§5 Hierarchical Models

Outline

1. Non-hierarchical models
2. Hierarchical models (hierarchical priors and exchangeability)
3. Using DAGs for hierarchical models
4. Summary

1. Non-hierarchical models

Example: Drug efficacy

Data:

$$y = 15 \text{ successes from}$$
$$n = 20 \text{ independent trials}$$

Likelihood:

$$Y \mid \theta \sim \text{Binomial}(n, \theta) \ ,$$

where θ is *true* success rate (ie probability of success)

Prior:

$$\theta \sim \text{Beta}(9.2, 13.8)$$

Posterior:

$$p(\theta \mid y) \propto p(y \mid \theta)\, p(\theta)$$
$$\theta \mid y \sim \text{Beta}(24.2, 18.8)$$

Example: Hospital death rates

Now suppose we observe N sets of binomial data, for example: N=12 hospitals performing cardiac surgery in babies

Number of failures (deaths) per hospital:

Hospital i	1	2	3	....	10	11	12
No. of ops. n_i	15	148	10	....	97	256	360
No. of deaths y_i	0	18	1	....	8	29	24

How would you model these data?

Assume that, given 'true' death rate θ_i (ie probability of death) in hospital i, operation outcomes within hospital i are independent.

$$Y_i \mid \theta_i \sim \text{Binomial}(n_i, \theta_i) \ \ (i = 1, \ldots, 12)$$

Using a common death rate θ

Assume true death rate in each hospital is the same (ie $\theta_i = \theta, \forall i$).

$$Y_i \mid \theta \sim \text{Binomial}(n_i, \theta) \ \ (i = 1, \ldots, 12)$$

Then, likelihood is

$$p(\mathbf{y} \mid \theta) = \prod_{i=1}^{12} p(y_i \mid \theta)$$
$$\propto \prod_{i=1}^{12} \theta^{y_i}(1-\theta)^{n_i - y_i} = \theta^{\sum y_i}(1-\theta)^{(\sum n_i - \sum y_i)}$$

This is equivalent to observing a single hospital with $\sum_i y_i$ deaths in $\sum_i n_i$ operations.

Assume Beta prior for θ with α, β fixed:

$$\theta \mid \alpha, \beta \sim \text{Beta}(\alpha, \beta)$$

Then the posterior for θ is

$$\theta \mid \mathbf{y} \sim \text{Beta}\left(\sum_{i=1}^{12} y_i + \alpha, \sum_{i=1}^{12}(n_i - y_i) + \beta\right)$$

But is it reasonable to assume a *common* probability θ of death for every hospital?

Using different death rates θ_i

In each hospital i (with 'true' death rate θ_i),

$$\begin{aligned} Y_i \mid \theta_i &\sim \text{Binomial}(n_i, \theta_i) \\ \theta_i \mid \alpha, \beta &\sim \text{Beta}(\alpha, \beta) \end{aligned}$$

- θ_i's are random sample from a common *population distribution*: Beta(α, β)
- So, hospital 'true' death rates are assumed to be **similar** but not identical.
  Is this reasonable?
  Suppose the only information you have is that 3 hospitals have 'true' death rates 5%, 4% and 9% respectively. Guess the death rate of a 4th hospital

How would you specify values for α and β?
How would you justify the values of α and β?

Empirical Bayes approach

Hospital i	1	2	3	....	10	11	12
No. of ops. n_i	15	148	10	....	97	256	360
No. of deaths y_i	0	18	1	....	8	29	24

1. Calculate observed death rates $\frac{y_i}{n_i}$
2. Calculate the mean and variance of these 12 values $\frac{y_i}{n_i}$
3. Find $\widehat{\alpha}$, $\widehat{\beta}$ such that Beta$(\widehat{\alpha}, \widehat{\beta})$ distribution has this mean and variance.
4. Use $\theta_i \sim$ Beta$(\widehat{\alpha}, \widehat{\beta})$ as a prior to obtain posterior $\theta_i \mid y_i$

Disadvantages of this approach are:
- We are using the data twice: once to estimate the prior; again in the likelihood. $\Rightarrow$ over-estimated precision of our inference
- Using a point estimate for α and β (and treating them as fixed) ignores some uncertainty about the population distribution of the θ_i's

2. Hierarchical models

Fundamental idea of Bayesian inference is to assume a probability distribution for uncertainty about any unknown quantities.

So, treat α and β as unknown and independent, and assign prior distributions to them independently, e.g.

$$\begin{aligned} \alpha &\sim \text{Exponential}(0.01) \\ \beta &\sim \text{Exponential}(0.01) \end{aligned}$$

Now, the unknown parameters are $(\alpha, \beta, \boldsymbol{\theta})$, where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_{12})$. Since $\theta_i \sim$ Beta(α, β) independently for each i given α and β, the joint prior distribution for the entire set of parameters is

$$p(\boldsymbol{\theta}, \alpha, \beta) = \left\{ \prod_{i=1}^{N} p(\theta_i \mid \alpha, \beta) \right\} p(\alpha)\, p(\beta)$$

Bayes Theorem gives us the joint posterior distribution of $(\alpha, \beta, \boldsymbol{\theta})$:

$$\begin{aligned} p(\boldsymbol{\theta}, \alpha, \beta \mid \mathbf{y}) &\propto \left\{ \prod_{i=1}^{N} p(\theta_i \mid \alpha, \beta) \right\} p(\alpha)\, p(\beta) \\ &\quad \times \left\{ \prod_{i=1}^{N} p(y_i \mid \theta_i) \right\} \end{aligned}$$

Advantages of this approach:
- The posterior distribution for each θ_i
  - *'borrows strength'* from the likelihood contributions of *all* hospitals, via their influence on the estimate of the unknown population parameters α, β
  - reflects our full uncertainty about the true values of α and β
- This latter is also useful if we are interested in α and β themselves (e.g. $\alpha/(\alpha + \beta)$ is mean death rate over population of hospitals)

Such models are also called *Random effect* or *Multilevel* models.

Example: Hospital death rates

In the 12 hospitals, there were a total of 2073 operations including 159 deaths; ie, the overall death rate is $159/2073 = 0.077$.

We fitted the following models:

1. MLE (non-Bayesian): y_i/n_i
2. Non-hierarchical Bayesian

$$\begin{aligned} Y_i \mid \theta_i &\sim \text{Binomial}(n_i, \theta_i) \\ \theta_i \mid \alpha, \beta &\sim \text{Beta}(\alpha = 1, \beta = 1) \end{aligned}$$

   The posterior distribution of θ_i for the non-hierarchical model is Beta(y_i+1, n_i-y_i+1). So, the posterior mean of θ_i is $E[\theta_i|\mathbf{y}] = \frac{y_i+1}{n_i+2}$.

3. Hierarchical Bayesian

$$\begin{aligned} Y_i \mid \theta_i &\sim \text{Binomial}(n_i, \theta_i) \\ \theta_i \mid \alpha, \beta &\sim \text{Beta}(\alpha, \beta) \\ \alpha &\sim \text{Exponential}(0.01) \\ \beta &\sim \text{Exponential}(0.01) \end{aligned}$$

   The hierarchical model was fitted by using WinBUGS.

Therefore, we obtained three estimates of θ_i:

1. the MLE $\frac{y_i}{n_i}$;
2. the posterior mean of θ_i for the non-hierarchical Bayesian model;
3. the posterior mean of θ_i for the hierarchical Bayesian model.

				Posterior mean for	
i	y_i	n_i	MLE	non-hier.	hier.
1	0	15	0.000	0.059	*0.075*
2	18	148	0.122	0.127	*0.102*
3	1	10	0.100	**0.167**	**0.085**
	⋮			⋮	
10	8	97	0.082	0.091	*0.081*
11	29	256	0.113	0.116	*0.102*
12	24	360	0.067	0.069	*0.072*

NB: Compared with the non-hierarchical model, the hierarchical Bayesian model

- moved estimates towards the overall death rates, 0.077
- made estimates more reliable for those hospitals with little data, ie small n_i

Hierarchical priors

We have specified a *hierarchical prior* for the surgical failure rates θ_i.

In general, suppose we have data $\mathbf{y}$ and parameters $\boldsymbol{\theta} = (\theta_1, ..., \theta_n)$

- Likelihood $p(\mathbf{y} \mid \boldsymbol{\theta})$ (1st level)
- Prior $p(\boldsymbol{\theta})$ depends on higher level parameter ϕ_2: $p(\boldsymbol{\theta} \mid \phi_2)$ (2nd level)
- $p(\phi_2)$ (3rd level)
  Marginal prior for $\boldsymbol{\theta}$ is then
  $$p(\boldsymbol{\theta}) = \int p(\boldsymbol{\theta} \mid \phi_2) p(\phi_2) d\phi_2$$
- We might add further levels
  $p(\phi_2 \mid \phi_3)$ (3rd level)
  ...
  $p(\phi_m)$ ($(m+1)$-th (top) level)
  Marginal prior for $\boldsymbol{\theta}$ is then
  $$\begin{aligned} p(\boldsymbol{\theta}) = & \int \cdots \int p(\boldsymbol{\theta} \mid \phi_2) \times p(\phi_2 \mid \phi_3) \times \ldots \\ & \times p(\phi_{m-1} \mid \phi_m) \times p(\phi_m)\, d\phi_2 \ldots d\phi_m \end{aligned}$$

- ϕ_k are called (kth level) *hyper-parameters*
- Theoretically there can be as many levels as necessary, but in practice it is usually hard to interpret parameters of level 3 or higher
- A non-informative prior is usually specified for the marginal distribution of the top-level parameters

For the hospital example:

$$\begin{aligned} Y_i \mid \theta_i &\sim \text{Binomial}(n_i, \theta_i) && \text{(Level 1)} \\ \theta_i \mid \alpha, \beta &\sim \text{Beta}(\alpha, \beta) && \text{(Level 2)} \\ \alpha &\sim \text{Exponential}(0.01) && \text{(Top level)} \\ \beta &\sim \text{Exponential}(0.01) && \text{(Top level)} \end{aligned}$$

Exchangeability

In our hierarchical model we assumed that

$$\theta_i \mid \alpha, \beta \sim \text{Beta}(\alpha, \beta) \quad (i = 1, \ldots, N)$$

So, conditional on (α, β), the θ_i's are independent of one another.

$$p(\boldsymbol{\theta} \mid \alpha, \beta) = \prod_{i=1}^{N} p(\theta_i \mid \alpha, \beta)$$

E.g. if $N = 4$ and I know the values of θ_1, θ_2, θ_3 and $\theta_i \sim$ Beta(3, 30), then this tells me nothing about θ_4.

The marginal distribution of $\boldsymbol{\theta}$ is

$$p(\boldsymbol{\theta}) = \int p(\alpha, \beta) \left\{ \prod_{i=1}^{N} p(\theta_i \mid \alpha, \beta) \right\} \, d\alpha \, d\beta$$

This cannot be factorised into separate functions of $\theta_1, \ldots, \theta_N$. So, unconditional on (α, β), the θ_i's are not (marginally) independent.

E.g., if $N = 4$ and I know the values of θ_1, θ_2 and θ_3, then this tells me something about θ_4.

That is, θ_i's are not marginally independent. However, they are *exchangeable*.

Definition of exchangeability

A sequence of random variables $\theta_1, ..., \theta_n$ is said to be *exchangeable* if, for any permutation $\{i_1, ..., i_n\}$ of $\{1, ..., n\}$, $(\theta_{i_1}, ..., \theta_{i_n})$ have the same n-dimensional joint probability distribution as $(\theta_1, ..., \theta_n)$. That is, $\forall a_1, \ldots, a_n$

$$p(\theta_1 = a_1, \ldots, \theta_n = a_n) = p(\theta_{i_1} = a_1, \ldots, \theta_{i_n} = a_n)$$

Notes

1. If $\theta_1, \ldots, \theta_n$ are marginally independent and have same marginal distribution, they are exchangeable.
2. If $\theta_1, \ldots, \theta_n$ are exchangeable, they have same marginal distribution, but are not necessarily marginally independent.

General representation theorem (De Finetti, 1937, 1970/1974; Hewitt and Savage, 1955; Diaconis and Freedman, 1984, 1987)

If $\theta_1, \theta_2, \ldots$ are exchangeable, then there exists a parametric model $p(\theta \mid \phi)$ with prior $p(\phi)$ for ϕ such that $\theta_i \perp\!\!\!\perp \theta_j \mid \phi$, ie,

$$p(\theta_1, \ldots, \theta_N, \phi) = \left[\prod_{i=1}^{N} p(\theta_i \mid \phi) \right] p(\phi)$$

That is, $\theta_1, \ldots, \theta_N$ is a random sample from some model $p(\theta \mid \phi)$ with prior $p(\phi)$.

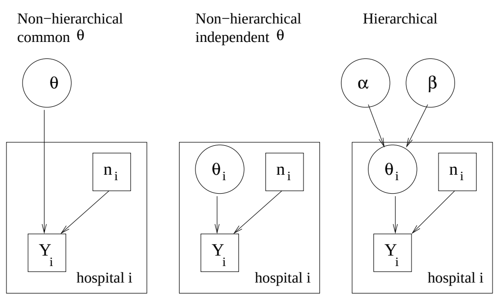
Thus, exchangeability implies a hierarchical model.

3. Using DAGs for hierarchical models

DAGs can be used to represent hierarchical models. Conventionally, it uses

- circle nodes to represent unknown rvs (e.g. parameters, missing data)
- square nodes to represent known rvs (e.g. data)
- rectangular boxes to represent repetitive structures (e.g. one box for each hospital)

Our hospital models can be represented:



Hierarchical models with covariates

Example: GLMM

- 1st level

$$Y_i \mid \theta_i \sim \text{Poisson}(C_i \theta_i)$$

- 2nd level

$$\begin{aligned} \log \theta_i &= \beta_0 + \beta_1 X_i + \lambda_i \\ \lambda_i \mid \tau &\sim \text{Normal}(0, \tau^{-1}) \\ \beta_0 &\sim \text{non-informative} \\ \beta_1 &\sim \text{non-informative} \end{aligned}$$

- 3rd level – hyper-priors

$$\tau \sim \text{non-informative}$$

Often known as a generalised linear mixed model (GLMM)

Example: Hepatitis B

Background

- Hepatitis B (HB) is endemic in Africa
- National programme of childhood vaccination against HB introduced in Gambia
- Program effectiveness depends on duration of immunity afforded by vaccination

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 occasions

Study objective

- To obtain a model for predicting an individual child's protection against HB after vaccination

A. Non-hierarchical LM

1. Probability distribution (likelihood) for responses:

$$Y_{ij} \mid \mu_{ij}, \tau \sim \text{Normal}(\mu_{ij}, \tau^{-1})$$

   where
   Y_{ij} = *log* of jth titre measurement for child i

2. Linear predictor:

$$\mu_{ij} = \alpha + \beta(t_{ij} - \bar{t}) + \gamma(Y_{i0} - \bar{Y}_0)$$

   where
   t_{ij} = *log* time (in days since vaccination) of the jth titre measurement for child i
   Y_{i0} = *log* baseline titre for child i

3. A vague but *proper* prior for the HB model:

$$\begin{aligned} \alpha &\sim \text{Normal}(0, 10000) \\ \beta &\sim \text{Normal}(0, 10000) \\ \gamma &\sim \text{Normal}(0, 10000) \\ \tau &\sim \text{Gamma}(0.001, 0.001) \end{aligned}$$

B. Hierarchical LM (LMM)

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

- Modify our LM to allow separate intercept and slope for each child:

$$Y_{ij} \mid \mu_{ij}, \tau \sim \text{Normal}(\mu_{ij}, \tau^{-1})$$

$$\mu_{ij} = \alpha_i + \beta_i(t_{ij} - \bar{t}) + \gamma(Y_{i0} - \bar{Y}_0)$$

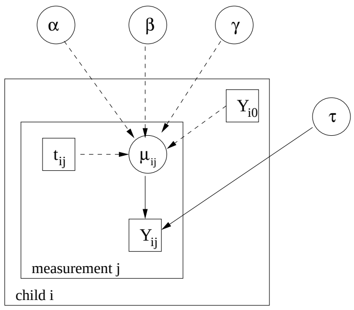
- What prior distributions should we choose for the α_i's and β_i's?
  Assume that the α_i's are exchangeable, and likewise for the β_i's. E.g.

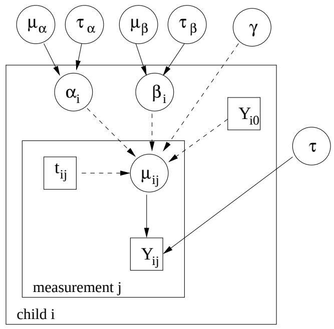
$$\begin{aligned} \alpha_i \mid \mu_\alpha, \tau_\alpha &\sim \text{Normal}(\mu_\alpha, \tau_\alpha^{-1}) \quad i = 1, ..., 106 \\ \beta_i \mid \mu_\beta, \tau_\beta &\sim \text{Normal}(\mu_\beta, \tau_\beta^{-1}) \quad i = 1, ..., 106 \end{aligned}$$

- We can assume vague priors for the *hyperparameters*, e.g.:

$$\begin{aligned} \mu_\beta, \mu_\alpha &\sim \text{Normal}(0, 10000) \\ \tau_\alpha, \tau_\beta &\sim \text{Gamma}(0.001, 0.001) \end{aligned}$$

DAGs for the LM and LMM





(Dashed arrows denote deterministic dependencies)

4. Summary

- *Hierarchical modelling involves breaking down the problem into layers and specifying a model for each layer*: a model for data given parameters; a model for parameters given hyper-parameters; maybe a model for hyper-parameters given higher-level hyper-parameters
- *It is useful when data obtained from similar-but-not-the-same units — parameters for different units are exchangeable*. Such models enables data on one unit to inform parameters of other units (borrowing strength). They move extreme estimates of units with little information towards population mean — this stabilises parameter estimates
- It is often difficult to specify informative priors for hyper-parameters, so we usually use non-informative (vague) priors for hyper-parameters

Obtaining marginal posterior distributions for parameters of a hierarchical model analytically is often not possible. We need MCMC.

Outline revisited

1. Non-hierarchical models
2. Hierarchical models (hierarchical priors and exchangeability)
3. Using DAGs for hierarchical models
4. Summary

Next week: MCMC