An Exploration of Latent Class Analysis for Data Representation

Andrew J. Otis & Hsing Yu Chen

COMP 4441

March 18th, 2022

Table of Contents

SUMMARY

Our group utilizes Latent Class Analysis(LCA) as a statistical model to determine if distinct unobservable classes exist within a dataset utilized in a study by Bertrand and Mullainathan(2004).

LCA is model in which individuals can be classified into mutually exclusive and exhaustive events known as latent classes, based on their pattern of response and measured with categorical variables.

Analysis was carried out in Rstudio using the poLCA function which estimates latent class models for polytomous outcomes. It can also estimate latent class regression models, but these results were not utilized in the analysis. Ultimately, the results were determined to not be valid; but served a pedagogical purpose to demonstrate the process of LCA

INTRODUCTION

The cross-sectional data was originally used as part of a larger study to aid researchers in determining if an individual's implied ethnicity had an effect on their chances of progressing through the application process, specifically African American or Caucasian. The data was from employer ads posted in Chicago and Boston, where the information on the resumes is randomly generated and assigned. This information included things like gender, ethnicity, and name, plus the information posted on the job ads. Our group is simply utilizing the dataset described to perform LCA. Even we wanted to perform a similar task to that of the researchers, LCA would not be the model to use to make predictions; it is a model for data representation.

The data source has come cleaned in columns separating the types of responses in the data set. For LCA, it was decided that we would look at the columns "equal"=employer offers equal opportunity, "reqeduc"=education requirement, "reqcomp"=computer competency requirement, "reqorg"=organizational requirement, "ethnicity"=implied ethnicity based on the sound of name, and "call"=whether or not the applicant proceeded in the application process. The columns chosen are details regarding requirements of the job in question and a few person-centered details regarding the applicant.

```
vars <- c("equal", "reqeduc", "reqcomp", "reqorg","call", "gender","ethnicity")
for (i in 1:length(vars)){
    ResumeNames[,vars[i]] <- ResumeNames[,vars[i]]+1
}

raw_data <- read_excel("raw.data.xlsx")
raw_data
```

```
## # A tibble: 5 × 6
##   ethnicity call  equal reqeduc reqcomp reqorg
##   <chr>     <chr> <chr> <chr>   <chr>   <chr>
## 1 cauc      no    yes   no      yes     no
## 2 cauc      no    yes   no      yes     no
## 3 afam      no    yes   no      yes     no
## 4 afam      no    yes   no      yes     no
## 5 cauc      no    yes   no      yes     yes
```

The isolated data then has responses transformed into 1's and 0's for probabilistic algorithms to run. These responses may be binomial (i.e. only two outcomes) or multimodal(i.e. more than two outcomes

```
trans_data <- read_excel("trans.data.xlsx")
trans_data
```

```
## # A tibble: 5 × 7
##   equal reqeduc reqcomp reqorg  call gender ethnicity
##   <dbl>   <dbl>   <dbl>  <dbl> <dbl>  <dbl>     <dbl>
## 1     1       0       1      0     0      0         1
## 2     1       0       1      0     0      0         1
## 3     1       0       1      0     0      0         0
## 4     1       0       1      0     0      0         0
## 5     1       0       1      1     0      0         1
```

We must now ask “does the data satisfy the requirements of LCA?”, this is important regarding the validity of our results.

The Assumptions of LCA are as follows,

1. Data is non-parametric
2. Data is categorical
3. Observations in each class must be independent from one another

The way our data is currently structured, we would meet the first two assumptions of LCA, however the data set fails to meet the final assumption of LCA. Due to a lack of company id present in responses, we cannot guarantee that responses in each class determined are independent of one another. For example, one company having more than one ad or more than one resume being sent in response to the same ad. Therefore, the results cannot be considered valid. For pedagogical purposes we continued with the analysis to demonstrate the process of LCA.

SUMMARY STATISTICS

```
classes_summary <- read_excel("classes_summary.xlsx")
classes_summary
```

```
## # A tibble: 6 × 6
##   classes    G2    X2    df    AIC    BIC
##     <dbl> <dbl> <dbl> <dbl>  <dbl>  <dbl>
## 1       2 391.  417.    112 32325. 32442
## 2       3 162.  150.    104 32112. 32261.47
## 3       4 109.   93.3    96 32075. 32277.52
## 4       5  78.7  66.4    88 32061. 32314.83
## 5       6  57.5  44.1    80 32055. 32360.48
## 6       7  42.9  33.3    72 32057. 32414.88
```

```
##  Chi-squared test for given probabilities
##
## data:  equal.c3.obs
## X-squared = 3669.1, df = 2, p-value < 2.2e-16

##  Chi-squared test for given probabilities
##
## data:  reqeduc.c3.obs
## X-squared = 3308.8, df = 2, p-value < 2.2e-16

##  Chi-squared test for given probabilities
##
## data:  reqcomp.c3.obs
## X-squared = 1382.1, df = 2, p-value < 2.2e-16

##  Chi-squared test for given probabilities
##
## data:  reqorg.c3.obs
## X-squared = 4602.7, df = 2, p-value < 2.2e-16

##  Chi-squared test for given probabilities
##
## data:  call.c3.obs
## X-squared = 9.1187, df = 2, p-value = 0.01047

##  Chi-squared test for given probabilities
##
## data:  gender.c3.obs
## X-squared = 1247.8, df = 2, p-value < 2.2e-16

##  Chi-squared test for given probabilities
##
## data:  ethnicty.c3.obs
## X-squared = 1.2488, df = 2, p-value = 0.5356
```

VISUALS

Figure 1

```
## f1 <- as.formula(cbind(equal, reqeduc, reqcomp, reqorg, call, gender, ethnicity)~1)

## LCA3 <- poLCA(f1, data=ResumeNames, nclass=3, maxiter=50000, nrep=5)

## plot(LCA3)
```

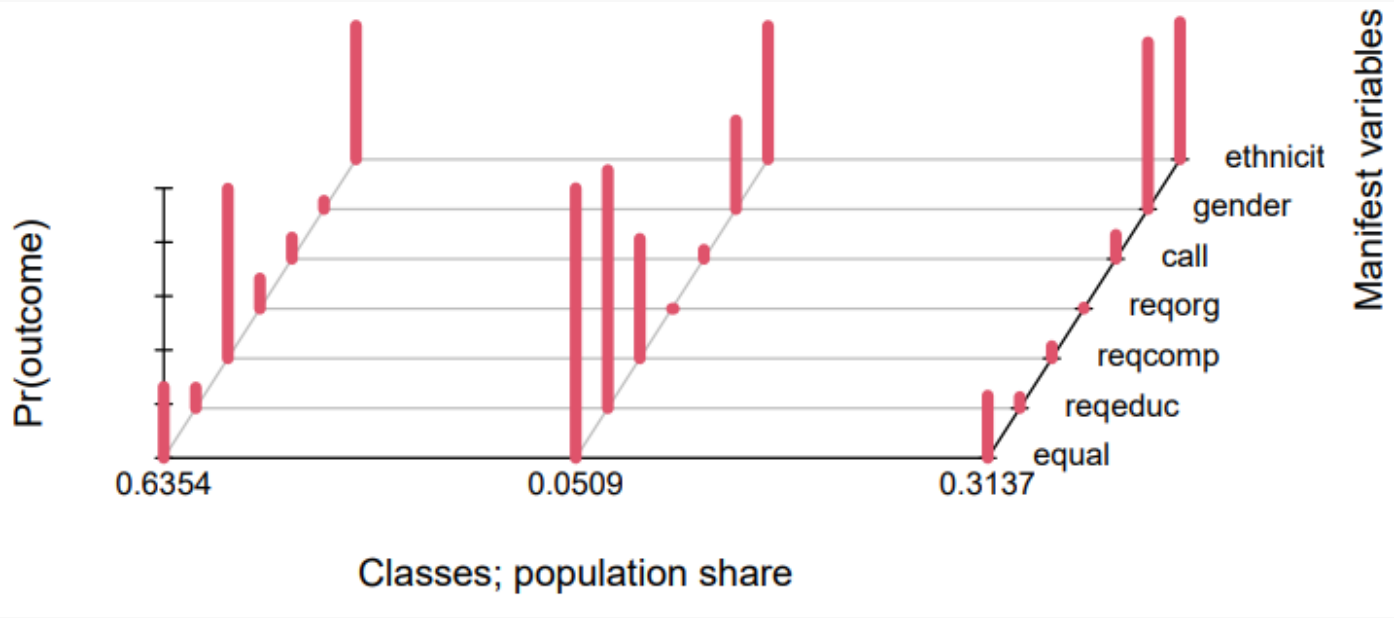

Fig. 1: The bar graph representation of the data from Table 1. It is a 3-class LCA, with a maxiter=50000 and nrep=5, where "maxiter"=the number of iterations for the algorithm to cycle and "nrep"=the number of times the model is ran.

Table 1

	Class 1 (74.97%)	Class 2 (5.91%)	Class 3 (19.12%)
Item	**Probability of "Yes"**		
equal	73.61%	***0%***	76.86%
reqeduc	92.33%	***11.73%***	95.81%
reqcom	***37.02%***	55.62%	95.4%
reqorg	88.79%	100%	99.53%
call	92.04%	96.57%	91.02%
gender	**96.87%**	67.09%	***38.12%***
ethnicity	50.43%	50.5%	49.05%

Table 1: For the 3-class LCA, directly under class number we have the probability of an individual's membership in a particular class. Below that we have the conditional probability of a particular response, given an individual's classification.

CONCLUSION

The LCA model chosen was on the basis of the Bayesian Information Criterion (BIC) for estimating how well a given LCA model would fit our dataset.

$$BIC = kln(n) - 2\ln(\hat{L})$$

BIC prefers models where the number of samples (n), far outnumber the number of parameters (k) and in our case, a lower BIC value is indicative for the best fit model. Referring to the Summary Statistics section of the paper, we can observe BIC stop decreasing and begin to increase during the transition from 3 to 4 class LCA models. Allowing us to conclude that a 2 or 3-class LCA model would be best fitted for the data set.

Recall, the original research question was if the data set can be split into distinct groups and to model callback response to possible gain some insight on that particular variable.

Now let's interpret the results of our 3-class LCA model. Referring to table 1, we can observe that no matter what class an individual is in, they have higher than a 90% chance of progressing through the application process. We can also observe that there exist other item differences that could possibly imply a distinction of classes. The resulting classes can be described as follows,

Class 1: Are jobs that rarely require computer skills and consisting of mostly male applicants

Class 2: Are Jobs that don't offer equal opportunity employment and rarely have an education requirement

Class 3: Are jobs whose applicant pool consisted of mostly female

It was determined in the introduction that the data set does not suit itself very well for LCA, for further confirmation, a Chi Squared test was ran regarding each item in Table 1 to compare the 3-Class LCA model values to the actual observed values collected from our data source. Referring to the summary statistics questions, we can observe that every single item was not very well represented by the model; except for "gender"; confirming the lack of validity of our model's results. Still, we hope the process of LCA makes a bit more sense and the implications it has on the field of statistics.

LCA is a model that is widely used to represent data in sociological and mental health settings. Such as the classification of drinking groups, or types of voters, etc. While we did not name our classes, doing so presents researchers with the unethical choice of purposefully giving names to classes that inaccurately represent class membership.

REFERENCES

1. Bertrand, M. and Mullainathan, S. (2004). Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination. American Economic Review, 94, 991–1013.

2. Durso, C. (2022), COMP4441 – Probability and Statistics 1.

3. P., B. (2014, October 31). Latent Class Analysis vs. Cluster Analysis - differences in inferences? Cross Validated. Retrieved March 7, 2022.

4. Petersen, K. J. (2019, May 29). The Application of Latent Class Analysis for Investigating Population Child Mental Health: A Systematic Review. Frontiers. Retrieved March 3, 2022.

5. Pratt, B. (2020, January 14). Latent Class Analysis Using R. Office of Population Research. Retrieved March 3, 2022.

6. Websites DataCamp. poLCA function - RDocumentation. RDocumentation. Retrieved March 3, 2022

7. Weller, B. E., Bowen, N. K., & Faubert, S. J. (2020). Latent Class Analysis: A Guide to Best Practice. Journal of Black Psychology, 46(4), 287–311.

8. Wikipedia contributors. (2022, January 18). Bayesian information criterion. Wikipedia. Retrieved March 5, 2022.