**INTRODUCTION TO LATENT CLASS ANALYSIS**

**by Simon Moss**

|  |
| --- |
| **Introduction** |

To introduce latent class analysis, consider the following example. Imagine that a cohort of 200 research candidates are invited to complete a survey on whether they engage in various unproductive behaviours during work hours—such as consuming alcohol, eating chocolate, playing computer games, shopping for clothes, and chatting with friends. In particular, they indicate which of these activities they had enacted in the last week. The following table presents the data.

* 1s indicate the participant had enacted this behavior;
* 2s indicates the participant had not enacted this behavior



 If these data are subjected to a latent class analysis, the software generates several tables of data. First, the software will indicate the number of classes—that is, subgroups of participants whose responses are similar. To illustrate

* in this instance, the latent class analysis would probably show the participants can be divided into three, or perhaps four, classes or groups
* for example, one class might primarily include Adam, Betty, Carl, Donna, and Edith because these individuals produced similar responses
* another class might primarily include Fred, Gloria, Helen, Ian, Kate, Len, and Molly because these individuals also produced similar responses

Second, and perhaps more importantly, the software will generate a table that indicates the proportion of individuals who engage in these behaviours—but for each class separately. The following table presents these responses. In this table

* proportions that exceed .4 are printed in bold
* people in Class 1 often consumed alcohol, smoking, drugs, and chocolate
* people in Class 2 often played games, engaged in social media, and chatted with friends
* people in Class 3 tended to engage in retail shopping, internet shopping, and gambling
* the final row indicates that approximately 30%, 30%, and 40% of individuals belonged to Classes 1, 2, and 3 respectively

|  |  |  |  |
| --- | --- | --- | --- |
| Unproductive behaviour  | Class 1 | Class 2 | Class 3 |
| Alcohol | **.78** | .08 | .04 |
| Smoking | **.86** | .14 | .09 |
| Drugs | **.67** | .12 | .12 |
| Chocolate | **.79** | .06 | .11 |
| Playing games | .18 | **.67** | .14 |
| Social media | .06 | **.87** | .06 |
| Chatting  | .11 | **.58** | .07 |
| Retail shopping | .07 | .06 | **.87** |
| Internet shopping | .11 | .14 | **.79** |
| Gambling | .07 | .07 | **.70** |

 Researchers can derive several conclusions from these results. In particular

* The first class primarily corresponds to unhealthy substances, such as alcohol and chocolate. Hence, almost a third of research candidates consume unhealthy substances
* The second class primarily corresponds to social distractions, such as social media and chatting. Hence, almost a third of research candidates engage in these distractions
* Finally, the third class primarily corresponds to impulsive behaviors, such as shopping or gambling. Hence, over a third of research candidates engage in these impulsive behaviors

|  |
| --- |
| **Implementation of latent class analysis** |

**Basic commands**

 Only a subset of statistical packages, such as R, can be utilized to implement latent class analyses. This section demonstrates how you can utilize R to conduct this technique. If you are not familiar with R, please read the document about the “basics of R” first. To conduct latent class analysis, first install the package called poLCA. Then, utilize the following R commands

* library(poLCA)
* f = cbind(item1, item2, item3, item4, item5, item6, item7, item8) ~ 1
* poLCA(f, Data.file, nclass = 2, verbose = FALSE)

You can enter these commands exactly, except

* replace item1, item2, and so forth, specify the names of your columns or items, such as alcohol, smoking, drugs, and so forth
* replace Data.file with the name of your data file
* latent class analysis is suitable whenever the items or variables are categorical; in this example, for instance, each item is assigned one of two categories: 1 or 2
* ensure the numbers in your data file begin with 1, such as 1, 2, or 3. Exclude zeros

**Determine number of classes**

 When you conduct a latent class analysis, your first task is to determine the number of classes or groups to which participants should be assigned. Should participants be divided into three classes, as the previous example demonstrated? Or, should participants be divided into two classes, four classes, five classes, or more classes? To answer this question

* conduct the analysis
* in the output that appears, towards the end, you will receive an index called AIC, as shown in the following extract of output
* repeat this analysis but instruct the software to examine three classes instead
* that is, replace “nclass = 2” with “nclass = 3”
* repeat this analysis several more times, but with 4, 5, 6, 7, and 8 classes respectively
* choose the number of classes that generates the lowest AIC

|  |  |  |  |
| --- | --- | --- | --- |
| 2 classes | 3 classes | 4 classes | 5 classes |
| ## AIC(2): 497.5768## BIC(2): 548.2894## G^2(2): 99.87539 ## X^2(2): 168.7594  | ## AIC(3): 491.5768## BIC(3): 541.2894## G^2(3): 95.87539 ## X^2(3): 165.7594 | ## AIC(4): 503.5768## BIC(4): 569.2894## G^2(4): 104.87539 ## X^2(4): 175.7594 | ## AIC(5): 514.5768## BIC(5): 598.2894## G^2(5): 112.87539 ## X^2(5): 179.7594 |

**Interpret the output**

Suppose, after scrutinizing the AIC values, you decide the participants should be assigned to three classes. You should now scan the output that corresponds to this model. As this data shows, if you scan the four rows after “Alcohol”, you would conclude that

* For people in class 1, 38.4% specified 1 and 61.6% specified 2
* For people in class 2, 0% specified 1 and 100% specified 2
* For people in class 3, 67% specified 1 and 33% specified 2
* You could then apply the same rationale for all the other items, to generate something like the previous table on page 2.

|  |
| --- |
| ## Conditional item response (column) probabilities,## by outcome variable, for each class (row)#### $Alcohol## Pr(1) Pr(2)## class 1: 0.3841 0.6159## class 2: 0.0000 1.0000## class 3: 0.6667 0.3333#### $Smoking## Pr(1) Pr(2)## class 1: 0.3841 0.6159## class 2: 0.0000 1.0000## class 3: 0.6667 0.3333#### $Drugs## Pr(1) Pr(2)## class 1: 0.4609 0.5391## class 2: 0.0000 1.0000## class 3: 0.1111 0.8889#### $Chocolate## Pr(1) Pr(2)## class 1: 0.5767 0.4233## class 2: 0.1650 0.8350## class 3: 0.2222 0.7778#### $Playing## Pr(1) Pr(2)## class 1: 0.0398 0.9602## class 2: 1.0000 0.0000## class 3: 0.1667 0.8333#### $SocialMedia## Pr(1) Pr(2)## class 1: 0.2319 0.7681## class 2: 1.0000 0.0000## class 3: 0.2222 0.7778#### $Chatting## Pr(1) Pr(2)## class 1: 0.4225 0.5775## class 2: 0.0000 1.0000## class 3: 0.1667 0.8333#### $Gambling## Pr(1) Pr(2)## class 1: 0.4997 0.5003## class 2: 0.1657 0.8343## class 3: 0.1111 0.8889#### Estimated class population shares## 0.5207 0.1193 0.36#### Predicted class memberships (by modal posterior prob.)## 0.52 0.12 0.36 |

|  |
| --- |
| **When to use latent class analysis** |

 Thus far, this document has illustrated and implemented basic latent class analysis. This section presents a few more insights about latent class analysis, such as when to apply this technique

**Applicability of this technique**

In general, researchers utilize latent class analysis whenever

* all the items or variables in the analysis comprise the same number of categories; in the previous example, all the items comprise two categories: yes or no
* the researcher assumes that subsets of participants might generate similar responses

Latent class analysis fulfils a similar aim to factor analyses and cluster analyses except

* to conduct factor analyses, the items or variables need to be numerical rather than categorical
* latent class analysis can provide more information than cluster analysis, such as goodness of fit measures such as the AIC

**Assumptions**

Latent class analysis assumes a multinomial distribution and that, within each class, the items or variables are independent of one another. To illustrate, in the previous example

* one class seemed to include primarily Adam, Betty, Carl, Donna, and Edith because these individuals produced similar responses
* yet, within this subset of participants, the likelihood someone consumes alcohol is unrelated to the likelihood this person consumes substances
* but, across classes, the items or variables are not independent of one another
* that is, across all participants, people who consume alcohol are more likely to consume other substances as well

In addition

* by default, participants with missing values are excluded; they can be retained, however, if you include the subcommand “na.rm = FALSE”

|  |
| --- |
| **References** |

Agresti, A. (2002). Categorical data analysis, second edition. Hoboken: John Wiley & Sons.

Linzer, D. A., & Lewis, J. B. (2011). poLCA: An R Package for Polytomous Variable Latent Class Analysis. Journal of Statistical Software, 42(10), 1-29.

Hagenaars, J. A., & McCutcheon, A. L. (2009). Applied latent class analysis. Cambridge University Press.