

Introduction to Latent Class Analysis

Exercises

We are going to use Mplus to conduct some analyses in a dataset. The dataset is provided with the exercises and it is called:

- `ess_ex1.dat`

In what follows, I will describe the Dataset. Two sets of exercises are then provided. Solutions are available in separate files.

The Dataset

This is an extract of some variables from the European Social Survey (Round 7). Many variables have been transformed before analyses. The key variables we are going to use in this exercise concern respondents' views about the possibility of influencing politics. The original variables names have been changed adding an "r" in front, to indicate variables have been recoded. These are the variables we will be using in the exercises:

rpspsgv:	"Political system allows people to have a say in what government does"
ractrolg:	"Able to take active role in political group"
rpsppi:	"Political system allows people to have influence on politics"
rcptppol:	"Confident in own ability to participate in politics"
rptcpplt:	"Politicians care what people think"
retapapl:	"Easy to take part in politics"

The original responses ranged from 0 "Not at all" , to 10 "Completely". For simplicity, all the items have been recoded into 3 categories (**0; 1; 2**), whereby:

- **0** includes the scores originally ranging from 0 to 3;
- **1** the scores originally ranging from 4 to 6;
- **2** the scores originally ranging from 7 to 10.

These variables are therefore ordered categorical.

The dataset includes only respondents from 6 countries, Austria, Belgium, Switzerland, Spain, and France. Overall, there are 8,938 respondents in the dataset

Other variables in the dataset set are:

ess7id: A numerical ID.

cou: A variable indicating the respondent's country (1 = Austria [AT]; 2= Belgium [BE]; 3=Switzerland [CH]; 8=Spain [ES]; 10=France [FR]).

nuts1en: An ID to identify different NUTS1 areas.

nuts2en: An ID to identify different NUTS2 areas.

Exercise #1 (for Mplus beginners) Descriptive statistics

It is always better to check the Mplus file to make sure the software is reading it properly and nothing went wrong while creating the file you are going to use for analyses (whether you are creating using STATA, SPSS, etc.).

If you use STATA, the process of creating an MPLUS data file and a standard input file is facilitated by the stata2mplus command (see <http://www.ats.ucla.edu/stat/stata/faq/stata2mplus.htm>)

In the files provided with this “Introduction to LCA” you will find a **BASIC.inp** file.

Open it with Mplus (from the main menu FILE → Open:)

The file has already the variable names and the missing value indicator.

TASKS:

1. Specify the variables **cou rpsppsgv ractrolg rpsppi1 rcpttpol rptcpplt retapapl** as the variables to be included in analyses (HINT: “usevar” ...)
2. Specify which variables are categorical (at least for descriptive analyses)
3. Use Analysis: TYPE = BASIC; to invoke descriptive analyses.
4. How many observations are in the dataset? _____
5. What is the proportion of respondents from France (category 5)? _____
6. What is the proportion of respondents providing the lowest score to item **rcpttpol** ? _____
7. Are there missing data in any of these variables ? _____

Exercise #2

Latent class analysis

We are going to define a latent class variable with 2 classes based on responses to the **rspspsgv ractrolg rpsppi1 rcptppol rptcp1t retapap1** variables.

We can call the latent variable *class* (or indeed, any other name) using the option CLASS in the command VARIABLE:

```
CLASSES = class(2);
```

The command above will estimate a latent class with two categories

Remember to invoke a mixture model algorithm: you need to specify:

```
ANALYSIS: TYPE=mixture;
```

Often data are clustered (e.g. participants within neighbourhoods). In this case it is possible to adjust standard errors allowing for the clustering at the country level (variable “cou”). To do this, specify the clustering variable using option CLUSTER in VARIABLE: command. E.g.:

```
CLUSTER = cou;
```

while in the ANALYSIS command, include COMPLEX after TYPE. E.g.:

```
ANALYSIS: TYPE = mixture complex;
```

Make sure you include in the OUTPUT options :

TECH1 →report arrays containing parameter specifications and starting values for all free parameters in the model;

TECH10→reports univariate, bivariate and response pattern model fit information for the categorical dependent variables in the model;

TECH11→reports the Vu-Luong-Mendell-Rubin likelihood ratio of test model fit that compares the estimated model with a model with one less class;

SVALUES →reports the start values in the model, which can be useful for building model constraints later;

(TECH14→ reports the parametric bootstrapped likelihood ratio test that compares the estimated model to a model with one less class than the estimated model: However, this is not available when the COMPLEX type of analysis is invoked, for example to control for clustering at the country level.)

TASKS:

1. Using as a basis the previous “BASIC.inp” file, specify a 2-class latent variable model based on the **rspspsgv ractrolg rpsppi1 rcptppol rptcp1t retapap1** variables. Ensure that the variable **cou** (country) is used as a clustering variable and clustering is accounted for in the analyses.
2. Check the log-likelihood values of the final stage optimisation: does it appear as a trustworthy solution?
3. What are the proportions of individuals in the two classes based on the estimated model?
4. How can you interpret the two classes?

5. What does the Vuong-Lo-Mendell-Rubin Likelihood Ratio test indicate?
6. Now, run a model with 5 classes
7. Check the loglikelihood values of the solution: does it appear as a trustworthy solution?
8. Choose the number of starts and final stage interactions to be 1000 and 100 respectively, and check the loglikelihood values of the solution
9. Compare the information criteria, and other parameters of the 2- and 5-class solutions: which appears to provide a better fit?
10. Inspect the graph of estimated probabilities for conditional item responses of item category 3 in the 5-class solution.
11. Estimate models with the number of classes increasing from 1 to 8, and compare the models based on fit statistics, information criteria, entropy, and the Vuong-Lo-Mendell-Rubin Likelihood Ratio test. Which model would you select?
12. A parallel indicators model is one where the all the categories of response to an indicator have the same probabilities within a class. For example, if the latent classes were “Depression” vs. “No Depression”, a parallel model would assume that people in the “Depression” class will have the same probability of reporting the symptoms of depression, e.g. they will have the same probability of reporting “Low Mood”, “Lack of Pleasure”, “Sleep Problems”, etc. Try to specify a similar model for the latent class more likely to report lower scores across all indicators in the 4-class solution (the “Sceptics”). Compare this constrained model with the unconstrained one using a likelihood-ratio test. Should we accept or reject the constrained “parallel indicators” model?