

# Hierarchical Models

John P. Burkett  
burkett@uri.edu

November 25, 2009

# Outline

Multilevel data structures

Multilevel models

- Basics

- Bayesian analysis

Example: heart transplant mortality

## Multilevel data structures I

Individuals (or other disaggregated observational units) are sometimes grouped in ways that suggest possible similarities in behavioral patterns. For example, the individuals in an educational testing data set might be students grouped by sex, ethnicity, teacher, parents, texts, or extra-curricular activities. The data set might also include higher level groupings. Teachers could be grouped by school, schools by school district, and districts by state. We might have group level covariates which would help to explain variation across groups in individual outcomes such as students' test scores. For instance, we might have data on teachers' education, school principals' experience, parents' income, school districts' expenditure per student, and states' population densities. A data structure may be called **multilevel** if it includes covariates measured at more than one level of aggregation—e.g. characteristics of individual students, classes, parents, states, etc.

## Multilevel data structures II

A multilevel data structure may be called **hierarchical** if its groups can be arranged in levels such that individuals who belong to the same group at one level belong to the same group at every higher level. For example, students who are in the same class are normally in the same school and district. In such cases, groups are said to be **nested**.

In contrast, a non-hierarchical or non-nested multilevel data structure includes groups of various types such that two individuals may belong to a single group of one type but different groups of another type. For example, a data set that records students' parents and classes would typically include some students who have the same parents but attend different classes and some who have the same class but different parents. Family groups are not nested with class groups, nor are class groups nested within family groups. As another example of a non-nested multilevel data structure, consider a data set that includes information about workers

## Multilevel data structures III

cross-classified by occupation and state of residence and further information about the occupations and states. Typically, occupations are not nested in states and states are not nested in occupations (Gelman and Hill 2007, p. 244).

## Multilevel models I

A multilevel model pertains to a multilevel data structure and contains corresponding levels of parameters. Relationships between individual level variables are expressed in terms of parameters which are in turn represented in terms of higher level variables and higher level parameters, termed **hyperparameters**. For example, at the individual level of an educational testing model, a student's achievement test score might be represented as a linear fn of a previous aptitude score. The intercept and slope coefficients might be represented as fns of the teacher's education, the parents' income, and the year's flu prevalence (Gelman and Hill 2007, pp. 1–2). A multilevel model that pertains to a hierarchical data structure may also be called hierarchical.

## Bayesian analysis of multilevel models I

The likelihood fn of a multilevel model, like that of a single level model, is proportional to the sampling distribution of the individual level:  $L(\alpha; y) \propto f(y|\alpha)$ , where  $y$  denotes an observable individual level variable and  $\alpha$  is an unknown parameter of its distribution, either or both of which could be a vector. The prior distribution of  $\alpha$  in a multilevel model is conditional on a hyperparameter (possibly a vector)  $\beta$ . Bayesians assign  $\beta$  a **hyperprior** distribution, which describes their beliefs about it based on information other than  $y$ . The hyperprior distribution for  $\beta$  may depend on higher level hyperparameters, either known or unknown. At the top level of the hyperprior distributions the hyperparameters are treated as known and often chosen to be noninformative.

The posterior density for a multilevel model is a product involving a likelihood fn, one or more prior densities conditional on hyperparameters, and a marginal density for the the highest level unknown hyperparameter. For example, a multilevel model with

## Bayesian analysis of multilevel models II

two levels of unknown parameters might be written as follows:  
 $p(\alpha, \beta | y) \propto L(\alpha; y)p(\alpha | \beta)p(\beta)$ . This specification of the posterior distribution indicates that we believe that  $y$  depends on  $\beta$  only to the extent that  $\beta$  affects the distribution of  $\alpha$  which in turn affects the distribution of  $y$ . The distribution for  $\beta$  involves no unknown parameters. It might, for instance, be a reference prior, such as a normal distribution with mean zero and precision .000001.

## Example: heart transplant mortality

Albert's data set `hearttransplants` contains observations on two variables for 94 hospitals:  $y_i$  denotes the number of deaths within 30 days of heart transplant surgery in hospital  $i$  and  $e_i$  denotes the "exposure" of hospital  $i$ —that is, the number of transplant patients adjusted for individual mortality risk estimated on the basis of information about gender, race, and "medical condition before surgery" (Albert 2009, p. 41).

Assuming that each hospital has a mortality rate that is roughly constant over time, we can consider various approaches to estimating such rates. Two simple possibilities spring to mind: (a) the mortality rate in hospital  $i$  could be estimated by the ratio of deaths to exposure,  $y_i/e_i$ , or (b) under the assumption that all hospitals have the same mortality rate, this common rate could be estimated by  $\sum_{i=1}^{94} y_i / \sum_{i=1}^{94} e_i$ . Both approaches have possible shortcomings.

Approach (a) could give rise to large sampling errors, particularly for the hospitals with small exposures. To assess the seriousness of this problem, we could compare standard deviations in  $y_i/e_i$  for hospitals with low and high exposures, using R code such as the following:

```
> library(LearnBayes)
> data(hearttransplants)
> attach(hearttransplants)
> ## sort data in order of increasing exposure
> ht <- hearttransplants
> htordered <- ht[order(ht[,1]),]
> ## calculate observed mortality rates
> htomr <- htordered[,2]/htordered[,1]
> ## calculate standard deviations in observed
> ## mortality rates in hospitals across hospitals
> ## with various exposure rates.
> sd47low <- sd(htomr[1:47])
> sd47high <- sd(htomr[48:94])
```

```
> sd10low <- sd(htomr[1:10])
> sd10high <- sd(htomr[85:94])
> ## calculate ratios of standard deviations
> sd47low/sd47high
```

```
[1] 1.842144
```

```
> sd10low/sd10high
```

```
[1] 3.55479
```

These results reinforce worries that hospital by hospital estimates of mortality rates may subject to large sampling errors particularly for hospitals with low exposure rates. These worries could motivate consideration of pooling data across hospitals, estimating a supposedly common mortality rate as

$\sum_{i=1}^{94} y_i / \sum_{i=1}^{94} e_i = 277 / 294681 = 0.0009399995$ . This approach is open to the objection that the mortality rates might differ substantially among hospitals. As evidence that such differences exist, Albert notes that the observed mortality rates in 15 of the 94

hospitals lie in the extreme tails of a predictive Poisson distribution fit to the overall mortality rates—so far out in the tails that the p-values are less than .10.

This use of p-values might be questioned by some Bayesians. As noted in an earlier class, Jeffreys (1983, p. 385) argued that departures of observed quantities from predicted ones are poorly expressed by p-values. A p-value “gives the probability of departures, measured in a particular way, equal to *or greater than* the observed set, and the contribution from the actual value is nearly always negligible.” Thus when hypotheses are assessed in terms of p-values “*a hypothesis that may be true may be rejected because it has not predicted observable results that have not occurred.*” This seems a remarkable procedure. On the face of it the fact that such results have not occurred might more reasonably be taken as evidence for the law, not against it.” Similarly, Geisser (1993, p. 107) argues that p-values can be changed by substituting one noninformative stopping rule for another and that use of

p-values to test models thus contravenes the Bayesian likelihood principle.

Whatever we may think of Albert's use of p-values, let us accept his conclusion that mortality rates probably differ across hospitals. In between the arguably untenable extremes of estimating the mortality rate in hospital  $i$  as  $y_i/e_i$  or as  $\sum y_j/\sum e_j$  is a continuum of more or less plausible compromise estimates of the form

$$(1 - \omega) \frac{y_i}{e_i} + \omega \frac{\sum y_j}{\sum e_j}, \quad (1)$$

where  $\omega$  is a parameter lying in the interval  $(0, 1)$  which serves to construct a weighted average of the two extreme estimates. The parameter  $\omega$  is sometimes called a **pooling factor** because it is the weight on the pooled estimate  $\sum y_j/\sum e_j$  (Gelman and Hill 2007, p. 477). Expressions of the form (1) are called **shrinkage estimators** because they shrink the range of estimates from that spanned by  $[\min(y_j/e_j), \max(y_j/e_j)]$  down to a narrower range around the the pooled estimate.

Let us assume that each hospital has its own Poisson distribution for deaths. Thus for hospital  $i$  the sampling distribution for  $y_i$  is a Poisson distribution with parameter  $\lambda_i$ . The conjugate distribution for a Poisson likelihood is a gamma distribution. Using Albert's parametrization, we can write the gamma density for  $\lambda_i$  as

$$g(\lambda_i|\alpha, \mu) = \frac{(\alpha/\mu)^\alpha \lambda_i^{\alpha-1} \exp(-\alpha\lambda_i/\mu)}{\Gamma(\alpha)}, \quad \lambda_i > 0, \quad (2)$$

The prior mean of  $\lambda_i$  is  $\mu$  and its prior variance is  $\mu^2/\alpha$ . Because  $\alpha$  is inversely proportional to the variance, Albert calls it a "precision parameter," meaning a hyperparameter proportional to the precision.

Next Albert assigns hyperpriors to the hyperparameters  $\mu$  and  $\alpha$ :

$$g(\mu) \propto 1/\mu, \quad \mu > 0, \quad (3)$$

$$g(\alpha) \propto \frac{z_0}{(\alpha + z_0)^2}, \quad \alpha > 0. \quad (4)$$

The hyperparameter  $z_0$  is the median of  $\alpha$ . As a basis for his simulations, Albert lets  $z_0$  take the value 0.53.

Each  $\lambda_i$  has a posterior gamma( $y_i + \alpha, e_i + \alpha/\mu$ ) distribution with mean

$$E(\lambda_i|y, \alpha, \mu) = \frac{y_i + \alpha}{e_i + \alpha/\mu} = (1 - \omega_i) \frac{y_i}{e_i} + \omega_i \mu, \quad (5)$$

where  $\omega_i = \alpha/(\alpha + e_i\mu)$ . Equation (5) clearly resembles (1). It indicates that a hospital with a near-zero exposure  $e_i$  (and therefore a near-zero death count  $y_i$ ) will have a posterior expectation  $E(\lambda_i|y, \alpha, \mu)$  close to the overall mean  $\mu$ . This relatively heavy weighting of  $\mu$  for a hospital with scant exposure alleviates the problem of high variance in small samples, a problem that aroused worries about the using  $y_i/e_i$  to estimate  $\lambda_i$ . At the opposite extreme, a hospital with a very large exposure will have  $E(\lambda_i|y, \alpha, \mu)$  close to  $y_i/e_i$ . Thus two hospitals with large exposures but sharply different observed death rates will get substantially different posterior means for their  $\lambda$  parameters. This relatively heavy weighting of  $y_i/e_i$  for hospitals with large exposure

alleviates the problem of imposing artificial homogeneity, a problem that raised questions about the appropriateness of  $(\sum y_j)/(\sum e_j)$  as an estimate for all  $\lambda_j$ .

In accordance with the two levels of unknown parameters, Albert adopts a two-step procedure for sampling from the posterior distribution: First simulate the hyperparameters  $(\mu, \alpha)$  from their marginal posterior distribution, as shown on p. 163. Second simulate  $\lambda_1, \dots, \lambda_{94}$  from their posterior gamma distribution conditioned on the simulated values of the hyperparameters obtained in the previous step.

The hyperparameters  $\alpha$  and  $\mu$  are both positive. To prevent their lower bound from retarding simulation, Albert adopts logarithmic transformations:  $\theta_1 = \log(\alpha)$  and  $\theta_2 = \log(\mu)$ . The inverse fns are of course  $\alpha = \exp(\theta_1)$  and  $\mu = \exp(\theta_2)$ . The marginal posterior density for  $(\theta_1, \theta_2)$  is  $p(\theta_1, \theta_2 | \text{data}) =$

$$K \frac{1}{[\Gamma(\alpha)]^{94}} \prod_{j=1}^{94} \left[ \frac{(\alpha/\mu)^\alpha \Gamma(\alpha + y_j)}{(\alpha/\mu + e_j)^{(\alpha+y_j)}} \right] \frac{z_0 \alpha}{(\alpha + z_0)^2}, \quad (6)$$

where  $K$  is a constant of proportionality.

To define the log posterior of  $\theta_1$  and  $\theta_2$ , Albert provides a function `poissgamexch` whose structure closely matches that of the marginal posterior density (6). This function is one input to `gibbs`, Albert's implementation of the Metropolis within Gibbs algorithm.

Another essential input to `gibbs` is a starting point in the  $(\theta_1, \theta_2)$  plane. To find a starting point near the distribution's mode, Albert uses his `laplace fn`, which itself needs a starting point. To obtain a starting point for  $\theta_1$  we might reason that this hyperparameter is  $\log(\alpha)$  and the prior median of  $\alpha$  has been set at 0.53. Thus we could pick  $\log(.53) = -0.6349$  as the starting value for  $\theta_1$ . To get a starting point for  $\theta_2$  we could interpret this hyperparameter as the logarithm of the mean of the  $\lambda_j$ , which we may expect to be close to  $(\sum y_j)/(\sum e_j) = 0.00094$ . Hence we could pick  $\log(.00094) = -6.9696$  as our starting point for  $\theta_2$ . From this starting point, we can search for the mode of the posterior distribution using the following R code:

```
> datapar = list(data = hearttransplants, z0 = 0.53)
> start=c(-0.6349, -6.9696)
> fit = laplace(poissgamexch, start, datapar)
> fit
```

```
$mode
```

```
[1] 1.887259 -6.955554
```

```
$var
```

```
          [,1]      [,2]
[1,] 0.234737682 -0.003089225
[2,] -0.003089225  0.005859393
```

```
$int
```

```
[1] -2208.501
```

```
$converge
```

```
[1] TRUE
```

The estimated mode shown above is reassuringly similar to that found by Albert (1.883954, -6.955446) from a different starting point.

Next we can use our estimated mode as a starting point for sampling from the posterior distribution. The R code to do that is as follows:

```
> start = c(1.887259, -6.955554)
> fitgibbs = gibbs(poissgamexch, start, 1000,
+   c(1,.15), datapar)
> fitgibbs$accept

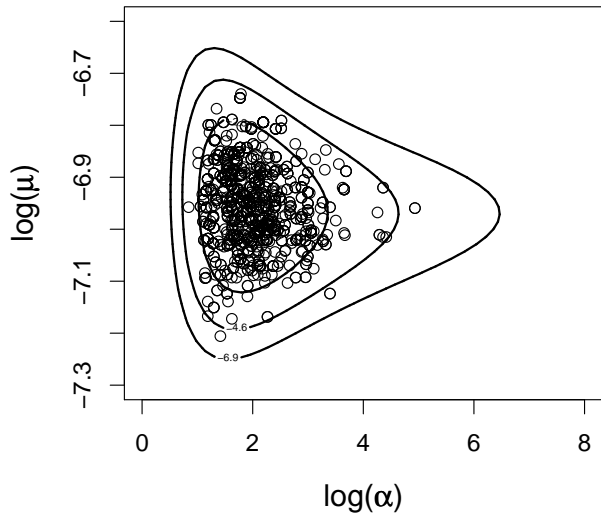
      [,1] [,2]
[1,] 0.335 0.327
```

To get an explanation for the output for `accept`, we can type `?gibbs` and scroll down to the “value” section,” where `accept` is characterized as a “vector of acceptance rates of the Metropolis steps of the algorithm.” If these acceptance rates had been very low and the algorithm very slow, we could have responded by thinking about alternative parametrization.

To visualize the sample of  $(\alpha, \mu)$ , we can plot the same and contour lines using the following R code:

```
> par(mfrow = c(1, 1))
> par(cex=1.3, cex.lab=1.3)
> mycontour(poissgamexch, c(0, 8, -7.3, -6.6), datapar,
+   xlab=expression(paste("log(", alpha, ")")),
+   ylab=expression(paste("log(", mu, ")")))
> points(fitgibbs$par[, 1], fitgibbs$par[, 2])
```

The resulting figure is shown in the next frame.



Now we can sample from the posterior distribution of  $\lambda_1, \dots, \lambda_{94}$  conditional on our simulated values of  $\alpha$  and  $\mu$ . Recalling that each  $\lambda_i$  has a posterior  $\text{gamma}(y_i + \alpha, e_i + \alpha/\mu)$  distribution, we can see that the following R code simulates 1000 values of  $\lambda_1$  for each draw of the hyperparameters.

```
> alpha = exp(fitgibbs$par[, 1])
> mu = exp(fitgibbs$par[, 2])
> lam1 = rgamma(1000, y[1] + alpha, e[1] + alpha/mu)
> mean(lam1)

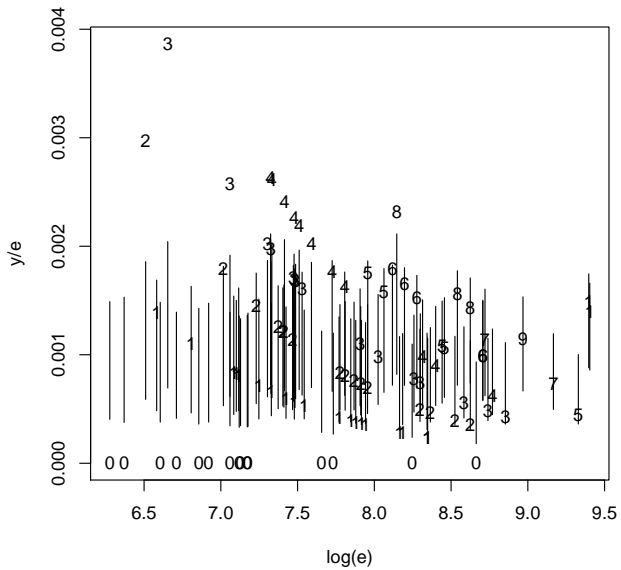
[1] 0.0008785182

> sd(lam1)

[1] 0.0003399624
```

Albert supplies the following code which does something similar for all 94 hospitals and plots posterior 90% credible intervals together with the original data:

```
> alpha = exp(fitgibbs$par[, 1])
> mu = exp(fitgibbs$par[, 2])
> plot(log(e), y/e, pch = as.character(y))
> for (i in 1:94) {
+   lami = rgamma(1000, y[i] + alpha, e[i] + alpha/mu)
+   probint = quantile(lami, c(0.05, 0.95))
+   lines(log(e[i]) * c(1, 1), probint)
+ }
```



## References

- Albert, J. (2009). *Bayesian Computation with R*. Springer, New York, second edition.
- Geisser, S. (1993). *Predictive Inference: An Introduction*. Chapman & Hall, New York.
- Gelman, A. and Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, Cambridge.
- Jeffreys, H. (1983). *Theory of Probability*. Oxford University Press, New York, third edition.