Diagnostics for generalised linear mixed models

Sophia Rabe-Hesketh, Institute of Psychiatry, King's College, London

Anders Skrondal, Norwegian Institute of Public Health, Oslo

UK Stata Users' Group Meeting
London, May 2003

Outline

- Example: Longitudinal epileptic seizure count data
- Influence
- Empirical Bayes (EB) prediction of higher-level residuals
- Detecting outliers by cross-validation
- Conclusions

Example: Longitudinal count data

- Famous epilepsy data from Thall & Vail (1990)
- 59 subjects j were randomized to receive progabide or placebo
- Outcomes:
  - Counts y_{ij} of epileptic seizures during the two weeks before each of four clinic visits, $i = 1, \cdots, n_j$, $n_j = 4$
- Between-subject covariates $\mathbf{x}_j$:
  - [Lbas] The logarithm of a quarter of the number of seizures in the eight weeks preceding entry into the trial
  - [Treat] Dummy variable for treatment group
  - [LbasTrt] Interaction between two variables above
  - [Lage] Logarithm of age
- Within-subject covariate z_{ij}:
  - [V4] Dummy for visit 4

Model and estimates

- Model II from Breslow & Clayton (1993)

$$y_{ij} \sim \mathrm{Poisson}(\mu_{ij}), \qquad \ln(\mu_{ij}) = \mathbf{x}'_j\boldsymbol{\beta} + \beta_5 z_{ij} + u_j, \qquad u_j \sim \mathrm{N}(0, \sigma^2)$$

```
gllamm y lbas treat lbas_trt lage v4, i(subj) fam(poiss) nip(15) adapt
gllamm, robust
```

			Robust
	Est	(SE)	(SE)
Fixed effects:			
β_0 [Cons]	2.11	(0.22)	(0.21)
β_1 [Lbas]	0.88	(0.13)	(0.11)
β_2 [Treat]	-0.93	(0.40)	(0.40)
β_3 [LbasTrt]	0.34	(0.20)	(0.20)
β_4 [Lage]	0.48	(0.35)	(0.30)
β_5 [V4]	-0.16	(0.05)	(0.07)
Random effect:			
σ	0.50	(0.06)	(0.06)
Log-likelihood		-665.29	

Influence of top-level unit j

- Influence on log-likelihood: Cook's D

$$D_j \;=\; -2\boldsymbol{s}_j'\boldsymbol{H}^{-1}\boldsymbol{s}_j,$$

- D_j can be interpreted as a quadratic approximation to twice the change in log-likelihood when parameters are estimated with and without cluster j
  - $\boldsymbol{s}_j$ is the score vector (first derivatives of log-likelihood contribution) for cluster j
  - $\boldsymbol{H}$ is the Hessian of the total log-likelihood
  - In `gllamm` (using numerical derivatives):

```
gllapred c, cooksd
```

Interpreting influence of top-level unit j

- Influence on particular parameter θ_p

$$\text{DFBETAS}_{pj} = \frac{\widehat{\theta}_p - \widehat{\theta}_{p(-j)}}{\mathrm{SE}(\widehat{\theta}_p)},$$

$\widehat{\theta}_{p(-j)}$ is the estimate of the pth parameter when cluster j is deleted

Influence for epilepsy data

							DFBETAS		
Subj.	[Base]	$\boldsymbol{y}_j$				Cook's D	[Treat]	[V4]	σ
Placebo									
126	13.0	40	20	23	12	1.10	-0.02	0.51	0.02
135	2.5	14	13	6	0	1.52	0.39	0.40	-0.34
227	13.8	18	24	76	25	1.46	-0.14	0.39	-0.33
Progabide									
207	37.8	102	65	72	63	1.68	0.58	0.24	-0.16
225	5.5	1	23	19	8	1.05	-0.23	0.18	-0.45
232	3.3	0	0	0	0	1.57	0.34	0.00	-0.44
Mean over all subjects									
	7.8	8.9	8.4	8.4	7.3	0.30			

- [Treat]
  - Deleting subjects with large counts in placebo group (135) and small counts in progabide group (232) will diminish the negative treatment effect
    $\Longrightarrow$ positive DFBETAS
  - Deleting subjects with small counts in placebo group and large counts in progabide group (225) will increase the negative treatment effect
    $\Longrightarrow$ negative DFBETAS
  - Subject 207 is complicated; due to the lage baseline value, this subject is responsible for the positive coefficient of [LbasTrt] with a DFBETAS of -0.71 (the coefficient becomes nearly 0)
- [V4]: Subjects 126, 135 and 227 have a large drop at visit 4, so that deleting them will diminish the negative coefficient of [V4]
  $\Longrightarrow$ positive DFBETAS
- σ: Deleting subjects with extreme counts, relative to baseline, (large: 135, 227, 225; small: 232) will decrease σ
  $\Longrightarrow$ negative DFBETAS

Estimation using adaptive quadrature

- Likelihood contribution for cluster j by Gaussian quadrature:

$$\ell_j(\boldsymbol{\beta},\sigma) = \int \underbrace{\phi(u_j;0,\sigma)\prod_i f(y_{ij} \mid u_j;\boldsymbol{\beta})}_{\propto \text{ posterior of } u_j} \mathrm{d}u_j \approx \sum_{r=1}^{R} W_r \prod_i f(y_{ij} \mid \sigma A_r;\boldsymbol{\beta})$$

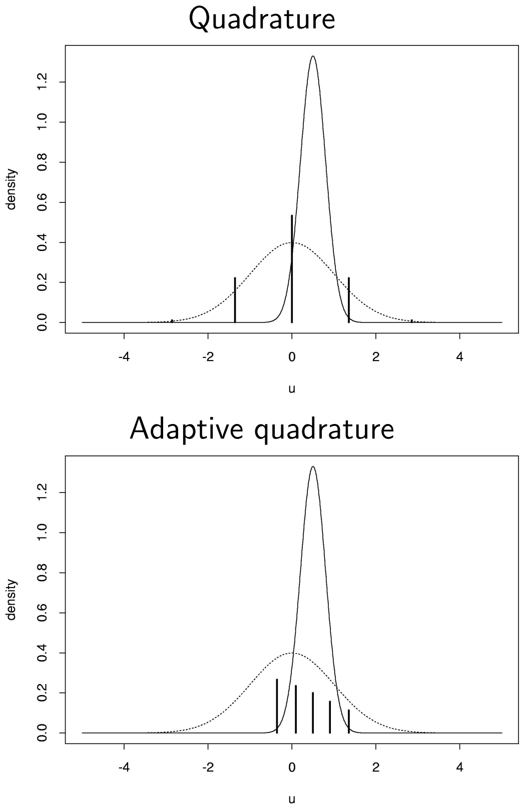
  - A_r: Quadrature locations
  - W_r: Quadrature weights

- Adaptive quadrature:

$$\ell_j(\boldsymbol{\beta},\sigma) \approx \sum_{r=1}^{R} \omega_{jr} \prod_i f(y_{ij} \mid \sigma\alpha_{jr};\boldsymbol{\beta})$$

  - α_{jr}: Adaptive quadrature location: $\widetilde{u}_j + \tau_j A_r$
    - $\widetilde{u}_j$: Posterior mean of u_j
      $\Longrightarrow$ Locations shifted to posterior mean $\approx$ peak of integrand
    - τ_j: Posterior standard deviation of u_j
      $\Longrightarrow$ Locations scaled by posterior sd $\approx$ width of peak
  - ω_{jr}: Adaptive quadrature weights: $\sqrt{2\pi}\tau_j \exp(A_r^2/2)\phi(\alpha_{jr})W_r$

Adaptive quadrature


Quadrature

Adaptive quadrature

Prior (dotted curve) and posterior (solid curve) densities

Empirical Bayes using adaptive quadrature

- Posterior mean and variance given $\boldsymbol{y}_j$ with $\widehat{\boldsymbol{\beta}}$ and $\widehat{\sigma}$ plugged in

$$
\begin{aligned}
\widetilde{u}_j = \mathrm{E}[u_j \mid \boldsymbol{y}_j, \mathbf{x}_j; \widehat{\boldsymbol{\beta}}, \widehat{\sigma}] &= \frac{\int u_j \phi(u_j; 0, \widehat{\sigma}) \prod_i f(y_{ij} \mid u_j; \widehat{\boldsymbol{\beta}}) \mathrm{d}u_j}{\ell_j(\widehat{\boldsymbol{\beta}}, \widehat{\sigma})} \\
\tau_j^2 = \mathrm{var}[u_j \mid \boldsymbol{y}_j, \mathbf{x}_j; \widehat{\boldsymbol{\beta}}, \widehat{\sigma}] &= \frac{\int u_j^2 \phi(u_j; 0, \widehat{\sigma}) \prod_i f(y_{ij} \mid u_j; \widehat{\boldsymbol{\beta}}) \mathrm{d}u_j}{\ell_j(\widehat{\boldsymbol{\beta}}, \widehat{\sigma})} - \widetilde{u}_j^2
\end{aligned}
$$

- Adaptive quadrature (in `gllamm`; similar to Naylor & Smith, 1988)
  - Start with $\widetilde{u}_j^0 = 0$ and $\tau_j^0 = 1$
  - In iteration k (between NR steps):

$$
\begin{aligned}
\ell_j(\widehat{\boldsymbol{\beta}}, \widehat{\sigma})^k &= \sum_{r=1}^{R} w_{jr}^{k-1} \prod_i f(y_{ij} \mid \widehat{\sigma}\alpha_{jr}^{k-1}; \widehat{\boldsymbol{\beta}}) \\
\widetilde{u}_j^k &= \frac{\sum_{r=1}^{R} (\widehat{\sigma}\alpha_{jr}^{k-1}) w_{jr}^{k-1} \prod_i f(y_{ij} \mid \widehat{\sigma}\alpha_{jr}^{k-1}; \widehat{\boldsymbol{\beta}})}{\ell_j(\widehat{\boldsymbol{\beta}}, \widehat{\sigma})^k} \\
(\tau_j^k)^2 &= \frac{\sum_{r=1}^{R} (\widehat{\sigma}\alpha_{jr}^{k-1})^2 w_{jr}^{k-1} \prod_i f(y_{ij} \mid \widehat{\sigma}\alpha_{jr}^{k-1}; \widehat{\boldsymbol{\beta}})}{\ell_j(\widehat{\boldsymbol{\beta}}, \widehat{\sigma})^k} - (\widetilde{u}_j^k)^2
\end{aligned}
$$

Variances for EB prediction & approximations

- Posterior variance (by numerical integration):

$$\mathrm{var}[u_j \mid \boldsymbol{y}_j, \mathbf{x}_j; \widehat{\boldsymbol{\theta}}]$$

- Marginal sampling variance:

$$\nu_j^2 \;\equiv\; \mathrm{var}_{\boldsymbol{y}}[\widetilde{u}_j^{\mathsf{EB}} \mid \mathbf{x}_j; \widehat{\boldsymbol{\theta}}] \;\approx\; \widehat{\sigma}^2 - \tau_j^2$$

'Diagnostic' variance

- Prediction error variance (marginal):

$$\mathrm{var}_{\boldsymbol{y}}[\widetilde{u}_j^{\mathsf{EB}} - u_j \mid \mathbf{x}_j; \widehat{\boldsymbol{\theta}}] \;\approx\; \tau_j^2$$

'Comparative' variance

Deletion residuals

- A large true residual will lead to a larger estimate of the random effects variance, making the residual appear more consistent with the model

- To avoid this problem, estimate EB residuals $\widetilde{u}_{j(-j)}$ using parameter estimates $\widehat{\boldsymbol{\theta}}_{(-j)}$ when the jth top-level cluster is deleted

$$\widetilde{u}_{j(-j)} \ = \ \mathrm{E}[u_j \mid \boldsymbol{y}_j, \mathbf{x}_j; \widehat{\boldsymbol{\theta}}_{(-j)}]$$

- Standardised deletion residual

$$\frac{\widetilde{u}_{j(-j)}}{\nu_{j(-j)}}$$

- In multilevel models, delete the top-level cluster to derive deletion residuals for all lower-level units in that cluster

EB prediction in gllamm

- Raw and standardised residuals:

```
gllapred res_, u       /* posterior mean and sd in res_m1 res_s1  */
gllapred stres_, ustd  /* stres_m1 =  $\widetilde{u}_j/\nu_j$  */
```

- Deletion residuals:

```
gllamm ... if subj~=126, i(subj) from(a) ...
gllapred dres if subj==126, u fsample       /* fsample to include 126 */
gllapred dstres if subj==126, ustd fsample
```

Level-2 residuals for epilepsy data

	DFBETAS			
Subj.	σ	$\frac{\widetilde{u}_{j(-j)}}{\nu_{j(-j)}}$	$\frac{\widetilde{u}_j}{\nu_j}$	$\widetilde{u}_j$
Placebo				
126	0.02	1.04	0.89	0.44
135	-0.34	2.23	1.97	0.93
206	-0.32	-2.11	-1.91	-0.88
227	-0.33	2.19	1.93	0.96
Progabide				
207	-0.16	1.97	1.37	0.69
112	-0.32	2.25	2.07	1.01
225	-0.46	2.47	2.26	1.09
232	-0.44	-2.92	-2.77	-0.97

Cross-validation by simulation

- Obtain sampling distribution of deletion statistic $S_{j(-j)}$ for cluster j under null hypothesis that the responses for cluster j come from the same distribution as for remaining clusters (Similar to Marshall & Spiegelhalter, 2001):
    - For cluster j, simulate new responses $\boldsymbol{y}_j^k$ from the model with parameters $\widehat{\boldsymbol{\theta}}_{(-j)}$
    - Obtain the statistic $S_{j(-j)}^k$ for the simulated responses

- Stata commands for simulating standardised deletion residuals under null hypothesis:

```
postfile file res using delres, replace
forvalues i=1/1000 {
   gllasim y1 if subj==126, fsample      /* simulate new responses  */
   replace y = y1 if subj==126
   gllapred b if subj==126, ustd fsample /* simulated std. del. res. */
   summ bm1
   post file (r(mean))
   drop y1 bm1 bs1
}
postclose file
```

- Obtain p-value using empirical sampling distribution

Cross-validation results

	Std. Deletion Residual $\frac{\widetilde{u}_{j(-j)}}{\nu_{j(-j)}}$				Del. Log-likelihood $\ell_{j(-j)}$			
			Power $\alpha = 0.05$				Power $\alpha = 0.05$	
Subj.	Obs.	p-value	$u_j = -1$	$u_j = 1$	Obs.	p-value	$u_j = -1$	$u_j = 1$
Placebo								
126	1.04	0.314	0.43	0.58	-19.1	0.005	0.00	0.55
135	2.23	0.026	0.26	0.47	-20.1	0.001	0.00	0.49
206	-2.11	0.058	0.33	0.44	-19.4	0.004	0.00	0.52
227	2.20	0.026	0.38	0.69	-39.9	0.001	0.01	0.63
Progabide								
207	1.98	0.068	0.50	0.40	-21.3	0.004	0.01	0.58
112	2.25	0.028	0.49	0.68	-13.8	0.043	0.00	0.63
225	2.47	0.020	0.35	0.46	-26.4	0.001	0.00	0.50
232	-2.92	0.002	0.25	0.57	-6.4	0.821	0.00	0.57

Conclusions

- Adaptive quadrature can be used to obtain reliable estimates and empirical Bayes predictions

- Cook's distances and DFBETAS are useful for identifying influential top-level clusters

- Standardized residuals (and their deletion counterparts) can flag potential outliers at any level

- Cross-validation is a useful method for testing for outliers/influential units at any level. This method is feasible for applications since the parameters do not need to be re-estimated in each simulation

- All diagnostics discussed, as well as simulations, are available in `gllamm` (from next update after 20 May 2003)

- `gllamm` can also be used to compute expected counts for categorical data. If there is a moderate number of response and covariate patters, these can be used to obtain the deviance, Pearson X^2 and various residuals

- `gllamm` can be downloaded from:

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

References to our work

- Generalized multilevel structural equation modeling. *Psychometrika*, in press. (S.Rabe-Hesketh, A.Skrondal & A.Pickles).

- *Generalized latent variable modeling: Multilevel, longitudinal and structural equation models*. Boca Raton, FL: Chapman & Hall/ CRC, to appear. (A.Skrondal & S.Rabe-Hesketh).

- *GLLAMM Manual*. London: Institute of Psychiatry, 2001.
  (S.Rabe-Hesketh, A.Pickles & A.Skrondal).

- Reliable estimation of generalized linear mixed models using adaptive quadrature. *The Stata Journal*, 2: 1-21, 2002.
  (S.Rabe-Hesketh, A.Skrondal & A.Pickles).

- Correcting for covariate measurement error in logistic regression using nonparametric maximum likelihood estimation. *Statistical Modelling*, in press. (S.Rabe-Hesketh, A.Pickles & A.Skrondal).

References

- Breslow & Clayton (1993). Approximate inference in generalized linear mixed models. JASA 88, 9-25.
- Marshall C. & Spiegelhalter D. (2001). Simulation-based tests for divergent behaviour in hierarchical models. Submitted.
- Naylor & Smith (1988). Econometric illustrations of novel numerical integration strategies for Bayesian inference. Journal of Econometrics 38, 103-125.
- Thall & Vail (1990). Some covariance models for longitudinal count data with overdispersion. Biometrics 46, 657-671.