Random Effects Models

Yingbo Li

Clemson University

MATH 9810

Hierarchical Models

- Often data are grouped, e.g., students within schools, patients within hospitals.
- Estimation can be done separately in each group, but...
- Arguably the outcomes in the groups are similar, so that when estimating for any one group, there is useful information in the other groups' outcomes
- Hierarchical models (also called multilevel models) provide a principled way to take advantage of that information.
- Typically need to use Gibbs samplers to fit the models, but we'll start off with simpler cases to elucidate the main ideas.

Example: Math Score in US Public Schools

- Math scores of students from the 10th grade.
- 100 large public schools
- 1993 students
- Within each school, a random sample of students is selected (4 to 32)

- Let's develop a hierarchical model for these data.
- In each school j, where $j = 1, \ldots, J$, we test random sample of n_j students.
- For each school j, let μ_j be the school-wide average test score, and let σ_j^2 be the school-wide variance of individual test scores.
- Let $\bar{y}_j$ be the sample average in school j, and s_j^2 be the sample variance in school j.

One-Way AOV Model

In the classical one-way analysis of variance model:

$$y_{ij} = \mu_j + \epsilon_{ij}, \quad \epsilon_{ij} \overset{\text{iid}}{\sim} N(0, \sigma^2)$$

- The μ_j are referred to as fixed effects.
- Interest is in the individual means or differences in means for the J groups. Hypothesis tests:
  - $H_0 : \mu_1 = \mu_2 = \cdots = \mu_{100}$ (distribution of math scores is the same across schools)
  - H_1: distributions do differ in locations

```
> anova(aov(mathscore ~ school));
Analysis of Variance Table

Response: mathscore
            Df Sum Sq Mean Sq F value    Pr(>F)
school      99  48825  493.18   5.834 < 2.2e-16 ***
Residuals 1893 160024   84.53
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Initial Model for Testing Example

- Suppose we believe that the ith student's score in school j, denoted y_{ij}, is iid with
$$y_{ij} \mid \mu_j, \sigma_j^2 \sim N(\mu_j, \sigma_j^2)$$
- Classical inference for each school based on large sample 95% CI: $\bar{y}_j \pm 1.96\sqrt{s_j^2/n_j}$
- Bayesian inference with noninformative prior distribution gives similar intervals (even without normal sampling model when n_j is large)
- We may assume that all σ_j^2 are equal (to σ^2).

Hierarchical Model

- Instead use informative prior distribution for each μ_j based on following: conceive of the schools themselves as a random sample from all possible schools.
- Suppose μ_0 is overall mean of all schools' average scores, and τ^2 is variance of all schools' average scores.
- We can think of each μ_j as coming from a distribution, e.g.,

$$p(\mu_j|\mu, \tau^2) \sim N(\mu, \tau^2)$$

- In this second-level model, the school level means vary about an overall mean μ with variance τ^2.
- View that the 100 schools are exchangeable, and students within a school are exchangeable.

Random Effects

In 2-level model, the school-level means are viewed as random effects arising from a normal population.

$$p(\mu_j|\mu, \tau^2) \sim N(\mu, \tau^2)$$

- μ is the overall population mean, a fixed e?ect
- σ^2 is the within-group variance or variance component
- τ^2 is the between-group variance
- 2 additional parameters versus the $J + 1$ in the fixed effects model.

Mixed Effects Model

We can write $\mu_j = \mu + s_j$ where each school mean is centered at the overall mean μ plus some normal random effect s_j. Substituting this into the distribution for y_{ij}, we arrive at the combined model:

$$y_{ij} = \mu + s_j + \epsilon_{ij}$$

with fixed effect μ and school level random effects s_j and individual random effects ϵ_{ij} , leading to what is known as a mixed effects model.

Marginal Model

Because linear combinations of normals are normally distributed we have the equivalent model:

$$y_{ij} \sim \mathsf{N}(\mu, \sigma^2 + \tau^2)$$

where

$$Cov(Y_{ij}, Y_{i'j}) = \tau^2, \quad Cov(Y_{ij}, Y_{i'j'}) = 0$$

This model that implies students within schools are exchangeable and that student achievements across different schools are independent given the school effect. (reasonable assumption?)

Bayesian Model

- We have

$$
\begin{aligned}
y_{ij}|\mu_j, \sigma^2 &\sim N(\mu_j, \sigma^2) \\
\mu_j|\mu, \tau^2 &\sim N(\mu, \tau^2)
\end{aligned}
$$

- Unknown parameters of interest: $\mu_j, \mu, \sigma^2, \tau^2$.
- Distribution for μ_j is given by the 2nd level model specification.
- Specify joint prior distribution for remaining unknowns μ, σ^2, τ^2.

Informative Prior Distributions

- Can use semi-conjugate prior distributions

$$
\begin{aligned}
\mu &\sim N(\mu_0, \gamma_0^2) \\
1/\sigma^2 &\sim \text{Gamma}(v_0/2, v_0\sigma_0^2/2) \\
1/\tau^2 &\sim \text{Gamma}(\eta_0/2, \eta_0\tau_0^2/2)
\end{aligned}
$$

- μ_0: best guess of average of school average test scores
- γ_0^2: set to accord with plausible ranges of values of μ
- σ_0^2: best guess of variance of individual test scores
- v_0: set based on how tight the prior for σ^2 is around σ_0^2
- τ_0^2: best guess of variance of school average test scores
- η_0: set based on how tight the prior for τ^2 is around τ_0^2

Example of setting $\mu_0, \gamma_0, \tau_0^2, \eta_0$

- The nationwide mean math score is 50. Set $\mu_0 = 50$.
- We are not fully sure that $\mu = \mu_0$. Rather, our prior belief is that $P(40 < \mu < 60) = .95$. Set $\gamma_0^2 = 5^2$.
- The math exam was designed to give a nationwide variance (includes both within-school and between-school variance) of 100. We let $\sigma_0^2 = \tau_0^2 = 100$ the upper bound (larger prior variance means weaker informative prior).
- To get relatively flat prior, let $\nu_0 = \eta_0 = 1$.

Posterior Inference

- Obtain a joint posterior distribution over all unknowns

$$p(\mu_1, \ldots, \mu_J, \mu, \sigma^2, \tau^2 \mid Y) \propto p(Y \mid \mu_1, \ldots, \mu_J, \sigma^2) p(\mu_1, \ldots, \mu_J \mid \mu, \tau^2) p(\mu, \sigma^2, \tau^2)$$

- Gibbs sampling using the full conditionals:

$$p(\mu_j \mid \mu, \sigma^2, \tau^2, Y) \propto p(\mu_j \mid \mu, \tau^2) \prod_{i=1}^{n_j} p(y_{ij} \mid \mu_j, \sigma^2)$$

$$p(\mu \mid \mu_j, \sigma^2, \tau^2, Y) \propto p(\mu) \prod_{j=1}^{J} p(\mu_j \mid \mu, \tau^2)$$

$$p(\tau^2 \mid \mu_j, \mu, \sigma^2, Y) \propto p(\tau^2) \prod_{j=1}^{J} p(\mu_j \mid \mu, \tau^2)$$

$$p(\sigma^2 \mid \mu_j, \mu, \tau^2, Y) \propto p(\sigma^2) \prod_{j=1}^{J} \prod_{i=1}^{n_j} p(y_{ij} \mid \mu_j, \sigma^2)$$

Default Priors

- In the normal (non-hierarchical) model, reference prior for $\phi = 1/\sigma^2$ was $p(\phi) \propto 1/\phi$, which is a limiting case of a Gamma(ϵ, ϵ).
- In hierarchical model, using $p(\phi) \propto 1/\phi$ can be OK.
- Let $\phi_\mu = 1/\tau^2$. In the hierarchical model, using $p(\phi_\mu) \sim 1/\phi_\mu$, i.e., $p(\phi_\mu) \sim$ Gamma(ϵ, ϵ), results in **improper posterior distributions.**
- Problem arises because data cannot refute possibility that $\tau^2 = 0$, which results in infinite integrals
- Note that a set of full conditionals where each is a proper distribution does not imply that their joint distribution is proper.

Legitimate Default Priors

- Some default (improper and proper) priors for hierarchical normal models

$$
\begin{aligned}
p(\mu) &\propto 1 & \qquad \mu &\sim N(0, \delta) \\
p(\phi) &\propto 1/\phi & \qquad \phi &\sim \text{Gamma}(\epsilon, \epsilon) \\
p(\tau^2) &\propto 1/\tau & \qquad \tau &\sim U(0, 1/\epsilon)
\end{aligned}
$$

- $p(\tau^2) \propto 1/\tau$ is equivalent to $p(\tau) \propto 1$; also equivalent to $p(\phi_\mu) \propto \phi_\mu^{-3/2}$.
- This prior is no longer noninformative when there is substantial support in the likelihood for τ^2 near zero
- Results in posteriors for τ^2 that have mass piled up near zero, increasingly so as $\epsilon \to 0$.
- Recommended default: $p(\tau^2) \propto 1/\tau$ is usually reasonable when $J \geq 5$.

Posterior for improper default prior

$$
\begin{aligned}
p(\mu_1, \ldots, \mu_J, \mu, \phi, \phi_\mu | Y) &\propto p(Y | \mu_1, \ldots, \mu_J, \phi, \mu, \phi_\mu) \\
&\times [\prod_j p(\mu_j | \mu, \phi_\mu, \phi)] p(\mu, \phi, \phi_\mu)
\end{aligned}
$$

Using the default priors and plugging in the normal distributions, for each piece we have

$$
p(Y \mid \mu_1, \ldots, \mu_J, \phi, \mu, \phi_\mu) \propto \prod_j \prod_i \sqrt{\phi} \exp\left\{ -\frac{\phi}{2} (y_{ij} - \mu_j)^2 \right\}
$$

$$
p(\mu_j | \mu, \phi_\mu) \propto \sqrt{\phi_\mu} \exp\left\{ -\frac{\phi_\mu}{2} (\mu_j - \mu)^2 \right\}
$$

$$
p(\mu, \phi, \phi_\mu) \propto (1/\phi)(1/\phi_\mu)^{3/2}
$$

Markov Chain Monte Carlo Sampling

Cannot obtain the posterior distributions in closed form; instead create a Gibbs sampler. Need the full conditional distributions of each of the following parameters.

$$\mu_j | \mu, \phi, \phi_\mu, Y \text{ for } j = 1, \ldots, J$$

$$\mu | \mu_1, \ldots, \mu_J, \phi, \phi_\mu, Y$$

$$\phi | \mu, \mu_1, \ldots, \mu_J, \phi_\mu, Y$$

$$\phi_\mu | \mu, \mu_1, \ldots, \mu_J, \phi, Y$$

- Only terms that involve parameter of interest are needed in writing the full conditional

Full conditional for each μ_j

We have $p(\mu_j|Y, \mu, \phi, \phi_\mu) = N(\mu_j^*, v_j^*)$, where

$$\mu_j^* = \frac{(n_j\phi)\bar{y}_j + \phi_\mu \mu}{n_j\phi + \phi_\mu}$$

$$v_j^* = 1/\left(n_j\phi + \phi_\mu\right)$$

Full conditional for μ

Using the "uniform" prior for μ

$$
\begin{aligned}
p(\mu|\mu_1, \ldots \mu_J, \phi, \phi_\mu, Y) &\propto (\phi_\mu)^{n/2} \exp(-\frac{\phi_\mu}{2} \sum_{j=1}^{J} (\mu_j - \mu)^2) \\
&\propto \exp\left(-\frac{J\phi_\mu}{2} (\mu - \sum_{j=1}^{J} \mu_j / J)^2\right)
\end{aligned}
$$

This is kernel of $N\left(\frac{\sum \mu_j}{J}, \frac{1}{\phi_\mu J}\right)$.

Since sampling distribution of each μ_j is normal around μ with precision ϕ_μ, we can think of the J values of μ_j as iid "data" from a normal distribution.

Full conditional for ϕ_μ

Using the “uniform” prior for τ

$$p(\phi_\mu \mid \mu_1, \ldots \mu_J, \mu, \phi, Y) \propto \prod_{j=1}^{J} \sqrt{\phi_\mu} \exp\left\{-\frac{\phi_\mu}{2} (\mu_j - \mu)^2\right\} \phi_\mu^{-3/2}$$
$$= \phi_\mu^{(J-3)/2} \exp\left\{-\frac{\phi_\mu}{2} \sum_{j=1}^{J} (\mu_j - \mu)^2\right\}$$

Full conditional for ϕ_μ is

$$\text{Gamma}\left(\frac{J-1}{2}, \frac{\sum_j (\mu_j - \mu)^2}{2}\right)$$

Full conditional for ϕ

Using the Jeffreys prior for ϕ, we have

$$
\begin{aligned}
& p(\phi \mid \mu_1, \ldots \mu_J, \mu, \phi_\mu, Y) \propto \\
& \prod_{j=1}^{J} \prod_{i=1}^{n_j} \sqrt{\phi} \exp\left\{-\frac{\phi}{2}(y_{ij} - \mu_j)^2\right\} (1/\phi) \\
= \quad & \phi^{n/2-1} \exp\left\{-\frac{\phi}{2} \sum_j \sum_i (y_{ij} - \mu_j)^2\right\}
\end{aligned}
$$

Full conditional for ϕ is

$$
\text{Gamma}\left(\frac{n}{2}, \frac{\sum_{j=1}^{J} \sum_{i=1}^{n_j} (Y_{ij} - \mu_j)^2}{2}\right)
$$