GLMM

Emily Huang and Leo Collado

March 10th, 2014

1. GLMM
    - LMM
    - GLMM
    - Estimation

2. Quadrature points exploration
    - cbpp data
    - Toe-nail infection data

What are GLMM's?

- Extension of linear mixed models to allow response variables from different distributions
- Extension of GLMs to allow random effects and within-subject correlation

LMM in the GLMM Framework

1) Distributional assumption

- $Y_{ij}|b_i \sim Normal(E[Y_{ij}|b_i], Var(Y_{ij}|b_i))$
- $Var(Y_{ij}|b_i) = \sigma^2$
- $Cov(Y_{ij}, Y_{ik}|b_i) = 0$

2) Systematic component

- $\eta_{ij} = X_{ij}\beta + Z_{ij}b_i$
- Linear predictor incorporates both fixed effects and subject-specific effects.
- $b_i \sim N(0, G)$, b_i is independent of the covariates

3) Link function

- g is the identity link, so we have $g\{E[Y_{ij}|b_i]\} = E[Y_{ij}|b_i] = X_{ij}\beta + Z_{ij}b_i$

General Setup of GLMM

1) Distribution assumption

- $Y_{ij}|b_i \sim$ Exponential family
- $Var(Y_{ij}|b_i) = v\{E(Y_{ij}|b_i)\}\phi$, where v is a known function
- $Cov(Y_{ij}, Y_{ik}|b_i) = 0$

2) Systematic component

- $\eta_{ij} = X_{ij}\beta + Z_{ij}b_i$

3) Link function

- $g\{E(Y_{ij}|b_i)\} = \eta_{ij} = X_{ij}\beta + Z_{ij}b_i$ for some known link function, g

4) Random effects

- Assumed to have some probability distribution, such as $b_i \sim MVN(0, G)$
- b_i are assumed to be independent of the covariates

Example: Logistic model with random intercept

1) Distributional Assumption:

- Given b_i, Y_{ij} are independent and have a Bernoulli distribution
- $Var(Y_{ij}|b_i) = E(Y_{ij}|b_i)\{1 - E(Y_{ij}|b_i)\}$

2) Systematic Component

- $\eta_{ij} = \beta_0 + b_i + \beta_1 age_{ij}$

3) Link function

- $logit\{E[Y_{ij}|b_i]\} = \eta_{ij} = \beta_0 + b_i + \beta_1 age_{ij}$

4) Random effects

- The single random effect b_i is assumed to be $N(0, g_{11})$

Likelihood-based Estimation

- The joint distributions of both $Y_i|b_i$ and b_i are fully specified
- We can base estimation and inference on the likelihood function (ML estimation for β, ϕ, and G)
- Data can be missing at random (GEE required missing completely at random)

Likelihood Function

$$\begin{aligned} f(Y_i, b_i) &= f(Y_i|b_i)f(b_i) \\ &= f(Y_{i1}|b_i)f(Y_{i2}|b_i)...f(Y_{in_i}|b_i)f(b_i) = \prod_{j=1}^{n_i} f(Y_{ij}|b_i)f(b_i) \\ L(\beta, \phi, G) &= \prod_{i=1}^{N} \int f(Y_i, b_i)db_i \\ &= \prod_{i=1}^{N} \int \{\prod_{j=1}^{n_i} f(Y_{ij}|b_i)\}f(b_i)db_i \end{aligned}$$

- Since b_i is unobserved, inference about β, ϕ, and G is based on the integrated likelihood function $L(\beta, \phi, G)$.

Gauss-Hermite Quadrature

$$\int_{-\infty}^{\infty} h(v)e^{-v^2}dv \approx \sum_{k=1}^{d} h(x_k)w_k$$

- d quadrature points (weights, w_k, and evaluation points, x_k)
- The more quadrature points used, the more accurate the approximation
- But computational burden increases with quadrature points, and grows exponentially with the number of random effects

Example for Random Intercept GLMM

$$
\begin{aligned}
L(\beta,\phi,\sigma_b^2) &= \prod_{i=1}^{N}\int_{-\infty}^{\infty}\{\prod_{j=1}^{n_i} f(Y_{ij}|b_i)\}f(b_i)db_i \\
&= \prod_{i=1}^{N}\int_{-\infty}^{\infty} exp\left\{\sum_{j=1}^{n_i}\frac{Y_{ij}\theta_{ij}-b(\theta_{ij})}{a(\phi)}+\sum_{j=1}^{n_i}c(Y_{ij},\phi)\right\}\frac{1}{\sqrt{2\pi\sigma_b^2}}exp\left\{-\frac{b_i^2}{2\sigma_b^2}\right\}db_i \\
&= \prod_{i=1}^{N}\int_{-\infty}^{\infty} exp\left\{\sum_{j=1}^{n_i}\frac{Y_{ij}\theta_{ij}-b(\theta_{ij})}{a(\phi)}+\sum_{j=1}^{n_i}c(Y_{ij},\phi)\right\}\frac{\sqrt{2}\sigma_b}{\sqrt{2\pi\sigma_b^2}}exp\left\{-\nu_i^2\right\}d\nu_i \\
&= \prod_{i=1}^{N}\int_{-\infty}^{\infty} h(\upsilon_i)exp\{-\nu_i^2\}d\nu_i \approx \prod_{i=1}^{N}\sum_{k=1}^{d}h(x_k)w_k
\end{aligned}
$$

Report

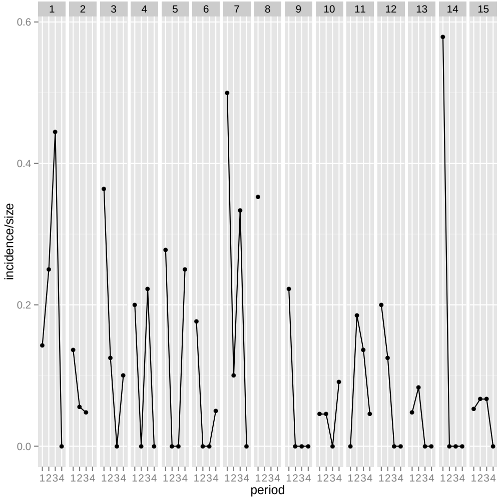
For more details and code check the report online at

`http://lcolladotor.github.io/BiostatLab5/`

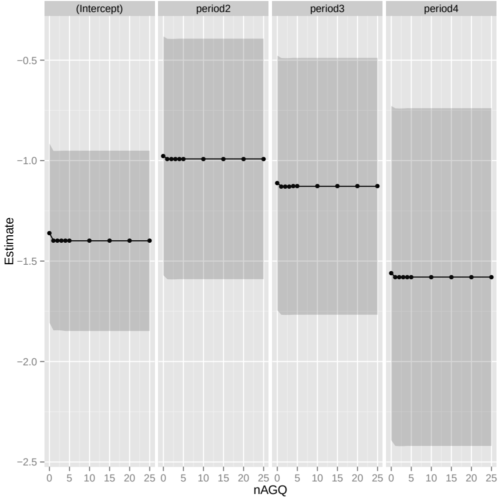
Data description

Contagious bovine pleuropneumonia (CBPP) is a major disease of cattle in Africa, caused by a mycoplasma. This dataset describes the serological incidence of CBPP in zebu cattle during a follow-up survey implemented in 15 commercial herds located in the Boji district of Ethiopia. The goal of the survey was to study the within-herd spread of CBPP in newly infected herds. Blood samples were quarterly collected from all animals of these herds to determine their CBPP status. These data were used to compute the serological incidence of CBPP (new cases occurring during a given time period). Some data are missing (lost to follow-up).

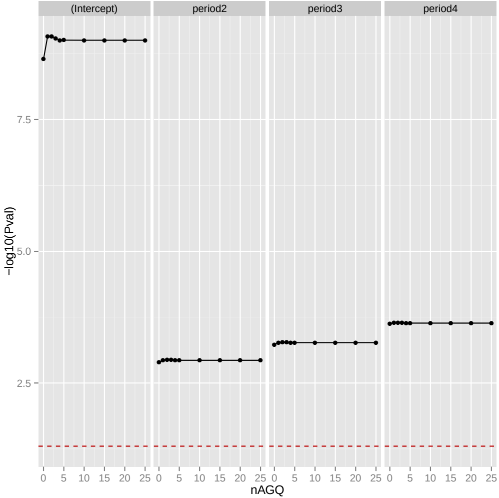
EDA


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
0.6
0.4
0.2
0.0
incidence/size
1234
period

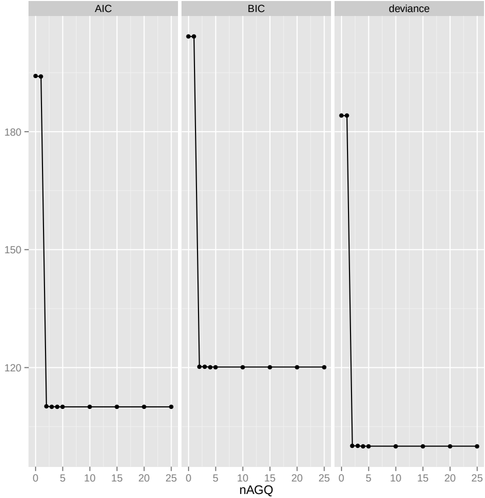
Estimates vs nAGQ



P-values vs nAGQ



Log Like vs nAGQ



Data description

The data from (Lesaffre & Spiessens, 2001) is available online. The researchers were interested in the degree of onycholysis which is related to the degree of separation of the nail plate from the nail-bed.

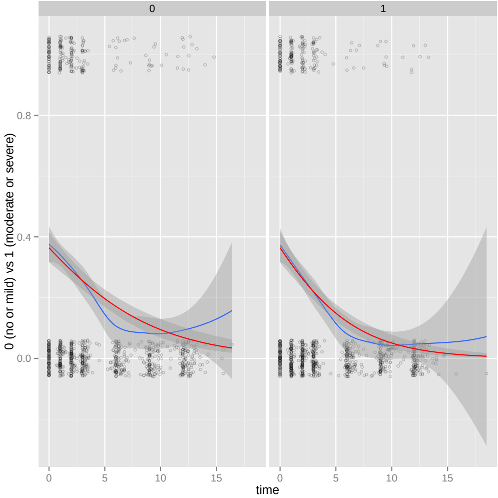
Model

$$\text{logit}\{P(Y_{ij} = 1 | b_i, \beta)\} = \beta_0 + \beta_1 \text{treatment}_i + \beta_2 t_{ij} + \beta_3 t_{ij} \times \text{treatment}_i + \beta_i$$

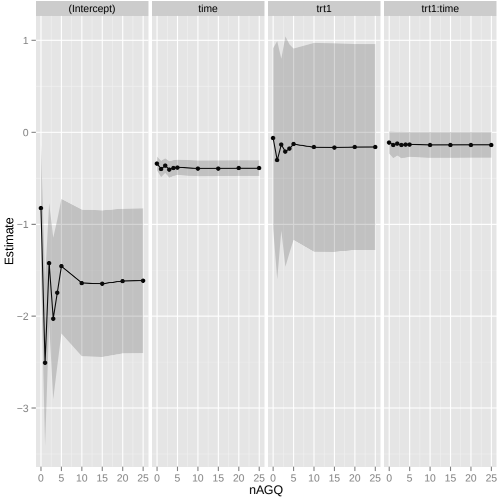
where

- Y_{ij} is the binary response at the jth visit of the ith subject
- $i = 1, \ldots, N$
- $j = 1, \ldots, n_i$
- $b_i = \sigma z_i$
- $z_i \sim N(0, 1)$

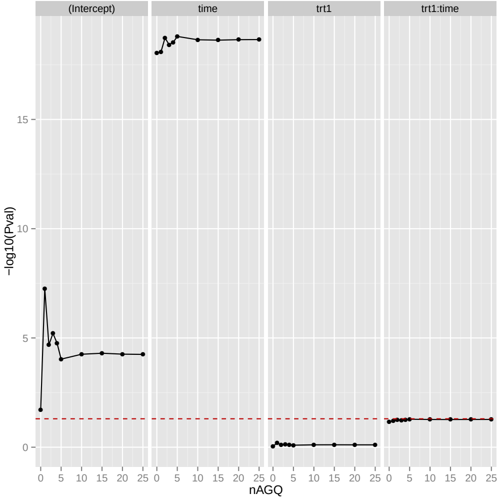
EDA


0
1
0 (no or mild) vs 1 (moderate or severe)
0.8
0.4
0.0
0
5
10
15
0
5
10
15
time

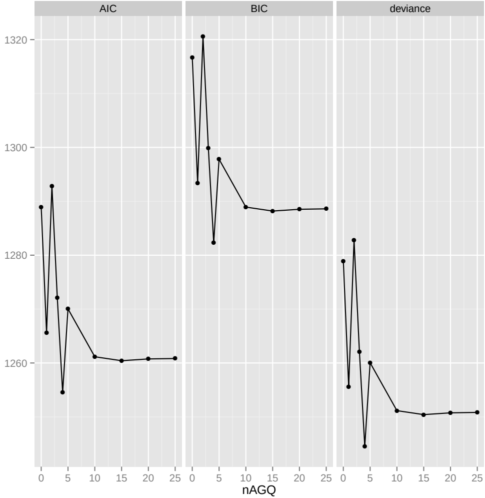
Estimates vs nAGQ



P-values vs nAGQ



Log Like vs nAGQ



Conclusions

- Quadrature points affect estimates, p-values, log likelihood calculations
- Do not assume that the default number of quadrature points will work with your data!
- Using adaptive seems better than non-adaptive

Thank you!