example 50g — Latent class model

Description

To demonstrate latent class models, we use the following data:

```
. use http://www.stata-press.com/data/r15/gsem_lca1
(Latent class analysis)

. describe

Contains data from http://www.stata-press.com/data/r15/gsem_lca1.dta
  obs:           216                          Latent class analysis
 vars:             4                          17 Jan 2017 12:52
 size:           864                          (_dta has notes)
-------------------------------------------------------------------------------
              storage   display    value
variable name   type    format     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

. 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".
```

See *Latent class models* in [SEM] **intro 5** for background.

Remarks and examples

A latent class model is characterized by having a categorical (rather than continuous) latent variable. The levels of the categorical latent variable represent groups in the population and are called classes. We are interested in identifying and understanding these classes.

Goodman (2002) fits a variety of latent class models to the dataset described above with a focus on understanding how groups of individuals differ in response to situations that require making a decision between helping a friend (a particularistic choice) and doing what is right for society (a universalistic choice). Individuals were asked how they would respond in four such situations, and their responses were recorded in the variables `accident`, `play`, `insurance`, and `stock`. These variables are coded such that 1 is a universalistic response and 0 is a particularistic response.

To fit a latent class model, we must specify the number of classes in the latent variable. In the basic form of the latent class model demonstrated here, we have one categorical latent variable with two classes. The parameters in the model, namely, the intercepts in logistic regression models for the four observed variables, are allowed to vary across the classes.

More specifically, the model that we will fit estimates an intercept, α, for each observed variable for the first class,

$$\Pr(\texttt{accident} = 1 \,|\, C = 1) = \frac{\exp(\alpha_{11})}{1 + \exp(\alpha_{11})}$$

$$\Pr(\texttt{play} = 1 \,|\, C = 1) = \frac{\exp(\alpha_{21})}{1 + \exp(\alpha_{21})}$$

$$\Pr(\texttt{insurance} = 1 \,|\, C = 1) = \frac{\exp(\alpha_{31})}{1 + \exp(\alpha_{31})}$$

$$\Pr(\texttt{stock} = 1 \,|\, C = 1) = \frac{\exp(\alpha_{41})}{1 + \exp(\alpha_{41})}$$

and a corresponding intercept for the second class,

$$\Pr(\texttt{accident} = 1 \,|\, C = 2) = \frac{\exp(\alpha_{12})}{1 + \exp(\alpha_{12})}$$

$$\Pr(\texttt{play} = 1 \,|\, C = 2) = \frac{\exp(\alpha_{22})}{1 + \exp(\alpha_{22})}$$

$$\Pr(\texttt{insurance} = 1 \,|\, C = 2) = \frac{\exp(\alpha_{32})}{1 + \exp(\alpha_{32})}$$

$$\Pr(\texttt{stock} = 1 \,|\, C = 2) = \frac{\exp(\alpha_{42})}{1 + \exp(\alpha_{42})}$$

We also estimate the probability of being in each class using multinomial logistic regression,

$$\Pr(C = 1) = \frac{e^{\gamma_1}}{e^{\gamma_1} + e^{\gamma_2}}$$

$$\Pr(C = 2) = \frac{e^{\gamma_2}}{e^{\gamma_1} + e^{\gamma_2}}$$

where γ_1 and γ_2 are intercepts in the multinomial logit model. By default, the first class will be treated as the base, so $\gamma_1 = 0$.

To fit this model, we type

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

No variables are listed on the right side of the arrow because we are fitting intercept-only models for each observed variable. `logit` specifies that we are fitting logistic regression models for all four variables. The `lclass(C 2)` option specifies that the name of our categorical latent variable is `C` and that it has two latent classes.

The result of typing our estimation command is

```
. gsem (accident play insurance stock <- ), logit lclass(C 2)
Fitting class model:
Iteration 0:   (class) log likelihood = -149.71979
Iteration 1:   (class) log likelihood = -149.71979
Fitting outcome model:
Iteration 0:   (outcome) log likelihood = -403.97142
Iteration 1:   (outcome) log likelihood = -398.15909
Iteration 2:   (outcome) log likelihood = -397.81953
Iteration 3:   (outcome) log likelihood =  -397.8164
Iteration 4:   (outcome) log likelihood =  -397.8164
Refining starting values:
Iteration 0:   (EM) log likelihood = -570.24204
Iteration 1:   (EM) log likelihood = -576.20485
Iteration 2:   (EM) log likelihood = -577.41464
Iteration 3:   (EM) log likelihood = -576.88554
Iteration 4:   (EM) log likelihood = -575.59242
Iteration 5:   (EM) log likelihood = -573.90567
Iteration 6:   (EM) log likelihood = -571.99868
Iteration 7:   (EM) log likelihood = -569.97482
Iteration 8:   (EM) log likelihood = -567.90955
Iteration 9:   (EM) log likelihood = -565.86392
Iteration 10:  (EM) log likelihood = -563.88815
Iteration 11:  (EM) log likelihood = -562.02165
Iteration 12:  (EM) log likelihood = -560.29231
Iteration 13:  (EM) log likelihood = -558.71641
Iteration 14:  (EM) log likelihood = -557.29974
Iteration 15:  (EM) log likelihood = -556.03949
Iteration 16:  (EM) log likelihood = -554.92679
Iteration 17:  (EM) log likelihood = -553.94914
Iteration 18:  (EM) log likelihood = -553.09241
Iteration 19:  (EM) log likelihood = -552.34233
Iteration 20:  (EM) log likelihood = -551.68539
Note: EM algorithm reached maximum iterations.
Fitting full model:
Iteration 0:   log likelihood = -504.62913
Iteration 1:   log likelihood = -504.47255
Iteration 2:   log likelihood = -504.46773
Iteration 3:   log likelihood = -504.46767
Iteration 4:   log likelihood = -504.46767
Generalized structural equation model           Number of obs     =        216
Log likelihood = -504.46767
```

	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

Response       : accident
Family         : Bernoulli
Link           : logit

Response       : play
Family         : Bernoulli
Link           : logit

Response       : insurance
Family         : Bernoulli
Link           : logit

Response       : stock
Family         : Bernoulli
Link           : logit
```

	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

```
Class          : 2

Response       : accident
Family         : Bernoulli
Link           : logit

Response       : play
Family         : Bernoulli
Link           : logit

Response       : insurance
Family         : Bernoulli
Link           : logit

Response       : stock
Family         : Bernoulli
Link           : logit
```

	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

Notes:

1. The output shows four iteration logs. The first three are for models that are fit to obtain good starting values. Starting values are challenging for latent class models, and `gsem` provides a variety of options for specifying and computing starting values. See [SEM] **gsem estimation options** and [SEM] **intro 12** for more information on these options.

2. The first table in the output provides the estimated coefficients in the multinomial logit model for `C`.

3. The next two tables are the results for the logistic regression models for the first and second classes.

To better understand this model, let's examine how the probabilities of giving a universalistic response differ across classes. The `estat lcmean` command reports class-specific marginal means for each variable. Because we are using logistic regression, these means are actually the predicted probabilities.

```
. estat lcmean
Latent class marginal means                     Number of obs     =        216
```

	Margin	Delta-method Std. Err.	[95% Conf.	Interval]
1				
accident	.7135879	.0403588	.6285126	.7858194
play	.3296193	.0496984	.2403573	.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

The first section of this table reports the probabilities for class 1. In this class, the probability of giving a universalistic response to the first question—the question that concerns testifying against a friend who was involved in an accident—is 0.714. The probability of giving a universalistic response to the last question—the question about warning a friend about a stock price that is about to fall—is 0.132.

The second section of the table reports the corresponding probabilities for class 2. We find that the probability of giving a universalistic response to each question is higher in class 2 than in class 1. Class 2 appears to represent a more universalistically inclined group.

We can estimate probabilities of being in each class using `estat lcprob`.

```
. estat lcprob
Latent class marginal probabilities             Number of obs     =        216
```

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

This indicates that 72% of individuals are expected to be in the less universalistic class and 28% are expected to be in the more universalistic class.

We can use the predictions of the posterior probability of class membership to evaluate an individual's probability of being in each class.

```
. predict classpost*, classposteriorpr
. list in 1, abbrev(10)
```

	accident	play	insurance	stock	classpost1	classpost2
1.	0	0	0	0	.999975	.000025

For the first individual in our dataset, who responded with a particularistic answer to all four questions, the probability of being in class 1, the less universalistic class, is almost 1.

We can determine the expected class for each individual based on whether the posterior probability is greater than 0.5.

```
. generate expclass = 1 + (classpost2>0.5)
. tabulate expclass
```

expclass	Freq.	Percent	Cum.
1	145	67.13	67.13
2	71	32.87	100.00
Total	216	100.00	

In our dataset, 145 individuals are expected to be in class 1 and 71 individuals are expected to be in class 2.

References

Goodman, L. A. 2002. Latent class analysis: The empirical study of latent types, latent variables, and latent structures. In *Applied Latent Class Analysis*, ed. J. A. Hagenaars and A. L. McCutcheon, 3–55. Cambridge: Cambridge University Press.

Stouffer, S. A., and J. Toby. 1951. Role conflict and personality. *American Journal of Sociology* 56: 395–406.

Also see

[SEM] **example 51g** — Latent class goodness-of-fit statistics

[SEM] **example 52g** — Latent profile model

[SEM] **gsem** — Generalized structural equation model estimation command

[SEM] **intro 5** — Tour of models

[SEM] **estat lcmean** — Latent class marginal means

[SEM] **estat lcprob** — Latent class marginal probabilities