

%LCA\_Covariates\_3Step SAS Macro Users’ Guide

(Version 1.0)

**John J. Dziak**

**Bethany C. Bray**

**Aaron T. Wagner**

Penn State

Copyright © 2020 The Pennsylvania State University

ALL RIGHTS RESERVED

Please send questions and comments to *MChelpdesk@psu.edu*.

The development of the SAS %LCA\_Covariates\_3Step macro was supported by National Institute on Drug Abuse Grants P50 DA10075 and P50 DA039838. The authors would like to thank Amanda Applegate for helpful comments.

Thank you for citing this users’ guide when you use this macro. Suggested citation:

Dziak, J. J., Bray, B. C., & Wagner, A. T. (2020). *LCA\_Covariates\_3Step SAS macro users’ guide* (Version 1.0). University Park, PA: The Methodology Center, Penn State. Retrieved from [http://methodology.psu.edu](http://methodology.psu.edu/downloads/move)

Contents

1 About the %LCA\_Covariates\_3Step Macro 2

2 System Requirements 4

3 The BCH Approach to Covariates 5

4 Using the %LCA\_Covariates\_3Step Macro 8

4.1 Argument Definitions 8

4.2 Preparation 9

4.3 Estimation of the Latent Class Model in PROC LCA 9

4.4 Macro Syntax and Input 10

4.5 Output 11

5 Demonstrations of the %LCA\_Covariates\_3Step Macro 12

5.1 Estimating a Binary Distal Outcome 12

5.2 Estimating a Continuous Distal Outcome 17

5.3 Estimating a Count Distal Outcome 20

5.4 Estimating a Categorical Distal Outcome 24

6 Demonstration of the %LCA\_Covariates\_3Step Macro for Multiple Groups 28

6.1 Example Data 28

6.2 Example Syntax 29

7 Demonstration of Assignment and Adjustment Options 34

References 37

# About the %LCA\_Covariates\_3Step Macro

The SAS %LCA\_Covariates\_3Step macro estimates the association between a latent class variable and a distal outcome using the app roach of Bolck, Croon, and Hagenaar (2004), as adapted by Vermunt (2010) and Vermunt and Magidson (2015). The %LCA\_Covariates\_3Step macro is designed to work with SAS Version 9.1 or higher and PROC LCA.

**NOTE: This macro does not estimate covariates for LCA with a distal outcome.** Two of our macros employ the approach developed by Bolck, Croon, and Hagenaar, the %LCA\_Covariates\_3Step macro for incorporating covariates in a latent class model and the %LCA\_Distal\_BCH macro for estimating the association between a latent class variable and a distal outcome. These two macros are not designed to work together.

The %LCA\_Covariates\_3Step macro

* uses simple, minimal syntax;
* estimates class-specific response probabilities and standard errors for binary and categorical distal outcomes;
* estimates class-specific means and standard errors for continuous and count distal outcomes;
* provides significance tests to compare distal outcome means or proportions between classes
* can accommodate distal outcomes for multiple demographic groups.

This guide assumes the user has a working knowledge of latent class analysis and PROC LCA. The book, *Latent class and latent transition analysis: With applications in the social, behavioral, and health sciences* (Collins & Lanza, 2010), provides a comprehensive introduction to the use of latent class analysis in applied research.

To use this macro, you must have PROC LCA version 1.3.2 or higher installed. PROC LCA and the accompanying users’ guide can be downloaded from <http://methodology.psu.edu/downloads>.

The variables designated as covariates, for purposes of this macro, does not literally have to represent an event that occurs later in time than the original measurements. The macro was originally intended for describing the proportions of later events among classes (e.g., predicting relapse from latent class of withdrawal symptoms), but the same method works for describing other covariates that characterize classes (e.g., comparing psychological profiles or demographic variables among latent classes of drug withdrawal symptoms). However, the current macro works for one outcome at a time and currently does not allow adjusting for other outcomes simultaneously. Instead, it is intended to be run separately for each outcome or covariate of interest.

# System Requirements

The %LCA\_Covariates\_3Step macro requires

* SAS Version 9.1 or higher (Windows version),
* PROC LCA & PROC LTA Version 1.3.2 or higher (to fit LCA models),

Note: SAS/STAT is sold separately from the base SAS package, but most university licenses include it. If you can run PROC LCA, you can run this macro.

# The BCH Approach to Covariates

Researchers are often interested in the relationship between a latent class variable, *C,* and a distal outcome, *Z*. Often, they wish to compare the class-specific expected value E(*Z*|*C*=c) for each class *c*. This expected value is the same as the mean (average) for count or continuous variables. For binary variables coded as 0 or 1, the expected value is the proportion of 1s in the population, or equivalently, the probability of a 1 rather than a 0 for a single randomly selected population member.

The BCH method is a kind of “three-step” method. This means that (1) the parameters of the LCA model first are estimated without the distal outcome, then (2) the posterior probabilities of class membership based on this model are used to compute a special weighting variable, and finally (3) the weighting variable is used to calculate a weighted average of *Z* for each class. The simplest approach to creating weights is to use either the posterior probabilities themselves as weights (“proportional assignment”) or to round the highest probability for each subject to 1 and the others to zero (“modal assignment”), and to apply no further adjustment. However, this treats the posterior probabilities as if they were known quantities measuring degrees of class membership, and does not take into account uncertainty introduced by possible misclassification when estimating the model parameters. Bolck, Croon, and Hagenaars (2004) proposed a more accurate method that accounts for misclassification probabilities. Although they first proposed this method only in the case of categorical outcomes, Vermunt (2010) explained how to adapt it to continuous outcomes as well.

This macro will calculate distal outcome estimates with either modal or proportional assignment, and either with BCH adjustment (“BCH” estimates) or without it (“naïve” or “unadjusted” estimates). It is generally better to use BCH adjustment rather than unadjusted estimates. However, as long as BCH adjustment is being used, it usually does not matter very much whether modal assignment or proportional assignment is used. Occasionally, BCH assignment has been found to give an uninterpretable value (such as a negative probability); in this case, it is better to revert to the unadjusted assignment. The three steps followed by this macro are described further below.

**Step 1.** Fit the LCA model to define latent class memberships, using only the indicator variables **Y=***Y***1,…,**,*Y*m, without including the distal outcome *Z* in the model. This will provide posterior probabilities of class membership, , for each individual *i=*1,…,*N* in the dataset and each class *c=*1,…,*n*c.

**Step 2.** Construct the weights for use in calculating weighted averages for each class on the distal outcome. The details depend on the options chosen.

* Unadjusted Modal Assignment. For each individual *i* and each possible class *c*, define the class weights . Specifically, let if *c* is the most likely class (the maximum among ) for a given individual, and 0 otherwise. For example, if individual *i* is estimated to have a 60% chance of belonging to class 2, then individual *i* will count as 100% of a member of class 2 and 0% of other classes.
* Unadjusted Proportional Assignment. Define the class weights as for each individual *i* and each possible class *c*. For example, if individual *i* is estimated to have a 60% chance of belonging to class 2, then individual *i* will count as 60% of a member of class 2 when calculating weighted averages; the remaining 40% of the membership of individual *i* is divided among the remaining classes.
* BCH-Adjusted Modal Assignment. Calculate the misclassification matrix **D**. The entry in row *a* and column *b* of **D** represents the estimated probability that a subject who truly belongs to class *a* would be labeled as belonging to class *b*. Specifically, **D**ab is calculated as , where *N* is the number of subjects, is the unadjusted modal weight for individual *i* in class *b*, and is the estimated overall class probability P(*C*=*a*). Then calculate the vector of BCH weights using linear algebra as , where is the*N*×*n*cmatrix of unadjusted modal weights *w*.
* BCH-Adjusted Proportional Assignment**.** Sameas BCH-adjusted modal, but use the proportional weights for *w* instead of using the modal weights.

**Step 3*.***Estimate the expected value of the distal outcome within each latent class by taking a weighted average of the observed values for all participants, weighted by each participant’s value of or, as requested by the user. Standard errors are calculated using Taylor linearization (“sandwich” covariance estimation).

**Standard errors and tests**. In principle, there are two ways of doing tests, or obtaining standard errors or confidence intervals, for non-normal distal outcomes. One is to treat them as simply averages and ignore the fact that they are not normally distributed. This is convenient and asymptotically valid, although not the most statistically efficient. The other is to assume a non-normal distribution (here we use Bernoulli for binary and Poisson for count) and construct the confidence intervals or tests for the underlying parameter (the logit probability or log mean) of this distribution. This macro mostly imitates the behavior of the LatentGOLD software, in that standard errors are provided using the simpler method, and tests are performed using the more complicated method. For the binary case, non-symmetric confidence intervals are additionally provided using the more complicated method (calculating standard errors and confidence interval limits for the logit, and then back-transforming the ends of this confidence interval to describe the observed mean).

**Pairwise and omnibus tests**. The macro provides Wald tests and p-values for comparing the expected values of the distal outcome between each pair of latent classes, testing the null hypothesis that the expected values are equal. The p-values are not adjusted for multiple comparisons, but a user who wishes to apply a Bonferroni correction could simply divide the alpha level used for comparison (e.g., .05) by the number of pairs being compared: specifically, by )/2. In addition to these tests, an omnibus test simultaneously comparing all of the expected values is also performed. For categorical outcomes in the current version of the macro, only an omnibus test, rather than pairwise tests, is performed.

**Sampling weights**. If complex survey sample weights are used in the LCA (the weight option in PROC LCA) then these must be specified in this macro also (using the sampling\_weight= optional argument). Sampling weights are implemented by multiplying each by the corresponding sampling weight *s­i*. This is done before postmultiplying by in the BCH method. Note that although survey weights can be accommodated, the current version of the macro does not account for clustering when calculating standard errors.

**Grouping variable.** The calculations of the macro can accommodate an observed grouping variable (usually gender or other demographic categories) as in the groups command in PROC LCA. The macro assumes measurement invariance across groups and performs calculations separately for each group. Separate output is also provided for each group.

# Using the %LCA\_Covariates\_3Step Macro

## Argument Definitions

Table 1. Argument Definitions for the %LCA\_Covariates\_3Step Macro.

|  |  |  |
| --- | --- | --- |
| Argument | Required | Description |
| input\_data | Y | Input data set. The ID variable and covariates must all be included among the variables. |
| covariates | Y | The name of the covariates to be used in the multinomial regression to predict class membership. |
| id | Y | The name of the subject identification variable used in the datasets specified in input\_data and post. |
| postprobs | Y | Name of the data set generated by PROC LCA as the OUTPOST output when analyzing the input\_data dataset. The data set contains estimates of the posterior probabilities. The ID variable, along with the covariates, must be included in the ID statement of PROC LCA. **The model itself must not use the COVARIATES statement in PROC LCA.** |
| groups | N | The name of the variable, if any, by which analyses should be done separately by category. If specified, it must also be used in the GROUPS statement in LCA. |
| adjustment | N | The method, if any, of adjusting the class membership weights for the possibility of misclassification. This may be “BCH” (default, recommended) or “unadjusted,” without quotes. |
| assignment | N | The method of generating class membership weights based on the posterior probabilities, before doing the BCH adjustment if any. This may be “modal” (default) or “proportional,” without quotes. |
| ref\_class | N | The reference class for the multinomial regression; this class’s regression parameters are not estimated because of identifiability constraints. All other classes are compared to this one. The default is 1 as in PROC\_LCA. |
| automatically\_add\_intercept | N | Whether to automatically include an intercept column (1) or not (0). The default is 1. |
| sampling\_weight | N | Name of the variable specifying survey weight. If used, it assumes that WEIGHT has also been used in the previous call to PROC LCA. |

## Preparation

A SAS macro is a special block of SAS commands. The block is first defined and then called when needed. Four steps need to be completed before you run the macro.

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

%INCLUDE “S:\myfolder\LCA\_Covariates\_3Step\_v10.sas”;

1. Direct SAS to the input data file. We assume the data set is a permanent file saved to a designated directory. If so, we recommend using a “libname” statement.The statement should give the libname command, name the library, and then identify the path to the data. For example,

libname sasf “s:\myfolder\”;

## Estimation of the Latent Class Model in PROC LCA

Use PROC LCA to generate the output needed for use by the %LCA\_Covariates\_3Step macro. First, you must select the LCA model. This process is described in Chapter 5 of the *PROC LCA & PROC LTA Users’ Guide* (Lanza, Dziak, Huang, Xu, & Collins, 2011).

Once model selection is complete, generate a file containing the posterior probabilities to be used in the macro by estimating the latent class model with the covariates. This file can be generated using the OUTPOST option in PROC LCA. (See section 5.3 of the *PROC LCA & PROC LTA Users’ Guide* for more information.) The PROC LCA syntax will be similar to the following:

PROC LCA DATA = my\_data OUTPOST = my\_post; /\* the input data set and the file to be generated containing the posterior probabilities \*/

NCLASS 5; /\*the number of latent classes \*/

ITEMS item001 item002 item003 item004 item005 item006 item007 item008; /\*indicator variablesused to measure the latent class variable \*/

CATEGORIES 2 2 2 2 2 2 2 2; /\* number of response categories for each indicator variable (in this case, all dichotomous) \*/

ID SubjectID /\*the unique integer representing each case \*/

SEED 54327;/\* an arbitrary number to be used as a seed for generating reproducible random starting values \*/

RUN;

The covariates statement should not be used. The group or weight statement may be used, if demographic groups or survey weights are required in the model. Other arguments available in PROC LCA, such as rho prior, maxiter, and criterion may be necessary for estimation of the latent class model. Refer to the *PROC LCA& PROC LTA Users’ Guide* for more information.

## Macro Syntax and Input

Call the macro using a percent sign, its name, and user-defined arguments in parentheses. The macro parameters are shown below.

%LCA\_Covariates\_3Step\_v10(

input\_data = *data set name,*

postprobs = *name of OUTPOST data set created by PROC LCA,*

covariates = *variables*

id = *variable,*

groups = *variable,*

sampling\_weight *= survey weighting variable name,*

adjustment *= word describing the misclassification adjustment method (BCH or unadjusted),*

assignment *= word describing the class membership weight assignment option (modal or proportional)*

ref\_class = *deafult = 1*

automatically\_add\_intercept = *1 or 0 (for automatically including an intercept)*

) ;

## Output

The macro produces both screen output and SAS datasets. The screen output first presents a table of estimates and standard errors for the expected value of the distal outcome within each class. In addition, for binary distal outcomes, a table of log odds estimates and asymmetric confidence intervals is provided. The macro then provides a table of Wald chi-squared tests for testing the equality of expected values between classes. These include both pairwise and omnibus tests, except for categorical distal outcomes, for which only omnibus tests are provided.

Two SAS datasets, Distal\_Estimates and Distal\_Tests, are also created. These contain similar information to what is shown on screen. For binary outcomes, a dataset called Distal\_Log\_Odds is also created. Although these datasets contain the same information that is shown on screen, they can be useful if you want to copy and save the results of many analyses into a larger compilation (e.g., in a simulation loop).

# Demonstrations of the %LCA\_Covariates\_3Step Macro

In this section, we first describe the structure of the data sets and the variables to be analyzed. Then, we illustrate how to estimate the distribution of the distal outcome within each latent class using the %LCA\_Covariates\_3Step macro and describe the output of the macro. Section 5.1 describes use of the macro with a binary distal outcome. Continuous, count, and categorical outcomes are discussed in sections 5.2, 5.3, and 5.4, respectively.

For demonstrations of the macro with multiple groups, see chapter 6.

## Estimating a Binary Distal Outcome

Before attempting to complete the following example, please download the file *%LCA\_Distal Examples* from the %LCA\_Distal macros download page at [http://methodology.psu.edu](http://methodology.psu.edu/). Also, verify that you are running PROC LCA v.1.3.2 or higher.

### Example Data

Below are the first 10 observations from the SAS data set **simdata\_binary.sas7bdat**, which is containedin the *%LCA\_Distal Examples* file.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ID | Item001 | Item002 | Item003 | Item004 | Item005 | Item006 | Item007 | Item008 | Z |
| 1 | 2 | 2 | 1 | 2 | 2 | 2 | 2 | 2 | 1 |
| 2 | 1 | 1 | 2 | 2 | 2 | 2 | 2 | 2 | 0 |
| 3 | 2 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 0 |
| 4 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 1 |
| 5 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 1 |
| 6 | 1 | 1 | 1 | 2 | 2 | 2 | 2 | 2 | 1 |
| 7 | 2 | 2 | 1 | 2 | 2 | 2 | 2 | 2 | 1 |
| 8 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 1 |
| 9 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 1 |
| 10 | 2 | 2 | 2 | 2 | 1 | 2 | 2 | 2 | 1 |

ID= subject’s identification variable,

Item001,…, Item008= 8 items used to measure the latent class variable

*Z*= the distal outcome (Note: binary distal outcome should be coded using 0s and 1s.)

### Example Syntax

Include a “libname” statement prior to running the macro to direct SAS to the data file.

libname sasf "S:\myfolder\";

Note: We suppose that the SAS data set exists in the folder S:\myfolder\. This path represents any user-specified folder.

Once the LCA model has been identified, estimate the LCA model using PROC LCA. Notice that Z is not included as a covariate in this step.

**PROC** **LCA** DATA = SimData\_Binary OUTPARAM = Binary\_param OUTPOST = Binary\_post;

ID id;

NCLASS **5**;

ITEMS item001-item008;

CATEGORIES **2** **2** **2** **2** **2** **2** **2** **2**;

SEED **12345**;

RHO PRIOR = **1**;

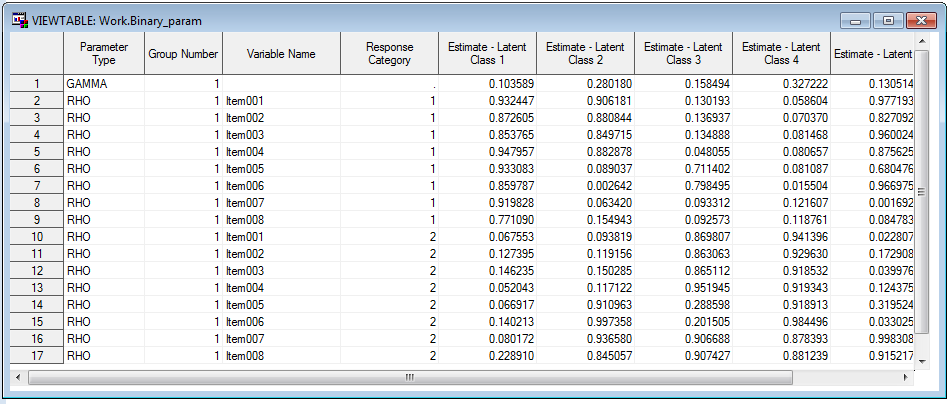
NSTARTS **20**;

MAXITER **5000**;

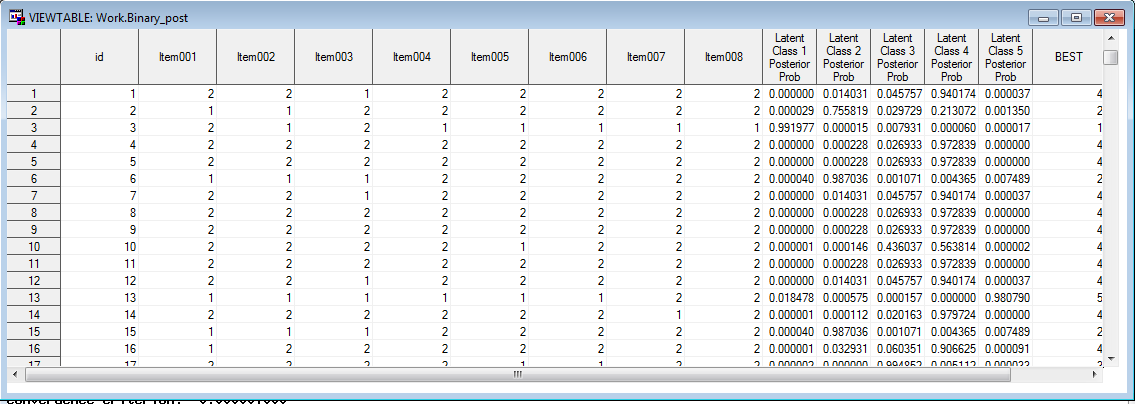
CRITERION **0.000001**;

**RUN**;

The output is described in the *PROC LCA & PROC LTA Users’ Guide.* It should include the files Binary\_param and Binary\_post in the WORK directory.



Binary\_param



Binary\_post

Now the distal outcomes macro can be run. Include the macro and enter the proper syntax in SAS.

%***LCA\_Distal\_BCH***(input\_data = SimData\_Binary,

param = Binary\_param,

post = Binary\_post,

id = id,

distal = z,

metric = binary );

The input\_dataargument identifies the data file. The param argument directs the macro to the parameters in the outparam file generated by PROC LCA. The post argument directs the macro to the posterior probabilities in the outpost file generated by PROC LCA. The id variable identifies the column in the dataset that uniquely identifies subjects. The distal argument identifies the distal outcome variable in the data set. The metric argument indicates that the distal outcome is binary.

In this example there were no survey weights. If there had been, it would be necessary to add a line such as

WEIGHT SurveyWeight;

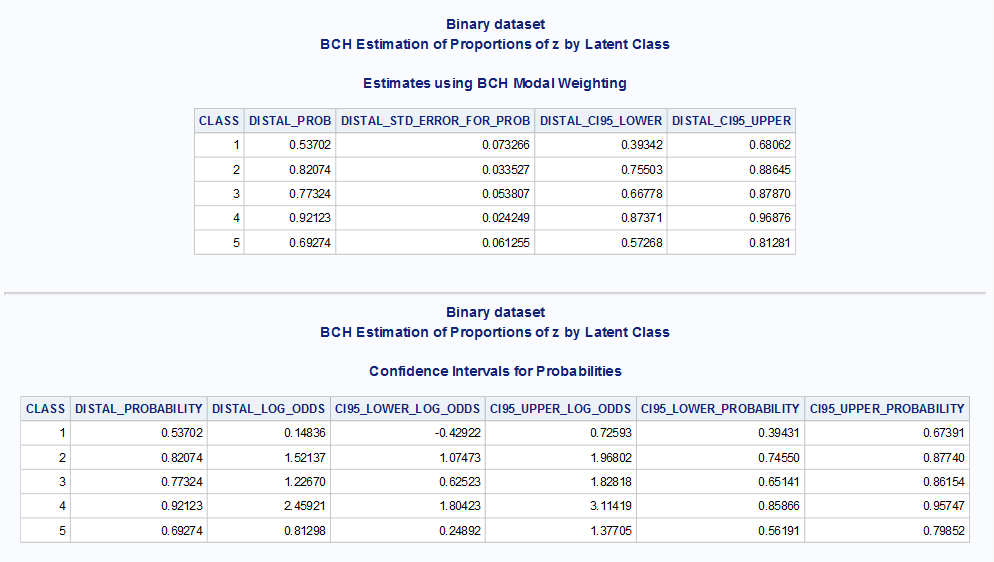
to the PROC LCA syntax and a line such as

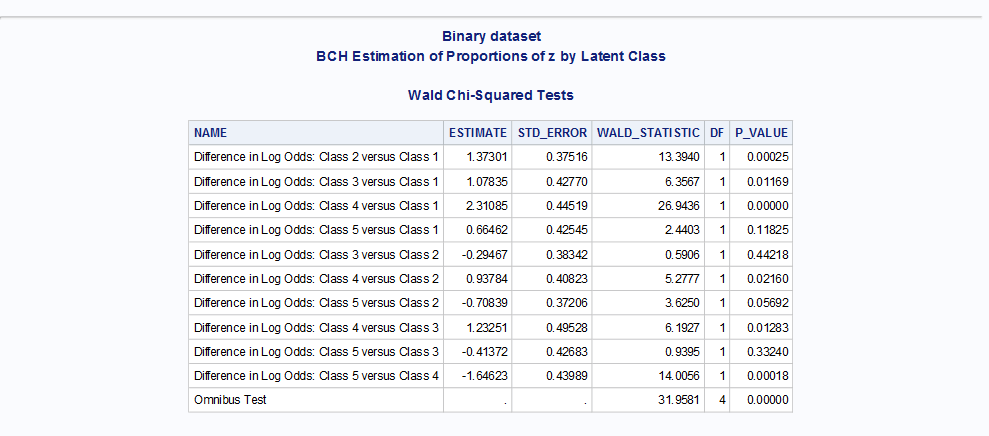
sampling\_weight=SurveyWeight,

to the macro syntax.

### Example Output

Below is the onscreen output. It includes the class-specific distribution estimates for the distal outcome, the estimated class-conditional probabilities, the Wald test statistic on class-conditional probabilities, and the p-value on class-conditional probabilities.





### Overall Response Proportions

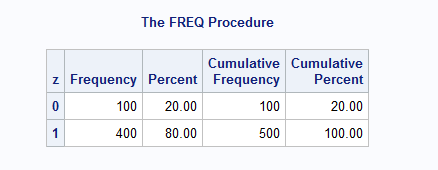
When interpreting the estimated response proportions within each of the latent classes, it may be useful to compare them to the overall estimated response proportion, ignoring latent class. This can be accomplished in the usual way using PROC FREQ (if survey weights are not used) or by using PROC SURVEYFREQ with its weight statement (if survey weights are being used). For example, one can use the syntax

PROC FREQ DATA= SimData\_Binary;

TABLES z;

RUN;

In the artificial dataset provided for this example, exactly 80% of the distal outcomes are yes (1).



Technical note: PROC LCA (and therefore %LCA\_Distal\_BCH) ignores participants who omit all of the answers to the indicators (items). If there are many subjects who omit all items, then the subsample being described by %LCA\_Covariates\_3Step may noticeably differ from the whole sample. If so, the user might consider omitting these subjects before running PROC FREQ or PROC SURVEYFREQ, for compatibility with the results found in %LCA\_Distal\_BCH. However, in most cases this will probably not be necessary, because most participants will answer at least some of the LCA items.

## Estimating a Continuous Distal Outcome

Before attempting to complete the following example, please download the file *%LCA\_Distal Examples* from the %LCA\_Covariates\_3Step macro download page.

### Example Data

In **simdata\_conti.sas7bdat**, the data structure is similar to the data set in section 5.1 of this document. However, instead of binary values for z, the values are continuous.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ID | Item001 | Item002 | Item003 | Item004 | Item005 | Item006 | Item007 | Item008 | Z |
| 1 | 2 | 1 | 2 | 2 | 2 | 2 | 2 | 2 | -1.8513098 |
| 2 | 1 | 1 | 2 | 1 | 2 | 2 | 2 | 2 | -0.5950087 |
| 3 | 2 | 2 | 2 | 1 | 1 | 1 | 2 | 2 | 1.55437269 |
| 4 | 2 | 1 | 2 | 2 | 2 | 2 | 2 | 2 | 0.89742276 |
| 5 | 1 | 1 | 1 | 1 | 2 | 2 | 2 | 2 | -0.3121734 |
| 6 | 1 | 1 | 1 | 1 | 2 | 2 | 2 | 2 | -1.5068341 |
| 7 | 2 | 2 | 1 | 2 | 2 | 2 | 2 | 2 | 0.73713821 |
| 8 | 1 | 1 | 1 | 2 | 1 | 1 | 2 | 2 | 1.8747736 |
| 9 | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 1 | -0.0463611 |
| 10 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | -0.1706686 |

ID= subject’s identification variable

Item001,…, Item008= 8 items used to measure the latent class variable

*Z*= the distal outcome (in this case a CONTINUOUS distal outcome)

### Example Syntax

Include a “libname” statement prior to running the macro to direct SAS to the data file.

libname sasf "S:\myfolder\";

Note: we suppose that the SAS data set exists in the folder S:\myfolder\. This path represents any user-specified folder.

Estimate the LCA model using PROC LCA.

**PROC** **LCA** DATA = SimData\_conti OUTPARAM = conti\_param OUTPOST = conti\_post ;

ID id;

NCLASS **5**;

ITEMS item001-item008;

CATEGORIES **2** **2** **2** **2** **2** **2** **2** **2**;

SEED **12345**;

RHO PRIOR = **1**;

NSTARTS **20**;

MAXITER **5000**;

CRITERION **0.000001**;

**RUN**;

Now, include the macro and enter the following syntax in SAS.

%***LCA\_Distal\_BCH***(input\_data = SimData\_conti,

param = conti\_param,

post = conti\_post,

id = id,

distal = z,

metric = Continuous );

The input\_data argument identifies the data file. The paramargument directs the macro to the parameters generated in the OUTPARAM file generated by PROC LCA. The id and distal arguments identify the subject identification variable and the distal outcome. The metric argument indicates that the distal outcome is continuous, and output\_dataset\_name names the macro’s output.

In this example there were no survey weights. If there had been, it would be necessary to add a line such as

WEIGHT SurveyWeight;

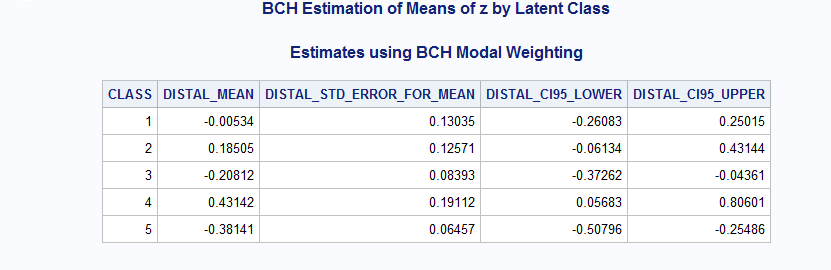
to the PROC LCA syntax and a line such as

sampling\_weight=SurveyWeight,

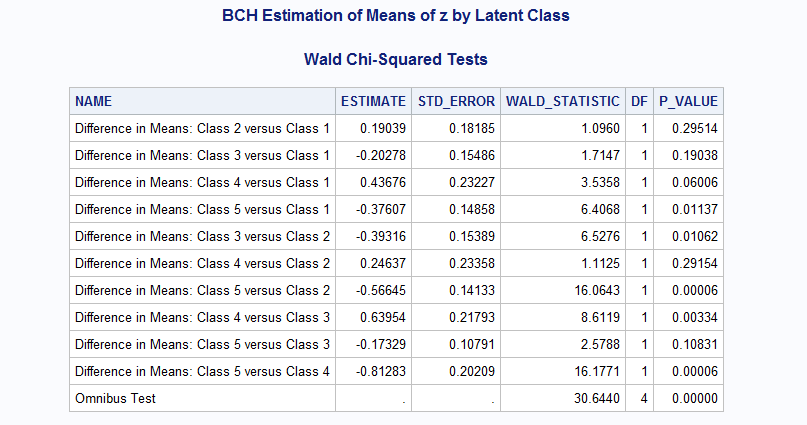
to the macro syntax.

### Example Output

The estimated means, along with standard errors and 95% confidence intervals, are shown in the output below.



Tests of the differences between means are shown in the output below.



These output tables are also generated as datasets, namely distal\_estimates and distal\_tests.

### Overall Response Means

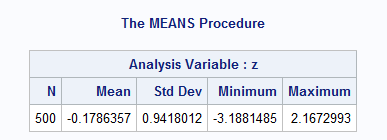
When interpreting the estimated response means within each of the latent classes, it may be useful to compare them to the overall estimated response mean, ignoring latent class. This can be accomplished using PROC MEANS (if survey weights are not used) or by using PROC SURVEYMEANS with the weight statement (if survey weights are being used). For example, one can use the syntax

PROC MEANS DATA= SimData\_conti;

VAR z;

RUN;

In the artificial dataset provided for this example, the mean of the distal outcome is -0.1786357.



Technical note: PROC LCA (and therefore %LCA\_Distal\_BCH) ignores participants who omit all of the answers to the indicators (items). If there are many subjects who omit all items, then the subsample being described by %LCA\_Covariates\_3Step may noticeably differ from the whole sample. If so, the user might consider omitting these subjects before running PROC MEANS or PROC SURVEYMEANS, for compatibility with the results found in %LCA\_Distal\_BCH. However, in most cases this will probably not be necessary, because most participants will answer at least some of the LCA items.

## **Estimating a Count Distal Outcome**

Before attempting to complete the following example, please download the file *%LCA\_Distal Examples* from the %LCA\_Distal macros download page.

### Example Data

In **simdata\_count.sas7bdat**, the data structure is similar to the dataset in section 5.1 of this document. However, the item *z* contains count responses with values from 0 to 4.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ID | Item001 | Item002 | Item003 | Item004 | Item005 | Item006 | Item007 | Item008 | Z |
| 1 | 2 | 2 | 2 | 2 | 1 | 1 | 2 | 1 | 2 |
| 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 0 |
| 3 | 2 | 1 | 1 | 1 | 2 | 1 | 2 | 2 | 0 |
| 4 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 1 | 0 |
| 5 | 2 | 2 | 1 | 2 | 2 | 2 | 2 | 2 | 0 |
| 6 | 1 | 1 | 1 | 1 | 2 | 1 | 2 | 2 | 1 |
| 7 | 2 | 2 | 2 | 2 | 1 | 1 | 2 | 2 | 0 |
| 8 | 1 | 1 | 1 | 1 | 2 | 2 | 2 | 1 | 0 |
| 9 | 2 | 2 | 1 | 1 | 2 | 2 | 2 | 2 | 1 |
| 10 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 1 |

ID= subject’s identification variable

Item001,…, Item008= 8 items used to measure the latent class variable

Z= the distal outcome (in this case a COUNT distal outcome)

### Example Syntax

Include a “libname” statement prior to running the macro to direct SAS to the data file.

libname sasf "S:\myfolder\";

Note: we suppose that the SAS data set exists in the folder S:\myfolder\. This path represents any user-specified folder.

Estimate the LCA model using PROC LCA.

**PROC** **LCA** DATA = SimData\_Count OUTPARAM = Count\_param OUTPOST = Count\_post;

ID id;

NCLASS **5**;

ITEMS item001-item008;

CATEGORIES **2** **2** **2** **2** **2** **2** **2** **2**;

SEED **12345**;

RHO PRIOR = **1**;

NSTARTS **20**;

MAXITER **5000**;

CRITERION **0.000001**;

**RUN**;

Then, call the macro.

%***LCA\_Distal\_BCH***(input\_data = SimData\_Count,

param = Count\_param,

post = Count\_post,

id=id,

distal = z,

metric = Count );

The input\_dataargument identifies the data file. The paramargument directs the macro to the parameters generated in the OUTPARAM file generated by PROC LCA. The id and distalarguments identify the subject identification variable and the distal outcome. The metric argument indicates that the distal outcome is continuous, and output\_dataset\_name names the macro’s output.

In this example there were no survey weights. If there had been, it would be necessary to add a line such as

WEIGHT SurveyWeight;

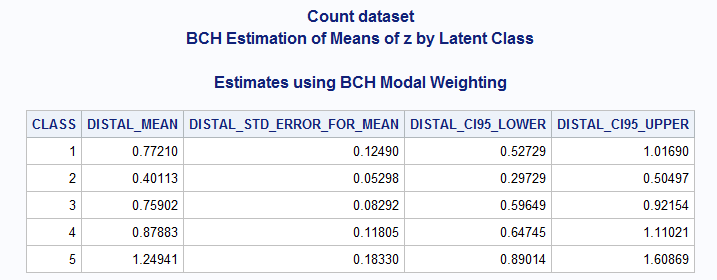
to the PROC LCA syntax and a line such as

sampling\_weight=SurveyWeight,

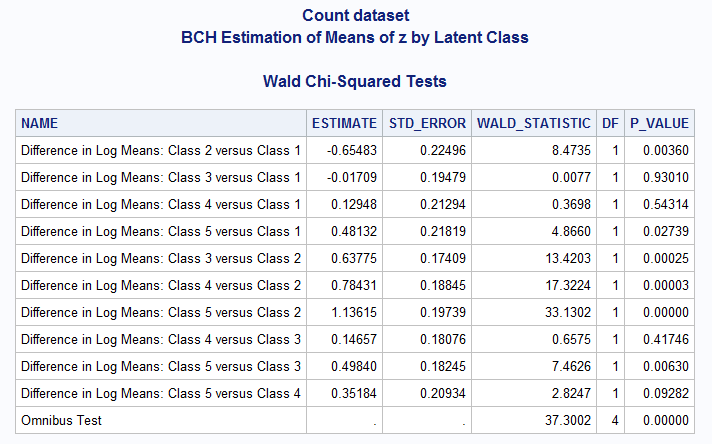
to the macro syntax.

### Example Output

The first table shows the estimated distal outcome means within each class.



The second shows tests of equality of the means between different classes.



These output tables are also generated as datasets, namely distal\_estimates and distal\_tests.

### Overall Response Means

When interpreting the estimated response means within each of the latent classes, it may be useful to compare them to the overall estimated response mean, ignoring latent class. This can be accomplished in the usual way using PROC MEANS (if survey weights are not used) or by using PROC SURVEYMEANS with the weight statement (if survey weights are being used). This is described further in the corresponding subsection for continuous outcomes; it is not necessary to specify to PROC MEANS or PROC SURVEYMEANS that the response is count rather than continuous.

## Estimating a Categorical Distal Outcome

Before attempting to complete the following example, please download the file *%LCA\_Distal Examples* from the %LCA\_Distal macros download page.

### Example Data

First, we will examine the structure of the database and the variables to be analyzed. Below are the first 10 observations from the SAS data set **simdata\_categ.sas7bdat**, which is containedin the *%LCA\_Distal Examples* file available at [http://methodology.psu.edu](http://methodology.psu.edu/)

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ID | Item001 | Item002 | Item003 | Item004 | Item005 | Item006 | Item007 | Item008 | Z |
| 1 | 2 | 2 | 2 | 2 | 1 | 2 | 2 | 2 | 2 |
| 2 | 2 | 1 | 2 | 2 | 2 | 2 | 2 | 2 | 3 |
| 3 | 1 | 1 | 2 | 1 | 1 | 2 | 2 | 1 | 2 |
| 4 | 2 | 1 | 1 | 1 | 2 | 1 | 2 | 2 | 3 |
| 5 | 2 | 2 | 2 | 1 | 2 | 2 | 2 | 2 | 1 |
| 6 | 2 | 2 | 2 | 2 | 1 | 2 | 2 | 2 | 1 |
| 7 | 1 | 1 | 1 | 1 | 1 | 2 | 2 | 2 | 2 |
| 8 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 1 | 3 |
| 9 | 2 | 2 | 2 | 2 | 2 | 2 | 1 | 2 | 2 |
| 10 | 2 | 2 | 2 | 1 | 1 | 1 | 2 | 1 | 1 |

**ID**= subject’s identification variable

**Item001,…, Item008**= 8 items used to measure the latent class variable

***Z***= the distal outcome (Note: The categorical distal outcome should be coded using 1, 2, 3, …, g, where g = the number of categories.)

### Example Syntax

Once the LCA model has been identified, estimate the LCA model using PROC LCA.

**PROC** **LCA** DATA = SimData\_Categ OUTPARAM = Categ\_param OUTPOST = Categ\_post;

ID id;

NCLASS **5**;

ITEMS item001-item008;

CATEGORIES **2** **2** **2** **2** **2** **2** **2** **2**;

SEED **12345**;

RHO PRIOR = **1**;

NSTARTS **20**;

MAXITER **5000**;

CRITERION **0.000001**;

**RUN**;

The output is described in the *PROC LCA & PROC LTA Users’ Guide*.

Then, include and run the macro.

%***LCA\_Distal\_BCH***(input\_data = SimData\_Categ,

param = Categ\_param,

post = Categ\_post,

id = id,

distal = z,

metric = categorical );

The input\_dataargument identifies the data file. The paramargument directs the macro to the parameters in the OUTPARAM file generated by PROC LCA. The id and distal argumentsidentifies the subject identification and distal outcome variable in the data set. The metric argument indicates that the distal outcome is categorical, and output\_dataset\_name names the macro’s output.

In this example there were no survey weights. If there had been, it would be necessary to add a line such as

WEIGHT SurveyWeight;

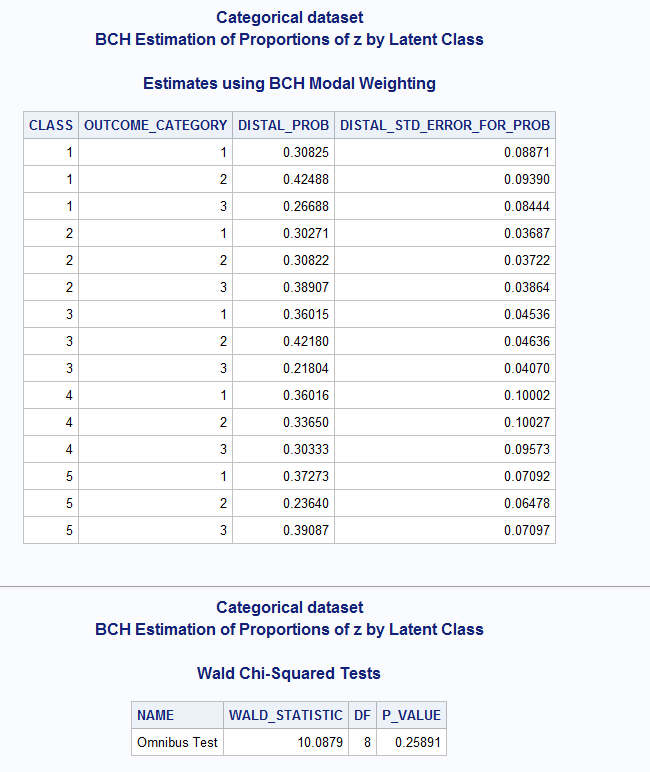
to the PROC LCA syntax, and a line such as

sampling\_weight=SurveyWeight,

to the macro syntax.

### Example Output

The onscreen output contains the estimated proportions of each response category within each latent class.



The contents of the output are stored in the distal\_estimates and distal\_tests datasets, respectively.

### Overall Response Proportions

When interpreting the estimated response proportions within each of the latent classes, it may be useful to compare them to the overall estimated response proportion, ignoring latent class. This can be accomplished in the usual way using PROC FREQ (if survey weights are not used) or by using PROC SURVEYFREQS with the weight statement (if survey weights are being used). This is described further in the corresponding subsection for binary outcomes; it is not necessary to specify to PROC FREQ or PROC SURVEYFREQ that the response is count rather than continuous.

# Demonstration of the %LCA\_Covariates\_3Step Macro for Multiple Groups

In this section, we first describe the structure of the data sets and the variables to be analyzed when there are multiple groups. Then, we illustrate how to estimate the distribution of the distal outcome within each latent class using the %LCA\_Covariates\_3Step macro and describe the output of the macro. This section describes use of the macro with a binary distal outcome. The results with other outcomes are very similar. Before attempting to complete the following example, please download the file *%LCA\_Distal Examples* from the %LCA\_Distal macros download page. Also, verify that you are running PROC LCA v.1.3.2 or higher.

## Example Data

Below are 10 putative observations from the SAS data set **simdata\_binary\_group.sas7bdat**, which is containedin the *%LCA\_Distal Examples* file available at [http://methodology.psu.edu](http://methodology.psu.edu/).

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ID | Item001 | Item002 | Item003 | Item004 | Item005 | Item006 | Item007 | Item008 | Z | Educ |
| 1 | 2 | 2 | 1 | 2 | 2 | 2 | 2 | 2 | 1 | 1 |
| 2 | 1 | 1 | 2 | 2 | 2 | 2 | 2 | 2 | 0 | 1 |
| 3 | 2 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 0 | 1 |
| 4 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 1 | 2 |
| 5 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 1 | 2 |
| 6 | 1 | 1 | 1 | 2 | 2 | 2 | 2 | 2 | 1 | 2 |
| 7 | 2 | 2 | 1 | 2 | 2 | 2 | 2 | 2 | 1 | 3 |
| 8 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 1 | 3 |
| 9 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 1 | 3 |
| 10 | 2 | 2 | 2 | 2 | 1 | 2 | 2 | 2 | 1 | 3 |

ID= subject’s identification variable,

Item001,…, Item008= 8 items used to measure the latent class variable,

*Z*= the distal outcome (Note: binary distal outcome should be coded using 0s and 1s.)

Educ=the variable for multiple groups.

## Example Syntax

Include a “libname” statement prior to running the macro to direct SAS to the data file.

libname sasf "S:\myfolder\";

Note: we suppose that the SAS data set exists in the folder S:\myfolder\. This path represents any user-specified folder.

Once the LCA model has been identified, estimate the LCA model including the distal outcome *Z* as a covariate and *Educ* as the grouping variable using PROC LCA.

**PROC** **LCA** DATA = simdata\_Binary\_group OUTPARAM = Binary\_param OUTPOST = Binary\_post ;

ID id;

NCLASS **5**;

ITEMS item001-item008;

CATEGORIES **2** **2** **2** **2** **2** **2** **2** **2**;

SEED **12345**;

RHO PRIOR = **1**;

NSTARTS **20**;

GROUP educ;

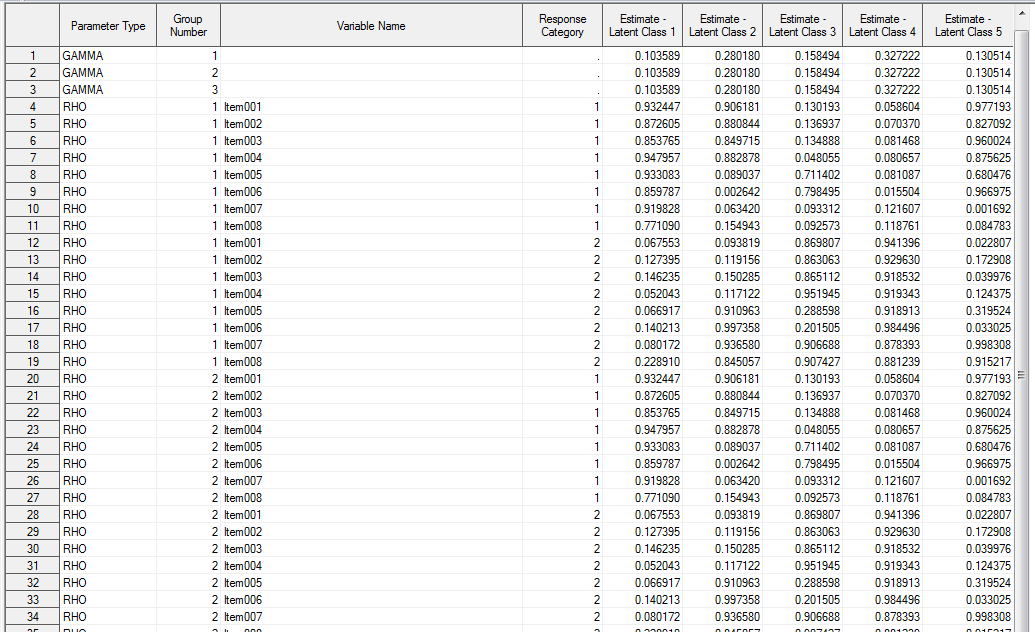
MAXITER **5000**;

CRITERION **0.000001**;

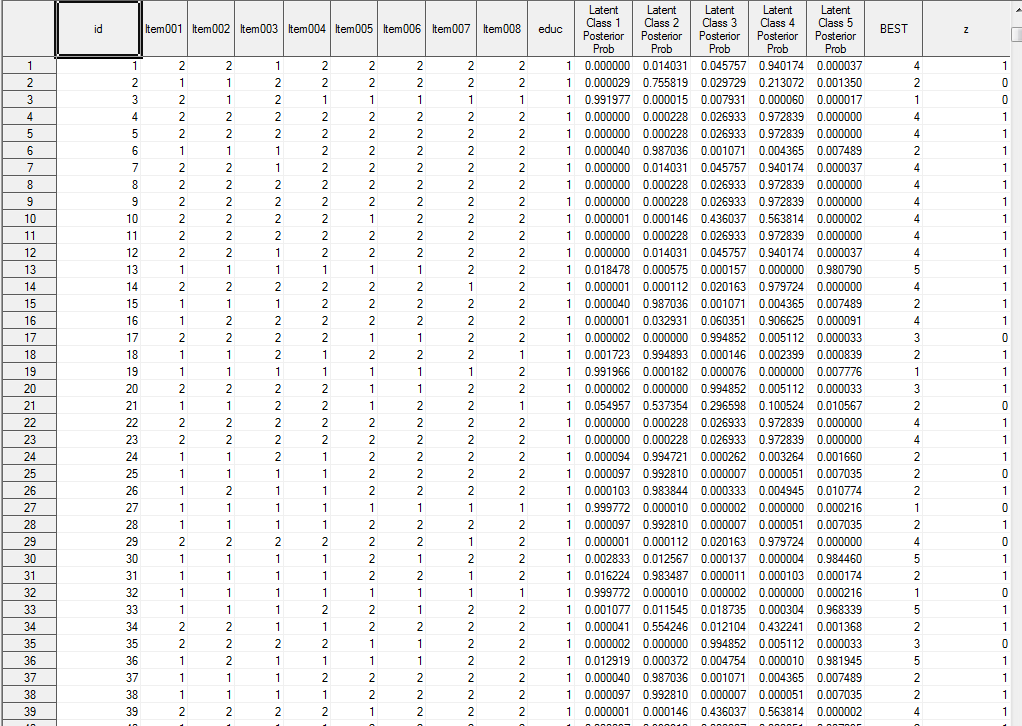
**RUN**;

The output is described in the *PROC LCA & PROC LTA Users’ Guide*.

The output will also include the files Binary\_param and Binary\_post in the WORK directory.



Binary\_param



Binary\_post

Now, include and run the macro:

%***LCA\_Distal\_BCH***(input\_data = simdata\_Binary\_group,

param = Binary\_param,

post = Binary\_post,

id = id,

group = educ,

distal = z,

metric = Binary );

The input\_dataargument identifies the data file. The param argument directs the macro to the parameters in the OUTPARAM file generated by PROC LCA. The id and distal argument identify the subject identification variable and distal outcome variable in the data set. The group argument identifies the variable for multiple groups. The metric argument indicates that the distal outcome is binary.

In this example there were no survey weights. If there had been, it would be necessary to add a line such as

WEIGHT SurveyWeight;

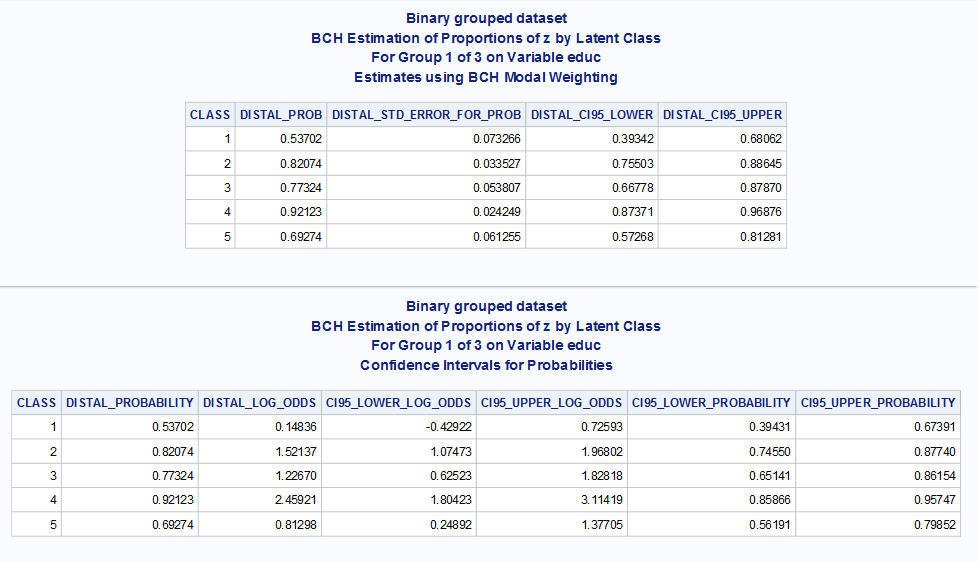
to the PROC LCA syntax, and a line such as

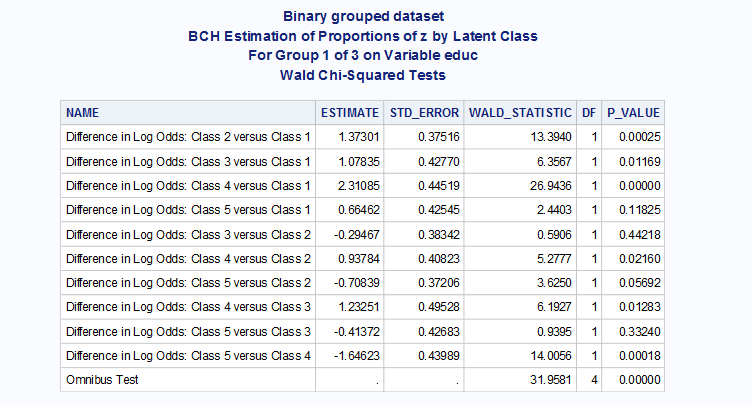
sampling\_weight=SurveyWeight,

to the macro syntax.

### Example Output

Below is the onscreen output for the first group on the educ variable. Similar output follows for the second and third groups.





# Demonstration of Assignment and Adjustment Options

Shown here are four different approaches to distal outcome analysis for the binary example. All give roughly similar answers in this example. Simulation studies suggest that the BCH answers may be more accurate than the unadjusted answers (see Chapter 3).

TITLE "Modal unadjusted";

%***LCA\_Distal\_BCH***(input\_data = SimData\_Binary,

param = Binary\_param,

post = Binary\_post,

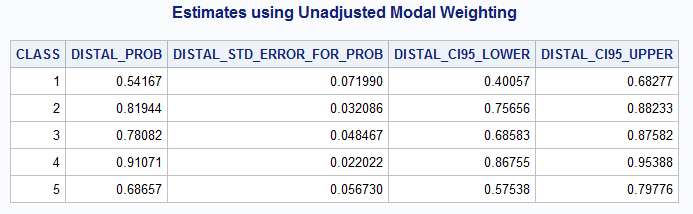
distal = z,

id = id,

metric = binary ,

adjustment\_method = unadjusted,

assignment = modal);



TITLE "Proportional unadjusted";

%***LCA\_Distal\_BCH***(input\_data = SimData\_Binary,

param = Binary\_param,

post = Binary\_post,

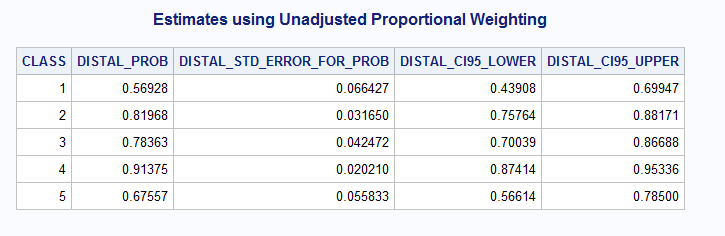
distal = z,

id = id,

metric = binary ,

adjustment\_method = unadjusted,

assignment = proportional);



TITLE "Modal BCH";

%***LCA\_Distal\_BCH***(input\_data = SimData\_Binary,

param = Binary\_param,

post = Binary\_post,

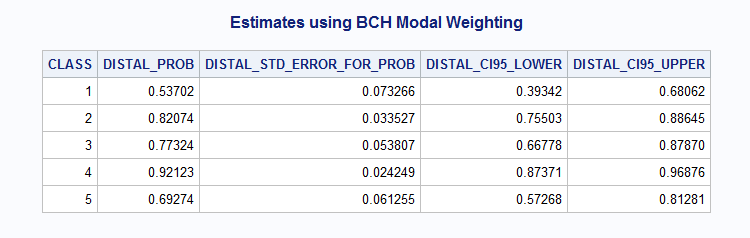
distal = z,

id = id,

metric = binary ,

adjustment\_method = BCH,

assignment = modal);



TITLE "Proportional BCH";

%***LCA\_Distal\_BCH***(input\_data = SimData\_Binary,

param = Binary\_param,

post = Binary\_post,

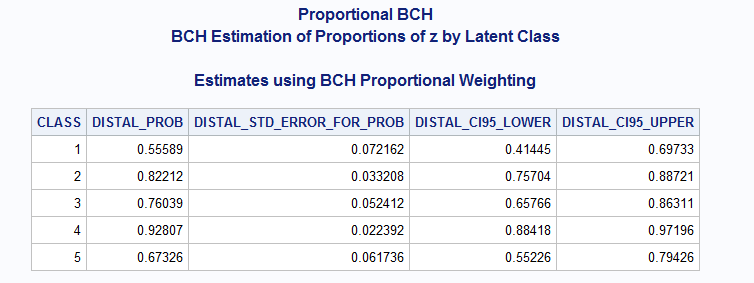
distal = z,

id = id,

metric = binary,

adjustment\_method = BCH,

assignment = proportional);



# References

Bakk, Z., Oberski, D. L., & Vermunt, J. K. (2014). Relating latent class assignments to external variables: Standard errors for corrected inference. *Political Analysis, 22*, 520-540. doi:10.1093/pan/mpu003

Bakk, Z., & Vermunt, J. K. (2016). Robustness of stepwise latent class modeling with continuous distal outcomes. *Structural Equation Modeling, 23,* 20-31*.* doi:10.1080/10705511.2014.955104

Bolck, A., Croon, M., & Hagenaars, J. (2004). Estimating latent structure models with categorical variables: One-step versus three-step estimators. *Political Analysis*, *12*(1), 3–27.

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.

Dziak, J. J., Bray, B. C., Zhang, J. - T., Zhang, M., & Lanza, S. T. (2016). Comparing the performance of improved classify-analyze approaches in latent profile analysis. Methodology: *European Journal of Research Methods for the Behavioral and Social Sciences, 12,* 107-116. http://doi.org/10.1027/1614-2241/a000114

Lanza, S. T., Dziak, J. J., Huang, L., Wagner, A. T., & Collins, L. M. (2015). *Proc LCA & Proc LTA users' guide* (Version 1.3.2). University Park: The Methodology Center, Penn State. Available from [methodology.psu.edu](http://methodology.psu.edu)

Lanza, S. T., Tan, X., & Bray, B. C. (2013). Latent class analysis with distal outcomes: A flexible model-based approach. *Structural Equation Modeling: A Multidisciplinary Journal, 20*, 1-20.

Schafer, J. L. (1997) *Analysis of incomplete multivariate data*. Boca Raton, FL: Chapman and Hall/CRC.

Vermunt, J. K. (2010). Latent class modeling with covariates: Two improved three-step approaches. *Political Analysis, 18,* 450–469.

Vermunt, J. K., & Magidson, J. (2015). *Upgrade manual for Latent GOLD 5.1*. Belmont, MA: Statistical Innovations, Inc.