Spatio-temporal Methods in Environmental Epidemiology

Alexandra M. Schmidt

Session 3 - Hierarchical Models

Department of Epidemiology, Biostatistics and Occupational Health
McGill University

alexandra.schmidt@mcgill.ca

Applied Bayesian Statistics School 2025

University of Genova, Department of Architecture and Design
03-06, June 2025

Hierarchical Priors

A good strategy to specify the prior distribution or to describe better the experimental situation, is to divide it into stages or into a hierarchy (Lindley & Smith, 1972). The prior specification is made in two phases:

1. structural, for the division into stages;
2. subjective, for quantitative specification at each stage.

Example: Let $\mathbf{Y} = (Y_1, \cdots, Y_n)$ with $Y_i \sim N(\theta_i, \sigma^2)$, with σ^2 known. There are many choices for specification of the prior for $\theta = (\theta_1, \cdots, \theta_n)$. Some options:

Hierarchical Priors

- θ_i's are independent $\Rightarrow p(\theta) = \prod_i p(\theta_i)$.
- θ_i's are a sample from a population with $p(\theta|\lambda)$ where λ contains the parameters describing the population
  $\Rightarrow p(\theta|\lambda) = \prod_{i=1}^n p(\theta_i \mid \lambda)$
  This specification corresponds to the first stage. To complete the prior setting, it is necessary to specify the second stage: the distribution of λ, $p(\lambda)$.
  Note that $p(\lambda)$ corresponds to the second stage and **does not** depend on the first stage. Recall that

$$p(\theta) = \int p(\theta, \lambda) d\lambda = \int p(\theta \mid \lambda) p(\lambda) d\lambda$$

$$= \int \prod_{i=1}^{n} p(\theta_i \mid \lambda) p(\lambda) d\lambda.$$

Note that θ_i's are supposed exchangeable.

Hierarchical Priors

Since the distribution of λ is independent of the first stage, it can be stated as:

1. Concentrated: $p(\lambda = \lambda) = 1$
2. Discrete: $p(\lambda = \lambda_j) = p_j$, $j = 1, \cdots, k$, with $\sum_j p_j = 1$. In this case the distribution of θ will be a finite mixture of the densities $p(\theta \mid \lambda_j)$ with weights p_j, $j = 1, \cdots, k$.
3. Continuous: as before, the distribution of θ will be a continuous mixture of $p(\theta \mid \lambda)$ with weights given by $p(\lambda)$.

In the example we can assume: $\theta_i \sim N(\mu, \tau^2), i = 1, 2, \cdots, n$ then $\lambda = (\mu, \tau^2)$. Assuming $p(\tau^2 = \tau_0) = 1$ and μ normally distributed then θ has a multivariate normal distribution.
OR assuming $p(\mu = \mu_0) = 1$ and τ^{-2} with a gamma prior distribution implies that θ has a multivariate Student t distribution.

Hierarchical Priors

Nothing prevents these ideas from going further into the hierarchy. For example, the distribution of λ can depend on ϕ, in this case,

$$p(\theta) = \int_{\Phi} \int_{\Lambda} p(\theta|\lambda) p(\lambda \mid \phi) p(\phi) d\lambda d\phi.$$

The parameters λ and ϕ are called hyperparameters and are introduced to ease the prior specification. In practice, it is very hard to interpret the parameters of third or higher stages, so it is common practice to use a non-informative prior for these levels.

A three-stage hierarchical model

Consider

$$\mathbf{y}_i = \mathbf{x}_i\beta + \mathbf{z}_i\mathbf{b}_i + \varepsilon_i$$

with $\mathbf{b}_i$ and ε_i independent and distributed as $\mathbf{b}_i \mid \mathbf{D} \sim N_{q+1}(\mathbf{0}, \mathbf{D})$, and $\varepsilon_i \mid \sigma_\varepsilon^2 \sim N_{n_i}(\mathbf{0}, \sigma_\varepsilon^2\mathbf{I})$, $i = 1, 2, \cdots, m$.

- The second stage assumption for $\mathbf{b}_i$ can be motivated using the concept of exchangeability. Then it can be shown that

  $$p(\mathbf{b}_1, \cdots, \mathbf{b}_m) = \int \prod_{i=1}^{m} p(\mathbf{b}_i \mid \phi)\pi(\phi)d\phi$$

  so that $[\mathbf{b}_1, \cdots, \mathbf{b}_m]$ are *conditionally* independent, given *hyperparameters* ϕ, with the hyperparameters having a distribution known as *hyperprior*.
- Hence, we have a two-stage hierarchical model

  $$\begin{aligned} \mathbf{b}_i \mid \phi &\sim_{\text{i.i.d}} p(\cdot \mid \phi) \\ \phi &\sim_{\text{i.i.d}} \pi(\cdot) \end{aligned}$$

A three-stage hierarchical model

- Parametric choices for $p(\cdot \mid \phi)$ and $\pi(\cdot)$ are based on the application, though computational convenience may also be a consideration. In general, if collections of units cluster due to an observed covariate that we believe will influence $\mathbf{b}_i$, then our prior should reflect this.
- This framework contrasts with the sampling theory approach in which the random effects are assumed to be a random sample from a hypothetical *infinite* population.

The three-stage model is:

- Stage one: Likelihood:

$$p(\mathbf{y}_i \mid \beta, \mathbf{b}_i, \sigma_\varepsilon^2) \qquad i = 1, 2, \cdots, m$$

- Stage two: Random Effects prior:

$$\mathbf{b}_i \mid \phi \sim_{\text{i.i.d}} p(\cdot \mid \phi)$$

- Stage three: Hyperprior

$$p(\beta, \mathbf{D}, \sigma_\varepsilon^2).$$

Hyperpriors : It is common to assume independent priors:

$$p(\beta, \mathbf{D}, \sigma_\varepsilon^2) = \pi(\beta)\pi(\mathbf{D})\pi(\sigma_\varepsilon^2).$$

A three-stage hierarchical model

Hyperpriors :

It is common to assume independent priors:

$$p(\beta, \mathbf{D}, \sigma_\varepsilon^2) = \pi(\beta)\pi(\mathbf{D})\pi(\sigma_\varepsilon^2).$$

- A multivariate normal distribution for β and an inverse gamma distribution for σ_ε^2 are often reasonable choices, since they are flexible
- These choices also lead to conditional distributions that have convenient forms for Gibbs sampling
- The prior specification of **D** is less straightforward
  - If **D** is a diagonal matrix with elements σ_k^2, $k = 0, 1, \cdots, q$ then an obvious choice is

    $$\pi(\sigma_0^2, \cdots, \sigma_q^2) = \prod_{k=0}^{q} \text{IGa}(a_k, b_k)$$

    $\text{IGa}(a_k, b_k)$ inverse gamma distribution with known parameters a_k and b_k
  - A prior for non-diagonal **D** is more troublesome; there are $(q+2)(q+1)/2$ elements, with the restriction that the matrix is positive definite → natural choice is an Inverse Wishart distribution

Inverse Wishart Distribution

Suppose $\mathbf{Z}_1, \cdots, \mathbf{Z}_r \sim_{i.i.d} N_p(\mathbf{0}, \mathbf{S})$, with $\mathbf{S}$ a non-singular variance-covariance matrix, and let

$$\mathbf{W} = \sum_{j=1}^{r} \mathbf{Z}_j \mathbf{Z}_j^T.$$

Then $\mathbf{W}$ follows a Wishart distribution, denoted $W(r, S)$, with probability density function

$$p(\mathbf{w}) = c^{-1} |\mathbf{w}|^{(r-p-1)/2} \exp\left\{-\frac{1}{2} tr(\mathbf{w}\mathbf{S}^{-1})\right\}$$

where $c = 2^{rp/2}\Gamma_p(r/2)|\mathbf{S}|^{r/2}$ with $\Gamma_p(r/2) = \pi^{p(p-1)/4} \prod_{j=1}^{p} \Gamma[(r+1-j)/2]$, the generalized gamma function. We require $r > p-1$ for a proper density. The mean is

$$E[\mathbf{W}] = r\mathbf{S}$$

Taking $p = 1$ yields a Gamma distribution with parameters $r/2$ and $1/(2S)$. Further, taking $S = 1$ gives a χ^2_r r.v.

Inverse Wishart Distribution

- If $\mathbf{W} \sim W(r, \mathbf{S})$, the distribution of $\mathbf{D} = \mathbf{W}^{-1}$ is known as the inverse Wishart distribution, denoted $InvW(r, \mathbf{S})$, with density

$$p(\mathbf{d}) = c^{-1}|\mathbf{d}|^{-(r+p+1)/2} \exp\left\{-\frac{1}{2}tr(\mathbf{d}^{-1}\mathbf{S})\right\}$$

  where c is like before.

- $E(\mathbf{D}) = \frac{\mathbf{S}^{-1}}{r-p-1}$ and is defined for $r > p+1$. If $p = 1$ we recover the inverse gamma distribution, $IGa(r/2, 1/2S)$, with $E(D) = \frac{1}{S(r-2)}$, and $Var(D) = \frac{1}{S^2(r-2)(r-4)}$, so that small values of r gives a more dispersed distribution (which is true for general p).
- One way of thinking about prior specification is to imagine that the prior data for the precision consists of observing r multivariate normal r.v. with empirical covariance matrices $\mathbf{R} = \mathbf{S}^{-1}$.
- We summarize samples from the Wishart via marginal distributions for σ_0,σ_1, and ρ since these are more interpretable.
- Example $\mathbf{D}^{-1} \sim W_2(r, \mathbf{R}^{-1})$, $r = 4$, $E[\mathbf{D}] = \frac{\mathbf{R}}{4-1-2} = \mathbf{R}$ with $\mathbf{R} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$.

McGill

A three-stage hierarchical model - Implementation

For simplicity, assume $\mathbf{x}_i = \mathbf{z}_i$. It is convenient to reparametrize in terms of $[\beta_1, \beta_2, \cdots, \beta_m, \tau, \beta, \mathbf{W}]$, where $\beta_i = \beta + \mathbf{b}_i$, $\tau = \sigma_\varepsilon^2$, and $\mathbf{W} = \mathbf{D}^{-1}$. The joint posterior is proportional to

$$p(\beta_1, \beta_2, \cdots, \beta_m, \tau, \beta, \mathbf{W} \mid \mathbf{y}) \propto \prod_{i=1}^{m} [p(\mathbf{y_i} \mid \beta_i, \tau) p(\beta_i \mid \beta, \mathbf{W})] \pi(\beta)\pi(\tau)\pi(\mathbf{W}) \tag{1}$$

with priors

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

- Marginal distributions, and summaries of these distributions are not available in closed form.
- Possibilities:
  - INLA (`http://www.r-inla.org/`) is ideally suited to the LMM
  - MCMC is an alternative, which we describe next.

A three-stage hierarchical model - MCMC Implementation

Posterior full conditional distributions

- $\beta \mid \beta_1, \cdots \beta_m, \mathbf{W} \sim N_{q+1}\left[(m\mathbf{W} + \mathbf{V}_0^{-1})\left(\mathbf{W}\sum_{i=1}^m \beta_i + \mathbf{V}_0^{-1}\beta_0\right), \left(m\mathbf{W} + \mathbf{V}_0^{-1}\right)^{-1}\right]$
- $\tau \mid \beta_i, \mathbf{y} \sim Ga\left[a_0 + \frac{\sum_{i=1}^m n_i}{2}, b_0 + \frac{1}{2}\sum_{i=1}^m (\mathbf{y}_i - \mathbf{x}_i\beta_i)^T(\mathbf{y}_i - \mathbf{x}_i\beta_i)\right]$
- $\beta_i \mid \tau, \mathbf{W}, \mathbf{y} \sim N_{q+1}\left[(\tau\mathbf{x}_i^T\mathbf{x}_i + \mathbf{W})^{-1}(\tau\mathbf{x}_i^T\mathbf{y}_i + \mathbf{W}\beta), (\tau\mathbf{x}_i^T\mathbf{x}_i + \mathbf{W})^{-1}\right]$
- $\mathbf{W} \mid \beta_1, \cdots, \beta_m, \beta \sim W_{q+1}\left[r + m, \left(\mathbf{R} + \sum_{i=1}^m (\beta_i - \beta)(\beta_i - \beta)^T\right)^{-1}\right]$

  Note that $E[\mathbf{D} \mid \beta_1, \cdots, \beta_m, \beta] = \frac{\mathbf{R} + \sum_{i=1}^m \left(\beta_i - \beta)(\beta_i - \beta\right)^T}{r + m - q - 2}$, suggesting that it is better to pick a small $\mathbf{R}$, since a large $\mathbf{R}$ will always dominate the sum of squares. If m is small the prior is always influential.

Example - dental growth data

- Table 1 records dental measurements of the distance in millimeters from the center of the pituitary gland to the pteryo-maxillary fissure in 11 girls and 16 boys at the ages of 8, 10, 12 and 14 years.
- Here we have an example of repeated measures or longitudinal data.
- Figure 1 plots these data and we see that dental growth for each child increases in an approximately linear fashion.
- One common aim of such studies is to identify the within-individual and between-individual sources of variability.

Example - dental growth data

Girls	8	10	12	14
1	21	20	21.5	23
2	21	21.5	24	25.5
3	20.5	24	24.5	26
4	23.5	24.5	25	26.5
5	21.5	23	22.5	23.5
6	20	21	21	22.5
7	21.5	22.5	23	25
8	23	23	23.5	24
9	20	21	22	21.5
10	16.5	19	19	19.5
11	24.5	25	28	28
Boys	8	10	12	14
1	26	25	29	31
2	21.5	22.5	23	26.5
3	23	22.5	24	27.5
4	25.5	27.5	26.5	27
5	20	23.5	22.5	26
6	24.5	25.5	27	28.5
7	22	22	24.5	26.5
8	24	21.5	24.5	25.5
9	23	20.5	31	26
10	27.5	28	31	31.5
11	23	23	23.5	25
12	21.5	23.5	24	28
13	17	24.5	26	29.5
14	22.5	25.5	25.5	26
15	23	24.5	26	30
16	22	21.5	23.5	25

Table: Dental growth data for girls and boys.

Example - dental growth data

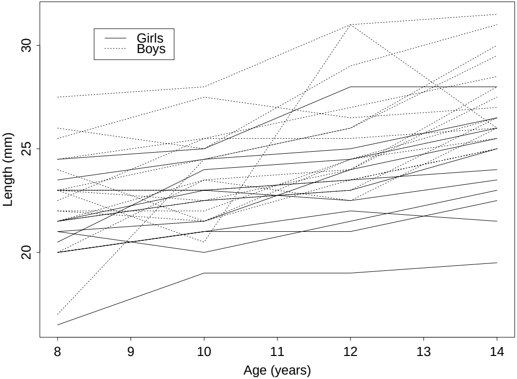


Figure: Dental growth data for girls and boys.

A three-stage hierarchical model - Dental Growth Example

The three-stage model is:

- Stage one: Likelihood:

$$y_{ij} = \beta_{i0} + \beta_{1i} t_j + \varepsilon_{ij}, \quad \text{with } \varepsilon_{ij} \mid \tau \sim_{iid} N(0, \tau^{-1})\, i = 1, 2, \cdots, m$$

- Stage two: Random Effects prior:

$$\beta_i \mid \beta, \mathbf{D} \sim_{\text{i.i.d}} N_2(\beta, \mathbf{D})$$

$$\beta_i = \begin{pmatrix} \beta_{i0} \\ \beta_{i1} \end{pmatrix}, \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \mathbf{D} = \begin{pmatrix} \sigma_0^2 & \sigma_{01} \\ \sigma_{10} & \sigma_1^2 \end{pmatrix}$$

- Stage three: Hyperprior

$$p(\beta, \mathbf{D}, \tau) = \pi(\tau) \times \pi(\beta) \times \pi(\mathbf{D})$$

If we assume improper priors for β and τ, we have that

$$p(\beta, \mathbf{D}, \tau) \propto \tau^{-1} \pi(\mathbf{D})$$

with $D^{-1} \sim W_2(r, \mathbf{R}^{-1})$

A three-stage hierarchical model - Dental Growth Example

See file `OrthogirlsRandomInterceptRandomSlopeModel.r`

Linear Mixed Effects Models

Recall that for Gaussian data, we have the following Linear Mixed Effects Model:
The **conditional** model is given by:

$$Y_i = X_i\beta + Z_i b_i + \varepsilon_i, \tag{2}$$

where ε_i is an $n_i \times 1$ zero mean vector of error terms.
Assuming $\varepsilon_i \sim \mathcal{N}(0, \Xi)$, then conditionally,

$$\text{Var}(Y_i) = \Xi.$$

The **marginal** model is given by:
First, assuming $b_i \sim \mathcal{N}(0, \Gamma)$, then

$$\begin{aligned} E[Y_i] &= \mu_i(\beta) = X_i\beta, \\ \text{Var}(Y_i) &= Z_i \Gamma Z_i' + \Xi, \\ \text{cov}(Y_i, Y_{i'}) &= 0,\ i \neq i'. \end{aligned}$$

Linear Mixed Effects Models

- When we speak about marginal or conditional models, we are referring to the random effects; conditional dependence on the covariates is assumed.
- Because the random effects have zero mean **on the outcome scale**, fixed effect parameters have the same interpretation under either a marginal or a conditional view of the model for Gaussian data.
- This is **not** the case for categorical data, and so the distinction between marginal and conditional models becomes more important.
- Which approach to choose will depend primarily on the research question. In general, conditional models are more appropriate when interest lies in particular clusters/individuals in the sample.

A note on interpretation

- The fixed effects model $\log E[Y_{ij} \mid \alpha] = \alpha_0 + \alpha_1 X_i$ provides population-average contrasts, and is a marginal model.
- The mixed effects model $\log E[Y_{ij} \mid \beta, b_i] = \beta_0 + \beta_1 X_{ij} + b_i Z_i$ provides subject-specific contrasts, and is a conditional model.
  - E.g.: for Poisson data, the marginal RR associated with a 1 unit change in X in a random intercepts model, not conditioning on the random effect gives:

$$\begin{aligned} \text{Expected RR for } X &= E\left[exp\{(\beta_0 + \beta_1(X+1) + b_i) - \right. \\ &\qquad \left. (\beta_0 + \beta_1 X + b_k)\}\right] \\ &= E\left[exp(\beta_1 + b_i - b_k)\right] \\ &= exp(\beta_1) E\left[exp(b_i - b_k)\right] \end{aligned}$$

which may not equal $exp(\beta_1)$ for $b_i \neq b_k$. In the typical case in which we assume $b_i \sim \mathcal{N}(0, \sigma_0^2)$, we can instead look at the ratio of expected risks and find $E\left[exp(\beta_0 + \beta_1(X+1) + b_i)\right] / E\left[exp(\beta_0 + \beta_1 X + b_k)\right] = exp(\beta_1)$.

A note on interpretation

- Note, however, that if we condition on b_i we have the following
  - for Poisson data, the RR associated with a 1 unit change in X in a random intercepts model given the random effect, we have:

$$\begin{aligned}\text{RR for } X &= E\left[exp(\beta_0 + \beta_1(X+1) + b_i) - \right.\\ &\qquad\qquad \left.(\beta_0 + \beta_1 X + b_i)\right]\\ &= E\left[exp(\beta_1)\right]\\ &= exp(\beta_1)\end{aligned}$$

- The mixed model parameters are subject-specific, and measure the change in conditional log of the mean response for units with covariates X in the group defined by the random effect b_i.
  - This can problematic for the interpretation of parameters associated with *between*-cluster covariates.

A note on interpretation

With reference to a random effects model for binomial data, Neuhaus, Kalbfleish, and Hauck (1991) write:

> Although the cluster-specific model seems to provide the more unified approach, parameter interpretation in these models is difficult. The cluster-specific model presupposes the existence of latent risk groups indexed by [the random effect], and parameter interpretation is with reference to these groups.

Generalized Linear Mixed Models

- We can use the usual properties of exponential families to consider the conditional model specified by the GLMM (just as we did for uncorrelated data in a GLM).
- We can also derive the properties of the marginal model, although these may be difficult to calculate in closed form.
- The basic idea of a GLMM is much like a LME: we add the random effects on the **linear** scale.

Generalized Linear Mixed Models

A GLMM is defined by

1. Random component: $Y_{ij}|\theta_{ij}, b_i, \phi \sim_{iid} p(\cdot)$ where $p(\cdot)$ is a member of the exponential family, that is

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

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

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

$$g(\mu_{ij}) = x_{ij}\beta + z_{ij}b_i,$$

   so that we have introduced random effects into the linear predictor. The above defines the **conditional** part of the model. If we wish to consider a marginal model, the random effects are then assigned a distribution; in a GLMM this is typically $b_i \sim_{iid} \mathcal{N}(0, \Gamma)$.

Generalized Linear Mixed Models

There are, in general, two approaches to inference for a GLMM from a likelihood perspective:

1. Conditional inference: Condition on random effects in order to eliminate them from from the likelihood.
2. Full likelihood inference: Make a distributional assumption about the random effects and then carry out the usual likelihood inference on the full distribution (some form of approximation will be required to evaluate the required integrals).

Conditional moments

Mean:

$$E[Y_{ij}|b_i] = E[\mu_{ij}] = g^{-1}(x_{ij}\beta + z_{ij}b_i).$$

Variance:

$$\text{Var}(Y_{ij}|b_i) = b''(\theta_{ij})a(\phi).$$

Covariance:

$$\text{Cov}(Y_{ij}, Y_{ik}|b_i) = 0.$$

Marginal moments

Mean:

$$\begin{aligned} E[Y_{ij}] &= E\{E[Y_{ij}|b_i]\} \\ &= E[\mu_{ij}] = E_b[g^{-1}(x_{ij}\beta + z_{ij}b_i)]. \end{aligned}$$

Variance:

$$\begin{aligned} \text{Var}(Y_{ij}) &= E[\text{Var}(Y_{ij}|b_i)] + \text{Var}(E[Y_{ij}|b_i]) \\ &= \phi E_b[\text{Var}\{g^{-1}(x_{ij}\beta + z_{ij}b_i)\}] + \text{Var}_b[g^{-1}(x_{ij}\beta + z_{ij}b_i)]. \end{aligned}$$

Covariance:

$$\begin{aligned} \text{Cov}(Y_{ij}, Y_{ik}) &= E[\text{Cov}(Y_{ij}, Y_{ik}|b_i)] + \text{Cov}(E[Y_{ij}|b_i], E[Y_{ik}|b_i]) \\ &= \text{Cov}\{g^{-1}(x_{ij}\beta + z_{ij}b_i), g^{-1}(x_{ik}\beta + z_{ik}b_i)\} \\ &\neq 0, \end{aligned}$$

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.

McGill

Example: Log-Linear regression for seizure data

Data on seizures were collected on 59 epileptics. For each patient the no. of seizures were recorded during a baseline period of 8 weeks, after which patients were randomized to treatment with the drug progabide or to placebo. The no. of seizures was then recorded in 4 consecutive 2-week periods. Patient age was also available. Let

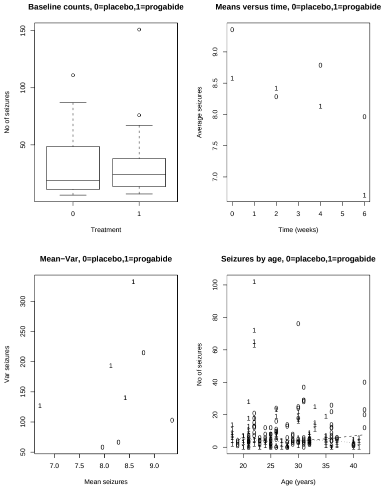
$$\begin{aligned} Y_{ij} &= \text{number of seizures on patient } i \text{ at occasion } j \\ t_{ij} &= \text{length of observation period on patient } i \text{ at occasion } j \\ x_{i1} &= \text{0/1 if patient } i \text{ was assigned placebo/progabide} \\ x_{ij2} &= \text{0/1 if } j = 0/1,2,3,4 \end{aligned}$$

with $t_{ij} = 8$ if $j = 0$ and $t_{ij} = 2$ if $j =\geq 1$, for all i.

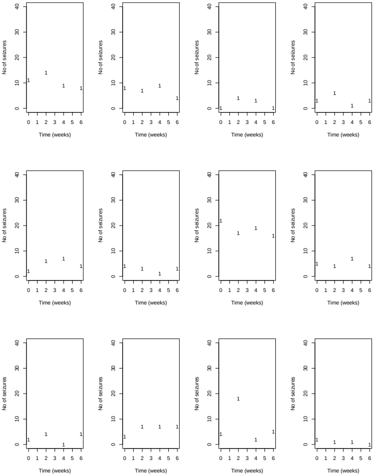
The question: does progabide reduce the number of seizures?

We begin with exploratory plots of the data.

Example: seizure data

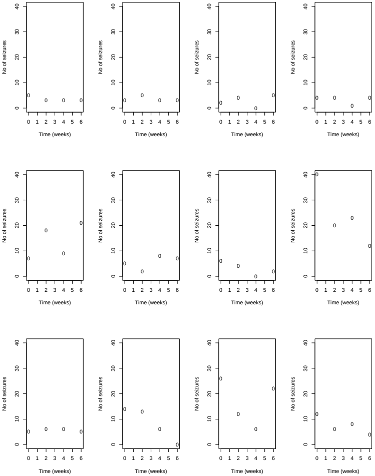


Example: Log-Linear regression for seizure data



No. of seizures for selected individuals over time (progabide).

Example: Log-Linear regression for seizure data



No. of seizures for selected individuals over time (placebo).

Generalized Mixed Effects Model for seizure data

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

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

where

$$x_{ij}\beta = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{ij2} + \beta_3 x_{i1} x_{ij2}.$$

Hence

$$E[Y_{ij}|b_i] = \mu_{ij} = t_{ij} \exp(x_{ij}\beta + b_i), \quad \text{Var}(Y_{ij}|b_i) = \mu_{ij}.$$

Stage 2?: $b_i \sim_{iid} \mathcal{N}(0, \sigma^2)$.
The marginal mean is given by

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

and the marginal median by $t_{ij} \exp(x_{ij}\beta)$.

Generalized Mixed Effects Model for seizure data

The marginal variance is given by

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

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

The marginal covariance is

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

Generalized Mixed Effects Model for seizure data

Hence for individual *i* we have variance-covariance matrix

$$\left[\begin{array}{cccc} \mu_{i1}+\mu_{i1}^2\kappa & \mu_{i1}\mu_{i2}\kappa & \ldots & \mu_{i1}\mu_{in_i}\kappa \\ \mu_{i2}\mu_{i1}\kappa & \mu_{i2}+\mu_{i2}^2\kappa & \ldots & \mu_{i2}\mu_{in_i}\kappa \\ \ldots & \ldots & \ldots & \ldots \\ \mu_{in_i}\mu_{i1}\kappa & \mu_{in_i}\mu_{i2}\kappa & \ldots & \mu_{in_i}+\mu_{in_i}^2\kappa \end{array}\right],$$

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

McGill

Model formulation

- A Bayesian approach to inference for a GLMM requires a prior distribution for β and α (α is the scale parameter)
- As with the linear mixed model, a proper prior is required for the matrix D.
- A proper prior is not always necessary for β, but care is required
- The exponential family and canonical link lead to a likelihood that is well behaved (in particular, with respect to tail behavior), though it is safer to specify a proper prior since impropriety of the posterior can occur in some cases (e.g., with noncanonical links or when counts are either equal to zero or to the denominator
- Closed-form inference is unavailable, but MCMC is almost as straightforward as in the LMM, and INLA is also available though the approximation is not always accurate for the GLMM

Model formulation

- We assume that

$$\begin{aligned} g(\mu_{ij}) &= x_{ij}\beta + z_{ij}b_i \\ b_i &\sim N(0, D) \end{aligned}$$

- Let $W = D^{-1}$, and assume there are no unknown parameters at stage one of the model (i.e. $\alpha = 1$). The posterior distribution is proportional to

$$p(\beta, W, b \mid y) \propto \prod_{i=1}^{m} [p(y_i \mid \beta, b_i) p(b_i \mid W)] \pi(\beta, W)$$

- Typically we assume independent hyperpriors:

$$\pi(\beta, W) = \pi(\beta)\pi(W),$$

  with $\beta \sim N_{p+1}N(\beta_0, V_0)$ and $W \sim Wish_{p+1}(r, R^{-1})$, where $Wish_{p+1}(r, R^{-1})$ denotes a Wishart distribution of dimension $p+1$ with r degrees of freedom and scale matrix R^{-1}
- The conditional distribution for W is unchanged from the LMM case
- There are no closed-form conditional distributions for β, or for b_i. Metropolis-Hastings steps can be used

Hyperpriors

- Specify priors for more meaningful parameters than the original elements of β
- For example, $\exp(\beta)$ is the relative risk/rate in a loglinear model and is the odds ratio in a logistic model
- It is convenient to specify lognormal priors for a generic parameter $\theta > 0$, since one may specify two quantiles of the distribution, and directly solve for the two parameters of the prior
- Let $\theta \sim LN(\mu, \sigma^2)$ such that $E(\log\theta) = \mu$ and $Var(\log\theta) = \sigma^2$, and let θ_1 and θ_2 e the q_1 and q_2 quantiles of this prior, then the parameters of the normal prior are

$$\mu = \frac{q_1\theta_2 - q_2\theta_1}{q_1 - q_2} \text{ and } \sigma^2 = \frac{\theta_1 - \theta_2}{q_1 - q_2}$$

Hyperpriors

- Consider $b_i \mid \sigma_0^2 \sim N(0, \sigma_0^2)$. Note that σ_0 is the standard deviation of the residuals on the linear predictor scale (not easy to interpret)
- Let $\tau_0 = 1/\sigma_0^2 \sim Ga(a, b)$, with a and b known $\Rightarrow$ marginal distribution for $b_i \sim t_d(0, \lambda^2)$ with $d = 2a$ and $\lambda^2 = b/a$
- These summaries allow prior specification based on beliefs concerning the residuals on a natural scale
- For a log-link, the above prior is equivalent to the residual relative risks following a log Student's t distribution
- We specify the range $\exp(\pm V)$ within which we expect the residual relative risks to lie with probability q and use the relationship $\pm t^d_{q/2}\lambda = \pm V$, where $t^d_{q/2}$ is the $q-$th quantile of a Student's t distribution with d degrees of freedom, to give $a = d/2$ and $b = \frac{V^2 d}{2(t^d_{q/2})^2}$
- For example, if we assume a priori that the residual relative risks follow a log Student's t distribution with $d = 2$ and that 95% of these risks fall in the interval $[0.5; 2]$ then the prior is $Ga(1; 0.026)$

Example: Seizure Data

We fit 3 models to these data:

M1

$$\begin{aligned} Y_{ij} \mid b_i &\sim Poisson[t_{ij}\exp(x_{ij}\beta + b_i)] \\ b_i \mid \sigma_0^2 &\sim^{iid} N(0, \sigma_0^2) \end{aligned}$$

Vague prior for β, and $\sigma_0^{-2} = \tau_0 \sim Ga(1, 0.026)$, which is equivalent to $b_i \sim t_2$ with 95% prior interval of $[0.5, 2]$

M2 $\tau_0 \sim Ga(2, 1.376)$ which is equivalent to $b_i \sim t_2$ with 95% prior interval of $[0.1, 10]$

M3

$$\begin{aligned} Y_{ij} \mid b_i &\sim Poisson[t_{ij}\exp(x_{ij}\beta + b_i + \varepsilon_{ij})] \\ b_i \mid \sigma_0^2 &\sim^{iid} N(0, \sigma_0^2) \\ \varepsilon_{ij} \mid \sigma_\varepsilon^2 &\sim N(0, \sigma_\varepsilon^2) \quad \text{with } \varepsilon_{ij} \perp b_i \end{aligned}$$

σ_0^2 captures between individual variability
σ_ε^2 captures within individual variability

Example: Seizure Data

- Note that under model M3 there is no simple marginal interpretation of σ_0^2 and σ_ε^2 since

$$\begin{aligned} E[Y_{ij}] &= \mu_{ij} = t_{ij}\exp(x_{ij}\beta + \sigma_0^2/2 + \sigma_\varepsilon^2/2) \\ Var[Y_{ij}] &= \mu_{ij}\left\{1 + \mu_{ij}[\exp(\sigma_0^2) - 1][\exp(\sigma_\varepsilon^2) - 1]\right\} \\ Cov[Y_{ij}, Y_{ik}] &= t_{ij}t_{ik}\exp[(x_{ij} + x_{ik})\beta]\exp(\sigma_0^2)[\exp(\sigma_0^2) - 1] \end{aligned}$$

- From the marginal model $Cov(\cdot;\cdot)$, σ_0^2 is controlling the within individual dependence in the model, with large values giving high dependence
- The marginal variance is quadratic in the mean and is controlled by both σ_0^2 and σ_ε^2, with large values corresponding to greater excess-Poisson variability
- We assign independent priors $\sigma_0^{-2} \sim Ga(1, 0.0260)$ and $\sigma_\varepsilon^{-2} \sim Ga(1, 0.0260)$

Example: Seizure Data

Parameter	Model 1	Model 2	Model 3
β_0	1.03 (0.016)	1.04 (0.16)	1.04(0.18)
β_1	-0.036 (0.21)	-0.030 (0.22)	0.062 (0.25)
β_2	0.11 (0.047)	0.11 (0.047)	0.0064 (0.10)
β_3	-0.10 (0.065)	-0.10 (0.065)	-0.29 (0.14)
σ_0	0.80 (0.078)	0.81 (0.077)	0.82 (0.084)
σ_ε	-	-	0.39 (0.033)

Table: Posterior mean and standard deviations (in brackets) for Bayesian analyses of the seizure data

Example: Seizure Data

- Note that model M3 shows substantive differences. β_3 is greatly reduced, with 95% credible interval for the rate being $[0.56, 0.99]$. This is because in the progabide group there is a single individual who is very influential. The introduction of measurement error accommodates this individual
- Also β_2 is now close to zero, whereas in models M1 and M2 it is 0.11. This shows that the aberrant individual's measurements were responsible for the high value of β_2 in the first two models
- The estimate of σ_ε is less than half the estimate for σ_0 so that between individual variability is greater than within-individual variability for these data

Conjugate Random Effects Model

Assume a random effect distribution that is conjugate to the likelihood

- Assume

$$\begin{aligned} Y_{ij} \mid \varepsilon_{ij} &\sim Poisson(t_{ij}\exp(x_{ij}\beta)\varepsilon_{ij}) \\ \varepsilon_{ij} \mid b &\sim^{iid} Ga(b,b) \end{aligned}$$

  Then

$$\begin{array}{rcl} Y_{ij} \mid b & \sim & NegBin(t_{ij}\exp(x_{ij}\beta), b) \\ E(Y_{ij}) = \mu_{ij} = t_{ij}\exp(x_{ij}\beta) & \text{and} & Var(Y_{ij}) = \mu_{ij}(1 + \mu_{ij}/b) \end{array}$$

- This model allows for excess Poisson variability but not for dependence of observations on the same patient
- The introduction of patient specific random effects allows for dependence on the same patient but looses the analytical tractability. Specifically,

$$\begin{aligned} Y_{ij} \mid \delta_i &\sim Poisson[\mu_{ij}\delta_i] \\ \delta_i \mid b &\sim^{iid} Ga(b,b) \qquad \text{note that } E(\delta_i) = 1 \end{aligned} \tag{3}$$

Conjugate Random Effects Model

- Model in equation (3) leads to a marginal model for the data of the *i*-th inidividual of

$$p(y_{i0}, y_{i1}, \cdots, y_{i4} \mid \mu_{ij}, b) = \prod_{j=0}^{4} \frac{\mu_{ij}^{y_{ij}}}{y_{ij}!} \frac{b^b}{\Gamma(b)} \frac{\Gamma(b + y_{i+})}{\Gamma(b + \mu_{i+})^{b + y_{i+}}}$$

which is not of negative binomial form

Hierarchical Models

- Hierarchical modeling provides a framework for building complex and high-dimensional models from simple and low-dimensional building blocks
- Of course, it is possible to analyze these models using non-Bayesian methods
- However, this modeling framework is popular in the Bayesian literature because MCMC is conducive to hierarchical models
- Both “divide and conquer” big problems by splitting them into a series of smaller problems in the same way

Hierarchical Models

Often Bayesian models can we written in the following layers of the hierarchy

1. Data layer: $[\mathbf{Y} \mid \theta, \alpha]$ is the likelihood for the observed data **Y** given the model parameters
2. Process layer: $[\theta \mid \alpha]$ is the model for the parameters θ that define the latent data generating process
3. Prior layer $[\alpha]$ prior for hyperparameters

Epidemiology example - Data layer

- Let S_t and I_t be the number of susceptible and infected individuals in a population, respectively, at time t
- The data Y_t is the number of observed cases at time t
- The data layer models our ability to measure the process I_t
- Data layer: $Y_t \mid I_t \sim Binomial(I_t, p)$
- This assumes no false positives and false negative probability p

Epidemiology example - Process layer

- Scientific understanding of the disease is used to model disease propagation

- We might select the simple Reed-Frost model

$$\begin{aligned} \text{Process layer: } I_{t+1} &\sim Binomial\left[S_t, 1-(1-q)^{I_t}\right] \\ S_{t+1} &= S_t - I_{t+1} \end{aligned}$$

- This assumes all infected individuals are removed from the population before the next time step

- Also that q is the probability of a non-infected person coming into contact with and contracting the disease from an infected individual

Epidemiology example - Prior layer

- The epidemiological process-layer model expresses the disease dynamics up to a few unknown parameters
- The Bayesian model is completed using priors, say,
- Prior layer:

$$\begin{aligned} I_1 &\sim Poisson(\lambda_1) \\ S_1 &\sim Poisson(\lambda_2) \\ p, q &\sim beta(a, b) \end{aligned}$$

McGill

When to stop adding layers?

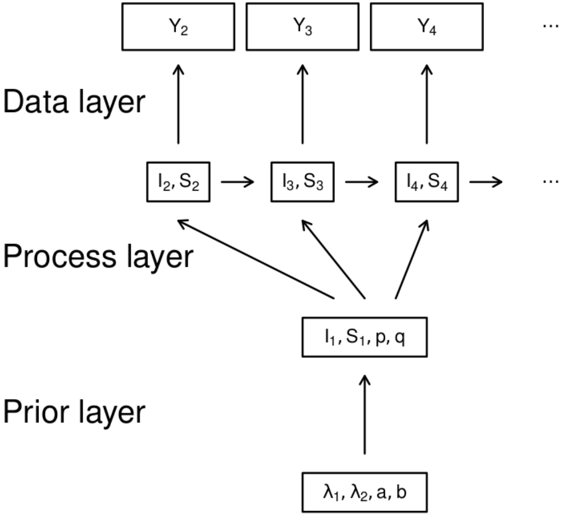
- In the previous example *a*, *b*, λ_1 and λ_2 are fixed
- But we will have uncertainty about the correct value
- Maybe replace a fixed value with another layer, say $a \sim \mathit{Uniform}(0, \theta)$?
- Then maybe $\theta \sim \mathit{Exponential}(\xi), \xi \sim \mathit{Uniform}(0, \eta)$, etc
- Rule of thumb: Be careful assigning priors to parameters in layers without replication
- For example, even if we knew p exactly this would be just one value and we couldn't hope to estimate the parameters of its beta distribution

Directed acyclic graphs (DAGs)

- A DAG is a graphical representation of a hierarchical model
- DAGS sometimes go by the name Bayesian networks
- Each observation and parameter is a node
- An arrow for X to Y means that the conditional distribution of Y depends on X
- “Directed” means that arrows only go one way
- Acyclic means there are no cycles, e.g.,

$$X \to Y \to Z \to X$$

Epidemiology example - DAG



Directed acyclic graphs (DAGs)

- Building models this way ensures we will always have a valid joint distribution

- For example, say we need to specify the joint distribution of (X, Y, Z)

- Any joint distribution can be written as

$$f(X, Y, Z) = f(X)f(Y|X)f(Z|X, Y)$$

- This is a fully-connected DAG

- Ad-hoc constructions like

$$f(X, Y, Z) = f(X|Z)f(Y|X)f(Z|X, Y)$$

  may or may not give a valid joint PDF

Hierarchical models and MCMC

- Consider the classic one-way random effects model:

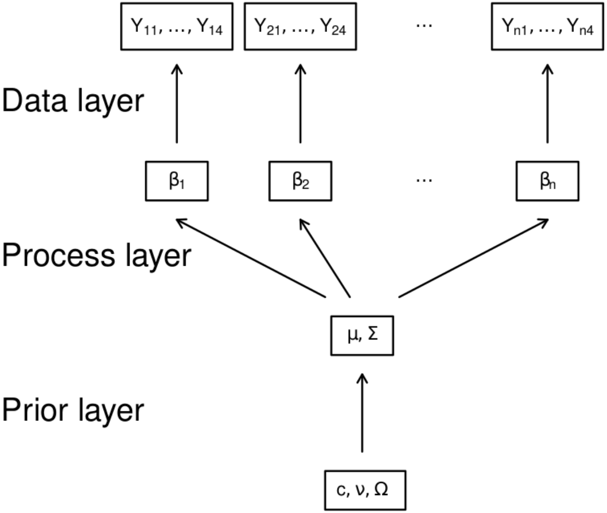
  $$Y_{ij} \sim N(\theta_i, \sigma^2) \text{ and } \theta_i \sim N(\mu, \tau^2)$$

  where Y_{ij} is the $j-th$ replicate for unit i and $\alpha = (\mu, \sigma^2, \tau^2)$ has

  an uninformative prior

- This hierarchy can be written using a directed acyclic graph

Random effects example - DAG



Hierarchical models and MCMC

- MCMC is efficient in this case even if the number of parameter or levels of the hierarchy is large

- You only need to consider "connected nodes" when you update each parameter

    1. $[\theta_i|\cdot]$

    2. $[\mu|\cdot]$

    3. $[\sigma^2|\cdot]$

    4. $[\tau^2|\cdot]$

- Each of these updates is a draw from a standard one-dimensional normal or inverse gamma

Worked examples from Reich and Ghosh

1. Analysis of tyrannosaurid growth curves

2. Species distribution mapping via data fusion