

Latent Class Analysis Using PROC LCA

Patricia A. Berglund, University of Michigan

ABSTRACT

This paper presents the use of latent class analysis (LCA) to base the identification of a set of mutually exclusive latent classes of individuals on responses to a set of categorical, observed variables. The LCA procedure, a user-defined SAS® procedure for conducting LCA and LCA with covariates, is demonstrated using data on substance use from Monitoring the Future, a nationally representative sample of high school seniors who are also followed at selected time points during adulthood. The demonstration includes guidance on data management prior to analysis, PROC LCA syntax requirements and options, and interpretation of output.

INTRODUCTION

The paper includes a broad overview of latent class analysis followed by an application using PROC LCA, (Lanza ST, Lemmon D, Schafer JL, Collins LM. 2006). The procedure is designed to work within SAS and is available at <https://methodology.psu.edu/>. For a comprehensive theoretical discussion of latent class analysis, see Collins and Lanza, (2013).

The application is a “step-by-step” demonstration of data preparation, baseline model selection and identification, and extensions of LCA such as multiple-group LCA and LCA with covariates. Other options including creation of output data sets and utilization of built-in SAS macros to prepare diagnostic plots are also covered.

The analysis application uses data from the 2014 Monitoring the Future (http://www.monitoringthefuture.org/pubs/monographs/mtf-vol1_2014.pdf) survey of high school seniors and concentrates on alcohol behavior variables (from Form 1 of the survey). Gender and an indicator of skipping school are also used in multiple group LCA and LCA with covariates. The data set is public release and can be obtained from University of Michigan, ICPSR (<https://www.icpsr.umich.edu>).

Three analytic goals are addressed:

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?**

LATENT CLASS ANALYSIS

Latent class analysis is a statistical method used to identify unobserved or latent classes of individuals from observed responses to categorical variables (Goodman, 1974). It is analogous to factor analysis which is commonly used to identify latent classes for a set of continuous variables (Gorsuch, R. L., 1974). This technique offers a method for defining and analyzing unobserved classes and allows the analyst to make sense of a large number of possible combinations of responses from manifest variables.

Two extensions of latent class analysis are multiple-group LCA and LCA with covariates. Multiple-group LCA permits class membership and item-response probabilities to vary across a group of interest where measurement invariance across groups can be tested. LCA with covariates extends the LCA model by including predictors (categorical or continuous) of class membership. For a more detailed discussion of LCA and other extensions, see Collins and Lanza (2013).

PROC LCA

The LCA procedure software and associated products including installation instructions, user documentation, analysis applications, SAS macros, recommended readings and advanced extensions to PROC LCA can be downloaded from <https://methodology.psu.edu/downloads/proclcalta>. PROC LCA is organized much like a production SAS procedure and is easy to use and code. For more on required and optional procedure statements, see the documentation. Note that since PROC LCA is considered a production procedure in SAS v9.4, use of ODS destinations such as ODS HTML, ODS RTF, and ODS PDF are not available, therefore, list output from PROC LCA is used in this paper.

APPLICATION

The application demonstrates use of PROC LCA and two previously mentioned extensions to LCA. Using 2014 12th grade MTF data, the analysis replicates a previous latent class analysis based on the same questions/variables, but from the 2004 12th grade MTF data set, see Lanza et al, (2007).

Along with required PROC LCA statements, optional techniques including summary reports and plots of LCA output are employed to assist in interpretation. Selected code is presented in the paper body and full code (except for slight modifications to PSU supplied macros) is included in Appendix A.

DATA PREPARATION AND DESCRIPTIVE ANALYSIS

Prior to use of PROC LCA, data preparation consisting of sample refinement, variable construction, and descriptive analysis of key variables was performed.

The alcohol behaviors variables are:

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 is based on a 6 or 7 point scale that was dichotomized by setting values of 1 (0 Occasions)=1, No Use and 2+ (1+ Occasions) to 2=Yes Used with missing data to system missing, “.”.

Here is a sample of the Form 1 survey questions for lifetime, past 12 months, and past 30 days alcohol use:

7. On how many occasions (if any) have you had alcohol to drink—more than just a few sips...
(Mark one circle for each line.)

	0 Occasions	1-2 Occasions	3-5 Occasions	6-9 Occasions	10-19 Occasions	20-39 Occasions	40 or More
a. ...in your lifetime?	<input type="radio"/> 1	<input type="radio"/> 2	<input type="radio"/> 3	<input type="radio"/> 4	<input type="radio"/> 5	<input type="radio"/> 6	<input type="radio"/> 7
b. ...during the last 12 months?	<input type="radio"/> 1	<input type="radio"/> 2	<input type="radio"/> 3	<input type="radio"/> 4	<input type="radio"/> 5	<input type="radio"/> 6	<input type="radio"/> 7
c. ...during the last 30 days?	<input type="radio"/> 1	<input type="radio"/> 2	<input type="radio"/> 3	<input type="radio"/> 4	<input type="radio"/> 5	<input type="radio"/> 6	<input type="radio"/> 7

After deletion of observations missing on all seven alcohol variables, the analysis sample is $n=2,101$. Two additional variables used in multiple group LCA and LCA with covariates are gender (SEX, 1=male, 2=female) and skipped school during past month (SKIP_30, 1=No, 2=Yes). We also use the weight variable, ARCHIVE_WT, in all analyses.

The code below produces weighted descriptive statistics included in Table 1. This analysis uses PROC TABULATE with CLASS, VAR, and TABLE statements with the weight variable applied in the TABLE

statement. Use of the weight in the TABLE statement rather than in a WEIGHT or FREQ statement permits use of non-integer weights:

```
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=' ' olpctsum='%'*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 ;
```

Table 1. Weighted Sample Descriptives

	N	%
LT Alcohol		
No	462.28	28.77
Yes	1144.53	71.23
Past YR Alcohol		
No	618.15	38.47
Yes	988.67	61.53
30 Day Alcohol		
No	990.42	61.64
Yes	616.39	38.36
Drunk LT		
No	831.63	51.76
Yes	775.18	48.24
Drunk Past YR		
No	965.07	60.06
Yes	641.74	39.94
Drunk Past 30 Days		
No	1271.83	79.15
Yes	334.99	20.85
5+ Drinks Past 2 Wks		
No	1409.05	87.69
Yes	197.76	12.31
Skipped School Past 30 Days		
No	1134.50	70.61
Yes	472.31	29.39
Sex 1=M, 2=F		
Male	766.52	47.70
Female	840.30	52.30

Based on Table 1, an estimated 71.23% of high school seniors drank alcohol during their lifetime, 61.53% during the past year, 38.36% during the past month, and 48.24%, 39.94%, and 20.85% drank enough to be drunk during those same time periods while 12.31% of the sample consumed 5+ drinks during the past

two weeks. During the past 30 days, 29.39% skipped school and the sample is 47.70% male and 52.30% female.

BASELINE MODEL SELECTION

To address our 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?**”, PROC LCA is used repeatedly to analyze models with 2-7 classes. The goal is selection of an optimal baseline model from the six LCA models tested. This process can be challenging given the interplay of different factors such as evaluation of model fit statistics, model identification, class membership probabilities, and interpretability of latent classes.

Initially, PROC LCA is executed six times using the user-defined macro code below. Each model uses 300 random starts and seven alcohol behavior variables (coded 1 or 2). The **%alc** macro includes required and optional statements such as PROC LCA with ORIG_WEIGHTS, OUTEST, OUTPOST statements along with WEIGHT, ID, NSTARTS, NCLASS, SEED, and RHO PRIOR statements. See the User’s Guide for more detail on syntax.

The PROC LCA statement includes use of the probability weight with the ORIG_WEIGHTS option and also requests output data sets of parameter estimates (OUTEST with a macro variable that resolves to the number of classes) and posterior probabilities (OUTPOST with same macro variable resolution). Additional options define the caseid (ID) needed for future data set manipulation, number of classes (NCLASS), number of random starts (NSTARTS), number of CPU cores used to process the job (CORES), a seed value (SEED), a weight variable (WEIGHT), the alcohol behavior items and corresponding number of categories (ITEMS, CATEGORIES), and a prior used in the calculation of the Rho values (RHO PRIOR):

```
%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. Default LCA Output for the 5 Class Model

Table 2a includes output from the 5 class LCA model, however, similar output is obtained for each run of PROC LCA. The Data Summary, Model Information, Fit Statistics, and Parameter Estimates (Table 2b) comprise the default 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
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
(Base on the pseudo-likelihood incorporating weights.)

```

Key sections of Table 2a are highlighted in red. The number of measurement items (7), categories per item (2), groups (1) and latent classes (5) are listed at the top of the output. Next, the seed selected for the best fitted model (262169154) and the percentage of seeds associated with the best model (85.33%) provide information about the random starts process. The high percentage of seeds associated with the best fitted model indicates that the model appears to be well-identified (more on this concept to come). The Fit Statistics section of output includes statistics to assist in comparison with the other LCA models in order to select an optimal baseline model.

Table 2b. Default LCA Output for 5 Class Model, continued

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
	(0.0113)	(0.0120)	(0.0116)	(0.0108)	(0.0075)
Item response probabilities: Rho estimates (standard errors)					
Response category 1: (RESPONSE CATEGORY 1 MEANS NO USE)					
Class:	1	2	3	4	5
alc_lt :	0.0001	0.0001	0.9972	0.0001	0.0003
	(0.0000)	(0.0000)	(0.0011)	(0.0000)	(0.0000)
alc_yr :	0.4110	0.0001	0.9998	0.0002	0.0004
	(0.0270)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
alc_mo :	0.9996	0.0002	0.9998	0.6198	0.0060
	(0.0000)	(0.0000)	(0.0000)	(0.0344)	(0.0017)
alc_drunk_lt:	0.7374	0.0002	0.9998	0.0003	0.7104
	(0.0248)	(0.0000)	(0.0000)	(0.0000)	(0.0397)
alc_drunk_yr:	0.9998	0.0003	0.9999	0.0023	0.9994
	(0.0000)	(0.0000)	(0.0000)	(0.0008)	(0.0001)
alc_drunk_mo:	0.9999	0.0986	0.9999	0.9998	0.9998
	(0.0000)	(0.0221)	(0.0000)	(0.0000)	(0.0000)
alc_5plus_2w:	0.9999	0.4924	1.0000	0.9999	0.8943
	(0.0000)	(0.0284)	(0.0000)	(0.0000)	(0.0269)

Response category 2: (RESPONSE CATEGORY 2 MEANS USED)					
Class:	1	2	3	4	5
alc_lt :	0.9999 (0.0000)	0.9999 (0.0000)	0.0028 (0.0011)	0.9999 (0.0000)	0.9997 (0.0000)
alc_yr :	0.5890 (0.0270)	0.9999 (0.0000)	0.0002 (0.0000)	0.9998 (0.0000)	0.9996 (0.0000)
alc_mo :	0.0004 (0.0000)	0.9998 (0.0000)	0.0002 (0.0000)	0.3802 (0.0344)	0.9940 (0.0017)
alc_drunk_lt:	0.2626 (0.0248)	0.9998 (0.0000)	0.0002 (0.0000)	0.9997 (0.0000)	0.2896 (0.0397)
alc_drunk_yr:	0.0002 (0.0000)	0.9997 (0.0000)	0.0001 (0.0000)	0.9977 (0.0008)	0.0006 (0.0001)
alc_drunk_mo:	0.0001 (0.0000)	0.9014 (0.0221)	0.0001 (0.0000)	0.0002 (0.0000)	0.0002 (0.0000)
alc_5plus_2w:	0.0001 (0.0000)	0.5076 (0.0284)	0.0000 (0.0000)	0.0001 (0.0000)	0.1057 (0.0269)

Table 2b includes parameter estimates of class membership probabilities (Gamma estimates and standard errors) followed by item response probabilities (Rho estimates and standard errors). Note that response category 1 (No Use) and Response Category 2 (Yes Used) sum to 100%.

Since endorsed alcohol use behaviors are of interest, we focus on response category 2 probabilities (highlighted in red). Inspection of the class membership probabilities and item response probabilities reveals five distinct classes. At this point, identification of item response probabilities \Rightarrow .5 and gamma estimates with good size, i.e. not near zero, are factors in the choice of an optimal baseline model.

Model Fit Comparison

The following code uses a macro called `%it` to create a new variable called NCLASS with values of 2-7 to assign a class number to each output data set. Then the six output datasets are concatenated to produce a summary data set called `allfit_alc`. Finally, PROC PRINT is used to produce a model fit comparison (Table 3):

```
%macro it (nc) ;
data sgf.outestst1&nc ;
  set sgf.outestst1&nc ;
  nclass=&nc ;
run ;
%mend ;
%it (2) ; %it(3) ; %it(4) ; %it(5) ; %it(6) ; %it(7) ;
data sgf.allfit_alc ;
  set sgf.outestst12 - sgf.outestst17 ;
run ;
proc print data=sgf.allfit_alc noobs label ;
  label nclass="Number of Classes"
  log_likelihood="LL" degrees_of_freedom="DF";
  var nclass LOG_LIKELIHOOD DEGREES_OF_FREEDOM G_SQUARED AIC BIC CAIC ABIC
  ENTROPY ;
run ;
```

Table 3. Model Fit for Baseline Models, 2-7 Classes

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 a comparative summary of model fit statistics. For the AIC, BIC, CAIC, ABIC and G^2 statistics, lower values generally indicate better model fit while higher values on Entropy indicate better separation of latent classes. Here, the five class model appears to be the best fit since the AIC, BIC, CAIC and ABIC are each the lowest for this model. Also, the $G^2=16.92$ with 88 degrees of freedom for the 5 class model shows a large drop from the 4 class model with $G^2=193.66$ and 96 degrees of freedom. Entropy equals 0.94 which suggests good class interpretability and separation.

Table 4 provides a repeat of class membership probabilities and parameter estimates for response category 2 from the 5 class model.

Table 4. Parameter Estimates from 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

Based on Table 4, each class is of good size with clearly defined characteristics. A preliminary effort to name each class is done by highlighting item probabilities => .5 (in red) and then assigning a descriptive label. Use of the .50 cutoff is a general guideline and not a strict rule. Though it is tempting to declare the five class model as optimal, a check of model identification and further inspection of latent classes and their meaning should be performed before making a final decision. Another key consideration is if the classes make sense theoretically.

As a preliminary step, we label the five latent classes to help clarify the apparent meaning of the classes. Despite the 10 year gap in data collection between 2004 and 2014, the latent classes in 2014 closely match those from the 2007 paper, based on 2004 MTF data. For example, as in 2004, 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%).

Item Response and Model Identification Plots

Two evaluation tools are used to assist in baseline model selection. The item response and model identification plots are produced using the `%itemresponseplot` and `%identificationplot` macros provided by the PSU Methodology Center. These macros were modified slightly for this application and full code is

available upon request.

The item response plot offers a visual display of the item response probabilities while the model identification plot shows how often the best fitted model likelihood is selected during the random starts process. Because we used 300 random starts in each LCA, it is important to evaluate how often the best fitted model occurs during the Expectation Maximization process. For statistical details, see the PROC LCA user guide or a text on latent class analysis.

Item Response Plot

The following code re-runs the 5 class LCA model (with the same seed) and saves an output data set called *outseeds_5c_alc* which serves as input to the *%itemresponseplot* macro:

```
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 ;
  seed 1232;
  rho prior=1 ;
run ;
%itemresponseplot(paramdataset=sgf.outparm_5c_alc) ;
```

Figure 1. Item Response Plot

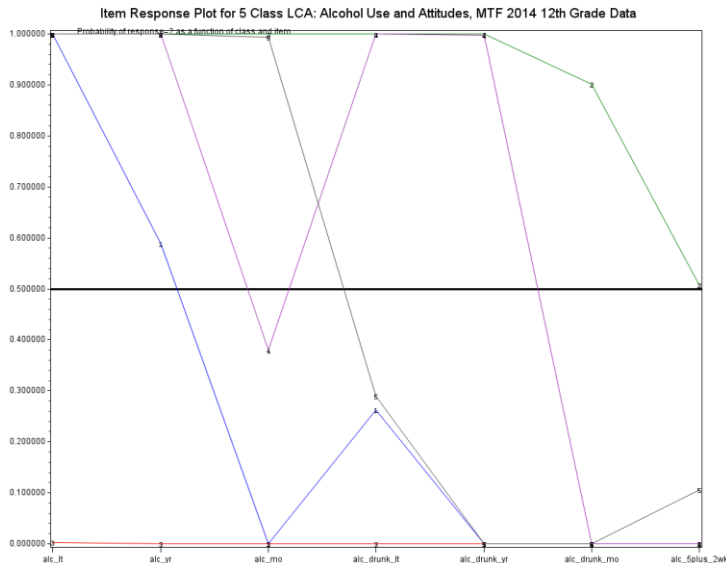


Figure 1 displays item response probabilities for response category 2 on the Y axis with alcohol behavior variables along the X axis. A separate line with joined points for each of the 5 latent classes is displayed using different colors. A horizontal reference line at 0.5 is also displayed, making identification of item response probabilities \Rightarrow .5 easier to identify. This plot is simply a visual representation of the information in Table 4 and may assist in class interpretation.

Model Identification

The next two figures illustrate use of evaluation tools related to model identification. Figure 2 is based on default output presented in Table 2a.

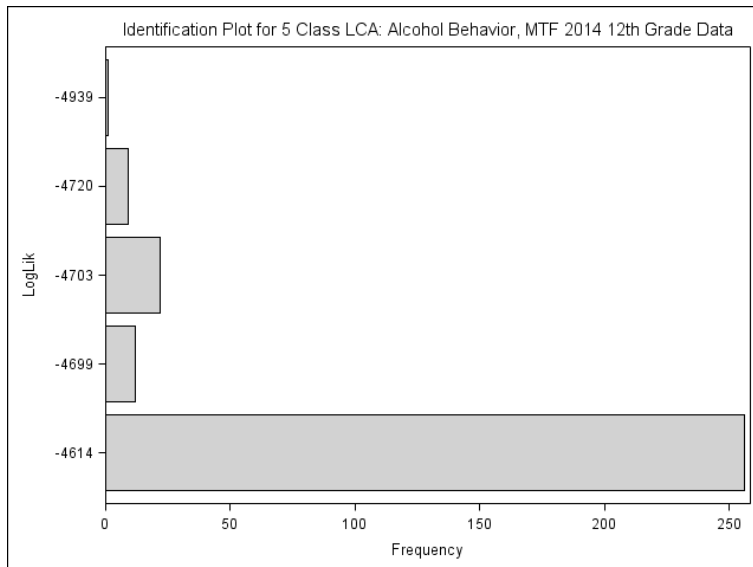
Figure 2. Percentage of Seeds Associated with Best Fitted Model

```
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



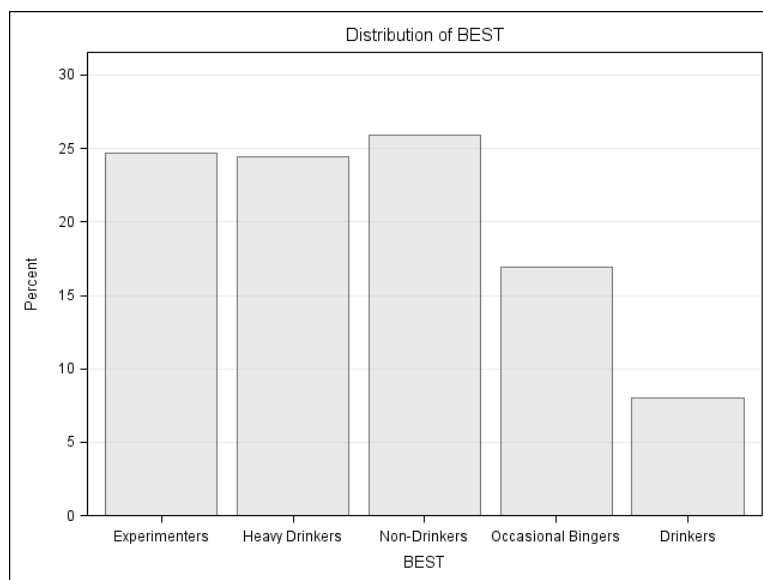
Figures 2 and 3 provide support for the conclusion that the five class model is identified, that is, a majority of random starts settled on one log-likelihood value of -4614 (Figure 3) and a high percentage of the seeds (85.3%) were associated with the best fitted model.

With model identification confirmed, a final evaluation of latent class membership assignments is performed. For example, the next block of code produces a weighted frequency table and bar chart (via ODS GRAPHICS with a PLOTS= option on the TABLES statement) of the variable called BEST. The BEST variable was saved in the *outposts15* data set from a previous PROC LCA run and represents the best class membership based on our five class model:

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

Figure 4. Frequency Distribution and Plot of Class Membership

BEST				
BEST	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 and addresses the second research question: **“Is measurement of latent classes invariant across gender?”**.

The strategy for executing multiple group LCA and subsequent testing of measurement invariance is to run 2 LCA models, one run without measurement restrictions and the second with measurement invariance specified in the PROC LCA syntax. An empirical test of measurement invariance can be performed by taking the difference in G^2 values and degrees of freedom with a comparison to a Chi-Square distribution. A significant p value suggests rejection of the null hypothesis of measurement invariance.

In the following code, SEX is used as a group variable and declared as such in the GROUPS statement along with names for each level of SEX in the GROUPNAMES statement. In the code for the second LCA model, the MEASUREMENT statement requests that PROC LCA apply measurement invariance during estimation. All other programming statements remain the same for both models:

```

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;

```

```

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

```

Model 1:
=====
Fit statistics:
=====
G-squared:           29.13
Degrees of freedom:   177

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

```

Using the two G^2 statistics and associated degrees of freedom, the data step code calculates the differences and the comparison to a Chi-Square distribution:

```

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

```

The value of the PROBDIFF variable is about .63 which is greater than 0.05 and suggests acceptance of the null hypothesis of measurement invariance. Given this, we can interpret class membership probabilities across gender without concern about differences between males and females.

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 and reveals minor differences in class membership. This output is drawn from the second model where measurement invariance is specified in the PROC LCA syntax. Class labels were added to the output manually to make interpretation easier. The results suggest that female adolescents have higher probabilities of belonging to the Experimenters and Occasional Bingers classes while males are more likely to belong to the Heavy Drinkers, Drinkers and Non-Users classes.

LCA WITH COVARIATES

The final section of the application demonstrates LCA with a categorical covariate and addresses research question 3: “**Does skipping school during the past month predict latent class membership?**”. Use of a continuous covariate is also an option, see Lanza et al (2007) for examples.

As a reminder, a binary variable representing skipped school during the past 30 days (SKIP_30) was coded in the data preparation step so that 1=no, did not skip and 2=yes, did skip. This variable serves as a binary covariate and is declared as such in the code (COVARIATES SKIP_30). Use of REFERENCE 3, is a new option used to declare the Non-Drinkers class as the reference category during the modeling process. Output data sets are saved with OUTPARAM, OUTSTDERR, and OUTPOST requests on the PROC LCA statement.

Other statements such as 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

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 parameter estimates, odds ratios and confidence limits, and significance tests. We examine the results with the question of if skipping school predicts class membership in mind. Indeed, the results suggest that 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, and 3.2 times more likely to be in the Drinkers class. All odds ratios are significant at the alpha=0.05 level. The Type III test is highly significant ($p=0.0000$) indicating that skipping school significantly predicts class membership.

The code below uses the *outparms_cov* and *outse_cov* data sets as macro parameters in the **%oddsratioplot** macro:

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

Figure 5. Odds Ratio Plot

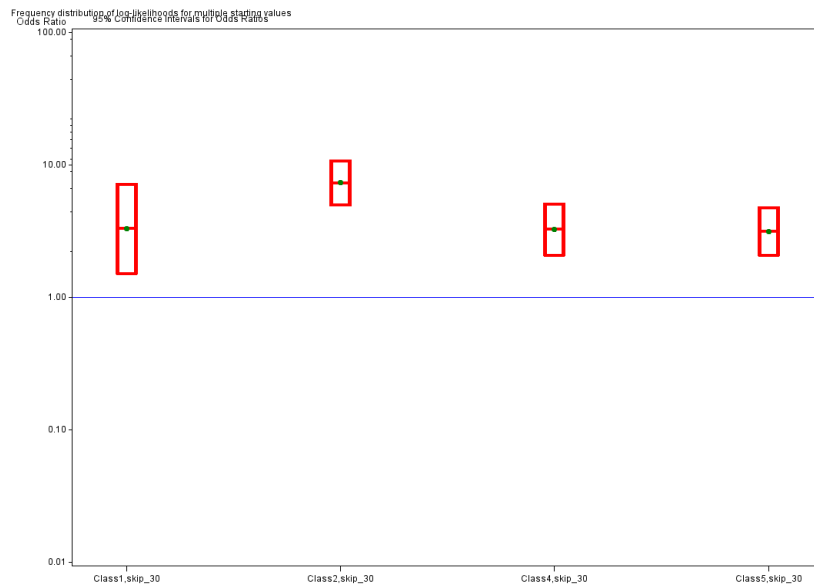


Figure 5 is a visual representation of the results presented in Table 7. The plot illustrates that each odds ratio is greater than 1.0 and because each set of confidence limits do not include 1.0, all are statistically significant. Again, the strongest finding is that conditional on having skipped school, high school seniors are about 7 times more likely to be in the Heavy Drinkers class, as compared to the Non-Drinkers class.

The final set of commands produce Table 8, a weighted cross-tabulation of the class membership assignment (BEST variable) and SKIP_30, the binary indicator of skipped school during the past 30 days. As a reminder the BEST variable is produced by PROC LCA and saved in the *outpost_cov* data set. The TABLE statement requests only column percentages:

```
proc freq data=sgf.outpost_cov ;
tables best*skip_30 / nocum norow nofreq nopercnt ;
weight archive_wt ;
format best bestf. skip_30 skipf. ;
run ;
```

Table 8. Class Membership by Skip School Status

Class	No, Did Not Skip %	Yes Skipped School Past Month
Experimenters	3.02%	3.94%
Heavy Drinkers	15.49%	39.27%
Non-Drinkers	41.76%	13.98%
Occasional Bingers	16.34%	18.76%
Drinkers	23.39%	24.05%

Table 8 shows that those who skipped school have higher probabilities of class membership in the Heavy Drinkers and Occasional Bingers classes, lower probability of membership in the Non-Drinkers class while class membership probabilities are about equal for the Experimenters and Drinkers classes.

CONCLUSION

This paper has provided an introduction to latent class analysis using PROC LCA. An overview of latent class analysis along with a detailed application including model evaluation and selection, multiple-group LCA, and LCA with covariates has been presented.

REFERENCES

- Bray, B. C., Lanza, S. T., & Tan, X. (2015). Eliminating bias in classify-analyze approaches for latent class analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, 22(1), 1-11. doi: 10.1080/10705511.2014.935265 PMID: PMC4299667
- Collins, L. M., & Lanza, S. T. (2010). *Latent class and latent transition analysis: With applications in the social, behavioral, and health sciences*. New York: Wiley.
- Gorsuch, R. L. (1974). *Factor Analysis*, Philadelphia: W. B. Saunders.
- Lanza, S. T., Bray, B. C., & Collins, L. M. (2013). An introduction to latent class and latent transition analysis. In J. A. Schinka, W. F. Velicer, & I. B. Weiner (Eds.), *Handbook of psychology* (2nd ed., Vol. 2, pp. 691-716). Hoboken, NJ: Wiley.

Lanza, S. T., Collins, L. M., Lemmon, D. R., & Schafer, J. L. (2007). PROC LCA: A SAS procedure for latent class analysis. *Structural Equation Modeling*, 14(4), 671-694. PMID: PMC2785099

Lanza, S. T., Dziak, J. J., Huang, L., Wagner, A., & Collins, L. M. (2015). PROC LCA & PROC LTA users' guide (Version 1.3.2). University Park: The Methodology Center, Penn State.

PROC LCA & PROC LTA (Version 1.3.2) [Software]. (2015). University Park: The Methodology Center, Penn State. Retrieved from <http://methodology.psu.edu>.

ACKNOWLEDGMENTS

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

RECOMMENDED READING

Refer to the Recommended Reading List from the PSU Methodology Center:

<https://methodology.psu.edu/ra/lcalta/bib>.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Patricia Berglund
University of Michigan, Institute for Social Research
pberg@umich.edu

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.

APPENDIX A

All code except the SAS plotting macros supplied by the PSU Methodology Center is included in this Appendix. The modified macro code is available upon request from the author (pberg@umich.edu).

```
Libname sgf 'p:\' ; * set libname ;

data f12014 ;
  set sgf.da36263p2 ; * public use : 2014 form 1 data set from ICPSR ;
  array c [*] v1214 v1215 v1216 v1805 v1806 v1807 v1243 ;
  do i=1 to 7 ;
    if c[i]=-9 then c[i]=. ;
  end ;

  if nmiss(v1214, v1215, v1216, v1805, v1806, v1807,v1243) =7 then delete ;

  if v1214 >=2 then alc_lt=2 ; else if v1214=1 then alc_lt=1 ;
  if v1215 >=2 then alc_yr=2 ; else if v1215=1 then alc_yr=1 ;
  if v1216 >=2 then alc_mo=2 ; else if v1216=1 then alc_mo=1 ;
  if v1805 >=2 then alc_drunk_lt =2 ; else if v1805=1 or (alc_lt=1 and alc_yr=1 and alc_mo=1) then
    alc_drunk_lt =1 ;
  if v1806 >=2 then alc_drunk_yr =2 ; else if v1806=1 or (alc_yr=1) then alc_drunk_yr =1 ;
  if v1807 >=2 then alc_drunk_mo =2 ; else if v1807=1 or (alc_mo=1) then alc_drunk_mo=1 ;
  if v1243 >=2 then alc_5plus_2wk =2 ; else if v1243=1 or (alc_lt=1 or alc_yr=1) then
    alc_5plus_2wk =1 ;
  if v1176 eq -9 then v1176=. ;
  if v1176 >=2 then skip_30=2 ; else if v1176=1 then skip_30=1 ;
  if v1179 =-9 then v1179=. ;
  if v1150=-9 then sex=. ; else sex=v1150 ;
run ;

data f12014_final ; * final working data set ;
set f12014 ;
  if v1150=1 then male=1 ; else if v1150=2 then male=0 ; else male=. ;
  if v1150=2 then female=1 ; else if v1150=1 then female=0 ; else female=. ;
label
  alc_lt='LT Alcohol'
  alc_yr='Past YR Alcohol'
  alc_mo='30 Day Alcohol'
  alc_drunk_lt='Drunk LT'
  alc_drunk_yr='Drunk Past YR'
  alc_drunk_mo='Drunk Past 30 Days'
  alc_5plus_2wk='5+ Drinks Past 2 Wks'
  skip_30='Skipped School Past 30 Days'
  sex='Sex 1=M, 2=F' ;
run ;

proc format ; * formats for descriptive table ;
  value yn 1='No' 2='Yes' ; value sexf 1='Male' 2='Female' ;
run ;

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 ;

%macro alc (nc) ;
proc lca data=f12014_final orig_weights outest=sgf.outests1&nc outpost=sgf.outposts1&nc;
  id caseid ;
  weight archive_wt ;
  title2 " Alcohol Use for 12th Grade Students, 2014: &nc Class LCA, weighted with archive
  weight, with random starting values: 2MAR2016" ;
  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) ;

%macro it (nc) ;
data sgf.outests1&nc ;
  set sgf.outests1&nc ;
  nclass=&nc ;
run ;
%mend ;
%it(2) ; %it(3) ; %it(4) ; %it(5) ; %it(6) ; %it(7) ;

data sgf.allfit_alc ;
  set sgf.outests12 - sgf.outests17 ;
run ;
proc print ;
run ;

proc print data=sgf.allfit_alc noobs label ;
  title "Model Fit for Baseline Models: Alcohol Use 12th Grade Students, 2014 (Form 1) " ;
  label nclass="Number of Classes" log_likelihood="LL" degrees_of_freedom="DF" ;
  var nclass LOG_LIKELIHOOD DEGREES_OF_FREEDOM G_SQUARED AIC BIC CAIC ABIC ENTROPY ;
run ;

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 ;

%IdentificationPlot(seedsdataset=sgf.outseeds 5c alc) ;
%itemresponseplot(paramdataset=sgf.outparm_5c_alc) ;

proc format ;
  value bestf 1='Experimenters' 2='Heavy Drinkers' 3='Non-Drinkers' 4='Occasional Bingers'
  5='Drinkers';
run ;

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

* 5 Class model with group variable: Gender ;
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;

proc lca data=f12014_final orig_weights ;
  title2 '5 Class Alcohol Use and Attitudes 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;

data _null_ ;
diffgsq=60.81-29.13 ;
diffdf=212-177 ;
probdiff=1-probchi(diffgsq, diffdf) ;
put probdiff ;
run ;

*5 Class model with covariate: Skipping school during past month is categorical covariate
(0=N, 1=Y);
proc lca data=f12014_final orig_weights outparam=sgf.outparms_cov outstderr=sgf.outse_cov
outpost=sgf.outpost_cov ;
title2 '5 Class Alcohol Use and Attitudes with Covariate : Skipped School During Past Month';
weight archive wt;
id caseid ; *ID is needed with out post type of data set ;
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;

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

proc format ;
value skipf 2='Yes Skipped School Past Month' 1='No, Did Not Skip School' ;
run;

proc freq data=sgf.outpost_cov ;
tables best*skip_30 / nocum norow nofreq nopercnt ;
weight archive_wt ;
format best bestf. skip_30 skipf. ;
run ;

```