Lecture 5: Random coefficient models for non-normal data

Antonello Maruotti

Lecturer in Medical Statistics, S3RI and School of Mathematics
University of Southampton

Southampton, 11 December 2014

Summary

Random coefficient Poisson model

Random coefficient Logistic model

Mixed models

- Mixed models state that observed data consist of two parts
  - fixed effects
  - random effects
- Fixed effects define the expected values of the observations
- Random effects result from variation between subjects and from variation within subjects.

Mixed models

- Measures on the same subject at different times almost always are correlated, with measures taken close together in time being more highly correlated than measures take far apart in time
- Observations on different subjects are often assume independent
- Mixed models are used with repeated measures data to accommodate the fixed effects of covariates and the covariation between observations on the same subject at different times

Mixed models: Random coefficients

- To model the mean structure in sufficient generality to ensure unbiasedness of the fixed effect estimates
- Random intercept models allow the overall level of response to vary over *clusters* after controlling for covariates
- We include random coefficient or random slope in addition to random intercepts, thus also allowing the effects of covariates to vary over *clusters*
- In longitudinal settings, we have occasions (level-1 units) measured for different subjects (clusters). These models are also referred to as growth-curve models.

Model specification

- Let Y_{ik} denote the value of the response measured at time k on subject i
- $g[E(Y_{ik} \mid \alpha_{i0}, \alpha_{i1})] = (\gamma_0 + \alpha_{i0}) + (\gamma_1 + \alpha_{i1})x_{ik}$
- $g[E(\mathbf{Y} \mid \cdot)] = \mathbf{X}\boldsymbol{\gamma} + \mathbf{Z}\boldsymbol{\alpha}$
  $\mathbf{X}$ is a matrix of known covariates
  $\boldsymbol{\beta}$ is the vector of fixed parameters
  $\mathbf{Z}$ is a matrix collecting random effects
  $\boldsymbol{\alpha}$ is the vector of random parameters.

Covariance structure

- We assume that, given the covariates $\mathbf{x}_{ik}$, the random intercept and random coefficient (collected in $\boldsymbol{\alpha}$) have a multivariate normal distribution with zero mean.
- Let us consider a model with a random intercept and a single random slope, the random terms have a bivariate normal distribution with zero mean and covariance matrix

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{bmatrix}, \quad \sigma_{12} = \sigma_{21}$$

- The correlation between the random intercept and the random coefficient becomes

$$\rho_{12} = \frac{\sigma_{12}}{\sqrt{\sigma_1^2 \sigma_2^2}}$$

Covariance structure

independent: one variance parameter per random effect, all covariances zero; the default unless a factor variable is specified

exchangeable: equal variances for random effects, and one common pairwise covariance

identity: equal variances for random effects, all covariances zero; the default for factor variables

unstructured: all variances-covariances distinctly estimated

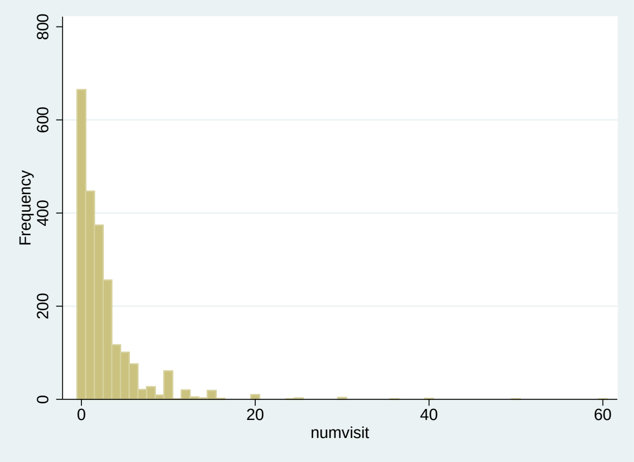
Health-care reform data

- We analyze data from the German Socio-Economic Panel (see SOEP Group, 2001; Winklmann, 2004)
- A major health-care reform took place in 1997.
- We will consider a subset of data, comprised of women working full time in 1996 panel wave preceding the reform and the 1998 panel wave following the reform
- The dataset `drvisits.dta` is considered

Data description

- `id`: person identifier
- `numvisits`: self-reported number of visits to a doctor during the 3 months before the interview
- `reform`: dummy variable for interview being during 1998
- `age`: age in years
- `educ`: education in years
- `married`: dummy variable for being married
- `badh`: dummy variable for self-reported current health
- `loginc`: logarithm of household income

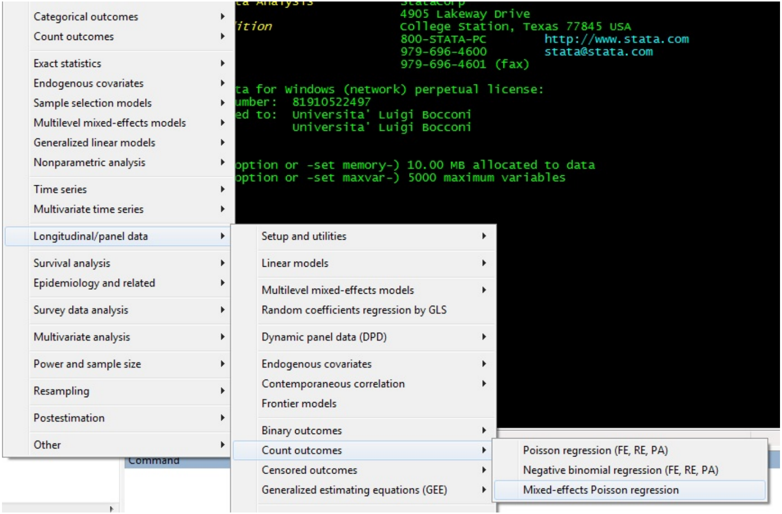
Data description



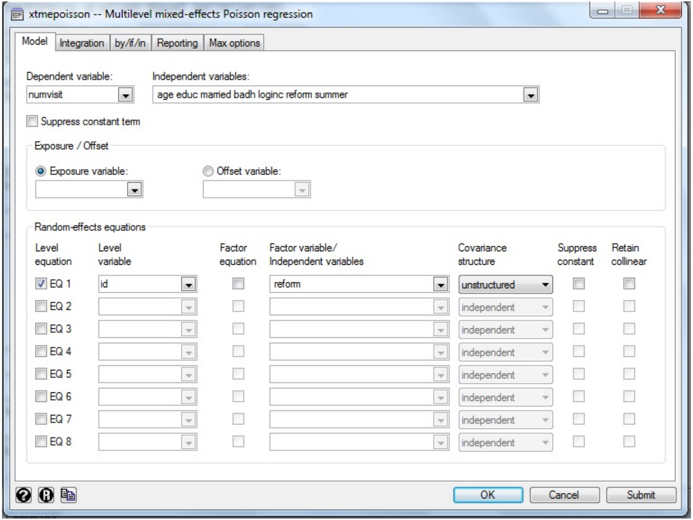
Do data fit the Poisson assumptions?

```
. summarize  numvisit, detail

                          numvisit
-------------------------------------------------------------
      Percentiles      Smallest
 1%            0              0
 5%            0              0
10%            0              0       Obs                2227
25%            0              0       Sum of Wgt.        2227

50%            2                      Mean           2.589133
                        Largest       Std. Dev.      4.016199
75%            3             40
90%            6             40       Variance       16.12985
95%           10             50       Skewness        4.92272
99%           20             60       Kurtosis       45.41174

.
```

xtmepoisson



xtmepoisson



xtmepoisson

```
Mixed-effects Poisson regression                Number of obs      =      2227
Group variable: id                              Number of groups   =      1518

                                                Obs per group: min =         1
                                                               avg =       1.5
                                                               max =         2

Integration points =   7                        Wald chi2(7)       =    241.11
Log likelihood = -4513.7299                     Prob > chi2        =    0.0000
```

numvisit	IRR	Std. Err.	z	P>|z|	[95% Conf.	Interval]
age	1.003456	.0028317	1.22	0.222	.997921	1.009021
educ	1.008894	.0128121	0.70	0.486	.984093	1.034321
married	1.086874	.064118	1.41	0.158	.9681982	1.220097
badh	3.028323	.2322949	14.44	0.000	2.605606	3.519619
loginc	1.135637	.0866512	1.67	0.096	.9778926	1.318827
reform	.9022789	.0483866	-1.92	0.055	.8122568	1.002278
summer	.9140245	.0741941	-1.11	0.268	.7795846	1.071649

Random-effects Parameters	Estimate	Std. Err.	[95% Conf.	Interval]
id: Unstructured				
sd(reform)	.9303105	.0561788	.8264685	1.0472
sd(_cons)	.9541109	.0357017	.8866412	1.026715
corr(reform,_cons)	-.4908238	.0506049	-.5835348	-.3854832

LR test vs. Poisson regression: chi2(3) = 2857.93 Prob > chi2 = 0.0000

Results

	Poisson		RI Poisson		RC Poisson	
	IRR	(95% CI)	IRR	(95% CI)	IRR	(95% CI)
Fixed part						
`reform`	0.87	(0.82; 0.91)	0.95	(0.90; 1.02)	0.90	(0.81; 1.00)
`age`	1.00	(1.00; 1.01)	1.10	(1.00;1.01)	1.00	(1.00;1.01)
`educ`	0.99	(0.98; 1.00)	0.99	(0.97;1.01)	1.01	(0.98;1.03)
`married`	1.04	(0.99; 1.10)	1.08	(0.97;1.20)	1.09	(0.97;1.22)
`badh`	3.11	(2.93; 3.30)	2.48	(2.19;2.78)	3.03	(2.61;3.52)
`loginc`	1.16	(1.08; 1.25)	1.10	(0.96;1.25)	1.14	(0.98;1.32)
`summer`	1.01	(0.93; 1.09)	0.87	(0.76;0.98)	0.91	(0.78;1.07)
Random part						
σ_1			0.90		0.95	
σ_1					0.93	
ρ_{12}					-0.49	
Log-likelihood	-5942.69		-4643.36		-4513.73	

Random intercept vs. random coefficient

- `xtmepoisson numvisit age educ married badh loginc reform summer || id:, irr`
- `estimates store xtri`
- `xtmepoisson numvisit age educ married badh loginc reform summer || id:reform, covariance(unstructured) irr`
- `estimates store xtrc`
- `lrtest xtri xtrc`

```
. lrtest xtri xtrc

Likelihood-ratio test                                 LR chi2(2)  =    259.10
(Assumption: xtri nested in xtrc)                     Prob > chi2 =    0.0000
```

Immunization data

- Data are available from the National Survey of Maternal and Child Health conducted in Guatemala in 1987
- A nationally representative sample of 5160 women aged between 15 and 44 were interviewed
- The questionnaire included questions determining the immunization status of children who were born in the previous 5 years and alive at the time of the interview
- Beginning 1986, the Guatemalan government undertook a series of campaign to immunize the population against major childhood diseases
- An important explanatory variable is whether the child was at least 2 years old at the time of the interview, in which case the child was old enough to be immunized during the 1986 campaign
- If this variable is associated with immunization, there is some indication that the government campaign work

Data description

- `immun`: Indicator variable for child receiving full set of immunizations
- `kid2p`: dummy variable for child being at least 2 years old at time of the interview
- `cluster`: identifier for communities
- `rural`: dummy variable for community being rural
- `pcInd81`: percentage of population that was indigenous in 1981

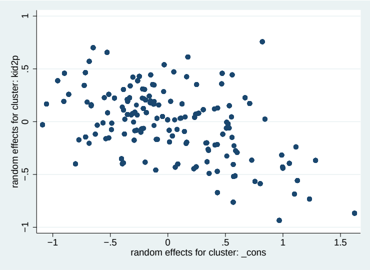
Results

xtmelogit

```
Mixed-effects logistic regression               Number of obs      =      2159
Group variable: cluster                         Number of groups   =       161

                                                Obs per group: min =         1
                                                               avg =      13.4
                                                               max =        55

Integration points =   7                        Wald chi2(3)       =     84.83
Log likelihood = -1378.2717                     Prob > chi2        =    0.0000

------------------------------------------------------------------------------
       immun | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       kid2p |   2.983037   .4716586     6.91   0.000     2.188125    4.066728
       rural |   .5294337    .086763    -3.88   0.000     .3839879    .7299708
     pcInd81 |   .3842706   .0782523    -4.70   0.000     .2578099    .5727625
------------------------------------------------------------------------------

------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
cluster: Unstructured        |
                   sd(kid2p) |    .775025   .2229342      .4410314    1.361952
                   sd(_cons) |   .9244355   .1961248      .6099425    1.401085
            corr(kid2p,_cons)|  -.6872667   .1609391     -.8938087   -.2401777
------------------------------------------------------------------------------
LR test vs. logistic regression:     chi2(3) =    68.64   Prob > chi2 = 0.0000

Note: LR test is conservative and provided only for reference.
```

Predict random effects

- `xtmelogit immun kid2p rural pcInd81 || cluster:kid2p, covariance(unstructured) or`
- `predict ri rs, reff`
- `scatter ri rs`



Final remarks

- `xtmelogit` and `xtmepoisson` extend previous commands `xtlogit` and `xtpoisson`, respectively
- `xtmelogit` and `xtmepoisson` take longer than `xtlogit` and `xtpoisson` $\Rightarrow$ computations are intensive
- `xtmelogit` and `xtmepoisson` allow one to get predicted (modal) random effects and to add random effects
- The Laplacian approximation can be used as an alternative to Adaptive Gaussian Quadrature
- Multilevel models can be easily fitted