ST 740: Introduction to Hierarchical Models

Alyson Wilson

Department of Statistics
North Carolina State University

October 28, 2012

Hierarchical/Multilevel Models

General Formulation

- Observational Model: $y_i \mid \theta_i \sim f_i(y_i \mid \theta_i)$
- Structural Model: $\theta_i \mid \alpha \sim g(\theta_i \mid \alpha)$
- Hyperparameter model: $\alpha \sim h(\alpha)$

Pump Data

System	s_i (failures)	t_i (thousand hours)
1	5	94.320
2	1	15.720
3	5	62.880
4	14	125.760
5	3	5.240
6	19	31.440
7	1	1.048
8	1	1.048
9	4	2.096
10	22	10.480

Table : Pump failure count data from Farley 1 U.S. nuclear power plant

Sampling Distribution

$$
\begin{aligned}
S_i \mid \lambda_i, t_i &\sim \text{Poisson}(\lambda_i t_i) \\
\lambda_i \mid \alpha, \beta &\sim \text{Gamma}(\alpha, \beta) \\
\alpha, \beta &\sim \pi(\alpha, \beta)
\end{aligned}
$$

Pump Failure Model

Let s_i denote the failures for the ith pump.

$$\begin{aligned} S_i \mid \lambda_i &\sim \text{Poisson}(\lambda_i t_i) \\ \lambda_i \mid \alpha, \beta &\sim \text{Gamma}(\alpha, \beta) \\ \alpha, \beta &\sim \pi(\alpha, \beta) \end{aligned}$$

We'll choose

$$\begin{aligned} \alpha &\sim \text{Gamma}(a_1, a_2) \\ \beta &\sim \text{Gamma}(b_1, b_2) \end{aligned}$$

Posterior Distribution

$$
\begin{aligned}
&\pi(\lambda_1, \ldots, \lambda_{10}, \alpha, \beta \mid \mathbf{s}) \\
&\quad = f(\mathbf{s} \mid \lambda_1, \ldots, \lambda_{10}, \alpha, \beta)\pi(\lambda_1, \ldots, \lambda_{10}, \alpha, \beta) \\
&\quad = f(\mathbf{s} \mid \lambda_1, \ldots, \lambda_{10})\pi(\lambda_1, \ldots, \lambda_{10}, \alpha, \beta) \\
&\quad = f(\mathbf{s} \mid \lambda_1, \ldots, \lambda_{10})\pi(\lambda_1, \ldots, \lambda_{10} \mid \alpha, \beta)\pi(\alpha, \beta) \\
&\quad = f(\mathbf{s} \mid \lambda_1, \ldots, \lambda_{10})\pi(\lambda_1 \mid \alpha, \beta) \cdots \pi(\lambda_{10} \mid \alpha, \beta)\pi(\alpha, \beta) \\
&\quad = f(\mathbf{s} \mid \lambda_1, \ldots, \lambda_{10})\pi(\lambda_1 \mid \alpha, \beta) \cdots \pi(\lambda_{10} \mid \alpha, \beta)\pi(\alpha)\pi(\beta)
\end{aligned}
$$

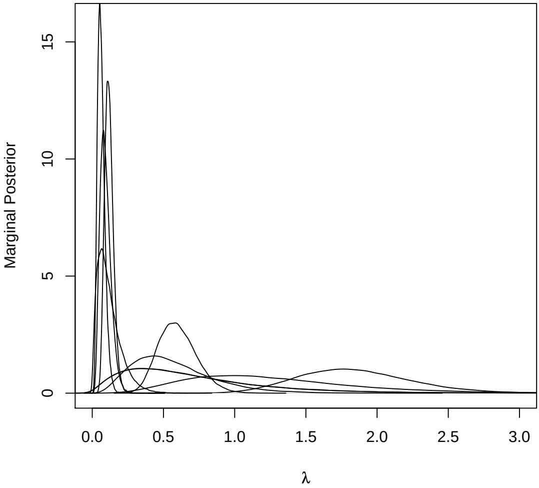
Posterior Distribution

$$
\begin{aligned}
&\pi(\lambda_1, \ldots, \lambda_{10}, \alpha, \beta \,|\, \mathbf{s}, \mathbf{t}) \\
&\quad \propto \left(\prod_{i=1}^{10} (\lambda_i t_i)^{s_i} \exp(-\lambda_i t_i) \right) \left(\prod_{i=1}^{10} \frac{\beta^\alpha}{\Gamma(\alpha)} \exp(-\beta \lambda_i) \lambda_i^{\alpha-1} \right) \pi(\alpha, \beta) \\
&\quad \propto \frac{\beta^{10\alpha}}{\Gamma^{10}(\alpha)} \left(\prod_{i=1}^{10} \exp(-\lambda_i(\beta + t_i)) \lambda_i^{s_i+\alpha-1} \right) \pi(\alpha, \beta) \\
&\quad \propto \frac{\beta^{10\alpha}}{\Gamma^{10}(\alpha)} \left(\prod_{i=1}^{10} \exp(-\lambda_i(\beta + t_i)) \lambda_i^{s_i+\alpha-1} \right) \\
&\qquad\quad \alpha^{a_1-1} \exp(-a_2 \alpha) \beta^{b_1-1} \exp(-b_2 \beta)
\end{aligned}
$$

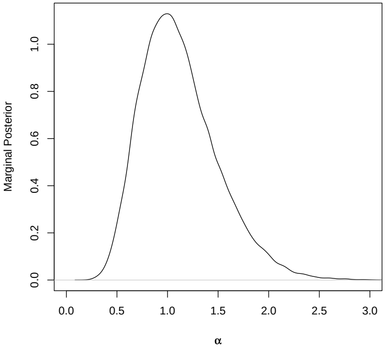
Full Conditional Distributions

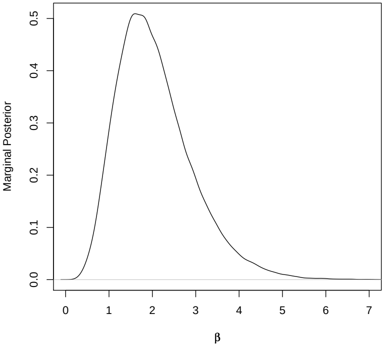
$$
\begin{aligned}
\lambda_i \mid \alpha, \beta &\sim \text{Gamma}(\alpha + s_i, \beta + t_i) \\
\beta \mid \lambda_1, \ldots, \lambda_{10}, \alpha &\sim \text{Gammma}(10\alpha + b_1, b_2 + \sum \lambda_i) \\
\alpha \mid \lambda_1, \ldots, \lambda_{10}, \beta &\sim \frac{\beta^{10\alpha}}{\Gamma^{10}(\alpha)} \left(\prod_{i=1}^{10} \lambda_i \right)^{\alpha} \alpha^{a_1 - 1} \exp(-a_2 \alpha)
\end{aligned}
$$

Posterior Distributions

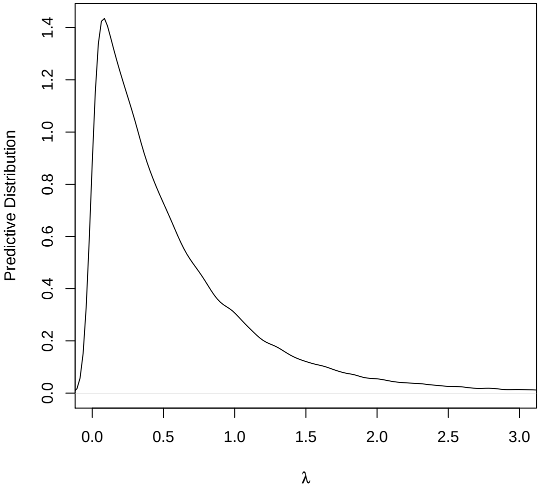


Posterior Distributions





Predictive Distribution for Rate



Another Example

Consider the following study:

- J hospitals are sampled from the population of hospitals in the U.S.
- Within each hospital, n_j patients are selected to participate in a drug trial
- Outcome y_{ij} corresponds to ith patient in jth hospital, and is modeled as $f(y_{ij} \mid \theta_j)$.
- We have two interests: estimating the θ_j and predicting the outcome of the trial at the next hospital we select.

Possible Approahces

1. Conduct J separate analyses and obtain $\pi(\theta_j \mid y_j)$. This assumes that all of the θ_j are independent.
2. Assume that all $\theta_j = \theta$ and get one posterior $\pi(\theta \mid y_1, y_2, ..., y_J)$
3. Model the θ_j hierarchically

Exchangeability

Let $f(y_1, \ldots y_n)$ be the joint density of $Y_1, \ldots, Y_n$. If $f(y_1, \ldots, y_n) = f(y_{\pi_1}, \ldots, y_{\pi_n})$ for all permutations π of $\{1, \ldots, n\}$, then $Y_1, \ldots, Y_n$ are *exchangeable*.

Roughly speaking, $Y_1, \ldots, Y_n$ are exchangeable if the subscript labels convey no information about the outcomes.

Claim: If $\theta \sim p(\theta)$ and $Y_1, \ldots, Y_n$ are conditionally i.i.d. given θ, then marginally (unconditionally on θ), $Y_1, \ldots, Y_n$ are exchangeable.

Exchangeability

Proof: Suppose $Y_1, \ldots, Y_n$ are conditionally i.i.d. given some unknown parameter θ. Then for any permutation π of $\{1, \ldots, n\}$ and any set of values $(y_1, \ldots, y_n)$,

$$\begin{aligned}
p(y_1, \ldots, y_n) &= \int p(y_1, \ldots, y_n \mid \theta) p(\theta) d\theta \\
&= \int \left[\prod_{i=1}^{n} p(y_i \mid \theta) \right] p(\theta) d\theta \\
&= \int \left[\prod_{i=1}^{n} p(y_{\pi_i} \mid \theta) \right] p(\theta) d\theta \\
&= p(y_{\pi_1}, \ldots, y_{\pi_n})
\end{aligned}$$

You can show that this claim is actually "if and only if." The other direction is called *de Finetti's Theorem*.

Exchangeability and Hierarchical/Multilevel Models

- Observational Model: $y_i \mid \theta_i \sim f_i(y_i \mid \theta_i)$
- Structural Model: $\theta_i \mid \alpha \sim g(\theta_i \mid \alpha)$
- Hyperparameter model: $\alpha \sim h(\alpha)$

In many hierarchical models, we're making a double assumption of exchangeability—once in the observational model and once in the structural model.

If no information (other than the data $\mathbf{y}$) is available to distinguish any of the θ_i's from any of the others, and no ordering or grouping of the parameters can be made, then we commonly assume symmetry among the parameters in their prior distribution.

In particular, we represent this symmetry probabilistically by modeling the parameters $(\theta_1, \ldots, \theta_J)$ as *exchangeable*.

Exchangeability and Hierarchical Models

The simplest form of an exchangeable prior distribution has the parameters θ_i as an independent sample from a prior distribution governed by some unknown parameter vector α;

$$\pi(\theta_1, \ldots, \theta_J \mid \alpha) = \prod_{i=1}^{J} \pi(\theta_i \mid \alpha)$$

In particular, the prior distribution for $\theta_1, \ldots, \theta_J$ is

$$\pi(\theta_1, \ldots, \theta_J) = \int \left[\prod_{i=1}^{J} \pi(\theta_i \mid \alpha) \right] \pi(\alpha) d\alpha$$

This mixture model characterizes $(\theta_1, ..., \theta_J)$ as independent draws from a *superpopulation* that is determined by unknown parameters α.

In general, you will see this model specified using all three components of the general formulation of the hierarchical model: observational model, structural model, hyperparameter model.

Basic Two-Level Normal Model

$$
\begin{aligned}
Y_{ij} &= \mu_j + \epsilon_{ij} \\
\epsilon_{ij} &\sim \text{Normal}(0, \sigma^2)
\end{aligned}
$$

Basic Two-Level Normal Model

"Random Effects"

$$\mu_i \mid \mu, \sigma_\alpha^2 \sim \text{Normal}(\mu, \sigma_\alpha^2)$$

We then also specify prior distributions for μ, σ_α^2, and σ^2.

Often μ and σ^2 have nonifnormative priors.

Basic Two-Level Normal Model

Prior distributions for σ_α^2

- There are a number of choices for σ_α^2. Common choices (Gelman (2006)) for $J > 5$ are improper uniform (either on the variance or log variance scale), proper uniform (with large upper bound), proper normal (with large variance).
- For $J < 5$, or in a situation where more prior information is helpful, consider the half-t. A reasonable starting point is the half-Cauchy with reasonably large scale parameter.
- In general, Gelman (2006) does not recommend the inverse-gamma(ϵ,ϵ) family of noninformative prior distributions because the inferences are often sensitive to the choice of ϵ.

Basic Two-Level Normal Model

"Fixed Effects"

$$\begin{aligned} Y_{ij} &= \mu_j + \epsilon_{ij} \\ \epsilon_{ij} &\sim \text{Normal}(0, \sigma^2) \end{aligned}$$

- The μ_j have independent improper (Jeffreys) prior distributions.
- Or $\mu = (\mu_1, \ldots, \mu_J) \sim \text{MVN}(m, V)$
- Or ....

Estimating the Risk of Tumor in a Group of Rats

Data from Gelman et al. (2003). We're insterested in estimating the proportion of tumor in a population of female laboratory rats that receive a zero dose of a particular drug. In our current experiment, we observe that 4/14 develop a tumor. We have previous data on 70 other similar groups of rats.

0/20	0/19	1/18	2/20	3/20	4/20	6/23
0/20	0/18	1/18	1/10	2/13	10/48	5/19
0/20	0/18	2/25	5/49	9/48	4/19	6/22
0/20	0/17	2/24	2/19	10/50	4/19	6/20
0/20	1/20	2/23	5/46	4/20	4/19	6/20
0/20	1/20	2/20	3/27	4/20	5/22	6/20
0/20	1/20	2/20	2/17	4/20	11/46	16/52
0/19	1/20	2/20	7/49	4/20	12/49	15/47
0/19	1/19	2/20	7/47	4/20	5/20	15/46
0/19	1/19	2/20	3/20	4/20	5/20	9/24

How do we analyze this data?

Let's model the data as having a binomial sampling distribution, and let's use the (conjugate) beta prior distribution to describe our knowledge about the probability of developing a tumor.

- We could essentially ignore the 70 data points, choose a prior distribution (maybe informative, maybe non-informative), and construct a posterior distribution based on our current experiment: $4/14$.
- We could use the 70 data points to choose a prior distribution for our current experiment, then calculate the posterior based on our current experiment. For example, the sample mean of the 70 values is 0.136 and the sample standard deviation is 0.103, which suggests a $\text{Beta}(1.4, 8.6)$ prior distribution.
- We could model the data hierarchically.

Stern and Sugano's Noninformative Prior

Given a beta distribution as the hyperprior for the p_i,

$$\frac{\alpha}{\alpha + \beta}$$

is the prior mean and

$$\alpha + \beta$$

is the prior "sample size".

If we assign independent uniform prior distributions to the prior mean and prior "sample size," we have an improper posterior density. However, assigning independent uniform prior distributions to $\frac{\alpha}{\alpha+\beta}$ and $(\alpha + \beta)^{-1/2}$ leads to a proper posterior distribution. This yields

$$\pi(\alpha, \beta) \propto (\alpha + \beta)^{-5/2}$$

Cool Book

Gelman and Hill (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.