Extensions to \texttt{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

\textbf{Extensions to \texttt{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
- \texttt{gllapred} for posterior means and probabilities
- A manual
Generalised Linear Latent and Mixed Models (GLLAMMs)

- Conditional expectation of response
  
  \[ g(E[y|x, u]) = \eta \]

  where \( g \) is a link function and \( \eta \) is the linear predictor.

- Linear predictor:
  
  \[ \eta = \beta'x + \sum_{l=2}^L \sum_{m=1}^M u_l^{(l)} \lambda^{(l)}_m z_l^{(l)} \]

  for identification, \( \lambda^{(l)}_{01} = 1 \)

- 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

Response Processes

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

<table>
<thead>
<tr>
<th>Links</th>
<th>Families</th>
<th>Polytamous responses</th>
</tr>
</thead>
<tbody>
<tr>
<td>identity</td>
<td>Gaussian</td>
<td>ordinal logit</td>
</tr>
<tr>
<td>reciprocal</td>
<td>gamma</td>
<td>ordinal probit</td>
</tr>
<tr>
<td>logarithm</td>
<td>Poisson</td>
<td>ordinal compl. log-log</td>
</tr>
<tr>
<td>logit</td>
<td>binomial</td>
<td>multinomial logit</td>
</tr>
<tr>
<td>probit</td>
<td></td>
<td></td>
</tr>
<tr>
<td>scaled probit</td>
<td></td>
<td></td>
</tr>
<tr>
<td>compl. log-log</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- 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
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

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

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 attractiveness or usefulness (voting/purchasing), cost-effectiveness (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 $\epsilon^a$ 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)

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

- Variables:
  - male, age, manual (father a manual worker)
  - rldist: distance between voter and party on left-right dimension constructed from respondent’s and party’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

<table>
<thead>
<tr>
<th>serialno</th>
<th>year</th>
<th>party</th>
<th>chosen</th>
<th>rldist</th>
</tr>
</thead>
<tbody>
<tr>
<td>10.</td>
<td>92</td>
<td>con</td>
<td>1</td>
<td>.5031</td>
</tr>
<tr>
<td>11.</td>
<td>92</td>
<td>lab</td>
<td>0</td>
<td>30.12</td>
</tr>
<tr>
<td>12.</td>
<td>92</td>
<td>lib</td>
<td>0</td>
<td>16.18</td>
</tr>
<tr>
<td>13.</td>
<td>87</td>
<td>con</td>
<td>0</td>
<td>22.86</td>
</tr>
<tr>
<td>14.</td>
<td>87</td>
<td>lab</td>
<td>1</td>
<td>.3622</td>
</tr>
<tr>
<td>15.</td>
<td>87</td>
<td>lib</td>
<td>0</td>
<td>1.894</td>
</tr>
<tr>
<td>16.</td>
<td>92</td>
<td>con</td>
<td>0</td>
<td>20.62</td>
</tr>
<tr>
<td>17.</td>
<td>92</td>
<td>lab</td>
<td>1</td>
<td>.0567</td>
</tr>
<tr>
<td>18.</td>
<td>92</td>
<td>lib</td>
<td>0</td>
<td>1.507</td>
</tr>
<tr>
<td>19.</td>
<td>87</td>
<td>con</td>
<td>0</td>
<td>25.32</td>
</tr>
</tbody>
</table>

- zrldist, zprice are standardised versions

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:
  \[ V_{ij} = \beta^T X_{ij} + (\alpha + u_i) d_{ij} \]
  where \( i \) indexes the voter, \( j \) indexes the election and \( d_{ij} \) is the distance between voter and party on the left-right political dimension.
  eq beta1: zrldist
gllamm party zrldist lab87 lib87 lab92 lib92 lab_man lib_man /*
  */ lab_age lib_age lab_man lab_zpri lib_zpri, nocons */
  */ i(serialno) expand(occ chosen o) f(binom) l(mlogit) eq(beta1)
### 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 $\epsilon^a$ has an extreme value distribution (Gumbel)
- Let $r^s$ be the alternative with rank $s$. Then the ranking $R = (r^1, r^2, \ldots, r^A)$ is obtained if
  \[ U^{r^1} > U^{r^2} > \cdots > U^{r^A} \]
- The probability of a ranking $R$ is (Luce, 1959)
  \[ \Pr(R) = \frac{\exp(V^{r^1})}{\sum_{i=1}^{A} \exp(V^{r^i})} \times \frac{\exp(V^{r^2})}{\sum_{i=2}^{A} \exp(V^{r^i})} \times \cdots \times \frac{\exp(V^{r^A})}{\sum_{i=A-1}^{A} \exp(V^{r^i})} \]
- 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
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

<table>
<thead>
<tr>
<th>serialno</th>
<th>year</th>
<th>occ</th>
<th>rank</th>
<th>party</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.</td>
<td>11</td>
<td>87</td>
<td>3</td>
<td>con</td>
</tr>
<tr>
<td>8.</td>
<td>11</td>
<td>87</td>
<td>3</td>
<td>lib</td>
</tr>
<tr>
<td>9.</td>
<td>11</td>
<td>87</td>
<td>3</td>
<td>lab</td>
</tr>
<tr>
<td>10.</td>
<td>11</td>
<td>92</td>
<td>4</td>
<td>con</td>
</tr>
<tr>
<td>11.</td>
<td>11</td>
<td>92</td>
<td>4</td>
<td>lab</td>
</tr>
<tr>
<td>12.</td>
<td>11</td>
<td>92</td>
<td>4</td>
<td>lib</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>serialno</th>
<th>year</th>
<th>occstage</th>
<th>party</th>
<th>chosen</th>
</tr>
</thead>
<tbody>
<tr>
<td>11.</td>
<td>11</td>
<td>87</td>
<td>con</td>
<td>1</td>
</tr>
<tr>
<td>12.</td>
<td>11</td>
<td>87</td>
<td>lab</td>
<td>0</td>
</tr>
<tr>
<td>13.</td>
<td>11</td>
<td>87</td>
<td>lib</td>
<td>0</td>
</tr>
<tr>
<td>14.</td>
<td>11</td>
<td>92</td>
<td>con</td>
<td>1</td>
</tr>
<tr>
<td>15.</td>
<td>11</td>
<td>92</td>
<td>lab</td>
<td>0</td>
</tr>
<tr>
<td>16.</td>
<td>11</td>
<td>92</td>
<td>lib</td>
<td>0</td>
</tr>
<tr>
<td>17.</td>
<td>11</td>
<td>92</td>
<td>lib</td>
<td>1</td>
</tr>
<tr>
<td>18.</td>
<td>11</td>
<td>92</td>
<td>lab</td>
<td>0</td>
</tr>
</tbody>
</table>
Analysing rankings

- There are a number of possible random structures for election within voter within constituency.

- Example: correlated random coefficients for lab and lib at voter level

\[ V_{ij}^a = \beta^a x_{ij} + \alpha d_{ij} + u_{1i} z_{1ij} + u_{2i} x_{2ij} \]

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

- random coefficients of Labour and Liberal induce longitudinal 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

<table>
<thead>
<tr>
<th></th>
<th>var(1): 18.918555 (2.2960479)</th>
<th>cov(1,2): 9.0408098 (1.1798509)</th>
<th>cor(1,2): .82567035</th>
</tr>
</thead>
<tbody>
<tr>
<td>var(2): 6.3374322</td>
<td></td>
<td></td>
<td>.88875039</td>
</tr>
</tbody>
</table>

Assuming uncorrelated random coefficients (using nocor option, gives a log-likelihood of -2800.2076
II. Regressions of latent variables on observed variables

Structural equations for the latent variables

Regress the latent variables on other latent and explanatory variables

\[ u = Bu + \Gamma w + \zeta \]

- factors
- random coefficients

\[ u = (u^{(2)}_1, u^{(2)}_2, \ldots, u^{(2)}_{M_2}, \ldots, u^{(l)}_1, \ldots, u^{(l)}_{M_l})' \quad (M \text{ elements}) \]

- 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)
**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 $i$th subject’s true dietary intake ($-\text{population mean}$)
- **Measurement model for fibre intake**: $y_{1i}, y_{2i}$ conditionally independently normally distributed with
  \[ E[y_{ij} | u] = \beta_j + u_i \lambda_j, \quad j = 1, 2 \quad (\beta_1 = \beta_2, \ \lambda_1 = \lambda_2 = 1) \]
- **Disease model**: $y_{3i}$ conditionally binomial with
  \[ \logit(E[y_{3i} | u]) = \beta_3 + u_i \lambda_3 \quad \lambda_3 \text{ is log(OR)} \]
- **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.
  \[ \eta_{ij} = \beta z_{ij} + u_i \lambda z_{ij} \quad z_{ij} = (z_{1ij}, z_{3ij})' \]

---

**Diet example in gllamm, see STB53, sg129**

<table>
<thead>
<tr>
<th>id</th>
<th>resp</th>
<th>diet</th>
<th>chd</th>
<th>var</th>
</tr>
</thead>
<tbody>
<tr>
<td>425</td>
<td>217</td>
<td>3.06</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>426</td>
<td>217</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>427</td>
<td>218</td>
<td>3.14</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>428</td>
<td>218</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>429</td>
<td>219</td>
<td>2.75</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>430</td>
<td>219</td>
<td>2.7</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>431</td>
<td>219</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

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** $y_3$

  Measurement model: $E[y_{ij}|u] = \beta_j + u_i$, $j = 1, 2$

  Disease model: $\text{logit}(E[y_{i3}|u]) = \beta_3 + \beta_4 x + u_{i3}$

- **Indirect effect of** $x$ **on** $y_3$

  $u_i = \gamma x + \zeta_i$,

  where $\zeta_i$ is a residual error term

  Measurement model: $E[y_{ij}|u] = \beta_j + \gamma x + \zeta_i$

  Disease model: $\text{logit}(E[y_{i3}|u]) = \beta_3 + \gamma \lambda_3 x + \zeta_i \lambda_3$

  $\Rightarrow$ would require nonlinear constraint for the coefficients if we couldn’t regress $u$ on explanatory variables

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

  Measurement model: $E[y_{ij}|u] = \beta_j + \gamma x + \zeta_i$

  Disease model: $\text{logit}(E[y_{i3}|u]) = \beta_3 + (\beta_4 + \gamma \lambda_3) x + \zeta_i \lambda_3$

  (would not require a constraint for the coefficients)

\[ \begin{align*}
\lambda_3 & \quad \beta_4 \\
\gamma & \quad \beta_3 \\
\beta_1 & \\
\zeta_i & \quad \lambda_3 \\
\end{align*} \]

Including effect of occupation (bus staff vs bank staff)

<table>
<thead>
<tr>
<th>id</th>
<th>resp</th>
<th>diet</th>
<th>chd</th>
<th>var</th>
</tr>
</thead>
<tbody>
<tr>
<td>425.</td>
<td>217</td>
<td>3.06</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>426.</td>
<td>217</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>427.</td>
<td>218</td>
<td>3.14</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>428.</td>
<td>218</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>429.</td>
<td>219</td>
<td>2.75</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>430.</td>
<td>219</td>
<td>2.7</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>431.</td>
<td>219</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

- **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)
Results

- direct and indirect effect of $x$ on $y_3$

- Indirect effect of $x$ on $y_3$

\[
\begin{align*}
\text{Log-likelihood} &= -186.90 \\
\begin{array}{l|l|l}
\text{Parameters} & \text{Estimates} & \text{SE} \\
\hline
\lambda_3 & -1.95 & 0.73 \\
\gamma & -0.12 & 0.03 \\
\beta_4 & -0.19 & 0.34 \\
\sigma^2 & 0.02 & 0.003 \\
\text{var}(u) & 0.07 & 0.007 \\
\end{array}
\end{align*}
\]

\[
\begin{align*}
\text{Log-likelihood} &= -187.05 \\
\begin{array}{l|l|l}
\text{Parameters} & \text{Estimates} & \text{SE} \\
\hline
\lambda_3 & -1.86 & 0.70 \\
\gamma & -0.12 & 0.03 \\
\sigma^2 & 0.02 & 0.003 \\
\text{var}(u) & 0.07 & 0.007 \\
\end{array}
\end{align*}
\]