Generalised Linear Mixed Models

The development of linear models for the outcomes, historically, moved from the normal linear model, through generalised linear models, to generalised linear mixed models (GLMMs). GLMMs are a set of models for discrete or continuous data characterised by link functions to linear predictors, where there are random coefficients or random effects in the linear model.

We have seen random coefficient models already:

- Random coefficient models: they are just the models with prior distributions for the parameters.
  - example:
  - $y_i \sim N(\mu_i, \sigma^2)$, dependent variable
  - $\mu_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}$ independent variables $x_{1,} x_2$
  - $\beta_0 \sim N(0, a)$,
  - $\beta_1 \sim N(0, a)$,
  - $\beta_2 \sim N(0, a)$,
  - where $a = 10000$.
- Here the regression parameters are random and have prior distributions
- The model example above is a special case of a GLMM. It is a normal (Gaussian) linear model, with identity link, and regression parameters which are random variables.

- A binomial example (y dependent variable: two independent variables x_1, x_2) :
  - the outcome/response variable is binomial here:
  - $y_i \sim bin(\theta_i, n_i)$,
  - we usually use a logit link to a linear predictor:
  - $\text{logit}(\theta_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}$
  - $\beta_0 \sim N(0, a)$,
  - $\beta_1 \sim N(0, a)$,
  - $\beta_2 \sim N(0, a)$,
  - where $a = 10000$.

Random Effects

- Most studies in medicine or bioinformatics are not complete.
- That is, we dont always have complete knowledge of the system under study
- The implication of this is that there could be extra variation in the study which adds noise to the observed data. There could be:
  - a known confounder or explanatory variable which you *havent* measured
  - an unknown factor which you havent measured which is related to the outcome.
- What do we do about this?
  - we could ignore it (many frequentist statisticians do this)
  - we could try to get rid of it
  - we could model or estimate it

Approaches to random effects

- normal example:
  - Ignore them: $y_i = \mu_i + e_i$
    - here e_i = *pure error+extra noise*
    - this leaves the estimate of $\widehat{e_i} = y_i - \widehat{\mu_i}$ with the extra noise
  - Get rid of the extra noise: assume $y_i = \mu_i + v_i + e_i$ where
    - v_i is the extra noise and e_i is the pure error
    - either estimate v_i or remove it
- Problem: we must now consider identifiability : how do we separate out v_i and e_i.

- This is a major problem. Each term must have a different form if we are to have a hope of recovering it (as they are similar in form). For Bayesians this is not difficult as we can assume *strong priors* to make sure they are separable. However even if we assume a distribution for one or both of these effects we must still decide how to ~~dewala~~ with the overall problem.
- To *get rid of* the v_i we could
  - a) assume a prior dsitribution and concentrate the parameter out of the problem (ie integrate over the parameter space)
  - This can be done by finding the marginal posterior distribution of *y* given the other parameters (θ):

$$Mp = \int_v f(y|v,\theta)p(v)dv$$

OR

- b) assume a prior distribution and estimate the parameter i.e. use the posterior distribution

$$P \propto f(y|v,\theta)p(v)p(\theta)$$

  - there are many ways to handle this estimation
  - If you are fully Bayesian you will sample the posterior distribution in the usual way
  - Some people use special non-Bayesian methods for this (quasi-likelihood, penalised likelihood, estimating equations)

Bayesian Random Effect Modeling

We can specifiy a range of random effect models:

- Gaussian example:
  - $y_i \sim N(\mu_i + z_i, \sigma^2)$, dependent variable
  - $\mu_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}$ independent variables $x_{1,} x_2$
  - $z_i = v_i + u_i$ random effects (v_i, u_i)
  - this is a mixed model and a *particular* example of a GLMM
  - a simple Bayesian RE model:
  - $y_i = \mu_i + z_i + e_i$
  - what distribution should the REs have?
- We must separate the terms

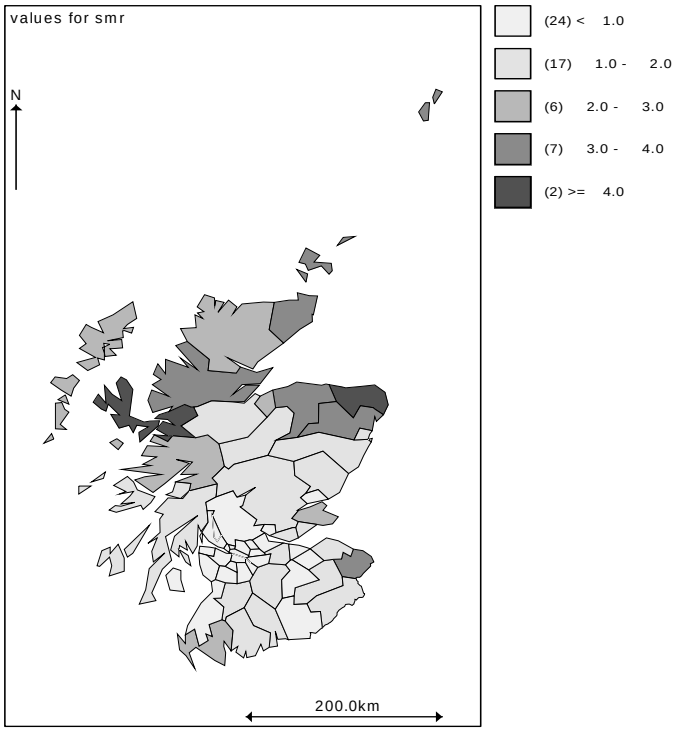
- What about
  - $z_i \sim N(0, \sigma_z^2)$
  - $e_i \sim N(0, \sigma_e^2)$
  - Assuming the variance are relatively different then we might be able to distinguish the terms
    - Can we use hyperpriors for this? YES!
    - $\sigma_z^2 \sim E(\pi)$
    - $\sigma_e^2 \sim E(\alpha\pi)$
    - where $\alpha \neq 1$ is a constant multiplier
    - This will make the variances different and so it will be more diffcult for any sampler to exchange the parameters

Another Example: Disease Mapping

- lets assume that we have a collection of m regions (eg counties, or zip codes or census tracts, municipalities). Withijn these you observe disease incidence. Let y_i be the count of disease in the i th small area.
- We often also calculate an expected rate for each small area e_i.
- This is calculated often by some standard rate eg
  - $e_i = p_i \sum_{ij} r_{ij} p_{ij}$ where r_{ij} is a strata specific rate in j th strata
  - and p_{ij} is the population of the j th strata in i th small area
  - Strata could be age groups or gender class or both

Mapping Models

- Basic model: $y_i \sim Poiss(e_i\theta_i)$
- θ_i is called the relative risk in the i th small area
- A crude ML estimate of θ_i is $\hat{\theta}_i = y_i/e_i$ the SMR for each region



- Scottish lip cancer example: lip cancer *mortality* for a period of years
- example given in WinBUGS manual

(GeoBUGS manual:lips)

- Lip cancer is relatively rare and is known to be related to exposure to sunlight (as one aetiological factor)

- **Data**:The rates of lip cancer in 56 counties in Scotland have been analysed by Clayton and Kaldor (1987) and Breslow and Clayton (1993). The form of the data includes the observed and expected cases (expected numbers based on the population and its age and sex distribution in the county), a covariate measuring the percentage of the population engaged in agriculture, fishing, or forestry, and the "position" of each county expressed as a list of adjacent counties.

```
model
{
 for (i in 1 : N) {
h[i]~dnorm(0.0,tau.h)
 O[i] ~dpois(mu[i])
 log(mu[i]) <- log(E[i]) + alpha0 + alpha1
*X[i]+h[i]
 RR[i] <- exp(alpha0 + alpha1 * X[i]+h[i]) #
Area-specific relative risk (for maps)
 }
```

```
 # Other priors:
 alpha0 ~dflat()
 alpha1 ~dnorm(0.0, 1.0E-5)
 tau.h ~dgamma(0.5, 0.0005)
 sigma.h <- sqrt(1 / tau.h) # standard
deviation
}
```

```
list(N = 56,
O = c( 9, 39, 11, 9, 15, 8, 26, 7, 6, 20,
 13, 5, 3, 8, 17, 9, 2, 7, 9, 7,
 16, 31, 11, 7, 19, 15, 7, 10, 16, 11,
 5, 3, 7, 8, 11, 9, 11, 8, 6, 4,
 10, 8, 2, 6, 19, 3, 2, 3, 28, 6,
 1, 1, 1, 1, 0, 0),
```

E = c(1.4, 8.7, 3.0, 2.5, 4.3, 2.4, 8.1, 2.3, 2.0, 6.6,

4.4, 1.8, 1.1, 3.3, 7.8, 4.6, 1.1, 4.2, 5.5, 4.4,

10.5,22.7, 8.8, 5.6,15.5,12.5, 6.0, 9.0,14.4,10.2,

4.8, 2.9, 7.0, 8.5,12.3,10.1,12.7, 9.4, 7.2, 5.3,

18.8,15.8, 4.3,14.6,50.7, 8.2, 5.6, 9.3,88.7,19.6,

3.4, 3.6, 5.7, 7.0, 4.2, 1.8),

X = c(16,16,10,24,10,24,10, 7, 7,16,

7,16,10,24, 7,16,10, 7, 7,10,

7,16,10, 7, 1, 1, 7, 7,10,10,

7,24,10, 7, 7, 0,10, 1,16, 0,

1,16,16, 0, 1, 7, 1, 1, 0, 1,

1, 0, 1, 1,16,10))

```
list( tau.h=1,alpha0 = 0, alpha1 = 0,
h=c(0,0,0,0,0,NA,0,NA,0,0,
 NA,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))
```

We can run this model for a large number of iterations (until convergence of course)

Then we can map the relevant RR and h components.

How can we simplify this model?

- Run without covariate and check DIC for both models?
- Does the covaraite explain any of the variation in lip cancer incidence?