Chapter 11

Hierarchical Models -I

11.1 Introduction

Consider the following problem based on data containing death in hospitals performing cardiac surgery in babies.

No of Ops (n_i)	47	148	119	810	211	196	148	215	207	97	256	360
No of Deaths (y_i)	0	18	8	46	8	13	9	31	14	8	29	24

The purpose of the study is to estimate the probability of deaths, to rank the hospitals according to these probabilities, and to make some predictions for new hospitals.

To estimate the p_i in each hospital i, $\quad i = 1, ..., 12$, we could fit a conjugate Beta-Binomial fixed effects model, which assumes that the true failure probabilities p_i are independent for each hospital. This is equivalent to assuming a standard non-informative prior distribution for the $p_i's$, namely $p_i \sim Beta(1, 1)$. The likelihood would be $\prod_{i=1}^{12} p_i^{y_i}(1 - p_i)^{(n_i - y_i)}$ and the posterior would be $\prod_{i=1}^{12} Beta_i(y_i + 1, n + 1)$. The marginals for each p_i would be independent betas.

Alternatively, we could consider that the failure rate, p_i, across hospitals is different in each hospital, but somehow related to each other. After all, all doctors go to similar medical schools and receive similar training. This calls for a random effects model, for example:

$$\begin{aligned} y_i &\sim Binomial(p_i, n_i) \qquad i = 1, ..., 10 \\ logit(p_i) &= b_i \\ b_i &\sim N(\mu.\tau) \end{aligned}$$

with μ and τ having non-informative priors. The distribution of the $b_i's$ is called the population distribution of the $b_i's$.

The latter is an example of a simple hierarchical model: a sampling distribution for the data, a population distribution for the parameters and then hyperpriors. Under this model, we would keep track of the posterior distribution of the 12 $p_i's$ and the μ and τ.

Notice what we did: we assumed a different random effect for each hospital, but all random effects came from the same distribution.

Other possible fixed effects models that we could consider are:

$$\begin{aligned} y_i &\sim Binomial(p_i, n_i) \qquad i = 1, ..., 10 \\ logit(p_i) &= b \\ b &\sim N(0, 0.0001) \end{aligned}$$

with common probability of death after surgery for all patients, or

$$\begin{aligned} y_i &\sim Binomial(p_i, n_i) \qquad i = 1, ..., 10 \\ logit(p_i) &= b_i \\ b_i &\sim N(0, 0.0001) \end{aligned}$$

independent probability of cardiac surgery mortality in each hospital.

The above examples illustrate the kind of reasoning that might go into modeling to answer our research questions. The expert in the area studied can aid in determining whether this or that random effect makes sense. Bayesian theory, and in particular, DeFinetti's theorem, tell us that if the observations are exchangeable, there is a distribution for the random effects that we could exploit.

As the complexity of the models and the structure in the data increases, the reasoning may not be more complicated but the number of parameters increases and there are many more things to take into account. The rest of this lecture will consider an additional example. Examples that will require programming will be done in the next lectures.

11.2 Reasoning for modeling

Hepatitis B is endemic in Africa. A national program of childhood vaccination against HB was introduced in Gambia. The program effectiveness depends on the duration of the immunity afforded by the vaccination.

The data: 106 children immunized against HB. For each child: anti-HB titre measured at time of vaccination (baseline) and on 2 or 3 follow-up occassions.

The objective of the study is to obtain a model useful for predicting an individual child's protection against HB after vaccination.

A similar study in Senegal found: anti-HB titre $\propto \frac{1}{T}$, where T = time since HB vaccination.

11.2.1 Specifying a Bayesian Generalized linear model for the HB data

Let y_{ij} be the *log* of the *jth* anti-HB titre measurement for child i.

$$y_{ij} \sim Normal(\mu_{ij}, \sigma^2)$$

Assume that response (log anti-HB titre) may depend on time and on baseline titre.

$$\mu_{ij} = \alpha + \beta(t_{ij} - \bar{t}) + \gamma(y0_i - \bar{y0})$$

where t_{ij} = log of the time (in days since vaccination) of the *jth* titre measurement for child i. $\bar{t}$ = mean of the $t'_{ij}s$. $y0_i$ = log of the baseline anti-HB titre for child i $\bar{y0}$ = mean of the $y0'_i s$

A standard "non-informative" prior distribution for the unknown parameters in the HB model is Uniform on $(\alpha, \beta, \gamma, \log \sigma)$. This is an improper prior density, i.e. it does not integrate to 1. A vague but proper prior for the HB model is

$$\begin{aligned} \alpha &\sim Normal(0, 10000) \\ \beta &\sim Normal(0, 10000) \\ \gamma &\sim Normal(0, 10000) \\ \frac{1}{\sigma^2}(= \tau) &\sim Gamma(0.001, 0.001) \end{aligned}$$

In Bayesian analysis, as you know, we need to find the joint posterior distribution of "all" the unknown quantities in the model, conditional on the quantities (data) we have observed (even though quite often we just require inference on a single parameter, say, θ_k). To obtain the latter, we just average (i.e., integrate) the joint posterior over all the other unknowns.

The joint posterior distribution in the example we are considering here is:

$$\begin{aligned} p(\alpha,\beta,\gamma,\sigma^2 \mid y,t,y0) &\propto \prod_{ij} p(y_{ij} \mid t_{ij}, y0_i, \alpha,\beta,\gamma,\sigma^2) \\ &\quad \times p(\alpha)p(\beta)p(\gamma)p(\sigma^2) \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \prod_{ij} \exp^{-\frac{1}{2\sigma^2}(y_{ij}-\mu_{ij})^2} \\ &\quad \times \frac{1}{\sqrt{2\pi \times 10000}} \exp^{-\frac{\alpha^2}{2\times 10000}} \\ &\quad \times \frac{1}{\sqrt{2\pi \times 10000}} \exp^{-\frac{\beta^2}{2\times 10000}} \\ &\quad \times \frac{1}{\sqrt{2\pi \times 10000}} \exp^{-\frac{\gamma^2}{2\times 10000}} \\ &\quad \times \left(\frac{1}{\sigma^2}\right)^{0.001-1} \exp^{-0.001\times\frac{1}{\sigma^2}} \end{aligned}$$

The marginal posterior distribution for β is:

$$\begin{aligned} p(\beta \mid y,t,y0) &\propto \int \prod_{ij} p(y_{ij} \mid t_{ij}, y0_i, \alpha,\beta,\gamma,\sigma^2) \\ &\quad \times p(\alpha)p(\beta)p(\gamma)p(\sigma^2) d\alpha d\gamma d\sigma^2 \end{aligned}$$

Remember that to obtain the marginal for β we do the following:

(a) Draw samples from the joint posterior using dependent sampling from a Markov chain that has our joint posterior distribution as its stationary (equilibrium) distribution.

(b) This gives us a matrix that has in each row a sample of $\alpha,\beta,\gamma,\sigma^2$. The column for the β, when summarized, gives us all the information we need about the marginal distribution for β.

To draw the samples from the joint posterior, we will need to find the full conditional distributions and use Gibbs sampling to obtain markov chains. Let our vector of unknowns be $\alpha,\beta,\gamma,\sigma^2$. Then

(a) Choose starting values $\alpha_1^{(0)}, \beta_1^{(0)}, \gamma_1^{(0)}, \sigma 2_1^{(0)}$ where we have written σ^2 as *sigma*2 to avoid confusion with the many superscripts below...

(b) Sample $\alpha_1^{(1)}$ from $p(\alpha \mid \beta_1^{(0)}, \gamma_1^{(0)}, \sigma_1^{(0)}$

Sample $\beta_1^{(1)}$ from $p(\beta \mid \alpha_1^{(1)}, \gamma_1^{(0)}, \sigma_1^{(0)}$

Sample $\gamma_1^{(1)}$ from $p(\gamma \mid \alpha_1^{(1)}, \beta_1^{(1)}, \sigma_1^{(0)}$

Sample $\sigma_1^{(1)}$ from $p(\sigma \mid \gamma_1^{(1)}, \alpha_1^{(1)}, \beta_1^{(1)}$.

(c) Repeat step (b) many 1000s times... Eventually, you will obtain the sample from the joint posterior.

Recall that burn in samples should be ignored when summarizing the samples for posterior inference.

In general, the full conditional density functions are one dimensional, of complex algebraic form and log-concave. It may be that we will need to use some Adaptive rejection sampling method at each iteration of the Gibbs sampler.

11.2.2 Inclusion of prior information

So far, we have just considered vague prior distributions for the regression coefficients (Normal with large variance). Recall that the Senegal study found: anti-HB titre $\propto \frac{1}{T}$, where T = time since HB vaccination. This gives a linear relationship on the log scale: log anti-HB titre = $\kappa_i - \log T$. This is equivalent to setting $\gamma = 1, \beta = -1$ in the linear predictor of our original model for the HB data, i.e,

$$\mu_{ij} = \alpha + \gamma y0_i + \beta t_{ij}$$

Based on the Senegal evidence, informative priors for β and γ may be:

$$\begin{aligned} \beta &\sim Normal(-1, 0.05^2) \\ \gamma &\sim Normal(1, 0.05^2) \end{aligned}$$

For β, this gives prior mean = -1.0 and prior 95% interval of $(-1.0 \pm 1.96 \times 0.05) \approx (-1.1, -0.9)$. For γ, this gives prior mean = 1.0 and prior 95% interval of $(1.0 \pm 1.96 \times 0.05) \approx (0.9, 1.1)$.

11.2.3 Prediction

Once we have obtained the joint posterior, we recall the original objective of the Gambian HB study:"To obtain a model useful for predicting an individual child's protection against HB after vaccination."

Suppose we want to predict the anti-HB titre 3 years (1095 days) after vaccination for a child with baseline titre of 665 mIU.

The predictive distribution will have the same form as the observed $y'_{ij}s$. At each iteration of the gibbs sampler, we can also sample from the likelihood conditional on the current values of the other model parameters. That is, for example, at iteration 100,

$$\begin{aligned} \mu_{i3}^{(100)} &= \alpha_{100} + \beta_{100}(1095 - \bar{t}) + \gamma_{100}(665 - \bar{y0}) \\ y_{i3}^{(100)} &\sim Normal(\mu_{i3}, \sigma^2_{100}) \end{aligned}$$

This will give us 1000 draws from the predictive distribution under these conditions, and we can summarize this distribution when done.

Suppose we also want to know the posterior probability that a child with baseline anti-HB titre of 10000 mIU will have an anti-HB titre above 1000 mIU 2 years (730 days) after vaccination. The solution to this would be the predicted 2-year log anti-HB titre for the child:

$$\begin{aligned} \mu_{new}^{(100)} &= \alpha_{100} + \beta_{100}(log(730) - \bar{t}) + \gamma_{100}(log(10000 - \bar{y0}) \\ y_{new}^{(100)} &\sim Normal(\mu_{new}^{(100)}, \sigma^2_{100}) \end{aligned}$$

With the 1000 draws of from this predictive distribution, we can obtain many quantities of interest (histogram, quantiles, etc.)

11.3 Is it resonable to assume a common regression line for all children?

After completing the above process, we may ask ourselves: Have we adequately modelled the random variation in responses? Are children's anti-HB titre trajectories more heterogeneous than our model assumes?

One can think of some reasons for excess variation in response:

(a) Individual heterogeneity, i.e., systematic differences between units not attributable to random variation.

(b) Repeated response measurements from the same unit tend to be correlated. 2 responses from the same unit will be more alike than 2 responses from different units. Variation in responses is not completely random.

(c) Failure to measure or include a relevant explanatory variable

(d) Inaccurate measurement of relevant explanatory variables

Perhaps we could modify our model to allow a separate intercept and slope for each child as follows:

$$\begin{aligned} y_{ij} &\sim Normal(\mu_{ij}, \sigma^2) \\ \mu_{ij} &= \alpha_i + \beta_i(t_{ij} - \bar{t}) + \gamma(y0_i - \bar{y0}) \end{aligned}$$

Assume independent (vague) priors for each α_i and β_i, e.g.,

$$\begin{aligned} \alpha_1 &\sim Normal(0, 10000) \\ \alpha_2 &\sim Normal(0, 10000) \\ ..... &.... .... \\ \alpha_{106} &\sim Normal(0, 10000) \\ .... &.... ... \\ \beta_{106} &\sim Normal(0, 10000) \end{aligned}$$

There are some problems with this approach:

(a) Overfitting: many children have only 2 anti-HB titre measurements, so model will fit data exactly

(b) Unrealistic to assume that each child's trajectory is totally unlike (independent) of any other child's

(c) How do we use such a model for prediction?

11.3.1 Random effects model

Assume that all of the $\alpha_i's$ follow a common population prior distribution, and likewise for the $\beta_i's$, e.g.

$$\begin{aligned} \alpha_i &\sim Normal(\mu_\alpha, \sigma_\alpha^2) \qquad i = 1, ..., 106 \\ \beta_i &\sim Normal(\mu_\beta, \sigma_\beta^2) \qquad i = 1, ..., 106 \end{aligned}$$

We may then assume vague priors for the hyper-parameters of the population distribution:

$$\begin{aligned} \mu_\alpha &\sim Normal(0, 10000) \\ \tau_\alpha = \frac{1}{\sigma_\alpha^2} &\sim Gamma(0.001, 0.001) \\ \mu_\beta &\sim Normal(0, 10000) \\ \tau_\beta = \frac{1}{\sigma_\beta^2} &\sim Gamma(0.001, 0.001) \end{aligned}$$

The $\alpha_i's$ and $\beta_i's$ are called random effects (they are viewed as a random sample from some larger population.)

Advantages of the random effects model

(a) Intercepts and slopes are allowed to vary between children, but are assumed to be similar (not completely independent).

(b) The estimates of each α_i and β_i borrow strength from all the other $\alpha_i's$ and $\beta_i's$ via the common population distribution.

(c) possible and sensible to fit random effects models with more parameters than data points!

(d) The responses y_{ij} within a child i are assumed to be conditionally independent given that child's random effects and any other covariate effects.

(e) Prediction is straightforward since all children are assumed to follow the same population distribution.

(f) Such models are also termed Generalized Linear Mixed Models or Hierarchical models or Multilevel models.

(g) Random effects modelling allows estimation of subject-specific parameters which borrow strength from data on all subjects, yet do not constraint every individual to follow identical trajectory.