A­dvanced Topics in Research Methods and Design

Yi Chen

Columbia University

Exercise Five: Latent Class Model

Note: The SPSS modeling exercises are finished by myself first, then checked and discussed with Jingru Zhang and Lesheng Xu. The proposal is finished completely by myself.

A­dvanced Topics in Research Methods and Design

# Problem 1: Exercises in Chapter 2

The syntax of the whole model is:

```````````````````````````````````````````````````````````````````````````````````````````````````````````

Title:

Fictitous Latent Class Analysis.

Data:

File is C:\Users\ATS-Workshop\Downloads\lca1.dat ;

Variable:

names = id item1 item2 item3 item4 item5 item6 item7 item8 item9;

usevariables = item1 item2 item3 item4 item5 item6 item7 item8 item9;

categorical = item1 item2 item3 item4 item5 item6 item7 item8 item9;

classes = c(3);

Analysis:

Type=mixture;

Plot:

type is plot3;

series is item1 (1) item2 (2) item3 (3) item4 (4) item5 (5)

item6 (6) item7 (7) item8 (8) item9 (9);

Savedata:

file is lca1\_save.txt ;

save is cprob;

format is free;

Output:

tech11 tech14;

```````````````````````````````````````````````````````````````````````````````````````````````````````````````

The model output includes:

* Summary of model varaibles

```````````````````````````````````````````````````````````````````````````````````````````````````````````````

Number of groups 1

Number of observations 1000

Number of dependent variables 9

Number of independent variables 0

Number of continuous latent variables 0

Number of categorical latent variables 1

Observed dependent variables

Binary and ordered categorical (ordinal)

ITEM1 ITEM2 ITEM3 ITEM4 ITEM5 ITEM6

ITEM7 ITEM8 ITEM9

Categorical latent variables

C

```````````````````````````````````````````````````````````````````````````````````````````````````````````````

Interpretation: latent class analysis (LCA) is unsupervised learning, which do not have independent variable and dependent variable as linear regression. The 9 dependent variables are the variables used for classify the 1000 observations into three classes. All of these variables are categorical and the categorical latent variable is the membership of class, which LCA estimated.

* Summary of the model estimations

```````````````````````````````````````````````````````````````````````````````````````````````````````````````

Estimator MLR

Information matrix OBSERVED

Optimization Specifications for the Quasi-Newton Algorithm for

Continuous Outcomes

Maximum number of iterations 100

Convergence criterion 0.100D-05

Optimization Specifications for the EM Algorithm

Maximum number of iterations 500

Convergence criteria

Loglikelihood change 0.100D-06

Relative loglikelihood change 0.100D-06

Derivative 0.100D-05

Optimization Specifications for the M step of the EM Algorithm for

Categorical Latent variables

Number of M step iterations 1

M step convergence criterion 0.100D-05

Basis for M step termination ITERATION

Optimization Specifications for the M step of the EM Algorithm for

Censored, Binary or Ordered Categorical (Ordinal), Unordered

Categorical (Nominal) and Count Outcomes

Number of M step iterations 1

M step convergence criterion 0.100D-05

Basis for M step termination ITERATION

Maximum value for logit thresholds 15

Minimum value for logit thresholds -15

Minimum expected cell size for chi-square 0.100D-01

Optimization algorithm EMA

Random Starts Specifications

Number of initial stage random starts 20

Number of final stage optimizations 4

Number of initial stage iterations 10

Initial stage convergence criterion 0.100D+01

Random starts scale 0.500D+01

Random seed for generating random starts 0

Link LOGIT

```````````````````````````````````````````````````````````````````````````````````````````````````````````````

Interpretation: latent class analysis (LCA) use the MLR estimation method, which means maximum likelihood parameter estimates with standard errors. The optimization algorithm is EMA, which combines EM steps with interspersed Quasi-Newton (QN) or Fisher scoring (FS) steps when they are needed to speed things along. The link function for classification is logit.

* Univariate Proportions and Counts for Categorical Variables

```````````````````````````````````````````````````````````````````````````````````````````````````````````````

ITEM1

Category 1 0.307 307.000

Category 2 0.693 693.000

ITEM2

Category 1 0.709 709.000

Category 2 0.291 291.000

ITEM3

Category 1 0.916 916.000

Category 2 0.084 84.000

ITEM4

Category 1 0.910 910.000

Category 2 0.090 90.000

ITEM5

Category 1 0.801 801.000

Category 2 0.199 199.000

ITEM6

Category 1 0.718 718.000

Category 2 0.282 282.000

ITEM7

Category 1 0.861 861.000

Category 2 0.139 139.000

ITEM8

Category 1 0.833 833.000

Category 2 0.167 167.000

ITEM9

Category 1 0.723 723.000

Category 2 0.277 277.000

```````````````````````````````````````````````````````````````````````````````````````````````````````````````

Interpretation: This is a descriptive analysis of the categorical variables. All of the input variable in this example are binary. Thus, this table shows for each input variable (item), how many observations are belongs to each category and what is the corresponding proportion. This will help us to check the overall distribution of the categorical variables and whether there is significant unbalance issues.

* Final stage loglikelihood values at local maxima, seeds, and initial stage start numbers

```````````````````````````````````````````````````````````````````````````````````````````````````````````````

-4231.696 68985 17

-4231.696 253358 2

-4231.696 903420 5

-4231.930 533738 11

```````````````````````````````````````````````````````````````````````````````````````````````````````````````

Interpretation: this shows the general convergence and robustness of the EM algorithm. As we can see, the best loglikelihood value has been replicated three times with random starts. This means the result is reliable and unlikely to have local maxima issues.

* Model Fit Information

```````````````````````````````````````````````````````````````````````````````````````````````````````````````

Loglikelihood

H0 Value -4231.696

H0 Scaling Correction Factor 1.1062

for MLR

Information Criteria

Akaike (AIC) 8521.392

Bayesian (BIC) 8663.716

Sample-Size Adjusted BIC 8571.611

(n\* = (n + 2) / 24)

Chi-Square Test of Model Fit for the Binary and Ordered Categorical

(Ordinal) Outcomes

Pearson Chi-Square

Value 488.512

Degrees of Freedom 482

P-Value 0.4089

Likelihood Ratio Chi-Square

Value 319.955

Degrees of Freedom 482

P-Value 1.0000

```````````````````````````````````````````````````````````````````````````````````````````````````````````````

Interpretation: this table show the model fit indicators of the final result. The saturated model Mplus assumed has 29 free parameters. The loglikelihood give the absolute model fit based on the latent class model we proposed without any penalty. Information criteria, on the other hand, combine the loglikelihood information with the number of parameter and number of observations. The AIC, BIC, and adjusted BIC is helpful to compare the model with other possibilities (e.g., do the LCA with more classes). Likelihood ratio test can be used to compare the model with the nested model framework.

* Final Classes Counts and Proportions

```````````````````````````````````````````````````````````````````````````````````````````````````````````````

Latent

Classes

1 79.29073 0.07929

2 557.55198 0.55755

3 363.15728 0.36316

Latent

Classes

1 66 0.06600

2 646 0.64600

3 288 0.28800

```````````````````````````````````````````````````````````````````````````````````````````````````````````````

Interpretation: the result is same foe the estimated model and posterior probabilities. And based on the most likely membership result, there are 66 observations (6.6%) belongs to the class one, 646 observations (64.6%) belongs to class two, and 288 observations belongs to class three (28.8%).

* Classification Quality

```````````````````````````````````````````````````````````````````````````````````````````````````````````````

CLASSIFICATION QUALITY

Entropy 0.548

Classification Probabilities for the Most Likely Latent Class Membership (Column)

by Latent Class (Row)

1 2 3

1 0.659 0.329 0.012

2 0.023 0.900 0.077

3 0.002 0.325 0.673

Logits for the Classification Probabilities for the Most Likely Latent Class Membership (Column)

by Latent Class (Row)

1 2 3

1 4.038 3.343 0.000

2 -1.202 2.464 0.000

3 -5.616 -0.729 0.000

```````````````````````````````````````````````````````````````````````````````````````````````````````````````

Interpretation: entropy indicates the overall information in the membership distribution. The classification probabilities tables show that most observations are classified into the class, which should be the most likely class. In particular, 90% of the class two observations are indeed the one which are most likely to be in class 2. This indicates a high confidence of the result.

* Model Result

```````````````````````````````````````````````````````````````````````````````````````````````````````````````

Two-Tailed

Estimate S.E. Est./S.E. P-Value

Latent Class 1

Thresholds

ITEM1$1 -2.488 0.695 -3.579 0.000

ITEM2$1 -0.185 0.371 -0.499 0.618

ITEM3$1 0.297 0.580 0.511 0.609

ITEM4$1 0.331 0.332 0.997 0.319

ITEM5$1 -1.183 0.707 -1.673 0.094

ITEM6$1 0.116 0.357 0.325 0.745

ITEM7$1 -0.050 0.459 -0.108 0.914

ITEM8$1 -0.486 0.652 -0.746 0.456

ITEM9$1 0.624 0.318 1.964 0.050

Latent Class 2

Thresholds

ITEM1$1 -2.293 1.542 -1.487 0.137

ITEM2$1 0.675 0.411 1.641 0.101

ITEM3$1 2.638 0.323 8.159 0.000

ITEM4$1 2.658 0.428 6.215 0.000

ITEM5$1 1.271 0.472 2.694 0.007

ITEM6$1 0.755 0.297 2.546 0.011

ITEM7$1 2.062 0.249 8.279 0.000

ITEM8$1 1.816 0.330 5.505 0.000

ITEM9$1 0.731 0.250 2.926 0.003

Latent Class 3

Thresholds

ITEM1$1 0.790 1.516 0.521 0.602

ITEM2$1 1.629 0.375 4.338 0.000

ITEM3$1 3.295 0.538 6.121 0.000

ITEM4$1 2.824 0.332 8.518 0.000

ITEM5$1 3.070 1.244 2.467 0.014

ITEM6$1 1.497 0.416 3.597 0.000

ITEM7$1 2.223 0.263 8.454 0.000

ITEM8$1 2.092 0.255 8.204 0.000

ITEM9$1 1.464 0.342 4.283 0.000

Categorical Latent Variables

Means

C#1 -1.522 1.152 -1.321 0.186

C#2 0.429 1.290 0.332 0.740

RESULTS IN PROBABILITY SCALE

Latent Class 1

ITEM1

Category 1 0.077 0.049 1.558 0.119

Category 2 0.923 0.049 18.756 0.000

ITEM2

Category 1 0.454 0.092 4.934 0.000

Category 2 0.546 0.092 5.937 0.000

ITEM3

Category 1 0.574 0.142 4.041 0.000

Category 2 0.426 0.142 3.004 0.003

ITEM4

Category 1 0.582 0.081 7.201 0.000

Category 2 0.418 0.081 5.170 0.000

ITEM5

Category 1 0.235 0.127 1.847 0.065

Category 2 0.765 0.127 6.029 0.000

ITEM6

Category 1 0.529 0.089 5.949 0.000

Category 2 0.471 0.089 5.297 0.000

ITEM7

Category 1 0.488 0.115 4.256 0.000

Category 2 0.512 0.115 4.473 0.000

ITEM8

Category 1 0.381 0.154 2.479 0.013

Category 2 0.619 0.154 4.031 0.000

ITEM9

Category 1 0.651 0.072 9.019 0.000

Category 2 0.349 0.072 4.831 0.000

Latent Class 2

ITEM1

Category 1 0.092 0.128 0.714 0.475

Category 2 0.908 0.128 7.069 0.000

ITEM2

Category 1 0.663 0.092 7.209 0.000

Category 2 0.337 0.092 3.671 0.000

ITEM3

Category 1 0.933 0.020 46.344 0.000

Category 2 0.067 0.020 3.314 0.001

ITEM4

Category 1 0.935 0.026 35.705 0.000

Category 2 0.065 0.026 2.502 0.012

ITEM5

Category 1 0.781 0.081 9.674 0.000

Category 2 0.219 0.081 2.715 0.007

ITEM6

Category 1 0.680 0.065 10.545 0.000

Category 2 0.320 0.065 4.954 0.000

ITEM7

Category 1 0.887 0.025 35.583 0.000

Category 2 0.113 0.025 4.525 0.000

ITEM8

Category 1 0.860 0.040 21.668 0.000

Category 2 0.140 0.040 3.524 0.000

ITEM9

Category 1 0.675 0.055 12.314 0.000

Category 2 0.325 0.055 5.926 0.000

Latent Class 3

ITEM1

Category 1 0.688 0.325 2.113 0.035

Category 2 0.312 0.325 0.959 0.337

ITEM2

Category 1 0.836 0.051 16.240 0.000

Category 2 0.164 0.051 3.187 0.001

ITEM3

Category 1 0.964 0.019 51.983 0.000

Category 2 0.036 0.019 1.926 0.054

ITEM4

Category 1 0.944 0.018 53.820 0.000

Category 2 0.056 0.018 3.195 0.001

ITEM5

Category 1 0.956 0.053 18.115 0.000

Category 2 0.044 0.053 0.841 0.400

ITEM6

Category 1 0.817 0.062 13.139 0.000

Category 2 0.183 0.062 2.942 0.003

ITEM7

Category 1 0.902 0.023 38.929 0.000

Category 2 0.098 0.023 4.214 0.000

ITEM8

Category 1 0.890 0.025 35.692 0.000

Category 2 0.110 0.025 4.406 0.000

ITEM9

Category 1 0.812 0.052 15.573 0.000

Category 2 0.188 0.052 3.601 0.000

LATENT CLASS ODDS RATIO RESULTS

(Est. - 1) Two-Tailed

Estimate S.E. / S.E. P-Value

Latent Class 1 Compared to Latent Class 2

ITEM1

Category > 1 1.215 2.408 0.089 0.929

ITEM2

Category > 1 2.363 1.643 0.830 0.407

ITEM3

Category > 1 10.394 5.390 1.743 0.081

ITEM4

Category > 1 10.247 4.890 1.891 0.059

ITEM5

Category > 1 11.630 7.224 1.472 0.141

ITEM6

Category > 1 1.895 1.093 0.819 0.413

ITEM7

Category > 1 8.264 3.690 1.968 0.049

ITEM8

Category > 1 9.998 5.767 1.560 0.119

ITEM9

Category > 1 1.113 0.478 0.237 0.813

Latent Class 1 Compared to Latent Class 3

ITEM1

Category > 1 26.518 51.201 0.498 0.618

ITEM2

Category > 1 6.133 3.565 1.440 0.150

ITEM3

Category > 1 20.060 13.020 1.464 0.143

ITEM4

Category > 1 12.094 5.692 1.949 0.051

ITEM5

Category > 1 70.304 83.109 0.834 0.404

ITEM6

Category > 1 3.977 2.372 1.255 0.210

ITEM7

Category > 1 9.707 4.707 1.850 0.064

ITEM8

Category > 1 13.173 9.847 1.236 0.216

ITEM9

Category > 1 2.317 1.060 1.243 0.214

Latent Class 2 Compared to Latent Class 3

ITEM1

Category > 1 21.819 24.935 0.835 0.404

ITEM2

Category > 1 2.595 1.090 1.463 0.143

ITEM3

Category > 1 1.930 1.143 0.814 0.416

ITEM4

Category > 1 1.180 0.751 0.240 0.810

ITEM5

Category > 1 6.045 5.641 0.894 0.371

ITEM6

Category > 1 2.098 0.879 1.250 0.211

ITEM7

Category > 1 1.175 0.473 0.369 0.712

ITEM8

Category > 1 1.318 0.685 0.464 0.643

ITEM9

Category > 1 2.081 0.634 1.705 0.088

```````````````````````````````````````````````````````````````````````````````````````````````````````````````

Interpretation: These results shows how each variable is used for identify the latent class. For each variable, the estimated effect each variable (set category 2 as the reference) in determining whether the observation belong to the corresponding class or not is first provided. More significant the effect of variables is, the more useful is the variable in identifying the class. For example, the item 1 is the one most significant for class 1, while item 2, 3,7, and 8 are the ones must useful for class 2.

Then, the probability scale result indicates that for each latent class, which is the probability that each item should be look like. For example, if the observation is belonging to latent class 1, the first item should be much more likely to be in the second category. This result is consistent with the estimated effect (aforementioned) and the likelihood ratio test. More significant effect of the item leads to more significant result in the likelihood ratio test, lead to more uneven probability across difference category. Finally, the tables are the likelihood ratio tests.

* Technical 11 Output and Technical 14 Output

```````````````````````````````````````````````````````````````````````````````````````````````````````````````

Random Starts Specifications for the k-1 Class Analysis Model

Number of initial stage random starts 20

Number of final stage optimizations 4

VUONG-LO-MENDELL-RUBIN LIKELIHOOD RATIO TEST FOR 2 (H0) VERSUS 3 CLASSES

H0 Loglikelihood Value -4251.208

2 Times the Loglikelihood Difference 39.025

Difference in the Number of Parameters 10

Mean 20.256

Standard Deviation 22.225

P-Value 0.1457

LO-MENDELL-RUBIN ADJUSTED LRT TEST

Value 38.468

P-Value 0.1500

Random Starts Specifications for the k-1 Class Analysis Model

Number of initial stage random starts 20

Number of final stage optimizations 4

Random Starts Specification for the k-1 Class Model for Generated Data

Number of initial stage random starts 0

Number of final stage optimizations for the

initial stage random starts 0

Random Starts Specification for the k Class Model for Generated Data

Number of initial stage random starts 40

Number of final stage optimizations 8

Number of bootstrap draws requested Varies

PARAMETRIC BOOTSTRAPPED LIKELIHOOD RATIO TEST FOR 2 (H0) VERSUS 3 CLASSES

H0 Loglikelihood Value -4251.208

2 Times the Loglikelihood Difference 39.025

Difference in the Number of Parameters 10

Approximate P-Value 0.0000

Successful Bootstrap Draws 49

```````````````````````````````````````````````````````````````````````````````````````````````````````````````

Interpretation: entropy indicates the overall information in the membership distribution. The classification probabilities tables show that most observations are classified into the class, which should be the most likely class. In particular, 90% of the class two observations are indeed the one which are most likely to be in class 2. This indicates a high confidence of the result. Lo-Mendell-Rubin adjusted LRT test and Parametric Bootstrapped Likelihood Ratio test are used for identify whether the proposed model is significantly better fit than the less complex model with fewer latent class. Based on the p-value, the result is not consistent for the two tests.

# Problem 2: Research Proposal

## Purpose & Research Question

The purpose of the proposed project idea is to investigate the empirical typology of early childhood program participation in United States. Previous study usually classifies the participation of early childhood education based on the specific program the child attend: care from relatives, care from non-relatives, care from day care centers or Preschool programs. However, there may be some other possibility of classify the early childhood education based on parents’ value of choosing care for their children. There are four main research questions that could answered by the three-step latent class analysis model: 1) Does there exist latent classes for the believe of early childhood education from parents? 2) What is the possible meaning of the latent classes? And how the latent classes related to their choose of care program (i.e., non-relatives, care from day care centers or Preschool programs)? 3) How children and family background related to the estimated latent classes? 4) How does the background of children and the latent class together effect children’s learning?

## Literature Review and Theory

Educators and policymakers are pressed to respond to the current crisis in American public education. These deficiencies are most severe in large, low-income urban settings, where children are living in neighborhoods that are characterized by a disproportionate number of family and community risk factors and overburdened school and public service agencies (Children's Defense Fund, 1997; U.S. Department of Education, 1996). Concerns about the crisis in public education have led to the establishment of the National Education Goals (U.S. Department of Education, 1992). The goals represent a strategic plan to enhance learning opportunities for all students by targeting what experts believe to be the most fundamental components of effective intervention. One of the cardinal foci of this strategic national plan is school readiness. This goal and its accompanying objectives highlight the need for quality early childhood educational programs and emphasize the importance of establishing and maintaining substantial parent involvement to promote student learning. However, there still limit information from the empirical typology analysis about the early childhood education.

## Data

The National Household Education Surveys Program (NHES), the flagship household survey program of the National Center for Education Statistics, collects nationally representative, descriptive data on the educational activities of children and adults in the United States. Data about young children's care and education before kindergarten are collected in the NHES's Early Childhood Program Participation Survey (ECPP). This survey collects data about children from birth through age 6 who are not yet enrolled in kindergarten. The surveys typically focus on children's participation in and barriers to participating in nonparental care arrangements; what activities the family does with the child, such as reading, singing, and arts and crafts; and what the child is learning, such as counting, recognizing the letters of the alphabet, and reading. Parents are the respondents. In ECPP, there are several subdomains, including parents choose of care of the children, children and family background, children’s learning status, children’s health condition, and detailed information about the care program that children participated.

## Method

Three-step latent class analysis can be used for the purpose of this project.

* The indicators (parents’ perceptions of what is importance for the early childhood education program): There are 10 indicators in total. The original indicators are all four-level Likert scale from not at all important to very important. I will dichotomize the indicators as not important and important.
* Contextual background (family and children’s background): children’s race (CHISPAN, CAMIND, CASIAN, CBLACK, CPACI, CWHITE), children’s sex (CSEX), children’s overall health (HDHEALTH), whether the child has any kinds of disability (HDINTDIS, HDSPEECHX, HDDIXTRBX, HDDEAFIMX, HDBLINDX, HDORTHOX, HDAUTISMX, HDPDDX, HDADDX, HDLEARNX, HDDELAYX, HDTRBRAIN, and HDOTHERX), number of people living in the family (HHTOTALXX), frequency of using Internet (USEINTRNT), total income of the family (TTLHHINC), house type (OWNRNTHB). Data transformation may need for these background variables since there may have collinearity issues.
* Distal outcome (children’s learning status): color identification (DPCOLOR), letter reorganization (DPLETTER), count ability (DPCOUNT), whether kid can write the first word even backward (DPNAME).