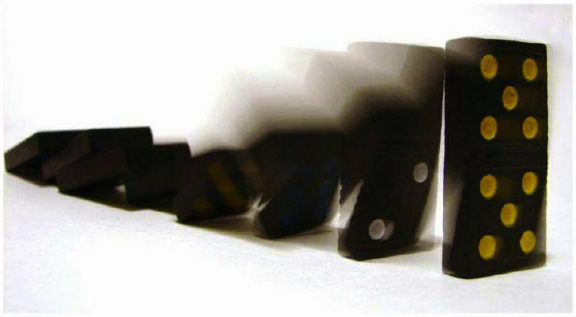
And from a behavioral point of view, although I am happy to assume the first model, the fact that car ownership can be explained by age, I have a hard time believing the second model, that the age of a person can be predicted by looking if he owns or not an electric car. So this is something which is very important. Even if mathematically we can write the model both ways, most of the time, only one of the two directions is actually making sense in terms of cause to consequence. And here, the causal model that I believe is the most meaningful is when age explains electric car ownership. So this is what I will call a behavioral model.

Notes

Summary

2m 48s

Model



Model

- identify stable causal relationships between the variables
- stability over time necessary to forecast
- here: we select $P(i|k)$ as an acceptable behavioral model

And because I want to predict, I would like this causal relationship to be stable over time. And this is something very important. Because when you predict, you have to assume that something does not change. That's what we need to be forecasting. So that's what we do here. We will select P(i|k), so explaining electric car ownership as a function of age, as an acceptable behavioral model which is stable over time. That's the behavioral assumption that we will make here. And then, we will select the first decomposition that we have seen in the previous slide.

Notes

Summary

3m 34s

Model

Specification

$$
\begin{aligned}
P(i = \text{yes} \mid k = 20\text{–}39) &= \pi_1, \\
P(i = \text{yes} \mid k = 40\text{–}64) &= \pi_2, \\
P(i = \text{yes} \mid k = 65+) &= \pi_3.
\end{aligned}
$$

Parameters

- π_1, π_2, π_3
- unknown
- must be estimated from data

So now we need to specify the model, to be more explicit. There are three things that we need to calculate: what is the probability that I own an electric car given that I belong to the first category, given that I belong to the second category, and given that I belong to the third category. The specification that we choose here is extremely simple. We simply associate a parameter to each of these values. So pi_1 is associated with this first probability, pi_2 the second one and pi_3 to the third one. And the idea is that these parameters are not known a priori, but must be estimated from the data. The model is specified. We have selected a behavioral causal relationship, we have associated each of them with a parameter, and the idea is that these parameters must be estimated from data.

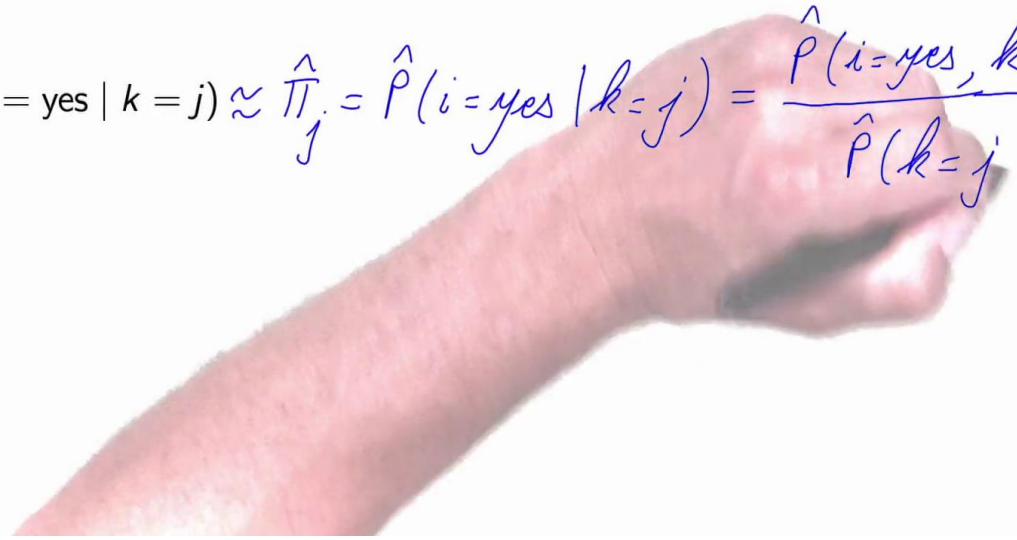
Notes

Summary

4m 10s

Model estimation

$$\pi_j = P(i = \text{yes} \mid k = j) \approx \hat{\pi}_j = \hat{P}(i = \text{yes} \mid k = j) = \frac{\hat{P}(i = \text{yes}, k = j)}{\hat{P}(k = j)}$$



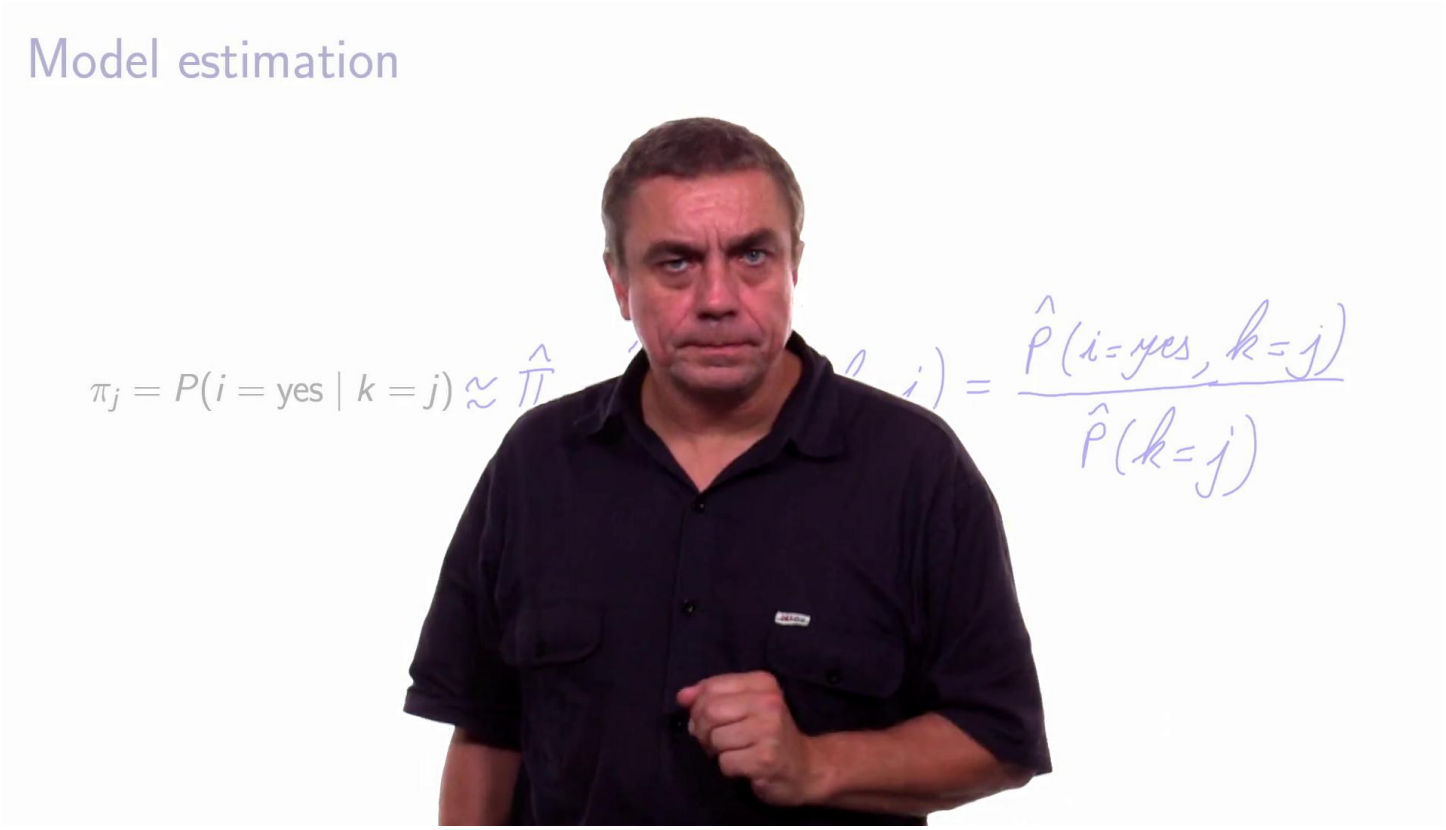
In order to calculate these parameters, we will use the same idea that we did for the current market shares: we will use statistical inference using the sample that we have collected. So we will approximate the true value of the parameter pi_j by an estimate pi hat j, which is equal to the proportion of people in the sample who have said "yes" given that they belong to age category j. And this can be calculated as the ratio between the number of people who said "yes" to the first question and j to the age category, divided by the total persons who are in age category j. And these two values can be actually collected from the contingency table.

Notes

Summary

5m 00s

Model estimation



And now I suggest that you calculate the estimates of these three parameters, pi_1, pi_2, pi_3 from the contingency table, using this formula. It's very easy, you will see.

Notes

Summary

5m 52s