

Latent Class Analysis Using PROC LCA

Pat Berglund

Institute for Social Research - University of Michigan

Michigan SAS User's Group Presentation

February 16, 2017

Presentation Outline

- Overview of latent class analysis, what is it, why use this tool, how to perform LCA in SAS?
- Detailed application using PROC LCA, user-defined SAS tool developed by Penn State University Methodology Center, see <https://methodology.psu.edu>
- Three techniques covered in analysis application
 - Baseline model selection
 - Multiple-group LCA
 - LCA with covariate

Overview of Latent Class Analysis

Introduction to Latent Class Analysis

- LCA used to identify latent classes from observed responses to categorical variables
 - Make sense of large number of possible combinations from manifest or observed variables:
 - » Consider 5 variables used in LCA, assume each is a binary indicator of Marijuana use, can have $2*2*2*2*2=32$ possible combinations, difficult to make sense of so many categories, LCA can assist in this process!
 - Similar to factor analysis used to identify latent classes for a set of continuous variables but LCA uses categorical variables
 - Baseline latent class model is usually primary interest but related techniques are possible
 - For more on statistical details, see Collins, L. M., & Lanza, S. T. (2010)

LCA Extensions

- Two extensions of LCA are demonstrated in application
 - **Multiple-group LCA** allows class membership and item-response probabilities to vary across a group of interest
 - **LCA with covariates** enables predictors of class membership in LCA model
 - Additional extensions such as Latent Transition Analysis, LCA with a Distal Outcome, LCA with Moderation and more are possible

PROC LCA

- PROC LCA software developed by Penn State Methodology Center, free tool that runs under SAS as user-defined procedure, not considered a production SAS procedure
- Other procedures are programmed and available from PSU Methodology Center
- Software and related products (documentation, related publications, tutorials etc.)
<https://methodology.psu.edu/downloads/proclcalta>

Analysis Application

Overview of LCA Application

- Step-by-step demonstration of a typical LCA session
- Data used is 2014 Monitoring the Future survey of high school seniors,
(http://www.monitoringthefuture.org/pubs/monographs/mtf-vol1_2014.pdf)
- Main analytic interest is alcohol use behavior with gender and indicator of skipping school used in multiple group LCA and LCA with covariates
- MTF data set is public release and can be obtained from University of Michigan, ICPSR (<https://www.icpsr.umich.edu>)

Research Questions

- Three questions addressed through use of LCA and extensions to LCA
 - 1) “What patterns of underlying alcohol behaviors exist, can latent class analysis help explain those patterns and, if so, what are the types and prevalences?”
 - 2) “Is latent class measurement invariant across gender?”
 - 3) “Does skipping school during the past month predict latent class membership?”

Data Source and SAS Code

- 2014 12th grade MTF data, replicates a previous LCA based on the same questions/variables but from the 2004 12th grade MTF data set (Lanza et al, 2007)
- Application provides demonstration of required and optional PROC LCA statements
- Selected code is presented in this presentation, full code (except for slight modifications to PSU supplied macros) is included in Appendix A of Paper 5500

Variable Construction

- Data management includes sample refinement, variable construction, and descriptive analysis of key variables
- Seven alcohol behavior variables used in LCA:
 1. lifetime alcohol use (ALC_LT)
 2. past year use (ALC_YR)
 3. past month use (ALC_MO)
 4. drunk during lifetime (ALC_DRUNK_LT)
 5. drunk past year (ALC_DRUNK_YR)
 6. drunk past month (ALC_DRUNK_MO)
 7. 5+ drinks during past 2 weeks (ALC_5PLUS_2WK)
- Each alcohol variable recoded/dichotomized to 1 “No Use” or 2=“Used 1+ times” with missing set to “.”

Sample Refinement

- Records missing on all alcohol variables deleted, analysis n=2,101
- Two additional variables used in multiple-group LCA and LCA with covariates:
 - Gender (SEX, 1=male, 2=female)
 - Skipped school during past month (SKIP_30, 1=No, 2=Yes)
- All analyses weighted with variable ARCHIVE_WT

Descriptive Statistics

- PROC TABULATE with CLASS, VAR, and TABLE statements used to create Table 1 (presented in next slide)
- Weight variable used in TABLE statement, permits use of non-integer weights (unlike FREQ/WEIGHT statements in TABULATE)

```
proc tabulate data=f12014_final;  
  class alc_lt alc_yr alc_mo alc_drunk_lt alc_drunk_yr  
        alc_drunk_mo alc_5plus_2wk skip_30 sex ;  
  var archive_wt ;  
  table alc_lt alc_yr alc_mo alc_drunk_lt alc_drunk_yr  
        alc_drunk_mo alc_5plus_2wk skip_30 sex,  
        (sum='N'*archive_wt=' ' colpctsum='%'*archive_wt=' ') / rts=30 ;  
  format alc_lt alc_yr alc_mo alc_drunk_lt alc_drunk_yr  
        alc_drunk_mo alc_5plus_2wk skip_30 yn. sex sexf. ;  
run ;
```


Baseline Model Selection

- Baseline model selection addresses first research question, “What patterns of underlying alcohol behaviors exist, can latent class analysis help explain those patterns and, if so, what are the types and prevalences?”
 - Repeated use of PROC LCA for 2-7 class models each with 7 alcohol indicators, point is to explore possible baseline models
 - Goal is selection of optimal baseline model from six LCA models tested while considering interplay of factors:
 - Evaluation of model fit statistics
 - Model identification
 - Class membership probabilities
 - Interpretability of latent classes

Model Testing

- PROC LCA executed six times (2-7 classes) using **%alc** macro, each LCA model includes the seven alcohol behavior variables
- The macro uses required and optional statements such as PROC LCA with ORIG_WEIGHTS, OUTEST and OUTPOST to create output data sets, WEIGHT, ID, NSTARTS, NCLASS, ITEMS, CATEGORIES, SEED, and RHO PRIOR statements to specify LCA options:

```
%macro alc (nc);  
proc lca data=f12014_final orig_weights outest=sgf.outests1&nc  
  outpost=sgf.outposts1&nc;  
  id caseid ;  
  weight archive_wt ;  
  nstarts 300 ;  
  nclass &nc ;  
  cores 4 ;  
  items alc_lt alc_yr alc_mo alc_drunk_lt alc_drunk_yr alc_drunk_mo  
  alc_5plus_2wk ;  
  categories 2 2 2 2 2 2 2 ;  
  seed 1232 ;  
  rho prior=1 ;  
run ;  
%mend alc ;  
%alc(2) ; %alc(3) ; %alc(4) ; %alc(5) ; %alc(6) ; %alc(7) ;
```

Table 2a Selected LCA Output

Data Summary, Model Information, and Fit Statistics (EM Algorithm)

Number of subjects in dataset: 2101
Number of subjects in analysis: 2101

Number of measurement items: 7
Response categories per item: 2 2 2 2 2 2 2
Number of groups in the data: 1
Number of latent classes: 5

The analysis includes sampling weights.
Weighting variable name: ARCHIVE_WT

NOTE: A data-derived prior was applied to the rho parameters to help avoid parameter estimates on boundary values of zero and one.

Rho starting values were randomly generated (seed = 1232).
No parameter restrictions were specified (freely estimated).

Seed selected for best fitted model: 262169154
Percentage of seeds associated with best fitted model: 85.33%

The model converged in 382 iterations.

Maximum number of iterations: 5000
Convergence method: maximum absolute deviation (MAD)
Convergence criterion: 0.000001000

=====
Fit statistics:
=====

Log-likelihood: -4613.88
G-squared: 16.93
AIC: 94.93
BIC: 315.29
CAIC: 354.29
Adjusted BIC: 191.38
Entropy: 0.94
Degrees of freedom: 88
(Based on the pseudo-likelihood incorporating weights.)

- Table 2a presents output from the 5 Class LCA model
- Data Summary shows 7 measurement items (alcohol behaviors) with 2 response categories (1=no use, 2=used) and 5 latent classes
- The Fit Statistics section includes information used to select an optimal baseline model from 2-7 class LCA models

Table 2b Estimates for Five Class LCA

```

Parameter Estimates
Class membership probabilities: Gamma estimates (standard errors)
Class:           1           2           3           4           5
                0.2363    0.2440    0.2582    0.1716    0.0900

Item Response Probabilities
Response category 2: (RESPONSE CATEGORY 2 MEANS USED)
Class:           1           2           3           4           5
alc_lt          :    0.9999    0.9999    0.0028    0.9999    0.9997
alc_yr          :    0.5890    0.9999    0.0002    0.9998    0.9996
alc_mo          :    0.0004    0.9998    0.0002    0.3802    0.9940
alc_drunk_lt   :    0.2626    0.9998    0.0002    0.9997    0.2896
alc_drunk_yr   :    0.0002    0.9997    0.0001    0.9977    0.0006
alc_drunk_mo   :    0.0001    0.9014    0.0001    0.0002    0.0002
alc_5plus_2w  :    0.0001    0.5076    0.0000    0.0001    0.1057
    
```

- Table 2b includes class membership and item response probabilities
- Endorsed alcohol use behaviors are of interest, focus on response category 2 probabilities
- Class membership probabilities and item response probabilities suggest five distinct classes
- Identification of item response probabilities => .5 and gamma estimates with good size, i.e. not near zero, considered in choice of optimal baseline model

Table 3 Model Fit Comparison

Number of Classes	LL	DF	G_SQUARED	AIC	BIC	CAIC	ABIC	ENTROPY
2	-5612.915461	112	2015.0000106	2045.0000106	2129.7525411	2144.7525411	2082.0960057	0.9391314531
3	-4986.690482	104	762.5500529	808.5500529	938.50393302	961.50393302	865.43057884	0.9115795257
4	-4702.24711	96	193.66330909	255.66330909	430.81853881	461.81853881	332.32836579	0.9111756793
5	-4613.880293	88	16.929674557	94.929674557	315.28625389	354.28625389	191.37926203	0.9427667744
6	-4607.257115	80	3.6833188242	97.683318824	363.24124776	410.24124776	213.91743706	0.8858524546
7	-4607.003522	72	3.1761316366	113.17613164	423.93541018	478.93541018	249.19478063	0.8918621285

- Table 3 provides comparative summary of model fit statistics
- For AIC, BIC, CAIC, ABIC, and G^2 lower values preferred but for Entropy, higher values indicate better separation/interpretability of classes
- In bold, 5 class model appears to be the best fit since AIC, BIC, CAIC and ABIC are lowest
- $G^2=16.92$ (88 df) for 5 class model is large drop from 4 class model, $G^2=193.66$ (96 df)
- Entropy=0.94, suggests good class separation

Review of Parameter Estimates for 5 Class Model

```
Parameter Estimates
Class membership probabilities: Gamma estimates
Class:           1           2           3           4           5
                0.2363    0.2440    0.2582    0.1716    0.0900

Response category 2: (RESPONSE CATEGORY 2 MEANS YES, ENDORSED)
Class:           1           2           3           4           5
alc_lt          : 0.9999    0.9999    0.0028    0.9999    0.9997
alc_yr          : 0.5890    0.9999    0.0002    0.9998    0.9996
alc_mo          : 0.0004    0.9998    0.0002    0.3802    0.9940
alc_drunk_lt    : 0.2626    0.9998    0.0002    0.9997    0.2896
alc_drunk_yr    : 0.0002    0.9997    0.0001    0.9977    0.0006
alc_drunk_mo    : 0.0001    0.9014    0.0001    0.0002    0.0002
alc_5plus_2w   : 0.0001    0.5076    0.0000    0.0001    0.1057
```

- For response category 2, meaning “Used Alcohol”, interpret by highlighting item probabilities => .5 (in orange) and consider descriptive labels for classes (>=.5 is a guideline, not a rule!)
- Class 1 might be labeled “Experimenters” (23.6%), Class 2 “Heavy Drinkers” (24.4%), Class 3 “Non-Drinkers” (25.8%), Class 4 “Occasional Bingers” (17.2%), and Class 5 “Drinkers” (9.0%)
- Before settling on five class model as optimal:
 - Check of model identification (is optimal model well identified?)
 - Interpretation of item response probabilities (good size?) and class characteristics (do they make sense theoretically?)

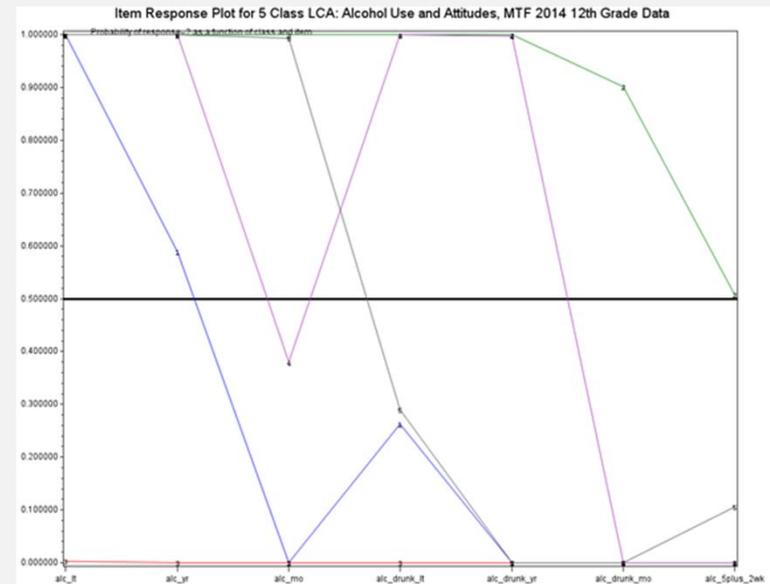
Item Response and Model Identification Plots

- Plots produced by two macros, ***%itemresponseplot*** and ***%identificationplot***, plots can assist in evaluation of models
- Item response plot offers visual display of item response probabilities
- Model identification plot shows how often the best fitted model likelihood selected random starts process, why does this matter?
 - Because 300 random starts used in each LCA model, important to evaluate how often the best fitted model occurs during the Expectation Maximization (EM) process

Figure 1 Item Response Plot for 5 Class Model

```
proc lca data=f12014 final orig_weights
  outparam=sgf.outparm_5c_alc outseeds=sgf.outseeds_5c_alc ;
  id caseid ;
  weight archive_wt ;
  title2 " Alcohol Use: 5 Class LCA" ;
  nclass 5 ;
  nstarts 300 ;
  cores 4 ;
  items alc_lt alc_yr alc_mo alc_drunk_lt alc_drunk_yr alc_drunk_mo
    alc_5plus_2wk ;
  categories 2 2 2 2 2 2 2 ;
  seed 1232 ;
  rho prior=1 ;
run ;
%itemresponseplot(paramdataset=sgf.outparm_5c_alc) ;
```

- Figure 1 displays item response probabilities for response category 2 (Used Alcohol) on Y axis with alcohol behavior variables on X axis
- Separate line with joined points for each of the 5 latent classes displayed using different colors
- Horizontal reference line at 0.5 displayed, helps to identify item response probabilities => .5



Figures 2 and 3 Model Identification Tools

Figure 2. Percentage of Seeds Associated with Best Fitted Model

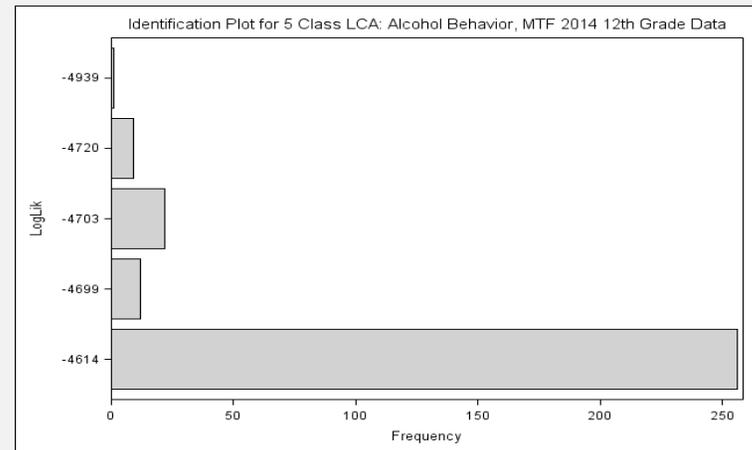
Seed selected for best fitted model: 262169154

Percentage of seeds associated with best fitted model: 85.33%

Invocation of the model identification macro with the *outseeds_5c_alc* data set produces Figure 3:

```
%IdentificationPlot(seedsdataset=sgf.outseeds_5c_alc) ;
```

Figure 3. Model Identification Plot



- Figures 2 and 3 support conclusion that the five class model is identified
- Majority of random starts settled on one log-likelihood value of -4614 (Figure 3)
- High percentage of the seeds (85.3%) were associated with the best fitted model

Figure 4 Latent Class Membership

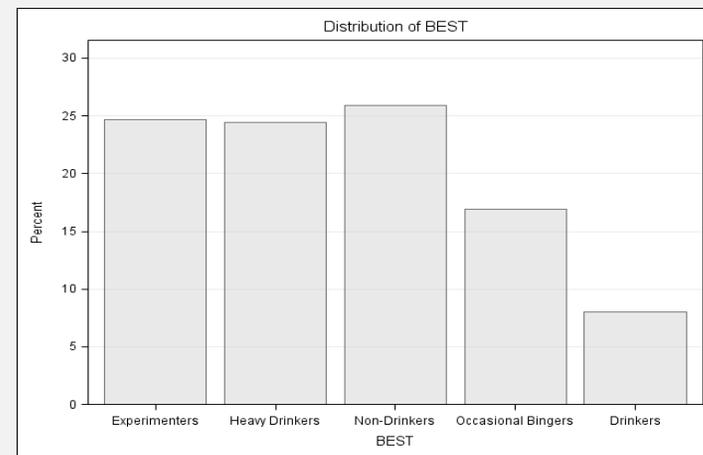
- The optimal baseline LCA model addresses the first research question, “What patterns of underlying alcohol behaviors exist, can latent class analysis help explain those patterns and, if so, what are the types and prevalences?”
- Provides way to make sense of large number of possible combinations of adolescent alcohol behaviors, and gives insight about otherwise unobserved relationships
- SAS code below produces Figure 4 (presented in next slide)
 - PROC FREQ used with output data set called “outposts15” which was saved in previous PROC LCA analysis, contains a variable produced by PROC LCA which contains “best” class membership assignment from 5 class model, use of /plots in PROC FREQ creates a barchart directly from the procedure:

```
proc freq data=sgf.outposts15 ;  
  tables best / plots=freqplot(type=barchart scale=percent) ;  
  format best bestf. ; weight archive_wt ;  
run ;
```

Figure 4 Latent Class Membership

- Figure 4 contains both frequency table and plot, both produced by PROC FREQ
- The 5 classes suggest:
 - 26% of adolescents are members of the Non-Drinkers class,
 - 25% in Experimenters class,
 - 24% in Heavy Drinkers class,
 - 17% in Occasional Bingers class and,
 - 8% in Drinkers class
 - Classes are mutually exclusive and can be used in subsequent analyses if desired, see references and/or PSU Methodology site for examples and relevant publications

Figure 4. Latent Class Membership Best Class Prediction				
	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Experimenters	522.3785	24.72	522.3785	24.72
Heavy Drinkers	516.2041	24.43	1038.583	49.15
Non-Drinkers	547.6265	25.92	1586.209	75.07
Occasional Bingers	357.5398	16.92	1943.749	91.99
Drinkers	169.3238	8.01	2113.073	100.00



Multiple Group LCA

- Multiple group LCA allows item response and class membership probabilities to vary by values of a group variable
- Addresses second research question, **“Is measurement of latent classes invariant across gender?”**
- Strategy for executing multiple group LCA and subsequent testing of measurement invariance is to run 2 LCA models with baseline 5 class model plus a group variable:
 - First LCA model run without measurement restrictions, allow measurement across classes to differ among gender
 - Second LCA model run with measurement invariance specified in the PROC LCA syntax
 - An empirical test of measurement invariance then performed by taking the difference in G^2 values and degrees of freedom using a comparison to a Chi-Square distribution
 - A significant p value suggests rejection of the null hypothesis of measurement invariance

Multiple Group LCA using Gender

- SAS code on right shows how to run two models with key features related to Multiple Groups LCA analysis:
 - SEX is used as a group variable and declared in the GROUPS statement
 - Levels of SEX specified in GROUPNAMES statement
- In the code for second LCA model:
 - MEASUREMENT statement requests PROC LCA apply measurement invariance during estimation
 - All other programming statements remain the same for both models

```
*Model 1;
proc lca data=f12014_final orig_weights;
  title2 '5 Class Alcohol Behavior with Gender Grouping Variable
        (No Measurement Invariance)';
  weight archive_wt;
  nclass 5;
  items alc_lt alc_yr alc_mo alc_drunk_lt alc_drunk_yr
  alc_drunk_mo alc_5plus_2wk ;
  categories 2 2 2 2 2 2 2;;
  groups sex ;
  groupnames male female ;
  seed 262169154;
  rho prior=1 ;
run;

*Model 2;
proc lca data=f12014_final orig_weights ;
  title2 '5 Class Alcohol Behavior with Gender Grouping Variable
        (With Measurement Invariance)';
  weight archive_wt;
  nclass 5;
  items alc_lt alc_yr alc_mo alc_drunk_lt alc_drunk_yr
  alc_drunk_mo alc_5plus_2wk ;
  categories 2 2 2 2 2 2 2;;
  groups sex ;
  measurement groups ;
  groupnames male female ;
  seed 262169154;
  rho prior=1 ;
run;
```

Table 5 Model Fit Statistics for Multiple Group LCA

- Summary of Model fit for Multiple Groups LCA models:

```
Table 5. Model Fit for Multiple Group LCA
Model 1:
-----
Fit statistics:
-----
G-squared:           29.13
Degrees of freedom:   177

Model 2:
-----
Fit statistics:
-----
G-squared:           60.81
Degrees of freedom:   212
```

```
data _null_ ;
diffgsq=60.81-29.13 ;
diffdf=212-177 ;
probdiff=1-probchi(diffgsq, diffdf) ;
put probdiff ;
*value from SAS log= 0.6291764886 , not significant at the alpha=0.05 level! ;
run ;
```

- Use G^2 statistics and df's in data step for tests of differences, compare to Chi-Square distribution to test measurement invariance hypothesis
- Value of the PROBDIFF variable is about **.63** which is **> 0.05**, suggests acceptance of the null hypothesis of measurement invariance
- Therefore, measurement of latent classes is invariant across gender, in other words, can interpret class membership probabilities across gender without concern about measurement differences between girls and boys

Table 6 Class Membership by Gender

Table 6. Class Membership by Gender						
Parameter Estimates						
Class membership probabilities: Gamma estimates (standard errors)						
Class:		Experimenters	Heavy Drinkers	Non-Users	Occasional Bingers	Drinkers
MALE	:	0.2232 (0.0174)	0.2458 (0.0173)	0.2732 (0.0182)	0.1553 (0.0146)	0.1026 (0.0129)
FEMALE	:	0.2434 (0.0175)	0.2333 (0.0178)	0.2450 (0.0163)	0.1892 (0.0168)	0.0891 (0.0120)

- Table 6 presents class membership probabilities and standard errors by gender (from 2nd LCA model)
- Female adolescents show slightly higher probabilities of belonging to Experimenters and Occasional Bingers classes, males more likely to belong to Heavy Drinkers, Drinkers, Non-Users classes

LCA with Covariates

- Final section of application demonstrates LCA with a categorical covariate, using the baseline LCA model with 5 classes
- Addresses research question 3, **“Does skipping school during the past month predict latent class membership?”**
- Continuous covariates also an option, see Lanza et al (2007) for examples

LCA with Indicator of Skipped School During Past 30 Days

- An indicator of having skipped school during past 30 days (SKIP_30) was coded 1="no, did not skip" and 2="yes, did skip"
- SAS code below includes new statements to include covariate:
 - COVARIATES SKIP_30, REFERENCE 3 (declare the Non-Drinkers class as the reference category during the modeling process), output data sets are saved with OUTPARAM, OUTSTDERR, and OUTPOST on the PROC LCA statement
 - NCLASS, WEIGHT, ITEMS, CATEGORIES, CORES, and SEED statements repeat previous LCA syntax:

```
proc lca data=f12014_final orig_weights outparam=sgf.outparms_cov outstderr=sgf.outse_cov
  outpost=sgf.outpost_cov ;
  title2 '5 Class Alcohol Use with Covariate, Skipped School During Past Month ' ;
  weight archive_wt; id caseid ;
  nclass 5; cores 4 ;
  items alc_lt alc_yr alc_mo alc_drunk_lt alc_drunk_yr alc_drunk_mo alc_5plus_2wk ;
  categories 2 2 2 2 2 2 2 ;
  seed 262169154 ;
  covariates skip_30 ;
  reference 3;
  rho prior=1;
run;
```

Table 7 Selected Output for LCA with Covariate

Table 7. Selected Output for LCA with Covariate

Beta estimates (standard errors)					
Class:	1	2	3	4	5
Intercept	-3.5232 (0.6284)	-2.8952 (0.2607)	Reference	-2.0952 (0.2777)	-1.7719 (0.2561)
skip_30 :	1.2001 (0.3965)	1.9916 (0.2002)		1.1832 (0.2268)	1.1477 (0.2099)

Odds Ratio estimates [95% Confidence Interval]					
Class:	1	2	3	4	5
Intercept(odds):	0.0295	0.0553	Reference	0.1230	0.1700
Lower bound	[0.0086]	[0.0332]		[0.0714]	[0.1029]
Upper bound	[0.1011]	[0.0922]		[0.2120]	[0.2809]
skip_30 :	3.3204	7.3276		3.2649	3.1510
Lower bound	[1.5263]	[4.9495]		[2.0930]	[2.0884]
Upper bound	[7.2232]	[10.848]		[5.0929]	[4.7542]

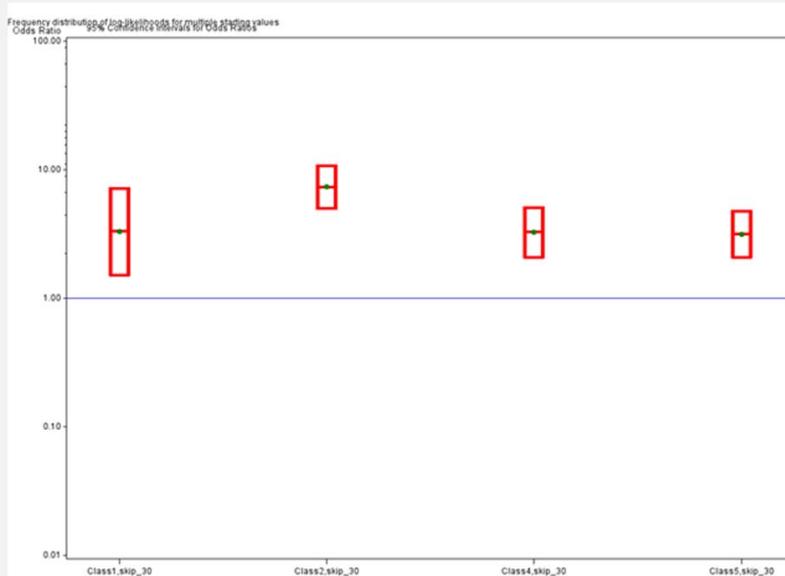
Significance Tests				
Beta parameter test (Type III):				
Covariate	Exclusion LL	Change in 2*LL	deg freedom	p-Value
skip_30	-3982.21	88.44	4	0.0000

- Table 7 includes Beta estimates, Odds ratio estimates, confidence limits, and significance tests
- Interpret with **if skipping school predicts class membership in mind**
- Given having skipped school in the past month and as compared to the Non-Drinkers class,
 - Adolescents were 3.3 times more likely to be in the Experimenters class
 - 7.3 times more likely to be in the Heavy Drinkers class,
 - 3.3 times more likely to be in the Occasional Bingers class,
 - 3.2 times more likely to be in the Drinkers class.
 - All odds ratios significant at the alpha=0.05 level
 - Type III test is highly significant ($p=0.0000$) indicating that skipping school significantly predicts class membership

Figure 5 Odds Ratio Plot

The code below uses the `%oddsratioplot` macro with output data sets from PROC LCA:

```
%OddsRatioPlot(ParamDataset=sgf.outparms_cov,  
StdErrDataset=sgf.outse_cov);
```



- Figure 5 is a visual representation of Table 7
- Each odds ratio is greater than 1.0 and because each set of confidence limits do not include 1.0, all are significant (as compared to omitted Non-Drinkers class)
- The strongest finding is conditional on having skipped school, high school seniors are about 7 times more likely to be members of the Heavy Drinkers class, as compared to the Non-Drinkers class

Conclusion

- This presentation has covered latent class analysis and how to execute this technique using PROC LCA
- The analysis application has demonstrated how to perform baseline model selection, multiple-group LCA, and LCA with covariates to address three research questions
- Presentation of PROC LCA syntax, key output and interpretation of results has been emphasized
- Many more features and extensions are available for Latent Class Analysis, see PSU website for more information

Your comments and questions are valued and encouraged!

Contact Patricia Berglund at pberg@umich.edu

Special thanks to Stephanie Lanza, Megan Patrick, Bethany Bray,
Deb Kloska, and Joy Jang for their guidance and support in preparation of this paper.

Research funded by the National Institute on Drug Abuse

R01DA037902 to M. E. Patrick for analyses/manuscript preparation

R01DA001411 & R01DA016575 to L. D. Johnston for data collection