Slide 1

Generalized latent class modeling using `gllamm`

Sophia Rabe-Hesketh
Institute of Psychiatry, London

Andrew Pickles
The University of Manchester

Anders Skrondal
Norwegian Institute of Public Health, Oslo

Second US Stata Users' Group Meeting, March 2003

Slide 2

Outline

- Latent class models as two-level GLLAMMs with discrete latent variables
- Syntax for latent class models
  - `gllamm` for estimation
  - post-estimation commands:
    - `gllapred` for prediction
    - `gllasim` for simulation
- Example 1: Diagnosis of myocardial infarction
- Example 2: Attitudes to abortion

Slide 3

Response model for two-level GLLAMMs

- Conditional on the latent variables, the response model is a generalized linear model with linear predictor

$$\nu_{ij} \;=\; \mathbf{x}'_{ij}\boldsymbol{\beta} + \sum_{m=1}^{M} \eta_{jm}\mathbf{z}'_{mij}\boldsymbol{\lambda}_m, \quad \lambda_{m1} = 1$$

  - i indexes the units at level 1 (e.g. items, $i = 1, \cdots, I$).
  - j indexes the units at level 2 (e.g. subjects, $j = 1, \cdots, N$).
  - $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}_m$ are parameters.
  - $\mathbf{x}_{ij}$ and $\mathbf{z}_{mij}$ are vectors of observed variables and known constants.
  - η_{jm} is the mth element of the latent variable vector $\boldsymbol{\eta}_j$.
- The usual links and distributions can be specified, so that the following response types can be modeled:
  - continuous
  - dichotomous
  - ordinal
  - nominal (polytomous or rankings)
  - counts
  - durations (discrete and continuous)

Slide 4

Discrete latent variables

- Latent variable vector $\boldsymbol{\eta}_j$ for unit j with discrete values (or locations) $\mathbf{e}_c, c=1, \cdots, C$ in M dimensions.
- Individuals in the same latent class share the same value or location $\mathbf{e}_c$.
- Probability that subject j is in latent class c is π_{jc}.
- This probability may depend on covariates $\mathbf{v}_j$ through a multinomial logit model

$$\pi_{jc} \;=\; \frac{\exp(\mathbf{v}'_j\boldsymbol{\alpha}^c)}{\sum_d \exp(\mathbf{v}'_j\boldsymbol{\alpha}^d)},$$

  where $\boldsymbol{\alpha}^c$ are parameters with $\boldsymbol{\alpha}^C = 0$ for identification.
- Two parameterizations:
  1. **non-centered**: $\mathbf{e}_c$, C locations freely estimated
  2. **centered**: $\tilde{\mathbf{e}}_c$, $C-1$ locations estimated, last location determined by constraint

$$\sum_c \pi_{0c}\tilde{e}_c = 0,$$

     where π_{0c} is the probability when all covariates $\mathbf{v}_j$ are zero (except the constant). This parameterization allows mean structure to be modeled using $\mathbf{x}'_{ij}\boldsymbol{\beta}$.

Slide 5

Three different types of latent class models

- Linear predictor:

$$\nu_{ij} \;=\; \mathbf{x}_{ij}'\boldsymbol{\beta} + \sum_{m=1}^{M} \eta_{jm}\mathbf{z}_{mij}'\boldsymbol{\lambda}_m, \quad \lambda_{m1} = 1$$

1. **Exploratory latent class model** (I latent variables):

$$\begin{aligned}\nu_{ijc} &= \sum_{m=1}^{I} e_{mc}d_{mi} \\ &= e_{ic},\end{aligned}$$

where

$$d_{mi} = \begin{cases} 1 & \text{if } m = i \\ 0 & \text{otherwise} \end{cases}$$

2. **Discrete one-factor model** (one latent variable):

$$\begin{aligned}\nu_{ijc} &= \mathbf{d}_i'\boldsymbol{\beta} + \tilde{e}_c\mathbf{d}_i'\boldsymbol{\lambda} \\ &= \beta_i + \tilde{e}_c\lambda_i,\end{aligned}$$

where $\mathbf{d}_i' = (d_{1i}, d_{2i}, \cdots, d_{Ii})$

3. **Discrete random coefficient model**
(i typically not items but lower-level units, $i = 1, \cdots, n_j$)

$$\nu_{ijc} \;=\; \mathbf{x}_{ij}'\boldsymbol{\beta} + \sum_{m=1}^{M} e_{mc}z_{ij}$$

- e.g. Longitudinal data, occasions i at times t_{ij} for units j

$$\nu_{ijc} \;=\; e_{1c} + e_{2c}t_{ij}$$

$\Longrightarrow$ Linear latent trajectory model

Slide 6

Basic `gllamm` syntax for latent class models

`gllamm` [*varlist*] [`if` *exp*] [`in` *range*] , `i`(*varname*) [
`nrf`(#) `eqs`(*eqnames*) `noconstant` `ip`(*string*)
`nip`(#) `peqs`(*eqnames*) `constraints`(*numlist*)
`family`(*family*) `link`(*link*) `weight`(*string*) ⋯]

`i`(*varname*) variable identifying the (level 2) units j.

`nrf`(#) number of latent variables M.

`eqs`(*eqnames*) M equations, one for each $\mathbf{z}_m'\boldsymbol{\lambda}_m$, $m = 1, \cdots, M$. No constant is assumed unless explicitly included in the equation definition.

`ip`(*string*) `ip(f)` gives centered latent classes $\tilde{\mathbf{e}}_c$ and `ip(fn)` gives non-centered latent classes $\mathbf{e}_c$.

`nip`(#) number of latent classes.

`peqs`(*eqnames*) equations for $\mathbf{v}_j'\boldsymbol{\alpha}^c$ in multinomial logit model for latent class probabilities - a constant is automatically included.

`constraints`(*numlist*) list of linear parameter constraints defined using the `constraint define` command.

`family`(*family*), `link`(*link*), `noconstant` as `glm`, plus `link(ologit)`, `link(mlogit)`, etc.

`weight`(*string*) frequency weights at levels 1 and 2 are in *string*1 and *string*2.

Slide 7

Basic gllapred syntax for latent class models

`gllapred` *varname* [`if` *exp*] [`in` *range*] [, `p` `mu`
`marginal` `us`(*varname*) `outcome`(#) `above`(#) `ll`
`from`(*matrix*) ···]

`p` posterior probabilities returned in *varname*1, *varname*2, etc.

`mu` mean response returned in *varname*. Without further options, mean w.r.t. posterior distribution.

`marginal` together with `mu`, causes marginal or population average mean to be returned (mean w.r.t. prior distribution).

`us`(*varname*) together with `mu`, causes conditional mean to be returned, conditional on latent variables being equal to the values in *varname*1, *varname*2, etc.

`outcome`(#) with `mlogit` link, causes `mu` option to return probability that the response equals #.

`above`(#) with ordinal links, causes `mu` option to return probability that the response exceeds #.

`ll` log-likelihood contributions of top-level clusters returned in *varname*. This can be used to compute expected counts.

`from`(*matrix*) causes predictions to be made for the model just estimated in `gllamm` but with parameter values in *matrix*.

Slide 8

Basic gllasim syntax for latent class models

`gllasim` *varname* [`if` *exp*] [`in` *range*] [, `u`
`us`(*varname*) `from`(*matrix*) ···]

By default, responses are simulated for the model just estimated and returned in *varname*.

`u` latent variables are simulated and returned in *varname*p1, *varname*p2, etc.

`us`(*varname*) response variables are simulated for latent variables equal to *varname*1, *varname*2, etc.

`from`(*matrix*) causes responses/latent variables to be simulated from the model just estimated in `gllamm` but with parameter values in *matrix*.

Slide 9

Example 1: Diagnosis of myocardial infarction

- 94 patients admitted for the purpose of ruling out myocardial infarction (MI) or 'heart attack'.
- Four diagnostic criteria:
  - [Q-wave] presence of a Q-wave in the ECG
  - [History] presence of a classical clinical history
  - [LDH] presence of flipped LDH
  - [CPK] presence of a CPK-MB
- Data:

patt	var	y	v1	v2	v3	v4	wt2
1	1	1	1	0	0	0	24
1	2	1	0	1	0	0	24
1	3	1	0	0	1	0	24
1	4	1	0	0	0	1	24
2	1	0	1	0	0	0	5
2	2	1	0	1	0	0	5
2	3	1	0	0	1	0	5

  - `patt` identifies the unique response patterns
  - `y` is the response
  - `var` is the diagnostic criterion, dummies `v1` to `v4`
  - `wt2` is the number of subjects with the response pattern

Slide 10

Estimation and prediction

- Exploratory latent class model (two classes):

$$\text{logit}[\Pr(y_{ij}=1|c)] = e_{ic}, \quad c = 1, 2$$

```
eq v1: v1
eq v2: v2
eq v3: v3
eq v4: v4
gllamm y, i(patt) ip(fn) nrf(4) eqs(v1 v2 v3 v4) /*
  */ weight(wt) nip(2) l(logit) f(binom) nocons
```

- Part of output:

```
loc1: -17.585, 1.1903
loc2: -1.4173, 1.3333
loc3: -3.5875, 1.5708
loc4: -1.4143, 16.857
prob: 0.5422, 0.4578
```

- Conditional response probabilities:

Sensitivity : $\Pr(y_{ij}=1|c=2)$ Specificity : $\Pr(y_{ij}=0|c=1)$

```
gen e1 = 1.1903  /* could use gllasim e, u */
gen e2 = 1.3333
gen e3 = 1.5708
gen e4 = 16.857
gllapred cprob, mu us(e) /* sensitivity */
```

Slide 11

Estimates

	Class 1 ('No MI')			Class 2 ('MI')		
			Prob.			Prob.
Parameter	Est	(SE)	(%)	Est	(SE)	(%)
			1-Spec.			Sens.
e_{1c} [Q-wave]	-17.58	*(953.49)	0	1.19	(0.42)	77
e_{2c} [History]	-1.42	(0.39)	30	1.33	(0.39)	79
e_{3c} [LDH]	-3.59	(1.01)	3	1.57	(0.47)	83
e_{4c} [CPK]	-1.41	(0.41)	20	16.86	*(706.04)	100
			1-Prev.			Prev.
α_0 [Cons]	0.17	0.22	54	–	–	46

* boundary solution

Slide 12

More prediction

- Posterior probabilities ("Positive & Negative predictive values"):

$$\Pr(c{=}1|\mathbf{y}_j) = \frac{\pi_1 \prod_{i=1}^4 \Pr(y_{ij}|c{=}1)}{\sum_c \pi_c \prod_{i=1}^4 \Pr(y_{ij}|c)}$$

$$\Pr(c{=}2|\mathbf{y}_j) = \frac{\pi_2 \prod_{i=1}^4 \Pr(y_{ij}|c{=}2)}{\sum_c \pi_c \prod_{i=1}^4 \Pr(y_{ij}|c)}$$

```
gllapred prob, p
(probabilities will be stored in prob1 prob2)
```

- Predicted counts:

$$\begin{aligned} & 94 \Pr(\mathbf{y}_j) \\ = \; & 94 \exp(\ell_j) \\ = \; & 94 \sum_c \pi_c \prod_{i=1}^4 \Pr(y_{ij}|c), \end{aligned}$$

where ℓ_j is the log-likelihood contribution of cluster j.

```
gllapred l, ll  /* log-likelihood contributions */
gen count = 94*exp(l)
```

  - Could calculate diagnostics and deviance as in loglinear modeling of contingency tabels

Slide 13

Posterior probabilities and expected counts

[Q-wave] ($i=1$)	[History] ($i=2$)	[LDH] ($i=3$)	[CPK] ($i=4$)	Obs. count	Exp. count	Prob. of MI ($c=2$)
1	1	1	1	24	21.62	**1.000**
0	1	1	1	5	6.63	**0.992**
1	0	1	1	4	5.70	**1.000**
0	0	1	1	3	1.95	**0.889**
1	1	0	1	3	4.50	**1.000**
0	1	0	1	5	3.26	0.420
1	0	0	1	2	1.19	**1.000**
0	0	0	1	7	8.16	0.044
1	1	1	0	0	0.00	0.017
0	1	1	0	0	0.22	0.000
1	0	1	0	0	0.00	0.001
0	0	1	0	1	0.89	0.000
1	1	0	0	0	0.00	0.000
0	1	0	0	7	7.78	0.000
1	0	0	0	0	0.00	0.000
0	0	0	0	33	32.11	0.000

Slide 14

Example 2 : Attitudes to abortion

- British Social Attitudes Survey 1983
- Respondents were asked whether or not abortion should be allowed by law if:

`[wom]` The woman decides on her own she does not wish to have the child

`[cou]` The couple agree that they do not wish to have the child

`[mar]` The woman is not married and does not wish to marry the man

`[fin]` The couple cannot afford any more children

`[gen]` There is a strong chance of a genetic defect in the baby

`[ris]` The woman's health is seriously endangered by the pregnancy

`[rap]` The woman became pregnant as a result of rape

- 720 respondents, 11% have some missing responses, in total responses to 7% of items are missing

Slide 15

Data structure

id	ab	wom	cou	mar	fin	gen	ris	rap	fem	pwt2	area83
1	1	1	0	0	0	0	0	0	0	.8281	102
1	1	0	1	0	0	0	0	0	0	.8281	102
1	1	0	0	1	0	0	0	0	0	.8281	102
1	1	0	0	0	1	0	0	0	0	.8281	102
1	1	0	0	0	0	1	0	0	0	.8281	102
1	1	0	0	0	0	0	1	0	0	.8281	102
1	1	0	0	0	0	0	0	1	0	.8281	102
2	0	1	0	0	0	0	0	0	0	.621075	102

- variables:
  - `id` identifies subjects
  - `ab` is the response
  - `wom` to `rap` are dummies for the items
  - `fem` is dummy for females
  - `pwt2` are inverse probability weights at level 2
  - `area83` identifies primary sampling units

Slide 16

Estimation

- Model 1: Discrete one-factor model

$$\text{logit}[\Pr(y_{ij}=1|c)] = \beta_i + \tilde{e}_c\lambda_i,$$

```
eq fac: wom cou mar fin gen ris rap
gllamm ab wom cou mar fin gen ris rap, nocons i(id)/*
  */ nrf(1) l(logit) f(binom) eqs(fac) ip(f) nip(2)
```

- Model 2: Class probabilities depend on sex (v_j=`[fem]`)

$$\pi_{j1} = \frac{\exp(\alpha_0^1 + \alpha_1^1 v_j)}{1+\exp(\alpha_0^1 + \alpha_1^1 v_j)},\ \pi_{j2} = 1 - \pi_{j1}.$$

```
eq fem: fem
gllamm ab wom cou mar fin gen ris rap, nocons i(id)/*
  */ nrf(1) l(logit) f(binom) eqs(fac) peqs(fem)   /*
  */ ip(f) nip(2)
```

- Model 2a: Include a direct effect of gender on the second item `[cou]`.

$$\text{logit}[\Pr(y_{2j}=1|c,v_j)] = \beta_{02} + \beta_{12}v_j + \lambda_i\tilde{e}_c.$$

```
gen femcou = fem*cou
gllamm ab wom cou femcou mar fin gen ris rap, ...
```

Slide 17

Estimates

Two classes	Model 1	Model 2
Intercepts:		
β_1 [wom]	-0.49 (0.12)	-0.46 (0.12)
β_2 [cou]	0.39 (0.24)	0.60 (0.28)
β_3 [mar]	-0.19 (0.15)	0.06 (0.17)
β_4 [fin]	0.22 (0.14)	0.43 (0.16)
β_5 [gen]	2.69 (0.26)	2.86 (0.29)
β_6 [ris]	3.48 (0.47)	3.66 (0.52)
β_7 [rap]	2.85 (0.22)	2.95 (0.24)
Factor loadings:		
λ_1 [wom]	1 (–)	1 (–)
λ_2 [cou]	1.62 (0.24)	1.64 (0.24)
λ_3 [mar]	1.33 (0.16)	1.32 (0.16)
λ_4 [fin]	1.16 (0.15)	1.15 (0.15)
λ_5 [gen]	0.94 (0.22)	0.93 (0.21)
λ_6 [ris]	1.05 (0.39)	1.04 (0.38)
λ_7 [rap]	0.61 (0.19)	0.60 (0.18)
Locations parameter:		
$\widetilde{e}_1$	-1.28 (0.14)	-1.47 (0.16)
Probability parameters (class 1):		
α_0^1 [cons]	0.24 (0.12)	-0.01 (0.17)
α_1^1 [fem]	–	0.48 (0.17)
Log-likelihood:	-1967.89	-1963.82

Slide 18

Three classes	Model 3	Model 4
Intercepts:		
β_1 [wom]	-0.73 (0.21)	-0.69 (0.32)
β_2 [cou]	0.15 (0.40)	0.22 (0.61)
β_3 [mar]	-0.49 (0.28)	-0.43 (0.45)
β_4 [fin]	-0.04 (0.25)	0.02 (0.39)
β_5 [gen]	2.68 (0.25)	2.73 (0.29)
β_6 [ris]	3.52 (0.31)	3.51 (0.34)
β_7 [rap]	2.90 (0.20)	2.91 (0.22)
Factor loadings:		
λ_1 [wom]	1 (–)	1 (–)
λ_2 [cou]	1.89 (0.33)	1.88 (0.32)
λ_3 [mar]	1.41 (0.17)	1.40 (0.17)
λ_4 [fin]	1.23 (0.16)	1.23 (0.16)
λ_5 [gen]	0.63 (0.24)	0.60 (0.27)
λ_6 [ris]	0.55 (0.24)	0.47 (0.25)
λ_7 [rap]	0.35 (0.15)	0.31 (0.16)
Locations parameters:		
$\widetilde{e}_1$	-8.03 (3.01)	-8.90 (3.99)
$\widetilde{e}_2$	-0.81 (0.19)	-0.86 (0.30)
Probability parameters:		
α_0^1 [cons]	-2.15 (0.26)	-2.08 (0.32)
α_1^1 [fem]	–	-0.07 (0.41)
α_0^2 [cons]	0.30 (0.10)	-0.01 (0.15)
α_1^2 [fem]	–	0.54 (0.17)
Log-likelihood:	-1921.29	-1916.16

Slide 19

Prediction

- `e1` contains the locations for Model 4

```
gllapred mup, mu us(e)
gllapred mu, mu marg
```

For three classes, Model 4:

	class 1	class 2	class 3		
Prior Probabilities (%)					
male	6	47	47		
female	4	60	36		
Conditional Probabilities (%)				Marginal Prob. (%)	
				male	female
[wom]	0	18	78	45	38
[cou]	0	20	98	56	47
[mar]	0	16	91	51	42
[fin]	0	26	92	56	48
[gen]	7	90	98	89	89
[ris]	33	96	99	94	94
[rap]	53	93	97	93	93

Slide 20

Models for complex survey data

- British Attitudes Survey not a simple random sample
- Pseudolikelihood estimation with inverse probability weights
- Robust standard errors (sandwich estimator) for cluster sampling with electoral ward as primary sampling uunit.
- `gllamm` options `pweight()`, `robust()` and `cluster()`:

```
gllamm ab wom cou mar fin gen ris rap, nocons     /*
  */ i(id) l(logit) f(binom) eqs(fac) ip(f) nip(2) /*
  */ peqs(fem) pweight(pwt) robust cluster(area83)
```

Slide 21

Estimates

Model 2	No pweights Model-based SE	pweights Robust SE	pweights Robust SE, cluster
Intercepts:			
β_1 [wom]	-0.46 (0.12)	-0.26 (0.15)	(0.16)
β_2 [cou]	0.60 (0.28)	0.82 (0.39)	(0.40)
β_3 [mar]	0.06 (0.17)	0.04 (0.21)	(0.24)
β_4 [fin]	0.43 (0.16)	0.35 (0.18)	(0.21)
β_5 [gen]	2.86 (0.29)	2.81 (0.31)	(0.30)
β_6 [ris]	3.66 (0.52)	3.72 (0.58)	(0.61)
β_7 [rap]	2.95 (0.24)	2.87 (0.31)	(0.32)
Factor loadings:			
λ_1 [wom]	1 (-)	1 (-)	
λ_2 [cou]	1.64 (0.24)	1.67 (0.29)	(0.30)
λ_3 [mar]	1.32 (0.16)	1.31 (0.18)	(0.21)
λ_4 [fin]	1.15 (0.15)	1.12 (0.18)	(0.19)
λ_5 [gen]	0.93 (0.21)	0.87 (0.24)	(0.27)
λ_6 [ris]	1.04 (0.38)	1.12 (0.46)	(0.45)
λ_7 [rap]	0.60 (0.18)	0.57 (0.26)	(0.25)
Location parameter:			
$\widetilde{e}_1$	-1.47 (0.16)	-1.40 (0.21)	(0.20)
Probability parameters (class 1):			
α_0^1 [cons]	-0.01 (0.17)	0.07 (0.21)	(0.19)
α_1^1 [fem]	0.48 (0.17)	0.43 (0.18)	(0.19)

Slide 22

Concluding remarks

- New classes can be introduced using the `gateaux()` option.
- Potential problems:
  - Local maxima can be a problem $\Longrightarrow$ try different sets of starting values.
  - Boundary solutions can be a problem.
- More information on `gllamm` and a manual can be found at www.iop.kcl.ac.uk/IoP/Departments/BioComp/programs/gllamm.html
  - A latent class model for rankings is described in Section 9.4 of the manual.
  - Slides of a talk at the RSS 'Half day meeting on latent class analysis and finite mixture models' are available under 'courses and presentations'.