PENNSTATE

The Methodology Center

SimulateLcaDataset SAS Macro Users' Guide Version 1.1.0

John J. Dziak
Stephanie T. Lanza
Shu Xu

Please send questions and comments to *MChelpdesk@psu.edu*.
The development of these SAS® macros was supported by National Institute on Drug Abuse, NIH, P50-DA10075 to The Center for Prevention and Treatment Methodology. We thank Dr. YoungKyoung Min for assistance in finding a bug in an earlier version of this macro.
The suggested citation for this users' guide is

Dziak, J. J., Lanza, S. T., & Xu, S. (2011). *SimulateLcaDataset SAS macro users' guide* (version 1.1.0). Univeristy Park: The Methodology Center, Penn State. Available from http://methodology.psu.edu.

The Methodology Center is pleased to release the %**SimulateLcaDataset** macro, which can assist in studying the performance of latent class analysis (LCA) models. It works in conjunction with the SAS®[1] software package (version 9.1 or higher) and the PROC LCA procedure (version 1.2.5 or higher). **%SimulateLcaDataset** generates a random dataset from a population assumed to be described by the LCA model with a given set of parameters.

1 About the %SimulateLcaDataset Macro

The **%SimulateLcaDataset** macro generates a dataset from a simulated LCA model (see Collins and Lanza 2010) without covariates. The data is assumed to consist of dichotomous items coded as 1 and 2. This macro performs a simulation. That is, it does not use empirical data to provide parameter estimates. Rather, it is given supposed population parameter values; it uses these to create artificial sample data.

The simulation model is described as follows. The population is assumed to consist of a number of classes, n_c, which the user can specify. The proportion of the population in a given class c is assumed to be a number, γ_c; the user also specifies these values. Last, there are M observed items (indicator variables), each dichotomous and coded 1 or 2. For an individual in class c, responding to item m, the probability of providing a response of 1 is $\rho_{m|c}$; the user also specifies these ρ values.

The current version of this macro is only for classic LCA models without covariates. Specifically, it is assumed that the model specified for data generation does not have COVARIATES, GROUPs, WEIGHTs, CLUSTERs, or special parameter restrictions (RESTRICT option), even though these features are allowed in PROC LCA (see PROC LCA & PROC LTA User's Guide; Lanza, Dziak, Huang, Xu, & Collins, 2011). Also, the current version of the macro cannot handle polytomous items (i.e., with more than 2 possible responses).

[1] 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.

2 Using the Macro

2.1 Preparation

A SAS macro is a special block of SAS commands that are first defined, and then called later when needed. The procedure for using a SAS macro is very straightforward. Several steps need to be completed before running the macro:

1. If you haven't already done so, download and save the macro at the designated path (say, *S:\myfolder\*).
2. Direct SAS to read the macro code from the path, using a SAS %INCLUDE statement such as

```
%INCLUDE “S:\myfolder\SimulateLcaDataset.sas”;
```

2.2 Syntax and Input

Call the macro using a percent sign, its name, and user-defined arguments in parentheses. For the %SimulateLcaDataset macro, the calling syntax is shown below.

```
%SimulateLcaDataset(    true_gamma_dataset = filename1,
                        true_rho_dataset = filename2,
                        output_dataset_name = filename3,
                        total_n = number );
```

The “arguments” or “parameters” of the macro (i.e., the information in parentheses to be provided to the macro) are listed below.

Arguments	**Required**	**Description**
true_gamma_dataset	Y	The name of a dataset contains the desired gammas (i.e., latent class preferences), listed as a single column (variable). The number of entries in this dataset also indicates the number of latent classes there will be. The entries should sum to 1.
true_rho_dataset	Y	The name of a dataset contains the desired rhos (i.e., item response probabilities) for the “1” response. (The items are assumed to be dichotomous, so the rhos for the “2” response are one minus the rhos for the “1” response.) This dataset should have M rows (one for each item) and columns (one for each class).
output_dataset_name	Y	The name of a dataset to be created which contains the simulated data. (Warning: If there is already a dataset with this name, the pre-existing file will be replaced.)
total_n	Y	A number which tells how many simulated individuals there should be in the simulated sample.

2.3 Output

If the input is appropriate, the output dataset should contain random data generated from an LCA model with the specified characteristics. The random variables are labeled Item001, Item002, … . The resulting SAS data file is structured in an aggregated format: individuals with the same pattern of responses to the items are grouped together. The variable COUNT is included in the last column, indicating the number of individuals with that particular response pattern.

2.4 Example

The following code generates data for 1000 individuals from an LCA model with 3 latent classes and 5 items, and then analyzes the results in PROC LCA. Note that it is necessary to create the datasets containing the hypothesized true γ and ρ values before calling the macro.

```
%INCLUDE "C:\Work\lca\SimulateLcaDataset.sas";
DATA gammas;
      INPUT gammas;
      DATALINES;
      .5
      .3
      .2
RUN;

DATA rhos;
/*columns respond to latent classes; rows correspond to items*/
      INPUT rhos1 rhos2 rhos3 ;
      DATALINES;
      .3  .7  .5
      .3  .7  .5
      .7  .3  .5
      .7  .3  .5
      .7  .7  .5
RUN;

%SimulateLcaDataset(  true_gamma_dataset = gammas,
                      true_rho_dataset = rhos,
                      output_dataset_name = data1,
                      total_n = 1000 );

PROC LCA DATA=data1;
    NCLASS 3;
    ITEMS Item001 Item002 Item003 Item004 Item005 ;
    FREQ Count;
    CATEGORIES 2 2 2 2 2 ;
    SEED 1000;
    NSTARTS 10;
RUN;
```

Note that the FREQ statement must be used when analyzing data generated from the %SimulateLcaDataset macro, as data are aggregated by response pattern. The variable Count should be specified in this statement, as it contains the frequency count associated with each response pattern. See the PROC LCA & PROC LTA user's guide for more information on aggregated data (Lanza, Dziak, Huang, Xu, & Collins, 2011).

When we analyze the simulated data in this way, with the same number of classes (NCLASS 3) in the estimation model as in the data-generating model, it is easier to compare the resulting estimates to the original values. However, the simulated dataset can be analyzed in any way you wish; for example, it may be of interest to find out what kinds of estimates would result from imposing a 2-class model on a 3-class population or vice versa. In fact, this is part of the idea behind the parametric bootstrap test, and a version of %SimulateLcaDataset is therefore part of the Methodology Center's %LcaBootstrap macro.

References

Collins, L. M., & Lanza, S. T. (2010). *Latent class and latent transition analysis: With applications in the social, behavioral, and health sciences.* New York, NY: Wiley.

Lanza, S. T., Dziak, J. J., Huang, L., Xu, S., and Collins, L. M. (2011). *PROC LCA & PROC LTA users' guide* (version 1.2.6). University Park: The Methodology Center, Penn State. Available from http://methodology.psu.edu.