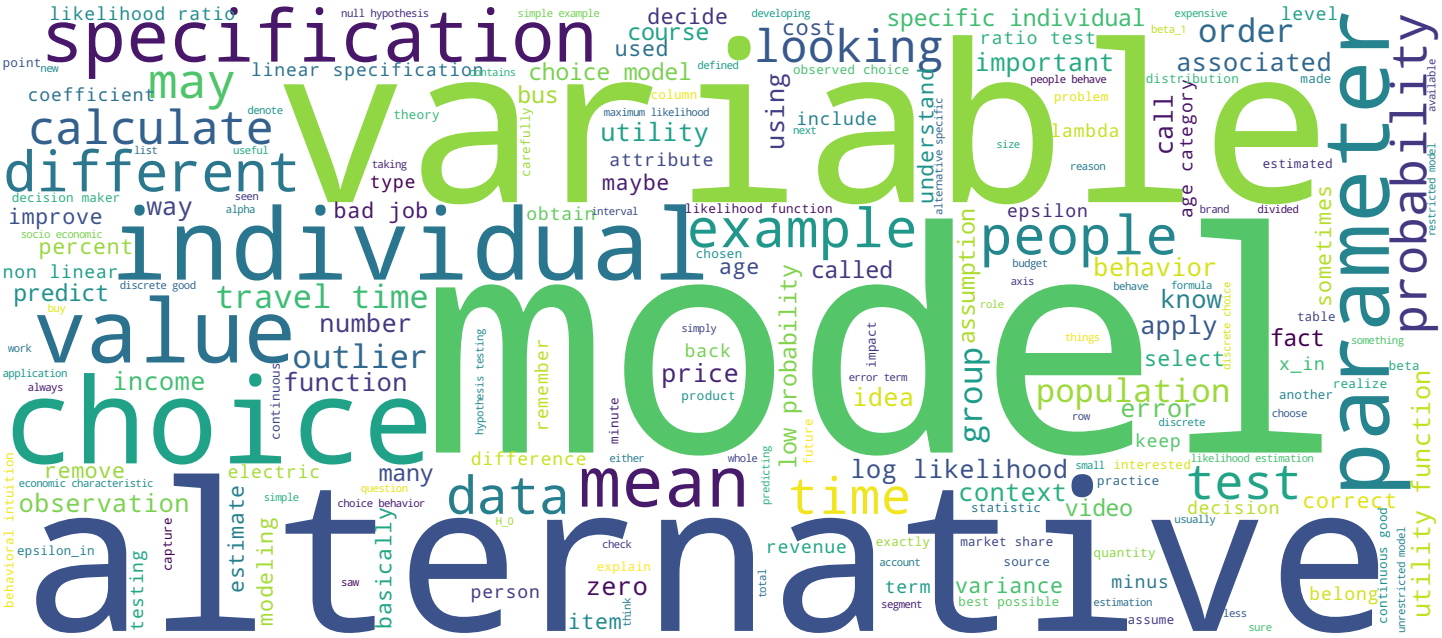
Testing

Prediction tests

Michel Bierlaire

Introduction to choice models

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE



Search MOOC

Video

EPFL

Outlier analysis

One of the main applications of our choice models is to be able to predict the behavior in the future. So one of the things we can use for testing the model is to first check that it is able to predict what we have already observed. Well it should do a good job in doing this because we have used maximum likelihood estimation. The parameters that we have used are the best possible in order to replicate the observed choices. But still, even if the model is doing the best possible job for the whole sample, there may be some specific individuals in the sample such that the model is doing a bad job in predicting their choices. This is what we call the "outliers". And actually, by looking at these outliers, we may actually get insight in why the model is not doing well. This is what we will do in this video.

Notes

Summary

0m 04s

Outlier analysis

Procedure

- Apply the model on the sample
- Examine observations where the predicted probability is the smallest for the observed choice
- Test model sensitivity to outliers, as a small probability has a significant impact on the log likelihood
- Potential causes of low probability:
  - Coding or measurement error in the data
  - Model misspecification
  - Inexplicable variation in choice behavior

So the procedure of this outlier analysis is that, you take your estimated model, and then you apply the model on the sample. It means that, basically, you calculate the probability predicted by the model, that the observed choice is actually made. And these are actually the components of the log likelihood function. And now, you sort them and you look at those for which this predicted probability is small. It means that you are looking at the specific individuals for which the model is doing a bad job. This is what we call the outliers. And then you want to test the model sensitivity to outliers. Suppose that you remove them from the sample. Does it affect the model a lot? And if so, then, there is a warning. Because these people are associated with a low probability in the log likelihood function, they have a significant impact on this log likelihood. So, it's important to test the sensitivity of the model to these outliers. What are the potential causes of low probability? There are many of them, but let's look into three of them: coding errors, model misspecification, and some inexplicable variation of choice behavior. Let's go through them one by one.

Notes

Summary

0m 53s

Coding or measurement error in the data

Look for signs of data errors

- Travel time is negative
- Number is coded as a string
- etc.

Correct or remove the observation

- Go back to the original survey
- Correct only if you are certain

When you look more carefully at the data, associated with these specific individuals, the outliers, those for which the model is doing a bad job, sometimes you realize that there are some mistakes in the coding of the data. You realize that a travel time was coded using a negative number, or maybe a number was coded as a string, and therefore the software interprets it as zero, or you know what. So, you look for errors in the data file. It may happen. In this case, either you can correct the error, because you can go back to the survey, and you know what is the value of the travel time and then you can correct it. But if you cannot correct it, it means that this data is corrupted, and then you remove it from the sample. So, in general, correct it only if you are sure about what you are doing. Because you don't want to add measurement errors in the data.

Notes

Summary

2m 04s

Model misspecification

Improve the specification

- Seek clues of missing variables from the observation.
- Why is the model associating such a low probability for this choice?
- Did we forget to account for age, income, or any other variable ?
- Should a nonlinear specification be investigated?
- Use a behavioral intuition.

The second source of low probability prediction, the second source of outliers, is that the model has specification problems. And then you want to improve the specification. So the idea is to look carefully at the variables of these individuals. You are looking for hints, for clues of possibly missing variables, by looking at the observation. The idea is to say: OK, look. My model is telling me that there is a high probability to select the bus, but this person is taking the car. And the car has a very low probability in my model to be selected. Maybe he belongs to an age category or an income category where the behavior is different, and we may want to account for this heterogeneity. Sometimes, you can also say: oh, but maybe a non linear specification would be useful. And this is where you use your behavioral intuition. By looking at specific individuals, in specific contexts, you try to understand why the model is performing badly in this context, and you use your behavioral intuition to improve the specification. So this is more of an art then a science, but this is the most exciting part of modeling. It's to try to improve the model in order to explain better the observed behavior.

Notes

Summary

2m 57s

Inexplicable variation in choice behavior

Keep the observation

- If no acceptable explanation is found, keep the observation.
- Avoid overfitting of the model to the data.
- The model should reflect how people behave, not how they should behave.

But sometimes, you cannot do that. Sometimes, you look at it, you look at all the variables and you do not understand. Why is this guy taking the bus? There is no reason. The bus is more expensive. It's slower. All the variables that you could think of are not explaining why this alternative has been chosen. Well, in this case, what you do is that you give up. But you keep the observation. This is extremely important. It's not because your model is not able to explain the choice of this individual that you should remove the observation from the sample. This is very important. Because, if you do so, you will overfit the model to the data. So, remember that the models that we are developing should reflect how people behave, not how we would like them to behave. And, in the model, the role of the epsilon is exactly designed to capture these discrepancies between the modeling assumptions and the actual behavior. So we definitely keep the observation, and we just live with it.

Notes

Summary

4m 20s