PennState The Methodology Center
advancing methods, improving health

LCA_Distal_LTB Stata function users' guide (Version 1.1)

Liying Huang
John J. Dziak
Bethany C. Bray
Aaron T. Wagner
Stephanie T. Lanza
Penn State

NOTE: the functionality of this function is identical to the (no longer distributed) LCA Distal Stata function (v. 1.0). For estimating LCA with a distal outcome, we recommend using the LCA_Distal_BCH Stata function. See section 1 for more information.

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

The development of the LCA_Distal_LTB Stata function was supported by National Institute on Drug Abuse Grant P50 DA039838. Jingyun Yang and Xianming Tan made significant contributions to earlier versions of this software. The authors would like to thank Amanda Applegate for her helpful comments.

Thank you for citing this users' guide when you use this macro. Suggested citation:
Huang, L., Dziak, J. J., Bray, B. C., Wagner, A. T., & Lanza, S. T. (2017). *LCA_Distal_LTB Stata function users' guide* (Version 1.0). University Park: The Methodology Center, Penn State. Retrieved from http://methodology.psu.edu

Contents

1 About the LCA_Distal_LTB Stata function

The LCA_Distal_LTB Stata function estimates the association between a latent class variable and a distal outcome using a model-based approach. The LCA_Distal_LTB Stata function is designed to work with Stata Version 11.0 or higher and the LCA Stata plugin, version 1.2.1 or higher. This function is based on an earlier function created by The Methodology Center, LCA Distal Stata function version 1.0. The name of the function has been changed to distinguish it from another the LCA_Distal_BCH Stata function.

Note: Currently, the recommended procedure for LCA with a distal outcome for any metric (e.g., binary, count, continuous) is a classification error adjusted approach, colloquially known as the BCH approach. The BCH approach can be implemented using the LCA_Distal_BCH Stata function, which is also available for free download at methodology.psu.edu. The LCA_Distal_LTB Stata function is provided primarily for methodological research, such as comparing results using the BCH approach and the LTB approach.

The LCA_Distal_LTB Stata function

- uses simple, minimal syntax;
- estimates class-specific probabilities for binary distal outcomes; and
- can accommodate distal outcomes for multiple groups.

This guide assumes the user has a working knowledge of latent class analysis and the LCA Stata plugin. 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. The LCA Stata plugin and the accompanying users' guide can be downloaded from http://methodology.psu.edu/downloads. A detailed explanation of the model-based estimation approach used in the LCA_Distal_LTB Stata function appears in Lanza, Tan, and Bray (2013).

Note: To use this function, you must also use the LCA Stata plugin Version 1.2.1 or higher.

2 System requirements

The LCA_Distal_LTB Stata function requires

- Stata Version 11.0 or higher (Windows version) and
- LCA Stata plugin Version 1.2.1 or higher (to fit LCA models).

3 The LTB approach to LCA with distal outcomes

3.1 Recommended method for estimating LCA with distal outcomes

Currently, the recommended procedure for LCA with a distal outcome for any metric (e.g., binary, count, continuous) is a classification error adjusted approach, colloquially known as the BCH approach. The BCH approach can be implemented using the LCA_Distal_BCH Stata function, which is also available for free download at methodology.psu.edu. The LCA_Distal_LTB Stata function is provided primarily for methodological research, such as comparing results between the BCH approach and the LTB approach.

3.2 LTB Approach to LCA with a distal outcome for a single group

Researchers are often interested in the relationship between a latent class variable, C, and a distal outcome, Z. The LCA_Distal_LTB Stata function provides a model-based approach to estimating the association between C and Z; that is, the conditional distribution of Z given C, $f\{Z|C\}$.

In order to estimate $f\{Z|C\}$, we assume conditional independence between an observed item in the model, X, and Z given C. Stated mathematically, we assume that $f\{X,Z|C\} = f\{X|C\}\, f\{Z|C\}$. This conditional independence assumption is similar to the local independence assumption essential to most LCA models (Collins & Lanza, 2010).

In LCA with a distal outcome, interest lies in the conditional distribution $f\{Z|C\}$. The distribution of $Z|C$ can be determined by applying Bayes' theorem:

$$f\{Z=z\,|\,C=c\} = \frac{f\{Z=z\}\times f\{C=c\,|\,Z=z\}}{f\{C=c\}}. \quad (1)$$

To obtain each element in the model,

- $f\{C = c\}$ is derived from the LCA model,
- $f\{C = c|Z = z\}$ is derived from the LCA model with Z included as a covariate, and
- $f\{Z = z\}$ is the marginal distribution of Z, and can be empirically estimated.

We include Z in the LCA model as a covariate, and assume conditional independence between X and Z given C. The conditional distribution of $f\{Z = z|C = c\}$ for Z can be calculated by re-expressing (1) as follows:

$$\Pr\{Z = z \mid C = c\} = \frac{\Pr\{Z = z\} e^{\beta_{0c} + \beta_{1c} z}}{\Pr\{C = c\}(1 + \sum_{c'=1}^{K-1} e^{\beta_{0c'} + \beta_{1c'} z})} . \quad (2)$$

Specifically, $\Pr\{Z = z\}$ is estimated from the empirical distribution of Z (i.e., from the proportions in the observed data); the estimates for $\{\beta_{0c}, \beta_{1c}; c = 1,2,\ldots, K-1\}$ are provided by the LCA with covariates model; and the marginal distribution $\Pr\{C = c\}$ can be obtained by multiplying $\Pr\{C = c|Z = z\}$ by the marginal empirical distribution $\Pr\{Z = z\}$. Thus, we can obtain $\Pr\{Z = z|C = c\}$ given these estimates for $\Pr\{Z = z\}$, $\Pr\{C = c\}$, and $\{\beta_{0c}, \beta_{1c}; c = 1,2,\ldots, K-1\}$.

The function handles binary distal outcomes. It does not accommodate count, categorical, or continuous outcomes.

An overview of the LCA model, a thorough explanation of the model-based approach to LCA with a distal outcome, and results from a Monte Carlo study for evaluating this approach can be found in a technical report by Lanza, Tan, and Bray (2011), which can be downloaded from http://methodology.psu.edu.

3.3 LTB appraoch to LCA with a distal outcome for multiple groups

In many cases, there are multiple groups G (e.g., male and female) in the data, and researchers are interested in conditional distribution of Z given C and G. Similarly, within each group, we assume conditional independence between any given observed item in the model and the outcome Z given C.

We are interested in the conditional distribution of $f(Z = z \mid C = c, G = g)$. Using Bayes' theorem, this can be expressed as

$$f(Z = z \mid C = c, G = g) = \frac{f(C = c \mid Z = z, G = g) f(Z = z \mid G = g) f(G = g)}{f(G = g) f(C = c \mid G = g)} = \frac{f(Z = z \mid G = g) f(C = c \mid Z = z, G = g)}{f(C = c \mid G = g)}. \quad (3)$$

To obtain each element in the model,

- $f(C=c \mid G=g)$ is derived from the LCA model conditional on *G,*
- $f(C=c \mid Z=z, G=g)$ is derived from the LCA model with Z included as a covariate conditional on G, and
- $f(Z=z \mid G=g)$, the marginal distribution of Z conditional on *G*, is estimated empirically.

4 Using the LCA_Distal_LTB Stata function

Table 1. Option definitions for the LCA_Distal_LTB Stata function

Option	Required	Description
`varname`	Yes	Distal outcome variable.
`gammaList`	Yes	List of estimates of the gamma parameters. Generated from the gamma matrix generated by LCA Stata plugin.
`betaList`	Yes	List of estimates of the beta parameters. Generated from the beta matrix generated by LCA Stata plugin.
`cov_blist`	Yes	List of estimates of the betas' covariance. Generated from the betas' covariance matrix generated by LCA Stata plugin.
`groups`	No	Variable for multiple groups. If no group argument is supplied, the function assumes there is only one group.
`alpha`	No	Significance level. Default = 0.05
`weight`	No	Name of the variable specifying survey weight. This option only works in the binary outcome case. It assumes that `weight` has also been used in the run of the LCA Stata plugin.

4.1 Managing files and preparing data

Three steps are required to set up the function before use.

1. Set up the LCA Stata plugin as described in the *LCA Stata plugin users' guide.*
2. Unzip the files in the LCA_Distal_LTB Stata function folder downloaded from methodology.psu.edu/downloads and place all the files in the desired folder on your computer.
3. Run an example:
    a. Open relevant ".do" file
    b. In the **4th line of code,** modify the path "D:\project\Stata_lca\distal_lca_ltb_32-bit_ to match the folder path where you placed the files. **NOTE: If there are any spaces in your directory path, you will need to put the path in double quotation marks, per Stata convention.**
    c. Save the changes.

The function is ready to use.

Note: Missingness in the distal outcome variable should be imputed before running the function (e.g., multiple imputation; Schafer, 1997). Otherwise, cases with missing values in the distal outcome variable must be removed from the analysis.

4.2 Estimation of the latent class model in the LCA Stata plugin

Use the LCA Stata plugin to generate the output needed for use by the LCA_Distal_LTB Stata function. First, select the LCA model. This process is described in chapter 5 of the *LCA Stata Plugin Users' Guide*.

Once model selection is complete, generate the matrices containing the parameter estimates to be used in the plugin by estimating the latent class model with the distal outcome included as a covariate.

```
qui doLCA Item001  Item002 Item003 Item004 Item005 Item006
    Item007 Item008, ///
      nclass(5) ///
       id(ID)                  ///
       maxiter(5000)              ///
       groups (educ)                   ///
       seed(10)               ///
       covariates( z )   ///
       categories(2 2 2 2 2 2 2 2)///
       criterion(0.000001)        ///
       rhoprior(1.0)
```

4.3 Converting matrices to lists

Note: Section 4.3 is a basic Stata operation and not part of our function, but we include it here as a convenient review for some users.

The LCA_Distal_LTB Stata function relies on inputs from the LCA Stata plugin, but the LCA Stata plugin generates matrices, and the LCA_Distal_LTB Stata function (by Stata convention) cannot accept matrices as input for options. This means that three matrices—r(gamma), r(beta) and r(covb)—must be converted to lists after running the LCA Stata plugin and before running the LCA_Distal_LTB Stata function. The following code can be used for this purpose.

```
mat G = r(gamma)
mat B = r(beta)
mat C = r(covb)

forvalues i=1/`=rowsof(G)' {
      forvalues j=1/`=colsof(G)' {
      local glist `glist' `=G[`i', `j']'
      }
}

forvalues i=1/`=rowsof(B)' {
      forvalues j=1/`=colsof(B)' {
      local blist `blist' `=B[`i', `j']'
      }
}

forvalues i=1/`=rowsof(C)' {
      forvalues j=1/`=colsof(C)' {
      local covblist `covblist' `=C[`i', `j']'
      }
}
```

Then, the lists—`‘glist’`, `‘blist’`, and `‘covblist’`—can be used in the LCA_Distal_LTB Stata function.

4.4 Function options and input

Run the function with user-defined options in parentheses. Only the name of the distal outcome variable and the `groups` and `weight` (if used) are likely to change.

```
    doLCA_Distal_LTB z,  ///
       gammaList(`glist') ///
       betaList(`blist') ///
       covbList(`covblist') ///
       groups(educ) /*INCLUDED ONLY WHEN GROUPS ARE IN THE DATA*/

return list
```

4.5 Output

If there is a single group, the output will always contain a set of estimates representing the association between membership in each latent class and the distal outcome. The onscreen results will include a row for each latent class, and the probability associated with each level of the distal outcome will be listed for each class.

The function will also display results onscreen. The results displayed onscreen include

- the estimated probability of each outcome within each latent class,
- confidence intervals for the class-specific probabilities, and
- a Wald chi-squared test for every pair of classes to test whether their class-specific probabilities are significantly different. (This is expressed as the Wald chi-squared test statistic and as the p-value associated with the test statistic.)

The additional output created in the binary case is also saved in three new matrices.

- r(pi_by_class): the probability that distal outcome = 1 given each class, with confidence bounds, either 95% or as specified in the `alpha` option.
- r(Wald): the test statistics of the class-conditional probabilities
- r(Wald_pvalue): the p-values for the class-conditional probabilities.

5 Empirical demonstration of the LCA_Distal_LTB Stata function

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_Distal_LTB Stata function and describe the output of the macro.

For more demonstrations, run the files included with the download of the LCA_Distal_LTB Stata function , available at methodology.psu.edu/downloads. This section of the users' guide is based on *distal-example0.do.*

Before attempting to complete the following example, please locate the example files that were included when you downloaded the LCA Distal Stata function from http://methodology.psu.edu. Also, verify that you are running LCA Stata plugin version 1.2.1 or higher.

5.1 Example data

Below are the first 10 observations from the data set **Example0.dct**, which is contained in the files downloaded with the function.

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: the distal outcome should be coded using 0s and 1s.)

5.2 Example syntax

First, estimate the latent class model in the LCA Stata plugin. Include the distal outcome as a covariate.

```
qui doLCA Item001  Item002 Item003 Item004 Item005 Item006 Item007
Item008, ///
      nclass(5) ///
       id(ID)                 ///
       maxiter(5000)  ///
       groups (educ) ///
       seed(10) ///
       covariates( z )   ///
       categories(2 2 2 2 2 2 2 2)///
       criterion(0.000001)  ///
       rhoprior(1.0)
```

The output is described in the *LCA Stata Plugin Users' Guide*.

Then, three matrices—r(gamma), r(beta), and r(covb)—need to be converted to lists so they can be used as input options in the LCA_Distal_LTB Stata function.

```
mat G = r(gamma)
mat B = r(beta)
mat C = r(covb)

forvalues i=1/`=rowsof(G)' {
      forvalues j=1/`=colsof(G)' {
      local glist `glist' `=G[`i', `j']'
      }
}

forvalues i=1/`=rowsof(B)' {
      forvalues j=1/`=colsof(B)' {
      local blist `blist' `=B[`i', `j']'
```

```
	}
}

forvalues i=1/`=rowsof(C)' {
	forvalues j=1/`=colsof(C)' {
	local covblist `covblist' `=C[`i', `j']'
	}
}
```

Next, the lists—(‘glist’), (‘blist’), and (‘covblist’)—can be used in the LCA_Distal_LTB Stata function.

```
doLCA_Distal_LTB z,   ///
        gammaList(`glist') ///
        betaList(`blist') ///
        covbList(`covblist') ///
        groups(educ) /*INCLUDED ONLY WHEN GROUPS ARE IN THE DATA*/

return list
```

The option `z` is the name of the distal outcome variable in this code. The `gammaList` option directs the function to the parameters in the list you just created. The `betaList` option directs the function to the estimates of the beta parameters in the list you just created. The `covbList` option directs the function to the betas’ covariance matrix in the list you just created. The `groups` option identifies the grouping variable in the data.

5.3 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. This information will be repeated for each group.

Estimated Class-Conditional Probabilities for P(z=1|Class=c,educ=1)

Class	Probability	Confidence Interval For Probability
1	0.9364	(0.6892, 0.9899)
2	0.8226	(0.7404, 0.8829)
3	0.5485	(0.3864, 0.7009)
4	0.6768	(0.5629, 0.7729)
5	0.7453	(0.6105, 0.8452)

Wald Test on Class-Conditional Probabilities

Classes	Wald	df	p-value
1 vs. 2	1.0160	1	0.3135
1 vs. 3	5.3061	1	0.0213
1 vs. 4	3.2849	1	0.0699
1 vs. 5	2.4524	1	0.1173
2 vs. 3	10.3763	1	0.0013
2 vs. 4	5.1812	1	0.0228
2 vs. 5	1.0841	1	0.2978
3 vs. 4	1.4320	1	0.2314
3 vs. 5	3.2440	1	0.0717
4 vs. 5	0.5937	1	0.4410

Estimated Class-Conditional Probabilities for P(z=1|Class=c,educ=2)

Class	Probability	Confidence Interval For Probability
1	0.5462	(0.4571, 0.6324)
2	0.4135	(0.3391, 0.4920)
3	0.5063	(0.3678, 0.6438)
4	0.4818	(0.3638, 0.6019)

The same information will also be returned as matrices:

- r(wald)
- r(wald_pvalue)
- r(pi_given_class)

The matrices are returned for technical use. They contain the data summarized in the onscreen output.

6 References

Dziak, J. D., Yang, J., Tan, X., Bray, B. C., Wagner, A. T., & Lanza, S. T. (2017). *LCA_Distal_LTB SAS macro users' guide* (Version 1.0). University Park: The Methodology Center, Penn State. Retrieved from 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.

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., Wagner, A. T., & Collins, L. M. (2015). *LCA Stata plugin users' guide* (Version 1.2.1). University Park: The Methodology Center, Penn State. Available from methodology.psu.edu

Lanza, S. T., Tan, X., & Bray, B. C. (2011) *A technical introduction: A model-based approach to latent class analysis with distal outcomes*. (Technical Report No. 11-116). University Park: The Methodology Center, Penn State. Retrieved from methodology.psu.edu

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