Slide 1

Extensions to `gllamm`

7th UK Stata Users' Meeting

Sophia Rabe-Hesketh
Department of Biostatistics and Computing
Institute of Psychiatry, London

Andrew Pickles
School of Epidemiology and Health Sciences and CCSR
The University of Manchester

Anders Skrondal
Department of Epidemiology
National Institute of Public Health, Oslo

Slide 2

Extensions to `gllamm`:

- More response processes
  - Ordinal responses
  - **I. Nominal responses and rankings**
- Structural equations for the latent variables
  - **II. Regressions of latent variables on observed variables**
  - Regressions of latent variables on other latent variables
- Parameter constraints
- `gllapred` for posterior means and probabilities
- A manual

Slide 3

Generalised Linear Latent and Mixed Models (GLLAMMs)

- Conditional expectation of response

$$g(E[y|\mathbf{x},\mathbf{u}]) = \eta$$

where g is a link function and η is the linear predictor.

- Linear predictor:

$$\eta = \boldsymbol{\beta}'\mathbf{x} + \sum_{l=2}^{L}\sum_{m=1}^{M_l} u_m^{(l)}\boldsymbol{\lambda}_m^{(l)\prime}\mathbf{z}_m^{(l)} \quad \text{for identification, } \lambda_{\mathrm{m1}}^{(\mathrm{l})} = 1$$

- Conditional distribution of response is from exponential family
- Latent variables can be factors or random coefficients:
  - Random coefficient: one explanatory variable multiplies the latent variable
  - Factor: The items are treated as level 1 units and a linear combination of dummy variables for the items multiplies the latent variable

Slide 4

Response Processes

- The response variables may be of mixed type - requiring mixed links and families:

Links
identity
reciprocal
logarithm
logit
probit
scaled probit
compl. log-log

Families
Gaussian
gamma
Poisson
binomial

Polytomous responses
ordinal logit
ordinal probit
ordinal compl. log-log
multinomial logit

- Heteroscedasticity: The dispersion parameter for the Gauss and gamma families can differ between responses or depend on covariates
- Offsets
- Many response processes: multivariate survival, discrete survival data, rankings, ceiling/floor effects

Slide 5

I. Nominal responses and rankings

- Nominal or unordered categorical responses:
  - Party voted for
  - Treatment selected for a patient
  - Brand of ketchup bought

  One of A alternatives is 'selected': *first choice* data.
- Multinomial logit model (polytomous logistic regression):
  - linear predictor for alternative a is V^a, e.g., $V^a = \beta_0^a + \beta_1^a \text{Age}$
  - The probability that f is the 'chosen' alternative is

$$\Pr(f) = \frac{\exp(V^f)}{\sum_{a=1}^{A} \exp(V^a)}$$

Slide 6

Latent Response Derivation of Multinomial Logit Model

- Associated with each alternative is an unobserved 'utility' U^a (latent response). The alternative with the highest utility is selected. Depending on the situation, utility means attractivess or usefulness (voting/purchasing), cost-efficitveness (treatments), etc. of the alternative.

$$U^a = V^a + \epsilon^a$$

- f is chosen if

$$U^f > U^g \text{ for all } g \neq f$$

  or

$$U^f - U^g = V^f - V^g + (\epsilon^f - \epsilon^g) > 0$$

- If the error term ϵ^a has an extreme value distribution of type I (Gumbel), then the differences $(\epsilon^f - \epsilon^g)$ have a logistic distribution and it follows that (McFadden, 1974)

$$\Pr(f) = \frac{\exp(V^f)}{\sum_{a=1}^{A} \exp(V^a)}$$

Slide 7

British Election Study

- voters who voted Conservative, Labour, Liberal in 1987 and 1992 elections.
- Variables:
  - `male`, `age`, `manual` (father a manual worker)
  - `rldist`: distance between voter and party on left-right dimension constructed from respondent's and paty's position on 4 scales, e.g. more effort to redistribute wealth→less effort
  - `price`: judgement how much prices have risen
  - Expanded or "exploded" data

	serialno	year	party	chosen	rldist
10.	11	92	con	1	.5031
11.	11	92	lab	0	30.12
12.	11	92	lib	0	16.18
13.	13	87	con	0	22.86
14.	13	87	lab	1	.3622
15.	13	87	lib	0	1.894
16.	13	92	con	0	20.62
17.	13	92	lab	1	.0567
18.	13	92	lib	0	1.507
19.	15	87	con	0	25.32

  - `zrldist`, `zprice` are standardised versions

Slide 8

Multinomial logit in gllamm

- Data in expanded form: alternative sets (analogous to risk sets)
- Dummy variables `lab` and `lib` for Labour and Liberal and interactions with all subject-specific explanatory variables

```
gen lab_age = lab*age
gen lib_age = lib*age
```

- Multinomial logit model with a random effect of `zrldist`:

$$V_{ij}^{a} = \boldsymbol{\beta}^{a\prime}\mathbf{x}_{ij} + (\alpha + u_i)d_{ij}^{a}$$

where i indexes the voter, j indexes the election and d_{ij}^{a} is the distance between voter and party on the left-right political dimension.

```
eq beta1: zrldist
gllamm party zrldist lab87 lib87 lab92 lib92 lab_mal lib_mal   /*
  */ lab_age lib_age lab_man lib_man lab_zpri lib_zpri, nocons /*
  */ i(serialno) expand(occ chosen o) f(binom) l(mlogit) eq(beta1)
```

Slide 9

party	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
zrldist	-1.060069	.0494498	-21.44	0.000	-1.156989	-.9631496
lab87	.5425648	.2526952	2.15	0.032	.0472912	1.037838
lib87	.157179	.2357233	0.67	0.505	-.3048302	.6191881
lab92	.6552893	.2613689	2.51	0.012	.1430156	1.167563
lib92	-.1244899	.2459711	-0.51	0.613	-.6065843	.3576046
lab_mal	-.8032313	.137831	-5.83	0.000	-1.073375	-.5330876
lib_mal	-.6890129	.1298547	-5.31	0.000	-.9435235	-.4345024
lab_age	-.3636854	.0465448	-7.81	0.000	-.4549116	-.2724592
lib_age	-.2127401	.0430409	-4.94	0.000	-.2970987	-.1283815
lab_man	.8029466	.1472146	5.45	0.000	.5144113	1.091482
lib_man	-.0675791	.1309883	-0.52	0.606	-.3243114	.1891532
lab_zpri	.573825	.0766719	7.48	0.000	.4235507	.7240992
lib_zpri	.4458234	.0706312	6.31	0.000	.3073889	.584258

```
Variances and covariances of random effects
------------------------------------------------------------------------------
***level 2 (serialno)

    var(1): .27935136 (.07560918)
------------------------------------------------------------------------------
```

Slide 10

Multinomial Logit Model For Rankings

- Rankings are orderings of alternatives (parties, treatments, brands) according to preference or some other characteristic.
- Associated with each alternative a is an unobserved utility U^a

$$U^a = V^a + \epsilon^a$$

  where ϵ^a has an extreme value distribution (Gumbel)
- Let r^s be the alternative with rank s. Then the ranking $R = (r^1, r^2, \cdots r^A)$ is obtained if

$$U^{r^1} > U^{r^2} > \cdots > U^{r^A}$$

- The probability of a ranking R is (Luce, 1959)

$$\Pr(R) = \frac{\exp(V^{r^1})}{\sum_{s=1}^{A} \exp(V^{r^s})} \times \frac{\exp(V^{r^2})}{\sum_{s=2}^{A} \exp(V^{r^s})} \times \cdots \times \frac{\exp(V^{r^A})}{\sum_{s=A-1}^{A} \exp(V^{r^s})}$$

- At each ‘stage’, a first choice is made among the remaining alternatives
- A subject’s contribution to the likelihood is identical to the contribution of a stratum to the partial likelihood in Cox’s regression

Slide 11

Rankings for British Election Study

- First choice: party voted for
- Rankings: the parties were rated on a five point scale

  strongly against → strongly in favour
- The parties not voted for are ranked into second and third place using the rating scales. (In 6.5% of votes, the party voted for did not have the highest score)
- original data

	serialno	year	occ	rank	party
7.	11	87	3	1	con
8.	11	87	3	2	lib
9.	11	87	3	2	lab
10.	11	92	4	1	con
11.	11	92	4	2	lib
12.	11	92	4	3	lab

Slide 12

Data preparation for rankings

- "Exploding the data to alternative sets" using `stsplit`

```
egen maxr = max(rank), by(occ)
gen chosen=1
gen id=_n
stset rank, fail(chosen) id(id)
stsplit , at(failures) strata(occ) riskset(occstage)
replace chosen=0 if chosen==.
drop if rank==maxr
```

	serialno	year	occstage	party	chosen
11.	11	87	7	con	1
12.	11	87	7	lab	0
13.	11	87	7	lib	0
14.	11	92	9	con	1
15.	11	92	9	lab	0
16.	11	92	9	lib	0
17.	11	92	10	lib	1
18.	11	92	10	lab	0

Slide 13

Analysing rankings

- There are a number of possible random structures for election within voter within consituency.
- Example: correlated random coefficients for lab and lib at voter level

$$V_{ij}^a = \boldsymbol{\beta}^{a\prime}\mathbf{x}_{ij} + \alpha d_{ij}^a + u_{1i} z_{1ij}^a + u2ix_{2ij}^a$$

where z_{1ij} and x_{2ij} are dummy variables for Labour and Liberal, respectively.

– random coefficients of Labour and Liberal induce longitudial correlations across elections for Labour and Liberal, respectively.

– correlation between random coefficients of Labour and Liberal induces both cross-sectional and longitudinal correlations between the utilities for Labour and Liberal.

```
eq lab: lab
eq lib: lib
gllamm party zrldist lab87 lib87 lab92 lib92 lab_mal lib_mal   /*
  */ lab_age lib_age lab_man lib_man lab_zpri lib_zpri, nocons /*
  */ i(serialno) expand(occstage chosen o) f(binom) l(mlogit) /*
  */ nrf(2) nip(10) eqs(lib lab)
```

Slide 14

```
log likelihood = -2647.2917

>>> fixed effects omitted

Variances and covariances of random effects
-----------------------------------------------------------


***level 2 (serialno)

    var(1): 18.918555 (2.2960479)
    cov(1,2): 9.0408098 (1.1798509) cor(1,2): .82567035

    var(2): 6.3374322 (.88875039)
-----------------------------------------------------------
```

Assuming uncorrelated random coefficients (using `nocor` option, gives a log-likelihood of -2800.2076

Slide 15

II. Regressions of latent variables on observed variables

Structural equations for the latent variables

Regress the latent variables on other latent and explanatory variables

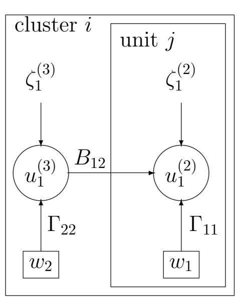
$$\mathbf{u} = \mathbf{B}\mathbf{u} + \mathbf{\Gamma}\mathbf{w} + \boldsymbol{\zeta}$$

- $\mathbf{u} = (u_1^{(2)}, u_2^{(2)}, \cdots, u_{M_2}^{(2)}, \cdots u_1^{(l)}, \cdots, u_{M_l}^{(l)}, \cdots, u_{M_L i}^{(L)})'$ (M elements)
  - factors
  - random coefficients
- $\mathbf{B}$ is an upper diagonal $M \times M$ matrix of regression coefficients
- $\mathbf{\Gamma}$ is an $M \times p$ matrix of regression coefficients
- $\mathbf{w}$ are p explanatory variables
- $\boldsymbol{\zeta}$ is an M dimensional vector of errors/disturbances (same level as corresponding elements in $\mathbf{u}$).

Slide 16

Theoretical example

$$\begin{bmatrix} u_1^{(2)} \\ u_1^{(3)} \end{bmatrix} = \begin{bmatrix} 0 & B_{12} \\ 0 & 0 \end{bmatrix} \begin{bmatrix} u_1^{(2)} \\ u_1^{(3)} \end{bmatrix} + \begin{bmatrix} \Gamma_{11} & 0 \\ 0 & \Gamma_{22} \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} + \begin{bmatrix} \zeta_1^{(2)} \\ \zeta_1^{(3)} \end{bmatrix}$$



- School level factor regressed on (and measured by) school level variables
- (a) School level factor affects pupil level factor (e.g. ability)
- (b) School level factor affects pupil level random coefficient (e.g. rate of increase in performance)

Slide 17

Example: Logistic regression with covariate measurement error

- Data and notation
  - Effect of fibre intake (continuous, measured twice on a subset of subjects) on coronary heart disease (CHD present/absent) (Morris, Marr and Clayton, 1977)
  - Responses are dietary fibre intake ($j = 1, 2$) and coronary heart disease ($j = 3$)
  - u_i is ith subject's true dietary intake (- population mean)
- Measurement model for fibre intake: y_{i1}, y_{i2} conditionally independently normally distributed with

$$E[y_{ij}|\mathbf{u}] = \beta_j + u_i\lambda_j, \quad j = 1, 2 \ (\beta_1 = \beta_2, \ \lambda_1 = \lambda_2 = 1)$$

- Disease model: y_{i3} conditionally binomial with

$$\text{logit}(E[y_{i3}|\mathbf{u}]) = \beta_3 + u_i\lambda_3 \quad \lambda_3 \text{ is } \log(\text{OR})$$

- GLLAMM
  - z_{1ij} is 1 for the element(s) corresponding to fibre and 0 otherwise.
  - z_{3ij} is 1 for the element corresponding to CHD, 0 otherwise.

$$\eta_{ij} = \boldsymbol{\beta}'\mathbf{z}_{ij} + u_i\boldsymbol{\lambda}'\mathbf{z}_{ij} \qquad \mathbf{z}_{ij} = (z_{1ij}, z_{3ij})'$$

Slide 18

Diet example in `gllamm`, see STB53, sg129

	id	resp	diet	chd	var
425.	217	3.06	1	0	1
426.	217	0	0	1	2
427.	218	3.14	1	0	1
428.	218	0	0	1	2
429.	219	2.75	1	0	1
430.	219	2.7	1	0	1
431.	219	0	0	1	2

`diet` is z_1 and `chd` is z_3

```
eq id: diet chd
gllamm resp diet chd, nocons i(id) eqs(id) link(ident logit) /*
*/ fam(gauss binom) lv(var) fv(var) nip(30)
```

Including other covariates

- **Direct effect of x on y_3**

$$\begin{aligned}\text{Measurement model}: E[y_{ij}|\mathbf{u}] &= \beta_j + u_i, \quad j = 1,2 \\ \text{Disease model}: \text{logit}(E[y_{i3}|\mathbf{u}]) &= \beta_3 + \beta_4 x + u_i\lambda_3\end{aligned}$$

- **Indirect effect of x on y_3**

$$u_i = \gamma x + \zeta_i,$$

where ζ_i is a residual error term

$$\begin{aligned}\text{Measurement model}: E[y_{ij}|\mathbf{u}] &= \beta_j + \gamma x + \zeta_i \\ \text{Disease model}: \text{logit}(E[y_{i3}|\mathbf{u}]) &= \beta_3 + \gamma\lambda_3 x + \zeta_i\lambda_3\end{aligned}$$

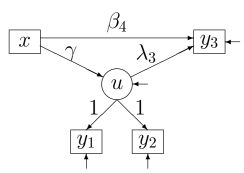
Slide 19

$\Rightarrow$ would requires nonlinear constraint for the coefficients if we couldn't regress $\mathbf{u}$ on explanatory variables

- **Direct and indirect effect of x on y_3**

$$\begin{aligned}\text{Measurement model}: E[y_{ij}|\mathbf{u}] &= \beta_j + \gamma x + \zeta_i \\ \text{Disease model}: \text{logit}(E[y_{i3}|\mathbf{u}]) &= \beta_3 + (\beta_4 + \gamma\lambda_3)x + \zeta_i\lambda_3\end{aligned}$$

(would not require a constraint for the coefficients)



Including effect of occupation (bus staff vs bank staff)

	id	resp	diet	chd	var
425.	217	3.06	1	0	1
426.	217	0	0	1	2
427.	218	3.14	1	0	1
428.	218	0	0	1	2
429.	219	2.75	1	0	1
430.	219	2.7	1	0	1
431.	219	0	0	1	2

Slide 20

- `gllamm` syntax without `occ`

```
eq id: diet chd
gllamm resp diet chd, nocons i(id) eqs(id) link(ident logit) /*
*/ fam(gauss binom) lv(var) fv(var) nip(30)
```

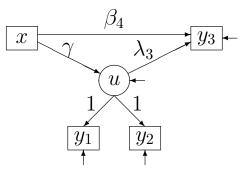
- `gllamm` syntax with direct and indirect effects of `occ` (dummy for bus staff)

```
eq f1: occ
gen occc=occ*chd
gllamm resp diet chd occc, nocons i(id) eqs(id) link(ident logit) /*
*/ fam(gauss binom) lv(var) fv(var) nip(30) geqs(f1)
```

Slide 21

Results

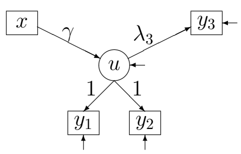
- direct and indirect effect of x on y_3



Log-likelihood=-186.90

Parameters	Estimates	SE
λ_3	-1.95	0.73
γ	-0.12	0.03
β_4	-0.19	0.34
σ^2	0.02	0.003
var(u)	0.07	0.007

- Indirect effect of x on y_3



Log-likelihood=-187.05

Parameters	Estimates	SE
λ_3	-1.86	0.70
γ	-0.12	0.03
σ^2	0.02	0.003
var(u)	0.07	0.007