Estimating chopit models in `gllamm`
Political efficacy example from King et al. (2002)

Sophia Rabe-Hesketh
Department of Biostatistics and Computing
Institute of Psychiatry
King's College London

Anders Skrondal
Division of Epidemiology
Norwegian Institute of Public Health
Oslo

September 2, 2002

1 Introduction

King et al. (2002) introduce a method for analyzing ordinal survey responses taking into account individual differences in interpretation of the survey questions. In addition to answering a survey question relating to their own situation (the 'self-assessment' question), respondents answer the same question in relation to a number of hypothetical individuals described by written vignettes. The responses to the vignettes are then used as anchors for the self-assessment question by specifying a joint 'chopit' (compound hierarchical ordinal probit) model for the self-assessment question and vignettes.

The purpose of this document is to show how the political efficacy example discussed in Section 8.1 of King et al. (2002) can be estimated using the Stata program `gllamm` (Rabe-Hesketh, Pickles, and Skrondal, 2001; Rabe-Hesketh, Skrondal, and Pickles, 2002b). The `gllamm` program, simulated data and do-file for this example are available from

http://www.iop.kcl.ac.uk/IoP/Departments/BioComp/programs/gllamm.html

2 Brief description of the political efficacy example

2.1 The data

Surveys were carried out for the WHO in China ($n = 371$) and Mexico ($n = 551$) in 2002. Respondents were asked

> How much say do you have in getting the government to address issues that interest you?

and given the following set of ordinal categories in which to respond: (1) 'no say at all', (2) 'little say', (3) 'some say', (4) 'a lot of say' and (5) 'unlimited say'.

The respondents were also given five vignettes describing hypothetical individuals with varying degrees of political efficacy and were asked the same question as above regarding each hypothetical individual (with appropriate substitutions for 'you'). Possible covariates include country, age, sex and years of education.

2.2 Model for self-assessment question

The response y_i for person i is modelled as an ordinal probit model with underlying response

$$y_i^* = \mathbf{x}_i'\boldsymbol{\beta} + \epsilon_i,$$

where $\mathbf{x}_i$ are covariates, $\boldsymbol{\beta}$ are fixed effects and ϵ_i is a residual error term

$$\epsilon_i \sim N(0,1).$$

The observed responses $k=1,\cdots,K$ are generated via a threshold model with person-specific thresholds τ_i^k

$$y_i = k \quad \text{if } \tau_i^{k-1} \leq y_i^* < \tau_i^k,$$

where $-\infty = \tau_i^0 < \tau_i^1 < \cdots < \tau_i^K = \infty$. The thresholds are modelled as

$$\begin{aligned} \tau_i^1 &= \boldsymbol{\gamma}^{1\prime}\mathbf{v}_i \\ \tau_i^k &= \tau_i^{k-1} + \exp(\boldsymbol{\gamma}^{k\prime}\mathbf{v}_i), \quad k = 2,\cdots,K, \end{aligned} \tag{1}$$

where $\mathbf{v}_i$ are covariates and $\boldsymbol{\gamma}^k$ are parameters.

Here the underlying response y_i^* can be interpreted as the true preceived political efficacy of respondent i, on a scale that is comparable across individuals. The observed responses result from different individuals applying different thresholds τ_i^k and are therefore no longer comparable.

2.3 Model for vignettes

In order to anchor the self-assessment questions against the vignettes, it must be assumed that there is a true political efficacy θ_j associated with the hypothetical person described in the jth vignette $j = 1,\cdots,J$. The true perception of the survey respondents differs from this only by a random error term

$$z_{ij}^* = \theta_j + u_{ij},$$

$$u_{ij} \sim N(0,\sigma^2).$$

Note that, in contrast to the self-assessment question, the standard deviation σ of the error term is now a free parameter. It is further assumed that the observed responses are generated by applying the same thresholds as for the self-assessment question, i.e.,

$$z_{ij} = k \quad \text{if } \tau_i^{k-1} \leq z_{ij}^* < \tau_i^k,$$

where the thresholds are modelled as in (1).

3 Estimation using gllamm

We will now analyze the original data, but note that the data provided in *effic1.dta* differ from these since the responses were simulated as discussed below. We have six responses or items per person; the self-assessment question y_i and the vignettes z_{i1} to z_{i5} with corresponding variables names `xsayself` and `xsay1` to `xsay5`.

```
. use effic0, clear
. list xsay* in 1/5
     xsayself     xsay1     xsay2     xsay3     xsay4     xsay5
  1.        1         5         1         2         5         2
  2.        0         3         1         1         3         4
  3.        1         1         1         5         5         5
  4.        2         3         2         2         1         1
  5.        2         4         3         2         2         1
```

Here the value 0 stands for 'missing'. In order to use `gllamm`, the responses must be stacked into a single variable. We will use the `reshape` command to achieve this and also delete items with missing values:

```
. rename xsayself xsay6
. reshape long xsay, i(id) j(item)
(note: j = 1 2 3 4 5 6)
Data                                   wide   ->   long
-----------------------------------------------------------------------------
Number of obs.                          981   ->    5886
Number of variables                      12   ->       8
j variable (6 values)                         ->   item
xij variables:
                  xsay1 xsay2 ... xsay6   ->   xsay
-----------------------------------------------------------------------------
. drop if xsay==0
(806 observations deleted)
. list id xsay item in 1/6
           id      xsay      item
  1.        1         5         1
  2.        1         1         2
  3.        1         2         3
  4.        1         5         4
  5.        1         2         5
  6.        1         1         6
```

The linear predictors in the probit models (or the means of the underlying responses) are $\mathbf{x}'\boldsymbol{\beta}$ for the self-assessment question and θ_1 to θ_5 for the vignettes. To define these linear predictors using a single set of covariates or 'design matrix', we need to generate dummy variables, `i1` to `i6` for items 1 to 6:

```
. tab item, gen(i)
       item |      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |        844       16.61       16.61
          2 |        842       16.57       33.19
          3 |        841       16.56       49.74
          4 |        845       16.63       66.38
          5 |        849       16.71       83.09
          6 |        859       16.91      100.00
------------+-----------------------------------
      Total |       5080      100.00
. sort id item
```

(*Continued on next page*)

```
. list id xsay i1-i6 in 1/6
          id    xsay      i1      i2      i3      i4      i5      i6
  1.       1       5       1       0       0       0       0       0
  2.       1       1       0       1       0       0       0       0
  3.       1       2       0       0       1       0       0       0
  4.       1       5       0       0       0       1       0       0
  5.       1       2       0       0       0       0       1       0
  6.       1       1       0       0       0       0       0       1
```

The covariates `x` (`china`, `age`, `male` and `educyrs`) must now be multiplied by the dummy variable for the self-assessment question which we will rename to `self`

```
. rename i6 self
. for var china age male educyrs: gen s_X=self*X
->  gen s_china=self*china
->  gen s_age=self*age
->  gen s_male=self*male
->  gen s_educyrs=self*educyrs
```

The linear predictors can now be defined in the `gllamm` command using

```
gllamm xsay s_china s_age s_male s_educyrs i1 i2 i3 i4 i5, ...
```

Each response is modelled as a scaled ordinal probit, with a separate scale for the self-assessment question ($\mathrm{sd}(\epsilon_i) = 1$) and for the vignettes ($\mathrm{sd}(u_{ij}) = \sigma$). In `gllamm`, we must therefore specify a 'scaled ordinal probit link' using `link(soprobit)` and introduce heteroscedasticity using the `s(het)` option where `het` is an equation for the log of the scale:

```
. gen vign = 1-self
. eq het: vign self
```

We can use Stata's `constraints` command to set the scale for the self-assessment question to 1. In `gllamm` the relevant parameter is the log standard deviation `[lns1]self` which must then be set to 0

```
. constraint def 1 [lns1]self=0
```

We now specify the model in (1) for the thresholds with the same covariates $\mathbf{v}_i = \mathbf{x}_i$ as for the linear predictor of the self-assessment question

```
. eq thresh: china age male educyrs
```

this model will be passed to `gllamm` using the `ethresh(thresh)` option. Since the six responses are treated as a single ordinal response, we only need to specify a single threshold model and there is no need to explicitly constrain the $\boldsymbol{\gamma}^k$ parameters to be the same across items.

Since the model does not contain any random effects or latent variables, we can estimate the model using the `init` option (stands for initial values, omitting any latent variables):

```
. gllamm xsay s_china s_age s_male s_educyrs i1 i2 i3 i4 i5,  /*
>   */ i(id) link(soprobit) s(het) constr(1) ethresh(thresh)  /*
>   */ init

number of level 1 units = 5080

Condition Number = 1643.3401

gllamm model with constraints:
 ( 1)  [lns1]self = 0.0

log likelihood = -7062.131187808051
```

(*Continued on next page*)

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
xsay						
s_china	-.3628872	.0904038	-4.01	0.000	-.5400754	-.1856989
s_age	.0058606	.0028128	2.08	0.037	.0003476	.0113735
s_male	.1134398	.0809884	1.40	0.161	-.0452945	.2721741
s_educyrs	.019557	.0082068	2.38	0.017	.0034721	.035642
i1	1.282264	.1606217	7.98	0.000	.9674518	1.597077
i2	1.193919	.1600941	7.46	0.000	.8801404	1.507698
i3	.8424784	.1589611	5.30	0.000	.5309204	1.154036
i4	.7919615	.1589452	4.98	0.000	.4804346	1.103488
i5	.6188468	.1590183	3.89	0.000	.3071766	.930517
_cut11						
china	-1.059206	.0591826	-17.90	0.000	-1.175202	-.9432106
age	.0019404	.0013004	1.49	0.136	-.0006083	.0044891
male	.0434652	.0363915	1.19	0.232	-.0278608	.1147912
educyrs	-.0010787	.0037971	-0.28	0.776	-.0085209	.0063636
_cons	.4389508	.1513253	2.90	0.004	.1423586	.735543
_cut12						
china	-.1607994	.0708768	-2.27	0.023	-.2997154	-.0218834
age	-.0020513	.0018631	-1.10	0.271	-.0057029	.0016003
male	-.0572983	.0504858	-1.13	0.256	-.1562487	.041652
educyrs	.0016812	.005509	0.31	0.760	-.0091162	.0124787
_cons	-.2621123	.1117188	-2.35	0.019	-.481077	-.0431475
_cut13						
china	.3439344	.0525479	6.55	0.000	.2409423	.4469264
age	-.0010884	.0016392	-0.66	0.507	-.0043012	.0021244
male	.0432661	.0472464	0.92	0.360	-.0493352	.1358675
educyrs	-.0023726	.0050643	-0.47	0.639	-.0122985	.0075533
_cons	-.4853311	.1039865	-4.67	0.000	-.6891409	-.2815214
_cut14						
china	.6279309	.0829145	7.57	0.000	.4654214	.7904403
age	.0042546	.0023514	1.81	0.070	-.0003541	.0088632
male	-.0977914	.072163	-1.36	0.175	-.2392283	.0436456
educyrs	.0266069	.0073255	3.63	0.000	.0122492	.0409646
_cons	-1.613507	.1485483	-10.86	0.000	-1.904656	-1.322357

```
Variance at level 1
------------------------------------------------------------------------------

    equation for log standard deviaton:

    vign: -.23828219 (.04228416)
    self: 0 (0)
------------------------------------------------------------------------------
```

In terms of the model parameters, the estimates are given in Table 1 under 'Real Data'. These estimates are very close to the estimates in Table 2 of King et al. (2002).

Since the real data cannot be made available, we created artificial data by simulating the responses from the model just estimated using a 'post-estimation' command for `gllamm`, `gllasim`:

```
. drop xsay
. set seed 12345
. gllasim xsay
```

The data file *effic1.dta* available from the `gllamm` webpage contains these simulated responses in the same 'wide' form as the original data, so that all commands described in this section can be repeated. However, the estimates will be a little different due to sampling variability (see also Table 1 under 'Simulated Data'):

Table 1: Estimates using `gllamm`

		Real Data		Simulated Data	
		coeff.	s.e.	coeff.	s.e.
$\boldsymbol{\beta}$	china	-0.363	0.090	-0.222	0.088
	age	0.006	0.003	0.002	0.003
	male	0.113	0.081	0.187	0.080
	education	0.020	0.008	0.011	0.008
$\boldsymbol{\gamma}^1$	china	-1.059	0.059	-1.068	0.059
	age	0.002	0.001	0.001	0.001
	male	0.043	0.036	0.038	0.036
	education	-0.001	0.004	0.000	0.004
	constant	0.439	0.151	0.218	0.152
$\boldsymbol{\gamma}^2$	china	-0.161	0.071	-0.127	0.070
	age	-0.002	0.002	-0.002	0.002
	male	-0.057	0.050	-0.083	0.050
	education	0.002	0.006	0.003	0.005
	constant	-0.262	0.112	-0.258	0.106
$\boldsymbol{\gamma}^3$	china	0.344	0.053	0.268	0.051
	age	-0.001	0.002	0.001	0.002
	male	0.043	0.047	0.017	0.047
	education	-0.002	0.005	-0.001	0.005
	constant	-0.485	0.104	-0.489	0.099
$\boldsymbol{\gamma}^4$	china	0.628	0.083	0.677	0.082
	age	0.004	0.002	0.002	0.002
	male	-0.098	0.072	-0.148	0.070
	education	0.027	0.007	0.030	0.007
	constant	-1.614	0.148	-1.498	0.146
θ_1	vignette 1	1.282	0.160	1.016	0.159
θ_2	vignette 2	1.194	0.160	0.947	0.159
θ_3	vignette 3	0.842	0.159	0.639	0.158
θ_4	vignette 4	0.792	0.159	0.554	0.158
θ_5	vignette 5	0.619	0.159	0.394	0.158
$\ln(\sigma)$	log scale for vignettes	-0.238	0.042	-0.237	0.041

```
. matrix a=e(b)

. gllamm xsay s_china s_age s_male s_educyrs i1 i2 i3 i4 i5, /*
>   */ i(id) link(soprobit) s(het) constr(1) ethresh(thresh) /*
>   */ init from(a) long

number of level 1 units = 5080

Condition Number = 1639.1856

gllamm model with constraints:
 ( 1)  [lns1]self = 0.0

log likelihood = -7055.990593194162

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
xsay         |
     s_china |  -.2220871   .0883742    -2.51   0.012    -.3952974   -.0488769
       s_age |   .0019784   .0027754     0.71   0.476    -.0034612     .007418
      s_male |    .187026     .08026     2.33   0.020     .0297194    .3443327
   s_educyrs |   .0110663   .0081896     1.35   0.177     -.004985    .0271177
          i1 |   1.016296   .1590159     6.39   0.000     .7046307    1.327962
          i2 |   .9471709    .158769     5.97   0.000     .6359893    1.258352
          i3 |   .6385328    .158199     4.04   0.000     .3284685    .9485972
          i4 |   .5536792   .1582356     3.50   0.000     .2435431    .8638153
          i5 |   .3936774   .1584527     2.48   0.013     .0831159    .7042389
-------------+----------------------------------------------------------------
_cut11       |
       china |  -1.067561   .0586146   -18.21   0.000    -1.182444   -.9526787
         age |   .0014611   .0012901     1.13   0.257    -.0010674    .0039895
        male |   .0380037   .0361159     1.05   0.293    -.0327821    .1087896
     educyrs |  -.0002502   .0037184    -0.07   0.946    -.0075381    .0070377
       _cons |   .2177559   .1520234     1.43   0.152    -.0802044    .5157163
-------------+----------------------------------------------------------------
_cut12       |
       china |   -.126661   .0695481    -1.82   0.069    -.2629728    .0096507
         age |  -.0025451   .0018357    -1.39   0.166    -.0061431    .0010528
        male |  -.0825773   .0504054    -1.64   0.101      -.18137    .0162155
     educyrs |   .0027462   .0051001     0.54   0.590    -.0072499    .0127423
       _cons |    -.25809    .105963    -2.44   0.015    -.4657737   -.0504063
-------------+----------------------------------------------------------------
_cut13       |
       china |   .2679189   .0508883     5.26   0.000     .1681796    .3676582
         age |   .0005807   .0016031     0.36   0.717    -.0025613    .0037228
        male |   .0170187   .0466373     0.36   0.715    -.0743887    .1084261
     educyrs |  -.0014223   .0047722    -0.30   0.766    -.0107756     .007931
       _cons |  -.4891784   .0992565    -4.93   0.000    -.6837176   -.2946392
-------------+----------------------------------------------------------------
_cut14       |
       china |   .6767611   .0815664     8.30   0.000     .5168938    .8366284
         age |   .0024734   .0023321     1.06   0.289    -.0020974    .0070442
        male |  -.1478691   .0701101    -2.11   0.035    -.2852824   -.0104558
     educyrs |   .0295294   .0067523     4.37   0.000     .0162952    .0427637
       _cons |  -1.498073    .145767   -10.28   0.000    -1.783771   -1.212375
------------------------------------------------------------------------------

Variance at level 1
------------------------------------------------------------------------------

    equation for log standard deviaton:

    vign: -.23655207 (.04105994)
    self: 0 (0)
------------------------------------------------------------------------------
```

King et al. (2002) describe a more general model for the situation when there are several self-assessment questions with a set of vignettes for one of them. Their general model includes a shared random effect for the self-assessment questions, a common threshold model for one self-assessment question and the corresponding vignettes and separate threshold models for the other self-assessment questions. Such models and various extensions can also be estimated in `gllamm`, since they are special cases of GLLAMMs (Generalized Linear Latent And Mixed Models), see for example Rabe-Hesketh et al. (2002a).

References

King, G., Murray, C. J. L., Salomon, J. A., and Tandon, A. 2002. Enhancing the validity of cross-cultural comparability of survey research. Submitted for publication.
Downloadable from `gking.harvard.edu/files/vign.pdf`.

Rabe-Hesketh, S., Pickles, A., and Skrondal, A. 2001. GLLAMM Manual. Tech. Rep. 2001/01, Department of Biostatistics and Computing, Institute of Psychiatry, King's College, University of London.
Downloadable from `www.iop.kcl.ac.uk/iop/departments/biocomp/programs/gllamm.html`.

Rabe-Hesketh, S., Skrondal, A., and Pickles, A. 2002a. Generalized multilevel structural equation modeling. *Psychometrika*. Conditionally accepted.

—. 2002b. Reliable estimation of generalized linear mixed models using adaptive quadrature. *The Stata Journal* 2:1–21.