A SHORT CLASSROOM-BASED WORKSHOP ON LATENT CLASS ANALYSIS

Dan Green and Eirini Koutoumanou
UCL Great Ormond Street Institute of Child Health,
University College London, London, WC1N 1EH, UK
daniel.green@ucl.ac.uk

Latent Class Analysis (LCA) is a statistical technique that differs to standard regression methods by attempting to identify patterns not directly observed within a population. LCA courses in the United Kingdom focus primarily on statistical software whilst detailed underlying theory is given relatively minor coverage. Our aim was to develop a non-computer based LCA course for non-statisticians that encourages a solid understanding of the theory by using lay language, thus providing a better basis to progress to utilising software and an improved overall student learning outcome. The target audience consists of participants that might never apply these techniques themselves in addition to those that may progress to application. In this presentation, I discuss student feedback and attitudes to the classroom-based approach.

BACKGROUND AND AIMS

The teaching team at CASC (Centre for Applied Statistics Courses), within UCL, offers a range of applied statistical courses that can take attendees from no knowledge of statistics, through to sample size estimates, analysis of missing data, regression models, and various specific statistical techniques such as survival, Bayesian and meta-analyses. Often delegates will come from a highly applied background, whether that being clinical, financial or education: attendees might simply want to learn about how a statistical method works, rather than becoming knowledgeable in carrying out that particular analysis.

Courses are primarily developed to further educate researchers within the GOS Institute of Child Health at UCL, with additional places marketed to individuals external to the department. One such new course theorised was the topic of Latent Class Analysis (LCA), which can be very technical to implement particularly for novices to statistical software (which is often the case for attendees on our courses). Investigating other similar courses available, predominantly advertised in the UK, the majority of courses on LCA were taught along with a software package. Often the structure of the course would be based around how to carryout LCA in Mplus, Latent GOLD or Stata. In fact, from internet searching, we could not find a course currently running within the UK that purely focused on the theory behind LCA without the inclusion of software. While we are not advocating that it is incorrect to include software whilst teaching LCA, the attendees we get on our courses would not in general, we felt, benefit from the content being taught alongside software. Therefore, we developed a theory-based LCA course. The aim of this study was to investigate the opinions of delegates, in relation to lack of software, who attended a one-day course on LCA, (without any teaching or practical 'hands-on' use of software).

LCA BACKGROUND AND COURSE CONTENT

Latent, from the Latin 'latere', means hidden or concealed; therefore, the purpose of latent methodology is to identify knowledge (here sub-groups, that are termed 'classes' in LCA) that are not previously 'known'. LCA is a commonly-used, 'person-centred' technique to explore categorical variables of interest, and depending on the specific model criteria defined, will create a particular number of classes. The characteristics of the variables within one class will then (or at least should) be distinctly different to the characteristics of another. Therefore, individuals within one class will have similar characteristics, compared to individuals in other classes.

The general basis of LCA is quite straightforward; it is a statistical method to segment your sample population into an appropriate number of sub-groups. However, putting this into practice can contain slightly different obstacles to a standard regression model. The main key points to LCA:

- People are put into a class (subgroup) where they share common characteristics to each other;
- People between classes differ distinctly on similar variables;

In M. A. Sorto, A. White, & L. Guyot (Eds.), *Looking back, looking forward. Proceedings of the Tenth International Conference on Teaching Statistics (ICOTS10, July, 2018), Kyoto, Japan.* Voorburg, The Netherlands: International Statistical Institute. iase-web.org

- Everyone included in the analysis is designated into a class;
- LCA is a 'probabilistic approach' meaning everyone gets a probability of belonging to each defined class;
- Once individuals are allocated to a class, the characteristics used to define the class are independent;
- Classes make up a 'latent variable'- an unobserved construct within the data that is often difficult to measure directly (such as happiness or social behaviours);

The researcher sets the number of classes to be defined within a given analysis, and then checks numerous criteria to decide what number of classes reflect the data 'best' (balancing statistical fit, parsimony and relevance). As with many areas of statistics, one of the most challenging aspects of creating a LCA model is deciding on the model: which variables to include in the model and deciding whether the model is a valid (and clinically/scientifically useful) representation of your research question.

The course was split into 5 main sections:

- Why LCA is used in research (real examples where using other standard methods would not be an efficient approach);
- Basic concepts of LCA with simple example (only 3 variables included, 3 latent classes identified);
- Applied example using real data (how to decide on number of latent classes with all the main model criteria introduced);
- LCA extensions (such as multi-group LCA, and the inclusion of covariates);
- Longitudinal extensions (such as Latent Class Growth Analysis and Latent Transition Analysis).

Delegates were also provided with a comprehensive set of printed notes that they could annotate and keep for their own use. All of the taught content is explained in detail in their course notes. In addition to the elements taught during the course, a couple of extra sections included further reading and also some software templates for various software packages. For example, screenshots of how to compute a LCA model with 3 latent classes for the main example provided throughout the course was provided for R, Stata, Mplus and LatentGOLD. While this content was not taught directly during the course, it was provided as a starting block for attendees that might want to venture into carrying out LCA software analyses in the future.

Also included in the course was a practical session where delegates were provided with a print-out containing a specific scenario, some variables of interest, and then the different LCA models for 2 to 7 classes, along with model-specific criteria for each (all presented in clean formatted tables, not software output). The task of the delegate was then to assess the different models (using only the printed material provided) to decide which of the models would be the most sufficient representation for this scenario. They were guided to focus on the different model criteria discussed during the course presentation, and to consider the interpretation of each class within each model. As a group, we then discussed the merits of considering each model, and debated on which would be the most optimum.

ASSESSMENT

After every course at CASC we ask for delegates' feedback collected through an online survey sent after the course finishes. All feedback is strictly anonymous. The feedback received is used as a proxy to see how delegates found the course (measured on a 5-point Likert scale representing 'poor' (1) through to 'excellent' (5)) and whether they have suggestions to improve the course. Delegates to the LCA were asked numerous questions regarding their satisfaction regarding various elements of the course, such as the content, presentation, usefulness, practical activities, timings, comfort, satisfaction with refreshments, with also space for comments provided throughout.

Delegates were also asked the question: "How useful did you find this course on Latent Class Analysis, without the use of software?". This question had five potential answers:

- I think this course should be taught with software, not in the classroom;
- I think this course was OK, but some software should be included;

- It was OK, no strong opinion on software;
- I think the theory based class was good, some extra time for software would also be useful (e.g. extra half day);
- I think the theory based class was great, there was no need for software.

The delegates were also given a further sixth option "N/A" and encouraged to offer their own alternative response. Finally, delegates were provided with a text box to enter any further comments about the use of software.

FEEDBACK

From the 25 attendees of the course, 20 individuals provided feedback. Generally, delegates were pleased with the course. Eighteen (90%) rated the content of the course as good or excellent (two rated it as poor). Similarly, 16 (80%) rated the usefulness of the course as good or excellent, and 17 (85%) rated the usefulness of the practical exercises as good or excellent.

With regard to the question on the use of software, 19 individuals provided a response to the multiple-choice question, of which two selected a "N/A" response. By far, the most common response was option iv: "I think the theory based class was good, some extra time for software would also be useful (e.g. extra half day)", which was selected by 13 of the responding delegates (74%). Of the remaining four delegates that gave a different response, three selected option v (no need for software) and one selected option ii (software should be included).

Some of the written comments (selected as most relevant from the 13 delegates who provided comments) provided highlight that delegates did feel they benefitted from the theory-based class:

Delegate 1 (selected option iv):

"[…] Very useful to learn the theory behind this type of statistical analysis and to have time to focus on it. Usually, when we focus on software use, we miss the importance of knowing the theory upon which the analysis is based, so good to have a whole day for that! The addition of software may be useful only as an additional element to this training (an extra day/half day perhaps)."

Delegate 4 (selected option iv):

"I had little/no knowledge of the latent class analytical method and found the course very helpful for understanding the theory and addressing questions on the different applications, test and limitations […] It was really helpful to have a review of the statistical software tools that could be used also and the different ways of assessing a good model fit […] If a practical session is added in later courses, it can be included at the end for those who are interested but the theory only session (without computers) met my objectives."

Delegate 2 (selected option iv):

"Explaining the theory was great, and it was clear and relatively easy to understand. It would be good to have a bit more discussion or practical examples of software […] I think one benefit to using software in the course would be to get a better idea of the kind of output you get […] it would be beneficial to see what actually comes out of the software and work with it."

Of the students that selected option v (felt that no software was needed), only one left a written comment:

Delegate 11 (selected option v):

"Having software outputs to interrogate was critical to making this approach work."

By this comment, we presume the delegate is referring to either the tables provided during the class and practical exercise (by getting them to understand the values that would be provided from the software), or by the template software code provided towards the back of the notes (not focussed on explicitly during the course).

As already alluded to earlier, a couple of the attendees (who rated the course as poor) did not find the course to their liking. However, their issues were more concerned with the fact they were hoping for a more advanced/mathematical course (despite our website clearly stating that our courses are 'applied' and aimed at lay non-statistically minded people).

CONCLUSIONS

In general, the reception to the course was positive. Taking the feedback from the delegates overall, the majority did in no way feel that the course required software to convey the main messages of LCA. For example, the feedback from delegate 1 submitted an ideal answer: “Usually, when we focus on software use, we miss the importance of knowing the theory upon which the analysis is based, so good to have a whole day for that!”.

However, many of the delegates did reaffirm that the inclusion of software would increase the appeal of the course. A further complication highlights which software could be used. Expense aside, the suitability of software considering the type of attendees often seen on CASC courses will have a large impact. Do you breach the complex, but flexible, coding system of Mplus that the majority of individuals will have never seen before? Or go down the line of the more user-friendly interface of Latent GOLD, which can have some limitations in terms of flexibility for LCA.

Another delegate suggests a potential solution: “I think one benefit to using software in the course would be to get a better idea of the kind of output you get”. Therefore, rather than presenting the ‘output’ tables in a clear formatted table, a solution could be to provide them with the raw output from different software and identify where the relevant results/ estimates are. Again, there would have to be a decision made on which particular software to go through during the class, but this would have less of an impact if it was built into the standard day class, rather than adding a half-day software practical extension. This alongside the basic code/ script files provided with the course notes could aid this additional function.

In summary, the viability of running a course on LCA (or other related methods) without the necessity of software definitely seems plausible. The success of the course was very well complemented with the use of a practical example that exposed delegates to different outputs (in neatly formatted tables here) to assist them in model selection and to re-affirm the concepts and content they have been introduced to throughout the course. The nature of the course, non-computer based, helped delegates appreciate the decisions required when settling on an optimum model and the fundamental principles underlying in Latent Class Analysis.