

CHAPTER 9: GENERAL REGRESSION MODELS

We begin by considering the class of *non-linear mixed effects models* (NLMEMs) before turning to *Generalized Linear Mixed Models* (GLMMs).

In this chapter we will again consider both a *conditional* approach to modeling, via the introduction of random effects, and a *marginal* approach using GEEs. Likelihood and Bayesian methods will be used for inference in the conditional approach.

Non-linear Mixed Effects Models

Example: Pharmacokinetics of Indomethacin

Six human volunteers received bolus intravenous doses (of the same size) of Indomethacin, and subsequently 11 blood samples were taken, and the drug concentrations recorded.

Figure 27 shows the concentration-time data – the curves follow a similar pattern but there is clearly person to person variability.

210

2006 Jon Wakefield, Stat/Biostat 571

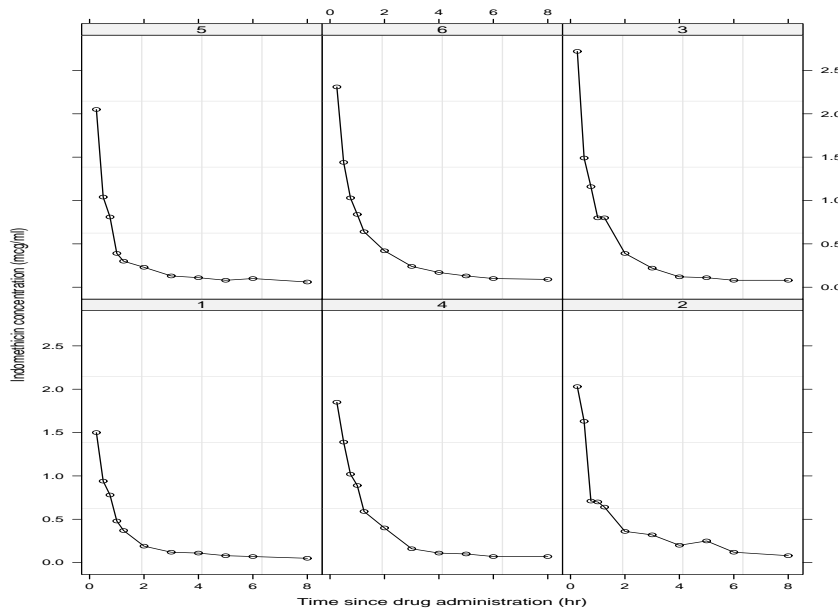


Figure 27: Concentration time data for Indomethacin.

211

Non-Linear Mixed Effects Models

Consider the two-stage model:

Stage 1: Response model, *conditional* on random effects:

$$\mathbf{y}_i = \mathbf{f}_i(\boldsymbol{\beta}, \mathbf{b}_i, \mathbf{x}_{ij}) + \boldsymbol{\epsilon}_i, \quad (43)$$

where $\mathbf{f}_i = [f_{i1}, \dots, f_{in_i}]^T$ are a set of functions that are non-linear in the parameters $\boldsymbol{\beta}$ and \mathbf{b}_i , and $\boldsymbol{\epsilon}_i$ is an $n_i \times 1$ zero mean vector of error terms.

Stage 2: Model for random terms:

$$\begin{aligned} \mathbf{E}[\boldsymbol{\epsilon}_i] &= \mathbf{0}, \quad \text{var}(\boldsymbol{\epsilon}_i) = \mathbf{E}_i(\boldsymbol{\alpha}), \\ \mathbf{E}[\mathbf{b}_i] &= \mathbf{0}, \quad \text{var}(\mathbf{b}_i) = \mathbf{D}(\boldsymbol{\alpha}), \\ \text{cov}(\mathbf{b}_i, \boldsymbol{\epsilon}_i) &= \mathbf{0} \end{aligned}$$

where $\boldsymbol{\alpha}$ is the vector of variance-covariance parameters.

A common model assumes

$$\boldsymbol{\epsilon}_i \sim_{ind} N(\mathbf{0}, \sigma_\epsilon^2 \mathbf{I}_{n_i}), \quad \mathbf{b}_i \sim_{iid} N(\mathbf{0}, \mathbf{D}),$$

212

A particular form that covers a lot of longitudinal situations is to assume $f_i(\eta_{ij}, t_{ij})$ where

$$\eta_{ij} = \mathbf{x}_{ij}\boldsymbol{\beta} + \mathbf{z}_{ij}\mathbf{b}_i,$$

where

- a $(k + 1) \times 1$ vector of fixed effects, $\boldsymbol{\beta}$,
- a $(q + 1) \times 1$ vector of random effects, \mathbf{b}_i , with $q \leq k$.
- $\mathbf{x}_i = (\mathbf{x}_{i1}, \dots, \mathbf{x}_{in_i})^T$, the design matrix for the fixed effect with $\mathbf{x}_{ij} = (1, x_{ij1}, \dots, x_{ijk})^T$, and
- $\mathbf{z}_i = (\mathbf{z}_{i1}, \dots, \mathbf{z}_{in_i})^T$, and design matrix for the random effects with $\mathbf{z}_{ij} = (1, z_{ij1}, \dots, z_{ijq})^T$.

Let $\boldsymbol{\alpha}$ represent σ_ϵ^2 and the parameters of \mathbf{D} and $N = \sum_i n_i$.

Example: Pharmacokinetics of Indomethacin

Let Y_{ij} represent the concentration of drug on individual i at time t_{ij} , $i = 1, \dots, 6$, $j = 1, \dots, 11$. The compartmental model that has previously been used for this drug is the two-compartment bi-exponential model:

$$E[Y_{ij} | \beta] = A_{1i} \exp\{-\alpha_{1i}t_{ij}\} + A_{2i} \exp\{-\alpha_{2i}t_{ij}\},$$

where Y_{ij} is concentration and $A_{1i}, A_{2i}, \alpha_{1i}, \alpha_{2i} > 0$.

An obvious NLMEM would take

$$\log A_{1i} = \beta_1 + b_{1i}$$

$$\log A_{2i} = \beta_2 + b_{2i}$$

$$\log \alpha_{1i} = \beta_3 + b_{3i}$$

$$\log \alpha_{2i} = \beta_4 + b_{4i}$$

with $\mathbf{b}_i = [b_{1i}, b_{2i}, b_{3i}, b_{4i}]^T \sim_{iid} N_4(\mathbf{0}, \mathbf{D})$.

214

Likelihood Inference

See Pinheiro and Bates (2000, Chapter 7).

The likelihood is, as usual, obtained by integrating out the random effects:

$$\begin{aligned} L(\boldsymbol{\beta}, \boldsymbol{\alpha}) &= (2\pi\sigma_\epsilon^2)^{-N/2} (2\pi)^{-m/2} |\mathbf{D}|^{-m/2} \\ &\times \prod_{i=1}^m \int \exp \left[-\frac{(\mathbf{y}_i - \mathbf{f}_i)^T (\mathbf{y}_i - \mathbf{f}_i)}{2\sigma_\epsilon^2} - \frac{\mathbf{b}_i^T \mathbf{D}^{-1} \mathbf{b}_i}{2} \right] d\mathbf{b}_i. \end{aligned}$$

where \mathbf{f}_i is made up of terms $f(\eta_{ij}, t_{ij})$, $i = 1, \dots, m$, $j = 1, \dots, n_i$.

215

Difficulties

1. The first difficulty is how to calculate the required integrals, which for non-linear models are analytically intractable, recall for linear models they were available in closed form. Even the first two moments are not available in closed form in general:

$$\begin{aligned} E[Y_{ij} \mid \boldsymbol{\beta}, \boldsymbol{\alpha}] &= E_{b_i|D}[f(\boldsymbol{\beta}, \mathbf{b}_i, \mathbf{x}_{ij})] \neq f(\boldsymbol{\beta}, \mathbf{0}, \mathbf{x}_{ij}) \\ \text{var}(Y_{ij} \mid \boldsymbol{\beta}, \boldsymbol{\alpha}) &= \sigma_\epsilon^2 + \text{var}_{b_i|D}[f(\boldsymbol{\beta}, \mathbf{b}_i, \mathbf{x}_{ij})] \\ \text{cov}(Y_{ij}, Y_{i'j'} \mid \boldsymbol{\beta}, \boldsymbol{\alpha}) &= \text{cov}_{b_i|D}(f(\boldsymbol{\beta}, \mathbf{b}_i, \mathbf{x}_{ij}), f(\boldsymbol{\beta}, \mathbf{b}_i, \mathbf{x}_{i'j'})) \\ \text{cov}(Y_{ij}, Y_{i'j'} \mid \boldsymbol{\beta}, \boldsymbol{\alpha}) &= 0, \quad i \neq i' \end{aligned}$$

Note that

$$E_{b_i|D}[f(\boldsymbol{\beta}, \mathbf{b}_i, \mathbf{x}_{ij})] \neq f(\boldsymbol{\beta}, \mathbf{0}, \mathbf{x}_{ij})$$

we had equality for the linear model.

The data do not have a known marginal distribution.

2. How do we then maximize the resultant likelihood? For the linear model we used EM or Newton-Raphson algorithms.

216

Overview of Integration Techniques

We describe a number of generic integration techniques, in particular:

- Laplace approximation (an analytical approximation).
- Quadrature (numerical integration).
- Importance sampling (a Monte Carlo method).

Before the MCMC revolution these techniques were used in a Bayesian context.

217

Laplace Approximation

Let

$$I = \int \exp\{ng(\theta)\}d\theta,$$

denote a generic integral of interest and suppose m is the maximum of $g(\cdot)$.

We have

$$ng(\theta) = n \sum_{k=0}^{\infty} \frac{(\theta - m)^k}{k!} g^{(k)}(m),$$

where $g^{(k)}(m)$ represents the k -th derivative of g evaluated at m . Hence

$$\begin{aligned} I &= \int \exp \left\{ n \sum_{k=0}^{\infty} \frac{(\theta - m)^k}{k!} g^{(k)}(m) \right\} d\theta \\ &\approx e^{ng(m)} \int \exp \left\{ \frac{(\theta - m)^2}{2/[ng^{(2)}(m)]} \right\} d\theta \\ &= e^{ng(m)} (2\pi v)^{1/2} n^{-1/2} \end{aligned}$$

where $v = -1/[g^{(2)}(m)]$, and we have ignored terms in cubics or greater in the Taylor series.

218

Laplace Approximation in the NLMEM

See Pinheiro and Bates, Chapter 7.

We wish to evaluate

$$p(\mathbf{y}_i | \boldsymbol{\beta}, \boldsymbol{\alpha}) = (2\pi\sigma^2)^{-n_i/2} (2\pi)^{-(q+1)/2} |\mathbf{D}|^{-1/2} \int \exp\{n_i g(\mathbf{b}_i)\} d\mathbf{b}_i,$$

where

$$-2n_i g(\mathbf{b}_i) = [\mathbf{y}_i - \mathbf{f}_i(\boldsymbol{\beta}, \mathbf{b}_i, \mathbf{x}_i)]^T [\mathbf{y}_i - \mathbf{f}_i(\boldsymbol{\beta}, \mathbf{b}_i, \mathbf{x}_i)] / \sigma_\epsilon^2 + \mathbf{b}_i^T \mathbf{D}^{-1} \mathbf{b}_i.$$

A Laplace approximation is a second-order Taylor series expansion of g about

$$\hat{\mathbf{b}}_i = \arg \min_{\mathbf{b}_i} -g(\mathbf{b}_i)$$

which will not be available in closed form for a non-linear model.

219

Gaussian Quadrature

A general method of integration is provided by quadrature (numerical integration) in which an integral

$$I = \int f(u) \, du,$$

is approximated by

$$\hat{I} = \sum_{i=1}^{n_w} f(u_i)w_i,$$

for design points u_1, \dots, u_{n_w} and weights w_1, \dots, w_{n_w} . Different choices of (u_i, w_i) lead to different integration rules.

In mixed model applications we have integrals with respect to a normal density, *Gauss-Hermite* quadrature is designed for problems of this type.

Specifically, it provides exact integration of

$$\int_{-\infty}^{\infty} g(u)e^{-u^2} \, du,$$

where $g(\cdot)$ is a polynomial of degree $2n_w - 1$.

220

The design points are the zeroes of the so-called Hermite polynomials.

Specifically, for a rule of n_w points, u_i is the i -th zero of $H_{n_w}(u)$, the Hermite polynomial of degree n_w , and

$$w_i = \frac{w^{n_w-1} n_w! \sqrt{\pi}}{n_w^2 [H_{n_w-1}(u_i)]^2}.$$

Now suppose $\boldsymbol{\theta}$ is two-dimensional and we wish to evaluate

$$I = \int f(\boldsymbol{\theta}) d\boldsymbol{\theta} = \int \int f(\theta_1, \theta_2) d\theta_2 d\theta_1 = \int f^*(\theta_1) d\theta_1,$$

where

$$f^*(\theta_1) = \int f(\theta_1, \theta_2) d\theta_2.$$

Now form

$$\hat{I} = \sum_{i=1}^{m_1} w_i \hat{f}^*(\theta_{1i}),$$

where

$$\hat{f}^*(\theta_{1i}) = \sum_{j=1}^{m_2} u_j f(\theta_{1i}, \theta_{2j}).$$

221

Then we have

$$\hat{I} = \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} w_i u_j f(\theta_{1i}, \theta_{2j}),$$

which is known as the *Cartesian Product*.

Scaling and reparameterization

To implement this method the function must be centered and scaled in some way, for example we could center and scale by the current estimates of the mean, \mathbf{m} , and variance-covariance matrix, \mathbf{V} – known as adaptive quadrature.

We then form

$$\mathbf{X} = \mathbf{L}(\boldsymbol{\theta} - \mathbf{m})$$

where $\mathbf{L}'\mathbf{L} = \mathbf{V}^{-1}$ and carry out integration in the space of \mathbf{X} .

There is no guarantee that the most efficient rule is obtained by scaling in terms of the posterior mean and variance, but we note that the ‘best’ normal approximation to a density (in terms of Kullback-Leibler divergence) has the same mean and variance.

222

Gauss-Hermite Code in R

Nodes and weights for $n = 4$:

```
> n <- 4
> quad <- gauss.quad(n,kind="hermite")
> quad$nodes
[1] -1.6506801 -0.5246476  0.5246476  1.6506801
> quad$weights
[1] 0.08131284 0.80491409 0.80491409 0.08131284
```

Nodes and weights for $n = 5$:

```
> n <- 5
> quad <- gauss.quad(n,kind="hermite")
> quad$nodes
[1] -2.0201829 -0.9585725  0.0000000  0.9585725  2.0201829
> quad$weights
[1] 0.01995324 0.39361932 0.94530872 0.39361932 0.01995324
```

223

Importance Sampling

Rather than deterministically selecting points we may randomly generate points from some density $h(\boldsymbol{\theta})$.

We have

$$I = \int f(\boldsymbol{\theta})d\boldsymbol{\theta} = \int \frac{f(\boldsymbol{\theta})}{h(\boldsymbol{\theta})}h(\boldsymbol{\theta})d\boldsymbol{\theta} = E[w(\boldsymbol{\theta})],$$

where $w(\boldsymbol{\theta}) = f(\boldsymbol{\theta})/h(\boldsymbol{\theta})$.

Hence we have the obvious estimator

$$\hat{I} = \sum_{i=1}^m w(\boldsymbol{\theta}_i),$$

where $\boldsymbol{\theta}_i \sim_{iid} h(\cdot)$. We have $E[\hat{I}] = I$ and

$$V = \text{var}(\hat{I}) = \frac{1}{m} \text{var}\{w(\boldsymbol{\theta})\}.$$

From this expression it is clear that a good $h(\cdot)$ produces an approximately constant $w(\boldsymbol{\theta})$.

224

We may estimate V via

$$\hat{V} = \frac{1}{m} \sum_{i=1}^m \frac{f^2(\boldsymbol{\theta}_i)}{h^2(\boldsymbol{\theta}_i)} - \frac{1}{m} \hat{I}^2,$$

and (appealing to the central limit theorem) \hat{I} is asymptotically normal and so a $100(1 - \alpha)\%$ confidence interval is given by

$$\hat{I} \pm Z_{\alpha/2} \hat{V}^{1/2}$$

where $Z_{\alpha/2}$ is the $\alpha/2$ point of an $N(0, 1)$ random variable.

Hence the accuracy of the approximation may be directly assessed, providing an advantage over analytical approximations and quadrature methods.

Notes on Importance Sampling

- We require an $h(\cdot)$ with heavier tails than the integrand. We can carry out importance sampling with any h but if the tails are lighter we will have an estimator with infinite variance (and hence an inconsistent procedure). Many suggestions for h have been made including Student t distributions and mixtures of Student t distributions.
- Iteration may again be used to obtain an estimator with good properties.

225

Notes on Implementation

- If the number of parameters is small then numerical integration techniques (e.g. quadrature) are highly efficient in terms of the number of function evaluations required. Hence if, for example, obtaining a point on the likelihood surface is computationally expensive (as occurs if a large simulation is required) then such techniques are preferable to Monte Carlo methods.
- The method employed will depend on whether it is for a one-off application, in which case ease-of-implementation is a consideration, or for a great deal of use, in which case an efficient method may be required.
- In general it is difficult to assess the accuracy of Laplace/numerical integration techniques.
- For simulation methods we note that independent samples are ideal for assessing Monte Carlo error since standard errors on expectations of interest may be simply calculated.
- Evans and Swartz (1995, *Statistical Science*) provide a good review of integration techniques.

226

The nlme algorithm

Within `nlme` an algorithm, introduced by Lindstrom and Bates (1990) is used.

The algorithm alternates between two steps:

Penalized Non-linear Least Squares (PNLS)

Condition on the current estimates of $\hat{\mathbf{D}}$ and $\hat{\sigma}_\epsilon^2$ and then minimize

$$\frac{1}{\hat{\sigma}_\epsilon^2} \sum_{i=1}^m (\mathbf{y}_i - \mathbf{f}_i)^\top (\mathbf{y}_i - \mathbf{f}_i) + \mathbf{b}_i \hat{\mathbf{D}}^{-1} \mathbf{b}_i,$$

to obtain estimates $\hat{\boldsymbol{\beta}}, \hat{\mathbf{b}}_1, \dots, \hat{\mathbf{b}}_m$, which may be viewed as finding the posterior mode for $\boldsymbol{\beta}$ and $\mathbf{b}_1, \dots, \mathbf{b}_m$.

Linear Mixed Effects (LME)

Carry out a first-order Taylor series of \mathbf{f}_i about $\widehat{\boldsymbol{\beta}}, \widehat{\mathbf{b}}_i$.

This results in a linear mixed effects model which can be maximized to obtain estimates of \mathbf{D} and σ_ϵ^2 .

We have likelihood

$$L(\boldsymbol{\beta}, \boldsymbol{\alpha}) = |\mathbf{D}|^{-m/2} \sigma_\epsilon^{-N} \int \exp \left\{ -\frac{1}{2} \sum_{i=1}^m (\mathbf{y}_i - \mathbf{f}_i)^\top (\mathbf{y}_i - \mathbf{f}_i) - \mathbf{b}_i^\top \mathbf{D}^{-1} \mathbf{b}_i \right\} d\mathbf{b}_i$$

where $\mathbf{f}_i = \mathbf{f}(\boldsymbol{\beta}, \mathbf{b}_i, \mathbf{x}_i)$, $i = 1, \dots, m$.

Carry out a first-order Taylor series expansion of \mathbf{f}_i about the estimates, obtained in the PNLs step at iteration k , of $\boldsymbol{\beta}$ and \mathbf{b}_i , call these $\widehat{\boldsymbol{\beta}}^{(k)}$ and $\widehat{\mathbf{b}}_i^{(k)}$.

228

Specifically

$$\mathbf{f}_i(\boldsymbol{\beta}, \mathbf{b}_i) \approx \mathbf{f}_i(\widehat{\boldsymbol{\beta}}^{(k)}, \widehat{\mathbf{b}}_i^{(k)}) + \widehat{\boldsymbol{\alpha}}_i^{(k)} (\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}^{(k)}) + \widehat{\boldsymbol{\alpha}}_i^{(k)} (\mathbf{b}_i - \widehat{\mathbf{b}}_i^{(k)})$$

where

$$\begin{aligned} \widehat{\boldsymbol{\alpha}}_i^{(k)} &= \left. \frac{\partial \mathbf{f}_i}{\partial \boldsymbol{\beta}^\top} \right|_{\widehat{\boldsymbol{\beta}}^{(k)}, \widehat{\mathbf{b}}_i^{(k)}} \\ \widehat{\boldsymbol{\alpha}}_i^{(k)} &= \left. \frac{\partial \mathbf{f}_i}{\partial \mathbf{b}_i^\top} \right|_{\widehat{\boldsymbol{\beta}}^{(k)}, \widehat{\mathbf{b}}_i^{(k)}} \end{aligned}$$

This gives

$$\mathbf{y}_i - \mathbf{f}_i(\boldsymbol{\beta}, \mathbf{b}_i) \approx \mathbf{y}_i^{(k)} - \widehat{\boldsymbol{\alpha}}_i^{(k)} \boldsymbol{\beta} - \widehat{\boldsymbol{\alpha}}_i^{(k)} \mathbf{b}_i$$

where

$$\mathbf{y}_i^{(k)} = \mathbf{y}_i - \mathbf{f}_i(\widehat{\boldsymbol{\beta}}^{(k)}, \widehat{\mathbf{b}}_i^{(k)}) + \widehat{\boldsymbol{\alpha}}_i^{(k)} \widehat{\boldsymbol{\beta}}^{(k)} + \widehat{\boldsymbol{\alpha}}_i^{(k)} \widehat{\mathbf{b}}_i^{(k)}$$

229

The integral can now be evaluated in closed-form to give the log-likelihood

$$l(\boldsymbol{\alpha}) = -\frac{1}{2} \sum_{i=1}^m \log |\widehat{\mathbf{V}}_i| - \frac{1}{2} \sum_{i=1}^m (\mathbf{y}_i^{(k)} - \widehat{\mathbf{x}}_i^{(k)} \boldsymbol{\beta})^\top \widehat{\mathbf{V}}_i^{-1} (\mathbf{Y}_i - \widehat{\mathbf{x}}_i \boldsymbol{\beta})$$

where

$$\widehat{\mathbf{V}}_i = \widehat{\mathbf{z}}_i^{(k)} \mathbf{D} \widehat{\mathbf{z}}_i^{(k)\top} + \sigma_\epsilon^2 \mathbf{I}_i,$$

which may be maximized to give ML estimates. REML estimates are obtained by adding the term

$$-\frac{1}{2} \sum_{i=1}^m \log | \widehat{\mathbf{x}}_i^{(k)\top} \widehat{\mathbf{V}}_i(\boldsymbol{\alpha}) \widehat{\mathbf{x}}_i^{(k)} |$$

The Laplace approximation is generally more accurate than the LB algorithm, it is, however, more computationally expensive.

230

Asymptotic Inference

Under the LB algorithm, the asymptotic distribution of the REML estimator $\widehat{\boldsymbol{\beta}}$ is

$$\left(\sum_{i=1}^m \widehat{\mathbf{x}}_i^\top \widehat{\mathbf{V}}_i^{-1} \widehat{\mathbf{x}}_i \right)^{1/2} (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \rightarrow_d N_{p+1}(\mathbf{0}, \mathbf{I}_{p+1}),$$

where $\widehat{\mathbf{x}}_i = \widehat{\mathbf{x}}_i^{(k)}$ with k the final iteration, $i = 1, \dots, m$

Similarly, the asymptotic distribution of $\boldsymbol{\alpha}$ is based on the information as calculated from the linear approximation to the likelihood.

The LB estimator is inconsistent if the n_i 's are fixed and $m \rightarrow \infty$.

Empirical Bayes estimates for the random effects are available, but caution should be given to using these for checking assumptions since they are strongly influenced by the assumption of normality being correct. If n_i is large then this will be less of a problem.

231

Approaches for NLMEMs

Various other approaches to likelihood inference have been suggested, we briefly summarize.

In general we need to carry out m integrals of dimension $q + 1$ for each likelihood evaluation, so with large m and q this can be computationally expensive.

First-Order Approximation

Let $\boldsymbol{\beta}_i = \mathbf{x}_i\boldsymbol{\beta} + \mathbf{b}_i$, and then carry out a first-order Taylor series about $E[\mathbf{b}_i] = \mathbf{0}$ to give

$$\mathbf{y}_i = \mathbf{f}_i(\boldsymbol{\beta}_i) + \boldsymbol{\epsilon}_i \approx \mathbf{f}_i(\mathbf{x}_i\boldsymbol{\beta}) + \frac{\partial \mathbf{f}_i}{\partial \boldsymbol{\beta}_i} \frac{\partial \boldsymbol{\beta}_i}{\partial \mathbf{b}_i} \mathbf{b}_i + \boldsymbol{\epsilon}_i.$$

In contrast to the LB algorithm which considered an expansion about the subject-specific mean, the expansion here is about the population-averaged mean. The first-order estimator is inconsistent and has bias even if n_i and m go to infinity, see Demidenko (2004, Chapter 8)

Adaptive Gaussian quadrature may also be used.

232

Example: Pharmacokinetics of Indomethacin

The compartmental model that has previously been used for this drug is the two-compartment bi-exponential model:

$$E[Y] = A_1 \exp\{-\alpha_1 t\} + A_2 \exp\{-\alpha_2 t\},$$

where Y is concentration, and t is time, and $A_1, A_2, \alpha_1, \alpha_2 > 0$.

Note: this model is unidentifiable since the parameter set $(A_1, \alpha_1, A_2, \alpha_2)$ gives the same fitted curve (and hence likelihood) as the set $(A_2, \alpha_2, A_1, \alpha_1)$. If this is a practical problem for a particular dataset (say $\alpha_1 \approx \alpha_2$) then we may parameterize in terms of α_1 and $\alpha_2 - \alpha_1$.

Figure 28 gives the log concentrations versus time – such a plot can be useful for picking the number of exponentials (and modeling the log concentration can provide initial estimates). Certainly not linear in time so more than a single exponential needed.

233

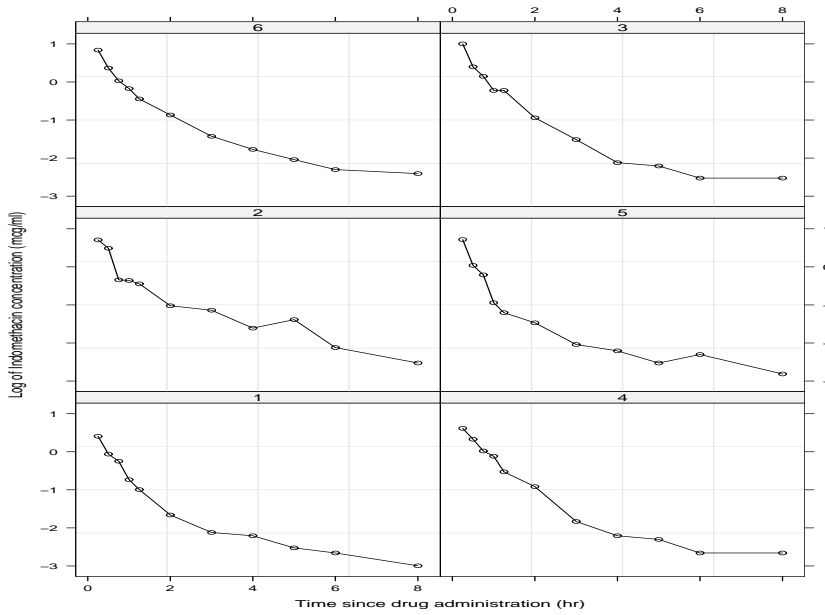


Figure 28: Log concentration time data for Indomethacin.

234

Individual fits

Let Y_{ij} be the drug concentration at time t_{ij} on individual i , $j = 1, \dots, 11$, $i = 1, \dots, 6$. We first fit bi-exponential models to each individual, using non-linear least squares.

We parameterize as

$$E[Y_{ij} | \beta_i] = \beta_{1i} \exp\{-e^{\beta_{3i}} t_{ij}\} + \beta_{2i} \exp\{-e^{\beta_{4i}} t_{ij}\},$$

for $i = 1, \dots, 6$.

Even though the data are balanced, the standard errors are different for different individuals, as we see in Figure 29.

R code for fitting individual models:

```
> indiv.lis <- nlsList( conc ~ SSbiexp(time,A1,lrc1,A2,lrc2),data=Indometh )
> indiv.lis
Call:
Model:conc~SSbiexp(time,A1,lrc1,A2,lrc2)|Subject
  Data: Indometh
Coefficients:
      A1      lrc1      A2      lrc2
1 2.029277 0.5793887 0.1915475 -1.7877849
4 2.198132 0.2423124 0.2545223 -1.6026859
2 2.827673 0.8013195 0.4989175 -1.6353512
5 3.566103 1.0407660 0.2914970 -1.5068522
6 3.002250 1.0882119 0.9685230 -0.8731358
3 5.468312 1.7497936 1.6757522 -0.4122004
Degrees of freedom: 66 total; 42 residual
Residual standard error: 0.0755502
> plot( intervals(indiv.lis) )
```

236

2006 Jon Wakefield, Stat/Biostat 571

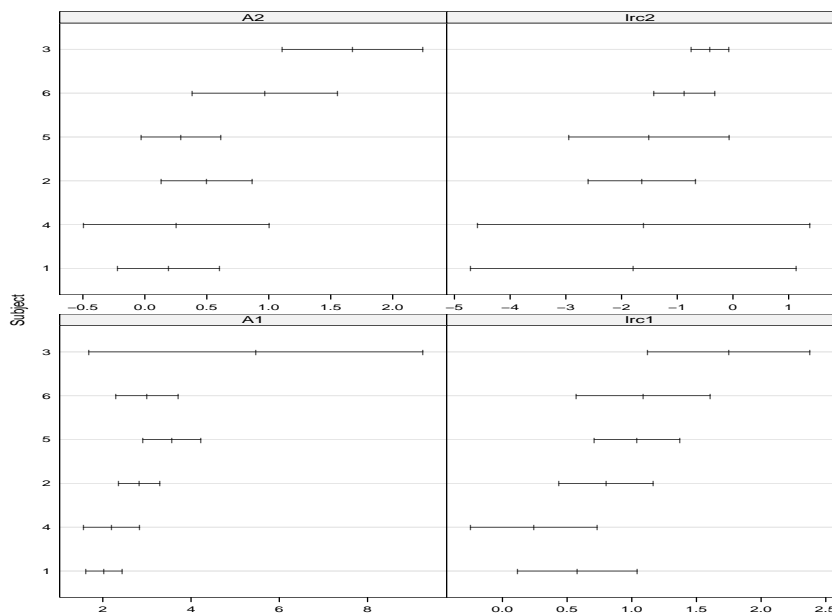


Figure 29: Asymptotic 95% CIs for elements of β_i , $i = 1, \dots, 6$.

237

Now we fit some NLMEMs, we first assume a diagonal D with random effects for first three elements only.

```
> nlme.indo <- nlme( indiv.lis,random=pdDiag(A1+lrc1+A2~1))
> summary(nlme.indo)
Nonlinear mixed-effects model fit by maximum likelihood
  Model: conc ~ SSbiexp(time, A1, lrc1, A2, lrc2)
Random effects:
  Formula: list(A1 ~ 1, lrc1 ~ 1, A2 ~ 1)
  Level: Subject
  Structure: Diagonal
           A1      lrc1      A2  Residual
StdDev: 0.57135 0.1581214 0.1115283 0.08149631
Fixed effects: list(A1 ~ 1, lrc1 ~ 1, A2 ~ 1, lrc2 ~ 1)
           Value Std.Error DF   t-value p-value
A1      2.8276029 0.2639744 57 10.711656 0e+00
lrc1    0.7732529 0.1100086 57  7.029021 0e+00
A2      0.4610197 0.1127560 57  4.088648 1e-04
lrc2   -1.3450041 0.2313139 57 -5.814627 0e+00
Correlation:
      A1      lrc1      A2
lrc1  0.055
A2    -0.102  0.630
lrc2 -0.139  0.577  0.834
```

238

Now assume a non-diagonal D for all four parameters.

```
> nlme2.indo2 <- update( nlme.indo, random=A1+lrc1+A2+lrc2~1)
> summary(nlme2.indo2)
  Model: conc ~ SSbiexp(time, A1, lrc1, A2, lrc2)
Random effects: Formula: list(A1 ~ 1, lrc1 ~ 1, A2 ~ 1, lrc2 ~ 1)
  Structure: General positive-definite, Log-Cholesky parametrization
           StdDev      Corr
A1      0.77583020 A1      lrc1  A2
lrc1    0.26863662 0.963
A2      0.38707000 0.459 0.682
lrc2    0.48253192 0.153 0.414 0.948
Residual 0.06962038
Fixed effects: list(A1 ~ 1, lrc1 ~ 1, A2 ~ 1, lrc2 ~ 1)
           Value Std.Error DF   t-value p-value
A1      2.8531611 0.3485825 57  8.185039 0e+00
lrc1    0.8755645 0.1253269 57  6.986245 0e+00
A2      0.6357872 0.1715520 57  3.706091 5e-04
lrc2   -1.2757709 0.2161119 57 -5.903288 0e+00
Correlation:
      A1      lrc1      A2
lrc1  0.907
A2    0.411  0.676
lrc2  0.108  0.378  0.912
```

239

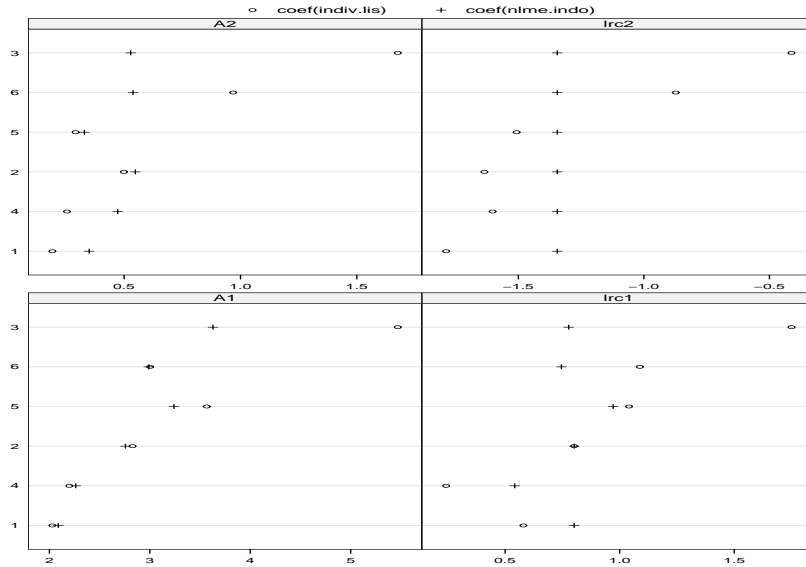


Figure 30: Comparison of non-linear LS and nlme estimates, with the latter from the model `nlme.indo` Created using the command `plot(compareFits(coef(indiv.lis),coef(nlme.indo)))`.

240

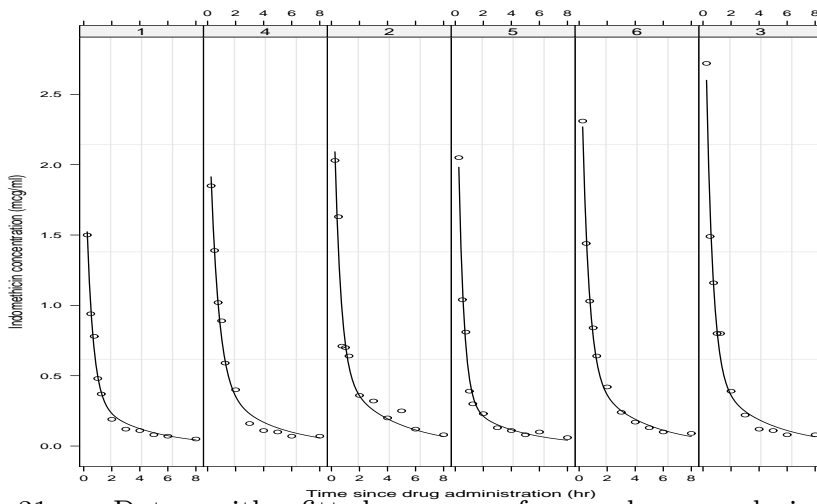


Figure 31: Data with fitted curves from nlme analysis superimposed from the model `nlme.indo`. Created with the command `plot(augPred(nlme.indo),aspect='xy',grid=T)`.

241

The following commands produced Figures 32–34.

```
> plot(nlme.indo, resid(., type="n")~fitted(.), id=0.05, adj=-1) # id=0.05 gives
# outliers outside of 95% of distn, adj=-1 adjusts the text which
# labels these outliers
> plot(nlme.indo, resid(., type="n")~time, id=0.05, adj=-1)
> qqnorm(nlme.indo)
> plot(augPred(nlme.indo, level=0:1)) # Obtain predictions at population and
# individual level of hierarchy
```

242

2006 Jon Wakefield, Stat/Biostat 571

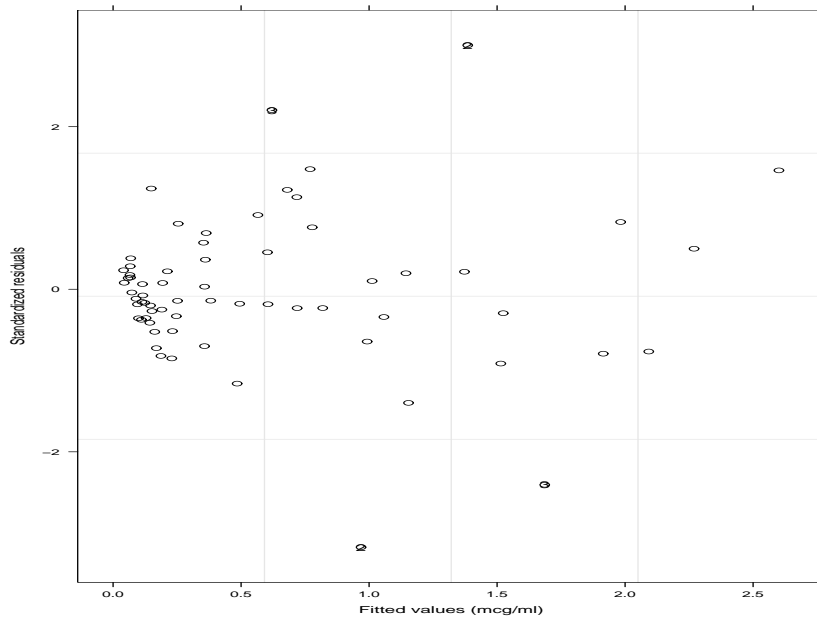


Figure 32: Standardized residuals versus fitted values.

243

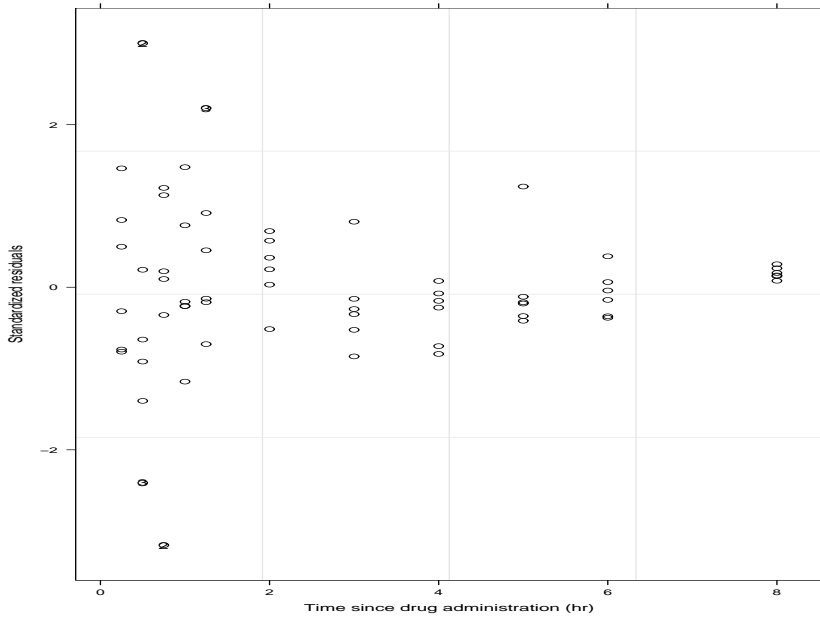


Figure 33: Standardized residuals versus time.

244

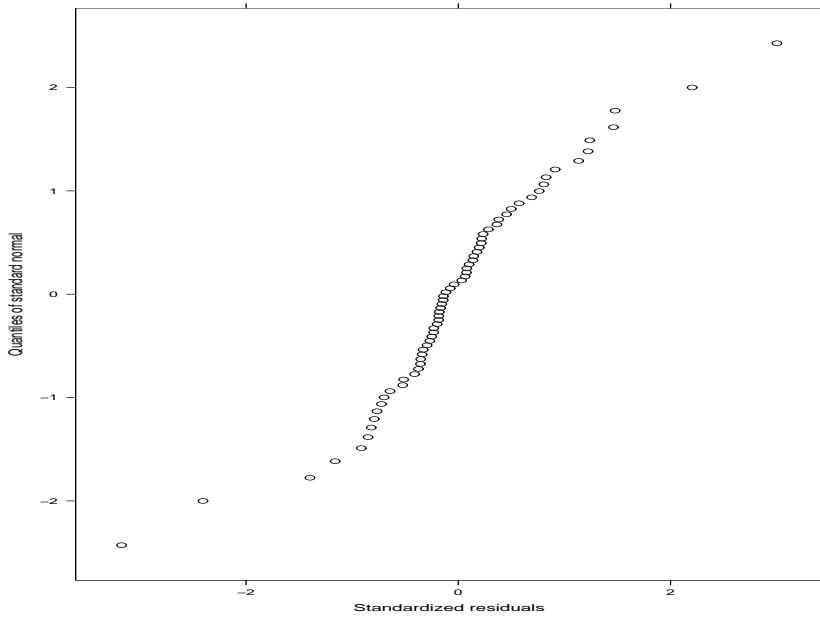


Figure 34: QQ plot of normalized residuals.

245

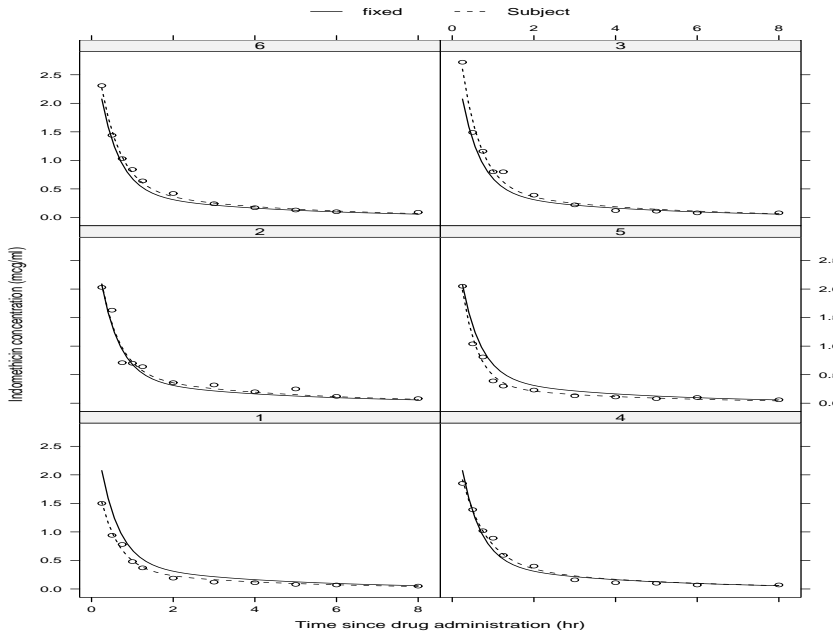


Figure 35: Solid lines are population predictions, dashed lines individual predictions.

246

Bayesian Approach

A Bayesian approach adds a prior distribution for β, α , to the likelihood $L(\beta, \alpha)$. As with the linear model proper prior is required for the matrix D . In general a proper prior is required for β also, to ensure the propriety of the posterior distribution. Closed-form inference is unavailable, but MCMC is almost as straightforward as in the LMEM case. The joint posterior is

$$p(\beta_1, \dots, \beta_m, \tau, \beta, \mathbf{W}, \mathbf{b} \mid \mathbf{y}) \propto \prod_{i=1}^m \{p(\mathbf{y}_i \mid \beta_i, \tau)p(\beta_i \mid \beta, \mathbf{W})\} \pi(\beta)\pi(\tau)\pi(\mathbf{W}).$$

Suppose we have priors:

$$\begin{aligned} \beta &\sim N_{q+1}(\beta_0, \mathbf{V}_0) \\ \tau &\sim \text{Ga}(a_0, b_0) \\ \mathbf{W} &\sim W_{q+1}(r, \mathbf{R}^{-1}) \end{aligned}$$

The conditional distributions for β, τ, \mathbf{W} are unchanged from the linear case. There is no closed form conditional distribution for β_i , which is given by:

$$p(\beta_i \mid \beta, \tau, \mathbf{W}, \mathbf{y}) \propto p(\mathbf{y}_i \mid \beta_i, \tau) \times p(\beta_i \mid \beta, \mathbf{W})$$

but a Metropolis-Hastings step can be used.

Generalized Estimating Equations

If interest lies in population parameters then we may use the estimator $\widehat{\boldsymbol{\beta}}$ that satisfies

$$\mathbf{G}(\boldsymbol{\beta}, \widehat{\boldsymbol{\alpha}}) = \sum_{i=1}^m \mathbf{D}_i^T \mathbf{W}_i^{-1} (\mathbf{Y}_i - \boldsymbol{\mu}_i) = \mathbf{0},$$

where $\mathbf{D}_i = \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}}$, $\mathbf{W}_i = \mathbf{W}_i(\boldsymbol{\beta}, \widehat{\boldsymbol{\alpha}})$ is the working covariance model, $\boldsymbol{\mu}_i = \boldsymbol{\mu}_i(\boldsymbol{\beta})$ and $\widehat{\boldsymbol{\alpha}}$ is a consistent estimator of $\boldsymbol{\alpha}$. Sandwich estimation may be used to obtain an empirical estimate of the variance, \mathbf{V}_β :

$$\left(\sum_{i=1}^m \mathbf{D}_i^T \mathbf{W}_i^{-1} \mathbf{D}_i \right)^{-1} \left\{ \sum_{i=1}^m \mathbf{D}_i^T \mathbf{W}_i^{-1} \text{cov}(\mathbf{Y}_i) \mathbf{W}_i^{-1} \mathbf{D}_i \right\} \left(\sum_{i=1}^m \mathbf{D}_i^T \mathbf{W}_i^{-1} \mathbf{D}_i \right)^{-1}.$$

We then have

$$\mathbf{V}_\beta^{-1/2} (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \rightarrow_d N(\mathbf{0}, \mathbf{I}).$$

In practice an empirical estimator of $\text{cov}(\mathbf{Y}_i)$ is substituted to give $\widehat{\mathbf{V}}_\beta$.

GEE has not been extensively used in a non-linear (non-GLM) setting. This is probably because in many settings (e.g. pharmacokinetic/pharmacodynamic) interest focuses on understanding between individual-variability, and explaining this in terms of individual-specific covariates.

248

Generalized Linear Mixed Models

A GLMM is defined by

1. *Random Component*: $Y_{ij} | \theta_{ij}, \alpha \sim p(\cdot)$ where $p(\cdot)$ is a member of the exponential family, that is

$$p(y_{ij} | \theta_{ij}, \alpha) = \exp[\{y_{ij} \theta_{ij} - b(\theta_{ij})\}] / a(\alpha) + c(y_{ij}, \alpha),$$

for $i = 1, \dots, m$ units, and $j = 1, \dots, n_i$, measurements per unit.

2. *Systematic Component*: If $\mu_{ij} = E[Y_{ij} | \theta_{ij}, \alpha]$ then we have a link function $g(\cdot)$, with

$$g(\mu_{ij}) = \mathbf{x}_{ij} \boldsymbol{\beta} + \mathbf{z}_{ij} \mathbf{b}_i,$$

so that we have introduced random effects into the linear predictor. The above defines the *conditional* part of the model. The random effects are then assigned a distribution, and in a GLMM this is assumed to be

$$\mathbf{b}_i \sim_{iid} N(\mathbf{0}, \mathbf{D}).$$

We also have

$$\text{var}(Y_{ij} | \theta_{ij}, \alpha) = \alpha v(\mu_{ij}).$$

249

Marginal Moments

Mean:

$$\begin{aligned} E[Y_{ij}] &= E\{E[Y_{ij}|\mathbf{b}_i]\} \\ &= E[\mu_{ij}] = E_b[g^{-1}(\mathbf{x}_{ij}\boldsymbol{\beta} + \mathbf{z}_{ij}\mathbf{b}_i)]. \end{aligned}$$

Variance:

$$\begin{aligned} \text{var}(Y_{ij}) &= E[\text{var}(Y_{ij}|\mathbf{b}_i)] + \text{var}(E[Y_{ij}|\mathbf{b}_i]) \\ &= \alpha E_b[v\{g^{-1}(\mathbf{x}_{ij}\boldsymbol{\beta} + \mathbf{z}_{ij}\mathbf{b}_i)\}] + \text{var}_b[g^{-1}(\mathbf{x}_{ij}\boldsymbol{\beta} + \mathbf{z}_{ij}\mathbf{b}_i)]. \end{aligned}$$

Covariance:

$$\begin{aligned} \text{cov}(Y_{ij}, Y_{ik}) &= E[\text{cov}(Y_{ij}, Y_{ik}|\mathbf{b}_i)] + \text{cov}(E[Y_{ij}|\mathbf{b}_i], E[Y_{ik}|\mathbf{b}_i]) \\ &= \text{cov}\{g^{-1}(\mathbf{x}_{ij}\boldsymbol{\beta} + \mathbf{z}_{ij}\mathbf{b}_i), g^{-1}(\mathbf{x}_{ik}\boldsymbol{\beta} + \mathbf{z}_{ik}\mathbf{b}_i)\} \\ &\neq 0, \end{aligned}$$

for $j \neq k$ due to shared random effects, and

$$\text{cov}(Y_{ij}, Y_{lk}) = 0,$$

for $i \neq l$, as there are no shared random effects.

250

Example: Log-Linear Regression for Seizure Data

Data on seizures were collected on 59 epileptics. For each patient the number of epileptic seizures were recorded during a baseline period of eight weeks, after which patients were randomized to treatment with the anti-epileptic drug progabide, or to placebo. The number of seizures was then recorded in four consecutive two-week periods. The age of the patient was also available.

Let

$$\begin{aligned} Y_{ij} &= \text{number of seizures on patient } i \text{ at occasion } j \\ t_{ij} &= \text{observation period on patient } i \text{ at occasion } j \\ x_{i1} &= 0/1 \text{ if patient } i \text{ was assigned placebo/progabide} \\ x_{ij2} &= 0/1 \text{ if } j = 0/1, 2, 3, 4 \end{aligned}$$

with $t_{ij} = 8$ if $j = 0$ and $t_{ij} = 2$ if $j = 1, 2, 3, 4$, $i = 1, \dots, 59$.

The question of primary scientific interest here is whether progabide reduces the number of seizures.

A marginal mean model is given by

$$E[Y_{ij}] = t_{ij} \exp(\beta_0 + \beta_1 x_{ij1} + \beta_2 x_{ij2} + \beta_3 x_{ij1} x_{ij2})$$

251

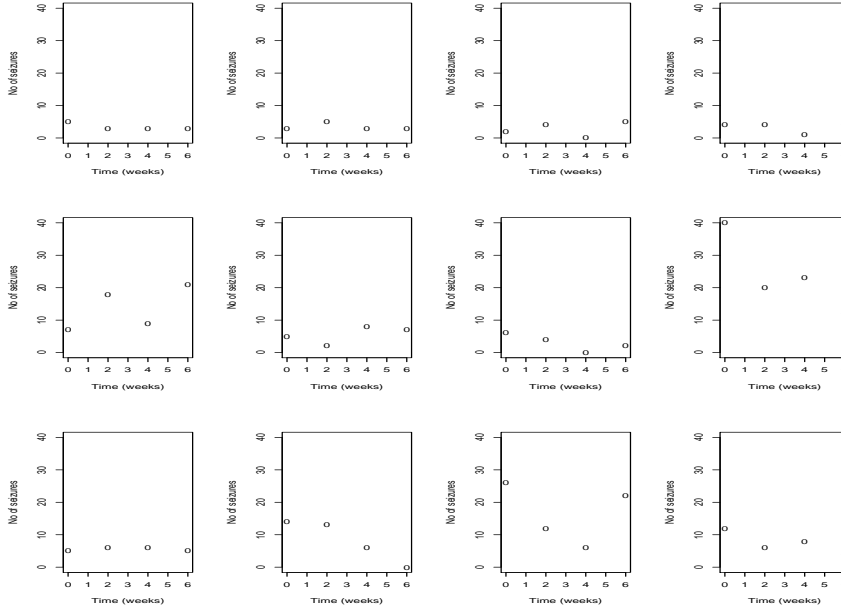


Figure 36: Number of seizures for selected individuals over time for placebo group.

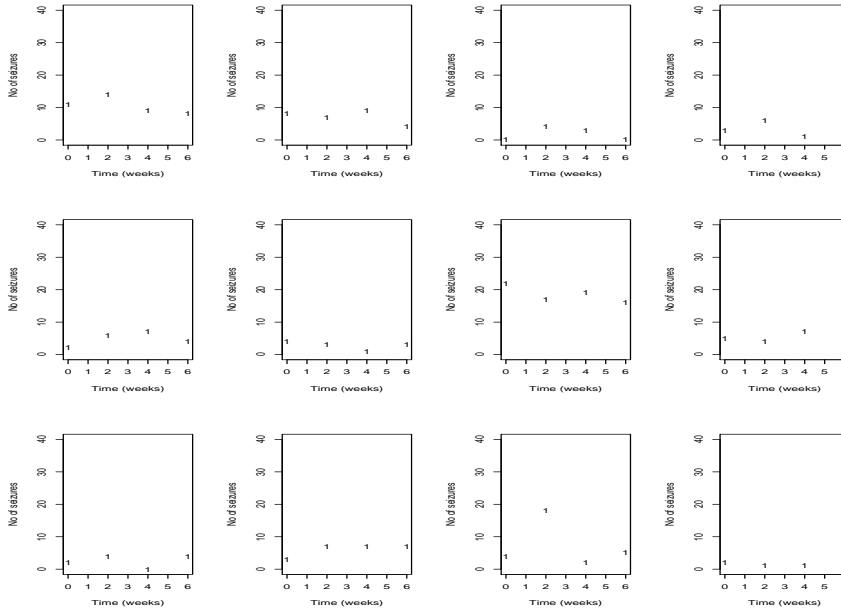


Figure 37: Number of seizures for selected individuals over time for progabide group.

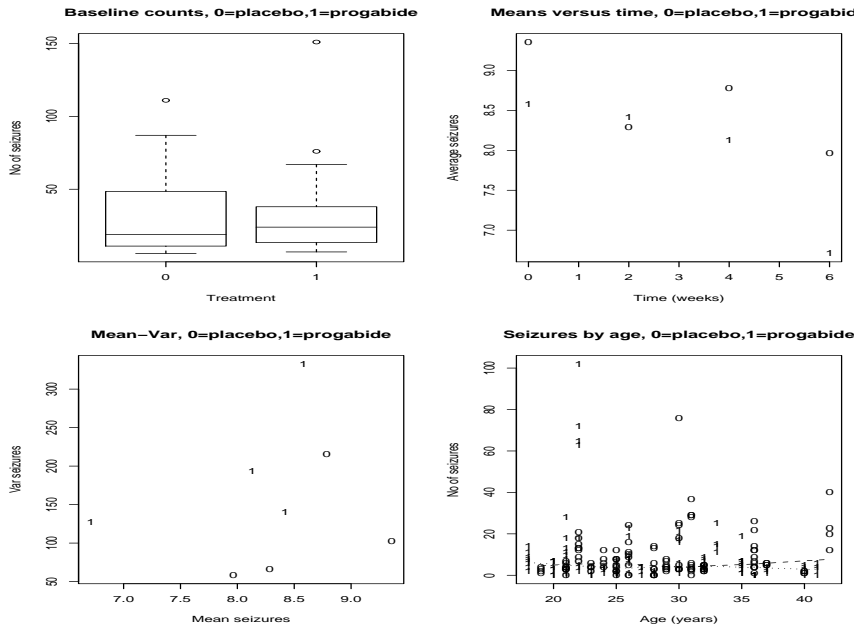


Figure 38: Summaries for seizure data.

254

Mixed Effects Model for Seizure Data

Stage 1: $Y_{ij} | \boldsymbol{\beta}, b_i \sim_{ind} \text{Poisson}(\mu_{ij})$, with

$$g(\mu_{ij}) = \log \mu_{ij} = \log t_{ij} + \mathbf{x}_{ij}\boldsymbol{\beta} + b_i,$$

where

$$\mathbf{x}_{ij}\boldsymbol{\beta} = \beta_0 + \beta_1 x_{ij1} + \beta_2 x_{ij2} + \beta_3 x_{ij1} x_{ij2}.$$

Hence

$$E[Y_{ij} | b_i] = \mu_{ij} = t_{ij} \exp(\mathbf{x}_{ij}\boldsymbol{\beta} + b_i), \quad \text{var}(Y_{ij} | b_i) = \mu_{ij}.$$

Stage 2: $b_i \sim_{iid} N(0, \sigma^2)$.

The marginal mean is given by

$$E[Y_{ij}] = t_{ij} \exp(\mathbf{x}_{ij}\boldsymbol{\beta} + \sigma^2/2),$$

and the marginal median by

$$t_{ij} \exp(\mathbf{x}_{ij}\boldsymbol{\beta}).$$

The marginal variance is given by

$$\begin{aligned}\text{var}(Y_{ij}) &= \text{E}[\mu_{ij}] + \text{var}(\mu_{ij}) \\ &= \text{E}[Y_{ij}]\{1 + \text{E}[Y_{ij}](e^{\sigma^2} - 1)\} = \text{E}[Y_{ij}](1 + \text{E}[Y_{ij}] \times \kappa)\end{aligned}$$

where $\kappa = e^{\sigma^2} - 1 > 0$ illustrating excess-Poisson variation which increases as σ^2 increases.

For the marginal covariance

$$\begin{aligned}\text{cov}(Y_{ij}, Y_{ik}) &= \text{cov}\{t_{ij} \exp(\mathbf{x}_{ij}\boldsymbol{\beta} + b_i), t_{ik} \exp(\mathbf{x}_{ik}\boldsymbol{\beta} + b_i)\} \\ &= t_{ij} \exp(\mathbf{x}_{ij}\boldsymbol{\beta} + \mathbf{x}_{ik}\boldsymbol{\beta}) \times e^{\sigma^2} \{e^{\sigma^2} - 1\} \\ &= \text{E}[Y_{ij}]\text{E}[Y_{ik}]\kappa.\end{aligned}$$

Hence for individual i we have variance-covariance matrix

$$\begin{bmatrix} \mu_{i1} + \mu_{i1}^2 \kappa & \mu_{i1}\mu_{i2}\kappa & \dots & \mu_{i1}\mu_{in_i}\kappa \\ \mu_{i2}\mu_{i1}\kappa & \mu_{i2} + \mu_{i2}^2 \kappa & \dots & \mu_{i2}\mu_{in_i}\kappa \\ \dots & \dots & \dots & \dots \\ \mu_{in_i}\mu_{i1}\kappa & \mu_{in_i}\mu_{i2}\kappa & \dots & \mu_{in_i} + \mu_{in_i}^2 \kappa \end{bmatrix},$$

where $\kappa = e^{\sigma^2} - 1 > 0$.

256

Likelihood Inference

In general there are two approaches to inference from a likelihood perspective:

1. Carry out conditional inference in order to eliminate the random effects.
2. Make a distributional assumption for \mathbf{b}_i , and then carry out likelihood inference (using some form of approximation to evaluate the required integrals).

We first consider the conditional likelihood approach.

257

Conditional Likelihood

Recall the definition of conditional likelihood. Suppose the distribution of the data may be factored as

$$p(\mathbf{y} \mid \boldsymbol{\beta}, \boldsymbol{\gamma}) = h(\mathbf{y}) \times p(\mathbf{t}_1, \mathbf{t}_2 \mid \boldsymbol{\beta}, \boldsymbol{\gamma}) = h(\mathbf{y}) \times p(\mathbf{t}_1 \mid \mathbf{t}_2, \boldsymbol{\beta}) \times p(\mathbf{t}_2 \mid \boldsymbol{\beta}, \boldsymbol{\gamma}),$$

where we choose to ignore the second term and consider the conditional likelihood

$$L_c(\boldsymbol{\beta}) = p(\mathbf{t}_1 \mid \mathbf{t}_2, \boldsymbol{\beta}) = \frac{p(\mathbf{t}_1, \mathbf{t}_2 \mid \boldsymbol{\beta}, \boldsymbol{\gamma})}{p(\mathbf{t}_2 \mid \boldsymbol{\beta}, \boldsymbol{\gamma})}.$$

Maximizing the conditional likelihood yields an estimator, $\widehat{\boldsymbol{\beta}}_c$ with the usual properties, for example

$$\mathbf{I}_c(\boldsymbol{\beta})^{1/2}(\widehat{\boldsymbol{\beta}}_c - \boldsymbol{\beta}) \rightarrow_d \mathbf{N}(\mathbf{0}, \mathbf{I}),$$

and $\mathbf{I}_c(\boldsymbol{\beta})$ is the expected information derived from the conditional likelihood.

258

Conditional Likelihood for GLMMs

In the context of GLMMs we have

$$L_c(\boldsymbol{\beta}) = \prod_{i=1}^m p(\mathbf{t}_{1i} \mid \mathbf{t}_{2i}, \boldsymbol{\beta}) = \prod_{i=1}^m \frac{p(\mathbf{t}_{1i}, \mathbf{t}_{2i} \mid \boldsymbol{\beta}, \mathbf{b}_i)}{p(\mathbf{t}_{2i} \mid \boldsymbol{\beta}, \mathbf{b}_i)}$$

where

$$p(\mathbf{t}_{1i}, \mathbf{t}_{2i} \mid \boldsymbol{\beta}, \mathbf{b}_i) \propto p(\mathbf{y}_i \mid \boldsymbol{\beta}, \mathbf{b}_i)$$

and

$$p(\mathbf{t}_{2i} \mid \boldsymbol{\beta}, \mathbf{b}_i) = \sum_{S_{2i}} p(\mathbf{u}_{1i}, \mathbf{t}_{2i} \mid \boldsymbol{\beta}, \mathbf{b}_i),$$

and S_{2i} is the set of values of \mathbf{y}_i such that $\mathbf{T}_{2i} = \mathbf{t}_{2i}$, a set of disjoint events.

The different notation is to emphasize that \mathbf{T}_{1i} takes on values different to \mathbf{t}_{1i} .

259

For simplicity we assume the canonical link function,

$$g(\mu_{ij}) = \theta_{ij} = \mathbf{x}_{ij}\boldsymbol{\beta} + \mathbf{z}_{ij}\mathbf{b}_i$$

and assume $\alpha = 1$. Viewing \mathbf{b}_i as fixed effects we have the likelihood

$$L(\boldsymbol{\beta}, \mathbf{b}) = \exp \left\{ \sum_{i=1}^m \sum_{j=1}^{n_i} y_{ij} \mathbf{x}_{ij} \boldsymbol{\beta} + y_{ij} \mathbf{z}_{ij} \mathbf{b}_i - b(\mathbf{x}_{ij} \boldsymbol{\beta} + \mathbf{z}_{ij} \mathbf{b}_i) \right\},$$

so that

$$\mathbf{t}_1 = \sum_{i=1}^m \mathbf{t}_{1i} = \sum_{i=1}^m \sum_{j=1}^{n_i} y_{ij} \mathbf{x}_{ij}$$

and

$$\mathbf{t}_{2i} = \sum_{j=1}^{n_i} y_{ij} \mathbf{z}_{ij}.$$

We emphasize that no distribution has been specified for the \mathbf{b}_i , and they are being viewed as fixed effects.

260

Conditional Likelihood for the Poisson GLMM

Assume for simplicity that $\mathbf{z}_{ij}\mathbf{b}_i = b_i$, so that we have the random intercepts only model. Also, in an obvious change in notation

$$\mathbf{x}_{ij}\boldsymbol{\beta} + \mathbf{x}_i\boldsymbol{\beta}_1 + b_i = \mathbf{x}_{ij}\boldsymbol{\beta} + \gamma_i$$

so that $\boldsymbol{\beta}$ are the regression associated with covariates that change within an individual.

Then

$$\begin{aligned} p(\mathbf{y} | \boldsymbol{\beta}, \boldsymbol{\gamma}) &= \prod_{i=1}^m p(\mathbf{y}_i | \boldsymbol{\beta}, \gamma_i) = \prod_{i=1}^m \frac{\exp \left(-\sum_{j=1}^{n_i} \mu_{ij} + \sum_{j=1}^{n_i} y_{ij} \log \mu_{ij} \right)}{\prod_{j=1}^{n_i} y_{ij}!} \\ &= c_1 \prod_{i=1}^m \exp \left(-\mu_{i+} + y_{i+} \gamma_i + \sum_{j=1}^{n_i} y_{ij} \log (t_{ij} \exp(\mathbf{x}_{ij}\boldsymbol{\beta})) \right) \end{aligned}$$

where $c_1^{-1} = \prod_i \prod_j y_{ij}!$ and $\mu_{i+} = \sum_{j=1}^{n_i} t_{ij} \exp(\mathbf{x}_{ij}\boldsymbol{\beta})$.

261

In this case the distribution of the conditioning statistic is straightforward:

$$y_{i+} \mid \boldsymbol{\beta}, \gamma_i \sim \text{Poisson}(\mu_{i+})$$

so that

$$\begin{aligned} p(y_{i+} \mid \boldsymbol{\beta}, \gamma_i) &= c_2 \prod_{i=1}^m \exp(-\mu_{i+} + y_{i+} \log \mu_{i+}) \\ &= c_2 \prod_{i=1}^m \exp\left(-\mu_{i+} + y_{i+} \gamma_i + y_{i+} \log\left(\sum_{j=1}^{n_i} t_{ij} \exp(\mathbf{x}_{ij} \boldsymbol{\beta})\right)\right) \end{aligned}$$

where $c_2^{-1} = y_{i+}!$

Hence

$$p(\mathbf{y} \mid y_{1+}, \dots, y_{n_i+}, \boldsymbol{\beta}) = \frac{p(\mathbf{y} \mid \boldsymbol{\beta}, \boldsymbol{\gamma})}{p(y_{1+}, \dots, y_{n_i+} \mid \boldsymbol{\beta}, \boldsymbol{\gamma})}$$

which is given by

$$\frac{c_1 \prod_i \exp\left(-\mu_{i+} + y_{i+} \gamma_i + \sum_{j=1}^{n_i} y_{ij} \log(t_{ij} \exp(\mathbf{x}_{ij} \boldsymbol{\beta}))\right)}{c_2 \prod_{i=1}^m \exp\left(-\mu_{i+} + y_{i+} \gamma_i + y_{i+} \log\left(\sum_{j=1}^{n_i} t_{ij} \exp(\mathbf{x}_{ij} \boldsymbol{\beta})\right)\right)}$$

262

After simplification:

$$\begin{aligned} p(\mathbf{y} \mid y_{1+}, \dots, y_{n_i+}, \boldsymbol{\beta}) &= \frac{c_1 \prod_{i=1}^m \prod_{j=1}^{n_i} (t_{ij} \exp(\mathbf{x}_{ij} \boldsymbol{\beta}))^{y_{ij}}}{c_2 \prod_{i=1}^m \left(\sum_{j=1}^{n_i} t_{ij} \exp(\mathbf{x}_{ij} \boldsymbol{\beta})\right)^{y_{i+}}} \\ &= \binom{y_{i+}}{y_{i1} \dots y_{in_i}} \prod_{i=1}^m \prod_{j=1}^{n_i} \left(\frac{t_{ij} \exp(\mathbf{x}_{ij} \boldsymbol{\beta})}{\sum_{l=1}^{n_i} t_{il} \exp(\mathbf{x}_{il} \boldsymbol{\beta})}\right)^{y_{ij}} \end{aligned}$$

which is a multinomial likelihood (we have conditioned a set of Poisson counts on their total so obvious!):

$$y_{ij} \mid y_{i+}, \boldsymbol{\beta} \sim \text{Mult}_{n_i}(y_{i+}, \boldsymbol{\pi}_i)$$

where $\boldsymbol{\pi}_i^T = (\pi_{i1}, \dots, \pi_{in_i})$ and

$$\pi_{ij} = \frac{t_{ij} \exp(\mathbf{x}_{ij} \boldsymbol{\beta})}{\sum_{l=1}^{n_i} t_{il} \exp(\mathbf{x}_{il} \boldsymbol{\beta})}.$$

263

Conditional Likelihood for the Seizure Data

Recall

$$\begin{aligned} Y_{ij} &= \text{number of seizures on patient } i \text{ at occasion } j \\ t_{ij} &= \text{observation period on patient } i \text{ at occasion } j \\ x_{i1} &= 0/1 \text{ if patient } i \text{ was assigned placebo/progabide} \\ x_{ij2} &= 0/1 \text{ if } j = 0/1, 2, 3, 4 \end{aligned}$$

with $t_{ij} = 8$ if $j = 0$ and $t_{ij} = 2$ if $j = 1, 2, 3, 4$, $i = 1, \dots, 59$.

A log-linear random intercept model is given by

$$\log E[Y_{ij} | b_i] = \log t_{ij} + \beta_0 + \beta_1 x_{i1} + \beta_2 x_{ij2} + \beta_3 x_{i1} x_{ij2} + b_i$$

Group	$j = 0$ period	$j = 1, 2, 3, 4$ period
Placebo	β_0	$\beta_0 + \beta_2$
Progabide	$\beta_0 + \beta_1$	$\beta_0 + \beta_1 + \beta_2 + \beta_3$

Table 9: Parameter interpretation.

264

Precise definitions:

- $\exp(\beta_0)$ is the expected number of seizures for a typical individual in the placebo group in time period 0;
- $\exp(\beta_1)$ is the ratio of the expected seizure rate in the progabide group, compared to the placebo group, for a typical individual, i.e. one with $b_i = 0$, in time period 0;
- $\exp(\beta_2)$ is the ratio of the expected seizure rate at times $j = 1, 2, 3, 4$, as compared to $j = 0$, for a typical individual in the placebo group;
- $\exp(\beta_3)$ is the ratio of the expected seizure rates in the progabide group in the $j = 1, 2, 3, 4$ period, as compared to the placebo group, in the same period for a typical individual. Hence $\exp(\beta_3)$ is the parameter of interest.

More colloquially:

- β_0 INTERCEPT
- β_1 BASELINE TREATMENT EFFECT
- β_2 PERIOD EFFECT
- β_3 TREATMENT \times PERIOD EFFECT

265

In the conditional likelihood notation:

$$\log E[Y_{ij} | \gamma_i] = \log t_{ij} + \gamma_i + \beta_2 x_{ij2} + \beta_3 x_{i1} x_{ij2}$$

where $\gamma_i = \beta_0 + \beta_1 x_{i1} + b_i$ so that we cannot estimate β_1 , which is not a parameter of primary interest.

Since $\mathbf{x}_{i1} = \mathbf{x}_{i2} = \mathbf{x}_{i3} = \mathbf{x}_{i4}$ and $t_{i0} = 8 = \sum_{j=1}^4 t_{ij}$, we effectively have two observation periods which we label (slightly abusing our previous notation), $j = 0, 1$. Let $Y_{i1} = \sum_{j=1}^4 Y_{ij}$.

For the placebo group:

$$Y_{i1} \sim_{ind} \text{Binomial}(Y_{i+}, \pi_{i1})$$

for $i = 1, \dots, 29$, with

$$\pi_{i1} = \frac{\exp(\beta_2)}{1 + \exp(\beta_2)}.$$

For the progabide group:

$$Y_{i1} \sim_{ind} \text{Binomial}(Y_{i+}, \pi_{i1})$$

for $i = 30, \dots, 59$, where

$$\pi_{i1} = \frac{\exp(\beta_2 + \beta_3)}{1 + \exp(\beta_2 + \beta_3)}.$$

.

266

This model is straightforward to fit in R:

```
> xcond <- c(rep(0,28),rep(1,31))
> condmod <- glm(cbind(y1,y0)~xcond,family=binomial)
> summary(condmod)
Call:
glm(formula = cbind(y1, y0) ~ xcond, family = binomial)
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.11080     0.04689   2.363  0.0181 *
xcond        -0.10368     0.06505  -1.594  0.1110
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for binomial family taken to be 1)
Null deviance: 306.50 on 58 degrees of freedom
Residual deviance: 303.96 on 57 degrees of freedom
```

Hence the treatment effect is $\exp(-.10) = 0.90$ so that the rate of seizures is estimated as 10% less in the progabide group, though this change is not statistically significant.

267

Conditional Likelihood for the Seizure Data

The overall fit of the random intercept model is poor (304 on 57 degrees of freedom).

Once possibility is to extend the model to allow a random slope for the effect of treatment x_{i2} , i.e. $\beta_{2i} = \beta_2 + b_{2i}$, but a conditional likelihood approach for this model will condition away the information relevant for estimation of β_3 .

We will examine such a model using a mixed effects approach.

268

Likelihood Inference in the Mixed Effects Model

As with the linear mixed effects model (LMEM) we maximize $L(\boldsymbol{\beta}, \boldsymbol{\alpha})$ where $\boldsymbol{\alpha}$ denote the variance components in \mathbf{D} , and

$$L(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \prod_{i=1}^m \int p(\mathbf{y}_i | \boldsymbol{\beta}, \mathbf{b}_i) \times p(\mathbf{b}_i | \boldsymbol{\alpha}) d\mathbf{b}_i.$$

As with the NLMEM the required integrals are not available in closed form and so some sort of analytical or numerical approximation is required.

269

Example: Log-linear Poisson regression GLMM

With a single random effect we have $\boldsymbol{\alpha} = \sigma^2$.

$$\begin{aligned}
 L(\boldsymbol{\beta}, \boldsymbol{\alpha}) &= \prod_{i=1}^m \int \prod_{j=1}^{n_i} \frac{\exp(-\mu_{ij}) \mu_{ij}^{y_{ij}}}{y_{ij}!} \times (2\pi\sigma^2)^{-1/2} \exp\left(-\frac{1}{2\sigma^2} b_i^2\right) db_i \\
 &= \prod_{i=1}^m (2\pi\sigma^2)^{-1/2} \exp\left(\sum_{i=1}^{n_i} y_{ij} x_{ij} \boldsymbol{\beta}\right) \\
 &\times \int \exp\left(-e^{b_i} \sum_{j=1}^{n_i} e^{\mathbf{x}_{ij} \boldsymbol{\beta}} + \sum_{j=1}^{n_i} y_{ij} b_i - \frac{1}{2\sigma^2} b_i^2\right) db_i \\
 &= \prod_{i=1}^m \exp\left(\sum_{i=1}^{n_i} y_{ij} x_{ij} \boldsymbol{\beta}\right) \times \int h(b_i) \frac{\exp\{-b_i^2/(2\sigma^2)\}}{(2\pi\sigma^2)^{-1/2}} db_i,
 \end{aligned}$$

an integral with respect to a normal random variable (which is analytically intractable).

270

Integration in the GLMM

As with the NLMEM there are a number of possible approaches for integrating out the random effects including:

- Analytical approximations, including Laplace, and the closely-related penalized quasi-likelihood approach.
- Gaussian quadrature.
- Importance sampling Monte Carlo

271

Penalized Quasi-Likelihood

Breslow and Clayton (1993) introduced the method of Penalized Quasi-Likelihood (PQL) which was an attempt to extend quasi-likelihood to GLMMs. One justification of the method is a Laplace approximation.

If we write the required integration in the form

$$\log L(\boldsymbol{\beta}, \boldsymbol{\alpha}) = -\frac{1}{2} \log |\mathbf{D}| + \log \left(\int \exp\{-g(\mathbf{b})\} d\mathbf{b} \right),$$

and use a Laplace approximation to the second term to obtain

$$\log L(\boldsymbol{\beta}, \boldsymbol{\alpha}) \approx -\frac{1}{2} \log |\mathbf{D}| - \frac{1}{2} \log |g''(\hat{\mathbf{b}})| - g(\hat{\mathbf{b}}),$$

where $\hat{\mathbf{b}}$ maximizes $g(\mathbf{b})$, and so satisfies $g'(\hat{\mathbf{b}}) = 0$.

PQL is very poor for binary data but may be OK for binomial and Poisson data (as long as the counts are not too small).

Within the MASS package the `glmmPQL` function obtains PQL estimates.

272

GLMM for the Seizure Data

We begin with the random intercepts model only:

```
> summary(glmmPQL(y~x1+x2+x3+offset(log(time)),random=~1|ID,family=poisson,data=seiz))
iteration 1
...
iteration 8
Linear mixed-effects model fit by maximum likelihood
Random effects:
  Formula: ~1 | ID
          (Intercept) Residual
StdDev:    0.7479965 1.670907
Variance function:
  Structure: fixed weights
  Formula: ~invwt
Fixed effects: y ~ x1 + x2 + x3 + offset(log(time))
              Value Std.Error DF  t-value p-value
(Intercept)  1.0762701 0.15622775 234  6.889110  0.0000
x1           -0.0196026 0.21561685  57 -0.090914  0.9279
x2            0.1107981 0.07888257 234  1.404596  0.1615
x3           -0.1036807 0.10943566 234 -0.947412  0.3444
Number of Observations: 295
Number of Groups: 59
```

273

Bayesian Inference for GLMMs

A Bayesian approach to inference for a GLMM adds a prior distribution for $\boldsymbol{\beta}, \boldsymbol{\alpha}$, to the likelihood $L(\boldsymbol{\beta}, \boldsymbol{\alpha})$. Again a proper prior is required for the matrix \mathbf{D} . In general a proper prior is not required for $\boldsymbol{\beta}$ – the exponential family and linear link lead to a likelihood that is well-behaved. Closed-form inference is unavailable, but MCMC is almost as straightforward as in the linear mixed model case. The joint posterior is

$$p(\boldsymbol{\beta}, \mathbf{W}, \mathbf{b} \mid \mathbf{y}) \propto \prod_{i=1}^m \{p(\mathbf{y}_i \mid \boldsymbol{\beta}, \mathbf{b}_i)p(\mathbf{b}_i \mid \mathbf{W})\} \pi(\boldsymbol{\beta})\pi(\mathbf{W}).$$

Suppose we have priors:

$$\begin{aligned} \boldsymbol{\beta} &\sim N_{q+1}(\boldsymbol{\beta}_0, \mathbf{V}_0) \\ \mathbf{W} &\sim W_{q+1}(r, \mathbf{R}^{-1}) \end{aligned}$$

The conditional distributions for $\boldsymbol{\beta}, \tau, \mathbf{W}$ are unchanged from the linear case. There is no closed form conditional distribution for $\boldsymbol{\beta}$, or for \mathbf{b}_i , but Metropolis-Hastings step can be used (or adaptive rejection sampling can be utilized, the conditional is log concave).

274

Priors for $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ in the GLMM

Lognormal Priors

It is convenient to specify lognormal priors for positive parameters θ , since one may specify two quantiles of the distribution, and directly solve for the two parameters of the prior. In a GLMM we can often specify priors for more meaningful parameters than elements of $\boldsymbol{\beta}$. For example, e^{β_1} is the relative risk in a log linear model, and is the odds ratio in a logistic model.

Suppose we wish to specify a lognormal prior for a generic parameter θ .

Denote by $\text{LN}(\mu, \sigma)$ the lognormal distribution with $E[\log \theta] = \mu$ and $\text{var}(\log \theta) = \sigma^2$, and let θ_1 and θ_2 be the q_1 and q_2 quantiles of this prior.

Then

$$\mu = \log(\theta_1) \left(\frac{z_{q_2}}{z_{q_2} - z_{q_1}} \right) - \log(\theta_2) \left(\frac{z_{q_1}}{z_{q_2} - z_{q_1}} \right), \quad \sigma = \frac{\log(\theta_1) - \log(\theta_2)}{z_{q_1} - z_{q_2}}. \quad (44)$$

As an example, suppose that for θ we believe there is a 50% chance that the relative risk is less than 1 and a 95% chance that it is less than 5; with $q_1 = 0.5, \theta_1 = 1.0$ and $q_2 = 0.95, \theta_2 = 5.0$, we obtain lognormal parameters $\mu = 0$ and $\sigma = \log 5/1.96 = 0.98$.

275

Gamma Priors

Consider the random intercepts model with $b_i \sim_{iid} N(0, \sigma^2)$.

It is not straightforward to specify a prior for σ , which represents the standard deviation of the residuals on the linear predictor scale, and is consequently not easy to interpret.

We specify a gamma prior $\text{Ga}(a, b)$ for the precision $\tau = 1/\sigma^2$, with parameters a, b specified *a priori*. The choice of a gamma distribution is convenient since it produces a marginal distribution for the residuals in closed form.

Specifically the two-stage model

$$b_i | \sigma \sim_{iid} N(0, \sigma^2), \quad \tau = \sigma^{-2} \sim \text{Ga}(a, b)$$

produces a marginal distribution for b_i which is $t_d(0, \lambda^2)$, a Student's t distribution with $d = 2a$ degrees of freedom, location zero, and scale $\lambda^2 = b/a$.

We now consider a log link, in which case the above is equivalent to the residual relative risks following a log t distribution.

We specify the range $\exp(\pm R)$ within which the residual relative risks will lie with probability q , and use the relationship $\pm t_{q/2}^d \lambda = \pm R$, where t_q^d is the q -th quantile of a Student t random variable with d degrees of freedom, to give $a = d/2$, $b = R^2 d/2 (t_{q/2}^d)^2$.

276

For example, if we assume *a priori* that the residual relative risks follow a log Student t distribution with 2 degrees of freedom, and that 95% of these risks fall in the interval (0.5,2.0) then we obtain the prior, $\text{Ga}(1, 0.0260)$.

In terms of σ this results in (2.5%, 97.5%) quantiles of (0.084,1.01) with posterior median 0.19.

It is important to assess whether the prior allows all reasonable levels of variability in the residual relative risks, in particular small values should not be excluded.

The prior $\text{Ga}(0.001, 0.001)$ which has previously been used (e.g. in the WinBUGS manual) should be avoided for this very reason (this corresponds to relative risks which follow a log Student t distribution with 0.002 degrees of freedom).

277

Bayesian Inference for the Seizure Data

We fit various models and begin with a discussion of prior specification.

We fit four models to the seizure data.

- Model 1 Random intercepts only, $\pi(\boldsymbol{\beta}) \propto 1$, $\tau \sim \text{Ga}(1, 0.260)$ – corresponds to Student t_2 residuals and 95% $\in (0.5, 2.0)$.
- Model 2 Random intercepts only, $\pi(\boldsymbol{\beta}) \propto 1$, $\tau \sim \text{Ga}(2, 1.376)$ – corresponds to Student t_4 residuals and 95% $\in (0.1, 10.0)$.
- Model 3 Random effects for intercept and for x_2 .
- Model 4 We allow a bivariate Student t distribution for the pair of random effects introduced in Model 3.

278

WinBUGS code

```

model
{
for (i in 1:n){
  for (j in 1:k){
    Y[i,j] ~ dpois(mu[i,j])
    log(mu[i,j]) <- log(t[j])+beta0+beta1*x1[i]+beta2*x2[j]+
                    beta3*x1[i]*x2[j]+b[i]
  }
  b[i] ~ dnorm(0,tau)
}
tau ~ dgamma(1,0.260)
sigma <- 1/tau
beta0 ~ dflat()
beta1 ~ dflat()
beta2 ~ dflat()
beta3 ~ dflat()
}

```

279

Generalized Estimating Equations (GEEs)

Liang and Zeger (1986, *Biometrika*), and Zeger and Liang (1986, *Biometrics*) considered GLMs with dependence within individuals (in the context of longitudinal data).

Theorem (Liang and Zeger, 1986): the estimator $\hat{\beta}$ that satisfies

$$\mathbf{G}(\beta, \hat{\alpha}) = \sum_{i=1}^m \mathbf{D}_i^T \mathbf{W}_i^{-1} (\mathbf{Y}_i - \boldsymbol{\mu}_i) = \mathbf{0},$$

where $\mathbf{D}_i = \frac{\partial \boldsymbol{\mu}_i}{\partial \beta}$, $\mathbf{W}_i = \mathbf{W}_i(\beta, \alpha)$ is the working covariance model, $\boldsymbol{\mu}_i = \boldsymbol{\mu}_i(\beta)$ and $\hat{\alpha}$ is a consistent estimator of α , is such that

$$\mathbf{V}_\beta^{-1/2} (\hat{\beta} - \beta) \rightarrow_d \mathbf{N}(\mathbf{0}, \mathbf{I}),$$

where \mathbf{V}_β is given by

$$\left(\sum_{i=1}^m \mathbf{D}_i^T \mathbf{W}_i^{-1} \mathbf{D}_i \right)^{-1} \left\{ \sum_{i=1}^m \mathbf{D}_i^T \mathbf{W}_i^{-1} \text{cov}(\mathbf{Y}_i) \mathbf{W}_i^{-1} \mathbf{D}_i \right\} \left(\sum_{i=1}^m \mathbf{D}_i^T \mathbf{W}_i^{-1} \mathbf{D}_i \right)^{-1}.$$

In practice an empirical estimator of $\text{cov}(\mathbf{Y}_i)$ is substituted to give $\widehat{\mathbf{W}}_\beta$.

282

Choice of Working Covariance Models

As in the linear case, various assumptions about the form of the working covariance may be assumed (what is a natural choice?); we write

$$\mathbf{W}_i = \boldsymbol{\Delta}_i^{1/2} \mathbf{R}_i(\alpha) \boldsymbol{\Delta}_i^{1/2},$$

where $\boldsymbol{\Delta}_i = \text{diag}[\text{var}(Y_{i1}), \dots, \text{var}(Y_{in_i})]^T$ and \mathbf{R}_i is a working correlation model, for example, independence, exchangeable, AR(1), unstructured.

- For small m the sandwich estimator will have high variability and so model-based variance estimators may be preferable (but would we trust asymptotic normality if m were small anyway?).
- Model-based estimators are more efficient if the model is correct.

Published comments:

- Liang and Zeger (1986): “little difference when correlation is moderate”.
- McDonald (1993): “The independence estimator may be recommended for practical purposes”.
- Zhao, Prentice and Self (1992): Assuming independence “can lead to important losses of efficiency”.
- Fitzmaurice, Laird and Rotnitsky (1993): “important to obtain a close approximation to $\text{cov}(\mathbf{Y}_i)$ in order to achieve high efficiency”.

283

GEE for the Seizure Data

We have the log-linear model is given

$$\log E[Y_{ij}] = \log \mu_{ij} = \log t_{ij} + \beta_0^* + \beta_1 x_{i1} + \beta_2 x_{ij2} + \beta_3 x_{i1} x_{ij2}$$

and $\text{var}(Y_{ij}) = \alpha \mu_{ij}$. Recall β_1 is baseline comparison of rates, β_2 is period effect and β_3 is treatment effect.

Both quasi-likelihood and working independence GEE have estimating equation

$$\mathbf{G}(\boldsymbol{\beta}, \hat{\boldsymbol{\alpha}}) = \sum_{i=1}^m \mathbf{x}_i^T (\mathbf{Y}_i - \boldsymbol{\mu}_i) = \mathbf{0},$$

but differ in the manner in which the standard errors are calculated.

284

	Estimates (standard errors)				
	Poisson	Quasi-Lhd	GEE Ind	GEE Exch	GEE AR(1)
β_0^*	1.35 (0.034)	1.35 (0.15)	1.35 (0.16)	1.35 (0.16)	1.31 (0.16)
β_1	0.0265 (0.047)	0.0265 (0.21)	0.0265 (0.22)	0.0265 (0.22)	0.0153 (0.21)
β_2	0.111 (0.047)	0.111 (0.21)	0.111 (0.12)	0.111 (0.12)	0.155 (0.11)
β_3	-0.104 (0.065)	-0.104 (0.29)	-0.104 (0.22)	-0.104 (0.22)	-0.131 (0.27)
α_1, α_2	1.0, 0	19.7, 0	19.4, 0	19.4, 0.78	20.0, 0.89

Table 11: Parameter estimates and standard errors under various models; α_1 is a variance parameter, and α_2 a correlation parameter.

The point estimates under Poisson, quasi-likelihood and GEE working independence will always agree. The Poisson standard errors are clearly much too small. The quasi-likelihood standard errors are increased by $\sqrt{19.7} = 4.4$, but do not acknowledge dependence on observations on the same individual (it is as if we have 59×5 independent observations). The standard errors of estimated parameters that are associated with time-varying covariates (β_2 and β_3) are reduced under GEE, since within-person comparisons are being made. The coincidence of the estimates and standard errors for independence and exchangeability is a quirk of the balanced design.

285

Interpretation of Marginal and Conditional Coefficients

In a marginal model (which we consider under GEE), we have

$$E[Y | x] = \exp(\gamma_0 + \gamma_1 x)$$

in which case e^{γ_1} is the change in the average response when we increase x by 1 unit in the population under consideration.

Under the conditional (mixed effects) model the interpretation of regression coefficients is conditional on the value of the random effect.

For the model

$$E[Y | x, b] = \exp(\beta_0 + \beta_1 x + b),$$

with $b \sim_{iid} N(0, \sigma^2)$, the marginal mean is given by:

$$E[Y | x] = E_{b|\sigma^2} \{E[Y | x, b]\} = \exp(\beta_0 + \sigma^2/2 + \beta_1 x).$$

Hence for the log-linear model, e^{β_1} has a very similar marginal interpretation to e^{γ_1} , though estimation of the latter via GEE produces a consistent estimator in more general circumstances (though there is an efficiency loss if the random effects model is correct).

286

In the model

$$E[Y | x, \mathbf{b}] = \exp\{\beta_0 + b_{0i} + (\beta_1 + b_{1i})x_i\}$$

e^{β_1} is the relative risk between two populations with the same \mathbf{b} but whose x values differ by one unit, that is:

$$\exp(\beta_1) = \frac{E[Y | x, \mathbf{b}]}{E[Y | x - 1, \mathbf{b}]}$$

An alternative interpretation is to say that it is the expected change between two “typical individuals”, that is, individuals with random effects, $\mathbf{b} = 0$.

With $\mathbf{b} \sim_{iid} N(\mathbf{0}, \mathbf{D})$ we have the marginal mean

$$E[Y | x] = \exp\{\beta_0 + D_{00}/2 + x(\beta_1 + D_{01}) + x^2 D_{11}/2\}$$

so that there is no useful marginal interpretation of $\exp(\beta_1)$.

287

Stochastic Covariates

In some longitudinal situations, the response at time t on individual i may depend on not just the current covariates, but also previous values.

For example, in an investigation into the health effects of recent air pollution we may believe that the response depends on not just today's exposure, but also the preceding days.

In such situations, obtaining the correct form of the model will in general be difficult, and instead we might decide to estimate the association for a simpler model.

As an example, suppose that we have a single covariate, and we decide to examine the *cross-sectional* association:

$$\mu_{ij} = E[Y_{ij} | X_{ij}]. \quad (45)$$

In such a situation great care must be taken to obtain a consistent estimator.

We demonstrate with a GEE approach, though the pitfalls of estimation apply equally to likelihood and Bayesian approaches.

288

With a GLM:

$$\eta_{ij} = g(\mu_{ij}) = \mathbf{x}_{ij}\boldsymbol{\beta},$$

and assume for simplicity $\boldsymbol{\beta} = (\beta_0, \beta_1)^\top$. The generalized estimating function is given by

$$\mathbf{G}(\boldsymbol{\beta}) = \sum_{i=1}^m \mathbf{D}_i^\top \mathbf{W}_i^{-1} (\mathbf{Y}_i - \boldsymbol{\mu}_i)$$

which has second row

$$G_2(\boldsymbol{\beta}) = \sum_{i=1}^m \left[\sum_{j=1}^{n_i} \sum_{k=1}^{n_i} X_{ij} W_{ijk}^* (Y_{ik} - \mu_{ik}) \right]$$

where (45) is the assumed model, i.e. $\mu_{ik} = E[Y_{ij} | X_{ik}]$, and $W_{ijk}^* = \frac{\partial \mu_{ij}}{\partial \eta_{ij}} W_i^{jk}$ with W_i^{jk} the (j, k) -th element of \mathbf{W}_i^{-1} .

289

To obtain consistency we require

$$\mathbf{E}[\mathbf{G}(\boldsymbol{\beta})] = \mathbf{0}.$$

Previously we have seen that if the mean specification is correct then we obtain consistency of $\widehat{\boldsymbol{\beta}}$.

Since now the estimating function depends on the random variables

$\mathbf{X} = (X_1, \dots, X_m)^T$ the expectation is with respect to \mathbf{X} and \mathbf{Y} . Specifically

$$\mathbf{E}_{Y, X}[G_2(\boldsymbol{\beta})] = \sum_{i=1}^m \mathbf{E}_{Y_i | X_i} \left[\sum_{j=1}^{n_i} \sum_{k=1}^{n_i} X_{ij} W_{ijk}^* (Y_{ik} - \mu_{ik}) \right]$$

and

$$\begin{aligned} \mathbf{E}_{Y_i, X_i} [X_{ij} W_{ijk}^* (Y_{ik} - \mu_{ik})] &= \mathbf{E}_{X_i} \left\{ \mathbf{E}_{Y_i | X_i} [X_{ij} W_{ijk}^* (Y_{ik} - \mu_{ik})] \right\} \\ &= \mathbf{E}_{X_i} \left\{ X_{ij} W_{ijk}^* (\mathbf{E}[Y_{ik} | X_{i1}, \dots, X_{in_i}] - \mu_{ik}) \right\} \end{aligned}$$

Hence to ensure an unbiased estimating function, and hence consistency of our estimator, we require

$$\mathbf{E}[Y_{ik} | X_{i1}, \dots, X_{in_i}] = \mu_{ik} = \mathbf{E}[Y_{ij} | X_{ik}],$$

otherwise we have bias.

290

Suppose we assume working independence, the above simplifies to

$$G_2(\boldsymbol{\beta}) = \sum_{i=1}^m \sum_{j=1}^{n_i} X_{ij} W_{ijj}^* (Y_{ij} - \mu_{ij}),$$

so that we only need to consider the expectation with respect to $\mathbf{Y}_i, \mathbf{X}_i$:

$$\mathbf{E}_{Y_{ij} | X_{ij}} [\mathbf{G}(\boldsymbol{\beta})] = \sum_{i=1}^m \sum_{j=1}^{n_i} \mathbf{E}_{X_i} [X_{ij} W_{ijj}^* (\mathbf{E}[Y_{ij} | X_{ij}] - \mu_{ij})] = \mathbf{0},$$

and we obtain a consistent estimator.

For more details see DHLZ, Section 12.3.1.

Cross-Sectional Versus Longitudinal Studies

Consider modeling $Y = \text{FEV}_1$ as a function of age. We might envisage that Y changes both over time within an individual, and that individuals may have different baseline levels of Y from which they begin, due to “cohort” effects. A birth cohort is a group of individuals who were born in the same year.

Cohort effects include the effects of environmental pollutants, and differences in lifestyle choices and medical treatment.

In a cross-sectional study a group of individuals are measured at a single time point. A great advantage of longitudinal studies, over cross-sectional studies is that both cohort and aging effects can be estimated.

292

To illustrate, consider the model:

$$E[Y_{ij} | \text{Age}_{ij}] = \beta_0 + \beta_c \text{Age}_{i1} + \beta_l (\text{Age}_{ij} - \text{Age}_{i1})$$

where Y_{ij} is the j -th FEV_1 measurement on individual i and Age_{ij} is the age of the individual when that measurement was taken, with Age_{i1} begin the age on a certain day (so that all the individuals are comparable).

Parameter interpretation:

We have

$$E[Y_{i1} | \text{Age}_{i1}] = \beta_0 + \beta_c \text{Age}_{i1},$$

so that β_c is the average change in Y between two populations who differ by one unit in their baseline ages; said another way we are examining the differences in Y between two birth cohorts a year apart.

Also

$$E[Y_{ij} | \text{Age}_{ij}] - E[Y_{i1} | \text{Age}_{i1}] = \beta_l (\text{Age}_{ij} - \text{Age}_{i1})$$

so that β_l is the longitudinal effect, that is the change in the average response between two populations who are in the same birth cohort, and whose ages differ by one year. In a cohort study with $E[Y_{ij} | x_{ij}] = \beta_0 + \beta_1 x_{ij}$ we have

$$E[\hat{\beta}_1] = \beta_l + \frac{\sum_{i=1}^m n_i (x_{i1} - \bar{x}_{i1})(\bar{x}_i - \bar{x})}{\sum_{i=1}^m \sum_{j=1}^{n_i} (x_{ij} - \bar{x})^2} (\beta_c - \beta_l).$$

293

GLMMs for the Seizure Data

New R functions have recently been released to fit GLMMs using MLE; the required integrals can be approximated using penalized quasi-likelihood, Laplace, or adaptive Gaussian quadrature.

```
> library(lme4)
> lmermod1 <- lmer(y ~ x1+x2+x3+(1|ID)+offset(log(time)),family=poisson,data=seiz,method="PQL")
> summary(lmermod1)
Generalized linear mixed model fit using PQL
Random effects:
      Groups      Name      Variance  Std.Dev.
      ID (Intercept)  0.20035   0.44761
Fixed effects:
      Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.076279  0.092852 11.5914 < 2e-16 ***
x1           -0.019602  0.128149 -0.1530  0.87843
x2            0.110798  0.046888  2.3630  0.01813 *
x3           -0.103681  0.065055 -1.5937  0.11099
> lmermod2 <- lmer(y ~ x1+x2+x3+(1|ID)+offset(log(time)),family=poisson,data=seiz,method="Laplace")
> summary(lmermod2)
Generalized linear mixed model fit using Laplace
Random effects:
      Groups      Name      Variance  Std.Dev.
      ID (Intercept)  0.60864   0.78015
Fixed effects:
      Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.037500  0.092852 11.1737 < 2e-16 ***
x1           -0.023061  0.128149 -0.1800  0.85719
x2            0.110799  0.046888  2.3630  0.01813 *
x3           -0.103680  0.065055 -1.5937  0.11099
> lmermod3 <- lmer(y ~ x1+x2+x3+(1|ID)+offset(log(time)),family=poisson,data=seiz,method="AGQ")
```

294

```
> summary(lmermod3)
Generalized linear mixed model fit using AGQ
Random effects:
      Groups      Name      Variance  Std.Dev.
      ID (Intercept)  0.60905   0.78042
# of obs: 295, groups: ID, 59
Fixed effects:
      Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.032596  0.092852 11.1209 < 2e-16 ***
x1           -0.023858  0.128149 -0.1862  0.85231
x2            0.110791  0.046888  2.3629  0.01813 *
x3           -0.103682  0.065055 -1.5938  0.11099
> lmermod4 <- lmer(y ~ x1+x2+x3+(1+x2|ID)+offset(log(time)),family=poisson,data=seiz,method="Laplace")
> summary(lmermod4)
Generalized linear mixed model fit using Laplace
Random effects:
      Groups Name      Variance Std.Dev. Corr
      ID (Intercept)  0.50071  0.70761
      x2            0.23285  0.48255  0.165
Fixed effects:
      Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.0767903  0.0980746 10.9793 < 2.2e-16 ***
x1           0.0494416  0.1351770  0.3658  0.714548
x2           0.0013726  0.0766903  0.0179  0.985721
x3          -0.3028290  0.1071637 -2.8259  0.004715 **
```

295