Study on LCA (Latent Class Analysis)

Id: 18200537 **Student Name : Prashanth Guntal**

Introduction

This report is providing the details analysis of clustering method, namely Latent Class Analysis. Here for our analysis we are using the *2012 Perception of Science and Technology* data for understanding on how LCA. We are using this dataset to cluster the respondents to uncover groups with similar attitudes to Science and Technology. Data has 392 observations and 7 different variables. We are using LCA for analyzing the relationships among the manifest data when some variables are unobserved. The unobserved variables are categorical, allowing the original data set to be segmented into a number of exclusive and exhaustive subsets: the latent class. Since LCA is based upon statistical model we can classify observations based on the posterior probabilities of class membership. Of the available 392 observations we don't have any observations with missing values/NA in any variable. Here we are performing our LCA analysis using the poLCA function. poLCA function returns error if any of the manifest variables contain zero's, negative or decimal values. Since our data is binary data we are starting our analysis report by first performing data preprocessing and then perform LCA on the processed data and detail analysis on output measurements.

Data Preprocessing

As mentioned above since our data is binary we are need to change to positive integer values. In our dataset we have seven variables of which four variables are positively worded (Comfort, Work, Future, Benefit) and negative worded (Technology, Environment, Industry). Since in our data "strongly disagree" and "disagree" for our negative variables are coded as 1, which means that respondent had positive opinion on Science and Technology. In our data respondents voted 1 for positive variables, hence in our data 1's indicate respondents had positive opinion on Science and Technology and 0's indicating respondents had negative opinion on Science and Technology. Since poLCA accepts only non-zero integer values we will add one to all the values in the dataset. Now, 1's indicated respondents having negative opinion and 2's indicated respondents have positive opinion.

LCA

Here we will apply poLCA function on processed data. An initial exploratory latent class analysis was conducted to narrow done the number of clusters, starting from number of clusters 2 to 10 so that we get different 9 solution and we compare which would be better solution for our analysis. Lower the BIC/AIC value the best is the model. *Table – 1* shows the log-likelihood, residual degree of freedom, BIC, G-Squared valued for six clustering solution starting from number of clusters 2 to 7. Similarly, lower the Chi-Squared values better the model, however no fewer than 10-20% of cells should contain fewer than five observations. So, going forwards we are analyzing the output of model with three cluster. The maximum log-likelihood of the model estimated is -1394.372.

Estimated Class-Conditional Probabilities:

Fig-1 shows the output estimated class-conditional probabilities for first two variables. First the estimated class conditional probabilities for each variable (Words), negative in the first column (Pr(1)) and positive in the second column(Pr(2)). Example, a respondent value belonging to third class has 0% chance of being rating negatively and 100% chance being providing positive rating to Science and Technology considering Comfort variable. Similarly, a respondent value belonging to second class has ~20% chance of being rating negatively and ~80% chance of being rating positive to Science and Technology considering Comfort variable and so forth.

Estimated mixing proportion(p(r)):

p(r) corresponds to the share of observations to each latent class. *Fig-2* shows the percentage of observations belonging to each class. Since there is a strong agreement between our estimated class population shares and Predicted class memberships we can say our model is good fit for the data. Values, 0.18, 0.33, 0.48 are the co-variance of the above estimated mixing proportion for each class.

Fit for 3 latent class:

The next set of output result shows that all the 392 observations where considered. Number of parameters estimated where 23. Since the number of degrees of freedom is non-negative our model won't return any output warnings. Latent class model is identified only of the number of estimated parameters is much less than total number of observations. Our estimated parameters are 23 which is much less that total number of observations (392), hence latent class model is identified.

Entropy of a fitted latent class model:

Entropy is a measure of concentration in a probability mass function. The entropy of the estimated probability mass function for our model with three clusters is 3.558.

Goodness of Fit:

As mentioned above the minimum the BIC and AIC value the better is the model. *Table-1* shows the different BIC for different number of class. With two classes, the BIC is 2946.75; with three classes and BIC is 2926.084 and with four clusters BIC is 2962.640. *Fig-3* shows us how BIC value is varying for different models, based on the this we can conclude that model with three classes is better than others for the Science and Technology data.

Clustering Uncertainty:

Here we are going to understand on uncertainty in the group/class assignment. Uncertainty of a group is calculated using the formula; (1 – max {conditional probability of observation *i* belonging to a particular group}). 0.2048, 0.0567, 0.044, 0.066, 0.1516 are the clustering uncertainty values for our first five observations.

Conclusion:

The above analysis appears to suggest that for the data set that doesn't have predefined classes/clusters, we need to try for different number of clusters/class so that we can decide what number of classes would be best for model based on few of the criteria's such as goodness fit test and other criteria as mentioned in the above report. If we are going to cluster using hierarchical clustering we would have faced difficulties in finding the parameters to provide us that number of chosen classes/clusters is best fit model. Since LCA provides us an easy access to BIC and AIC values which are used to access the better model, we would prefer LCA is better than hierarchical clustering in analyzing the binary data. From our above report we conclude saying that since the BIC and AIC of three class cluster is better than other models with different number of clusters, even the Chi-Squared test suggest the same. However, there is some minor uncertainty in the classes as shown in above. Our model had much less parameters to estimate than number of observations. poLCA provides us cell percentages and posterior probabilities of latent class membership. poLCA provides good visualization has shown in output of the RMD file. Considering all the above points, since our analysis is only limited to this dataset we conclude saying that LCA with three classes is better than other number of clusters. However, if we try to work on other dataset we can find LCA with more than three clusters would be better based on the chosen data set.

Plots and Tables:

	Modell	log_likelihood	df	BIC	Chi
1	Modell 2	-1428.592	112	2946.752	185.24800
2	Modell 3	-1394.372	104	2926.084	110.81916
3	Modell 4	-1388.766	96	2962.640	100.87735
4	Modell 5	-1382.971	88	2998.821	69.00209
5	Modell 6	-1380.196	80	3041.042	91.84977
6	Modell 7	-1375.277	72	3078.973	52.37098

Table – 1

```
$Comfort
             Pr(1)  Pr(2)
class 1:    0.1462 0.8538
class 2:    0.2023 0.7977
class 3:    0.0000 1.0000

$Environment
             Pr(1)  Pr(2)
class 1:    0.6146 0.3854
class 2:    0.2794 0.7206
class 3:    0.1998 0.8002
```

Fig-1

```
Estimated class population shares
 0.1865 0.3318 0.4817

Predicted class memberships (by modal posterior prob.)
 0.2092 0.273 0.5179
```

Fig-2

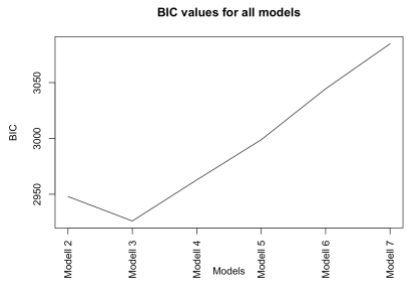

Fig-3