GENERAL REGRESSION MODELS

We consider the class of *Generalized Linear Mixed Models* (GLMMs) and *non-linear mixed effects models* (NLMEMs).

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.

First we briefly review generalized linear models and non-linear models for independent data.

Generalized Linear Models

GLMs provide a very useful extension to the linear model class. GLMs specify stochastic and deterministic components:

The responses y_i follow an exponential family so that the distribution is of the form

$$p(y_i \mid \theta_i, \alpha) = \exp(\{y_i\theta_i - b(\theta_i)\}/\alpha + c(y_i, \alpha)), \tag{41}$$

where θ_i and α are scalars. This is sometimes referred to as a *linear* or *natural* exponential family. It is straightforward to show that

$$\mathrm{E}[Y_i \mid \theta_i, \alpha] = \mu_i = b'(\theta_i)$$

and

$$\mathrm{var}(Y_i \mid \theta_i, \alpha) = b''(\theta_i)\alpha,$$

for $i = 1, ..., n$, with $\mathrm{cov}(Y_i, Y_j \mid \theta_i, \theta_j, \alpha) = 0$ for $i \neq j$.

The *link function* $g(\cdot)$ provides the connection between $\mu = \mathrm{E}[Y \mid \theta, \alpha]$ and the *linear predictor* $\boldsymbol{x}\boldsymbol{\beta}$, via

$$g(\mu) = \boldsymbol{x}\boldsymbol{\beta},$$

where $\boldsymbol{x}$ is a $(k+1) \times 1$ vector of explanatory variables (including a 1 for the intercept) and $\boldsymbol{\beta}$ is a $1 \times (k+1)$ of regression parameters.

If α is known this is a one-parameter (natural) exponential family model and there is a $(k+1)$-dimensional sufficient statistic for $\boldsymbol{\beta}$. If α is unknown then the distribution may or may not be a two-parameter exponential family model.

Likelihood Inference

We now derive the score vector and information matrix. For an independent sample from the exponential family (41), we have

$$l(\boldsymbol{\theta}) = \sum_{i=1}^{n} \frac{y_i\theta_i - b(\theta_i)}{\alpha} + c(y_i, \alpha) = \sum_{i=1}^{n} l_i(\boldsymbol{\theta}),$$

where $\boldsymbol{\theta} = [\theta_1(\boldsymbol{\beta}), ..., \theta_n(\boldsymbol{\beta})]$ is the vector of canonical parameters.

Using the chain rule we may write

$$\begin{aligned} S(\boldsymbol{\beta}) = \frac{\partial l}{\partial \boldsymbol{\beta}} &= \sum_{i=1}^{n} \frac{\partial l_i}{\partial \theta_i} \frac{d\theta_i}{d\mu_i} \frac{\partial \mu_i}{\partial \boldsymbol{\beta}} \\ &= \sum_{i=1}^{n} \frac{Y_i - b'(\theta_i)}{\alpha} \frac{1}{V_i} \frac{d\mu_i}{d\boldsymbol{\beta}} \end{aligned}$$

where $V_i = \mathrm{var}(Y \mid \boldsymbol{\beta})/\alpha$ and

$$\frac{d^2 b}{d\theta_i^2} = \frac{d}{d\theta_i}\mu_i = V_i.$$

Hence

$$\boldsymbol{S}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left(\frac{\partial \mu_i}{\partial \boldsymbol{\beta}} \right)^{\mathrm{T}} \frac{\{Y_i - \mathrm{E}[Y_i \mid \mu_i]\}}{\mathrm{var}(Y_i \mid \mu_i)} = \boldsymbol{D}^{\mathrm{T}} \boldsymbol{V}^{-1} \{\boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{\beta})\}, \tag{42}$$

where $\boldsymbol{D}$ is the $n \times p$ matrix with elements $\partial \mu_i / \partial \beta_j$, $i = 1, ..., n$, $j = 1, ..., p$, and $\boldsymbol{V}$ is the $n \times n$ diagonal matrix with $i-$th diagonal element $\mathrm{var}(Y_i \mid \mu_i)$.

The MLE has asymptotic distribution

$$\boldsymbol{I}_n(\boldsymbol{\beta})^{1/2}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \ \rightarrow_d \ \mathrm{N}_p(\boldsymbol{0}, \boldsymbol{I}_p),$$

where

$$\boldsymbol{I}_n(\boldsymbol{\beta}) = \mathrm{E}[\boldsymbol{S}(\boldsymbol{\beta})\boldsymbol{S}(\boldsymbol{\beta})^{\mathrm{T}}] = \boldsymbol{D}^{\mathrm{T}} \boldsymbol{V}^{-1} \boldsymbol{D}.$$

In practice we use

$$\boldsymbol{I}_n(\widehat{\boldsymbol{\beta}}) = \widehat{\boldsymbol{D}}^{\mathrm{T}} \widehat{\boldsymbol{V}} \widehat{\boldsymbol{D}},$$

where $\widehat{\boldsymbol{V}}$ and $\widehat{\boldsymbol{D}}$ are $\boldsymbol{V}$ and $\boldsymbol{D}$ evaluated at $\widehat{\boldsymbol{\beta}}$. Hence an estimator $\widehat{\boldsymbol{\beta}}$ defined through $\boldsymbol{S}(\widehat{\boldsymbol{\beta}}) = \boldsymbol{0}$ will be consistent so long as the mean function is appropriate. The variance of the estimator is appropriately estimated if the second moment is correctly specified.

Nonlinear Regression Models

Consider models of the form

$$Y_i = \mu_i(\boldsymbol{\beta}) + \epsilon_i,$$

$i = 1, ..., n$, where $\mu_i(\boldsymbol{\beta}) = \mu(\boldsymbol{x}_i, \boldsymbol{\beta})$ is nonlinear in $\boldsymbol{x}_i$, $\boldsymbol{\beta}$ is assumed to be of dimension $k \times 1$, and $\mathrm{E}[\epsilon_i \mid \mu_i] = 0$, $\mathrm{var}(\epsilon_i \mid \mu_i) = \sigma^2 f(\mu_i)$ with $\mathrm{cov}(\epsilon_i, \epsilon_j \mid \mu_i) = 0$.

Such models are often used for positive responses, and if such data are modeled on the original scale it is common to find that the variance is of the form $f(\mu) = \mu$ or $f(\mu) = \mu^2$.

An alternative approach that is appropriate in the case of $f(\mu) = \mu^2$ is to log transform the responses and then assume constant errors.

Likelihood Inference

To obtain the likelihood function the probability model for the data must be fully specified. Assume

$$Y_i \mid \boldsymbol{\beta}, \sigma^2 \sim_{ind} \mathrm{N}\{\mu_i(\boldsymbol{\beta}), \sigma^2 \mu_i(\boldsymbol{\beta})^r\},$$

for $i = 1, ..., n$, and known $r \geq 0$ to give the likelihood function

$$l(\boldsymbol{\beta}, \sigma) = -n \log \sigma - \frac{r}{2} \sum_{i=1}^{n} \log \mu_i(\boldsymbol{\beta}) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} \frac{\{Y_i - \mu_i(\boldsymbol{\beta})\}^2}{\mu_i^r(\boldsymbol{\beta})}.$$

Differentiation with respect to $\boldsymbol{\beta}$ and σ yields the score equations

$$\begin{aligned}
\boldsymbol{S}_1(\boldsymbol{\beta}, \sigma) &= -\frac{r}{2} \sum_{i=1}^{n} \frac{\partial \mu_i}{\partial \boldsymbol{\beta}}(\boldsymbol{\beta}) \frac{1}{\mu_i(\boldsymbol{\beta})} + \frac{1}{\sigma^2} \sum_{i=1}^{n} \frac{\{Y_i - \mu_i(\boldsymbol{\beta})\}}{\mu_i(\boldsymbol{\beta})^r} \frac{\partial \mu_i}{\partial \boldsymbol{\beta}} \\
&- \frac{r}{2\sigma^2} \sum_{i=1}^{n} \frac{\{Y_i - \mu(\boldsymbol{\beta})\}^2}{\mu_i^{r+1}(\boldsymbol{\beta})} \frac{\partial \mu_i}{\partial \boldsymbol{\beta}}. \\
S_2(\boldsymbol{\beta}, \sigma) &= -\frac{n}{\sigma} + \frac{1}{\sigma^3} \sum_{i=1}^{n} \frac{\{Y_i - \mu(\boldsymbol{\beta})\}^2}{\mu_i^r(\boldsymbol{\beta})}.
\end{aligned}$$

Notice that we have a pair of quadratic estimating functions here and if the first two moments are correctly specified then $\mathrm{E}[\boldsymbol{S}_1] = \boldsymbol{0}$ and $\mathrm{E}[S_2] = 0$.

Under the usual regularity conditions

$$\boldsymbol{I}(\boldsymbol{\theta})^{1/2}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}) \rightarrow_d \mathrm{N}_{k+1}(\mathbf{0},\boldsymbol{I}).$$

where $\boldsymbol{\theta}=(\boldsymbol{\beta},\sigma)$ and $\boldsymbol{I}(\theta)$ is Fisher's expected information. In the case of $r=0$ we obtain

$$\begin{aligned}
l(\boldsymbol{\beta},\sigma) &= -n\log\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^n\{Y_i-\mu_i(\boldsymbol{\beta})\}^2 \\
\boldsymbol{S}_1(\boldsymbol{\beta},\sigma) &= \frac{1}{\sigma^2}\sum_{i=1}^n\{Y_i-\mu_i(\boldsymbol{\beta})\}\frac{\partial\mu_i}{\partial\boldsymbol{\beta}} \\
S_2(\boldsymbol{\beta},\sigma) &= -\frac{n}{\sigma^2}+\frac{1}{\sigma^4}\sum_{i=1}^n\{Y_i-\mu_i(\boldsymbol{\beta})\}^2 \\
\boldsymbol{I}_{11} &= -\mathrm{E}\left[\frac{\partial\boldsymbol{S}_1}{\partial\boldsymbol{\beta}}\right]=\frac{1}{\sigma^2}\sum_{i=1}^n\left(\frac{\partial\mu_i}{\partial\boldsymbol{\beta}}\right)\left(\frac{\partial\mu_i}{\partial\boldsymbol{\beta}}\right)^{\mathrm{T}} \\
\boldsymbol{I}_{12} &= -\mathrm{E}\left[\frac{\partial\boldsymbol{S}_1}{\partial\sigma}\right]=\mathbf{0}^{\mathrm{T}} \\
\boldsymbol{I}_{21} &= -\mathrm{E}\left[\frac{\partial S_2}{\partial\boldsymbol{\beta}}\right]=\mathbf{0} \\
\boldsymbol{I}_{22} &= -\mathrm{E}\left[\frac{\partial S_2}{\partial\sigma}\right]=\frac{2n}{\sigma^2}
\end{aligned}$$

Identifiability

For many nonlinear models identifiability is an issue. Specifically the same curve may be obtained with different parameter values. For example, consider the sum-of-exponentials model

$$\mu(x,\boldsymbol{\beta})=\beta_0\exp(-x\beta_1)+\beta_2\exp(-x\beta_3),$$

where $\boldsymbol{\beta}=(\beta_0,\beta_1,\beta_2,\beta_3)$ and $\beta_j>0$, $j=0,1,2,3$. The same curve results under the parameter sets $(\beta_0,\beta_1,\beta_2,\beta_3)$ and $(\beta_2,\beta_3,\beta_0,\beta_1)$ and so we have non-identifiability. A solution to this problem, and to ensure that the parameters can only take admissible values, is to work with the set

$$\boldsymbol{\gamma}=(\log\beta_0,\log(\beta_3-\beta_1),\log\beta_2,\log\beta_1)$$

which constrains $\beta_3>\beta_1>0$.

Example of a non-linear model: One compartmental open model

- Let $w_i(x)$ represent the amount of drug in compartment i, $i = 0, 1$, at time x.
- Assume:

$$\begin{aligned} \frac{dw_0}{dx} &= -k_a w_0 \\ \frac{dw_1}{dx} &= k_a w_0 - k_e w_1 \end{aligned}$$

  where k_a is the absorption rate, and k_e is the elimination rate.
- Leads to

$$\mu(x) = \frac{Dk_a}{V(k_a - k_e)} \{\exp[-k_e x] - \exp[-k_a x]\}$$

  Note: non-identifiability (flip-flop).
- Assume

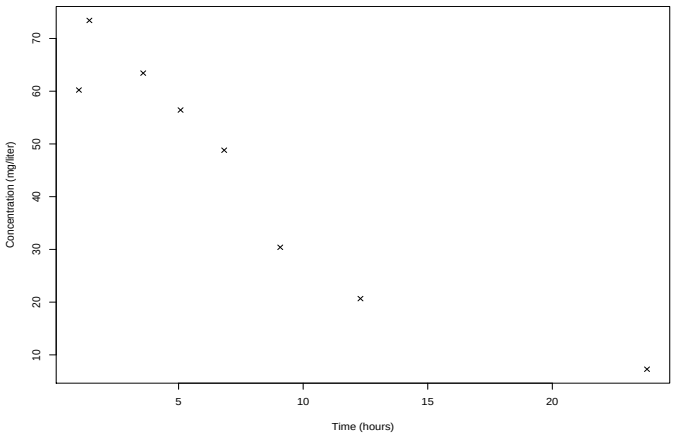
$$Y_i|\boldsymbol{\beta}, \sigma^2 = \text{LogNorm}\{\mu_i(x_i), \sigma^2\}.$$

  Mimics assay precision $\approx$ constant CV.

Pharmacokinetic Data Analysis

Concentration (y) as a function of time (x), obtained from a new-born baby following the administration of a 1mg dose of Theophylline.

Time	1.00	1.42	3.58	5.08	6.83	9.08	12.3	23.8
Conc	60.22	73.41	63.43	56.43	48.81	30.40	20.67	7.28



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

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

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

   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

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

   We also have

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

Marginal Moments

Mean:

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

Variance:

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

Covariance:

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

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

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

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

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.

Figures 18-20 contain summaries.

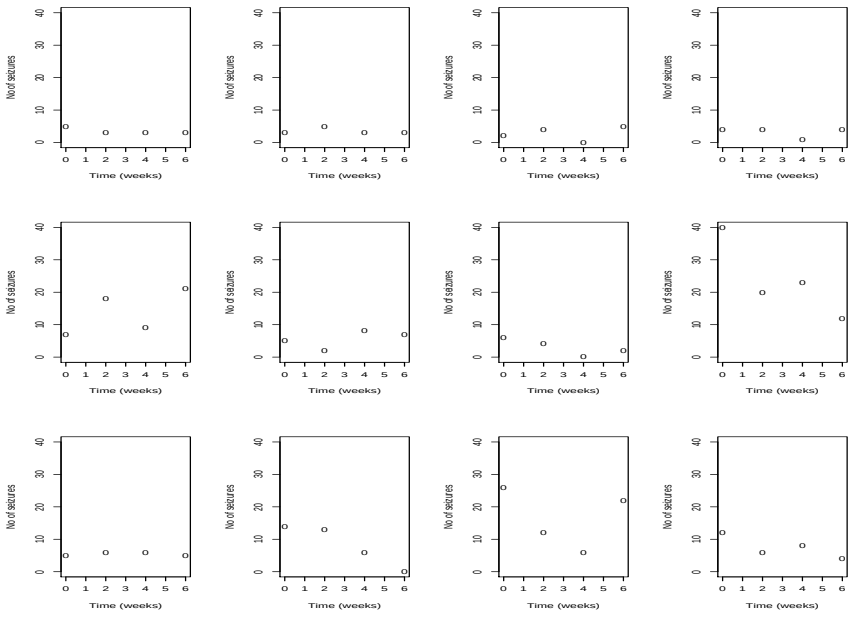


Figure 18: Number of seizures for selected individuals over time for placebo group.

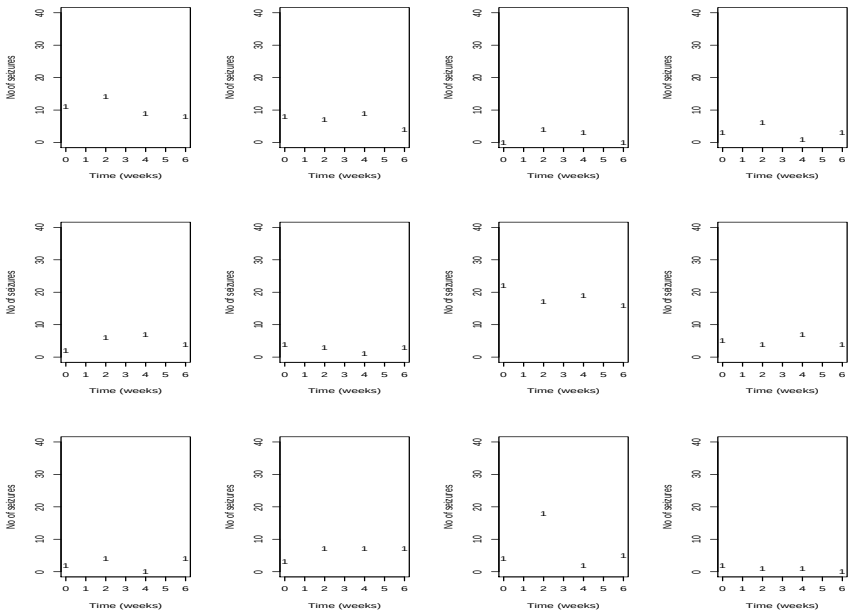


Figure 19: Number of seizures for selected individuals over time for progabide group.

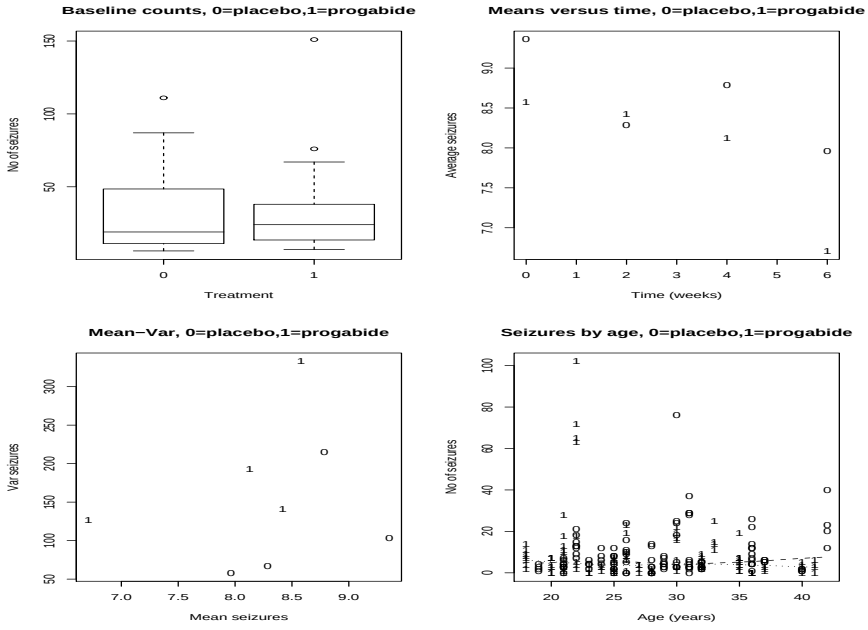


Figure 20: Summaries for seizure data.

A model for the seizure data

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 $j = 1, 2, 3, 4$, $i = 1, ..., 59$.

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

A marginal mean model is given by

$$E[Y_{ij}] = t_{ij} \exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{ij2} + \beta_3 x_{i1} x_{ij2})$$

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 11: Parameter interpretation.

Precise definitions:

- $\exp(\beta_0)$ is the expected number of seizures 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, 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 $j = 0$, 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. Hence $\exp(\beta_3)$ is the parameter of interest.

More colloquially:

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

Mixed Effects Model for Seizure Data

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

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

where

$$\boldsymbol{x}_{ij}\boldsymbol{\beta} = \beta_0 + \beta_1 x_{ij1} + \beta_2 x_{i2} + \beta_3 x_{ij1} x_{i2}.$$

Hence

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

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

The marginal mean is given by

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

and the marginal median by

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

The marginal variance is given by

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

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

For the marginal covariance

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

Hence for individual i we have variance-covariance matrix

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

where $\kappa = \mathrm{e}^{\sigma^2} - 1 > 0$. A deficiency of this model is that we only have a single parameter (σ^2) to control both excess-Poisson variability and dependence.

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 $\boldsymbol{b}_i$, and then carry out likelihood inference (using some form of approximation to evaluate the required integrals).

We first consider the conditional likelihood approach.

Conditional Likelihood

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

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

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

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

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

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

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

Conditional Likelihood for GLMMs

In the context of GLMMs we have

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

where

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

and

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

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

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

For simplicity we assume the canonical link function,

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

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

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

so that

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

and

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

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

Conditional Likelihood for the Poisson GLMM

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

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

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

Then

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

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

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

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

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

Hence

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

which is given by

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

After simplification:

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

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

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

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 $j = 1, 2, 3, 4$, $i = 1, ..., 59$.

A log-linear random intercept model is given by

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

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.

In the conditional likelihood notation:

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

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 $\boldsymbol{x}_{i1} = \boldsymbol{x}_{i2} = \boldsymbol{x}_{i3} = \boldsymbol{x}_{i4}$ 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 $i = 1, ..., 29$, 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 $i = 30, ..., 59$, where

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

.

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.

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_{ij2}, 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.

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 $\boldsymbol{D}$, and

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

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

Example: Log-linear Poisson regression GLMM

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

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

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

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

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.

Laplace Approximation

Let

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

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

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

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

Gaussian Quadrature

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

$$I = \int f(u)\ \mathrm{d}u,$$

is approximated by

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

for design points $u_1, ..., u_{n_w}$ and weights $w_1, ..., w_{n_w}$. 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

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

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

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

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

where

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

Now form

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

where

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

Then we have

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

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, $\boldsymbol{m}$, and variance-covariance matrix, $\boldsymbol{V}$ – known as adaptive quadrature.

We then form

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

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

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 Kullbach-Leibler divergence) has the same mean and variance.

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
```


Importance Sampling

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

We have

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

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

Hence we have the obvious estimator

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

where $\theta_i \sim_{iid} h(\cdot)$. We have $\mathrm{E}[\widehat{I}] = I$ and

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

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

We may estimate V via

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

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.

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.

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 first-order Laplace approximation.

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 `lme4` package the `lmer` function may be used to fit GLMMs using MLE/REML; the required integrals can be approximated using penalized quasi-likelihood, Laplace, or adaptive Gaussian quadrature.

GLMMs for the Seizure Data

PQL standard error for β_1 looks off here (doesn't tie in with later analyses). Adaptive quadrature option is not available for this model.

```
> library(lme4) # Need Matrix package version 0.995-5
> 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
```

```
Formula: y ~ x1 + x2 + x3 + (1 | ID) + offset(log(time))
   Data: seiz
 Family: poisson(log link)
      AIC      BIC    logLik deviance
 970.2882 988.7231 -480.1441 960.2882
Random effects:
     Groups        Name    Variance    Std.Dev.
         ID (Intercept)     0.60832     0.77995
# of obs: 295, groups: ID, 59
Estimated scale (compare to 1)  1.671041
Fixed effects:
             Estimate Std. Error z value  Pr(>|z|)
(Intercept)  1.032640   0.152524  6.7703 1.285e-11 ***
x1          -0.023848   0.210494 -0.1133   0.90980
x2           0.110798   0.046895  2.3627   0.01814 *
x3          -0.103681   0.065065 -1.5935   0.11105
```

The Laplace approach gives significantly different (and more reliable estimates).

Random intercepts and slopes

We may also allow the treatment effect to vary between individuals.

```
> 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
 802.2693 828.0782 -394.1347 788.2693
Random effects:
 Groups Name        Variance Std.Dev. Corr
 ID     (Intercept) 0.49990  0.70704
        x2          0.23189  0.48155  0.166
# of obs: 295, groups: ID, 59
Estimated scale (compare to 1)  1.403177
Fixed effects:
              Estimate Std. Error z value  Pr(>|z|)
(Intercept)  1.0712501  0.1398516  7.6599 1.861e-14 ***
x1           0.0494975  0.1927053  0.2569   0.79729
x2          -0.0023708  0.1078657 -0.0220   0.98246
x3          -0.3072281  0.1501527 -2.0461   0.04075 *
```

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 $\boldsymbol{D}$. 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.

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

Then

$$\mu = \log(\theta_1)\left(\frac{z_{q_2}}{z_{q_2} - z_{q_1}}\right) - \log(\theta_2)\left(\frac{z_{q_1}}{z_{q_2} - z_{q_1}}\right), \; \sigma = \frac{\log(\theta_1) - \log(\theta_2)}{z_{q_1} - z_{q_2}}. \quad (43)$$

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 $\sigma = \log 5/1.96 = 0.98$.

Gamma Priors

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

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 $\mathrm{Ga}(a, b)$ 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

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

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 $\pm t^d_{q/2}\lambda = \pm R$, where t^d_q is the q-th quantile of a Student t random variable with d degrees of freedom, to give $a = d/2$, $b = R^2 d/2(t^d_{q/2})^2$.

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, $\mathrm{Ga}(1, 0.0260)$.

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 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).

Implementation

Closed-form inference is unavailable, but MCMC is almost as straightforward as in the linear mixed model case. The joint posterior is

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

Suppose we have priors:

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

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

Integrated Nested Laplace Approximation (INLA)

Recently an approach has emerged that combines Laplace approximations and numerical integration in a very efficient manner, see Rue, Martino and Chopin (2008) for more detail.

The method is designed for "latent Gaussian model" — we describe in the context of a GLMM.

Suppose Y_{ij} is of exponential family form: $Y_{ij}|\theta_{ij}, \alpha \sim p(\cdot)$ 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 $i = 1, ..., m$ units, and $j = 1, ..., n_i$, measurements per unit.

Lwt $\mu_{ij} = \mathrm{E}[Y_{ij}|\theta_{ij}, \alpha]$ with

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

where $\boldsymbol{b}_i \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{D})$, and $\boldsymbol{\beta}$ is assigned a normal prior.

We also have priors for α (if not a constant) and $\boldsymbol{D}$ — these priors are non-normal.

Let $\boldsymbol{\gamma} = (\boldsymbol{b}, \boldsymbol{\beta})$ denote the $p \times 1$ vector of Gaussian parameters with $\pi(\boldsymbol{\gamma} \mid \boldsymbol{\theta}_1) \sim \mathrm{N}\{\mathbf{0}, \boldsymbol{Q}(\boldsymbol{D})\}$ and $\boldsymbol{\alpha} = (\alpha, \boldsymbol{D})$ be the hyperparameters which are not Gaussian.

Then

$$\begin{aligned}\pi(\boldsymbol{\gamma}, \boldsymbol{\alpha} \mid \boldsymbol{y}) &\propto \pi(\boldsymbol{\alpha})\pi(\boldsymbol{\gamma} \mid \boldsymbol{\alpha}) \prod_i p(\boldsymbol{y}_i \mid \boldsymbol{\gamma}, \boldsymbol{\alpha}) \\ &\propto \pi(\boldsymbol{\alpha}) \mid \boldsymbol{Q}(\boldsymbol{D}) \mid^{p/2} \exp\left\{-\frac{1}{2}\boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{Q}(\boldsymbol{D})\boldsymbol{\gamma} + \sum_i \log p(\boldsymbol{y}_i \mid \boldsymbol{\gamma}, \boldsymbol{\alpha})\right\}\end{aligned}$$

The INLA Algorithm

We wish to obtain the posterior marginals $\pi(\gamma_i \mid \boldsymbol{y})$ and $\pi(\alpha_i \mid \boldsymbol{y})$. We have

$$\pi(\gamma_i \mid \boldsymbol{y}) = \int \pi(\gamma_i \mid \boldsymbol{\alpha}, \boldsymbol{y}) \times \pi(\boldsymbol{\alpha} \mid \boldsymbol{y}) d\boldsymbol{\alpha}$$

which is evaluated via the approximation

$$\begin{aligned}\widetilde{\pi}(\gamma_i \mid \boldsymbol{y}) &= \int \widetilde{\pi}(\gamma_i \mid \boldsymbol{\alpha}, \boldsymbol{y}) \times \widetilde{\pi}(\boldsymbol{\alpha} \mid \boldsymbol{y}) d\boldsymbol{\alpha} &(44)\\ &= \sum_k \widetilde{\pi}(\gamma_i \mid \boldsymbol{\alpha}_k, \boldsymbol{y}) \times \widetilde{\pi}(\boldsymbol{\alpha}_k \mid \boldsymbol{y}) \times \Delta_k &(45)\end{aligned}$$

where

- $\widetilde{\pi}(\gamma_i \mid \boldsymbol{\alpha}, \boldsymbol{y})$ is approximated by Laplace (or other analytical approximations).
- For $\widetilde{\pi}(\boldsymbol{\alpha}_k \mid \boldsymbol{y})$ the mode is located and then the Hessian is approximated, from which a grid of points is found that cover the density.

Pros and Cons of INLA

Advantages:

- Quite widely applicable: GLMMs including temporal and spatial error terms.
- Very fast.
- An `R` package is available.

Disadvantages:

- Can't do NLMEMs (though could in principle).
- Difficult to access when the approximation is failing.
- `R` package is new, and so there are wrinkles.

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.

Model 5 We introduce "measurement error" into the model.

WinBUGS for Model 1

```
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()
}
```



```
# Data
list(k = 5, n = 59, t = c(8,2,2,2,2), x2 = c(0,1,1,1,1),
x1 = c(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
     1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1),
Y = structure(.Data = c(11,5,3,3,3,11,3,5,3,3,6,2,4,0,5,8,4,4,1,4,66,7,18,9,21,
27,5,2,8,7,12,6,4,0,2,52,40,20,23,12,23,5,6,6,5,10,14,13,6,0,52,26,12,6,22,
33,12,6,8,4,18,4,4,6,2,42,7,9,12,14,87,16,24,10,9,50,11,0,0,5,18,0,0,3,3,
111,37,29,28,29,18,3,5,2,5,20,3,0,6,7,12,3,4,3,4,9,3,4,3,4,17,2,3,3,5,
28,8,12,2,8,55,18,24,76,25,9,2,1,2,1,10,3,1,4,2,47,13,15,13,12,76,11,14,9,8,
38,8,7,9,4,19,0,4,3,0,10,3,6,1,3,19,2,6,7,4,24,4,3,1,3,31,22,17,19,16,
14,5,4,7,4,11,2,4,0,4,67,3,7,7,7,41,4,18,2,5,7,2,1,1,0,22,0,2,4,0,13,5,4,0,3,
46,11,14,25,15,36,10,5,3,8,38,19,7,6,7,7,1,1,2,3,36,6,10,8,8,11,2,1,0,0,
151,102,65,72,63,22,4,3,2,4,41,8,6,5,7,32,1,3,1,5,56,18,11,28,13,24,6,3,4,0,
16,3,5,4,3,22,1,23,19,8,25,2,3,0,1,13,0,0,0,0,12,1,4,3,2),.Dim = c(59,5)))

# Initial estimates
list(b=c(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0),
beta0=0, beta1=0, beta2=0, beta3=0, tau=1)
```

	Estimates (standard deviations)				
	Model 1	Model 2	Model 3	Model 4	Model 5
β_0	1.03 (0.15)	1.03 (0.15)	1.08 (0.13)	0.92 (0.15)	1.00 (0.18)
β_1	-0.024 (0.21)	-0.034 (0.21)	0.042 (0.19)	0.16 (0.20)	0.091 (0.24)
β_2	0.11 (0.047)	0.11 (0.047)	0.0045 (0.11)	-0.030 (0.11)	0.012 (0.10)
β_3	-0.11 (0.065)	-0.10 (0.065)	-0.31 (0.15)	-0.32 (0.15)	-0.30 (0.14)
σ_0	0.64 (0.13)	0.66 (0.13)	0.71 (0.072)	0.71 (0.10)	0.82 (0.084)
σ_1	–	–	0.473 (0.062)	0.399 (0.078)	
ρ	–	–	0.19 (0.16)	0.21 (0.21)	
σ_e	–	–	–	–	0.39 (0.033)

Table 12: Posterior means and standard deviations for Bayesian analysis of seizure data; σ_0 is the standard deviation of the random intercepts, σ_1 is the standard deviation of the random period effect, and ρ is the correlation between these random effects; σ_e is the standard deviation of the measurement error.

Poisson Model with a "nugget" effect

Recall the model

$$\begin{aligned} Y_{ij}|b_i &\sim \text{Poisson}(t_{ij}\exp(\boldsymbol{x}_{ij}\beta + b_i)) \\ b_i &\sim \text{N}(0, \sigma_0^2) \end{aligned}$$

has a single parameter only, σ_0 to allow for excess-Poisson variability *and* between-individual variability.

In the LMEM model we have

$$\begin{aligned} \text{E}[Y_{ij}|b_i] &= \boldsymbol{x}_{ij}\beta + b_i + \epsilon_{ij} \\ b_i &\sim \text{N}(0, \sigma_0^2) \\ \epsilon_{ij} &\sim \text{N}(0, \sigma_e^2) \end{aligned}$$

with b_i and ϵ_{ij} independent.

By analogy we might consider the model:

$$\begin{aligned} Y_{ij}|b_i, b_{ij} &\sim \text{Poisson}(t_{ij}\exp(\boldsymbol{x}_{ij}\beta + b_i + b_{ij})) \\ b_i &\sim \text{N}(0, \sigma_0^2) \\ b_{ij} &\sim \text{N}(0, \sigma_e^2) \end{aligned}$$

with b_i and b_{ij} independent.

We now two parameters to allow for between-individual variability, σ_0, and excess-Poisson variability, σ_e.

Unfortunately there is no simple marginal interpretation of σ_0 and σ_e:

$$\begin{aligned} \text{E}[Y_{ij}] &= t_{ij}\exp(\boldsymbol{x}_{ij}\boldsymbol{\eta} + \sigma_e^2/2 + \sigma_o^2) = \mu_{ij} \\ \text{var}(Y_{ij}) &= \mu_{ij} + \mu_{ij}^2(\text{e}^{\sigma_e^2} - 1)(\text{e}^{\sigma_0^2} - 1) \\ \text{cov}(Y_{ij}) &= t_{ij}t_{ik}\exp(\boldsymbol{x}_{ij}\boldsymbol{\beta} + \boldsymbol{x}_{ik}\boldsymbol{\beta})\text{e}^{\sigma_0^2}(\text{e}^{\sigma_0^2} - 1) \end{aligned}$$

Another possibility would be to start with a negative binomial distibution, and then introduce a random effect, b_i. This reveals the "heaven and hell" of mixed-effects models — we have a lot of flexibility in the models we can fit, but many formulations that are similar produce different marginal mean and covariance structures, and often there is no obvious "right" choice.

WinBUGS for Model 1

```
# Model 3 - Poisson lognormal for nugget also
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]+be[i,j]
        be[i,j] ~ dnorm(0,taue)
    }
    b[i] ~ dnorm(0,tau)
}
taue ~ dgamma(1,0.26)
tau ~ dgamma(1,0.26)
sigma <- sqrt(1/tau)
sigmae <- 1/sqrt(taue)
beta0 ~ dflat()
beta1 ~ dflat()
beta2 ~ dflat()
beta3 ~ dflat()
}
```

Seizure Data

Patient 49 had counts 151,102,65,72,63 under progabide — very surprising.

In DHLZ dropping this individual gave a parameter of interest of -0.30.

Posterior medians of b_{ij} for this individual ($i = 49, j = 0, 1, 2, 3, 4$) are:

-0.61, 0.61, 0.18, 0.27, 0.65, 0.15

Conclusions: there is evidence of a statistically significant treatment effect, under Model 4 the 95% credible interval on β_3 is (-0.60,-028).

Under model 5 the 95% credible interval on β_3 is (-0.59,-0.030).

Seizure Data: INLA implementation

```
> require (glmmAK); data(epileptic)
> epileptic$Y = epileptic$seizure; epileptic$x1 = epileptic$trt
> epileptic$x2 = as.integer(epileptic$visit>0)
> epileptic$t = 8-6*epileptic$x2
> epileptic$rand = 1:nrow(epileptic)
> formula1 = Y ~ x1 + x2 + I(x1*x2) + + f(id,model="iid",param=c(1,.26)) +
f(rand,model="iid",param=c(1,.26))
> inla1 = inla (formula1, family="poisson", data=epileptic,offset=I(log(t)))
> summary(inla1)
Fixed effects:
                 mean        sd 0.025quant   0.975quant    kld dist.
x1         0.05656008 0.2435636 -0.4230038  0.535726536 0.000000e+00
x2         0.03227324 0.1009346 -0.1668694  0.229812987 0.000000e+00
i(x1*x2)  -0.28425506 0.1402305 -0.5609546 -0.009858437 2.465190e-32
intercept  1.05710114 0.1767259  0.7086818  1.404250922 0.000000e+00
Model hyperparameters:
                     mean  sd    0.025quant 0.975quant
Precision.for..id.   1.535 0.305 1.007      2.201
Precision.for..rand. 6.766 1.110 4.846      9.199
```

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 $\widehat{\boldsymbol{\beta}}$ that satisfies

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

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

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

where $\boldsymbol{V}_\beta$ is given by

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

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

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

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

where $\boldsymbol{\Delta}_i = \mathrm{diag}[\mathrm{var}(Y_{i1}), ..., \mathrm{var}(Y_{in_i})]^{\mathrm{T}}$ and $\boldsymbol{R}_i$ 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 reccomended for practical purposes".
  - Zhao, Prentice and Self (1992): Assuming independence "can lead to important losses of efficiency".
  - Fitmaurice, Laird and Rotnitsky (1993): "important to obtain a close approximation to $\mathrm{cov}(\boldsymbol{Y}_i)$ in order to achieve high efficiency".

GEE for the Seizure Data

We have the log-linear model is given

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

and $\mathrm{var}(Y_{ij}) = \alpha \mu_{ij}$. Recall β_1 is baseline comparison of rates, β_2 is period effect in the placebo group and β_3 is treatment $\times$ period effect of interest.

Both quasi-likelihood and working independence GEE have estimating equation

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

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

	Estimates (standard errors)				
	Poisson	Quasi-Lhd	GEE Ind	GEE Exch	GEE AR(1)
$\beta_0^\star$	1.35 (0.034)	1.35 (0.15)	1.35 (0.16)	1.35 (0.16)	1.31 (0.16)
β_1	0.027 (0.047)	0.027 (0.21)	0.027 (0.22)	0.027 (0.22)	0.015 (0.21)
β_2	0.11 (0.047)	0.11 (0.21)	0.11 (0.12)	0.11 (0.12)	0.16 (0.11)
β_3	-0.10 (0.065)	-0.10 (0.29)	-0.10 (0.22)	-0.10 (0.22)	-0.13 (0.27)
α_1, α_2	1.0, 0	19.7, 0	19.4, 0	19.4, 0.78	20.0, 0.89

Table 13: 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 consequence of the balanced design.

Interpretation of Marginal and Conditional Coefficients

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

$$\mathrm{E}[Y \mid x] = \exp(\gamma_0 + \gamma_1 x)$$

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

$$\mathrm{E}[Y \mid x, b] = \exp(\beta_0 + \beta_1 x + b),$$

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

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

Hence for the log-linear model, e^{β_1} has the same marginal interpretation to e^{γ_1} (the marginal intercept is $\gamma_0 = \beta_0 + \sigma^{/}2$), 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).

In the model

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

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

$$\exp(\beta_1) = \frac{\mathrm{E}[Y \mid x, \boldsymbol{b}]}{\mathrm{E}[Y \mid x-1, \boldsymbol{b}]}.$$

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

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

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

so that there is no marginal mean interpretation of $\exp(\beta_1)$ (the latter is the marginal median).

Second Extension to GEE: Connected Estimating Equations, GEE2[a]

In GEE2, there are a *connected* set of joint estimating equations for $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$. Such an approach was proposed by Zhao and Prentice (1990), and Prentice and Zhao (1991).

This approach is particularly appealing if the variance-covariance model is of interest.

To motivate such a pair, consider the following model for a single individual with n *independent* observations:

$$Y_i|\boldsymbol{\beta},\alpha \sim_{ind} N\left\{\mu_i(\boldsymbol{\beta}),\Sigma_i(\boldsymbol{\beta},\alpha)\right\},$$

where, for example, we may have $\Sigma_i(\boldsymbol{\beta},\alpha)=\alpha\mu_i(\boldsymbol{\beta})^2$, $i=1,...,n$.

We have the likelihood

$$l(\boldsymbol{\beta},\alpha)=-\frac{1}{2}\sum_{i=1}^n \log\Sigma_i-\frac{1}{2}\sum_{i=1}^n\frac{(Y_i-\mu_i)^2}{\Sigma_i}.$$

[a]GEE1 is the method in which we have a single estimating equation, and a consistent estimator of $\boldsymbol{\alpha}$.

The score equations are given by

$$\begin{aligned}\frac{\partial l}{\partial\boldsymbol{\beta}} &= -\frac{1}{2}\sum_{i=1}^n\left(\frac{\partial\Sigma_i}{\partial\boldsymbol{\beta}}\right)^{\mathrm{T}}\frac{1}{\Sigma_i}+\sum_{i=1}^n\left(\frac{\partial\mu_i}{\partial\boldsymbol{\beta}}\right)^{\mathrm{T}}\frac{(Y_i-\mu_i)}{\Sigma_i}+\frac{1}{2}\sum_{i=1}^n\left(\frac{\partial\Sigma_i}{\partial\boldsymbol{\beta}}\right)^{\mathrm{T}}\frac{(Y_i-\mu_i)^2}{\Sigma_i^2}\\ &= \sum_{i=1}^n\left(\frac{\partial\mu_i}{\partial\boldsymbol{\beta}}\right)^{\mathrm{T}}\frac{(Y_i-\mu_i)}{\Sigma_i}+\sum_{i=1}^n\left(\frac{\partial\Sigma_i}{\partial\boldsymbol{\beta}}\right)^{\mathrm{T}}\frac{\left[(Y_i-\mu_i)^2-\Sigma_i\right]}{2\Sigma_i^2} \qquad (46)\end{aligned}$$

and

$$\begin{aligned}\frac{\partial l}{\partial\alpha} &= -\frac{1}{2}\sum_{i=1}^n\left(\frac{\partial\Sigma_i}{\partial\alpha}\right)^{\mathrm{T}}\frac{1}{\Sigma_i}+\frac{1}{2}\sum_{i=1}^n\left(\frac{\partial\Sigma_i}{\partial\alpha}\right)^{\mathrm{T}}\frac{(Y_i-\mu_i)^2}{\Sigma_i^2}\\ &= \sum_{i=1}^n\left(\frac{\partial\Sigma_i}{\partial\alpha}\right)^{\mathrm{T}}\frac{\left[(Y_i-\mu_i)^2-\Sigma_i\right]}{2\Sigma_i^2} \qquad (47)\end{aligned}$$

This pair of quadratic estimating functions, are unbiased given correct specification of the first two moments – so note that if the variance model is wrong, we are no longer guaranteed a consistent estimator of $\boldsymbol{\beta}$.

If it's correct, however, there will be a gain in efficiency.

Let

$$S_i = (Y_i - \mu_i)^2$$

with, under the model,

$$\begin{aligned} \mathrm{E}[S_i] &= \Sigma_i \\ \mathrm{var}(S_i) &= \mathrm{E}[S_i^2] - \mathrm{E}[S_i]^2 = 3\Sigma_i^2 - \Sigma_i^2 = 2\Sigma_i^2 \end{aligned}$$

Hence we can rewrite (46) and (47)

$$\begin{aligned} \frac{\partial l}{\partial \boldsymbol{\beta}} &= \sum_{i=1}^n \boldsymbol{D}_i^{\mathrm{T}} V_i^{-1}(Y_i - \mu_i) + \sum_{i=1}^n \boldsymbol{E}_i W_i^{-1}(S_i - \Sigma_i) \\ \frac{\partial l}{\partial \alpha} &= \sum_{i=1}^n \boldsymbol{F}_i W_i^{-1}(S_i - \Sigma_i) \end{aligned}$$

where $\boldsymbol{D}_i = \partial \mu_i / \partial \boldsymbol{\beta}$, $\boldsymbol{E}_i = \partial \Sigma_i / \partial \boldsymbol{\beta}$ and $\boldsymbol{F}_i = \partial \Sigma_i / \partial \alpha$.

This can be compared with the usual estimating equation specification:

$$\frac{\partial l}{\partial \boldsymbol{\beta}} = \sum_{i=1}^n \boldsymbol{D}_i^{\mathrm{T}} V_i^{-1}(Y_i - \mu_i).$$

GEE2 Continued

The general form of estimating equations, in the dependent data setting, is given by

$$\sum_{i=1}^m \begin{bmatrix} \boldsymbol{D}_i & \mathbf{0} \\ \boldsymbol{E}_i & \boldsymbol{F}_i \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} \boldsymbol{V}_i & \boldsymbol{C}_i \\ \boldsymbol{C}_i^{\mathrm{T}} & \boldsymbol{W}_i \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{Y}_i - \boldsymbol{\mu}_i \\ \boldsymbol{S}_i - \boldsymbol{\Sigma}_i \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}$$

where $\boldsymbol{D}_i = \partial \mu_i / \partial \boldsymbol{\beta}$, $\boldsymbol{E}_i = \partial \Sigma_i / \partial \boldsymbol{\beta}$ and $\boldsymbol{F}_i = \partial \Sigma_i / \partial \alpha$, and we have "working" variance-covariance structure

$$\begin{aligned} \boldsymbol{V}_i &= \mathrm{var}(\boldsymbol{Y}_i) \\ \boldsymbol{C}_i &= \mathrm{cov}(\boldsymbol{Y}_i, \boldsymbol{S}_i) \\ \boldsymbol{W}_i &= \mathrm{var}(\boldsymbol{S}_i) \end{aligned}$$

When $\boldsymbol{C}_i = \boldsymbol{0}$ we obtain

$$\begin{aligned}
\boldsymbol{G}_1(\boldsymbol{\beta}, \boldsymbol{\alpha}) &= \sum_{i=1}^m \boldsymbol{D}_i^{\mathrm{T}} \boldsymbol{V}_i^{-1} (\boldsymbol{Y}_i - \boldsymbol{\mu}_i) + \sum_{i=1}^m \boldsymbol{E}_i \boldsymbol{W}_i^{-1} (\boldsymbol{S}_i - \boldsymbol{\Sigma}_i) \\
\boldsymbol{G}_2(\boldsymbol{\beta}, \boldsymbol{\alpha}) &= \sum_{i=1}^m \boldsymbol{F}_i \boldsymbol{W}_i^{-1} (\boldsymbol{S}_i - \boldsymbol{\Sigma}_i)
\end{aligned}$$

which are the dependent data version of the normal score equations we obtained earlier.

Prentice and Zhao show that these equations arise from a so-called *quadratic exponential model*:

$$p(\boldsymbol{Y}_i | \boldsymbol{\theta}_i, \boldsymbol{\lambda}_i) = k_i^{-1} \exp[\boldsymbol{Y}_i^{\mathrm{T}} \boldsymbol{\theta}_i + \boldsymbol{S}_i^{\mathrm{T}} \boldsymbol{\lambda}_i + c_i(\boldsymbol{Y}_i)].$$

For example, $c_i = 0$ gives the multivariate normal.

For consistency of $\widehat{\boldsymbol{\beta}}$ we require models for both $\boldsymbol{Y}_i$ and $\boldsymbol{S}_i$ to be correct — increased eficiency if models are correct.