An introduction to latent class analysis using Mplus

University of Ulster, Magee (18.11.11)

Dr. Orla McBride

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Writing Syntax for Mplus**

Some general notes

* One helpful feature in Mplus is the use of an exclamation mark (!) - Any text in Mplus input that is preceded by ! will turn to **GREEN** text and will not be read by Mplus as an input command/subcommand.
* This feature to useful for writing notes for yourself about the analysis you are conducting etc.
* This feature can also be used to exclude certain pieces of syntax you may not want Mplus to include (examples will be shown during this workshop).
* Every single line of text that you do not want to include as an input instruction must be preceded by an !

Mplus input instructions

* Mplus has several command headings that, when typed into an input file and followed by a colon (:), will turn **BLUE** automatically.
* The main command headings are: **TITLE, DATA, VARIABLES, ANALYSIS, MODEL, SAVEDATA, OUTPUT,** and **PLOT.**
* Mplus recognises these commands as key elements of the input to run the programme.
* Generally, all of these command headings are used when running a latent class model.
* Many of the command headings have subcommands, which are used to tell Mplus exactly how to run the model you are specifying.
* Chapter 15 in the Mplus User Guide Version 6 is very useful in helping you to familiarise yourself with the different commands and subcommands.
* Neither commands, nor subcommands, need to be in block capital letters - this is your choice. Block capital letters are generally used in the Mplus manual.
* As you get more experienced at using Mplus, most of the commands and subcommands can be abbreviated to 4 characters. This approach will not be used in this workshop.
* Every subcommand must end in a semi-colon (;) to enable Mplus to run. It is important to remember, however, that lines of syntax in Mplus are limited to approximately 80 characters so it may be necessary to spread your input for a specific subcommand over many lines. A semi-colon is not required at the end of each line, it is required at the end of the entire subcommand (examples will be shown during this workshop).
* Generally, the terms ‘are’ ‘is’ and ‘=’ for writing subcommands can be used interchangeably.

Specific details for the various commands and subcommands used for a latent class model in Mplus Version 6

| **Command** | **Subcommand(s)** | **Information** |
| --- | --- | --- |
| **TITLE:** | None | The **TITLE** command is optional. There are no subcommands for this command. Writing a **TITLE** is useful to help you record what type of survey data you have used, the type of analysis you are conducting, what subsample you have used, etc. The **TITLE** you enter does not need to end in a semi-colon and can be anything of your own choice. |
| **DATA:** |  | The **DATA** command is used to provide information about the data set to be analysed. |
|  | FILE IS | Under the **DATA** command, you will need to include the subcommand 'FILE IS’. This basic instruction tells Mplus where the data file you want to analyse is located.  For example, if your data is on the C drive, the instruction would be FILE IS C:\FOLDERNAME\FILENAME.DAT;  Mplus will read ASCII and DAT files. It is easiest to save your SPSS or STATA file as a .DAT file - if you do this, the 'FILE IS' subcommand is sufficient to run your analysis.  If you save your data as an ASCII file, you will need to include an additional 'FORMAT IS' subcommand - use this if you are more comfortable with Mplus or if you have a very large data file. When analysing a very large data file, Mplus will find it easier to read data with a ‘FORMAT IS’ subcommand. |
| **VARIABLE:** |  | The **VARIABLE** command is used to provide information about the variables in the data file to be analysed. The **VARIABLE** command has options for naming and describing the variables in the data file to be analysed, subsetting the data file on observations, subsetting the data file on variables, and specifying missing values for each variable.  There are a number of important subcommands that will be needed under the **VARIABLE** command. Generally, when using the following subcommands, the ordering does not matter. |
|  | NAMES are | For this subcommand, include the name of each variable in the data file. |
|  | USEVARIABLES are | For this subcommand, include the name of each variable you want to include in the analysis.  (*Note. This is not required if you are including all of the variables in the data file in the analysis*) |
|  | CATEGORICAL are | For this subcommand, specify which dependent variables (if any) are treated as binary or ordered categorical (ordinal) variables in the model and its estimation. Ordered categorical dependent variables cannot have more than 10 categories. |
|  | CLASSES = | For this subcommand, include the names of categorical latent variable(s) and the number of latent classes.  (*Note. This option is required for TYPE=MIXTURE*) |
|  | IDVARIABLE IS | This subcommand is used to provide an identifier for each observation in the data set that is saved using the **SAVEDATA** command. This subcommand is optional, and is not always used depending on the data file. If there is an idvariable in the data file, try and get into the habit of using it because it can be useful for subsequent analyses. |
|  | ¹STRATIFICATION IS | Subcommands marked (¹) can only be used in conjunction with TYPE=COMPLEX - i.e. the survey data design is complex.  (*Note. Subcommands marked ¹ are NOT required for all survey types and the use of 1* *is not needed in the actual file*) |
|  | ¹CLUSTER IS | Use this subcommand to identify the variable in the data set that contains clustering information. |
|  | ¹WEIGHT IS | Use this subcommand to identify the variable that contains sampling weight information. Sampling weights are used when data have been collected with unequal probabilities of being selected. Sampling weight values must be non-negative real numbers. |
|  | ¹SUBPOPULATION is | Use this subcommand if you have a complex survey design and you want to select a subsample for your analysis.  *(Note.* *Logical operators are used for this subcommand – see Mplus User Guide Version 6 p. 513)*. |
|  | USEOBSERVATIONS is | Use this subcommand if you do not have a complex survey design and you want to select a subsample for your analysis.  (*Note. This subcommand is an alternative subcommand for SUBPOPULATION – it is optional and not always used in an analysis. Logical operators are used for this subcommand – see Mplus User Guide Version 6 p. 513*). |
|  | MISSING are all | Use this subcommand if there is missing data in the data file – this subcommand tells Mplus how you have coded missing data e.g. 999 or -9 |
| **ANALYSIS:** |  | The **ANALYSIS** command is used to describe the technical details of the analysis including the type of analysis and the statistical estimator. The four main subcommands that are used in a latent class model, described next. |
|  | TYPE= | Use this subcommand to describe the type of analysis. Analyses using ‘TYPE=MIXTURE’ include models with categorical latent variables which are also referred to as latent class or finite mixture models.  (*Note. For latent class models, use TYPE=MIXTURE*). |
|  | ESTIMATOR= | Use this subcommand to specify the estimator to be used in the analysis.  *(Note. This is not always necessary to include if you are using the default estimator for a specific type of analysis- check Mplus User Guide 531-533*).  For ‘TYPE=MIXTURE’, Maximum Likelihood (ML) is the default estimator.  For ‘TYPE=MIXTURE COMPLEX’, Maximum Likelihood Robust (MLR) is the default estimator. |
|  | STARTS= | Use this subcommand to specify the number of random sets of starting values to generate in the initial stage and the number of optimizations to use in the final stage. For ‘TYPE=MIXTURE’, the default is 10 random sets of starting values in the initial stage and 2 optimizations in the final stage. |
|  | STITERATIONS= | Use this subcommand to specify the maximum number of iterations allowed in the initial stage. The default number of iterations is 10. For a more thorough investigation, 20 iterations can be requested. |
| **MODEL:** |  | The **MODEL** command is used to describe the model to be estimated. It has options for defining latent variables, describing relationships among variables in the model, and specifying details of the model. |
|  | %OVERALL% | Technically not a subcommand (i.e. it does not need to end in a semi-colon). At a very basic level, this is the main ‘subcommand’ to use for a latent class model. |
| **SAVEDATA:** |  | The **SAVEDATA** command is used to save the analysis data, auxiliary variables, and a variety of analysis results. |
|  | SAVE = | When ‘SAVE=CPROBABILITIES’ is used in conjunction with ‘TYPE=MIXTURE’ in the **ANALYSIS** command, individual posterior probabilities for each class are saved along with the other analysis variables. In addition, a variable is saved that contains the most likely class membership, that is, the class with the highest posterior probability for each individual. |
|  | FILE IS | Use this subcommand to specify the name of the ASCII file in which the individual-level data used in the analysis will be saved. Following this subcommand with FILENAME.DAT means the file will be saved in the same location as the input file. If you want to save it in another location, you need to specify the file to that specific directory. |
| **OUTPUT:** |  | The OUTPUT command is not a required command. The default output for all analyses includes a listing of the input setup, a summary of the analysis specifications, and a summary of the analysis results. The more information you request in the output, the longer Mplus takes to run the analysis and produce the output. |
|  | CINTERVAL | Use this subcommand to request confidence intervals for model parameter estimates. |
|  | TECH10 | Use this subcommand to request univariate, bivariate, and response pattern model fit information for the categorical dependent variables in the model. This includes observed and estimated (expected) frequencies and standardized residuals. TECH10 is available for ‘TYPE=MIXTURE’ and categorical and count variables with maximum likelihood estimation. |
|  | TECH11 | Use this subcommand in conjunction with TYPE=MIXTURE to request the Lo-Mendell-Rubin likelihood ratio test of model fit that compares the estimated model with a model with one less class than the estimated model.  (*Note. TECH11 is available only for ESTIMATOR=MLR*). |
| **PLOT:** |  | The PLOT command is used to request graphical displays of observed data and analysis results. These graphical displays can be viewed after the analysis is completed using a post-processing graphics module. |
|  | TYPE IS | PLOT 3 is useful to request when estimating a latent class model. This will give you a graph of the estimated probabilities for each item in the analysis across the various classes. |
|  | SERIES IS | Use this subcommand to list the names of the variables to be used in plots. This can be done in two ways:   1. List the variables individually 2. List the first and last variables separated by a hyphen (-)   Then, use and asterisk (\*) to request integer values starting with 0 and increasing by 1. This means that the first variable in the list will correspond to a value of 0 on the x-axis, the second variable will correspond to a value of 1 on the x-axis, and so forth. |