Research in Social and Administrative Pharmacy 13 (2017) 1196–1201

Contents lists available at ScienceDirect

Research in Social and Administrative Pharmacy

journal homepage: www.rsap.org

Latent Class Analysis: An example for reporting results

James B. Schreiber

Duquesne University, USA

ARTICLE INFO

Article history:
Received 22 November 2016
Accepted 25 November 2016

Keywords:
Latent Class Analysis
Reporting results

ABSTRACT

Objective: The purpose of this paper is to provide a brief non-mathematical introduction to Latent Class Analysis (LCA) and a demonstration for researchers new to the analysis technique in pharmacy and pharmacy administration. LCA is a mathematical technique for examining relationships among observed variables when there may be collections of unobserved categorical variables. Traditionally, LCA focused on polytomous observed variables, but recent work has extended the types of data that can be utilized. Included in this introduction are basic guidelines for the information that should be part of a manuscript submitted for review. For the analysis, LatentGold is used, but I also include basic R code for running LCA and LC Regressions with the poLCA package.

1. Introduction

Latent classes are unobserved, or latent, segments. Participants, or more generally, cases, within the same latent class are considered homogenous based on certain pieces of information. Latent Class Analysis (LCA) was developed over 60 years ago as a way to characterize latent variables while analyzing dichotomous items.[1] During the past several years, it has expanded to all for all types of data. In the literature, LCA is referred to in different ways. It has been called Latent Structure Analysis,[2] Mixture Likelihood Clustering,[3,4] Model Based Clustering,[5–7] Mixture-model Clustering,[8] Bayesian Classification,[9] and Latent Class Cluster Analysis.[10]

LCA allows researchers to create or characterize a multidimensional discrete latent variable based on a cross-classification of two more observed categorical variables.[2] Because of the categorical nature of the latent classes, LCA is different from other latent approached such as factor analysis and structural equation modeling. LCA provides the possibility to develop typologies for understanding and, if desired, can be used in predictive models. In addition to the ability to analyze relationships in categorical data, numeric and non-numeric data can be utilized.

Classes or clusters are a common need for researchers and data analysists in general. K-means is a common technique used, but has several problems. The potential difficulties include sensitivity to outliers (outliers are extreme values that can skew the results) as well as the need to use interval or ratio data—which means that, in calculating distances, you have to know whether the numbers actually add up—and there are some concerns about the order in which data is assembled.[11] In some cases, data may not be appropriate for the K-means method. More fundamentally, the stability of clusters cannot be assumed because traditionally there has been no objective set of criteria to judge the suitability of solutions. K-means will always produce a solution, and *some of those solutions are likely to fit your expectations.*

LCA quite easily overcomes all of the problems with K-means clustering. The increase of computing power in the 1990s made LCA a very efficient technique. The best way to distinguish between LCA and cluster analysis techniques is to note that LCA is *model-based* and cluster analysis is not. “Model-based,” means there is a statistical model that is assumed to come from the population from which the data was gathered.[11] Both cluster and LCA are seeking divisions that maximize the between-cluster differences and minimize the within-cluster differences. But in traditional cluster analysis this decision is arbitrary or subjective. In LCA, a statistical model allows the comparison to be statistically tested, so that the decision to adopt a particular model is less subjective, or at least has some grounding for comparison. In addition, the items used in the analysis do not need to have the same scale or have equal variances. Finally, LCA allows for the examination of the residuals between items used in the analysis. In other words, LCA is useful in examining the data that does not fit the model or models, thus allowing the analyst to judge the overall quality.

A simulation study to compare K-means analysis and LCA against discriminant function analysis with known groups, a method generally considered to be the “gold standard,” tested how

http://dx.doi.org/10.1016/j.sapharm.2016.11.011
1551-7411/

well variables predict group membership.[10] In the study, group membership was known in advance and the authors applied the three methods to the data to examine classification performance. They argued that they used data that favored K-means analysis. Even so, the results of the comparison showed that K-means had an 8% misclassification rate versus 1.3% for LCA.[10]

The remainder of this article focuses on procedures for running LC models, decision points, and what should be included in a manuscript for review.

2. Example

2.1. Sample

This data comes from a larger data set with military and non-military personnel. The full data set involved 1100 individuals who completed several surveys (e.g., Reading Interests, Cognition, along with some pattern recognition tasks, such as the Raven's Matrices). The data was never published and is quite old now, but did provide the foundation for other studies. For this example, there are five hundred participants that provided demographic data, some health history data, and depression data (Center for Epidemiological Studies Depression Scale, CESD). Basic Demographics are in Table 1.

2.2. Analysis step 1

Given the nature of this article, there is no theoretically expected number of clusters. Therefore, an initial run of 1–5 clusters will be analyzed, with seven variables of interest. Latent Gold (v. 5.1.0.16288),[12] will be used for this example. In Appendix A, R code is provided for the poLCA package[13] for running LCA models along with an explanation guide of the code based on the poLCA code book.[14]

The variables of interest are:

Race-Three categories (Black-Non-Hispanic, Hispanic, White-Non-Hispanic).

Education Five categories-No High School Diploma to Bachelor).

Currently Working- Dichotomous- Yes (1), No (0).

Marital Status- Five Categories (Married, Separated, Divorced, Widow, Never).

Poverty Level- Dichotomous- Below (1), Above (0).

Poor Health- Dichotomous- Poor Health (1), Good Health (0).

CESD (Score Category- Three Categories (Low Risk, Moderate Risk, High Risk).

From the five models, there are several pieces of information which can be examined to begin to determine a useful number of classes. Table 2 provides evaluative information. In general, the initial analysis of the possible models focuses first on three evaluative indicators: LL, p-value, and BIC. LL is Log Likelihood, the logarithm of the likelihood ratio, a test that compares the fit of two models by examining how much more likely the data are predicted by one model compared to the other. The Log Likelihood can be used to compute a p-value, the main measure of statistical significance. BIC is the Bayes Information Criterion, which is a statistic created to aid model selection by penalizing the number of factors in a model. There are other indicators available such as AIC (Akake Information Criterion) and CAIC, but for explanation purposes, the displayed indicators work well. Recent simulation work has indicated that AIC and CAIC are not good indicators for choosing the correct number of *k* classes.[15]

Looking at these statistics, the two-cluster model might be the best fit because it has the lowest BIC value. But the log-likelihood value for the five-cluster model is the lowest. Thus, other pieces of information may be useful. The p-values of 0.99 and 1 indicate three candidates for best fit, the three-, four- and the five-cluster models. Npar, which is number of parameters in the model can also be part of the evaluation for all models where $p > 0.05$. There are other pieces of information discussion below that could also be used such as a residual analysis and the pseudo R-Squared values.

A boot-strap comparison within LatentGold can be run to statistically test the different cluster models. The results from the comparison indicate that model 4 is better than 3, but model 5 is not better than 4. This function allows the researcher to examine which model is best and acts as a guard against picking one because it appears to confirm the current theory or belief. It is important not to reify our models.[16] At this point, a deeper examination of model 4 is warranted.

2.3. Step 2: specific model examination

There are multiple pieces of information to examine for a specific model. For any model, error examination is important and in LCA an examination of bi-variate residuals is a first step. The bi-variate residual is similar to a Pearson chi-squared value divided by the degrees of freedom. The chi-square calculation is computed from the expected counts in the traditional two-way tables from the estimated model.[12] For 1° of freedom, residual values greater than 3.84 are statistically significant at the 0.05 level. At the value of 3.84, the model is not explaining the relationship between the variables well. From a modeling perspective, a value around 2.00 might be an indicator that the model (number of clusters chosen) is not explaining the relationship well. If one the relationships was above 2, the relationship can be set as a direct effect, "fixed," and the residual will be zero or close to zero in the re-analysis.

In Table 3, the bi-variate residuals from the four-class model are provided. None of the residuals reach two, so with this information, the four cluster-model appears to work well. In the three-cluster model, the bi-variate residual between marital status and race

Table 1
Descriptive statistics.

	Percent		Percent
Race		**Poor health**	
White/other, not Hispanic	78.3	Not in poor health	93.4
Hispanic	13.7	In poor health	6.6
Black, not Hispanic	8	**Poverty Status**	
Educational attainment		Below poverty	73.9
No high school diploma	46.6	Above poverty	26.1
Diploma or GED	49.4	**CESD Categories**	
AA degree	3.7	Low Risk	46.1
BA degree	0.3	Moderate Risk	23.6
Currently employed?		High Risk	30.3
No	55.6	**Marital status**	
Yes	44.4	Married	6.4
		Separated	11.2
		Divorced	16.6
		Widowed	1.7
		Never married	64.2

Table 2
Model fit evaluation information.

	LL	BIC(LL)	Npar	L^2	df	p-value	Class error
1-Cluster	−2517.52	5121.89	14	610.79	480	4.60E-05	0
2-Cluster	−2434.41	5005.28	22	444.57	472	0.81	0.09
3-Cluster	−2412.31	5010.70	30	400.37	464	0.99	0.21
4-Cluster	−2401.90	5039.49	38	379.54	456	1	0.18
5-Cluster	−2396.88	5079.08	46	369.50	448	1	0.19

*BIC for LCA models is a good indicator for which model to choose.[15]

Table 3
Bivariate residuals from four-class model.

Indicators	1	2	3	4	5	6
1. Race						
2. Education	0.87					
3. Currently Working	0.35	1.21				
4. Marital Status	1.82	0.92	0.13			
5. Below Poverty Line	0.16	0.00	0.16	0.07		
6. Poor Health	0.62	0.43	0.71	0.56	0.01	
7. CESD	0.05	1.77	0.29	0.25	0.01	0.34

Table 4
Four class model fit regressions for each variable.

	Class 1	Class 2	Class 3	Class 4	Wald	p-value	R^2
Race	−0.54	0.28	−0.15	0.41	9.79	0.02	0.05
Education	−0.13	−0.70	0.58	0.25	19.25	0.00	0.06
Currently Working	−1.37	−2.08	2.41	1.03	24.68	0.00	0.34
Marital Status	0.65	0.12	0.39	−1.16	18.38	0.00	0.29
Below Poverty Line	−2.99	−1.08	3.28	0.79	14.13	0.00	0.61
Poor Health	−3.70	4.10	−3.10	2.70	3.74	**0.29**	0.13
CESD-Category	0.20	2.16	0.05	−2.42	29.69	0.00	0.41

The *p*-value (in bold) indicates that Poor Health is not helping to discriminate between the clusters. The other variables are all less than 0.05, thus indicating that they are helping to discriminate among the clusters. Thus, the variable Poor Health could be considered for removal.

Table 5
Cluster/class response percentages within variables.

	Cluster1	Cluster2	Cluster3	Cluster4
Class Size	0.46	0.28	0.19	0.07
Indicators				
Race				
White/other, not Hispanic	0.86	0.67	0.79	0.63
Hispanic	0.10	0.18	0.14	0.20
Black, not Hispanic	0.04	0.15	0.07	0.18
Education				
No high school diploma	0.47	*0.61*	0.29	0.37
Diploma or GED	0.50	*0.37*	0.62	0.57
AA degree	0.03	*0.01*	0.08	0.05
BA degree	0.00	*0.00*	0.01	0.00
Currently Working				
No	**0.67**	0.81	0.04	0.16
Yes	0.33	0.19	0.96	0.84
Marital Status				
Married	0.01	0.06	0.03	0.48
Separated	0.05	0.16	0.09	0.36
Divorced	0.12	0.22	0.17	0.14
Widowed	0.01	0.01	0.01	0.00
Never married	**0.81**	0.54	0.70	0.03
Mean	4.57	3.82	4.27	1.74
Below Poverty Line				
Below poverty	0.97	0.85	0.07	0.46
Above poverty	0.03	0.15	0.93	0.54
Poor Health				
Not in poor health	1.00	0.81	1.00	0.94
In poor health	0.00	0.19	0.00	0.06
CESD Category				
LE 15 low risk of depression	0.58	0.06	0.63	0.97
16-23 mod risk of depression	0.27	0.19	0.25	0.03
GE 24 high risk of depression	0.15	0.75	0.12	0.00

*This table is commonly termed Profile of Clusters/Classes.

Table 6
Cluster/class response probability across levels of each variable.

	Cluster1	Cluster2	Cluster3	Cluster4
Class Size	0.46	0.28	0.19	0.07
Indicators				
Race				
White/other, not Hispanic	0.51	0.24	0.19	0.06
Hispanic	0.32	0.39	0.17	0.12
Black, not Hispanic	0.22	0.47	0.18	0.13
Education				
No high school diploma	**0.46**	**0.37**	**0.11**	**0.06**
Diploma or GED	0.47	0.21	0.24	0.08
AA degree	0.41	0.14	0.33	0.12
BA degree	0.00	0.00	0.99	0.01
Currently Working				
No	0.56	0.41	0.01	0.02
Yes	0.34	0.12	0.40	0.14
Marital Status				
Married	0.06	0.27	0.07	0.60
Separated	0.22	0.41	0.17	0.20
Divorced	0.34	0.39	0.20	0.07
Widowed	0.38	0.41	0.20	0.01
Never married	0.57	0.23	0.20	0.00
Below Poverty Line				
Below poverty	0.61	0.33	0.02	0.05
Above poverty	0.04	0.16	0.65	0.14
Poor Health				
Not in poor health	0.49	0.24	0.20	0.07
In poor health	0.00	0.93	0.00	0.07
CESD Category				
LE 15 low risk of depression	0.57	0.03	0.26	0.15
16-23 mod risk of depression	0.55	0.25	0.18	0.01
GE 24 high risk of depression	0.23	0.69	0.08	0.00

*Note this table is commonly class Probability Means.

was 5.10. One bi-variate residual, in the four-class model (Marital Status and Race Ethnicity) was 1.82. Setting that relationship to zero and re-analyzing the four-cluster model did not statistically significantly improve the model.

The residuals for each individual case can also be examined. In the analysis, there are a few individual cases where the model does not explain the individual case data well. Cook's D was used with a cut-off value based on 4 times the number of parameters (38) divided by the number of cases (500). Two cases surpassed the cut of value of 0.31 and were removed; this did not change the results. There are a large number of evaluation components and pieces of information that are useful in examining an individual model.

2.4. Step 3: parameter estimate evaluation

Once the residuals are examined, and any issues resolved, the model can be examined with the results for each manifest variable. Table 4, provides the results from this example data.

In addition, the Profile of the Class related to the variables can be examined. The first level examination is the class size. Class 1 has 46% of the sample in it and Class 4 has 7% (Table 5). Class four is small, but so far the evidence indicates that the model should have four classes. The second level examination of the profile is with each variable and the categories or values in each variable (Table 6). You can also look at conditional probabilities. For example in Class 1 individuals are likely to respond that they are “never married” and not currently working. Notice that in Cluster (Class) 2, that 61% have no high school diploma, 37% have a high school diploma, 1% have an

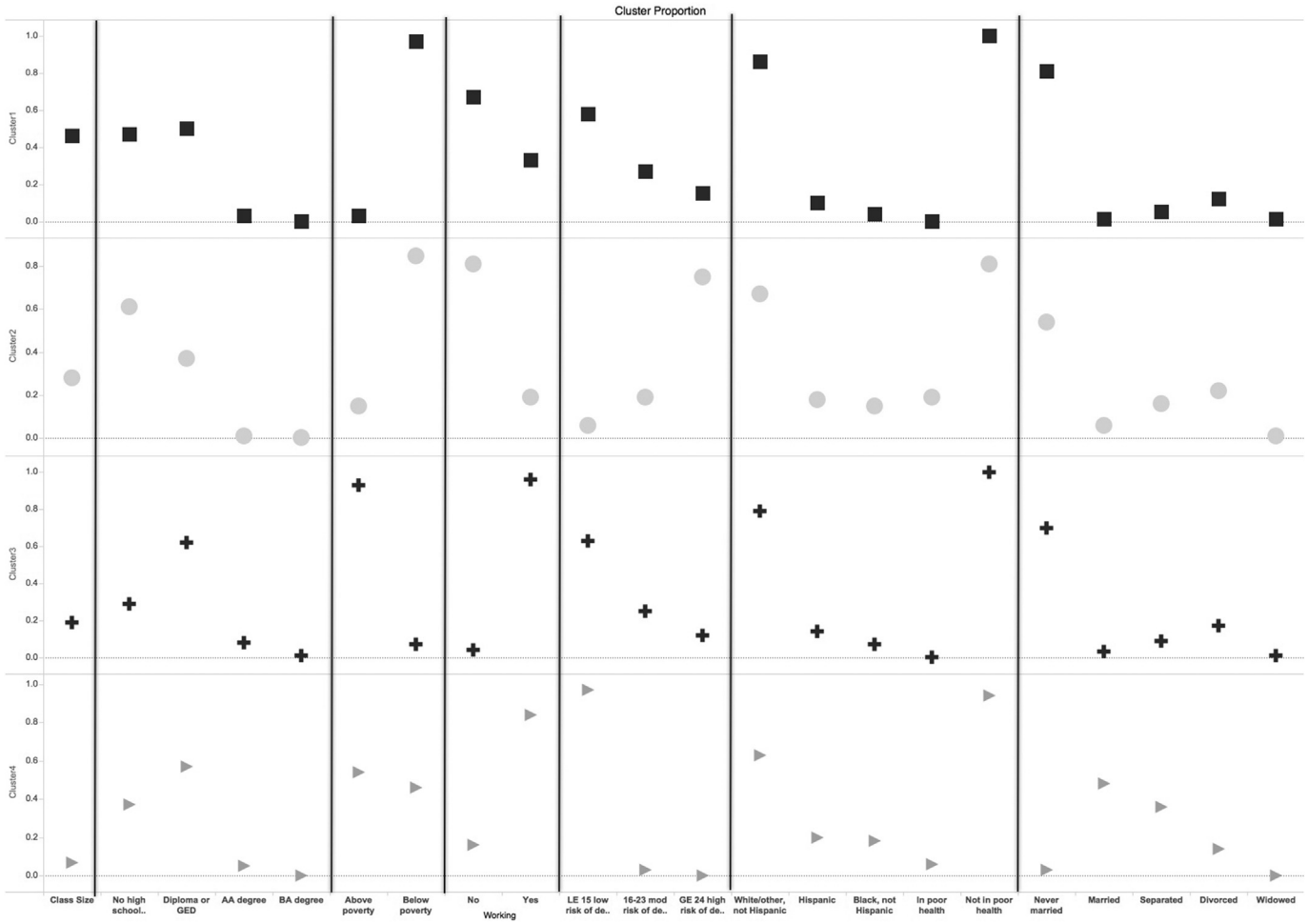

Fig. 1. Graphical Display of Profile Probabilities by Class.

associates degree. These sum to 1; with a bit of rounding error in the table for readability.

As can be seen for Poor Health, which was not helping discriminate across the classes, almost everyone in each class, is not in poor health. Again, this variable may be worth removing and re-running the model. At this point, it is common for researchers to name the classes. For Cluster 2, I might call this the "high risk/ negative indicators group." The classes can also be displayed graphically, based on the Profile of Classes Probabilities (Fig. 1).

A final look of the results occurs with an examination of the probabilities for each response category across the classes. In Table 6, for example, of all respondents who choose the category No High School Diploma, 46% are in Class 1, 37% in Class 2, 11% in Class 3, and 6% in Class 4. Note, these sum up to one (some rounding error in using two digits in the chart) if you move horizontally across the table (0.46 +0.37+0.11 +0.06 = 1.00).

There are other indicators that can be examined for model adequacy. There is a Standard R-Squared (0.63) and an Entropy R-Squared (0.64) and both of these are considered pseudo R-Squared values. The closer to 1, the more accurate the classification. Finally there is the classification error, which estimated the proportion of cases misclassified. The four-cluster model has 18% misclassified but the 3 and 5 cluster models had 19% and 21% misclassified.

There are a number of indicators to be used when deciding how well the data fits the model. There is also a balance issue of not trying to just statistically get the best fitting model, it must also

Table 7
Covariance matrix for four class model.

Inter-Item Covariance Matrix							
	Race	Education	Currently employed?	Marital Status	Poverty Status	CESD Categories	Poor health self rating
Race	0.38	0.01	0.01	−0.14	0.02	0.04	0.00
Education	0.01	0.33	0.02	−0.03	0.04	−0.09	−0.01
Currently employed?	0.01	0.02	0.25	−0.07	0.10	−0.11	−0.02
Marital Status	−0.14	−0.03	−0.07	1.80	−0.07	0.03	−0.04
Poverty Status	0.02	0.04	0.10	−0.07	0.20	−0.06	−0.01
CESD Categories	0.04	−0.09	−0.11	0.03	−0.06	0.75	0.05
Poor health self rating	0.00	−0.01	−0.02	−0.04	−0.01	0.05	0.06

Table 8
Model fit information for LC-Regression models.

	LL	BIC(LL)	Npar	L^2	df	*p*-value	Class.Err.	R^2
1-Class Regression	−483.76	1017.14	8	195.25	206	0.7	0.00	0.14
2-Class Regression	−478.79	1063.02	17	185.31	197	0.72	0.26	0.45
3-Class Regression	−466.88	1095.02	26	161.49	188	0.92	0.23	0.57
4-Class Regression	−460.10	1137.29	35	147.94	179	0.96	0.31	0.93

make sense. Finally, and as a reminder with any advanced technique, may represent the real world or they may not.[16] Over time, you model or parts of your model will be shown to be incorrect.

3. Basic reporting guidelines

1. Show Evaluative Information for all models tested, all of it.
   a. Covariance Matrix (Table 7)
   b. Profiles
   c. Percentages in Each Class
   d. Evaluative criteria used for choosing a specific model
   e. Residual analysis
2. Explain reasoning choices for choosing a specific model
3. Explain choices related to "fixing" a bivariate residual relationship to zero.
4. Provide software used and the version number. Software changes and you must let your reader know.
5. Submit the traditional descriptive and frequency data along with correlation and variance/covariance matrix with the review at a minimum. The matrix may not be published due to page space concerns, but you should provide it.
6. If in doubt, put the information into the manuscript or a table if you are worried about word space. As a former journal editor, there is little more frustrating than a lack of analytical information when reviewing a manuscript.
7. Many read the research from a theoretical point of view for their domain. Read the analysis literature with the same fervor. LCA and those working on it continue to improve the technique and it important to be current.

Table 9
Proportion of cases by class.

	Class1	Class2	Class3
Class Size	0.54	0.24	0.22
CESD Category			
LE 15 low risk of depression	0.49	0.21	0.68
16-23 mod risk of depression	0.01	0.74	0.21
GE 24 high risk of depression	0.51	0.06	0.11
Mean	2.02	1.85	1.43

Table 10
R-Squared values by class.

Model for Dependent				
	Class1	Class2	Class3	Overall
R^2	0.27	0.78	0.85	0.57

4. LC-regression

If examining by classes is the starting point, but the desire to predict an outcome based on the classes, then running a latent class regression analysis technique is needed. One could create the classes and save them by individual and then use the class designation in a regression or logistic regression or run the regression model all at once. The example below utilizes the same data and variables, with one switch, CESD-Category is now the outcome variable.

The evaluation of the models begins in a similar pattern with examining the fit data for several models. Table seven provides the evaluation data for LC-Regression models-1 to 4 classes-with CESD-Categories as the outcome variable (see Table 8).

Looking at the fit data, improvements in LL and L^2 from Classes 1 to 2, 2 to 3, and then just a small movement from 3 to 4 are visible. The BIC values are increasing which is a concern, but the R-Squared values along with the *p*-values indicate the larger number of classes. Examining all of this information, 3 or 4 Class LC-Regression models should be considered. Related, and of concern, the Classification Error rate has jumped quite a bit from three two four.

A bootstrap comparison between 3 and 4 Class regressions was run. The results indicate a 4 class regression model is not better than 3 (-2LL Diff = 9.92 p = 0.47. The residuals can also be examined. In LatentGold, you have to do a bit of work because the residuals are in a frequency chart, but careful examination of them will display any peaks (large number of residuals at one point) or very large standardized residuals. The residuals from this data do not appear to have any peaks or extremely large residuals. You can also examine Cook's D; in this example, two cases have values indicative for removal.

The model has several components that can be examined and still tested for fit and predictive validity (Table 9). The first is simply the proportion of cases in each Class (Table 9).

As is indicated, Class 1 has the largest proportion of cases with 54% and a split between those with low risk and high risk of depression. One could term this the split group. Additionally, in Class 2, 74% are in the moderate risk category, thus this may be termed the moderate group.

Next, the R-Squared values for each Class and the model overall are good, with the lowest being 0.27 for Class 1. The low value

Table 11
Parameter Estimates for the 3-Class model with CESD category as Outcome.

Predictors	Class1	Class2	Class3	Wald	p-value	Wald (=)	p-value	Mean	sd
Race	0.31	0.25	−3.30	7.80	0.05	4.07	0.13	−0.49	1.48
Education	−0.54	−4.17	4.05	13.02	< **0.01**	2.89	**0.24**	−0.43	2.79
Working	−0.61	−2.35	−5.20	16.41	<0.01	3.11	0.21	−2.03	1.81
Marital	−0.10	1.61	2.27	7.79	0.05	6.31	0.04	0.83	1.03
Poverty	−0.42	−5.26	14.19	10.91	0.01	7.28	0.03	1.57	6.93
Health	2.64	9.40	11.22	8.53	0.04	2.69	0.26	6.15	3.84

indicates the model is not working well for Class 1. This lower value may be due to the bimodal split of proportions in Class 1 (Table 10).

In Table 11, the beta parameters are a measure of the influence on CESD categorization. For Class 3, Poor Health and Poverty status are a heavy influences in the CESD categories. Interestingly, Class 1, does not have strong predictors, but that may also be due to the bimodal aspects of the outcome described earlier.

Examining a single predictor, such as Education, the Wald statistic (13.02) indicates statistical significance $p < 0.01$, and the Wald (=) indicates a non-statistically significant result. The first Wald indicates that Education is a statistically significant predictor. The second Wald (=) indicates that the differences across these beta effects is not statistically significant. Thus, a test of Education being

maxiter = 2000, graphs = FALSE, tol = 1e-10, na.rm = TRUE, probs.start = NULL, nrep = 2, verbose = TRUE, calc.se = TRUE).

#LC-Regression.

data (SRSAPLCACESD500) #Note: I had this data open in R-Studio.

*f.depress < - cbind(***racethn, educatn, worknow=worknow+1, marital, poverty, poorhlth=poorhlth+1***)~cesdcat*

nes.depress < - poLCA (f.depress, **SRSAPLCACESD500,**nclass = 3,verbose = F).

probs.start < - poLCA.reorder (nes.depress$probs.start, order (nes.depress$P,decreasing = T)).

nes.depress < - poLCA (f.depress, **SRSAPLCACESD500**,nclass = 3,probs.start = probs.start).

Code	Meaning
#	Tells R that this is a comment statement and not to process
f < - cbind (Y1,Y2,Y3)~1	Codes for the variables to be included in analysis. The +1 on the variables above has been added because zero is a value and in poLCA you cannot have a zero value.
~1	Instructs poLCA to estimate the basic latent class model.
nclass	The number of latent classes to assume in the model
maxiter	The maximum number of iterations through which the estimation algorithm will cycle.
graphs	Logical, for whether poLCA should graphically display the parameter estimates at the completion of the estimation algorithm. The default is FALSE.
tol	A tolerance value for judging when convergence has been reached.
na.rm	Logical, for how poLCA handles **cases** with missing values on the manifest variables. If TRUE, those **cases** are removed (listwise deleted) before estimating the model. If FALSE, cases with missing values are retained. The default is TRUE.
Probs.start	A list of matrices of class-conditional response probabilities to be used as the starting values for the EM estimation algorithm. Each matrix in the list corre- sponds to one manifest variable, with one row for each latent class, and one column for each possible outcome. The default is NULL, meaning that starting values are generated randomly. Note that if nrep >1, then any user-specified probs.start values are only used in the first of the nrep attempts.
nrep	Number of times to estimate the model, using different values of probs.start. The default is one. Setting nrep >1 initiates the search for the global maximum of the log-likelihood function. Setting nrep to 1 will allow the search for just the local maximum. poLCA returns only the parameter estimates corresponding to the model producing the greatest log-likelihood.
verbose	Tells poLCA output to the screen the results of the model. If FALSE, no output is produced. The default is TRUE.
calc.se	Tells poLCA to calculate the standard errors of the estimated class-conditional response probabilities and mixing proportions. The default is TRUE; can only be set to FALSE if estimating a basic model with no concomitant variables specified in formula.

class independent would be appropriate (as with Race, Working, and Poor Health).

Additionally, examining the z-distributions of each predictor for each class can be accomplished. Those that are not over the value of two are not statistically significant and the 0.05 level and can be set to zero. In this model, several of the predictors would be set to zero.

5. Conclusions

LCA and the advancements within LCA have provided a powerful alternative to previously used statistical techniques such as K-means for clustering. The ability to use a wide variety of data with different variances along with regression and factor analysis options makes latent class analysis an appealing option for many researchers. As with the changes in ease of use and understanding of the analysis technique, I expect many more LC analyses to be published in the near future.

Appendix A

R-Script-This can be cut and pasted and then adjusted to your variables.

The variables/data set from this example are in **bold**.

The R-code for poLCA is in italics.

#LCA.

f <- cbind(**racethn, educatn, worknow=worknow+1, marital, poverty, poorhlth=poorhlth+1, cesdcat**)~*1*

poLCA (f, data = **SRSAPLCACESD500**, nclass = 4,

References

1. McCutcheon A. *Latent Class Analysis. Quantitative Applications in the Social Sciences Series*. Newbury Park, London, and New Delhi: Sage Publications; 1987. Number 07–064.
2. Lazarsfeld P, Henry N. *Latent Structure Analysis*. Boston: Houghton Mifflin; 1968.
3. McLachlan G, Basford K. *Mixture Models: Inference and Application to Clustering*. New York: Marcel Dekker; 1988.
4. Everitt B. *Cluster Analysis*. London: Edward Arnold; 1993.
5. Banfield J, Raftery A. Model- based Gaussian and non-Gaussian clustering. *Biometrics*. 1993;49:803–821.
6. Bensmail H, Celeux G, Raftery A, Robert C. Inference in model based clustering. *Statistics Comput*. 1997;7:1–10.
7. Fraley C, Raftery A. *MCLUST: Software for Model-based Cluster and Discriminant Analysis*. Department of Statistics, University of Washington; 1998a. Technical Report No. 342.
8. McLachlan G, Peel D, Basford K, Adams P. The EMMIX software for the fitting of mixtures of normal and t-components. *J Stat, Soft*. 1999;4(2):1–15.
9. Cheeseman P, Stutz J. Bayesian classification (autoclass): theory and results. In: Fayyad U, Piatesky-Shapiro G, Smyth Uthurusamy P, eds. *Advances in Knowledge Discovery and Data Mining*. AAAI Press; 1995:153–180.
10. Magidson J, Vermunt J. Latent class models for clustering: a comparison with K-means. *Can J Mark Res*. 2002;20:37–44.
11. SPSS Technical Note: Can be found at http://www-01.ibm.com/support/docview.wss?uid=swg21476878.
12. Vermunt J, Magidson J. *Technical Guide for Latent GOLD 5.0: Basic, Advanced, and Syntax*. Belmont, MA: Statistical Innovations Inc; 2013.
13. R-package poLCA can be retrieved at https://CRAN.R-project.org/package=poLCA.
14. Linzer D, Lewis J. *Polytomous Variable Latent Class Analysis 1.4.1*; 2014. Retrieved at https://cran.r-project.org/web/packages/poLCA/poLCA.pdf.
15. Nylund K, Asparouhov T, Muthén B. Deciding on the number of classes in latent class analysis and growth mixture modeling: a Monte Carlo simulation study. *Struct Eq Model*. 2007;14(4):535–569.
16. Kline R. *Principles and Practice of Structural Equation Modeling*. Guilford Publications; 2015.