Latent class analysis with Stata

Isabel Canette

Principal Mathematician and Statistician

StataCorp LLC

2018 Mexican Stata Users Group Meeting
Tlaxcala, August 16-17, 2018

Introduction

"Latent class analysis" (LCA) comprises a set of techniques used to model situations where there are different subgroups of individuals, and group memebership is not directly observed, for example:.

- Social sciences: a population where different subgroups have different motivations to drink.
- Medical sciences: using available data to identify subgroups of risk for diabetes.
- Survival analysis: subgroups that are vulnerable to different types of risks (competing risks).
- Education: identifying groups of students with different learning skills.
- Market research: identifying different kinds of consumers.

The scope of the term “latent class analysis” varies widely from source to source.

Collin and Lanza (2010) discuss some of the models that are usually considered LCA. Also, they point out: “ In this book, when we refer to latent class models we mean models in which the latent variable is categorical and the indicators are treated as categorical”.

In Stata, we use " LCA" to refer to a wide array of models where there are two or more unobserved classes

- Dependent variables might follow any of the distributions supported by **gsem**, as logistic, Gaussian, Poisson, multinomial, negative binomial, Weibull, etc.(**help gsem family and link options**)
- There might be covariates (categorical or continuos) to explain the dependent variables
- There might be covariates to explain class membership

Stata adopts a model-based approach to LCA. In this context, we can see LCA as group analysis where the groups are unknown.

Let's see an example, first with groups and then with classes:

Below we use **group()** option fit regressions to the childweight data, weight vs age, different regressions per sex:

```
. gsem (weight <- age), group(girl) ginvariant(none) ///
>   vsquish nodvheader noheader nolog
```

Group : boy — Number of obs = 100

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
weight						
age	3.481124	.1987508	17.52	0.000	3.09158	3.870669
_cons	5.438747	.2646575	20.55	0.000	4.920028	5.957466
var(e.weight)	2.4316	.3438802			1.842952	3.208265

Group : girl — Number of obs = 98

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
weight						
age	3.250378	.1606456	20.23	0.000	2.935518	3.565237
_cons	4.955374	.2152251	23.02	0.000	4.533541	5.377207
var(e.weight)	1.560709	.2229585			1.179565	2.06501

Group analysis allows us to make comparisons between these equations, and easily set some common parameters. (**help gsem group options**)

Now let's assume that we have the same data, and we don't have a group variable. We suspect that there are two groups that behave different.

```
. gsem (weight <- age),  lclass(C 2)  lcinvariant(none) ///
>   vsquish nodvheader noheader nolog
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
1.C	(base outcome)					
2.C						
_cons	.5070054	.2725872	1.86	0.063	-.0272557	1.041267

Class : 1

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
weight						
age	5.938576	.2172374	27.34	0.000	5.512798	6.364353
_cons	3.8304	.2198091	17.43	0.000	3.399582	4.261218
var(e.weight)	.6766618	.1817454			.3997112	1.145505

Class : 2

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
weight						
age	2.90492	.2375441	12.23	0.000	2.439342	3.370498
_cons	5.551337	.4567506	12.15	0.000	4.656122	6.446551
var(e.weight)	1.52708	.2679605			1.082678	2.153893

The second table on the LCA model same structure as the output from the group model.

In addition, the LCA output starts with a table corresponding to the class estimation. This is a binary (**logit**) model used to find the two classes.

In the latent class model all the equations are estimated jointly and all parameters affect each other, even when we estimate different parameters per class.

How do we interpret these classes? We need to analyze our classes and see how they relate to other variables in the data. Also, we might interpret our classes in terms of a previous theory, provided that our analysis is in agreement with the theory. We will see post-estimation commands that implement the usual tools used for this task.

Let's compute the class predictions based on the posterior probability.

```
. predict postp*, classposteriorpr
. generate pclass = 1 + (postp2>0.5)
. tabulate pclass
```

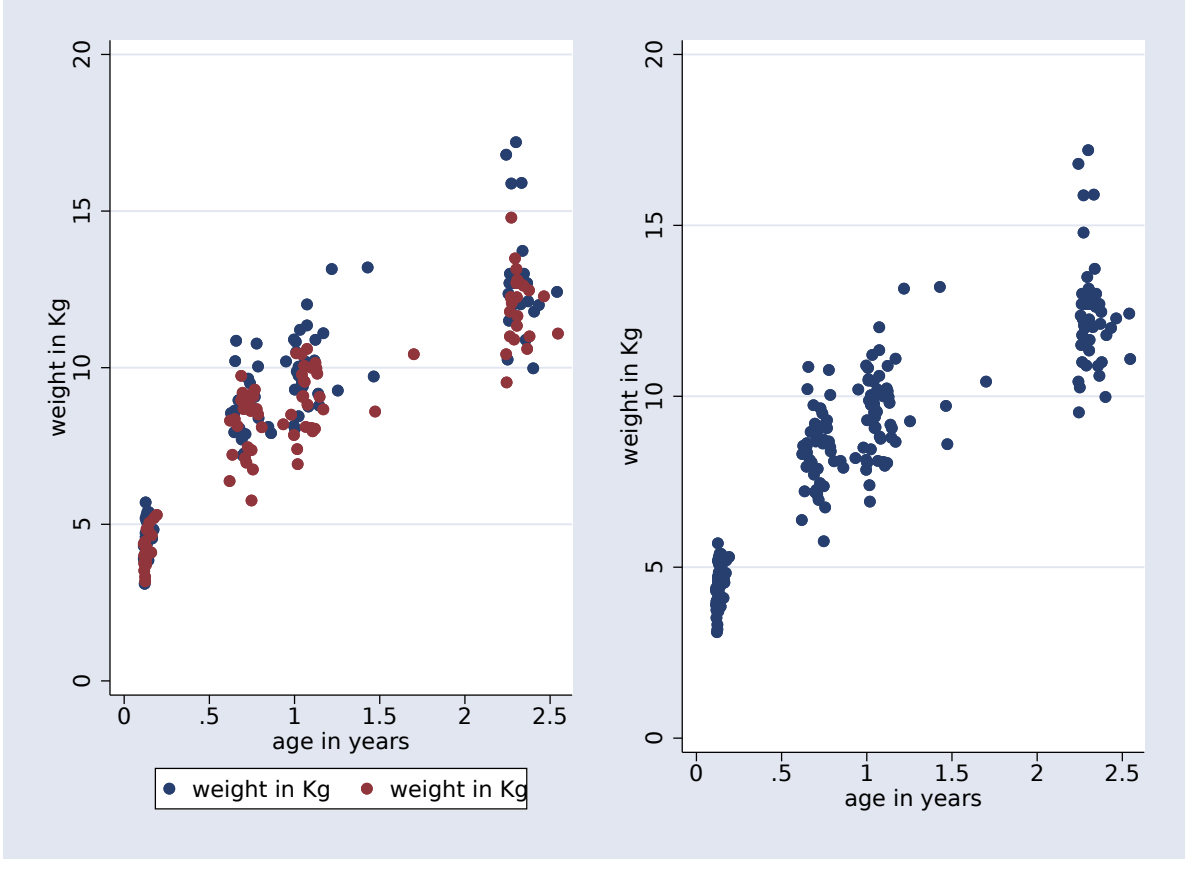
pclass	Freq.	Percent	Cum.
1	78	39.39	39.39
2	120	60.61	100.00
Total	198	100.00	

```
. tabulate pclass girl
```

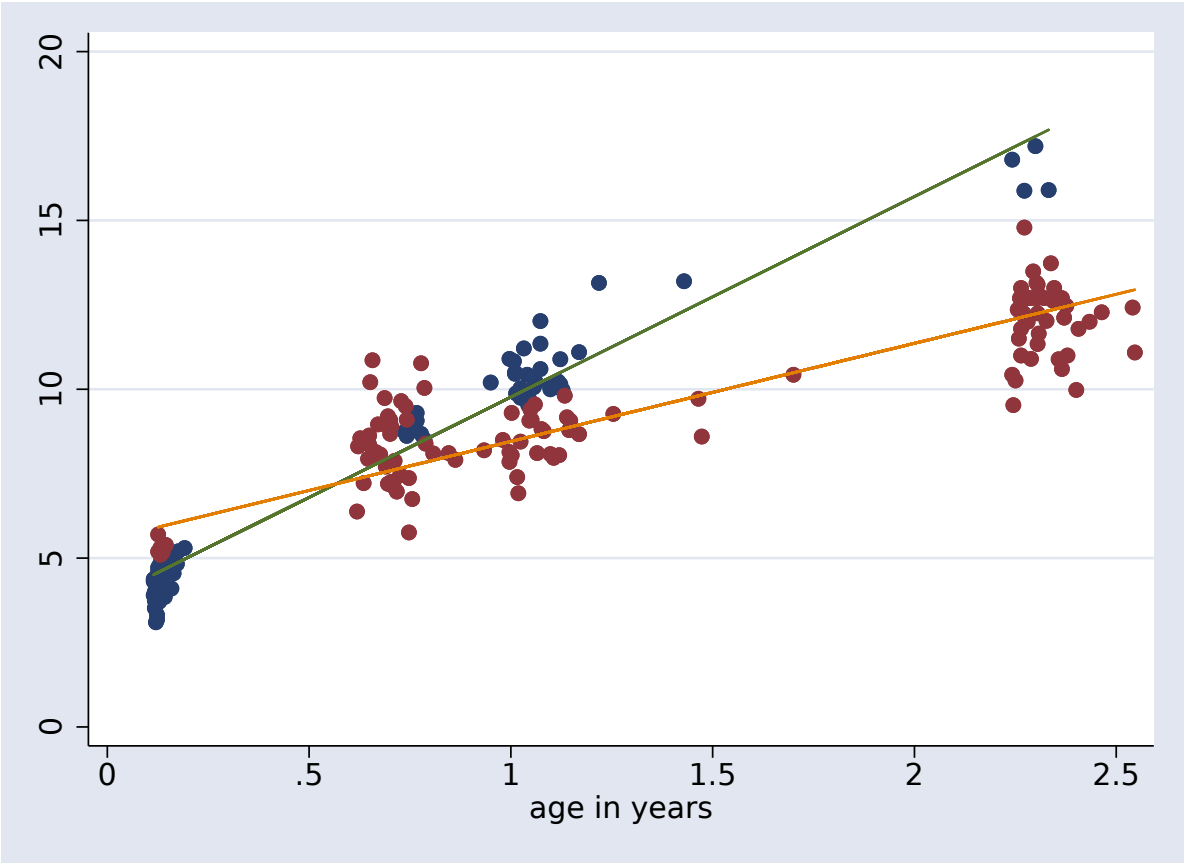
	gender		
pclass	boy	girl	Total
1	40	38	78
2	60	60	120
Total	100	98	198

Let's see some graphs.

```
. twoway scatter weight age if girl == 0 || ///
>     scatter weight age if girl == 1, saving(weighta, replace)
(file weighta.gph saved)

. twoway scatter weight age, saving(weightb, replace)
(file weightb.gph saved)

. graph combine weighta.gph weightb.gph
```



```
. predict mu*, mu

. twoway scatter weight age if pclass ==1 || ///
>        scatter weight age if pclass ==2 || ///
>        line mu1  age if pclass ==1 || ///
>        line mu2  age if pclass ==2 , legend(off)
```



gsem did exactly what we asked for: tell me what are the two more likely groups for two different linear regressions.

This approach allows us to generalize LCA in different directions, for example, if we had more information:

- we could incorporate more than one equation:

```
.  gsem (weight <- age ) (height <- age ), ///
> lclass(C 3) lcinvariant(none)
```

- we could incorporate class predictors:

```
.  gsem (weight <- age ) (height <- age ), ///
(C <- diet_quality) lclass(C 2) lcinvariant(none)
```

Estimation

For a dependent variables $\mathbf{y} = y_1, \ldots y_n$ and g groups for a given observation (i.e. no observation index below), the likelihood is computed as:

$$f(\mathbf{y}) = \sum_{i=1}^{g} \pi_i f_i(\mathbf{y}|\mathbf{z}_i), where :$$

- $\mathbf{z}_i$ is the vector of linear forms for class i, i.e., $z_{iji} = \mathbf{x}'\beta_{ij}$, where x are the dependent variables, and β_{ij} are the coefficients for main equation j, (conditional on) class i.
- f_i is the joint likelihood of $y = y_1, \ldots y_n$ conditional on class i
- the probabilities of belonging to each class $\pi_i, i = 1, \ldots, g$ are computed using a multinomial model,

$$\pi_i = \frac{\exp(\gamma_i)}{\sum_{k=1}^{g} \exp(\gamma_k)}.$$

$\gamma_k, k = 2, \ldots g$ is the linear form class k in the latent class equation, $\gamma_1 = 1$.

Classic LCA Example: Role conflict dataset

This is a classic example of LCA, where researchers use 4 binary variables to classify a sample.

```
. use gsem_lca1
(Latent class analysis)

. notes in 1/4

_dta:
  1.  Data from Samuel A. Stouffer and Jackson Toby, March 1951, "Role conflict
      and personality", _The American Journal of Sociology_, vol. 56 no. 5,
      395-406.
  2.  Variables represent responses of students from Harvard and Radcliffe who
      were asked how they would respond to four situations. Respondents
      selected either a particularistic response (based on obligations to a
      friend) or universalistic response (based on obligations to society).
  3.  Each variable is coded with 0 indicating a particularistic response and 1
      indicating a universalistic response.
  4.  For a full description of the questions, type "notes in 5/8".
```

```
. describe

Contains data from gsem_lca1.dta
  obs:           216                          Latent class analysis
 vars:             4                          10 Oct 2017 12:46
 size:           864                          (_dta has notes)
```

variable name	storage type	display format	value label	variable label
accident	byte	%9.0g		would testify against friend in accident case
play	byte	%9.0g		would give negative review of friend´s play
insurance	byte	%9.0g		would disclose health concerns to friend´s insurance company
stock	byte	%9.0g		would keep company secret from friend

```
Sorted by: accident  play  insurance  stock
```

```
. list in 120/121
```

	accident	play	insura~e	stock
120.	1	0	1	1
121.	1	1	0	0

For each observation, we have a vector of responses
$\mathbf{Y} = (Y_1, Y_2, Y_2, Y_4)$ (I am omitting an observation index)
The traditional approach deals with models that involve only categorical variables, so within each class we have 2^n cells with zeros and ones, and probabilities are estimates nonparametrically.

Stata (Model-based) approach

Now, how do we do it in Stata?

```
.  gsem (accident play insurance stock <- ), ///
> logit lclass(C 2)
```

We are fitting a logit model for each class, with no covariates. Because there are no covariates, estimating the constant is equivalent to estimating the probability: $p = F(constant)$, where F is the inverse logit function.

```
. gsem(accident play insurance stock <- ),logit lclass(C 2) ///
>     vsquish nodvheader noheader nolog
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
1.C	(base outcome)					
2.C						
_cons	-.9482041	.2886333	-3.29	0.001	-1.513915	-.3824933

Class : 1

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
accident						
_cons	.9128742	.1974695	4.62	0.000	.5258411	1.299907
play						
_cons	-.7099072	.2249096	-3.16	0.002	-1.150722	-.2690926
insurance						
_cons	-.6014307	.2123096	-2.83	0.005	-1.01755	-.1853115
stock						
_cons	-1.880142	.3337665	-5.63	0.000	-2.534312	-1.225972

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Class	: 2					
	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
accident						
_cons	4.983017	3.745987	1.33	0.183	-2.358982	12.32502
play						
_cons	2.747366	1.165853	2.36	0.018	.4623372	5.032395
insurance						
_cons	2.534582	.9644841	2.63	0.009	.6442279	4.424936
stock						
_cons	1.203416	.5361735	2.24	0.025	.1525356	2.254297

From the output, parameters for the second class are larger than those on the first class. Postestimation commands will help us to interpret this output.

To interpret the classes, we could compare the mean of the (counter-factual) conditional probabilities for each answer on each class; (the ones we get with **predict** by default) **estat lcmean** will do that.

```
. estat lcmean
```

Latent class marginal means — Number of obs = 216

	Delta-method Margin	Std. Err.	[95% Conf.	Interval]
1				
accident	.7135879	.0403588	.6285126	.7858194
play	.3296193	.0496984	.2403572	.4331299
insurance	.3540164	.0485528	.2655049	.4538042
stock	.1323726	.0383331	.0734875	.2268872
2				
accident	.9931933	.0253243	.0863544	.9999956
play	.9397644	.0659957	.6135685	.9935191
insurance	.9265309	.0656538	.6557086	.9881667
stock	.769132	.0952072	.5380601	.9050206

Also, we compute the predicted probabilities for each class.

Prior probabilities are the ones predicted by the logistic model for the latent class, which (with no covariates) will have no variations across the data.

```
. predict classpr*, classpr
. summ classpr*
```

Variable	Obs	Mean	Std. Dev.	Min	Max
classpr1	216	.7207538	0	.7207538	.7207538
classpr2	216	.2792462	0	.2792462	.2792462

This is an estimator of the population expected means for these variables. These estimates, and their confidence intervals can be obtained with **estat lcprob**.

```
. estat lcprob

Latent class marginal probabilities             Number of obs     =        216
```

	Delta-method Margin	Std. Err.	[95% Conf.	Interval]
C				
1	.7207539	.0580926	.5944743	.8196407
2	.2792461	.0580926	.1803593	.4055257

Stata provides some tools to evaluate goodness of fit:

```
. estat lcgof
```

Fit statistic	Value	Description
Likelihood ratio		
chi2_ms(6)	2.720	model vs. saturated
p > chi2	0.843	
Information criteria		
AIC	1026.935	Akaike´s information criterion
BIC	1057.313	Bayesian information criterion

Concluding remarks:

- **gsem** offers a framework where we can fit models accounting for latent classes.
- Responses might take one or more of the distributions supported by **gsem**.
- Discrete latent variables might have more than two groups, and more than one latent variable also might be included.
- Latent class models that have one dependent variable, can be seen as finite mixture models. The **fmm** prefix allows us to easily fit finite mixture models for a variety of distributions.

Appendix 1: Using **predict** after the role conflict dataset

Prior probability of class membership, $P(Ck)$ [1]	
$P(\mathbf{Y} \in C2)$	`predict newar, classpr class(2)`

Posterior probability of class membership, (Bayes formula)	
$P_{post}(\mathbf{Y} \in C_2)$	`predict newvar, class(2) classposteriorpr`

Probabilities of positive outcome, conditional on class (default) [2]	
$P(Y_1 = 1 \mid C2)$	`predict new, mu outcome(accident) class(2)`

Probabilities of positive outcome, marginal on (prior) class probability [3]	
$P(Y1 = 1)$	`predict newvar, mu outcome(1) marginal`

Probabilities of positive outcome, marginal on posterior class probability	
$P_{post}(Y1 = 1)$	`predict newvar, mu outcome(1) pmarginal`

[1] in this modle will be constant, because of no covariates in LC equation

[2] constant, because there are no covariates in `accident` equation

[3] constant, because there are no predictors at all

Appendix 2: Predictions after LCA, general case

(Prior) probability of class membership	
$\hat{\pi}_i = P(C = i, \mathbf{z}, \gamma, \Theta)$, for each i	`predict p*, classpr [class(i)]` Creates g variables by default
Posterior probability of class membership	
$\tilde{\pi}_i = P(C = i, \mathbf{Y}, \mathbf{z}, \gamma, \Theta)$, for each i	`predict postp*, classpostpr [class(i)]` Creates g variables by default
Expected value of $\mathbf{Y}$, conditional on class	
$\hat{\mu}_i = E(\mathbf{Y}\|C = i, \mathbf{z}, \Theta)$, for each i ;$(\mathbf{Y} = Y_1, \ldots Y_n)$	`predict m*, mu [outcome(j) class(i)]` Creates $n \times g$ variables by default
Expected value of $\mathbf{Y}$, marginal on (prior) class probability	
$\hat{\mu} = E(\mathbf{Y}\|\mathbf{z}, \gamma, \Theta)$, all j (based on $\hat{\pi}_i$)	`predict m*, mu marginal [outcome(j)]` Creates n variables by default
Expected value of $\mathbf{Y}$, marginal on posterior class probability	
$\tilde{\mu} = E_{post}(\mathbf{Y}\|\mathbf{z}, \gamma, \Theta)$, all j (based on $\tilde{\pi}_i$)	`predict m*, mu pmarginal [outcome(j)]` Creates n variables by default