Segmentation IV

Introduction to latent class analysis

Morten Berg Jensen

Department of Economics and Business Economics

April 18, 2024

Outline

Outcome

- This lecture will help you to understand
  - The relative position of latent class models and measures of interrelationships for categorical (binary) data
  - The latent class model for binary data and estimation of such models
  - Assessment and allocation to classes for such a model

Latent class models

- This lecture focuses on latent variable models where we assume that the manifest variables/indicators are categorical (binary) and the latent variable is categorical (nominal)
- Thus, contrasted with factor analysis we change the latent as well as the manifest variables from continuous to nominal – notice you can also do factor analysis for categorical indicators
- Thus we proceed as follows
  - Explore potential interrelationships for the binary indicators
  - Define a probabilistic model relating the binary indicators to a latent nominal factor
  - Assess whether the suggested model can reproduce the original relationships
  - Eventually assign a “factor score” to each individual – predicted class membership

Latent class vs. cluster analysis

- Latent class analysis can also due to its second stage be seen as a form of cluster analysis
- However, the latent class model is based on a probability model
- Cluster analysis is rooted in the similarities between rows of the data matrix
- Latent class analysis is based on the probabilities of the elements in the rows
- Seeking to form clusters of similar probabilities, the latent class analysis focuses on rows of similar expectations
- This is accomplished via the assumption of local independence

Measures of association for binary data

- The most natural way of assessing association between two binary variables is via a contingency table
- Thus for a set of binary variables we would look at all possible pairwise associations
- The general idea is still to assess whether the presence of some strong pairwise associations can be attributed to a common latent factor

The data matrix

- In the case of binary indicators we let each row be the answers from each of the n respondents
- In terms of values we use 1 to indicate the "success" outcome and 0 to indicate the "failure" outcome
- Any row of the data matrix is referred to as a score pattern
- For $p = 3$ indicators we have the following possible score patterns

$$000, 001, 010, 011, 100, 101, 110, 111$$

- As such, the sum of responses for each respondent corresponds to the number of positive responses – referred to as the total score

The data matrix (cont'd)

- In many situations we use the frequencies for the observed score patterns instead of the raw data
- This is an efficient way of reporting the answers from many respondents – provided the number of patterns isn't too large
- The columns for the raw data hold the answers to each indicator and hence an average will tell you how large a fraction of the respondents agree on this indicator

Assumptions

- The responses to the p observed binary items are independent given the latent variable, y, – conditional independence
- This assumption can only be tested indirectly via an assessment of the fit of the model to the data

The J-class model

- We let π_{ij} denote the probability of a positive response on variable i for a person from latent class j, $i = 1, \dots, p$ and $j = 1, \dots, J$
- We let η_j denote the prior probability that a randomly chosen individual is in class j, naturally $\sum_{j=1}^{J} \eta_j = 1$
- The joint probability of observing the response vector $\mathbf{x}$ then becomes

$$f(\mathbf{x}) = \sum_{j=1}^{J} \eta_j \prod_{i=1}^{p} \pi_{ij}^{x_i} (1 - \pi_{ij})^{1-x_i}$$

- The posterior probability that an individual with response vector $\mathbf{x}$ belongs to category j is then

$$h(j|\mathbf{x}) = \eta_j \prod_{i=1}^{p} \pi_{ij}^{x_i} (1 - \pi_{ij})^{1-x_i} / f(\mathbf{x})$$

Extensions

- The latent class model can also accommodate nominal indicators (more than two outcomes) and ordinal indicators
- And any combination

Maximum likelihood estimation

- The log-likelihood for a sample of size n is

$$\mathcal{L} = \sum_{h=1}^{n} \ln\{\sum_{j=1}^{J} \eta_j \prod_{i=1}^{p} \pi_{ij}^{x_{ih}} (1 - \pi_{ij})^{1-x_{ih}}\}$$

  which has to be maximized subject to $\sum_j \eta_j = 1$

- We form the Lagrangian

$$\phi = \mathcal{L} + \theta \sum_{j=1}^{J} \eta_j$$

Maximum likelihood estimation (cont'd)

- We get the following partial derivatives

$$\frac{\partial\phi}{\partial\eta_j} = \sum_{h=1}^{n}\{\prod_{i=1}^{p}\pi_{ij}^{x_{ih}}(1-\pi_{ij})^{1-x_{ih}}/f(\mathbf{x}_h)\} + \theta$$
$$= \sum_{h=1}^{n}\{g(\mathbf{x}_h|j)/f(\mathbf{x}_h)\} + \theta$$

and also (after a few manipulations)

$$\frac{\partial\phi}{\partial\pi_{ij}} = \{\eta_j/\pi_{ij}(1-\pi_{ij})\}\sum_{h=1}^{n}(x_{ih}-\pi_{ij})g(\mathbf{x}_h|j)/f(\mathbf{x}_h)$$

Maximum likelihood estimation (cont'd)

- Using Bayes' theorem, $h(j|\mathbf{x}_h) = \eta_j g(\mathbf{x}_h|j)/f(\mathbf{x}_h)$ we get

$$\hat{\eta}_j = \sum_{h=1}^{n} h(j|\mathbf{x}_h)/n \tag{1}$$

and

$$\hat{\pi}_{ij} = \sum_{h=1}^{n} x_{ih} h(j|\mathbf{x}_h)/n\hat{\eta}_j \tag{2}$$

- This looks simpler than it is because

$$h(j|\mathbf{x}_h) = \frac{\eta_j \prod_{i=1}^{p} \pi_{ij}^{x_{ih}}(1-\pi_{ij})^{1-x_{ih}}}{\sum_{k=1}^{J} \eta_j \prod_{i=1}^{p} \pi_{ik}^{x_{ih}}(1-\pi_{ik})^{1-x_{ih}}} \tag{3}$$

Maximum likelihood estimation (cont'd)

- The E-M-algorithm exploits the fact that if $h(j|\mathbf{x}_h)$ were known then (1) and (2) would be easy to solve
- Hence we proceed as follows:
  1. Choose an initial set of posterior probabilities, $\{h(j|\mathbf{x}_h)\}$
  2. Use (1) and (2) to get and approximation to $\{\hat{\eta}_j\}$ and $\{\hat{\pi}_{ij}\}$
  3. Substitute these approximations into (3)
  4. Return to 2. and continue until convergence

Standard errors

- Finding the second derivatives and cross-derivatives of $\mathcal{L}$ is easy but cumbersome
- As is well-known we get the asymptotic variance-covariance matrix as the inverse of the expected negative Hessian
- For p small this is doable but for larger p we run into numerical problems – the observed second derivatives may be used as substitutes
- A more concerning issue is the fact that the asymptotic approximation is rather poor for standard sample sizes – instead a parametric bootstrap has been suggested

Goodness-of-fit – global tests

- As is typically done when assessing statistical models involving categorical data a comparison of observed frequencies with estimated expected frequencies is carried out
- Thus, across the 2^p different score patterns we can calculate the log-likelihood ratio statistic

$$G^2 = 2\sum_{r=1}^{2^p} O(r) \log \frac{O(r)}{E(r)}$$

where $O(r)$ is the observed number of observations for pattern r and $E(r)$ is the estimated expected number of observations

Goodness-of-fit – global tests (cont'd)

- Alternatively, we can calculate the Pearson chi-squared goodness-of-fit statistic

$$X^2 = \sum_{r=1}^{2^p} \frac{(O(r) - E(r))^2}{E(r)}$$

- Under the null, that the model fits the data, both statistics follow a χ^2 distribution with degrees of freedom equal to $2^p - J(p+1)$
- Both statistics operate under the assumption that the expected number of observations are above 5 – if not aggregation of cells might be a solution
- However, that solution can run into problems with a non-positive degrees of freedom

Goodness-of-fit – local tests

- Instead of looking at the global set of response patterns we can look locally at all the possible 2×2 contingency tables that can be constructed
- For each table we look at the chi-squared residuals from each cell – i.e. from $(0,0), (0,1), (1,0)$ and $(1,1)$
- For each table the sum of all chi-squared residuals make up the overall chi-squared statistic for independence
- This idea can be extended to a $2 \times 2 \times 2$ table

Goodness-of-fit – local tests (cont'd)

- Unfortunately, we cannot aggregate the contributions across tables due to dependence between tables but more advanced methods exist that can calculate an overall test
- Meanwhile, as a rule of thumb, chi-square residuals for each cell can be assumed to have a χ^2 distribution with one degree of freedom
- Thus, chi-square residuals above 4 (sometimes 3) are seen as indications of a poor fit

Goodness-of-fit – model comparisons

- An alternative to assessing how well a particular model fits the data is obtained by comparing the fit of a particular model to that of a reference model
- This idea corresponds to the analysis in "standard" factor analysis and PCA of assessing the amount of variance explained
- In the current situation, the reference model is the independence model
- This reference model is the pertinent model if there were no associations between the indicators

Goodness-of-fit – model comparisons (cont'd)

- We compare the fit of the two models in terms of their log-likelihood ratio statistics, G^2

$$\% G^2 = \frac{G_0^2 - G_J^2}{G_0^2} \times 100$$

  where G_0^2 and G_J^2 are the log-likelihood ratio statistics for the independence model and the latent J-class model respectively
- There are no rules of thumb for assessing what constitutes a large fraction of G^2 explained
- Finally, statistical models can be compared using the generic model selection criteria AIC, BIC, etc.

Determining the number of clusters

- See discussion in relation to the same issue for model-based clustering

Posterior analysis

- As a byproduct of the E-M-algorithm we get the required posterior probabilities
- For response patterns not in the sample, these probabilities can easily be calculated using (3)

Background and deciding (not) to segment

- McDonald's would like to know whether consumer segments exist with distinctly different images of McDonald's
- Such an understanding would inform McDonald's which segment(s) to focus on if any and what kind of communication to use
  - McDonald's can choose to cater to the entire market and hence ignore systematic differences across segments
  - The can also choose to focus market segments with a positive perception and strengthen this perception
  - Or, focus on the segment with a negative perception and try to modify the drivers of the negative perception

Collecting data

- We have information from 1453 adult Australian consumers regarding their perception of McDonald's
- Specifically, they have indicated whether they feel McDonald's do or do not possess the following 11 attributes:
YUMMY, CONVENIENT, SPICY, FATTENING, GREASY, FAST, CHEAP, TASTY, EXPENSIVE, HEALTHY, and DISGUSTING
- Given the data limitations we will use "liking McDonald's" and "frequently eating at McDonald's" as attractiveness criteria
- Finally, in addition to the two attractiveness criteria we have information about gender and age
- Had the data been collected for segmentation, additional information should have been collected about for instance dining out behaviour and use of information channels

Extracting segments

- In order to investigate the optimal number of segments we begin by estimating a mixture of binary distributions for all possible number of segments between two and eight
- Ten random restarts of the EM algorithm are used for each value of the number of clusters
- The decision regarding the number of segments is guided by the values of AIC, BIC, and ICL

Extracting segments (cont'd)

- Adhering strictly to the information criteria BIC and ICL suggest a seven segment solution
- AIC suggests an eight segment solution (which could, in principle, also be a nine or ten segment solution)
- However, from the plot of the information criteria as a function of the number of segments not much is gained beyond a four segment solution
- A cross table of the four segment solutions from the k-means and the mixture of distributions models can be used to assess the stability

Extracting segments (cont'd)

- The stable segments in the k-means solution (segments 2 and 3) are very similar to segments 4 and 2 in the finite mixture model solution
- This is even more apparent when we use the k-means solution as initialization values for the finite mixture model where segments 2,3, and 4 in the k-means solution are very similar to segments 2,3, and 4 in the finite mixture model
- Notice that the log-likelihood values for the two mixture solutions are very similar suggesting that we have indeed found a global solution for the maximization problem

The remaining steps

- See previous slide set for considerations regarding
  - Profiling segments
  - Describing segments
  - Selecting (the) target segment(s)
  - Customising the marketing mix
  - Evaluation and monitoring